UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

DIVISIÓN DE ESTUDIOS DE POSGRADO

PRONÓSTICOS MEDIANTE ANÁLISIS DE REGRESIÓN

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN INGENIERÍA

(INVESTIGACIÓN DE OPERACIONES)

P R E S E N T A:

A C T. M A R Í A I S A B E L
E S C A L A N T E M E M B R I L L O

DIRECTORA DE TESIS: M. en I. ISABEL PATRICIA AGUILAR JUÁREZ

MÉXICO, D. F. 2005

ÍNDICE

INTRODUCCIÓN

En la antigüedad, conocer el futuro siempre fue una de las inquietudes del hombre, y trataron de adivinarlo y anticiparse a él. Todo tipo de personajes intentaron predecir los hechos venideros, desde el destino de una persona, la ocurrencia de un desastre natural, hasta el resultado de un conflicto armado.

Conforme pasaron los años, los siglos, y el mundo fue evolucionado, esta preocupación del hombre por saber el futuro se fue extendiendo, y aún cuando se han podido pronosticar algunos acontecimientos naturales, sociales, económicos, etc, todavía no se dispone de herramientas que proporcionen el 100% de certeza en las predicciones.

La importancia de la elaboración de un pronóstico es grande, por ejemplo, para la economía de una nación es primordial conocer entre otros aspectos, el comportamiento de la inflación, el crecimiento en el producto interno bruto, la tasa de desempleo, etc., puesto que conociéndolos de antemano, el gobierno podría realizar cambios oportunos para optimizar las metas a alcanzar o para mejorar la situación de los ciudadanos.

Asimismo, siempre ha sido una parte integral de casi todos los tipos de toma de decisiones dentro de las organizaciones tanto productivas como sin fines de lucro. Algunas de las áreas en donde se utilizan pronósticos en la industria son la planeación y control de inventarios, producción, finanzas, ventas, comercialización, entre muchas otras.

Una de las técnicas de pronóstico cuantitativo, de mayor aplicación es la basada en el análisis de regresión. Ello se debe a una serie de factores tales como su facilidad de aplicación, su respaldo estadístico y la confianza que genera el método, ya que los principales resultados son familiares para cualquier persona que cuente con conocimientos básicos de inferencia estadística.

El objetivo principal del presente trabajo es presentar de manera sencilla el papel de las técnicas de regresión lineal clásica (simple y múltiple) en la elaboración de un pronóstico.

Para lograr lo anterior, y con la idea de introducir los conceptos básicos es necesario saber qué es un pronóstico, cuáles son los diversos métodos que se pueden utilizar para desarrollar un pronóstico, así como su clasificación; en qué consisten estas técnicas, cuándo se recomiendan utilizarlas, etc. Todos estos temas son abordados en el capítulo uno.

En el capítulo dos se explica qué es una ecuación de regresión lineal simple, los supuestos básicos que se consideran en su elaboración, las pruebas de hipótesis y estadísticos que ayudan a determinar si el modelo ajustado es el adecuado o no, la estimación de intervalos de confianza y de predicción.

De igual manera, en el capítulo tres, se desarrolla el modelo de regresión múltiple, los supuestos de qué parte, la estimación por intervalos de confianza, la predicción de nuevas observaciones, pruebas de hipótesis, gráficas de residuales, se abordan los problemas de multicolinealidad y de autocorrelación, entre otros.

El capítulo cuatro, está destinado a la parte aplicativa del trabajo, que es la utilización de los pronósticos y la interpretación de resultados, en dos problemas económicos.

El anexo A indica la estimación de los coeficientes de la regresión simple como de la múltiple, a partir del método de mínimos cuadrados ordinarios, así como de las propiedades que poseen éstos, y el anexo B muestra las tablas de resultados de los casos prácticos, emitidos por un programa estadístico.

CAPÍTULO 1

MARCO TEÓRICO DEL PRONÓSTICO

1.1. LA INVESTIGACIÓN DE OPERACIONES

Después de la Revolución Industrial, el mundo fue testigo de un crecimiento en el tamaño y la complejidad de las organizaciones. Los pequeños talleres artesanales se convirtieron en las corporaciones actuales. Una parte integral de este cambio revolucionario fue el gran aumento en la división del trabajo y en la separación de las responsabilidades administrativas en estas organizaciones. Junto con los beneficios, el aumento en el grado de especialización creó nuevos problemas que ocurren hasta la fecha en muchas empresas. Uno de estos problemas es que conforme la complejidad y la especialización crecen, se vuelve más difícil asignar los recursos disponibles a las diferentes actividades de la manera más eficaz para la organización como un todo. Este tipo de problemas, y la necesidad de encontrar la mejor forma de resolverlos proporcionaron el ambiente adecuado para el surgimiento de la **investigación de operaciones.**

Sus orígenes se remontan a muchas décadas, cuando se hicieron los primeros intentos para emplear el enfoque científico en la administración de una empresa. No obstante, su inicio casi siempre se atribuye a los servicios militares prestados a principios de la Segunda Guerra Mundial. Debido a los esfuerzos bélicos, existía una necesidad urgente de asignar recursos escasos a las distintas operaciones militares y a las actividades dentro de cada operación, en la forma más efectiva.

Por todo esto, las administraciones militares americana e inglesa hicieron un llamado a un gran número de científicos para que aplicaran el enfoque científico a éste y a otros problemas de estrategia y táctica, es decir, se les pidió que hicieran investigación sobre operaciones militares, convirtiéndose así en los primeros equipos de investigación de operaciones.

Estimulados por el evidente éxito en lo militar, los industriales comenzaron a interesarse en este nuevo campo. Como la explosión industrial seguía su curso al terminar la guerra, los problemas causados por el aumento de la complejidad y especialización dentro de las organizaciones pasaron a primer plano. Comenzó a ser evidente para un gran número de personas, incluyendo a los consultores industriales que habían trabajado con o para los equipos científicos durante la guerra, que estos problemas eran básicamente los mismos que los enfrentados por la milicia, pero en un contexto diferente. De esta forma, comenzó a introducirse en la industria, los negocios y el gobierno.

Dos factores que jugaron un papel importante durante este periodo, en el desarrollo de la investigación de operaciones fueron:

- El gran progreso que ya se había hecho en el mejoramiento de las técnicas disponibles en esta área. Posteriormente a la guerra, muchos científicos que habían participado en los primeros equipos o que tenían información sobre este trabajo, se encontraban motivados a buscar resultados sustanciales en este campo; resultando de esto avances importantes. Un ejemplo destacado es el *método simplex* para resolver problemas de programación lineal, desarrollado en 1947 por George Dantzig. Muchas de las herramientas propias, hoy en día de la investigación de operaciones, como programación lineal, programación dinámica, líneas de espera y teoría de inventarios, entre otras, tuvieron su principal impulso antes del término de la década de 1950.

- Un segundo factor fue la revolución de las computadoras. El manejo efectivo de los complejos problemas, casi siempre requiere un gran número de cálculos, efectuarlos a mano puede resultar casi imposible. Por lo tanto, el desarrollo de la computadora, con su capacidad para hacer cálculos aritméticos, miles o quizá millones de veces más rápido que los seres humanos, fue una gran ayuda. Un avance más tuvo lugar en la década de 1980 con el desarrollo de computadoras personales cada vez más rápidas, acompañado de buenos paquetes de software para resolver problemas de esta disciplina.

La investigación de operaciones puede describirse como un enfoque científico de la toma de decisiones que requiere la acción de sistemas organizacionales. Significa "*hacer investigación sobre las operaciones*". Esto dice algo tanto del enfoque como del área de aplicación; así, se emplea a problemas que se refieren a la conducción y coordinación de actividades dentro de una organización. La naturaleza de esta última es esencialmente inmaterial y, de hecho, se ha aplicado en los negocios, la industria, la milicia, el gobierno, los hospitales, la industria manufacturera, las telecomunicaciones, el transporte, la construcción, la planeación financiera, el cuidado de la salud, etc. Así, la gama de aplicaciones es extraordinariamente amplia.

El enfoque de la investigación de operaciones es el mismo del método científico. En particular, el proceso comienza por la observación cuidadosa y la formulación de la situación, le sigue la construcción de un modelo que intenta abstraer la esencia del problema. En este punto se propone la hipótesis de que se trata de una representación lo suficientemente precisa de las características esenciales de la situación, el modelo se compara contra la realidad para cerciorarse de que los elementos del mismo realmente están representando características o variables propias del problema o del entorno que se está contemplando.

Una vez que el paso anterior se ha superado, se resuelve el modelo logrado. Para esto se considera que el planteamiento realizado encaja o se ubica perfectamente en una de las múltiples técnicas ya desarrolladas.

Si no fuera el caso, se deberá revisar el modelo logrado para ver que cosas nuevas, matemáticamente hablando, se tendrán que desarrollar.

En cualquiera de las dos posibilidades, habrá que encontrar la solución que satisface el objetivo y una vez logrado lo anterior se verifica la viabilidad y similitud con la realidad.

Si lo anterior no plantea circunstancias que lleven a concebir ideas de que lo realizado hasta este punto no está bien, se debe de proceder a implantar la solución obtenida del modelo en la realidad.

En resumen, la investigación de operaciones se ocupa de la toma de decisiones óptima y del modelado de sistemas, ya sea que su perspectiva sea **determinista** o **probabilista** que se originan en la vida real.

Los problemas con perspectiva *determinista* son aquellos en los que cada alternativa de solución (hay más de dos) tiene sólo un resultado final. Como son varias alternativas, hay varias soluciones, cada una con diferente eficiencia y/o efectividad asociada al objetivo del problema. Por lo tanto, no se sabe qué decisión adoptar y se enfrenta una incertidumbre.

Los modelos determinístas se ubican en técnicas tales como la programación lineal, la teoría de redes, la programación de metas, la programación entera, programación dinámica, teoría de inventarios, entre otras.

Por otra parte, los modelos probabilístas se encuentran en técnicas como el análisis de decisiones, la teoría de juegos, los modelos de pronósticos, la programación dinámica probabilísta, la teoría de inventarios probabilísta, la teoría de colas y la simulación, entre otros.

Los problemas de este tipo se caracterizan porque cada alternativa de solución desencadena diferentes resultados y no se sabe cual será la más eficiente, la incertidumbre se da por no saber que alternativa de solución escoger y el riesgo existe porque dada una selección no se está cierto hasta que el resultado deseado se obtenga.

1.2. ANTECEDENTES DEL PRONÓSTICO

La idea de mirar hacia el futuro no es nueva; desde la antigüedad se realizaban profecías y vaticinios de carácter esotérico y religioso, para averiguar acerca de lo que debía esperarse en el futuro; la peculiaridad era que el destino no podía modificarse y sólo había que esperar su llegada o, prepararse lo mejor posible para lo que necesariamente habría de venir.

Esta concepción fatalista permaneció durante muchos siglos como algo asociado a aspectos religiosos y a procesos de adivinos; sin embargo, con el desarrollo de la sociedad y la ciencia en su conjunto se ha buscado y logrado (en cierta medida) superar esta idea.

Desgraciadamente tal panorama es cierto para las instituciones de gran tamaño y que tienen la posibilidad de aplicar alguna parte de sus recursos financieros al esfuerzo que representan tales posibilidades.

Sin embargo, aún queda mucho por hacer puesto que existen muchas entidades económicas y el hombre mismo que están sujetos todavía al "sentimiento" , "esperanza" de "alguien" que les predice o augura tal o cual situación con muy pocos elementos de información, sin consistencia y que son la causa de la desaparición de unos y otros.

La preocupación por el futuro ha adquirido una importancia de primer orden, lo que, por una parte, es resultado de que se observa una realidad que a pocos tiene satisfechos, cuya perspectiva es que las cosas empeoran cada vez más; y, por la otra, a que se considera que el futuro es transformable, elegible, moldeable hasta determinados puntos, razones que han servido de base y estímulo para el desarrollo y popularización de la planeación y, con ello, del pronóstico.

En general, se puede decir que existe una necesidad de generar acciones para diseñar el futuro, lo que parece ser la raíz del pronóstico.

El tema del pronóstico cada vez adquiere una mayor importancia, no obstante que en su desarrollo se ha dado un énfasis a la parte de técnicas y modelos, eludiendo lo relativo a la metodología y bases de tal ejercicio y considerando que resulta insuficiente el conocer una serie de técnicas para el pronóstico, si no se tiene una conciencia clara acerca de cuándo son aplicables unas u otras, si no es posible definir con precisión qué debe ser pronosticado y para qué, etcétera.

Para comprender la amplitud de los enfoques de predicción actualmente disponibles, es útil presentar una breve sinopsis histórica.

Antes del decenio de 1950 había pocos o ningunos esfuerzos sistemáticos de pronósticos en las empresas. Aunque se disponía de un puñado de metodologías, como la regresión lineal y la descomposición de series de tiempo, sus aplicaciones estaban limitadas a departamentos de economía de avanzada en las universidades y dependencias gubernamentales. La aplicación difundida y sistemática de tales técnicas se hallaba severamente obstaculizada por la carencia de datos apropiados y lo tedioso de los cálculos requeridos.

A mediados de la década de 1950 dos hechos importantes cambiaron el campo de los pronósticos. El primero fue la introducción de una amplia gama de técnicas de atenuación exponencial; las cuales, en un principio fueron empleadas, sólo por militares, pero poco a poco se extendieron hacia las organizaciones comerciales. Las principales ventajas de estos métodos, fueron su simplicidad conceptual y su facilidad de cálculo. Aun cuando estas metodologías eran muy aceptadas por los profesionistas en el campo, la mayor parte de los académicos y pronosticadores profesionales consideraron que tales métodos tan simples no podían ser lo suficientemente exactos como para merecer una seria atención.

A pesar de que se requieren únicamente unas cuantas ecuaciones y relativamente pocos cálculos aritméticos, en la década de 1950, la aplicación de los métodos de atenuación exponencial todavía se dificultaba. Si se necesitaban los pronósticos para varios miles de artículos, se requería una enorme cantidad de trabajo para mantener los archivos de datos, hacer los cálculos pertinentes y simplemente transcribir los resultados. Sin embargo, un

segundo gran acontecimiento en esta época resolvió muchos de estos problemas: la introducción de la *computadora*. Ésta no sólo permitió la atenuación exponencial sino también el uso continuo de un sinnúmero de otros métodos de pronóstico. Asimismo, ha permitido un amplio uso de éstos, tanto en las empresas como en el gobierno.

Se han desarrollado diversas variaciones y extensiones del estudio de los métodos de atenuación, las más notables de éstas son las de Brown, Holt y Winters. Más tarde se desarrollaron las técnicas de atenuación exponencial de parámetros adaptativos, los cuales permiten el uso mecánico y automatizado de los métodos de atenuación. En estas técnicas más recientes, el usuario no necesita especificar los valores de los parámetros para el modelo de atenuación exponencial; por el contrario, se les puede calcular y actualizar automáticamente. No mucho después de que los métodos de atenuación empezaran a despertar interés, los métodos de descomposición comenzaron a llamar la atención. En este grupo destacó el método II del censo, desarrollado por Julius Shiskin de la Oficina de Censos del gobierno de los Estados Unidos. Si bien estos métodos de descomposición tuvieron poco apoyo estadístico, sí interesaron a los profesionales en el área.

Conforme se abarató el costo de las computadoras y al aumento de su disponibilidad en la década de 1960, los métodos estadísticos de pronóstico más complejos aparecieron. Técnicas tales como los métodos econométricos se volvieron prácticos y se les utilizó para cuantificar y probar la teoría econométrica con los datos empíricos.

Durante las décadas de 1950 y 1960, los académicos todavía se encontraban buscando una teoría unificadora del pronóstico, un método que incorporara muchos de los elementos. Finalmente, con el trabajo de los profesores George Box y Gwilym Jenkins se hizo realidad. La metodología Box–Jenkins (como se le dio a conocer), proporcionó un procedimiento sistemático para el análisis y pronóstico de series de tiempo que fue lo bastante general como para manejar prácticamente todos los patrones de datos acerca de las series de tiempo observados en forma empírica. Se dio un importante impulso al método cuando varios estudios comparativos de pronósticos demostraron que el método Box-Jenkins era al menos tan exacto como los métodos econométricos.

A mediados de la década de 1970 surgió una variante del procedimiento del promedio móvil autorregresivo (ARIMA) desarrollado por Box y Jenkins. Estos fueron los métodos ARIMA de parámetro adaptativo. Una de las dificultades asociadas con éstos era su complejidad estadística, la cual dificultaba su comprensión a los profesionistas no especializados. A finales de la década de 1970 se desarrollaron técnicas más eficientes para el modelado de los procesos ARIMA por gente como Parzen, logrando así su aceptación general.

En el aspecto cualitativo, los métodos tecnológicos de pronóstico tuvieron mucha aceptación durante las décadas de 1960 y 1970. A principios de la década de 1980, en varias organizaciones se usaron procedimientos tales como el método Delphi y el de las matrices de impacto en el costo. Estas técnicas tecnológicas o cualitativas para la elaboración de pronósticos intentaron manejar las tendencias a largo plazo de las variables cuando no se disponía de los datos y patrones históricos necesarios para aplicar los métodos estadísticos de pronóstico o cuando esos datos y patrones no eran aplicables. Simultáneamente, se llevaron a cabo amplias investigaciones en el campo de la mercadotecnia relacionado con los nuevos productos y el pronóstico de nuevos mercados, los cuales compartían esta carencia de datos históricos.

Uno de los adelantos más interesantes en el campo de los pronósticos se llevó a cabo a finales de la década de 1970, fue la demostración de que los pronósticos solos son inútiles, a menos que se les aplique a la planeación y la toma de decisiones. Diversos estudios señalaron que los problemas de organización frecuentemente dificultarían el uso de ellos, aun cuando pudieron demostrar tener un desempeño bastante exacto en el tiempo. Durante este mismo período varios estudios identificaron las características individuales de comportamiento que podrían dificultar el uso de métodos "comprobados" de pronóstico. Estas investigaciones, indicaron que a menudo, las revisiones de los predicciones por parte de la gerencia tuvieron como punto de partida sueños, ilusiones equivocadas, y la influencia política, en vez de la realidad objetiva.

A finales de la década de 1970, esta disciplina se transformó en un campo con derecho propio tanto para los profesionales en ejercicio como para los académicos, a medida que se reconoció su importancia para todas las formas de planeación y toma de decisiones, en áreas tan diversas como los negocios, el gobierno, las instituciones no lucrativas y las organizaciones militares.

Actualmente los pronósticos son una de las herramientas fundamentales para la toma de decisiones dentro de las organizaciones tanto productivas como sin fines de lucro. Algunas de las áreas en donde se utilizan pronósticos en la industria son la planeación y control de inventarios, producción, finanzas, ventas, comercialización, entre muchas otras.

1.3. CONCEPTOS GENERALES DE LOS PRONÓSTICOS

1.3.1. ¿QUÉ ES UN PRONÓSTICO?

Cuando cualquier empresa o individuo, hace una afirmación acerca de

- la ocurrencia o no de un evento en el futuro
- la fecha en que va a suceder algo
- la intensidad de un evento futuro

está realizando un pronóstico.

La palabra **pronóstico** se deriva del griego *prognôstikon* (latín *pronosticum*), que significa conjetura acerca de lo que puede suceder.

Los pronósticos son predicciones que pueden acontecer o esperarse, son premisas o suposiciones básicas en que se basan la planeación y la toma de decisiones. Su objetivo es reducir la incertidumbre acerca de lo que ocurrirá en el futuro proporcionando información cercana a la realidad que permita elegir decisiones sobre los cursos de acción a tomar, tanto en el presente como en el futuro.

1.3.2. LOS PRONÓSTICOS Y SU RELACIÓN CON OTRAS DISCIPLINAS

1. Con la *toma de decisiones*. El uso principal de los pronósticos está relacionado con esta ciencia. Algunas son "pequeñas" decisiones, por ejemplo las vinculadas al mantenimiento de un adecuado nivel de inventario. Otras son "grandes", como las asociadas con inversiones. En todo caso los pronósticos, para ser útiles, deben estar enlazados con la toma de decisiones y se deben presentar con oportunidad.

2. Con la *planeación*. Al desempeño de una empresa lo afectan eventos externos e internos. Los externos son los que afectan a la empresa pero sobre los cuales tiene poco o nulo control; éstos son incontrolables y, por lo general son aquellos que se tratan con las técnicas de pronósticos. Por otra parte los eventos internos son los relacionados a decisiones que la empresa toma; sobre éstos la empresa tiene completo control. La planeación relaciona tanto los internos como los externos.

1.3.3. CLASIFICACIÓN DE LOS PRONÓSTICOS

1. *Corto, mediano y largo plazos*. Un criterio de división de los pronósticos atiende al tiempo para el cual se prepara.

   - Para programación de producción, transporte, efectivo, personal, etc., generalmente se hacen pronósticos de corto plazo. Un ejemplo es la programación del efectivo, una cuenta bancaria debe tener dinero suficiente para afrontar los compromisos previstos. Si tiene de más, se pierde la oportunidad de invertirlo, si tiene menos no habrá la liquidez suficiente para enfrentarlos y puede haber pérdida de oportunidades.
   - Para adquisiciones de materia prima, equipo, etc., así como para la contratación de personal, se hacen pronósticos de mediano plazo. La obtención de algunos bienes no son inmediatas y necesitan un cierto tiempo para realizarse (un tipo de equipo especializado no se adquiere rápidamente, en algunos casos), es preciso planear adecuadamente y para eso es útil tener un pronóstico de necesidades.

- Para algunos aspectos de presupuestación, planeación de inversiones, planeación estratégica, etc., se requieren pronósticos de largo plazo. Las inversiones de capital siempre requieren de predicciones, tanto de la necesidad de las mismas como de plazos en los que se espera recuperar la inversión, estos últimos requieren de estimaciones de otras cosas como ventas, tasas de interés y otros factores importantes.

2. *Complejidad de la técnica*. Hay pronósticos muy simples en cuanto a la técnica para elaborarlos. El más simple es "lo mismo que hoy". En el otro extremo se encuentran aquéllos que hacen uso de computadoras y algoritmos sofisticados, o los que reúnen la opinión de varios expertos y que pueden llevar meses para su conclusión.

3. *Integración de los pronósticos*. Dado que se realizan en diferentes departamentos de las empresas como producción, ventas, desarrollo del producto, presupuestos, tesorería y dirección general, entre otros, e incluso pueden ser contradictorios, es necesario sistematizar la forma de obtenerlos y planear cómo se van a integrar para darles coherencia.

1.3.4. CARACTERÍSTICAS DE LOS PRONÓSTICOS

1. Todas las situaciones en que se requiere un pronóstico, tratan con el *futuro* y el **tiempo** está directamente involucrado. Así, debe pronosticarse para un punto específico, tomando en cuenta que un cambio de éste generalmente altera la predicción.

2. Otro elemento siempre presente es la *incertidumbre*. Si el pronosticador tuviera certeza sobre las circunstancias que existirán en un tiempo dado, la preparación de un pronóstico seria trivial.

3. El tercer elemento, es la *confianza* de la persona que hace el pronóstico sobre la información contenida en *datos históricos*.

1.4. MÉTODOS PARA HACER PRONÓSTICOS

La *técnica* o *método de pronóstico* es el procedimiento por medio del cual se lleva a cabo una predicción. Disminuye la incertidumbre sobre el futuro, permitiendo estructurar planes y acciones congruentes con los objetivos de la organización, además de tomar acciones correctivas apropiadas y a tiempo cuando ocurren situaciones fuera de lo pronosticado.

Para la selección del método de pronóstico, se deben considerar los siguientes factores:

- El contexto del pronóstico.
- La relevancia y disponibilidad de datos históricos.
- El grado de exactitud deseado.
- El periodo de tiempo que se va a pronosticar.
- El análisis de costo – beneficio del pronóstico.
- El punto del ciclo de vida en que se encuentra el producto.

La elección del método correcto dependerá principalmente de la cantidad y calidad de los antecedentes disponibles, así como de los resultados esperados. Su efectividad se evaluará en función de su precisión, sensibilidad y objetividad.

Las conclusiones que se obtienen de estas técnicas son sólo indicadores de referencia para una predicción definitiva, la cual, deberá completarse con el juicio y las apreciaciones cualitativas del analista, quien quizás utilizará más de un método en la búsqueda del pronóstico más certero.

Se distribuyen principalmente en dos tipos:

* Cuantitativos.
* Cualitativos.

A su vez, cada uno se subdivide en varios más, quedando su clasificación como se ilustra en la figura 1.1. A continuación, se describirán las características principales de ellos.

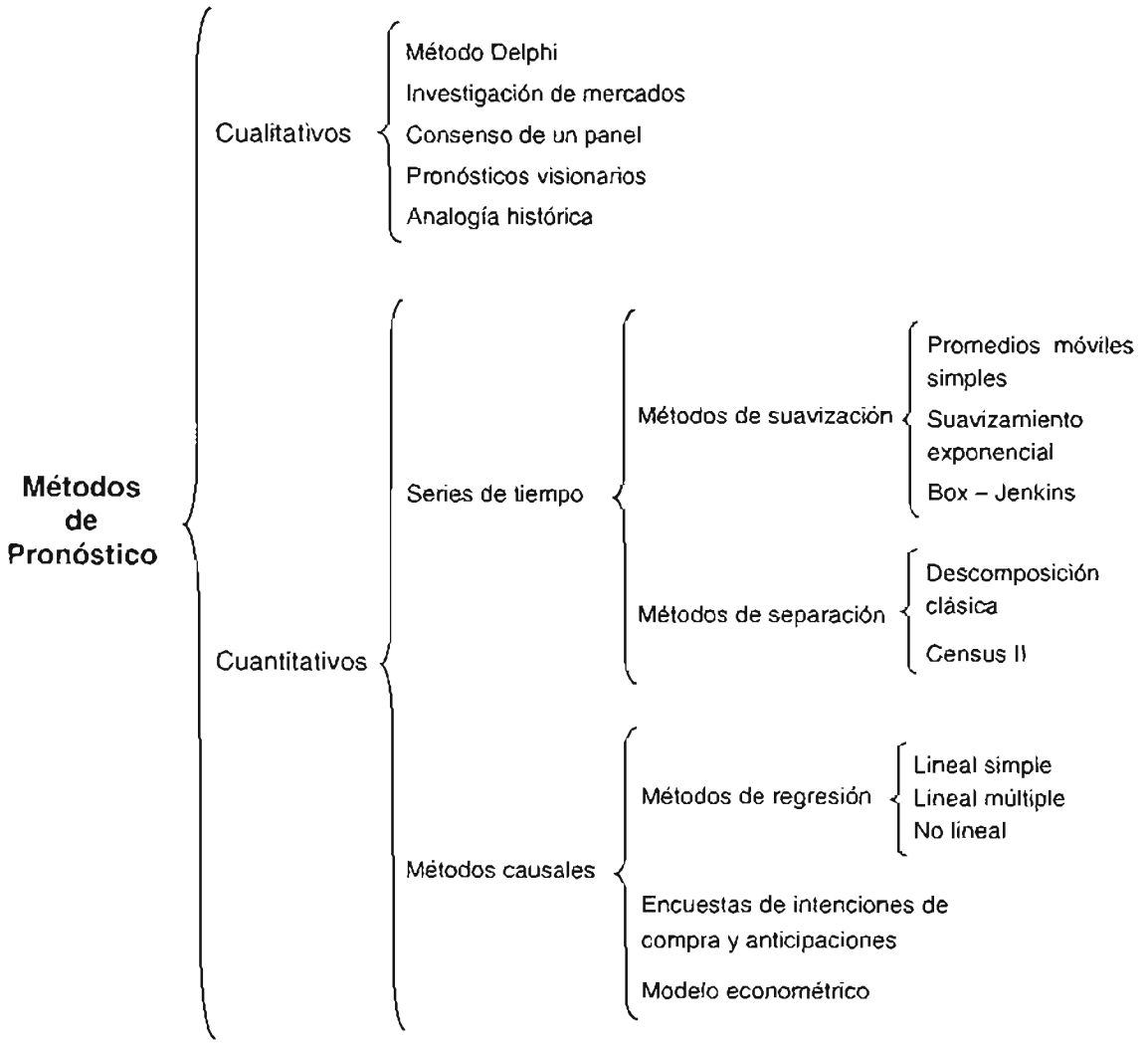


Figura 1.1

1.4.1. MÉTODOS CUALITATIVOS

Las técnicas cualitativas o subjetivas (como también se les conoce) se usan cuando los datos son escasos o bien la información disponible no es confiable para predecir algún comportamiento futuro, así mismo, si el tiempo para elaborar el pronóstico es insuficiente; por ejemplo, cuando se introduce un producto nuevo al mercado. Sin embargo, aun cuando se tengan buenos datos, algunos tomadores de decisiones prefieren alguno de estos procedimientos y no uno formal, y en otros casos, se usa una combinación de ellos.

