41
2ej.

UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES
CUAUTITLAN

APLICACION, EN EL AREA FARMACEUTICA, DEL ANALISIS DE REGRESION LINEAL.

T E S I S

QUE PARA OBTENER EL TITULO DE:

QUIMICO FARMACEUTICO BIOLOGO

P R E S E N T A :

FERNANDO MARTINEZ ROMERO

ASESOR: M.C. ARMANDO CERVANTES SANDOVAL

CUAUTITLAN IZCALLI, EDO. DE MEX. 1997

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLAN
UNIDAD DE LA ADMINISTRACION ESCOLAR
DEPARTAMENTO DE EXAMENES PROFESIONALES

U. N. A. M.
FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLAN
Departamento de Exámenes Profesionales

ASUNTO: VOTOS APROBATORIOS

DR. JAIME KELLER TORRES
DIRECTOR DE LA FES-CUAUTITLAN
P R E S E N T E .

AT'N: Ing. Rafael Rodriguez Ceballos
Jefe del Departamento de Exámenes
Profesionales de la F.E.S. - C.

Con base en el art. 28 del Reglamento General de Exámenes, nos permitimos comunicar a usted que revisamos la TESIS:

Aplicación, en el Area Farmacéutica,
del Análisis de Regresión Lineal

que presenta el pasante: Fernando Martínez Romero

con número de cuenta: 8738541-6 para obtener el TITULO de:

Químico Farmacéutico Biologo .

Considerando que dicha tesis reune los requisitos necesarios para ser discutida en el EXAMEN PROFESIONAL correspondiente, otorgamos nuestro VOTO APROBATORIO.

A T E N T A M E N T E .
"POR MI RAZA HABLARA EL ESPIRITU"
Cuautitlan Izcalli, Edo. de Méx., a 15 de Abril de 1997

PRESIDENTE	I.Q.I. Guadalupe Sevilla Díaz	
VOCAL	M. en C. Armando Cervantes Sandoval	
SECRETARIO	M. en C. Virginia Lara Sagahón	
PRIMER SUPLENTE	Q.F.B. José A. Garduño Rosas	
SEGUNDO SUPLENTE	Q.F.B. Hector Coss Garduño	

UAE/DEP/VAP/01

A mis padres
Por la confianza que depositaron en mi y
porque este es la conclusión de una etapa
en que ellos siempre estuvieron presentes.

A mis hermanos
Por todo su apoyo y comprensión
en aquellas noches de desvelo.

A Lucy
Por el gran apoyo que recibí de ella
en la finalización de este trabajo.

A mis amigos
Quienes me hicieron pasar momentos
inolvidables durante mis estudios.

GRACIAS

ÍNDICE

	Página
1. RESUMEN	1
2. INTRODUCCIÓN: *Objetivos, hipótesis y material y métodos.*	2
3. INTRODUCCIÓN AL ANÁLISIS DE REGRESIÓN	
3.1. Definición	7
3.2. Importancia	7
3.3. Aplicación del Análisis de Regresión	7
3.4. Técnicas de Análisis de Regresión	10
3.5. Propiedades matemáticas del modelo de regresión lineal	11
4. ANÁLISIS DE REGRESIÓN LINEAL SIMPLE	
4.1. Observación gráfica del problema	13
4.2. Supuestos estadísticos para un modelo de Regresión	
4.2.1. Distribución de probabilidad	14
4.2.2. Independencia	15
4.2.3. Linealidad	15
4.2.4. Homocedasticidad	16
4.2.5. Normalidad de Y	17
4.3. Línea recta de mejor ajuste	
4.3.1. Método de Mínimos Cuadrados	18
4.4. Solución al problema de mejor ajuste	
4.4.1. Solución por mínimos cuadrados	20
4.4.2. Solución del sistema de ecuaciones normales	24
4.5. Medidas de la calidad de una línea recta	
4.5.1. Estimación de la varianza	29
4.6. Relación entre la tabla de análisis de varianza y el análisis de regresión	31

Índice

Página

4.7. Inferencias acerca de la pendiente e intercepto 36
4.7.1. Prueba para pendiente cero 38
4.7.2. Prueba para el intercepto cero 39
4.8. Inferencia acerca de la línea de regresión
4.8.1. Intervalos de predicción 40
4.8.2. Intervalos de confianza 42
4.9. Prueba de falta de ajuste 43

5. COEFICIENTE DE CORRELACIÓN Y DE DETERMINACIÓN
5.1. Coeficiente de Correlación
5.1.1. Definición del coeficiente de correlación 49
5.1.2. Propiedades matemáticas del coeficiente de correlación 51
5.1.3. Interpretación del coeficiente de correlación 52
5.1.4. Prueba de hipótesis para el coeficiente de correlación 53
5.1.5. Intervalos de confianza para el coeficiente de correlación 55
5.2. Coeficiente de Determinación
5.2.1. Definición del coeficiente de determinación 57
5.2.2. Interpretación del coeficiente de determinación 58
5.2.3. Relación con el coeficiente de correlación 59
5.2.4. Conceptos erróneos del coeficiente de determinación 60

6. ESTUDIO DE CASO DE REGRESIÓN LINEAL SIMPLE
6.1. Problema 63
6.2. Resultados experimentales 65
6.3. Diagramas de dispersión 69
6.4. Cálculo de coeficientes de regresión 73
6.5. Uso de SAS para ajustar un modelo lineal simple 74
6.6. Prueba de la pendiente e intercepto 83
6.7. Intervalos de confianza y de predicción 88
6.8. Coeficiente de correlación y determinación 93
6.9. Prueba de falta de ajuste 97

Índice

Página

7. ANÁLISIS DE REGRESIÓN MÚLTIPLE

7.1. Definición 101

7.2. Observación gráfica del problema 102

7.3. Suposiciones del análisis de regresión múltiple 104

7.4. Determinación de la ecuación de regresión múltiple

7.4.1. Método de Mínimos Cuadrados 107

7.5. Tabla de análisis de varianza. Prueba de F Total y de F Parcial

7.5.1. Prueba de significancia para la regresión total 110

7.5.2. Prueba de F parcial 111

7.6. Correlaciones Múltiple, Parcial y Múltiple-parcial

7.6.1. Matriz de correlación 116

7.6.2. Coeficiente de correlación múltiple 117

7.6.3. Coeficiente de correlación parcial 120

7.6.4. Coeficiente de correlación Múltiple-parcial 126

7.7. Concepto de interacción 127

7.8. Métodos de selección de variables 131

7.8.1. Procedimiento de todas las posibles regresiones 132

7.8.2. Procedimiento de Eliminación (Backward) 132

7.8.3. Procedimiento de Selección Forward 133

7.8.4. Procedimiento de Regresión Stepwise 134

8. ESTUDIOS DE CASO DE REGRESIÓN MÚLTIPLE

8.1. Estudio de caso No. 1. 136

8.1.1. Problema 137

8.1.2. Resultados experimentales 139

8.1.3. Ajuste del Modelo de Regresión Múltiple 140

8.1.3.1. Matriz de correlación 141

8.1.3.2. Procedimiento Backward 143

8.1.3.3. Procedimiento Forward 149

8.1.3.4. Procedimiento Stepwise 151

8.2. Estudio de caso No. 2. 156
8.2.1. Problema 156
8.2.2. Resultados experimentales y ajuste del modelo 156
8.3. Estudio de caso No. 3. 164
8.3.1. Problema 164
8.3.2. Resultados experimentales y ajuste del modelo 164
8.4. Estudio de caso No. 4. 171
8.4.1. Problema 171
8.4.2. Resultados experimentales y ajuste del modelo 171
9. DISCUSIÓN 178
10. CONCLUSIONES 192
11. BIBLIOGRAFÍA 194
12. ANEXOS

ÍNDICE DE FIGURAS

		Página
3.1.	Propiedades de una recta	12
4.1.	Diagrama de dispersión	13
4.2	Distribución de probabilidad de Y	14
4.3.	Suposición de linealidad	16
4.4.	Componente del error	17
4.5.	Desviación de los puntos observados sobre la línea de regresión	18
4.6.	Varianza de la línea de regresión	31
4.7.	Variación explicada y no explicada por la regresión	33
4.8.	Intervalos de predicción y de confianza	41
4.9.	Diagrama de flujo para probar falta de ajuste	46
5.1.	Coeficiente de correlación como una medida de asociación	53
5.2.	Predicción de Y usando X y sin usar X	58
5.3.	Coeficiente de determinación (no mide la pendiente)	61
5.4.	Coeficiente de correlación no mide la conveniencia	62
6.1.	Cinética de disolución de orden cero	64
6.2.	Varianza de *Disolu* en cada nivel de *Min*	72
6.3.	Varianza de *LDIS* en cada nivel de *Min*	72
6.4.	Región crítica para la prueba de F	82
6.5.	Intervalos de Confianza y de Predicción para los modelos del caso de estudio de regresión lineal simple	91
6.6.	Gráfica de residuales para tres modelos	93
7.1.	Gráfica de dispersión para una sola variable independiente	103
7.2.	Plano tridimensional	104
7.3.	"No interacción" contra "Interacción"	128
7.4.	Ilustración gráfica de "no interacción"	129
7.5.	Métodos de Selección de Variables	135
8.1.	Diagrama de dispersión de *FR-predichos* vs *FR-observados*	153
8.2.	Residuales para los valores predichos	153

8.3. Gráfica de contorno para la friabilidad de gránulos 154
8.4. Superficie de respuesta para la friabilidad de gránulos 155
8.5. Diagrama de dispersión de valores predichos vs valores observados para el estudio de caso No. 2. 163
8.6. Residuales para los valores predichos de la viscosidad en la suspensión de rifampicina 163
8.7. Diagrama de dispersión de valores predichos vs valores observados para el estudio de caso No. 3. 170
8.8. Residuales para los valores predichos de la friabilidad de tabletas de compresión directa 170
8.9. Diagrama de dispersión de valores predichos vs valores observados para el estudio de caso No. 4. 177
8.10. Residuales para los valores predichos del estudio de caso No. 4. 177
9.1. Ubicación del Análisis de Regresión en una Planeación Experimental 180
9.2. Modelos del análisis de regresión 182
9.3. Análisis de regresión lineal simple 185
9.4. Análisis de regresión múltiple 189

ÍNDICE DE CUADROS, PROGRAMAS Y SALIDAS DE SAS

	Página
Cuadro 4.1. Tabla de ANOVA para regresión lineal	32
Cuadro 4.2. Tabla de ANOVA para regresión lineal (menos común)	35
Cuadro 4.3. Tabla de ANOVA incluyendo la prueba de Falta de Ajuste	47
Programa 6.1. Programa SAS para imprimir datos en pantalla	66
Salida 6.1. Salida SAS (Proc Print) del lote 8 de tabletas de Furosemida	67
Programa 6.2. Programación de SAS para obtener diagramas de dispersión	69
Programa 6.3. Programación de SAS para obtener estadísticos básicos	70
Salida 6.2. Diagramas de dispersión	71
Salida 6.3. Anexo A "Resultados del Proc univariate plot"	199
Programa 6.4. Programación de SAS para obtener modelos de regresión lineal simple	75
Salida 6.4. Análisis de regresión lineal simple para tres modelos diferentes	76
Cuadro 6.1. Valores de *t* para pruebas de hipótesis	84
Programa 6.5. Programación de SAS para obtener modelos de regresión lineal simple sin intercepto	85
Salida 6.5. Reajuste de modelos sin intercepto	87
Programa 6.6. Programación de SAS para obtener modelo de regresión lineal simple con sus correspondientes intervalos de confianza, de predicción y residuales.	90
Salida 6.6. Anexo B Intervalos de confianza y de predicción	215
Cuadro 6.2. Coeficientes de correlación y determinación para los 3 modelos propuestos	96
Cuadro 6.3. Datos para realizar la prueba de falta de ajuste (modelo 1)	97
Cuadro 6.4. Datos para realizar la prueba de falta de ajuste (modelo 3)	99
Cuadro 7.1. Tabla de análisis de varianza para regresión múltiple	109
Cuadro 7.2. Tabla de ANOVA para prueba de *F*-parcial	112
Cuadro 8.1. Niveles de las variables independientes del caso de estudio	137
Cuadro 8.2. Matriz experimental del caso de estudio No. 1.	138
Cuadro 8.3. Resultados del porcentaje de friabilidad	140
Programa 8.1. Programa para obtener modelos de Análisis de Regresión Múltiple a través de diversas técnicas	141
Salida 8.1. Matriz de correlación entre todas las variables en estudio	142

Índice

Página

Salida 8.2. Modelo de regresión múltiple por el procedimiento de eliminación 145
Salida 8.3. Modelo de regresión múltiple por el procedimiento Forward 150
Salida 8.4. Modelo de regresión múltiple por el procedimiento Stepwise 152
Salida 8.5. Modelo ajustado para el caso de estudio No. 1. 153
Programa 8.2. Resultados experimentales del estudio de caso No. 2 y ajuste del modelo de regresión múltiple a través del método Stepwise. 157
Salida 8.6. Matriz de correlación entre la viscosidad de la suspención de Rifampicina y las variables en estudio 158
Salida 8.7. Modelo de regresión múltiple a través de Stepwise para la viscosidad de una suspensión de rifampicina. 160
Salida 8.8. Modelo ajustado para la viscosidad. 163
Programa 8.3. Resultados experimentales del estudio de caso No. 3 y ajuste del modelo de regresión múltiple a través del método Stepwise.
Salida 8.9. Matriz de correlación entre la friabilidad de tabletas por compresión directa y las variables en estudio. 166
Salida 8.10. Modelo de regresión múltiple a través de Stepwise para la friabilidad de tabletas elaboradas por compresión directa 168
Salida 8.11. Modelo ajustado para la friabilidad de tabletas de compresión directa. 170
Programa 8.4. Resultados experimentales y ajuste del modelo de regresión múltiple a través del método Stepwise para el estudio de caso No. 4. 172
Salida 8.12. Matriz de correlación entre la dureza de tabletas y variables del estudio de caso No. 4 173
Salida 8.13. Modelo de regresión múltiple a través de Stepwise para la dureza de tabletas del ejemplo No. 4 175
Salida 8.14. Modelo ajustado para la dureza de tabletas del ejemplo No. 4. 177

I. RESUMEN

Este trabajo reúne los aspectos teóricos y prácticos del análisis de regresión, tanto lineal simple como múltiple, enfocados a la investigación farmacéutica. Se comienza con la descripción general del análisis de regresión, mostrando su concepto, importancia y aplicación dentro de la investigación farmacéutica, ya que en esta área se ha utilizado ampliamente para describir la influencia de uno o más factores sobre una respuesta en particular o sobre todo un proceso de fabricación de alguna forma farmacéutica.

Se describen los fundamentos teóricos del análisis de regresión lineal simple y posteriormente se muestra un estudio de caso del ámbito farmacéutico, donde se aplica esta técnica estadística, en el cual se trata de encontrar la mejor relación funcional entre la cantidad de principio activo disuelto de tabletas de furosemida en función del tiempo. Primero se realiza un análisis de regresión sin transformar las variables en estudio, donde se encuentra que la relación lineal no es muy buena, por lo que fue necesario hacer una transformación de las variables, obteniéndose así un modelo que predice la respuesta con menos error.

Posteriormente se describen los aspectos teóricos del análisis de regresión múltiple así como las técnicas de selección de variables que más se emplean, las cuales son: método de selección Backward (de eliminación), método de selección Forward y método de selección Stepwise (paso a paso). La aplicación de estas técnicas se muestra a través de un caso de estudio del área farmacéutica en donde se trata de encontrar la relación funcional entre la friabilidad de gránulos elaborados por fluidización y diversas variables independientes como: Temperatura del aire de entrada, Presión del aire de atomización y Cantidad de solución aglutinante. Sobre este caso se aplican las diversas técnicas de selección de variables y se llega a un modelo de regresión múltiple donde se observa que no todas las variables involucradas desde un principio tienen influencia sobre la variable de respuesta.

En los estudios de caso se aplican herramientas estadísticas y gráficas con ayuda del paquete estadístico SAS, se muestran tanto los programas como las salidas y la interpretación de los resultados, obteniéndose con esto una optimización del tiempo de análisis que permita al investigador ocupar su tiempo en hacer investigación farmacéutica sin llevarse mucho tiempo en los cálculos y análisis.

En este trabajo se logra reunir en un solo texto una guía práctica para el investigador que emplea el análisis de regresión.

II. INTRODUCCIÓN

La estadística es una herramienta para los investigadores de diferentes áreas, la cual ayuda a presentar resultados de forma resumida, ya sea en cuadros o gráficos, facilitando su interpretación. Cuando los resultados de los experimentos conducen a la proposición de modelos teóricos, una de las herramientas estadísticas que más se emplea es el Análisis de Regresión, el cual permite evaluar la relación de una o más variables independientes y al menos una sola variable dependiente continua, es por ello que en la investigación farmacéutica se han realizado numerosos trabajos apoyados en el análisis de regresión, por ejemplo, evaluar la influencia de algunos excipientes de una formulación farmacéutica sobre los aspectos de liberación y características técnicas de la forma farmacéutica final, caracterizar un proceso de fabricación de tabletas en función de las variables que lo afectan, como el tiempo de granulación, concentración del agente aglutinante, humedad, fuerza de compresión y tiempo de mezclado, entre otras. Existen muchos otros casos donde se aplica el análisis de regresión, debido a que los resultados de experimentos del área farmacéutica pueden conducir a la proposición de modelos teóricos, lo cual es un requisito para poder aplicar esta técnica estadística.

Las aseveraciones estadísticas que pueden hacerse en base al análisis de regresión u otro análisis estadístico deben distinguirse de aseveraciones determinísticas. Las aseveraciones estadísticas contemplan la posibilidad de error en la descripción de la relación, a través del uso de la probabilidad y la teoría estadística toman en cuenta las irregularidades del mundo real que son asociadas con el error de medición y la variabilidad individual, a diferencia de las aseveraciones determinísticas las cuales no contemplan la posibilidad de error.

Por otro lado, en muchos de los experimentos publicados, donde se aplica el análisis de regresión, muchas veces no se consideran todos los parámetros importantes relacionados con el análisis de regresión, tales como el cuadrado medio del error, el coeficiente de correlación y de determinación, el coeficiente de variación del modelo y el valor del estadístico F, así como también se omiten los criterios para realizar las pruebas estadísticas que conducen a un modelo que ajuste de manera adecuada los resultados experimentales, para llegar a conclusiones más confiables. Por ello, en este trabajo se destaca la importancia de realizar un análisis estadístico más profundo teniendo como base los fundamentos teóricos del análisis de regresión, los cuales permiten tener más elementos o criterios de decisión para interpretar los resultados del software estadístico SAS que se empleó como apoyo para facilitar los cálculos matemáticos.

En base a esto se plantearon los siguientes objetivos.

OBJETIVO GENERAL.

Revisar los fundamentos teóricos del Análisis de Regresión, tanto lineal simple como múltiple, mostrar su aspecto práctico a través de estudios de caso como una herramienta al investigador farmacéutico y así obtener una ayuda teórico práctica de la aplicación de esta técnica estadística en un conjunto de datos.

OBJETIVOS PARTICULARES.

1. Revisar los fundamentos teóricos del Análisis de Regresión Lineal Simple.
2. Mostrar la aplicación, en el área farmacéutica, del análisis de regresión lineal simple a través de un estudio de caso.
3. Revisar los fundamentos teóricos del Análisis de Regresión Múltiple.
4. Revisar las técnicas de selección de variables del Análisis de Regresión Múltiple.
5. Presentar un enfoque práctico del Análisis de Regresión Múltiple y las diversas técnicas de selección de variables a través de estudios de caso del área farmacéutica.
6. Mostrar la importancia del cumplimento de los supuestos en el análisis de regresión.
7. Interpretar los resultados que se obtienen de las herramientas computacionales al aplicar un Análisis de Regresión, como es el caso del paquete estadístico SAS para Windows.

Al cumplir los objetivos antes mencionados, se puede comprobar que:

La revisión y comprensión de los fundamentos teóricos del Análisis Estadístico de Regresión llevan a la selección más adecuada de análisis y a la interpretación de resultados en forma más rápida y confiable..

MATERIAL Y MÉTODOS.

Para lograr los objetivos, se revisó y conjuntó en forma didáctica y resumida los fundamentos teóricos del análisis de regresión de las siguientes fuentes:

- Bohidar N. R., (1993), *"Relative efficiency of R^2 and β^2 in regression analysis for calibration and formulation"*. **Drug Development and Industry Pharmacy, (19:12, 1447-59).**
- Box George E. P., Hunter William G., Hunter Stuart, (1988), *"Estadistica para investigadores"*. España, Ed. Reverté. 653 pp.
- Cochran William G., (1974), *"Análisis de Regresión múltiple"*. México, Ed. Continental. 703 pp.
- Daniel Wayne W., (1990). *"Bioestadistica: Base para el análisis de las ciencias de la salud"*. 4a reimpresión, México, Ed. Limusa. 667 pp.
- Dick, F. Christopher., Kassen, A. Roger and Amidon, E. Gregory., (1987), *"Determination of the sensitivity of a tablet formulation to variations in excipient levels and processing conditions using optimization techniques"*. **International Journal of Pharmaceutics., (38, 23-31).**
- Draper N. R., Smith H., (1981), *"Applied Regression Analysis"*. 2ª edition. U.S.A., Ed. Jhon Wiley & Sons. 709 pp.
- Elkheshen, A. Seham, Badawi, S. Sabry and Badawi, A. Alia., (1996), *"Optimization of a reconstitutable suspension of Rifampicina using 2^4 factorial desing"*. **Drug Development and Industry Pharmacy, (22:7, 623-630).**
- Harris, M. R.., Schwartz, J. B. and McGinity, (1985), *"Optimization of a slow-release tablet formulation containing sodium sulfathiazole and a montmorillonite clay"*. **Drug Development and Industry Pharmacy, (11:5, 1089-1110).**
- Hamilton Lawrence C., (1990), *"Modern data analysis. A first course in applied statistics"*. U.S.A., Ed. New hamsphire Reooks/ Cole Publishing Company Pacific Grove California. 684 pp.
- Harnett L. Donald, Murphy L. James, (1987), *"Introducción al análisis estadistico"*. México, Ed. Addison-Wesley Iberoamericana. p 486, 525.
- Kleinbaum G. David; Kupper Lawrence L., (1978), *"Applied Regression Analysis and Other Multivariable Methods"*. U.S.A., Ed. Duxbury press. 556 pp.
- Kreyzing Erwin, (1981), *"Introducción a la estadistica matemática. Principios y métodos"*. 6a. edición. México, Ed. Limusa. 505 pp.
- Merkku Pasi, Antikainen Osmo and Yliruusi Jouko, (1993), *"Use of 3^3 Factorial Desing and*

Multilinear Stepwise Regression Analysis in Studying the Fluidized Bed Granulation Process. Part II". **Eur. J. Pharm. Biopharm.** (39:3, 112-116).

- Merkku Pasi, Yliruusi Jouko, (1993), *"Use of 3^3 Factorial Desing and Multilinear Stepwise Regression Analysis in Studying the Fluidized Bed Granulation Process, Part I"*. **Eur. J. Pharm. Biopharm.** (39:2, 75-81).
- Montgomery Douglas C., [Trad. Jaime Saldivar Delgado], (1991), *"Diseño y Análisis de Experimentos"*.México, Ed. Iberoamérica. 589 pp.
- Montgomery Douglas C., Peck Elizabeth A., (1992), *"Introduction to linear regression analysis"*. 2nd ed. U.S.A., Ed. John Wiley & Sons. 527 pp.
- Myers Raymond H., (1989), *"Classical and Modern Regression with applications"*. 2a. edition. U.S.A., 488 pp.
- SAS Institute Inc., SAS/LAB® (1993) *"Software: Graphics editor", version 6, First edition*, cary, NC: SAS Institute. pp.
- SAS Institute Inc., SAS/STAT® (1989) *"Software: User's guide", version 6, First edition, vol. 1* cary, NC: SAS Institute. 943 pp.
- Seber G. A. F., (1977), *"Linear Regression Analysis"*. U.S.A., Ed. John Wiley & Sons. 465 pp.
- Stetsko G., (1986), *"Statistical experimental design and its application to pharmaceutical development problems"*. **Drug Development and Industry Pharmacy**, (12, 1109-23).

Así, toda la información se presenta en varios capítulos, tanto de aspectos teóricos como de aspectos prácticos del análisis de regresión, como se describe a continuación:

El capitulo 3 presenta una introducción al análisis de regresión lineal, donde se define; se menciona su importancia y se enlistan algunos ejemplos de la aplicación de esta técnica en el área farmacéutica.

En el capítulo 4 se encuentran los fundamentos teóricos del Análisis de Regresión Lineal Simple, se describen los diagramas de dispersión, los supuestos para un modelo de regresión, el cálculo de la línea recta a través del método de mínimos cuadrados, inferencias estadísticas y la prueba de falta de ajuste.

En el capítulo 5 se describen los coeficientes de correlación y de determinación, los cuales se presentan en un capítulo aparte debido a que estos coeficientes son de interés tanto para regresión lineal simple como para regresión múltiple.

En el capítulo 6 se analiza un estudio de caso donde se describe la aplicación práctica del análisis de regresión lineal simple. Este caso trata sobre un estudio de disolución, donde se busca el mejor modelo que relacione el porcentaje de principio activo (Furosemida) disuelto de una tableta en función del tiempo.

En el capítulo 7 se presentan los fundamentos teóricos del Análisis de Regresión Múltiple, se describen las pruebas de F-total y F-parcial, las correlaciones Múltiple, Parcial y Múltiple-parcial así como los diversos métodos de Selección de Variables.

En el capítulo 8 se aplican los diversos métodos de selección de variables del análisis de regresión múltiple a un estudio de caso, donde se trata de encontrar el mejor modelo que relacione la friabilidad de gránulos elaborados por fluidización y diversas variables independientes, las cuales fueron: Temperatura del aire de entrada, Presión del aire de atomización y Cantidad de solución aglutinante. También se presentan otros tres ejemplos para mostrar el ajuste de modelos de regresión múltiple a través del método de selección Stepwise.

En los estudios de caso se aplican herramientas estadísticas y gráficas con ayuda del software de análisis estadístico SAS, por lo que se muestra la interpretación y conclusiones de cada salida del programa.

Finalmente se presentan la discusión, conclusiones y bibliografía del trabajo.

III. INTRODUCCIÓN AL ANÁLISIS DE REGRESIÓN

En este capítulo se mencionan las generalidades del análisis de regresión: su definición, importancia y aplicación en el área farmacéutica, así como también se describen las diferentes técnicas de análisis del regresión.

3.1.- DEFINICIÓN.

El análisis de regresión es una herramienta estadística que ayuda a evaluar la relación de una o más variables independientes y al menos una sola variable dependiente continua, es decir, proporciona una descripción de como se relacionan las variables en estudio.

3.2.- IMPORTANCIA.

Cuando se sospecha que las variables "*X's*" son una causa del comportamiento de la variable de respuesta "*Y*", se puede preguntar lo siguiente:

- ¿Si alguna *X* cambia, *Y* cambia?, ¿En cuánto?
- ¿ El cambio de *Y* es el mismo a niveles altos o bajos de las *X*'s?
- ¿ Qué tan exactamente se pueden predecir valores de *Y*, si se conocen valores de *X*?

El análisis de regresión responde a todas estas preguntas directamente, por lo tanto, proporciona una forma de relacionar las variables en estudio para: predecir, optimizar o controlar un experimento o proceso, en cualquier área de experimentación.

3.3.- APLICACIONES DEL ANÁLISIS DE REGRESIÓN.

En la investigación farmacéutica existen diversas posibilidades para aplicar un análisis de regresión ya sea simple o múltiple, entre estas se encuentran los siguientes casos:

1) Cuando se quiere caracterizar la relación entre la variable dependiente e independiente, para determinar la extensión, dirección y fuerza de asociación entre las variables, por ejemplo:

- Determinar el efecto de la solución aglutinante, tanto cantidad como composición, sobre el tamaño promedio de partícula y la distribución del tamaño del gránulo empleado en la elaboración de formas farmacéuticas sólidas [32].

- Caracterizar la turbidez y el punto de turbidez máximo de una solución surfactante no-iónica de administración oral en función de algunos componentes de la formulación como el alcohol USP, polisorbato 80, propilenglicol y sacarosa [25].

2) Para generar el mejor modelo matemático interpretativo, que describa la relación entre una variable dependiente como una función de una o más variables independientes. Por ejemplo:

- Describir un proceso por medio de una ecuación simple que relacione la respuesta Y para cada valor fijo de X, tal como en la predicción de la estabilidad de medicamentos, en la cual existe una relación entre la concentración del fármaco y el tiempo.

- Caracterizar un proceso de elaboración de tabletas en función de las variables del proceso como son: Tiempo de granulación, concentración del agente aglutinante, humedad después del secado, cantidad y tiempo de lubricación. Como variables de respuesta se pueden evaluar la distribución del tamaño de partícula de los gránulos, friabilidad tanto en porcentaje de peso perdido como en porcentaje de tabletas rotas después de la prueba, tiempo de desintegración, apariencia de las tabletas y características de compresión de la mezcla [6].

3) Cuando se necesita describir cuantitativa o cualitativamente la relación entre las variables independientes y la variable dependiente. Por ejemplo:

- Para describir la relación entre variables donde la relación funcional se conoce que es lineal, tal como sucede con la ley de Beer, donde se representa en una gráfica la densidad óptica en función de la concentración del fármaco. Esto se emplea principalmente en la elaboración de curvas de calibración, en el desarrollo y validación de métodos analíticos.

- Cuando se desea caracterizar la estabilidad química de un fármaco como el Diazepan en fase líquida, esto es importante debido a que en la preparación de formas farmacéuticas la apropiada estabilidad del principio activo es fundamental. De acuerdo a la estructura química del fármaco, muchos factores influyen la estabilidad del principio activo los cuales pueden ser cantidades y calidades de excipientes e incluso los empaques del producto farmacéutico. Se ha realizado un estudio para caracterizar la estabilidad del Diazepan en función de la composición del solvente, la temperatura y pH; la variable de respuesta fue la constante de velocidad de descomposición [8].

4) Para determinar cuales variables independientes son importantes en la descripción ó predicción del comportamiento de una variable dependiente, o para clasificarlas en orden de importancia. Por ejemplo:

- Cuando se investiga la influencia de algunos excipientes de la formulación de una matriz de liberación prolongada como pueden ser: celulosa microcristalina, fosfato dicálcico y lactosa, sobre la velocidad de liberación y caracteristicas técnicas del producto farmacéutico tales como dureza, friabilidad y variación de peso [34].

- Cuando se quiere optimizar la friabilidad de una tableta, tomando como variables de estudio la perdida de peso durante el secado del granulado, la distribución del tamaño de partícula, la concentración del lubricante, la fuerza de compresión y precompresión [19].

5) Cuando se desean comparar diversas relaciones de regresión, para seleccionar aquella que mejor describa el comportamiento del fenómeno en estudio.

Sin embargo, es importante ser cauteloso ácerca de los resultados de un análisis de regresión debido a que si se encuentra una fuerte relación entre las variables en estudio no necesariamente implica que las variables independientes sean la causa de la variable dependiente. Aunque la causalidad no puede ser necesariamente inferida de un análisis de regresión, una interpretación significativa de la relación entre variables puede ser descrita en un sentido estadístico.

3.4.- TÉCNICAS DE ANÁLISIS DE REGRESIÓN.

En general, si se supone que hay una sola variable dependiente (Y) o de respuesta; que depende de k variables independientes o de regresión $(X_1, X_2, \ldots, X_k)$, la regresión entre estas variables se caracteriza por un modelo matemático conocido como ecuación de regresión, el cual se ajusta a un conjunto de datos muestrales; en la mayoría de los casos la verdadera relación funcional se desconoce y el investigador debe elegir una función apropiada. Sin embargo, en un experimento diseñado, el análisis de varianza ayuda a determinar que factores son importantes, usando la regresión para construir un modelo cuantitativo que relacione los factores importantes con la respuesta.

Dentro del análisis de regresión existen diversas técnicas cuya aplicación depende de la cantidad de factores o variables que se desean analizar, las cuales se pueden clasificar en:

a) Análisis de regresión lineal:

- Análisis de regresión lineal simple, donde se estudia la relación que existe entre una sola variable independiente y una variable de respuesta.

- Análisis de regresión lineal múltiple, en el cual las variables en estudio son dos o más variables independientes y al menos una variable dependiente (univariado) o más de una variable dependiente (multivariado).

b) Técnicas de análisis de regresión que no necesariamente son lineales, es decir, los datos pueden ajustarse a modelos cuadráticos (simples, múltiples univariados o múltiples multivariados), exponenciales, logarítmicos o cúbicos. Sin embargo, generalmente se busca un modelo de regresión lineal a través de una transformación de las variables, debido a que su manejo e interpretación es más fácil.

3.5.- PROPIEDADES MATEMÁTICAS DEL MODELO DE REGRESIÓN LINEAL.

El modelo de análisis de regresión lineal más sencillo, es decir con una sola variable independiente, se representa por la ecuación de una línea recta, la cual se puede describir matemáticamente de la forma $Y = \beta_0 + \beta_1 X$, donde Y es la variable dependiente, X la variable independiente; el parámetro β_0 es el valor del intercepto y β_1 es el valor de la pendiente. Los símbolos β_0 y β_1 tienen valores constantes para una determinada línea por lo que no se consideran como variables.

El intercepto, β_0, es el valor de Y cuando X=0. La pendiente, β_1, es la cantidad de cambio en Y por cada unidad de cambio en X. Para una línea recta esta relación de cambio es siempre constante. Las propiedades de la línea recta se representan en la figura 3.1., en la cual se puede observar que para la ecuación Y=5-2X, Y decrece cuando X se incrementa, tal línea se dice que tiene una pendiente negativa $(\beta_1 = -2)$, cuyo intercepto es igual a 5; similarmente, la línea Y=-4+X se dice que tiene una pendiente positiva porque Y aumenta cuando X se incrementa, $(\beta_1 = 1)$ y su intercepto es igual a -4.

Para describir las propiedades matemáticas de la ecuación lineal, las variables X e Y se tratan en un sentido matemático, más que en un contexto estadístico.

Para obtener la gráfica de una recta, mínimo se necesitan dos puntos, uno de los cuales puede ser el punto correspondiente al intercepto, mientras que el otro se puede determinar al seleccionar arbitrariamente un valor de X y encontrar el correspondiente valor de Y, o se puede emplear el punto $(\overline{X}, \overline{Y})$, ya que la línea siempre pasa a través de este punto. Sin embargo, si sólo existieran únicamente dos puntos, el análisis de regresión es innecesario ya que en ese caso siempre se obtiene una línea recta.

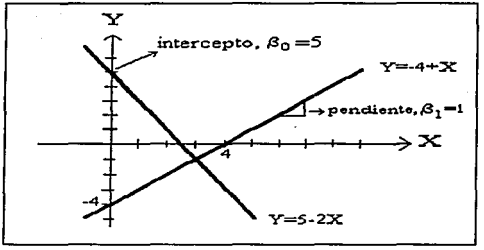


Figura 3.1. Propiedades de una recta.

La forma más simple del problema general de regresión parte con una sola variable dependiente "*Y*" y una variable independiente "*X*", donde al modelo que se encuentra se le llama Regresión Lineal Simple, el cual se revisa en el siguiente capítulo.

IV. ANÁLISIS DE REGRESIÓN LINEAL SIMPLE

El análisis de regresión lineal simple es una técnica estadística que define la relación funcional entre una variable de regresión (X) y la respuesta (Y). Donde se supone que X es una variable continua y controlable por el experimentador.

4.1. OBSERVACIÓN GRÁFICA DEL PROBLEMA.

Dada una muestra de n individuos (u otras unidades de estudio), se observa para cada valor de X uno o más valores de Y, así, se tienen n pares de observaciones, las cuales pueden ser representadas por $(X_1, Y_1)(X_2, Y_2),\ldots,(X_n, Y_n)$ donde los subíndices se refieren a los diferentes individuos. Estos pares se pueden considerar como puntos en un espacio bidimensional y se pueden representar en un diagrama de dispersión, el cual se dibuja con dos ejes y por convención, el eje horizontal o eje-x es una escala para la variable-X, independiente o "causa" y el eje-y es una escala para la variable-Y, dependiente o "efecto" [13].

Estos diagramas ayudan a observar en forma gráfica la posible relación entre las dos variables en estudio, como se muestra en la figura 4.1, de donde se puede inferir que la relación entre X e Y es lineal.

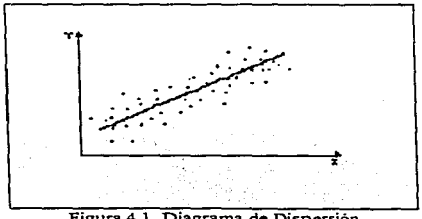


Figura 4.1. Diagrama de Dispersión

4.2. SUPUESTOS ESTADÍSTICOS PARA UN MODELO DE REGRESIÓN LINEAL SIMPLE.

Antes de comenzar a describir los procedimientos de análisis es importante conocer las suposiciones estadísticas para un modelo de regresión lineal. Se debe considerar que la línea recta que se busca es una aproximación al estado verdadero de trabajo. El hecho de que la línea se determina de los datos muestrales y no de la población requiere considerar problemas estadísticos concernientes a la estimación de parámetros poblacionales desconocidos de una línea recta de forma matemática general:

$$Y = \beta_0 + \beta_1 X \qquad (4.2.1)$$

donde β_0 es el intercepto y β_1 la pendiente.

Para hacer inferencias estadísticas acerca de la línea poblacional mediante pruebas de hipótesis se requiere la validez de las siguientes suposiciones estadísticas:

4.2.1.- Distribución de Probabilidad de la variable *Y*. Para un valor fijo de la variable X, Y es una variable aleatoria con una cierta distribución de probabilidad. La media (poblacional) de esta distribución se representa en la figura 4.2. como la media poblacional de Y dado un valor de X, (μ_{Y/X_i}), y la varianza como σ^2_{Y/X_i}, que es la varianza de Y al i-ésimo valor de X.

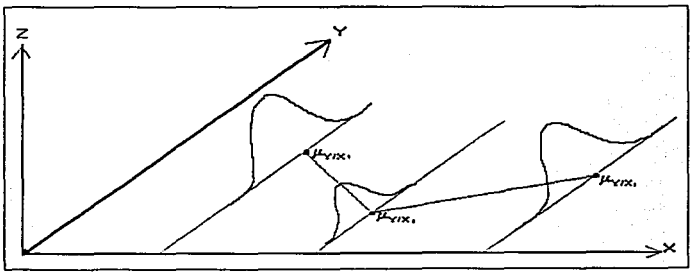


Figura 4.2. Distribución de probabilidad de la variable aleatoria Y.

4.2.2.-Independencia. Los valores de Y son estadísticamente independientes uno de otro. Esta suposición generalmente se viola cuando se hacen varias observaciones sobre un mismo individuo a diferentes tiempos, por ejemplo, en los estudios de disolución, donde se toman muestras del mismo vaso a diferentes tiempos.

4.2.3.- Linealidad. El valor medio de Y, $\mu_{Y/X}$, es una línea recta en función de X, esto es, si se conectan los puntos denotando los diferentes valores medios $\mu_{Y/X}$ se obtendrá una línea recta, como se muestra en la figura 4.3.

Usando símbolos matemáticos, se puede describir esta suposición por la ecuación:

$$\mu_{Y/X} = \beta_0 + \beta_1 X \qquad (4.2.2)$$

donde β_0 y β_1 son el intercepto y la pendiente de la línea recta.

De manera equivalente se puede expresar la ecuación (4.2.2) en la forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (4.2.3)$$

donde ε denota una variable aleatoria la cual tiene media cero a un valor de X, $\left(\mu_{\varepsilon/X} = 0\right)$. Ya que el valor de X es fijo y no aleatorio, la ecuación (4.2.3) representa la variable dependiente Y como la suma de un término constante $\left(\beta_0 + \beta_1 X\right)$ y una variable aleatoria (ε). Así, la distribución de probabilidad de Y y ε difieren sólo en el valor de este término constante, esto es, como ε tiene media cero, Y puede tener media $\left(\beta_0 + \beta_1 X\right)$.

La ecuación (4.2.3) que describe un modelo estadístico es diferente al modelo matemático lineal (4.2.1) debido a que este último no considera a Y como una variable aleatoria, es decir, no contempla la posibilidad de error en la predicción de Y, y el modelo estadístico se basa en la probabilidad de que suceda un evento, es decir, toma en cuenta las irregularidades que se relacionan con la variabilidad individual.

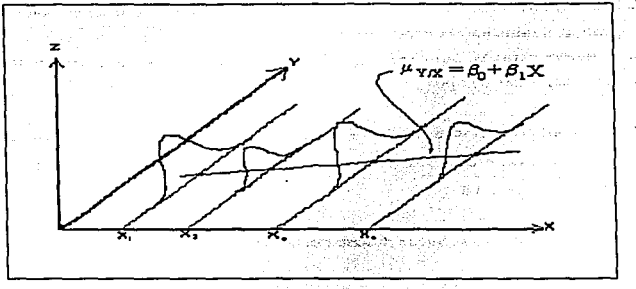


Figura 4.3. Suposición de linealidad.

La variable ε se representa en la figura 4.4, donde se observa que esta variable describe que tan lejos está una respuesta individual de la línea de regresión poblacional, es decir, una observación a un determinado valor de X es un error sobre la media $\mu_{Y/X}$ por una cantidad ε, la cual es aleatoria y varia de individuo a individuo, por lo que ε es referido como el componente del error en el modelo, el cual está dado por:

$$\varepsilon = Y - (\beta_0 + \beta_1 X) \quad \text{o} \quad \varepsilon = Y - \mu_{Y/X} \tag{4.2.4}$$

El concepto de componente del error es importante para definir un buen ajuste de la línea ya que esta debe tener desviaciones (o errores) pequeñas entre lo que se observa y lo que predice el modelo ajustado.

4.2.4.- Homocedasticidad. La varianza de Y es la misma en cada nivel de X. Esta suposición se conoce como homocedasticidad, donde "homo" significa "la misma" y "cedasticidad" significa "dispersión". La violación a esta suposición se le llama heterocedasticidad. En términos matemáticos, la suposición de homocedasticidad se puede escribir como $\sigma^2_{Y/X} = \sigma^2$, para toda X. Un número de técnicas de estadística sofisticada se pueden emplear para determinar si esta suposición se satisface.

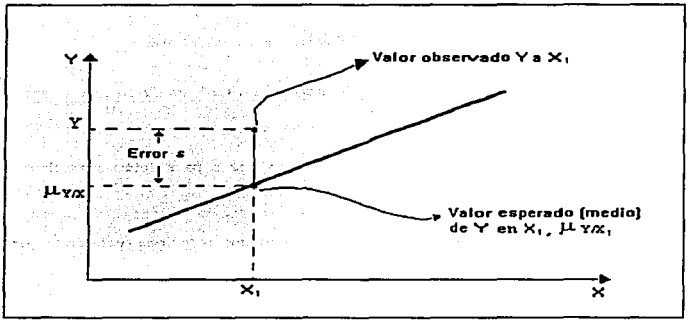


Figura 4.4. Componente del error.

4.2.5.- Distribución normal de Y en cada valor fijo de X. Esta suposición hace conveniente evaluar la significación estadística (por medio de intervalos de confianza y pruebas de hipótesis) de la relación de X e Y reflejada por la línea recta. Es importante enfatizar que si no existe mucha desviación a la normalidad, las conclusiones que se obtienen por un análisis de regresión en el que se asume normalidad generalmente serán confiables y ciertas. Esta propiedad de estabilidad con respecto a las desviaciones de normalidad es un tipo de robusticidad.

Si la suposición de normalidad no se cumple, se puede hacer un intento para transformar las observaciones usando logaritmo, raíz cuadrada u otra función para tratar de hacer al nuevo conjunto de observaciones aproximadamente normal. Debe tenerse cuidado cuando se usan tales transformaciones para que las otras suposiciones, tales como homogeneidad de varianzas, se cumplan con la variable transformada.

4.3. LÍNEA RECTA DE MEJOR AJUSTE.

En el análisis de regresión hay dos cuestiones básicas que deben tratarse:

1) ¿Cuál es el modelo matemático más apropiado para usar?, es decir, ¿Se usara una línea recta, una parábola o una función logaritmica?

2) Dado un modelo especifico, ¿Qué se debe hacer y cómo para determinar el modelo que mejor ajusta a los datos?

A continuación se describen dos formas analiticas de encontrar la línea recta de mejor ajuste:

4.3.1- Método de Mínimos Cuadrados.

El método de minimos cuadrados determina la línea recta que mejor ajusta a los datos en la cual la suma de cuadrados de las distancias entre los puntos de los datos observados sobre el diagrama de dispersión hacia la línea ajustada es minima como se muestra en la figura 4.5.

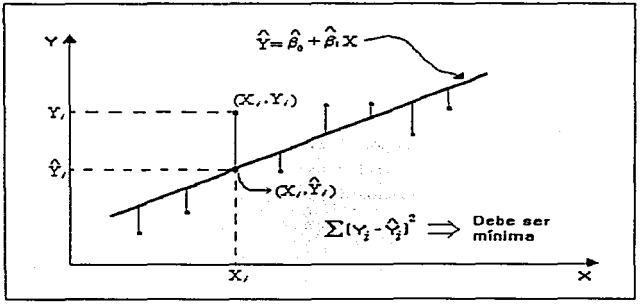


Figura 4.5. Desviaciones de los puntos observados sobre la línea de regresión.

El método de mínimos cuadrados se describe de la siguiente manera: Denotar a $\hat{Y}_i$ como la respuesta estimada a X_i basada sobre la línea de regresión ajustada, en otras palabras $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. La distancia vertical entre el punto observado (X_i, Y_i) y el correspondiente punto $(X_i, \hat{Y}_i)$ sobre la línea fijada está dada por el valor absoluto $|Y_i - \hat{Y}_i|$ ó $\left|Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right|$. La definición de Suma de Cuadrados de las distancias verticales de cada punto de la línea ajustada se escribe matemáticamente de la siguiente forma:

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 \qquad (4.3.1)$$

La solución de mínimos cuadrados se define por ser quien elige a $\hat{\beta}_0$ y $\hat{\beta}_1$ para los cuales la suma de cuadrados mencionada es mínima. A $\hat{\beta}_0$ y $\hat{\beta}_1$ se les llama estimadores de mínimos cuadrados de los parámetros β_0 y β_1 respectivamente y a la línea que se construye de acuerdo a esta definición se le llama línea de mínimos cuadrados.

La suma de cuadrados mínima correspondiente a los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ se les llama suma de cuadrados acerca de la regresión lineal, suma de cuadrados residual o suma de cuadrados debida al error (SCE).

Matemáticamente, la propiedad esencial de la cantidad SCE puede ser expuesta de la siguiente manera: Si β_0^* y β_1^* denotan algunos otros estimadores de β_0 y β_1, se tiene:

$$SCE = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 \le \sum_{i=1}^{n}\left(Y_i - \beta_0^* - \beta_1^* X_i\right)^2 \qquad (4.3.2)$$

es decir, los estimadores mínimo cuadráticos de β_0 y β_1 deben proporcionar menor suma de cuadrados que cualquier otra combinación de estimadores.

En la siguiente sección se describe la solución al problema de encontrar los estimadores de mínimos cuadrados de los parámetros β_0 y β_1.

4.4. SOLUCIÓN AL PROBLEMA DE MEJOR AJUSTE.

Cuando la relación entre X e Y es una línea recta y la observación Y a cada nivel de X es una variable aleatoria, el valor esperado de Y para cada valor de X se puede definir como en la ecuación (4.4.1), donde los parámetros de la recta, β_0 y β_1, son constantes desconocidas.

$$E(Y/X) = \beta_0 + \beta_1 X = \mu_{Y/X} \tag{4.4.1}$$

Se supone que cada observación, Y, puede describirse mediante el modelo (4.4.2), donde ε es un error aleatorio con media cero y varianza σ^2. Se supone que $\{\varepsilon\}$ constituye un conjunto de variables aleatorias no correlacionadas.

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{4.4.2}$$

Los parámetros del modelo β_0 y β_1 pueden estimarse mediante Mínimos Cuadrados si se tienen n pares de datos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

4.4.1. Solución por mínimos cuadrados.

Al aplicar la ecuación 4.4.2 se obtiene la siguiente expresión:

$$Y_i = \mu_{Y_{X_i}} + \varepsilon_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad ; \quad i = 1,2,\ldots,n \tag{4.4.3}$$

donde Y_i es la i-ésima observación, $\mu_{Y_{X_i}}$ es la media de Y en el i-ésimo nivel de X, β_0 es la ordenada al origen, β_1 la pendiente (coeficiente de regresión), X_i el i-ésimo valor de la variable independiente y ε_i el i-ésimo valor no observable de la variable aleatoria ε (error).

Luego entonces, para cada observación el modelo queda:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ &\bullet \\ &\bullet \\ &\bullet \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned} \qquad (4.4.4)$$

de manera que se tiene un sistema de ecuaciones que se puede escribir en forma matricial como:

$$\underline{\mathbf{y}} = \underline{\beta}\underline{\mathbf{X}} + \underline{\mathbf{e}} \qquad (4.4.5)$$

donde

$$\underline{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad \underline{\mathbf{X}} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \underline{\mathbf{e}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

En general, $\underline{\mathbf{y}}$ es un vector de respuesta (n x 1); $\underline{\mathbf{X}}$ es una matriz (n x p) de los niveles de las p variables de regresión, $\underline{\beta}$ es un vector de coeficientes de regresión (p x 1) y $\underline{\mathbf{e}}$ es un vector de errores aleatorios (n x 1).

Obsérvese que:

$$\underline{\mathbf{X}}\,\underline{\beta} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_0 + \beta_1 x_n \end{bmatrix} \qquad (4.4.6)$$

de manera que

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \qquad (4.4.7)$$

A partir de este sistema de ecuaciones se pueden calcular los estimadores de β_0 y β_1 mediante Mínimos Cuadrados de la siguiente manera:

Sea $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. la estimación de la respuesta a un X_i de la línea de regresión ajustada, la distancia vertical entre el punto (X_i, Y_i) y el punto $\left(X_i, \hat{Y}_i\right)$ de la recta ajustada está dada por el valor absoluto $\left|Y_i - \hat{Y}_i\right|$ ó $\left|Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right|$. La suma de cuadrados de tales distancias es:

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 \qquad (4.4.8 = 4.3.1)$$

La solución de mínimos cuadrados se define por la elección de $\hat{\beta}_0$ y $\hat{\beta}_1$ para los cuales la suma de cuadrados de la distancia es mínima. Esto conduce al siguiente problema:

Encontrar los valores de β_0 y β_1 (llamados $\hat{\beta}_0$ y $\hat{\beta}_1$) tales que $\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_i\right)^2$ sea mínima.

La solución a este problema es:

Sea $Q = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$, la función que se va a minimizar.

Puesto que se trata de hallar los valores de β_0 y β_1 que minimizan la función, se toman las derivadas parciales de Q respecto a estas variables y se igualan a cero cada una de ellas.

Las derivadas parciales son:

$$\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-1) \qquad (4.4.9)$$

y

$$\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) \qquad (4.4.10)$$

Si se igualan a cero cada una de estas dos expresiones, se obtienen dos ecuaciones llamadas **ecuaciones normales**, cuya solución permite obtener los valores de β_0 y β_1.

Al igualar a cero la derivada parcial (4.4.9) se obtiene:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\beta_0 - \sum_{i=1}^{n}\beta_1 X_i = 0$$

$$\sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

Nótese que al igualar a cero esta primera derivada parcial equivale a requerir que la suma de los residuales sea cero, ya que la expresión entre paréntesis es el residual $e_i = (Y_i - \beta_0 - \beta_1 X_i)$.

Para la segunda derivada parcial (4.4.10) se tiene:

$$\frac{\partial Q}{\partial \beta_1} = 2\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)\right)(-X_i) = 0$$

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

$$\sum_{i=1}^{n} Y_i(-X_i) - \sum_{i=1}^{n} \beta_0(-X_i) - \sum_{i=1}^{n} \beta_1(X_i)(-X_i) = 0$$

$$-\sum_{i=1}^{n} Y_i(X_i) + \sum_{i=1}^{n} \beta_0(X_i) + \sum_{i=1}^{n} \beta_1(X_i)(X_i) = 0$$

$$-\sum_{i=1}^{n} Y_i(X_i) + \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i(X_i)$$

Por lo tanto se tendrá el siguiente sistema de Ecuaciones Normales:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \qquad (4.4.11)$$

$$\beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i(X_i) \qquad (4.4.12)$$

4.4.2.- Solución del sistema de ecuaciones normales.

