03043 4 29.

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESPECIALIZACION EN ESTADISTICA APLICADA
DE LA U A C P y P DEL C C H
CON SEDE EN EL I I M A S

USO DE LAS VARIABLES INDICADORAS "DUMMY" EN LOS MODELOS DE REGRESION

TESTIMONAL DE LA CONFERENCIA

QUE PARA OBTENER EL DIPLOMA DE LA

ESPECIALIZACION EN ESTADISTICA APLICADA

P R E S E N T A

GERARDO JESUS VARELA HERNANDEZ

BAJO LA DIRECCION DEL

M. EN C. INOCENCIO RAFAEL MADRID RIOS

MEXICO, D. F. 1998

1998

TESIS CON
FALLA DE ORIGEN

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

DERECHOS RESERVADOS © PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

A mis padres:

Lydia Hernández

Ignacio Varela

A mis suegros:

Angelina Roa

Manuel Ramírez

A mi compañera de toda la vida:

Angélica Ramírez

A mi hijo por nacer

A mis "nenes"

Junior
Monchito
Chiquito
Angelito

A mis hermanos, cuñados y sobrinos.

A mis maestros, compañeros y alumnos

A mi Universidad:

la U N A M

AGRADECIMIENTOS

Agradezco sinceramente al M. en C. Inocencio Rafael Madrid Ríos el haberme permitido desarrollar este trabajo bajo su tutela, así como todo su apoyo académico y personal.

A los miembros del jurado, Dr. Ignacio Méndez Ramírez, Dr. Raúl Rueda Díaz del Campo, Dr. Víctor Fenton Navarro y M. en C. Martín Romero Martínez, por sus interesantes y atinadas observaciones y correcciones en la revisión del presente trabajo.

A todos mis maestros de la Especialización en Estadística Aplicada, por los conocimientos y experiencias compartidas en clase y extraclase.

A los coordinadores, en su tiempo, del Posgrado en Estadística, Dr. Federico O'Reilly y M. en C. Rafael Madrid, por permitirme participar en el posgrado desde otro punto de vista, y muy especialmente a la Dra. Guillermina Eslava por brindarme la oportunidad de ser su ayudante y así iniciar esa enriquecedora experiencia del "otro punto de vista".

A todo el personal de la UACPyP y del IIMAS con los que convivi durante este tiempo y que me permitieron llegar hasta donde estoy.

A mis maestros, compañeros alumnos, compañeros profesores y alumnos por los conocimientos, experiencias, necesidades e inquietudes compartidas, que tanto han contribuido a mi desarrollo académico y profesional, incluyendo este nuevo paso.

A todos mis familiares y amigos, por el apoyo brindado desde mi ingreso a la especialización.

PREFACIO

El presente trabajo es el testimonial de la conferencia que será dictada para obtener el diploma de la Especialización en Estadística Aplicada.

Ya que esta modalidad de examen consistente en una exposición y la respuesta a las preguntas que sobre la misma surjan, es menester elaborar un testimonio escrito del tema que se abordó durante la misma.

Dada la imposibilidad de transcribir la conferencia en el momento mismo de su presentación, y de la necesidad de evaluar el desarrollo del tema previo a su exposición, el testimonial es sólo una constancia escrita del desarrollo del trabajo que de manera resumida constituye el material de la conferencia.

Puesto que la presentación se apoyará en acetatos, se incluyen los mismos al final del escrito. Sin embargo, se pretendió ilustrar la mayoría de ellos dentro del mismo texto, aun cuando algunos comentarios que pudieran hacerse durante la exposición no necesariamente estarán incluidos en el escrito.

Con relación a la temática, se refiere a conceptos básicos de regresión con variables indicadoras y de su equivalencia con los modelos de diseño de experimentos, procurando que su tratamiento fuese mas bien intuitivo y con un mínimo de álgebra.

La razón de enfocarse a aspectos básicos es la falta de entendimiento de los mismos por gente de áreas diferentes a la matemática, lo que dificulta el aprendizaje y aplicación de aspectos más profundos. A todo lo anterior hay que añadir el poco agrado que los alumnos de cualquier nivel profesan por las matemáticas.

Dentro de la misma Especialización en Estadística Aplicada, vi cómo algunos conceptos no nos quedaban lo suficientemente claros, a pesar de que se comentaran en más de una materia. No es de extrañar esta situación, pues al no ser estos conceptos el tema central de muchas de esas clases, se comentaban más que tratarse profundamente, ya fuera porque se consideraran obvios o ya conocidos, ya porque fueran a revisarse en otras materias.

Por ello, sentí que era compromiso de cada uno de los que aquí hemos estudiado, una vez que hayamos comprendido un tema, al menos en parte, tratar de compartir la experiencia que llevó a su comprensión o, mejor aún, simplificar a otros el camino que a uno tanto trabajo costó. De esta manera, nos repartiremos el esfuerzo y la recompensa será mayor.

No pretendo que cualquier persona, por el simple hecho de leer este trabajo o de haber asistido a su exposición, pueda manejar el tema. La intención es contribuir a ese proceso de comprensión con las herramientas de las que uno ha hecho acopio durante ese mismo proceso.

Lo que al principio fueron unos ejemplos manejados de la manera más sencilla posible para el que escribe, fue haciéndose más elaborado conforme se pretendía fuese de utilidasd para un mayor número de usuarios.

Finalmente, se trabajaron varias ideas, dominando la temática de las variables indicadoras y la preocupación por llegar a un público variado, de muchas áreas, pero sin formación matemática sólida.

A lo largo del escrito se sigue una forma bastante uniforme de presentación, la cual consiste básicamente en lo siguiente:

- un conjunto de datos asociados a un problema,
- un manejo intuitivo del problema,
- la proposición de un modelo,
- la explicación de cada término del modelo,
- el ajuste del modelo,
- la interpretación de los datos a través del resultado del ajuste del modelo.

La primera parte del trabajo es tal vez la que difiere un poco del patrón anterior, pues sólo hasta el final se trabaja un modelo de regresión. En esta primera parte, con un mismo conjunto de datos, se resaltan diferencias en el tratamiento de los datos e interpretación de los resultados con base en diferentes situaciones alrededor de dichos datos.

En la segunda y tercera parte, la secuencia de la presentación prácticamente se repite para cada juego de hipótesis que se pretende probar.

En la cuarte parte, además de diferentes juego de hipótesis para un mismo conjunto de datos, se manejan tanto modelos de diseño de experimentos como sus "equivalentes" modelos de regresión.

Esta forma reiterada de presentación, si bien llega a ser cansada para aquéllos con bases matamáticas o con conocimiento de este tema en particular, es un medio para facilitar la conprensión y el aprendizaje de aquéllos sin tales antecedentes.

Los años de estudiante y de experiencia docente me han mostrado que al enseñar no basta con presentar los conceptos de manera clara y sencilla; es necesario exponerlos repetidas veces y trabajarlos mediante ejemplos (mejor aún con ejercicios) para que maduren en el conocimiento de los educandos, máxime cuando se trata de un área nueva o considerada tradicionalmente difícil, como es el caso de las matemáticas.

Además, se han hecho algunos señalamientos con el fin de facilitar la lectura y la comprensión. En primer lugar, se han puesto en letra cursiva y de menor tamaño párrafos que, si bien puede omitirse su lectura sin perder la continuidad del tema, están dirigidos preferentemente, pero NO excusivamente, a los lectores con menos antecedentes estadísticos y de áreas diferentes a la matemática.

En segundo lugar, al final de cada sección se presentan algunos comentarios, limitados por líneas dobles, que expresan las princiapales ideas de la sección, con uno o dos párrafos en cursivas a manera de conclusión.

En tercer lugar, se incluyen notas en cuadros sombreados (listados en la página *v*) dirijidas a aquéllos que deseen profundizar un poco más en ciertos aspectos particularmente importantes y de interés, pero que requieren un poco más de antecedentes o esfuerzo para su entendimiento.

Por último, considerando que en ocasiones no es el prefacio lo primero que se lee y que muchas veces se consulta sólo parte de un trabajo sin revisarlo en su totalidad, al principio de cada sección se remite al lector al prefacio invitándolo a leerlo, para que comprenda la estructura y el formato del escrito y se le facilite su consulta. Así mismo, en las secciones posteriores, se le invita a revisar los conceptos generales y la explicación de la salida del paquete estadístico que se encuentran en las primeras secciones.

El material tal cual está escrito, dada su extensión, no podrá exponerse en una conferencia como se planeó originalmente. Sin embargo, puede ser adaptado de varias maneras con diferentes propósitos. Lo que se pretende en última instancia, más allá de cumplir con un requisito de titulación, es el de ofrecer un material de utilidad a estudiantes e incluso a profesores, en la medida que unos y otros lo encuentren accesible y manejable.

Dentro de las posibilidades de uso se pueden mencionar, a manera de ejemplo, los siguientes temas:

- *Introducción a las variables indicadoras* (II -una sesión- o II y III -dos sesiones).
- *El ajuste de diferentes modelos y su efecto en la interpretación* (III).
- *Reexpresión de modelos de diseño de experimentos como modelos de regresión* (IV).

Espero que este trabajo pueda servir a quien busque una respuesta a cuetiones básicas que no siempre son tan sencillas, como es el tema que nos ocupa.

Gerardo Varela Hdez.
Agosto de 1997

ÍNDICE

LISTA DE CUADROS DE NOTAS

I. NECESIDAD DEL CONOCIMIENTO DE LA HERRAMIENTA ESTADÍSTICA PARA SU USO ADECUADO

Estimado lector:

Antes de entrar en tema, es recomendable que lea primero el prefacio, donde no solamente se explica el contenido de este trabajo, sino también el formato que se ha seguido en su presentación. Tales explicaciones le permitirán una mayor comprensión del documento, así como una lectura más fluída.

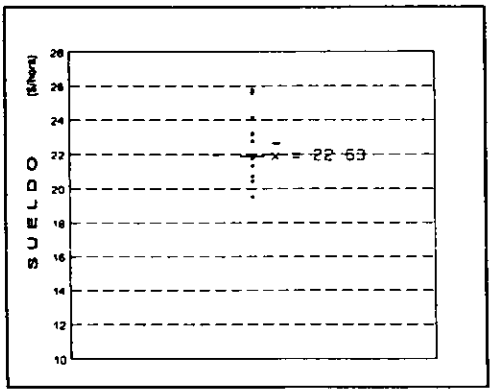
Si usted ya leyó el prefacio, lo invito a continuar la lectura de esta página.

Cuando pretendemos analizar datos queremos saber algo de ellos, queremos que nos hablen, que nos revelen propiedades, características o regularidades del fenómeno en estudio, que nos lo describan, si no completamente, al menos lo suficiente para contestar las preguntas que sobre él nos formulemos.

La estadística representa una herramienta muy útil para tal propósito. Por ejemplo, de los siguientes datos referentes a sueldo, en nuevos pesos por hora,

Sueldo ($/hora)			
19.51	21.34	23.16	24.14
20.44	21.76	23.20	25.59
20.71	22.75	23.21	25.76

podemos ver cuál es la media del sueldo en $/hora ($\bar{x}$ = 22.63) o incluso calcular un intervalo de confianza para la media del sueldo de la poblacion de donde se extrajeron tales datos.

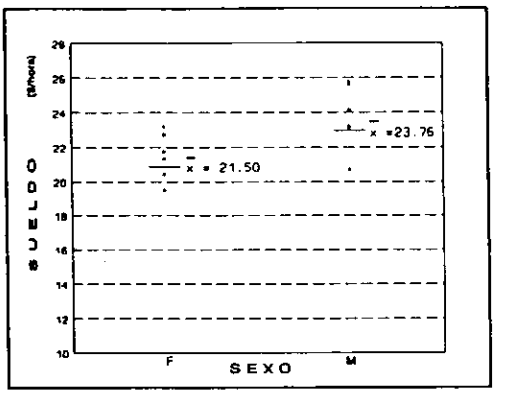


Contando con más información y sabiendo que los datos se colectaron entre hombres y mujeres, registrándose el sexo, tendríamos:

Sueldo ($/hora)			
Mujeres		Hombres	
19.51	22.75	20.71	24.14
20.44	23.20	23.16	25.76
21.34	21.76	23.21	25.59

Podemos preguntarnos si los hombres ganan más que las mujeres, es decir, ver si los datos dan evidencia de que la media del sueldo de los hombres ($\overline{x}$ = 23.76 en la muestra) es diferente que al de las mujeres ($\overline{x}$ = 21.50 en la muestra) y si es mayor o menor, en la población de donde se tomó la muestra.

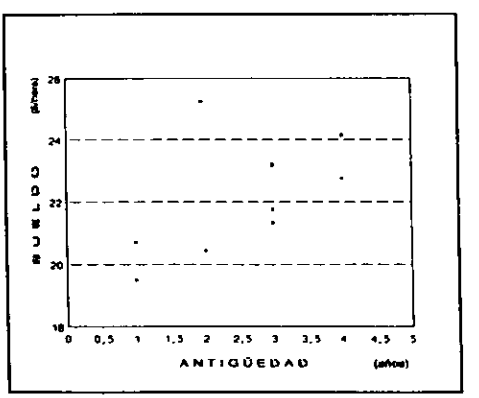
Gráficamente podría expresarse



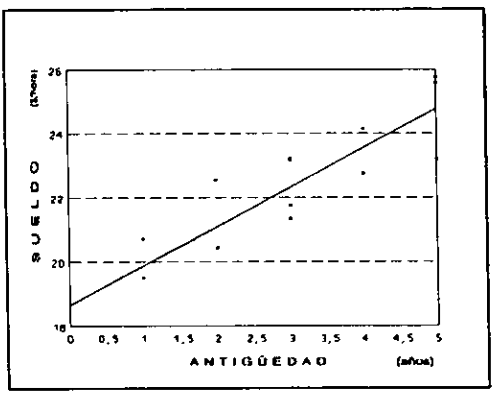
Por otro lado, la información adicional pudiera ser sobre la antigüedad laboral, en años,

Años de exper.	1	2	3	4	5
SUELDO ($/hora)	19.51	20.44	22.34	22.75	23.26
			21.76		25.59
	20.71		23.16	24.14	25.76
			23.21		

y entonces podríamos preguntarnos si están relacionados el sueldo y la antigüedad, es decir, si en general al aumento de la antigüedad corresponde un aumento o disminución en el sueldo, o es irrelevante en la variación de éste último.



La gráfica del sueldo contra la antigüedad sugiere la existencia de una relación entre el sueldo (variable respuesta) y la antigüedad (variable explicativa), pues se aprecia que, en general, al aumento de la antigüedad corresponde un aumento en el sueldo, y este aumento es "constante". De hecho, podría dibujarse una recta inclinada en medio de la nube de puntos, que mostraría dicho comportamiento general de los datos, es decir, la tendencia del aumento del sueldo conforme aumenta la antigüedad.



Con este sencillo ejemplo, puede verse la importancia de hacer una gráfica con los datos (en este caso un "diagrama de dispersión") antes de plantear un modelo o una prueba de hipótesis.

La ecuación de una recta se emplea precisamente para relacionar el valor de una variable (la variable respuesta "Y") con el valor de otra variable (la variable explicativa "X"), cuando se observa una tendencia como la que siguen los datos del ejemplo, conocida como "lineal", donde la variable respuesta "Y" sería el sueldo y la variable explicativa "X" sería la antigüedad.

$$Y = b + m X$$

La ecuación de una recta muestra el valor de Y cuando X vale cero, nombrándolo ordenada al origen y representándolo ("denotándolo", en lenguaje más formal) como "b". También muestra cuántas unidades aumenta la variable respuesta "Y" cuando la variable explicativa "X" aumenta una unidad, lo que se conoce como pendiente, representada o denotada por "m". La idea que se introdujo arriba de aumento "constante" es que independientemente del valor que partamos de X para incrementarlo en una unidad, el aumento correspondiente en Y siempre será el mismo valor, "m".

Probar que en general exista una relación lineal entre el sueldo y la antigüedad, equivale a probar que en general la variación del sueldo (variable respuesta) puede ser expicada con la variación de la antigüedad (variable explicativa); es decir, que la relación entre el sueldo y la antigüedad puede ser descrita globalmente como una recta donde la pendiente fuese "diferente de cero".

Una recta con pendiente "igual a cero" es una recta horizontal, y mostraría un mismo valor de "Y" independientemente del valor de "X". En nuestro ejemplo querría decir que el sueldo es el mismo sin importar la antiguedad.

Estadísticamente lo que se hace es "ajustar" a los datos un modelo que involucra tanto una recta (como la que mencionamos para la gráfica), como un término que mide el alejamiento de los datos a la recta propuesta.

El modelo empleado para ajustar una recta a los datos anteriores, es el siguiente:

$$\text{SUELDO}_{ir} = \beta_0 + \beta_1 \text{ ANTIGÜEDAD}_i + \varepsilon_{ir} \qquad \begin{matrix} i = 1,2,\ldots,5. \\ r = 1,2,\ldots,n_i. \end{matrix}$$

ó

$$Y_{ir} = \beta_0 + \beta_1 X_i + \varepsilon_{ir}$$

con

$\varepsilon_{ir} \sim NIID(0,\sigma^2)$

(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada valor de i de X)
(los errores son **independientes**, el valor de un error no tiene que ver que ver con el valor de otro error)
(todos los errores se **distribuyen** de manera **idéntica**, todos provienen de una misma distribución normal con media 0 y varianza σ^2)

donde

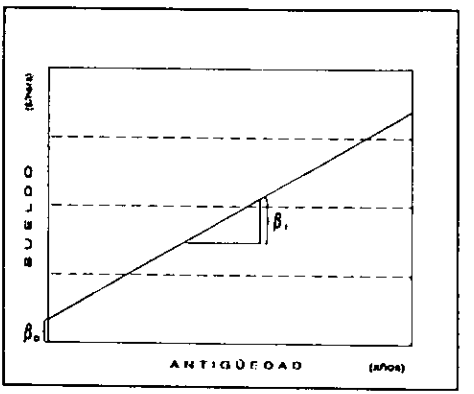
- i = i-ésimo valor de anigüedad considerado (desde 1, hasta 5).
- r = r-ésima observación (repetición) del i-ésimo valor de antigüedad (desde 1, hasta n_i).
- Y_{ir} = S U E L D O (en $) del profesor de la r-ésima observación (repetición) del i-ésimo valor de antigüedad.
- X_i = i-ésimo valor de años de A N T I G Ü E D A D.
- ε_{ir} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de antigüedad.

en el cual la primera parte ($Y_{ir} = \beta_0 + \beta_1 X$) corresponde a la ecuación de una recta ($Y = b + mX$), mientras que la segunda parte (ε_{ir}) es el componente aleatorio que corresponde a la separación entre los puntos y la recta descrita por la primera parte. Estas desviaciones de la recta se conocen como "errores" aleatorios y

deben cumplir con las características enunciadas (ε_{ir}~NIID(0,σ^2)) para que permitan el ajunte del modelo mediante técnicas estadísticas. Tales características son lo que se conocen como "supuestos del modelo" (ver pag. 8).

Por lo tanto, β_0 representa la ordenada al origen o intercepto ("b"), esto es, el sueldo recién se ingresa al trabajo (antigüedad de "cero años"), y β_1 la pendiente ("m") de la recta de ajuste, es decir, el cambio en el sueldo en pesos por cada año de antigüedad.

$$SUELDO_{ir} = \beta_0 + \beta_1 \, ANTIGÜEDAD_i + \varepsilon_{ir}$$



La estimación de β_0 y de β_1, es decir, la suposición que hacemos del valor de β_0 y del de β_1, que se representan con β_0 y β_1 en los modelos, se calculan a través de los datos y las técnicas estadísticas, y se obtendrán con el paquete estadístico "SYSTAT5" para todos los ejemplos que se presenten.

De esta manera, la relación entre el sueldo y la antigüedad se ajusta con el modelo

$$\hat{Y}_{ir} = \beta_0 + \beta_1 X_i$$

donde $\hat{Y}_{ir}$ es el sueldo estimado para el profesor correspondiente a la r-ésima observación en el i-ésimo valor de antigüedad, tomando en cuenta su antigüedad X_i.

En las salidas del paquete estadístico las β's serán los coeficientes (listados bajo el encabezado **"COEFFICIENT"**) asociados a las variables enlistadas bajo el encabezado **"VARIABLE"**, donde

β_0 es el coeficiente asociado con el término "*CONSTANT*", mientras que β_1 es el coeficiente asociado con el nombre suministrado al paquete (en el ejemplo es "*AN*") de la variable con la que se pretende explicar la variación de "Y".

A continuación se presenta una "salida" del paquete correspondiente a nuestro ejemplo, la cual ha sido editada resaltando con letras sombreadas o en negritas cursivas y agregando los símbolos "← β_0" y "← β_1", para señalar lo mencionado en el párrafo anterior:

DEP VAR: SAL N: 12 MULTIPLE R: 0.886 SQUARED MULTIPLE R: 0.785
ADJUSTED SQUARED MULTIPLE R: .763 STANDARD ERROR OF ESTIMATE: 0.956

VARIABLE	COEFFICIENT		STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	*18.655*	← β_0	.714	0.000	.	26.122	0.000
AN	*1.223*	← β_1	.203	0.886	1.000	6.036	0.000

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	33.294	1	33.294	36.431	0.000
RESIDUAL	9.139	10	.914		

Previo a la obtención de concluciones sobre las β's (de preferencia antes de ajustar modelo alguno) es necesario verificar que todo aquello en lo que se basa el modelo propuesto, lo que se conoce como "supuestos del modelo", realmente se cumpla. Esto puede realizarse con análisis gráficos o pruebas de normalidad y aleatoriedad sobre los residuos.

En nuestro ejemplo, el residuo para cada dato es la diferencia entre el valor del sueldo de un profesor y el valor del sueldo estimado con el modelo y el dato de antigüedad de ese profesor.

Para entender la importancia de la verificación de los "supuestos" veamos una analogía con un ejemplo de escuela:

> Un alumno aprueba un curso bajo el supuesto de que aprendió durante el mismo (lo que puede verificarse mediante un examen o un trabajo, por ejemplo); si no se cumpliera el supueto del aprendizaje, entonces el alumno no debería aprobar el curso. Si se le diera una nota

aprobatoria o reprobatoria sin la previa verificación, la calificación podría no corresponder con la realidad, y cualquier conclusión sobre el aprovechamiento del alumno basada en tal nota sería incierta.

Con el afán de hacer más sencillo el desarrollo del tema, no se explicará ni se llevará a cabo la verificación de los supuestos de los modelos propuestos para cada conjunto de datos; ya que la presentación y justificación de las pruebas del cumplimiento de tales supuestos podría ser en sí el tema de otro trabajo semejante al que aquí se expone.

Una vez que el modelo propuesto se ha ajustado a los datos y que se han verificado los supuestos (normalidad, aleatoriedad, homogeneidad de varianzas), nuestro interés se centra en ver si se puede concluir que exista tal "relación" entre la antigüedad y el sueldo, como lo establece el modelo.

Con ese propósito realizaremos una prueba (conocida como prueba de "t") para ver si $\beta_1 \neq 0$, es decir, si los datos aportan suficiente evidencia para concluir que la variación del sueldo de acuerdo con la antigüedad puede describirse con una recta con pendiente diferente de cero.

Recordemos que β_1 representa la pendiente, y que si es cero, esto es, la línea es horizontal, no importando cual sea la antigüedad, el sueldo será siempre el mismo).

En la salida del paquete nos concretaremos en ver el valor de P correspondiente a la "t" de la prueba para β_1; de tal manera que si el valor de P es igual o menor que 0.05, concluiremos que los datos dan evidencia estadística para decir que $\beta_1 \neq 0$, lo que en este caso equivale a decir que sí hay relación lineal entre la antigüedad y el sueldo de los profesores de nuestro ejemplo.

La forma en que puede interpretarse esta "P" es como "la probabilidad de que con nuestros datos β_1 fuera igual a cero", o como "el riesgo que corremos al decir, con los datos de la muestra, que β_1 es diferente de cero". Por ello, el valor "P" recibe el nombre de "nivel de significancia descriptivo (de la muestra)". Puesto que no queremos correr un gran riesgo de equivocar en nuestras conclusiones (en este caso la relación entre sueldo y antiguedad), fijaremos como límite un valor de 0.05 para "P".

El **valor límite de P** para aceptar los resultados como evidencia estadísticamente significativa es fijado por quien esté realizando la investigación o estudio. Tal valor generalmente se toma entre **0.01** y **0.05**, se conoce como **"α" (alfa)**, **"nivel de significancia de la prueba"** o **"probabilidad de cometer el error tipo I"** y representa un límite al riego que queremos correr al confiar en datos de una muestra para concluir sobre una población. De ahí que su nombre incluya el término "error"[1].

Dados el valor de "P" (0.000) calculado con los datos, resultado de la prueba de "t" para ver si β_1 es diferente de cero, podríamos considerar que los datos sí muestran evidencia estadística para suponer una relación lineal entre los años de antigüedad y el sueldo (tomando un $\alpha = 0.05$).

DEP VAR: SAL N: 12 MULTIPLE R: 0.886 SQUARED MULTIPLE R: 0.785
ADJUSTED SQUARED MULTIPLE R: .763 STANDARD ERROR OF ESTIMATE: 0.956

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	18.655 ← β_0	.714	0.000	.	26.122	0.000
AN	1.223 ← β_1	.203	0.886	1.000	6.036	*0.000*

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	33.294	1	33.294	36.431	0.000
RESIDUAL	9.139	10	.914		

Dado que el valor de "P" de la salida (P = 0.000) es menor a lo más que estábamos dispuestos a arriesgarnos (0.05), podemos concluir, sin mucho temor a equivocarnos pero nunca con plena seguridad (la informació es tan solo de una muestra), que hay una relación lineal entre sueldo y antiguedad en la población de donde se obtuvieron los datos. Por ello, las conclusiones se redactan generalmente como "...los datos aportan evidencia estadística de que...", en vez de ser categóricas.

El valor P = 0.000 no significa que P sea "0", sino que es un valor tan pequeño, pero mayor que cero, que redondeado a tres decimales resulta cero. De ahí que se manejen tres ceros como decimales.

[1]Estrictamente, el "error" hace referencia al rechazo de Ho siendo verdadera, y α representa el límite al riesgo que queremos correr al rechazar Ho siendo verdadera. Considérese Ho: $\beta_1 = 0$ y vuelva a leer desde dos párrafos antrás, donde se explica el significado de "P".

Como pudimos ver, a través del análisis de los datos colectados durante un estudio, podemos lograr que aquéllos "nos hablen", nos revelen propiedades, características o regularidades del fenómeno de estudio (media del sueldo por hora, media del sueldo por hora para cada sexo, relación entre sueldo y antigüedad laboral, en los ejemplos), describiéndolo lo suficiente como para contestar algunas de las preguntas que sobre él nos formulemos.

Sin embargo, no se trata de ver qué información puede sacarse de los datos; sino que, antes de realizar el análisis estadístico, es indispensable que se determine el propósito del estudio, para entonces establecer la estrategia de colecta de datos y de su análisis estadístico, bien sea para describir el fenómeno, bien para apoyar o refutar las hipótesis que se hubiesen formulado para explicarlo, y entonces sí "dejar que hablen los datos".