Estos métodos dependen, por naturaleza, del juicio personal y pueden incluir cualidades como intuición, opinión de expertos y experiencia. En general, usan el criterio de la persona y ciertas relaciones para transformar información cualitativa en estimados cuantitativos. Aunque la gama de estos métodos es bastante amplia, es prácticamente imposible emitir algún juicio sobre la eficacia de sus estimaciones finales.

En seguida, se presenta una idea general breve de los principales métodos subjetivos.

Método Delphi.

Consiste en reunir a un grupo de expertos en calidad de panel, a quienes se les somete a una serie de cuestionarios de forma independiente, con un proceso de retroalimentación controlada después de cada serie de respuestas, para tratar de obtener un consenso confiable entre diversos especialistas para usarlo como base para pronosticar. Como puede observarse, esta técnica funciona básicamente por rondas, y una vez que se ha logrado una dispersión relativamente pequeña en las conclusiones, los tomadores de decisiones evalúan los datos para desarrollar el pronóstico.

El reunir a varias personas especializadas en el tema de interés, para que emitan sus opiniones, conlleva a tener factores psicológicos que afectan al consenso. Algunos de ellos, pueden tener mayor claridad en sus exposiciones, ser más persuasivos, o mejores polemistas que otros, sin que por ello tengan mayor razón. Por esto la técnica Delphi

funciona evitando que los expertos se reúnan, toda la comunicación se hace mediante un coordinador.

Este proceso complejo se usa sólo en niveles altos de la corporación o el gobierno para desarrollar pronósticos a largo plazo de tendencias generales, como ventas de productos nuevos y tecnológicos.

Investigación de Mercados.

Es más sistemático y objetivo, que se vale del método científico. Se usa principalmente para evaluar y probar hipótesis acerca de mercados reales, a través de encuestas a clientes, experimentos, mercados prueba u otra forma.

Esta técnica podría decirse que constituye un paso necesario para la aplicación y uso de cualquiera de los restantes métodos, dado que suministra información sistematizada y objetiva.

La flexibilidad para seleccionar e incluso diseñar la metodología que más se adecue al problema en estudio, requiriendo una investigación ya sea exploratoria, descriptiva o explicativa, es quizás su principal característica.

Consenso de un Panel.

Básicamente tiene los mismos usos que el método Delphi, dado que es muy similar a éste, su diferencia consiste en que no existen secretos sobre la identidad del emisor de las opiniones, y en que no hay retroalimentación dirigida desde el exterior.

Pronósticos Visionarios.

Se usa para hacer una profecía del futuro usando la intuición personal. Es obvio que este método presenta ventajas en cuando a costo y rapidez, dado que no necesita de destrezas especiales, pero presenta algunas insuficiencias derivadas de la influencia dominante de las experiencias más recientes y de la falta de unidades de medida que den exactitud a la estimación.

Tabla 1.1

Método	Tiempo estimado	Exactitud
Delphi	Más de dos meses	De regular a muy buena
Investigación de Mercados	Más de tres meses	Excelente, dependiendo del cuidado que se haya puesto en el trabajo
Consenso de un Panel	Más de dos semanas	De baja a regular
Pronósticos Visionarios	Una semana	Mala
Analogía Histórica	Más de un mes	De regular a buena

Analogía Histórica.

Se usa para productos nuevos, basándose en el análisis comparativo de la introducción y crecimiento de productos similares. La desventaja que presenta, es la de suponer que las variables .determinantes en el comportamiento pasado tomadas como referencia se mantendrán en el futuro y, tendrán el mismo efecto sobre el producto en estudio.

En la tabla 1.1, se especifican el tiempo estimado, así como, de la exactitud de cada uno de los métodos antes descritos.

1.4.2. MÉTODOS CUANTITATIVOS

Estas técnicas se basan principalmente en datos históricos. Esta información pasada se encuentra en forma numérica, donde las fuentes usuales son los registros de la propia empresa o información oficial de diverso origen: gobierno, asociaciones de empresarios o profesionistas, organismos internacionales, etcétera.

Se debe tener precaución, sobre todo cuando la información procede de la propia empresa (aunque en la proveniente de otras fuentes también hay que cuidarse), que haya sido

cuantificada de manera uniforme. Para información sobre costos, por ejemplo, hay que asegurarse de que éstos incluyan los mismos conceptos en todos los años que se va a utilizar; de no ser así es preciso tratar previamente los datos.

Para aplicar los métodos cuantitativos es preciso convencerse, razonablemente, de que se cumple la llamada *hipótesis de continuidad.* Este supuesto consiste en considerar que los factores externos en los que se dieron los datos históricos, no cambiarán en el futuro para él que se está pronosticando, los cuales son, en forma destacada:

- Economía en general.
- Competencia en el mercado (oferta).
- Estado del mercado (demanda).
- Estado tecnológico del producto ("ciclo de vida del producto").

Esta continuidad del ambiente nunca se da en forma perfecta, sino que se produce de manera gradual. Se requiere buen juicio para suponer que las violaciones a la hipótesis antes señalada, no van a afectar a los resultados de la aplicación del método de pronóstico.

Las técnicas cuantitativas son de dos tipos, según la información en que se basan:

- *Métodos de series de tiempo.* Se usa información de la misma variable que se va a pronosticar.
- *Métodos causales.* Se utiliza información de la variable que se va a pronosticar y de otras variables que influyen o que están relacionadas con ella y cuyo pronóstico sea más simple.

1.4.2.1. Métodos de series de tiempo.

El análisis consiste en encontrar el patrón de comportamiento del pasado y proyectarlo al futuro de la variable deseada, a través de un modelo matemático que sea representativo de proceso, basándose en datos históricos de una *serie de tiempo*[1]. La forma de éste será de la siguiente manera

$$y_{t+h} = f\left(y_t, y_{t-1}, y_{t-2}, \ldots, y_{t-d}\right),$$

donde f( ) es una función que permite calcular el valor futuro de *y*, a partir de un conjunto de hechos pasados, tomando en cuenta factores de variación cíclica, tendencias, estacionalidad, autocorrelaciones, entre otros..

Existen cuatro elementos básicos encontrados frecuentemente en una serie de datos, que son:

- *Efecto tendencial.* Se refiere al crecimiento o disminución en el largo plazo del valor promedio de la variable analizada, por ejemplo la demanda. Su importancia se origina de considerar fluctuaciones en el nivel de la variable en el tiempo, con lo cual la investigación del nivel promedio de ésta a lo largo de todo el periodo, es mejor que su estudio en un momento específico de tiempo. Las ventas de muchas compañías, el producto interno bruto, los precios y otros indicadores económicos y empresariales siguen un patrón en sus movimientos como el que se observa en la figura 1.2

- *Efecto estacional.* Es el que exhibe fluctaciones que se repiten en forma periódica y que normalmente dependen de factores como el clima (ropa de verano, bebidas gaseosas) y la tradición (tarjetas de navidad), entre otros. La figura 1.3 presenta un patrón en el cual las estaciones corresponden a los cuatro trimestres del calendario para la primavera, el varano, el otoño y el invierno.

[1] Una serie de tiempo es un conjunto de observaciones en un periodo de tiempo de alguna cantidad de interés (variable aleatoria).

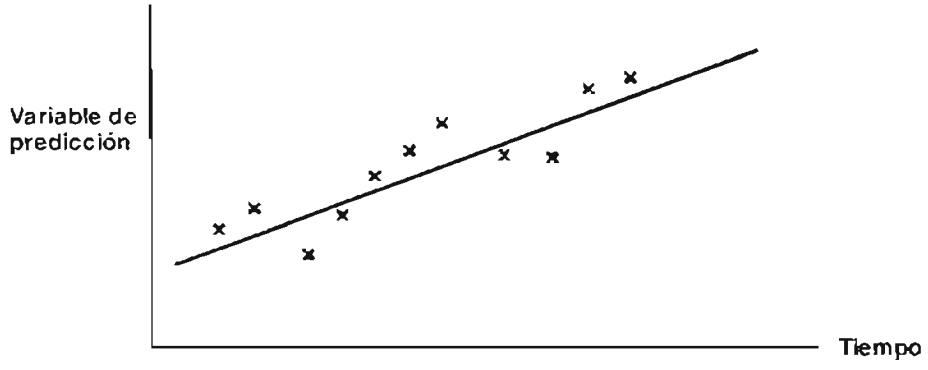


Figura 1.2 Patrón de tendencia

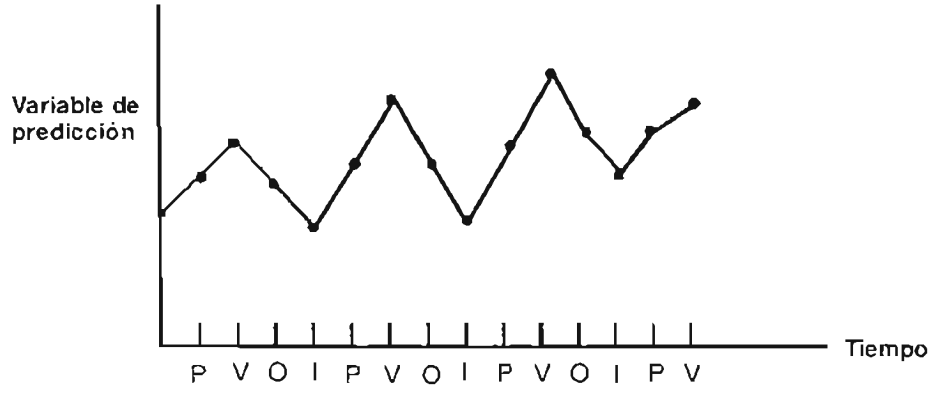


Figura 1.3 Patrón de estacionalidad

- *Efecto cíclico.* Es la divergencia significativa entre la línea de tendencia proyectada y el valor real que exhiba la variable, y se admite entre sus causas el comportamiento del efecto combinado de fuerzas económicas, sociales, políticas, tecnológicas, culturales y otras existentes en el mercado. Cabe señalar, que la mayoría de estos ciclos no tienen patrones constantes que permitan prever su ocurrencia, magnitud y duración, lo cual los hace difícil de pronosticar. La figura 1.4 indica un patrón cíclico.

- *Aleatoriedad.* Aunque se conozcan lo tres componentes anteriores, una variable puede tener todavía un componente real distinto del previsible por su línea de tendencia y por los factores cíclicos y estacionales. A esta desviación se le asigna el carácter de no sistemática y corresponde al llamado elemento aleatorio. En la figura 1.5 se muestran los cuatros integrantes de una serie cronológica.

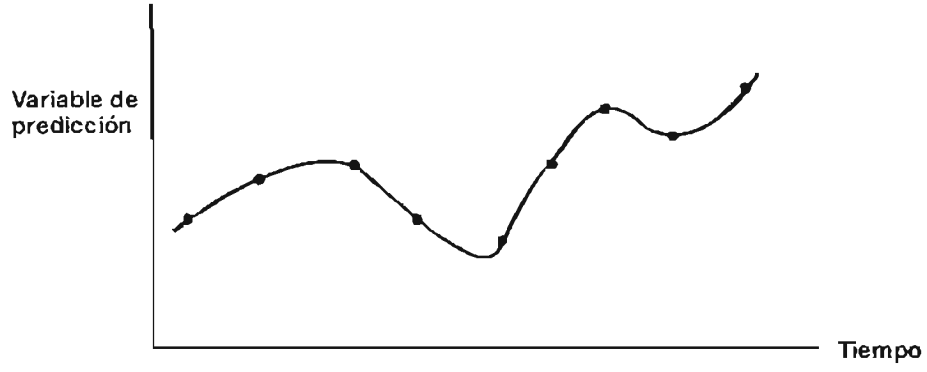


Figura 1.4 Patrón cíclico.

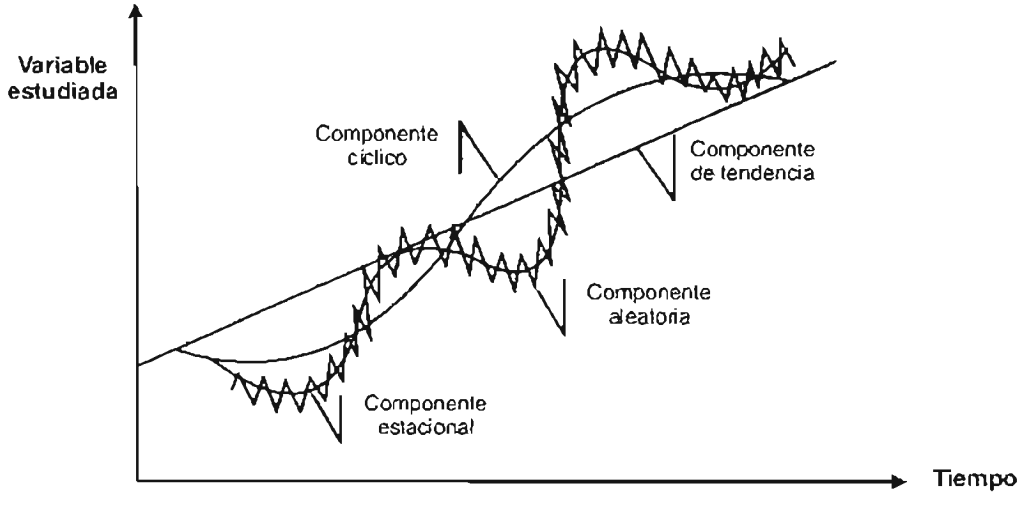


Figura 1.5

Estos métodos a su vez se dividen en:

*1. **Métodos de proyección.***

Son aquellos que tratan de encontrar el patrón total de los datos para proyectarlos al futuro, los principales son:

- *Promedios Móviles*
- *Suavización Exponencial*
- *Box-Jenkins*

2. Métodos de separación.

Se les conoce también como *técnicas de descomposición de series de tiempo.* Son aquellos que separan la serie en sus elementos, para identificar el patrón de cada componente, para posteriormente preparar un pronóstico para un cierto período, el cual simplemente es F = estacional × tendencial × cíclico. Siguiendo los pasos anteriores, se tendrá lo que algunos autores llaman la "**descomposición clásica**"; sin embargo, es posible aplicar variedades, como puede ser el **Census II**, que en esencia es semejante al primero, pero contiene refinamientos y elaboraciones que hacen a los resultados más apropiados para ciertos tipos de aplicaciones.

1.4.2.2. Métodos Causales.

Se basan en identificar y determinar cuales son las relaciones existentes entre la variable dependiente de interés a pronosticar y las independientes que la determinan al ejercer su influencia sobre ella. A continuación se dará una breve descripción de las principales técnicas de este tipo.

1. Métodos de Regresión.

Es una técnica estadística para investigar y modelar la relación funcional entre variables. La idea básica consiste en identificar cuál es la curva que mejor se ajusta a un conjunto de N datos, a partir de ciertas variables relacionadas x', y con ello establecer una ecuación que permita estimar el valor de la variable dependiente y. El modelo que expresa dicha interdependencia funcional es

$$y = f\left(x_1, x_2, \ldots\ldots, x_k\right)$$

donde:

y es la variable por estimar, conocida como variable dependiente.

x_i es la i-ésima variable relacionada, conocida como variable independiente.

$f(\ )$ es la función que indica la relación que guarda la y con las x's.

Básicamente este tipo de técnicas se dividen en:

- Regresión lineal simple
- Regresión lineal múltiple
- Regresión no lineal

2. Encuestas de intenciones de compra y anticipaciones.

Estas encuestas que se hacen al público, determinan:

a. Las intenciones de compra de ciertos productos.

b. Derivan un índice que mide el sentimiento general sobre el consumo presente y futuro y estiman como afectan estos sentimientos a los hábitos de consumo. Este enfoque para hacer pronósticos es más útil que otras técnicas para seguir el desarrollo de la demanda y para señalar puntos de peligro.

3. Modelo econométrico.

Algunos lo consideran un sistema de ecuaciones estadísticas que interrelacionan a las actividades de diferentes sectores de la economía y ayudan a evaluar la repercusión sobre la demanda de un producto o servicio. Por otra parte, también se puede definir como un modelo para estimar la demanda de un producto, que parte de la base de que el precio se determina por la interacción de la oferta y la demanda.

CAPÍTULO 2

ANÁLISIS DE REGRESIÓN LINEAL SIMPLE

En algunos casos, la variable que se va a pronosticar tiene una relación bastante directa con una o más variables cuyos valores se pueden conocer con anterioridad. Si es así, tendría sentido basar la predicción en esta relación. Este tipo de enfoque se llama *pronóstico causal.*

"*Un* ***pronóstico causal*** *obtiene un pronóstico de la cantidad de interés (la variable dependiente) relacionándola en forma directa con una o más cantidades (variables independientes) que impulsan a la cantidad de interés.*"[2]

En este capítulo se describirá un tipo de predicción causal donde se supone que la relación matemática entre las variables dependiente e independiente(s) es lineal (más alguna fluctuación aleatoria). El análisis en este caso se conoce como ***regresión lineal.***

2.1. LOS ORÍGENES DEL ANÁLISIS DE REGRESIÓN [3]

El análisis de regresión nace y se desarrolla en dos contextos culturales distintos: la francesa y la inglesa. En la primera vinculado a la astronomía y en la segunda a estudios eugenésicos.

El desarrollo del análisis de regresión, en Francia, estuvo asociado a tres problemas planteados en el siglo XIX:

---

[2] HILLER, Frederick S. y Gerard J. LIEBERMAN. Investigación de Operaciones. México, p. 1028.

[3] CORTÉS, Fernando. Regresión logística en la investigación social: potencialidades y limitaciones. http://www.rau.edu.uy/fcs/soc/revista_13/cortes13.htm

1. representar y determinar matemáticamente los movimientos de la luna,
2. dar cuenta de una desigualdad no periódica en los movimientos de los planetas Júpiter y Saturno, y
3. determinar la forma de la tierra.

Todos ellos envolvían, por una parte, observaciones astronómicas y por la otra, la teoría de la gravitación. Se trataba entonces de ajustar las ecuaciones derivadas de ésta a los resultados de las observaciones astronómicas. El ajuste implicaba más ecuaciones que incógnitas en la medida que se tenía un número apreciable de observaciones y las segundas se reducían a unos pocos parámetros.

En estos estudios no se dudaba acerca de la hipótesis, ésta era tomada por buena, las desviaciones entre los resultados que arrojaba el modelo y los datos observados, se suponía, tenían su origen en los errores de medición, los cuales, eran considerados aleatorios.

Considérese entonces que por alguna razón se postula que hay una relación lineal entre la variable explicativa x y la explicada y, de manera que si se tiene un conjunto de n pares ordenados (x, y) se genera un sistema de n ecuaciones con dos incógnitas: la ordenada al origen a y el coeficiente angular b. La solución matemática a este problema se debe a Adrien Legendre quien en el apéndice a su trabajo "*Nouvelles méthodes pour la determination des orbites des cométes*", publicado el 6 de marzo de 1805, se propone para resolver este problema la "técnica mínimo cuadrática".

Este tipo de estimación ordinaria, así como sus variantes, se han utilizado profusamente en el ajuste de modelos de regresión, ya sea lineales o susceptibles de ser linealizados.

El desarrollo del análisis de regresión a mediados del siglo XIX en Inglaterra, se dio en un ambiente cultural distinto al francés; y se enmarcó en el debate sobre las diferencias de clase, de raza y de inteligencia.

No por casualidad fue Francis Galton (1822 – 1921) quién acuñó el término *eugenesia.* Su trabajo en estadística, genética y psicología de las diferencias individuales estuvo marcado por su interés en mejorar la raza. La técnica de regresión le permitió predecir las características de los hijos a partir de los rasgos de los padres. Éste trabajo, lo continuaron Karl Pearson y R. A. Fisher.

El planteamiento clásico del modelo de regresión suponía que todas las variables pertenecieran a las escalas de intervalos. Sin embargo, esta limitación fue superada al introducir primero, *variables explicativas ficticias* (dicotómicas) que dieran cuenta de la presencia o ausencia de un evento particular (por ejemplo, el impacto sobre la función consumo de los años de paz o de guerra). Esta aproximación se generalizó a una o más variables pluricotómicas.

Fue así como la regresión ganó en ductilidad, haciéndose más atractiva a los ojos de los científicos. No obstante, el modelo, desarrollado hasta este punto, aún requería que la variable dependiente fuese continua o tuviera un conjunto grande de valores posibles, que pudiera asumir.

En efecto, el ajuste de un modelo de regresión lineal cuando la variable dependiente es dicotómica conlleva una serie de anomalías en el modelo de regresión estándar. Estos problemas se superaron aplicando la transformación *"logit"* a la variable dependiente (H. Theil; Aldrich; N. Forrest; A. Agreste; etc.).

A pesar de que la teoría estadística se ha generalizado, para el caso en que la variable dependiente adopta más de dos categorías, su aplicación se ve restringida por la escasa disponibilidad de paquetes de cómputo que incluyan las rutinas de cálculo que permitan su empleo.

Como puede apreciarse, sucesivos avances liberaron al modelo de regresión de lo que alguna vez se consideró su limitación más importante, para aplicarlo al análisis de problemas de tipo categórico.

2.2. NOCIONES DE REGRESIÓN LINEAL

El objetivo primordial del análisis de regresión es estimar el valor de una variable aleatoria (*variable dependiente* o *de respuesta*) dado que el valor de una o más variables asociadas (*variables independientes* o *de predicción*) es conocido.

La *ecuación de regresión* es la fórmula algebraica de un hiperplano en $\mathbb{R}^k$, por la cual se determina el valor estimado de la variable dependiente. La fórmula general de regresión entre la variable dependiente y la(s) variable(s) independiente(s) está dada como

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

donde b_0, b_1, ..., b_k son parámetros desconocidos. El error estadístico ε, es una variable aleatoria que explica por qué el modelo no ajusta exactamente los datos.

Las constantes b_0, b_1, ..., b_k se determinan de los datos con base en el **método de mínimos cuadrados** que busca minimizar la suma del cuadrado de las diferencias entre los valores observados y los estimados. Este método se describe en el anexo A.

Por otra parte, la existencia o no de una relación lineal entre las dos variables vectoriales se investiga por lo general mediante la elaboración de un diagrama de dispersión o de una gráfica de residuales. Estos diagramas también se emplean para observar si la dispersión vertical (varianza) es aproximadamente igual a lo largo de la línea de regresión. La forma de la relación representada por el diagrama de dispersión puede ser *curvilínea* más que lineal. En el caso de relaciones no lineales, un enfoque frecuente consiste en determinar un método de transformación de valores de una o ambas variables a fin de que la relación de los valores transformados sea lineal. Así, el análisis de regresión lineal puede aplicarse a los valores transformados, y los valores estimados de la variable dependiente pueden transformarse a la escala de medición original.

El término análisis de regresión simple indica que el valor de una variable dependiente se estima con base en una variable independiente, de predicción o regresora. El análisis de

regresión múltiple se ocupa de la estimación del valor de una variable de respuesta con base en dos o más variables independientes.

En la práctica se dan muchas situaciones en las que el análisis de regresión se puede aplicar convenientemente y con todo éxito. Cuando se aplica para hacer pronósticos, sus dos principales ventajas son:

- Se puede utilizar para explicar lo que sucede a la variable dependiente cuando la(s) variable(s) independiente(s) cambia(n).
- El uso de un modelo estadístico para poner al descubierto y medir la relación, si es que existe.

2.3. MODELO DE REGRESIÓN LINEAL SIMPLE

La forma más sencilla del modelo de regresión, como ya se mencionó, es cuando solo existe una variable independiente. Así, la ecuación de regresión se reduce a una línea recta, es decir,

$$y = b_0 + b_1 x + \varepsilon$$

donde b_0 y b_1 son parámetros desconocidos que definen la posición e inclinación de la recta, y ε es un componente aleatorio del error.

Cabe indicar, que las consideraciones básicas del modelo del análisis de regresión que se presenta en este capítulo son que

- El error aleatorio ε tiene media cero y varianza σ^2. Además, se supone que los $\{\varepsilon\}$ son variables aleatorias no correlacionadas, es decir, que el valor de cualquiera de ellos, no depende de otro.
- La variable de respuesta y es una variable aleatoria, cuya media para cada valor de x es

  $$E(y|x) = b_0 + b_1 x$$

  con varianza

$$Var(y|x) = Var(b_0 + b_1 x + \varepsilon) = \sigma^2$$

Así, la media de y es una función lineal de x, aunque su varianza no depende del valor de la variable regresora. Además, dado que los errores no están correlacionados, las respuestas tampoco lo están.

- La variable independiente x es una variable determinista, que está controlada por el analista de datos, y se puede medir con error despreciable.[4]

En la figura 2.1 se puede apreciar gráficamente el segundo supuesto: cada observación de la variable dependiente tiene una distribución normal y su desviación estándar es la misma.

A las constantes b_0 y b_1 se les suele llamar **coeficientes de regresión**. Éstos tienen una interpretación simple y, frecuentemente útil. El parámetro b_0, conocido como la *ordenada en el origen*, indica cuánto es y cuando $x = 0$. La *pendiente* b_1, es el cambio de la media de la distribución de y producido por un cambio unitario en x.

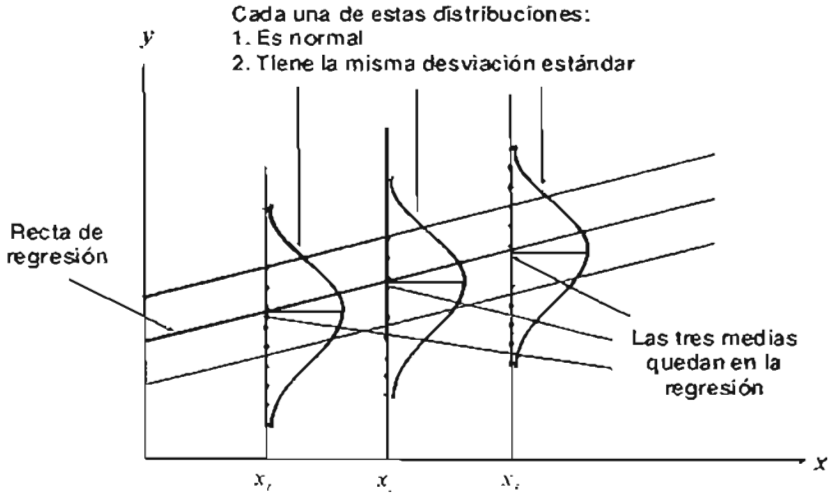


Figura 2.1

[4] El modelo de regresión lineal que se describe en esta tesis, se basa en este supuesto (x es una variable matemática continua). Sin embargo, si x es una variable aleatoria, bajo ciertas circunstancias (como que y y x son variables aleatorias independientes con distribución conjunta), la estimación de parámetros, pruebas y predicción, descritos más adelante, siguen siendo válidos.

Por otro lado, el análisis de regresión, tradicionalmente ha recurrido al método de mínimos cuadrados (ver anexo A.1), para obtener las estimaciones de estos coeficientes a partir de una muestra de observaciones sobre las variables y y x, que como ya se indicó, se puede adquirir por un experimento controlado, diseñado en forma específica para recolectar los datos, a través de un estudio observacional, o a partir de registros históricos existentes.

Las fórmulas que se tienen para calcular los valores b_0 y b_1 son

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

con $\bar{x} = \frac{\sum_{i=1}^{n}(x_i)}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n}(y_i)}{n}$, y n el número de observaciones con los que se estima la regresión.