El sistema de ecuaciones normales se puede escribir como

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix} \qquad (4.4.13)$$

obsérvese que

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_1 & x_2 & . & . & . & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 + y_2 + \ldots + y_n \\ x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix} \qquad (4.4.14)$$

además

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_1 & x_2 & . & . & . & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} 1+1+\ldots+1 & x_1 + x_2 + \ldots + x_n \\ x_1 + x_2 + \ldots + x_n & x_1^2 + x_2^2 + \ldots + x_n^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix} \qquad (4.4.15)$$

Por lo tanto el sistema de ecuaciones normales se puede escribir como:

$$(\mathbf{X}'\mathbf{X})\underline{\beta} = \mathbf{X}'\underline{\mathbf{y}} \tag{4.4.16}$$

con $$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

De aqui que la solución a las ecuaciones normales sea:

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{\mathbf{y}} \tag{4.4.17}$$

por lo que el estimador de minimos cuadrados es:

$$\underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - \overline{X})^2} & \dfrac{-\sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ \dfrac{-\sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n}(X_i - \overline{X})^2} & \dfrac{n}{n\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix}$$

$$\underline{\hat{\beta}} = \begin{bmatrix} \overline{Y} - \hat{\beta}_1 \overline{X} \\ \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{bmatrix} = \begin{bmatrix} \overline{Y} - \hat{\beta}_1 \overline{X} \\ \dfrac{\sum_{i=1}^{n} X_i Y_i - \dfrac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \dfrac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

(4.4.18)

La recta de regresión estimada tiene la forma

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad ; \qquad i = 1,2,\ldots,n \qquad (4.4.19)$$

Como $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ se puede escribir:

$$\hat{Y} = \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) + \hat{\beta}_1 X_i$$

$$\hat{Y} = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i$$

$$\hat{Y} = \bar{Y} + \hat{\beta}_1 (X_i - \bar{X})$$

que en forma matricial puede expresarse como: $\hat{\underline{\mathbf{y}}} = \underline{\mathbf{X}}\,\hat{\underline{\beta}}$

(4.4.20)

Esto es :

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \qquad (4.4.21)$$

De (4.4.18) se obtiene que

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad (4.4.22)$$

y

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad (4.4.23)$$

Se observa que el cálculo de los coeficientes de regresión lineal simple son en cierta forma complicados y cuando el número de variables independientes aumenta, este cálculo se vuelve más complejo, por lo que se recomienda el empleo de un software estadístico para evaluar estos parámetros. Sin embargo se determina que la estimación de los coeficientes de regresión a través de mínimos cuadrados proporciona la mejor estimación de la linea recta que pasa a través de los datos experimentales, lo cual se verifica en la siguiente sección donde se analiza la calidad de la recta.

4.5. MEDIDA DE LA CALIDAD DE UNA LÍNEA RECTA: CME, R^2 Y FALTA DE AJUSTE.

4.5.1. Estimación de la varianza (σ^2).

Una vez que se determina la línea recta se debe evaluar si ayuda o no a predecir el comportamiento de la variable de respuesta Y. Una forma de responder a esta cuestión está dada por la Suma de Cuadrados del Error (SCE):

$$SCE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad (4.5.1 = 4.3.1)$$

donde $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Si SCE = 0, la línea recta ajusta perfectamente, esto es, todos los puntos observados caen sobre la línea teórica, $Y_i = \hat{Y}_i$ para cada i. Cuando el ajuste no es bueno, la SCE es grande, ya que las desviaciones de los puntos de la línea de regresión también son grandes.

Existen dos posibles factores que contribuyen a la SCE. El primero es que puede haber un lote con variación en los datos, esto es, σ^2 puede ser grande. El segundo factor es que la suposición de un modelo lineal no sea adecuada. Por lo tanto, es importante determinar los efectos separados de cada uno de estos componentes ya que cada uno conduce a resultados diferentes. Por ahora se puede asumir que el segundo factor no ocurre, es decir, si el modelo lineal es apropiado, se puede obtener una estimación de la varianza (σ^2) usando la SCE. Tal estimación es necesaria para hacer inferencias estadísticas concernientes a la verdadera relación entre X e Y. La estimación de σ^2 está dada por la ecuación 4.5.2.

$$S^2_{Y/X} = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{1}{n-k-1}SCE \qquad (4.5.2)$$

Donde:

n= Número de observaciones.

k = Número de variables independientes.

Otra forma de expresar la varianza es:

$$S^2_{Y/X} = \frac{n-1}{n-k-1}\left(S^2_Y - \hat{\beta}_1^2\, S^2_X\right) \qquad (4.5.3)$$

donde $S^2_Y = \dfrac{\sum_{i=1}^{n} Y_i^2 - \dfrac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}}{n-1}$ (4.5.3a); es la varianza muestral de las Y's observadas

y $S^2_X = \dfrac{\sum_{i=1}^{n} X_i^2 - \dfrac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}$ (4.5.3b); es la varianza muestral de las X's.

La ecuación usual para la varianza muestral está dada por $\sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$, la cual es apropiada cuando las Y's son independientes con la misma media μ y varianza σ^2. Puesto que en este caso μ es desconocida se puede dividir por n-1 en lugar de n, igual que para la varianza muestral es un estimador insesgado de σ^2. Este concepto se ilustra en la figura 4.6.

Si el modelo lineal es apropiado, la respuesta media poblacional $\mu_{Y/X}$ cambia con X. Por ejemplo, en disolución la respuesta media del porcentaje disuelto del principio activo cambia con respecto al tiempo. Al usar la línea de mínimos cuadrados como una aproximación a la línea poblacional, la media estimada de las Y's a un determinado valor de X es muy aproximado al valor real. Además, en lugar de sustraer $\bar{Y}$ de cada Y_i cuando se estima σ^2, se puede sustraer $\hat{Y}_i$ de Y_i debido a que $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ es el estimador de $\mu_{Y/X}$.

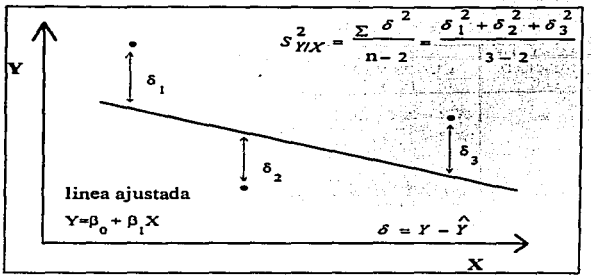


Figura 4.6. Varianza de la línea de regresión.

El coeficiente de determinación y la prueba de falta de ajuste, los cuales son también una medida de la calidad de la línea recta, se describen más adelante.

4.6. RELACIÓN ENTRE LA TABLA DE ANÁLISIS DE VARIANZA Y EL ANÁLISIS DE REGRESIÓN LINEAL.

Los resultados de un análisis de regresión lineal se pueden resumir en un cuadro de Análisis de Varianza (ANOVA). La tabla de ANOVA se puede emplear para contestar preguntas esenciales del análisis de regresión lineal, tales como: ¿La pendiente β_1 es verdaderamente cero?, ¿Cuál es la fuerza de la relación lineal? y ¿Es apropiado el modelo lineal?.

El análisis de varianza y el análisis de regresión están en estrecha relación, tanto que un problema de análisis de varianza se puede expresar en un contexto de regresión. La forma más común de presentar la tabla de ANOVA para la regresión lineal se muestra en el cuadro 4.1., el cual aparece en las salidas del paquete estadístico SAS.

Cuadro 4.1. Tabla de Análisis de Varianza para Regresión Lineal.

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD (g.l)	SUMA DE CUADRADOS (SC)	CUADRADO MEDIO (CM)	F	R^2
Regresión	k	$SCR = SCY - SCE$ $SCR = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$	$CMR = \frac{SCR}{k}$	$F = \frac{CMR}{CME}$	$R^2 = \frac{SCY - SCE}{SCY}$
Residual	n-k-1	$SCE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	$CME = \frac{SCE}{n-k-1}$		
Total	n-1	$SCY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			
Hipótesis a probar: *Ho: No existe relación lineal significativa de Y sobre X.* *Ha: Existe relación lineal entre X e Y.*					

k=Número de variables independientes. n=número de observaciones totales. $\overline{Y}$ =promedio de los valores observados. Y_i =*i*-ésimo valor observado. $\hat{Y}_i$ = *i*-ésimo valor predicho por el modelo

En esta tabla, el cuadrado medio se obtiene al dividir la suma de cuadrados por sus grados de libertad, el cociente de varianzas o estadístico F se obtiene al dividir el cuadrado medio de la regresión (*CMR*) entre el cuadrado medio residual o del error (*CME*). Es importante advertir que si bien las sumas de cuadrados y los grados de libertad son aditivos, los cuadrados medios no lo son.

El coeficiente de determinación, r^2, se obtiene al dividir la suma de cuadrados de la regresión entre la suma de cuadrados total, esto es,

$$r^2 = \frac{SCY - SCE}{SCY} \tag{4.6.1}$$

donde *SCY* es la suma de cuadrados de las desviaciones de las observaciones de *Y* con respecto a la media $\overline{Y}$, y *SCE* es la suma de cuadrados de las desviaciones entre las observaciones de *Y* y la línea de regresión ajustada. A *SCY* se le llama variación total no-explicada o suma de cuadrados total corregida por la media, debido a que representa la variación total de *Y* calculada antes de considerar el efecto lineal de la variable *X*; a *SCY-SCE* se le llama suma de cuadrados debida o explicada por la regresión porque *SCE* mide la variación de las *Y*'s observadas tomando en cuenta el efecto lineal de la variable *X*, por lo que (*SCY-SCE*) es matemáticamente equivalente a la expresión (4.6.2), la cual

representa la suma de cuadrados de las desviaciones entre los valores predichos y la media $\overline{Y}$, así se tiene como resultado la ecuación (4.6.3).

$$\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 \tag{4.6.2}$$

Variación total no-explicada = Variación debida a la regresión
+ Variación residual no-explicada

o
$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{4.6.3}$$

A la expresión anterior se le llama ecuación fundamental del análisis de regresión, la cual se ilustra en la figura 4.7.

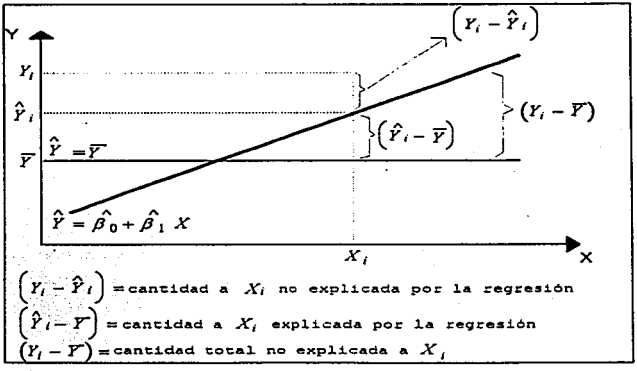


Figura 4.7. Variación explicada y no-explicada por la regresión

En el cuadro de análisis de varianza, el término cuadrado medio del error es el estimador de $S^2_{Y/X}$, el cual es un estimador de la varianza σ^2 cuando el modelo lineal es adecuado. Por otro lado, el término cuadrado medio de regresión $(SCY\text{-}SCE)/k$ proporciona una estimación de la varianza σ^2, cuando la variable X no interviene en la predicción de la variable dependiente Y, esto es, si la hipótesis nula $H_0: \beta_1 = 0$ no se rechaza, ya que cuando $\beta_1 \neq 0$, el valor del cuadrado medio de regresión se eleva en proporción a la magnitud de β_1 y por consecuencia hay una sobrestimación de la varianza.

Debido a que los términos cuadrado medio del error (*CME*) y cuadrado medio de regresión (*CMR*) son estadisticamente independientes uno de otro, si $H_0: \beta_1 = 0$ no se rechaza, el cociente de estos términos representa la relación de dos estimadores independientes de la misma varianza σ^2. Bajo la suposición de normalidad de las Y's, tal cociente tiene una distribución F y el valor del estadístico F se puede emplear para probar la hipótesis nula H_0: no existe relación lineal significativa de Y sobre X.

La prueba de F acerca del cociente de las variaciones es comparable con la prueba t referente a la pendiente β_1, que se analiza en la sección 4.7., pues en ambas se trata de probar el grado o fuerza de la relación entre X e Y en una regresión lineal. De hecho, se puede ver que la prueba t y la prueba F son pruebas equivalentes para determinar si es significativa la relación lineal entre dos variables. El valor de F_c es igual al cuadrado del valor de t_c, lo cual se muestra en la sección del estudio de caso. La ventaja de la prueba F es que puede ser general para cubrir los casos en que hay más de una variable independiente, cosa que no se puede hacer con la prueba t. La ventaja de la prueba t es que se puede utilizar en los casos en que β_1 toma valores diferentes de cero, mientras que la F sólo se emplea para la hipótesis nula $H_0: \beta_1 = 0$.

Otra forma menos común de representar la tabla de ANOVA se presenta en el cuadro 4.2.

Esta tabla difiere de la primera sólo en que parte la suma de cuadrados total corregida por la media, SCY, en sus dos componentes, la suma de cuadrados total sin corregir, $\sum_{i=1}^{n} Y_i^2$ y el factor de corrección $\frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$.

La relación entre estos dos componentes es:

$$SCY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \qquad (4.6.4)$$

La suma de cuadrados total sin corregir $\sum_{i=1}^{n} Y_i^2$ considera las n observaciones sobre Y antes de considerar la media poblacional de Y. El termino "regresión $\overline{Y}$" en el cuadro 4.2. se refiere a la variabilidad explicada al usar un modelo que sólo involucre β_0 que es estimada mediante $\overline{Y}$. Esta es la misma variabilidad explicada al usar sólo $\overline{Y}$ para estimar Y, sin incluir la contribución lineal de X. El término "regresión $(X/\overline{Y})$" describe la contribución de la variable X en la predicción de Y. Generalmente "regresión $(X/\overline{Y})$" se escribe como "regresión X".

Cuadro 4.2. Tabla de ANOVA menos común.

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD (gl)	SUMA DE CUADRADOS (SC)	CUADRADO MEDIO (CM)	COCIENTE DE VARIANZAS (F)
REGRESIÓN				
Variabilidad explicada por $\overline{Y}$	k	$\frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$		
Variabilidad explicada tomando en cuenta X, $(X/\overline{Y})$	k	$SCY\text{-}SCE$	$CMR = \frac{SCY - SCE}{gl}$	$\frac{CMR}{CME}$
RESIDUAL	n-k-1	SCE	$CME = \frac{SCE}{gl}$	
TOTAL (no corregida por la media) (r^2 =)	n-1	$\sum_{i=1}^{n} Y_i^2$		
Hipótesis a probar: *Ho: No existe relación lineal significativa de Y sobre X.* *Ha: Existe relación lineal entre X e Y.*				

4.7. INFERENCIAS ACERCA DE LA PENDIENTE E INTERCEPTO.

Una vez que se estima la línea recta por mínimos cuadrados se pueden realizar ciertas inferencias para determinar si la línea ajustada ayuda o no a predecir Y, y con que grado de confiabilidad lo hace. Para esto es de ayuda práctica calcular intervalos de confianza o probar hipótesis estadísticas acerca de los parámetros desconocidos con la suposición del modelo lineal, tomando en cuenta la incertidumbre de usar una muestra. Tales intervalos de confianza y pruebas de hipótesis requieren la suposición de que la variable aleatoria Y tiene una distribución normal a cada valor fijo de X. Bajo esta suposición se puede deducir que los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ son normalmente distribuidos cada uno con medias β_0 y β_1 respectivamente cuando se mantiene la ecuación $\mu_{y|x} = \beta_0 + \beta_1 X$ y con varianzas fácilmente derivables. Estos estimadores al igual que sus estimadores de varianzas, se pueden usar para formar intervalos de confianza y pruebas estadísticas basadas en la distribución t.

Para probar la hipótesis $H_o : \beta_1 = \beta_1^A$, donde β_1^A es algún valor hipotético para β_1, el estadístico de prueba es:

$$T = \frac{\hat{\beta}_1 - \beta_1^A}{S_{y|x} / S_x \sqrt{n-1}} \qquad (4.7.1)$$

Este estadístico presenta una distribución t, con n-2 grados de libertad (g.l), cuando H_0 es verdadera, y relaciona una variable aleatoria que se distribuye normalmente, dividida por un estimador de su desviación estándar.

Por ejemplo, cuando se hace un ensayo para encontrar la potencia de un fármaco donde se obtiene una gráfica de la cantidad recuperada contra cantidad conocida, la magnitud de la pendiente tiene un significado físico especial. Una pendiente de 1 indica que la cantidad recuperada es igual a la cantidad en la muestra después de la corrección por el blanco. Una pendiente diferente de 1 indica que la cantidad que se recupera es algún porcentaje constante de la potencia de la muestra. Así puede ser de interés probar las siguientes hipótesis: $H_o : \beta_1 = 1$ contra $H_a : \beta_1 \neq 1$.

Para probar la hipótesis $H_o : \beta_0 = \beta_0^{\wedge}$ el estadístico de prueba que se usa es:

$$T = \frac{\hat{\beta}_0 - \beta_0^{\wedge}}{S_{Y/X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}}} \qquad (4.7.2)$$

el cual también tiene distribución *t* con n-2 g.l cuando H_0 es verdadera.

Ambos estadísticos de prueba tienen n-2 g.l ya que involucran a la varianza muestral, $S^2_{Y/X}$, la cual por sí misma tiene n-2 g.l y es el único componente aleatorio en el denominador de cada uno de los estadísticos.

Al realizar cada uno de los procedimientos de pruebas de hipótesis a niveles de significancia α, H_0 se rechaza siempre que ocurra lo siguiente:

$$\begin{cases} T \geq t_{n-2,1-\alpha} & \text{para una prueba unilateral superior} \\ T \leq -t_{n-2,1-\alpha} & \text{para una prueba unilateral inferior} \\ |T| \geq t_{n-2,1-\alpha/2} & \text{para una prueba bilateral} \end{cases}$$

donde $t_{n-2,1-\alpha}$ denota el punto 100(1-α)% de la distribución *t* con n-2 grados de libertad.

Debe tenerse especial cuidado para interpretar correctamente los resultados de las pruebas relacionadas con la pendiente e intercepto ya que se cometen errores con frecuencia en este aspecto, ya que en muchas ocasiones no se tiene claro el criterio de decisión y puede ser rechazada una hipótesis que no se debe rachazar.

A continuación se discuten las conclusiones que se pueden hacer sobre el rechazo o no rechazo de las hipótesis nulas que se prueban comúnmente acerca de la pendiente y el intercepto. Se asume que los supuestos de normalidad, independencia y homogeneidad de varianza se cumplen.

4.7.1.- Prueba para pendiente cero.

La prueba más importante de las hipótesis tratadas con los parámetros del modelo lineal es aquella en la cual se determina si la pendiente de la línea de regresión es significativamente diferente de cero o no, lo cual equivale a decir si las X's ayudan a predecir Y usando un modelo lineal. La hipótesis nula apropiada para esta prueba es $H_o : \beta_1 = 0$. Si se ignora por ahora la posibilidad de cometer un error tipo I (rechazar una H_0 verdadera) o un error tipo II (no rechazar una H_0 falsa), se pueden hacer las siguientes interpretaciones:

1) Si H_0 no se rechaza esto significa que:

a) Para un modelo lineal verdadero, X provee poca o nada de ayuda en la predicción de los valores de Y.

b) La verdadera relación entre X e Y no es lineal, el modelo verdadero puede involucrar funciones cuadráticas, cúbicas u otras más complejas.

Combinando (a) y (b), se puede decir que no rechazar H_0 implica que el modelo lineal en X no es el mejor modelo para usar porque no provee mucha ayuda para predecir Y.

2) Si H_0 se rechaza, esto significa que:

a) X provee información significante para la predicción de Y; esto es, el modelo lineal $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ es adecuado para estimar los valores de $\hat{Y}$, conociendo los valores de X_i.

b) Un mejor modelo tal vez puede tener un término curvilíneal.

Al combinar (a) y (b), se puede decir que rechazar H_0 implica que el modelo lineal al tomar en cuenta X es mejor que cuando no se incluye esta variable, pero también puede ser que el modelo sea sólo una buena "aproximación lineal" cuando la verdadera relación es no lineal.

4.7.2.- Prueba para el intercepto cero.

Esta prueba se aplica para determinar si la línea recta poblacional pasa por el origen. La hipótesis nula es $H_o : \beta_0 = 0$. Si esta hipótesis nula no se rechaza, es posible eliminar la constante del modelo ya que se aporta evidencia para sugerir que la línea pasa a través del origen y que hay observaciones tomadas alrededor del origen para mejorar la estimación de la recta. En algunos casos esta hipótesis no es de interés, porque no se tienen los datos cerca del origen.

Por ejemplo, se sabe que muchas curvas de calibración pasan a través del origen; esto es, la respuesta del ensayo debe ser cero si la concentración del fármaco es cero. El cálculo de la pendiente se simplifica si se obliga a la línea a pasar por el punto (0,0).

4.8. INFERENCIAS ACERCA DE LA LÍNEA DE REGRESIÓN $\mu_{Y/X} = \beta_0 + \beta_1 X$

Uno de los usos más importantes de la recta de regresión muestral es la obtención de predicciones de la variable dependiente para algún valor de la variable independiente. El valor estimado $\hat{Y}_i = \beta_0 + \beta_1 X_i$ es la mejor estimación que se puede obtener de μ_{Y/X_i} (valor medio de Y, dado un valor X_i) y de Y_i (valor real de Y correspondiente al valor dado X_i). Frecuentemente se requieren ambos tipos de predicciones.

Para obtener la mejor estimación puntual de las predicciones del valor medio y del valor real de Y, el valor de la variable independiente (X_0) se sustituye en la ecuación de la línea recta de regresión muestral, con lo cual resulta que el valor de predicción es $\hat{Y}_{X_0} = \beta_0 + \beta_1 X_0$. Así, $\hat{Y}_{X_0}$ es una estimación tanto de μ_{Y/X_0} como de Y_{X_0}.

Aunque ambas estimaciones tienen el mismo valor, se debe subrayar que se interpretan de forma diferente. Estas interpretaciones distintas son importantes cuando se describen las estimaciones mediante intervalos. Se analizará en concreto que el intervalo de confianza para estimar un valor

aislado será necesariamente mayor que el correspondiente al valor medio, ya que el primero tendrá siempre un error estándar mayor.

4.8.1. Intervalos de predicción.

Una estimación por intervalos tiene su centro, en términos probabilísticos en la estimación puntual, por consiguiente, los extremos del intervalo se obtienen empleando la información que se refiere a la distribución probabilística del estimador puntual y su error estándar. La estimación puntual para las predicciones del valor medio y del valor real de Y es siempre el mismo valor de $\hat{Y}_{x_0}$. Sin embargo, los extremos de los intervalos de estimación de estos dos tipos de predicción, basados en un nuevo valor X_0 serán diferentes porque sus errores estándar son diferentes.

Para el intervalo de predicción del valor real de Y_{x_0}, el error estándar que se usa para la construcción de este intervalo se llama error estándar de la predicción y se denota por S_f. Su calculo se obtiene con la ecuación (4.8.1).

$$S_f = S_{Y/X}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} \qquad (4.8.1)$$

Se observa que S_f será siempre mayor que $S_{Y/X}$ puesto que la expresión bajo el símbolo radical siempre será mayor que uno, además S_f depende del valor particular que se asigna a X_0. Mientras más alejado este el valor de X_0 de la media de los valores usados para determinar la recta de regresión muestral, menos exacta será la predicción basada en esa recta. Se debe tener cuidado al hacer predicciones que vayan más allá del recorrido de los valores observados. En la figura 4.8. se observa que las franjas o regiones de confianza se limitan por curvas, y no por líneas rectas. Finalmente se observa que si n es grande y si el valor X_0 está cerca de $\bar{X}$, entonces la expresión bajo el símbolo radical tendrá un valor cercano a 1.0 por lo cual $S_{Y/X}$ y S_f serán aproximadamente iguales, es decir, mientras mayor sea la muestra y menor la desviación de X_0 respecto a $\bar{X}$, más confianza se tendrá en los resultados muestrales y en la predicción subsiguiente. La predicción será más exacta en las proximidades del punto $(\bar{X}, \bar{Y})$.

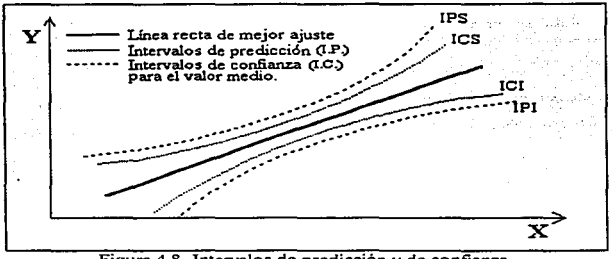


Figura 4.8. Intervalos de predicción y de confianza.

Ahora se puede usar el estimador puntual $\hat{Y}_{x_0}$ y el error estándar S_f para construir un intervalo de predicción de 100(1-α)% para Y_{x_0}. El estadístico adecuado para este caso tiene distribución t con (n-2) grados de libertad, como se muestra en la ecuación (4.8.2).

$$IP = \hat{Y}_{x_0} \pm t_{n-2,1-\alpha/2} S_f \tag{4.8.2}$$

En la predicción de una respuesta Y hay dos fuentes de error en la operación: el error individual medido por la varianza (σ^2) y el error en la estimación de μ_{Y/x_0} al usar $\hat{Y}_{x_0}$. Esto puede ser expresado por la ecuación (4.8.3).

$$\underset{\text{error en la predicción de cada } Y \text{ a } X_0}{Y - \hat{Y}_{X_0}} = \underset{\text{desviación de las Y's individuales de la media verdadera a } X_0}{(Y - \mu_{Y/X_0})} + \underset{\text{desviación de } \hat{Y}_{X_0} \text{ de la media verdadera a } X}{(\hat{Y}_{X_0} - \mu_{Y/X_0})} \tag{4.8.3}$$

Esta representación permite escribir la varianza de una respuesta individual predicha al nivel X_0 como:

$$Var\, Y + Var\, \hat{Y}_{x_0} = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2} \right] \tag{4.8.4}$$

esta expresión de varianza se estima al reemplazar σ^2 por su estimador $S^2_{Y/X}$.

4.8.2. Intervalos de confianza.

Se puede encontrar para un valor de $X=X_0$ el intervalo de confianza para el valor medio de Y a X_0, μ_{Y/X_0}. En este caso, el error estándar adecuado se denota por el símbolo $S_{\bar{Y}/X_0}$ y se expresa por la ecuación (4.8.5).

$$S_{\bar{Y}/X_0} = S_{Y/X}\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} \qquad (4.8.5)$$

Como en el caso de $S_{\bar{Y}}$, también $S_{\bar{Y}/X_0}$ depende de n, X_0 y $S_{Y/X}$. Sin embargo, el valor $S_{\bar{Y}/X_0}$ será siempre menor que el de $S_{\bar{Y}}$ ya que contiene un termino positivo adicional bajo el radical. También en este caso el estadístico apropiado tiene una distribución t con *(n-2)* grados de libertad. Así, los extremos del intervalo de 100(1-α)% son:

$$\hat{Y}_{X_0} \pm t_{n-2,1-\alpha/2}\, S_{\bar{Y}/X_0} \qquad (4.8.6)$$

Para hacer inferencias acerca de puntos específicos sobre la línea de regresión, se utiliza un intervalo de confianza para la línea de regresión sobre el rango de valores de X. La manera más conveniente es hacer una gráfica de los límites de confianza obtenidos para los distintos valores específicos de X y entonces esbozar las dos curvas que conecten estos puntos. Tales curvas son llamadas bandas de confianza para la línea de regresión como se muestran en la figura 4.8.

4.9. PRUEBA DE FALTA DE AJUSTE.

A menudo, los modelos de regresión se adecuan a los datos cuando no se conoce la relación funcional real. En temas anteriores se señala que la estrategia general en el análisis de regresión es considerar la suposición que un modelo lineal es apropiado para emplearse en la predicción de la variable dependiente, sin embargo, es importante conocer si el orden del modelo tentativamente supuesto es correcto o si existe un modelo más complejo que pueda describir mejor los datos. Un método para determinar si la suposición de modelo lineal es razonable, es decir, probar la validez del modelo, se basa en una prueba de *falta de ajuste* o *bondad de ajuste*. A pesar de que se usa una sola variable independiente, la generalización para K variables de regresión es directa.

Las hipótesis que se desean probar son las siguientes:

H_0: El modelo se ajusta adecuadamente a los datos (No existe falta de ajuste)
H_a: El modelo no se ajusta adecuadamente a los datos (Existe falta de ajuste)

El principal estadístico que se emplea es la suma de cuadrados residual (SCE).

Como se mencionó, existen dos posibles razones para obtener un valor grande de SCE, la primera es que haya una gran variabilidad dentro de los mismos datos (σ^2 es grande); el segundo es que la suposición del modelo lineal no es completamente apropiado, es decir, la suma de cuadrados residual contiene un componente que describe el error puro (σ^2 no relacionada a la regresión cuando se emplea X) y un componente que describe la extensión de falta de ajuste en la suposición del modelo lineal.

Para estimar estos dos componentes, primero se debe estimar el error puro de σ^2 por medio de "replicar las observaciones", es decir, repetir las observaciones en un mismo valor de X, las cuales serán observaciones independientes de una distribución con varianza σ^2, así, para cualquier X se puede obtener una estimación de σ^2 al aplicar la ecuación usual de varianza muestral a las observaciones de Y que se toman a un valor de X. De esta forma, una estimación se puede considerar como *pura* debido a que no depende del modelo que se considera. Todas las X's asociadas con dos o

más observaciones de Y se pueden usar para tener una estimación pura de la varianza σ^2, ya que una sola observación por si misma no proporciona información de la variabilidad.

A continuación se describe una forma general para obtener la suma de cuadrados del error puro y la estimación correspondiente de la varianza del error puro cuando la suposición de homocedasticidad se cumple.

Se supone que

$$\begin{array}{ll} Y_{11}, Y_{12}, \ldots, Y_{1n_1} & \text{son observaciones con } n_1 \text{ replicas a } X_1 \\ Y_{21}, Y_{22}, \ldots, Y_{2n_2} & \text{son observaciones con } n_2 \text{ replicas a } X_2 \\ \vdots \quad \vdots & \vdots \quad \vdots \quad \vdots \\ Y_{k1}, Y_{k2}, \ldots, Y_{kn_k} & \text{son observaciones con } n_k \text{ replicas a } X_k \end{array} \qquad (4.9.1)$$

por lo que se tienen k conjuntos de observaciones con replicas. La contribución a la suma de cuadrados del error puro de X_i y la estimación correspondiente de σ^2 (con $n_i - 1$ g.l) está dada por la ecuación (4.9.2) y (4.9.3) respectivamente.

$$\sum_{u=1}^{n_i} \left(Y_{iu} - \overline{Y}_i\right)^2 \qquad (4.9.2) \qquad \text{y} \qquad \frac{1}{n_i - 1} \sum_{u=1}^{n_i} \left(Y_{iu} - \overline{Y}_i\right)^2 \qquad (4.9.3)$$

De esto se observa que la suma de cuadrados del error puro se obtiene al sumar la ecuación (4.9.2) sobre todos los niveles de X:

$$\text{Suma de Cuadrados del error puro } \left(SC_{ep}\right) = \sum_{i=1}^{k} \sum_{u=1}^{n_i} \left(Y_{iu} - \overline{Y}_i\right)^2 \qquad (4.9.4)$$

con $\sum_{i=1}^{k} (n_i - 1) = (n_1 - 1) + (n_2 - 1) + \ldots + (n_{k-1} - 1) = N - k$ grados de libertad.

y el estimador de la varianza del error puro es:

$$S_{ep}^2 = \frac{1}{n_1 + n_2 + \ldots + n_k - k} SC_{ep} = \frac{1}{n_1 + n_2 + \ldots + n_k - k} \sum_{i=1}^{k} \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y}_i)^2 \qquad (4.9.5)$$

La suma de cuadrados de falta de ajuste se obtiene al restar SC_{ep} de SCE, esto es:

$$\begin{array}{l} SC(\text{falta de ajuste}) = \text{suma de cuadrados residual} - \text{suma de cuadrados del error puro} \\ SC(\text{falta de ajuste}) = SCE - SC_{ep} \end{array} \qquad (4.9.6)$$

Los grados de libertad para SC(falta de ajuste) se obtiene al sustraer los grados de libertad del error puro de los grados de libertad residual:

$$\begin{array}{l} \text{gl (falta de ajuste)} = \text{gl (residual)} - \text{gl (error puro)} \\ \text{gl (falta de ajuste)} = (n-2) - (n-k) = k-2 \end{array} \qquad (4.9.7)$$

El siguiente paso es determinar el cuadrado medio del error puro (CM_{ep}) y el cuadrado medio para la falta de ajuste (CM_{falta}) dividiendo cada suma de cuadrados por los correspondientes grados de libertad. El CM_{ep} es simplemente S_{ep}^2, la estimación de la varianza del error puro.

El paso final es la comparación del estadístico F_0 (4.9.8) con el punto $100(1-\alpha)\%$ de la distribución F, con los grados de libertad de falta de ajuste como numerador y los grados de libertad del error puro como denominador $\left(F_{gl_{falta}, gl_{ep}, \alpha}\right)$, y se rechaza la hipótesis de idoneidad del modelo si $F_0 > \left(F_{gl_{falta}, gl_{ep}, \alpha}\right)$.

$$F = \frac{CM_{falta}}{CM_{ep}} \qquad (4.9.8)$$

Si la hipótesis nula de la adecuación del modelo se rechaza, el modelo se debe descartar y buscar otro que resulte más apropiado. Si la hipótesis nula no se rechaza, no existe una razón

aparente para dudar de la adecuación del modelo y por lo tanto se puede concluir que el modelo tiene un buen ajuste.

La figura 4.9 muestra el procedimiento para probar si el modelo lineal es apropiado. En este diagrama no sólo se aplica la prueba de ajuste para un modelo lineal simple sino para cualquier modelo

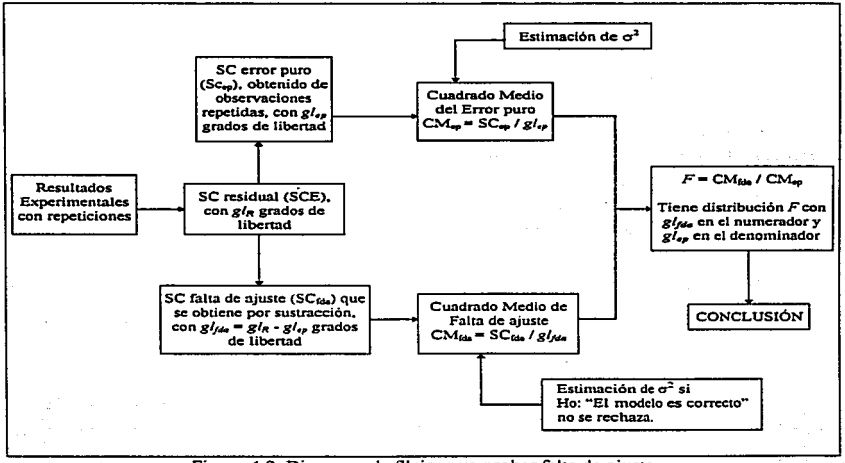


Figura 4.9. Diagrama de fllujo para probar falta de ajuste.

La prueba de falta de ajuste se puede incorporar en la tabla de Análisis de Varianza para regresión lineal, como se muestra en el cuadro 4.3, en el cual el residual se divide en falta de ajuste y error puro para obtener la información esencial con respecto a la prueba de falta de ajuste.

Cuadro 4.3. Tabla de ANOVA incluyendo la prueba de falta de ajuste

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD (gl))	SUMA DE CUADRADOS (SC)	CUADRADO MEDIO (CM)	(F)
REGRESIÓN (X)	k	SCR=SCY-SCE		
RESIDUAL				
falta de ajuste	k-2	$SC_{(fda)} = SCE - SC_{ep}$	$CM_{fda} = \frac{SC_{fda}}{gl_{fda}}$	$F = \frac{CM_{fda}}{CM_{ep}}$
error puro	N-k	$SC_{ep} = \sum_{i=1}^{k}\sum_{u=1}^{n_i}(Y_{iu} - \overline{Y}_i)^2$	$CM_{ep} = \frac{SC_{ep}}{gl_{ep}}$	
TOTAL (corregida por la media) (r^2 =)	N-1			

Hipótesis a probar:
H_0: El modelo se ajusta adecuadamente a los datos (No existe falta de ajuste)
H_A: El modelo no se ajusta adecuadamente a los datos (Existe falta de ajuste)

Por otro lado cuando se determina que existe un punto extremo en los datos que se van a analizar, se pueden llevar a cabo dos procedimientos razonables para un ajuste por la presencia de este punto extremo.

Primero, tal vez no es correcto eliminar por completo esta observación, ya que este punto extremo puede contener información con respecto a la falta de ajuste, no obstante, su efecto sobre la suma de cuadrados del error puro es considerable, por lo que se deberá retener pero no se debe emplear en el cálculo de la suma de cuadrados del error puro, es decir, no contemplar su contribución sobre este parámetro. Esto se debe hacer observando el supuesto de homogeneidad de varianza ya que causa un incremento en el cuadrado medio de la falta de ajuste y un descenso en el cuadrado medio del error puro.

La prueba de F para falta de ajuste sin el punto extremo proporciona un valor de F mayor que el valor de F cuando se consideran todos los datos, por lo que la hipótesis nula de que el modelo lineal es apropiado (no hay falta de ajuste) corre el riesgo de ser rechazada, a diferencia de alguna decisión anterior.

Sin embargo, la hipótesis nula tal vez no se rechace a un nivel de significancia diferente, por lo que se puede pensar que este primer procedimiento es contradictorio. Así que se cuestiona sobre la demostración de la "falta de ajuste" tomando en cuenta el punto extremo. Esto permite considerar un segundo procedimiento, el cual consiste en descartar por completo el punto extremo y reanalizar los datos. Tal vez la relación lineal resulte más fuerte cuando el punto extremo sea removido.

Otro punto interesante es decidir si se remueve o no el punto extremo, lo cual es muy difícil, ya que un valor de r^2 se considera "alto" en un sentido práctico. La respuesta depende del propósito de la investigación y de la experiencia previa. Si la investigación se realiza en un ambiente natural, un valor pequeño de r^2 como 0.4 ó 0.2 se puede considerar alto en términos de que ayuda a ilustrar la variable independiente. También algunos estudios previos pueden tener resultados con valores de r^2 menores a los resultados que se obtienen en un experimento reciente.

Al finalizar el análisis se puede llegar a la conclusión de que la falta de ajuste no es significativa y por lo tanto un modelo lineal es apropiado para describir la relación entre las variables de estudio.

V. COEFICIENTE DE CORRELACIÓN Y DE DETERMINACIÓN.

5.1. COEFICIENTE DE CORRELACIÓN.

Los métodos de correlación se utilizan para medir la "asociación" de dos o más variables. Generalmente se desea determinar si dos o más variables están relacionadas, en el sentido que una variable se podría predecir al conocer la otra. Por ejemplo, si se puede predecir la disolución de una tableta en base a su dureza, se puede decir que la disolución y la dureza están correlacionadas. En el análisis de correlación se asume una relación lineal entre las variables. La correlación se aplica generalmente a la relación de variables continuas, y su mejor visualización es a través de diagramas de dispersión o diagramas de correlación.

5.1.1.- Definición del coeficiente de correlación.

Una medida de la relación poblacional entre dos variables aleatorias es su covarianza como se muestra en la ecuación 5.1.1.

$$C[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] \qquad (5.1.1)$$

Si bien la covarianza tiene muchos usos estadísticos importantes, por lo general no es un buen indicador de la fuerza relativa de la relación entre dos variables, debido a que su magnitud depende mucho de las unidades que se emplean para medir las variables. Por esta razón es necesario "estandarizar" la covarianza de dos variables para disponer de una buena medida de ajuste. Esta estandarización se obtiene al dividir $C[X,Y]$ por el producto de las desviaciones estandar σ_X y σ_Y. La medida resultante se llama coeficiente de correlación poblacional o de Pearson, y se denota por la letra ρ (ecuación 5.1.2).

$$\rho = \frac{\text{covarianza de X e Y}}{\text{(desv. est. de X) (desv. est. de Y)}} = \frac{C[X,Y]}{\sigma_X\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \qquad (5.1.2)$$

Así, el coeficiente de correlación es un estadístico que proporciona una medida de la asociación lineal entre dos variables que se consideran aleatorias aunque en la regresión, mínimo hay

una variable que fija el investigador. El coeficiente de correlación también tiene propiedades que revelan la aproximación de regresión lineal, más no es una medida de linealidad. Como en todos los problemas de estimación, para calcular el parámetro poblacional ρ se usan los datos muéstrales. En este caso, el estadístico muestral recibe el nombre de coeficiente de correlación muestral, y se denota con la letra r, su valor se define de igual manera que ρ. Así para dos variables X e Y, el coeficiente de correlación muestral se define por:

$$Cov[X,Y] \Rightarrow S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{5.1.3}$$

$$\hat{\sigma}_x^2 \Rightarrow S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{5.1.4}$$

$$\hat{\sigma}_y^2 \Rightarrow S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{5.1.5}$$

Al sustituir estas estimaciones en la ecuación de ρ (5.1.2) se obtiene:

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\text{covarianza de los valores muestrales de X e Y}}{\text{(desv. est. muestral de X) (Desv. est. muestral de Y)}} \tag{5.1.6}$$

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left[\left(\frac{1}{n-1}\right)^2\right]^{1/2}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]^{1/2}} \tag{5.1.7}$$

de manera que el coeficiente de correlación muestral (r) se define por la siguiente ecuación:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]^{1/2}} \tag{5.1.8}$$

Otra forma de expresar esta ecuación matemática es:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)/n}{\left\{\left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right]\right\}^{1/2}} \qquad (5.1.8a)$$

Una expresión equivalente para r, la cual ilustra su relación matemática con la estimación por mínimos cuadrados de la pendiente de una línea recta de regresión es:

$$r = \frac{S_x}{S_y}\hat{\beta}_1 \qquad (5.1.9)$$

5.1.2. Propiedades matemáticas del coeficiente de correlación.

Las propiedades matemáticas del coeficiente de correlación (r) son las siguientes:

1) El rango de valores posibles de r es de -1 a 1.

2) El coeficiente de correlación se representa con una cantidad adimensional, esto es, r es independiente de las unidades en que se miden X e Y.

3) El coeficiente de correlación es positivo, negativo o cero de acuerdo si $\hat{\beta}_1$ es positivo, negativo o cero. Esta propiedad proviene de la ecuación (5.1.9), es decir, del valor de la pendiente.

5.1.3. Interpretación del coeficiente de correlación.

En las suposiciones estadísticas para el análisis de regresión lineal no se considera a la variable X como una variable aleatoria, sin embargo es importante considerar a X e Y como variables aleatorias ya que en este contexto, el valor de r se interpreta como un índice de asociación entre X e Y en el siguiente sentido [14]:

1) Cuando la asociación es positiva, el valor de r también es positivo.(fig. 5.1a). Si todos lo pares de valores X e Y están sobre la recta con pendiente positiva creciente, se dice que hay una relación lineal positiva directa entre X e Y, el valor de $C[X,Y]$ será exactamente igual al producto σ_X por σ_Y de modo que ρ será igual a +1.

2) Cuando la asociación es negativa decreciente, el valor de r es negativo. (fig. 5.1b). Cuando la relación es negativa indirecta, $C[X,Y]$ será exactamente igual a $-(\sigma_X)(\sigma_Y)$ y ρ será igual a -1.

3) Si r es cercano a cero, existe poca o ninguna asociación lineal entre X e Y (fig. 5.1c).

La falta de asociación significa que el valor de una variable no se puede predeterminar linealmente con seguridad si se conoce el valor de la otra variable.

Existen dos aspectos interesantes en el análisis de correlación, los cuales son:

a) En problemas típicos de correlación, X e Y son variables aleatorias, en contraste al caso de regresión lineal, donde X se considera fija, seleccionada a priori por el investigador.

b) En una distribución normal bivariada, X e Y están relacionadas linealmente. La regresión de X e Y y Y sobre X es una linea recta. Así, cuando se prueban estadísticamente los coeficientes de correlación, no se prueba la linealidad.

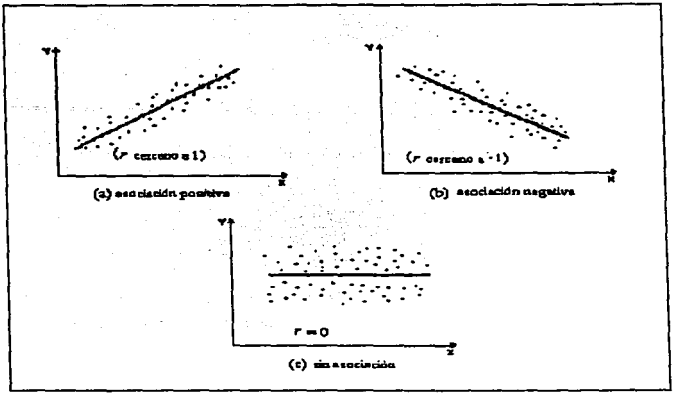


Figura 5.1. Coeficiente de correlación como una medida de asociación

5.1.4. Prueba de hipótesis para el coeficiente de correlación.

El coeficiente de correlación es una medida aproximada del grado de asociación de dos o más variables. Una cuestión importante desde el punto de vista estadístico es que si un coeficiente de correlación es "real" o debido a un cambio. Si dos variables no están correlacionadas, el coeficiente de correlación es cero, por lo tanto es útil probar la hipótesis $H_0: \rho = 0$ para evaluar la asociación entre las variables X e Y; esta prueba es equivalente a la prueba de hipótesis $H_0: \beta_1 = 0$. Esta equivalencia se proporciona a partir de las ecuaciones $\beta_1 = \rho\sigma_Y/\sigma_X$ y $\hat{\beta}_1 = rS_Y/S_X$ (5.1.9), de las cuales se dice que β_1 es positiva, negativa o cero conforme ρ es positivo, negativo o cero, y existe una relación análoga entre $\hat{\beta}_1$ y r. De esta manera es posible escribir la prueba estadística mediante la hipótesis $H_0: \rho = 0$ únicamente en términos de r y n. El estadístico de prueba es:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (5.1.10)$$

el cual tiene una distribución t con n-2 grados de libertad cuando la hipótesis nula no se rechaza.

El estadístico 5.1.10 proporciona la misma respuesta numérica que (5.1.11).

$$t = \frac{\hat{\beta}_1 S_X \sqrt{n-1}}{S_{Y|X}} \qquad (5.1.11)$$

Por otra parte, una prueba de hipótesis nula H_0: $\rho = \rho_0$, $\rho_0 \neq 0$ no está en relación directa a la prueba de β_1, así como la hipótesis H_0: $\rho = \rho_0$, $\rho_0 \neq 0$ no es equivalente a la hipótesis H_0: $\beta_1 = \beta_1^{(0)}$ para algún valor de $\beta_1^{(0)}$. Sin embargo la prueba de hipótesis para H_0: $\rho = \rho_0$, $\rho_0 \neq 0$ es importante cuando las experiencias previas o la teoria sugieren un valor particular de ρ_0.

En este caso se puede obtener un estadístico de prueba al considerar la distribución del coeficiente de correlación muestral r. Esta distribución es simétrica, como la distribución normal, solo cuando ρ es cero, cuando ρ es diferente de cero, la distribución de r es sesgada. Esta falta de normalidad no permite usar la forma general del estadistico de prueba, el cual tiene un estimador normalmente distribuido en el numerador y una estimación de la desviación estándar en el denominador. Sin embargo, a través de una transformación Z de Fisher, r se puede cambiar a un estadístico que sea aproximadamente normal. La ecuación para esta transformación es:

$$\frac{1}{2}\log_e \frac{1-r}{1-r} \qquad (5.1.12)$$

Esta cantidad tiene aproximadamente una distribución normal con media $1/2\log_e(1+\rho)/(1-\rho)$ y varianza $1/(n-3)$ cuando n es grande $n \geq 20$. Por lo tanto para probar la hipótesis H_0: $\rho = \rho_0$, $\rho_0 \neq 0$ se puede emplear el estadistico 5.1.13.

$$Z = \frac{\frac{1}{2}\log_e\frac{(1+r)}{(1-r)} - \frac{1}{2}\log_e\frac{(1+\rho_0)}{(1-\rho_0)}}{\frac{1}{\sqrt{n-3}}} \qquad (5.1.13)$$

Este estadístico de prueba tiene aproximadamente una distribución normal estándar y cuando se prueba $H_0: \rho = \rho_0$, $\rho_0 \neq 0$ se emplea una de las siguientes regiones criticas para el nivel de significancia α :

$Z \geq Z_{1-\alpha}$	Hipotesis alternativa	$H_A: \rho > \rho_0$	unilateral superior
$Z \leq -Z_{1-\alpha}$	Hipotesis alternativa	$H_A: \rho < \rho_0$	unilateral inferior
$\|Z\| \geq Z_{1-\alpha/2}$	Hipotesis alternativa	$H_A: \rho \neq \rho_0$	bilateral

donde $Z_{1-\alpha}$ denota el punto $100(1-\alpha)\%$ de la distribución normal estándar.

5.1.5. Intervalos de confianza para el coeficiente de correlación.

Un intervalo de confianza de $100(1-\alpha)\%$ para ρ se puede obtener mediante el empleo de la transformación Z de Fisher de la siguiente manera.

Primero, calcular un intervalo de confianza al $100(1-\alpha)\%$ para el parámetro $1/2\log_e(1+\rho)/(1-\rho)$ con la ecuación (5.1.14).

$$\frac{1}{2}\log_e\frac{1+r}{1-r} \pm Z_{1-\alpha/2}/\sqrt{n-3} \qquad (5.1.14)$$

donde previamente se define $Z_{1-\alpha/2}$.

Posteriormente se escribe el límite inferior del intervalo de confianza como L_I y el límite superior como L_S; así el intervalo de confianza es:

$$L_I < \frac{1}{2} Log_e \frac{1+\rho}{1-\rho} < L_S \qquad (5.1.15)$$

donde $$L_I = \frac{1}{2}\log_e \frac{1+\rho_I}{1-\rho_I} \qquad (5.1.15a)$$

y $$L_S = \frac{1}{2}\log_e \frac{1+\rho_S}{1-\rho_S} \qquad (5.1.15b)$$

Para transformar este intervalo de confianza, se determinan aquellos valores de ρ_I y ρ_S que satisfagan a los limites de confianza L_I y L_S, como se muestra a continuación.

$$L_I \Rightarrow \frac{1}{2}\log_e \frac{1+r}{1-r} - Z_{1-\alpha/2}/\sqrt{n-3}$$

$$L_I = \frac{1}{2}\log_e \frac{1+r_I}{1-r_I}$$

$$2L_I = \log_e \frac{1+r_I}{1-r_I}$$

$$e^{2L_I} = \frac{1+r_I}{1-r_I}$$

$$e^{2L_I}(1-r_I) = 1+r_I$$

$$e^{2L_I} - e^{2L_I} r_I = 1+r_I$$

$$e^{2L_I} - 1 = r_I + e^{2L_I} r_I$$

$$e^{2L_I} - 1 = r_I(1+e^{2L_I})$$

$$r_I = \frac{e^{2L_I}-1}{e^{2L_I}+1}$$

y

$$L_S \Rightarrow \frac{1}{2}\log_e \frac{1+r}{1-r} + Z_{1-\alpha/2}/\sqrt{n-3}$$

$$L_S = \frac{1}{2}\log_e \frac{1+r_S}{1-r_S}$$

$$2L_S = \log_e \frac{1+r_S}{1-r_S}$$

$$e^{2L_S} = \frac{1+r_S}{1-r_S}$$

$$e^{2L_S}(1-r_S) = 1+r_S$$

$$e^{2L_S} - e^{2L_S} r_S = 1+r_S$$

$$e^{2L_S} - 1 = e^{2L_S} r_S + r_S$$

$$e^{2L_S} - 1 = r_S(1+e^{2L_S})$$

$$r_S = \frac{e^{2L_S}-1}{e^{2L_S}+1}$$

Por lo tanto, el intervalo de confianza al $100(1-\alpha)\%$ es:

$$\rho_I < \rho < \rho_S \qquad (5.1.16a)$$

o $$r_I < r < r_S \qquad (5.1.16b)$$

5.2. COEFICIENTE DE DETERMINACIÓN.

5.2.1. Definición del coeficiente de determinación.

Para cuantificar la magnitud de la fuerza de relación lineal entre X e Y, primero se considera cual es el predictor de Y si no se emplean las X's; el mejor predictor en este caso puede ser simplemente $\overline{Y}$, la media muestral de las Y's, así la suma de cuadrados de las desviaciones asociada con la nueva variable de predicción la proporciona (5.2.1).

$$SCY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{5.2.1}$$

Pero si la variable X ayuda a predecir el comportamiento de la variable Y, la suma de cuadrados de los residuales es

$$SCE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{5.2.2}$$

la cual es considerablemente menor que SCY.

Si el modelo de mínimos cuadrados $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ ajusta mejor los datos que la línea horizontal $\hat{Y} = \overline{Y}$ (Fig. 5.2), el cuadrado del coeficiente de correlación muestral r proporciona una medida cuantitativa del mejoramiento al fijar una línea que considere a la variable X, el cual se puede escribir de la siguiente forma:

$$r^2 = \frac{SCY - SCE}{SCY} \tag{5.2.3}$$

Esta cantidad varia entre 0 y 1 porque r varia entre -1 y 1.

5.2.2. Interpretación del coeficiente de determinación.

En relación a la interpretación de r^2, primero se debe resaltar que la diferencia o reducción de SCY cuando se usa X en la predicción de Y se puede medir por (SCY - SCE) la cual siempre es positiva, además, la reducción proporcionada por SCY al usar X es dividida por SCY. De esta forma, r^2 mide la fuerza de la relación lineal entre X e Y ya que proporciona la reducción en la suma de cuadrados de las desviaciones verticales que se obtiene cuando se usa la línea de mínimos cuadrados $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ en relación al modelo $\hat{Y} = \bar{Y}$, que es predictor de Y sin tomar en cuenta a X. Por lo tanto, entre mayor sea r^2 es mayor la reducción en SCE en relación a $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, y más fuerte es la relación lineal entre X e Y.

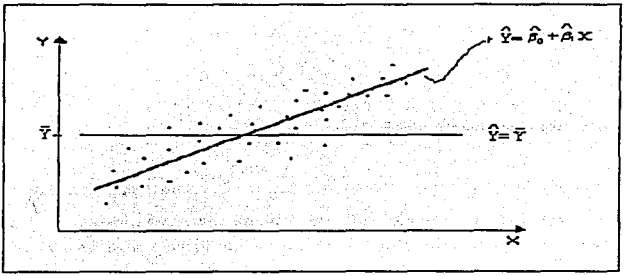


Figura 5.2. Predicción de Y usando X y sin usar X.