De aquí se desprende, ya no tanto la necesidad del uso de la herramienta estadística por parte del investigador, sea cual fuere el área de su especialidad, sino de contar al menos con un conocimiento de tal herramienta para (i) la elaboración del proyecto de investigación que desde un principio contemple el elemento estadístico, (ii) la interpretación de los resultados del análisis estadístico; todo ello independientemente de que el tratamiento estadístico de los datos lo realice el mismo investigador o un estadístico.

A continuación se presentan algunos modelos de regresión, la interpretación de cada uno de sus términos y las hipótesis que pueden probarse a través de los mismos.

II. COMPARACIÓN DE MODELOS DE REGRESIÓN SIMPLE ENTRE DOS GRUPOS

Estimado lector:

Antes de entrar en tema, es recomendable que lea primero el prefacio, donde no solamente se explica el contenido de este trabajo, sino también el formato que se ha seguido en su presentación. Tales explicaciones le permitirán una mayor comprensión del documento, así como una lectura más fluída.

Además, es conveniente que revise la sección I "NECESIDAD DEL CONOCIMIENTO DE LA HERRAMIENTA ESTADÍSTICA PARA SU USO ADECUADO", donde se exponen conceptos generales, se señalan resultados importantes generados por el paquete estadístico, y se explica su interpretación, lo cual será de gran utilidad para entender el desarrollo del resto de este trabajo.

Si usted ya leyó el prefacio y la sección I, lo invito a continuar la lectura de esta página.

Si recordamos que los datos del ejemplo de la sección anterior se registraron tanto de hombres como de mujeres, dado que la media del sueldo resultó diferente para los sexos, sería de interés saber si la relación entre la antigüedad y el sueldo es igual o diferente para ambos sexos, es decir, si la regresión es igual o diferente para hombres y mujeres.

Años de exper.		1	2	3	4	5
SUELDO ($)	Mujeres	19.51	20.44	21.34	22.75	23.20
				21.76		
	Hombres	20.71		23.16	24.14	25.76
				23.21		25.59

Considerando que ese hubiese sido el objetivo de quien colectara los datos, veamos como puede probarse la diferencia en las regresiones.

En el modelo de regresión simple están involucrados dos coeficientes; de haber diferencia en los modelos, podría ser en el intercepto u ordenada al origen (β_0), en la pendiente (β_1) o en ambos.

Como primera opción, podemos ajustar un modelo de regresión simple, es decir, de una recta, para los datos separados por sexo:

Mujeres		Hombres
$Y_{ir} = \beta_0 + \beta_1 X_i + \varepsilon_{ir}$		$Y_{ir} = \beta_0 + \beta_1 X_i + \varepsilon_{ir}$
$\hat{Y}_{ir} = \beta_0 + \beta_1 X_i$		$\hat{Y}_{ir} = \beta_0 + \beta_1 X_i$
$\beta_0 = 18.593$	(Coeficiente de CONSTANT)	$\beta_0 = 19.459$
$\beta_1 = 0.969$	(Coeficiente de AN)	$\beta_1 = 1.229$

VARIABLE	COEFF.	T	P(2 TAIL)
CONSTANT	18.593	73.5	0.00
AN	0.969	12.5	0.00

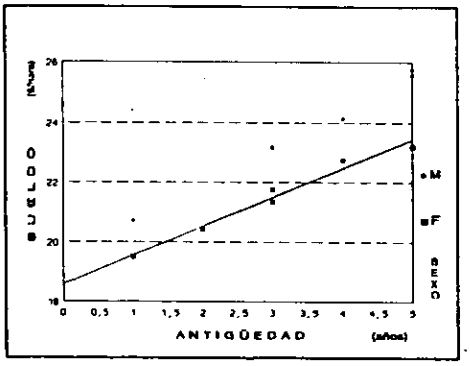
ANALYSIS OF VARIANCE

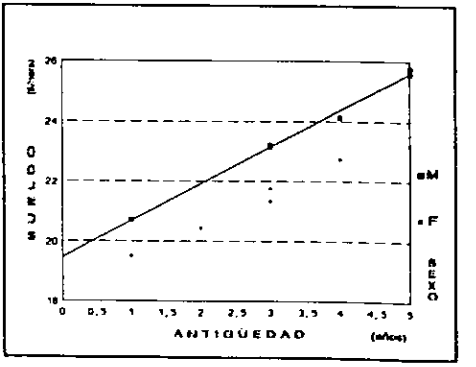
SOURCE	DF	MEAN-SQ	F	P
REGRES	1	9.389	156	0.00
RESID	4	0.059		

VARIABLE	COEFF.	T	P(2 TAIL)
CONSTANT	19.459	120.6	0.00
AN	1.229	28.6	0.00

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	1	17.37	822	0.00
RESID	4	0.021		





De esta forma podemos realizar la prueba de si las pendientes son diferentes de cero, pero no de si son diferentes entre sí. Lo mismo sucede para los interceptos (ordenadas al origen).

La manera de expresar los dos modelos en un modelo único, para poder probar si las dos pendientes y las dos ordenadas al origen (interceptos) son iguales o diferentes, es involucrando la variable sexo, como variable explicativa, al modelo de regresión que relaciona "Y" (sueldo) con "X" (años de antigüedad).

La forma de involucrar una variable categórica en un modelo de regresión, como es el caso del "sexo", es declarando explícitamente **cada una de sus categorías menos una** a través de variables indicadoras o "dummy" (de ahí su nombre de "indicadora, pues "indica" una categoría particular de la variable), que representaremos (denotaremos) como "Z_{ij}", las cuales sólo podrán tomar valores de 1 (Z_{i1}=1) y 0 (Z_{i2}=0). La categoría no representada por las variables "dummy" queda declarada implícitamente por eliminación.

El subíndice "i" indica el número de categoría representada por la variable indicadora, mientras que el subíndice "j" indica el número posible de valores que toma Z (que ya se dijo sólo pueden ser dos: 1 ó 0). Dado que la variable "sexo" sólo presenta dos categorías y que el número de variables indicadoras por variable categórica debe ser menor en uno a las categorías, sólo es necesaria una variable indicadora, por lo que se omitirá el subíndice "i".

Conviniendo en que si j=1 nos referiremos al sexo masculino, la variable Z_j es en realidad indicadora del sexo masculino. Sin embargo, el sexo femenino queda representado implícitamente pues si j=2 no nos referiremos al sexo masculino, no quedando otra posibilidad que la de sexo femenino. Debe quedar claro que es igualmente posible emplear la variable indicadora para representar explícitamente al sexo femenino, quedando definido implícitamente el sexo masculino.

Considerando a la vez el subíndice y los valores de Z_j para el caso que nos ocupa, quedó convenido que Z tome el valor de 1 cuando nos refiramos al sexo masculino ($j = 1$, $Z_j = Z_1 = 1$), y 0 si nos referimos al sexo femenino ($j = 2$, $Z_j = Z_2 = 0$).

Por tanto, en nuestro ejemplo tendremos que

$$Z_j = \begin{cases} 1 \text{ si SÍ es hombre} & (j=1) \\ 0 \text{ si NO es hombre (es mujer)} & (j=2) \end{cases}$$

Nota 1. ***Declaración de una variable categórica a través de variables indicadoras.***

De manera general, para involucrar o declarar una variable categórica en un modelo de regresión, se emplean tantas variables indicadoras "dummy" como categorías tenga la variable categórica original menos una. De este modo, si "a" es el número de categorías de la variable cualitativa, "a-1" es el número de variables indicadoras necesarias para representar la variable cualitativa, es decir, cada categoría individual será declarada explícitamente con una variable "indicadora" (indicadora de esa categoría en particular) excepto una categoría que quedará expresada implícitamente. (El subíndice "i" servirá para reconocer cada una de estas "a-1" categorías representadas explícitamente).

Cada variable "dummy" o indicadora sólo puede tomar dos valores (j=1,2): toma el valor de 1 cuando representa una categoría de la variable original (j=1), y toma el valor de 0 si la variable categórica tiene un valor diferente a la categoría que representan la variable indicadora (j=2). Esto es, cada variable indicadora representa una categoría cuando vale 1 (j=1). La categoría no declarada explícitamente a través de una variable indicadora con valor de 1, quedará expresada implícitamente cuando todas las variables "dummy" valgan 0.

Por ejemplo, para representar la variable categórica "color", con las cuatro categorías (a=4) "azul", "rojo", "amarillo" y "verde", se requieren tan solo tres (a-1=4-1=3) variables indicadoras (Z_{ij} con i=1,2,3 ó Z_{1j}, Z_{2j} y Z_{3j}), como se presenta en la siguiente tabla:

Color	Variables Indicadoras de Categoría		
	Z_{1j}(azul)	Z_{2j}(rojo)	Z_{3j}(amarillo)
Azul	1 (j=1)	0 (j=2)	0 (j=2)
Rojo	0	1	0 (j=2)
Amarillo	0	0	1 (j=1)
Verde	0	0	0 (j=2)

La primera variable "dummy" (Z_{1j}), indicadora del color azul, tomará el valor de 1 cuando el color sea "azul" (j=1, $Z_{1j}=Z_{11}=1$) y el valor 0 cuando NO sea "azul" (j=2, $Z_{1j}=Z_{12}=0$); la segunda (Z_{2j}), indicadora de color rojo, valdrá 1 cuando el color sea "rojo" (j=1, $Z_{2j}=Z_{21}=1$) y el valor 0 cuando NO sea "rojo" (j=2, $Z_{2j}=Z_{22}=0$); mientras que la tercera (Z_{3j}), indicadora del color "amarillo", será igual a 1 cuando el color sea "amarillo" (j=1, $Z_{3j}=Z_{31}=1$) y el valor 0 cuando NO sea "amarillo" (j=2, $Z_{3j}=Z_{32}=0$). El color "verde" quedará representado cuando las tres variables indicadoras tomen el valor de 0, esto es, no es azul, ni rojo, ni amarillo, sino verde ($Z_{i2}=0$, $Z_{22}=0$, $Z_{32}=0$, o $Z_{i2}=0$ para toda "i").

Nota 2. ***Declaración de más una variable categórica a través de variables indicadoras.***

En caso de necesitar declarar variables con más de dos categorías, dado que se emplearían dos o más variables indicadoras para variable categórica, resultaría más conveniente emplear letras diferentes para una de las variables indicadoras de las categorías de las diferentes variables cualitativas, y así asociar el primer subíndice a cada categoría de la variable cualitativa.

Por ejemplo, para declarar simultáneamente cinco tipos de fertilizante (F_1, F_2, F_3, F_4 y F_5), cinco tipos de fungicida (G_1, G_2, G_3, G_4 y G_5) y cinco tipos de insecticida (I_1, I_2, I_3, I_4 e I_5), se podría de emplear S_{ij} (con i=1,2,3,4) para las cuatro categorías a declarar de fertilizante, U_{kj} (con k=1,2,3,4) para las cuatro categorías a declarar de fingicida, y V_{lj} (con l=1,2,3,4) para las cuatro categorías a declarar de insecticida, teniendo todas el primer subíndice igual a la categoría de la variable cualitativa original que representan.

Tal ejemplo se ilustra a continuación:

Variables Categóricas	Variables Indicadoras (Una letra diferente correspondientes a cada variables categórica original)	
Fertilizante (F1, F2 F3, F4 y F5)	$S_{1j}=\begin{cases}1 \text{ si es } F_1 & (j=1)\\ 0 \text{ si NO es } F_1 & (j=2)\end{cases}$	$S_{2j}=\begin{cases}1 \text{ si es } F_2 & (j=1)\\ 0 \text{ si NO es } F_2 & (j=2)\end{cases}$
	$S_{3j}=\begin{cases}1 \text{ si es } F_3 & (j=1)\\ 0 \text{ si NO es } F_3 & (j=2)\end{cases}$	$S_{4j}=\begin{cases}1 \text{ si es } F_4 & (j=1)\\ 0 \text{ si NO es } F_4 & (j=2)\end{cases}$
Fungicida (G1, G2 G3, G4 y G5)	$U_{1j}=\begin{cases}1 \text{ si es } G_1 & (j=1)\\ 0 \text{ si No es } G_1 & (j=2)\end{cases}$	$U_{2j}=\begin{cases}1 \text{ si es } G_2 & (j=1)\\ 0 \text{ si NO es } G_2 & (j=2)\end{cases}$
	$U_{3j}=\begin{cases}1 \text{ si es } G_3 & (j=1)\\ 0 \text{ si No es } G_3 & (j=2)\end{cases}$	$U_{4j}=\begin{cases}1 \text{ si es } G_4 & (j=1)\\ 0 \text{ si NO es } G_4 & (j=2)\end{cases}$
Insecticida (I1, I2 I3, I4 e I5)	$V_{1j}=\begin{cases}1 \text{ si es } G_1 & (j=1)\\ 0 \text{ si No es } G_1 & (j=2)\end{cases}$	$V_{2j}=\begin{cases}1 \text{ si es } I_2 & (j=1)\\ 0 \text{ si NO es } I_2 & (j=2)\end{cases}$
	$V_{1j}=\begin{cases}1 \text{ si es } G_1 & (j=1)\\ 0 \text{ si No es } G_1 & (j=2)\end{cases}$	$V_{2j}=\begin{cases}1 \text{ si es } I_2 & (j=1)\\ 0 \text{ si NO es } I_2 & (j=2)\end{cases}$

(continúa en la siguiente página)

Nota 2 *(continuación)*

Por otro lado, en caso de necesitar declarar únicamente variables con tan sólo dos categorías, dado que se emplearía una sola variable indicadora para cada variable categórica, podría resultar útil conservar un sólo índice con las funciones de ambos. Aún más, en véz de usar una letra diferente para cada variable indicadora, podría emplearse una sola letra indexada (además del subíndice mencionado anteriormente).

En ambos casos la variación de la letra del último subíndice (el único en el primer caso) se hace necesaria para indicar en la respuesta las diferentes variables y sus categorías.

Por ejemplo, para declarar simultáneamente dos tipos de fertilizante (F1 y F2), dos tipos de fungicida (G1 y G2) y dos tipos de insecticida (I1 e I2), en vez de emplear S_{1j} (S_{ij} con i=1, j=1,2) para la única categoría a declarar de fertilizante, U_{1j} (U_{kj} con k=1, j=1,2) para la única categoría a declarar de fingicida, y V_{1j} (V_{lj} con l=1, j=1,2) para la única categoría a declarar de insecticida, teniendo todas el primer subíndice igual a "1", podrían "combinarse" ambos subíndices "suprimiendo" el primero ("i", "j" y "k") y variando el segundo ("j" pasaría a ser "i", "k" o "l") para poderlo expresar en la respuesta; o lo que es lo mismo, "eliminar" el segundo ("j"), y que ahora el primero tome valores "1" y "2" para representar el valor de la variable indicadora).

Además, en vez de "S", "U" y "V" se podría emplear Z_1, Z_2 y Z_3 (Z_m con m=1,2,3).

Ambas opciones se ilustran a continuación:

Variables Categóricas	Variables Indicadoras: Varias Letras	Variables Indicadoras: Una Letra
Fertilizante (F1, F2)	$S_i = \begin{cases} 1 \text{ si es F1} & (i=1) \\ 0 \text{ si NO es F1 (es F2)} & (i=2) \end{cases}$	$Z_{1i} = \begin{cases} 1 \text{ si es F1} & (i=1) \\ 0 \text{ si NO es F1 (es F2)} & (i=2) \end{cases}$
Fungicida (G1, G2)	$U_k = \begin{cases} 1 \text{ si es G1} & (k=1) \\ 0 \text{ si No es G1 (es G2)} & (k=2) \end{cases}$	$Z_{2k} = \begin{cases} 1 \text{ si es G1} & (k=1) \\ 0 \text{ si NO es G1 (es G2)} & (k=2) \end{cases}$
Insecticida (I1, I2)	$V_l = \begin{cases} 1 \text{ si es I1} & (l=1) \\ 0 \text{ si No es I1 (es I2)} & (l=2) \end{cases}$	$Z_{3l} = \begin{cases} 1 \text{ si es I1} & (l=1) \\ 0 \text{ si NO es I1 (es I2)} & (l=2) \end{cases}$

En todos los casos el uso de las letras para los subíndices es arbitrario. Por ejemplo, en vez de "m" podría haberse usado "j", pero no se empleó para evitar confusiones dado que se había estado empleando para los dos valores que toman las variables indicadoras.

1. *Modelo para detectar diferencia en Ordenada al Origen (Intercepto).*

Una diferencia en intercepto u "ordenada al origen" es una diferencia en el valor de Y (sueldo) cuando X (antigüedad) es cero, es decir, es una diferencia en el sueldo inicial (cuando recién se entra a trabajar) que recibiría un hombre del que recibiría inicialmente una mujer.

Con el modelo que se explica a continuación, no se podrá ver diferencia en pendiente, es decir, no se modelará ni se probará que haya diferencia entre hombres y mujeres en el aumento de sueldo dado por antiguedad; tal efecto se contemplará en modelos posteriores.

Para representar una diferencia en intercepto (ordenada al origen) entre los dos modelos particulares por sexo, esto es, diferencia en el sueldo base (considerando que no hay antiguedad) entre hombres y mujeres, el modelo general a ajustar sería

$$\text{SUELDO}_{ijr} = \beta_0 + \beta_1\ \text{ANTIGÜEDAD}_i + \beta_2\ \text{SEXO}_j + \varepsilon_{ijr}$$

ó

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 Z_j + \varepsilon_{ijr}$$

con

$i = 1,2,\ldots,5.$
$j = 1,2.$
$r = 1,2,\ldots,n_i.$

$$Z_j = \begin{cases} 1 \text{ si es hombre} & (j=1) \\ 0 \text{ si NO es hombre (es mujer)} & (j=2) \end{cases}$$

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

- i = i-ésimo valor de antigüedad considerado (desde 1 hasta 5).
- j = j-ésimo valor de la variable indicadora de sexo.
- r = r-ésima observación (repetición) del i-ésimo valor de antigüedad (desde 1, hasta n_i).
- Y_{ijr} = S U E L D O (en $) del profesor de la r-ésima observación (repetición) del i-ésimo valor de antigüedad y j-ésimo valor de S E X O.
- X_i = i-ésimo valor de años de A N T I G Ü E D A D.
- Z_j = j-ésimo valor de la variable indicadora de S E X O (Z_1=1, Z_2=0).
- ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de antigüedad y j-ésimo valor de SEXO.

Si la categoría de sexo es "mujer" (no es "hombre"), y por tanto $Z_j = Z_2 = 0$, el modelo se reduciría a

$$Y_{i2r} = \beta_0 + \beta_1 X_i + \varepsilon_{i2r}$$

porque $\beta_2(0) = 0$.

Si la categoría de sexo es "hombre", y por tanto $Z_j = Z_1 = 1$, el modelo quedaría

$$Y_{i1r} = \beta_0 + \beta_1 X + \beta_2 + \varepsilon_{i1r}$$

porque $\beta_2(1) = \beta_2$, o reagrupando

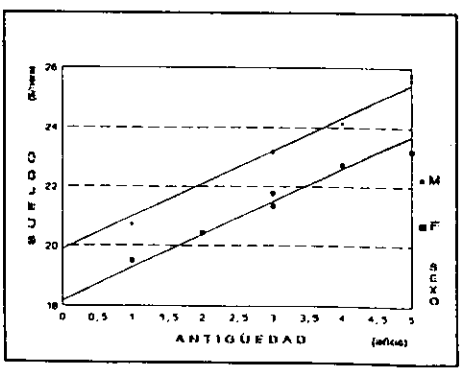
$$Y_{i1r} = (\beta_0+\beta_2) + \beta_1 X_i + \varepsilon_{i1r}$$

De acuerdo con la salida del paquete estadístico (SYSTAT5) correspondiente al modelo anterior, podemos decir que, los datos de nuestro ejemplo aportan evidencia estadísticamente significativa de diferencia en intercepto (ordenada al origen), ya que P = 0.000 (para ver si β_2 es diferente de cero) y considerando un $\alpha = 0.05$[2]. (La gráfica se elaboró con los coeficientes estimados reportados en la misma salida).

VARIABLE	COEFF.	T	P(2 TAIL)
CONSTANT	18.176	86.0	0.00
AN	1.108	18.6	0.00
SEXO	1.708 β_2	10.5	*0.00*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	2	20.87	274	0.00
RESID	9	0.08		



[2]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

Resumiendo, el modelo general estimado que involucra la especificación del sexo para la diferencia en intercepto (ordenada al origen)

$$\hat{Y}_{ij} = \beta_0 + \beta_1 X + \beta_2 Z_j$$

una vez definido si se trata de mujeres u hombres, se puede ver como

$$\hat{Y}_{i2} = \underbrace{\beta_0}_{\llcorner\_\lrcorner} + \beta_1 X_i \qquad \text{para mujeres} \quad (j = 2)$$

y

$$\hat{Y}_{i1} = \underbrace{(\beta_0 + \beta_2)}_{\llcorner\_\_\_\_\_\lrcorner} + \beta_1 X_i \qquad \text{para hombres} \quad (j = 1)$$

De las últimas dos expresiones se desprende que la diferencia entre ambas es la ordenada al origen (intercepto), señalada con $\llcorner\_\lrcorner$, y en ella la diferencia la da β_2, el coeficiente asociado originalmente a "Z", la variable indicadora.

Esto es, β_2 representa el cambio estimado en la ordenada al origen cuando se pasa de la categoría de sexo "mujer" (Z_2=0) a la categoría "hombre" (Z_1=1).

En nuestro ejemplo β_2 representa el cambio estimado en el sueldo inicial (cuando recién se entra a trabajar) al pasar en el modelo de la categoría de sexo "mujer" (Z_2=0) a la categoría "hombre" (Z_1=1), es decir, es una diferencia entre el sueldo inicial que recibiría un hombre del que recibiría una mujer.

De tal forma que si la hipótesis "H_0: $\beta_2 = 0$" no se rechaza (esto es si $\beta_2 = 0$), ambos modelos serían iguales; mientras que si la hipótesis "H_0: $\beta_2 = 0$" sí se rechaza (esto es si $\beta_2 \neq 0$) los modelos son diferentes en intercepto (ordenada al origen).

A continuación se presenta un resumen gráfico de esto mismo.

$$y_{i2r} = \beta_0 + \beta_1 x_i + \varepsilon_{i2r}$$

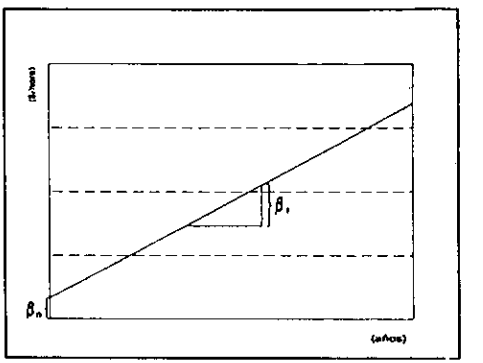
igual a

$$y_{i1r} = (\beta_0 + \beta_2) + \beta_1 x_i + \varepsilon_{i1r}$$

$\Leftrightarrow \quad \beta_2 = 0$

N O R E C H A Z O

H_0: $\beta_2 = 0$



$$y_{i2r} = \beta_0 + \beta_1 x_i + \varepsilon_{i2r}$$

diferente de

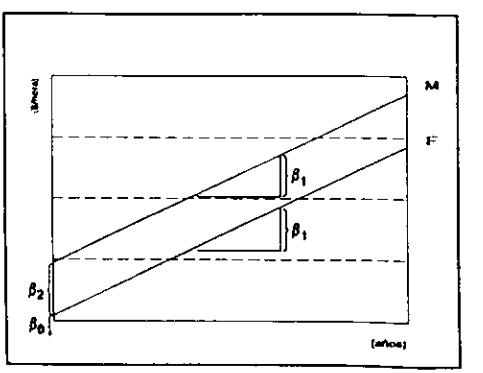
$$y_{i1r} = (\beta_0 + \beta_2) + \beta_1 x_i + \varepsilon_{i1r}$$

$\Leftrightarrow \quad \beta_2 \neq 0$

S Í R E C H A Z O

H_0: $\beta_2 = 0$

(aquí→ $\beta_2 > 0$)



2. *Modelo para detectar diferencia en Pendiente.*

Una diferencia en pendiente es una diferencia en el número de unidades que aumenta la variable respuesta "Y" (sueldo) cuando la variable explicativa "X" (antigüedad) aumenta una unidad, es decir, es una diferencia entre hombres y mujeres en el aumento de sueldo dado por un año más de antigüedad.

Con el modelo que se explica a continuación, no se podrá ver diferencia en intercepto ("ordenada al origen"), es decir, no se modelará ni se probará que haya diferencia en el sueldo inicial (cuando recién se entra a trabajar) que recibiría una hombre del que recibiría inicialmente una mujer; tal efecto ya se revisó en el modelo anterior y se verá combinado con la pendiente en modelos posteriores.

Por otro lado, para representar una diferencia en pendientes entre los dos modelos particulares por sexo, es decir, una diferencia en la variación por antigüedad del sueldo por sexo, el modelo general a ajustar sería

$$SUELDO_{ijr} = \beta_0 + \beta_1\ ANTIGÜEDAD_i + \beta_2\ X_i\ SEXO_j + \varepsilon_{ijr}$$

ó

$$Y_{ijr} = \beta_0 + \beta_1\ X_i + \beta_2\ X_i\ Z_j + \varepsilon_{ijr}$$

con

$i = 1,2,\ldots,5.$
$j = 1,2.$
$r = 1,2,\ldots,n_i.$

$$Z_j = \begin{cases} 1 \text{ si es hombre} & (j=1) \\ 0 \text{ si NO es hombre (es mujer)} & (j=2) \end{cases}$$

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

- i = i-ésimo valor de anigüedad considerado (desde 1, hasta 5).
- j = j-ésimo valor de la variable indicadora de sexo.
- r = r-ésima observación (repetición) del i-ésimo valor de antigüedad (desde 1, hasta n_i).
- Y_{ijr} = S U E L D O (en $) del profesor de la r-ésima observación (repetición) del i-ésimo valor de antigüedad y jésimo valor de SEXO.
- X_i = i-ésimo valor de años de A N T I G Ü E D A D.
- Z_j = j-ésimo valor de la variable indicadora de S E X O (Z_1=1, Z_2=0).
- ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de ANTIGÜEDAD y j-ésimo valor de SEXO.

Si la categoría de sexo es "mujer" (no es "hombre"), y por tanto $Z_j = Z_2 = 0$, el modelo se reduciría a

$$Y_{i2r} = \beta_0 + \beta_1 X_i + \varepsilon_{i2r}$$

porque $\beta_3 X(0) = 0$.

Si la categoría de sexo es "hombre", y por tanto $Z_j = Z_1 = 1$, el modelo quedaría

$$Y_{i1r} = \beta_0 + \beta_1 X_i + \beta_2 X_i + \varepsilon_{i1r}$$

porque $\beta_2(1) = \beta_2$, o reagrupando

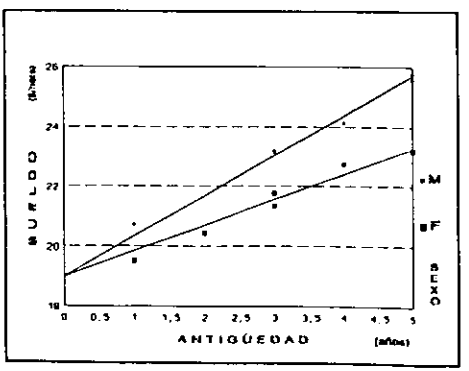
$$Y_{i1r} = \beta_0 + (\beta_1 + \beta_2)\ X_i + \varepsilon_{i1r}$$

De acuerdo con la salida del paquete estadístico (SYSTAT5) correspondiente al modelo anterior, podemos decir que, los datos de nuestro ejemplo aportan evidencia estadísticamente significativa de diferencia en la pendiente, ya que P = 0.000 (para ver si β_2 es diferente de cero) y considerando un $\alpha = 0.05$[3]. (La gráfica se elaboró con los coeficientes estimados reportados en la misma salida).