Por lo tanto, el modelo ajustado de regresión lineal simple es

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

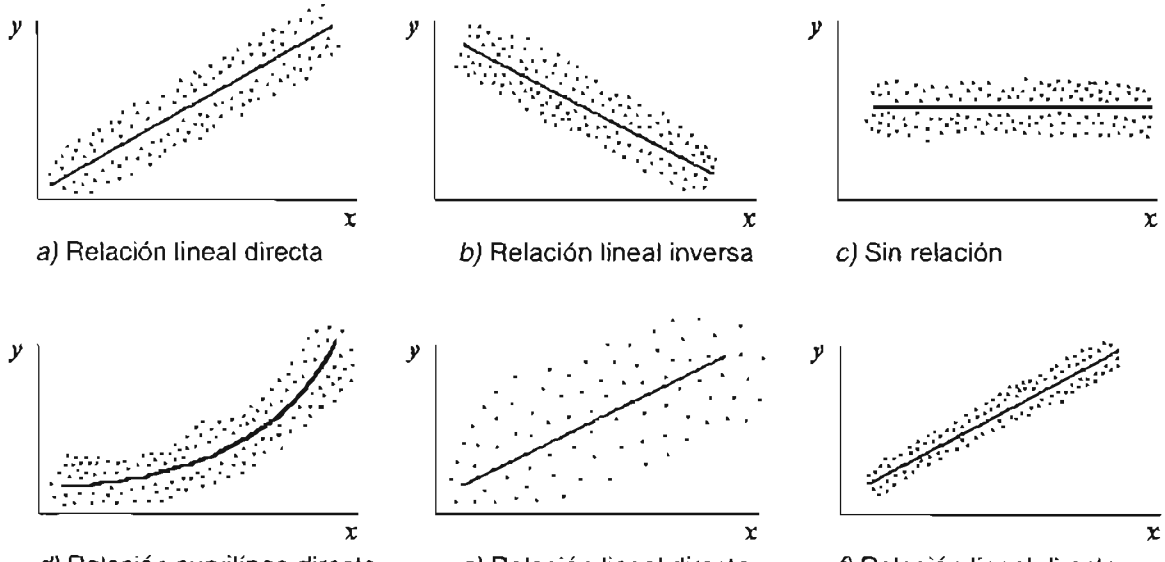
Cabe señalar que la última ecuación produce un estimado puntual, de la media de y para una determinada x.

Es común encontrar en algunos libros, que al numerador de la fórmula del estimador de b_1, se le llame la suma de cuadrados de los productos cruzados de x_i y y_i, y al denominador la suma de cuadrados de las x_i.

2.4. DIAGRAMA DE DISPERSIÓN

Un *diagrama de dispersión* es una gráfica en la que cada punto trazado representa un par de valores observados de las variables independientes y dependiente. El eje horizontal contiene el conjunto de valores necesarios de la variable independiente x, mientras que el vertical tiene una escala adecuada para los valores de la variable dependiente y.

Si el diagrama de dispersión indica en general una relación lineal, se ajusta una recta a los datos. La ubicación precisa de ésta es determinada, como ya se ha señalado, por el método de mínimos cuadrados. Tal como se ilustra a continuación, una línea de regresión con pendiente positiva indica una dependencia directa entre las variables, si es negativa muestra una relación inversa y si es cero demuestra que x e y no tienen conexión entre sí. Además, el grado de dispersión vertical de los puntos trazados respecto de la recta de regresión indica el grado de asociación entre las dos variables.



a) Relación lineal directa

b) Relación lineal inversa

c) Sin relación

d) Relación curvilínea directa

e) Relación lineal directa con menor grado de relación que en *a*.

f) Relación lineal directa con mayor grado de relación que en *a*.

Figura 2.2

2.5. RESIDUALES Y SU GRÁFICA

Como ya se indicó, en el modelo de regresión lineal simple, existe un error ε que representa la inexactitud del modelo, ya sea por falta de variables de predicción que deben de incluirse, o a otras causas. La diferencia entre el valor observado y el ajustado correspondiente recibe el nombre de **residual** de ese dato, se denota con e y matemáticamente, es

$$e = y - \hat{y}.$$

Una *gráfica de residuales* se obtiene trazando los residuales e respecto de la variable independiente x (ver ejemplo en la figura 2.3) o alternativamente, referente a los valores $\hat{y}$ de la línea de regresión ajustada. Estas representaciones sirven como opción o complemento al uso del diagrama de dispersión para investigar si los supuestos sobre linealidad e igualdad de las varianzas parecen satisfacerse, es decir, si la desviación estándar es aproximadamente igual a lo largo de la línea de regresión. Este tipo de gráficas son particularmente importantes en el análisis de regresión múltiple, como se indica en el siguiente capítulo.

Además, el conjunto de residuales de los datos muestrales también sirve de base para calcular el error estándar del estimador.

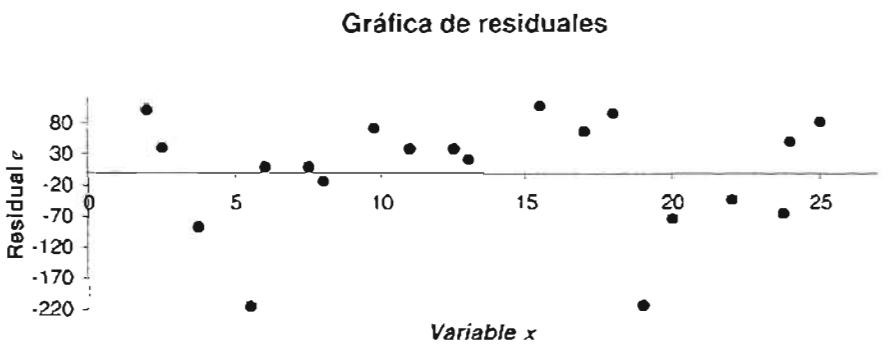


Figura 2.3

2.6 ERROR ESTÁNDAR DE LA ESTIMACIÓN

Además de calcular $\hat{b}_0$ y $\hat{b}_1$, se requiere un estimador insesgado de σ^2 para probar hipótesis y elaborar intervalos pertinentes al modelo de regresión. En el caso ideal, éste no debería depender de la adecuación del modelo ajustado. Eso sólo es posible cuando hay varias observaciones de y para cuando menos un valor de x, o cuando se dispone de información anterior acerca de σ^2. Cuando no se puede usar este método, $\hat{\sigma}^2$ se obtiene de la suma de cuadrados de residuales, o suma de cuadrados del error[5], la cual tiene $n-2$ grados de libertad, porque estos se asocian con los estimadores $\hat{b}_0$ y $\hat{b}_1$ que se usan como estimaciones de b_0 y b_1 en la ecuación de regresión muestral para obtener $\hat{y}_i$. Por tanto,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(e_i^2\right)}{n-2} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2} = MS_{Res}.$$

La cantidad se llama **cuadrado medio residual**. La raiz cuadrada de esta medida se llama a veces, el **error estándar de la regresión**, que es la dispersión o desviación estándar de los valores de y observados en la muestra, alrededor de la recta de regresión.

De tal forma, la fórmula de desviaciones por la cual se estima este último valor, con base en datos muestrales es

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}}$$

[5] MONTGOMERY, Douglas C., PECK, Elizabeth A. y Geoffrey VINNING. Introducción al análisis de regresión lineal. México, p. 21, 22, 537 y 538. En estas páginas, se ven los pasos a seguir para obtener este estimador, así como la demostración de que el valor esperado de la suma de cuadrados del error es $(n-2)\,\sigma^2$.

2.7. COEFICIENTE DE CORRELACIÓN

Por lo general, se desea medir el grado de dependencia entre x e y, al igual que observarla en un diagrama de dispersión. La medida relativa que se usa para este fin es el ***coeficiente de correlación***,[6] que es un valor numérico entre –1 y +1 y que mide la fuerza de la relación lineal entre dos variables cuantitativas.

Éste existe para una población de valores y para cada muestra que se extrae de ella. El símbolo que se utiliza para una población es ρ, la letra griega rho, para una muestra, este coeficiente se representa por la letra r.

Una correlación de +1 indica una dependencia lineal perfecta, un valor de 0 indica que no existe y un valor de –1 habla de una relación negativa óptima. Estos valores rara vez aparecen en situaciones reales, aunque constituyen excelentes puntos de referencia para evaluar el coeficiente de cualquier conjunto de datos.

Para calcularlo, se utiliza la siguiente fórmula:

$$r = \frac{n\sum_{i=1}^{n}(x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n}\left(x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}\left(y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

En el caso de la regresión simple, el signo del coeficiente de correlación siempre es el mismo que el del parámetro b_1.

[6] Para considerar esta medida, es necesario que y y x sean variables aleatorias independientes, con distribución normal bivariada. Asimismo, que la distribución condicional de y dada x es normal, con media condicional $E(y|x) = b_0 + b_1 x$ y varianza condicional $\sigma^2_{y.x}$.

2.8. COEFICIENTE DE DETERMINACIÓN

Otro estadístico que se consulta con frecuencia en el análisis de regresión es el coeficiente de determinación simple (r^2). Éste es útil porque mide la proporción o porcentaje de variabilidad en la variable dependiente, y, que se puede explicar por medio de la variable predictora, x.

No es coincidencia que se use el mismo símbolo para éste (r^2), que para el coeficiente de correlación (r). De hecho, el primero es igual al cuadrado de la segunda medida.

Para calcular r^2 (como ya se indicó), basta con elevar al cuadrado el coeficiente de correlación, r. No obstante, también se pueden utilizar las siguientes fórmulas:

$$r^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

o

$$r^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Una parte importante de la evaluación de la suficiencia del modelo de regresión lineal simple es la prueba de hipótesis estadísticas en torno a los parámetros de éste y la construcción de ciertos intervalos de confianza. Las primeras se estudian en la siguiente sección y los segundos en la posterior.

Estos procedimientos requieren hacer la suposición adicional de que los errores ε_i del modelo estén distribuidos normalmente e independientes, lo cual se abrevia $NID\left(0,\ \sigma^2\right)$.

2.9. PRUEBA DE HIPÓTESIS

Una vez realizada la estimación de los parámetros, ha de verificarse si efectivamente el ajuste será el adecuado o no. Para ello, se emplean las pruebas descritas a continuación.

2.9.1. PRUEBAS *t*

Un estadístico importante en el análisis de regresión, es *t*, por ejemplo, supóngase que se desea probar la hipótesis que la pendiente es igual a una constante, dígase $b_{1,0}$. Las hipótesis apropiadas son

$$H_0: b_1 = b_{1,0}$$

$$H_1: b_1 \neq b_{1,0}$$

donde se ha inferido una alternativa de dos lados. El estadístico de prueba[7] es

$$t_0 = \frac{\hat{b}_1 - b_{1,0}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}}$$

La distribución muestral adecuada para esta prueba es la *t* con $(n - 2)$ grados de libertad.[8] Se rechazará H_0, si $|t_0| > t_{(\alpha/2,\, n-2)}$. Esta situación se ilustra en la figura 2.4.

Cabe hacer notar que el denominador en la ecuación de t_0 se le llama el **error estándar estimado de la pendiente**. Esto es,

---

[7] HINES, William W. y Douglas C. MONTGOMERY. Probabilidad y estadística para ingeniería y administración. México, p. 532 y 533. En estas páginas se puede encontrar el procedimiento que se sigue para encontrar este estadístico de prueba.

[8] Se pierden dos grados de libertad porque se estiman los parámetros poblacionales (b_0 y b_1) usando los estadísticos muestrales $\hat{b}_0$ y $\hat{b}_1$.

$$se\left(\hat{b}_1\right)=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-2)\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}$$

También, se puede probar la hipótesis acerca de la ordenada al origen:

$$H_0: b_0 = b_{0,0}$$

$$H_1: b_0 \neq b_{0,0}$$

donde el estadístico a usar es

$$t_0=\frac{\hat{b}_0-b_{0,0}}{\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{n-2}\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right]}}$$

Se rechazará la hipótesis nula, si $|t_0| > t_{(\alpha/2,\, n-2)}$. En la figura 2.4 se puede apreciar la región de rechazo y de aceptación de esta prueba.

Además, $se\left(\hat{b}_0\right)=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{n-2}\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right]}$ es llamado el **error estándar de la ordenada al origen**.

Figura 2.4

2.9.2. PRUEBA DE SIGNIFICANCIA DE LA REGRESIÓN

2.9.2.1. Prueba de hipótesis para la pendiente.

Un caso especial muy importante de la primera prueba, es el de probar la hipótesis nula que considera que el coeficiente de x es nulo, lo cual implica que ante cambios de la variable independiente la característica y no se ve afectada, es decir, x e y no tienen correlación en la población. La hipótesis nula y la alternativa de dos colas a probar son:

$$H_0: b_1 = 0$$

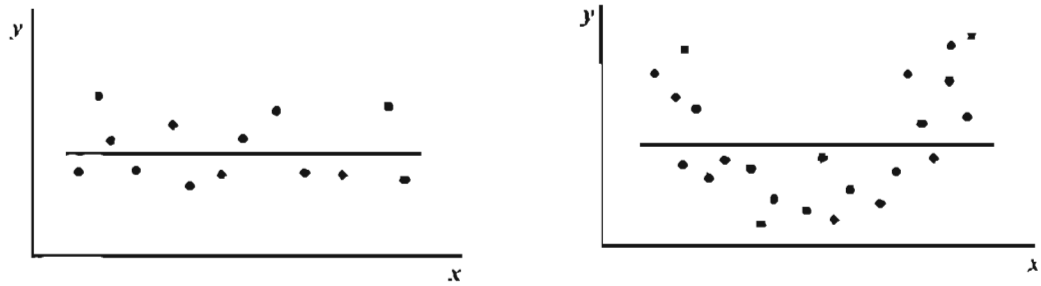
$$H_1: b_1 \neq 0$$

El estadístico de prueba se calcula de la siguiente manera.

$$t_0 = \frac{\hat{b}_1}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}}$$

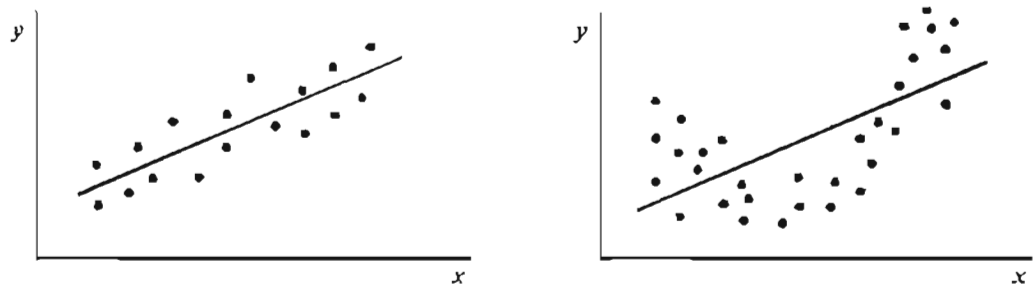
Si se utiliza un nivel de significación α, la regla de decisión señala que si la t_0 calculada se localiza en el área entre $-t_{(\alpha/2,\, n-2)}$ y $+t_{(\alpha/2,\, n-2)}$, se aceptará la hipótesis nula, en caso contrario se rechazará. Lo anterior se ve en la figura 2.4.

Esta hipótesis se relaciona con la *significancia de la regresión*. El hecho de no desechar H_0 es equivalente a concluir que no existe dependencia lineal entre y y x. Este caso se muestra en la figura 2.5. Obsérvese que esto puede implicar que x tiene muy poco valor para explicar la variación de y y que un mejor estimador para cualquier valor de la variable de respuesta es $\hat{y} = \bar{y}$ (gráfica 2.5a), o que la verdadera relación entre y y x no es lineal (gráfica 2.5b).



a) b)

Figura 2.5



a) b)

Figura 2.6

Alternativamente, si se rechaza la hipótesis nula, esto implica que x sí tiene valor para explicar la variabilidad de y. Esto se ilustra en la figura 2.6. Sin embargo, lo anterior puede significar dos cosas: que el modelo de regresión lineal es adecuado (gráfica 2.6a), o que aun cuando hay un efecto rectilíneo de x, se podrían obtener mejores resultados con la adicción de términos de polinomios de mayor orden en la variable independiente (gráfica 2.6b).

2.9.2.2. Prueba F.

Para probar la significancia de la regresión, también se puede usar un método de **análisis de la varianza**. Éste se basa en una partición de la variabilidad total de la variable de respuesta. Para producir esta división se comienza con la identidad

$$y_i - \bar{y} = \left(\hat{y}_i - \bar{y}\right) + \left(y_i - \hat{y}_i\right)$$

Se elevan al cuadrado ambos lados de la ecuación y se suma para todas las n observaciones, obteniéndose

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 + 2\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)\left(y_i - \hat{y}_i\right)$$

Como

$$2\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)\left(y_i - \hat{y}_i\right) = 2\sum_{i=1}^{n}\hat{y}_i\left(y_i - \hat{y}_i\right) - 2\sum_{i=1}^{n}\bar{y}\left(y_i - \hat{y}_i\right) = 2\sum_{i=1}^{n}\left(\hat{y}_i e_i\right) - 2\bar{y}\sum_{i=1}^{n}\left(e_i\right) = 0$$

porque $\sum_{i=1}^{n}\left(e_i\right) = 0$, así como la suma de los residuales ponderados por el valor ajustado $\hat{y}_i$ correspondiente también es igual a cero. Por consiguiente,

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

El lado izquierdo de la anterior ecuación, es la **suma corregida de cuadrados de las observaciones**, que cuantifica la variación total de las observaciones. Los dos componentes de la derecha miden, respectivamente, la cantidad de variabilidad en los datos y_i explicada por la línea de regresión, y la residual que queda sin explicar por ésta. Se le suele llamar a $\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$ la **suma de cuadrados del modelo** o **de la regresión** y a $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ la **suma de cuadrados del error** o **de los residuales**.

El primer elemento de la última ecuación tiene $(n - 1)$ grados de libertad. El segundo y tercer término tienen 1 y $(n - 2)$, respectivamente.

El procedimiento de prueba suele arreglarse en una **tabla de análisis de la varianza (ANOVA)**, como la del cuadro 2.1.

Por lo tanto, para probar las hipótesis

$$H_0: b_1 = 0$$

$$H_1: b_1 \neq 0$$

se calcula el estadístico de prueba

$$F_0 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \Big/ n-2}$$

y se rechazará H_0, si este último tiene un valor grande, es decir, $F_0 > F_{(\alpha,\, 1,\, n-2)}$. En la figura 2.7 se indican las regiones de aceptación y de rechazo de esta prueba.

Tabla 2.1

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0
Regresión	$\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$	1	$\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$	$F_0 = \dfrac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \Big/ n-2}$
Residual	$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$	$n - 2$	$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \Big/ n-2$	
Total	$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$	$n - 1$		

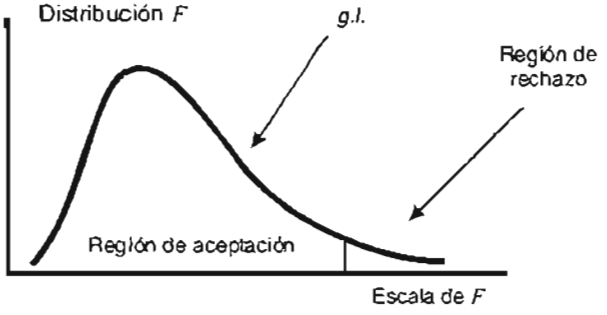


Figura 2.7

No obstante que en la regresión simple, esta prueba es equivalente a la t, esta última es algo más adaptable, porque se puede usar para probar hipótesis alternativas unilaterales (H_0: $b_1 > 0$ o H_0: $b_1 < 0$), mientras que la F únicamente considera la opción bilateral. Cabe señalar que actualmente, todos los programas de cómputo para regresión producen, tanto el estadístico t, como una tabla ANOVA parecida a la del cuadro 2.1. Sin embargo, la verdadera utilidad del análisis de regresión está en los modelos de regresión múltiple.

2.10. ESTIMACIÓN DE INTERVALOS DE CONFIANZA Y DE PREDICCIÓN

Dado que una estimación puntual no proporciona información suficiente acerca del parámetro poblacional, el cálculo de un intervalo podría resultar más útil. Se puede elegir entre dos tipos: **intervalo de predicción** (para una evaluación específica de y) o **intervalo de confianza** (para el valor esperado de la variable de respuesta). Los primeros se usan para pronosticar un valor individual de y dada una x, y es por tanto un intervalo de probabilidad. Los segundos se utilizan para estimar el valor medio de la variable dependiente para una observación determinada de x.

También es posible elaborar estimaciones del intervalo de confianza de la pendiente y la ordenada al origen, como se describe a continuación.

2.10.1. INTERVALOS DE CONFIANZA DE b_1 Y b_2

Si los errores se distribuyen en forma normal e independiente, entonces la distribución de muestreo tanto de

$$\frac{\hat{b}_1 - b_1}{se\left(\hat{b}_1\right)} \qquad \text{y} \qquad \frac{\hat{b}_0 - b_0}{se\left(\hat{b}_0\right)}$$

es t, con $(n - 2)$ grados de libertad. Por consiguiente, un intervalo de confianza de $100(1 - \alpha)\%$ para la pendiente b_1 se determina con

$$t_{(\alpha/2,\, n-2)} - \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} \le \hat{b}_1 \le t_{(\alpha/2,\, n-2)} + \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

y para la ordenada al origen b_0 es

$$t_{(\alpha/2,\, n-2)} - \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]} \le \hat{b}_0 \le t_{(\alpha/2,\, n-2)} + \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]}$$

Una interpretación usual que poseen estos intervalos es la de frecuencia, por tanto, si se tomaran muestras aleatorias de igual tamaño, a los mismos valores de x, y se construyeran, por ejemplo, intervalos de confianza de 99% de b_0 en cada una de ellas, entonces el 99% de éstos contendrán el valor verdadero de la ordenada al origen.

2.10.2. INTERVALOS DE CONFIANZA DE LA RESPUESTA MEDIA

Una aplicación importante de un modelo de regresión es estimar la respuesta media, $E(y)$, para un determinado valor de la variable regresora x, denótese éste con x_0. Puesto que $E(y|x_0) = b_0 + b_1x_0$, se puede obtener un estimador puntual insesgado[9] de éste, a partir del modelo ajustado como sigue

$$\widehat{E(y|x_0)} = \hat{y}_0 = \hat{b}_0 + \hat{b}_1 x_0$$

Adviértase que $\hat{y}_0$ se distribuye normalmente, porque $\hat{b}_0$ y $\hat{b}_1$ lo hacen de ese mismo modo. Entonces, la varianza de $\hat{y}_0$ es

$$V(\hat{y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

Por tanto, un intervalo de confianza de $100(1 - \alpha)\%$ para la respuesta media o en torno a la línea de regresión verdadera en el punto $x = x_0$ puede calcularse a partir de

$$\hat{y}_0 - t_{(\alpha/2,\, n-2)} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n-2)} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}$$

$$\leq E(y|x_0) \leq \hat{y}_0 + t_{(\alpha/2,\, n-2)} \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n-2)} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}$$

[9] $\hat{y}_0$ es un estimador puntual insesgado de $E(y|x_0)$, puesto que $\hat{b}_0$ y $\hat{b}_1$ lo son de b_0 y b_1.

Nótese que el ancho de este intervalo es una función de x_0, además de que es un mínimo en $x_0 = \bar{x}$, y se ensancha conforme $\left|x_0 - \bar{x}\right|$ aumenta.

2.10.3. INTERVALOS DE PREDICCIÓN

Ahora se estimará un intervalo para una observación futura y_0. El intervalo para la respuesta media en $x = x_0$, es inapropiado para este problema, dado que éste es para el valor esperado de y, un parámetro de población, y no es una declaración de probabilidad sobre futuros datos, a partir de esa distribución.

Supóngase que y_0 la observación futura en $x = x_0$, y sea $\hat{y}_0$ su estimador. Adviértase que la variable aleatoria

$$\Psi = y_0 - \hat{y}_0$$

tiene una distribución normal, con media cero y varianza

$$Var\left(y_0 - \hat{y}_0\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]$$

debido a que y_0 es independiente de $\hat{y}_0$. De modo que, el intervalo de predicción del $100(1 - \alpha)$ por ciento de confianza para una observación futura en x_0 es

$$\hat{y}_0 - t_{(\alpha/2,\, n-2)}\sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)}\left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)}$$

$$\leq y_0 \leq \hat{y}_0 + t_{(\alpha/2,\, n-2)}\sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)}\left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)}$$

Como puede notarse la diferencia entre las ecuaciones para estimara un intervalo de confianza y de predicción, es únicamente que en la última se incluye el 1 dentro de la raíz. De esta manera, el intervalo de una observación futura resulta ser más amplio que el de un valor medio. Lo anterior es lógico, dado que predecir únicamente un valor para una sola variable es más difícil que estimar el promedio de una población de éstos.

La figura 2.8 ilustra la diferencia entre los intervalos de predicción y los intervalos de confianza.

2.11. REGRESIÓN POR EL ORIGEN

Existen casos de regresión, en los que la recta debe pasar por el origen para un mejor ajuste de las observaciones, por ejemplo, para analizar datos de procesos químicos y de manufactura. Es decir, se tendrá un **modelo de regresión sin ordenada al origen**, el cual es

$$y = b_1 x + \varepsilon$$

Dadas n observaciones (y_i, x_i), $i = 1, 2, \ldots, n$, la función de mínimos cuadrados es

$$S(b_1) = \sum_{i=1}^{n} (y_i - b_1 x_i)^2$$

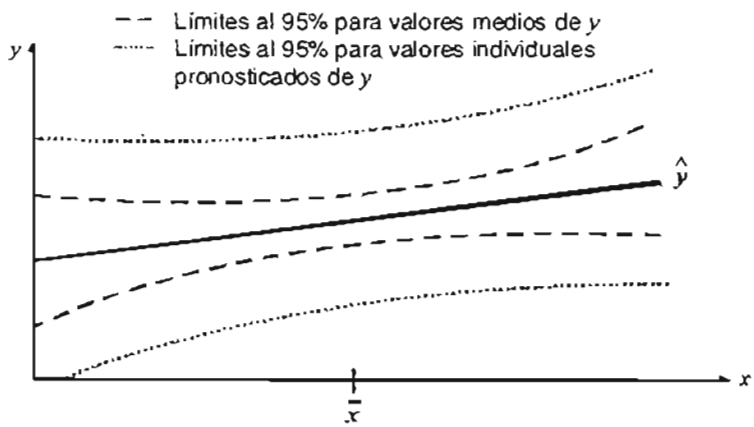


Figura 2.8

La única ecuación normal es

$$\hat{b}_1 \sum_{i=1}^{n} \left(x_i^2\right) = \sum_{i=1}^{n} \left(y_i x_i\right)$$

y el **estimador insesgado de la pendiente por mínimos cuadrados** es

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

Por tanto, el **modelo de regresión ajustado** es

$$\hat{y} = \hat{b}_1 x$$

El estimador de σ^2 con $n-1$ grados de libertad, es

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n-1} = \frac{\sum_{i=1}^{n} \left(y_i^2\right) - \hat{b}_1 \sum_{i=1}^{n} \left(y_i x_i\right)}{n-1}$$

También se pueden probar hipótesis y establecer intervalos de confianza y de predicción, para ello se supone normalidad en los errores.

Para la prueba de significancia de la regresión, el estadístico es

$$t_0 = \frac{\hat{b}_1}{\sqrt{\dfrac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{(n-1)\sum_{i=1}^{n} \left(x_i^2\right)}}}$$

y para la F

$$F_0 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i^2\right)}{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \Big/ n-1}$$

Se rechazará H_0, si $|t_0| > t_{(\alpha/2,\, n-1)}$, para la primer prueba y para la segunda, si $F_0 > F_{(\alpha,\, 1,\, n-1)}$

Así, el **intervalo de confianza de 100(1 – α) por ciento para b_1** es

$$\hat{b}_1 - t_{(\alpha/2,\, n-1)}\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-1)\sum_{i=1}^{n}\left(x_i^2\right)}} \le b_1 \le \hat{b}_1 + t_{(\alpha/2,\, n-1)}\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-1)\sum_{i=1}^{n}\left(x_i^2\right)}}$$

y para **la respuesta media en $x = x_0$** es

$$\hat{y}_0 - t_{(\alpha/2,\, n-1)}\sqrt{\frac{x_0^2\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-1)\sum_{i=1}^{n}\left(x_i^2\right)}} \le E\left(y|x_0\right) \le \hat{y}_0 + t_{(\alpha/2,\, n-1)}\sqrt{\frac{x_0^2\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-1)\sum_{i=1}^{n}\left(x_i^2\right)}}$$

Además, el **intervalo de predicción de 100(1 – α) por ciento para una observación futura en $x = x_0$** es

$$\hat{y}_0 - t_{(\alpha/2,\, n-1)}\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-1)}\left[1+\frac{x_0^2}{\sum_{i=1}^{n}\left(x_i^2\right)}\right]}$$

$$\le y_0 \le \hat{y}_0 + t_{(\alpha/2,\, n-1)}\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-1)}\left[1+\frac{x_0^2}{\sum_{i=1}^{n}\left(x_i^2\right)}\right]}$$

Por último, el análogo de r^2 en el modelo sin ordenada al origen será

$$r_0^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i^2\right)}{\sum_{i=1}^{n}\left(y_i^2\right)} \qquad \text{o} \qquad r_0^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i^2\right)}$$

el cual indica la proporción de variabilidad respecto al origen (cero) que explica la regresión. A menudo, se encuentra que r_0^2 es mayor que r^2, aún cuando el cuadrado medio residual (que es una medida razonable de la calidad general de ajuste) para el modelo con ordenada al origen sea menor que el de sin b_1, Este se debe a que r_0^2 se calcula con valores de sumas de cuadrados no corregidas.