El valor más grande que puede alcanzar r^2 es 1, lo cual se logra cuando $\hat{\beta}_1$ es diferente de cero y cuando la SCE=0, lo que significa que hay una relación lineal perfecta entre X e Y, es decir, todos los puntos caen sobre la línea recta ajustada. En otras palabras, cuando $Y_i = \hat{Y}_i$ para toda i, se

puede tener que la línea de regresión se ajusta perfectamente a todos los puntos, por tanto los residuales serán igual a cero, esto es:

$$SCE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2 = 0 \qquad (5.2.4)$$

Así que $$r^2 = \frac{SCY - SCE}{SCY} = \frac{SCY}{SCY} = 1$$

A medida que el ajuste se hace más preciso, disminuye la variación en Y que está explicada por la relación con X (es decir, SCE decrece), lo cual significa que r^2 decrece también. El valor más pequeño de r^2 es cero, lo que significa que no hay mejora en la predicción al usar X; esto es, $SCE = SCY$, y se observa que un coeficiente de correlación cero implica una pendiente cero y por consecuencia ausencia de alguna relación lineal [16]. Por lo tanto el coeficiente de determinación también es una medida de la bondad del ajuste, que sirve para interpretar la cantidad relativa de la variación que explica la recta de regresión.

5.2.3. Relación con el coeficiente de correlación.

Existen muchas vinculaciones entre el análisis de regresión y el de correlación. Se puede observar, por ejemplo, la conexión existente entre el valor de r y el de la pendiente $\hat{\beta}_1$, comparando las ecuaciones:

$$\beta_1 = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

y $$r = \frac{S_{XY}}{S_X S_Y}$$

Debido a que $S_X^2 = S_X S_X$, sustituyendo se tiene que:

$$S_{XY} = \beta_1 S_X^2$$

y

$$r = \frac{\beta_1 S_X^2}{S_X S_Y} = \beta_1 \frac{S_X S_X}{S_X S_Y} = \beta_1 \frac{S_X}{S_Y} \qquad (5.2.5)$$

Puesto que S_X y S_Y nunca son negativas, el signo de r depende del signo de la pendiente. La ecuación (5.2.5) sólo es válida si el numerador es igual a β_1 multiplicado por la desviación estándar de la variable independiente.

Si se considera la relación entre el coeficiente de correlación r y el coeficiente de determinación, se tiene: $r^2 = \frac{SCR}{SCY} = \frac{SCY - SCE}{SCY} = \frac{S_{XY}^2}{S_X^2 S_Y^2}$ lo que es exactamente igual al cuadrado del coeficiente de correlación. Con esto se comprende la razón por la cual se usa la letra r para denotar el coeficiente de correlación y r^2 para representar el coeficiente de determinación. En la mayoría de los casos, r^2 es mucho más fácil de interpretar.

Para que r sea un estimador insesgado de ρ, la distribución conjunta de X e Y debe ser normal. Esto significa que el valor del coeficiente de correlación no depende de cuál sea la variable designada por X y cuál designada por Y. Sin embargo, esta distinción es importante en el análisis de regresión, pues la distribución condicional de Y, dada X, da lugar a una recta de regresión diferente de la que corresponde a la distribución condicional de X, dada Y. Por lo tanto en el análisis de regresión, la variable independiente X no tiene que ser necesariamente aleatoria, sino que puede ser fija [10].

5.2.4. Conceptos erróneos de r^2.

Existen dos conceptos erróneos acerca de r^2, que ocasionalmente conducen a interpretaciones falsas acerca de la relación entre X e Y los cuales son:

1.- El valor de r^2 no es una medida de la magnitud de la pendiente de la línea de regresión, esto es, si el valor de r^2 es alto, no necesariamente la magnitud de la pendiente $\hat{\beta}_1$ es grande. Esto se pude observar en la figura 5.3. Otra forma de observar esto es a través de la ecuación (5.1.9), donde $\hat{\beta}_1^2 = \frac{S_Y^2}{S_X^2}$ cuando $r^2=1$ en ambos casos, a pesar de que las pendientes son diferentes.

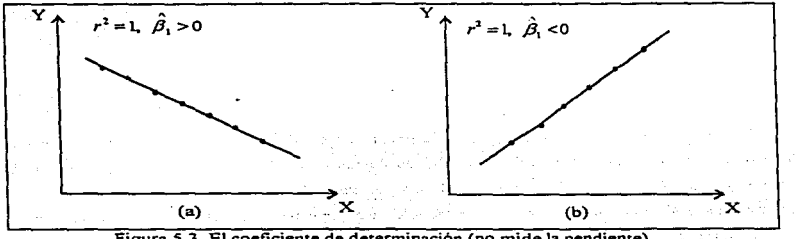


Figura 5.3. El coeficiente de determinación (no mide la pendiente).

Asi, si dos conjuntos de datos diferentes tienen la misma variación con respecto a X pero el primer conjunto tiene menos variación con respecto a Y que el segundo, la magnitud de la pendiente para el primer conjunto es más pequeña que para el segundo conjunto de datos.

2.- El valor de r^2 no es una medida de la conveniencia del modelo lineal, es decir r^2 puede ser igual a cero cuando no hay evidencia de alguna relación entre X e Y (Fig. 5.4a) ó cuando existe una fuerte evidencia de alguna asociación no lineal (Fig. 5.4b). Por otro lado, r^2 puede ser grande cuando un modelo lineal es completamente apropiado (Fig. 5.4c) ó cuando no es muy apropiado (Fig. 5.4d) [16].

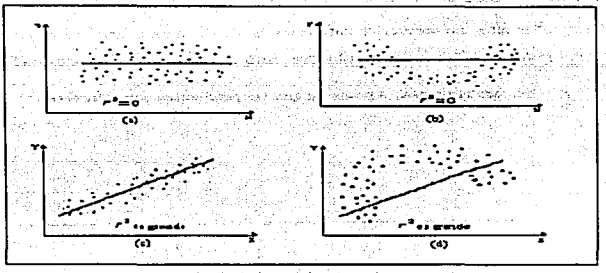


Figura 5.4. El coeficiente de correlación no mide la conveniencia

El coeficiente de determinación debe emplearse con cuidado porque siempre es posible hacerlo igual a uno al agregar el suficiente número de términos al modelo. Por ejemplo, es posible obtener un ajuste "perfecto" de n puntos al obtener un polinomio de grado n-1. Asimismo, r^2 siempre aumenta si se agrega una variable al modelo, aunque esto no significa necesariamente que el nuevo modelo sea mejor que el anterior.

A continuación se describe un estudio de caso del área farmacéutica donde se aplica la teoría del análisis de regresión lineal simple para encontrar un modelo que describe mejor los resultados experimentales en base a los criterios de decisión, el caso trata sobre un estudio de disolución de tabletas de furosemida, donde se quería encontrar la relación entre el tiempo de disolución y la cantidad de principio activo disuelto.

VI. ESTUDIO DE CASO DE REGRESIÓN LINEAL SIMPLE.

Se realizó un estudio para evaluar el efecto del grado de viscosidad de 4 tipos de carboximetilcelulosa sódica y las condiciones del mezclado (tiempo y velocidad) sobre el perfil de disolución de comprimidos de furosemida de liberación prolongada en matrices hidrofílicas, a través de un diseño experimental. Para ello, se tomaron al azar los resultados del lote 8 de diecinueve que se elaboraron para describir el porcentaje de disolución de furosemida en función del tiempo a través de un modelo de regresión lineal simple [15].

6.1.- PROBLEMA.

Uno de los objetivos del estudio fue determinar un modelo de regresión para describir el comportamiento del perfil de disolución de las tabletas de furosemida en función del tiempo. Para esto se realizó un estudio del perfil de disolución para cada lote por sextuplicado durante 8 horas (480 min.). La respuesta fue el porcentaje de furosemida disuelto y la variable independiente fue el tiempo. Se eligieron arbitrariamente los resultados del lote 8 para ejemplificar el cálculo de los parámetros de un modelo lineal mediante el uso del paquete estadístico SAS para Windows, así como también para la aplicación de la teoría descrita en capítulos anteriores.

En la literatura se reportan varios tipos de modelos que pueden describir el perfil de disolución de una tableta, entre los cuales se encuentran los siguientes:

Modelo de orden cero : $Disolución = kt$, en el cual disolución representa el porcentaje de principio activo disuelto a un tiempo t y k es la constante de velocidad de disolución de orden cero. Esta expresión tiene la forma de un modelo de análisis de regresión lineal simple $Y = \beta_0 + \beta_1 X$ e indica que la velocidad con que el sólido se disuelve en el medio de disolución es constante con el tiempo e independiente de la concentración del soluto. Al representar la cantidad que se disuelve a diferentes tiempos en función del tiempo, se obtiene una gráfica recta cuya pendiente es la constante de velocidad del proceso, como se indica en la figura 6.1.

En este modelo, la cantidad total de fármaco disuelto a tiempo infinito corresponde teóricamente a la cantidad que se agrega al medio de disolución al inicio. Sin embargo, no siempre es igual, ya que cuando se trata de una forma farmacéutica pueden ocurrir dos casos:

a) el principio activo no es cedido por completo a la solución por existir cierto grado de retención por parte de los excipientes.

b) si bien se conoce la cantidad teórica que lleva la forma farmacéutica, ésta en la práctica puede experimentar variaciones propias de la manufactura o de las maquinarias procesadoras u otras variables.

A pesar de que estas fluctuaciones suelen ser de pequeña importancia, en la expresión de los resultados pueden ser importantes, sobre todo si se quiere determinar con precisión los parámetros de disolución.

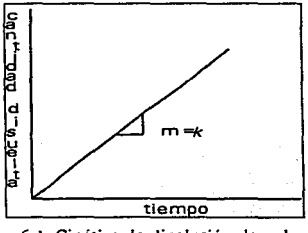


Figura 6.1. Cinética de disolución de orden cero

Por otro lado, cuando se trata de comprimidos farmacéuticos, la disolución suele comenzar de inmediato cuando el comprimido se pone en contacto con el líquido, sin embargo, a menudo los puntos correspondientes a los tiempos iniciales no resultan ser lineales debido a la desintegración retardada de los comprimidos, por lo que se recomienda no tomar en cuenta los valores de disolución en los primeros minutos cuando se busca ajustar un modelo lineal. A este tiempo inicial se le llama de latencia o de inducción, que representa el tiempo necesario para que la forma farmacéutica comience a ceder su principio activo [3].

También existe un modelo de la forma $Disolución = kt^n$, el cual se le conoce como modelo de Pephas y puede ser linealizado al obtener el logaritmo natural (ln) de ambos términos con lo que se

llega al siguiente modelo: $\ln Dis = \ln k + n \ln t$ que también tiene forma de lineal simple. En este caso, el valor del exponente o la pendiente del modelo linealizado indica el orden del perfil de disolución.

Así, el objetivo es determinar cual de los dos modelos ajusta mejor los resultados experimentales, lo cual se facilita con ayuda del paquete estadístico SAS, como se muestra a continuación.

Existen otros modelos, que también se evaluaron, pero se descartaron por no ajustar en forma adecuada los resultados experimentales del estudio de caso en cuestión.

6.2.- RESULTADOS EXPERIMENTALES.

Los resultados del porcentaje de disolución del lote 8 de las tabletas de furosemida se introducen al SAS como se muestra en el programa 6.1. Se comienza asignando un nombre al conjunto de datos, en este caso llamado *lote 8*; en la siguiente linea se indica en que orden se introducen los valores de las variables en estudio; debido a que también se necesita el logaritmo natural tanto del porcentaje de disolución como de los minutos, se deben crear nuevas variables llamadas *LMIN* y *LDIS;* posteriormente con la palabra *CARDS* se le indica a SAS que se comenzaran a escribir los valores de las variables en el orden indicado. Finalmente se escribe el procedimiento de imprimir (*Proc Print)*, para que muestre los valores de todas las variables, lo cual se observa en la salida 6.1, donde se presentan tanto valores normales como el logaritmo de minutos y disolución (variables transformadas). Estos resultados son los que se emplearon para ajustar un modelo de regresión adecuado.

Programa 6.1. Programa SAS para Imprimir datos en pantalla.

```
OPTIONS PS=60 NODATE;
DATA lote8;
INPUT lote vaso min disolu @@;
LMIN=Log(min);
LDIS=Log(disolu);
CARDS;
8 1  10  1.79     8 1  20   3.66     8 1  30   6.00     8 1  60 12.60
8 1 120 30.14     8 1 180  47.79     8 1 240  63.43     8 1 300 79.89
8 1 360 91.53     8 1 420  96.58     8 1 480  98.98     8 2  10  1.40
8 2  20  2.97     8 2  30   4.32     8 2  60   9.70     8 2 120 22.83
8 2 180 38.23     8 2 240  52.96     8 2 300  70.61     8 2 360 88.12
8 2 420 98.23     8 2 480 107.76     8 3  10   1.20     8 3  20  2.45
8 3  30  3.70     8 3  60   8.22     8 3 120  19.44     8 3 180 31.57
8 3 240 43.92     8 3 300  56.18     8 3 360  67.57     8 3 420 75.60
8 3 480 82.24     8 4  10   1.87     8 4  20   3.66     8 4  30  5.61
8 4  60 12.58     8 4 120  28.95     8 4 180  48.57     8 4 240 68.23
8 4 300 83.74     8 4 360  95.38     8 4 420  98.30     8 4 480 98.83
8 5  10  1.35     8 5  20   2.75     8 5  30   4.57     8 5  60 11.98
8 5 120 27.31     8 5 180  42.45     8 5 240  58.54     8 5 300 73.96
8 5 360 85.75     8 5 420  90.77     8 5 480  94.21     8 6  10  1.48
8 6  20  2.88     8 6  30   4.50     8 6  60  10.56     8 6 120 26.05
8 6 180 41.90     8 6 240  58.55     8 6 300  75.52     8 6 360 89.62
8 6 420 100.5     8 6 480 101.29
;
PROC PRINT;
TITLE1 'OBSERVACIONES DE DISOLUCION';
TITLE2 'LOTE 8';
RUN;
```

Salida 6.1. Salida SAS (PROC PRINT) del lote 8 de tabletas de Furosemida.

OBSERVACIONES DE DISOLUCION 1
'LOTE 8'

OBS	LOTE	VASO	MIN	DISOLU	LMIN	LDIS
1	8	1	10	1.79	2.30259	0.58222
2	8	1	20	3.66	2.99573	1.29746
3	8	1	30	6.00	3.40120	1.79176
4	8	1	60	12.60	4.09434	2.53370
5	8	1	120	30.14	4.78749	3.40585
6	8	1	180	47.79	5.19296	3.86682
7	8	1	240	63.43	5.48064	4.14994
8	8	1	300	79.89	5.70378	4.38065
9	8	1	360	91.53	5.88610	4.51667
10	8	1	420	96.58	6.04025	4.57037
11	8	1	480	98.98	6.17379	4.59492
12	8	2	10	1.40	2.30259	0.33647
13	8	2	20	2.97	2.99573	1.08856
14	8	2	30	4.32	3.40120	1.46326
15	8	2	60	9.70	4.09434	2.27213
16	8	2	120	22.83	4.78749	3.12808
17	8	2	180	38.23	5.19296	3.64362
18	8	2	240	52.96	5.48064	3.96954
19	8	2	300	70.61	5.70378	4.25717
20	8	2	360	88.12	5.88610	4.47870
21	8	2	420	98.23	6.04025	4.58731
22	8	2	480	107.76	6.17379	4.67991
23	8	3	10	1.20	2.30259	0.18232
24	8	3	20	2.45	2.99573	0.89609
25	8	3	30	3.70	3.40120	1.30833
26	8	3	60	8.22	4.09434	2.10657
27	8	3	120	19.44	4.78749	2.96733
28	8	3	180	31.57	5.19296	3.45221
29	8	3	240	43.92	5.48064	3.78237
30	8	3	300	56.18	5.70378	4.02856
31	8	3	360	67.57	5.88610	4.21316
32	8	3	420	75.60	6.04025	4.32546
33	8	3	480	82.24	6.17379	4.40964
34	8	4	10	1.87	2.30259	0.62594
35	8	4	20	3.66	2.99573	1.29746
36	8	4	30	5.61	3.40120	1.72455

Salida 6.1. Continuación

OBSERVACIONES DE DISOLUCION 2
'LOTE 8'

OBS	LOTE	VASO	MIN	DISOLU	LMIN	LDIS
37	8	4	60	12.58	4.09434	2.53211
38	8	4	120	28.95	4.78749	3.36557
39	8	4	180	48.57	5.19296	3.88301
40	8	4	240	68.23	5.48064	4.22288
41	8	4	300	83.74	5.70378	4.42772
42	8	4	360	95.38	5.88610	4.55787
43	8	4	420	98.30	6.04025	4.58802
44	8	4	480	98.83	6.17379	4.59340
45	8	5	10	1.35	2.30259	0.30010
46	8	5	20	2.75	2.99573	1.01160
47	8	5	30	4.57	3.40120	1.51951
48	8	5	60	11.98	4.09434	2.48324
49	8	5	120	27.31	4.78749	3.30725
50	8	5	180	42.45	5.19296	3.74833
51	8	5	240	58.54	5.48064	4.06971
52	8	5	300	73.96	5.70378	4.30352
53	8	5	360	85.75	5.88610	4.45144
54	8	5	420	90.77	6.04025	4.50833
55	8	5	480	94.21	6.17379	4.54553
56	8	6	10	1.48	2.30259	0.39204
57	8	6	20	2.88	2.99573	1.05779
58	8	6	30	4.50	3.40120	1.50408
59	8	6	60	10.56	4.09434	2.35707
60	8	6	120	26.05	4.78749	3.26002
61	8	6	180	41.90	5.19296	3.73529
62	8	6	240	58.55	5.48064	4.06988
63	8	6	300	75.52	5.70378	4.32440
64	8	6	360	89.62	5.88610	4.49558
65	8	6	420	100.50	6.04025	4.61016
66	8	6	480	101.29	6.17379	4.61799

6.3.- DIAGRAMAS DE DISPERSIÓN.

Como se mencionó en la sección 4.1, para obtener una información gráfica de la posible relación entre las variables de estudio, es conveniente observar los resultados de la salida 6.1. en un diagrama de dispersión. Para esto, se le indica a SAS que realice el gráfico a través de las indicaciones del programa 6.2, donde el *"Proc Plot"* representa un procedimiento para hacer una gráfica, y después se le indica lo que debe graficar en cada eje. Como se observa, en el programa 6.2 se indica que se quieren 2 gráficas, *DISOLU vs MIN* y *LDIS vs LMIN*.

Programa 6.2. Programación de SAS para obtener diagramas de dispersión.

```
OPTIONS NODATE;
DATA lote8;
INPUT lote vaso min disolu;
LMIN=Log(min);
LDIS=Log(disolu);
CARDS;
8 1  10   1.79
8 1  20   3.66
8 1  30   6.00
. .   .    .
. .   .    .
. .   .    .
8 6 360  89.62
8 6 420 100.50
8 6 480 101.29
;
PROC PLOT;
PLOT DISOLU*MIN/HAXIS= BY 50 VAXIS= BY 10;
PLOT LDIS*LMIN/HAXIS= BY 0.5 VAXIS=BY 0.5;
TITLE1 'DIAGRAMA DE DISPERSION';
TITLE2 'LOTE 8';
RUN;
```

De esta forma se obtiene la salida de SAS 6.2, donde se observa en el diagrama de dispersión DISOLU*MIN (salida 6.2-1) que conforme aumenta el tiempo de disolución, aumenta la dispersión de los resultados de porcentaje de disolución de la tableta de furosemida, por lo que la suposición de homocedasticidad no se cumple, a diferencia del diagrama de dispersión LDIS*LMIN (salida 6.2-2) en el cual todos los datos se mantienen con una variación relativamente constante. Esto se puede ratificar al calcular los parámetros estadístico básicos, como son la media, desviación estándar, varianza y coeficiente de variación en cada punto de la variable independiente, lo cual se realiza a través de un procedimiento UNIVARIATE PLOT con lo que además de calcular las variaciones en

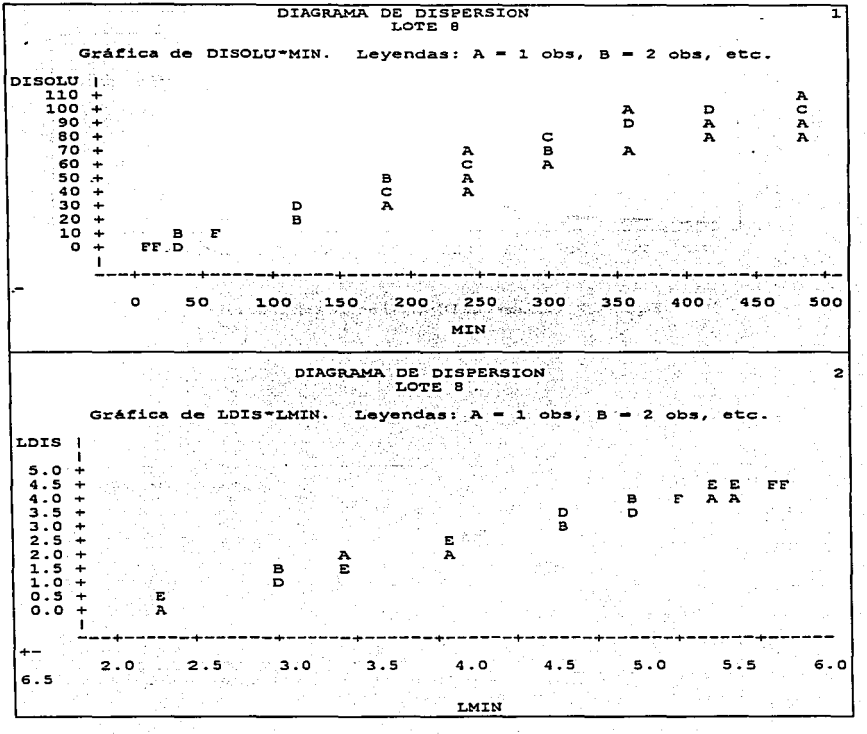
cada punto, se obtiene la forma de la distribución de la variable de respuesta en cada nivel de X, por lo que se puede observar si se cumplen o no los supuestos estadísticos básicos de regresión lineal, descritos en la sección 4.2. Sin embargo ambos diagramas de la salida 6.2 muestran que existe una relación entre las dos variables en estudio ya que la variable independiente aumenta al incrementar el valor de la variable independiente.

Para obtener los estadísticos básicos de los resultados experimentales se emplea el programa 6.3, con el cual se obtiene la salida 6.3, la cual se encuentra en el anexo A, donde se presentan los resultados necesarios para construir las figuras 6.2. y 6.3.

Programa 6.3. Programación de SAS para obtener estadísticos básicos.

```
OPTIONS NODATE;
DATA lote8;
INPUT lote vaso min disolu;
LMIN=Log(min);
LDIS=Log(disolu);
CARDS;
8 1  10   1.79
8 1  20   3.66
8 1  30   6.00
. .   .    .
. .   .    .
. .   .    .
8 6 360  89.62
8 6 420 100.50
8 6 480 101.29
;
PROC SORT;
BY MIN;
PROC UNIVARIATE PLOT;
VAR DISOLU;
BY MIN;
PROC SORT;
BY LMIN;
PROC UNIVARIATE PLOT;
VAR LDIS;
BY MIN;
TITLE1 'OBSERVACIONES DE DISOLUCION';
TITLE2 'DISPERSION EN CADA NIVEL';
RUN;
```

Salida 6.2. Diagramas de Dispersión.



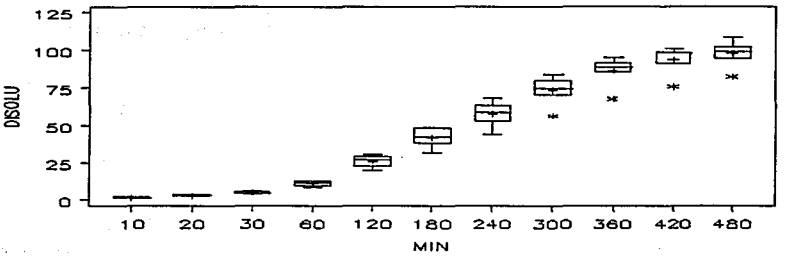


Figura 6.2. Varianzas de DISOLU en cada nivel de MIN

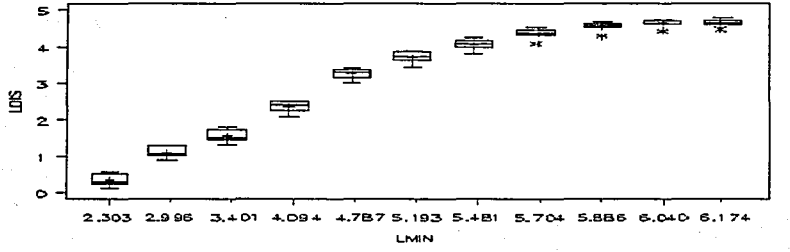


Figura 6.3. Varianzas de LDIS en cada nivel de LMIN

De la salida 6.3 del proc univariate plot (anexo-A) se observa que en cada nivel de la variable independiente existe una distribución aproximadamente normal y se observa mayor aproximación cuando las variables se transforman a su logaritmo. También se observa un comportamiento lineal en las figuras 6.2 y 6.3, sin embargo, en la primera se observa un comportamiento aproximadamente lineal a partir del minuto 30, pero conforme aumenta el tiempo, aumenta la dispersión de los resultados del porcentaje de disolución, es decir, la suposición de homocedasticidad no se cumple, a diferencia de lo que se observa en la figura 6.3, donde el comportamiento lineal comienza desde los primeros minutos hasta un tiempo de 300 minutos aproximadamente (*LMIN* = 5.704), y también se observa que la variación en cada nivel de la variable independiente es constante, lo cual queda representado por los diagramas de caja en cada nivel de X, por lo que se puede decir que si se cumplen los supuestos de homocedasticidad y linealidad, por lo tanto se puede comenzar a ajustar un modelo de regresión lineal simple.

A continuación se ejemplifica el cálculo de los parámetros de un modelo de regresión con ayuda del paquete estadistico SAS, tanto para las variables no transformadas (unidades reales) como para las variables transformadas.

6.4.- CALCULO DE COEFICIENTES DE REGRESIÓN.

En esta sección se ejemplifica el cálculo de los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ del modelo de regresión 2, $Disolu = \beta_0 + \beta_1 t$, para el caso de estudio, descrito en la sección 6.1., sin tomar en cuenta los resultados del porcentaje de disolución en los minutos 10 y 20, es decir se eliminan 12 observaciones por lo que n=54. Para esto se tiene que:

X= Tiempo (t) en minutos.
Y= Disolución (Disolu).

$$\sum_{i=1}^{n} X_i Y_i = 985028.4$$

$$\sum_{i=1}^{n} X_i = 13140$$

$$\sum_{i=1}^{n} Y_i = 2946.36$$

$$\sum_{i=1}^{n} X_i^2 = 4411800$$

$$n = 54$$

Por lo tanto, al sustituir en la ecuación (4.4.23) se obtiene:

$$\hat{\beta}_1 = \frac{985028.4 - \frac{(13140)(2946.36)}{54}}{4411800 - \frac{(13140)^2}{54}} = 0.220751646$$

Para el calculo del intercepto se tiene:

$$\overline{Y} = 54.5622$$
$$\overline{X} = 243.3333$$
$$\hat{\beta}_1 = 0.220751646$$

y al sustituir en la ecuación (4.4.22) se obtiene:

$$\hat{\beta}_0 = 54.5622 - 0.220751646(243.3333) = 0.845988$$

De esta forma se puede proponer el siguiente modelo de regresión lineal simple:

$$\hat{Y} = 0.845988 + 0.220752\,X$$

donde $\hat{Y}$ es el porcentaje de disolución predicho por el modelo y X es el tiempo en minutos, sin embargo no se puede asegurar que es el mejor modelo, ya que aun no se analizan los parámetros importantes como CME, r^2 y residuales.

6.5.- USO DE SAS PARA AJUSTAR UN MODELO LINEAL SIMPLE.

El cálculo de los coeficientes del modelo de regresión se puede hacer con la ayuda del SAS, con el cual se obtienen los resultados acerca del modelo de regresión lineal.

Para los datos del caso de estudio de la sección 6.1. se pueden obtener los modelos propuestos a continuación a través del análisis de regresión, incluyendo el modelo descrito en la sección 6.4. Para obtener el modelo $Disolu = \beta_0 + \beta_1 t$ sin eliminar ninguna observación se debe especificar el modelo en la programación de SAS, como se indica en la línea `MODEL1` del programa

6.4. Sin embargo para obtener el mismo modelo, pero sin las observaciones de disolución al minuto 10 y 20 se vuelve a especificar el modelo y se le indica cuales observaciones debe eliminar con un REWEIGHT, como se muestra en la línea MODEL2 del programa 6.4.

Por otro lado, para obtener el modelo $LnDIS = LnK + nLnt$, se indica dicho modelo al programar SAS como se muestra en el programa 6.4 en la línea MODEL3 y se le ordena un REWEIGHT UNDO para deshacer la última acción "REWEIGHT", en el que se indica eliminar las observaciones de los minutos 10 y 20.

Programa 6.4. Programación de SAS para obtener modelos de Regresión Lineal Simple.

```
OPTIONS PS= 60 NODATE;
DATA lote8;
INPUT lote vaso min disolu;
LMIN=Log(min);
LDIS=Log(disolu);
CARDS;
8 1  10    1.79
8 1  20    3.66
8 1  30    6.00
. .   .      .
. .   .      .
. .   .      .
8 6 360   89.62
8 6 420  100.50
8 6 480  101.29
;
PROC REG;
TITLE1 "ANALISIS DE REGRESION SIMPLE";
TITLE2 "LOTE 8";
MODEL1 DISOLU=MIN;
MODEL2 DISOLU=MIN;
         REWEIGHT MIN<=20;
MODEL3 LDIS=LMIN;
         REWEIGHT UNDO;
RUN;
```

Con el Programa 6.4. se obtiene la salida 6.4. en la cual se describen las características de los modelos de regresión lineal simple que se proponen para representar la disolución de las tabletas de furosemida en función del tiempo expresado en minutos.

Salida 6.4. Análisis de Regresión para tres modelos diferentes.

```
                    ANALISIS DE REGRESION SIMPLE                    1
                              LOTE 8
Modelo: MODEL1
Variable Depend: DISOLU
                      Análisis de Varianza

Fuente de                    Suma de          Cuadrado          Valor
Variación          GL        Cuadrados        Medio             de F
Prob>F

Modelo              1      85989.68733      85989.68733       1516.310
0.0001
Error              64       3629.43017         56.70985
C Total            65      89619.11750

        Raiz CME            7.53059      R-cuadrada          0.9595
        Prom Dep           45.05788      R-cd ajust          0.9589
        C.V.               16.71316

                    Parámetros Estimados

                    Parámetros       Error          T para HO:
Variable    GL      Estimados        Estándar       Parámetro=0     Prob>|T|

INTERCEP    1       0.018759         1.48224230       0.013         0.9899
MIN         1       0.223167         0.00573107      38.940         0.0001

                    ANALISIS DE REGRESION SIMPLE                    2
                              LOTE 8
Modelo: MODEL2
Variable Depend: DISOLU
                      Análisis de Varianza

Fuente de                    Suma de          Cuadrado          Valor
Variación          GL        Cuadrados        Medio             de F
Prob>F

Modelo              1      59179.27810      59179.27810       854.260
0.0001
Error              52       3602.32603         69.27550
C Total            53      62781.60413

        Raiz CME            8.32319      R-cuadrada          0.9426
        Prom Dep           54.56222      R-cd ajust          0.9415
        C.V.               15.25449

                    Parámetros Estimados

                    Parámetros       Error          T para HO:
Variable    GL      Estimados        Estándar       Parameter=0
Prob>|T|

INTERCEP    1       0.845988         2.15883766       0.392
0.6968
MIN         1       0.220752         0.00755292      29.228
0.0001
```

Salida 6.4. Continuación

```
                    ANALISIS DE REGRESION SIMPLE                          3
                              LOTE 8

Modelo: MODEL3
Variable Depend: LDIS

                        Análisis de Varianza

Fuente de                    Suma de          Cuadrado          Valor
Variación          GL       Cuadrados          Medio            de F
Prob>F

Modelo              1      137.05033        137.05033        5433.655
0.0001
Error              64        1.61424          0.02522
C Total            65      138.66457

        Raiz CME          0.15882      R-cuadrada          0.9884
        Prom Dep          3.11755      R-cd ajust          0.9882
        C.V.              5.09425

                      Parámetros Estimados

                    Parámetros          Error          T para H0:
Variable     GL      Estimados        Estándar       Parámetro=0     Prob>|T|

INTERCEP      1      -2.223076      0.07504235           -29.624       0.0001
LMIN          1       1.128471      0.01530891            73.713       0.0001
```

Se observa que los resultados del modelo de regresión calculados en la sección 6.4 corresponden a los que se encuentran en la salida 6.4-2, es decir, al modelo 2, además se obtienen otros parámetros adicionales para cada modelo, los cuales ayudan a determinar el modelo que mejor ajusta los resultados experimentales.

En la salida de SAS se obtiene primero el resultado del análisis de varianza, con el cual se pueden obtener parámetros importantes del modelo, tales como la desviación estándar, varianza y coeficiente de determinación, así como también se prueba la hipótesis nula Ho: "No existe relación lineal significativa entre las variables en estudio".

Se observa en la salida 6.4 que el modelo 3 presenta un coeficiente de determinación mayor, así como el valor de F más grande, sin embargo, también el valor del coeficiente de determinación para el modelo 1 es grande por lo que se podría decir que sí es un buen modelo pero no es el único parámetro que se debe tomar en cuenta para decidir si un modelo se ajusta o no a los datos experimentales por lo que a continuación se hace una discusión para decidir cual es el modelo que mejor ajusta a los resultados del porciento de disolución de las tabletas de Furosemida.

Como se mencionó en la sección 4.5, para comenzar se tiene que evaluar uno de los parámetros que miden la calidad de la linea recta, el cual es la suma de cuadrados del error (SCE) y debe tener un valor pequeño. El cálculo de la suma de cuadrados del error en base al modelo 2 (sin las observaciones de los minutos 10 y 20) propuesto para el caso de estudio es:

$$SCE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = 3602.32603$$

este valor se obtiene en la salida 6.4. de SAS. Como se observa este valor resulto ser muy grande para los modelo 1 y 2, a diferencia del valor que se obtiene en el modelo 3 que fue de 1.61424, sin embargo una mejor medida de la dispersión o grado de ajuste una vez que se tiene el valor de SCE es la varianza del modelo, $S^2_{Y/X}$, así por ejemplo, para el modelo 2, el cual es $Disolu = \beta_0 + \beta_1 t$ sin tomar en cuenta los minutos 10 y 20, se puede obtener $S^2_{Y/X}$ a partir de la ecuación (4.5.2), que se encuentra en la sección 4.5:

$$S^2_{Y/X} = \frac{1}{52}(3602.32603) = 69.2755$$

Este valor se encuentra en la salida de SAS 6.4. y representa el Cuadrado Medio del Error (CME) y también se observa que el valor más pequeño lo tiene el modelo 3, por lo que hasta este punto es el modelo que tiene menor dispersión al ajustar los resultados experimentales.

También se puede obtener a través de la ecuación 4.5.3 la estimación de la varianza, por ejemplo para el mismo modelo 2 se tiene:

De (4.5.3a):

$$\sum_{i=1}^{n} Y_i^2 = 223541.5532$$

$$\sum_{i=1}^{n} Y_i = 2946.36$$

$$n = 54$$

$$S_Y^2 = \frac{223541.5532 - \frac{(2946.36)^2}{54}}{53} = 1184.5585$$

y de (4.5.3b) se tiene:

$$\sum_{i=1}^{n} X_i^2 = 4411800$$

$$\sum_{i=1}^{n} X_i = 13140$$

$$n = 54$$

$$S_X^2 = \frac{4411800 - \frac{(13140)^2}{54}}{53} = 22913.20755$$

que al sustituir en (4.5.3.) se obtiene:

$$S_{Y/X}^2 = \frac{53}{52}\left(1184.5585 - (0.220752)^2 (22913.20755)\right) = 69.2755$$

el cual es el mismo valor calculado anteriormente.

En base al cálculo de la varianza se puede obtener el valor numérico de la desviación estándar al obtener la raiz cuadrada de la varianza como se muestra a continuación:

$$S_{Y/X} = \sqrt{69.2755} = 8.3231905$$

Este valor también aparece en la Salida de SAS 6.4. con el nombre de Raíz del Cuadrado Medio del Error (CME) y por consecuencia el modelo 3 es quien tiene el menor valor.

Con la estimación de la desviación estándar se puede obtener el coeficiente de variación para el modelo de la siguiente manera:

$$C.V. = \frac{S_{Y|X}}{\bar{Y}} * 100$$

Asi, para el modelo 2 se tiene:

$$C.V. = \frac{8.3231905}{54.56222} \times 100 = 15.25449$$

el cual también aparece en la salida 6.4 de SAS.

En resumen, de la salida 6.4 de SAS se observa que para los modelo 1 y 2 es mucho mayor la varianza, desviación estándar y el coeficiente de variación que para el modelo 3, lo cual implica que en el modelo $Disolu = \beta_0 + \beta_1 t$ existe mayor variación en la predicción de los resultados, es decir, hay más variabilidad en los datos lo cual aumenta el grado de error en el modelo, a diferencia del modelo 3, $\ln Dis = \ln k + n \ln t$, en el cual el coeficiente de variación es menor, por lo que se puede decir que la media poblacional del logaritmo del porciento de disolución a un determinado valor del logaritmo de tiempo será muy aproximado al valor real. Sin embargo se debe tener precaución en esta interpretación ya que se estan comparando modelos de regresión en diferentes escalas, no obstante se ha determinado que el mejor modelo es el de variables transformadas debido a que posteriormente se calculó el valor predicho y se transformó a la misma dimensión, con lo que el residual que se obtiene es menor que el producido por el modelo original.

Por otro lado, en los resultados de SAS se obtiene un cuadro de análisis de varianza, en el cual se resume toda la información del análisis de regresión, como se describió en la sección 4.6. En este cuadro se incluye una prueba de hipótesis que evalúa si existe relación lineal o no entre las variables.

Por ejemplo para el modelo1 del caso de estudio, el estadístico $F = CMR/CME$ que aparece en la salida 6.4. se obtiene a través del CMR=85989.68733 y CME=56.70985 por lo tanto F=1516.310 donde se observa que el valor del CMR es muy elevado en relación con el valor del CME, lo cual indica que una buena parte de la variabilidad en la respuesta es explicada por la recta de regresión, es decir, se debe rechazar la hipótesis nula. Si el valor del CMR fuera pequeño en relación al CME, ello indicaria que la recta de regresión no explica mucho la variabilidad existente en los valores de Y, razón por la cual no se debe rechazar la hipótesis nula.

Para probar la hipótesis nula *Ho: No existe relación lineal significativa de Y sobre X vs Ha: Existe relación lineal entre X e Y* tomando en cuenta los resultados del modelo 1 de la salida 6.4, se tiene que F_c=1516.310 y $F_{(1,52,0.05)} = 4.00$, y como la regla de decisión consiste en rechazar H_0 (es decir, no rechazar que la recta de regresión contribuye a explicar la variabilidad en Y) si el valor de F que se calcula en base a la muestra es mayor de 4.00, se rechaza la hipótesis nula y se llega a la conclusión de que el modelo de regresión lineal ayuda a explicar la variación en el porcentaje de disolución en función del tiempo. En la salida 6.4. se observa el parámetro $Prob>F$ que es el valor de probabilidad de F el cual evita recurrir a valores de tablas, ya que para su interpretación se debe elegir el nivel de significancia con el que se quiere trabajar y con ese valor se compara $Prob>F$, así por ejemplo, si $\alpha = 0.05$, basta que el valor de $Prob>F$ sea menor a 0.05 para que se encuentre en la zona de rechazo de la hipótesis nula. En el caso del modelo 1 se tiene un valor de $Prob>F = 0.0001$, con lo que se concluye que existe relación lineal entre X e Y. Este criterio de decisión se muestra en la figura 6.4.

Lo mismo sucede para los modelos 2 y 3 de la salida 6.4., por lo que se puede suponer que los tres modelos lineales son adecuados para explicar la Disolución de las tabletas de Furosemida en función del Tiempo, sin embargo se debe realizar un análisis más profundo para llegar al mejor modelo de regresión.

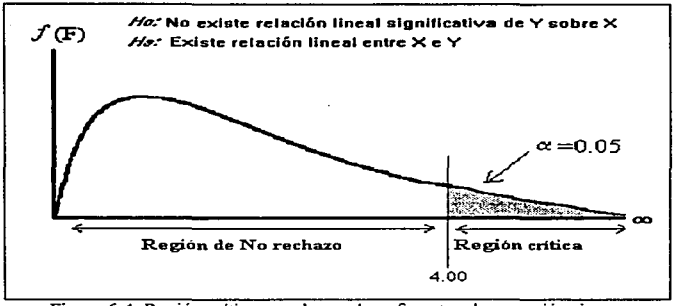


Figura 6.4. Región crítica para la prueba referente a la regresión tiempo-disolución efectuada mediante el análisis de varianza.

Se puede comprobar que la prueba de F es comparable con la prueba de t referente a la pendiente, como se mencionó en la sección 4.6, así para el modelo 1 de la salida 6.4 el valor de t calculado es t_c=38.940, de modo que $t_c^2 = (38.940)^2 = 1516.3236$, valor que sólo difiere de F_C = 1516.310 por errores de redondeo. Así se puede establecer que el punto $100(1-\alpha)\%$ de la distribución F es el mismo que el cuadrado del punto $100(1-\alpha)\%$ de la distribución t con los mismos grados de libertad.

6.6. PRUEBA DE LA PENDIENTE E INTERCEPTO.

Una vez que se estima la línea recta por mínimos cuadrados, se deben realizar inferencias acerca de la pendiente y el intercepto, como se describió en la sección 4.7, para asegurar que el modelo ajustado ayuda a predecir el valor de la respuesta. Por ejemplo, para el estudio de caso de la sección 6.1. y para el modelo lineal 2, que es $Disolu = \beta_0 + \beta_1 t$ sin tomar en cuenta las observaciones de los minutos 10 y 20, se calculan los estadísticos t para hacer una prueba de hipótesis tanto de la pendiente como de la ordenada al origen.

Para la prueba de pendiente cero, de la ecuación (4.7.1) se tiene:

$$\hat{\beta}_1 = 0.220752$$
$$\beta_1^{(0)} = 0.0$$
$$S_{Y/X} = 8.32319$$
$$S_X = 151.37109$$

$$t_c = \frac{0.220752 - 0.0}{\frac{8.32319}{151.37109\sqrt{54-1}}} = 29.22777$$

y para la prueba del intercepto cero se tiene de la ecuación (4.7.2) lo siguiente:

$$\hat{\beta}_0 = 0.845988$$
$$\beta_0^{(0)} = 0.0$$
$$\overline{X} = 243.3333$$
$$S_X^2 = 22913.20755$$

$$t_c = \frac{0.845988 - 0.0}{8.32319\sqrt{\frac{1}{54} + \frac{(243.3333)^2}{(54-1)22913.20755}}} = 0.391872$$

Para probar la hipótesis de la pendiente cero del modelo 2 se obtiene un valor de $t = 29.228$, y como es mayor a $t_{32,1-0.05/2} = 2.01$ se rechaza Ho y se puede concluir que la pendiente del modelo es diferente de cero, y en base a los parámetros evaluados anteriormente se puede decir que X provee información significativa para la predicción de Y, es decir, el modelo lineal es adecuado para estimar los valores de $\hat{Y}_i$ si se conocen valores de X_i. Sin embargo, de la salida 6.4 se observa que el valor más grande de t lo tiene el modelo 3, y se ha comprobado que este modelo es hasta ahora el que tiene menor variación en la predicción, es decir, es mejor modelo que los dos primeros.

Por otro lado, en cuanto a la prueba del intercepto cero, para el modelo 2 se tiene un valor de $t=0.392$ el cual es menor a $t_{32,1-0.05/2} = 2.021$ por lo que no se rechaza Ho, de tal forma se puede concluir que el intercepto es igual a cero, por lo tanto, es posible eliminar la constante $\hat{\beta}_0$ del modelo ya que se aporta evidencia estadística para sugerir que la línea pasa a través del origen.

Estos estadísticos aparecen en la salida 6.4. de SAS, así para los tres modelos, los valores del estadístico t para probar las hipótesis del intercepto y pendiente cero son los siguientes:

Cuadro 6.1. Valores de t para pruebas de hipótesis.

Modelo	Valor de t para intercepto cero / Prob > T	Valor de t para pendiente cero / Prob > T
1	0.013 / 0.9899	38.940 / 0.0001
2	0.392 /0.6968	29.228 / 0.0001
3	-29.624 /0.0001	73.713 /0.0001

Se observa para todos los modelos, que el valor de t para la prueba de hipótesis de pendiente cero es grande, y su valor de probabilidad es menor de 0.05 por lo que se puede concluir que el valor de la pendiente es diferente de cero en todos los casos, es decir, existe evidencia estadística que la variable independiente ayuda a predecir los valores de la variable dependiente.

Por otro lado, los valores de t para probar la hipótesis del intercepto cero, en los dos primeros casos proporcionan evidencia estadística de que el valor del intercepto es cero ya que el valor de t es pequeño y el valor de probabilidad mayor de 0.05, a diferencia del modelo 3, en el cual el valor de t es grande y la probabilidad menor de 0.05, por lo que se concluye que en este modelo no se debe eliminar el intercepto ya que es diferente de cero.

En base a los resultados de la prueba de hipótesis acerca del intercepto cero se puede proponer eliminar la constante β_0 de los dos primeros modelos, es decir, se tiene que reajustar dichos modelos, lo cual se puede realizar desde el paquete estadístico con el programa 6.5., donde se escribe la opción NOINT en el renglón que especifica el modelo para indicar a SAS que no incluya el intercepto en el modelo.

Programa 6.5. Programación de SAS para obtener modelos de Regresión Lineal Simple sin Intercepto.

```
OPTIONS NODATE;
DATA lote8;
INPUT lote vaso min disolu;
CARDS;
8 1   10    1.79
8 1   20    3.66
8 1   30    6.00
. .    .      .
. .    .      .
. .    .      .
8 6 360   89.62
8 6 420 100.50
8 6 480 101.29
;
PROC REG;
TITLE1 "ANALISIS DE REGRESION SIMPLE";
TITLE2 "SIN INTERCEPTO EN EL MODELO";
MODEL1 DISOLU=MIN/ NOINT;
MODEL2 DISOLU=MIN / NOINT;
          REWEIGHT MIN<=20;
RUN;
```

Así se obtiene la salida 6.5 de SAS, en la que se observa que hay un reajuste de los modelos, en los cuales aumenta el valor de R^2 y también aumenta el valor de F del modelo por lo que lo hace mucho más probable, sin embargo no se observa prácticamente ningún cambio en la desviación estándar del modelo ($S_{Y/X}$ ó Raíz del CME), ni en el coeficiente de variación.

Se observa que el modelo 3 de la salida 6.4. aún mantiene el valor más grande del coeficiente de determinación, el valor más pequeño del coeficiente de variación y el mayor valor del estadístico F para el modelo, por lo que aún es el mejor modelo para representar el porcentaje de disolución de las tabletas de Furosemida en función del tiempo.

En la salida 6.5. se observa que el valor de los coeficiente de regresión de los modelos prácticamente no cambian en relación a los valores que se obtienen en la salida 6.4., sin embargo los valores de t para la prueba de hipótesis acerca de la pendiente cero aumentan, en el modelo 1 aumenta de 38.940 a 62.767 y para el modelo 2 aumenta de 29.228 a 56.800 por lo que aumenta también la probabilidad de estos coeficiente. Pero el valor de t para el coeficiente de regresión del modelo 3 de la salida 6.4 aún es mayor (73.713) por lo que también es mayor su probabilidad que la de los coeficientes de los modelos de la salida 6.5.

Por lo anterior se puede decir que un buen modelo lineal para representar la disolución de las tabletas de Furosemida, en función del tiempo es:

$$\ln Dis = \ln k + n \ln t$$

Donde: $\ln K = -2.223076$

$n = 1.128471$ (cinética de la disolución)

es decir, $$\ln Dis = -2.223076 + 1.128471 \ln t$$

Salida 6.5. Reajuste de modelos sin intercepto

```
                    ANALISIS DE REGRESION SIMPLE                        1
                      SIN INTERCEPTO EN EL MODELO

Modelo: MODEL1
NOTA: No intercepto en el modelo. R-cuadrada es redefinida.
Variable Dependiente: DISOLU

                          Análisis de Varianza

Fuente de                    Suma de          Cuadrado          Valor
Variación          GL       Cuadrados          Medio            de F      Prob>F

Modelo              1    219983.69935    219983.69935        3939.711    0.0001
Error              65      3629.43925        55.83753
U Total            66    223613.13860

      Raiz CME          7.47245      R-cuadrada          0.9838
      Prom Dep         45.05788      R-cd ajust          0.9835
      C.V.             16.58412

                       Parámetros Estimados

                    Parámetro          Error          T para H0:
   Variable   GL    Estimado          Estándar       Parámetro=0      Prob>|T|

   MIN         1    0.223223       0.00355638          62.767          0.0001
```

```
                    ANALISIS DE REGRESION SIMPLE                        2
                      SIN INTERCEPTO EN EL MODELO

Modelo: MODEL2
NOTA: No intercepto en el modelo. R-cuadrada es redefinida.
Variable Dependiente: DISOLU

                          Análisis de Varianza

Fuente de                    Suma de          Cuadrado          Valor
Variación          GL       Cuadrados          Medio            de F      Prob>F

Modelo              1    219928.58897    219928.58897        3226.219    0.0001
Error              53      3612.96423        68.16914
U Total            54    223541.55320

      Raiz CME          8.25646      R-cuadrada          0.9838
      Prom Dep         54.56222      R-cd ajust          0.9835
      C.V.             15.13219

                       Parámetros Estimados

                    Parámetro          Error          T para H0:
   Variable   GL    Estimado          Estándar       Parámetro=0      Prob>|T|

   MIN         1    0.223271       0.00393085          56.800          0.0001
```

6.7. INTERVALOS DE CONFIANZA Y DE PREDICCIÓN.

Como se explicó en la sección 4.8, existen dos clases de estimaciones importantes a partir de un modelo de regresión, éstos son el valor medio de Y dado un valor de X ($\mu_{Y|X}$) y el valor real de Y correspondiente a un valor dado de X, así por ejemplo, si en el caso del modelo 3 de la salida 6.4 se desea obtener la predicción correspondiente al porcentaje de disolución del minuto 60, es decir LMIN=4.09434, se emplean los coeficientes estimados $\hat{\beta}_0 = -2.223070$ y $\hat{\beta}_1 = 1.128471$, a partir de los cuales se obtiene:

$$LDIS_{60} = \hat{Y}_{60} = -2.223070 + 1.128471(4.09434)$$
$$LDIS_{60} = 2.397274$$
$$DISOLU_{60} = 10.993167\%$$

Así, la mejor estimación del porcentaje de disolución de las tabletas de furosemida del lote 8 al minuto 60 es $\hat{Y}_{60} = 10.993167$. Análogamente, la estimación del porcentaje de disolución medio de todas las tabletas cuyo tiempo de disolución sea igual a 60 min es también $\hat{Y}_{60} = 10.993167$.

Esto se puede verificar al obtener el residual en este punto, entre el valor promedio observado y el valor predicho por el modelo, como se muestra a continuación:

Con variables transformadas:

$LDIS_{60\,(observación)}$ =2.3808 $\qquad$ $LDIS_{60\,(predicción)}$ =2.3972 $\qquad$ *Residual* = -0.01647

haciendo conversión a variables originales:

$DISOLU_{60\,(observacion)}$ =10.94 $\qquad$ $DISOLU_{60\,(predicción)}$ =10.999317 $\qquad$ *Residual* = -0.05316

Como se observa, se tiene un residual muy pequeño, esto significa que el modelo tiene poco error al predecir los valores de *LDIS* en función de *LMIN*. Así como también se observa que al convertir los valores en unidades reales, se obtiene un pequeño error en la predicción. A continuación se obtienen las predicciones para todos los niveles de la variable independiente y sus correspondientes intervalos a través de SAS.

En la sección 4.8. se menciona que existen dos tipos de intervalos, uno para la predicción de un valor aislado de la variable independiente, llamado intervalo de predicción; y un intervalo de confianza para el valor medio de la variable de respuesta en cada nivel de la variable independiente. Este último será más pequeño, debido a que tendrá un error estándar menor.

Así por ejemplo, para construir un intervalo de predicción de 95% para el porcentaje de disolución de las tabletas de furosemida al minuto 60, se tiene que:

$t_{(64,0.025)} = 2.021$

$LMIN_{60} = 4.09434$

$LDIS_{60} = \hat{Y}_{60} = 2.397274 \quad (Disolu = 10.9932)$

$S_{Y/X} = 0.15882$

Al sustituir estos valores en la ecuación (4.8.2), se obtienen los siguientes extremos del intervalo de predicción:

$$2.397274 \pm 2.021(0.15882)\sqrt{1+\frac{1}{66}+\left(\frac{(4.09434-4.732629)^2}{107.621435}\right)}$$

$$2.397274 \pm 0.3209752\sqrt{1.009220689}$$

$$2.397274 \pm 0.3225$$

$$\therefore \quad I.P._{X=4.06434} = 2.074774 - 2.719774$$

y convertidos a valores reales es: $I.P._{X=60} = 7.96275 \quad - \quad 15.17689$

Empleando este método se puede esperar que el intervalo de predicción determinado incluya al valor verdadero del logaritmo natural del porcentaje de disolución el 95% de las veces.