VARIABLE	COEF.	T	P(2 TAIL)
CONSTANT	19.00	93.4	0.00
AN	0.86	12.9	0.00
AN*SEXO	0.49 ← β_2	10.8 →	*0.00*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	2	20.89	289	0.00
RESID	9	0.07		



[3] Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

Resumiendo, puede verse que el modelo general estimado que involucra la especificación del sexo para la diferencia en pendiente

$$\hat{Y}_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i Z_j$$

una vez definido si se trata de mujeres u hombres, se transforma en

$$\hat{Y}_{i2} = \beta_0 + \underbrace{\beta_1 X_i}$$ para mujeres (j = 2)

y

$$\hat{Y}_{i1} = \beta_0 + \underbrace{(\beta_1+\beta_2)\ X_i}$$ para hombres (j = 1)

De las últimas dos expresiones se desprende que la diferencia entre ambas es la pendiente, señalada con $\underbrace{\ }$, y en ella la diferencia la da β_2, el coeficiente asociado originalmente al producto de "X" y "Z", la variable indicadora.

Esto es, β_2 representa el cambio estimado en la pendiente cuando se pasa de la categoría de sexo "mujer" (Z_2=0) a la categoría "hombre" (Z_1=1).

En nuestro ejemplo β_2 representa el cambio estimado en el aumento de sueldo dado por un año más de antigüedad al pasar en el modelo de la categoría de sexo "mujer" (Z_2=0) a la categoría "hombre" (Z_1=1), es decir, es una diferencia entre el aumento de sueldo que recibiría un hombre por un año más de antigüedad del aumento que recibiría una mujer.

De tal forma que si la hipótesis "H_0: $\beta_2 = 0$" no se rechaza (esto es si $\beta_2 = 0$), ambos modelos serían iguales; mientras que si la hipótesis "H_0: $\beta_2 = 0$" sí se rechaza (esto es si $\beta_2 \neq 0$) los modelos son diferentes en pendiente.

A continuación se presenta un resumen gráfico de esto mismo.

$$y_{i2r} = \beta_0 + \beta_1 x_i + \varepsilon_{i2r}$$

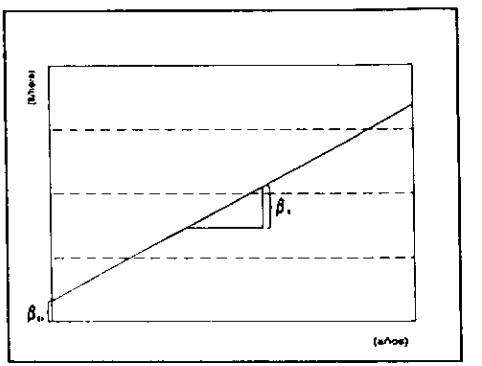
igual a

$$y_{i1r} = \beta_0 + (\beta_1+\beta_2)\, x_i + \varepsilon_{i1r}$$

$\Leftrightarrow \quad \beta_2 = 0$

N O R E C H A Z O

H_0: $\beta_2 = 0$



$$y_{i2r} = \beta_0 + \beta_1 x_i + \varepsilon_{i2r}$$

diferente de

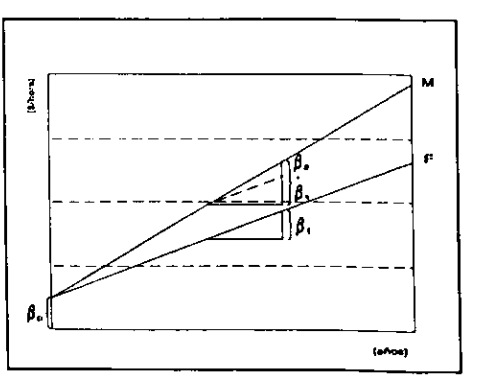
$$y_{i1r} = \beta_0 + (\beta_1+\beta_2)\, x_i + \varepsilon_{i1r}$$

$\Leftrightarrow \quad \beta_2 \neq 0$

S Í R E C H A Z O

H_0: $\beta_2 = 0$

(aquí $\beta_2 > 0$)



3. *Modelo para detectar simultáneamente diferencia en Ordenada al origen (Itercepto) y diferencia en Pendiente.*

De manera más general, en un solo modelo podemos incluir las especificaciones para diferencia en ordenada al origen (intercepto) y en pendiente de acuerdo con el sexo, es decir, la diferencia en el sueldo base (considerando que no hay antigüedad) entre hombres y mujeres (intercepto u ordenada al origen), y la diferencia entre hombres y mujeres en la variación del sueldo por antigüedad (pendiente).

Con el modelo que se revisó en el primer caso (diferencia en ordenada al origen-sueldo inicial, pag. 17) no se podía ver diferencia en pendiente, es decir, no se modelaba ni se probaba que hubiera diferencia entre hombres y mujeres en el aumento de sueldo dado por antiguedad.

Con el modelo que revisó en el segundo caso (diferencia en pendiente-variación del sueldo por antigüedad, pag. 21) no se podía ver diferencia en intercepto ("ordenada al origen"), es decir, no se modelaba ni se probaba que hubiera diferencia en el sueldo inicial (cuando recién se entra a trabajar) que recibiría una hombre del que recibiría inicialmente una mujer.

Para poder representar en un solo modelo lo que hemos hecho a través de dos modelos separados, es necesario emplear otro modelo que contemple ambas diferencias (en ordenada al origen y en pendiente), que es el que se considera a continuación.

Para representar en un modelo general la diferencia en intercepto (ordenada al origen) y pendiente entre los dos modelos particulares por sexo, el modelo a ajustar sería

$$SUELDO_{ir} = \beta_0 + \beta_1 ANTIGÜEDAD_i + \beta_2 SEXO_j + \beta_3 X_i SEXO_j + \varepsilon_{ijr}$$

ó

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 Z_j + \beta_3 X_i Z_j + \varepsilon_{ijr}$$

con

$i = 1,2,\ldots,5.$
$j = 1,2.$
$r = 1,2,\ldots,n_i.$

$$Z_j = \begin{cases} 1 \text{ si es hombre} & (j=1) \\ 0 \text{ si NO es hombre (es mujer)} & (j=2) \end{cases}$$

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

- i = i-ésimo valor de anigüedad considerado (desde 1, hasta 5).
- j = j-ésimo valor de la variable indicadora de sexo.
- r = r-ésima observación (repetición) del i-ésimo valor de antigüedad (desde 1, hasta n_i).
- Y_{ijr} = S U E L D O (en $) del profesor de sexo j, de la r-ésima observación (repetición) del i-ésimo valor de antigüedad.
- X_i = i-ésimo valor de años de A N T I G Ü E D A D.
- Z_j = j-ésimo valor de la variable indicadora de S E X O (Z_1=1, Z_2=0).
- ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de antigüedad y j-ésimo valor de SEXO.

Si la categoría de sexo es "mujer" (no es "hombre), y por tanto $Z_j = Z_2 = 0$, el modelo se reduciría a

$$Y_{i2r} = \beta_0 + \beta_1 X_i + \varepsilon_{i2r}$$

porque $\beta_2(0) = 0$ y $\beta_3 X(0) = 0$.

Si la categoría de sexo es "hombre", y por tanto $Z_j = Z_2 = 1$, el modelo quedaría

$$Y_{i1r} = \beta_0 + \beta_1 X_i + \beta_2 + \beta_3 X_i + \varepsilon_{i1r}$$

porque $\beta_2(1) = \beta_2$ y $\beta_3(1) = \beta_3$, o reagrupando

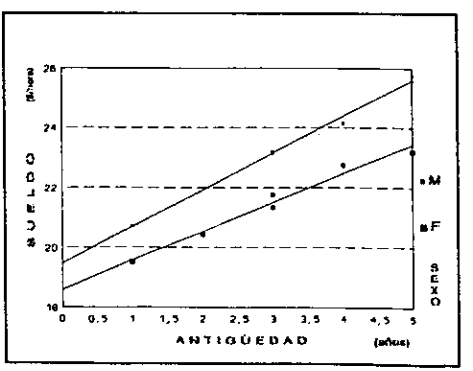
$$Y_{i1r} = (\beta_0+\beta_2) + (\beta_1+\beta_3)\ X_i + \varepsilon_{i1r}$$

De acuerdo con la salida del paquete estadístico (SYSTAT5) correspondiente al modelo anterior, podemos decir que, los datos de nuestro ejemplo aportan evidencia estadísticamente significativa de diferencia tanto en intercepto (ordenada al origen) como en pendiente, ya que P = 0.02 para ver si β_2 es diferente de cero, P = 0.02 para ver si β_3 es diferente de cero, y considerando un $\alpha = 0.05$[4]. (La gráfica se elaboró con los coeficientes estimados reportados en la misma salida).

VARIABLE	COEF.	T	P(2 TAIL)
CONSTANT	18.59	89.4	0.00
AN	0.97	15.2	0.00
SEXO	0.88←β_2	2.8	→ *0.02*
AN*SEXO	0.26←β_3	3.0	→ *0.02*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	3	14.04	346	0.00
RESID	8	0.04		



[4]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

Resumiendo, puede verse que el modelo general estimado que involucra la especificación del sexo para las diferencias en intercepto (ordenada al origen) y pendiente

$$\hat{Y}_{ij} = \beta_0 + \beta_1 X_i + \beta_2 Z_j + \beta_3 X_i Z_j$$

una vez definido si se trata de mujeres u hombres, se transforma en

$$\hat{Y}_{i2} = \underbrace{\beta_0} + \underbrace{\beta_1 X_i} \qquad \text{para mujeres} \quad (j = 2)$$

y

$$\hat{Y}_{i1} = \underbrace{(\beta_0+\beta_2)} + \underbrace{(\beta_1+\beta_3)\ X_i} \qquad \text{para hombres} \quad (j = 1)$$

De las últimas dos expresiones se desprende que las diferencias entre ambas son la ordenada al origen (intercepto), señalada con $\underbrace{\ }$, y la pendiente, señalada con $\underbrace{\ }$, diferencias dadas respectivamente por β_2, el coeficiente asociado originalmente a la variable indicadora sola, y β_3, el coeficiente asociado originalmente al producto de "X" y la variable indicadora; esto es, coeficientes de términos asociados a la variable "Z" indicadora de las categorías del sexo.

Así, mientras que β_2 representa el cambio estimado en la ordenada al origen (intercepto) cuando se pasa de la categoría de sexo "mujer" (Z_2=0) a la categoría "hombres" (Z_1=1), β_3 representa el cambio estimado en la pendiente cuando se pasa de la categoría de sexo "mujer" (Z_2=0) a la categoría "hombre" (Z_1=1).

En nuestro ejemplo β_2 representa el cambio estimado en el sueldo inicial (cuando recién se entra a trabajar) al pasar en el modelo de la categoría de sexo "mujer" (Z_2=0) a la categoría "hombre" (Z_2=1), es decir, es una diferencia entre el sueldo inicial que recibiría un hombre del que recibiría una mujer.

Por otro lado, β_3 representa el cambio estimado en el aumento de sueldo dado por un año más de antigüedad al pasar en el modelo de la categoría de sexo "mujer" (Z_2=0) a la categoría "hombre" (Z_1=1), es decir, es una diferencia entre el aumento de sueldo que recibiría un hombre por un año más de antigüedad del aumento que recibiría una mujer.

De tal forma que cuando la hipótesis "doble" "H_0: $\beta_2 = 0$, $\beta_3 = 0$" no se rechaza (esto es si $\beta_2 = 0$ y $\beta_3 = 0$), ambos modelos serían iguales; si se rechaza H_0 sólo porque $\beta_2 \neq 0$, ambos modelos son diferentes sólo en intercepto (ordenada al origen); si se recha H_0 sólo porque $\beta_3 \neq 0$, ambos modelos son diferentes sólo en pendiente; mientras que si H_0 se rechaza porque $\beta_2 \neq 0$ y $\beta_3 \neq 0$, entonces ambos modelos son diferentes tanto en intercepto (ordenada al origen) como en pendiente.

A continuación se presenta un resumen gráfico de esto mismo.

$$y_{i2r} = \beta_0 + \beta_1 X_i + \varepsilon_{i2r}$$

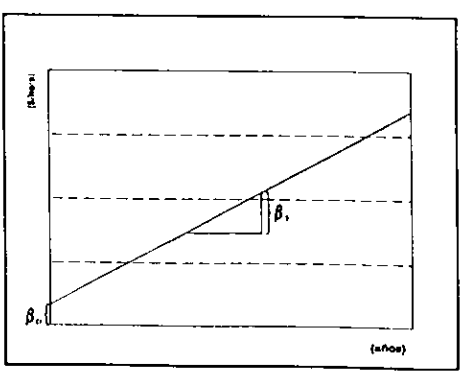
igual a

$$y_{i1r} = (\beta_0+\beta_2) + (\beta_1+\beta_3) X_i + \varepsilon_{i1r}$$

$\Leftrightarrow \quad \beta_2 = 0 \quad y \quad \beta_3 = 0$

N O R E C H A Z O

H_0: $\beta_2 = 0$, $\beta_3 = 0$



$$y_{i2r} = \beta_0 + \beta_1 x_i + \varepsilon_{i2r}$$

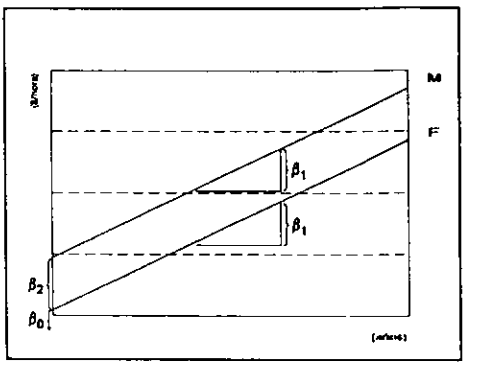
diferente de

$$y_{i1r} = (\beta_0+\beta_2) + (\beta_1+\beta_3)\, x_i + \varepsilon_{i1r} \quad \Leftrightarrow \quad \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0$$

R E C H A Z O S Ó L O

H_0: $\beta_2 = 0$

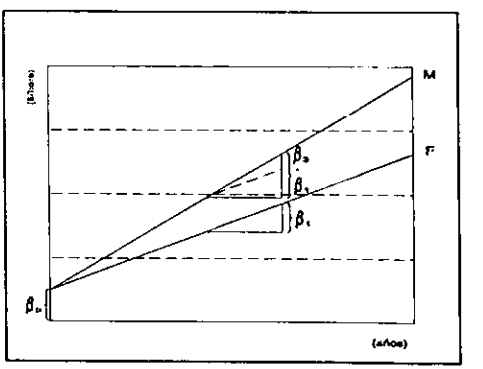
(aquí→ $\beta_2 > 0$)



R E C H A Z O S Ó L O

H_0: $\beta_3 = 0$

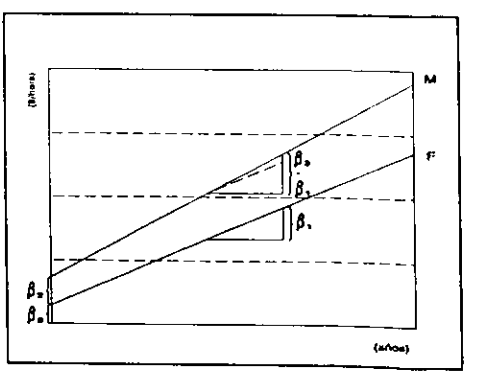
(aquí→ $\beta_3 > 0$)



S Í R E C H A Z O

H_0: $\beta_2 = 0$ y $\beta_3 = 0$

(aquí→ $\beta_2 > 0$) $\beta_3 > 0$)



Cuando se declare un modelo en la computadora para el análisis de los datos, es importante determinar el modelo de interés, es decir, qué hipótesis nos interesa probar. De esa forma, cuando obtengamos "la salida de la corrida" sabremos cómo interpretar cada resultado, y a la vez no faltará alguna especificación en el modelo que impida tener resultados sobre alguna hipótesis o comportamiento de interés.

En nuestro ejemplo se vió la significancia de los coeficientes asociados a la ordenada al origen o intercepto (β_0), a la pendiente (β_1), a la variación en la ordenada al origen (intercepto) por cambio en categoría de sexo (β_2) y a la variación en la pendiente por el cambio en categoría de sexo (β_3).

Cuando, en la declaración del modelo al paquete estadístico, se omite algún término asociado a un comportamiento o hipótesis de interés, no se podrán obtener conclusiones sobre lo no declarado. Por otro lado, cuando sean declarados los términos, si resultasen significativos, no podrían interpretarse tales resultados si no somos capaces de asociar cada coeficiente a un efecto o característica particular en el contexto del área del problema que se trata de modelar.

III. COMPARACIÓN DE MODELOS DE REGRESIÓN DE GRADO DOS ENTRE DOS GRUPOS

Estimado lector:

Antes de entrar en tema, es recomendable que lea primero el prefacio, donde no solamente se explica el contenido de este trabajo, sino también el formato que se ha seguido en su presentación. Tales explicaciones le permitirán una mayor comprensión del documento, así como una lectura más fluída.

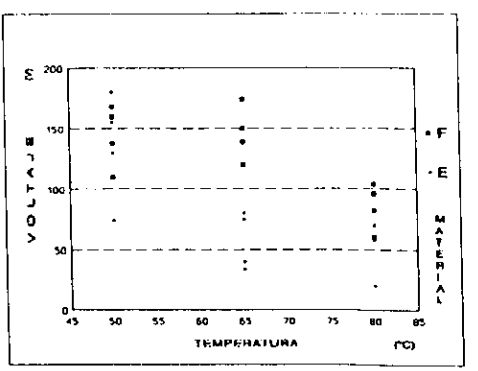
Además, es conveniente que revise la sección I "NECESIDAD DEL CONOCIMIENTO DE LA HERRAMIENTA ESTADÍSTICA PARA SU USO ADECUADO", donde se exponen conceptos generales, se señalan resultados importantes generados por el paquete estadístico, y se explica su interpretación, lo cual será de gran utilidad para entender el desarrollo del resto de este trabajo. A su vez, en la sección II, se introduce el concepto de las variables indicadoras o "dummy".

Si usted ya leyó el prefacio y la sección I, lo invito a continuar la lectura de esta página.

No todos los fenómenos nos dan igual tipo de información, ni pueden ser descritos de la misma forma. Análogamente, no todos los datos pueden analizarse o modelarse igual.

Por ejemplo, los siguientes datos representan la salida de voltaje de dos tipos de pilas, diferentes en el tipo de material, e instaladas en lugares a tres diferentes temperaturas.

		Voltaje					
		Temperatura (°C)					
		50		65		80	
MATERIAL	E	130	155	34	40	20	70
		74	180	80	75	82	58
	F	138	110	174	120	96	104
		168	160	150	139	82	60



A tales datos se les podría ajustar un modelo de regresión simple (una recta), con la temperatura como variable explicativa (X) y la salida de voltaje como variable respuesta (Y).

$$\text{VOLTAJE}_{ir} = \beta_0 + \beta_1 \text{ TEMPERATURA}_i + \varepsilon_{ir}$$

ó

$$Y_{ir} = \beta_0 + \beta_1 X_i + \varepsilon_{ir}$$

con

$i = 1, 2, 3.$
$r = 1, 2, 3, 4.$

$\varepsilon_{ir} \sim NIID(0, \sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada valor de i de X)

donde

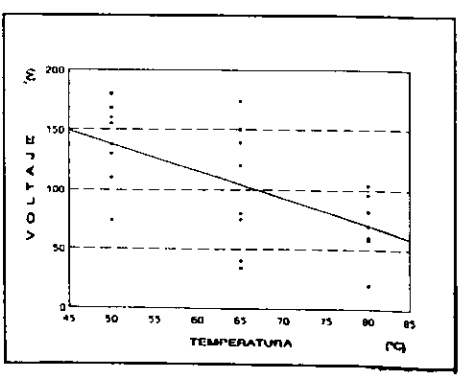
- i = i-ésimo valor de temperatura considerado (desde 1, hasta 3).
- r = r-ésima observación (repetición) del i-ésimo valor de temperatura (desde 1, hasta 4).
- Y_{ir} = VOLTAJE de la pila de la r-ésima observación (repetición) del i-ésimo valor de temperatura.
- X_i = i-ésimo valor de TEMPERATURA.
- ε_{ir} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de temperatura.

Por lo tanto, β_1 (pendiente) representa la variación en la salida del voltaje (Y) por cada grado centígrado (°C) que varíe la temperatura (X) a la que se encuentre la pila. Por otro lado, aun cuando la fórmula de la parábola así lo sugiere, no es conveniente interpretar β_0 (la ordenada al origen o intercepto) como la salida de voltaje (Y) de la pila a temperatutra (X) de 0°C, pues los datos de voltaje y temperatura, que son la base del cálculo de la parábola, fueron tomados muy lejos de los 0°C y no sabemos cómo son a esa temperatura y si correspondieran a otro modelo. Por lo tanto, no tiene interpretación dentro de nuestro problema (en el rango que conocemos) y sólo es un dato en la fórmula.

VARIABLE	COEF.	T	P(2 TAIL)
CONSTANT	251.2	5.9	0.00
TEM	-2.3 $\leftarrow \beta_1$	-3.5	→ *0.00*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	1	18428	12.6	0.002
RESID	22	1461		



La salida del paquete estadístico SYSTAT5 para el modelo anterior, nos dice que podemos considerar que los datos de nuestro ejemplo aportan evidencia estadísticamente significativa de que sí hay relación entre la temperatuta y el voltaje, ya que P = 0.000 (para ver si β_1 es diferente de cero) y considerando un $\alpha = 0.05$[5], y que tal relación es negativa ($\beta_1 = -2.3$), es decir, que el voltaje de salida de la pila decrece conforme aumenta la temperatura de donde se coloca la misma. (La gráfica se elaboró con los coeficientes estimados reportados en la misma salida).

Como en la sección anterior se hizo con relación al sexo, a los datos anteriores (pag. 31) se les podría ajustar un modelo de regresión que incluyera la diferencia en intercepto (ordenada al origen) y pendiente (componente lineal) dada por el tipo de material. Dado que sólo son dos categorías de material, únicamente se necesita una variable indicadora.

$$VOLTAJE_{ijr} = \beta_0 + \beta_1 TEM_i + \beta_2 MAT_j + \beta_3 TEM_i MAT_j + \varepsilon_{ijr}$$

ó

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 Z_j + \beta_3 X_i Z_j + \varepsilon_{ijr}$$

con

$i = 1,2,3.$
$j = 1,2.$
$r = 1,2,3,4.$

$$Z_j = \begin{cases} 1 \text{ si es material F} & (j=1) \\ 0 \text{ si NO es material F (es material E)} & (j=2) \end{cases}$$

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

- i = i-ésimo valor de temperatura considerado (desde 1, hasta 3).
- j = j-ésimo valor de la variable indicadora de material.
- r = r-ésima observación (repetición) del i-ésimo valor de temperatura (desde 1, hasta 4).
- Y_{ijr} = VOLTAJE de la pila de material j, de la r-ésima observación (repetición) del i-ésimo valor de temperatura.
- X_i = i-ésimo valor de TEMPERATURA.
- Z_j = j-ésimo valor de la variable indicadora de MATERIAL (Z_1=1, Z_2=0).
- ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de temperatura y j-ésimo material.

Aquí β_3 (diferencia en pendiente de las rectas para cada material) representa la diferencia, dada por el material de la pila, en la variación de la salida del voltaje por cada grado centígrado (°C) que aumente la temperatura a la que se encuentra la pila. Por su parte, β_2 es tan solo la diferencia en ordenada al origen o intercepto de las rectas para cada material y no debe interpretarse como "la diferencia en la salida de voltaje (Y) de la pila a temperatutra (X) de 0°C debida a la diferencia en material de la pila", por las razones expuestas en la página 30 (el rango del estudio y de la colecta de datos).

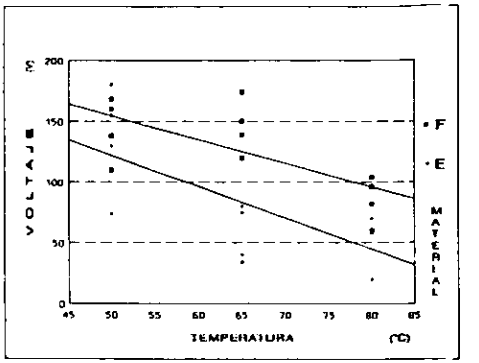
---

[5]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

VARIABLE	COEF.	T	P(2 TAIL)
CONSTANT	250.5	4.9	0.00
TEM	$-2.6 \leftarrow \beta_1$	-3.4	→ *0.00*
MAT	$\beta_2 \leftarrow 1.3$	0.0	→ *0.99*
TEM*MAT	$0.6 \leftarrow \beta_3$	0.6	→ *0.57*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	3	9773.9	9.2	0.00
RESID	20	1063.0		



La salida del paquete estadístico SYSTAT5 para el modelo anterior, que involucra las diferencias por material, nos dice que los datos no aportan suficiente evidencia para considerar que haya diferencia estadísticamente significativa (considerando un $\alpha = 0.05$) en la relación entre la temperatuta y el voltaje para cada tipo de material (β_2 y β_3), es decir, no podemos decir que entre las rectas que relacionan la temperatura y el voltaje para cada tipo de material haya diferencia en intercepto (ordenada al origen, β_2) o pendiente (β_3), (con valores de P = 0.99 y P =0.57 respectivamente[6]).

Con los resultados de ajustar el segundo modelo no se modifica la descripción del fenómeno que habíamos logrado a través del primer modelo, es decir, que hay una relación entre la temperatura y el voltaje ($\beta_1 \neq 0$)[7]. Sin embargo, mientras que con el primero conocemos nada sobre la diferencia de tal relación con respecto al material de la pila (no se hacen hipótesis con respecto a β_2 y a β_3), con el segundo ya conocemos algo sobre tal diferencia: con un $\alpha = 0.05$, no se puede considerar que sí la haya (no se rechaza la hipótesis nula de "H_0: $\beta_2 = 0$, $\beta_3 = 0$").

Ahora que hemos concluido que los datos sí apoyan la existencia de una relación lineal entre la temperatura y el voltaje, sería deseable saber si la tendencia en la disminución del voltaje debida al aumento de la temperatura se puede describir con una curva cuadrática en vez de una recta. Para ello es necesario

[6]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

[7]**No siempre es así**, ya que la introducción de nuevos términos al modelo modifica la estimación de los coeficientes asociados a los términos previamente declarados (ver pags. 34 y 45).

incluir la variable explicativa cuantitativa (temperatura) al cuadrado (X_i^2).