2.12. EJEMPLO

Un analista toma una muestra aleatoria de 10 embarques recientes por camión realizados por una compañía y registra la distancia en millas y el tiempo de entrega al medio día más cercano a partir del momento en que el embarque estuvo listo para su carga. La tabla 2.2 muestra estos datos.[10]

A continuación se buscará ajustar estos datos a una recta, se calcularán todos los parámetros del análisis de regresión lineal antes descritos, y se explicará su significado.

Tabla 2.2

Embarque muestreado i	Distancia (millas) x_i	Tiempo de entrega (días) y_i
1	825	3.5
2	215	1.0
3	1,070	4.0
4	550	2.0
5	480	1.0
6	920	3.0
7	1,350	4.5
8	325	1.5
9	670	3.0
10	1,215	5.0

[10] KAZMIER, Leonard J. Estadística aplicada a la administración y a la economía. México, p. 256.

Primero se elaborará el diagrama de dispersión de los datos (figura 2.9), para averiguar si los puntos trazados parecen tener una tendencia lineal, y de esta manera efectuar el análisis de regresión.

Una vez que se ha ilustrado la dispersión y tendencia de los datos, que resulta ser en general lineal, se procede a calcular la ecuación de regresión, para esto se utilizan los datos de la tabla 2.3.

Entonces,

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i y_i) - n\overline{xy}}{\sum_{i=1}^{n}(x_i^2) - n\overline{x}^2} = \frac{26370 - (10)(762)(2.85)}{7104300 - (10)(762)^2} = 0.003585132 \approx 0.0036$$

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x} = 2.85 - (0.003585132)(762) = 0.1181$$

Por lo tanto, la ecuación de regresión es

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x = 0.1181 + 0.0036x$$

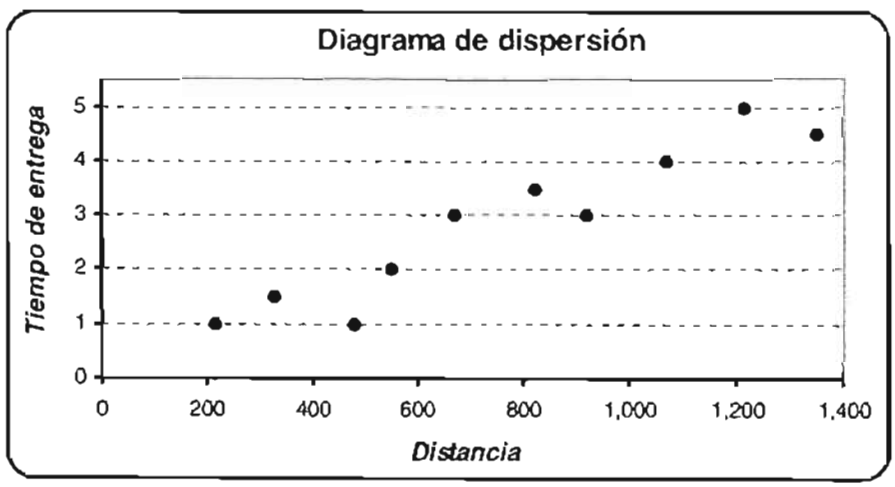


Figura 2.9

Tabla 2.3

Embarque muestreado i	Distancia (millas) x_i	Tiempo de entrega (días) y_i	xy	x^2	y^2
1	825	3.5	2,887.5	680,625.0	12.3
2	215	1.0	215.0	46,225.0	1.0
3	1,070	4.0	4,280.0	1,144,900.0	16.0
4	550	2.0	1,100.0	302,500.0	4.0
5	480	1.0	480.0	230,400.0	1.0
6	920	3.0	2,760.0	846,400.0	9.0
7	1,350	4.5	6,075.0	1,822,500.0	20.3
8	325	1.5	487.5	105,625.0	2.3
9	670	3.0	2,010.0	448,900.0	9.0
10	1,215	5.0	6,075.0	1,476,225.0	25.0
Totales	7,620	28.5	26,370.0	7,104,300.0	99.8
Medias	$\bar{x} = 762$	$\bar{y} = 2.85$			

La ordenada al origen (0.1181) puede interpretarse como un tiempo de entrega fijo mínimo. La pendiente de la ecuación de regresión indica que por cada milla de distancia adicional, el tiempo de entrega aumentará en promedio 0.0036 días.

Una vez que se ha encontrado la ecuación de regresión se procede a calcular las estimaciones de la variable y. Estos valores se presentan en la tabla 2.4 y en la figura 2.10, se aprecia la línea de regresión de los datos muestrales sobre el diagrama de dispersión.

Asimismo, una vez que se obtiene el valor ajustado de la variable y, se calcula cada uno de los residuales. En la última columna de la tabla 2.4, se alcanzan a distinguir.

Como se puede ver en la figura 2.11, el monto de la desviación entre cada valor muestreado y el estimado correspondiente, se reduce al mínimo.

Tabla 2.4

Embarque muestreado i	Distancia (millas) x_i	Tiempo de entrega (días) y_i	Valor Ajustado $\hat{y}_i$	Residual $e = y_i - \hat{y}_i$
1	825	3.50	3.08	0.42
2	215	1.00	0.89	0.11
3	1,070	4.00	3.95	0.05
4	550	2.00	2.09	-0.09
5	480	1.00	1.84	-0.84
6	920	3.00	3.42	-0.42
7	1,350	4.50	4.96	-0.46
8	325	1.50	1.28	0.22
9	670	3.00	2.52	0.48
10	1,215	5.00	4.47	0.53

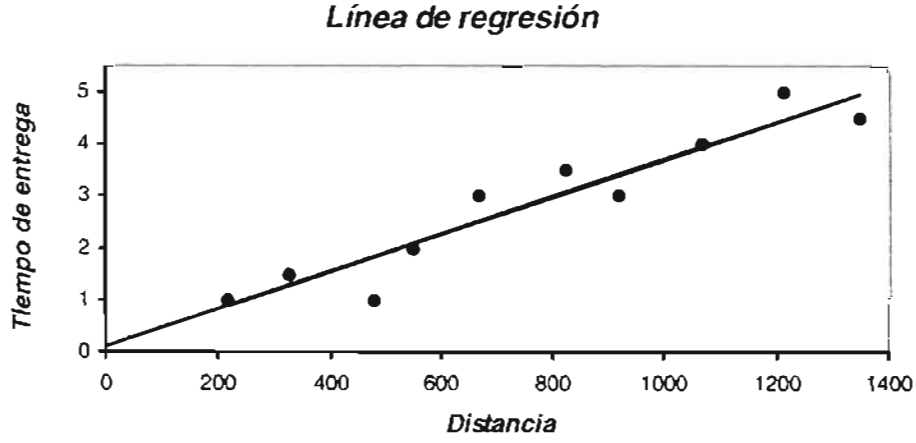


Figura 2.10

Como la distancia y el tiempo de entrega son variables aleatorias, se puede calcular el coeficiente de correlación de los datos muestrales, éste es

$$r = \frac{(10)(26370) - (7620)(28.5)}{\sqrt{(10)(7104300) - (7620)^2}\sqrt{(10)(99.75) - (28.5)^2}} = +0.9489 \approx 0.95$$

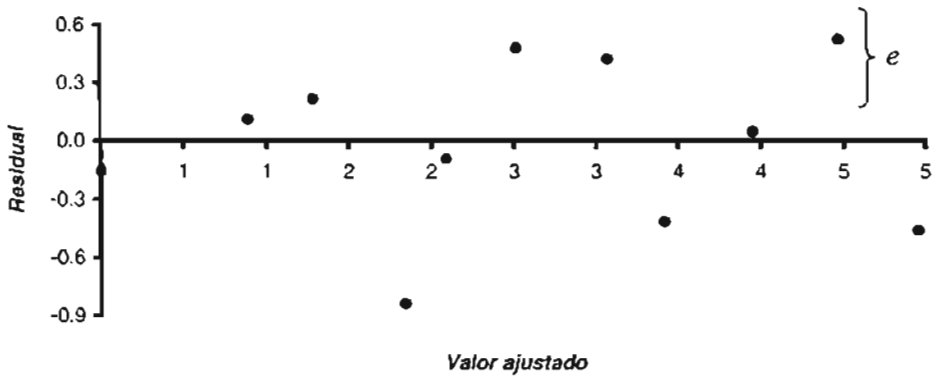


Figura 2.11

Dado que el valor de r es igual a 0.95 (muy cercano a 1), se puede decir que existe una muy fuerte correlación (relación o asociación) lineal positiva entre la distancia a recorrer y el tiempo de entrega del embarque. Es decir, a medida que una sube, la otra también lo hará y viceversa.

Por otro lado, si se quiere saber qué proporción de la varianza en la variable dependiente estadísticamente es explicada por la ecuación de regresión, se utiliza la medida de la bondad de ajuste r^2. Así,

$$r^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = \frac{16.682}{18.525} = 0.9005 \approx 0.90$$

Puesto que, el coeficiente de determinación es 0.90, se puede concluir que la recta de regresión $\hat{y} = 0.1181 + 0.0036x$ estadísticamente explica alrededor del 90% de la variación total en el tiempo de entrega mediante la distancia implicada. Además, se puede concluir que sólo alrededor de 10% de la varianza queda sin explicar.

Se calcula ahora el error estándar del estimador.

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}\left(e_i^2\right)}{n-2}} = \sqrt{\frac{1.8434}{8}} = 0.480026 \approx 0.48$$

Se interpreta el error estándar de la estimación, de la siguiente manera: la desviación estándar entre los valores muestrales y_i y sus estimaciones mediante la recta de regresión es 0.48. Así, para valores pronosticados de y, se esperará que tengan un error similar.

A continuación se probará la hipótesis de que el valor del coeficiente de la pendiente es significativamente diferente a cero a un nivel de significancia de 5%. Se tiene que:

$$H_0: b_1 = 0 \quad \text{vs} \quad H_1: b_1 \neq 0$$

Estadístico de prueba

$$t_0 = \frac{\hat{b}_1}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}} = \frac{0.0036}{\sqrt{\dfrac{1.8434}{8(1297860)}}} = +8.54$$

t crítica

$$t_{(\alpha/2=0.025,\ n-2=8)} = \pm 2.306$$

Como puede apreciarse en la figura 2.12, el valor calculado del estadístico de prueba +8.54 se encuentra en la región de rechazo, así la hipótesis nula de que pendiente de la ecuación de regresión es igual a cero se rechaza, y se concluye que existe una relación lineal entre distancia y tiempo de entrega.

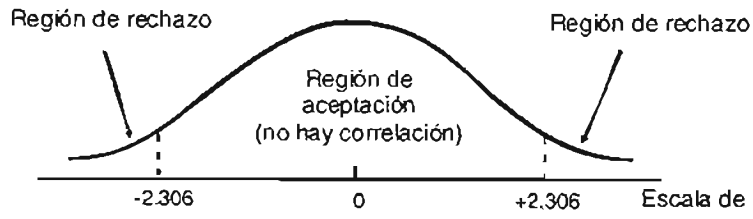


Figura 2.12

En seguida se probará la hipótesis nula que establece que x e y no tienen correlación en la población, mediante la prueba F con nivel de significancia de 5%. Se tiene que:

$$H_0: b_1 = 0 \quad \text{vs} \quad H_1: b_1 \neq 0$$

Estadístico de prueba $$F_0 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \Big/ n-2} = \frac{16.8616}{1.8434/8} = 73.18.$$

F crítico $$F_{(\alpha = 0.05,\ 1,\ n-2 = 8)} = 5.32.$$

Puesto que 73.18 > 5.32, se rechaza la hipótesis nula. Se concluye entonces que la relación entre la distancia y el tiempo de entrega es efectivamente lineal.

Con base a los resultados de las dos pruebas anteriores, al coeficiente de determinación, así como del análisis de residuales, se puede concluir que efectivamente el ajuste de los datos es bueno, y que se pueden hacer pronósticos con este modelo.

Usando la ecuación de regresión anteriormente dada, se puede pronosticar el tiempo de entrega a partir del momento en que el embarque esté listo para su carga de un embarque de por ejemplo 1,500 millas. Si $x_0 = 1500$, entonces

$$\hat{y}_0 = 0.1181 + 0.0036(1500) = \textbf{5.52 días}.$$

Asimismo, se puede estimar este tiempo de entrega a través de un intervalo de predicción[11] y de confianza al 95%.

Sean

$\hat{y}_0 = 5.52$, para $x_0 = 1500$

[11] Como se recordará un intervalo de predicción es un intervalo de probabilidad, que sirve para estimar un *valor individual* de la variable independiente.

$$t_{(\alpha/2,\, n-2)} = t_{(0.025,\, 8)} = 2.306$$

$$\sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)}\left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)} = \sqrt{\frac{1.8434}{8}}\sqrt{1 + \frac{1}{10} + \frac{(1500 - 762)^2}{1297860}}$$

$$= 0.48\sqrt{1.5196} = 0.5917 \approx 0.59$$

$$\sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)}\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)} = \sqrt{\frac{1.8434}{8}}\sqrt{\frac{1}{10} + \frac{(1500 - 762)^2}{1297860}}$$

$$= 0.48\sqrt{0.5196} = 0.3460 \approx 0.35$$

Entonces, el intervalo de predicción de 95% es:

$$5.52 - 2.306(0.59) \le y_0 \le 5.52 + 2.306(0.59)$$

$$5.52 - 1.36 \le y_0 \le 5.52 + 1.36$$

$$4.16 \le y_0 \le 6.88$$

Es decir, el pronóstico de tiempo de entrega de un embarque de 1,500 millas, será de entre **4.16 a 6.88 días**.

Por otro lado, el intervalo de confianza para estimar el tiempo de entrega medio respecto de una distancia de acarreo de 1500 millas es:

$$5.52 - 2.306(0.35) \le E\left(y|x_0\right) \le 5.52 + 2.306(0.35)$$

$$5.52 - 0.81 \le E\left(y|x_0\right) \le 5.52 + 0.81$$

$$4.71 \le E\left(y|x_0\right) \le 6.33$$

Lo anterior significa que el tiempo de entrega medio estimado a partir de del momento en que el embarque está listo es de entre **4.71 y 6.33 días**, con una confianza de 95%.

Cabe hacer notar que el intervalo de predicción es levemente más amplio que el intervalo de confianza.

Una hipótesis que también se puede probar es que la ordenada al origen es igual a cero, que se dejó a propósito al final, por el resultado que este emite. Se usa un nivel de significancia de 5%, para

$$H_0: b_0 = 0 \quad \text{vs} \quad H_1: b_0 \neq 0$$

Donde el estadístico de prueba es

$$t_0 = \frac{\hat{b}_0}{\sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]}} = \frac{0.1181}{\sqrt{\frac{1.8434}{8}\left(\frac{1}{10} + \frac{(762)^2}{1297860}\right)}} = 0.3325 \approx 0.33$$

y $t_{(\alpha/2=0.025,\, n-2=8)} = \pm 2.306$

Dado que 0.33 > 2.306, la hipótesis nula no se rechaza. Se concluye que la ordenada al origen es igual a cero. Esto conlleva a tener que ajustar el modelo de regresión a uno sin ordenada al origen (ver sección 2.11), lo anterior suena lógico, dado que si no existe distancia recorrida, no tendría que haber nada de tiempo de embarque.

Utilizando los datos de la tabla 2.3, se tiene que

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}\left(y_i x_i\right)}{\sum_{i=1}^{n}\left(x_i\right)^2} = \frac{26370}{7104300} = 0.003711836 \approx 0.0037$$

Entonces, la ecuación de regresión ajustada es

$$\hat{y} = \hat{b}_1 x = 0.0037x$$

A continuación, se calculan las estimaciones de la variable y, así como los residuales correspondientes (ver tabla 2.5), para posteriormente hacer el diagrama de dispersión y trazar línea de regresión (ver figura 2.13).

Tabla 2.5

Embarque muestreado i	Distancia (millas) x_i	Tiempo de entrega (días) y_i	Valor Ajustado $\hat{y}_i$	Residual $e = y_i - \hat{y}_i$
1	825	3.50	3.06	0.44
2	215	1.00	0.80	0.20
3	1,070	4.00	3.97	0.03
4	550	2.00	2.04	-0.04
5	480	1.00	1.78	-0.78
6	920	3.00	3.41	-0.41
7	1,350	4.50	5.01	-0.51
8	325	1.50	1.21	0.29
9	670	3.00	2.49	0.51
10	1,215	5.00	4.51	0.49

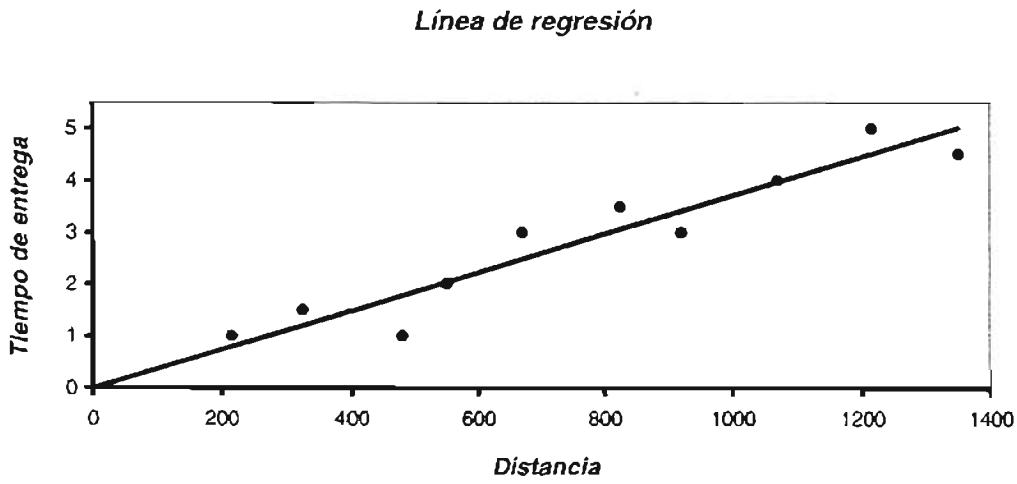


Figura 2.13

En la figura 2.14, se tiene una gráfica de residuales. Se puede apreciar en ella, que éstos se distribuyen aleatoriamente.

El error estándar del modelo es

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-1}} = \sqrt{\frac{1.8689}{9}} = 0.45569 \approx 0.46$$

Se calcula el coeficiente de determinación

$$r_0^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i^2\right)}{\sum_{i=1}^{n}\left(y_i^2\right)} = \frac{97.8811}{99.75} = 0.981264 \approx 0.98$$

Así, esta recta de regresión explica alrededor del 98% de la variación total.

Se procede a calcular el estadístico t para probar H_0: $b_1 = 0$, a un nivel de significancia de 5%, entonces

Figura 2.14

$$t_0 = \frac{\hat{b}_1}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-1)\sum_{i=1}^{n}\left(x_i^2\right)}}} = \frac{0.0037}{\sqrt{\dfrac{1.8689}{9(7104300)}}} \approx 21.64 \qquad \text{y} \qquad t_{(\alpha/2=0.025,\, n-1=9)} = 2.262.$$

Como puede observarse, $t_0 > t$, lo que indica que la hipótesis nula se rechaza, por ende, existe una relación lineal entre distancia y tiempo de entrega.

De igual forma, se calcula el estadístico $F_0 = \dfrac{\sum_{i=1}^{n}\left(\hat{y}_i^2\right)}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 / n-1} = \dfrac{97.8811}{1.8689/9} \approx 471.36$, para comprobar H_0: $b_1 = 0$. Como $F_0 > F_{(\alpha = 0.05,\ 1,\ n-1=9)} = 5.12$, se rechaza la hipótesis nula.

Después de considerar las anteriores pruebas, el coeficiente de determinación, así como que el error estándar de este ajuste, es menor que el que considera ordenada al origen, se llega a la conclusión de que el modelo sin ordenada al origen es mejor.

En base a lo anterior, se pueden calcular futuras observaciones. Por ejemplo, sea

$$x_0 = 1700 \text{ millas}$$

entonces

$$\hat{y}_0 = 0.0037(1700) = \mathbf{6.29\ días}.$$

Además, sean

$\hat{y}_0 = 6.92$, para $x_0 = 1700$

$t_{(\alpha/2,\, n-2)} = t_{(0.025,9)} = 2.262$

$$\sqrt{\frac{x_0^2\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-1)\sum_{i=1}^{n}\left(x_i^2\right)}}=\sqrt{\frac{(1700)^2(1.8689)}{9(7104300)}}\approx 0.29$$

$$\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-1)}\left[1+\frac{x_0^2}{\sum_{i=1}^{n}\left(x_i^2\right)}\right]}=\sqrt{\frac{1.8689}{9}\left(1+\frac{(1700)^2}{7104300}\right)}\approx 0.54$$

Así, el intervalo de confianza de 95% para la respuesta media en $x = x_0$ es

$$6.92 - 2.262\,(0.29) \le E\left(y|x_0\right) \le 6.92 + 2.262(0.29)$$

$$6.92 - 0.66 \le E\left(y|x_0\right) \le 6.92 + 0.66$$

$$6.26 \le E\left(y|x_0\right) \le 7.58$$

Lo anterior significa que el tiempo de entrega medio estimado a partir de del momento en que el embarque está listo es de entre **6.26 y 7.58 días**, con una confianza de 95%.

Además, el intervalo de predicción de 95% es:

$$6.92 - 2.262(0.54) \le y_0 \le 6.92 + 2.262(0.54)$$

$$6.92 - 1.22 \le y_0 \le 6.92 + 1.22$$

$$5.70 \le y_0 \le 8.14$$

Es decir, el pronóstico de tiempo de entrega de un embarque de 1,500 millas, será de entre **5.70 a 8.14 días**.

CAPÍTULO 3

ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

En muchas situaciones de toma de decisiones se precisa utilizar más de una variable para predecir o explicar cierta variable dependiente, por lo que la regresión simple no es adecuada, y se generalizan los planteamientos anteriores mediante la regresión múltiple, la cual permite incluir más de una variable independiente.

En forma general, la ecuación de regresión lineal múltiple es:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

donde

y = variable dependiente o de respuesta.

$x_1, x_2, \ldots, x_k$ = variables independientes o de predicción.

$b_0, b_1, \ldots, b_k$ = parámetros o coeficientes de regresión.

ε = error aleatorio.

El objetivo general de la ecuación de regresión lineal múltiple es identificar el hiperplano de mejor ajuste, el cual es una línea a través de un espacio (k+1) dimensional (tridimensional en el caso de dos variables independientes, figura 3.1).

Para determinar los valores de las estimaciones de los parámetros $b_0, b_1, \ldots, b_k$ y del error estándar asociado, a partir de un conjunto de datos conocidos, se aplica el método de mínimos cuadrados ordinarios (ver anexo A.3). Sin embargo, estos cálculos son muy complejos e implican por lo general el manejo del álgebra matricial; no obstante, se dispone fácilmente de muchos programas de computadora (*software* de cómputo, como SPSS, STATISTICA, STATA, MINITAB, etc.) para la realización de tales cálculos.

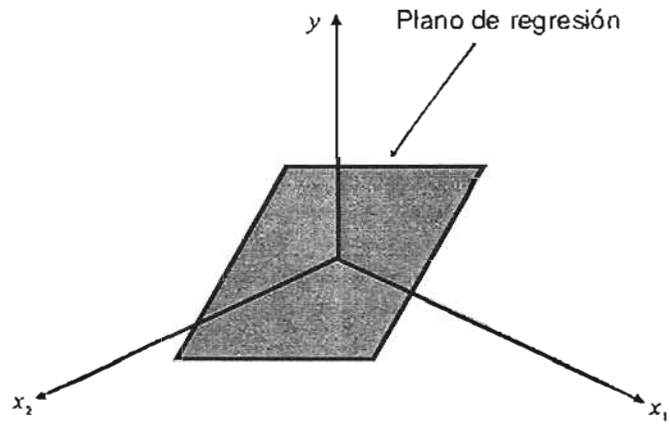


Figura 3.1

Ciertamente, casi nunca se conocen los valores exactos de los parámetros de regresión, sino que sólo se da una aproximación de ellos. A partir de los datos muestrales, se encuentran las estimaciones de los coeficientes y se determina el hiperplano en el espacio $(k + 1)$–dimensional de las variables regresoras $\{x_j\}$ y de y, que mejor ajuste al conjunto de datos, llamado **hiperplano de regresión muestral**. Así, la ecuación de regresión múltiple ajustada quedará

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_k x_k$$

donde

$\hat{y}$ = valor estimado de la variable dependiente.

$x_1, x_2, \ldots, x_k$ = variables de predicción.

$\hat{b}_0, \hat{b}_1, \hat{b}_2, \cdots, \hat{b}_k$ = estimaciones muestrales de los parámetros de regresión o coeficientes de regresión estimados.

Cada coeficiente $\hat{b}_0, \hat{b}_1, \hat{b}_2, \cdots, \hat{b}_k$ mide el cambio promedio en y, debido a un incremento de una unidad de cambio en x_i, manteniendo constantes las otras variables de predicción. Los parámetros $\hat{b}_j$, $j = 0, 1, \ldots, k$, se denominan algunas veces **coeficientes de regresión parciales**, porque ellos describen el efecto parcial de una variable

independiente cuando los otras regresoras, como ya se indicó, se conservan constantes en el modelo.

3.1. SUPUESTOS DEL ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

Siempre que la regresión lineal múltiple se emplea en la práctica, se hacen varios postulados básicos, los cuales son similares a los del caso simple.

En cuanto a la estimación puntual, los principales supuestos son:

- La variable dependiente y es una variable aleatoria.

- La relación entre las diversas variables regresoras y la dependiente es lineal, al menos en forma aproximada. (Técnicamente este supuesto se refiere a la linealidad de los coeficientes).

- Las variables independientes no deben de estar correlacionadas (**multicolinealidad**).

  Si dos variables de predicción en una regresión múltiple tienen una correlación muy fuerte, interfieren entre sí explicando la misma varianza de la variable dependiente. Es indeseable que exista multicolinealidad, porque indica que las variables de predicción no son independientes y, por ende, es difícil distinguir qué cantidad del efecto observado se debe a una variable de predicción individual. Es decir, si dos variables están fuertemente correlacionadas, proporcionan casi la misma información en el pronóstico, por consiguiente los coeficientes de regresión estimados no son confiables, y una de las dos variables debe eliminarse.

Los supuestos para la inferencia estadística (estimación o prueba de hipótesis) son:

- Los residuales están distribuidos normalmente. Si no se cumple este supuesto, las pruebas de significación y los intervalos de confianza, desarrollados a partir de los mismos, pueden estar incorrectos.

- El error aleatorio ε del modelo tiene $E(\varepsilon)=0$ y $Var(\varepsilon)=\sigma^2$. Asimismo, se supone que los $\{\varepsilon\}$ no están correlacionados. Si esta última consideración no se cumple, la situación se denomina **autocorrelación**, la cual ocurre con frecuencia cuando se recopilan datos durante periodos o intervalos de tiempo.

3.2. ERROR ESTÁNDAR MÚLTIPLE DE LA ESTIMACIÓN

Al igual que en la regresión simple, el **error estándar de la estimación** mide la variabilidad o dispersión de los valores observados que se obtienen a partir de la ecuación de regresión. Se puede desarrollar un estimador de σ^2 a partir de la suma de cuadrados de los residuales,[12] obteniéndose la siguiente ecuación, la cual indica cómo se calcula el error estándar de la estimación en regresión lineal múltiple:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-k-1} = MS_{Res}$$

donde y_i representa el i–ésimo valor muestral, $\hat{y}_i$ denota el valor estimado de la variable dependiente, k es número de variables independientes que se están considerando y n es el tamaño de la muestra.