Por otro lado, se puede encontrar el intervalo de confianza para el mismo valor de la variable independiente, de la siguiente manera:

Al sustituir los valores adecuados en la ecuación (4.8.6) y usando la ecuación (4.8.5) para determinar S_{Y/X_o}, se obtienen los extremos del intervalo de confianza para t=60 min.

$\hat{Y}_{4.09434} = 2.397274$

$t_{(64,0.025)} = 2.021$

$S_{Y|X} = 0.15882$

$n = 66$

$$I.C. = 2.397274 \pm (2.021)(0.15882)\sqrt{\frac{1}{66} + \frac{(4.09434 - 4.732624)^2}{107.621435}}$$

$I.C. = 2.397274 \pm 0.04417$

$I.C. =$ De 2.353104 a 2.441444

y en valores reales: $I.C._{X=60} =$ De 10.5182 a 11.4896

Tal como se espera, el intervalo para $\mu_{Y|X_o}$ es más pequeño que el correspondiente a $\hat{Y}_{X_o}$, estos resultados se obtienen con facilidad al emplear una herramienta computacional como se muestra a continuación.

Para obtener los intervalos de predicción y de confianza a través de SAS se tiene el programa 6.6. con el cual además del cuadro de análisis de varianza que se muestra en las salidas 6.4 y 6.5, se obtiene la salida 6.6 (Anexo-B), donde se enlistan los valores predichos de la variable dependiente, los intervalos de confianza para un valor medio y los intervalos de predicción para un valor puntual.

Programa 6.6. Programación de SAS para obtener modelos de Regresión Lineal Simple y sus correspondientes intervalos de confianza y residuales

```
OPTIONS NODATE;
DATA lote8;
INPUT lote vaso min disolu;
CARDS;
8 1  10   1.79
8 1  20   3.66
8 1  30   6.00
. .   .     .
. .   .     .
. .   .     .
8 6 360  89.62
8 6 420 100.50
8 6 480 101.29
;
PROC REG;
TITLE1 'ANALISIS DE REGRESION SIMPLE';
TITLE2 'LOTE 8';
MODEL1 DISOLU=MIN/NOINT R CLI CLM;
MODEL2 DISOLU=MIN/NOINT R CLI CLM;
          REWEIGHT MIN<=20;
MODEL3 LDIS=LMIN/R CLI CLM;
          REWEIGHT UNDO;
RUN;
```

Los resultados de la salida 6.6 (Anexo B) se pueden observar mejor en las figuras 6.5a-b y 6.6, en las primeras se observa que la mejor relación lineal se encuentra cuando se emplean las variables transformadas (*LMIN vs LDIS*), y se observa que los intervalos de predicción son más amplios que el intervalo de confianza, debido a que en el de predicción se deben incluir a todas las observaciones. Sin embargo en la gráfica de *Min vs Disolu*,se observa que al transcurrir el tiempo aumenta la dispersión lo que provoca que en los intervalos de confianza no se incluyan todos los puntos experimentales, y que los intervalos de predicción sean muy amplios, aunque al inicio no sea necesario tener un intervalo tan grande. Por otro lado, en la figura 6.6 se observa que en la gráfica de Residuales vs Valores Predichos de variables no transformadas es evidente un efecto de embudo, es decir, al aumentar los valores del porcentaje de disolución aumentan los valores del residual, lo cual indica que uno de los supuestos estadísticos no se cumple, por lo que es necesario encontrar una transformación, donde la relación lineal sea mejor, lo cual ocurre al obtener el logaritmo de ambas variables, y como se observa en el gráfico de residuales, todos los valores se distribuyen de manera más homogénea alrededor del cero, lo cual indica que los supuestos del análisis de regresión lineal, tales como homogeneidad de varianzas, linealidad y normalidad se cumplen, por lo que el modelo que se obtiene a partir de estos datos tiene menor error en la predicción de la variable de respuesta..

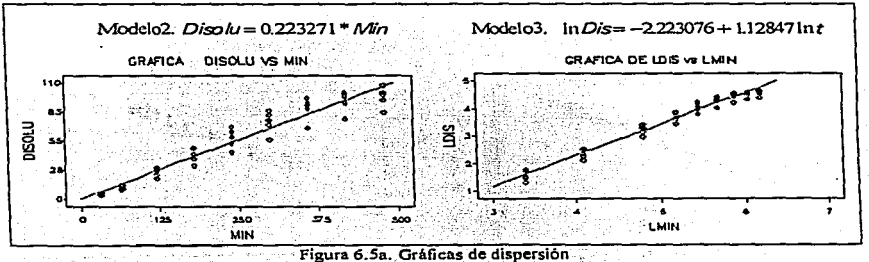
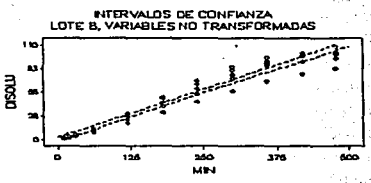
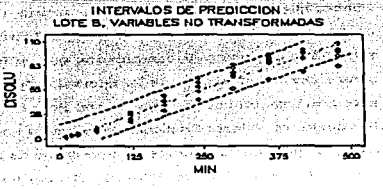


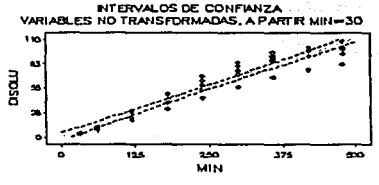
Figura 6.5a. Gráficas de dispersión

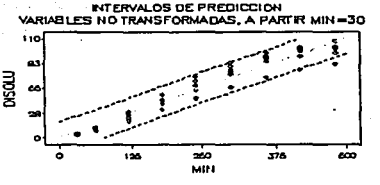
Modelo1. $Disolu = 0.223223 * Min$



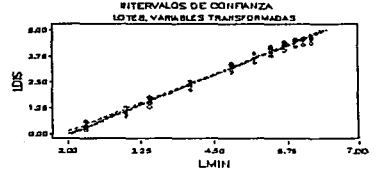


Modelo2. $Disolu = 0.223271 * Min$





Modelo3. $\ln Dis = -2.223076 + 1.12847 \ln t$



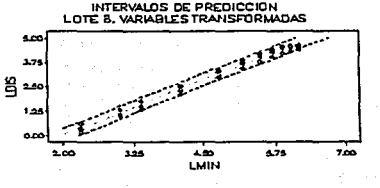


Figura 6.5b. Intervalos de confianza y de predicción (para 3 modelos)

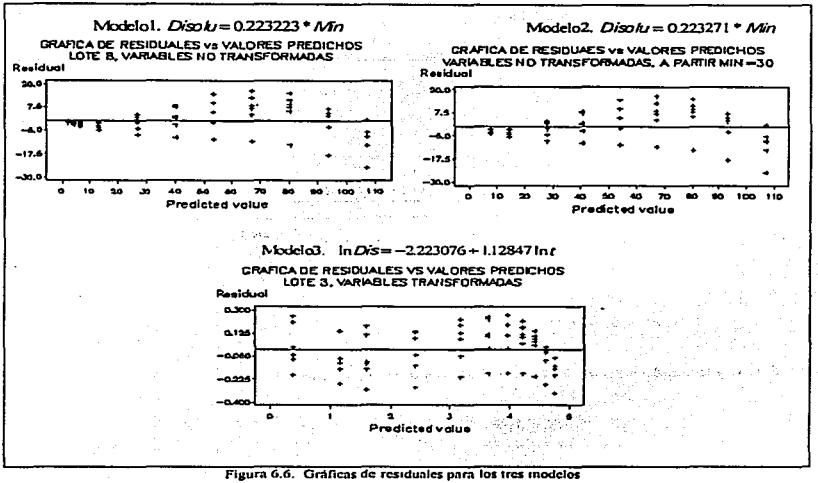


Figura 6.6. Gráficas de residuales para los tres modelos

6.8. COEFICIENTE DE CORRELACIÓN Y DETERMINACIÓN.

Por otro lado, para evaluar el porcentaje de predicción del modelo de regresión, se debe calcular el **coeficiente de correlación muestral** y posteriormente el **coeficiente de determinación**. Así, para ilustrar el cálculo del coeficiente de correlación muestral se emplean los datos del estudio de caso, y se considera el modelo 2, de la salida 6.4, donde se toman en cuenta las observaciones a partir del minuto 30. Para esto se tiene:

$$\sum_{i=1}^{n} X_i Y_i = 985028.4 \qquad \sum_{i=1}^{n} X_i^2 = 4411800$$

$$\sum_{i=1}^{n} X_i = 13140 \qquad \sum_{i=1}^{n} Y_i^2 = 223541.5332$$

$$\sum_{i=1}^{n} Y_i = 2946.36 \qquad n = 54$$

Así, de la ecuación 5.1.8a se obtiene:

$$r = \frac{985028.4 - \dfrac{(13140)(2946.36)}{54}}{\left\{\left[4411800 - \dfrac{(13140)^2}{54}\right]\left[223541.5332 - \dfrac{(2946.36)^2}{54}\right]\right\}^{1/2}}$$

$$r = 0.970886869$$

$$r^2 = 0.9426213$$

ó de (5.1.9) y tomando en cuenta el valor de la pendiente se tiene:

$$S_x = 151.37109$$

$$S_Y = 34.41741665$$

$$\hat{\beta}_1 = 0.220752$$

Por lo tanto $r = \dfrac{151.37109}{34.41741665}(0.220752) = 0.970886869$

y $r^2 = 0.9426213$

Sin embargo, al considerar el reajuste del modelo cuando se elimina el intercepto, el valor de la pendiente cambia a: $\hat{\beta}_1 = 0.223271$

Por lo que $r = \dfrac{151.37109}{34.41741665}(0.223271) = 0.981967225$

y $r^2 = 0.964259$

Estos resultados se obtienen en las salidas 6.4 y 6.5 de SAS y se observa que en ambos casos existe una buena correlación, sin embargo es mejor el coeficiente de determinación en el segundo caso, es decir, el modelo sin intercepto puede predecir un mayor porcentaje del comportamiento de los datos de disolución de las tabletas de furosemida.

Debido a que una prueba de hipótesis para el coeficiente de correlación es equivalente a la prueba de la pendiente, como se indica en la sección 5.1.4, esta prueba se obtiene en las salidas de SAS, donde se observa que en todos los casos el valor del coeficiente de correlación es diferente de cero, y los valores que se obtienen para los coeficientes de determinación indican que existe un alto porcentaje de predicción en todos los modelos ajustados, por lo que fue necesario evaluar otros parámetros, tales como CME, C.V., Residuales, para determinar el modelo que mejor ajusta a los resultados experimentales.

Para ejemplificar los cálculos necesarios para realizar una prueba de hipótesis del coeficiente de correlación, se tiene del caso de estudio, para el modelo 2 sin intercepto: $r = \sqrt{0.9838}$, y el valor del estadístico t es:

$$t = \frac{\sqrt{0.9838}\sqrt{54-2}}{\sqrt{1-(0.9838)}} = 56.194987$$

el cual es el mismo valor que se obtiene para la prueba de la pendiente, y a través del cual se puede concluir que el valor del coeficiente de correlación es diferente de cero.

Una vez que se tiene el valor del coeficiente de correlación y se esta seguro que es diferente de cero, se calcula el valor del coeficiente de determinación. Así por ejemplo, para el modelo 2 del caso de estudio, al tomar en cuenta el intercepto, el valor de r^2 es, según la ecuación (5.2.3):

$$SCY=62781.60413$$

$$SCE=3602.32603$$

$$r^2 = \frac{62781.60413 - 3602.32603}{62781.60413} = 0.942621404$$

Este resultado se encuentra en la salida de SAS 6.4, para el modelo 2, e indica que el modelo puede explicar el 94.3 porciento de la variación en los datos, sin embargo, cuando el intercepto se hace cero, el valor de r^2 es:

$$SCY = 223541.55320$$

$$SCE = 3612.96423$$

$$r^2 = \frac{223541.5532 - 3612.96423}{223541.5532} = 0.983837616$$

También este resultado se encuentra en la salida de SAS 6.5 (modelo 2) y se observa que el porcentaje de variación explicada aumenta, por lo que se puede concluir que este es un mejor modelo que aquel donde se considera el valor del intercepto.

Debido a que el valor de r^2 es positivo y cercano a 1 en todos los casos, esto indica que existe una asociación positiva entre las variables en estudio y la relación es aproximadamente lineal. Sin embargo, al comparar los tres coeficientes de determinación de cada uno de los modelos propuestos (cuadro 6.2), se concluye que existe mejor linealidad y poder de predicción en el modelo 3, debido a que su coeficiente de determinación tiene el valor más cercano a 1, por lo que se considera el modelo que mejor ajusta los resultados del estudio de disolución de las tabletas de furosemida, lo cual se fundamenta también con los criterios estadísticos antes descritos.

Cuadro 6.2. Coeficientes de correlación y determinación para los modelo del caso de estudio

	r	r^2
Modelo 1 (Sin intercepto)	0.9917	0.9835
Modelo 2 (Sin intercepto)	0.9917	0.9835
Modelo 3 (Variables transformadas)	0.9941	0.9882

6.9. PRUEBA DE FALTA DE AJUSTE

Una vez que se encuentran los parámetros importantes del modelo lineal como hasta ahora se ha mostrado, es importante probar la validez del modelo, lo cual se realiza a través de una prueba de falta de ajuste o bondad de ajuste, donde se prueba la hipótesis nula H_0 : El modelo se ajusta adecuadamente a los datos (no existe falta de ajuste). Esta prueba se describe en la sección 4.9, donde se menciona la estrategia a seguir para encontrar los estadísticos necesarios, los cuales se muestran a continuación. Para estimar el componente del error puro y el componente que describe la falta de ajuste es necesario tener réplicas en cada nivel del factor en estudio

En el cuadro 6.3 se muestran los valores que se emplearon para ajustar el modelo 1, el cual describe la disolución en función del tiempo, sin transformar ni eliminar los datos.

Cuadro 6.3. Datos para realizar la prueba de falta de ajuste (Modelo 1)

X (Min)	$\overline{Y}$ (Disolu Promedio)	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	$\Sigma(Y_n - \overline{Y})^2$	σ
10	1.515000	1.79	1.40	1.20	1.87	1.35	1.48	0.3426	0.069
20	3.061667	3.66	2.97	2.45	3.66	2.75	2.88	1.2287	0.246
30	4.783333	6.00	4.32	3.70	5.61	4.57	4.50	3.6778	0.736
60	10.94000	12.60	9.70	8.22	12.58	11.98	10.56	15.6072	3.121
120	25.78667	30.14	22.83	19.44	28.95	27.31	26.05	80.3701	16.074
180	41.75167	47.79	38.23	31.57	48.57	42.45	41.90	199.5293	39.906
240	57.60500	63.43	52.96	43.92	68.23	58.54	58.55	357.4438	71.489
300	73.31667	79.89	70.61	56.18	83.74	73.96	75.52	458.1145	91.623
360	86.32833	91.53	88.12	67.57	95.38	85.75	89.62	475.2447	95.049
420	93.33000	96.58	98.23	75.60	98.30	90.77	100.50	431.5889	86.318
480	97.21833	98.98	107.76	82.24	98.83	94.21	101.29	366.8067	73.361
								Σ=2389.9543	

Con los datos del cuadro 6.3 se obtienen los siguientes parámetros:

Suma de cuadrados del error puro: $SC_{ep} = 2389.9543$

Grados de libertad del error puro: $gl_{ep} = 66 - 11 = 55$ gl

Cuadrado Medio del error puro: $CM_{ep} = Sc_{ep} / gl_{ep} = 43.453715$

Suma de cuadrados de falta de ajuste:

$$SC_{fda} = SCE - SC_{ep} = 3629.43017 - 2389.9543 = 1239.47587$$

Grados de libertad del error puro:

$$gl_{fda} = 11 - 2 = 9 \text{ gl}$$

Cuadrado Medio del error puro:

$$CM_{fda} = Sc_{fda} / gl_{fda} = 137.719541$$

$$F_0 = \frac{CM_{fda}}{CM_{ep}} = \frac{137.719541}{43.453715} = 3.169$$

$$F_{9,55,0.05} = 2.12$$

Hipótesis a probar:

H_0: El modelo se ajusta adecuadamente a los datos (No existe falta de ajuste)

H_a: El modelo no se ajusta adecuadamente a los datos (Existe falta de ajuste)

Como $F_0 = 3.169 > F_{9,55,0.05} = 2.12$, se rechaza la hipótesis nula y se concluye que el modelo tiene falta de ajuste, por lo que es conveniente encontrar un modelo que ajuste mejor los datos.

En base a lo anterior, se transformaron las variables en estudio, obteniendo el logaritmo natural de cada variable y se ajusto el modelo 3 y sus parámetros se encuentran en la salida de SAS 6.4. Así, se tienen los datos del cuadro 6.4 para calcular los estadísticos de la prueba de falta de ajuste.

Cuadro 6.4. Datos para realizar la prueba de falta de ajuste (Modelo 3)

X (LMIN)	$\overline{Y}$ (LDIS Promedio)	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	$\Sigma(Y_i - \overline{Y})^2$	σ
2.30258	0.403182	0.5822	0.3365	0.1823	0.6259	0.3001	0.3920	0.14567	0.0291
2.99573	1.108161	1.2974	1.0886	0.8961	1.2975	1.0116	1.0578	0.12889	0.0258
3.40119	1.551915	1.7917	1.4633	1.3083	1.7245	1.5195	1.5041	0.15786	0.0316
4.09434	2.380802	2.5337	2.2721	2.1066	2.5321	2.4832	2.3571	0.14434	0.0289
4.78749	3.239017	3.4058	3.1281	2.9673	3.3656	3.3072	3.2600	0.13507	0.0270
5.19296	3.721544	3.8668	3.6436	3.4522	3.8830	3.7483	3.7353	0.12670	0.0253
5.48064	4.044053	4.1499	3.9695	3.7824	4.2229	4.0697	4.0699	0.11855	0.0237
5.70378	4.287004	4.3806	4.2572	4.0286	4.4277	4.3035	4.3244	0.09793	0.0196
5.88610	4.452236	4.5167	4.4787	4.2132	4.5578	4.4514	4.4956	0.07505	0.0150
6.04025	4.531608	4.5704	4.5873	4.3255	4.5880	4.5083	4.6102	0.05699	0.0114
6.17379	4.573564	4.5949	4.6799	4.4096	4.5934	4.5455	4.6180	0.04179	0.0084
								Σ=1.22884	

Así, se obtienen los siguientes parámetros:

Suma de cuadrados del error puro: $SC_{ep} = 1.22884$

Grados de libertad del error puro: $gl_{ep} = 66 - 11 = 55$ gl

Cuadrado Medio del error puro: $CM_{ep} = Sc_{ep} / gl_{ep} = 0.0223425$

Suma de cuadrados de falta de ajuste:

$$SC_{fda} = SCE - SC_{ep} = 1.61424 - 1.22884 = 0.3854$$

Grados de libertad del error puro:

$$gl_{fda} = 11 - 2 = 9 \text{ gl}$$

Cuadrado Medio del error puro:

$$CM_{fda} = SC_{fda} / gl_{fda} = 0.04282222$$

Hipótesis a probar:

H_0: El modelo se ajusta adecuadamente a los datos (No existe falta de ajuste)
H_a: El modelo no se ajusta adecuadamente a los datos (Existe falta de ajuste)

$$F_0 = \frac{CM_{fda}}{CM_{ep}} = \frac{0.04282222}{0.0223425} = 1.9166$$

$$F_{9,55,0.05} = 2.12$$

Como $F_0 = 1.9166 < F_{9,55,0.05} = 2.12$, no se rechaza la hipótesis nula y se concluye que el modelo no tiene falta de ajuste, por lo que es un modelo que ajusta mejor los datos.

Por lo tanto se puede decir que el modelo que mejor ajusta los resultados experimentales del logaritmo del porcentaje de disolución de tabletas de furosemida en función del tiempo es:

$$\ln Dis = -2.223076 + 1.128471 \ln t$$

Como se observa, es necesario aplicar todos los elementos de decisión cuando se realiza un análisis de regresión, ya que su conjunción conduce a la propuesta de modelos más confiables, y que sin el conocimiento de los fundamentos teóricos e interpretación, el análisis se vuelve más difícil. También cabe mencionar la importancia de contar con un software estadístico debido a que proporciona gran ayuda en los cálculos y análisis del modelo.

De esta forma se concluye el estudio de caso para el análisis de regresión lineal simple y a continuación se describe lo referente al análisis de regresión múltiple.

VII. ANÁLISIS DE REGRESIÓN MÚLTIPLE

7.1.- DEFINICIÓN.

El análisis de regresión múltiple ayuda a evaluar la relación funcional entre dos o más variables de regresión o independientes $(X_1, X_2, \ldots, X_k)$ y una variable de respuesta (Y). En muchos problemas de regresión intervienen más de una variable independiente, por ejemplo, el rendimiento de una reacción química puede depender de la temperatura, presión y concentración del catalizador. En este caso existen al menos tres variables de regresión, por lo que es importante conocer los fundamentos del método de regresión múltiple, el cual se describe en esta sección, se comienza con las suposiciones que se requieren, se describen los procedimientos para estimar parámetros importantes, se explica como realizar e interpretar inferencias acerca de esos parámetros y se dan ejemplos del ámbito farmacéutico que ilustran el manejo de las técnicas y herramientas disponibles para este tipo de análisis. Antes de pasar a los detalles de la técnica es importante mencionar que trabajar con diversas variables independientes simultáneamente en un análisis de regresión es considerablemente más difícil que trabajar con una sola variable independiente, por las siguientes razones: [16,22]

1) Es más difícil la elección del mejor modelo, ya que casi siempre hay varios candidatos razonables a elegirse.

2) Es más difícil de visualizar el modelo a considerar, ya que no es posible representar de manera gráfica más de tres dimensiones.

3) Los cálculos son virtualmente imposibles sin el acceso a una computadora de alta velocidad y un programa computacional eficiente.

Por otro lado, cuando se adicionan al modelo términos de mayor orden como por ejemplo x^2 o x^3, puede considerarse que se están introduciendo nuevas variables independientes. Así, si se renombra a x como x_1 y a x^2 como x_2, el modelo de segundo orden

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \qquad (7.1.1)$$

se puede escribir como

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (7.1.2)$$

Por lo general en regresión polinomial solamente se tienen pocas variables independientes básicas, y las otras son simples funciones matemáticas de las variables básicas. La forma general de un modelo de regresión para k variables independientes está dado por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + \varepsilon \qquad (7.1.3)$$

Donde $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_K$ son los coeficientes de regresión a estimarse. Las variables independientes $X_1, X_2, X_3, ..., X_K$ pueden ser todas ellas variables básicas separadas o algunas de ellas pueden ser funciones de otra variable básica.

7.2.- OBSERVACIÓN GRÁFICA DEL PROBLEMA.

Cuando se trabaja con una sola variable independiente, el problema se puede describir en forma gráfica para encontrar la curva que mejor ajusta la dispersión de un conjunto de n puntos $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$. En este caso se tiene una representación bidimensional, como se muestra en la figura 7.1. [16]

Sin embargo, cuando el número de variables independientes básicas es mayor a dos, la dimensión gráfica del problema se incrementa y la ecuación de regresión no es una curva en el espacio bidimensional sino una "hipersuperficie" en un espacio $(K - 1)$-dimensional, donde K es el número de variables independientes básicas, por lo que no es posible representar en una gráfica simple la dispersión de los datos o la ecuación de regresión. En el caso especial K=2, el problema es encontrar la superficie en tres dimensiones que mejor ajuste la dispersión de puntos $(X_{11}, X_{21}, Y_1), (X_{12}, X_{22}, Y_2), ..., (X_{1n}, X_{2n}, Y_n)$, donde (X_{1i}, X_{2i}, Y_i) denotan los valores X_1, X_2 e Y para el i-ésimo individuo en la muestra. La ecuación de regresión en este caso es por consiguiente la superficie descrita por los valores medios de Y a varias combinaciones de X_1 y X_2, es decir, para cada par distinto de X_1 y X_2 existe una distribución de valores de Y con media $\mu_{Y/X_1,X_2}$ y varianza $\sigma^2_{Y/X_1,X_2}$, al igual que para el análisis de regresión lineal simple, donde en cada nivel de X existe una variable aleatoria Y con cierta distribución de probabilidad.

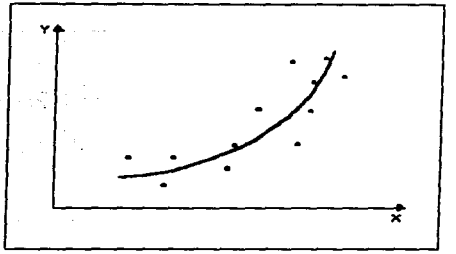


Figura 7.1. Gráfica de dispersión para una sola variable independiente.

En el espacio bidimensional la curva más sencilla es una línea recta y en el espacio tridimensional la curva más simple es un plano, el cual tiene un modelo estadístico de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (7.2.1)$$

De manera que encontrar el plano más adecuado es el primer paso en la determinación de la superficie de mejor ajuste en el espacio tridimensional, cuando existen dos variables independientes. Una representación gráfica de un plano que ajusta los datos en tres dimensiones se presenta en la figura 7.2.

Para el caso tridimensional, la solución de mínimos cuadrados para obtener el plano de mejor ajuste se determina minimizando la suma de cuadrados de las distancias entre los valores observados Y_i y los correspondientes valores estimados por $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ que se basan en el plano ajustado, esto es, la cantidad (7.2.2) se minimiza para encontrar los estimadores de mínimos cuadrados $\hat{\beta}_0$ de β_0, $\hat{\beta}_1$ de β_1 y $\hat{\beta}_2$ de β_2.

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)^2 \qquad (7.2.2)$$

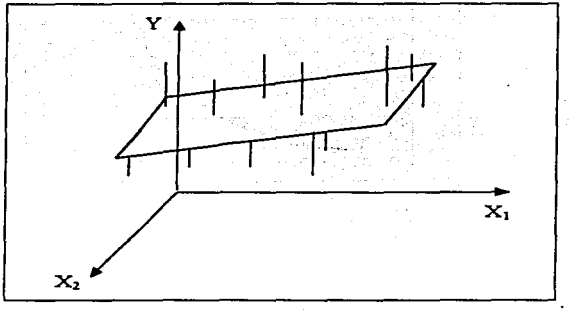


Figura 7.2. Plano tridimensional.

Por otro lado, en el análisis de regresión múltiple existen diversas estrategias para estudiar la relación de dos o más variables, entre las más importantes se encuentran las siguientes:

- Método Forward.
- Método Backward.
- Método Stepwise.

La elección de la estrategia depende esencialmente del tipo de problema, y de los datos que se tienen. De manera que se pueden probar más de uno de estos caminos y usar aquel cuyos resultados describa de la manera más razonable la relación entre las variables dependiente e independientes. [16,29]

7.3.- SUPOSICIONES DE LA REGRESIÓN MÚLTIPLE.

Ya se ha descrito el problema de regresión múltiple en general y se han mencionado implícitamente algunas de las suposiciones que se involucran, las cuales se enumeran a continuación.

Estas suposiciones coinciden con las del análisis de regresión lineal simple, pero toman en cuenta más de una variable independiente.

1) Para cada combinación especifica de valores de las variables independientes (básicas) $X_1, X_2, \ldots, X_K$, Y es una variable aleatoria (univariada) con una cierta distribución de probabilidad.

2) Las observaciones de Y son estadísticamente independientes.

3) El valor medio de Y para cada combinación específica de $X_1, X_2, \ldots, X_K$ es una función lineal de $X_1, \ldots, X_K$; esto se puede representar con la ecuación (7.3.1) o (7.3.2), donde, ε es el componente del error que refleja la diferencia entre una respuesta observada individual de Y y la respuesta media verdadera $\left(\mu_{Y/X_1, X_2, X_3, \ldots, X_K}\right)$.

$$\mu_{Y/X_1, X_2, X_3, \ldots, X_K} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \qquad (7.3.1)$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon \qquad (7.3.2)$$

Algunas consideraciones con respecto a la suposición tres son: [16]

a) A la superficie descrita por la ecuación (7.3.1), se le conoce como ecuación de regresión, superficie de respuesta o superficie de regresión.

b) Si algunas de las variables independientes son funciones de mayor orden de unas cuantas variables independientes básicas, por ejemplo: $X_3 = X_1^2$, $X_3 = X_1 X_2$, la expresión $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$ no es realmente lineal en las variables básicas por lo que el término "superficie" es mejor que "plano".

c) Al igual que en regresión lineal simple, ε es la cantidad diferencial que una respuesta individual observada se desvía de la superficie de respuesta, es decir, ε es el componente del error en el modelo.

4) Suposición de homocedasticidad. La varianza de Y es la misma para cualquier combinación fija de $X_1..X_2.....X_K$, esto es,

$$\sigma^2_{Y/X_1..X_2....X_K} = Var(Y/X_1..X_2.....X_K) = \sigma^2 \qquad (7.3.3)$$

Esta suposición quizá parece muy restrictiva, sin embargo, la homocedasticidad sólo se cuida cuando los datos muestran muy obvias y significativas desviaciones de homogeneidad; en general, pequeñas desviaciones no tienen efectos adversos sobre los resultados.

5) Para una combinación fija de $X_1..X_2.....X_K$, Y tiene una distribución normal. Esto es,

$$Y \sim N\left(\mu_{Y/X_1..X_2....X_K}, \sigma^2\right) \qquad (7.3.4)$$

Estas suposiciones no son necesarias para fijar el modelo de regresión por mínimos cuadrados, pero se requieren, en general, para propósitos de inferencia. En este aspecto, las pruebas de hipótesis e intervalos de confianza usuales en un análisis de regresión son "robustos" en el sentido de que algunas desviaciones extremas de la distribución de Y de la normalidad puede proporcionar falsos resultados.

7.4. DETERMINACIÓN DE LA ECUACIÓN DE REGRESIÓN MÚLTIPLE.

Al igual que en regresión lineal simple, existen dos formas básicas para determinar la mejor estimación de una ecuación de regresión múltiple: Mínimos Cuadrados y Varianza Mínima. Ambos proporcionan la misma solución.

7.4.1.- Método de Mínimos Cuadrados.

El método de mínimos cuadrados elige el modelo de mejor ajuste como aquel en donde la suma de cuadrados de las distancias entre la respuesta observada y la predicha por el modelo es mínima. Así, la ecuación (7.4.1) denota el modelo de regresión ajustado, y la ecuación (7.4.2.) proporciona la suma de las desviaciones de los valores de Y observados de los correspondientes valores predichos usando el modelo de regresión ajustado.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_K X_K \qquad (7.4.1)$$

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_k X_{ik}\right)^2 \qquad (7.4.2)$$

La solución de mínimos cuadrados consiste en encontrar los valores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$, llamados estimadores de mínimos cuadrados, para los cuales la suma en (7.4.2) es mínima. A esta suma de cuadrados se le llama suma de cuadrados residual, suma de cuadrados acerca de la regresión o suma de cuadrados del error (SCE).

Es importante mencionar algunas propiedades de la solución de mínimos cuadrados:

a) Cada uno de los estimadores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ es una función lineal de los valores de Y. Esta propiedad de linealidad facilita determinar las propiedades estadísticas de los estimadores. En particular, ya que se asume que los valores de Y son normalmente distribuidos, cada uno de los

estimadores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$ serán normalmente distribuidos, con desviaciones estándar fácilmente calculables.

b) La ecuación de regresión de mínimos cuadrados $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_K X_K$ es aquella única combinación lineal de las variables independientes $X_1, X_2, \ldots, X_K$, que tienen máxima correlación posible con la variable dependiente. En otras palabras, de todas las elecciones posibles de las combinaciones lineales de la forma $b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_K X_K$, la combinación lineal $\hat{Y}$ es tal que la correlación

$$r_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})\left(\hat{Y}_i - \bar{\hat{Y}}\right)}{\sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}} \qquad (7.4.3)$$

es máxima, donde $\hat{Y}_i$ es el valor predicho de Y para el i-ésimo individuo y $\bar{\hat{Y}}$ es la media de las $\hat{Y}_i$'s.

Otro método es el de varianza mínima que determina la superficie de mejor ajuste a través de los estimadores insesgados de mínima varianza $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ de $\beta_0, \beta_1, \ldots, \beta_K$ respectivamente. Sin embargo el más común es el de mínimos cuadrados.

A continuación se describen las pruebas estadísticas y criterios de descición del análisi de regresión múltiple.

7.5. TABLA DE ANÁLISIS DE VARIANZA, PRUEBA DE F TOTAL Y PRUEBA DE F PARCIAL.

Al igual que en el análisis de regresión lineal simple, se usa una tabla de análisis de varianza para proporcionar un resumen completo del análisis de regresión múltiple, como se muestra en el cuadro 7.1.

Cuadro 7.1. Tabla de Análisis de varianza para regresión múltiple

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD (g.l)	SUMA DE CUADRADOS (SC)	CUADRADO MEDIO (CM)	F	R^2
Regresión	k	$SCR = SCY - SCE$ $SCR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$	$CMR = \frac{SCR}{k}$	$F = \frac{CMR}{CME}$	$R^2 = \frac{SCY - SCE}{SCY}$
Residual	n-k-1	$SCE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	$CME = \frac{SCE}{n-k-1}$		
Total	n-1	$SCY = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$			

Hipótesis a probar:

Ho: "Todas las *k* variables independientes consideradas al mismo tiempo NO explican una cantidad significativa de la variación en la respuesta."

H_A: " Las *k* variables independientes explican una cantidad significativa de la variación en la respuesta"

Como antes, al término *SCY* se le llama suma de cuadrados total, y representa la variabilidad total en las observaciones de Y antes de tomar en cuenta el efecto de las variables independientes. El término *SCE* es la suma de cuadrados residual o suma de cuadrados debida al error, y representa la variación no explicada cuando las variables independientes se toman en cuenta en la ecuación de regresión para predecir Y. Al término *SCR* se le llama suma de cuadrados de la regresión y mide la reducción en la variación o variación explicada, debido a la presencia de las variables independientes en la ecuación de regresión. Así se tiene la siguiente representación:

Suma de cuadrados total = suma de cuadrados de la regresión
+ suma de cuadrados residual

$$SCY = SCR + SCE$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad (7.5.1)$$

7.5.1. Prueba de significancia para la regresión total.

Se puede probar la hipótesis nula general "H_0: todas las k variables independientes consideradas al mismo tiempo no explican una cantidad significativa de la variación en Y" calculando el estadístico F mediante la ecuación (7.5.2).

$$F = \frac{CMR}{CME} = \frac{\text{Cuadrado Medio de la Regresión}}{\text{Cuadrado Medio del Error}} \qquad (7.5.2)$$

El valor de F calculado se compara con el punto crítico $F_{k,n-k-1,1-\alpha}$, donde k es el número de variables independientes, n es el tamaño muestral y α es un nivel de significación predeterminado. Se rechaza H_0 si el valor de F calculado excede al valor de $F_{k,n-k-1,1-\alpha}$.

El término cuadrado medio del error que es el denominador en (7.5.2) está dado por la ecuación (7.5.4) y proporciona una estimación de la varianza del modelo.

$$CM_{residual} = \frac{1}{n-k-1}SCE = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad (7.5.4)$$

El cuadrado medio de la regresión, el numerador en (7.5.2), está dado por la ecuación (7.5.5) y da una estimación independiente de σ^2, sólo si la hipótesis nula H_0:"Regresión total no significativa" no se rechaza. De lo contrario existiría una sobrestimación de la varianza en proporción a la magnitud de los coeficientes de regresión $\beta_1, \beta_2, \ldots, \beta_k$, y esto es porque un valor de F que es

"grande" favorece el rechazo de H_0. Así, el estadístico F (7.5.2) es la razón de dos estimadores independientes de la misma varianza sólo si la hipótesis nula H_0 es verdadera.

$$CM_{regresión} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{k} \tag{7.5.5}$$

El valor del coeficiente de determinación, R^2, en la tabla de ANOVA provee una medida cuantitativa de como la combinación de las variables independientes predicen la variable dependiente, su cálculo es mediante la ecuación (7.5.6) que es igual a la ecuación (5.2.3) para el caso de regresión lineal simple.

$$R^2 = \frac{SCY - SCE}{SCY} \tag{7.5.6}$$

Esta información se obtiene cuando se ajusta un modelo de regresión múltiple en forma general, sin embargo cuando el ajuste del modelo se realiza mediante alguna técnica de selección de variables, como Forward, Backward o Stepwise, es necesario conocer los valores del estadístico F Parcial para realizar pruebas de hipótesis e inferencias acerca de diversos modelos que pueden elegirse. A continuación se describe la Prueba de F parcial.

7.5.2. Prueba de F parcial.

Cuando se trata de ajustar un modelo con k variables independientes, se puede obtener información adicional importante con respecto al modelo de regresión ajustado cuando se parte la suma de cuadrados de la regresión en sus componentes, por ejemplo si k=3, los componentes de la suma de cuadrados de regresión son:

1. $SC(X_1)$: suma de cuadrados explicada por usar solo a X_1 en la predicción de Y.

2. $SC(X_2/X_1)$: suma de cuadrados extra explicada por usar X_2 dado que X_1 ya está en el modelo de la predicción de Y.

3. $SC(X_3/X_1,X_2)$: suma de cuadrados extra explicada por usar X_3 dado que X_1 y X_2 ya están en el modelo de la predicción de Y.

Esta información se puede resumir en una tabla de ANOVA, como se muestra en el cuadro 7.2 y se puede emplear para contestar las siguientes preguntas:

1. ¿ Se puede predecir Y utilizando solo a X_1?

2. ¿ Adicionar X_2 en el modelo contribuye significativamente en la predicción de Y una vez que se toma en cuenta o se controla la contribución de X_1?

3. ¿Contribuye significativamente la adición de X_3 en la predicción de Y después de tomar en cuenta la contribución de X_1 y X_2?

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD (g.l)	SUMA DE CUADRADOS (SC)	CUADRADO MEDIO (CM)	*F parcial*
Regresión				
X_1	1	$SC(X_1)$	$CM_{(X_1)}\dfrac{SC(X_1)}{1}$	$F(X_1)$
X_2/X_1	1	$SC(X_2/X_1)$	$CM_{(X_2/X_1)}=\dfrac{SC(X_2/X_1)}{1}$	$F_{(X_2/X_1)}$
$X_3/X_1,X_2$	1	$SC(X_3/X_1,X_2)$	$CM_{(X_3/X_1,X_2)}=\dfrac{SC(X_3/X_1,X_2)}{1}$	$F_{(X_3/X_1,X_2)}$
Residual	n-k-1	$SCE=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$	$CME(X_1,X_2,X_3)$	
Total	n-1	$SCY=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$		

Cuadro 7.2. Tabla de ANOVA para prueba de F parcial

La pregunta 1 involucra un modelo de regresión lineal simple, donde la única variable independiente es X_1. Para responder las preguntas 2 y 3 se emplea una prueba de F parcial, la cual asegura si la adición de alguna variable independiente, teniendo otras en el modelo, contribuye significativamente en la predicción de Y. La prueba, por ello, permite eliminar las variables que no ayudan en la predicción de Y y así hace posible reducir el conjunto de variables independientes a un pequeño conjunto de predictores " importantes".

Por ejemplo, al realizar una prueba de F parcial sobre la variable X^*, una vez que ya se tienen en el modelo a las variables $X_1, X_2, \ldots, X_k$, primero se calcula la "suma de cuadrados extra al adicionar X^*, dadas $X_1, X_2, \ldots, X_k$", la cual se puede colocar en la tabla de ANOVA con el nombre de "Regresión $X^*/X_1, X_2, \ldots, X_k$ ". Esta suma de cuadrados se calcula por la ecuación (7.5.7).

suma de cuadrados extra adicionando X^*, dados $X_1, X_2, \ldots, X_k$ = suma de cuadrados de regresión cuando $X_1, X_2, \ldots, X_k$ y X^* están en el modelo − suma de cuadrados de regresión cuando $X_1, X_2, \ldots, X_k$ está n en el modelo, sin X^*

o

$$\text{SC extra}\ (X^*/X_1, X_2, \ldots, X_k) = \text{SC regresión}\ (X_1, X_2, \ldots, X_k, X^*) - \text{SC regresión}\ (X_1, X_2, \ldots, X_k) \qquad (7.5.7)$$

Para probar la hipótesis nula "H_0: la adición de X^* al modelo incluyendo $X_1, X_2, \ldots, X_k$ no mejora significativamente la predicción de Y" se emplea el estadístico F parcial, el cual se calcula con la ecuación (7.5.8).

$$F(X^*/X_1, X_2, \ldots, X_k) = \frac{\text{suma de cuadrados o cuadrado medio extra al adicionar } X^*, \text{ dadas } X_1, X_2, \ldots, X_k}{\text{cuadrado medio residual para el modelo que contiene todas las variables } X_1, X_2, \ldots, X_k, X^*}$$

o

$$F(X^*/X_1, X_2, \ldots, X_k) = \frac{\text{SC}\ (X^*/X_1, X_2, \ldots, X_k)}{\text{CME}\ (X_1, X_2, \ldots, X_k, X^*)} \qquad (7.5.8)$$

Este estadístico tiene una distribución F con 1 y $n-k-2$ grados de libertad bajo H_O, así se rechaza H_O si el valor de F calculado excede al valor de $F_{1,n-k-2,1-\alpha}$.

Existe una forma equivalente para realizar la prueba de F parcial, en la cual se involucra una prueba de hipótesis nula $H_0: \beta^* = 0$, donde β^* es el coeficiente de X^* en la ecuación de regresión $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta^* X^* + \varepsilon$, así el estadístico de prueba equivalente es (7.5.9).

$$t = \frac{\hat{\beta}^*}{S_{\hat{\beta}^*}} \qquad (7.5.9)$$

donde $\hat{\beta}^*$ es el estimador del coeficiente β^* y $S_{\hat{\beta}^*}$ es la estimación del error estándar de $\hat{\beta}^*$, ambos se calculan por programas de regresión.

Al hacer esta prueba, se rechaza $H_0: \beta^* = 0$ si

$$\begin{cases} |T| > t_{n-p-2,1-\alpha/2} & \text{(prueba bilateral; } H_A: \beta^* = 0\text{)} \\ T > t_{n-p-2,1-\alpha} & \text{(prueba unilateral superior; } H_A: \beta^* > 0\text{)} \\ T < -t_{n-p-2,1-\alpha} & \text{(prueba unilateral inferior; } H_A: \beta^* < 0\text{)} \end{cases}$$

7.6. CORRELACIONES MÚLTIPLE, PARCIAL Y MÚLTIPLE-PARCIAL

En temas anteriores se comentó que las características esenciales de la regresión lineal, además del modelo cuantitativo de predicción proporcionado por la ecuación de regresión ajustada por mínimos cuadrados, se puede describir en términos del coeficiente de correlación r, cuyas características se resumen a continuación:

1. El coeficiente de correlación elevado al cuadrado r^2 (coeficiente de determinación) mide la fuerza de la relación lineal entre la variable dependiente Y y la variable independiente X. Si r^2 es cercano a 1, es mayor la fuerza de la relación lineal; si r^2 es cercano a 0, la relación lineal es muy débil.

2. $r^2 = (SCY - SCE)/SCY$ es la reducción en la suma de cuadrados total al usar un modelo lineal en X para predecir Y.

3. $r = \hat{\beta}_1(S_X/S_Y)$, donde $\hat{\beta}_1$ es la pendiente estimada de la línea de regresión.

4. r o r_{XY}, es una estimación del parámetro poblacional ρ o ρ_{XY}, el cual describe la correlación entre X e Y, considerándolas como variables aleatorias.

5. Si se asume que X e Y tienen una distribución normal bivariada con parámetros $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ y ρ_{XY} la distribución condicional de Y dada X es $N(\mu_{Y/X}, \sigma_{Y/X}^2)$, donde

$$\mu_{Y/X} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) \tag{7.6.1}$$

y

$$\sigma_{Y/X}^2 = \sigma_Y^2(1 - \rho^2) \tag{7.6.2}$$

Aquí $r^2 = (SCY - SCE)/SCY$ estima ρ^2, el cual se puede expresar como:

$$\rho^2 = \frac{\sigma_Y^2 - \sigma_{Y/X}^2}{\sigma_Y^2} \tag{7.6.3}$$

La conexión entre regresión y correlación se puede extender al caso de regresión múltiple. Sin embargo, cuando se involucran diversas variables independientes, la característica esencial de regresión se describe no sólo por un coeficiente de regresión único como en el caso de regresión

lineal simple, sino por diversas correlaciones. Estas incluyen un conjunto de correlaciones de orden cero, así como un grupo de índices adicionales de mayor orden llamados correlaciones múltiples, correlaciones parciales y correlaciones múltiple-parciales. El examen de estas correlaciones de mayor orden permite responder muchas de las preguntas que se plantean al generar un modelo de regresión múltiple.

7.6.1.- Matriz de Correlación.

Cuando se trata con más de una variable independiente, la colección de todos los coeficientes de correlación de orden cero (es decir, las r's entre todos los posibles pares de variables) se puede representar de manera compacta en forma de una matriz de correlación. Por ejemplo, cuando K=3, es decir, las variables independientes X_1, X_2, X_3 y una variable dependiente Y, existen $C_2^4 = 6$ correlaciones de orden cero, y la matriz de correlación tiene la forma general (7.6.4).

$$\begin{array}{c} \\ Y \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{cccc} Y & X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} 1 & r_{Y1} & r_{Y2} & r_{Y3} \\ & 1 & r_{12} & r_{13} \\ & & 1 & r_{23} \\ & & & 1 \end{bmatrix} \end{array} \qquad (7.6.4)$$

Aquí r_{Yj} $(j = 1,2 \text{ ó } 3)$ es la correlación entre Y y X_j, y r_{ij} $(i, j = 1,2,3)$ es la correlación entre X_i y X_j. Sin embargo, el uso de todas las correlaciones de orden cero no describen:

1) la relación total de la variable dependiente Y con las variables independientes X_1, X_2 y X_3 considerando todas al mismo tiempo. Tal relación se puede medir con el coeficiente de correlación múltiple de Y sobre X_1, X_2 y X_3.

2) tampoco describe la relación entre Y y X_2 después de controlar el efecto de X_1. La medida que describe esta relación se le llama coeficiente de correlación parcial entre Y y X_2 tomando en cuenta a X_1.

3) No describe la relación entre Y y los efectos combinados de X_2 y X_3 después de controlar los efectos de X_1. Esta relación se mide mediante el coeficiente de correlación múltiple-parcial entre Y y los efectos combinados de X_2 y X_3 controlando a X_1.

A continuación se describe cada uno de los coeficientes de correlación.

7.6.2.- Coeficiente de Correlación Múltiple.

El coeficiente de correlación múltiple, se denota como $R_{Y/X_1,X_2,\ldots,X_k}$, y es una medida de la asociación lineal total de una variable dependiente Y con diversas variables independientes $X_1, X_2, \ldots, X_k$, es decir, asociación entre Y y la combinación lineal de mejor ajuste de las X's. El valor de $R_{Y/X_1,X_2,\ldots,X_k}$ es siempre positivo. El coeficiente de correlación múltiple es una generalización directa del coeficiente de correlación simple r al caso de diversas variables independientes.

Existen dos ecuaciones que proporcionan gran interpretación del coeficiente de correlación múltiple $R_{Y/X_1,X_2,\ldots,X_k}$ y su cuadrado, las cuales son (7.6.5) y (7.6.6).

$$R_{Y/X_1,X_2,\ldots,X_k} = \frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(\hat{Y}_i-\bar{\hat{Y}}\right)}{\left[\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{\hat{Y}}\right)^2\right]^{1/2}} \qquad (7.6.5)$$

y

$$R^2_{Y/X_1,X_2,\ldots,X_k} = \frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2-\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2} = \frac{SCY-SCE}{SCY} \qquad (7.6.6)$$

donde $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$ es el valor predicho para el i-ésimo individuo y $\bar{\hat{Y}} = \sum_{i=1}^{n}\hat{Y}_i/n$. La ecuación (7.6.6) es la que se emplea para probar el ajuste del modelo de regresión.

Si se compara la ecuación (7.6.5) con la ecuación (5.1.8), del capitulo 5, se puede observar que el coeficiente de correlación múltiple es la correlación lineal simple entre los valores observados de Y y los valores predichos $\hat{Y}$, es decir, $R_{Y/X_1,X_2,\ldots,X_k} = r_{Y\hat{Y}}$.

Como $\bar{Y} = \bar{\hat{Y}}$, la mejor ecuación para calcular $R_{Y/X_1,X_2,\ldots,X_k}$ es (7.6.7).

$$R_{Y/X_1,X_2,\ldots,X_k} = \frac{\sum_{i=1}^{n} Y_i \hat{Y}_i - n\bar{Y}^2}{\sqrt{\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)\left(\sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}^2\right)}} \tag{7.6.7}$$

El coeficiente de correlación múltiple muestral $R_{Y/X_1,X_2,\ldots,X_k}$ se puede considerar como un estimador del parámetro poblacional que caracteriza la distribución conjunta de todas las variables $Y, X_1, X_2, \ldots, X_k$ al mismo tiempo. Cuando se tienen dos variables X e Y y se asume que su distribución conjunta es normal bivariada $N_2\left(\mu_Y, \mu_X, \sigma_Y^2, \sigma_X^2, \rho_{XY}\right)$, se dice que r_{XY} estima a ρ_{XY}, lo cual satisface la ecuación $\rho_{XY}^2 = \left(\sigma_Y^2 - \sigma_{Y/X}^2\right)/\sigma_Y^2$, donde $\sigma_{Y/X}^2$ es la varianza de la distribución condicional de Y dado un valor de X. De la misma forma, cuando se tienen K variables independientes y una variable dependiente, se llega a un resultado análogo si se asume que su distribución conjunta es normal multivariada. Así, por ejemplo para el caso K=2, la distribución normal trivariada de, Y, X_1 y X_2 se puede describir como

$$N_3\left(\mu_Y, \mu_{X1}, \mu_{X2}, \sigma_Y^2, \sigma_{X1}^2, \sigma_{X2}^2, \rho_{Y1}, \rho_{Y2}, \rho_{12}\right) \tag{7.6.8}$$

donde μ_Y, μ_{X1} y μ_{X2} son las tres medias (no condicionadas), $\sigma_Y^2, \sigma_{X1}^2$ y σ_{X2}^2 son las tres varianzas (no condicionadas) y ρ_{Y1}, ρ_{Y2} y ρ_{12} son los tres coeficientes de correlación. La distribución condicional de Y dados X_1 y X_2 es una distribución normal univariada con una media (condicional) que se denota por $\mu_{Y/X_1,X_2}$ y una varianza (condicional) que se denota por $\sigma_{Y/X_1,X_2}^2$; lo cual se puede escribir en forma reducida como:

$$Y/X_1, X_2 \approx N\left(\mu_{Y/X_1,X_2}, \sigma_{Y/X_1,X_2}^2\right) \tag{7.6.9}$$

De hecho,

$$\mu_{Y/X_1,X_2} = \mu_Y + \rho_{Y1}\frac{\sigma_{Y/X_2}}{\sigma_{X_1/X_2}}\left(X_1 - \mu_{X_1}\right) + \rho_{Y2}\frac{\sigma_{Y/X_1}}{\sigma_{X_2/X_1}}\left(X_2 - \mu_{X_2}\right) \tag{7.6.10}$$

y
$$\sigma^2_{Y/X_1,X_2} = \left(1 - \rho^2_{Y,\mu_{Y/X_1,X_2}}\right)\sigma^2_Y \tag{7.6.11}$$

donde $\rho_{Y,\mu_{Y/X_1,X_2}}$ es el coeficiente de correlación poblacional entre las variables aleatorias Y y $\mu_{Y/X_1,X_2} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, donde se considera a X_1 y X_2 como variables aleatorias. También $\sigma^2_{Y/X_1}, \sigma^2_{Y/X_1}, \sigma^2_{X_2/X_1}$ y $\sigma^2_{X_2/X_1}$ son las varianzas condicionadas de Y dada X_2, de Y dada X_1, de X_1 dada X_2 y X_2 dada X_1, respectivamente.

El parámetro $\rho_{Y,\mu_{Y/X_1,X_2}}$ es el análogo poblacional del coeficiente de correlación múltiple muestral $R_{Y/X_1,X_2}$; y se puede escribir simplemente como $\rho_{Y/X_1,X_2}$. Además de la ecuación (7.6.11) se obtiene la ecuación (7.6.12), donde se observa que $\rho^2_{Y/X_1,X_2}$ es la reducción proporcionada en la varianza no condicionada de Y debido al condicionamiento de X_1 y X_2.

$$\rho^2_{Y/X_1,X_2} \left(\text{ó } \rho^2_{Y,\mu_{Y/X_1,X_2}}\right) = \frac{\sigma^2_Y - \sigma^2_{Y/X_1,X_2}}{\sigma^2_Y} \tag{7.6.12}$$

Generalizando para el caso de K variables independientes, se pueden resumir las características del coeficiente de correlación múltiple $R_{Y/X_1,X_2,\ldots,X_k}$ de la siguiente manera:

1. $R^2_{Y/X_1,X_2,\ldots,X_k}$ mide la reducción en la suma de cuadrados total $\sum(Y_i - \overline{Y})^2$ a $\sum\left(Y_i - \hat{Y}_i\right)^2$ debido a la regresión lineal múltiple de Y sobre $X_1, X_2, \ldots, X_k$.

2. $R_{Y/X_1,X_2,\ldots,X_k}$ es la correlación $r_{Y,\hat{Y}}$ de los valores observados de (Y) con los valores predichos $\left(\hat{Y}\right)$.

3. $R_{Y/X_1,X_2,\ldots,X_k}$ es un estimador de $\rho_{Y/X_1,X_2,\ldots,X_k}$ el cual es la correlación de Y con la media de la distribución condicional de Y dados $X_1, X_2, \ldots, X_k$, es decir, la correlación de Y con la ecuación de regresión $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$, donde las X's se consideran variables aleatorias.