El modelo para describir la situación anterior sería el de la fórmula de una parábola más el término aleatorio (ε_{ir}), es decir

$$Y_{ir} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_{ir}$$

con

$i = 1, 2, 3.$
$r = 1, 2, 3, 4.$

ε_{ir}-NIID(0,σ^2) (Normales, Independientes e Idénticamente Distribuidos) (los errores tienen distribución Normal en cada valor de i de X)

donde

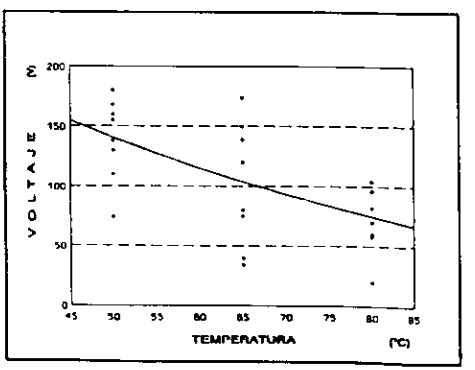
- i = i-ésimo valor de temperatura considerado (desde 1, hasta 3).
- r = r-ésima observación (repetición) del i-ésimo valor de temperatura (desde 1, hasta 4).
- Y_{ir} = VOLTAJE de la pila de la r-ésima observación (repetición) del i-ésimo valor de temperatura.
- X_i = i-ésimo valor de TEMPERATURA.
- X_i^2 = i-ésimo valor de TEMPERATURA al cuadrado.
- ε_{ir} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de temperatura.

En el modelo anterior $\beta_1 X_i$ es el término que describe el "componente lineal" y $\beta_2 X_i^2$ es el término que describe el "componente cuadrático" de la tendencia.

VARIABLE	COEF.	T	P(2 TAIL)
CONSTANT	322.50	1.04	0.31
TEM	-4.54 ← β_1	-0.46	→ *0.65*
TEM*TEM	0.02 ← β_2	0.23	→ *0.82*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	2	9255	6.06	0.01
RESID	21	1527		



Estos resultados no aportan suficiente evidencia para considerar que la relación entre el voltaje de una pila (Y) y la

temperatura (X) a la que se encuentra estén relacionados a través de una curva cuadrática, pues no se rechazó la hipótesis de que el componente cuadrático fuese cero (H_0: $\beta_2 = 0$), ya que P = 0.82 (para ver si β_2 es diferente de cero) y considerando un $\alpha = 0.05$. Lo curioso es que ¡tampoco se rechazó la hipótesis de que el componente lineal fuese cero (H_0: $\beta_1 = 0$)!, ya que P = 0.65 (para ver si β_2 es diferente de cero) y considerando un $\alpha = 0.05$[8], a pesar de que con ambos modelos que sólo consideran el comportamiento lineal ¡sí se vió la existencia de tal relación!

Esto último nos permite ver que el resultado del análisis variará no sólo porque se pueda o no inferir a partir de términos no declarados en el modelo (por ejemplo la diferencia por material), sino que incluso lo que con un modelo parece ser de una manera, ¡con otro modelo puede resultar ser diferente! Tal es el caso de la relación entre la variación del voltaje y la variación lineal de la temperatura, que parece existir según los primeros modelos vistos (de una y de dos rectas), pero no parece existir según el último modelo (componente lineal del modelo de la curva cuadrática).

De aquí se desprende que la significancia estadística de los coeficientes (β's) asociados a cada variable, es decir, la capacidad de explicación de una variable en un modelo se debe ver e interpretar en presencia (o ausencia, según sea el caso) de las demás variables involucradas en el modelo).

En nuestro ejemplo se ve que la significancia del coeficiente β_1 que describe la relación de aumento lineal del voltaje (Y) con respecto al aumento en la temperatura (X) varía de significarivo, cuando sólo ajustamos tal comportamiento (con o sin material) con los dos primeros modelos de esta sección (pag. 30 y 31), a no significativo, cuando se ajusta a la vez el comportamiento lineal y el comportamiento cuadrático (curvo) con el último modelo (pag.33).

Por lo anterior, podríamos preguntarnos si al ajustar un modelo que describiera la relación voltaje/temperatura como una curva parabólica para cada material entonces sí encontraríamos relación entre el voltaje y la temperatura de la pila.

---

[8]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

1. *Modelo Completo.*

Recordemos que en el caso del ejemplo del sueldo de los profesores, era de interés saber si existía diferencia en el sueldo base (considerando que no hay antiguedad) entre hombres y mujeres (intercepto u ordenada al origen), y la diferencia entre hombres y mujeres en la variación del sueldo por antigüedad (pendiente).

Recordemos también que esto lo podíamos hacer viendo si había diferencia en ordenada al origen y en pendiente, y que tal diferencia se modelaba a través de agregar nuevos términos de ordenada al origen y pendiente asociados a una variable indicadora del sexo, la variable "Z".

En este nuevo ejemplo para ver si la relación entre el voltaje y la temperatura a la que se encuentra la pila es diferente para distintos materiales, se procederá de la misma manera: se incluirán nuevos términos de ordenada al origen, componente lineal y componente cuadrático asociados a una variable indicadora de material, la variable "Z".

Para construir un modelo único que considere a la vez ambos materiales, recordemos que hay que asociar a cada término no aleatorio del modelo original la variable indicadora de tipo de material (que hemos representado o denotado "Z" para indicar que es o no el material "F"), y agregarlos al modelo arriba mostrado (pag. 35), el cual no consideraba tipo de material, dando como resultado el modelo "completo"

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_j + \beta_4 X_i Z_j + \beta_5 X_i^2 Z_j + \varepsilon_{ijr}$$

con

$i = 1,2,3.$
$j = 1,2.$
$r = 1,2,3,4.$

$$Z_j = \begin{cases} 1 \text{ si es material F} & (j=1) \\ 1 \text{ si NO es material F (es material E)} & (j=2) \end{cases}$$

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Indepedientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

- i = i-ésimo valor de temperatura considerado (desde 1, hasta 3).
- j = j-ésimo valor de la variable indicadora de material.
- r = r-ésima observación (repetición) del i-ésimo valor de temperatura (desde 1, hasta 4).
- Y_{ijr} = VOLTAJE de la pila de material j, de la r-ésima observación (repetición) del i-ésimo valor de temperatura.
- X_i = i-ésimo valor de TEMPERATURA.
- X_i^2 = i-ésimo valor de TEMPERATURA al cuadrado.
- Z_j = j-ésimo valor de la variable indicadora de MATERIAL (Z_1=1, Z_2=0).
- ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de temperatura y j-ésimo material.

Si se tratara del material "E" (no es el material "F"), y por tanto $Z_j = Z_2 = 0$, el modelo se reduciría a

$$Y_{i2r} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_{i2r}$$

porque $\beta_3(0) = 0$, $\beta_4(0) = 0$ y $\beta_5(0) = 0$.

Si se tratara del material "F", y por tanto $Z_j = Z_1 = 1$, el modelo quedaría

$$Y_{i1r} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 + \beta_4 X_i + \beta_5 X_i^2 + \varepsilon_{i1r}$$

porque $\beta_3(1) = \beta_3$, $\beta_4(1) = \beta_4$ y $\beta_5(1) = \beta_5$, o reagrupando

$$Y_{i1r} = (\beta_0+\beta_3) + (\beta_1+\beta_4)\ X_i + (\beta_3+\beta_5)\ X_i^2 + \varepsilon_{i1r}$$

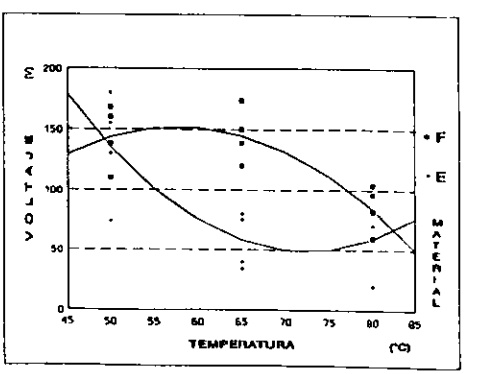
A continuación se muestra parte de la salida del paquete estadístico SYSTAT5 para el modelo anterior, donde puede verse claramente que los coeficientes (β's) asociados a todos los términos del modelo "completo" son significativamente diferentes de cero (de acuerdo con los valores de "P" y considerando un $\alpha = 0.05$[9]).

Esto es, podemos considerar que los datos de nuestro ejemplo aportan evidencia estadísticamente significativa de que sí hay relación entre la temperatuta y el voltaje, y que tal relación es diferente (en intercepto u ordenada al origen, componente lineal y componente cuadrático - β_3, β_4 y β_5 respectivamente) para cada tipo de material (con un $\alpha = 0.05$)[9].

VARIABLE	COEF.	T	P(2 TAIL)
CONSTANT	954.6 ← β_0	2.99	→ *0.01*
TEM	-25.0 ← β_1	-2.47	→ *0.02*
TEM*TEM	0.2 ← β_2	2.22	→ *0.04*
MAT	-1264.2 ← β_3	-2.80	→ *0.01*
TEM*MAT	41.0 ← β_4	2.86	→ *0.01*
TEM*TEM*MAT	-0.3 ← β_5	-2.83	→ *0.01*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	5	7183	8.8	0.00
RESID	18	815		



[9]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

Resumiendo, puede verse que el modelo general estimado que involucra la especificación del material para las diferencias en ordenada al origen (intercepto), componente lineal y componente cuadrático

$$\hat{Y}_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_j + \beta_4 X_i Z_j + \beta_5 X_i^2 Z_j$$

una vez definido si se trata del material "E" o "F", se transforma en

$$\hat{Y}_{i2} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \qquad \text{para el material "E"} \quad (j = 2)$$

y

$$\hat{Y}_{i1} = (\beta_0+\beta_3) + (\beta_1+\beta_4)\, X_i + (\beta_2+\beta_5)\, X_i^2 \qquad \text{para el "F"} \quad (j = 1)$$

De esa manera tenemos modelos de regresión de grado 2 (parábola) para los dos grupos en una sola expresión, y la diferencia entre las dos formas la dan β_3 (diferencia en el coeficiente del intercepto u ordenada al origen, señalado con ⌴), β_4 (diferencia en el coeficiente del componente lineal, señalado con ⌴), y β_5 (diferencia en el coeficiente del componente cuadrático, señalado con ⌴), todas estas β's coeficientes asociados originalmente a términos en "Z", la variable indicadora de la categoría del material.

Dicho de otro modo, β_3 representa el cambio en la ordenada al origen (intercepto) cuando se pasa de la categoría del material "E" (Z_2=0) al "F" (Z_1=1), β_4 representa el cambio en el componente lineal cuando se pasa del material "E" (Z_2=0) al "F" (Z_1=1), y β_5 representa el cambio en el componente cuadrático cuando se pasa del material "E" (Z_2=0) al "F" (Z_1=1).

De tal forma que cuando la hipótesis "H_0: $\beta_3 = \beta_4 = \beta_5 = 0$" no se rechaza (esto es $\beta_3 = 0$, $\beta_4 = 0$ y $\beta_5 = 0$ - todas iguales a cero), ambos modelos serían iguales; si se rechaza H_0 sólo porque $\beta_3 \neq 0$, ambos modelos son diferentes sólo en intercepto (ordenada al origen); si se rechaza H_0 sólo porque $\beta_4 \neq 0$, ambos modelos son diferentes sólo en componente lineal; si se rechaza H_0 sólo porque $\beta_5 \neq 0$, ambos modelos son diferentes sólo en componente cuadrático; mientras que si se rechaza H_0 porque $\beta_3 \neq 0$, $\beta_4 \neq 0$ y $\beta_5 \neq 0$ (todas diferentes de cero), entonces ambos modelos son diferentes tanto en intercepto (ordenada al origen) como en componentes lineal y cuadrático.

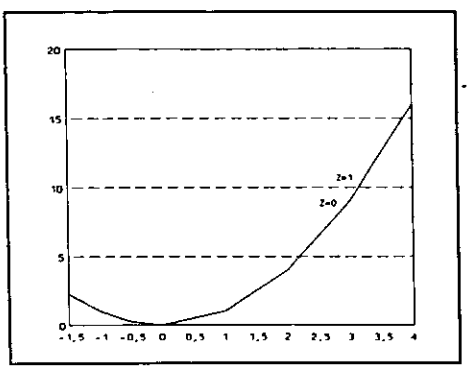
Para que esto quede más claro, se muestran las gráficas de lo mencionado en el párrafo anterior, pero con datos diferentes a nuestro ejemplo (para simplificar), acompañadas de las ecuación resultantes para el cada grupo, es decir, cuando "Z = 0" y cuando "Z = 1".

MODELO COMPLETO

$$\hat{Y}_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_j + \beta_4 X_i Z_j + \beta_5 X_i^2 Z_j$$

$Z = 0$ $\hat{Y}_{i2} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$

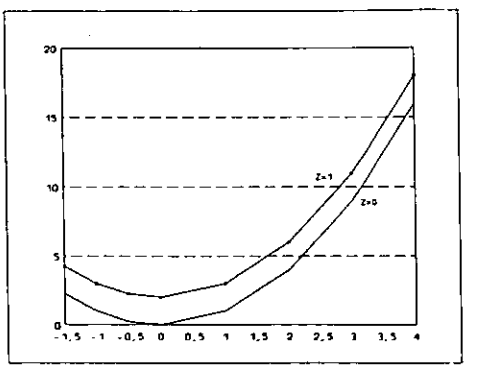
$Z = 1$ $\hat{Y}_{i1} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$



Ninguna Diferencia

$Z = 0$ $\hat{Y}_{i2} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$

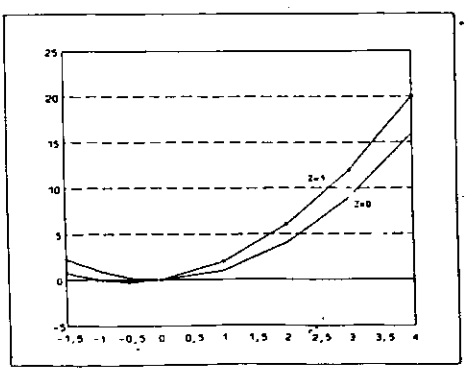
$Z = 1$ $\hat{Y}_{i1} = (\beta_0 + \beta_3) + \beta_1 X_i + \beta_2 X_i^2$



Diferencia sólo en Intercepto

$Z = 0$ $\hat{Y}_{i2} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$

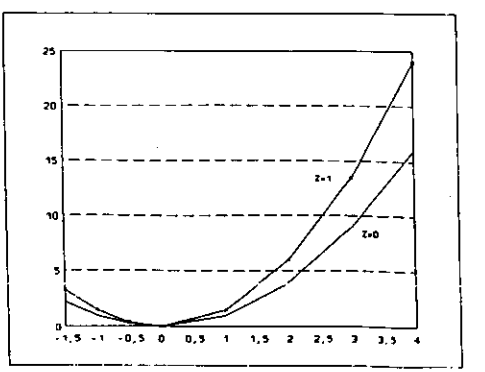
$Z = 1$ $\hat{Y}_{i1} = \beta_0 + (\beta_1 + \beta_4) X_i + \beta_2 X^2$



Diferencia sólo en Componente Lineal

$Z = 0$ $\hat{Y}_{i2} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$

$Z = 1$ $\hat{Y}_{i1} = \beta_0 + \beta_1 X_i + (\beta_2 + \beta_5) X_i^2$



Diferencia sólo en Componente Cuadrático

De esta forma no sólo puede describirse el comportamiento para cada grupo, sino que pueden hacerse comparaciones entre los dos grupos en cuanto a intercepto (ordenada al origen) y componentes lineal y cuadrático, pudiéndose probar diferencias en cualquiera de los términos mencionados: intercepto (ordenada al origen) y componentes lineal y cuadrático.

De nuevo, vale la pena mencionar que cuando se declare un modelo en la computadora para el análisis de los datos, es muy importante saber qué modelo es el de interés, es decir, qué hipótesis nos interesa probar.

En nuestro ejemplo es de interés probar la existencia de efecto lineal y cuadrático en la tendencia de la respuesta del voltaje con relación a la temperatura. De igual manera es de interés saber si tal tendencia es igual o diferente para los dos tipos de material.

Si no agregamos al modelo los términos donde asociamos la variable indicadora a los términos que representan los componentes lineal y cuadático de la respuesta, no se podrá ver si tales componentes son diferentes dependiendo del material. De igual manera, si no se incluyen los términos cuadrático y cuadrático con variable indicadora, no se podrá ver el efecto de la variación cuadrática de la variable explicativa en la respuesta para cada material.

Aún más, el no incluir términos de mayor orden (un término cuadrático es de mayor orden que uno lineal, uno cúbico es de mayor orden que uno cuadrático), puede llevarnos a diferentes conclusiones sobre los términos de menor grado que si los incluyéramos.

Además de lo ya visto, tal comportamiento es más evidente si al modelo de regresión (una parábola en este caso) vamos agregando uno a uno los términos para la diferenciación de material ("otros modelos" mostrados a continuación) hasta llegar al modelo "completo" (el que se acaba de analizar).

2. Otros Modelos.

- Modelo para detectar diferencia en ordenada al origen (intercepto).

Para representar una diferencia en intercepto (ordenada al origen) entre los dos modelos particulartes por material, el modelo general a ajustar sería

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_j + \varepsilon_{ijr}$$

con

$i = 1,2,3.$
$j = 1,2.$
$r = 1,2,3,4.$

$$Z_j = \begin{cases} 1 \text{ si es material F} & (j=1) \\ 0 \text{ si NO es material F (es material E)} & (j=2) \end{cases}$$

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Indepedientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

- i = i-ésimo valor de temperatura considerado (desde 1, hasta 3).
- j = j-ésimo valor de la variable indicadora de material.
- r = r-ésima observación (repetición) del i-ésimo valor de temperatura (desde 1, hasta 4).
- Y_{ijr} = VOLTAJE de la pila de material j, de la r-ésima observación (repetición) del i-ésimo valor de temperatura.
- X_i = i-ésimo valor de TEMPERATURA.
- X_i^2 = i-ésimo valor de TEMPERATURA al cuadrado.
- Z_j = j-ésimo valor de la variable indicadora de MATERIAL (Z_1=1, Z_2=0).
- ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de temperatura y j-ésimo valor de material.

Si se tratara del material "E" (no es el material "F"), y por tanto $Z_j = Z_2 = 0$, el modelo se reduciría a

$$Y_{i2r} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_{i2r}$$

porque $\beta_3(0) = 0$.

Si se tratara del material "F", y por tanto $Z_j = Z_1 = 1$, el modelo quedaría

$$Y_{i1r} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 + \varepsilon_{i1r}$$

porque $\beta_3(1) = \beta_3$, o reagrupando

$$Y_{i1r} = (\beta_0+\beta_3) + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_{i1r}$$

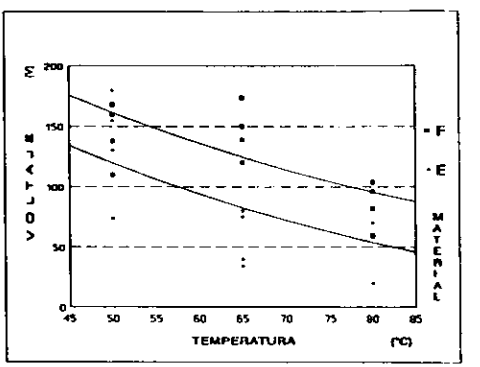
De aquí que la diferencia en ordenada al origen (intercepto) esté representada por β_3.

La salida del paquete estadístico SYSTAT5 para el modelo anterior, nos dice que podemos considerar que los datos de nuestro ejemplo sólo aportan evidencia estadísticamente significativa de que haya diferencia en la salida media de voltaje entre las pilas de un material y las del otro material, ya que P = 0.01 (para ver si β_3 es diferente de cero) y considerando un $\alpha = 0.05$[10].

VARIABLE	COEF.	T	P(2 TAIL)
CONSTANT	301.5	1.16	0.26
TEM	-4.5 ← β_1	-0.55 →	*0.59*
TEM*TEM	0.01 ← β_2	0.28 →	*0.79*
MAT	β_3 → 41.9	3.13 →	*0.01*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	3	9684	9.0	0.00
RESID	20	1076		



Los resultados no aportan suficiente evidencia para considerar que la relación entre el voltaje de las pilas (Y) y la temperatura (X) a la que se encuentran estén relacionados a través de una curva cuadrática, pues no se rechazó la hipótesis de que el componente cuadrático fuese cero (H_0: $\beta_2 = 0$), ya que P = 0.79 y considerando un $\alpha = 0.05$; ni tampoco se rechazó la hipótesis de que el componente lineal fuese cero (H_0: $\beta_1 = 0$), ya que P = 0.59 y considerando un $\alpha = 0.05$[10]. (La gráfica se elaboró con los coeficientes estimados reportados en la misma salida).

Al igual que lo visto en el modelo cuadrático sin distinción de material (pag. 35 y 36), no se aprecia relación (ni lineal ni al cuadrado) entre el voltaje y la variación de temperatura.

[10] Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

- Modelo para detectar diferencia en ordenada al origen (intercepto) y en componente lineal.

Para representar una diferencia en intercepto (ordenada al origen) y en componente lineal entre los dos modelos particulartes por material, el modelo general a ajustar sería

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_j + \beta_4 X_i Z_j + \varepsilon_{ijr}$$

con

$i = 1,2,3.$
$j = 1,2.$
$r = 1,2,3,4.$

$$Z_j = \begin{cases} 1 \text{ si es material F} & (j=1) \\ 2 \text{ si NO es material F (es material E)} & (j=2) \end{cases}$$

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Indepenidientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

- i = i-ésimo valor de temperatura considerado (desde 1, hasta 3).
- j = j-ésimo valor de la variable indicadora de material.
- r = r-ésima observación (repetición) del i-ésimo valor de temperatura (desde 1, hasta 4).
- Y_{ijr} = VOLTAJE de la pila de material j, de la r-ésima observación (repetición) del i-ésimo valor de temperatura.
- X_i = i-ésimo valor de TEMPERATURA.
- X_i^2 = i-ésimo valor de TEMPERATURA al cuadrado.
- Z_j = j-ésimo valor de la variable indicadora de MATERIAL (Z_1=1, Z_2=0).
- ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de temperatura y j-ésimo valor de material.

Si se tratara del material "E" (no es el material "F"), y por tanto $Z_j = Z_2 = 0$, el modelo se reduciría a

$$Y_{i2r} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_{i2r}$$

porque $\beta_3(0) = 0$ y $\beta_4(0) = 0$.

Si se tratara del material "F", y por tanto $Z_j = Z_1 = 1$, el modelo quedaría

$$Y_{i1r} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 + \beta_4 X_i + \varepsilon_{i1r}$$

porque $\beta_3(1) = \beta_3$ y $\beta_4(1) = \beta_4$, o reagrupando

$$Y_{i1r} = (\beta_0+\beta_3) + (\beta_1+\beta_4) X_i + \beta_3 X_i + \varepsilon_{i1r}$$

De aquí que la diferencia en ordenada al origen (intercepto) esté representada por β_3 y la diferencia en componente lineal por β_4.

La salida del paquete estadístico SYSTAT5 para el modelo anterior, nos dice que los datos de nuestro ejemplo no aportan suficiente evidencia de que haya diferencia en la salida media de voltaje entre las pilas de un material y las del otro material (P = 0.99 para ver si β_3 es diferente de cero y considerando un $\alpha = 0.05$[11]), ¡lo que sí se podía decir con el modelo que sólo involucraba el término para diferencia en ordenada al origen o intercepto! (pag. 43).

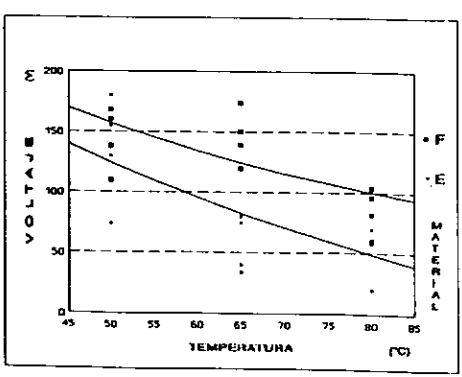
La salida del paquete también revela que los datos no aportan suficiente evidencia de que haya diferencia en el componente lineal para cada tipo de material (P = 0.58 para β_4, con un $\alpha = 0.05$[11]).

De igual manera, los resultados no aportan suficiente evidencia para considerar que la relación entre el voltaje de las pilas (Y) y la temperatura (X) a la que se encuentran estén relacionados a través de una curva cuadrática, pues no se rechazó la hipótesis de que el componente cuadrático fuese cero (H_0: $\beta_2 = 0$), ya que P = 0.79 y considerando un $\alpha = 0.05$; ni tampoco se rechazó la hipótesis de que el componente lineal fuese cero (H_0: $\beta_1 = 0$), ya que P = 0.59 y considerando un $\alpha = 0.05$[11]. (La gráfica se elaboró con los coeficientes estimados reportados en la misma salida).

VARIABLE	COEF	T	P(2 TAIL)
CONSTANT	321.9	1.21	0.24
TEM	$-4.9 \leftarrow \beta_1$	-0.58 →	*0.57*
TEM*TEM	$0.01 \leftarrow \beta_2$	0.27 →	*0.79*
MAT	$1.3 \leftarrow \beta_3$	0.02 →	*0.99*
TEM*MAT	$0.6 \leftarrow \beta_4$	0.56 →	*0.58*

ANALYSIS OF VARIANCE

SOURCE	DF	MEAN-SQ	F	P
REGRES	4	7351	6.60	0.00
RESID	19	1115		



[11]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

Como se vio en el modelo cuadrático sin distinción de material (pag. 35 y 36), no se aprecia relación entre el voltaje y la variación de temperatura (ni lineal ni al cuadrado); pero además, ya ni siquiera se aprecia la diferencia en la salida media de voltaje entre las pilas de un material y las del otro material, que sí se veía en el modelo que sólo involucraba el término para diferencia en ordenada al origen o intercepto (pag. 43).

Por otro lado, sabemos que ajustando el modelo completo a los datos resulta que ¡todos los coeficientes "β" pueden considerarse diferentes de cero! (pag. 36).

Parte de la explicación de que a veces puedan considerarse diferentes de cero y a veces no, es que estamos ajustando modelos a los datos, es decir, estamos evaluando la posibilidad de que los datos se comporten de acuerdo con esos modelos, lo que no quiere decir que forzosamente se tengan que comportar así.

Tal situación sería equivalente a "la prueba de la zapatilla de cristal", que se emplea en el cuento de Cenicienta, para descubrir a la joven que la perdió en palacio. Si a una joven le queda, le viene bien o le ajusta la zapatilla, es probable que sea la citada joven, pero ¡no quiere decir que realmente lo sea! Si la zapatilla le ajustara a dos doncellas, ¡ni modo que ambas sean la misma joven del palacio!

Además, el comportamiento de variación de los estimadores de los coeficientes de los términos en los modelos no es raro cuando se incluye una misma variable varias veces en el mismo modelo con diferentes exponentes (elevada a diferentes potencias). De ahí que la estimación del coeficiente de un término no sea independiente de la estimación del coeficiente de otro término en el modelo de nuestro ejemplo. De hecho, cuando una variable es la potencia de otra variable, ambas variables no son independientes estre sí (aunque no estén correlacionadas linealmente).

Aunque generalmente los modelos se "mejoran" (explican más y/o mejor) al ir agregando términos de manera secuencial, ya que se van haciendo más "flexibles" para ajustarse a la información contenida en los datos, esto no siempre sucede, como se pudo apreciar en el ejemplo que acabamos de ver; excepto en el caso del modelo "completo".