[12] MONTGOMERY, op. cit., p. 75, 535 a 538. En estas páginas, se tiene la demostración de que el estimador insesgado de σ^2 es MS_{Res}.

3.3. COEFICIENTE DE CORRELACIÓN MÚLTIPLE (R)

El **coeficiente de correlación múltiple** mide la fuerza de asociación lineal entre la variable dependiente y dos o más independientes. Puede tomar cualquier valor entre 0 y 1 inclusive y se denota por R. La R múltiple siempre es positiva. La figura 3.2 ilustra cómo puede ser la correlación.

En la regresión múltiple es importante que una vez que se ha identificado cuál es la variable dependiente y las predictoras que han de incluirse en el modelo, se determine si existe una relación entre éstas últimas. Lo anterior se puede hacer mediante el análisis del coeficiente de correlación individual para cada par de variables. Estos datos se presentan en una **matriz de correlación**, en la intersección de cada fila con cada columna correspondiente a cada combinación de dos variables.

La tabla 3.1 de la siguiente página, es un ejemplo de una matriz de correlación de tres variables. En ella, el coeficiente de correlación lineal que indica la relación lineal entre las variables 2 y 3 se representa por r_{23}. El primer subíndice especifica la fila y el segundo la columna. Cabe señalar que la relación entre las variables 2 y 3 es la misma que hay entre la 3 y la 2 (r_{32}). Los elementos en la diagonal principal (r_{11}, r_{22}, r_{33}) siempre serán 1 puesto que una variable siempre tendrá una relación positiva perfecta consigo misma.

Un segundo uso de la matriz de correlación es para verificar si existe multicolinealidad, la cual ocurre cuando las variables independientes están correlacionadas entre sí. Esto distorsiona el error estándar de la estimación y por ende se llegará a conclusiones incorrectas concernientes a qué variables son significativas y cuáles no.

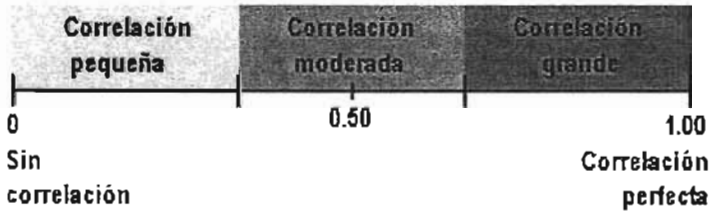


Figura 3.2

Tabla 3.1

Variables	1	2	3
1	r_{11}	r_{12}	r_{13}
2	r_{21}	r_{22}	r_{23}
3	r_{31}	r_{32}	r_{33}

Dado que la matriz de correlaciones es de gran valor para la regresión múltiple, porque muestra los coeficientes simples de correlación entre todas las variables, la mayoría de los programas estadísticos para computadora incluyen el cálculo de ésta.

3.4. PRUEBAS DE HIPÓTESIS

Una vez que se ha obtenido una muestra aleatoria, se han estimado los parámetros de la ecuación de regresión y se ha examinado la matriz de correlación para determinar aquellas combinaciones de variables que son de interés, se analizan los modelos con el mejor potencial.

Como ya se mencionó el objetivo del análisis de regresión múltiple es encontrar la mejor ecuación para predecir y, y a continuación decidir si ésta satisface las necesidades de exactitud del analista.

3.4.1. PRUEBA GLOBAL

Puede probarse la capacidad general de las variables independientes x_1, x_2, ..., x_k, para explicar el comportamiento de la variable y. La prueba utilizada se conoce como **prueba global** o **de la significación de la regresión**. Básicamente, investiga si todas las variables independientes tienen coeficientes netos de regresión iguales a cero. Es decir, se probará que la cantidad de variación explicada, por R^2, ocurre o no, al azar.

La hipótesis nula es

$$H_0: b_1 = b_2 = \cdots = b_k = 0$$

y la hipótesis alternativa es

$$H_1: b_j \neq 0 \text{ para al menos una } j$$

Si la hipótesis nula resulta ser verdadera, significará que todos los coeficientes de regresión son estadísticamente iguales a cero y, por consiguiente, no son de utilidad para pronosticar la variable dependiente. En este caso se tendrían que buscar otras variables independientes (o adoptar un enfoque distinto) para pronosticar y.

El rechazo de H_0: $b_j = 0 \quad \forall j = 1, \cdots, k$ implica que al menos una de las variables independientes $x_1, x_2, \ldots, x_k$ contribuye significativamente al modelo.

Para probar esta hipótesis nula, se aplica la prueba F.

El valor de F_0 se calcula de la siguiente manera:

$$F_0 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 \Big/ k}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \Big/ (n-k-1)}$$

Se rechazará H_0 si $F_0 > F_{\alpha, k, n-k-1}$.

En el diagrama 3.3 se muestra la región de aceptación y de rechazo, y el procedimiento de prueba[13] se resume en una **tabla de análisis de varianza**, como la del cuadro 3.2 que se aprecia en la siguiente página.

[13] MONTGOMERY, op. cit., p. 79.

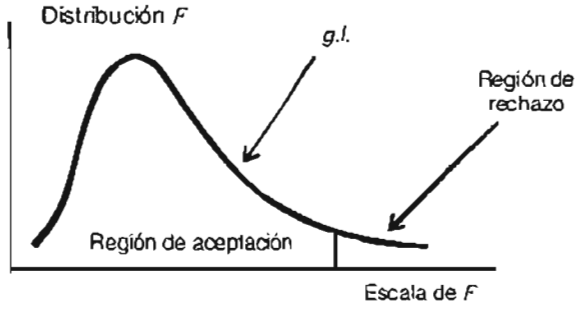


Figura 3.3

Tabla 3.2

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio
Regresión	$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2-\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$	k	$\dfrac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2-\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{k}$
Error	$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$	$n-k-1$	$\dfrac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{n-k-1}$
Total	$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$	$n-1$	

3.4.2. PRUEBA *t* DE REGRESIÓN

Con la prueba anterior se puede considerar el hecho de que todos los coeficientes de regresión son iguales a cero o que no necesariamente todos, son iguales a cero. El siguiente paso para determinar si el modelo es bueno o no, consiste en examinar las variables individualmente para determinar cual de ellas es importante.

El valor de t en el análisis de regresión prueba la hipótesis de que los parámetros de regresión poblacionales son iguales a 0. Si hay coeficientes para los cuales H_0 no puede rechazarse, se considerará su eliminación de la ecuación de regresión. Las k hipótesis para probar la significación de cualquier coeficiente de regresión individual, dígase b_j, son:

$$H_0: b_j = 0 \qquad j = 1, 2, \ldots, k$$
$$H_1: b_j \neq 0$$

La hipótesis nula establece que para la población, independientemente de los resultados muestrales, cuando x_j aumenta en uno, y no queda afectado por este aumento y adquiere un valor aleatorio. Es decir, la contribución de x_j a la habilidad predictiva de la ecuación de regresión es nula. Si H_0: $b_j = 0$ no se rechaza, entonces esto significa que x_j puede ser eliminada del modelo.

La estadística de prueba para esta hipótesis es

$$t_0 = \frac{\hat{b}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

donde

$\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-k-1}$ es la estimación de la varianza del error.

C_{jj} es el j–ésimo elemento de la diagonal principal de $(X'X)^{-1}$ correspondiente a $\hat{b}_j$, con

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}$$

La hipótesis nula H_0: $b_j = 0$ se rechaza si $|t_0| > t_{\alpha/2,\, n-k-1}$. Nótese que ésta es en realidad una prueba parcial o marginal, debido a que el coeficiente de regresión $\hat{b}_j$ depende de todas

las demás variables regresoras x_i $(i \neq j)$ que están en el modelo.

Cabe indicar que por lo general un valor grande de t_0, conducirá al rechazo de la hipótesis nula, y un valor cercano a 0 no lo hará. Por consiguiente, antes de usar la ecuación de regresión, se debe de asegurar que todos los valores de t_0 sean mayores que los valores críticos de la tabla t, para los grados de libertad apropiados al nivel de significancia deseado.

Los valores de t_0 calculados son de particular importancia en la regresión múltiple, porque constituyen la forma principal de detectar multicolinealidad. Si un analista no está seguro de incluir dos variables en la misma regresión debido a su alta interrelación, se verifican los valores de t_0. Si son suficientemente grandes, la correlación entre las dos variables predictoras no es un problema. Si uno o ambos valores t_0 son menores que los valores de la tabla para t, la multicolinealidad está presente, lo que produce coeficientes de regresión poco confiables.

3.5. INTERVALOS DE CONFIANZA

Los intervalos de confianza de los coeficientes de regresión individuales y de la respuesta media, para valores específicos de los regresores, juega un papel muy importante al igual que en la regresión simple.

3.5.1. INTERVALOS DE CONFIANZA DE LOS COEFICIENTES DE REGRESIÓN

A menudo es preciso construir estimaciones del intervalo de confianza para los coeficientes de regresión $\{b_j\}$, para obtenerlos se necesita suponer que los errores $\{\varepsilon_i\}$ se distribuyen normal e independientemente con media cero y varianza σ^2, como ya se mencionó anteriormente. En consecuencia, las respuestas y_i están distribuidas en forma normal e independiente, con media $b_0 + \sum_{j=1}^{n}(b_j x_{ij})$, y varianza σ^2. Dado que el estimador $\hat{b}$ por

mínimos cuadrados es una combinación lineal de las observaciones, igualmente está distribuido normalmente, con vector medio b y matriz de covarianza $\sigma^2 (X'X)^{-1}$. Así, cada una de las estadísticas

$$\frac{\hat{b}_j - b_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad j = 0, 1, 2, \ldots, k$$

se distribuye como t con $n - k - 1$ grados de libertad, donde C_{jj} es el j–ésimo elemento de la diagonal principal de $(X'X)^{-1}$ y $\hat{\sigma}^2$ es la estimación de la varianza del error.

De esta manera, un intervalo de confianza del $100(1 - \alpha)\%$ para el coeficiente de regresión b_j, con $j = 0, 1, \ldots, k$, está dado por

$$\hat{b}_j - t_{\alpha/2,\, n-k-1}\sqrt{\hat{\sigma}^2 C_{jj}} \le b_j \le \hat{b}_j + t_{\alpha/2,\, n-k-1}\sqrt{\hat{\sigma}^2 C_{jj}}$$

3.5.2. INTERVALO DE CONFIANZA DE LA RESPUESTA MEDIA

Una vez determinado el modelo de regresión múltiple, ésta se puede utilizar para hacer un pronóstico de y, esto se logra sustituyendo un conjunto dado de valores $\{x_i\}$ en la ecuación de regresión. Este valor se conoce como **estimación puntual de y**, el cual no necesariamente proporciona conocimiento sobre el grado de exactitud del pronóstico.

Sin embargo, se puede también construir un intervalo de confianza respecto a la respuesta media en un punto en particular, por ejemplo $x_{01}, x_{02}, \ldots, x_{0k}$, el cual puede resultar más útil (en algunos casos) que una simple estimación puntual. Para efectuar lo anterior, primero se define el vector

$$x_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

Así la respuesta media estimada en ese punto está dado por $\hat{y}_0 = x_0' \hat{b}$.

Entonces, un intervalo de confianza del 100(1 – α)% respecto a la respuesta media en el punto $x_{01}, x_{02}, \ldots, x_{0k}$ será determinado por

$$\hat{y}_0 - t_{\alpha/2,\, n-k-1}\sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,\, n-k-1}\sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0}$$

3.6. INTERVALOS DE PREDICCIÓN DE NUEVAS OBSERVACIONES

El modelo de regresión como ya se indicó anteriormente, puede utilizarse para predecir observaciones futuras respecto a y, que corresponde a valores particulares de las variables independientes, por ejemplo, $x_{01}, x_{02}, \ldots, x_{0k}$. Si $x'_0 = \{1, x_{01}, x_{02}, \ldots, x_{0k}\}$, entonces una estimación puntual de la observación futura y_0 en el punto $x_{01}, x_{02}, \ldots, x_{0k}$ es

$$\hat{y}_0 = x_0' \hat{b}$$

Un intervalo de predicción del 100(1 – α)% para esta observación futura es proporcionada por:

$$\hat{y}_0 - t_{\alpha/2,\, n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,\, n-k-1}\sqrt{\hat{\sigma}^2\left(1 + x_0'(X'X)^{-1}x_0\right)}$$

La expresión anterior es una generalización del intervalo de predicción para una observación futura en regresión lineal simple (capítulo dos).

Al predecir nuevas observaciones y estimar la respuesta media en un punto dado $x_{01}, x_{02}, \ldots, x_{0k}$, debe tenerse cuidado en cuanto a extrapolar más allá de la región que contienen las observaciones originales. Es muy posible que un modelo que ajusta bien en la región de los datos originales no sea adecuado fuera de ella.

En la regresión múltiple es fácil extrapolar o hacer estimaciones fuera de la zona que determinan los datos, el problema se debe principalmente a que no es fácil imaginarse el área que establecen valores observados, dado que se está tratando en un espacio (k+1)–dimensional y el conocimiento humano a lo más que llega es a imaginarse $\mathbb{R}^4$, y mayores dimensiones son incomprensibles.

Obsérvese que en los modelos de regresión lineal múltiple, las fórmulas para calcular los intervalos de confianza y de predicción, son bastantes complejas, pero éstos se pueden obtener mediante paquetes de computación para regresión.

3.7. COMPROBACIÓN DE LA ADECUACIÓN DEL MODELO

La validación del modelo es una parte importante del proceso de construcción de la ecuación de regresión múltiple lineal. Es posible utilizar diversas técnicas para medir la adecuación de un modelo de regresión, como el estudio del coeficiente de determinación o el análisis de los residuales.

Por otro lado, al estudiar la regresión múltiple lineal, se tienen varios supuestos como la que los errores son variables aleatorias independientes[14], sin embargo, siempre se debe de tener en cuenta que la validez de éstos es dudosa, por lo que es conveniente hacer un análisis para diagnosticar violaciones de estas premisas, y por consiguiente tener la confirmación o no del modelo que se ha desarrollado en forma tentativa.

Por lo general, no se pueden detectar desviaciones respecto a las proposiciones básicas examinando solamente los estadísticos estándar de resumen, como lo son t, F o R^2, porque éstos por sí mismos no aseguran la adecuación del modelo.

[14] Estos supuestos se encuentran escritos en la página 65 del presente trabajo.

A continuación se presentan varios métodos de utilidad para la validación del modelo y para diagnosticar violaciones a los supuestos básicos de regresión múltiple lineal, los cuales se basan principalmente en el estudio de los residuales.

3.7.1. COEFICIENTE DE DETERMINACIÓN MÚLTIPLE (R^2)

Un estadístico que se consulta con frecuencia en el análisis de regresión lineal múltiple es el **coeficiente de determinación múltiple**, representado por el símbolo R^2. Es útil para medir el porcentaje de la variación total en la variable dependiente que se puede explicar por medio del conjunto de variables independientes $x_1, x_2, \ldots, x_k$, a través del modelo.

La ecuación para calcular R^2 en una regresión múltiple es:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Al igual que en la regresión lineal simple, se tiene que $0 \leq R^2 \leq 1$. Si un plano de regresión se ajusta perfectamente al conjunto de puntos muestrales, la suma de cuadrados del error es igual a 0 y R^2 es igual a 1, lo que significa que la ecuación es muy exacta porque explica una gran porción de la variabilidad de y.

Sin embargo, al añadir una variable al modelo siempre aumentará R^2, independientemente de si la variable adicional es o no estadísticamente significativa, por lo cual se debe extremar precauciones, ya que esto no necesariamente implica que el modelo sea bueno. En este caso conviene también revisar la prueba de significancia de la regresión y de los coeficientes individuales, puesto que con frecuencia se comete el error de sólo ver R^2 al evaluar una ecuación de regresión múltiple.

Nótese que la raíz cuadrada positiva de R^2 es el **coeficiente de correlación múltiple** entre la respuesta y los regresores (apartado 3.3).

3.7.2. ANÁLISIS RESIDUAL

Los residuales de la ecuación de regresión múltiple estimado, definidos por $e_i = y_i - \hat{y}_i$, desempeñan un importante papel al juzgar la suficiencia del modelo, del mismo modo que lo hacen cuando se tiene una sola variable predictiva.

El análisis de los residuales es un modo muy efectivo de descubrir diversos tipos de inadecuación del modelo de regresión, además de ser una forma eficaz de investigar si el modelo se ajusta correctamente a los datos o no. Para comprobar los supuestos básicos de la regresión es necesario dibujar las gráficas de los residuales.

Dado que las violaciones a las premisas del modelo están con más probabilidad en los puntos extremos,[15] éstos pueden ser difíciles de detectar por inspección de los residuales ordinarios (e_i), es por esto, que se recomienda examinar también los escalonados como son: los estandarizados (d_i), los estudentizados (r_i), PRESS ($e_{(i)}$), R de Student, etcétera. En las siguientes gráficas, pueden verse ejemplos de valores atípicos.

Para los fines del presente trabajo, solo se mencionarán los residuales estandarizados y los estudentizados, los cuales aportan con frecuencia información equivalente; sin embargo, básicamente por su estructura, se recomienda examinar los r_i.

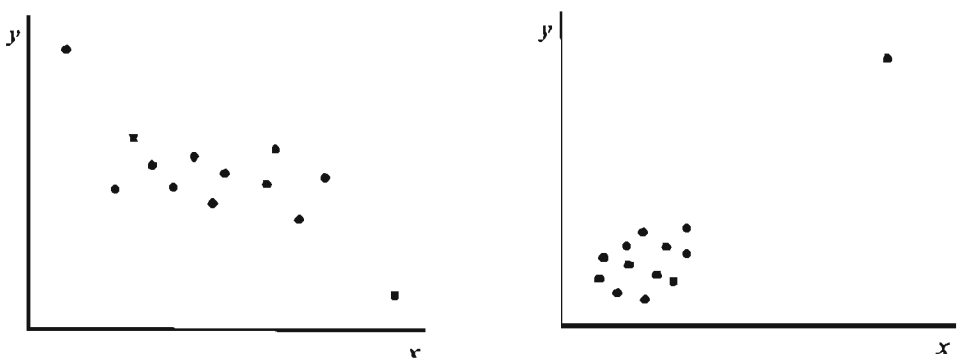


Figura 3.4

[15] Las **observaciones atípicas** o **valores extremos**, son aquellos datos que por alguna característica están separados del resto.

Los **residuales estandarizados** se obtienen mediante la fórmula:

$$d_i = \frac{e_i}{\sqrt{\frac{\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2}{n-k-1}}} \qquad i = 1, 2, \ldots, n$$

Éstos tienen media cero y varianza aproximadamente a la unidad, en consecuencia, un residual estandarizado grande (por ejemplo $d_i > 3$) indicará que se tratará de un valor atípico potencial.

Para examinar los **residuales estudentizados** se utiliza la siguiente expresión:

$$r_i = \frac{e_i}{\sqrt{\frac{\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2}{n-k-1}\left(1-h_{ii}\right)}} \qquad i = 1, 2, \ldots, n$$

siendo h_{ii} el i–ésimo elemento de la diagonal de la matriz $\mathbf{H} = \mathbf{X(X'X)^{-1}X'}$. Cuando la forma del modelo es correcta los r_i tienen varianza igual a 1.

3.7.3. GRÁFICAS DE RESIDUALES

Como se indicó al inicio de este apartado, el análisis gráfico de los residuales es otra forma efectiva de investigar la adecuación del ajuste de un modelo de regresión, y para comprobar los supuestos básicos. Hay varias gráficas que son a menudo útiles, por ejemplo en secuencia de tiempo (si se conoce), contra $\hat{y}_i$, y contra una variable independiente x_i.

3.7.3.1. Gráficas de probabilidad normal.

Las pequeñas desviaciones respecto a las hipótesis de normalidad no afectan mucho al modelo, pero una grande es potencialmente más seria, porque la prueba de la significación de la regresión, los intervalos de confianza y de predicción dependen de ella. Un método

muy sencillo de comprobar esta suposición es trazando una gráfica de probabilidad normal de los residuales.

Sean $e_{[1]} < e_{[2]} < \cdots < e_{[n]}$ los residuales ordenados en orden creciente. Si se dibujan en una gráfica $e_{[i]}$ en función de la probabilidad acumulada $P_i = \dfrac{i - \frac{1}{2}}{n}$, i = 1, 2, ... , n, los puntos que resulten deberían estar aproximadamente sobre una línea recta. Las diferencias apreciables respecto a ella indican que la distribución no es normal. La figura 3.5a indica una gráfica de probabilidad normal "idealizada", porque los puntos caen aproximadamente en una recta. Las partes b y c muestran otros problemas característicos.

El apartado b muestra una curva que va bruscamente hacia arriba y abajo en los dos extremos, lo que indica que las colas de esta distribución son demasiado gruesas para poder considerarla como normal, por consiguiente, el modelo de regresión por mínimos cuadrados será sensible a un subconjunto menor de datos, porque existen valores atípicos que "jalan" demasiado en su dirección el ajuste proporcionado con el método antes mencionado.

Por otro lado, la parte c muestra un patrón asociado con asimetría positiva.

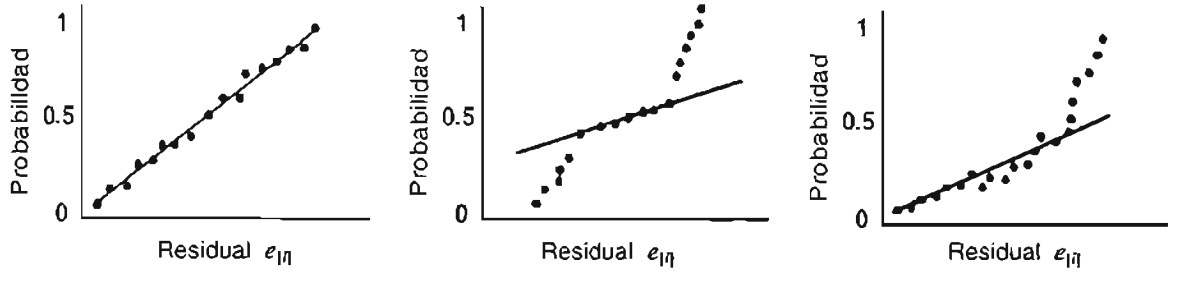


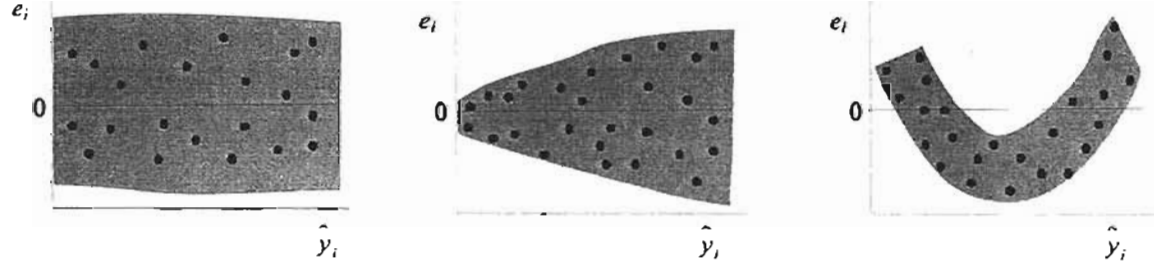
Figura 3.5

3.7.3.2. Gráficas de residuales en función de los valores ajustados $\hat{y}_i$.

También es útil una gráfica de los residuales e_i (o los escalonados d_i o r_i) en función de los valores ajustados correspondientes $\hat{y}_i$, para detectar algunos tipos frecuentes de inadecuaciones del modelo, como el mostrar que uno o más de estos puntos son anormalmente grandes, lo que significa que, son naturalmente valores atípicos potenciales.

Si el diagrama se parece a la de la figura 3.6a, que indica que los e_i se pueden encerrar en una banda horizontal, se tendrá que no hay defectos obvios del modelo; pero si se asemejan a cualquiera de los patrones de las partes b y c, entonces habrá síntomas de deficiencias en el ajuste de la ecuación de regresión.

La representación de **embudo abierto hacia afuera** en el apartado b implica que la varianza es función creciente de la variable dependiente. También es posible tener un embudo abierto hacia dentro, la que indica que Var(ε) aumenta a medida que y disminuye. La gráfica en curva, como en la parte c, muestra **no linealidad**. Lo anterior, señala que se necesitan otras variables regresoras en el modelo, o bien, podría ser necesario un término al cuadrado. Las transformaciones de la variable regresora y/o la de respuesta también podrían ayudar en estos casos.



a) b) c)

Figura 3.6

3.7.3.3. Gráfica de residuales en función de un regresor

También puede ayudar el dibujar la gráfica de los residuales en función de los valores correspondientes de cada variable predictiva. Éstas suelen presentar patrones como los de la figura 3.6, en donde una vez más es deseable de una banda horizontal contenga a todos los puntos. Los esquemas en embudo indican varianza no constante, y los de banda de curva señalan que no es correcto suponer una relación lineal entre la variable de respuesta y las regresoras.

3.7.3.4. Gráfica de residuales en el tiempo

Si se conoce la secuencia temporal de recolección de los datos, se aconseja representar los residuales en función de su orden en el tiempo. En el caso ideal, esa gráfica se parecerá a la de la figura 3.6a, es decir, una franja horizontal que abarque todos los puntos, en donde éstos varíen más o menos en forma aleatoria. Sin embargo, si se asemeja a los patrones b y c, puede señalar, entre otras cosas, que se deben agregar al modelo términos lineales o cuadráticos.

También este último diagrama puede indicar que los errores en un lapso de tiempo se correlacionan con los de otros periodos (**autocorrelación**), lo que es una violación potencialmente grave a las premisas básicas de la regresión. En la figura 3.7a se ejemplifican datos con autocorrelación positiva, mientras que en la parte b, con negativa.

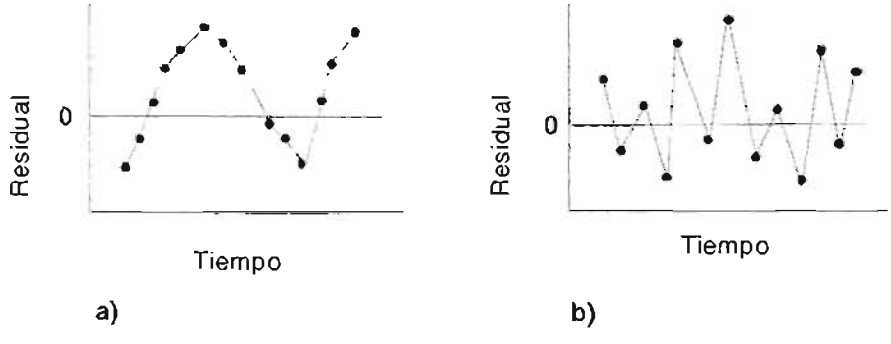


Figura 3.7

3.8. PROBLEMAS EN LA REGRESIÓN MÚLTIPLE

Hay diversos problemas que se encuentran con frecuencia al emplear la regresión múltiple. A continuación se hará un breve análisis de tres problemas: el efecto de puntos influyentes o atípicos, la multicolinealidad en el modelo, y la autocorrelación en los errores.

3.8.1. PUNTOS ATÍPICOS

Un valor atípico es una observación extrema, que no es representativo del resto de los datos. De acuerdo con su ubicación en el espacio x, puede tener efectos de moderados a graves sobre el modelo de regresión.

Existen varios tipos de observaciones atípicas que se presentan en el contexto de modelos de regresión, algunos de ellos son:

Valor atípico de regresión. Es un punto que se desvía de la ecuación de regresión lineal, que se determina con las $n-1$ observaciones restantes.

Valor atípico residual. Es aquel que tiene un residual estandarizado o estudentizado grande (por ejemplo, tres o cuatro desviaciones estándar), cuando se usa en la muestra de n observaciones con que se ajusta un modelo. Nótese que un punto puede ser un valor atípico residual sin que haya fuerte indicación que sea uno de regresión, o viceversa (esto último sucede con frecuencia cuando el punto es muy influyente).