4. $R^2_{Y/X_1,X_2,\ldots,X_k}$ es una estimación de la reducción proporcionada en la varianza no-condicionada de Y debido al condicionamiento sobre $X_1, X_2, \ldots, X_k$, esto es, estima

$$\rho^2_{Y/X_1,X_2,\ldots,X_k} = \frac{\sigma^2_Y - \sigma^2_{Y/X_1,X_2,\ldots,X_k}}{\sigma^2_Y} \tag{7.6.13}$$

Por otro lado, se puede definir un coeficiente de correlación múltiple ajustado, el cual se obtiene mediante la ecuación 7.6.14.

$$R^2_{ajustada} = 1-\left(\frac{n-i}{n-p}\right)\left(1-R^2\right) \qquad (7.6.14)$$

donde: n = Número de observaciones.

i = 1 si hay un intercepto, 0 si no lo hay.

p = Número de parámetros, incluyendo el intercepto.

La ventaja del coeficiente de correlación múltiple ajustado es que no aumenta automáticamente cada vez que se introduce una nueva variable de regresión al modelo (Montg). Por lo tanto este parámetro se puede considerar más confiable aún que el propio valor del coeficiente de correlación múltiple, aunque su valor siempre sea menor [22,29].

7.6.3. Coeficiente de Correlación Parcial

El coeficiente de correlación parcial es una medida de la fuerza de la relación lineal entre dos variables después de haber controlado los efectos de otras variables. Si las dos variables de interés son X e Y, y las variables que se controlan son $Z_1, Z_2, \ldots, Z_p$, entonces el correspondiente coeficiente de correlación parcial se escribe como $r_{Y,X/Z_1,Z_2,\ldots,Z_p}$. El orden del coeficiente de correlación parcial depende del número de variables que se controlan. Así,

parciales de primer orden tienen la forma $r_{Y,X/Z}$

parciales de segundo orden tienen la forma $r_{Y,X/Z_1,Z_2}$

y en general

parciales de p-ésimo orden tienen la forma $r_{Y,X/Z_1,Z_2,\ldots,Z_p}$

La forma más fácil de calcular un coeficiente de correlación parcial es mediante un programa de computación.

Con la ayuda de la matriz de correlación se puede obtener la variable que más se correlaciona con la variable de respuesta, es decir, la variable "más importante" en base a la fuerza de su relación lineal con Y, sin embargo, el coeficiente de correlación parcial ayuda a encontrar la variable que le sigue en importancia, una vez que se toma en cuenta la más importante y así sucesivamente.

Para realizar una prueba de significancia del coeficiente de correlación parcial, es decir, cuando se quiere probar si la adición de una variable al modelo de regresión es importante, una vez que otras variables están ya en el modelo, se emplea la prueba de F parcial, como se analizó en la sección 7.5.2. Así, para probar si $r_{Y,X/Z_1,\ldots,Z_p}$ es significativamente diferente de cero, se debe calcular la correspondiente F parcial, $F(X/Z_1,\ldots,Z_p)$, y rechazar la hipótesis nula si el estadístico F excede al valor crítico adecuado de la distribución $F_{1,n-p-2}$.

La hipótesis nula para esta prueba se puede proponer de manera formal al considerar la analogía entre el coeficiente de correlación parcial muestral $r_{YX/Z_1\ldots Z_p}$ y el coeficiente de correlación parcial poblacional $\rho_{YX/Z_1\ldots Z_p}$. La hipótesis nula se puede proponer como $H_0: \rho_{YX/Z_1,\ldots,Z_p}=0$ y la hipótesis alternativa como $H_A: \rho_{YX/Z_1,\ldots,Z_p}\neq 0$.

La estructura de la correlación parcial poblacional ayuda a relacionar la correlación de mayor orden con la regresión. Por ejemplo, si se considera la relación de Y con dos variables independientes, la ecuación para el cuadrado de ρ_{YX_1/X_2} es (7.6.15).

$$\rho^2_{YX_1/X_2}=\frac{\sigma^2_{Y/X_2}-\sigma^2_{Y/X_1,X_2}}{\sigma^2_{Y/X_2}} \qquad (7.6.15)$$

Así, el cuadrado de la correlación parcial muestral r_{YX_1/X_2} es una estimación de la reducción proporcionada en la varianza condicional de Y dado X_2 debido al control sobre X_1 y X_2.

La correlación parcial ρ_{YX_1/X_2} se puede describir como una correlación de orden cero para una distribución condicional bivariada. Si la distribución conjunta de Y, X_1 y X_2 es normal trivariada,

la distribución condicional de Y y X_1 dado X_2 es normal bivariada. La correlación de orden cero entre Y y X_1 para esta distribución condicional se escribe como ρ_{YX_1/X_2}, que es exactamente la correlación parcial entre X_1 e Y controlando X_2.

Entonces, una ecuación análoga para el cuadrado del coeficiente de correlación parcial muestral es

$$r^2_{YX_1/X_2} = \frac{SC \text{ residual (usando solo } X_2 \text{ en el modelo)} - SC \text{ residual (usando } X_1 \text{ y } X_2 \text{ en el modelo)}}{SC \text{ residual (usando solo } X_2 \text{ en el modelo)}} \qquad (7.6.16)$$

$$r^2_{YX_1/X_2} = \frac{\text{suma de cuadrados extra debido a la adición de } X_1 \text{ al modelo estando } X_2 \text{ en el modelo}}{SC \text{ residual (usando solo } X_2 \text{ en el modelo)}} \qquad (7.6.17)$$

Si se compara la ecuación (7.6.17) con la ecuación del estadístico F parcial (7.5.8), se observa que ambas tienen como numerador la suma de cuadrados extra debido a la adición de una variable cuando ya se tienen otras en el modelo, por lo que la prueba de $H_0: \rho_{YX/Z_1,\ldots,Z_p} = 0$ se hace empleando el valor de F parcial, $F(X_1/X_2)$, como el estadístico de prueba.

Otra forma de calcular una correlación parcial de primer orden es mediante la siguiente ecuación:

$$r_{YX \cdot Z} = \frac{r_{YX} - r_{YZ}r_{XZ}}{\sqrt{(1-r^2_{YZ})(1-r^2_{XZ})}} \qquad (7.6.18)$$

Se observa que la primera correlación en el numerador es la correlación muestral de orden cero entre Y y X. La variable de control Z aparece en la segunda expresión del numerador (donde se correlaciona por separado con las variables Y y X) y en ambos términos del denominador. Al usar (7.6.18) se puede interpretar al coeficiente de correlación parcial como un ajuste del coeficiente de correlación simple tomando en cuenta el efecto de la variable control. En particular, si r_{YZ} y r_{XZ} tienen el mismo signo, entonces el control en Z reduce la correlación de orden cero r_{YX} entre Y y X. Por otro lado, si r_{YZ} y r_{XZ} tienen signos contrarios, el control sobre Z incrementa r_{YX}.

Para calcular correlaciones parciales de mayor orden, se aplica la ecuación (7.6.18) pero empleando las correlaciones parciales de menor orden siguiente. Por ejemplo, la correlación parcial de segundo orden es un ajuste del coeficiente de correlación parcial de primer orden, el cual, a su vez, es un ajuste de la correlación simple de orden cero. Se tiene la siguiente fórmula general para una correlación parcial de segundo orden:

$$r_{YX/Z,W} = \frac{r_{YX/Z} - r_{YW/Z}r_{XW/Z}}{\sqrt{\left(1-r^2_{YW/Z}\right)\left(1-r^2_{XW/Z}\right)}} = \frac{r_{YX/W} - r_{YZ/W}r_{XZ/W}}{\sqrt{\left(1-r^2_{YZ/W}\right)\left(1-r^2_{XZ/W}\right)}} \qquad (7.6.19)$$

Existe una interpretación adicional importante que se puede hacer al tomar en cuenta las correlaciones parciales. Para las variables Y, X y Z, se pueden ajustar dos ecuaciones de regresión lineal $Y = \beta_0 + \beta_1 Z + \varepsilon$ y $X = \beta_0^* + \beta_1^* Z + \varepsilon$. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 Z$ es la línea ajustada de Y sobre Z, $\hat{X} = \hat{\beta}_0^* + \hat{\beta}_1^* Z$ es la línea ajustada de X sobre Z. Entonces, las desviaciones o residuales $\left(\hat{Y} - Y\right)$ y $\left(\hat{X} - X\right)$ representan los residuales después que la variable Z ha explicado toda la variación en las variables Y y X por separado.

Si se correlacionan estos residuales, es decir, se encuentra $r_{\hat{Y}-Y,\hat{X}-X}$, se obtiene una medida que es "independiente" de los efectos de Z, por lo que la correlación parcial entre Y y X controlando a Z se puede definir como la correlación de los residuales de las regresiones lineales de Y sobre Z y X sobre Z, esto es, $r_{YX/Z} = r_{\hat{Y}-Y,\hat{X}-X}$.

Se pueden resumir en forma general las características del coeficiente de correlación parcial de la siguiente manera:

1. El coeficiente de correlación parcial $r_{YX/Z_1,Z_2,\ldots,Z_p}$ mide la fuerza de la relación lineal entre dos variables X e Y controlando a otras variables $Z_1, Z_2, \ldots, Z_p$.

2. El cuadrado del coeficiente de correlación parcial $r_{YX/Z_1,Z_2,\ldots,Z_p}$ mide la proporción de la suma de cuadrados residual que se calcula (explicada) por la adición de X al modelo de regresión involucrando a $Z_1, Z_2, \ldots, Z_p$, esto es,

$$r^2_{YX/Z_1,Z_2,\ldots,Z_p} = \frac{\text{suma de cuadrados extra debido a la adición de } X \text{ al modelo que contiene } Z_1, Z_2, \ldots, Z_p}{SC \text{ residual (usando solo } Z_1, Z_2, \ldots, Z_p \text{ en el modelo)}} \tag{7.6.20}$$

3. El coeficiente de correlación parcial $r_{YX/Z_1,Z_2,\ldots,Z_p}$ es una estimación del parámetro poblacional $\rho_{YX/Z_1,Z_2,\ldots,Z_p}$, el cual es la correlación entre Y y X en la distribución condicional conjunta de Y y X dadas $Z_1, Z_2, \ldots, Z_p$. El cuadrado de este coeficiente de correlación parcial está dado por la ecuación (7.6.21), donde $\sigma^2_{Y/Z_1,Z_2,\ldots,Z_p}$ es la varianza de la distribución condicional de Y dadas $Z_1, Z_2, \ldots, Z_p$.

$$\rho^2_{YX/Z_1,Z_2,\ldots,Z_p} = \frac{\sigma^2_{Y/Z_1,Z_2,\ldots,Z_p} - \sigma^2_{Y/X,Z_1,Z_2,\ldots,Z_p}}{\sigma^2_{Y/Z_1,Z_2,\ldots,Z_p}} \tag{7.6.21}$$

4. El estadístico F parcial, $F(X/Z_1, Z_2, \ldots, Z_p)$, se usa para probar la hipótesis nula $H_0: \rho_{YX/Z_1,Z_2,\ldots,Z_p} = 0$.

5. El coeficiente de correlación parcial (de primer orden) $r_{YX/Z}$ es un ajuste de la correlación de orden cero r_{YX}, el cual toma en cuenta el efecto de la variable en control Z, lo cual se observa en la ecuación 7.6.22.

$$r_{YX/Z} = \frac{r_{YX} - r_{YZ}r_{XZ}}{\sqrt{(1-r^2_{YZ})(1-r^2_{XZ})}} \tag{7.6.22}$$

Las correlaciones parciales de mayor orden se calculan al aplicar esta ecuación, pero se usan los coeficientes de correlación parciales de menor orden inmediato.

6. La correlación parcial $r_{rx/z}$ se puede definir como la correlación de los residuales de la regresión lineal de Y sobre Z y de X sobre Z, esto es, $r_{rx/z} = r_{\hat{r}-r,\hat{x}-x}$.

La analogía de la correlación y la regresión múltiple permite expresar el modelo de regresión $\mu_{r/x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$ en términos de los coeficientes de correlación parcial y varianzas condicionales. Por ejemplo, cuando K=3, la representación es

$$\mu_{r/x_1,x_2,x_3} = \mu_r + \rho_{r1/23}\frac{\sigma_{r/23}}{\sigma_{1/23}}\left(X_1 - \mu_{x_1}\right) + \rho_{r2/13}\frac{\sigma_{r/13}}{\sigma_{2/13}}\left(X_2 - \mu_{x_2}\right) + \rho_{r3/12}\frac{\sigma_{r/12}}{\sigma_{3/12}}\left(X_3 - \mu_{x_3}\right) \qquad (7.6.23)$$

donde, por ejemplo, $\rho_{r1/23} = \rho_{rx_1/x_2x_3}$, y donde se define

$$\beta_1 = \rho_{r1/23}\frac{\sigma_{r/23}}{\sigma_{1/23}}, \qquad \beta_2 = \rho_{r2/13}\frac{\sigma_{r/13}}{\sigma_{2/13}}, \qquad \beta_3 = \rho_{r3/12}\frac{\sigma_{r/12}}{\sigma_{3/12}}$$

Se observa la similitud entre esta representación y aquella del caso lineal, donde β_1 es igual a $\rho(\sigma_r/\sigma_x)$. También se puede observar que

$$\beta_0 = \mu_r - \beta_1\mu_{x_1} - \beta_2\mu_{x_2} - \beta_3\mu_{x_3} \qquad (7.6.24)$$

Para encontrar los coeficientes $\beta_0, \beta_1, \beta_2$ y β_3 se pueden sustituir los parámetros poblacionales en las ecuaciones anteriores, por sus correspondientes estimadores, los cuales son:

$$\hat{\mu}_r = \bar{Y}, \quad \hat{\beta}_1 = r_{r1/23}\frac{S_{r/23}}{S_{1/23}}, \quad \hat{\mu}_{x_1} = \bar{X}_1, \quad \hat{\beta}_2 = r_{r2/13}\frac{S_{r/13}}{S_{2/13}},$$

$$\hat{\mu}_{x_2} = \bar{X}_2, \quad \hat{\beta}_3 = r_{r3/12}\frac{S_{r/12}}{S_{3/12}}, \quad \hat{\mu}_{x_3} = \bar{X}_3$$

Este método proporciona los mismos resultados que el método de mínimos cuadrados.

7.6.4.- Coeficiente de Correlación Multiple-Parcial.

El coeficiente de correlación parcial-múltiple se emplea para describir la relación total entre una variable dependiente y dos o más variables independientes mientras se controlan aun otras variables. Por ejemplo, si se consideran las variables independientes X_1, X_1^2, X_2, X_2^2 y X_1X_2; el modelo de regresión completo es de la forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon \qquad (7.6.25)$$

Este modelo es completo de segundo orden ya que incluye todas las variables posibles, hasta términos de segundo orden. En estos modelos se necesita conocer si algunos de los términos de segundo orden es importante o si el modelo de primer orden $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ es adecuado. Existen dos maneras para representar estas cuestiones como un problema de prueba de hipótesis, una forma es probar $H_0: \beta_{11} = \beta_{22} = \beta_{12} = 0$ (todos los coeficientes de segundo orden son cero) y la otra forma es probar la hipótesis $H_0 = \rho_{Y(X_1^2, X_2^2, X_1X_2)/X_1, X_2} = 0$, donde $\rho_{Y(X_1^2, X_2^2, X_1X_2)/X_1, X_2}$ es la correlación parcial-múltiple poblacional de Y con las variables de segundo orden, controlando los efectos de las variables de primer orden. La correlación parcial múltiple poblacional se estima por la correlación parcial-múltiple muestral $r_{Y(X_1^2, X_2^2, X_1X_2)/X_1, X_2}$, el cual describe la contribución múltiple total de la adición de términos de segundo orden al modelo después de que los efectos de los términos de primer orden se controlan. Dos ecuaciones equivalentes para calcular $r^2_{Y(X_1^2, X_2^2, X_1X_2)/X_1, X_2}$ son:

$$r^2_{Y(X_1^2, X_2^2, X_1X_2)/X_1, X_2} = \frac{SC \text{ residual (sólo } X_1 \text{ y } X_2 \text{ en el modelo)} - SC \text{ residual (todos los términos de primer y segundo orden en el modelo)}}{SC \text{ residual (sólo } X_1 \text{ y } X_2 \text{ en el modelo)}} \qquad (7.6.26)$$

$$r^2_{Y(X_1^2, X_2^2, X_1X_2)/X_1, X_2} = \frac{\text{suma de cuadrados extra debido a la adición de los términos de segundo orden } X_1^2, X_2^2 \text{ y } X_1X_2 \text{ al modelo que contiene solo los términos de primer orden } X_1 \text{ y } X_2}{SC \text{ residual (sólo } X_1 \text{ y } X_2 \text{ en el modelo)}}$$

y

$$r^2_{Y(X_1^2, X_2^2, X_1X_2)/X_1, X_2} = \frac{R^2_{Y/X_1, X_2, X_1^2, X_2^2, X_1X_2} - R^2_{Y/X_1, X_2}}{1 - R^2_{Y/X_1, X_2}} \qquad (7.6.27)$$

En general, el coeficiente de correlación parcial-múltiple poblacional se escribe como:

$$\rho_{Y(X_1,X_2,\ldots,X_k)/Z_1,Z_2,\ldots,Z_p} \qquad (7.6.28a).$$

y el correspondiente coeficiente de correlación parcial múltiple muestral como:

$$r_{Y(X_1,X_2,\ldots,X_k)/Z_1,Z_2,\ldots,Z_p} \qquad (7.6.28b).$$

Para probar ya sea $H_0: \rho_{Y(X_1,X_2,\ldots,X_k)/Z_1,Z_2,\ldots,Z_p} = 0$ o $H_0: \beta_{X_1} = \beta_{X_2} = \ldots = \beta_{X_k} = 0$, se calcula el estadístico F_c, ecuación (7.6.29), y se rechaza H_0: al nivel de significancia α si $F_c \geq F_{k,n-p-k-1,1-\alpha}$.

$$F_c = \frac{[SC \text{ residual (sólo } Z_1, \ldots, Z_p \text{ en el modelo)} - SC \text{ residual } (X_1, \ldots, X_k \text{ y } Z_1, \ldots, Z_p \text{ en el modelo)}]/k}{SC \text{ residual } (X_1, \ldots, X_k \text{ y } Z_1, \ldots, Z_p \text{ en el modelo})/(n-p-k-1)} \qquad (7.6.29)$$

Esta prueba es muy útil cuando se quiere probar si los términos de mayor orden o las interacciones de segundo, tercer, y mayor orden son importantes, una vez que se consideran los términos de primer orden en el modelo.

7.7.- CONCEPTO DE INTERACCIÓN.

En esta sección se hace una breve discusión del concepto de "interacción" estadística. Cuando dos variables independientes "interactuan" para afectar una variable dependiente, esta interacción se puede representar en términos de un modelo de regresión.

Para ayudar a ilustrar este concepto, se puede considerar el siguiente ejemplo simple: Es de interés determinar como dos variables independientes (T) y (C), afectan conjuntamente a (Y), además, si sólo se evalúan dos niveles particulares de T (T_0 y T_1) y dos niveles de C (C_0 y C_1) y se obtienen observaciones de Y para cada una de las cuatro combinaciones de los niveles de T y C, $(T_0,C_0),(T_0,C_1),(T_1,C_0)$, y (T_1,C_1), a este experimento se le llama experimento factorial completo,

completo porque las observaciones sobre Y se obtienen para todas las combinaciones de todas las variables independientes (o factores). Con el experimento factorial es fácil detectar y medir los efectos de interacción cuando existen.

Si se consideran dos gráficas como en la figura 7.3, las cuales se construyen a partir de dos conjuntos de datos que se obtienen a través de una vía de experimentación como la descrita anteriormente, se observa en la figura 7.3a que la razón de cambio en Y como una función de T es la misma para cada nivel de C, es decir, la relación entre Y y T no depende de C. Es importante resaltar que no se dice Y y C no están relacionadas, sino que la relación entre Y y T es independiente de la relación entre Y y C. Cuando este es el caso, se dice que T y C no interactuan o que no hay efecto de interacción *T-por-C*. Esto significa que se pueden investigar los efectos de T y C sobre Y independientemente uno de otro y que se puede hablar de los efectos separados, algunas veces llamado "efectos principales", de T y C sobre Y.

Una manera de cuantificar la relación descrita en la figura 7.3.a es con un modelo de regresión de la forma

$$\mu_{Y|T,C} = \beta_0 + \beta_1 T + \beta_2 C \qquad (7.7.1)$$

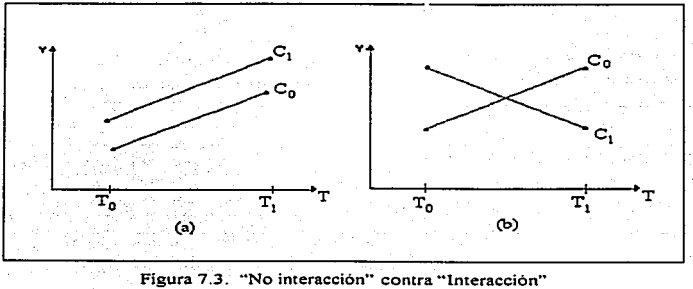


Figura 7.3. "No interacción" contra "Interacción"

Aquí, el cambio en la media de Y por un cambio unitario en T es igual a β_1, sin considerar los niveles de C. De hecho, si se cambian los niveles de C en la ecuación (7.7.1) se tiene sólo el efecto de desviar la línea recta que relaciona $\mu_{Y|T,C}$ y T ya sea hacia arriba o hacia abajo sin afectar el valor de la pendiente β_1, como se observa en la figura 7.3a. En particular, $\mu_{Y|T,C_0} = (\beta_0 + \beta_2 C_0) + \beta_1 T$ y $\mu_{Y|T,C_1} = (\beta_0 + \beta_2 C_1) + \beta_1 T$.

En general, se puede decir que la "no interacción" es sinónimo de "paralelismo" en el sentido que las curvas de respuesta Y contra T para diferentes valores fijos de C son paralelas, es decir, todas las curvas de respuesta (las cuales pueden ser lineales o no lineales) tienen la misma forma general y son diferentes una de otra solo por las constantes independientes aditivas de T, como se muestra en la figura 7.4.

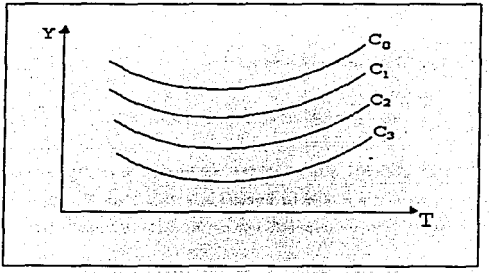


Figura 7.4. Ilustración gráfica de "no interacción"

En contraste, la figura 7.3b describe la situación donde la relación entre Y y T depende de C, se observa que Y incrementa al aumentar T con $C = C_0$ pero decrece al aumentar T cuando $C = C_1$. En este caso el comportamiento de Y como una función de T no se puede considerar independientemente de C, por lo tanto se dice que T y C interactuan o que existe un efecto de interacción *T-por-C*. Esto significa que no tiene sentido hablar acerca de los efectos separados o principales de T y C sobre Y, ya que T y C no operan independientemente uno de otro con respecto a sus efectos sobre Y.

Una manera de calcular matemáticamente tales efectos de interacción es considerar un modelo de regresión de la forma

$$\mu_{Y/T,C} = \beta_0 + \beta_1 T + \beta_2 C + \beta_{12} TC \qquad (7.7.2)$$

Aquí, el cambio en el valor medio de Y cuando sucede un cambio unitario en T es igual a $\beta_1 + \beta_{12} C$, el cual depende del nivel de C, es decir, la introducción de término $\beta_{12} TC$ al modelo de regresión es una manera de calcular la interacción de dos variables independientes.

Para el ejemplo, cuando $C = C_0$, el modelo (7.7.2) se puede escribir como

$$\mu_{Y/T,C_0} = (\beta_0 + \beta_2 C_0) + (\beta_1 + \beta_{12} C_0) T \qquad (7.7.2a)$$

y, cuando $C = C_1$, el modelo (7.7.2) se escribe como:

$$\mu_{Y/T,C_1} = (\beta_0 + \beta_2 C_1) + (\beta_1 + \beta_{12} C_1) T \qquad (7.7.2b)$$

En la figura II.3b se observa que el efecto de interacción β_{12} es negativo, el efecto lineal $(\beta_1 + \beta_{12} C_0)$ de T a C_0 es positivo y el efecto lineal $(\beta_1 + \beta_{12} C_1)$ de T a C_1 es negativo. Se espera un efecto de interacción negativo, porque la pendiente de la relación lineal entre Y y T decrece cuando C cambia de C_0 a C_1.

7.8.- MÉTODOS DE SELECCIÓN DE VARIABLES.

En esta sección se analizan los procedimientos para seleccionar el mejor modelo de regresión cuando se tiene una variable dependiente Y y un conjunto de k variables independientes $X_1, X_2, \ldots, X_k$. En este caso se busca determinar el mejor subconjunto de variables, es decir, las más importantes, de las k variables independientes y el correspondiente modelo de regresión de mejor ajuste para describir la relación entre Y y las X's.

Existen diversos procedimientos estadísticos básicos para seleccionar la mejor ecuación de regresión, sin embargo, los que más se emplean son:

1.- Procedimiento de todas las posibles regresiones.
2.- Procedimiento de eliminación backward.
3.- Procedimiento de selección forward.
4.- Procedimiento de regresión stepwise.

Antes de proceder a detallar los procedimientos, algunas consideraciones importantes son:

a) Algunas de las k variables independientes pueden consistir de funciones de mayor orden de unas cuantas variables básicas. En la practica, el uso de algún procedimiento de selección de variables da resultados más aceptables cuando no existen muchos términos de mayor orden, la razón es que el modelo que contiene tales términos es difícil de interpretar.

b) Es posible llegar a diferentes soluciones al usar los cuatro diferentes métodos. Cuando esto sucede, es necesario reflexionar acerca de los resultados y elegir el mejor modelo basándose en consideraciones prácticas observando las variables bajo estudio, la naturaleza de los datos y las interpretaciones que pueden hacerse con los diferentes modelos candidatos.

c) Algunas veces hasta un simple procedimiento proveerá un número de modelos razonablemente buenos, de los cuales se tendrá que hacer una elección.

7.8.1.- Procedimiento de todas las posibles regresiones.

El procedimiento de todas las posibles regresiones requiere fijar todas las posibles ecuaciones de regresión asociadas con todas las combinaciones posibles de las variables independientes. En particular, si $k=3$, es decir, X_1, X_2, X_3, se tienen que fijar siete modelos correspondientes a las siguientes siete combinaciones de las variables independientes: 1) X_1; 2) X_2; 3) X_3; 4) X_1, X_2; 5) X_1, X_3; 6) X_2, X_3; 7) X_1, X_2, X_3. Para k variables independientes, el número de modelos a fijar es $2^k - 1$. Posteriormente se reparten las diferentes ecuaciones obtenidas dentro de subconjuntos de 1,2,3..., y k variables, y se ordenan los modelos dentro de cada grupo en base a algún criterio como por ejemplo R^2. Se seleccionan los modelos más importantes de cada subgrupo y se elige el mejor.

7.8.2.- Procedimiento de Eliminación Backward.

La técnica de eliminación backward comienza calculando estadísticos para el modelo, incluyendo todas las variables independientes.. Entonces las variables son eliminadas del modelo una por una hasta que todas las variables que queden en el modelo produzcan un estadístico F al nivel de significancia determinado. En cada paso, la variable que muestra la más baja contribución al modelo se elimina.

En el procedimiento de eliminación bakward, se realizan los siguientes pasos: [16]

1) Determinar la ecuación de regresión ajustada que contenga todas las variables independientes.

2) Calcular el estadístico F parcial para cada una de las variables en el modelo como si fuera la última variable que entra al modelo.

3) Enfocarse sobre el valor de F parcial más bajo (F_b).

4) Comparar este valor con un valor crítico preseleccionado de la distribución F (F_s), es decir, probar la significancia del valor de F parcial más bajo. a) Si $F_b < F_s$, remover del modelo la variable

en consideración, recalcular la ecuación de regresión para las variables restantes y repetir los pasos 2,3 y 4. b) Si $F_b > F_c$, adoptar la ecuación de regresión completa.

7.8.3.- Procedimiento de Selección Forward.

La técnica de selección Forward comienza sin ninguna variable en el modelo. Para cada una de las variables independientes, FORWARD calcula los estadísticos F que reflejan la contribución de la variable al modelo si es incluida. Si el estadístico F no tiene un nivel de significancia mayor que el valor crítico seleccionado, el procedimiento forward termina. De otra manera, Forward adiciona la variable que tiene el mayor estadístico F al modelo, calcula el estadístico F de nuevo para las variables que aun quedan en el modelo, y el proceso de evaluación se repite. Así, las variables se adicionan una por una al modelo hasta que no queden variables que produzcan un estadístico F significante. Una vez que la variable esta en el modelo, se queda.

En este procedimiento se realizan los siguientes pasos: [16]

1) Seleccionar la primer variable a entrar en el modelo, la cual debe tener la mayor correlación con la variable dependiente, y entonces fijar la ecuación de regresión lineal correspondiente. Se calcula el estadístico F y si no es significativo, se termina el procedimiento con la conclusión de que no existe una variable independiente importante para la predicción de la variable dependiente. Si el estadístico F es significativo, se incluye esta variable en el modelo y se realiza el paso 2.

2) Calcular el estadístico F parcial asociado con cada una de las variables restantes basadas sobre una ecuación de regresión que contiene aquella variable y la variable inicialmente seleccionada. Por ejemplo, si k=3, y en el paso 1 se eligió a la variable X_1, las F's parciales a calcular son:

$$F(X_2/X_1) \text{ parcial} \quad \text{y} \quad F(X_3/X_1) \text{ parcial}$$

3) Enfocarse a la variable con un valor más grande del estadístico F parcial.

4) Probar la significancia del estadístico F parcial asociado con la variable seleccionada en el paso 3. a) Si esta prueba es significante, adicionar la nueva variable a la ecuación de regresión. b) Si esta prueba no es significante, usar en el modelo solamente la variable adicionada en el paso 1.

5) En cada paso subsecuente, determinar el estadístico F parcial para aquellas variables que aun no están en el modelo y adicionar al modelo la variable que tenga el valor de F parcial más grande si es estadísticamente significante. En algún paso, si la F parcial no es significante, no incluir más variables en el modelo y el proceso se termina.

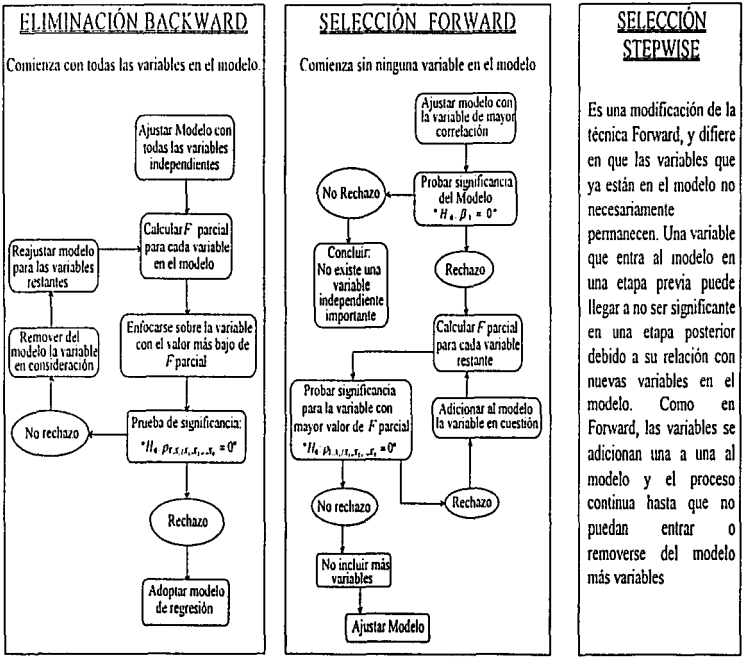
7.8.4.- Procedimiento de Regresión Stepwise

El método Stepwise es una modificación de la técnica de selección forward y difiere en que las variables que ya están en el modelo no necesariamente permanecen ahí, es decir, stepwise hace una reexaminación de las variables incorporadas en el modelo en un paso previo. Una variable que entra en una etapa temprana puede llegar a ser superflua en una etapa posterior debido a si relación con nuevas variables en el modelo. Como en el método de selección forward, las variables se adicionan una a una al modelo, y en cada etapa se hace una prueba del estadístico F parcial para cada variable presente en el modelo, tratándola como si fuera la variable que se adiciona al último. La variable con el estadístico F parcial más pequeño no significante se elimina y el modelo se reajusta con las variables restantes, se obtienen las F's parciales y se examinan similarmente y así sucesivamente. El proceso continua hasta que no puedan entrar más variables o removerse del modelo.

En la figura 7.5 se presenta un resumen de los métodos de selección de variables.

Figura 7.5.- MÉTODOS DE SELECCIÓN DE VARIABLES

Para determinar el mejor subconjunto de variables, de las k variables independientes, que proporcionan el modelo de regresión de mejor ajuste, se emplea alguno de los siguientes procedimientos:



VIII. ESTUDIOS DE CASO PARA REGRESIÓN MÚLTIPLE.

Con el objetivo de mostrar las herramientas del análisis de regresión múltiple que se revisaron en el capitulo anterior, a continuación se presentan diversos estudios de caso para ilustrar algunas posibles aplicaciones de este análisis estadístico.

8.1. *ESTUDIO DE CASO No. 1*

Este primer caso se trata de un experimento sobre un proceso farmacéutico que es influenciado por diversos factores, el cual es una granulación en lecho fluidizado. En este proceso influyen factores tales como: Temperatura del aire de entrada al granulador de lecho fluido, presión del aire de atomización, cantidad de solución aglutinante, entre otros. En este estudio sólo se tomaron en cuenta los tres factores mencionados, para evaluar su influencia sobre la friabilidad de gránulos de α-Lactosa monohidratada, elaborados en un granulador de lecho fluidizado. La friabilidad se determinó como peso perdido de los gránulos después de 100 ciclos en el fragilizador (aparato que ayuda a determinar la friabilidad). [21]

La granulación en lecho fluidizado se ha estudiado ampliamente, pero en la mayoría de los casos sólo se evalúa el efecto de una variable, y en algunos otros se emplearon diseños factoriales 2^n y 3^2, los cuales también son limitados. Se ha mostrado que el empleo de la metodología de superficie de respuesta con diseños factoriales es un método efectivo que proporciona la máxima información con un limitado número de experimentos. Diversos autores han empleado análisis de regresión, por ejemplo Lindber et al (1985-1987) empleo análisis de regresión en el estudio de la influencia de la composición y variables de proceso sobre el tiempo de desintegración, dureza y friabilidad de tabletas. En granulación, Wehrlé et al (1989) empleo el análisis de regresión stepwise para comparar diferentes granuladores. Posteriormente, Bos et al (1991 a,b,c) aplicó el análisis de regresión para estudiar tabletas elaboradas por compresión directa [21].

A continuación se plantea el estudio de caso, en el cual se aplican las diversas técnicas de análisis de regresión sobre un problema del ámbito farmacéutico.

8.1.1.- PROBLEMA.

El objetivo del estudio de Merkku y col. (1993) sobre el proceso de granulación húmeda en lecho fluido fue mostrar la aplicación del método stepwise del análisis de regresión múltiple para encontrar el mejor modelo de predicción, sin embargo, en esta sección se presentan los diferentes métodos del análisis de regresión aplicados a un proceso farmacéutico, mostrando los diferentes criterios e interpretaciones para lograr obtener un modelo de regresión múltiple que ajuste de manera adecuada a los resultados experimentales.

El estudio se realizó en base a un diseño factorial 3^3, donde las variables independientes fueron: Temperatura del aire de entrada, presión del aire de atomización y cantidad de solución aglutinante, mientras que la variable de respuesta fue el porcentaje de peso perdido por friabilidad. Los niveles de las variables independientes se muestran en el cuadro 8.1 y la matriz de experimentación en el cuadro 8.2. Es importante mencionar que en los extremos del diseño los granulados se hicieron por duplicado y el punto central se realizo por cuadruplicado, por lo tanto fueron 38 experiencias en total.

Variable	Niveles			
	-1	0	+1	Dimensión
Temperatura del aire de entrada (T)	40	50	60	(°C)
Presión del aire de atomización (p)	1.0	1.5	2.0	(bar)
Cantidad de solución aglutinante (m)	150	300	450	(g)

Cuadro 8.1. Niveles de las variables independientes.

Experimento	Variables		
	T	p	m
1*	-1	-1	-1
2	-1	-1	0
3*	-1	-1	+1
4	-1	0	-1
5	-1	0	0
6	-1	0	+1
7*	-1	+1	-1
8	-1	+1	0
9*	-1	+1	+1
10	0	-1	-1
11	0	-1	0
12	0	-1	+1
13	0	0	-1
14**	0	0	0
15	0	0	+1
16	0	+1	-1
17	0	+1	0
18	0	+1	+1
19*	+1	-1	-1
20	+1	-1	0
21*	+1	-1	+1
22	+1	0	-1
23	+1	0	0
24	+1	0	+1
25*	+1	+1	-1
26	+1	+1	0
27*	+1	+1	+1
* Duplicado	** Cuadruplicado		

Cuadro 8.2. Matriz experimental

Una vez que se construye el diseño factorial y se obtienen los resultados, se puede estudiar, a través de regresión múltiple, la dependencia de la friabilidad, en función de las variables independientes *T, p* y *m* en el diseño factorial 3^3. El método Stepwise de regresión múltiple se ha empleado en estudios de granulación con dos variables independientes. Bos et al, aplicó la misma técnica en un diseño factorial 3^4 en la evaluación de estabilidad de tabletas. Por lo tanto, en este caso de estudio se llegará a un modelo de regresión múltiple mostrando e interpretando cada etapa en que se realiza. La forma general del modelo de regresión que describe el comportamiento de la friabilidad de los gránulos elaborados en un granulador de lecho fluidizado en función de las tres variables independientes en estudio es la ecuación (8.1.1), la cual se deberá simplificar lo más posible hasta obtener un modelo que contenga sólo aquellos factores que influyen de manera significativa sobre la respuesta.

$$FR = \beta_0 + \beta_1 T + \beta_2 p + \beta_3 m + \beta_{12} Tp + \beta_{13} Tm + \beta_{23} pm + \beta_{11} T^2 + \beta_{22} p^2 + \beta_{33} m^2 + \beta_{123} Tpm \qquad (8.1.1)$$

Donde *T, p, m* son las variables independientes. *FR* es la variable de respuesta (friabilidad) y las β's son los diferentes coeficientes de regresión que se deben estimar a partir de los resultados experimentales, los cuales se presentan a continuación. [21]

8.1.2.- RESULTADOS EXPERIMENTALES.

El análisis estadístico, es decir, el ajuste de los resultados a un modelo de regresión que mejor describe la dependencia de la friabilidad en función de *T, p, m*, se realizó con los resultados que se presentan en el cuadro 8.3; se observa que se tomaron las unidades reales de las variables independientes, esto se debe a que no se controlaron estrictamente los niveles de cada variable, y como se desea llegar a un modelo de predicción, es mejor considerar estas variaciones.

Experiencia	Temperatura del aire de entrada (°C)	Presión del aire de atomización (bar)	Cantidad de solución aglutinante (g)	Friabilidad (%)	Experiencia	Temperatura del aire de entrada (T) (°C)	Presión del aire de atomización (bar)	Cantidad de solución aglutinante (g)	Friabilidad (%)
1a	41.4	1.0	162	24.8	14c	51.0	1.4	308	47.8
1b	42..4	1.0	152	21.0	14d	52.7	1.5	308	37.2
2	44.6	0.9	307	18.6	15	48.9	1.5	457	24.3
3a	44.1	1.0	459	5.6	16	51.6	1.9	157	47.5
3b	41.4	1.0	457	15.0	17	51.1	2.0	306	29.0
4	41.2	1.5	156	36.9	18	49.5	2.0	453	24.2
5	42.6	1.4	306	16.1	19a	61.4	1.0	160	45.5
6	44.2	1.5	459	8.3	19b	59.0	1.0	157	45.5
7a	43.8	1.9	167	45.3	20	61.5	1.0	313	14.1
7b	41.5	2.0	158	53.3	21a	58.0	0.9	459	13.0
8	44.6	2.0	305	37.4	21b	59.6	1.0	459	6.2
9a	43.4	2.0	461	23.3	22	59.7	1.5	157	37.8
9b	41.3	2.0	459	23.9	23	58.8	1.5	309	31.3
10	53.7	1.0	162	24.2	24	58.9	1.6	457	20.9
11	51.9	1.0	306	11.8	25a	61.4	1.9	162	64.8
12	49.9	1.0	457	8.8	25b	61.0	2.0	159	47.9
13	51.7	1.5	158	37.4	26	59.7	1.9	309	38.2
14a	51.2	1.5	300	44.2	27a	59.3	2.0	456	31.6
14b	50.5	1.4	310	28.9	27b	58.6	2.0	456	51.0

Cuadro 8.3. Resultados del porcentaje de friabilidad

8.1.3.- AJUSTE DEL MODELO DE REGRESIÓN MÚLTIPLE.

A continuación se presenta la aplicación de las diferentes técnicas del análisis de regresión múltiple descritas en el capítulo 7, y con ayuda del paquete estadístico SAS se determinan los modelos de regresión. En el programa 8.1. se muestran los diferentes procedimientos para lograr lo anterior, y cada uno se describe posteriormente. Se deben crear en el programa todas las variables necesarias que sean función de otras variables básicas que se quieran probar en el modelo, antes de introducir los resultados experimentales.

Programa 8.1. Programa para obtener un Modelo de Análisis de Regresión, a través de diversas técnicas.

```
OPTIONS PS=60 nodate nonumber;
DATA REGMUL1;
INPUT    Batch $
         T  /* Temperatura del aire de entrada (°C)  -1=40  0=50  +1=60  */
         p  /* Presión del aire de atomización (Bar) -1=1.0 0=1.5 +1=2.0 */
         m  /* Cantidad de solución aglutinante (g)  -1=150 0=300 +1=450 */
         Fr /* Friabilidad (%), perdida de masa en porcentaje            */
;
Tp=T*p; Tm=T*m; pm=p*m; TT=T*T; pp=p*p; mm=m*m; Tpm=T*p*m;

CARDS;
1a  41.4 1.0 162 24.8   1b  42.4 1.0 152 21.0   2  44.6 0.9 307 18.6
3a  44.1 1.0 459  5.6   3b  41.4 1.0 457 15.0   4  41.2 1.5 156 36.9
5   42.6 1.4 306 16.1   6   44.2 1.5 459  8.3   7a 43.8 1.9 167 45.3
7b  41.5 2.0 158 53.3   8   44.6 2.0 305 37.4   9a 43.4 2.0 461 23.3
9b  41.3 2.0 459 23.9   10  53.7 1.0 162 24.2   11 51.9 1.0 306 11.8
12  49.9 1.0 457  8.8   13  51.7 1.5 158 37.4   14a 51.2 1.5 300 44.2
14b 50.5 1.4 310 28.9   14c 51.0 1.4 308 47.8   14d 52.7 1.5 308 37.2
15  48.9 1.5 437 24.3   16  51.6 1.9 157 47.5   17  51.1 2.0 306 29.0
18  49.5 2.0 453 24.2   19a 61.4 1.0 160 45.5   19b 59.0 1.0 157 45.5
20  61.5 1.0 313 14.1   21a 58.0 0.9 459 13.0   21b 59.6 1.0 459  5.2
22  59.7 1.5 157 37.8   23  58.8 1.5 309 31.3   24  58.9 1.6 457 20.9
25a 61.4 1.9 162 64.8   25b 61.0 2.0 159 47.9   26  59.7 1.9 309 38.2
27a 59.3 2.0 456 31.6   27b 58.6 2.0 456 51.0
;
proc corr;
proc reg;
model Fr=t p m Tp Tm pm TT pp mm Tpm  /SELECTION=BACKWARD
                                       SLE=0.05 SLS=0.05;
model Fr=t p m Tp Tm pm TT pp mm Tpm  /SELECTION=FORWARD
                                       SLE=0.05 SLS=0.05;
model Fr=t p m Tp Tm pm TT pp mm Tpm  /SELECTION=STEPWISE
                                       SLE=0.05 SLS=0.05;
run;
```

8.1.3.1.- Matriz de correlación

Debido a que existen más de dos variables independientes, no se puede observar de manera gráfica la posible relación entre dichas variables y la variable de respuesta, sin embargo, como se mencionó en la sección 7.6.1. se puede comenzar por encontrar una matriz de correlación entre todas las variables, a través de la cual se obtienen todos los coeficientes de correlación de orden cero para tener una idea de la posible correlación entre las variables independientes (Temperatura del aire de entrada, presión del aire de atomización y cantidad de solución aglutinante) y la variable de respuesta (friabilidad de los gránulos). Esta matriz se muestra en la salida 8.1.

Salida 8.1. Matriz de correlación entre todas las variables

The SAS System
Correlation Analysis

11 'VAR' Variables: T P M FR TP TM PM TT PP MM TPM

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
T	38	51.2395	7.2172	1947	41.2000	61.5000
P	38	1.4789	0.4154	56.2000	0.9000	2.0000
M	38	307.4474	124.5938	11683	152.0000	461.0000
FR	38	30.0421	14.9980	1142	5.2000	64.8000
TP	38	75.7676	23.9641	2879	40.1400	122.0000
TM	38	15707	6652	596864	6427	27356
PM	38	455.8079	235.7149	17321	152.0000	922.0000
TT	38	2676	741.0779	101696	1697	3782
PP	38	2.3553	1.2327	89.5000	0.8100	4.0000
MM	38	109639	77382	4166281	23104	212521
TPM	38	23277	12375	884510	6445	54082

The SAS System
Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 38

	T	P	M	FR	TP	TM
T	1.00000	-0.00440	-0.05312	0.26481	0.44160	0.26654
	0.0	0.9791	0.7515	0.1081	0.0055	0.1058
P	-0.00440	1.00000	0.02202	0.54776	0.88560	0.01741
	0.9791	0.0	0.8956	0.0004	0.0001	0.9174
M	-0.05312	0.02202	1.00000	-0.59782	-0.00619	0.93763
	0.7515	0.8956	0.0	0.0001	0.9706	0.0001
FR	0.26481	0.54776	-0.59782	1.00000	0.60947	-0.49251
	0.1081	0.0004	0.0001	0.0	0.0001	0.0017
TP	0.44160	0.88560	-0.00619	0.60947	1.00000	0.13272
	0.0055	0.0001	0.9706	0.0001	0.0	0.4270
TM	0.26654	0.01741	0.93763	-0.49251	0.13272	1.00000
	0.1058	0.9174	0.0001	0.0017	0.4270	0.0
PM	-0.04756	0.57694	0.79671	-0.16236	0.48795	0.74403
	0.7767	0.0001	0.0001	0.3301	0.0019	0.0001
TT	0.99868	-0.00434	-0.06135	0.26897	0.44121	0.25795
	0.0001	0.9794	0.7144	0.1025	0.0056	0.1179
PP	-0.00499	0.99536	0.03341	0.53520	0.88041	0.02822
	0.9763	0.0001	0.8422	0.0005	0.0001	0.8665
MM	-0.05905	0.02457	0.99081	-0.59157	-0.00653	0.92612
	0.7247	0.8836	0.0001	0.0001	0.9690	0.0001
TPM	0.20697	0.55834	0.75865	-0.08512	0.59540	0.80525
	0.2125	0.0003	0.0001	0.6114	0.0001	0.0001

En la salida 8.1 se obtienen los coeficientes de correlación de orden cero, llamados también de Pearson, en forma equivalente a la matriz de correlación 7.6.4. De estos coeficientes, los de mayor interés son aquellos que relacionan a las variables independientes con la variable de respuesta, los cuales se resaltan en la salida 8.1. De estos se observa, en función de la hipótesis nula que se prueba (Ho:Rho=0), que las variables que pueden estar correlacionadas con la Friabilidad son: P, M, TP, TM, PP y MM; debido a que todas ellas tiene un nivel de significancia menor al 0.05 %, por lo que caen en la zona de rechazo con un nivel de significancia del 5%. Así, la variable que mayor correlación tiene con la friabilidad es la interacción TP, debido a que tiene el coeficiente de correlación más cercano a 1 (0.60947) cuya probabilidad de que sea diferente de cero es alta. Los factores que le siguen en importancia de correlación de orden cero con la friabilidad son: M, MM, P, PP y TM. . Sin embargo, esto no significa que todas las variables mencionadas sean estadísticamente importantes en la predicción de la friabilidad, por lo que se requiere evaluar correlaciones de mayor orden para determinar cuales variables realmente influyen sobre la respuesta y en que magnitud, lo cual se obtiene a través de los diferentes procedimientos del análisis de regresión múltiple y pruebas de F total y parcial.

8.1.3.2.- Procedimiento Backward.

De acuerdo al procedimiento del programa 8.1. se comienza con el análisis de regresión múltiple, obteniéndose un modelo a través del procedimiento de eliminación Backward, el cual se describe en la sección 7.8.2. Para este procedimiento se obtiene la salida 8.2. donde se pueden observar el número de pasos necesarios para llegar al modelo que involucra sólo a las variables que ayudan a predecir la friabilidad.

Comienza ajustando un modelo con todas las variables, proporcionando un cuadro de Análisis de Varianza, semejante al que se describe en la sección 7.5, donde se resume la prueba de hipótesis acerca del modelo completo, es decir, se evalúan las 10 variables independientes al mismo tiempo, y se observa que el modelo es estadisticamente significativo debido a que el valor *Prob>F* es de 0.0001, además de tener un coeficiente de determinación diferente de cero (0.7645)

Posteriormente proporciona información acerca de las pruebas de hipótesis de *F's* parciales (con la suma de cuadrados tipo II) donde se observa que el valor de *F* parcial más bajo lo tiene la variable TT, cuyo efecto no es significativamente importante, por lo que en el siguiente paso es eliminada, lo cual se observa en la etapa 1 del proceso de eliminación de la salida 8.2, donde se obtiene un modelo con 9 variables independientes y aun es estadisticamente significativo, debido a que el valor de *Prob>F* es de 0.0001, de hecho, los parámetros del ANOVA son muy parecidos a los que se obtienen en el modelo con todas las variables independientes, sin embargo aún se tienen términos en el modelo cuyo efecto no influye sobre la variable de respuesta que se está evaluando, por lo que el proceso de eliminación continúa hasta obtener un modelo en que todas las variables independientes tengan un efecto significativo sobre la respuesta, lo cual se consigue hasta la etapa 7, donde sólo quedan tres variables en el modelo. Al final de la salida 8.2 se resumen los pasos para llegar al modelo de regresión múltiple, donde se describen los valores de los coeficientes de correlación parciales, los cuales proporcionan el mejoramiento en la predicción de la respuesta por incluir al modelo la variable que se está eliminando, así se observa que las variables que se eliminaron no ayudan a predecir la respuesta en forma significativa.