Es menester mencionar una vez más que cuando se declare un modelo en la computadora para el análisis de los datos, es muy importante saber qué modelo es el de interés, es decir, qué hipótesis nos interesa probar.

En nuestro ejemplo es de interés descubrir la posibilidad de que la tendencia de la respuesta del voltaje se relacione con la variaciónen la temperatura y si tal relación depende o no del material de la pila.

Si no agregamos al modelo los términos donde asociamos la variable indicadora a los términos que representan los componentes lineal y cuadático de la respuesta, no se podrá ver si tales componentes son diferentes dependiendo del material. De igual manera, si no se incluyen los términos cuadrático y cuadrático con variable indicadora, no se podrá ver el efecto de la variación cuadrática de la variable explicativa en la respuesta para cada material.

Aún más, el no incluir términos de mayor orden puede llevarnos a diferentes conclusiones sobre los términos de menor grado que si los incluyéramos. El desconocimiento de tal comportamiento puede conducirnos a conclusiones erróneas.

Recordemos que la significancia estadística de los coeficientes (β's) asociados a cada variable, es decir, la capacidad de explicación de una variable en un modelo se debe ver e interpretar en presencia (o ausencia, según sea el caso) de las demás variables involucradas en el modelo (de los demás coeficientes asociados a tales variables).

En nuestro ejemplo, el modelo "completo" muestra evidencias de que la relación voltaje/temperatura sigue una curva cuadrática, y los modelos "incompletos" no muestran tal relación entre el voltaje y la temperatura.

IV. CORRESPONDENCIA DEL CONCEPTO DE INTERACCIÓN DE DOS FACTORES EN LOS MODELOS DE DISEÑO DE EXPERIMENTOS CON LOS MODELOS DE REGRESIÓN

Estimado lector:

Antes de entrar en tema, es recomendable que lea primero el prefacio, donde no solamente se explica el contenido de este trabajo, sino también el formato que se ha seguido en su presentación. Tales explicaciones le permitirán una mayor comprensión del documento, así como una lectura más fluída.

Además, es conveniente que revise la sección I "NECESIDAD DEL CONOCIMIENTO DE LA HERRAMIENTA ESTADÍSTICA PARA SU USO ADECUADO", donde se exponen conceptos generales, se señalan resultados importantes generados por el paquete estadístico, y se explica su interpretación, lo cual será de gran utilidad para entender el desarrollo del resto de este trabajo. A su vez, en la sección II, se introduce el concepto de las variables indicadoras o "dummy".

Si usted ya leyó el prefacio y la sección I, lo invito a continuar la lectura de esta página.

Cuando un mismo fenómeno puede describirse de más de una forma, es deseable establecer la equivalencia entre los diferentes puntos de cada una de las descripciones. Tal es el caso de los modelos de diseño de experimentos y los modelos de regresión, sobre todo con relación a la idea de interacción, la cual, por su gran uso en la práctica, es muy clara en los modelos de diseño de experimentos.

Si nuestro interés se centra en saber si una respuesta numérica a un fenómeno varía según los niveles "i" de un factor explicativo "A" (sean categorías de una variable cualitativa o valores determinados de una variable cuantitativa), entonces la respuesta "Y" para cada nivel de i del factor A, puede ser descrita a través de un modelo de diseño de experimentos[12], como el siguiente:

---

[12] Para ilustrar los modelos en esta parte introductoria (excepto la gráfica de la pag. 50), se emplearán, invirtiendo sólo los niveles de cada factor, los datos y resultados presentados más adelante en el "caso 2" (pag. 68).

$$Y_{ir} = \mu + \alpha_i + \varepsilon_{ir}$$

$i = 1, 2, \ldots, a.$ (i-ésimo nivel de A).
$r = 1, 2, \ldots, n_i.$ (r-ésima repetición).

con

$\sum_{i=1}^{a} \alpha_i = 0$[13] (la suma de los efectos de los a niveles de A es igual a 0, $i = 1,2, \ldots, a$)

$\varepsilon_{ir} \sim NIID(0,\sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada nivel i de A)
(los errores son independientes, el valor de un error no tiene que ver que ver con el valor de otro error)
(todos los errores se distribuyen de manera idéntica, todos provienen de una misma distribución normal con media 0 y varianza σ^2)

donde

Y_{ir} = respuesta para el tratamiento con el i-ésimo nivel de "A", r-ésima repetición.
μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A", ($\alpha_i = \mu_i - \mu$).
ε_{ir} = error aleatorio no observable de la r-ésima repetición del i-ésimo nivel de "A".
n_i = número de repeticiones en el nivel i-ésimo de "A".

Como puede observarse es costumbre representar los factores en los modelos con letras griegas. Así, para un primer factor "A" se emplea la letra griega alfa "α", para un segundo factor "B" se emplearía la letra "β", pero aquí no se usa para no confundirlo con los coeficientes empleados en los modelos de regresión, nombrando al segundo factor como "C" y representandolo con la letra gama "γ" (pag. 51). Tampoco se debe confundir α, efecto del factor "A", con el α, probabilidad de cometer el error tipo I, descrito en la página 9. A éste último solo se hará referencia cuando se interpreten las salidas del paquete estadístico.

Con respecto a los factores y sus niveles, en los ejemplos de esta parte introductoria corresponden a variables cualitativas y sus respectivas categorías. Aunque en tales ejemplos no se hace referencia a ello, dado que fueron tomados del caso 3 (pag. 84), con excepción de la gráfica de la página 50, el factor "A" son fungicidas y los niveles son el tipo (categoría) de fungicida particular empleado (sea el "1" o el "2"). El factor "C" son insecticidas y los niveles son el tipo (categoría) de insecticida particular empleado (sea el "1" o el "2").

Una variable cuantitativa como factor sería el caso del sueldo de los profesores de acuerdo con su antigüedad, donde los valores determinados de antigüedad (1, 2, 3, 4 y 5 años) representarían los niveles de ese factor (pag. 3 y 12). En el caso de la salida de voltaje de acuerdo con la temperatura a la que se encuentra la pila, los valores de temperatura 50°C, 65°C y 80°C (pag. 29 y 59) representan los niveles del factor temperatura; de hecho, este último ejemplo es el que se presenta en el caso 1 (pag. 57).

---

[13] Esta condición no es realmente un supuesto del modelo, sino una "restricción colateral sobre los estimadores de los parámetros", esto es, un artificio conveniente para poder obtener los estimadores.

Si ahora en vez de considerar un valor de Y_{ir}, nos centramos en la media para el tratamiento i, dado que $\varepsilon_{ir} \sim N(0,\sigma^2)$, podemos reescribir el modelo como

$$\mu_i = \mu + \alpha_i \qquad i = 1, 2, \ldots, a. \quad \text{(i-ésimo nivel de A).}$$

con

$$\sum_{i=1}^{a} \alpha_i = 0^{(14)} \quad \text{(la suma de los efectos de los a niveles de A es igual a 0, } i = 1,2, \ldots, a)$$

donde

μ_i = media poblacional de la respuesta para el tratamiento con el i-ésimo nivel de "A"

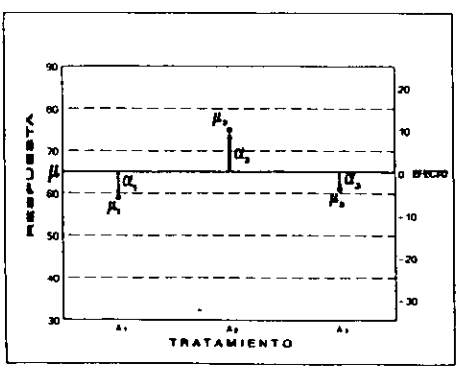
μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.

α_i = efecto (principal) del nivel i de "A".

En este caso, dado que el único efecto involucrado que modifica al valor de la media general para dar como resultado la media de la respuesta presentada en el nivel i de "A" es el correspondiente a dicho nivel, el efecto del nivel i de "A" es igual a la desviación de la media de la respuesta en tal nivel con respecto a la media general

$$\alpha_i = \mu_i - \mu$$

(que de hecho es como se define α_i), lo que se representa a continuación con tres niveles de "A" (a=3), es decir α_1, α_2 y α_3:



De ahí que la media de la respuesta para cada nivel i de "A", pueda entenderse como la media general más el efecto del nivel i del factor "A", al que se han sometido las unidades experimentales ($\mu_i = \mu + \alpha_i$).

---

(14) Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

Con el modelo original, que se presenta a continuación

$$Y_{ir} = \mu + \alpha_i + \varepsilon_{ir}$$

$i = 1, 2, \ldots, a.$ (i-ésimo nivel de A).
$r = 1, 2, \ldots, n_i.$ (r-ésima repetición).

y a través del análisis de varianza correspondiente al mismo, se puede probar el efecto del factor "A", donde H_a se plantea como que al menos el efecto de uno de los niveles i de A es "significativamente" diferente de los demás, y H_0 como que el efecto de ningún nivel es significativamente de cialquier otro:

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_a$
ó
H_0: $\alpha_i = \alpha_{i'}$ para toda i,i'

vs.

H_a: al menos una α_i diferente de las demás α_i
H_a: $\alpha_i \neq \alpha_{i'}$ para alguna i,i'

Si nuestro interés involucra no uno, sino dos factores explicativos "A" y "C", nuestro modelo contemplará para cada tratamiento tanto el nivel i de "A" como el nivel j de "C", y se expresaría:

$$Y_{ijr} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijr}$$

$i = 1, 2, \ldots, a.$ (i-ésimo nivel de A).
$j = 1, 2, \ldots, c.$ (j-ésimo nivel de C).
$r = 1, 2, \ldots, n_{ij}.$ (r-ésima repetición).

con

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{c} \gamma_j = 0$$ [15]

(la suma de los efectos de los "a" niveles de A es igual a 0, i = 1,2, ... , a)
(la suma de los efectos de los "c" niveles de C es igual a 0, j = 1,2, ... , c)

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$ (Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de niveles de i de A y j de C)
(los errores son **independientes**, el valor de un error no tiene que ver que ver con el valor de otro error)
(todos los errores se **distribuyen** de manera **idéntica**, todos provienen de una misma distribución normal con media 0 y varianza σ^2)

donde

- Y_{ijr} = respuesta para el tratamiento con el i-ésimo nivel de "A", j-ésimo nivel de "C" y r-ésima repetición.
- μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.
- α_i = efecto (principal) del nivel i de "A".
- γ_j = efecto (principal) del nivel j de "C".
- ε_{ijr} = error aleatorio no observable de la r-ésima repetición del tratamiento con el i-ésimo nivel de "A" y el j-ésimo nivel de "C".
- n_{ij} = número de repeticiones en el tratamiento con el i-ésimo nivel de "A" y el j-ésimo nivel de "C".

---

[15] Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

Considerando la media por tratamiento (combinación de niveles de los diferentes factores), el modelo sería

$$\mu_{ij} = \mu + \alpha_i + \gamma_j \qquad \begin{array}{l} i = 1, 2, \ldots, a. \\ \quad \text{(i-ésimo nivel de A).} \\ j = 1, 2, \ldots, c. \\ \quad \text{(j-ésimo nivel de C).} \end{array}$$

con

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{c} \gamma_j = 0^{(16)} \quad \begin{array}{l} \text{(la suma de los efectos de los "a" niveles de A es igual a 0)} \\ \text{(la suma de los efectos de los "c" niveles de C es igual a 0)} \end{array}$$

donde

μ_{ij} = respuesta media para el tratamiento con el i-ésimo nivel de "A" y j-ésimo nivel de "C".
μ = media general poblacional, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A".
γ_j = efecto (principal) del nivel j de "C".

Para este caso, donde tenemos más de un factor, los efectos α_i y γ_j quedan definidos como $\alpha_i = \mu_{i.} - \mu$ y $\gamma_j = \mu_{.j} - \mu$, donde $\mu_{i.}$ es el promedio, sobre los "c" niveles de C, de las μ_{ij}, mientras que $\mu_{.j}$ es el promedio, sobre los "a" niveles de A, de las μ_{ij} (ver también la página 58).

Con el modelo original de la página anterior y a través del análisis de varianza correspondiente al mismo, se puede probar el efecto de los factores "A" y "C", donde H_{a1} y H_{a2} se plantean como que al menos el efecto de uno de los niveles i de "A" o j de "C" es, "significativamente" diferente de los demás, y H_{01} y H_{02} como que el efecto de ningún nivel de "A" o de "C", respectivamente, es significativamente diferente de cualquier otro:

- *Para el efecto (principal) de "A":*

H_{01}: $\alpha_1 = \alpha_2 = \ldots = \alpha_a$ ó H_{01}: $\alpha_i = \alpha_{i'}$, para toda i,i' vs. H_{a1}: al menos una α_i diferente de las demás α_i H_{a1}: $\alpha_i \neq \alpha_{i'}$, para al menos una i,i'

- *Para el efecto (principal) de "C":*

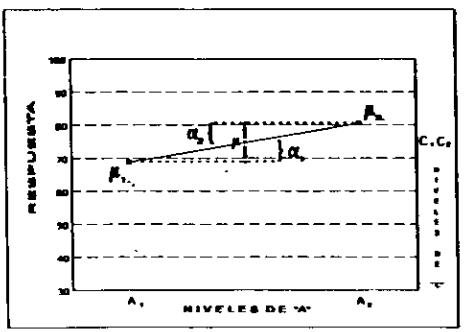
H_{02}: $\gamma_1 = \gamma_2 = \ldots = \gamma_c$ H_{02}: $\gamma_j = \gamma_{j'}$, para toda j,j' vs. H_{a2}: al menos una γ_j diferente de las demás γ_j H_{a2}: $\gamma_j \neq \gamma_{j'}$, para al menos unaj,j'

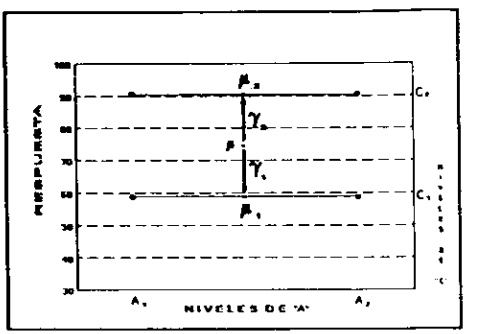
En el análisis de varianza correspondiente a dicho modelo, puede resultar que ninguna hipótesis nula se rechace, que sólo se rechace una de ellas (cualquiera), o que se rechacen ambas. Se analizarán brevemente los dos últimos casos.

---

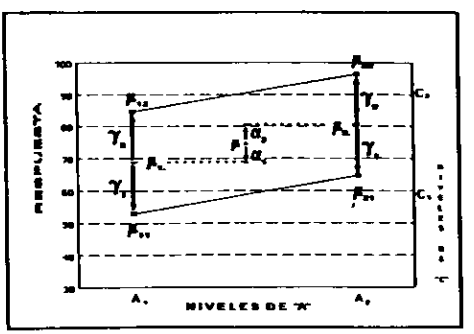
(16) Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

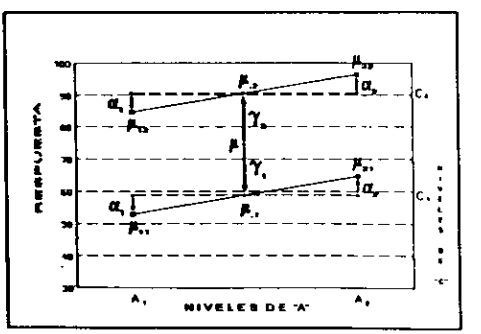
Entonces, puede que sólo sea estadísticamente significativo el efecto del factor "A", es decir, se rechace "H_{01}: $\alpha_i = \alpha_{i'}$ para toda i,i'" pero no "H_{02}: $\gamma_j = \gamma_{j'}$ para toda j,j'" (gráfica de la izquierda a continuación). También puede resultar que sólo sea estadísticamente significativo el efecto del factor "C", es decir, se rechace "H_{02}: $\gamma_j = \gamma_{j'}$ para toda j,j'" pero no "H_{01}: $\alpha_i = \alpha_{i'}$ para toda i,i'" (gráfica de la derecha a continuación).





En el caso de que ambas hipótesis nulas se rechacen el modelo mostrará que los efectos de los nivel 1 y 2 de C (γ_1 y γ_2) son diferentes entre sí y los efectos de los nivel 1 y 2 de A (α_1 y α_2) también lo son.





Sin embargo, de acuerdo con el modelo las diferencias $\mu_{11} - \mu_{1.}$ y $\mu_{21} - \mu_{2.}$ son ambas iguales a γ_1, y $\mu_{12} - \mu_{1.}$ y $\mu_{22} - \mu_{2.}$ son ambas iguales a γ_2 (gráfica anterior a la izquierda), esto es, los efectos de los niveles 1 y 2 de C son iguales en los diferentes niveles de A (modelándose tan solo con γ_1 y γ_2 respectivamente). De manera similar, las diferencias $\mu_{11} - \mu_{.1}$ y $\mu_{12} - \mu_{.2}$ son ambas iguales a α_1, y $\mu_{21} - \mu_{.1}$ y $\mu_{22} - \mu_{.2}$ son ambas iguales a α_2 (gráfica anterior a la derecha), esto es, los efectos de los niveles 1 y 2 de A son iguales en los diferentes niveles de C (modelándose tan solo con α_1 y α_2 respectivamente). Lo anterior es una forma de expresar la ausencia de interacción, ¡pero no se declaró un término en el modelo para representarla!.

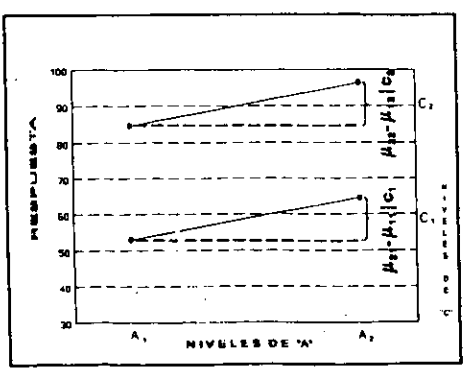
Esta situación de ausencia de interacción no siempre corresponde con el comportamiento de nuestros datos en un problema real, y de ahí la necesidad de manejar el concepto de **interacción** y de involucrarlo en los modelos para poder plantear hipótesis sobre su presencia o ausencia.

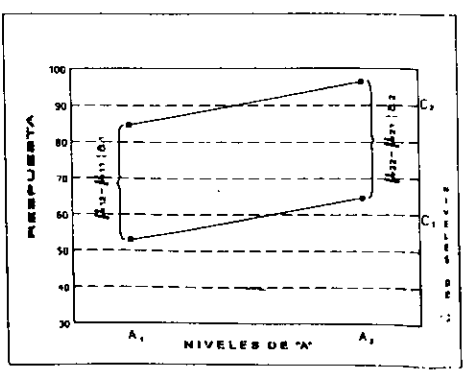
Una forma sencilla para explicar el concepto de interacción en Diseño de Experimentos, es con un arreglo de tratamientos factorial 2x2 completo y balanceado.

Un diseño "factorial 2x2 completo y balanceado", esto es, "factorial 2x2" porque involucra dos factores con dos niveles para cada factor (2factoresx2niveles, de ahí el nombre); "completo" porque se consideran todos los niveles de cada factor en todos los demás factores, en este caso, los dos niveles del primer factor están presentes en los dos niveles del otro factor (con un total de cuatro tratamientos); y "balanceado" porque para cada tratamiento (combinación de niveles de los diferentes factores) existe el mismo número de repeticiones.

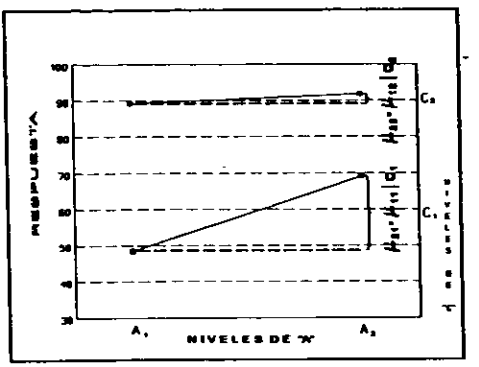
En tal experimento, cuando el cambio de la respuesta de un nivel de un factor a otro nivel de ese mismo factor **es diferente** para los dos niveles del otro factor, se dice que ***hay interacción***. Por el contrario, cuando el cambio de la respuesta de un nivel de un factor a otro nivel de ese mismo factor **es igual** para los dos niveles del otro factor, se dice que *no hay interacción*.

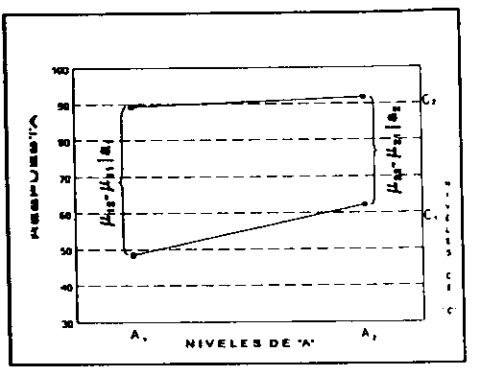
Por ejemplo, cuando el cambio de la respuesta de A_1 a A_2 (los dos niveles de un factor, el "A") **es igual** para C_1 y C_2 (los dos niveles del otro factor, el "C") *no hay interacción* entre los factores "A" y "C". Equivalentemente, cuando el cambio de la respuesta de C_1 a C_2 (niveles del "C") **es igual** para A_1 y A_2 (niveles del otro factor, el "A") *no hay interacción* entre "A" y "C".



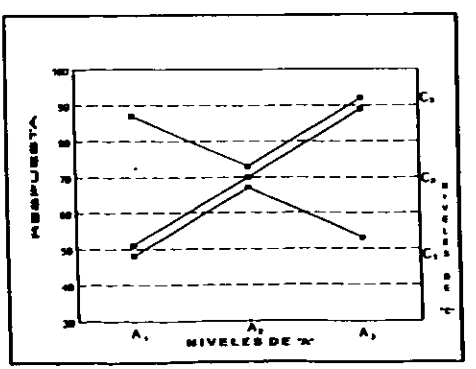


Por otro lado, cuando el cambio de la respuesta de A_1 a A_2 (dos niveles de un factor, el "A") **es diferente** para C_1 y C_2 (los dos niveles del otro factor, el "C") *hay interacción*. Equivalentemente, cuando el cambio de la respuesta de C_1 a C_2 (dos niveles de un factor, el "C") **es diferente** para A_1 y A_2 (los dos niveles del otro factor, el "A") *hay interacción*.





De manera más general, considerando dos o más niveles por factor (en la gráfica a continuación se presentan 3 niveles por factor), "A" y "C" interactúan si al cambiar del mismo modo los niveles de "A" al menos un patrón de efectos obtenidos es diferente al cambiar los niveles de "C". Equivalentemente, "A" y "C" interactúan si al cambiar del mismo modo los niveles de "C" al menos un patrón de efectos obtenidos es diferente al cambiar los niveles de "A". O lo que es lo mismo, si los efectos de un factor cambian según los niveles de otro factor, tales factores interactúan entre sí.



Si vemos el cambio de la media cuando pasamos por C_1, C_2 y C_3, todo en el nivel A_3, la respuesta sube y luego baja, mientras que en A_2 la respuesta siempre sube. Aunque al pasar de C_1 a C_2 el cambio en la respuesta es igual en ambos casos, al pasar de C_2 a C_3 baja en A_3 y sube en A_2.

La manera de modelar esta situación de interacción es a través de un término que implique el efecto conjunto de A y C para cada tratamiento (combinación de niveles de A y C), como se muestra a continuación:

$$Y_{ijr} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \varepsilon_{ijr}$$

$i = 1, 2, \ldots, a.$ (i-ésimo nivel de A).
$j = 1, 2, \ldots, c.$ (j-ésimo nivel de C).
$r = 1, 2, \ldots, n_{ij}.$ (r-ésima repetición).

con

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{c} \gamma_j = \sum_{i=1}^{a} \alpha\gamma_{ij} = \sum_{j=1}^{c} \alpha\gamma_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{c} \alpha\gamma_{ij} = 0^{(17)}$$

(la suma de los efectos de los "a" niveles de A es igual a 0, $i = 1,2, \ldots, a$)

(la suma de los efectos de los "c" niveles de C es igual a 0, $j = 1,2, \ldots, c$)
(la suma de los efectos de los "a" niveles de la interacción entre A y C es igual a 0, $i = 1,2,\ldots,a$)
(la suma de los efectos de los "c" niveles de la interacción entre A y C es igual a 0, $j = 1,2,\ldots,c$)
(la suma de los efectos de los "a*c" niveles de la interacción entre A y C es igual a 0, $i = 1,2,\ldots,a$ $j = 1,2,\ldots,c$)

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de niveles de i de A y j de C)
(los errores son **independientes**, el valor de un error no tiene que ver que ver con el valor de otro error)
(todos los errores se **distribuyen** de manera **idéntica**, todos provienen de una misma distribución normal con media 0 y varianza σ^2)

donde

Y_{ijr} = respuesta para el tratamiento con el i-ésimo nivel de "A", j-ésimo nivel de "C" y r-ésima repetición.
μ = media general poblacional, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A".
γ_j = efecto (principal) del nivel j de "C".
$\alpha\gamma_{ij}$ = efecto de la interacción del nivel i de "A" y el nivel j de "C".
ε_{ijr} = error aleatorio no observable de la r-ésima repetición del tratamiento con el i-ésimo nivel de "A" y el j-ésimo nivel de "C".
n_{ij} = número de repeticiones en el tratamiento con el i-ésimo nivel de "A" y el j-ésimo nivel de "C".

Considerando la media por tratamiento (combinación de niveles de los diferentes factores), el modelo sería

$$\mu_{ij} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij}$$

$i = 1, 2, \ldots, a.$ (i-ésimo nivel de A).
$j = 1, 2, \ldots, c.$ (j-ésimo nivel de C).

---

(17) Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

con

$$\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{c}\gamma_j = \sum_{i=1}^{a}\alpha\gamma_{ij} = \sum_{j=1}^{c}\alpha\gamma_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{c}\alpha\gamma_{ij} = 0^{(18)}$$

(la suma de los efectos, por niveles de A, de C o de ambos es igual a 0)

donde

μ_i = respuesta media poblacional para el tratamiento con el i-ésimo nivel de "A" y j-ésimo nivel de "C".
μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A".
γ_j = efecto (principal) del nivel j de "C".
$\alpha\gamma_{ij}$ = efecto de la interacción del nivel i de "A" y el nivel j de "C".