Valor atípico en el espacio x. Es una observación remota en una o más coordenadas x. Un valor atípico en el espacio x también puede ser un atípico de regresión y/o atípico residual.

Valor atípico en el espacio y. Es una observación con coordenada y inusual. El efecto que tiene sobre el modelo de regresión depende de su correspondiente valor en x, y de la

disposición general de los demás datos de la muestra. Un punto de este tipo, puede ser atípico residual y/o posiblemente también de regresión.

Los valores atípicos se deben investigar con cuidado, para ver si se puede encontrar una razón de su comportamiento extraordinario, como puede ser el análisis incorrectos, el registro erróneo de los datos, la falla de un instrumento de medición, entre otros.

Es evidente que el eliminar puntos extremos es conveniente, porque los mínimos cuadrados jalan la ecuación ajustada hacia el valor atípico, sin embargo, puede ser peligroso quitar éstos para "mejorar el ajuste", porque puede dar al usuario una sensación falsa de precisión de la estimación o la predicción; debe tenerse una fuerte evidencia de que el valor atípico es malo. Para comprobar con facilidad el efecto de este tipo de observaciones, se excluyen del modelo y se vuelve a ajustar la ecuación de regresión. Se podrá encontrar que los valores de los coeficientes de regresión, o de los estadísticos de como t, F o R^2, y que el cuadrado medio de residuales pueden ser muy sensibles a los valores atípicos, así si dichos puntos son realmente malos y se debieran omitir, mejoraría la precisión de los estimadores de parámetros y reduciría bastante el ancho de los intervalos de confianza y de predicción, entre otras cosas.

Otro excelente diagnostico para detectar observaciones extremas, es la medida de **distancia de Cook**. La estadística de distancia de Cook es

$$D_i = \frac{f_i}{p}\frac{h_{ii}}{(1-h_{ii})} \qquad i = 1,\ 2, \cdots, n$$

donde $f_i = \dfrac{e_i}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-k-1)}(1-h_{ii})}}$ y h_{ii} es el i–ésimo elemento de la diagonal de la matriz $H = X(X'X)^{-1}X'$. Un valor de $D_i > 1$ indicará que el punto es influyente.

3.8.2. MULTICOLINEALIDAD

Si no hay relación lineal entre las variables independientes, se dice que éstas son **ortogonales**, y por ende, se pueden hacer inferencias como la predicción de valores de respuesta; desafortunadamente en la mayor parte de las aplicaciones de regresión, esto no sucede. Algunas veces no es grave la falta de ortogonalidad, sin embargo, en varios casos los regresores tienen una relación lineal casi perfecta (el problema de **multicolinealidad** está presente), haciendo que las inferencias basadas en el modelo de regresión sean erróneas.

Existen cuatro fuentes de multicolinealidad principales:

1. El método de recolección de datos que se empleó.
2. Restricciones en el modelo o en la población.
3. Especificación del modelo.
4. Un modelo sobredefinido.

Se poseen varias técnicas para detectar la multicolinealidad, las que indicarán el grado del problema y proporcionarán información de utilidad para determinar qué regresores están implicados.

Una medida muy sencilla es la inspección de los elementos r_{ij} no diagonales de la matriz de correlación. Si las variables independientes x_i y x_j son casi linealmente dependientes entonces $|r_{ij}|$ será próximo a la unidad. Sin embargo, lo anterior sirve sólo para detectar la dependencia casi lineal entre pares de regresores, por consiguiente cuando intervienen más de dos variables predictivas en una dependencia, no existe la seguridad de que alguna de las correlaciones r_{ij} sea grande. En general, este método no es suficiente para detectar cosas más complejas que la multicolinealidad por pares.

3.8.3. AUTOCORRELACIÓN

Los modelos desarrollados hasta este momento han supuesto que las componentes de error ε_i del ajuste son variables aleatorias no correlacionadas, sin embargo en muchas aplicaciones del análisis de regresión se incluyen datos para los cuales esta suposición puede resultar inapropiada. Con frecuencia en economía, los negocios, la agricultura, la administración y algunos campos de la ingeniería, muchos problemas de regresión involucran datos de series de tiempo donde la suposición de errores no correlacionados es a menudo insostenible.

Existen varias causas de autocorrelación, quizá la principal, en los problemas donde intervienen datos de series de tiempo es el no incluir uno o más regresores importantes en el modelo.

La ocurrencia de errores autocorrelacionados en forma positiva tiene varias consecuencias potencialmente serias sobre el procedimiento ordinario de regresión por mínimos cuadrados, como son:

1. Los coeficientes de regresión ordinaria por mínimos cuadrados siguen siendo insesagdos, pero ya no son estimados con varianza mínima, y por consiguiente, son estimaciones ineficientes.

2. El error cuadrático medio, puede subestimar mucho a σ^2, en consecuencia, los errores estándar de los coeficientes de regresión pueden ser muy pequeños, asimismo, los intervalos de confianza son más cortos de lo que deberían ser, y las pruebas de hipótesis acerca de los parámetros individuales de regresión pueden indicar que uno o más de los regresores contribuyen en forma significativa al modelo, cuando en realidad no lo hacen.

Existen varios procedimientos estadísticos que pueden emplearse para determinar si los términos del error en el ajuste no se correlacionan, uno que se usa ampliamente es la prueba de **Durbin–Watson**, la cual se basa en la hipótesis de que un **proceso**

autorregresivo de primer orden genera los datos, que se observa a intervalos de tiempo igualmente espaciados, esto es,

$$\varepsilon_t = \rho e_{t-1} + a_t$$

donde ε_t es el término de error en el modelo, en el periodo t; a_t es una variable aleatoria NID(0, σ^2); ρ $(|\rho|<1)$ es un parámetro desconocido, llamado **coeficiente de autocorrelación**. Así, un modelo de regresión lineal simple con errores autorregresivos de primer orden será:

$$y_t = b_0 + b_1 x_t + \varepsilon_t$$
$$\varepsilon_t = \rho \varepsilon_{t-1} + a_t$$

siendo y_t y x_t las observaciones de las variables de respuesta y regresión en el tiempo t, respectivamente.

Como la mayor parte de los problemas de regresión donde intervienen las series de tiempo tienen autocorrelación positiva, las hipótesis que se suelen considerar en la prueba de Durbin – Watson son:

$$H_0: \rho = 0$$

$$H_1: \rho > 0$$

Obsérvese que si H_0: $\rho = 0$ no se rechaza, se estará diciendo que no hay autocorrelación alguna en los errores, y que el modelo de regresión lineal ordinario es apropiado. Para probarla, primero hace el ajuste, después se calcula la estadística de prueba

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} (e_t)^2}$$

donde e_t es el residuo t–ésimo de un análisis de mínimos cuadrados ordinarios aplicado a los datos (y_t, x_t), $t = 1, 2, \ldots, n$.

Desafortunadamente, la distribución de d depende de la matriz **X**, sin embargo, se ha demostrado que d está entre dos cotas d_L y d_U, tales que si d sale de esos límites, se puede llegar a una conclusión acerca de las hipótesis antes dadas. Para un valor adecuado de α, el procedimiento de decisión es el siguiente:

Si $d > d_U$, no se rechaza H_0

Si $d < d_L$, se rechaza H_0 y se concluye que los errores se autocorrelacionan en forma positiva.

Si $d_L \leq d \leq d_U$, la prueba es inconcluyente, lo que implica que deben de colocarse más datos.

Si la autocorrelación aparente se debe a regresores faltantes, y si se pueden identificar e incorporar al ajuste, se podrá eliminar ésta. Si no se puede resolver el problema incluyendo factores omitidos antes, se debe buscar una ecuación de regresión que incorpore en forma específica la estructura de errores correlacionados. Esos modelos suelen requerir técnicas especiales de estimación de parámetros. A continuación se describen tres métodos para manejar la autocorrelación, en términos de tener solo variable predictora, pero su extensión a regresión lineal múltiple es directa.

Método de Cochrane y Orcutt

Considérese el modelo de regresión lineal simple con errores autocorrelacionados que siguen la ecuación

$$y_t = b_0 + b_1 x_t + \varepsilon_t$$
$$\varepsilon_t = \rho \varepsilon_{t-1} + a_t$$

Supóngase que la variable de respuesta se transforma en $y_t' = y_t - \rho y_{t-1}$. Sustituyendo y_t y y_{t-1}, se produce el modelo

$$\begin{aligned} y_t' &= y_t - \rho y_{t-1} = b_0 + b_1 x_t + \varepsilon_t - \rho\left(b_0 + b_1 x_{t-1} + \varepsilon_{t-1}\right) \\ &= b_0\left(1-\rho\right) + b_1\left(x_t - \rho x_{t-1}\right) + \varepsilon_t - \rho\varepsilon_{t-1} \\ &= b_0' + b_1' x_t' + a_t \end{aligned}$$

que satisface las suposiciones acostumbradas de la regresión y en el que se puede aplicar mínimos cuadrados ordinarios. Sin embargo, no se puede usar en forma directa el modelo reparametrizado, dado que y_t' y x_t' son funciones del parámetro desconocido ρ; pero, este coeficiente se puede estimar usando los residuales e_t y e_{t-1} de una regresión por mínimos cuadrados ordinarios, de y_t respecto a x_t, mediante

$$\hat{p} = \frac{\sum_{t=2}^{n}(e_t e_{t-1})}{\sum_{t=2}^{n}(e_{t-1})^2}$$

Así, las variables regresoras y de respuesta transformadas estimadas son:

$$x_t' = x_t - \hat{\rho} x_{t-1}$$
$$y_t' = y_t - \hat{\rho} y_{t-1}$$

y por consiguiente, se aplica mínimos cuadrados ordinarios a los datos transformados.

Por último se debe aplicar la prueba de Durbin–Watson a los residuales del modelo reparametrizado. Si de acuerdo con ésta, no existe autocorrelación positiva, no se necesita más análisis; pero, si todavía hay, será necesaria otra iteración, para la cual se estima de nuevo ρ con los nuevos e_i obtenidos. Se continua este proceso iterativo, hasta que los términos de error en el modelo reparametrizado no esté correlacionado. Si una o dos iteraciones producen errores no correlacionados, se deberá considerar otras técnicas de estimación.

Método de Hildreth–Lu

En este procedimiento el valor de ρ se obtiene minimizando

$$S(b_0, b_1, \rho) = \sum_{t=2}^{n}\left[y_t - \rho y_{t-1} - b_0(1-\rho) - b_1(x_t - \rho x_{t-1})\right]^2$$

es decir, se debe estimar en forma simultánea ρ, b_0 y b_1, lo cual es un problema no lineal de mínimos cuadrados.

Se pueden aplicar procedimientos directos de búsqueda para minimizar $S(b_0, b_1, \rho)$, como por ejemplo, seleccionar valores de ρ, para después aplicar mínimos cuadrados lineales para estimar b_0 y b_1. Este proceso se repite hasta llegar a un valor mínimo de $S(b_0, b_1, \rho)$.

Método de las primeras diferencias

Este procedimiento es más sencillo que los métodos Cochrane–Orcutt y Hildreth–Lu. Supone que ρ =1 en el modelo transformado, así $b_0' = b_0(1-\rho) = 0$, y por consiguiente se tendrá que:

$$y_t' = b_1' x_t' + a_t'$$

donde $y_t' = y_t - y_{t-1}$ y $x_t' = x_t - x_{t-1}$. De esta manera se tendrá que hacer la regresión de y_t' respecto a x_t', pasando por el origen, obteniéndose el coeficiente de regresión $b_1' = b_1$ por el método de mínimos cuadrados ordinarios.

La función de regresión ajustada con las variables transformadas es

$$\hat{y}_t' = b_1' x_t'$$

que puede ser llevarse a su vez, a las variables originales como:

$$\hat{y}_t = b_0 + b_1 x_t$$

donde $b_0 = \bar{y} - b_1' \bar{x}$ y $b_1 = b_1'$.

3.9. CONSTRUCCIÓN DE MODELOS

La **construcción del modelo** es la clave para el éxito o el fracaso de un analista de regresión, si él no representa la verdadera naturaleza de la relación entre la variable dependiente y las variables independientes, los esfuerzos por crearlo, en general, serán improductivos. El proceso debe conducir a la adopción de un patrón que ofrezca buenos pronósticos de valores futuros de y, para observaciones dados de x.

En el análisis de regresión, las dos actividades primordiales en la construcción de un modelo son: determinar la forma del mismo y seleccionar las variables que deben incluirse en él.

Los modelos que son más complejos en apariencia, pueden con frecuencia seguir siendo analizados mediante técnicas de regresión lineal múltiple. Por ejemplo, considérese el modelo polinomial cúbico en una variable independiente:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \varepsilon$$

Si se deja $x_1 = x$, $x_2 = x^2$, y $x_3 = x^3$, entonces la ecuación anterior puede escribirse como

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon$$

que es una ecuación de regresión lineal múltiple con tres variables regresoras. Los modelos que incluyen efectos de interacción también pueden analizarse por medio de métodos de regresión lineal múltiple. Por ejemplo, si

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + \varepsilon$$

y haciéndose $x_3 = x_1x_2$, y $b_{12} = b_3$, la ecuación anterior puede reescribirse como

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon$$

que es un patrón de regresión lineal.

El término *modelo de regresión no lineal* se usa para designar cualquier modelo cuyos

parámetros son no lineales (b^k está presente) y que pueden convertirse linealizarse mediante una transformación. Por ejemplo,

$$y = b_0 e^{b_1 x} \varepsilon$$

es un modelo exponencial con un término aditivo de error, que es no lineal en los parámetros b_0 y b_1. Este modelo puede transformarse a la forma lineal usando una transformación logarítmica:

$$\ln y = \ln b_0 + b_1 x + \ln \varepsilon$$

Su gráfica se puede observar en la figura 3.8.

Otros ejemplos de modelos no lineales se enuncian a continuación, con su respectiva gráfica (ver página 90).

La curva de crecimiento logístico:

$$y = \frac{b_0}{1 + b_1 p^x} + \varepsilon$$

Este es un caso especial de la regresión lineal, en la que la variable dependiente es dicotómica, esto es, puede asumir sólo dos valores (si/no, 0/1). Los valores pronosticados por el modelo de regresión logística se interpretan como *probabilidades de ocurrencia* de un evento.

Figura 3.8

La curva de crecimiento asintótico:

$$y = b_0 - b_1 p^x + \varepsilon$$

3.9.1. PROBLEMAS DE LA CONSTRUCCIÓN DEL MODELO

Otro problema en muchas aplicaciones del análisis de regresión, es la selección del conjunto de variables independientes que se utilizarán en el modelo. En ocasiones la experiencia previa o las consideraciones teóricas básicas pueden ayudar al analista a especificar el conjunto de regresores.

Lo que interesa es separar las variables candidatas para obtener un modelo de regresión que contenga el "mejor" subconjunto éstas, con el menor número de ellas.

No hay un algoritmo que produzca siempre una buena solución al problema de la selección de variables. La mayor parte de los procedimientos de que se dispone actualmente son técnicas de búsqueda. Para desempeñar en forma satisfactoria, requieren interactuar con ellos, así como el juicio del analista.

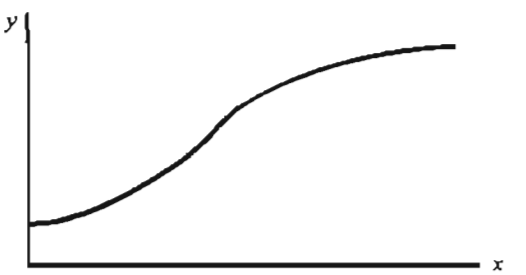


Figura 3.9 Gráfica del crecimiento logístico

Figura 3.10 Gráfica del crecimiento asintótico

3.9.2. VARIABLES CUALITATIVAS EN LA REGRESIÓN

Hasta ahora las variables utilizadas han sido **cuantitativas**; esto es, su naturaleza ha sido numérica. En ocasiones se desea utilizar otras cuyos valores son no numéricos, como en caso de la regresión logística. A tales variables se les denomina **variables cualitativas**.

Por ejemplo, podría ser de interés estimar el sueldo de un ejecutivo con base en los años de experiencia en el trabajo y el que tenga o no un título universitario. Se presupone que un graduado ganará un sueldo mayor que alguien que no lo es. El tener un título universitario puede ser sólo una de dos condiciones: sí o no. De esta forma se considera como variable cualitativa.

El método usual de explicar los diferentes niveles de una variable cualitativa es utilizando variables indicadoras.

En general, una variable cualitativa con m niveles se representa mediante $m - 1$ variables indicadoras, a las cuales se les asignan los valores de 0 ó 1.

La función más importante del análisis de regresión es probar varios modelos posibles, cada uno con un sólido fundamento. Cuando se sigue este enfoque, las pruebas de regresión simplemente sirven para determinar qué modelo es mejor.

3.9.3. PROCEDIMIENTOS POR COMPUTADORA PARA LA SELECCIÓN DE VARIABLES

Selección hacia delante. Este procedimiento se basa en el principio de que los regresores deben agregarse al modelo uno a la vez hasta que no haya variables candidatas restantes que produzcan un aumento significativo en la suma de cuadrados de la regresión.

Eliminación hacia atrás. Este algoritmo se inicia con todas las k variables candidatas en el modelo. Después se elimina la que tiene la estadística F parcial más pequeña o

insignificante. Luego, se estima de nuevo con $k-1$ regresores, y se encuentra el siguiente predictor que es factible de quitar. Este procedimiento termina cuando no pueden excluirse más variables.

Todas las regresiones posibles. Este método requiere que el analista ajuste todas las ecuaciones de regresión que involucren una variable candidata, las que impliquen dos regresores, etc. Luego estas ecuaciones se evalúan de acuerdo con algún criterio adecuado para seleccionar el "mejor" modelo.

3.10. EJEMPLO

Se piensa que la energía eléctrica (y) que consume mensualmente una planta química se relaciona con la temperatura ambiental promedio (x_1), el número de días en el mes (x_2), la pureza promedio del producto (x_3), y las toneladas del producto elaborado (x_4). Se disponen los datos históricos al año pasado que se presentan en la tabla 3.3. Ajuste estos datos a una ecuación de regresión múltiple.[16]

Tabla 3.3

i	y	x_1	x_2	x_3	x_4
1	240	25	24	91	100
2	236	31	21	90	95
3	290	45	24	88	110
4	274	60	25	87	88
5	301	65	25	91	94
6	316	72	26	94	99
7	300	80	25	87	97
8	296	84	25	86	96
9	267	75	24	88	110
10	276	60	25	91	105
11	288	50	25	90	100
12	261	38	23	89	98

[16] HINES, op. cit., p. 639.

Para obtener la ecuación de regresión se hace uso del paquete estadístico SPSS, utilizando el procedimiento de eliminación hacia atrás. En la tabla 3.4 se resumen los cuatro modelos obtenidos, en la que puede apreciarse el coeficiente de correlación y de determinación, el error estándar y la estadística de Durbin–Watson para probar autocorrelación.

Para determinar cuál es el mejor ajuste, dado que el cuadro 3.4 no proporciona la información adecuada para esto, se obtienen las estimaciones de los coeficientes de regresión. Utilizando nuevamente SPSS, se obtienen estos parámetros, así como su error estándar, la estadística t para cada uno de ellos y la significancia mínima para no rechazar la hipótesis de que éste es igual a cero, registrándolos en la tabla 3.5. Para el primer modelo que incluye las cuatro variables, se observa que b_4, y b_3 no son significativamente diferentes de cero, dado que el valor proporcionado en la última columna, así lo indica; por consiguiente las variables x_4 y x_3 se deben excluir de la ecuación de regresión. Algo similar sucede en los ajustes 2, y 3, en el segundo se señala que se debe quitar x_3 y en el tercero x_1.

Además, por la prueba t (sexta y séptima columna de la tabla 3.5), queda claro que la hipótesis H_0: $b_0 = 0$ no debe rechazarse, en consecuencia el término constante debe eliminarse del modelo. En base a lo anterior, lo que procede inmediatamente es calcular la ecuación de regresión lineal múltiple excluyendo b_0, para esto de nuevo se utiliza el método de eliminación hacia atrás, produciéndose tres ajustes (cuadro 3.6).

Tabla 3.4 RESUMEN DEL MODELO

Modelo	Variables incluidas	R	R^2	R^2 ajustada	Error estándar de la estimación	Estadística de Durbin–Watson
1	x_4, x_3, x_2, x_1	0.863	0.745	0.599	15.579	1.772
2	x_3, x_2, x_1	0.863	0.745	0.649	14.574	1.782
3	x_2, x_1	0.855	0.731	0.672	14.095	1.807
4	x_2	0.803	0.645	0.609	15.381	1.646

Tabla 3.5 ESTIMACIÓN DE LOS COEFICIENTES

Modelo	Variables incluidas	Coeficientes Ordinarios B	Error estándar	Coeficientes Estandarizados Beta	*t*	Signifi-cancia
1	Constante	−102.713	207.859		−0.494	0.636
	x_1	0.605	0.369	0.478	1.641	0.145
	x_2	8.924	5.301	0.473	1.684	0.136
	x_3	1.437	2.392	0.133	0.601	0.567
	x_4	0.014	0.734	0.004	0.019	0.986
2	Constante	−101.610	186.306		−0.545	0.600
	x_1	0.606	0.345	0.478	1.756	0.117
	x_2	8.919	4.952	0.472	1.801	0.109
	x_3	1.441	2.228	0.133	0.647	0.536
3	Constante	0.529	95.635		0.006	0.996
	x_1	0.497	0.292	0.392	1.705	0.122
	x_2	10.267	4.345	0.544	2.363	0.042
4	Constante	−90.161	86.741		−1.039	0.323
	x_2	15.161	3.560	0.803	4.259	0.002

Tabla 3.6 RESUMEN DEL MODELO SIN TÉRMINO CONSTANTE

Modelo	Variables incluidas	*R*	R^2	R^2 ajustada	Error estándar de la estimación	Estadística de Durbin–Watson
1	x_4, x_3, x_2, x_1	0.999	0.998	0.997	14.825	1.861
2	x_3, x_2, x_1	0.999	0.998	0.997	13.993	1.786
3	x_2, x_1	0.999	0.998	0.998	13.372	1.807

Como puede apreciarse en el último cuadro, el coeficiente de determinación aumentó considerablemente, pasando de ser de 0.7 a aproximadamente igual 1, lo que significa que este nuevo ajuste es mejor que el anterior. Se procede ahora a calcular las estimaciones de los parámetros de regresión para determinar cuál es el mejor modelo de los tres planteados.

Tabla 3.7 ESTIMACIÓN DE LOS COEFICIENTES SIN b_0

Modelo	Variable incluida	Coeficientes Ordinarios B	Error estándar	Coeficientes Estandarizados Beta	t	Signifi-cancia
1	x_1	0.580	0.348	0.125	1.669	0.134
	x_2	8.585	5.002	0.748	1.716	0.124
	x_3	0.512	1.415	0.163	0.362	0.727
	x_4	–0.090	0.669	–0.032	–0.135	0.896
2	x_1	0.577	0.327	0.124	1.763	0.112
	x_2	8.594	4.721	0.749	1.821	0.102
	x_3	0.411	1.136	0.131	0.362	0.726
3	x_1	0.496	0.230	0.107	2.159	0.056
	x_2	10.291	0.567	0.896	18.163	0.000

De acuerdo con los resultados del cuadro 3.7, el mejor modelo para el conjunto de datos es el tercero, que incluye los regresores temperatura ambiental promedio (x_1) y número de días en el mes (x_2), además de no incluir término constante en éste. Lo anterior se concluye al analizar el nivel de significancia para la estadística t, además de considerar que los coeficientes de regresión, al igual que la estimación, poseen el menor error estándar. Cabe señalar que las variables independientes: pureza promedio del producto (x_3) y las toneladas del producto elaborado (x_4) no contribuyen en forma significativa al ajuste.

Por lo tanto, la ecuación de regresión lineal múltiple es:

$$\hat{y} = \hat{b}_1 x_1 + \hat{b}_2 x_2 = 0.496 x_1 + 10.291 x_2$$

Dado que el coeficiente de determinación, como ya se señalo, indica que el ajuste es muy bueno, porque estadísticamente explica alrededor del 99% de la variación total de la energía consumida mediante temperatura ambiental promedio y número de días en el mes, se procede ahora a determinar la validación del modelo mediante el análisis de los residuales, éstos se presentan en la tabla 3.8.

Tabla 3.8 RESIDUALES

i	Residuales		
	Ordinarios	Estandarizados	Estudentizados
1	−19.390	−1.450	−1.841
2	4.503	0.337	0.369
3	20.680	1.547	1.650
4	−13.058	−0.977	−1.023
5	11.460	0.857	0.902
6	12.694	0.949	1.016
7	3.013	0.225	0.255
8	−2.973	−0.222	−0.260
9	−17.214	−1.287	−1.421
10	−11.058	−0.827	−0.866
11	5.907	0.442	0.469
12	5.446	0.407	0.443

Gráfica de probabilidad normal

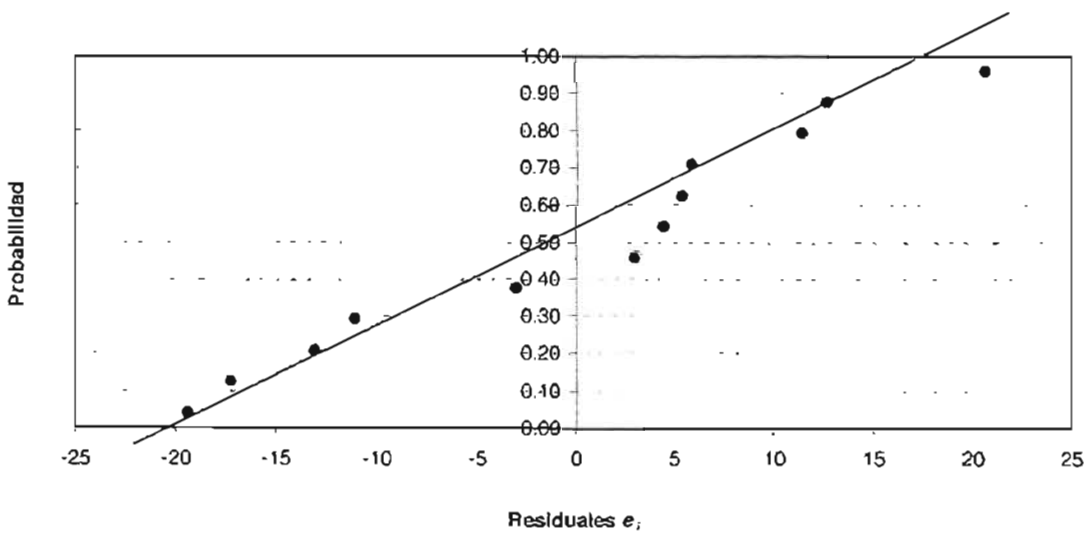


Figura 3.11

La anterior gráfica indica que los puntos caen alrededor de una recta, y la 3.12, muestra que los residuales, a través del tiempo se presentan dentro de una franja horizontal, por consiguiente las hipótesis de normalidad parecen no ser violadas, ni se aprecian valores

atípicos, en los datos.

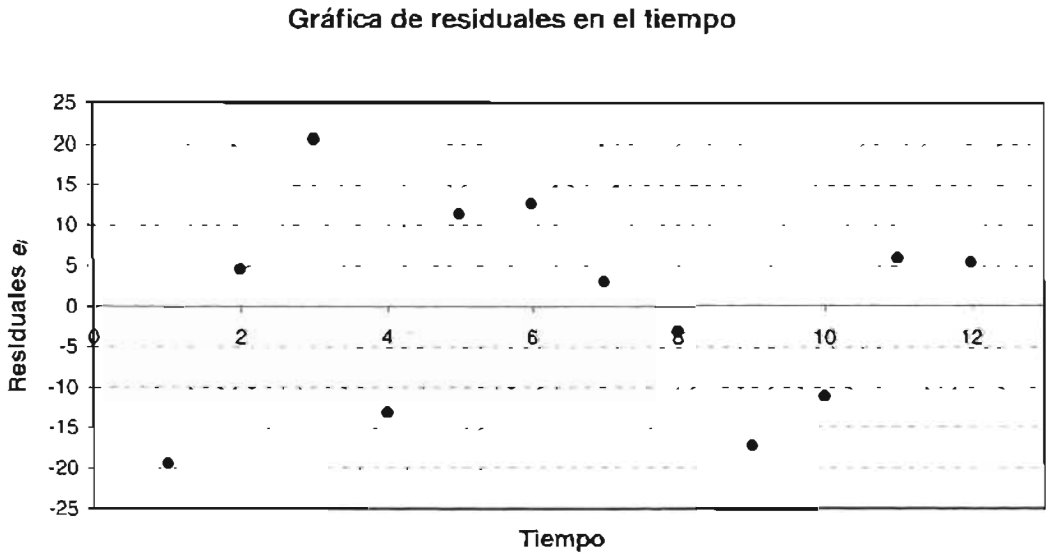


Figura 3.12

A continuación se determinará si los términos del error en el ajuste no se correlacionan, para ello se usa la prueba de Durbin–Watson. Si se opta por $\alpha = 0.01$, los valores críticos[17] correspondientes para $n = 12$ y dos regresores son aproximadamente $d_L = 0.70$ y $d_U = 1.25$ y la estadística, según lo indica la tabla 3.6 es $d = 1.807$. Como $d > d_U$, no se rechaza H_0: $\rho = 0$, por ende los errores no están correlacionados.