De esta forma se llega al siguiente modelo de regresión múltiple, a través del procedimiento Backward:

$$Fr = -24.004883 + 0.89739 * T + 20.236 * P - 0.001392 * TM$$

Salida 8.2. Modelo de regresión múltiple por procedimiento de eliminación.

```
                                The SAS System

        Backward Elimination Procedure for Dependent Variable FR

Step 0    All Variables Entered      R-square = 0.76452509   (p) =11.00000000

                  DF        Sum of Squares       Mean Square        F      Prob>F

Regression        10     6362.96849157      636.29684916         8.77      0.0001
Error             27     1959.80414001       72.58533852
Total             37     8322.77263158

                Parameter       Standard                Type II
Variable         Estimate          Error     Sum of Squares        F      Prob>F

INTERCEP     -147.29114119   127.36956724      97.06682275     1.34      0.2576
T               3.08468134     4.40212245      35.64062258     0.49      0.4895
P             122.07272736    67.72711932     235.80933543     3.25      0.0827
M               0.44075570     0.28345765     175.49648869     2.42      0.1316
TP             -1.82171206     1.09374435     201.36163347     2.77      0.1074
TM             -0.01024674     0.00515318     286.99243775     3.95      0.0570
PM             -0.33175573     0.17120127     272.56607387     3.76      0.0632
TT              0.00264552     0.04076029       0.30577182     0.00      0.9487
PP             -2.96610042    12.63133800       4.00241295     0.06      0.8161
MM              0.00001630     0.00014498       0.91795375     0.01      0.9113
TPM             0.00651017     0.00330405     281.79923076     3.88      0.0591

Bounds on condition number:        852.2396,       48381.22
------------------------------------------------------------------------------

Step 1    Variable TT Removed      R-square = 0.76448835   C(p) = 9.00421258

                  DF        Sum of Squares       Mean Square        F      Prob>F

Regression         9     6362.66271975      706.96252442        10.10      0.0001
Error             28     1960.10991183       70.00392542
Total             37     8322.77263158

                Parameter       Standard                Type II
Variable         Estimate          Error     Sum of Squares        F      Prob>F

INTERCEP     -153.12103100    88.68410994     208.68943314     2.98      0.0953
T               3.34863288     1.65500385     286.58895683     4.09      0.0527
P             121.27011174    65.39380947     240.74466610     3.44      0.0742
M               0.43791454     0.27503214     177.47378641     2.54      0.1226
TP             -1.81606002     1.07070951     201.39072703     2.88      0.1010
TM             -0.01022719     0.00505206     286.87865663     4.10      0.0526
PM             -0.33096684     0.16770517     272.64559374     3.89      0.0584
PP             -2.79290811    12.12473610       3.71442075     0.05      0.8195
MM              0.00001904     0.00013621       1.36844839     0.02      0.8898
TPM             0.00649481     0.00323643     281.91785975     4.03      0.0545

Bounds on condition number:        847.8646,       34933.49
------------------------------------------------------------------------------
```

Step 2 Variable MM Removed R-square = 0.76432393 C(p) = 7.02306554

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	8	6361.29427136	795.16178392	11.76	0.0001
Error	29	1961.47836022	67.63718484		
Total	37	8322.77263158			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-153.68081296	87.08319254	210.64741663	3.11	0.0881
T	3.34711233	1.62675143	286.34111216	4.23	0.0487
P	119.90071390	63.55380447	240.73872393	3.56	0.0693
M	0.44980147	0.25710195	207.02200689	3.06	0.0908
TP	-1.81398794	1.05235342	200.96992888	2.97	0.0954
TM	-0.01023238	0.00496579	287.18550491	4.25	0.0484
PM	-0.33078039	0.16484063	272.35571039	4.03	0.0542
PP	-2.36513648	11.53232892	2.84487963	0.04	0.8389
TPM	0.00649173	0.00318118	281.66391392	4.16	0.0505

Bounds on condition number: 847.8254, 29976.1

Step 3 Variable PP Removed R-square = 0.76398211 C(p) = 5.06225913

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	7	6358.44939173	908.34991310	13.87	0.0001
Error	30	1964.32323985	65.47744133		
Total	37	8322.77263158			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-149.45452516	83.24802337	211.03863894	3.22	0.0827
T	3.35729216	1.59982331	288.35396629	4.40	0.0444
P	113.29984760	53.91917407	289.11051635	4.42	0.0441
M	0.45293861	0.25251574	210.66553171	3.22	0.0829
TP	-1.82023947	1.03498115	202.52744101	3.09	0.0888
TM	-0.01029317	0.00487715	291.64695356	4.45	0.0433
PM	-0.33296130	0.16184966	277.11220791	4.23	0.0484
TPM	0.00653048	0.00312445	286.04555555	4.37	0.0452

Bounds on condition number: 844.8346, 24694.57

Step 4 Variable TP Removed R-square = 0.73964798 C(p) = 5.85245678

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	6155.92195072	1025.98699179	14.68	0.0001
Error	31	2166.85068086	69.89840906		
Total	37	8322.77263158			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-10.56187896	27.20632497	10.53440233	0.15	0.7005
T	0.65650365	0.46343653	140.26878928	2.01	0.1666
P	19.65166844	8.76270620	351.55205942	5.03	0.0322
M	0.06936765	0.13149743	19.45121181	0.28	0.6016
TM	-0.00282097	0.00247441	90.84898392	1.30	0.2630
PM	-0.07549793	0.07131752	78.33303854	1.12	0.2980
TPM	0.00151636	0.00132070	92.14322703	1.32	0.2597

Bounds on condition number: 149.5895, 3536.616

Step 5 Variable M Removed R-square = 0.73731087 C(p) = 4.12043392

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	6136.47073891	1227.29414778	17.96	0.0001
Error	32	2186.30189267	68.32193415		
Total	37	8322.77263158			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-2.07560885	21.69176957	0.62554883	0.01	0.9244
T	0.52578770	0.38718595	125.99166397	1.84	0.1840
P	18.54384053	8.41085153	332.10819046	4.86	0.0348
TM	-0.00157583	0.00073412	314.80894672	4.61	0.0395
PM	-0.04447763	0.03989439	84.92190931	1.24	0.2732
TPM	0.00097849	0.00082990	94.97821007	1.39	0.2471

Bounds on condition number: 57.12231, 643.8249

Step 6 Variable PM Removed R-square = 0.72710731 C(p) = 3.29039346

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	6051.54882960	1512.88720740	21.98	0.0001
Error	33	2271.22380198	68.82496370		
Total	37	8322.77263158			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-18.92754119	15.61535950	101.11845295	1.47	0.2341
T	0.89846940	0.19609174	1444.88724755	20.99	0.0001
P	16.76532908	8.28853047	281.58860694	4.09	0.0513
TM	-0.00170896	0.00072700	380.31109133	5.53	0.0249
TPM	0.00021395	0.00046912	14.31514984	0.21	0.6513

Bounds on condition number: 18.11915, 152.5636

Step 7 Variable TPM Removed R-square = 0.72538731 C(p) = 1.48761166

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	6037.23367976	2012.41122659	29.94	0.0001
Error	34	2285.53895182	67.22173388		
Total	37	8322.77263158			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-24.00488304	10.82142769	330.78040528	4.92	0.0333
T	0.89738860	0.19378022	1441.62370471	21.45	0.0001
P	20.23600365	3.24580205	2612.85402364	38.87	0.0001
TM	-0.00139191	0.00021027	2945.72046114	43.82	0.0001

Bounds on condition number: 1.076877, 9.46152

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination Procedure for Dependent Variable FR

Step	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	TT	9	0.0000	0.7645	9.0042	0.0042	0.9487
2	MM	8	0.0002	0.7643	7.0231	0.0195	0.8898
3	PP	7	0.0003	0.7640	5.0623	0.0421	0.8389
4	TP	6	0.0243	0.7396	5.8525	3.0931	0.0888
5	M	5	0.0023	0.7373	4.1204	0.2783	0.6016
6	PM	4	0.0102	0.7271	3.2904	1.2430	0.2732
7	TPM	3	0.0017	0.7254	1.4876	0.2080	0.6513

8.1.3.3.- Procedimiento Forward.

A continuación se presenta la salida 8.3. donde se obtiene un modelo de regresión múltiple a través del procedimiento FORWARD, el cual comienza por introducir al modelo la variable que tiene mayor correlación con la respuesta, lo cual se obtiene con la matriz de correlaciones de orden cero, en este caso es la variable *TP*, así esta variable es la primera que aparece en el modelo proporcionando un valor de coeficiente de determinación no muy grande, sin embargo la prueba de hipótesis de significancia del modelo (Ho: $\beta o = 0$) se rechaza, por lo que es necesario probar las hipótesis de las variables restantes en orden de importancia, y de esta forma continuar hasta que la hipótesis nula que se prueba no sea rechazada, en ese momento se detiene la introducción de mas variables al modelo concluyendo que todas las variables que están en el modelo son estadísticamente significativas para predecir la respuesta. Así se observa que el modelo que se encuentra sólo contiene dos variables independientes, las cuales son TP y M, quedando el modelo con un coeficiente de determinación de 0.7244. Al final de la salida 8.3 se obtienen los coeficientes de correlación parciales y sus respectivas pruebas de hipótesis, los cuales se observa que son significativos en la predicción de la friabilidad.

Salida 8.3 . Modelo de regresión múltiple por el procedimiento Forward.

The SAS System

Forward Selection Procedure for Dependent Variable FR

Step 1 Variable TP Entered R-square = 0.37145428 C(p) = 38.07024426

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	3091.52955457	3091.52955457	21.28	0.0001
Error	36	5231.24307701	145.31230769		
Total	37	8322.77263158			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	1.14137108	6.56381653	4.39383012	0.03	0.8629
TP	0.38143906	0.08269698	3091.52955457	21.28	0.0001

Bounds on condition number: 1, 1

Step 2 Variable M Entered R-square = 0.72436498 C(p) = -0.39516588

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	6028.72504814	3014.36252407	45.99	0.0001
Error	35	2294.04758344	65.54421667		
Total	37	8322.77263158			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	23.30177864	5.51288240	1170.99506913	17.87	0.0002
M	-0.07151178	0.01068263	2937.19549357	44.81	0.0001
TP	0.37913861	0.05554103	3054.23526521	46.60	0.0001

Bounds on condition number: 1.000038, 4.000153

No other variable met the 0.0500 significance level for entry into the model.

The SAS System

Summary of Forward Selection Procedure for Dependent Variable FR

Step	Variable Entered	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	TP	1	0.3715	0.3715	38.0702	21.2751	0.0001
2	M	2	0.3529	0.7244	-0.3952	44.8124	0.0001

8.1.3.4.- Procedimiento Stepwise.

Por último se presenta la salida 8.4. donde se obtiene el modelo de regresión lineal múltiple a través del procedimiento STEPWISE, y se observa que en este caso se obtiene el mismo modelo que con el procedimiento Forward, por lo que también se realizaron los mismo pasos. Cabe mencionar que no siempre se obtiene el mismo modelo a través de los dos procedimientos.

Al final se presentan los parámetros del modelo obtenido por el procedimiento de selección STEPWISE, donde se obtienen los resultados del análisis de varianza y las pruebas de hipótesis descritas anteriormente.

En la figura 8.1. se observa que existe una fuerte correlación entre los valores predichos por el modelo y los valores. observados, por lo que el modelo ayuda a predecir la friabilidad de los gránulos en un 70 % aproximadamente. También se observa en la figura 8.2 que los valores residuales se distribuyen de manera homogenea, por lo que se supone que se cumplen los supuestos del análisis.

En las figuras 8.3 y 8.4 se representa en forma gráfica el modelo de regresión múltiple ajustado a través de una gráfica de contorno y una superficie de respuesta respectivamente.

Salida 8.4 . Modelo de regresión múltiple por procedimiento Stepwise.

```
                                The SAS System

          Stepwise Procedure for Dependent Variable FR

Step 1    Variable TP Entered      R-square = 0.37145428    C(p) = 38.07024426

                 DF        Sum of Squares          Mean Square          F     Prob>F

Regression        1        3091.52955457        3091.52955457      21.28     0.0001
Error            36        5231.24307701         145.31230769
Total            37        8322.77263158

                Parameter       Standard          Type II
Variable        Estimate           Error     Sum of Squares          F     Prob>F

INTERCEP      1.14137108      6.56381653       4.39383012       0.03     0.8629
TP            0.38143906      0.08269698    3091.52955457      21.28     0.0001

Bounds on condition number:             1,              1
--------------------------------------------------------------------------------

Step 2    Variable M Entered       R-square = 0.72436498    C(p) = -0.39516588

                 DF        Sum of Squares          Mean Square          F     Prob>F

Regression        2        6028.72504814        3014.36252407      45.99     0.0001
Error            35        2294.04758344          65.54421667
Total            37        8322.77263158

                Parameter       Standard          Type II
Variable        Estimate           Error     Sum of Squares          F     Prob>F

INTERCEP     23.30177864      5.51288240    1170.99506913      17.87     0.0002
M            -0.07151178      0.01068263    2937.19549357      44.81     0.0001
TP            0.37913861      0.05554103    3054.23526521      46.60     0.0001

Bounds on condition number:      1.000038,       4.000153
--------------------------------------------------------------------------------
All variables left in the model are significant at the 0.0500 level.
No other variable met the 0.0500 significance level for entry into the
model.

                                The SAS System
        Summary of Stepwise Procedure for Dependent Variable FR

           Variable            Number     Partial      Model
Step     Entered  Removed        In       R**2         R**2        C(p)          F
Prob>F

1        TP            1       0.3715       0.3715     38.0702     21.2751     0.0001
2        M             2       0.3529       0.7244     -0.3952     44.8124     0.0001
```

Salida 8.5.-MODELO AJUSTADO PARA EL CASO DE ESTUDIO No. 1

The SAS System

Model: MODEL1
Dependent Variable: FR

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	6028.72505	3014.36252	45.990	0.0001
Error	35	2294.04758	65.54422		
C Total	37	8322.77263			

Root MSE	8.09594	R-square	0.7244
Dep Mean	30.04211	Adj R-sq	0.7086
C.V.	26.94864		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	23.301779	5.51288240	4.227	0.0002
TP	1	0.379139	0.05554103	6.826	0.0001
M	1	-0.071512	0.01068263	-6.694	0.0001

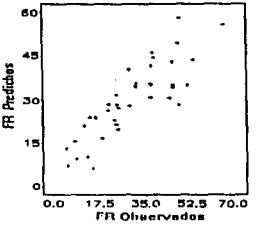


Figura 8.1.-Diagrama de dispersión de Valores Predichos VS Valores Observados

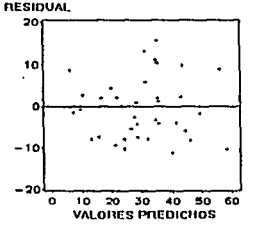
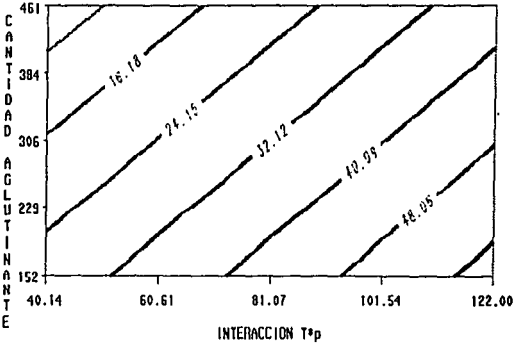


Figura 8.2.-Residuales para los valores predichos

$$FR=23.301779 - 0.071512\ M + 0.379139\ T*P$$

FIG. 8.3.- GRÁFICA DE CONTORNO PARA LA FRIABILIDAD DE GRÁNULOS

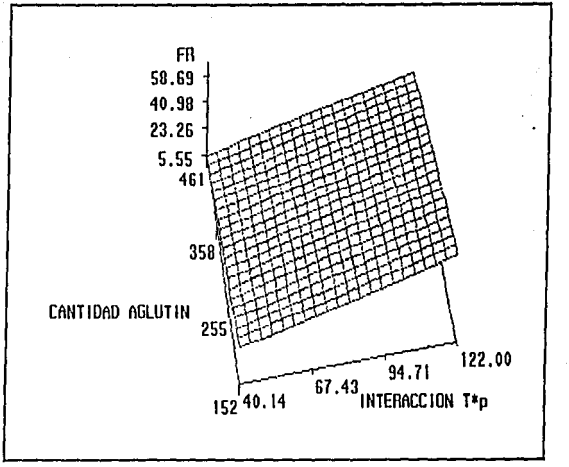


VARIABLES INDEPENDIENTES
M= CANTIDAD DE SOLUCIÓN AGLUTINANTE (g)
T = TEMPERATURA DEL AIRE DE ENTRADA (°C)
P = PRESIÓN DEL AIRE DE ATOMIZACIÓN (bar)

VARIABLE DE RESPUESTA:
FR = PORCENTAJE DE PESO PERDIDO POR FRIABILIDAD DE GRÁNULOS.

Se observa que el porcentaje de peso perdido por friabilidad de los gránulos disminuye cuando la cantidad de aglutinante aumenta o cuando el valor del producto de la interacción T*p disminuye, por lo que se puede obtener una friabilidad menor ya sea manteniendo fija la temperatura del aire de entrada y disminuyendo la presión del aire de atomización o viceversa. También se puede mantener fijo el valor de T*p y aumentar la cantidad de solución aglutinante para obtener gránulos con el menor porcentaje de peso perdido por friabilidad, ya que esta es una característica de calidad para proseguir con el proceso de tableteado.

FIG. 8.4.- SUPERFICIE DE RESPUESTA PARA LA FRIABILIDAD DE GRÁNULOS



En cualquier punto de la cantidad de aglutinante se observa que existe un aumento significativo en la friabilidad cuando el valor de la interacción T*p aumenta, sin embargo, cuando el valor de T*p se mantiene fijo y se incrementa la cantidad de aglutinante se observa sólo una pequeña disminución en la friabilidad, por lo que se puede decir que la interacción T*p es un factor que explica la mayor variación en la respuesta.

8.2.- ESTUDIO DE CASO No. 2.

8.2.1. *PROBLEMA*

En este caso se realizó un análisis de regresión múltiple para encontrar un modelo que describa la viscosidad de una suspensión reconstituible de Rifampicina. Este estudio fue realizado por Seham A. Elkheshem y col. y sus resultados se emplearon para la construcción del modelo. Ellos estudiaron el efecto del porcentaje de sacarosa, avicel RC-591, aerosil hidrofílico y aerosol-OT sobre algunas caracteristicas de la suspensión de Rifampicina, tales como: volumen de sedimentación, viscosidad y redispersabilidad de la suspensión. El estudio se realizo siguiendo un diseño factorial 2^4 con cinco réplicas al centro del diseño, obteniéndose así 21 experiencias.

8.2.2. *RESULTADOS EXPERIMENTALES Y AJUSTE DEL MODELO*

Para este estudio de caso sólo se tomo la viscosidad como variable de respuesta, así, para ajustar el modelo de regresión se tiene en el programa de SAS 8.2 la matriz de experimentación, es decir, las variables y los niveles a los que se trabajaron. También se tienen los resultados correspondientes a la viscosidad asi como el procedimiento necesario para obtener un modelo de regresión a través del método Stepwise. Cabe mencionar que se realizo el ajuste del modelo a través de los tres métodos de regresión analizados en la sección 7.8, los cuales produjeron los mismos resultados, por lo que únicamente se consideró el método Stepwise. En este programa se crean todas las variables necesarias que son funciones de las variables básicas.

Con el programa 8.2. primero se obtiene una matriz de correlación, la cual se observa en la salida 8.6. en donde se encuentran los coeficientes de correlación simple y la posible relación entre la variable de respuesta y las variables independientes

Programa 8.2.- Resultados experimentales del estudio de caso No. 2 y ajuste del modelo de regresión múltiple a través del método Stepwise.

```
OPTIONS PS=60 NODATE NONUMBER;
DATA SEHAM;
INPUT EXP X1 X2 X3 X4 Y ;
/*      NIVELES DE LAS VARIABLES          -1        0       +1
 EXP = EXPERIENCIA                      ----------------------
 X1 = CONCENTRACION DE SACAROSA          30.0%   45.0%   60.0%
 X2 = AVICEL RC 591                       1.0%    1.5%    2.0%
 X3 = AEROSIL 200                         0.0%    0.5%    1.0%
 X4 = AEROSOL-OT                          0.0%    0.05%   0.1%

 Y = VISCOSIDAD DE SUSPENSION 24 hr DESPUES DE RESUSPENDER (cps)
*/
X11=X1*X1; X22=X2*X2; X33=X3*X3; X44=X4*X4; X12=X1*X2; X13=X1*X3; X14=X1*X4;
X23=X2*X3; X24=X2*X4; X34=X3*X4; X123=X1*X2*X3; X124=X1*X2*X3; X234=X2*X3*X4;
CARDS;
1     -1      -1      +1      +1     6.0
2     -1      -1      -1      +1     3.5
3      0       0       0       0     9.0
4     -1      +1      +1      -1     6.5
5     +1      -1      +1      -1    22.5
6     +1      -1      +1      +1    22.5
7     -1      +1      +1      +1     8.0
8      0       0       0       0     9.5
9     -1      +1      -1      +1     4.5
10     0       0       0       0     9.5
11    +1      +1      -1      -1    18.0
12    +1      +1      +1      +1    27.0
13    +1      -1      -1      -1    13.0
14    -1      +1      -1      -1     4.0
15    +1      -1      -1      +1    16.5
16     0       0       0       0     9.0
17    -1      -1      -1      -1     3.5
18    -1      -1      +1      -1     6.0
19     0       0       0       0     8.5
20    +1      +1      -1      +1    21.0
21    +1      +1      +1      -1    25.5
;
PROC CORR;
PROC REG;
MODEL Y = X1 X2 X3 X4 X11 X22 X33 X44 X12 X13 X14
          X23 X24 X34 X123 X124 X234  / SELECTION=STEPWISE
                                        SLE=0.05 SLS=0.05;
RUN;
```

Salida 8.6. Matriz de correlación entre la viscosidad de la suspensión de Rifampicina y las variables en estudio.

The SAS System
Correlation Analysis

19 'VAR' Variables: EXP Y X12 X34 X1 X11 X13 X123 X2 X22 X14 X124 X3 X33 X23 X234 X4 X44 X24

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 21

	X1	X2	X3	X4	Y	X11	X22
Y	0.89803 0.0001	0.15208 0.5105	0.28969 0.2028	0.07242 0.7551	1.00000 0.0000	0.22051 0.3368	0.22051 0.3368

	X33	X44	X12	X13	X14	X23
Y	0.22051 0.3368	0.22051 0.3368	0.09415 0.6848	0.13036 0.5733	0.04345 0.8516	-0.00724 0.9751

	X24	X34	X123	X124	X234
Y	0.02173 0.9255	-0.02897 0.9008	-0.02173 0.9255	-0.02173 0.9255	0.02173 0.9255

Como se observa en la salida 8.6, donde se tienen los coeficientes de correlación de orden cero, la variable que tiene mayor correlación con la viscosidad es la que se muestra resaltada, X1 (concentración de sacarosa), y al parecer es la única que se relaciona con la respuesta, ya que el valor de Prob > R es menor a 0.05, por lo que se considera un coeficiente de correlación entre estas dos variables diferente de cero, sin embargo al hacer el ajuste del modelo se observó que no es la única variable que ayuda a predecir la respuesta ya que conforme se van adicionando las variables al modelo, su correlación parcial cambia, como se describe a continuación.

De acuerdo al programa 8.2, el ajuste del modelo se realiza a través del procedimiento de selección Stepwise, el cual comienza por introducir al modelo la variable que tiene mayor correlación con la respuesta (viscosidad de suspensión de Rifampicina), como se observó en la matriz de correlación, por lo que en la salida 8.7 aparece en el paso 1 del procedimiento Stepwise la variable X1 (concentración de sacarosa). De esta forma se observa que el coeficiente de determinación del modelo es mayor de 0.8 y la prueba de hipótesis acerca del modelo indica que éste es significativo,

es decir, $\beta o \neq 0$.

De esta forma se van introduciendo al modelo las variables que son significativamente importantes para la predicción de la respuesta como se observa en la salida 8.7, donde se presentan los resultados estadísticos en cada paso del procedimiento hasta que él mismo no encuentra alguna otra variable que sea significativa en la predicción de la respuesta.

En este procedimiento se realizan 7 pasos, por lo que en el modelo se encuentran 7 variables importantes en vez de 17 como se tenia al principio. Sin embargo, se puede observar que una sola variable (X1) es la que tiene mayor peso en el modelo, ya que su R^2 parcial es de 0.8064 a comparación de las otras variables cuyas R^2 parciales son menor a 0.1, no obstante las siete variables en conjunto generan un modelo de regresión con alto porcentaje de predicción ya que el R^2 del modelo es 0.9932, un valor muy cercano a 1.

En la salida 8.8. se comprueba que el modelo de regresión ajustado es bueno, ya que el valor del R^2 ajustado es también cercano a 1, y este parámetro confirma que no sólo es debido a la adición de variables independientes. También se observa que todos los coeficientes estimados para el modelo son significativos ya que el valor de Prob > |T| para todos los coeficientes es menor de 0.05, por lo que se puede concluir que el modelo es significativo con un valor de $\alpha = 0.05\%$.

Por otro lado se observa en las figuras 8.5 y 8.6 que sí se cumplen los supuestos estadisticos, ya que en la gráfica de valores observados *vs* valores predichos, los puntos se disponen en línea recta, y en la gráfica de residuales no se observa ninguna tendencia anormal en los puntos graficados.

De esta forma se concluye que la viscosidad de una suspensión de Rifampicina se puede modelar a través de las variables básicas que se estudiaron en este caso, y con algunas interacciones de estas variables, como se muestra en la salida 8.8.

Salida 8.7. Modelo de regresión múltiple a través de Stepwise para la viscosidad de una suspensión de Rifampicina.

The SAS System

Stepwise Procedure for Dependent Variable Y

Step 1 Variable X1 Entered R-square = 0.80644968 C(p) = 508.04065041

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	961.00000000	961.00000000	79.17	0.0001
Error	19	230.64285714	12.13909774		
Total	20	1191.64285714			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	12.07142857	0.76029749	3060.10714286	252.09	0.0001
X1	7.75000000	0.87103020	961.00000000	79.17	0.0001

Bounds on condition number: 1, 1

Step 2 Variable X3 Entered R-square = 0.89036744 C(p) = 282.39837398

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	1061.00000000	530.50000000	73.09	0.0001
Error	18	130.64285714	7.25793651		
Total	20	1191.64285714			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	12.07142857	0.58789117	3060.10714286	421.62	0.0001
X1	7.75000000	0.67351394	961.00000000	132.41	0.0001
X3	2.50000000	0.67351394	100.00000000	13.78	0.0016

Bounds on condition number: 1, 4

Step 3 Variable X11 Entered R-square = 0.93899179 C(p) = 152.49593496

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	1118.94285714	372.98095238	87.22	0.0001
Error	17	72.70000000	4.27647059		
Total	20	1191.64285714			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	9.10000000	0.92482113	414.05000000	96.82	0.0001
X1	7.75000000	0.51699073	961.00000000	224.72	0.0001
X3	2.50000000	0.51699073	100.00000000	23.38	0.0002
X11	3.90000000	1.05951571	57.94285714	13.55	0.0019

Bounds on condition number: 1, 9

Salida 8.7. Modelo de regresión múltiple a través de Stepwise para la viscosidad de la suspensión de Rifampicina. *(Continuación)*

Step 4 Variable X2 Entered R-square = 0.96212162 C(p) = 91.75203252

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	1146.50535714	286.62633929	101.60	0.0001
Error	16	45.13750000	2.82109375		
Total	20	1191.64285714			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	9.10000000	0.75114496	414.05000000	146.77	0.0001
X1	7.75000000	0.41990280	961.00000000	340.65	0.0001
X2	1.31250000	0.41990280	27.56250000	9.77	0.0065
X3	2.50000000	0.41990280	100.00000000	35.45	0.0001
X11	3.90000000	0.86054466	57.94285714	20.54	0.0003
Bounds on condition number:		1,	16		

Step 5 Variable X13 Entered R-square = 0.97911497 C(p) = 47.65447154

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	1166.75535714	233.35107143	140.64	0.0001
Error	15	24.88750000	1.65916667		
Total	20	1191.64285714			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	9.10000000	0.57604977	414.05000000	249.55	0.0001
X1	7.75000000	0.32202161	961.00000000	579.21	0.0001
X2	1.31250000	0.32202161	27.56250000	16.61	0.0010
X3	2.50000000	0.32202161	100.00000000	60.27	0.0001
X11	3.90000000	0.65994791	57.94285714	34.92	0.0001
X13	1.12500000	0.32202161	20.25000000	12.20	0.0033
Bounds on condition number:		1,	25		

Step 6 Variable X12 Entered R-square = 0.98797878 C(p) = 25.60975610

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	1177.31785714	196.21964286	191.77	0.0001
Error	14	14.32500000	1.02321429		
Total	20	1191.64285714			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	9.10000000	0.45237469	414.05000000	404.66	0.0001
X1	7.75000000	0.25288514	961.00000000	939.20	0.0001
X2	1.31250000	0.25288514	27.56250000	26.94	0.0001
X3	2.50000000	0.25288514	100.00000000	97.73	0.0001
X11	3.90000000	0.51826031	57.94285714	56.63	0.0001
X12	0.81250000	0.25288514	10.56250000	10.32	0.0063
X13	1.12500000	0.25288514	20.25000000	19.79	0.0006
Bounds on condition number:		1,	36		

Salida 8.7. Modelo de regresión múltiple a través de Stepwise para la viscosidad de la suspensión de Rifampicina. *(Continuación)*

Step 7 Variable X4 Entered R-square = 0.99322364 C(p) = 13.38211382

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	7	1183.56785714	169.08112245	272.20	0.0001
Error	13	8.07500000	0.62115385		
Total	20	1191.64285714			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	9.10000000	0.35246386	414.05000000	666.58	0.0001
X1	7.75000000	0.19703328	961.00000000	1547.12	0.0001
X2	1.31250000	0.19703328	27.56250000	44.37	0.0001
X3	2.50000000	0.19703328	100.00000000	160.99	0.0001
X4	0.62500000	0.19703328	6.25000000	10.06	0.0074
X11	3.90000000	0.40379807	57.94285714	93.28	0.0001
X12	0.81250000	0.19703328	10.56250000	17.00	0.0012
X13	1.12500000	0.19703328	20.25000000	32.60	0.0001

Bounds on condition number: 1, 49

All variables left in the model are significant at the 0.0500 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable Y

Step	Variable Entered	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X1		1	0.8064	0.8064	508.0407	79.1657	0.0001
2	X3		2	0.0839	0.8904	282.3984	13.7780	0.0016
3	X11		3	0.0486	0.9390	152.4959	13.5492	0.0019
4	X2		4	0.0231	0.9621	91.7520	9.7701	0.0065
5	X13		5	0.0170	0.9791	47.6545	12.2049	0.0033
6	X12		6	0.0089	0.9880	25.6098	10.3229	0.0063
7	X4		7	0.0052	0.9932	13.3821	10.0619	0.0074

Salida 8.8.- MODELO AJUSTADO PARA LA VISCOSIDAD

The SAS System

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	1183.56786	169.08112	272.205	0.0001
Error	13	8.07500	0.62115		
C Total	20	1191.64286			

Root MSE	0.78813	R-square	0.9932
Dep Mean	12.07143	Adj R-sq	0.9896
C.V.	6.52891		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	9.100000	0.35246386	25.818	0.0001
X1	1	7.750000	0.19703328	39.333	0.0001
X2	1	1.312500	0.19703328	6.661	0.0001
X3	1	2.500000	0.19703328	12.688	0.0001
X4	1	0.625000	0.19703328	3.172	0.0074
X11	1	3.900000	0.40379807	9.658	0.0001
X12	1	0.812500	0.19703328	4.124	0.0012
X13	1	1.125000	0.19703328	5.710	0.0001

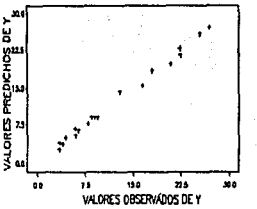


Figura 8.5.- Diagrama de dispersión de Valores Predichos VS Valores Observados

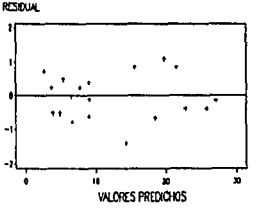


Figura 8.6 - Residuales para los valores predichos de la viscosidad en la suspensión de Rifampicina.

$$\text{Viscosidad} = 9.1 + 7.75\,X_1 + 1.3125\,X_2 + 2.5\,X_3 + 0.625\,X_4 + 3.9\,X_{11} + 0.8125\,X_{12} + 1.125\,X_{13}$$

8.3.- ESTUDIO DE CASO No. 3.

8.3.1. *PROBLEMA.*

En este caso se tiene un estudio realizado por Chistopher F. Dick et al (1987) quienes evaluaron los efectos de ciertos componentes de una formulación de tabletas elaboradas por compresión directa y algunas variables de proceso. Estudiaron a través de un diseño factorial fraccionado el efecto de cinco variables de formulación y proceso sobre las características finales de las tabletas. Las variables independientes que se estudiaron fueron: (X1) = relación lactosa / celulosa microcristalina, (X2) = Cantidad de estearato de magnesio, (X3) = Número de revoluciones del mezclador, (X4) = Fuerza de compresión, (X5) = Velocidad de tableteadora. Las variables de respuesta fueron: Disolución, Dureza, Espesor, Variación de Peso, Tiempo de Desintegración, Uniformidad de Contenido y Friabilidad. Sin embargo se observó que no todas las respuestas se pueden ajustar a un buen modelo, por lo que se tomó como ejemplo el ajuste del modelo de Regresión para la Friabilidad de las tabletas en función de los factores en estudio.

8.3.2. *RESULTADOS EXPERIMENTALES Y AJUSTE DEL MODELO.*

La matriz de experimentación y los resultados experimentales necesarios para el ajuste del modelo se muestran en el programa de SAS 8.3., donde también se encuentran los niveles de cada variable y el procedimiento para obtener el modelo de mejor ajuste a través del método de regresión múltiple Stepwise. Al igual que en el ejemplo anterior, se ajusto el modelo a través de los tres métodos de regresión múltiple obteniéndose los mismos resultados.

Con el programa 8.3. se obtiene en primer lugar una matriz de correlación, la cual se observa en la salida 8.9 en donde se encuentran los coeficientes de correlación simple y la posible relación entre la variable de respuesta y las variables independientes

Programa 8.3.- Resultados experimentales del estudio de caso No. 3 y ajuste del modelo de regresión múltiple a través del método Stepwise.

```
OPTIONS PS=60 NODATE NONUMBER;
DATA DICK;
INPUT EXP X1 X2 X3 X4 X5 Y;
/*
 EXP = EXPERIENCIA                          -1              0              +1
 ---------------------------------------------------------------------------
 X1 = RELACION LACTOSA/CELULOSA           6.5:1           7.5:1          8.5:1
    (relacion en mg)                  (147.3:22.7)    (150:30)       (152.1:17.9)
 X2 = ESTEARATO DE MAGNESIO (mg)          1.35            1.50           1.65
 X3 = NUMERO DE REV. DE MEZCLADO           378.00          450.00         522.00
 X4 = FUERZA DE COMPRESION                1800.00         2000.00        2200.00
 X5 = VELOCIDAD DE TABLETEADO (rpm)         32.00           36.00          40.00

 Y = FRIABILIDAD (%)

*/
X11=X1*X1; X22=X2*X2; X33=X3*X3; X44=X4*X4; X55=X5*X5; X12=X1*X2; X23=X2*X3;
X34=X3*X4; X45=X4*X5; X13=X1*X3; X14=X1*X4; X15=X1*X5; X24=X2*X4; X25=X2*X5;
X35=X3*X5;
CARDS;
1  -1   +1   +1   +1   -1   0.1353
2  -1   -1   -1   +1   -1   0.1492
3  +1   +1   +1   +1   +1   0.1782
4  -1   +1   +1   -1   +1   0.2235
5  -1   -1   -1   -1   +1   0.1815
6  +1   -1   -1   +1   +1   0.1800
7  +1   -1   +1   +1   -1   0.2084
8  -1   +1   -1   +1   +1   0.2100
9  -1   +1   -1   -1   -1   0.2248
10 -1   -1   +1   -1   -1   0.2688
11 +1   -1   +1   -1   +1   0.2602
12  0    0    0    0    0   0.2531
13 +1   -1   -1   -1   -1   0.2244
14  0    0    0    0    0   0.2203
15 -1   -1   +1   +1   +1   0.1735
16 +1   +1   -1   +1   -1   0.2192
17 +1   +1   -1   -1   +1   0.2512
18 +1   +1   +1   -1   -1   0.2749
19  0    0    0    0    0   0.2221
;
PROC CORR;
PROC REG;
MODEL Y = X1 X2 X3 X4 X5 X11 X22 X33 X44 X55 X12 X23 X34
          X45 X13 X14 X15 X24 X25 X35 / SELECTION=STEPWISE
                                        SLE=0.05 SLS=0.05;
RUN;
```

Salida 8.9. Matriz de correlación entre la Friabilidad de tabletas por compresión directa y las variables en estudio.

The SAS System
Correlation Analysis

22 'VAR' Variables: EXP X1 X2 X3 X4 X5 Y X11 X22 X33 X44 X55 X12 X23 X34 X45 X24 X13 X14 X25 X35

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 19

	X1	X2	X3	X4	X5
Y	0.34485 0.1482	0.10665 0.6639	0.12375 0.6137	-0.68325 0.0013	-0.07035 0.7747

	Y	X11	X22	X33	X44	X55
Y	1.00000 0.0	-0.20637 0.3966	-0.20637 0.3966	-0.20637 0.3966	-0.20637 0.3966	-0.20637 0.3966

	X12	X23	X34	X45	X13	X14
Y	0.04485 0.8553	-0.40365 0.0866	-0.31275 0.1923	0.15915 0.5152	0.01695 0.9451	0.00855 0.9723

	X15	X24	X25	X35
Y	-0.10155 0.6791	-0.01185 0.9616	0.09645 0.6945	-0.08565 0.7274

Como se observa en la salida 8.9, la variable que tiene mayor correlación con la friabilidad de las tabletas es X4 (fuerza de compresión) y al parecer es la única que se correlaciona con la respuesta, debido a que el valor de R (coeficiente de correlación) es -0.68 y su correspondiente valor de Prob > |R| es menor de 0.05, sin embargo al hacer el ajuste del modelo se encuentran otras variables que también están relacionadas con la respuesta y por consiguiente ayudan a modelarla. A continuación se describe el ajuste del modelo de regresión múltiple a través del método Stepwise, que como es de esperarse, debe iniciar introduciendo la variable X4 al modelo.

En base al programa 8.3, el ajuste del modelo se presenta en la salida 8.10., donde se observa que se emplea el procedimiento Stepwise, el cual incluye a la variable X4 en el primer paso, ya que esta es la que tiene mayor correlación con la respuesta (Friabilidad de tabletas). Así, se obtiene un primer modelo con una sola variable independiente cuyo ajuste no es muy bueno ya que el valor de R^2

es de 0.46, sin embargo los coeficientes calculados para el modelo son significativos, debido a que sus respectivos valores de Prob>F son menores de 0.05.

En el segundo paso del procedimiento se introduce la variable X23 (interacción entre cantidad de estearato de magnesio y número de revoluciones de mezclado), con lo cual aumenta el valor del coeficiente de determinación del modelo, es decir, el modelo de regresión mejora, pero no lo suficiente por lo que es necesario introducir más variables.

De esta forma se van introduciendo las variables al modelo hasta que al final se encuentran sólo aquellas variables que ayudan en la predicción de la respuesta. Así, se observa que el modelo es ajustado en cinco pasos, por lo que el modelo que mejor ajusta los resultados contiene cinco variables y su coeficiente de determinación es 0.889.

En la salida 8.11 se comprueba que el modelo de regresión ajustado para este caso es bueno, ya que se observa que todos los coeficientes estimados son significativamente importantes debido a que el valor de Prob>T es menor de 0.5 en todos los casos.

Por otro lado se observa en las figuras 8.7 y 8.8 que no existe violación en los supuestos debido a que en el gráfico de dispersión existe una alineación lineal de los puntos y en el gráfico de residuales no existen tendencias o puntos extremos que puedan salirse de la normalidad.

Así, se concluye que la friabilidad de las tabletas se puede modelar en función de las variables en estudios, como se muestra en la salida 8.11.

Salida 8.10. Modelo de regresión múltiple a través de Stepwise para la friabilidad de tabletas elaboradas por compresión directa.

The SAS System

Stepwise Procedure for Dependent Variable Y

Step 1 Variable X4 Entered R-square = 0.46683224 C(p) = 28.55759265

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	0.01296752	0.01296752	14.88	0.0013
Error	17	0.01481016	0.00087119		
Total	18	0.02777768			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.21361053	0.00677140	0.86695968	995.15	0.0001
X4	-0.02846875	0.00737897	0.01296752	14.88	0.0013

Bounds on condition number: 1, 1

Step 2 Variable X23 Entered R-square = 0.62976615 C(p) = 17.24656870

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	0.01749344	0.00874672	13.61	0.0004
Error	16	0.01028424	0.00064276		
Total	18	0.02777768			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.21361053	0.00581633	0.86695968	1348.80	0.0001
X4	-0.02846875	0.00633820	0.01296752	20.17	0.0004
X23	-0.01681875	0.00633820	0.00452593	7.04	0.0173

Bounds on condition number: 1, 4

Step 3 Variable X1 Entered R-square = 0.74868810 C(p) = 9.53113903

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	0.02079682	0.00693227	14.90	0.0001
Error	15	0.00698086	0.00046539		
Total	18	0.02777768			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.21361053	0.00494917	0.86695968	1862.86	0.0001
X1	0.01436875	0.00539323	0.00330338	7.10	0.0177
X4	-0.02846875	0.00539323	0.01296752	27.86	0.0001
X23	-0.01681875	0.00539323	0.00452593	9.73	0.0070

Bounds on condition number: 1, 9

Salida 8.10. Modelo de regresión múltiple a través de Stepwise para la friabilidad de tabletas elaboradas por compresión directa. *(Continuación)*

Step 4 Variable X34 Entered R-square = 0.84650101 C(p) = 3.54022998

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	0.02351383	0.00587846	19.30	0.0001
Error	14	0.00426385	0.00030456		
Total	18	0.02777768			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.21361053	0.00400369	0.86695968	2846.59	0.0001
X1	0.01436875	0.00436291	0.00330338	10.85	0.0053
X4	-0.02846875	0.00436291	0.01296752	42.58	0.0001
X23	-0.01681875	0.00436291	0.00452593	14.86	0.0018
X34	-0.01303125	0.00436291	0.00271702	8.92	0.0098

Bounds on condition number: 1, 16

Step 5 Variable X11 Entered R-square = 0.88908923 C(p) = 2.06094920

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	0.02469683	0.00493937	20.84	0.0001
Error	13	0.00308084	0.00023699		
Total	18	0.02777768			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.23183333	0.00888797	0.16124008	680.37	0.0001
X1	0.01436875	0.00384860	0.00330338	13.94	0.0025
X4	-0.02846875	0.00384860	0.01296752	54.72	0.0001
X11	-0.02163958	0.00968544	0.00118300	4.99	0.0437
X23	-0.01681875	0.00384860	0.00452593	19.10	0.0008
X34	-0.01303125	0.00384860	0.00271702	11.46	0.0049

Bounds on condition number: 1, 25

All variables left in the model are significant at the 0.0500 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable Y

Step	Variable Entered	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X4		1	0.4668	0.4668	28.5576	14.8849	0.0013
2	X23		2	0.1629	0.6298	17.2466	7.0413	0.0173
3	X1		3	0.1189	0.7487	9.5311	7.0981	0.0177
4	X34		4	0.0978	0.8465	3.5402	8.9211	0.0098
5	X11		5	0.0426	0.8891	2.0609	4.9918	0.0437

Salida 8.11.- MODELO AJUSTADO PARA LA FRIABILIDAD DE TABLETAS DE COMPRESIÓN DIRECTA

```
                        The SAS System
Model: MODEL1
Dependent Variable: Y

                      Analysis of Variance

                            Sum of          Mean
Source          DF        Squares        Square      F Value      Prob>F

 Model           5        0.02470       0.00494       20.842      0.0001
Error           13        0.00308       0.00024
C Total         18        0.02778

       Root MSE          0.01539     R-square         0.8891
       Dep Mean          0.21361     Adj R-sq         0.8464
       C.V.              7.20677

                    Parameter Estimates

                       Parameter      Standard    T for H0:
Variable    DF          Estimate         Error   Parameter=0    Prob > |T|

INTERCEP     1          0.231833    0.00888797        26.084        0.0001
X1           1          0.014369    0.00384860         3.733        0.0025
X4           1         -0.028469    0.00384860        -7.397        0.0001
X11          1         -0.021640    0.00968544        -2.234        0.0437
X23          1         -0.016819    0.00384960        -4.370        0.0008
X34          1         -0.013031    0.00384860        -3.386        0.0049
```

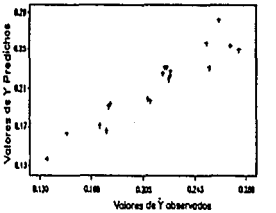


Figura 8.7.- Diagrama de dispersión de Valores Predichos VS Valores Observados.

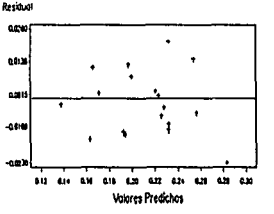


Figura 8.8.- Residuales para los valores predichos de la friabilidad de tabletas de compresión directa.

$$\text{Friabilidad} = 0.23183 + 0.014369\,X_1 - 0.028469\,X_4 - 0.02164\,X_{11} - 0.016819\,X_{23} - 0.013031\,X_{34}$$

8.4.- ESTUDIO DE CASO No. 4.

8.4.1. *PROBLEMA*

En este caso se analizó un estudio realizado por M. R. Harris et al (1985) quienes desarrollaron y optimizaron una formulación de tabletas de liberación controlada, para lo cual se basaron en un diseño experimental con cinco factores. Las variables en estudio o independientes fueron algunos componentes de la formulación y la fuerza de compresión durante el tableteo, tales variables fueron: (X1) = Fuerza de compresión, (X2) = Nivel de Montmorillonite (Veegum F®), (X3) = Nivel de dextrosa-maltosa, (X4) = Nivel de estearato de magnesio y (X5) = cantidad de almidón. Las respuestas o variables dependientes fueron la uniformidad del peso de las tabletas, dureza, friabilidad y porcentaje de fármaco en solución después de 3 horas. [12]

8.4.2. *RESULTADOS EXPERIMENTALES Y AJUSTE DEL MODELO*

Cada respuesta fue ajustada a un modelo de regresión, sin embargo se observó que no todas las respuestas se pueden ajustar a un buen modelo, por lo que se tomo como ejemplo el ajuste del modelo de Regresión para la Dureza de las tabletas en función de los factores en estudio.

La matriz de experimentación y los resultados experimentales necesarios para el ajuste del modelo se muestran en el programa de SAS 8.4., donde también se encuentran los niveles de cada variable y el procedimiento para obtener el modelo de mejor ajuste a través del método Stepwise. Al igual que en el ejemplo anterior, se ajusto el modelo a través de los tres métodos de regresión múltiple obteniéndose los mismos resultados, los cuales se describen a continuación..

Programa 8.4.- Resultados experimentales y ajuste del modelo de regresión múltiple a través del método Stepwise para el estudio de caso No. 4.

```
OPTIONS PS=60 NODATE NONUMBER;
DATA HARRIS;
INPUT EXP X1 X2 X3 X4 X5 Y;
/*                                        -1.547   -1      0     +1    +1.547
 EXP = EXPERIENCIA                       ------------------------------------
 X1 = FUERZA DE COMPRESION                 0.25    0.5    1.0    1.5    1.75
 X2 = MONTOMORILLONITE (Veegum)           22.50   25.5   30.0   35.5   37.50
 X3 = DEXTROSA                            10.50   11.5   14.0   16.5   17.75
 X4 = ESTEARATO DE MAGNESIO                0.25    0.5    1.0    1.5    1.75
 X5 = ALMIDON                             12.50   15.0   20.0   25.0   27.50

 Y = DUREZA (Kg)

*/
X11=X1*X1; X22=X2*X2; X33=X3*X3; X44=X4*X4; X55=X5*X5; X12=X1*X2; X23=X2*X3;
X34=X3*X4; X45=X4*X5; X13=X1*X3; X14=X1*X4; X15=X1*X5; X24=X2*X4; X25=X2*X5;
X35=X3*X5;
CARDS;
1  -1      -1      -1      -1      +1      5.00
2  +1      -1      -1      -1      -1      9.75
3  -1      +1      -1      -1      -1      5.53
4  +1      +1      -1      -1      +1      9.90
5  -1      -1      +1      -1      -1      5.50
6  +1      -1      +1      -1      +1      11.33
7  -1      +1      +1      -1      +1      6.68
8  +1      +1      +1      -1      -1      8.95
9  -1      -1      -1      +1      -1      4.93
10 +1      -1      -1      +1      +1      9.70
11 -1      +1      -1      +1      +1      6.10
12 +1      +1      -1      +1      -1      8.18
13 -1      -1      +1      +1      +1      4.83
14 +1      -1      +1      +1      -1      7.60
15 -1      +1      +1      +1      -1      6.08
16 +1      +1      +1      +1      +1      10.40
17 -1.547  0       0       0       0       4.48
18 +1.547  0       0       0       0       10.70
19 0       -1.547  0       0       0       8.88
20 0       +1.547  0       0       0       6.45
21 0       0       -1.547  0       0       6.58
22 0       0       +1.547  0       0       7.18
23 0       0       0       -1.547  0       7.68
24 0       0       0       +1.547  0       6.48
25 0       0       0       0       -1.547  5.95
26 0       0       0       0       +1.547  8.40
27 0       0       0       0       0       7.78
;
PROC CORR;
PROC REG;
MODEL Y = X1 X2 X3 X4 X5 X11 X22 X33 X44 X55 X12 X23 X34
          X45 X13 X14 X15 X24 X25 X35 / SELECTION=STEPWISE
                                        SLE=0.05 SLS=0.05;
RUN;
```

Salida 8.12. Matriz de correlación entre la Dureza de tabletas y las variables en estudio del caso No. 4.

```
                    The SAS System
                  Correlation Analysis

 22 'VAR' Variables:  EXP      X1      X2      X3      X4
                      X5       Y       X11     X22     X33
                      X44      X55     X12     X23     X34
                      X45      X13     X14     X15     X24
                      X25      X35

 Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 27

          X1          X2          X3          X4          X5

Y      0.88818     -0.01261     0.06987    -0.14540     0.24414
       0.0001       0.9502      0.7291      0.4693      0.2197

          Y           X11         X22         X33         X44         X55

Y      1.00000     0.05957     0.07010    -0.04014    -0.01206     0.00129
       0.0         0.7679      0.7283      0.8424      0.9524      0.9949

          X12         X23         X34         X45         X13         X14

Y     -0.12610     0.06255    -0.05660     0.02631    -0.01936    -0.08142
       0.5308      0.7566      0.7792      0.8963      0.9236      0.6864

          X15         X24         X25         X35

Y      0.15589     0.10475     0.03128     0.06951
       0.4375      0.6031      0.8769      0.7305
```

En la salida 8.12 se puede observar que la variable independiente que tiene mayor correlación con la respuesta es X1 (fuerza de compresión), ya que el valor de Prob>R es menor de 0.05, por lo que se puede considerar que el coeficiente de correlación correspondiente es significativamente diferente de cero. Así, al hacer el ajuste del modelo, la variable que tiene mayor importancia en la predicción es X1 por lo tanto debe aparecer en el primer paso del método de ajuste que se seleccionó (Stepwise).

En la salida 8.13 se obtienen los resultados del programa 8.4, donde se indica un procedimiento Stepwise para encontrar el modelo que mejor describa los resultados experimentales, teniendo como variable de respuesta la dureza de tabletas de liberación controlada.

Como se mencionó anteriormente, la variable que debe estar presente en el primer paso es X1 ya que tiene el mayor coeficiente de correlación simple, sin embargo se observa que existen otras variables que mejoran la predicción del modelo, las cuales aparecen en etapas posteriores del procedimiento stepwise. Asi, en el paso 2 se introduce al modelo la variable X5 (cantidad de almidón), en el paso tres se toma en cuenta la variable X15 (interacción entre fuerza de compresión y cantidad de almidón) y en el paso cuatro se introduce la variable X4 (cantidad de estearato de magnesio).

Como se observa, el modelo final contiene 4 variables que ayudan en la predicción de la respuesta, sin embargo se puede considerar que la variable más importante es la fuerza de compresión ya que tiene el coeficiente de correlación parcial mas grande.

En la salida 8.14 se obtienen los resultados del modelo propuesto con 4.variables y se observa que el modelo ajustado es bueno ya que tiene un valor de R ajustada de 0.87 lo cual indica que el modelo tiene un alto porcentaje de predicción. Además se observa que todos los parámetros del modelo son significativos debido a que su valor de Prob>|T| es menor de 0.05 por lo que se puede concluir que el modelo es significativo con un valor de α=0.05%.

También se puede observar en la figura 8.9 que existe tendencia lineal en los puntos que conforman la gráfica, asi como también no se observa ninguna tendencia de los residuales en la figura 8.10. De esta forma se puede concluir que la dureza de tabletas de liberación controlada esta más influenciada por la fuerza de compresión, sin embargo existen algunos componentes de la formulación que también afectan la dureza, tales como cantidad de almidón y cantidad de estearato de magnesio. Asi el modelo que se presenta en la salida 8.14 corresponde a este estudio de caso.

Salida 8.13. Modelo de regresión múltiple a través de Stepwise para la Dureza de tabletas del ejemplo No. 4.

The SAS System

Stepwise Procedure for Dependent Variable Y

Step 1 Variable X1 Entered R-square = 0.78886872 C(p) = 6.23991815

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	80.01375013	80.01375013	93.41	0.0001
Error	25	21.41472395	0.85658896		
Total	26	101.42847407			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	7.44518519	0.17811658	1496.63112593	1747.20	0.0001
X1	1.96197055	0.20300017	80.01375013	93.41	0.0001

Bounds on condition number: 1, 1

Step 2 Variable X5 Entered R-square = 0.84847380 C(p) = -0.01488770

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	86.05940277	43.02970139	67.19	0.0001
Error	24	15.36907130	0.64037797		
Total	26	101.42847407			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	7.44518519	0.15400553	1496.63112593	2337.11	0.0001
X1	1.96197055	0.17552071	80.01375013	124.95	0.0001
X5	0.53930167	0.17552071	6.04565265	9.44	0.0052

Bounds on condition number: 1, 4

Step 3 Variable X15 Entered R-square = 0.87277565 C(p) = -1.38049141

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	88.52430277	29.50810092	52.59	0.0001
Error	23	12.90417130	0.56105093		
Total	26	101.42847407			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	7.44518519	0.14415153	1496.63112593	2667.55	0.0001
X1	1.96197055	0.16429007	80.01375013	142.61	0.0001
X5	0.53930167	0.16429007	6.04565265	10.78	0.0033
X15	0.39250000	0.18725833	2.46490000	4.39	0.0473

Bounds on condition number: 1, 9

Salida 8.13. Modelo de regresión múltiple a través de Stepwise para la friabilidad de tabletas elaboradas por compresión directa *(Continuación)*.

Step 4 Variable X4 Entered R-square = 0.89391761 C(p) = -2.30847550

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	90.66869903	22.66717476	46.35	0.0001
Error	22	10.75977504	0.48908068		
Total	26	101.42847407			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	7.44518519	0.13458863	1496.63112593	3060.09	0.0001
X1	1.96197055	0.15339119	80.01375013	163.60	0.0001
X4	-0.32119050	0.15339119	2.14439626	4.38	0.0480
X5	0.53930167	0.15339119	6.04565265	12.36	0.0019
X15	0.39250000	0.17483576	2.46490000	5.04	0.0352

Bounds on condition number: 1, 16

All variables left in the model are significant at the 0.0500 level.
No other variable met the 0.0500 significance level for entry into the model.