Con el modelo original del principio de la página 56, y a través del análisis de varianza correspondiente al mismo, se puede probar, además del efecto principal de los factor "A" y "C" (α_i y γ_j), su interacción ($\alpha\gamma_{ij}$):

- *Para el efecto (principal) de "A":*

H_{01}: $\alpha_1 = \alpha_2 = \ldots = \alpha_a$ ó H_{01}: $\alpha_i = \alpha_{i'}$, para toda i,i' vs. H_{a1}: al menos una α_i diferente de las demás α_i H_{a1}: $\alpha_i \neq \alpha_{i'}$, para al menos unai,i'

- *Para el efecto (principal) de "C":*

H_{02}: $\gamma_1 = \gamma_2 = \ldots = \gamma_c$ ó H_{02}: $\gamma_j = \gamma_{j'}$, para toda j,j' vs. H_{a2}: al menos una γ_j diferente de las demás γ_j H_{a2}: $\gamma_j \neq \gamma_{j'}$, para al menos unaj,j'

- *Para el efecto de interacción entre "A" y "C".*

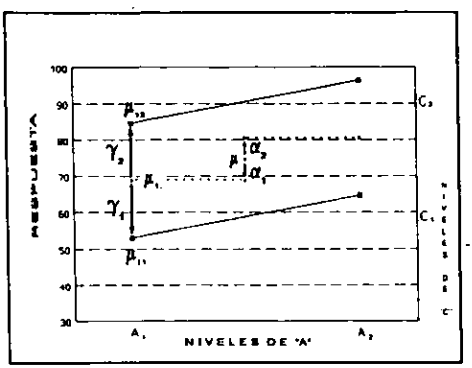
H_{03}: $\alpha\gamma_{ij}=0$ para toda combinación ij ó H_{03}: $\alpha\gamma_{11}=\alpha\gamma_{12}=\ldots=\alpha\gamma_{21}=\alpha\gamma_{22}=\ldots=\alpha\gamma_{ac}=0$ vs. H_{a3}: $\alpha\gamma_{ij}\neq 0$ al menos para una combinación ij

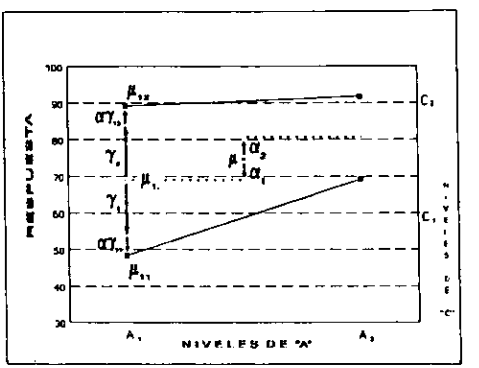
El efecto α_i es el efecto principal de "A" en el nivel "i", es decir, el efecto promedio del nivel "i" de "A" en la respuesta al considerar todos los niveles de "C" (definido en la página 52 como $\alpha_i = \mu_{i.} - \mu$). Lo mismo opera para γ_j, efecto promedio del nivel "j" de "C" en la respuesta al considerar todos los niveles de "A" (definido en la página 52 como $\gamma_j = \mu_{.j} - \mu$).

---

(18) Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

Los términos $\alpha\gamma_{ij}$ permiten modelar la diferencia de los efectos de los niveles de A en cada nivel de C, así como la diferencia de los efectos de los niveles de C en cada nivel de A, es decir, el efecto de A y C para cada combinación de niveles "ij". De manera que si $\alpha\gamma_{ij} = 0$ para toda combinación ij, entonces no hay interacción; si $\alpha\gamma_{ij} \neq 0$ para alguna combinación "ij", existe interacción entre "A" y "C" y $\alpha\gamma_{ij}$ mide el efecto de la interacción, para la combinación de niveles i de A y j de C, como la diferencia entre la respuesta con y sin interacción (es decir, la desviación del patrón de no interacción, como se muestra a continuación para $\alpha\gamma_{11}$ y $\alpha\gamma_{12}$).





De esta forma, para el caso del experimento 2x2 que nos sirvió de ejemplo, considerando que para toda "ij" $\alpha\gamma_{ij} \neq 0$, las diferencias $\mu_{11} - \mu_{1.}$ y $\mu_{21} - \mu_{2.}$ ya no son ambas iguales a γ_1, sino son $\gamma_1 + \alpha\gamma_{11}$ y $\gamma_1 + \alpha\gamma_{21}$ respectivamente, ni $\mu_{12} - \mu_{1.}$ y $\mu_{22} - \mu_{2.}$ son ambas iguales a γ_2, sino a $\gamma_2 + \alpha\gamma_{12}$ y $\gamma_2 + \alpha\gamma_{22}$ respectivamente.

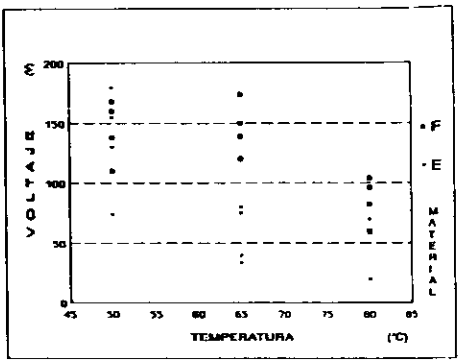
De manera similar, las diferencias $\mu_{11} - \mu_{.1}$ y $\mu_{12} - \mu_{.2}$ ya no son ambas iguales a α_1, sino son $\alpha_1 + \alpha\gamma_{11}$ y $\alpha_1 + \alpha\gamma_{12}$, ni $\mu_{21} - \mu_{.1}$ y $\mu_{22} - \mu_{.2}$ son ambas iguales a α_2, sino a $\alpha_2 + \alpha\gamma_{21}$ y $\alpha_2 + \alpha\gamma_{22}$ respectivamente. Con todo lo anterior se ve claramente que la diferencia en efectos está representada por los términos $\alpha\gamma_{ij}$.

Para comparar la diferencia en interpretación de los modelos de Diseño de Experimentos y de Regresión, sobre todo con relación a la interacción, se analizarán dos casos, según el tipo de variables explicativas (factores) cuando se involucran sólo dos.

CASO 1. UNA VARIABLE O FACTOR CUANTITATIVO Y EL OTRO CUALITATIVO

Regresando al ejemplo de la salida del voltaje de unas pilas y la relación con el material del que están hechas y la temperatura a la que se encuentran (pag. 31),

		Voltaje					
		Temperatura (°C)					
		50		65		80	
MATERIAL	E	130	155	34	40	20	70
		74	180	80	75	82	58
	F	138	110	174	120	96	104
		168	160	150	139	82	60



si queremos modelar la respuesta "salida del voltaje" en función de los factores "temperatura" donde se instala la pila ("A", con niveles de 50, 65 y 80 °C) y "material" del que está hecha ("C", con niveles E y F), a través de un modelo de diseño de experimentos, considerando el modelo más sencillo que contempla sólo los efectos principales, el modelo resultante sería

$$Y_{ijr} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijr}$$

$i = 1, 2, 3.$ (i-ésimo nivel de A).
$j = 1, 2.$ (j-ésimo nivel de C).
$r = 1, 2, 3, 4.$ (r-ésima repetición).

con

$$\sum_{i=1}^{3} \alpha_i = \sum_{j=1}^{2} \gamma_j = 0^{(19)}$$

(la suma de los efectos de los 3 niveles de A es igual a 0, la suma de los efectos de los 2 niveles de C es igual a 0)

$\varepsilon_{ijr} \sim NIID(0, \sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de niveles de i de A y j de C)

donde

- Y_{ijr} = respuesta para el tratamiento con el i-ésimo nivel de "A" (temperatura), j-ésimo nivel de "C" (material) y r-ésima repetición.
- μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.
- α_i = efecto (principal) del nivel i de "A" (temperatura).
- γ_j = efecto (principal) del nivel j de "C" (material).
- ε_{ijr} = error aleatorio no observable de la r-ésima repetición del tratamiento con el i-ésimo nivel de "A" y el j-ésimo nivel de "C".

---

(19) Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

y las hipótesis a contrastar serían:

- *Para el efecto (principal) de "A"* (temperatura).

H_{01}: $\alpha_1 = \alpha_2 = \alpha_3$ vs. H_{a1}: al menos una α_i diferente de las demás α_i

(H_{01}: $\alpha_i = \alpha_{i'}$, para toda i,i') (H_{a1}: $\alpha_i \neq \alpha_{i'}$, para al menos una i,i')

- *Para el efecto (principal) de "C"* (material).

H_{02}: $\gamma_1 = \gamma_2$ vs. H_{a2}: $\gamma_1 \neq \gamma_2$

(H_{02}: $\gamma_j = \gamma_{j'}$, para toda j,j') (H_{a2}: $\gamma_j \neq \gamma_{j'}$, para al menos una j,j')

A continuación se muestra parte de la salida del paquete estadístico SYSTAT5 para el modelo anterior, donde puede verse claramente que los datos aportan evidancia estadísticamente significativa de que hay un efecto de temperatura (diferente voltaje según la temperatura, **TEM**) y un efecto de material (diferente voltaje según el material, **MAT**), de acuerdo con los valores de "**P**" (0.002 y 0.005 respectivamente) y considerando un $\alpha = 0.05$[20].

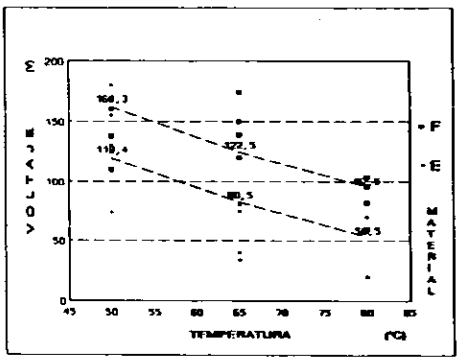
ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
TEM ←	18510.750	2	9255.375	8.598 →	*0.002*
MAT ←	10542.042	1	10542.042	9.793 →	*0.005*
ERROR	21529.833	20	1076.492		

En este caso el valor de **P** no corresponde a una prueba de "t" para ver si la estimación de un coeficiente de regresión β_i es diferente de cero (si $\beta_i \neq 0$), como se explicó en la página 8, sino a una prueba "F" para ver si la estimación de los efectos de los niveles de uno o varios factores son diferentes de entre sí (en este ejemplo "al menos un $\hat{\alpha}_i$ diferente de las demás $\hat{\alpha}_i$" y "$\hat{\gamma}_1 \neq \hat{\gamma}_2$").

[20] Para una explicación de la interpretación de "α" consulte la página 9, y las páginas 49 y 50 para la diferencia con "α_i".

Con la estimación de las medias μ_{ij} ($\hat{\mu}_{ij}$) o de los efectos α_i y γ_j ($\hat{\alpha}_i$ y $\hat{\gamma}_j$) podemos ver que el voltaje disminuye al aumentar la temperatura y que voltaje es menor para el material "E" que para el "F". (La gráfica se elaboró con las medias estimadas reportadas en la misma salida).



Además puede verse que el cambio del voltaje con respecto a la temperatura no es lineal pues, siendo los niveles de la temperatura equidistantes, la disminución del voltaje no es igual al pasar de 50 a 65 °C que al pasar de 65 a 80 °C.

Todo lo anterior también puede apreciarse cuando se ajusta el modelo de regresión

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_j + \varepsilon_{ijr}$$

$i = 1,2,3.$
$j = 1,2$
$r = 1,2,3,4.$

$$Z_j = \begin{cases} 1 \text{ si es material F} & (j=1) \\ 0 \text{ si NO es material F (es material E)} & (j=2) \end{cases}$$

con

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

Y_{ijr} = VOLTAJE de la pila de material j, de la r-ésima observación (repetición) del i-ésimo valor de temperatura.
X_i = i-ésimo valor de TEMPERATURA.
Z_j = j-ésimo valor de la variable indicadora de MATERIAL (Z_1=1, Z_2=0).
ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de temperatura para el material j.

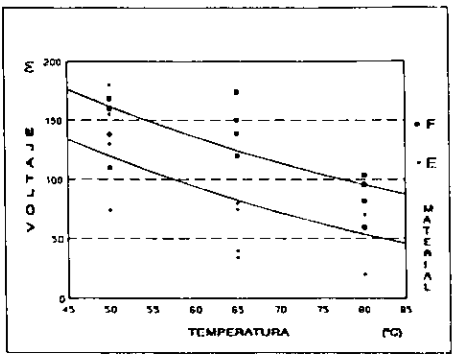
Dado que la temperatura es una variable cuantitativa su efecto (α_i en diseño de experimentos) puede modelarse a través de una parábola, es decir, con los coeficientes asociados a un término cuadrático y a uno lineal ambos en X ($\beta_1 X_i$ y $\beta_2 X_i^2$).

El efecto del material (γ_j en diseño de experimentos) se puede expresar como el cambio en ordenada al origen (mueve a la parábola verticalmente) al pasar del material "E" al "F".

DEP VAR: VOL N: 24
MULTI. R: 0.76 SQ.MULTI.R: 0.57
ADJUSTED SQUARED MULTI. R: 0.511

STANDARD ERROR OF ESTIMATE: 32.81

VARIABLE	COEFF.		T	P(2 TAIL)
CONSTANT	301.5		1.16	0.26
TEM	-4.5	$\leftarrow\beta_1$	-0.55	→ *0.59*
TEM*TEM	0.01	$\leftarrow\beta_2$	0.28	→ *0.79*
MAT	41.9	$\leftarrow\beta_3$	3.13	→ *0.01*



La salida del paquete estadístico SYSTAT5 para el modelo anterior, nos dice que podemos considerar que los datos de nuestro ejemplo sólo aportan evidencia estadísticamente significativa de que haya diferencia en la salida media de voltaje entre las pilas de un material y las del otro material, ya que P = 0.01 (para ver si β_3 es diferente de cero) y considerando un $\alpha = 0.05$[21].

Los resultados no aportan suficiente evidencia para considerar que la relación entre el voltaje de las pilas (Y) y la temperatura (X) a la que se encuentran estén relacionados a través de una curva cuadrática, pues no se rechazó la hipótesis de que el componente cuadrático fuese cero (H_0: $\beta_2 = 0$), ya que P = 0.79 y considerando un $\alpha = 0.05$; ni tampoco se rechazó la hipótesis de que el componente lineal fuese cero (H_0: $\beta_1 = 0$), ya que P = 0.59 y considerando un $\alpha = 0.05$[14]. (La gráfica se elaboró con los coeficientes estimados reportados en la misma salida).

[21]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

Estos son los mismos resultados ya revisados en la sección III página 43. Sin embargo, como se vió anteriormente en la página 38, con el modelo completo de la misma sección sí se puede decir que exista diferencia en las parábolas que describen el comportamiento del voltaje a partir de la temperatura para los diferentes materiales, es decir, los patrones de comportamiento no son iguales (las parábolas no son paralelas).

En un modelo de diseño de experimentos tal situación (comportamiento diferente por material) se expresaría con el término de la "interacción", quedando el modelo completo como sigue:

$$Y_{ijr} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \varepsilon_{ijr}$$

$i = 1, 2, 3.$ (i-ésimo nivel de A).
$j = 1, 2.$ (j-ésimo nivel de C).
$r = 1, 2, 3, 4.$ (r-ésima repetición).

con

$$\sum_{i=1}^{3} \alpha_i = \sum_{j=1}^{2} \gamma_j = \sum_{i=1}^{3} \alpha\gamma_{ij} = \sum_{j=1}^{2} \alpha\gamma_{ij} = \sum_{i=1}^{3}\sum_{j=1}^{2} \alpha\gamma_{ij} = 0^{(22)}$$

(suma de efectos, por niveles de A, de C o de ambos es igual a 0)

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de niveles de i de A y j de C)

donde

Y_{ijr} = respuesta para el tratamiento con el i-ésimo nivel de "A" (temperatura), j-ésimo nivel de "C" (material) y r-ésima repetición.
μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A" (temperatura).
γ_j = efecto (principal) del nivel j de "C" (material).
$\alpha\gamma_{ij}$ = efecto de la interacción del nivel i de "A" y el nivel j de "C".
ε_{ijr} = error aleatorio no observable de la r-ésima repetición del tratamiento con el i-ésimo nivel de "A" y el j-ésimo nivel de "C".

y las hipótesis a contrastar serían:

- *Para el efecto (principal) de "A"* (temperatura).

$H_{01}: \alpha_1 = \alpha_2 = \alpha_3$ vs. H_{a1}: al menos una α_i diferente de las demás α_i

($H_{01}: \alpha_i = \alpha_{i'}$, para toda i,i') ($H_{a1}: \alpha_i \neq \alpha_{i'}$, para al menos unai,i')

- *Para el efecto (principal) de "C"* (material).

$H_{02}: \gamma_1 = \gamma_2$ vs. $H_{a2}: \gamma_1 \neq \gamma_2$

($H_{02}: \gamma_j = \gamma_{j'}$, para toda j,j'). ($H_{a2}: \gamma_j \neq \gamma_{j'}$, para al menos unaj,j')

---

(22) Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

- *Para el efecto de interacción entre "A" y "C".*

H_{03}: $\alpha\gamma_{11}=\alpha\gamma_{12}=\alpha\gamma_{21}=\alpha\gamma_{22}=\alpha\gamma_{31}=\alpha\gamma_{32}=0$ vs. H_{a3}: $\alpha\gamma_{ij}\neq 0$ al menos para una combinación ij

(H_{03}: $\alpha\gamma_{ij}=0$ para toda combinación ij)

A continuación se muestra parte de la salida del paquete estadístico SYSTAT5 para el modelo anterior, donde puede verse claramente que los datos aportan evidencia estadísticamente significativa de que, además de haber un efecto de temperatura (diferente voltaje según la temperatura, **TEM**) y un efecto de material (diferente voltaje según el material, **MAT**), existe un efecto de interacción (**MAT*YEM**), de manera que la variación del voltaje dada por el incremento en temperatura es diferente para cada material, de acuerdo con los valores de **"P"** (0.002, 0.001 y 0.032 respectivamente) y considerando un $\alpha = 0.05$[23].

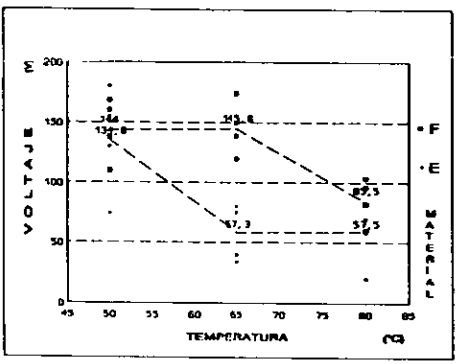
ANALYSIS OF VARIANCE

SOURCE		SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
TEM	←	18510.750	2	9255.375	11.358 →	***0.001***
MAT	←	10542.042	1	10542.042	12.937 →	***0.002***
TEM*MAT	←	6861.583	2	3430.792	4.210 →	***0.032***

Recordemos que este caso, como se describió en la página 60, el valor de **P** no corresponde á una prueba de "t" para ver si la estimación de un coeficiente de regresión β_i es diferente de cero (si $\beta_i \neq 0$), como se explicó en la página 8, sino a una prueba "F" para ver si la estimación de los efectos de los niveles de uno o varios factores son diferentes entre sí, o si su interacción es diferente de cero (en este ejemplo "al menos un $\hat{\alpha}_i$ diferente de las demás $\hat{\alpha}_i$", "$\hat{\gamma}_1 \neq \hat{\gamma}_2$" y "al menos un $\hat{\alpha\gamma}_{ij} \neq 0$").

[23]Para una explicación de la interpretación de "α" consulte la página 9, y las páginas 49 y 50 para la diferencia con "α_i".

A partir de las estimaciones de μ_{ij} o de las estimaciones de los efectos α_i, γ_j y $\alpha\gamma_{ij}$, podemos ver que al incrementar la temperatura el voltaje disminuye considerablemente (de 134.8 a 57.3) y luego aumenta muy ligeramente (de 57.3 a 57.5) para el material "E", mientras que aumenta muy ligeramente (de 144 a 145.8) y luego disminuye drásticamente (de 145.8 a 85.5) para el material "F".



Gráficamente puede verse que la respuesta para el material "E" asemeja una parábola cóncava hacia arriba, mientras que la respuesta para el material "F" asemeja una parábola cóncava hacia abajo.

Para el caso de regresión, este mismo ajuste se realizó a través del modelo

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_j + \beta_4 X_i Z_j + \beta_5 X_i^2 Z_j + \varepsilon_{ijr}$$

$i = 1,2,\ldots,t.$
$j = 1,2$
$r = 1,2,\ldots,n_i.$

$$Z_j = \begin{cases} 1 \text{ si es material F} & (j=1) \\ 0 \text{ si NO es material F (es material E)} & (j=2) \end{cases}$$

con

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$

(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de valores de i de X y j de Z)

donde

Y_{ijr} = VOLTAJE de la pila de material j, de la r-ésima observación (repetición) del i-ésimo valor de temperatura.
X_i = i-ésimo valor de TEMPERATURA.
Z_j = j-ésimo valor de la variable indicadora de MATERIAL (Z_1=1, Z_2=0).
ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo valor de temperatura para el material j.

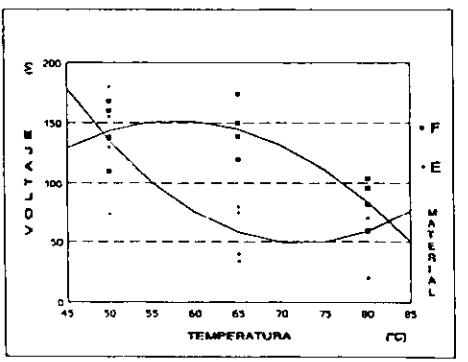
en el cual los coeficientes asociados a X (β_1 y β_2) pueden revelar un efecto de temperatura (lineal y cuadrático respectivamente), β_3 puede revelar el efecto de material y β_4 y β_5 revelan la existencia de la interacción.

A continuación se muestra parte de la salida del paquete estadístico SYSTAT5 para el modelo anterior, donde puede verse claramente que los coeficientes (β's) asociados a todos los términos del modelo son significativamente diferentes de cero (de acuerdo con los valores de "P" y considerando un $\alpha = 0.05$[24]).

```
DEP VAR:  VOL            N:   24
MULTI. R: 0.84   SQ.MULTI.R: 0.71
ADJUSTED SQUARED MULTIPLE R: .56
STANDARD ERROR OF ESTIMATE: 28.5

VARIABLE     COEF.      T  P(2 TAIL)

CONSTANT     954.6 <-β0   2.99  → 0.01
TEM          -25.0 <-β1  -2.47  → 0.02
TEM*TEM        0.2 <-β2   2.22  → 0.04
MAT        -1264.2 <-β3  -2.80  → 0.01
TEM*MAT       41.0 <-β4   2.86  → 0.01
TEM*TEM*MAT   -0.3 <-β5  -2.83  → 0.01
```



Esto es, podemos considerar que los datos de nuestro ejemplo aportan evidencia estadísticamente significativa de que sí hay relación entre la temperatuta y el voltaje, y que tal relación es diferente (en intercepto u ordenada al origen, componente lineal y componente cuadrático - β_3, β_4 y β_5 respectivamente) para cada tipo de material (con un $\alpha = 0.05$)[16], así que la relación entre la temperatura y el voltaje puede considerase diferente para cada material (presencia de interacción).

[24]Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

Para el caso de una variable cuantitativa y una cualitativa, dado que la interacción en modelos de diseños de experimentos es la diferencia en el patrón de variación en la respuesta dada por cambio de niveles de un factor al cambiar los niveles del otro factor, y que al ser una de las variables explicativas (factor) de tipo cuantitativo, el patrón de variación en la respuesta debida a los niveles de esa variable está dado por los coeficientes asociados al término en X en una curva, y la diferencia del patrón de cambio al variar el nivel del otro factor (cualitativo) está dado por el cambio en los términos en X asociados a una variable indicadora.

En nuestro ejemplo la variable explicativas (factor) de tipo cuantitativo es la temperatura, y el patrón de cambio en la respuesta debida a los niveles de esta variable cuantitativa está dado por los coeficientes asociados al término en X en una parábola, y la diferencia del patrón de cambio al variar el nivel del otro factor (material) está dado por el cambio en los términos en los términos lineal y cuadrático en X de la parábola asociados a la variable "Z".

De esta manera, si β_4 y/o β_5 son diferentes de cero, entonces podemos decir que sí hay interacción entre temperatura (A en diseños) y material (C en diseños) cuando se trate del modelo

$$Y_{ijr} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_j + \beta_4 X_i Z_j + \beta_5 X_i^2 Z_j + \varepsilon_{ijr}$$

equivalente a

$$Y_{ijr} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \varepsilon_{ijr}$$

Para los modelos anteriores

$\beta_4 \neq 0$ y/o $\beta_5 \neq 0$ $\longleftrightarrow$ $\alpha\gamma_{ij} \neq 0$

(diferencia en efecto lineal) *(diferencia en efecto cuadrático)* *al menos para una combinación "ij"*

CASO 2. DOS VARIABLES O FACTORES CUALITATIVOS

Es semejante al caso del sueldo de los profesores o al del voltaje de acuerdo a la temperatura y el material de la pila, sólo que ahora ambas variables (factores) son cualitativas. En lo que compete a diseño de experimentos, no hay modificación alguna en los modelos.

Con respecto a los modelos de regresión, es semejante a lo ya descrito, sólo que ahora se emplearán variables indicadoras para representar ambas variables o factores.

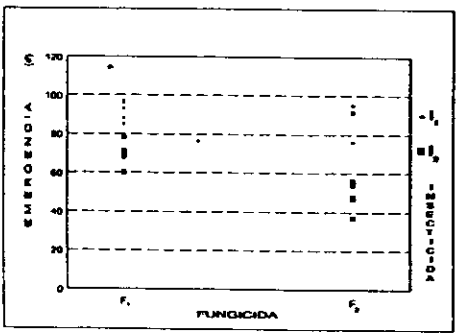
Consideremos los siguientes datos de un arreglo de tratamientos factorial 2x2 completo y balanceado[25], donde se quiere investigar la relación entre dos fungicidas ("A" con dos niveles F1 y F2) y dos insecticidas ("C" con dos niveles I1 y I2) con la variación en la cantidad de plántulas que emergieron (en cada repetición de tratamiento se emplearon 100 semillas).

Se pretende investigar si un fungicida es mejor que el otro (mayor efecto en el porcentaje de emergencia de plántulas, es decir major porcentaje de emergencia) y si un insecticidas es mejor que el otro (mismo criterio en la emergencia o formación de plántulas a partir de semillas).

Emergencia de Plántulas (#)

FUNGICIDA		F_1		F_2	
INSECTICIDA	I_1	85	93	91	76
		97		92	
		88	96	92	95
	I_2	60	69	47	37
		71		56	
		78	68	48	54

[25]Para una explicación sobre un diseño "factorial 2x2 completo y balanceado" consulte la página 54.