Por último, se verificará si existe multicolinealidad en los datos, para ello se calcula la matriz de correlación (tabla 3.9). En ésta se observa que no hay dependencia lineal entre las variables x_1 y x_2 porque su coeficiente de correlación así lo indica.

Por lo tanto, se concluye que la ecuación de regresión múltiple $\hat{y} = 0.496x_1 + 10.291x_2$ es el mejor ajuste para el conjunto de datos.

[17] HINES, op. cit., p. 613. Tabla 15.16 Valores críticos de la estadística de Durbin - Watson.

Tabla 3.9 MATRIZ DE CORRELACIÓN

Variables	x_1	x_2
x_1	1	–0.96
x_2	–0.96	1

Una vez que se ha obtenido y analizado el modelo, se procede a calcular los intervalos de confianza para los coeficientes de regresión y de predición de la variable de respuesta correspondiente a valores particulares de los regresores, por ejemplo, x_{01} = 86 y x_{02} = 27, utilizando nuevamente el paquete estadístico con un nivel de significancia de α = 0.05.

Así, un intervalo de confianza del 95% para los coeficientes de regresión son:

$$b_1 \in (-0.016,\ 1.009)$$
$$b_2 \in (9.028,\ 11.553)$$

Si x'_0 = [86 27], entonces una estimación puntual de la observación futura y_0 en este punto es

$$\hat{y}_0 = x_0' \hat{b} = 320.547$$

Un intervalo de confianza del 95% respecto a la respuesta media en el punto anterior x_0 es

$$y_0 \in (305.736,\ 335.359)$$

Lo anterior significa que energía eléctrica estimada, está entre **305.736** y **335.359**.

Además, un intervalo de predicción para esta misma observación futura es:

$$y_0 \in (287.277,\ 353.818)$$

es decir, el pronóstico de energía eléctrica consumida, con una temperatura ambiental promedio de 86 y 27días en el mes, será de entre **287.277** a **353.818** unidades.

CAPÍTULO 4

CASOS PRÁCTICOS

4.1. LEY DE OKUN

4.1.1. Antecedentes

En economía, la *tasa de desempleo* es el porcentaje de la fuerza laboral que está desempleado:

$$\text{Tasa de desempleo} = \frac{\text{desempleados}}{\text{fuerza laboral}} \times 100$$

donde fuerza laboral está conformada por todas las personas capaces y dispuestas a trabajar, incluye a empleados y quienes están desempleados pero buscan una vacante activamente.

Los mercados de trabajo se dice que están en equilibrio, cuando el número de personas que buscan empleo es igual al de vacantes. Si no se posee esta igualdad, es porque existe una deficiencia o un exceso en el gasto total. En la primera situación, las plazas disponibles son insuficientes, y por ende la demanda total de bienes y servicios disminuye, el empleo baja y el desempleo sube. En la segunda, aparece un faltante de demandantes (sobran puestos), entonces la tasa de desempleo baja.

En 1962 el economista norteamericano Arthur M. Okun (1928 – 1980) planteó un patrón macroeconómico para explicar las variaciones en la tasa de desempleo. Según este modelo, que se conoce hoy en día como la **ley de Okun**, existe una relación lineal entre el cambio en ésta y el crecimiento del producto nacional bruto (PIB) real.

Esta ley está basada en datos de la década de 1950 sobre desempleo y crecimiento económico en los Estados Unidos. En el estudio original, se obtuvo que la producción aumentaba alrededor de un 3% por cada 1% de incremento en el empleo.

Como varias leyes económicas, la de Okun es sólo la observación de una regularidad empírica que no se basó en ningún razonamiento económico fuerte, que sin embargo, ha soportado bien el paso del tiempo y actualmente se usa como medición de cómo los movimientos del producto nacional afectan al nivel de empleo y a la tasa de desempleo.[18]

No obstante las críticas, los trabajos de medición efectuados en distintos países muestran una relación lineal inversa entre el crecimiento del PIB y la tasa de desempleo.[19]

En términos generales, la ley puede escribirse de la siguiente manera:

Variación del desempleo = desviación de la producción del crecimiento normal

o bien

$$u_t - u_{t-1} = -p\left(g_t - c_t\right)$$

donde

u_t = tasa de desempleo al tiempo t

p = coeficiente de Okun, indica cómo se traduce un crecimiento mayor de lo normal en una reducción de la tasa de desempleo.

[18] Enciclopedia Multimedia Virtual de Economía.

[19] En las siguientes páginas se pueden ver los estudios que se han llevado a cabo en países como Argentina, Perú, Costa Rica y Puerto Rico, utilizando esta ley.

- DI PIETRO, Sergio R. Desempleo 2001: una odisea nacional. http://www.vaneduc.edu.ar/uai/comuni/pulso/numero-07/pi07_07.htm
- GARAVITO, Cecilia. La ley de Okun en el Perú: 1970 – 2000. http://www.pucp.edu.pe/economia/pdf/DDD212.pdf
- CUBILLO Arias, Eilyn, Ana Cecilia VALVERDE Kikut y Jorge MADRIGAL Badillo. Estimación de la ley de Okun para Costa Rica. http://www.bccr.fi.cr/udie/Documentos/DIE-03-2002-NTESTIMACION%20LA%20LEY%20DE%20OKUN.pdf
- LEMOIS, Félix A. Estimación de la ley de Okun para Puerto Rico. http://www.jpops01.jp.gobierno.pr:7778/pls/portal/url/ITEM/B423B085DF794407B7C5B7924C836904

g_t = crecimiento porcentual del PIB al tiempo t

c_t = porcentaje de crecimiento del producto que se necesita para mantener el mismo nivel de desempleo al tiempo t.

Por último, hay que señalar que Okun tuvo la precaución de advertir que la ley era válida solamente para tasas de desempleo entre 3 y 7.5%.

4.1.2. Ley de Okun en México.

Un aspecto muy importante en la macroeconomía es la relación positiva entre empleo y PIB, ya que el aumento en los indicadores económicos se traduce en una mejora en el bienestar. Dada la tecnología y las instituciones laborales, un incremento del producto debe traducirse en un crecimiento de las vacantes. Cuando ambos factores se mantienen constantes es posible encontrar una relación estable y positiva entre ellos.

Sin embargo, la pregunta importante es cuánto baja la tasa de desempleo con este crecimiento económico, que es el otro lado de la moneda: la relación negativa entre la tasa de desempleo y la variación del producto, que como ya se indicó anteriormente, se puede medir con la ley de Okun. Si un incremento en el nivel de actividad, se traduce en un aumento del empleo en todos los sectores, debe llevar a una caída en la tasa de desempleo siempre que el efecto del mayor ingreso de fuerza laboral al mercado sea menor al ritmo de creación de empleos.[20]

Un trabajo del Banco Mundial encontró que la relación entre el PIB y el desempleo existe. Además, indica: "*la respuesta del desempleo ante cambios en el producto en Latinoamérica es más baja y volátil que en Estados Unidos de Norteamérica. En el caso de América Latina, por cada punto de caída del crecimiento económico la tasa de desempleo se eleva en ¼ de punto porcentual.*"[21]

[20] GARAVITO, op. cit., p. 3

[21] Ibid, p. 15

Estudiar la dependencia entre el desempleo y el producto interno bruto para México, es importante para la política económica, por lo cual se consideró recomendable utilizar la ley, y calcular el **coeficiente de Okun**. La estimación de éste se hará sobre la base de datos de la variación del PIB y la tasa de desempleo anual, para el periodo de tiempo que abarca de 1987 hasta 2003, ver tabla 4.1 (página 105).

El coeficiente de Okun mide el efecto de la desviación de la tasa de crecimiento del producto con respecto a la normal, la cual es la suma del aumento de la productividad del trabajo y de la fuerza laboral. Este parámetro es importante para el diseño de políticas macroeconómicas y se asume que la tecnología y la legislación laboral tienden a ser relativamente estables en el tiempo, dado que éstos complican el análisis.

Así mismo, se partirá de un modelo donde el desempleo existe, el cual se debe a un desbalance entre la oferta y la demanda global. Desde el punto de vista macroeconómico, esto puede deberse a una falla en el funcionamiento del mercado o a las características estructurales del mismo.

De esta manera, la ecuación lineal a estimar es:

$$y = b_0 + b_1 x + \varepsilon$$

donde x y y son las tasas de crecimiento del PIB real de México y del desempleo abierto, respectivamente.

Antes de encontrar esta ecuación, se dará una descripción del comportamiento de los datos en este tiempo.

Como puede verse en la figura 4.1 (página 106), el desempleo es relativamente estable en el tiempo. Los valores del mismo oscilan entre un mínimo de 1.92% para el cuarto trimestre del 2000 y un máximo de 7.40% en el tercer trimestre de 1995, siendo de 3.38% la tasa promedio para este periodo de tiempo.

Tabla 4.1

Fecha	Tasa de desempleo trimestral[22]	Variación trimestral del PIB[23]	Fecha	Tasa de desempleo trimestral[22]	Variación trimestral del PIB[23]
1987/1	4.40	–1.0	1995/2	6.30	–4.9
1987/2	4.00	–0.4	1995/3	7.40	–5.9
1987/3	4.00	0.6	1995/4	6.10	–6.2
1987/4	3.30	1.7	1996/1	6.00	0.1
1988/1	3.50	2.6	1996/2	5.60	3.2
1988/2	3.70	1.8	1996/3	5.50	4.4
1988/3	4.00	1.3	1996/4	4.70	5.1
1988/4	3.20	1.3	1997/1	4.29	4.6
1989/1	3.20	2.9	1997/2	3.86	6.5
1989/2	3.00	3.8	1997/3	3.67	6.8
1989/3	3.30	4.5	1997/4	3.11	6.8
1989/4	2.50	4.1	1998/1	3.50	7.5
1990/1	2.50	4.3	1998/2	3.20	5.9
1990/2	2.80	4.2	1998/3	3.17	5.7
1990/3	3.10	4.4	1998/4	2.77	4.9
1990/4	2.60	5.2	1999/1	2.91	2.0
1991/1	2.70	3.8	1999/2	2.57	2.7
1991/2	2.30	4.7	1999/3	2.33	3.2
1991/3	2.90	4.3	1999/4	2.20	3.7
1991/4	2.60	4.2	2000/1	2.29	7.4
1992/1	2.90	4.7	2000/2	2.23	7.4
1992/2	2.80	3.5	2000/3	2.36	7.2
1992/3	2.90	3.8	2000/4	1.92	6.6
1992/4	2.70	3.5	2001/1	2.38	2.0
1993/1	3.50	3.0	2001/2	2.38	1.1
1993/2	3.20	1.9	2001/3	2.38	0.3
1993/3	3.70	1.8	2001/4	2.53	–0.1
1993/4	3.30	1.9	2002/1	2.80	–2.4
1994/1	3.70	2.3	2002/2	2.57	–0.2
1994/2	3.60	4.0	2002/3	2.99	0.3
1994/3	3.90	4.2	2002/4	2.46	0.7
1994/4	3.60	4.5	2003/1	2.87	2.5
1995/1	5.10	–0.4	2003/2	2.95	1.3

Fuente: INEGI. Encuesta nacional de empleo urbano y sistema de cuentas nacionales de México.

[22] La tasa general de desempleo abierto trimestral corresponde a 48 áreas urbanas (cobertura anterior).

[23] El producto interno bruto trimestral es la variación promedio anual porcentual a precios de 1993.

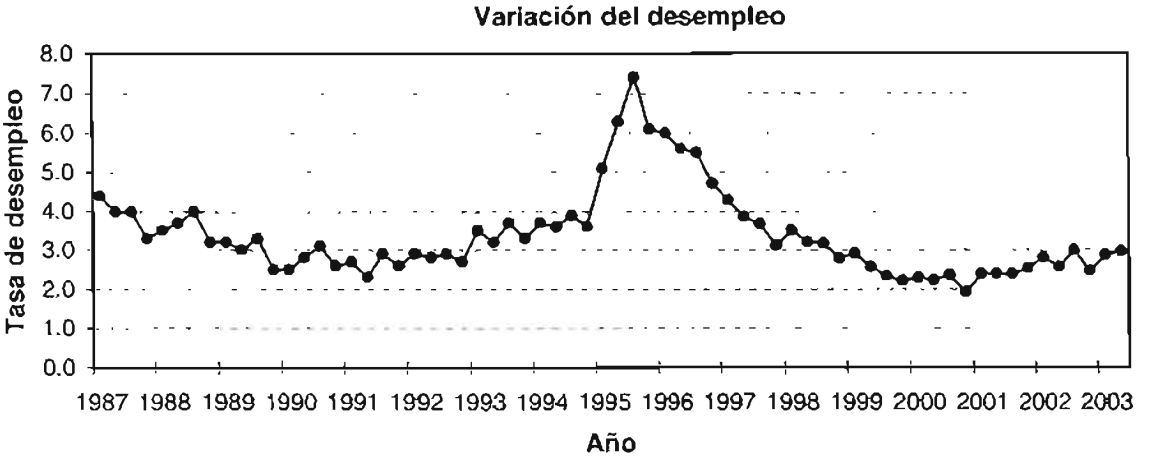


Figura 4.1

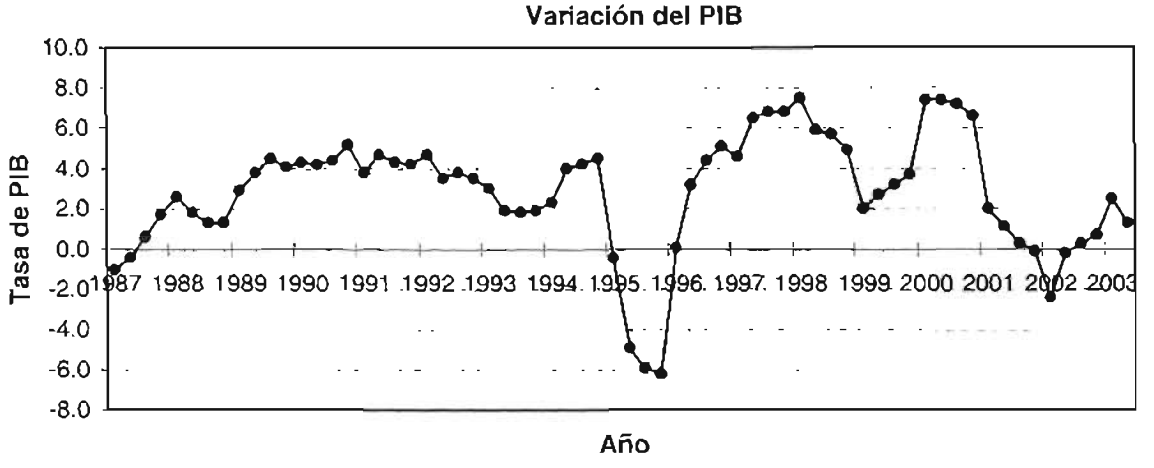


Figura 4.2

Por otro lado, en la figura anterior se puede apreciar que el producto interno bruto ha fluctuado entre –6.20% (cuarto trimestre de 1995,) y 7.50% (primer trimestre de 1998), siendo el promedio de 2.84%.

Así, la ecuación de la línea recta de mejor ajuste es

$$\hat{y} = 0.0394 - 0.1977x$$

con r^2 = 0.2745 y un error estándar igual a 0.0095. Como puede observarse, se encontró una correlación negativa entre el PIB y el desempleo.

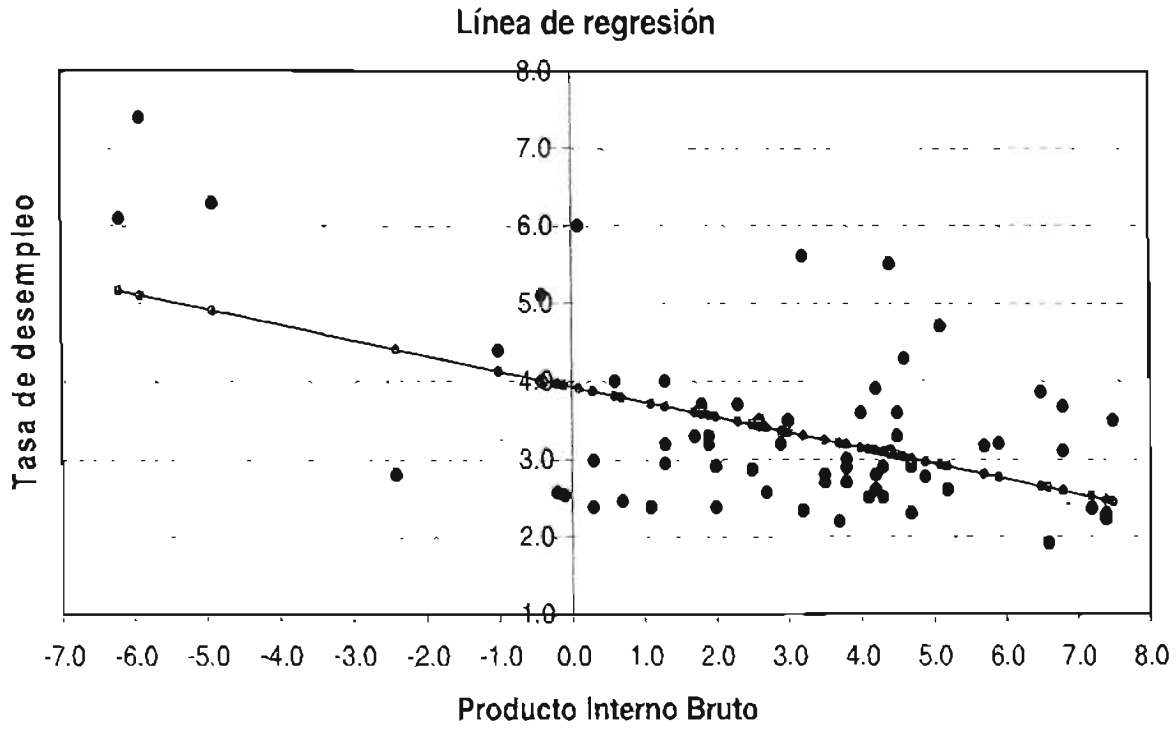


Figura 4.3

En la figura anterior, se ven la línea recta ajustada y los datos originales. Puede apreciarse que los datos no presentan una tendencia lineal, por lo cual el coeficiente de determinación es bastante "pobre"; para corregir lo anterior, lo que se podría hacer es agregar más variables al modelo y determinar si existe un mejor ajuste de los datos. Pero lo que se desea mostrar con este caso práctico es una regresión simple y la ley de Okun, principalmente.

También hay que mencionar que la baja sensibilidad de la tasa de desempleo a los cambios en el PIB se debe principalmente al gran crecimiento del sector informal en el país, así como a la definición y forma de medirlo, dado que éstos han sido modificados en los últimos años.

En el anexo B.1 se presentan los resultados emitidos por el paquete STATISTICA, para encontrar la línea de mejor ajuste, utilizando estos datos.

Si bien las variables tienen una correlación no muy significativa (0.5239), el coeficiente determinado de Okun es significativo, es de 0.1977, el cual es en alguna forma similar a los coeficientes de Chile y Brasil (1981 – 1994), y al de Gran Bretaña (1960 – 1980), como puede verse en la tabla 4.2

Factorizando el lado derecho de la ecuación, se obtiene lo siguiente:

$$\hat{y} = -0.1977(x - 0.1993)$$

donde 0.1993 es la suma de las tasas de crecimiento de la fuerza laboral más la productividad del trabajo. Esto significa que el producto interno bruto deberá crecer al menos a una tasa de 19.93% para que el desempleo no se eleve, lo cual es una utopía dadas las condiciones actuales del país.

Tabla 4.2

	1960 – 80	1981 – 94	1981 – 98
Países desarrollados*			
USA	0.39	0.47	0.42
Canadá			
Gran Bretaña	0.15	0.49	0.51
Alemania	0.20	0.42	0.32
España			0.98
Japón	0.10	0.23	0.20
Países en desarrollo**			
Argentina		0.11	
Bolivia		0.08	
Brasil		0.19	
Chile		0.16	
Colombia		0.37	
Perú		0.10	

Fuente: * CEBRIÁN López. Inmaculada. La ley de Okun. http//:www2.uah.es/icebrian/cap10_1.pdf. Para Alemania el coeficiente del periodo de 1981–98, es realmente de 1981–89.

** GARAVITO, op. cit., cuadro 5, p. 22.

Si se considera que la cifra preliminar del INEGI, del PIB para el primer trimestre de 2004 fue de 3.7%, se tendría que el pronóstico de la tasa de desempleo para ese periodo hubiera sido de 3.21%, que no está muy alejada de la cifra preliminar de 3.92% correspondiente a 32 ciudades (cobertura actual).

En conclusión, el PIB y la tasa de desempleo en México, sí poseen una relación lineal negativa, aunque no poseen una fuerte correlación, dado que al perder su empleo, los trabajadores en su mayoría se retiran del mercado, encaminándose al sector informal. Por otro lado, el coeficiente de Okun, es explicativo y debería tomarse en cuenta si se desea que el desempleo disminuya con el tiempo, o al menos que no se incremente y se mantenga estable.

4.2. INFLACIÓN EN MÉXICO

4.2.1. Teoría cuantitativa del dinero

La teoría cuantitativa del dinero (también conocida como **monetarismo**), es aquella rama de la economía política cuyo objetivo es explicar y definir el comportamiento del dinero y su influencia sobre el funcionamiento del sistema económico. Se basa en los trabajos desarrollados en la Universidad de Chicago, por el premio nobel de economía Milton Friedman, hacia finales de la década de los cincuenta y comienzos de los sesenta del siglo pasado, y puede considerarse como una crítica a la economía keynesiana[24] predominante durante esa época.

Esta teoría hace énfasis en las particularidades de la moneda y en los efectos que tiene la política monetaria sobre la demanda agregada, destacando la fuerte dependencia que existe entre el nivel de precios y el tamaño y la tasa de crecimiento de la masa monetaria.

[24] Básicamente Keynes planteaba que dada la rigidez de los salarios para ajustarse a la baja, los sistemas económicos no tendían al equilibrio con pleno empleo. Proponiendo remediar esa situación con emisión de dinero y con un aumento del gasto público.

Asimismo, su preocupación máxima es la inflación la cual, se dice, es un problema estrictamente monetario.

La inflación se produce, según el monetarismo, porque hay más dinero en circulación (en la calle, en manos de la gente), del que debería haber de acuerdo a las reservas del Banco Central y a la actividad económica en general. Por ejemplo, si la cantidad de moneda circulante por el país, supera sus reservas, entonces ese capital no tiene respaldo y vale menos. Para evitar lo anterior, se propone que la oferta monetaria crezca a un porcentaje fijo, constante e inamovible, y que sea bajo para evitar la inflación, pero que se ajuste al crecimiento del país a largo plazo (ya que no se puede controlar la circulación monetaria día a día de acuerdo a la actividad económica real).

Al mismo tiempo, la teoría monetaria sustenta que se debe eliminar el déficit público y evitar, con una buena legislación, a los monopolios, oligopolios y a los sindicatos porque interfieren en el funcionamiento del mercado de trabajo (que debe ser libre y sin intervención estatal).

4.2.2. Inflación en México

La teoría cuantitativa del dinero, postula a largo plazo una relación estable entre los cambios porcentuales de tres variables macroeconómicas muy importantes: el índice general (es decir, la inflación), la masa monetaria (crecimiento monetario), y el PIB a precios constantes (el crecimiento real). Según esta doctrina, la inflación estará positivamente relacionada con el incremento monetario, e inversamente asociada con la tasa de crecimiento económico real.

La tabla 4.3 exhibe las tasas anuales promedio de inflación, crecimiento monetario y variación real de México durante 69 trimestres (de enero de 1987 a marzo de 2004). La inflación está medida por medio del INPC (Índice Nacional de Precios al Consumidor); el crecimiento monetario se basa en el agregado monetario conocido como M_1 (efectivo fuera de bancos + depósitos en bancos comerciales), y la variación del PIB está calculado en base a precios constantes de 1993.

Tabla 4.3

Trimestre	Inflación*	M_1*	PIB**
1987/01	109.26	72.01	–1.0
1987/02	124.20	92.83	–0.4
1987/03	134.17	114.19	0.6
1987/04	147.94	130.42	1.7
1988/01	177.46	138.44	2.6
1988/02	148.36	134.10	1.8
1988/03	107.98	109.27	1.3
1988/04	67.92	79.10	1.3
1989/01	27.17	45.92	2.9
1989/02	18.44	30.90	3.8
1989/03	16.99	31.83	4.5
1989/04	18.68	34.82	4.1
1990/01	23.48	44.37	4.3
1990/02	25.14	58.57	4.2
1990/03	27.96	53.05	4.4
1990/04	29.59	68.32	5.2
1991/01	26.54	71.53	3.8
1991/02	24.35	64.43	4.7
1991/03	20.99	72.80	4.3
1991/04	19.48	128.73	4.2
1992/01	17.36	119.42	4.7
1992/02	16.26	104.33	3.5
1992/03	15.45	86.15	3.8
1992/04	13.24	20.97	3.5
1993/01	10.89	19.79	3.0
1993/02	9.99	18.44	1.9
1993/03	9.60	22.22	1.8
1993/04	8.62	16.52	1.9
1994/01	7.26	20.20	2.3
1994/02	6.93	13.12	4.0
1994/03	6.75	10.22	4.2
1994/04	6.94	7.61	4.5
1995/01	14.99	– 9.08	–0.4
1995/02	33.75	–12.35	–4.9
1995/03	41.65	– 7.63	–5.9

Trimestre	Inflación*	M_1*	PIB**
1995/04	48.70	0.47	–6.2
1996/01	48.14	22.30	0.1
1996/02	34.19	42.23	3.2
1996/03	30.54	43.49	4.4
1996/04	28.15	41.70	5.1
1997/01	25.51	42.07	4.6
1997/02	21.30	41.38	6.5
1997/03	19.21	42.36	6.8
1997/04	17.24	37.57	6.8
1998/01	15.30	29.88	7.5
1998/02	15.13	25.20	5.9
1998/03	15.61	20.64	5.7
1998/04	17.56	18.18	4.9
1999/01	18.61	18.27	2.0
1999/02	17.88	18.07	2.7
1999/03	16.48	19.48	3.2
1999/04	13.72	23.89	3.7
2000/01	10.55	22.90	7.4
2000/02	9.54	24.37	7.4
2000/03	9.02	22.37	7.2
2000/04	8.91	17.48	6.6
2001/01	7.46	14.78	2.0
2001/02	6.88	10.65	1.1
2001/03	5.98	14.74	0.3
2001/04	5.23	19.77	–0.1
2002/01	4.75	22.71	–2.4
2002/02	4.77	23.00	–0.2
2002/03	5.25	17.50	0.3
2002/04	5.34	12.29	0.7
2003/01	5.44	12.45	2.5
2003/02	4.74	11.78	1.3
2003/03	4.07	10.96	1.1
2003/04	3.97	12.22	1.3
2004/01	4.32	12.68	3.7

Fuente: * Banco de México. Indicadores económicos. Estas tasas de crecimiento estaban en forma mensual porcentual con variación anual, se calculó su promedio trimestral, porque el PIB sólo se maneja en forma trimestral y anual.

** INEGI. Sistema de cuentas nacionales de México. El producto interno bruto trimestral es la variación promedio anual porcentual a precios constantes de 1993.