The SAS System

Summary of Stepwise Procedure for Dependent Variable Y

Step	Variable Entered	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X1		1	0.7889	0.7889	6.2399	93.4097	0.0001
2	X5		2	0.0596	0.8485	-0.0149	9.4408	0.0052
3	X15		3	0.0243	0.8728	-1.3805	4.3934	0.0473
4	X4		4	0.0211	0.8939	-2.3085	4.3845	0.0480

Salida 8.14.- MODELO AJUSTADO PARA LA DUREZA DE TABLETAS DEL EJEMPLO No. 4

```
                        The SAS System
Model: MODEL1
Dependent Variable: Y

                      Analysis of Variance

                          Sum of          Mean
Source          DF       Squares        Square      F Value      Prob>F

Model            4      90.66870      22.66717       46.346      0.0001
Error           22      10.75978       0.48908
C Total         26     101.42847

    Root MSE          0.69934     R-square       0.8939
    Dep Mean          7.44519     Adj R-sq       0.8746
    C.V.              9.39323

                    Parameter Estimates

                    Parameter       Standard      T for H0:
Variable    DF      Estimate        Error         Parameter=0     Prob>|T|

INTERCEP     1      7.445185      0.13458863         55.318        0.0001
X1           1      1.961971      0.15339119         12.791        0.0001
X4           1     -0.321191      0.15339119         -2.094        0.0480
X5           1      0.539302      0.15339119          3.516        0.0019
X15          1      0.392500      0.17483576          2.245        0.0352
```

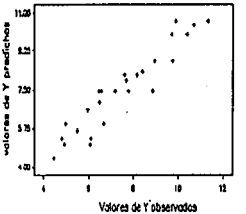


Figura 8.9 - Diagrama de dispersión de Valores Predichos VS Valores Observados

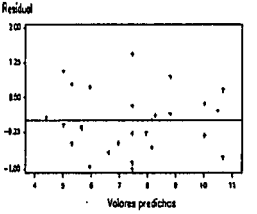


Figura 8.10 - Residuales para los valores predichos de la friabilidad de tabletas de compresión directa.

$$\text{Dureza} = 7.445105 + 1.961971\,X_1 - 0.321191\,X_4 + 0.539302\,X_5 + 0.3925\,X_{15}$$

IX. DISCUSIÓN

En la investigación farmacéutica existen diversas posibilidades para aplicar un análisis de regresión, algunas de ellas se mencionan en el capítulo 3, sección 3.3, donde se listan algunas investigaciones que se han publicado en el área farmacéutica, sin embargo, en todas estas no se explican a detalle los criterios estadísticos que se toman en cuenta para la aplicación del análisis de regresión que conlleva a la proposición de modelos para describir y predecir de manera significativa, en un sentido estadístico, una respuesta en función de las variables que la afectan en el dominio de estudio. Por lo tanto, es importante tomar en cuenta aquellos criterios estadísticos que conducen a la proposición de un modelo adecuado, lo cual se consigue cuando se conocen las bases teóricas del análisis de regresión, así como también es importante conocer, cuándo, dónde, porqué y quién puede aplicarlo.

Así, el análisis de regresión es una herramienta estadística que se emplea cuando se quiere encontrar la relación funcional entre una variable dependiente y al menos una variable independiente. Se aplica cuando se tiene un conjunto de datos que se han recolectado como resultado ya sea de un experimento de laboratorio, de un proceso de fabricación o de una optimización de un método analítico o proceso, entre otros. Este análisis estadístico se puede aplicar en cualquier área de la investigación, sin embargo, en este trabajo se muestran estudios de caso enfocados a la investigación farmacéutica, en donde se puede mencionar que el análisis de regresión se ha empleado en la construcción de curvas de calibración dentro del análisis químico cuantitativo y validación de técnicas y métodos analíticos; en la caracterización y optimización de procesos de fabricación de formas farmacéuticas en función de las variables que lo afectan; en el desarrollo y optimización de

formulaciones farmacéuticas en función de tipos, cantidades de excipientes y equipo de fabricación, entre otros.

Durante una planeación experimental, en cualquier área, incluyendo la investigación farmacéutica, se debe realizar una serie de actividades que conducen a conclusiones objetivas y confiables y así encontrar causas asignables de variación y proponer una mejora o encontrar un punto óptimo sobre el estudio que se realiza. Dentro de la planeación experimental se puede ubicar el análisis de regresión como una etapa del análisis de resultados, donde la aplicación de este análisis sólo es un paso para llegar a las conclusiones de la experimentación, es decir, no es la única herramienta que se debe emplear, más bien es recomendable emplear todas aquellas que estén al alcance del experimentador para obtener mayor información y tener más justificación para rechazar o no una hipótesis propuesta. Esta ubicación se muestra en la figura 9.1.

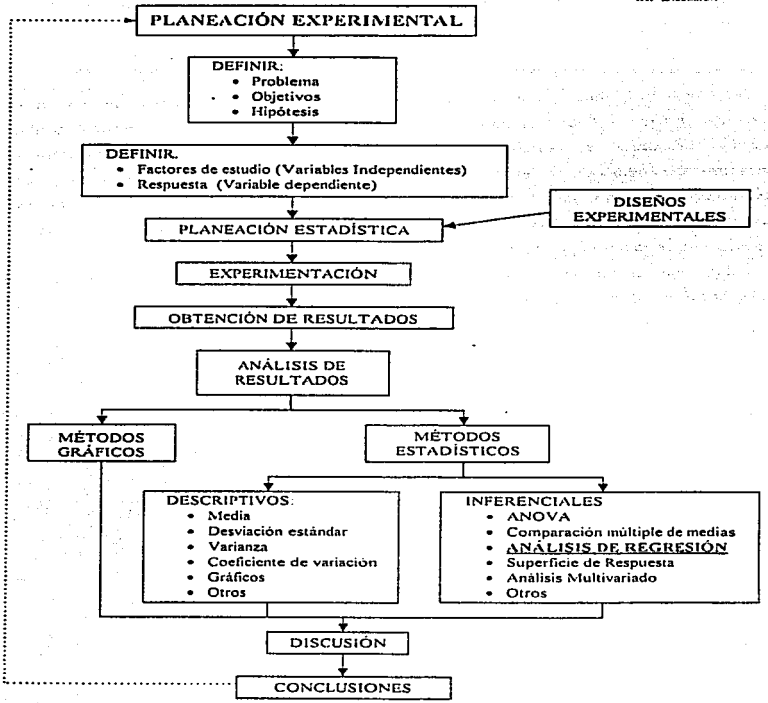


Figura 9.1. Ubicación del Análisis de Regresión en una Planeación Experimental

Durante el desarrollo de este trabajo se muestra la importancia del análisis de regresión, la cual radica en la caracterización y predicción de alguna variable de interés en función de variables que la afectan y que se pueden controlar por el experimentador. Para aplicar dicha herramienta estadística primero se analizaron los fundamentos teóricos del Análisis de Regresión lineal simple, los cuales sirven de base para entender el análisis de regresión múltiple. Es importante distinguir cuando se aplica la regresión lineal simple y cuando regresión múltiple, lo cual radica en el número de variables independientes, ya que cuando se tiene una sola variable "X" y una respuesta "Y" es necesario un análisis de regresión simple, por ejemplo en la construcción de una curva de calibración, donde la variable independiente es la concentración de una sustancia "A" la cual se puede controlar por el experimentador y la variable dependiente puede ser la respuesta analítica.

Por otro lado, cuando se tienen 2 o más variables independientes y una respuesta, es necesario aplicar un análisis de regresión múltiple, por ejemplo, en la caracterización de las propiedades farmacotécnicas de comprimidos, donde las variables independientes pueden ser aquellas relacionadas con el proceso de fabricación, como abertura de malla del tamizado, tipo y tiempo de mezclado, tiempo y temperatura de secado, velocidad y fuerza de compresión, o pueden ser en base a la formulación, como tipo y cantidad de lubricante, cantidad de excipientes, diluentes, desintegrantes, etc. y la variable de respuesta puede ser el tamaño de particula del granulado, dureza, tiempo de desintegración, friabilidad o tiempo de disolución de las tabletas. En el caso de regresión múltiple, las variables independientes pueden ser variables básicas o funciones de las variables básicas, un ejemplo de esta es:

$X_2 = X_1^2$ donde la variable X_1 es básica y la variable X_2 es función de X_1.

La representación matemática de ambos modelos se presenta en la figura 9.2.

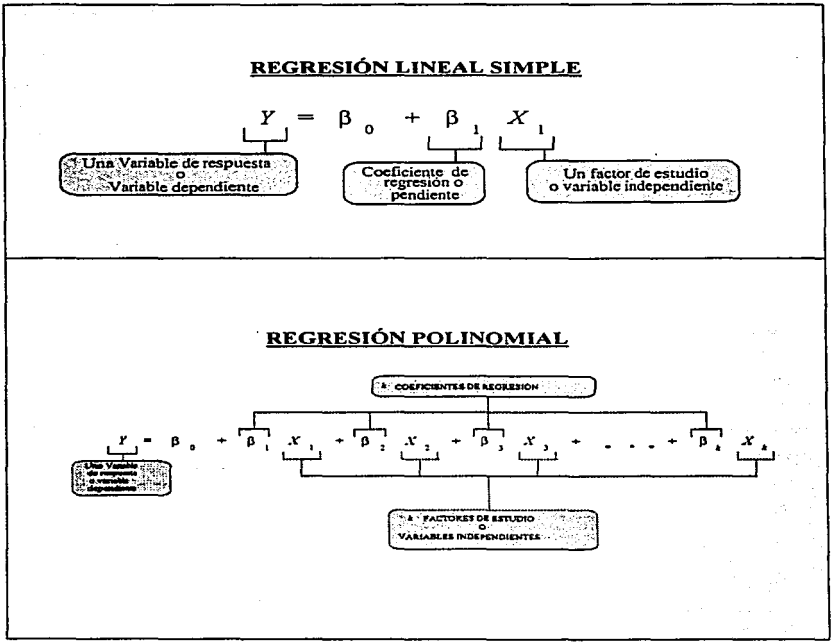


Figura 9.2. Modelos del análisis de regresión.

En este trabajo se analizaron los fundamentos teóricos de la técnica de regresión lineal simple, ya que es importante conocer la forma en que se ajusta un modelo de regresión a través de mínimos cuadrados, porque la mayoría de software estadísticos se basan en este método, así como también es importante conocer todas las pruebas estadísticas que se realizan para evaluar el modelo de regresión que se obtiene, ya que cada prueba proporciona un criterio de decisión sobre el mismo.

Por otro lado, fue necesario analizar lo referente a la regresión lineal simple debido a que es la base para entender o comprender la regresión múltiple, ya que ésta es en muchos aspectos una extensión de la regresión simple, la cual tiene la ventaja de que se puede representar en forma gráfica solo en dos dimensiones, así como también es más fácil el cálculo e interpretación de los parámetros relacionados a la regresión, y una vez que se entiende, se puede extender a regresión múltiple independientemente de que sean 2 o más variables independientes.

Así, una vez que se analizaron los fundamentos del análisis de regresión fue posible mostrar un estudio del área farmacéutica, con el cual se comprendió con mayor claridad como aplicar e interpretar la teoría del análisis de regresión lineal simple.

A través del estudio de caso se observaron los posibles problemas que se pueden presentar al aplicar la técnica de regresión, por ejemplo, se observó que no se cumplían todos los supuestos estadísticos, sin embargo, el modelo ajustado aparentemente era bueno, ya que la r^2 era grande y la pendiente era diferente de cero, pero se observó que el CME era demasiado grande, es decir, existe una gran variación en los datos por lo que no se cumplía el supuesto de homogeneidad de varianza, por lo tanto se realizó una transformación de las variables obteniéndose así un modelo que cumplía con los supuestos estadísticos, así como también con los demás criterios de decisión (r^2, CME, $\beta_1 = 0$ y residuales).

Por lo tanto, es de gran importancia analizar todos los parámetros o criterios de la regresión y no solo confiarse del coeficiente de determinación y CME, así, cuando el coeficiente de regresión β_1 es diferente de cero es recomendable analizar los supuestos estadísticos, debido a que pueden dar información acerca de la calidad o confiabilidad del modelo ajustado, ya que como se observó en el

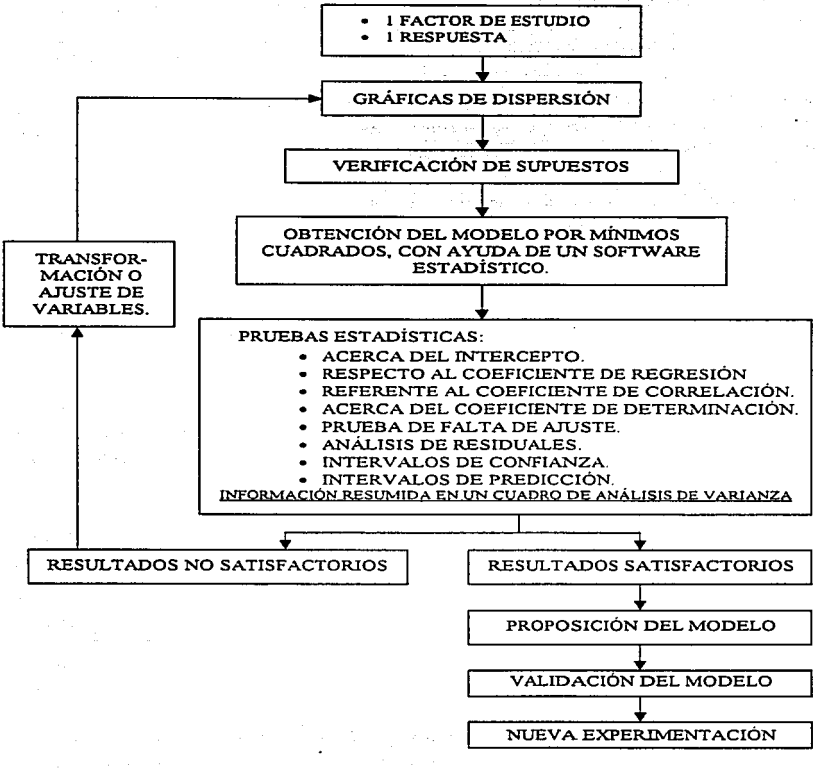
estudio de caso β_1 era diferente de cero, r^2 aparentemente cumplió, pero los supuestos no, por lo que se buscó un modelo que cumpliera con todos los criterios y fuera estadísticamente más confiable.

De esta forma, para que el análisis de regresión simple se lleve a cabo de forma exitosa, se recomienda seguir los siguientes pasos:

1. Observar la posible relación de las variables en forma gráfica, a través de un diagrama de dispersión.
2. Verificar supuestos estadísticos en forma gráfica.
3. Ajustar el modelo a través de mínimos cuadrados
4. Realizar pruebas estadísticas sobre el modelo ajustado.
5. Analizar residuales (verificación de supuestos).
6. Prueba de falta de ajuste.
7. Proposición del modelo de regresión.

Esta estrategia se muestra en la figura 9.3.

Figura 9.3. ANÁLISIS DE REGRESIÓN LINEAL SIMPLE



Por otro lado, se analizaron los fundamentos teóricos del análisis de regresión múltiple, los cuales se basan en gran parte sobre la teoría del análisis de regresión simple, ya que en este caso también es necesario evaluar el cumplimiento de supuestos y el modelo se ajusta a través de mínimos cuadrados. Sin embargo, no se puede considerar solamente una extensión de regresión simple debido a que la estrategia de aplicación cambia, como se comenta a continuación. También se realizan pruebas acerca del coeficiente de determinación y de los diversos coeficientes de regresión que permiten evaluar cuales variables son significativamente importantes en el modelo. Para esto existen diversas técnicas de selección de variables tales como:

a) Selección backward: Comienza con todas las variables en el modelo y se eliminan aquellas que no son significativas en la predicción de la variable de respuesta. Así el modelo queda solo con las variables que sí ayudan en la predicción de la respuesta.

b) Selección forward: Comienza con un modelo que incluye la variable de mayor influencia y posteriormente se adicionan de una en una en orden de importancia aquellas variables que son significativas para el modelo.

c) Selección Stepwise: Es una combinación de las dos técnicas anteriores ya que adiciona al modelo las variables de mayor importancia y elimina o adiciona otras variables al mismo tiempo, quedando el modelo con las variables que son más significativamente importantes para la predicción de la respuesta.

En el manejo de dos o más variables independientes es necesario considerar o tomar en cuenta aquellos criterios de selección de variables, además de los parámetros importantes que se tienen en regresión lineal simple, tales como r^2, CME y residuales.

Dentro del análisis de regresión múltiple, los parámetros que se deben evaluar son los valores de F total, que prueba si todas las variables independientes que se consideran al mismo tiempo explican significativamente la variación en Y, así como también el valor de F parcial, el cual evalúa si la adición de alguna variable independiente al modelo contribuye significativamente en la predicción de Y, por lo tanto, con este valor se puede evaluar la importancia de cada variable y así eliminar aquellas que no ayudan en la predicción de Y. Esto es de gran importancia, debido a que es la base de

las diferentes técnicas de selección de variables y gracias a ello es posible encontrar de manera más rápida y eficiente un modelo de regresión confiable.

Anteriormente lo que se hacía era seleccionar variables de forma manual, es decir, evaluar todas las variables y eliminar de una en una hasta quedar con las variables significativamente importante, pero ahora todo ese proceso se puede evitar si se tienen herramientas computacionales donde se puede realizar cualquier técnica de selección de variables y así llegar al mejor modelo de regresión.

Sin embargo, la decisión de qué técnica de selección aplicar depende del investigador y de sus objetivos y no siempre es necesario adoptar el modelo final propuesto por la técnica porque no siempre es el mejor, tal vez, al ver el análisis conviene quedarse un paso antes ya que también deben intervenir criterios de orden práctico. Por ejemplo, si en el último paso de la técnica que se utiliza, el valor de r^2 parcial es muy pequeño no tiene caso adicionar la variable en cuestión, ya que siempre lo que se busca es encontrar un modelo con el menor número de variables independientes posible. Por lo tanto, se observa que es necesario conocer los fundamentos teóricos de la regresión múltiple para que se tengan más elementos de decisión y poder proponer un mejor modelo.

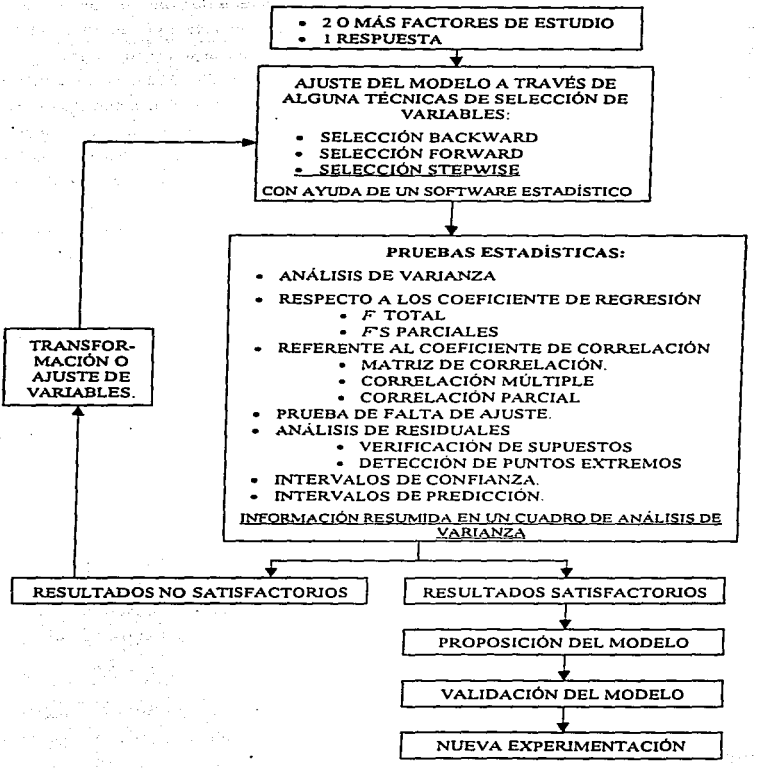
Otro aspecto importante en el análisis de regresión múltiple es la existencia de correlaciones múltiples, correlaciones parciales y correlaciones parcial-multiples, la primera es una generalización directa del coeficiente de correlación simple, ya que es una medida de la asociación lineal total de Y con todas las variables independientes. El coeficiente de correlación parcial es una medida de la mejora que puede tener el modelo al adicionar una variable tomando en cuenta los efectos de otras variables, por lo tanto para realizar una prueba de significancia de dicha variable se emplea el valor de F parcial. La correlación parcial-múltiple describe la relación total entre una variable dependiente y dos o más variables independientes tomando en cuenta los efectos de otras variables.

Una vez que se analizaron las herramientas teóricas de regresión múltiple, fue necesario aplicarlas en algunos estudios de caso, ya que la estrategia a seguir es en cierta forma diferente a la de regresión lineal simple.

Por lo tanto, para comenzar o decidir aplicar un análisis de regresión múltiple se deben tener dos o más variables independientes y una variable de respuesta. Una vez que se identifican dichas variables y se tienen los resultados experimentales, es de interés obtener una matriz de correlación, para observar cuales variables están más correlacionadas con la respuesta, ya que esta es la pauta para comenzar a seleccionar variables, para lo cual se debe escoger alguna técnica de regresión múltiple. En base al aspecto teórico-práctico, puede ser recomendable utilizar la técnica Stepwise, ya que esta es una combinación de Backward y Forward y por lo tanto puede encontrar la mejor combinación de variables que afecten o ayuden a predecir la respuesta. Sin embargo, si existe alguna duda en los resultados, se puede corroborar o apoyar en las otras técnicas de selección de variables, pero no es recomendable hacerlo de forma manual, o paso por paso, porque además de que se pierde información se emplea mucho tiempo de análisis.

De esta forma, se tiene que al aplicar un análisis de regresión múltiple, es importante conocer los criterios de selección de variables (r's parciales, F's-parciales, residuales), ya que en la práctica se debe tener un modelo con el menor número posible de variables. Así, en la figura 9.4 se propone una secuencia lógica para la aplicación del análisis de regresión múltiple que ayuda a encontrar el mejor modelo.

Figura 9.4. ANÁLISIS DE REGRESIÓN MÚLTIPLE



Una vez que se obtiene el mejor modelo de regresión múltiple, uno de los análisis más importantes es el de residuales (diferencia entre valor predicho por el modelo y valor observado), con los cuales se construyen gráficas para observar si existe algún incumplimiento de los supuestos estadísticos, esto se detecta por ejemplo cuando se presenta un efecto de embudo o una distribución no homogénea de los residuales. Si se presenta un efecto de embudo se considera que el supuesto de homogeneidad de varianza no se cumple por lo que quizá es necesario hacer una transformación de las variables; o si se presenta una tendencia curvilínea en los residuales, puede ser necesario incluir en el modelo un término cuadrático o cúbico. Así, se tiene que a través del análisis de residuales se pueden verificar los supuestos de regresión, así como también pueden detectarse aberraciones de la técnica como son puntos extremos (outliers), autocorrelación y multicolinealidad, por lo que se sugiere que además de la verificación de supuestos se realice este análisis ya que este tipo de situación podría conducir a un modelo que no es el mejor.

Otra prueba importante dentro del análisis de regresión es la de Falta de Ajuste o Bondad de Ajuste, la cual es común tanto para regresión simple como múltiple. Con esta prueba se puede determinar si el modelo propuesto ajusta de manera correcta los resultados experimentales o se debe buscar otro modelo mejor, ya sea incluyendo más términos o transformando variables. Para realizar esta prueba es necesario contar con observaciones repetidas, porque se debe estimar un error en cada nivel de las variables independientes.

Finalmente, una vez que se tiene el modelo que mejor ajusta los resultados, si se desea se puede probar o validar dicho modelo, siempre y cuando existan los recursos suficientes, ya que para esto se debe realizar una nueva experimentación dentro del dominio de estudio, debido a que el modelo ajustado no es válido para extrapolar, sino sólo se debe interpolar en el rango estudiado. Una vez que se experimenta se deben comparar los nuevos resultados con los valores predichos por el modelo y se debe observar un pequeña diferencia entre ambos, de lo contrario el modelo de regresión no es lo suficientemente confiable.

Por lo anterior, la revisión de los fundamentos teóricos del análisis de regresión y su aplicación a diversos casos de estudio proporcionan de manera conjunta todas las herramientas necesarias para la interpretación de resultados de un software estadístico, como SAS para Windows,

y no dejar a un lado algún criterio de decisión que pudiera ser determinante y por no conocer como se interpreta no se considere en la toma de una decisión importante. Por ejemplo, en la selección de una variable, que radica principalmente en la prueba de hipótesis de correlación parcial y que está directamente relacionado con el valor de *F*-parcial, se debe emplear la suma de cuadrados Tipo II. En SAS también aparece un valor de R^2 ajustada, el cual siempre es menor a R^2 normal, pero es más confiable debido a que se hace un ajuste con respecto al número de variables independientes y tiene la bondad de no aumentar solo por la adición de una variable independiente al modelo. Todos estos elementos de decisión los proporciona en forma rápida y veraz el software estadístico por lo que no deben ser omitidos en la interpretación.

Así, a través de este trabajo se muestra la gama de pruebas que se pueden y deben realizar para encontrar un buen modelo de predicción que posteriormente ayude a encontrar causas asignables de variación de algún proceso o en alguna experimentación del área farmacéutica, y así servir de apoyo para encontrar un punto óptimo de la variable de respuesta en función de los factores que la afectan.

Como se observa, este trabajo no sólo está dirigido al investigador farmacéutico sino al de cualquier área, sin embargo, los casos de estudio son del ámbito farmacéutico debido a que en esta área no existe un texto que trate resultados de este tipo, por lo que al farmacéutico se le dificulta más la comprensión de las interpretaciones, pero se espera que con este texto se familiarice con todos los términos estadísticos. Por otro lado, se han encontrado publicaciones en los que se aplica regresión lineal, sin mencionarlo como tal, ya que la mayoría comienza con un plan experimental a través de un diseño estadístico y termina con la proposición de un modelo de regresión, sin considerar todas las pruebas estadísticas que se necesitan para proponer el mejor modelo de regresión lineal.

X. CONCLUSIONES

- La conjunción de los aspectos teóricos con la aplicación de la técnica de regresión mostrando resultados concretos e interpretación resulta una buena estrategia para aplicar el análisis de regresión con más sentido estadístico y práctico, siendo más reales en los alcances y limitaciones de la técnica.

- Si se tiene una variable independiente o factor de estudio y una respuesta, se debe aplicar un análisis de regresión lineal simple.

- Si se tienen dos o más variables independientes y una variable de respuesta se debe aplicar un análisis de regresión múltiple.

- El conocimiento de los fundamentos y la visualización de manera gráfica del comportamiento de una regresión lineal simple ayuda a explorar la regresión lineal múltiple.

- La aplicación correcta de una secuencia lógica o estrategia de análisis con fundamentos teóricos conduce a un modelo que cumple con todos los criterios del análisis de regresión.

- Los parámetros importantes del análisis de regresión lineal simple son: coeficiente de determinación (r^2), cuadrado medio del error (CME), análisis de residuales, estimación por intervalos de confianza, verificación de supuestos estadísticos, prueba de falta de ajuste, prueba de significancia del intercepto y la pendiente del modelo.

- Los parámetros importantes del análisis de regresión múltiple son: matriz de correlación, coeficiente de correlación múltiple elevado al cuadrado (r^2-múltiple), cuadrado del coeficiente de correlación parcial (r^2-parcial), cuadrado del coeficiente de correlación múltiple-parcial (r^2-múltiple-parcial), valor de los estadísticos F total y F parcial, interacción entre variables, intervalos de confianza, análisis de residuales, prueba de falta de ajuste, verificación de supuestos.

- Los métodos de selección de variables más comunes son:
  - Selección Backward
  - Selección Forward
  - Selección Stepwise

  Se recomienda emplear el método de selección Stepwise, por ser una combinación de los dos primeros.

- En regresión múltiple, el método de selección de una variable depende de los objetivos del investigador y la posibilidad de manejarlo en un software le permite analizar las tres estrategias de manera rápida y utilizar los resultados que le sean más acordes.

- El análisis de la teoría de regresión ayuda a entender más los elementos de interpretación que aporta la paqueteria estadística.

- Se logró mostrar la importancia del análisis de regresión en la investigación farmacéutica así como también la necesidad de comprender los fundamentos teóricos de la técnica para su aplicación en un conjunto de datos.

XI. BIBLIOGRAFÍA

1. Bohidar N. R., (1993), *"Relative efficiency of R^2 and β^2 in regression analysis for calibration and formulation"*. **Drug Development and Industry Pharmacy**, (19:12, 1447-59).

2. Box George E. P., Hunter William G., Hunter Stuart, (1988), *"Estadistica para investigadores"*. España, Ed. Reverté. 653 pp.

3. Cid Cárcamo Edison, (1991), *"Cinética de disolución de Medicamentos"*. Secretaria General de la Organización de los Estados Americanos. Programa Regional de Desarrollo. Santiago de Chile.

4. Cochran William G., (1974), *"Análisis de Regresión múltiple"*. México, Ed. Continental. 703 pp.

5. Colombo, M. Bruno., (1976), *"Control of Phisical Properties. Farmaceutical Forms"*. Italia, Ed. Organizzazione. pp. 159 - 179.

6. Charlot M., Lewis G. A., Mathieu D., (1988), *"Experimental desing for pharmaceutical characterisation and optimisation using an exchange algorithm"*. **Drug Development and Industry Pharmacy**, (14:15-17, 2535-56).

7. Daniel Wayne W., (1990). *"Bioestadistica: Base para el análisis de las ciencias de la salud"*. 4a reimpresión, México, Ed. Limusa. 667 pp.

8. Dévay A., Kovács P. and Rácz Y., (1985), *"Optimization of chemical stability of diazepam in the liquid phase by means of factorial experimental desing"*. **Int. J. Pharm. Tech. & Prod. Mfr.**, (6:3, 5-9).

9. Dick, F. Christopher., Kassen, A. Roger and Amidon, E. Gregory., (1987), *"Determination of the sensitivity of a tablet formulation to variations in excipient levels and processing conditions using optimization techniques"*. **International Journal of Pharmaceutics.**, (38, 23-31).

10. Draper N. R., Smith H., (1981), *"Applied Regression Analysis"*. 2ª edition. U.S.A., Ed. Jhon Wiley & Sons. 709 pp.

11. Elkheshen, A. Seham, Badawi, S. Sabry and Badawi, A. Alia., (1996), *"Optimization of a reconstitutable suspension of Rifampicina using 2^4 factorial desing"*. **Drug Development and Industry Pharmacy**, (22:7, 623-630).

12. Harris, M. R.., Schwartz, J. B. and McGinity, (1985), *"Optimization of a slow-release tablet formulation containing sodium sulfathiazole and a montmorillonite clay"*. **Drug Development and Industry Pharmacy**, (11:5, 1089-1110).

13. Hamilton Lawrence C., (1990), *"Modern data analysis. A first course in applied statistics"*. U.S.A., Ed. New hamsphire Reooks/ Cole Publishing Company Pacific Grove California. 684 pp.

14. Harnett L. Donald, Murphy L. James, (1987), *"Introducción al análisis estadistico"*. México, Ed. Addison-Wesley Iberoamericana. p 486, 525.

15. Jiménez Ortega Alicia Natalia, (1997), *"Estudio de la constante del perfil de disolución de comprimidos de furosemida por matrices hidrofílicas"*. Tesis para obtener el titulo de Químico Farmacéutico Biologo. UNAM (FES-Zaragoza).

16. Kleinbaum G. David; Kupper Lawrence L., (1978), *"Applied Regression Analysis and Other Multivariable Methods"*. U.S.A., Ed. Duxbury press. 556 pp.

17. Kreyzing Erwin, (1981), *"Introducción a la estadística matemática. Principios y métodos"*. 6a. edición. México, Ed. Limusa. 505 pp.

18. Lachman, Leon and Schwartz, B. J., (1990), *"Pharmaceutical Dosage forms. Tablets: vol. 2"*. 2a. edition. U.S.A., Ed. Mercel Dekker. pp. 309-316.

19. Lindberg N-O, Holmquist B., (1987), *"Optimizing the friability of a tablet formulation"*. **Drug Development and Industry Pharmacy**, (13:6, 1063-67).

20. Merkku Pasi, Antikainen Osmo and Yliruusi Jouko, (1993), *"Use of 3^3 Factorial Desing and Multilinear Stepwise Regression Analysis in Studying the Fluidized Bed Granulation Process, Part II"*. **Eur. J. Pharm. Biopharm.** (39:3, 112-116).

21. Merkku Pasi, Yliruusi Jouko, (1993), *"Use of 3^3 Factorial Desing and Multilinear Stepwise Regression Analysis in Studying the Fluidized Bed Granulation Process, Part I"*. **Eur. J. Pharm. Biopharm.** (39:2, 75-81).

22. Montgomery Douglas C., [Trad. Jaime Saldivar Delgado], (1991), *"Diseño y Análisis de Experimentos"*.México, Ed. Iberoamérica. 589 pp.

23. Montgomery Douglas C., Peck Elizabeth A., (1992), *"Introduction to linear regression analysis"*. 2nd ed. U.S.A., Ed. John Wiley & Sons. 527 pp.

24. Myers Raymond H., (1989), *"Classical and Modern Regression with applications"*. 2a. edition. U.S.A., 488 pp.

25. Sanderak E., Merrell Down Marion, et al., (1993), *"Response surface methodology as an aproach to optimization of an oral solution"*. **Drug Development and Industry Pharmacy**, (19:4, 405-424).

26. SAS Institute Inc., SAS/LAB® (1993) *"Softwere: Graphics editor", version 6, First edition,* cary, NC: SAS Institute. págs. 95-130.

27. SAS Institute Inc., SAS/STAT® (1987) *"Guide for Personal Computers", version 6, First edition,* cary, NC: SAS Institute Inc. 1028 pp.

28. SAS Institute Inc., SAS/STAT® (1989) *"Softwere: User's guide", version 6, First edition, vol. 1* cary, NC: SAS Institute. 943 pp.

29. SAS Institute Inc., SAS/STAT® (1989) *"Softwere: User's guide", version 6, First edition , vol. 2* cary, NC: SAS Institute. 705 pp.

30. SAS Institute Inc., SAS®, (1990) *"Softwere: Procedures guide", version 6, Third edition,* cary, NC: SAS Institute. 705 pp.

31. Seber G. A. F., (1977), *"Linear Regression Analysis"*. U.S.A., Ed. John Wiley & Sons. 465 pp.

32. Shirakura O. Yolanda M., Hashimoto M., et al., (1991), *"Particule size desing using computer optimization technique"*. **Drug Development and Industry Pharmacy,** (17:4, 471-483).

33. Stetsko G., (1986), *"Statistical experimental design and its application to pharmaceutical development problems"*. **Drug Development and Industry Pharmacy,** (12, 1109-23).

34. Waaler P. J., Graffner C. and Müller B. W., (1992), *"Optimization of a matriz tablet formulation using a mixyure desing"*. **Acta Pharmaceutica Nordica,** (4:1, 9-16).

ANEXOS

Anexo A: Salida de SAS 6.3 del programa 6.3: Programación de SAS para obtener estadísticos básicos (Capítulo 6).

Anexo B: Salida de SAS 6.6. del programa 6.6: Programación de SAS para obtener modelos de Regresión Lineal Simple y sus correspondientes intervalos de confianza y residuales (Capítulo 6).

ANEXO-A

SALIDA SAS 6.3.

(Resultados del programa 6.3)

```
OPTIONS PS=60 NODATE;
DATA lote0;
INPUT lote$ min disolu @@;
lmin=Log(min);
ldis=Log(disolu);
CARDS;
0 1  10  1.79    0 1  20  3.66    0 1  30  6.00    0 1  60 12.60
0 1 120 30.14    0 1 180 47.79    0 1 240 63.43    0 1 300 79.09
0 1 360 91.53    0 1 420 96.58    0 1 480 98.98    0 2  10  1.40
0 2  20  2.97    0 2  30  4.32    0 2  60  9.70    0 2 120 22.83
0 2 180 38.23    0 2 240 52.96    0 2 300 70.61    0 2 360 88.12
0 2 420 98.23    0 2 480 107.76   0 3  10  1.20    0 3  20  2.45
0 3  30  3.70    0 3  60  8.22    0 3 120 19.44    0 3 180 31.57
0 3 240 43.92    0 3 300 56.18    0 3 360 67.57    0 3 420 75.60
0 3 480 82.24    0 4  10  1.87    0 4  20  3.66    0 4  30  5.61
0 4  60 12.58    0 4 120 28.95    0 4 180 48.57    0 4 240 68.23
0 4 300 83.74    0 4 360 95.38    0 4 420 98.30    0 4 480 98.83
0 5  10  1.35    0 5  20  2.75    0 5  30  4.57    0 5  60 11.98
0 5 120 27.31    0 5 180 42.45    0 5 240 58.54    0 5 300 73.96
0 5 360 85.75    0 5 420 90.77    0 5 480 94.21    0 6  10  1.40
0 6  20  2.88    0 6  30  4.50    0 6  60 10.56    0 6 120 26.05
0 6 180 41.90    0 6 240 58.55    0 6 300 75.52    0 6 360 89.62
0 6 420 100.5    0 6 480 101.29
;
proc sort;
by min;
PROC univariate plot;
var disolu;
by min;
proc sort;
by lmin;
PROC univariate plot;
var ldis;
by lmin;
TITLE1 'OBSERVACIONES DE DISOLUCION';
TITLE2 'LOTE 0';
RUN;
```

```
                          The SAS System                                1
------------------------- MIN=10 -------------------------------------------
                        Univariate Procedure
Variable=DISOLU

                              Moments

N                          6          Sum Wgts                   6
Mean                   1.515          Sum                     9.09
Std Dev            0.261744          Variance            0.068531
Skewness           0.472191          Kurtosis           -1.62318
USS                  14.1139          CSS                 0.34255
CV                 17.27684          Std Mean            0.106857
T:Mean=0           14.17788          Pr>|T|                0.0001
Num ^= 0                 6          Num > 0                    6
M(Sign)                  3          Pr>=|M|               0.0313
Sgn Rank              10.5          Pr>=|S|               0.0313

                          Quantiles(Def=5)

                  100% Max        1.87          99%        1.87
                   75% Q3         1.79          95%        1.87
                   50% Med        1.44          90%        1.87
                   25% Q1         1.35          10%         1.2
                    0% Min         1.2          5%          1.2
                                                1%          1.2
                  Range           0.67
                  Q3-Q1           0.44
                  Mode             1.2

                            Extremes

                    Lowest    Obs      Highest    Obs
                       1.2(    3)      1.35(    5)
                      1.35(    5)      1.4(     2)
                       1.4(    2)      1.44(    6)
                      1.44(    6)      1.79(    1)
                      1.79(    1)      1.87(    4)

Stem Leaf                          #             Boxplot
  18 7                             1                |
  17 9                             1                +-----+
  16                               0                |     |
  15                               0                |  +  |
  14 00                            2                *-----*
  13 5                             1                +-----+
  12 0                             1                  |
      ----+----+----+----+
  Multiply Stem.Leaf by 10**-1

                      Normal Probability Plot
1.85+                                          *+**
    |                                      * ***
    |                                  ****
    |                              ****
1.55+                          ****
    |                      ***  *
    |                  ***
    |              ****
1.25+          ****
    +----+----+----+----+----+----+----+----+----+----+
        -2        -1         0        +1        +2
```

MIN=20

Univariate Procedure

Variable=DISOLN

Moments

N	6	Sum Wgts	6
Mean	3.061667	Sum	18.37
Std Dev	0.495718	Variance	0.245737
Skewness	0.400251	Kurtosis	-1.53316
USS	57.4715	CSS	1.228683
CV	16.19113	Std Mean	0.202376
T:Mean=0	15.12859	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	3.66	99%	3.66
75% Q3	3.66	95%	3.66
50% Med	2.925	90%	3.66
25% Q1	2.75	10%	2.45
0% Min	2.45	5%	2.45
		1%	2.45
Range	1.21		
Q3-Q1	0.91		
Mode	3.66		

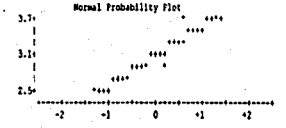
Extremes

Lowest	Obs	Highest	Obs
2.45(	3)	2.75(	5)
2.75(	5)	2.88(	6)
2.88(	6)	2.97(	2)
2.97(	2)	3.66(	1)
3.66(	4)	3.66(	4)

```
Stem Leaf                 #          Boxplot
 36 66                    2          +------+
 34                                  |      |
 32                                  |      |
 30                                  |  +   |
 28 87                    2          +------+
 26 5.                    1          +------+
 24 5                     1          |
    ----+----+----+----+
```

Multiply Stem.Leaf by 10**+1

Normal Probability Plot



MIN=30

Univariate Procedure

Variable=DISOLN

Moments

N	6	Sum Wgts	6
Mean	4.783333	Sum	28.7
Std Dev	0.85764	Variance	0.735547
Skewness	0.457773	Kurtosis	-0.95798
USS	140.9594	CSS	3.677733
CV	17.92976	Std Mean	0.35013
T:Mean=0	13.66159	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

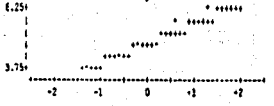
100% Max	6	99%	6
75% Q3	5.61	95%	6
50% Med	4.535	90%	6
25% Q1	4.32	10%	3.7
0% Min	3.7	5%	3.7
		1%	3.7
Range	2.3		
Q3-Q1	1.29		
Mode	3.7		

Extremes

Lowest	Obs	Highest	Obs
3.7(	3)	4.32(	2)
4.32(	2)	4.56(	6)
4.5(	6)	4.57(	5)
4.57(	5)	5.61(	4)
5.61(	4)	6(	1)

```
Stem Leaf                 #          Boxplot
 6 0                      1          |
 5 6                      1          +------+
 5                                   |      |
 4 56                     2          +--+---+
 4 3                      1          +------+
 3 7                      1          |
   ----+----+----+----+
```

Normal Probability Plot



---------- MIN=60 ----------

Univariate Procedure

Variable=DISOLU

Moments

N	6	Sum Wgts	6
Mean	10.94	Sum	65.64
Std Dev	1.76676	Variance	3.12144
Skewness	-0.65143	Kurtosis	-0.9969
USS	733.7088	CSS	15.6072
CV	16.14954	Std Mean	0.721277
T:Mean=0	15.16755	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	12.6	99%	12.6
75% Q3	12.58	95%	12.6
50% Med	11.27	90%	12.6
25% Q1	9.7	10%	8.22
0% Min	8.22	5%	8.22
		1%	8.22

Range	4.38
Q3-Q1	2.88
Mode	8.22

Extremes

Lowest	Obs	Highest	Obs
8.22(	3)	9.7(	2)
9.7(	5)	10.56(	6)
10.56(	6)	11.98(	5)
11.98(	5)	12.58(	4)
12.58(	4)	12.6(	1)

Stem	Leaf	#	Boxplot
12	66	2	+------+
12	0	1	| |
11			| |
11			+------+
10	6	1	| * |
10			| |
9	7	1	+------+
9			|
8			|
8	2	1	|

----+----+----+----+

Normal Probability Plot

```
12.75+                                  * *** *
     |                                 ***
     |                              * **
11.25+                            ***
     |                         ****
     |                      ***
     |                   ***
 9.75+                **
     |              ***
     |            **
     |         ***
 8.25+      ****
     +----+----+----+----+----+----+----+----+----+
               -2        -1         0        +1        +2
```

---------- MIN=120 ----------

Univariate Procedure

Variable=DISOLU

Moments

N	6	Sum Wgts	6
Mean	25.78667	Sum	154.72
Std Dev	4.009243	Variance	16.07403
Skewness	-0.76243	Kurtosis	-0.38928
USS	4070.083	CSS	80.37013
CV	15.54774	Std Mean	1.636766
T:Mean=0	15.75464	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	30.14	99%	30.14
75% Q3	28.95	95%	30.14
50% Med	26.68	90%	30.14
25% Q1	22.83	10%	19.44
0% Min	19.44	5%	19.44
		1%	19.44

Range	10.7
Q3-Q1	6.12
Mode	19.44

Extremes

Lowest	Obs	Highest	Obs
19.44(	3)	22.83(	2)
22.83(	2)	26.05(	6)
26.05(	6)	27.31(	5)
27.31(	5)	28.95(	4)
28.95(	4)	30.14(	1)

Stem	Leaf	#	Boxplot
30	1	1	|
28	9	1	+------+
26	03	2	+------+
24			| * |
22	8	1	+------+
20			|
19	4	1	|

----+----+----+----+

Variable=DISOLU

Normal Probability Plot

```
31+                                 *****
  |                              *****
  |                           ***
25+                        *****
  |                     *****
  |                  ****
  |               ***
19+            *****
  |         *****
  |      ****
  +----+----+----+----+----+----+----+----+----+
            -2        -1         0        +1        +2
```

---------- MIN=180 ----------

Univariate Procedure

Variable=DISOLU

Moments

N	6	Sum Wgts	6
Mean	41.75167	Sum	250.51
Std Dev	6.317108	Variance	39.90566
Skewness	-0.66843	Kurtosis	0.074719
USS	10658.74	CSS	199.5283
CV	15.13019	Std Mean	2.578949
T:Mean=0	16.18941	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	48.57	99%	48.57
75% Q3	47.79	95%	48.57
50% Med	42.175	90%	48.57
25% Q1	38.23	10%	31.57
0% Min	31.57	5%	31.57
		1%	31.57
Range	17		
Q3-Q1	9.56		
Mode	31.57		

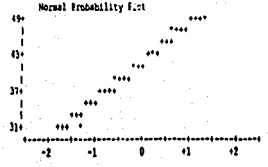
Extremes

Lowest	Obs	Highest	Obs
31.57(	3)	38.23(	2)
38.23(	2)	41.9(	6)
41.9(	6)	42.45(	5)
42.45(	5)	47.79(	1)
47.79(	1)	48.57(	4)

Stem	Leaf	#	Boxplot
48	6	1	|
46	8	1	+-----+
44			| |
42	4	1	*-----*
40	9	1	| + |
38	2	1	+-----+
36			|
34			|
32			|
30	6	1	|

----+----+----+----+

Normal Probability Plot



---------- MIN=240 ----------

Univariate Procedure

Variable=DISOLU

Moments

N	6	Sum Wgts	6
Mean	57.605	Sum	345.63
Std Dev	8.455102	Variance	71.48875
Skewness	-0.62753	Kurtosis	0.521229
USS	20267.46	CSS	357.4437
CV	14.67772	Std Mean	3.451781
T:Mean=0	16.68849	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	68.23	99%	68.23
75% Q3	63.43	95%	68.23
50% Med	58.345	90%	68.23
25% Q1	52.96	10%	43.92
0% Min	43.92	5%	43.92
		1%	43.92
Range	24.31		
Q3-Q1	10.47		
Mode	43.92		

Extremes

Lowest	Obs	Highest	Obs
43.92(	3)	52.96(	2)
52.96(	2)	58.54(	5)
58.54(	5)	58.55(	6)
58.55(	6)	63.43(	1)
63.43(	1)	68.23(	4)

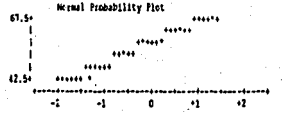
Stem	Leaf	#	Boxplot
6	8	1	|
6	3	1	+-----+
5	99	2	*-----*
5	2	1	+-----+
4			|
4	4	1	|

----+----+----+----+

Multiply Stem.Leaf by 10**+1

Variable=DISOLU

Normal Probability Plot



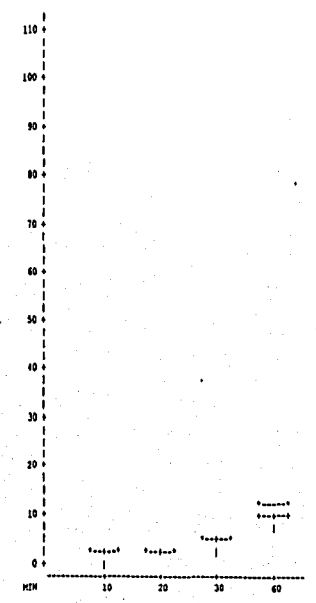
```
------------------------- MIN=300 ----------------------------------
                    Univariate Procedure
Variable=DISOLU
                          Moments

              N               6  Sum Wgts          6
              Mean     73.31667  Sum           439.9
              Std Dev  9.571986  Variance   91.62291
              Skewness -1.23995  Kurtosis   2.087509
              USS      32710.12  CSS        458.1145
              CV       13.05567  Std Mean   3.907747
              T:Mean=0 18.76188  Pr>|T|       0.0001
              Num ^= 0        6  Num > 0           6
              M(Sign)         3  Pr>=|M|      0.0313
              Sgn Rank     10.5  Pr>=|S|      0.0313

                       Quantiles(Def=5)

               100% Max     83.74       99%     83.74
                75% Q3      79.89       95%     83.74
                50% Med     74.74       90%     83.74
                25% Q1      70.61       10%     56.18
                 0% Min     56.18        5%     56.18
                                         1%     56.18

               Range        27.56
               Q3-Q1         9.28
               Mode         56.18

                          Extremes
                Lowest    Obs    Highest    Obs
                 56.18(     3)     70.61(     2)
                 70.61(     2)     73.96(     5)
                 73.96(     5)     75.52(     6)
                 75.52(     6)     79.89(     1)
                 79.89(     1)     83.74(     4)

          Stem Leaf                      #             Boxplot
             8 04                        2             +-----+
             7 6                         1             *-----*
             7 14                        2             +--+--+
             6
             6
             5 6                         1                0
               ----+----+----+----+
          Multiply Stem.Leaf by 10**+1

                       Normal Probability Plot
        82.5+                                    ++++++*
            |                              *+++*
            |                     *   +*+++
            |                 +++++
            |             +++++
        57.5+     ++++++*
             +----+----+----+----+----+----+----+----+----+----+
                 -2        -1         0        +1        +2
```

```
------------------------- MIN=360 ----------------------------------
                    Univariate Procedure
Variable=DISOLU
                          Moments

              N               6  Sum Wgts          6
              Mean     86.32833  Sum          517.97
              Std Dev  9.749304  Variance   95.04894
              Skewness -1.83789  Kurtosis   3.930341
              USS      45190.73  CSS        475.2447
              CV       11.29328  Std Mean   3.980137
              T:Mean=0 21.68979  Pr>|T|       0.0001
              Num ^= 0        6  Num > 0           6
              M(Sign)         3  Pr>=|M|      0.0313
              Sgn Rank     10.5  Pr>=|S|      0.0313

                       Quantiles(Def=5)

               100% Max     95.38       99%     95.38
                75% Q3      91.53       95%     95.38
                50% Med     88.87       90%     95.38
                25% Q1      85.75       10%     67.57
                 0% Min     67.57        5%     67.57
                                         1%     67.57

               Range        27.81
               Q3-Q1         5.78
               Mode         67.57

                          Extremes
                Lowest    Obs    Highest    Obs
                 67.57(     3)     85.75(     5)
                 85.75(     5)     88.12(     2)
                 88.12(     2)     89.62(     6)
                 89.62(     6)     91.53(     1)
                 91.53(     1)     95.38(     4)

          Stem Leaf                      #             Boxplot
             9 5                         1                |
             9 02                        2             +-----+
             8 68                        2             *--+--*
             8
             7
             7
             6 8                         1                *
               ----+----+----+----+
          Multiply Stem.Leaf by 10**+1

                       Normal Probability Plot
        97.5+                                      ++++*+
            |                                 ++*++
            |                     *   *+++*+
        82.5+                     +++++
            |                 +++++
            |            +++++
        67.5+     +++++   *
             +----+----+----+----+----+----+----+----+----+----+
                 -2        -1         0        +1        +2
```

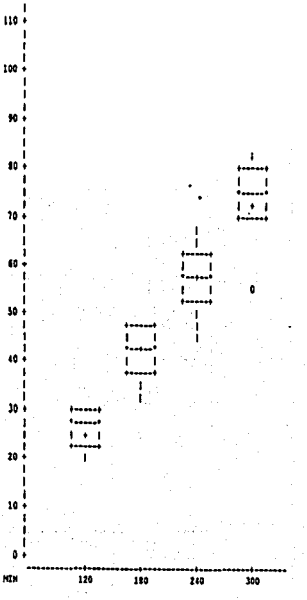
```
------------------------- MIN=420 ---------------------------------
                    Univariate Procedure
Variable=DISOLU

                              Moments

              N                 6  Sum Wgts          6
              Mean          93.33  Sum          559.98
              Std Dev    9.290735  Variance   86.31776
              Skewness   -1.85057  Kurtosis   3.408245
              USS        52694.57  CSS        431.5888
              CV         9.954735  Std Mean   3.792927
              T:Mean=0   24.60633  Pr>|T|       0.0001
              Num ^= 0          6  Num > 0           6
              M(Sign)           3  Pr>=|M|      0.0313
              Sgn Rank       10.5  Pr>=|S|      0.0313

                           Quantiles(Def=5)

               100% Max     100.5          99%     100.5
                75% Q3       98.3          95%     100.5
                50% Med    97.405          90%     100.5
                25% Q1      90.77          10%      75.6
                 0% Min      75.6           5%      75.6
                                            1%      75.6

               Range         24.9
               Q3-Q1         7.53
               Mode          75.6

                              Extremes
                 Lowest    Obs     Highest    Obs
                   75.6(     3)      90.77(     5)
                  90.77(     5)      96.58(     1)
                  96.58(     1)      98.23(     2)
                  98.23(     2)       98.3(     4)
                   98.3(     4)      100.5(     6)

          Stem Leaf                     #             Boxplot
            10 0                        1                |
             9 788                      3             +-----+
             9 1                        1             +--+--+
             8
             8
             7 6                        1                0
               ----+----+----+----+
          Multiply Stem.Leaf by 10**+1

                      Normal Probability Plot
        102.5+                                   +++++*
             |                       * . *++++*+
             |                   *  -++++
             |                 +++++
             |           ++++++
         77.5+      +++++ *
              +----+----+----+----+----+----+----+----+----+----+
                  -2        -1         0        +1        +2
```