El modelo de diseño de experimentos para probar las hipótesis de efectos principales sería:

$$Y_{ijr} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijr}$$

$i = 1, 2,.$ (i-ésimo nivel de A).
$j = 1, 2.$ (j-ésimo nivel de C).
$r = 1, 2, 3, 4, 5.$ (r-ésima repetición).

con

$$\sum_{i=1}^{2} \alpha_i = \sum_{j=1}^{2} \gamma_j = 0^{(26)}$$

(la suma de los efectos de los 2 niveles de A es igual a 0, la suma de los efectos de los 2 niveles de C es igual a 0)

$\varepsilon_{ijr} \sim NIID(0, \sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de niveles de i de A y j de C)

donde

Y_{ijr} = respuesta para el tratamiento con el i-ésimo nivel de "A" (FUNGICIDA), j-ésimo nivel de "C" (INSECTICIDA) y r-ésima repetición.
μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A" (FUNGICIDA).
γ_j = efecto (principal) del nivel j de "C" (INSECTICIDA).
ε_{ijr} = error aleatorio no observable de la r-ésima repetición del tratamiento con el i-ésimo nivel de "A" y el j-ésimo nivel de "C".

y las hipótesis a contrastar serían:

- *Para el efecto (principal) de "A" (fungicida).*

$H_{01}: \alpha_1 = \alpha_2$ vs. $H_{a1}: \alpha_1 \neq \alpha_2$

($H_{01}: \alpha_i = \alpha_{i'}$, para toda i,i') ($H_{a1}: \alpha_i \neq \alpha_{i'}$, para al menos una i,i')

---

(26) Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

- *Para el efecto (principal) de "C" (insecticida).*

H_{02}: $\gamma_1 = \gamma_2$ vs. H_{a2}: $\gamma_1 \neq \gamma_2$

(H_{02}: $\gamma_j = \gamma_{j'}$ para toda j,j') (H_{a2}: $\gamma_j \neq \gamma_{j'}$ para al menos una j,j')

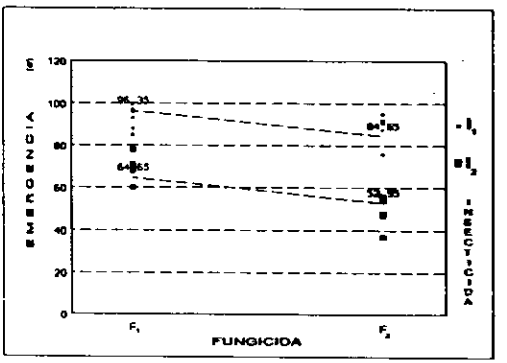
A continuación se muestra parte de la salida del paquete estadístico SYSTAT5 para el modelo anterior, donde puede verse claramente que los datos aportan evidancia estadísticamente significativa de que existe efecto de fungicidas (diferente porcetaje de emergencia de plántulas para los diferentes fungicidas, **FUN**) y efecto de insecticidas (diferente porcentaje de emergencia de plántulas según los diferentes insecticidas, **IN**), de acuerdo con los valores de **"P"** (0.005 y 0.000 respectivamente) y considerando un $\alpha = 0.05$[27].

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
FUN ←	684.450	1	684.450	10.246 →	*0.005*
IN ←	5024.450	1	5024.450	75.21 →	*0.000*
ERROR	1135.650	17	66.803		

Con la estimación de las medias μ_{ij} o de los efectos α_i y γ_j podemos ver que el porcentaje de emergencia de plántulas es mayor con el fungicida "2" (F_2) que con el fungicida 1 (F_1), y también es mayo con el insecticida "2" (I_2) que con el insecticida 1 (I_1). (La gráfica se elaboró con las medias estimadas reportadas en la misma salida).

[27]Para una explicación de la interpretación de "α" consulte la página 9, y las página 49 y 50 para la diferencia con "α_i". Para una explicación de la interpretación de "P" en diseño de experimentos, consulte las páginas 60 o 64.



Lo anterior también puede apreciarse cuando se ajusta el modelo de regresión

$$Y_{ijr} = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2j} + \varepsilon_{ijr}$$

$i = 1,2$[28]

$j = 1,2$

$r = 1,2,3,4,5$

$Z_{1i} = \begin{cases} 1 \text{ si es el fungicida F1} & (i=1) \\ 0 \text{ si NO es el fungicida F1 (es el fungicida F2)} & (i=2) \end{cases}$

$Z_{2j} = \begin{cases} 1 \text{ si es el insecticida I1} & (j=1) \\ 0 \text{ si NO es el insecticida I1 (es el insecticida I2)} & (j=2) \end{cases}$

con

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$

(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de niveles de i de Z_1 y j de Z_2)

donde

Y_{ijr} = EMERGENCIA DE PLÁNTULAS de la r-ésima observación (repetición) del i-ésimo nivel de FUNGICIDA y j-ésimo nivel de INSECTICIDA).
Z_{1i} = i-ésimo nivel de FUNGICIDA ($Z_{11}=1$, $Z_{12}=0$).
Z_{2j} = j-ésimo nivel de INSECTICIDA ($Z_{21}=1$, $Z_{22}=0$).
ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo nivel de FUNGICIDA y j-ésimo nivel de INSECTICIDA.

Dado que ambas variables son cualitativas, el efecto de la variación de ninguna de ellas en la respuesta puede ser modelada por una curva (como sí se hizo en los ejemplos hasta ahora revisados).

---

[28] El uso de las letras y subíndices en las variables indicadoras, es como se describió en la "NOTA 2", página 16, para varias variables cualitativas con sólo dos categorías cada una.

Sin embargo, así como anteriormente se emplearon las variables indicadoras para señalar un cambio en la posición vertical de una curva o recta, ahora señalarán el cambio en posición vertical de un punto, que representa la respuesta para un tratamiento dado, al cambiar el nivel de un factor (o varios a la vez).

De esta manera β_1 representa cambio sobre el eje vertical, el eje de la respuesta, al pasar del fertilizante F2 al F1 (considerando, en principio, los tratamientos con I2), o equivalentemente, como el cambio promedio en la respuesta al pasar del fertilizante F2 al F1 (ambos con I2). Esto es, β_1 representa la diferencia $\mu_{12} - \mu_{22}$, que en ausencia de interacción es igual a $\mu_{1.} - \mu_{2.}$[29].

Por otro lado, β_2 representa el cambio sobre el eje vertical, el eje de la respuesta, al pasar del insecticida I2 al I1 (considerando, en principio, los tratamientos con F2), o equivalentemente, como el cambio promedio en la respuesta al pasar del insecticida I2 al I1 (ambos con F2). Esto es, β_2 representa la diferencia $\mu_{21} - \mu_{22}$, que en ausencia de interacción es igual a $\mu_{.1} - \mu_{.2}$[30].

Por ello β_1, coeficiente asociado a la variable indicadora de fungicida ($Z_{1i}=Z_{11}=1$ si es F1), se interpreta directamente como

$$\text{(regresión)} \qquad \beta_1 = \mu_{12} - \mu_{11} \;::\; \mu_{1.} - \mu_{2.} = \alpha_1 - \alpha_2 \qquad \text{(diseño)}$$

$$\beta_1 = \alpha_1 - \alpha_2$$

y β_2, coeficiente asociado a la variable indicadora de insecticida ($Z_{2j}Z_{21}=1$ si es I1), se interpreta directamente como

$$\text{(regresión)} \qquad \beta_2 = \mu_{21} - \mu_{22} \;::\; \mu_{.1} - \mu_{.2} = \gamma_1 - \gamma_2 \qquad \text{(diseño)}$$

$$\beta_2 = \gamma_1 - \gamma_2$$

Mientras β_0, coeficiente no asociado a variables indicadoras, se interpreta como la respuesta media con el fungicida 2 ($Z_{1i}=Z_{12}=0$) y el insecticida 2 ($Z_{2j}=Z_{22}=0$), esto es "μ_{22}", dado que si $Z_{12}=0$ y $Z_{22}=0$ entonces

$$\mu_{ij} = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2j}$$

---

[29] Dada la ausencia de interacción se cumple que $\mu_{12} - \mu_{22} = \mu_{11} - \mu_{21} = \mu_{1j} - \mu_{2j} = \mu_{1.} - \mu_{2.}$ (página 78).

[30] Dada la ausencia de interacción se cumple que $\mu_{21} - \mu_{22} = \mu_{11} - \mu_{12} = \mu_{i1} - \mu_{i2} = \mu_{.1} - \mu_{.2}$ (página 78).

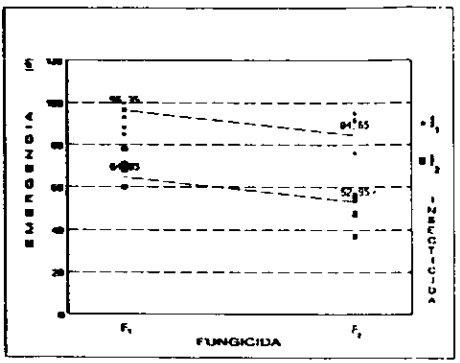
se convierte en $\mu_{22} = \beta_0 + \beta_1 z_{12} + \beta_2 z_{22}$

$\mu_{22} = \beta_0 + \beta_1(0) + \beta_2(0)$

$\mu_{22} = \beta_0$

DEP VAR: EMER N: 20
MULTIPLE R: 0.913
SQUARED MULTIPLE R: 0.834
ADJUSTED SQUARED MULTIPLE R: .815
STANDARD ERROR OF ESTIMATE: 8.173

VARIABLE	COEFF.	T	P(2 TAIL)
CONSTANT	52.95 ← β_0	16.73	→ 0.00
FUN	11.70 ← β_1	3.20	→ 0.01
IN	31.70 ← β_2	8.67	→ 0.00



La salida del paquete estadístico SYSTAT5 para el modelo anterior, nos dice que podemos considerar que los datos de nuestro ejemplo aportan evidencia estadísticamente significativa de que hay diferencia en el porcentaje de emergencia de plántulas debido al efecto de fungicida 1 con respecto al 2 ($P = 0.01$ para ver si β_1 es diferente de cero y considerando un $\alpha = 0.05$), y de que hay diferencia en el porcentaje de emergencia de plántulas debido al efecto de insecticida 1 con respecto al 2 ($P = 0.00$ para ver si β_2 es diferente de cero y considerando un $\alpha = 0.05$)[31].

Sin embargo, no hemos explorado la posibilidad de la interacción, es decir, que el patrón de efectos en la respuesta debida al fungicida varíe de un insecticida al otro, es decir que la diferencia en la respuesta al cambiar el fungicida varíe dependiendo del insecticida empleado.

En un modelo de diseño de experimentos la interacción se expresaría de manera semejante a lo ya visto con variables cualitativas y con variables cuantitativas:

[31] Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

$$Y_{ijr} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \varepsilon_{ijr}$$

$i = 1, 2.$ (i-ésimo nivel de A).
$j = 1, 2.$ (j-ésimo nivel de C).
$r = 1, 2, 3, 4, 5.$ (r-ésima repetición).

con

$$\sum_{i=1}^{2} \alpha_i = \sum_{j=1}^{2} \gamma_j = \sum_{i=1}^{2} \alpha\gamma_{ij} = \sum_{j=1}^{2} \alpha\gamma_{ij} = \sum_{i=1}^{2}\sum_{j=1}^{2} \alpha\gamma_{ij} = 0^{(32)}$$

(suma de efectos, por niveles de A, de C o de ambos es igual a 0)

$\varepsilon_{ijr} \sim NIID(0, \sigma^2)$
(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de niveles de i de A y j de C)

donde

Y_{ijr} = respuesta para el tratamiento con el i-ésimo nivel de "A" (FUNGICIDA), j-ésimo nivel de "C" (INSECTICIDA) y r-ésima repetición.
μ = media general poblacional para la respuesta, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A" (FUNGICIDA).
γ_j = efecto (principal) del nivel j de "C" (INSECTICIDA).
$\alpha\gamma_{ij}$ = efecto de la interacción del nivel i de "A" y el nivel j de "C".
ε_{ijr} = error aleatorio no observable de la r-ésima repetición del tratamiento con el i-ésimo nivel de "A" y el j-ésimo nivel de "C".

y las hipótesis a contrastar serían:

- *Para el efecto (principal) de "A" (fungicida).*

$H_{01}: \alpha_1 = \alpha_2$ vs. $H_{a1}: \alpha_1 \neq \alpha_2$

($H_{01}: \alpha_i = \alpha_{i'}$ para toda i,i') ($H_{a1}: \alpha_i \neq \alpha_{i'}$ para al menos una i,i')

- *Para el efecto (principal) de "C" (insecticida).*

$H_{02}: \gamma_1 = \gamma_2$ vs. $H_{a2}: \gamma_1 \neq \gamma_2$

($H_{02}: \gamma_j = \gamma_{j'}$ para toda j,j') ($H_{a2}: \gamma_j \neq \gamma_{j'}$ para al menos una j,j')

- *Para el efecto de interacción entre "A" y "C".*

$H_{03}: \alpha\gamma_{11}=\alpha\gamma_{12}=\alpha\gamma_{21}=\alpha\gamma_{22}=0$ vs. $H_{a3}: \alpha\gamma_{ij}\neq 0$ al menos para una combinación ij

($H_{03}: \alpha\gamma_{ij}=0$ para toda combinación ij)

---

(32) Esto no es un supuesto del modelo, sino una "restricción colateral" (ver nota 13, página 49).

A continuación se muestra parte de la salida del paquete estadístico SYSTAT5 para el modelo anterior, donde puede verse claramente que los datos aportan evidancia estadísticamente significativa de que, además de haber un efecto de fungicidas (diferente porcentaje de emergencia de plántulas para los diferentes fungicidas, **FUN**) y de insecticidas (diferente porcentaje de emergencia de plántulas para los diferentes insecticidas, **IN**), existe un efecto de interacción (**FUN*IN**), de acuerdo con los valores de "P" (0.001, 0.000 y 0.008 respectivamente) y considerando un $\alpha = 0.05$[33].

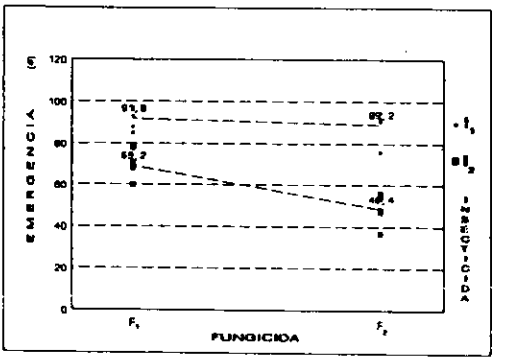
Esto es, la variación en el porcentaje de emergencia de plántulas cuando se emplea uno u otro fungicida es diferente para cada insecticida empleado, y viceversa, la variación en el porcentaje de emergencia de plántulas cuando se emplea uno u otro insecticida es diferente para cada fungicida empleado.

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
FUN ←	684.450	1	684.450	15.176 →	*0.001*
IN ←	5024.450	1	5024.450	111.407 →	*0.000*
FUN*IN ←	414.050	1	414.050	9.181 →	*0.008*
ERROR	721.600	16	45.100		

A partir de las estimaciones de μ_{ij} o de los efectos α_i, γ_j y $\alpha\gamma_{ij}$, podemos ver que el patrón de variación en la respuesta de F1 a F2 en el tipo "2" de insecticida (I2) es mínimo (casi el mismo valor de respuesta), mientras que nivel el patrón de variación en la respuesta de F1 a F2 en el tipo "1" de insecticida (I1) es mucho mayor (casi se duplica el valor de la respuesta). (La gráfica se elaboró con las medias estimadas reportadas en la misma salida).

[33] Para una explicación de la interpretación de "α" consulte la página 9, y las página 49 y 50 para la diferencia con "α_i". Para una explicación de la interpretación de "P" en diseño de experimentos, consulte las páginas 60 o 64.



Para el caso del modelo de regresión, este mismo ajuste se realizó a través de

$$Y_{ijr} = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2j} + \beta_3 Z_{1i} Z_{2j} + \varepsilon_{ijr}$$

$i = 1,2$[34]

$j = 1,2$

$r = 1,2,3,4,5$

$$Z_{1i} = \begin{cases} 1 \text{ si es el fungicida F1} & (i=1) \\ 0 \text{ si NO es el fungicida F1 (es el fungicida F2)} & (i=2) \end{cases}$$

$$Z_{2j} = \begin{cases} 1 \text{ si es el insecticida I1} & (j=1) \\ 0 \text{ si NO es el insecticida I1 (es el insecticida I2)} & (j=2) \end{cases}$$

con

$\varepsilon_{ijr} \sim NIID(0,\sigma^2)$

(Normales, Independientes e Idénticamente Distribuidos)
(los errores tienen distribución Normal en cada combinación de niveles de i de Z_1 y j de Z_2)

donde

Y_{ijr} = EMERGENCIA DE PLÁNTULAS de la r-ésima observación (repetición) del i-ésimo nivel de FUNGICIDA y j-ésimo nivel de INSECTICIDA).
Z_{1i} = i-ésimo nivel de FUNGICIDA.
Z_{2j} = j-ésimo nivel de INSECTICIDA.
ε_{ijr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo nivel de FUNGICIDA y j-ésimo nivel de INSECTICIDA.

donde los coeficientes asociados a Z_{1i} pueden revelar un efecto de fungicidas (diferente porcetaje de emergencia de plántulas para los diferentes fungicidas)[35], los coeficientes asociados a Z_{2j} pueden revelar un efecto de insecticidas (diferente porcentaje de emergencia de plántulas según los diferentes insecticidas)[35], y β_3 revela la existencia de la interacción, pues es el coeficiente asociado a las dos variables, y por ello el término que lo contiene varía tanto con cambio en fungicida como con cambio en insecticida.

---

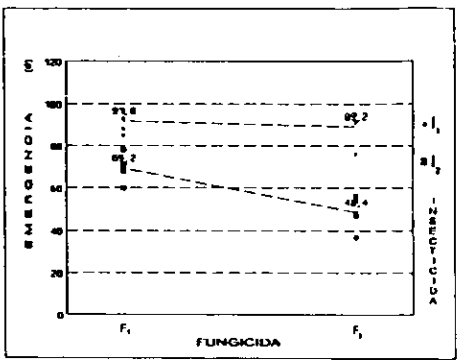
[34]El uso de las letras y subíndices en las variables indicadoras, es como se describío en la "NOTA 2", página 16, para varias variables cualitativas con sólo dos categorías cada una.

[35]Consulte la "Nota 5", páginas 85-87, para la equivalencia de tales coeficientes con los términos en los modelos de diseño de experimentos.

Dado que la interacción en modelos de diseños de experimentos es la diferencia en el patrón de la respuesta dada por cambio de niveles de un factor al cambiar los niveles de otro factor, $\beta_3 Z_{1i} Z_{2j}$ representa la interacción.

DEP VAR: EMER N: 20
MULTI. R: 0.95 SQ. MULTI.R: 0.90
ADJUSTED SQUARED MULTIPLE R: .875
STANDARD ERROR OF ESTIMATE: 6.716

VARIABLE	COEFF.	T	P(2 TAIL)
CONSTANT	48.4 ← β_0	16.12	→ *0.00*
FUN	20.8 ← β_1	4.90	→ *0.00*
IN	40.8 ← β_2	9.61	→ *0.00*
FUN*IN	-18.2 ← β_3	-3.03	→ *0.01*



La salida del paquete estadístico SYSTAT5 para el modelo anterior, nos dice que podemos considerar que los datos de nuestro ejemplo aportan evidencia estadísticamente significativa de que hay diferencia en el porcentaje de emergencia de plántulas debido al efecto de fungicida 1 con respecto al 2 (P = 0.00 para ver si β_1 es diferente de cero) en tratamientos con insecticida 2, de que hay diferencia en el porcentaje de emergencia de plántulas debido al efecto de insecticida 1 con respecto al 2 (P = 0.00 para ver si β_2 es diferente de cero) en tratamientos con fungicida 2, y que tales diferencias varían en los tratamientos con insecticida 1 y fungicida 1 respectivamente (P = 0.01 para ver si β_3 es diferente de cero), lo que representa un efecto de interacción entre fungicidas e insecticidas (todo considerando un $\alpha = 0.05$[36]). (La gráfica se elaboró con los coeficientes estimados reportados en la misma salida).

Esto es, la variación en el porcentaje de emergencia de plántulas cuando se emplea uno u otro fungicida es diferente para cada insecticida empleado, y viceversa, la variación en el porcentaje de emergencia de plántulas cuando se emplea uno u otro insecticida es diferente para cada fungicida empleado.

[36] Para una explicación de la interpretación de "P" y "α" consulte las páginas 8 y 9.

Si no hubiese interacción, y tomando el modelo para la "media por tratamiento"

$$\mu_{ij} = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2j}$$

con

$i = 1,2$[37]

$$z_{1i} = \begin{cases} 1 \text{ si es el fungicida F1} & (i=1) \\ 0 \text{ si NO es el fungicida F1 (es el fungicida F2)} & (i=2) \end{cases}$$

$j = 1,2$

$$z_{2j} = \begin{cases} 1 \text{ si es el insecticida I1} & (j=1) \\ 0 \text{ si NO es el insecticida I1 (es el insecticida I2)} & (j=2) \end{cases}$$

la variación en la respuesta por cambio de fungicida sería β_1 en cualquier nivel de insecticida:

F_2, I_2

$\mu_{22} = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$ *Diferencia por cambio en F considerando I_2:* β_1

F_1, I_2

$\mu_{12} = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$

F_2, I_1

$\mu_{21} = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$ *Diferencia por cambio en F considerando I_1:* β_1

F_1, I_1

$\mu_{11} = \beta_0 + \beta_1(1) + \beta_2(1) = \beta_0 + \beta_1 + \beta_2$

Igualmente, de no haber interacción la variación en la respuesta por el cambio de insecticida sería β_2 en cualquier nivel de fungicida:

F_2, I_2

$\mu_{22} = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$ *Diferencia por cambio en I considerando F_2:* β_2

F_2, I_1

$\mu_{21} = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$

F_1, I_2

$\mu_{12} = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$ *Diferencia por cambio en I considerando F_1:* β_2

F_1, I_1

$\mu_{11} = \beta_0 + \beta_1(1) + \beta_2(1) = \beta_0 + \beta_1 + \beta_2$

---

[37] El uso de las letras y subíndices en las variables indicadoras, es como se describió en la "NOTA 2", página 16, para varias variables cualitativas con sólo dos categorías cada una.

Con la presencia del término $\beta_3 z_{1i} z_{2j}$ al modelo anterior, la variación en la respuesta por cambio de fungicida es β_1 sólo en el nivel "2" de insecticida, mientras que es de $\beta_1+\beta_3$ en el nivel "1" de insecticida:

- *Diferencia por cambio en F considerando* I_2:

β_1

F_2, I_2

$$\mu_{22} = \beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0)(0) = \beta_0$$

F_1, I_2

$$\mu_{12} = \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(1)(0) = \beta_0 + \beta_1$$

- *Diferencia por cambio en F considerando* I_1:

$\beta_1 + \beta_3$

F_2, I_1

$$\mu_{21} = \beta_0 + \beta_1(0) + \beta_2(1) + \beta_3(0)(1) = \beta_0 + \beta_2$$

F_1, I_1

$$\mu_{11} = \beta_0 + \beta_1(1) + \beta_2(1) + \beta_3(1)(1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

De manera correspondiente, la variación en la respuesta por el cambio de insecticida es β_2 sólo en el nivel "2" de fungicida, mientras que es de $\beta_2+\beta_3$ en el nivel "1" de fungicida.

- *Diferencia por cambio en I considerando* F_2:

β_2

F_2, I_2

$$\mu_{22} = \beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0)(0) = \beta_0$$

F_2, I_1

$$\mu_{21} = \beta_0 + \beta_1(0) + \beta_2(1) + \beta_3(0)(1) = \beta_0 + \beta_2$$

- *Diferencia por cambio en I considerando* F_1:

$\beta_2 + \beta_3$

F_1, I_2

$$\mu_{12} = \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(1)(0) = \beta_0 + \beta_1$$

F_1, I_1

$$\mu_{11} = \beta_0 + \beta_1(1) + \beta_2(1) + \beta_3(1)(1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

Por consiguiente, si β_3 es diferentes de cero, entonces podemos decir que sí hay interacción entre fungicida (A en diseños) e insecticida (C en diseños) para el modelo de la página 76.

ESTA TESIS NO DEBE
SALIR DE LA BIBLIOTECA

Dado que la interacción en modelos de diseños de experimentos es la diferencia en el patrón de variación en la respuesta dada por cambio de niveles de un factor al cambiar los niveles de otro factor, y que el patrón de variación en la respuesta respecto a una variable cualitativa (categórica) puede modelarse en regresión con una varible indicadora o "dummy" (para cada categoría menos una, pag. 13 y 14), la interacción se podrá modelar con un término que haga variar el patrón de respuesta al variar simultáneamente ambas variables, esto es, un término que involucre a las variables indicadoras de ambas variables categóricas.

De esta manera, si el coeficiente asociado al término que incluye a ambas variables es diferentes de cero, podemos decir que sí hay interacción entre fungicida (A en diseños) e insecticida (B en diseños), dado que tal término varía tanto con el cambio en Z_1 como con el cambio en Z_2, es decir, si β_3 es diferente de cero entonces el modelo

$$Y_{ijr} = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2j} + \beta_3 Z_{1i} Z_{2j} + \varepsilon_{ijr}$$

equivale a

$$Y_{ijr} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \varepsilon_{ijr}$$

Para los modelos anteriores

$\beta_3 \neq 0$	$\longleftrightarrow$	$\alpha\gamma_{ij} \neq 0$
(efecto por cambio simultáneo de nivel en ambos factores)		*al menos para una combinación "ij"*

Recapitulando sobre las diferencias en modelado e interpretación de la interacción entre Diseño de Experimentos y Regresión encontradas en los tres ejemplos anteriores, pudo apreciarse que en Diseño de Experimentos la interacción se modelaba de la misma forma independientemente del tipo de variable, mientras que en los modelos de regresión la interacción se modelaba de diferentes maneras dependiendo del tipo de variable (y la forma de la curva, en el caso de variables cuantitativas).

Nota 3. *Uso de variables indicadoras para representar un diseño de experimentos completamente al azar.*

De manera general, para ***reexpresar un modelo de diseño de experimentos completamente al azar*** con un factor cualitativo (variable categórica) en un modelo de regresión, **se emplean tantas variables indicadoras** "dummy" como categorías o niveles tenga la variable categórica original menos una (pag. 14).

El término en que se encuentre una variable indicadora señalará el cambio en posición vertical de un punto, punto que representa el valor de la respuesta para un tratamiento dado, al cambiar al nivel del factor (tratamiento) descrito por esa variable (página 72).

Dado que una de las categorías o tratamientos no está expresado por variables indicadoras, tal tratamiento será descrito, junto con la media general del modelo de diseño de experimentos, con el término independiente en el modelo de regresión (páginas 72 y 73).

De esta manera, el modelo de diseño de experimentos

$$Y_{ir} = \mu + \tau_i + \varepsilon_{ir}$$

$i = 1, 2, \ldots, t.$ (i-ésimo tratamiento o nivel de T)
$r = 1, 2, \ldots, n_i.$ (r-ésima repetición).

con

$$\sum_{i=1}^{t} \tau_i = 0^* \qquad \varepsilon_{ir} \sim NIID(0, \sigma^2)$$

*Restricción colateral (ver nota 13, pag. 49)

donde

Y_{ir} = respuesta para el tratamiento del i-ésimo tratamiento (nivel de "T"), r-ésima repetición.
μ = media general poblacional, dada por los factores comunes no en estudio.
τ_i = efecto del i-ésimo tratamiento (nivel de "T").
ε_{ir} = error aleatorio no observable de la r-ésima repetición del i-ésimo tratamiento (nivel de "T").

puede reexpresarse en el modelo de regresión

$$Y_{ir} = \beta_0 + \beta_1 z_{1j} + \beta_2 z_{2j} + \beta_3 z_{3j} + \cdots + \beta_{(t-1)} z_{(t-1)j} + \varepsilon_{ir}$$

ó

$$Y_{ir} = \beta_0 + \sum_{i=1}^{t-1} \beta_i z_{ij} + \varepsilon_{ir}$$

$i = 1, 2, \ldots, t$ $\quad j = 1, 2$ $\quad r = 1, 2, \ldots, n_i$

$$z_{1j} = \begin{cases} 1 \text{ si es el tratamiento 1} & (j=1) \\ 0 \text{ si NO es el trat. 1} & (j=2) \end{cases} \qquad z_{2j} = \begin{cases} 1 \text{ si es el tratamiento 2} & (j=1) \\ 0 \text{ si NO es el trat. 2} & (j=2) \end{cases}$$

$$\ldots\ldots \qquad z_{(t-1)j} = \begin{cases} 1 \text{ si es el tratamiento t-1} & (j=1) \\ 0 \text{ si NO es el trat. t-1} & (j=2) \end{cases}$$

con

$$\varepsilon_{ir} \sim NIID(0, \sigma^2)$$

(continúa en la siguiente página)

Nota 3 *(continuación)*

donde

Y_{ir} = RESPUESTA de la r-ésima observación (repetición) del i-ésimo tratamiento.
Z_i = variable indicadora del i-ésimo tratamiento.
ε_{ir} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo tratamiento.

De esta manera β_0, coeficiente no asociado a variables indicadoras, se interpreta como la respuesta media con el tratamiento "t" (μ_t), ya que si $Z_i=0$ para toda "i", entonces

$$\mu_i = \beta_0 + \sum_{i=1}^{t-1} \beta_i Z_{ij}$$

se convierte en

$$\mu_t = \beta_0 + \sum_{i=1}^{t-1} \beta_i Z_{i1}$$

$$\mu_t = \beta_0 + \sum_{i=1}^{t-1} \beta_i (0)$$

$$\mu_t = \beta_0$$

y dado que

$$\mu_t = \mu + \tau_t$$

tenemos también

$$\beta_0 = \mu + \tau_t$$

Mientras que β_k, coeficiente asociado a la variable indicadora del tratamiento i=k, se interpreta como el cambio promedio en la respuesta al pasar del tratamiento "t" al tratamiento k-ésimo, ya que si $Z_i=0$ para toda $i \neq k$, entonces

$$\mu_i = \beta_0 + \sum_{i=1}^{t-1} \beta_i Z_{ij}$$

se convierte en

$$\mu_k = \beta_0 + \sum_{i=1}^{t-1} \beta_i Z_{i2} + \beta_k Z_{k1} \quad \text{con } i \neq k$$

$$\mu_k = \beta_0 + \sum_{i=1}^{t-1} \beta_i (0) + \beta_k (1) \quad \text{con } i \neq k$$

$$\mu_k = \beta_0 + \beta_k$$

y dado que

$$\beta_0 = \mu_t$$

tenemos las equivalencias

$$\beta_k = \mu_k - \mu_t$$

$$\beta_k = \tau_k - \tau_t$$

Por ello β_i, coeficiente asociado a la variable indicadora del tratamiento "i", se interpreta directamente como

(regresión) $\beta_i = \mu_i - \mu_t$ (diseño)
$= \tau_i - \tau_t$

donde "i" es cualquiera de los t-1 tratamientos representados con las variables indicadoras, diferencia que en diseño de experimentos se conoce como "contraste".

Nota 4. ***Uso de variables indicadoras para representar un diseño de experimentos factorial "a x c" sin interacción.***

De manera general, para ***reexpresar un modelo de diseño de experimentos factorial "a x c" sin interacción***, dos factores cualitativo (variables categóricas) con "a" categorías o niveles para la primer variable o factor (A) y "c" para la segunda (C), en un modelo de regresión, se emplean tantas variables indicadoras "dummy" como categorías o niveles tenga la variable categórica original menos una por variable, es decir "(a-1)+(c-1)" ó "a+c-2" (pág. 14).

El término en que se encuentre una variable indicadora señalará el cambio en posición vertical de un punto, punto que representa el valor de la respuesta para un tratamiento dado por la combinación de niveles de los dos factores, al cambiar al nivel del factor descrito por esa variable (página 72).

Dado que una de las categorías o niveles de cada factor no está expresada por variables indicadoras, tales niveles será descrito, junto con la media general del modelo de diseño de experimentos, con el término independiente en el modelo de regresión (páginas 72 y 73).

De esta manera, el modelo de diseño de experimentos

$$Y_{ikr} = \mu + \alpha_i + \gamma_k + \varepsilon_{ikr} \qquad \begin{matrix} i = 1, 2, \ldots, a. \\ k = 1, 2, \ldots, c. \\ r = 1, 2, \ldots, n_{ik} \end{matrix}$$

con $\sum_{i=1}^{a}\alpha_i = \sum_{k=1}^{c}\gamma_j = 0^*$ $\qquad \varepsilon_{ir} \sim NIID(0,\sigma^2)$

*Restricción colateral (ver nota 13, pag. 49)

donde

Y_{ikr} = respuesta para el tratamiento con el i-ésimo nivel de "A", k-ésimo nivel de "C" y r-ésima repetición.
μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A".
γ_k = efecto (principal) del nivel k de "C".
ε_{ikr} = error aleatorio no observable de la r-ésima repetición del tratamiento con el i-ésimo nivel de "A" y el k-ésimo nivel de "C".

puede reexpresarse en el modelo de regresión

$$Y_{ikr} = \beta_0 + \underbrace{\beta_1 z_{1j} + \cdots + \beta_{(a-1)} z_{(a-1)j}}_{\text{para a-1 categorías de A}} + \underbrace{\beta_a s_{1j} + \cdots + \beta_{(a+c-2)} s_{(c-1)j}}_{\text{para c-1 categorías de C}} + \varepsilon_{ikr}$$

ó

$$Y_{ikr} = \beta_0 + \sum_{i=1}^{a-1}\beta_i z_{ij} + \sum_{k'=a,k=1}^{a+c-2,c-1}\beta_{k'} s_{kj} + \varepsilon_{ikr}$$

$i = 1,2,\ldots,a \quad j = 1,2 \quad k' = a,\ldots a+c-2 \quad k = 1,2,\ldots,c \quad r = 1,2,\ldots,n_{ik}$

$z_{1j} = \begin{cases} 1 \text{ si es la categ. 1 de A (j=1)} \\ 0 \text{ si NO es la cat.1 de A (j=2)} \end{cases} \cdots z_{(a-1)j} = \begin{cases} 1 \text{ si es la categ. a-1 de A (j=1)} \\ 0 \text{ si NO es la cat.a-1 de A (j=2)} \end{cases}$

$s_{1j} = \begin{cases} 1 \text{ si es la categ. 1 de C (j=1)} \\ 0 \text{ si NO es la cat.1 de C (j=2)} \end{cases} \cdots s_{(c-1)j} = \begin{cases} 1 \text{ si es la categ. c-1 de C (j=1)} \\ 0 \text{ si NO es la cat.c-1 de C (j=2)} \end{cases}$

con $\varepsilon_{ikr} \sim NIID(0,\sigma^2)$

(continúa el la siguiente página)

donde
Y_{ikr} = RESPUESTA de la r-ésima observación (repetición) del nivel i de "A" y k de "C".
Z_{ij} = i-ésimo nivel de "A".
Z_{kj} = k-ésimo nivel de "C".
ε_{ikr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo nivel de "A" y k-ésimo nivel de "C".

De esta manera β_0, coeficiente no asociado a variables indicadoras, se interpreta como la respuesta media con el tratamiento resultado de la combinación de los niveles "a" de la variable "A" y "c" de la variable "C" (μ_{ac}), ya que si $Z_i=0$ para toda "i" y $S_k=0$ para toda "k", entonces

$$\mu_{ij} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{ij} + \sum_{k=1}^{c-1} \beta_k S_{kj}$$

se convierte en

$$\mu_{ac} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{i2} + \sum_{k=1}^{c-1} \beta_k S_{k2}$$

$$\mu_{ac} = \beta_0 + \sum_{i=1}^{a-1} \beta_i (0) + \sum_{k=1}^{c-1} \beta_k (0)$$

$$\mu_{ac} = \beta_0$$

y dado que

$$\mu_{ac} = \mu + \alpha_a + \gamma_c$$

tenemos también

$$\beta_0 = \mu + \alpha_a + \gamma_g$$

Mientras que $\beta_{i'}$, coeficiente asociado a la variable indicadora del tratamiento i=i', se interpreta como el cambio promedio en la respuesta al pasar del tratamiento "ac" al tratamiento "i'c", ya que si $Z_i=0$ para toda $i \neq i'$ y $S_k=0$ para toda "k", entonces

$$\mu_{ij} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{ij} + \sum_{k=1}^{c-1} \beta_k S_{kj}$$

se convierte en

$$\mu_{i'c} = \beta_0 + \sum_{i=1 \text{ con } i \neq i'}^{a-1} \beta_i Z_{i2} + \beta_{i'} Z_{i'2} + \sum_{k=1}^{c-1} \beta_k S_{k2}$$

$$\mu_{i'c} = \beta_0 + \sum_{i=1 \text{ con } i \neq i'}^{a-1} \beta_i (0) + \beta_{i'} (1) + \sum_{k=1}^{c-1} \beta_k (0)$$

$$\mu_{i'c} = \beta_0 + \beta_{i'}$$

y dado que

$$\beta_0 = \mu_{ac}$$

tenemos las equivalencias

$$\beta_{i'} = \mu_{i'c} - \mu_{ac}$$

$$\beta_{i'} = \alpha_{i'} - \alpha_a$$

Por ello β_i, coeficiente asociado a la variable indicadora del nivel "i", se interpreta directamente como

(regresión) $\beta_i = \mu_{ic} - \mu_{ac} = \alpha_i - \alpha_a$ (diseño)

donde "i" es cualquiera de los a-1 niveles de "A" representados con las variables indicadoras, diferencia que se conoce en diseños de experimentos como "contraste".

De la misma forma β_k, coeficiente asociado a la variable indicadora del nivel "k", se interpreta directamente como

(regresión) $\beta_k = \mu_{ak} - \mu_{ac} = \gamma_k - \gamma_a$ (diseño)

donde "k" es cualquiera de los c-1 niveles de "C" representados con las variables indicadoras.

Nota 5. ***Uso de variables indicadoras para representar un diseño de experimentos factorial "a x c" con interacción.***

De manera general, ***para reexpresar un modelo de diseño de experimentos factorial "a x c" con interacción,*** dos factores cualitativo (variables categóricas) con "a" categorías o niveles para la primer variable o factor (A) y "c" para la segunda (C), en un modelo de regresión, se emplean tantas variables indicadoras "dummy" como categorías o niveles tenga la variable categórica original menos una por variable para los efectos principales, es decir "(a-1)+(c-1)", y "(a-1)x(c-1)" variables indicadores para las interacciones entre tales categorías.

El término en que se encuentre una o más variable indicadoras señalará el cambio en posición vertical de un punto, punto que representa el valor de la respuesta para un tratamiento dado por la combinación de niveles de los dos factores, al cambiar al nivel del factor o factores descrito(s) por la(s) variable(s) en ese término.

Dado que una de las categorías o niveles de cada factor no está expresada por variables indicadoras, tales niveles será descrito, junto con la media general del modelo de diseño de experimentos, con el término independiente en el modelo de regresión.

De esta manera, el modelo de diseño de experimentos

$$Y_{ikr} = \mu + \alpha_i + \gamma_k + \alpha\gamma_{ik} + \varepsilon_{ikr}$$

$i = 1, 2, \ldots, a.$ (i-ésimo nivel de A).
$k = 1, 2, \ldots, c.$ (k-ésimo nivel de C).
$r = 1, 2, \ldots, n_{ik}.$ (r-ésima repetición).

con

$$\sum_{i=1}^{a} \alpha_i = \sum_{k=1}^{c} \gamma_k = \sum_{i=1}^{a} \alpha\gamma_{ik}^{*} = \sum_{k=1}^{c} \alpha\gamma_{ik}^{*} = \sum_{i=1}^{a}\sum_{k=1}^{c} \alpha\gamma_{ik}^{*} = 0^{*} \qquad \varepsilon_{ijr} \sim NIID(0, \sigma^2)$$

*Restricción colateral (ver nota 13, pag. 49)

donde

Y_{ikr} = respuesta para el tratamiento con el i-ésimo nivel de "A", j-ésimo nivel de "C" y r-ésima repetición.
μ = media general poblacional de la respuesta, dada por los factores comunes no en estudio.
α_i = efecto (principal) del nivel i de "A".
γ_k = efecto (principal) del nivel j de "C".
$\alpha\gamma_{ik}$ = efecto de la interacción del nivel i de "A" y el nivel k de "C".
ε_{ikr} = error aleatorio no observable de la r-ésima repetición del tratamiento con el i-ésimo nivel de "A" y el k-ésimo nivel de "C".

puede reexpresarse en el modelo de regresión

(continúa en la siguiente página)

Nota 5 *(continuación)*

$$Y_{ikr} = \beta_0 + \underbrace{\beta_1 Z_{1j} + \cdots + \beta_{(a-1)} Z_{(a-1)j}}_{\text{para a-1 categorías de A}} + \underbrace{\beta_a S_{1j} + \cdots + \beta_{(a+c-2)} S_{(c-1)j}}_{\text{para g-1 categorías de C}} +$$

$$+ \underbrace{\beta_{(a+c-1)} Z_{ij} S_{1j} + \cdots + \beta_{(a*c-1)} Z_{(a-1)j} S_{(c-1)j}}_{\text{para (a-1)*(c-1) términos de interacción}} + \varepsilon_{ikr}$$

ó

$$Y_{ikr} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{ij} + \sum_{k'=a,k=1}^{a+g-2,c-1} \beta_{k'} S_{kj} + \sum_{\substack{i=1 \\ l=a+c-1}}^{a*c-1} \sum_{k=1}^{a-1\ c-1} \beta_l Z_{ij} S_{kj} + \varepsilon_{ikr}$$

$i = 1, 2, \ldots, a \qquad j = 1, 2 \qquad k = 1, 2, \ldots, g \qquad r = 1, 2, \ldots, n_{ik}$

$Z_{1j} = \begin{cases} 1 \text{ si es la categ. 1 de A } (j=1) \\ 0 \text{ si NO es la cat.1 de A } (j=2) \end{cases} \cdots Z_{(a-1)j} = \begin{cases} 1 \text{ si es la categ. a-1 de A } (j=1) \\ 0 \text{ si NO es la cat.a-1 de A } (j=2) \end{cases}$

$S_{1j} = \begin{cases} 1 \text{ si es la categ. 1 de C } (j=1) \\ 0 \text{ si NO es la cat.1 de C } (j=2) \end{cases} \cdots S_{(g-1)j} = \begin{cases} 1 \text{ si es la categ. c-1 de C } (j=1) \\ 0 \text{ si NO es la cat.c-1 de C } (j=2) \end{cases}$

con

$$\varepsilon_{ikr} \sim NIID(0, \sigma^2)$$

donde

- Y_{ikr} = RESPUESTA de la r-ésima observación (repetición) del i-ésimo nivel de A y k-ésimo nivel de C.
- Z_{ij} = i-ésimo nivel de A.
- Z_{kj} = j-ésimo nivel de C.
- ε_{ikr} = error aleatorio no observable de la r-ésima observación (repetición) del i-ésimo nivel de A y k-ésimo nivel de C.

De esta manera β_0, coeficiente no asociado a variables indicadoras, se interpreta como la respuesta media con el tratamiento resultado de la combinación de los niveles "a" de la variable "A" y "c" de la variable "C" (μ_{ac}), ya que si $Z_i=0$ para toda "i" y $S_k=0$ para toda "k", entonces

$$\mu_{ik} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{ij} + \sum_{k=1}^{c-1} \beta_k S_{kj} + \sum_{i=1k=1}^{a-1c-1} \sum \beta_l Z_{ij} S_{kj}$$

se convierte en

$$\mu_{ac} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{i2} + \sum_{k=1}^{c-1} \beta_k S_{k2} + \sum_{i=1k=1}^{a-1c-1} \sum \beta_l Z_{i2} S_{k2}$$

$$\mu_{ac} = \beta_0 + \sum_{i=1}^{a-1} \beta_i (0) + \sum_{k=1}^{c-1} \beta_k (0) + \sum_{i=1k=1}^{a-1c-1} \sum \beta_l (0)(0)$$

$$\mu_{ac} = \beta_0$$

y dado que

$$\mu_{ac} = \mu + \alpha_a + \gamma_c + \alpha\gamma_{ac}$$

tenemos también

$$\beta_0 = \mu_{ac}$$

$$\beta_0 = \mu + \alpha_a + \gamma_c + \alpha\gamma_{ac}$$

Mientras que $\beta_{i'}$, coeficiente asociado a la variable indicadora del nivel i=i', se interpreta como el cambio promedio en la respuesta al pasar del tratamiento "ac" al tratamiento "i'c", ya que si $Z_i=0$ para toda i≠i' y $S_k=0$ para toda "k", entonces

(continúa en la siguiente página)

$$\mu_{ij} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{ij} + \sum_{k=1}^{c-1} \beta_k S_{kj} + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l Z_{ij} S_{kj}$$

se convierte en

$$\mu_{i'c} = \beta_0 + \sum_{\substack{i=1 \\ \text{con } i \neq i'}}^{a-1} \beta_i Z_{i2} + \beta_{i'} Z_{i'1} + \sum_{k=1}^{c-1} \beta_k S_{k2} + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l Z_{i2} S_{k2} + \beta_l Z_{i'1} S_{k2}$$

$$\mu_{i'c} = \beta_0 + \sum_{\substack{i=1 \\ \text{con } i \neq i'}}^{a-1} \beta_i (0) + \beta_{i'} (1) + \sum_{k=1}^{c-1} \beta_k (0) + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l (0)(0) + \beta_l (1)(0)$$

$$\mu_{i'c} = \beta_0 + \beta_{i'}$$

y dado que

$$\beta_0 = \mu_{ac} = \mu + \alpha_a^* + \gamma_c + \alpha\gamma_{ac} \qquad \mu_{i'c} = \mu + \alpha_{i'} + \gamma_c + \alpha\gamma_{i'c}$$

tenemos las equivalencias

$$\beta_{i'} = \mu_{i'c} - \beta_0 = \mu_{i'c} - \mu_{ac} \qquad \beta_{i'} = \alpha_{i'} - \alpha_a + \alpha\gamma_{i'c} - \alpha\gamma_{ac}$$

De manera similar $\beta_{k'}$, coeficiente asociado a la variable indicadora del nivel k=k', se interpreta como el cambio promedio en la respuesta al pasar del tratamiento "ac" al tratamiento "ak'", ya que si $S_k=0$ para toda $k \neq k'$ y $Z_i=0$ para toda "i", entonces

$$\mu_{ij} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{ij} + \sum_{k=1}^{c-1} \beta_k S_{kj} + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l Z_{ij} S_{kj}$$

se convierte en

$$\mu_{ak'} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{i2} + \sum_{\substack{k=1 \\ \text{con } k \neq k'}}^{c-1} \beta_k S_{k2} + \beta_{k'} S_{k'1} + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l Z_{i2} S_{k2} + \beta_l Z_{i2} S_{k'1}$$

$$\mu_{ak'} = \beta_0 + \sum_{i=1}^{a-1} \beta_i (0) + \sum_{\substack{k=1 \\ \text{con } k \neq k'}}^{c-1} \beta_k (0) + \beta_{k'} (1) + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l (0)(0) + \beta_l (0)(1)$$

$$\mu_{ak'} = \beta_0 + \beta_{k'}$$

y dado que

$$\beta_0 = \mu_{ac} = \mu + \alpha_a + \gamma_c + \alpha\gamma_{ac} \qquad \mu_{ak'} = \mu + \alpha_a + \gamma_{k'} + \alpha\gamma_{ak'}$$

tenemos las equivalencias

$$\beta_{k'} = \mu_{ik'} - \beta_0 = \mu_{ik'} - \mu_{ac} \qquad \beta_{k'} = \gamma_{k'} - \gamma_c + \alpha\gamma_{ik'} - \alpha\gamma_{ac}$$

(continúa en la siguiente página)

Nota 5 *(continuación)*

Por su parte β_i (o $\beta_{ik'}$), coeficiente asociado a la variable indicadora del nivel i=i' y a la variable indicadora de nivel k=k', se interpreta como el cambio promedio en la respuesta al pasar del tratamiento "ac" al tratamiento "i'k'", ya que si $Z_i=0$ para toda i≠i' y $S_k=0$ para toda k≠k', entonces

$$\mu_{ij} = \beta_0 + \sum_{i=1}^{a-1} \beta_i Z_{ij} + \sum_{k=1}^{c-1} \beta_k S_{kj} + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l Z_{ij} S_{kj}$$

se convierte en

$$\mu_{i'k'} = \beta_0 + \sum_{\text{con } i\neq i'}^{a-1} \beta_i Z_{i2} + \beta_{i'} Z_{i'1} + \sum_{\text{con } k\neq k'}^{c-1} \beta_k S_{k2} + \beta_{k'} S_{k'1} + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l Z_{i2} S_{k2} + \beta_l Z_{i'1} S_{k'1}$$

$$\mu_{i'k'} = \beta_0 + \sum_{\text{con } i\neq i'}^{a-1} \beta_i (0) + \beta_{i'} (1) + \sum_{\text{con } k\neq k'}^{c-1} \beta_k (0) + \beta_{k'} (1) + \sum_{i=1}^{a-1}\sum_{k=1}^{c-1} \beta_l (0)(0) + \beta_l (1)(1)$$

$$\mu_{i'k'} = \beta_0 + \beta_{i'} + \beta_{k'} + \beta_l$$

y dado que

$$\beta_0 = \mu_{ac} \qquad \beta_{i'} = \alpha_{i'} - \alpha_a + \alpha\gamma_{i'c} - \alpha\gamma_{ac}$$

$$\beta_{k'} = \gamma_{k'} - \gamma_c + \alpha\gamma_{ak'} - \alpha\gamma_{ac} \qquad \mu_{i'k'} = \mu + \alpha_{i'} + \gamma_{k'} + \alpha\gamma_{i'k'}$$

tenemos las equivalencias

$$\beta_l = \mu_{i'k'} - \beta_0 - \beta_{i'} - \beta_{k'} \qquad \beta_l = \alpha\gamma_{i'k'} + \alpha\gamma_{ac} - \alpha\gamma_{i'c} - \alpha\gamma_{ak'}$$

Por ello β_i, coeficiente asociado a la variable indicadora de la categoría "i", se interpreta directamente como

(regresión) $\beta_i = \mu_{ic} - \mu_{ac} = \alpha_i - \alpha_a + \alpha\gamma_{ic} - \alpha\gamma_{ac}$ (diseño)

donde "i" es cualquiera de los a-1 niveles de "A" representados con las variables indicadoras.

De la misma forma β_k, coeficiente asociado a la variable indicadora de la categoría "k", se interpreta directamente como

(regresión) $\beta_k = \mu_{ak} - \mu_{ac} = \gamma_k - \gamma_c + \alpha\gamma_{ak} - \alpha\gamma_{ac}$ (diseño)

donde "k" es cualquiera de los c-1 niveles de "C" representados con las variables indicadoras.

Por su parte β_l, coeficiente asociado a la variable indicadora de la categoría "i" de "A" y a la variable indicadora de la categoría "k" de "C", se interpreta directamente como

(regresión) $\beta_l = \mu_{ik} - \mu_{ac} = \alpha\gamma_{ik} + \alpha\gamma_{ac} - \alpha\gamma_{ic} - \alpha\gamma_{ak}$ (diseño)

Una vez más vale la pena mencionar que cuando se declare un modelo en la computadora para el análisis de los datos, es muy importante saber cuáles hipótesis nos interesa probar.

En los ejemplos anteriores fue de interés probar la existencia de efecto de "tratamientos" y modelar la tendencia en el cambio de la respuesta al cambiar las variables explicativas a través de curvas, en el caso de variables explicativas cuantitativas, o a través de diferencias, en el caso de variables explicativas cualitativas.

De mayor interés aún fue ver si las tendencias o diferencias en la respuesta debidas a una variable explicativa, variaban con el cambio en una segunda variable explicativa; es decir, detectar la presencia de la interacción de las variables explicativas en la respuesta.

También pudo apreciarse que la interacción que se modelaba de la misma forma, para los diferentes ejemplos, en los modelos de diseño de experimentos, se modelaba de diferentes maneras en los modelos de regresión.

Mientras que en los modelos de diseño de experimentos la interacción se modela como la desviación de la respuesta por tratamiento de los valores predichos por los efectos principales, los modelos de regresión describen el patrón de variación de la respuesta con el cambio en una variable, y si este patrón de variación es igual o diferente en los niveles de otra variable.

Además, tratándose de uno o varios factores (variables) cuantitativos, los modelos de regresión nos permiten modelar la interacción en varios componentes, lo que es muy útil cuando el interés del usuario va más allá de sólo detectar la interacción, cuando es de interés describirla, caracterizarla.

Por otro lado, las variables indicadoras ("dummy") nos permiten involucrar variables (factores) categóricas a los modelos de regresión y con ello permiten al usuario de la estadística modelar de manera diferente lo que ya se modelaba a través de los modelos de diseño de experimentos, ampliando las posibilidades de análisis, descripción e interpretación de los fenómenos de interés al usuario.

Por ello debe quedar claro que si no declaramos convenientemente los modelos, sean éstos de diseño de experimentos o de regresión, omitiendo términos que impliquen comportamientos de interés, no será factible describir, modelar y entender tal comportamiento de interés del fenómeno bajo estudio, obteniéndose conclusiones parciales o erróneas, con el mismo costo de recursos y trabajo que implicaría la obtención de mejores descripciones y conclusiones a través de modelos más apegados a lo que se desea describir.

EXTENSIONES

Son varios los aspectos en los que se pueden extender las ideas aquí vertidas sobre variables indicadoras, considerándose como inmediatas las que se mencionan a continuación.

- *Comparación de más de dos modelos de regresión simple.*

- *Comparación de modelos de regresión de grado dos en más de dos grupos.*

- *Interacción entre más de dos factores.*

Algunos aspectos más que se pueden tomar como extensiones, considerando o no a las variables indicadoras, y que de alguna manera se dejaron etrever en lo ya descrito son:

- *Selección de modelos.*

- *Superficie de respuesta.*

Un tema más que seguiría a lo ya mencionado, aunque pareciera más alejado que los anteriores, sería:

- *Otras reparametrizaciones.*

- No se puede concluir sobre términos no declarados en un modelo o sobre hipótesis no formuladas o a prueba (en contraste).

"Sánto que no es visto no es adorado"

- No se puede interpretar la significancia de términos en los modelos o de hipótesis estadísticas en el contexto del problema o fenómeno bajo estudio, si no se conoce o entiende la contraparte o el significado de tales términos o hipótesis en dicho estudio.

"Pa' los toros del jaral,
los caballos de allá 'mesmo'"

- La significancia estadística de los coeficientes (β's) asociados a cada variable, es decir, la capacidad de explicación de una variable en un modelo se debe ver e interpretar en presencia (o ausencia, según sea el caso) de las demás variables involucradas en el modelo.

"...todo es según el color
del cristal con que se mira"

IV

- Mientras que los Modelos de Diseño de experimentos modelan la interacción a través del alejamiento que hay entre la respuesta media y la respuesta predicha por los efectos principales, los Modelos de Regresión modelan directamente el comportamiento diferencial de los patrones de cambio en la respuesta.

"No es lo mismo 'las calles de General Prim'
que 'las primas del General Calles'"

GENERALES

- *Las recomendaciones del procedimiento le permiten al usuario de la estadística claridad en las inferencias realizadas.*

- *Las diferencias en los conceptos de interacción en los modelos de regresión y de diseño de experimentos permiten desterrar errores fatales.*