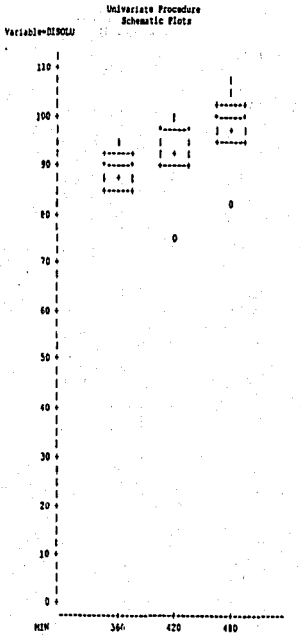
```
------------------------- MIN=480 ---------------------------------
                    Univariate Procedure
Variable=DISOLU
                              Moments
              N                 6  Sum Wgts          6
              Mean       97.21833  Sum          583.31
              Std Dev    8.565123  Variance   73.36134
              Skewness   -1.02125  Kurtosis   1.951361
              USS        57075.23  CSS        366.8067
              CV         8.810193  Std Mean   3.496697
              T:Mean=0   27.80292  Pr>|T|       0.0001
              Num ^= 0          6  Num > 0           6
              M(Sign)           3  Pr>=|M|      0.0313
              Sgn Rank       10.5  Pr>=|S|      0.0313

                           Quantiles(Def=5)

               100% Max    107.76          99%    107.76
                75% Q3     101.29          95%    107.76
                50% Med    98.905          90%    107.76
                25% Q1      94.21          10%     82.24
                 0% Min     82.24           5%     82.24
                                            1%     82.24
               Range        25.52
               Q3-Q1         7.08
               Mode         82.24

                              Extremes
                 Lowest    Obs     Highest    Obs
                  82.24(     3)      94.21(     5)
                  94.21(     5)      98.83(     4)
                  98.83(     4)      98.98(     1)
                  98.98(     1)     101.29(     6)
                 101.29(     6)     107.76(     2)

          Stem Leaf                     #             Boxplot
            10 8                        1                |
            10 1                        1             +-----+
             9 99                       2             *--+--*
             9 4                        1             +-----+
             8
             8 2                        1                0
               ----+----+----+----+
          Multiply Stem.Leaf by 10**+1

Variable=DISOLU
                      Normal Probability Plot
        107.5+                                      +++*+
             |                                 ++*+++
             |                            *+++*+
             |                      ++*+++
             |                 ++++++
         82.5+      ++++++ *
              +----+----+----+----+----+----+----+----+----+----+
                  -2        -1         0        +1        +2
```

Variable=DISOLU



Variable=DISOLU




Univariate Procedure
Schematic Plots
Variable=DIBOLW
110
100
90
80
70
60
50
40
30
20
10
0
MIN
360
420
480

LOTE 6'

---------- LMIN=2.302585093 ----------

Univariate Procedure

Variable=LDIS

Moments

N	6	Sum Wgts	6
Mean	0.403182	Sum	2.419095
Std Dev	0.170676	Variance	0.02913
Skewness	0.279057	Kurtosis	-1.25915
USS	1.120948	CSS	0.145632
CV	42.33226	Std Mean	0.069678
T:Mean=0	5.78634	Pr>|T|	0.0022
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	0.625938	99%	0.625938
75% Q3	0.582216	95%	0.625938
50% Med	0.364257	90%	0.625938
25% Q1	0.300105	10%	0.182322
0% Min	0.182322	5%	0.182322
		1%	0.182322

Range	0.443617		
Q3-Q1	0.282111		
Mode	0.182322		

Extremes

Lowest	Obs	Highest	Obs
0.182322(	3)	0.300105(	5)
0.300105(	5)	0.336472(	2)
0.336472(	2)	0.392042(	4)
0.392042(	4)	0.582216(	1)
0.582216(	1)	0.625938(	6)

```
Stem Leaf                     #             Boxplot
 6 3                          1                |
 5 8                          1            +------+
 5                                         |      |
 4                                         |      |
 4                                         |  +   |
 3 9                          1            *------*
 3 04                         2            +------+
 2                                         |      |
 2                                         |      |
 1 8                          1            |      |
    ----+----+----+----+----+
```

Multiply Stem.Leaf by 10**-1

Normal Probability Plot

```
0.625+                                  +**
     |                                *  ***
     |                              ***
0.475+                           ***
     |                        ***
     |                     ***  *
0.325+                  +**  *
     |                ***
     |             ***
0.175+          ***
     +----+----+----+----+----+----+----+----+
          -2   -1    0   +1   +2
```

---------- LMIN=2.9957322736 ----------

Univariate Procedure

Variable=LDIS

Moments

N	6	Sum Wgts	6
Mean	1.108161	Sum	6.648967
Std Dev	0.160556	Variance	0.025778
Skewness	0.216396	Kurtosis	-1.31523
USS	7.497019	CSS	0.128891
CV	14.48848	Std Mean	0.065547
T:Mean=0	16.90646	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	1.297463	99%	1.297463
75% Q3	1.297463	95%	1.297463
50% Med	1.073176	90%	1.297463
25% Q1	1.011601	10%	0.896088
0% Min	0.896088	5%	0.896088
		1%	0.896088

Range	0.401375		
Q3-Q1	0.285862		
Mode	1.297463		

Extremes

Lowest	Obs	Highest	Obs
0.896088(	3)	1.011601(	5)
1.011601(	5)	1.05779(	6)
1.05779(	6)	1.088562(	2)
1.088562(	2)	1.297463(	1)
1.297463(	4)	1.297463(	4)

```
Stem Leaf                     #             Boxplot
13 00                        2            +------+
12                                        |      |
12                                        |      |
11                                        |      |
11                                        |  +   |
10 69                        2            *------*
10 1                         1            +------+
 9                                         |      |
 9 0                         1            |      |
 8                                         |      |
    ----+----+----+----+----+
```

Multiply Stem.Leaf by 10**-1

Normal Probability Plot

```
1.325+                                  ++++
     |                                *  *** *
     |                              ***
1.175+                           ***
     |                        ***
     |                     ***  *
1.025+                  +**  *
     |                ***
     |             ***
0.875+          ***
     +----+----+----+----+----+----+----+----+
          -2   -1    0   +1   +2
```

---------- LMIN=3.401197817 ----------

Univariate Procedure

Variable=LDIS

Moments

N	6	Sum Wgts	6
Mean	1.551915	Sum	9.311489
Std Dev	0.177685	Variance	0.031572
Skewness	0.195171	Kurtosis	-0.79234
USS	14.6065	CSS	0.15786
CV	11.4494	Std Mean	0.07254
T:Mean=0	21.39405	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	1.791759	99%	1.791759
75% Q3	1.724551	95%	1.791759
50% Med	1.511795	90%	1.791759
25% Q1	1.463255	10%	1.308333
0% Min	1.308333	5%	1.308333
		1%	1.308333
Range	0.483427		
Q3-Q1	0.261295		
Mode	1.308333		

Extremes

Lowest	Obs	Highest	Obs
1.308333(	3)	1.463255(	2)
1.463255(	2)	1.504077(	6)
1.504077(	6)	1.519513(	5)
1.519513(	5)	1.724551(	4)
1.724551(	4)	1.791759(	1)

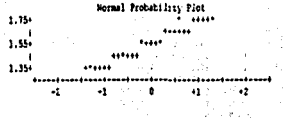
Stem	Leaf	#	Boxplot
17	29	2	+------+
16			| |
15	0C	2	*------*
14	6	1	+------+
13	1	1	|

----+----+----+----+

Multiply Stem.Leaf by 10**-1

Variable=LDIS

Normal Probability Plot



---------- LMIN=1.094341622 ----------

Univariate Procedure

Variable=LDIS

Moments

N	6	Sum Wgts	6
Mean	2.388802	Sum	14.28481
Std Dev	0.169906	Variance	0.028868
Skewness	-0.84449	Kurtosis	-0.39037
USS	34.15365	CSS	0.14434
CV	7.136503	Std Mean	0.069366
T:Mean=0	34.32339	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	2.533697	99%	2.533697
75% Q3	2.532108	95%	2.533697
50% Med	2.420156	90%	2.533697
25% Q1	2.272126	10%	2.10657
0% Min	2.10657	5%	2.10657
		1%	2.10657
Range	0.427127		
Q3-Q1	0.259982		
Mode	2.10657		

Extremes

Lowest	Obs	Highest	Obs
2.10657(	3)	2.272126(	2)
2.272126(	2)	2.357073(	6)
2.357073(	6)	2.483239(	5)
2.483239(	5)	2.532108(	4)
2.532108(	4)	2.533697(	1)

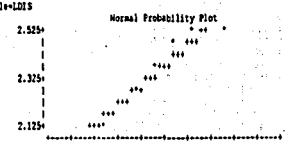
Stem	Leaf	#	Boxplot
25	33	2	+------+
24	8	1	| |
24			*------*
23	6	1	| + |
23			| |
22	7	1	+------+
22			|
21			|
21	1	1	|

----+----+----+----+

Multiply Stem.Leaf by 10**-1

Variable=LDIS

Normal Probability Plot



```
---------------------- LMIN=4.7874917428 ----------------------------
                    Univariate Procedure
Variable=LDIS
                           Moments
              N                 6  Sum Wgts           6
              Mean       3.239017  Sum          19.4341
              Std Dev    0.164358  Variance    0.027014
              Skewness   -0.97445  Kurtosis    0.131832
              USS        63.08246  CSS         0.135068
              CV         5.074313  Std Mean    0.067099
              T:Mean=0   48.27234  Pr>|T|        0.0001
              Num ^= 0          6  Num > 0            6
              M(Sign)           3  Pr>=|M|       0.0313
              Sgn Rank       10.5  Pr>=|S|       0.0313
                        Quantiles(Def=5)
               100% Max  3.405853          99%  3.405853
                75% Q3    3.36557          95%  3.405853
                50% Med  3.283635          90%  3.405953
                25% Q1   3.128075          10%  2.967333
                 0% Min  2.967333           5%  2.967333
                                            1%  2.967333
               Range      0.43852
               Q3-Q1     0.237495
               Mode      2.967333
                           Extremes
                 Lowest    Obs     Highest    Obs
               2.967333(      3) 3.128075(      2)
               3.128075(      2) 3.260018(      6)
               3.260018(      6) 3.307253(      5)
               3.307253(      5)  3.36557(      4)
                3.36557(      4) 3.405853(      1)
          Stem Leaf                     #             Boxplot
            34 1                        1                |
            33 7                        1             +------+
            33 1                        1             |      |
            32 6                        1             *------*
            32                                        |  +   |
            31                                        |      |
            31 3                        1             +------+
            30                                           |
            30                                           |
            29 7                        1                |
               ----+----+----+----+
          Multiply Stem.Leaf by 10**-1

                    Normal Probability Plot
      3.425+                                        +++*
           |                                    *+++
           |                                * +++
      3.275+                            *  +++
           |                             +++
           |                         +++
      3.125+                     ++*
           |                  +++
           |               +++
      2.975+            +++ *
            +----+----+----+----+----+----+----+----+----+----+
                 -2        -1         0        +1        +2
```

```
---------------------- LMIN=5.1929568509 ----------------------------
                    Univariate Procedure
Variable=LDIS
                           Moments
              N                 6  Sum Wgts           6
              Mean       3.721544  Sum         22.32926
              Std Dev    0.159182  Variance    0.025339
              Skewness   -0.94566  Kurtosis    0.754594
              USS        83.22603  CSS         0.126695
              CV         4.277311  Std Mean    0.064986
              T:Mean=0   57.26705  Pr>|T|        0.0001
              Num ^= 0          6  Num > 0            6
              M(Sign)           3  Pr>=|M|       0.0313
              Sgn Rank       10.5  Pr>=|S|       0.0313

                        Quantiles(Def=5)
               100% Max  3.883006          99%  3.883006
                75% Q3   3.866816          95%  3.883006
                50% Med  3.741806          90%  3.883006
                25% Q1   3.643621          10%  3.452207
                 0% Min  3.452207           5%  3.452207
                                            1%  3.452207
               Range     0.430799
               Q3-Q1     0.223196
               Mode      3.452207
                           Extremes
                 Lowest    Obs     Highest    Obs
               3.452207(      3) 3.643621(      2)
               3.643621(      2) 3.735286(      6)
               3.735286(      6) 3.748327(      5)
               3.748327(      5) 3.866816(      1)
               3.866816(      1) 3.883006(      4)

          Stem Leaf                     #             Boxplot
            38 78                       2             +------+
            38                                        |      |
            37 5                        1             |      |
            37 4                        1             *--+---*
            36                                        |      |
            36 4                        1             +------+
            35                                           |
            35                                           |
            34 5                        1                |
               ----+----+----+----+
          Multiply Stem.Leaf by 10**-1

                    Normal Probability Plot
      3.875+                                     *  +++ *
           |                                    ++++
           |                                 +++
           |                             ++++*
      3.675+                          +++
           |                       *++
           |                    +++
           |                 +++
      3.475+             ++++*
            +----+----+----+----+----+----+----+----+----+----+
                 -2        -1         0        +1        +2
```

---------- LNIN=5.4804389233 ----------

Univariate Procedure

Variable=LDIS

Moments

N	6	Sum Wgts	6
Mean	4.044053	Sum	24.26432
Std Dev	0.153979	Variance	0.02371
Skewness	-0.93621	Kurtosis	1.058765
USS	98.24475	CSS	0.118548
CV	3.80755	Std Mean	0.062862
T:Mean=0	64.33244	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	4.222884	99%	4.222884
75% Q3	4.149937	95%	4.222884
50% Med	4.069796	90%	4.222884
25% Q1	3.969537	10%	3.78237
0% Min	3.78237	5%	3.78237
		1%	3.78237
Range	0.440515		
Q3-Q1	0.1804		
Mode	3.78237		

Extremes

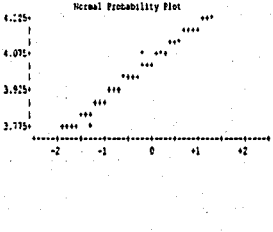
Lowest	Obs	Highest	Obs
3.78237(	3)	3.969537(	2)
3.969537(	2)	4.069721(	5)
4.06977(	5)	4.069881(	6)
4.069881(	6)	4.149937(	1)
4.149937(	1)	4.222884(	4)

Stem	Leaf	#	Boxplot
42	2	1	|
41	5	1	+-----+
41			| |
40	77	2	*-----*
40			| * |
39	7	1	+-----+
39			|
38			|
38			|
37	8	1	|

----+----+----+----+

Multiply Stem.Leaf by 10**-1

Normal Probability Plot



---------- LNIN=5.7037824747 ----------

Univariate Procedure

Variable=LDIS

Moments

N	6	Sum Wgts	6
Mean	4.267004	Sum	25.72202
Std Dev	0.135946	Variance	0.018565
Skewness	-1.49128	Kurtosis	2.752771
USS	110.3603	CSS	0.092824
CV	3.264415	Std Mean	0.055133
T:Mean=0	75.03611	Pr>|T|	0.0001
Num ^= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313

Quantiles(Def=5)

100% Max	4.427717	99%	4.427717
75% Q3	4.380651	95%	4.427717
50% Med	4.313961	90%	4.427717
25% Q1	4.257172	10%	4.028561
0% Min	4.028561	5%	4.028561
		1%	4.028561
Range	0.399156		
Q3-Q1	0.123479		
Mode	4.028561		

Extremes

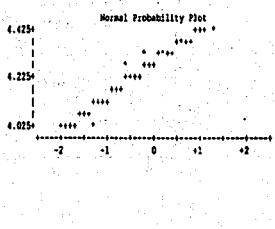
Lowest	Obs	Highest	Obs
4.028561(	3)	4.257172(	2)
4.257172(	2)	4.303524(	5)
4.303524(	5)	4.324398(	6)
4.324398(	6)	4.380651(	1)
4.380651(	1)	4.427717(	4)

Stem	Leaf	#	Boxplot
44	3	1	|
43	8	1	+-----+
43	02	2	*-----*
42	6	1	+--+--+
42			
41			
41			
40			
40	3	1	0

----+----+----+----+

Multiply Stem.Leaf by 10**-1

Normal Probability Plot



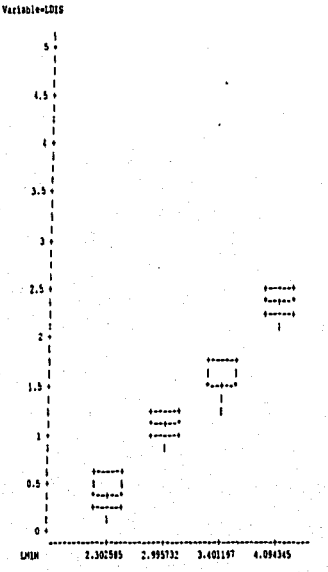
```
-------------------- LMIN=5.8861040315 --------------------------
                    Univariate Procedure
Variable=LDIS
                              Moments
                N              6  Sum Wgts          6
                Mean    4.452236  Sum        26.71341
                Std Dev 0.122511  Variance   0.015009
                Skewness -1.97711 Kurtosis   4.354008
                USS     119.0095  CSS        0.075045
                CV       2.75167  Std Mean   0.050015
                T:Mean=0 89.01829 Pr>|T|       0.0001
                Num ^= 0       6  Num > 0           6
                M(Sign)        3  Pr>=|M|      0.0313
                Sgn Rank    10.5  Pr>=|S|      0.0313

                         Quantiles(Def=5)
                 100% Max  4.557869        99%  4.557869
                  75% Q3   4.516667        95%  4.557869
                  50% Med  4.487139        90%  4.557869
                  25% Q1   4.451436        10%  4.213164
                   0% Min  4.213164         5%  4.213164
                                            1%  4.213164
                 Range     0.344705
                 Q3-Q1     0.065231
                 Mode      4.213164
                           Extremes
                  Lowest    Obs    Highest    Obs
                 4.213164(    3) 4.451436(     5)
                 4.451436(    5)   4.4787(     2)
                   4.4787(    2) 4.495579(     6)
                 4.495579(    6) 4.516667(     1)
                 4.516667(    1) 4.557869(     4)

            Stem Leaf                   #              Boxplot
              45 6                      1                 |
              45 02                     2              +-----+
              44 58                     2              *--+--*
              44
              43
              43
              42
              42 1                      1
                 ----+----+----+----+
             Multiply Stem.Leaf by 10**-1

                           Normal Probability Plot
          4.575+                              +++++*
               |                          ++*+
               |                  *  * *+*+
               |                     ++++
               |                ++++
               |            ++++
               |        ++++
          4.225+     ++++  *
                +----+----+----+----+----+----+----+----+----+----+
                    -2        -1         0        +1        +2
```

```
-------------------- LMIN=6.0402547113 --------------------------
                    Univariate Procedure
Variable=LDIS
                              Moments
                N              6  Sum Wgts          6
                Mean    4.531608  Sum        27.18965
                Std Dev  0.10677  Variance     0.0114
                Skewness -1.94188 Kurtosis   3.803987
                USS     123.2698  CSS        0.056999
                CV       2.35611  Std Mean   0.043589
                T:Mean=0 103.9633 Pr>|T|       0.0001
                Num ^= 0       6  Num > 0           6
                M(Sign)        3  Pr>=|M|      0.0313
                Sgn Rank    10.5  Pr>=|S|      0.0313

                         Quantiles(Def=5)
                 100% Max  4.610158        99%  4.610158
                  75% Q3   4.588024        95%  4.610158
                  50% Med  4.578842        90%  4.610158
                  25% Q1   4.508329        10%  4.325456
                   0% Min  4.325456         5%  4.325456
                                            1%  4.325456
                 Range     0.284701
                 Q3-Q1     0.079695
                 Mode      4.325456

                           Extremes
                  Lowest    Obs    Highest    Obs
                 4.325456(    3) 4.508329(     5)
                 4.508329(    5) 4.570372(     1)
                 4.570372(    1) 4.587312(     2)
                 4.587312(    2) 4.588024(     4)
                 4.588024(    4) 4.610158(     6)

            Stem Leaf                   #              Boxplot
              46 1                      1                 |
              45 799                    3              +-----+
              45 1                      1              *--+--*
              44
              44
              43
              43 3                      1                 0
                 ----+----+----+----+
             Multiply Stem.Leaf by 10**-1

                           Normal Probability Plot
          4.625+                                   +++++ *
               |                          *  *+++*
               |                      *  ++++
          4.475+                     +++++
               |               +++++
               |          +++++
          4.325+     ++++    *
                +----+----+----+----+----+----+----+----+----+----+
                    -2        -1         0        +1        +2
```

```
-------------------- LMIN=6.1737661029 --------------------------
                    Univariate Procedure
Variable=LDIS
                              Moments
                 N                  6  Sum Wgts          6
                 Mean        4.573564  Sum        27.44138
                 Std Dev      0.09142  Variance   0.008358
                 Skewness    -1.22458  Kurtosis   2.316103
                 USS         125.5467  CSS        0.041768
                 CV          1.998884  Std Mean   0.037322
                 T:Mean=0    122.5429  Pr>|T|       0.0001
                 Num ^= 0           6  Num > 0           6
                 M(Sign)            3  Pr>=|M|      0.0313
                 Sgn Rank        10.5  Pr>=|S|      0.0313

                        Quantiles(Def=5)
                  100% Max  4.679907        99%  4.679907
                   75% Q3   4.617988        95%  4.679907
                   50% Med   4.59416        90%  4.679907
                   25% Q1   4.545526        10%  4.409642
                    0% Min  4.409642         5%  4.409642
                                             1%  4.409642
                  Range     0.270265
                  Q3-Q1     0.072463
                  Mode      4.409642

                              Extremes
                     Lowest     Obs     Highest     Obs
                   4.409642(      3) 4.545526(       5)
                   4.545526(      5) 4.593401(       4)
                   4.593401(      4) 4.594918(       1)
                   4.594918(      1) 4.617988(       6)
                   4.617988(      6) 4.679907(       2)

              Stem Leaf                   #           Boxplot
                46 8                      1              |
                46 2                      1           +-----+
                45 599                    3           +-----+
                45
                44
                44 1                      1              0
                   ----+----+----+----+
              Multiply Stem.Leaf by 10**-1

Variable=LDIS

                              Normal Probability Plot
             4.675+                                        ++++*
                  |                                   +++*++
                  |                              *++++
                  |                         ++*+++
                  |                    +++++
             4.425+               +++++*
                   +----+----+----+----+----+----+----+----+----+----+
                       -2        -1         0        +1        +2
```

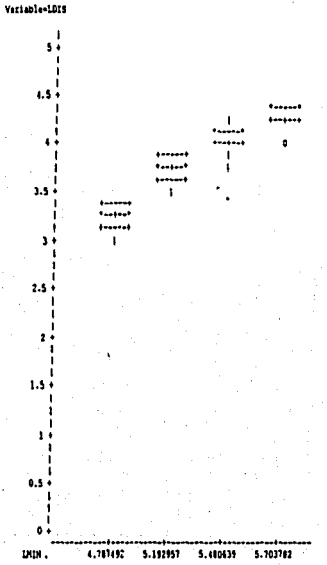
OBSERVACIONES DE DISOLUCION
'LOTE B'

Univariate Procedure
Schematic Plots



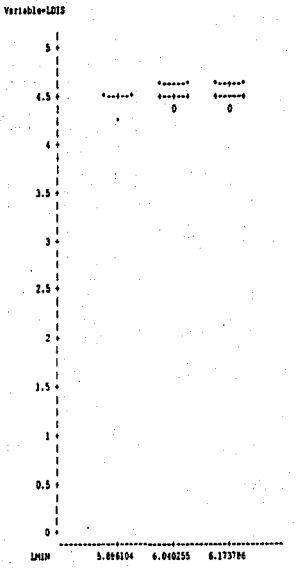
OBSERVACIONES DE DISOLUCION
'LOTE B'

Univariate Procedure
Schematic Plots



OBSERVACIONES DE DISOLUCION
'LOTE 6'

Univariate Procedure
Schematic Plots



ANEXO B

(Resultados del programa 6.6.)

Salida 6.6. Intervalos de confianza para tres modelos.

ANALISIS DE REGRESION SIMPLE
'LOTE 8'

Model: MODEL1
NOTE: No intercept in model. R-square is redefined.
Dependent Variable: DISOLU

Obs	Dep Var DISOLU	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
1	1.7900	2.2322	0.036	2.1612	2.3033	-12.6915	17.1559
2	3.6600	4.4645	0.071	4.3224	4.6065	-10.4597	19.3887
3	6.0000	6.6967	0.107	6.4836	6.9098	-8.2283	21.6217
4	12.6000	13.3934	0.213	12.9673	13.8196	-1.5362	28.3230
5	30.1400	26.7868	0.427	25.9345	27.6391	11.8390	41.7346
6	47.7900	40.1802	0.640	38.9018	41.4587	25.2020	55.1584
7	63.4300	53.5736	0.854	51.8690	55.2782	38.5531	68.5942
8	79.8900	66.9670	1.067	64.8363	69.0978	51.8922	82.0419
9	91.5300	80.3604	1.280	77.8035	82.9174	65.2194	95.5014
10	96.5800	93.7538	1.494	90.7708	96.7369	78.5351	109.0
11	98.9800	107.1	1.707	103.7	110.6	91.8393	122.5
12	1.4000	2.2322	0.036	2.1612	2.3033	-12.6915	17.1559
13	2.9700	4.4645	0.071	4.3224	4.6065	-10.4597	19.3887
14	4.3200	6.6967	0.107	6.4836	6.9098	-8.2283	21.6217
15	9.7000	13.3934	0.213	12.9673	13.8196	-1.5362	28.3230
16	22.8300	26.7868	0.427	25.9345	27.6391	11.8390	41.7346
17	38.2300	40.1802	0.640	38.9018	41.4587	25.2020	55.1584
18	52.9600	53.5736	0.854	51.8690	55.2782	38.5531	68.5942
19	70.6100	66.9670	1.067	64.8363	69.0978	51.8922	82.0419
20	88.1200	80.3604	1.280	77.8035	82.9174	65.2194	95.5014
21	98.2300	93.7538	1.494	90.7708	96.7369	78.5351	109.0
22	107.8	107.1	1.707	103.7	110.6	91.8393	122.5
23	1.2000	2.2322	0.036	2.1612	2.3033	-12.6915	17.1559
24	2.4500	4.4645	0.071	4.3224	4.6065	-10.4597	19.3887
25	3.7000	6.6967	0.107	6.4836	6.9098	-8.2283	21.6217
26	8.2200	13.3934	0.213	12.9673	13.8196	-1.5362	28.3230
27	19.4400	26.7868	0.427	25.9345	27.6391	11.8390	41.7346
28	31.5700	40.1802	0.640	38.9018	41.4587	25.2020	55.1584
29	43.9200	53.5736	0.854	51.8690	55.2782	38.5531	68.5942
30	56.1800	66.9670	1.067	64.8363	69.0978	51.8922	82.0419
31	67.5700	80.3604	1.280	77.8035	82.9174	65.2194	95.5014
32	75.6000	93.7538	1.494	90.7708	96.7369	78.5351	109.0
33	82.2400	107.1	1.707	103.7	110.6	91.8393	122.5
34	1.8700	2.2322	0.036	2.1612	2.3033	-12.6915	17.1559
35	3.6600	4.4645	0.071	4.3224	4.6065	-10.4597	19.3887
36	5.6100	6.6967	0.107	6.4836	6.9098	-8.2283	21.6217
37	12.5800	13.3934	0.213	12.9673	13.8196	-1.5362	28.3230
38	28.9500	26.7868	0.427	25.9345	27.6391	11.8390	41.7346
39	48.5700	40.1802	0.640	38.9018	41.4587	25.2020	55.1584
40	68.2300	53.5736	0.854	51.8690	55.2782	38.5531	68.5942
41	83.7400	66.9670	1.067	64.8363	69.0978	51.8922	82.0419
42	95.3800	80.3604	1.280	77.8035	82.9174	65.2194	95.5014
43	98.3000	93.7538	1.494	90.7708	96.7369	78.5351	109.0
44	98.8300	107.1	1.707	103.7	110.6	91.8393	122.5
45	1.3500	2.2322	0.036	2.1612	2.3033	-12.6915	17.1559
46	2.7500	4.4645	0.071	4.3224	4.6065	-10.4597	19.3887
47	4.5700	6.6967	0.107	6.4836	6.9098	-8.2283	21.6217
48	11.9800	13.3934	0.213	12.9673	13.8196	-1.5362	28.3230
49	27.3100	26.7868	0.427	25.9345	27.6391	11.8390	41.7346
50	42.4500	40.1802	0.640	38.9018	41.4587	25.2020	55.1584

Obs	Dep Var DISOLU	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
51	58.5400	53.5736	0.854	51.8690	55.2782	38.5531	68.5942
52	73.9600	66.9670	1.067	64.8363	69.0978	51.8922	82.0419
53	85.7500	80.3604	1.280	77.8035	82.9174	65.2194	95.5014
54	90.7700	93.7538	1.494	90.7708	96.7369	78.5351	109.0
55	94.2100	107.1	1.707	103.7	110.6	91.8393	122.5
56	1.4800	2.2322	0.036	2.1612	2.3033	-12.6915	17.1559
57	2.8800	4.4645	0.071	4.3224	4.6065	-10.4597	19.3887
58	4.5000	6.6967	0.107	6.4836	6.9098	-8.2283	21.6217
59	10.5600	13.3934	0.213	12.9673	13.8196	-1.5362	28.3230
60	26.0500	26.7868	0.427	25.9345	27.6391	11.8390	41.7346
61	41.9000	40.1802	0.640	38.9018	41.4587	25.2020	55.1584
62	58.5500	53.5736	0.854	51.8690	55.2782	38.5531	68.5942
63	75.5200	66.9670	1.067	64.8363	69.0978	51.8922	82.0419
64	89.6200	80.3604	1.280	77.8035	82.9174	65.2194	95.5014
65	100.5	93.7538	1.494	90.7708	96.7369	78.5351	109.0
66	101.3	107.1	1.707	103.7	110.6	91.8393	122.5

Obs	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
1	-0.4422	7.472	-0.059	| | | |	0.000
2	-0.8045	7.472	-0.108	| | | |	0.000
3	-0.6967	7.472	-0.093	| | | |	0.000
4	-0.7934	7.469	-0.106	| | | |	0.000
5	3.3532	7.460	0.449	| | | |	0.001
6	7.6098	7.445	1.022	| |** |	0.008
7	9.8564	7.424	1.328	| |** |	0.023
8	12.9230	7.396	1.747	| |*** |	0.064
9	11.1696	7.362	1.517	| |*** |	0.070
10	2.8262	7.322	0.386	| | | |	0.006
11	-8.1672	7.275	-1.123	| **| |	0.069
12	-0.8322	7.472	-0.111	| | | |	0.000
13	-1.4945	7.472	-0.200	| | | |	0.000
14	-2.3767	7.472	-0.318	| | | |	0.000
15	-3.6934	7.469	-0.494	| | | |	0.000
16	-3.9568	7.460	-0.530	| * | |	0.001
17	-1.9502	7.445	-0.262	| | | |	0.001
18	-0.6136	7.424	-0.083	| | | |	0.000
19	3.6430	7.396	0.493	| | | |	0.003
20	7.7596	7.362	1.054	| |** |	0.034
21	4.4762	7.322	0.611	| |* |	0.016
22	0.6128	7.275	0.084	| | | |	0.000
23	-1.0322	7.472	-0.138	| | | |	0.000
24	-2.0145	7.472	-0.270	| | | |	0.000
25	-2.9967	7.472	-0.401	| | | |	0.000
26	-5.1734	7.469	-0.693	| * | |	0.003
27	-7.3468	7.460	-0.985	| * | |	0.010
28	-8.6102	7.445	-1.157	| ** | |	0.022
29	-9.6536	7.424	-1.300	| ** | |	0.044
30	-10.7870	7.396	-1.459	| ** | |	0.091
31	-12.7904	7.362	-1.737	| *** | |	0.156
32	-18.1538	7.322	-2.479	| **** | |	0.645
33	-24.9072	7.275	-3.424	| ****** | |	0.000
34	-0.3622	7.472	-0.048	| | | |	0.000
35	-0.8045	7.472	-0.108	| | | |	0.000
36	-1.0867	7.472	-0.145	| | | |	0.000
37	-0.8134	7.469	-0.109	| | | |	0.000
38	2.1632	7.460	0.290	| | | |	0.000
39	8.3898	7.445	1.127	| |** |	0.009
40	14.6564	7.424	1.974	| |*** |	0.052
41	16.7730	7.396	2.268	| |**** |	0.107
42	15.0196	7.362	2.040	| |**** |	0.126

Obs	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
43	4.5462	7.322	0.621	| |	0.016
44	-8.3172	7.275	-1.143	** | |	0.072
45	-0.8822	7.472	-0.118	| |	0.000
46	-1.7145	7.472	-0.229	| |	0.000
47	-2.1267	7.472	-0.285	| |	0.000
48	-1.4134	7.469	-0.189	| |	0.000
49	0.5232	7.460	0.070	| |	0.000
50	2.2698	7.445	0.305	| |	0.001
51	4.9664	7.424	0.669	| * |	0.006
52	6.9930	7.396	0.946	| * |	0.019
53	5.3896	7.362	0.732	| * |	0.016
54	-2.9838	7.322	-0.408	| |	0.007
55	-12.9372	7.275	-1.778	*** | |	0.174
56	-0.7522	7.472	-0.101	| |	0.000
57	-1.5845	7.472	-0.212	| |	0.000
58	-2.1967	7.472	-0.294	| |	0.000
59	-2.8334	7.469	-0.379	| |	0.000
60	-0.7368	7.460	-0.099	| |	0.000
61	1.7198	7.445	0.231	| |	0.000
62	4.9764	7.424	0.670	| * |	0.006
63	8.5530	7.396	1.156	| ** |	0.028
64	9.2596	7.362	1.258	| ** |	0.048
65	6.7462	7.322	0.921	| * |	0.035
66	-5.8572	7.275	-0.805	| * |	0.036

Sum of Residuals	0.48421
Sum of Squared Residuals	3629.4393
Predicted Resid SS (Press)	3862.3341

ANALISIS DE REGRESION SIMPLE
'LOTE 8'

Model: MODEL2
NOTE: No intercept in model. R-square is redefined.
Dependent Variable: DISOLU

Obs	Weight	Dep Var DISOLU	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict
1	0	1.7900	2.2327	0.039	2.1539	2.3116	.
2	0	3.6600	4.4654	0.079	4.3077	4.6231	.
3	1.0000	6.0000	6.6981	0.118	6.4616	6.9347	-9.8639
4	1.0000	12.6000	13.3963	0.236	12.9232	13.8693	-3.1708
5	1.0000	30.1400	26.7926	0.472	25.8464	27.7387	10.2052
6	1.0000	47.7900	40.1888	0.708	38.7697	41.6080	23.5678
7	1.0000	63.4300	53.5851	0.943	51.6929	55.4773	36.9170
8	1.0000	79.8900	66.9814	1.179	64.6161	69.3467	50.2530
9	1.0000	91.5300	80.3777	1.415	77.5393	83.2160	63.5758
10	1.0000	96.5800	93.7740	1.651	90.4626	97.0854	76.8858
11	1.0000	98.9800	107.2	1.887	103.4	111.0	90.1830
12	0	1.4000	2.2327	0.039	2.1539	2.3116	.
13	0	2.9700	4.4654	0.079	4.3077	4.6231	.
14	1.0000	4.3200	6.6981	0.118	6.4616	6.9347	-9.8639
15	1.0000	9.7000	13.3963	0.236	12.9232	13.8693	-3.1708
16	1.0000	22.8300	26.7926	0.472	25.8464	27.7387	10.2052
17	1.0000	38.2300	40.1888	0.708	38.7697	41.6080	23.5678
18	1.0000	52.9600	53.5851	0.943	51.6929	55.4773	36.9170
19	1.0000	70.6100	66.9814	1.179	64.6161	69.3467	50.2530
20	1.0000	88.1200	80.3777	1.415	77.5393	83.2160	63.5758

Obs	Weight	Dep Var DISOLU	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict
21	1.0000	98.2300	93.7740	1.651	90.4626	97.0854	76.8858
22	1.0000	107.8	107.2	1.887	103.4	111.0	90.1830
23	0	1.2000	2.2327	0.039	2.1539	2.3116	.
24	0	2.4500	4.4654	0.079	4.3077	4.6231	.
25	1.0000	3.7000	6.6981	0.118	6.4616	6.9347	-9.8639
26	1.0000	8.2200	13.3963	0.236	12.9232	13.8693	-3.1708
27	1.0000	19.4400	26.7926	0.472	25.8464	27.7387	10.2052
28	1.0000	31.5700	40.1888	0.708	38.7697	41.6080	23.5678
29	1.0000	43.9200	53.5851	0.943	51.6929	55.4773	36.9170
30	1.0000	56.1800	66.9814	1.179	64.6161	69.3467	50.2530
31	1.0000	67.5700	80.3777	1.415	77.5393	83.2160	63.5758
32	1.0000	75.6000	93.7740	1.651	90.4626	97.0854	76.8858
33	1.0000	82.2400	107.2	1.887	103.4	111.0	90.1830
34	0	1.8700	2.2327	0.039	2.1539	2.3116	.
35	0	3.6600	4.4654	0.079	4.3077	4.6231	.
36	1.0000	5.6100	6.6981	0.118	6.4616	6.9347	-9.8639
37	1.0000	12.5800	13.3963	0.236	12.9232	13.8693	-3.1708
38	1.0000	28.9500	26.7926	0.472	25.8464	27.7387	10.2052
39	1.0000	42.5700	40.1888	0.708	38.7697	41.6080	23.5678
40	1.0000	68.2300	53.5851	0.943	51.6929	55.4773	36.9170
41	1.0000	83.7400	66.9814	1.179	64.6161	69.3467	50.2530
42	1.0000	95.3800	80.3777	1.415	77.5393	83.2160	63.5758
43	1.0000	98.3000	93.7740	1.651	90.4626	97.0854	76.8858
44	1.0000	98.8300	107.2	1.887	103.4	111.0	90.1830
45	0	1.3500	2.2327	0.039	2.1539	2.3116	.
46	0	2.7500	4.4654	0.079	4.3077	4.6231	.
47	1.0000	4.5700	6.6981	0.118	6.4616	6.9347	-9.8639
48	1.0000	11.9800	13.3963	0.236	12.9232	13.8693	-3.1708
49	1.0000	27.3100	26.7926	0.472	25.8464	27.7387	10.2052
50	1.0000	42.4500	40.1888	0.708	38.7697	41.6080	23.5678
51	1.0000	58.5400	53.5851	0.943	51.6929	55.4773	36.9170
52	1.0000	73.9600	66.9814	1.179	64.6161	69.3467	50.2530
53	1.0000	85.7500	80.3777	1.415	77.5393	83.2160	63.5758
54	1.0000	90.7700	93.7740	1.651	90.4626	97.0854	76.8858
55	1.0000	94.2500	107.2	1.887	103.4	111.0	90.1830
56	0	1.4800	2.2327	0.039	2.1539	2.3116	.
57	0	2.8800	4.4654	0.079	4.3077	4.6231	.
58	1.0000	4.5000	6.6981	0.118	6.4616	6.9347	-9.8639
59	1.0000	10.5600	13.3963	0.236	12.9232	13.8693	-3.1708
60	1.0000	26.0500	26.7926	0.472	25.8464	27.7387	10.2052
61	1.0000	41.9000	40.1888	0.708	38.7697	41.6080	23.5678
62	1.0000	58.5500	53.5851	0.943	51.6929	55.4773	36.9170
63	1.0000	75.5200	66.9814	1.179	64.6161	69.3467	50.2530
64	1.0000	89.6200	80.3777	1.415	77.5393	83.2160	63.5758
65	1.0000	100.5	93.7740	1.651	90.4626	97.0854	76.8858
66	1.0000	101.3	107.2	1.887	103.4	111.0	90.1830

Obs	Upper95% Predict	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
1	.	-0.4427	.	.		.
2	.	-0.8054	.	.		.
3	23.2602	-0.6981	8.256	-0.085	| | |	0.000
4	29.9634	-0.7963	8.253	-0.096	| | |	0.000
5	43.3799	3.3474	8.243	0.406	| | |	0.001
6	56.8099	7.6012	8.226	0.924	| * |	0.006
7	70.2532	9.8449	8.202	1.200	| ** |	0.019
8	83.7098	12.9086	8.172	1.580	| *** |	0.052
9	97.1795	11.1523	8.134	1.371	| ** |	0.057
10	110.7	2.8060	8.090	0.347	| | |	0.005
11	124.2	-8.1902	8.038	-1.019	| ** |	0.057
12	.	-0.8327	.	.		.

Obs	Upper95% Predict	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
13	.	-1.4954	.	.		.
14	23.2602	-2.3781	8.256	-0.288		0.000
15	29.9634	-3.6963	8.253	-0.448		0.000
16	43.3799	-3.9626	8.243	-0.481		0.001
17	56.8099	-1.9588	8.226	-0.238		0.000
18	70.2532	-0.6251	8.202	-0.076		0.000
19	83.7098	3.6286	8.172	0.444		0.004
20	97.1795	7.7423	8.134	0.952	|*	0.027
21	110.7	4.4560	8.090	0.551	|*	0.013
22	124.2	0.5898	8.038	0.073		0.000
23	.	-1.0327	.	.		.
24	.	-2.0154	.	.		.
25	23.2602	-2.9981	8.256	-0.363		0.000
26	29.9634	-5.1763	8.253	-0.627	*|	0.000
27	43.3799	-7.3526	8.243	-0.892	*|	0.003
28	56.8099	-8.6188	8.226	-1.048	**|	0.008
29	70.2532	-9.6651	8.202	-1.178	**|	0.018
30	83.7098	-10.8014	8.172	-1.322	**|	0.036
31	97.1795	-12.8077	8.134	-1.575	***|	0.075
32	110.7	-18.1740	8.090	-2.247	****|	0.210
33	124.2	-24.9302	8.038	-3.102	*******|	0.530
34	.	-0.3627	.	.		.
35	.	-0.8054	.	.		.
36	23.2602	-1.0881	8.256	-0.132		0.000
37	29.9634	-0.8163	8.253	-0.099		0.000
38	43.3799	2.1574	8.243	0.262	|*	0.000
39	56.8099	8.3812	8.226	1.019	|**	0.008
40	70.2532	14.6449	8.202	1.785	|***	0.042
41	83.7098	16.7586	8.172	2.051	|****	0.088
42	97.1795	15.0023	8.134	1.844	|***	0.103
43	110.7	4.5260	8.090	0.559	|*	0.013
44	124.2	-8.3402	8.038	-1.038	**|	0.059
45	.	-0.8827	.	.		.
46	.	-1.7154	.	.		.
47	23.2602	-2.1281	8.256	-0.258		0.000
48	29.9634	-1.4163	8.253	-0.172		0.000
49	43.3799	0.5174	8.243	0.063		0.000
50	56.8099	2.2612	8.226	0.275	|	0.001
51	70.2532	4.9549	8.202	0.604	|*	0.005
52	83.7098	6.9786	8.172	0.854	|*	0.015
53	97.1795	5.3723	8.134	0.660	|*	0.013
54	110.7	-3.0040	8.090	-0.371	|	0.006
55	124.2	-12.9602	8.038	-1.612	***|	0.143
56	.	-0.7527	.	.		.
57	.	-1.5854	.	.		.
58	23.2602	-2.1981	8.256	-0.266		0.000
59	29.9634	-2.8363	8.253	-0.344		0.000
60	43.3799	-0.7426	8.243	-0.090		0.000
61	56.8099	1.7112	8.226	0.208	|	0.000
62	70.2532	4.9649	8.202	0.605	|*	0.005
63	83.7098	8.5386	8.172	1.045	|**	0.023
64	97.1795	9.2423	8.134	1.136	|**	0.039
65	110.7	6.7260	8.090	0.831	|*	0.029
66	124.2	-5.8802	8.038	-0.732	*|	0.029

Sum of Residuals 12.57488
Sum of Squared Residuals 3612.9642
Predicted Resid SS (Press) 3846.2059
NOTE: The above statistics use observation weights or frequencies.

ANALISIS DE REGRESION
'LOTE 8'

Model: MODEL3
Dependent Variable: LDIS

Obs	Dep Var LDIS	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
1	0.5822	0.3753	0.042	0.2914	0.4593	0.0471	0.7035
2	1.2975	1.1575	0.033	1.0916	1.2235	0.8335	1.4816
3	1.7918	1.6151	0.028	1.5587	1.6715	1.2928	1.9373
4	2.5137	2.3973	0.022	2.3536	2.4409	2.0770	2.7175
5	3.4059	3.1795	0.020	3.1404	3.2186	2.8598	3.4991
6	3.8668	3.6370	0.021	3.5955	3.6785	3.3170	3.9570
7	4.1499	3.9617	0.023	3.9164	4.0069	3.6412	4.2821
8	4.3807	4.2135	0.025	4.1644	4.2625	3.8924	4.5345
9	4.5167	4.4192	0.026	4.3666	4.4718	4.0976	4.7408
10	4.5704	4.5932	0.028	4.5373	4.6491	4.2710	4.9153
11	4.5949	4.7439	0.029	4.6850	4.8027	4.4212	5.0666
12	0.3365	0.3753	0.042	0.2914	0.4593	0.0471	0.7035
13	1.0886	1.1575	0.033	1.0916	1.2235	0.8335	1.4816
14	1.4633	1.6151	0.028	1.5587	1.6715	1.2928	1.9373
15	2.2721	2.3973	0.022	2.3536	2.4409	2.0770	2.7175
16	3.1281	3.1795	0.020	3.1404	3.2186	2.8598	3.4991
17	3.6436	3.6370	0.021	3.5955	3.6785	3.3170	3.9570
18	3.9695	3.9617	0.023	3.9164	4.0069	3.6412	4.2821
19	4.2572	4.2135	0.025	4.1644	4.2625	3.8924	4.5345
20	4.4787	4.4192	0.026	4.3666	4.4718	4.0976	4.7408
21	4.5873	4.5932	0.028	4.5373	4.6491	4.2710	4.9153
22	4.6799	4.7439	0.029	4.6850	4.8027	4.4212	5.0666
23	0.1823	0.3753	0.042	0.2914	0.4593	0.0471	0.7035
24	0.8961	1.1575	0.033	1.0916	1.2235	0.8335	1.4816
25	1.3083	1.6151	0.028	1.5587	1.6715	1.2928	1.9373
26	2.1066	2.3973	0.022	2.3536	2.4409	2.0770	2.7175
27	2.9673	3.1795	0.020	3.1404	3.2186	2.8598	3.4991
28	3.4522	3.6370	0.021	3.5955	3.6785	3.3170	3.9570
29	3.7824	3.9617	0.023	3.9164	4.0069	3.6412	4.2821
30	4.0286	4.2135	0.025	4.1644	4.2625	3.8924	4.5345
31	4.2132	4.4192	0.026	4.3666	4.4718	4.0976	4.7408
32	4.3255	4.5932	0.028	4.5373	4.6491	4.2710	4.9153
33	4.4096	4.7439	0.029	4.6850	4.8027	4.4212	5.0666
34	0.6259	0.3753	0.042	0.2914	0.4593	0.0471	0.7035
35	1.2975	1.1575	0.033	1.0916	1.2235	0.8335	1.4816
36	1.7246	1.6151	0.028	1.5587	1.6715	1.2928	1.9373
37	2.5321	2.3973	0.022	2.3536	2.4409	2.0770	2.7175
38	3.3656	3.1795	0.020	3.1404	3.2186	2.8598	3.4991
39	3.8830	3.6370	0.021	3.5955	3.6785	3.3170	3.9570
40	4.2229	3.9617	0.023	3.9164	4.0069	3.6412	4.2821
41	4.4277	4.2135	0.025	4.1644	4.2625	3.8924	4.5345
42	4.5579	4.4192	0.026	4.3666	4.4718	4.0976	4.7408
43	4.5880	4.5932	0.028	4.5373	4.6491	4.2710	4.9153
44	4.5934	4.7439	0.029	4.6850	4.8027	4.4212	5.0666
45	0.3001	0.3753	0.042	0.2914	0.4593	0.0471	0.7035
46	1.0116	1.1575	0.033	1.0916	1.2235	0.8335	1.4816
47	1.5195	1.6151	0.028	1.5587	1.6715	1.2928	1.9373
48	2.4832	2.3973	0.022	2.3536	2.4409	2.0770	2.7175
49	3.3073	3.1795	0.020	3.1404	3.2186	2.8598	3.4991
50	3.7483	3.6370	0.021	3.5955	3.6785	3.3170	3.9570
51	4.0697	3.9617	0.023	3.9164	4.0069	3.6412	4.2821
52	4.3035	4.2135	0.025	4.1644	4.2625	3.8924	4.5345
53	4.4514	4.4192	0.026	4.3666	4.4718	4.0976	4.7408
54	4.5083	4.5932	0.028	4.5373	4.6491	4.2710	4.9153
55	4.5455	4.7439	0.029	4.6850	4.8027	4.4212	5.0666

Obs	Dep Var LDIS	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
56	0.3920	0.3753	0.042	0.2914	0.4593	0.0471	0.7035
57	1.0578	1.1575	0.033	1.0916	1.2235	0.8335	1.4816
58	1.5041	1.6151	0.028	1.5587	1.6715	1.2928	1.9373
59	2.3571	2.3973	0.022	2.3536	2.4409	2.0770	2.7175
60	3.2600	3.1795	0.020	3.1404	3.2186	2.8598	3.4991
61	3.7353	3.6370	0.021	3.5955	3.6785	3.3170	3.9570
62	4.0699	3.9617	0.023	3.9164	4.0069	3.6412	4.2821
63	4.3244	4.2135	0.025	4.1644	4.2625	3.8924	4.5345
64	4.4956	4.4192	0.026	4.3666	4.4718	4.0976	4.7408
65	4.6102	4.5932	0.028	4.5373	4.6491	4.2710	4.9153
66	4.6180	4.7439	0.029	4.6850	4.8027	4.4212	5.0666

Obs	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
1	0.2069	0.163	1.351	| |** |	0.069
2	0.1399	0.155	0.901	| |* |	0.018
3	0.1767	0.156	1.131	| |** |	0.021
4	0.1364	0.157	0.867	| |* |	0.007
5	0.2264	0.158	1.436	| |** |	0.016
6	0.2298	0.157	1.459	| |** |	0.019
7	0.1883	0.157	1.198	| |** |	0.015
8	0.1672	0.157	1.065	| |** |	0.014
9	0.0974	0.157	0.622	| |* |	0.005
10	-0.0228	0.156	-0.146	| | |	0.000
11	-0.1489	0.156	-0.954	|* | |	0.016
12	-0.0389	0.153	-0.254	| | |	0.002
13	-0.0690	0.155	-0.444	| | |	0.004
14	-0.1518	0.156	-0.971	|* | |	0.015
15	-0.1251	0.157	-0.796	|* | |	0.006
16	-0.0514	0.158	-0.326	| | |	0.001
17	0.00660	0.157	0.042	| | |	0.000
18	0.00787	0.157	0.050	| | |	0.000
19	0.0437	0.157	0.278	| | |	0.000
20	0.0595	0.157	0.380	| | |	0.001
21	-0.00586	0.156	-0.038	| | |	0.000
22	-0.0640	0.156	-0.410	| | |	0.000
23	-0.1930	0.153	-1.260	|** | |	0.003
24	-0.2614	0.155	-1.683	|*** | |	0.060
25	-0.3067	0.156	-1.963	|*** | |	0.064
26	-0.2907	0.157	-1.848	|*** | |	0.063
27	-0.2121	0.158	-1.346	|** | |	0.033
28	-0.1848	0.157	-1.174	|** | |	0.014
29	-0.1793	0.157	-1.141	|** | |	0.012
30	-0.1849	0.157	-1.179	|** | |	0.014
31	-0.2061	0.157	-1.316	|** | |	0.017
32	-0.2677	0.156	-1.713	|*** | |	0.044
33	-0.3342	0.156	-2.142	|**** | |	0.047
34	0.2506	0.153	1.636	| |*** |	0.082
35	0.1399	0.155	0.901	| |* |	0.101
36	0.1095	0.156	0.700	| |* |	0.018
37	0.1348	0.157	0.857	| |* |	0.008
38	0.1861	0.158	1.181	| |** |	0.007
39	0.2460	0.157	1.562	| |*** |	0.011
40	0.2612	0.157	1.662	| |*** |	0.021
41	0.2142	0.157	1.365	| |** |	0.029
42	0.1386	0.157	0.885	| |* |	0.023
43	-0.00515	0.156	-0.033	| | |	0.011
44	-0.1505	0.156	-0.964	|* | |	0.000
45	-0.0752	0.153	-0.491	| | |	0.017
46	-0.1459	0.155	-0.939	|* | |	0.009
					0.020

Obs	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
47	-0.0956	0.156	-0.611	*|	0.006
48	0.0860	0.157	0.546	|*	0.003
49	0.1278	0.158	0.811	|*	0.005
50	0.1113	0.157	0.707	|*	0.004
51	0.1080	0.157	0.687	|*	0.005
52	0.0900	0.157	0.574	|*	0.004
53	0.0322	0.157	0.206	|	0.001
54	-0.0848	0.156	-0.543	*|	0.005
55	-0.1983	0.156	-1.271	*|	0.029
56	0.0167	0.153	0.109	|	0.000
57	-0.0997	0.155	-0.642	*|	0.009
58	-0.1110	0.156	-0.710	*|	0.008
59	-0.0402	0.157	-0.256	|	0.001
60	0.0805	0.158	0.511	|*	0.002
61	0.0983	0.157	0.624	|*	0.003
62	0.1082	0.157	0.688	|*	0.005
63	0.1109	0.157	0.707	|*	0.006
64	0.0764	0.157	0.488	|	0.003
65	0.0170	0.156	0.109	|	0.000
66	-0.1259	0.156	-0.807	*|	0.012

Sum of Residuals	0
Sum of Squared Residuals	1.6142
Predicted Resid SS (Press)	1.7182