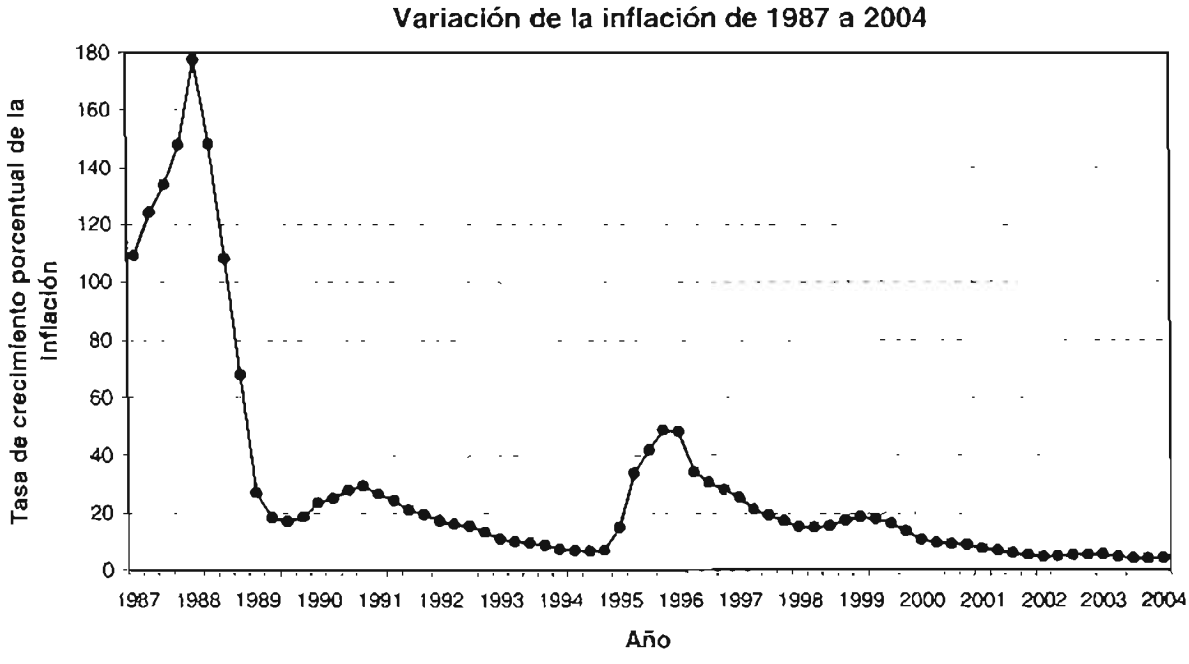


Figura 4.4

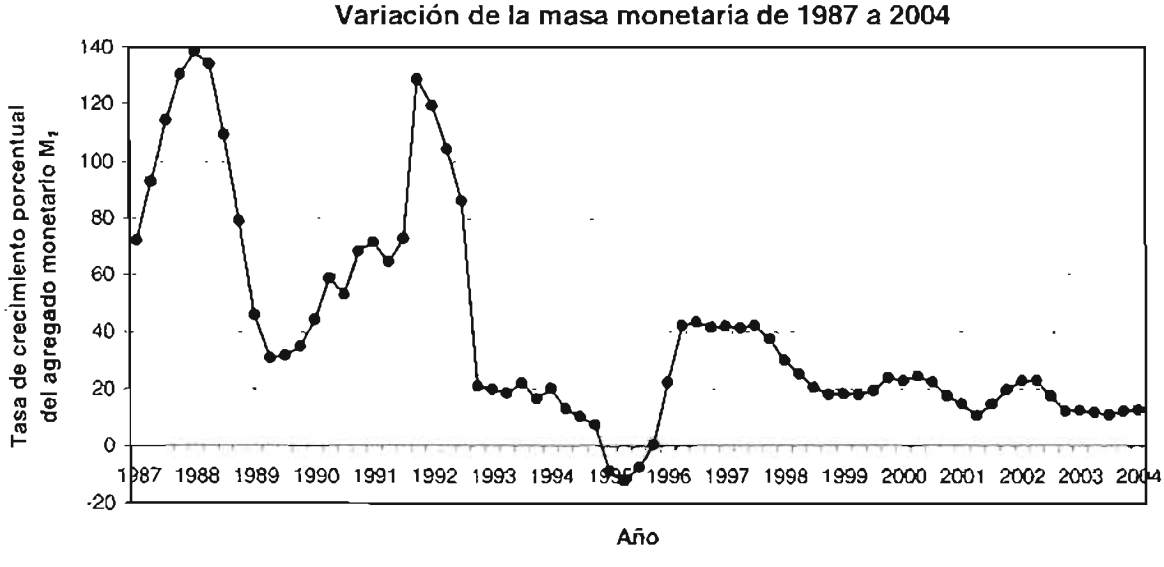


Figura 4.5

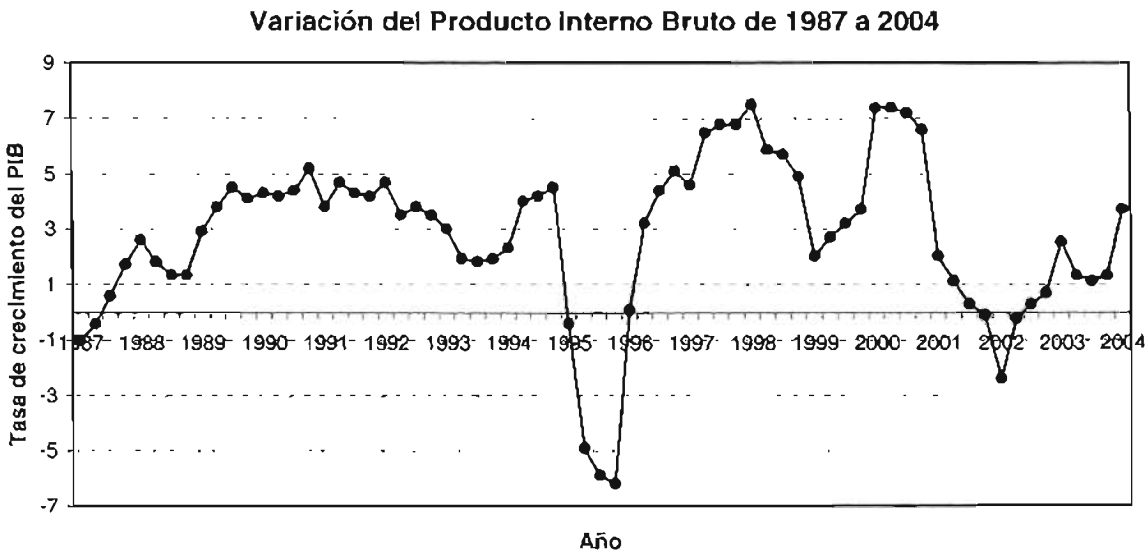


Figura 4.6

En las gráficas anteriores (4.4, 4.5 y 4.6) se puede apreciar la variación que han tenido estas variables a lo largo de tiempo. La mayor inflación se registra en los años ochentas, mientras que a partir del nuevo siglo, se ha notado una estabilidad en el crecimiento de la inflación.

Por otro lado, se observa más cambios en la masa monetaria a lo largo del tiempo, alcanzando su nivel más alto en el primer trimestre de 1988 y el más bajo en el segundo de 1995. De igual manera, la tasa de crecimiento del producto interno bruto es muy variante, la menor se registra en el tercer trimestre de 1995 y la mayor en el primer trimestre de 1998.

Utilizando el paquete estadístico STATISTICA y empleando los datos del último cuadro, se obtiene el coeficiente de Pearson para verificar la relación entre la inflación y la masa monetaria (ver tabla 4.4, página 114), así como con el PIB (tabla 4.5, página 114). Según los resultados obtenidos, se observa que efectivamente, existe una correlación positiva (0.712) entre el alza de precios y el agregado M_1. De igual manera, se comprueba, no obstante que ésta no es significativa, que la correlación entre el aumento de precios y el producto interno bruto es negativa (–0.246).

Ahora, con los mismos datos se estimará la regresión

$$y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$$

donde

y = tasa de inflación

x_1 = tasa de crecimiento monetario

x_2 = tasa de crecimiento del PIB real

ε = error aleatorio.

El coeficiente b_0 indicará la parte de la inflación que no varía directamente con la tasa de crecimiento monetario y del PIB. Se presume que éste sea positivo.

Asimismo, se espera que b_1 sea también mayor a cero, dado que indicará en cuánto se incrementa la inflación por trimestre, por cada punto adicional en el crecimiento de la tasa del agregado monetario M_1.

Tabla 4.4 Correlaciones

		INFLACIÓN	M_1
INFLACIÓN	Coeficiente de correlación de Pearson	1.000	0.712
	Significancia (2 – colas)	.	0.000
	n	69	69
M_1	Coeficiente de correlación de Pearson	0.712	1.000
	Significancia (2 – colas)	0.000	.
	n	69	69

Tabla 4.5 Correlaciones

		INFLACIÓN	PIB
INFLACIÓN	Coeficiente de correlación de Pearson	1.000	–0.246
	Significancia (2 – colas)		0.042
	n	69	69
PIB	Coeficiente de correlación de Pearson	–0.246	1.000
	Significancia (2 – colas)	0.042	
	n	69	69

Por otro lado, b_2 estará midiendo el incremento en la inflación que resulta del aumento del producto interno bruto, el cual se espera que sea negativo.

En el anexo B.2, se encuentran los resultados emitidos por el paquete estadístico STATISTICA, concernientes al cálculo de esta regresión. En éstos se obtiene la siguiente ecuación:

$$\hat{y} = 0.1098 + 0.7974x_1 - 4.8896x_2$$

con un coeficiente de determinación igual a 0.6373.

Se puede apreciar en primer lugar, que los valores estimados tienen el signo que se esperaba. Así, 10.98% de la tasa de crecimiento de la inflación no varía directamente del M_1 y PIB. Por cada aumento unitario en la tasa del M_1, la inflación crece 0.7974 unidades, y disminuye 4.8896 por cada incremento en el producto interno bruto.

Por otro lado, el grado de poder explicativo de esta regresión es bajo, la variación conjunta de estas dos variables justifica cerca del 64% del cambio de la inflación. Como se recordará, el sólo examinar este estadístico no asegura la adecuación o no del modelo y para poder hacer un pronóstico, se tienen que considerar otros criterios, como los estadísticos t y F, así como el análisis de los residuales.

Como se puede observar en los resultados presentados en el anexo B.2, el valor de F es 57.9798 y el de p es igual a cero, equivale a decir que la probabilidad de que el valor de b_0, b_1 y b_2 sean iguales a cero, es nula, al menos uno de ellos es diferente de este valor. Para corroborar esto se utiliza la prueba t.

Los cálculos de la prueba t, indican que los tres coeficientes, analizando cada uno, son significativamente diferentes de 0. Es decir, el aumento o disminución de crecimiento monetario y del producto interno bruto influyen en el alza de los precios, así mismo la constante b_0.

Por último, se hará un análisis de los residuales, puesto que es una forma muy efectiva de investigar la adecuación del ajuste de un modelo de regresión y para comprobar las premisas básicas.

Primero se comprobará la suposición de normalidad, trazando una gráfica de probabilidad normal de los residuales (figura 4.7), para determinar en forma visual si los puntos resultantes están aproximadamente sobre una línea recta o no. Como se observa, existe un aplanamiento en los extremos, lo que indica una muestra con distribución con colas más delgadas que la normal, consecuentemente, existen valores atípicos que "jalan" demasiado en su dirección el ajuste proporcionado con el método de mínimos cuadrados.

De igual manera, es útil una gráfica de los residuales e_i en función de los valores ajustados correspondientes $\hat{y}_i$, para detectar algunas inadecuaciones del modelo (figura 4.8). Dado que la gráfica asemeja un embudo abierto hacia afuera, lo que implica que la varianza de los errores no es constante, sino que es función creciente de y (heteroscedasticidad). Lo anterior, puede señalar que uno o más residuos son anormalmente grandes (valores atípicos).

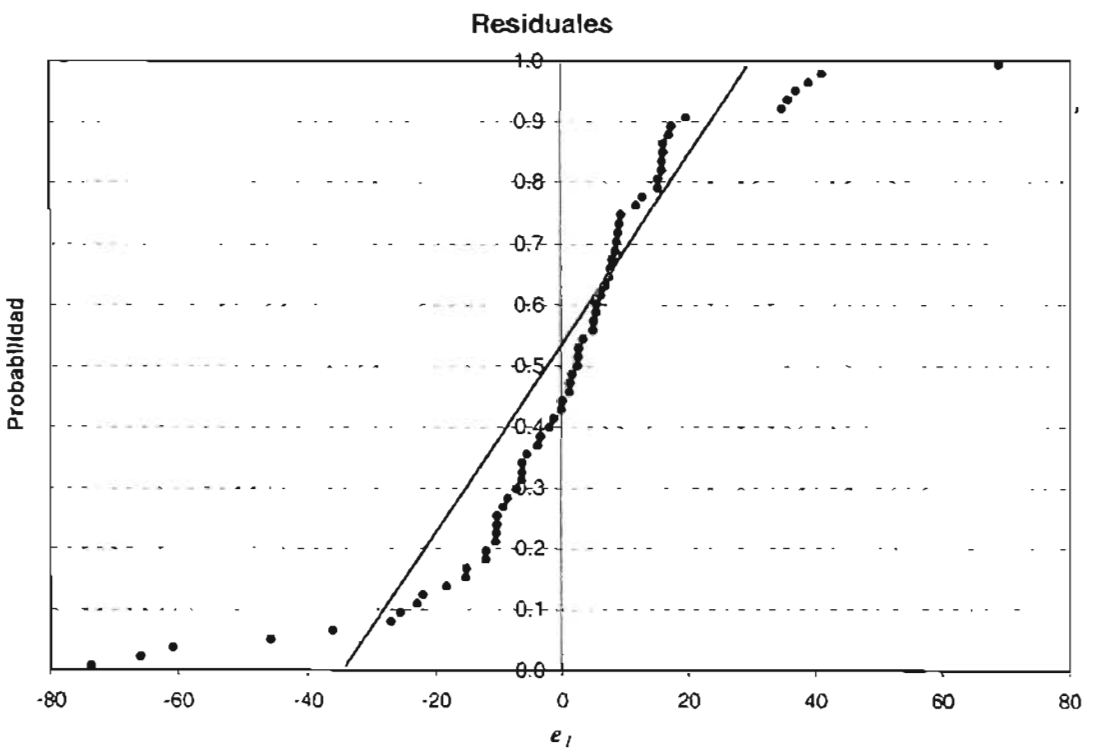


Figura 4.7

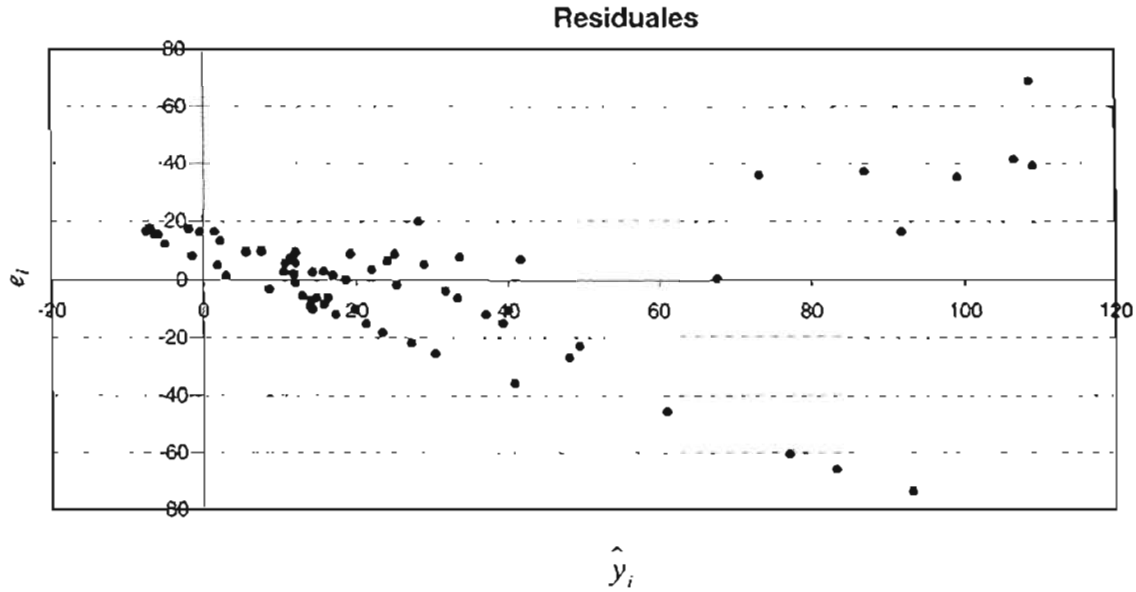


Figura 4.8

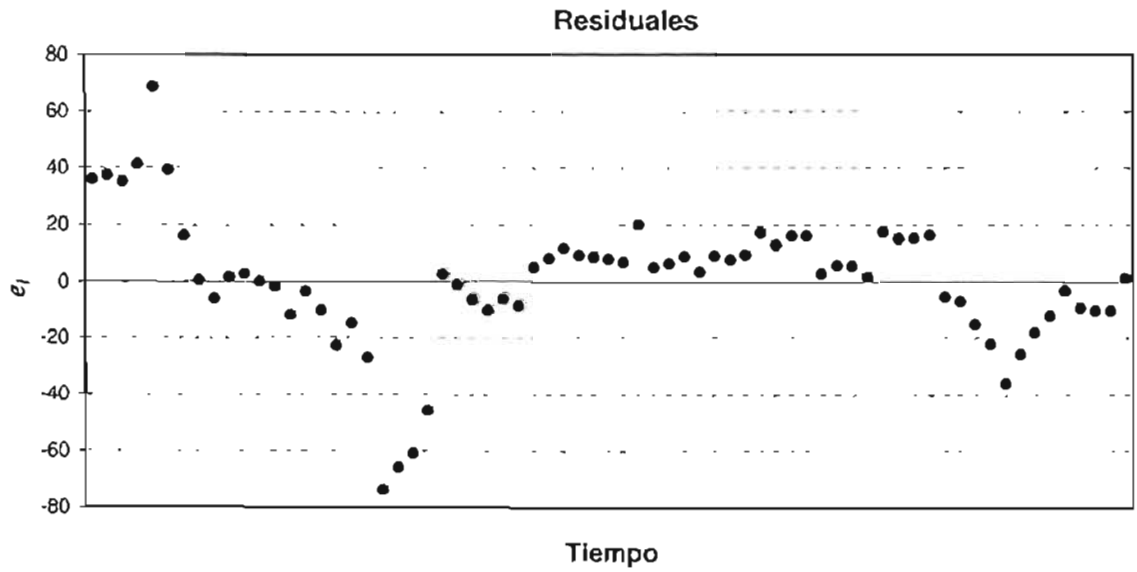


Figura 4.9

Dado que se conoce la secuencia de recolección de los datos, se traza la gráfica de los residuales en función de su orden en el tiempo. En la figura anterior se ve que una banda horizontal (aproximadamente de –40 a 20) abarca a la mayoría de los residuales, pero también se observa que existe una autocorrelación positiva, la cual es una violación grave

a las premisas básicas de la regresión. Existen varios métodos para manejar el problema de la autocorrelación, como por ejemplo incluir un regresor en el modelo, utilización de técnicas especiales de estimación de los parámetros (ver el apartado 3.8.3), hacer uso de mínimos cuadrados ponderados o generalizados si se conociera suficientemente la estructura de autocorrelación.

En base a todo lo anterior, se concluye que el modelo no es el adecuado para describir la inflación, se tiene que agregar otras variables o bien se tendrá que hacer una transformación a la variable de respuesta o a alguno de los regresores. Sin embargo, esto último se descarta, porque de acuerdo a la gráfica de los residuales en función de los valores ajustados (figura 4.8) no existe indicios de no linealidad. Para comprobar lo anterior, se realizan algunas transformaciones a las variables independientes (ver la siguiente tabla). Como se puede distinguir, en algunos modelos el coeficiente de determinación es menor al del modelo original, asimismo se sigue teniendo autocorrelación positiva, dado que el estadístico de Durbin–Watson (d) es menor al valor crítico para $\alpha = 1\%$ y $n = 80$, $d_L = 1.44$.[25]

Tabla 4.6

Modelo	Variables incluidas	R^2	Estadística de Durbin–Watson
1	M_1, PIB, M_1^2	0.6520	0.2910
2	M_1^2, PIB	0.6464	0.2968
3	M_1, M_1^2	0.5877	0.2630
4	M_1, M_1^2, M_1^3	0.5882	0.2599
5	log(M_1), PIB	0.4872	0.3715
6	$\sqrt{M_1}$, PIB	0.6059	0.2908
7	M_1, PIB, M_1^2, PIB^2	0.6581	0.3107

[25] HINES, op. cit., p. 613.

Tabla 4.7 Residuales

No.	Ordinarios	Estandarizados
1	35.9807	1.5261
2	37.2447	1.5797
3	35.0764	1.4878
4	41.2823	1.7510
5	68.8089	2.9186
6	39.2577	1.6651
7	16.2320	0.6885
8	0.2288	0.0097
9	−6.2426	−0.2648
10	1.4077	0.0597
11	2.6360	0.1118
12	−0.0157	−0.0007
13	−1.8509	−0.0785
14	−11.9958	−0.5088
15	−3.7978	−0.1611
16	−10.4441	−0.4430
17	−22.8931	−0.9710
18	−15.0172	−0.6370
19	−27.0031	−1.1453
20	−73.6092	−3.1222
21	−65.8637	−2.7936
22	−60.7945	−2.5786
23	−45.6466	−1.9361
24	2.6542	0.1126
25	−1.1973	−0.0508
26	−6.4006	−0.2715
27	−10.2967	−0.4367
28	−6.2347	−0.2644
29	−8.5768	−0.3638
30	5.0431	0.2139
31	8.1667	0.3464
32	11.8983	0.5047
33	9.2933	0.3942
34	8.6627	0.3674
35	7.9094	0.3355

No.	Ordinarios	Estandarizados
36	7.0328	0.2983
37	19.8733	0.8429
38	5.1866	0.2200
39	6.4003	0.2715
40	8.8573	0.3757
41	3.4843	0.1478
42	9.1134	0.3865
43	7.7081	0.3269
44	9.5623	0.4056
45	17.1638	0.7280
46	12.9076	0.5475
47	16.0451	0.6806
48	16.0430	0.6805
49	2.8389	0.1204
50	5.6962	0.2416
51	5.6172	0.2383
52	1.7809	0.0755
53	17.4922	0.7419
54	15.3130	0.6495
55	15.4111	0.6537
56	16.2688	0.6900
57	−5.5251	−0.2344
58	−7.2146	−0.3060
59	−15.2806	−0.6481
60	−22.0011	−0.9332
61	−36.0705	−1.5299
62	−25.5220	−1.0825
63	−18.2180	−0.7727
64	−12.0082	−0.5093
65	−3.2395	−0.1374
66	−9.2776	−0.3935
67	−10.2674	−0.4355
68	−10.3939	−0.4409
69	1.3199	0.0560

Puesto que hay indicios de que la varianza de los errores no es constante, porque uno o más datos son anormalmente grandes, se procede ahora a abordar este problema para "mejorar" el modelo de regresión lineal múltiple, para llevar a cabo lo anterior, se eliminarán aquellos valores que se consideren atípicos, utilizando para esto los residuales estandarizados. Posteriormente, se calculará la nueva ecuación de regresión, así como los nuevos d_i; este procedimiento se continuará hasta que todos los puntos atípicos sean suprimidos.

La condición que se tomará en cuenta para excluir del modelo aquellos valores que sean atípicos es que su residual estandarizado sea menor o igual a –2, o bien sea mayor o igual a 2 (es decir, $|d_i| \geq 2$). En la tabla 4.7 (página 119) se exhiben los residuales ordinarios y estandarizados, emitidos por el paquete STATISTICA, en ella se aprecia que los datos número 5, 20, 21 y 22 cumplen con el anterior requisito, y por ende se eliminarán.

La ecuación de regresión que se obtiene con los 65 datos, indica que ésta explica aproximadamente el 79% de la variación de la inflación. En el cuadro 4.8, se ilustra lo anterior, además de que se indican los otros ajustes que se producen al ir quitando los puntos considerados como atípicos, hasta tener únicamente 56, en donde el coeficiente de determinación es 0.9197. En la tabla 4.9 (página 121) se visualizan estos datos, así como sus residuales estandarizados y ordinarios, y la gráfica de probabilidad normal (figura 4.10, página 122), muestra que estos últimos están aproximadamente sobre una línea recta, lo que significa que ya no hay valores anormalmente grandes.

Tabla 4.8

Modelo	Número de datos	Estadístico		Error estándar de la estimación
		R^2	Durbin - Watson	
1	69	0.6373	0.2861	23.5764
2	65	0.7867	0.4518	16.4451
3	62	0.8618	0.4030	13.5028
4	60	0.8869	0.4757	12.3836
5	59	0.8966	0.4962	11.9438
6	57	0.9115	0.5042	11.2092
7	56	0.9197	0.5107	10.7822

Tabla 4.9

Fecha	Inflación	M1	PIB	Ordinarios	Estandarizados
1987/01	109.26	72.01	−1.0	19.62	1.82
1987/02	124.20	92.83	−0.4	15.30	1.42
1987/03	134.17	114.19	0.6	7.36	0.68
1987/04	147.94	130.42	1.7	9.15	0.85
1988/02	148.36	134.10	1.8	6.14	0.57
1988/03	107.98	109.27	1.3	−10.26	−0.95
1988/04	67.92	79.10	1.3	−18.27	−1.69
1989/01	27.17	45.92	2.9	−16.13	−1.50
1989/02	18.44	30.90	3.8	−4.61	−0.43
1989/03	16.99	31.83	4.5	−3.70	−0.34
1989/04	18.68	34.82	4.1	−7.10	−0.66
1990/01	23.48	44.37	4.3	−11.49	−1.07
1990/03	27.96	53.05	4.4	−15.75	−1.46
1992/04	13.24	20.97	3.5	−0.70	−0.06
1993/01	10.89	19.79	3.0	−4.18	−0.39
1993/02	9.99	18.44	1.9	−8.90	−0.83
1993/03	9.60	22.22	1.8	−13.79	−1.28
1993/04	8.62	16.52	1.9	−8.23	−0.76
1994/01	7.26	20.20	2.3	−11.59	−1.07
1994/02	6.93	13.12	4.0	3.71	0.34
1994/03	6.75	10.22	4.2	7.58	0.70
1994/04	6.94	7.61	4.5	11.97	1.11
1995/01	14.99	−9.08	−0.4	14.33	1.33
1995/02	33.75	−12.35	−4.9	15.07	1.40
1995/03	41.65	−7.63	−5.9	13.18	1.22
1995/04	48.70	0.47	−6.2	10.20	0.95
1996/01	48.14	22.30	0.1	16.55	1.54
1996/02	34.19	42.23	3.2	−3.76	−0.35
1996/03	30.54	43.49	4.4	−3.02	−0.28
1996/04	28.15	41.70	5.1	−0.16	−0.02
1997/01	25.51	42.07	4.6	−5.58	−0.52
1997/02	21.30	41.38	6.5	0.02	0.00
1997/03	19.21	42.36	6.8	−1.68	−0.16
1997/04	17.24	37.57	6.8	1.45	0.13
1998/01	15.30	29.88	7.5	11.00	1.02
1998/02	15.13	25.20	5.9	8.17	0.76

Fecha	Inflación	M1	PIB	Ordinarios	Estandarizados
1998/03	15.61	20.64	5.7	12.53	1.16
1998/04	17.56	18.18	4.9	13.27	1.23
1999/01	18.61	18.27	2.0	0.37	0.03
1999/02	17.88	18.07	2.7	3.20	0.30
1999/03	16.48	19.48	3.2	2.69	0.25
1999/04	13.72	23.89	3.7	–2.37	–0.22
2000/01	10.55	22.90	7.4	13.19	1.22
2000/02	9.54	24.37	7.4	10.62	0.99
2000/03	9.02	22.37	7.2	11.27	1.05
2000/04	8.91	17.48	6.6	13.49	1.25
2001/01	7.46	14.78	2.0	–7.07	–0.66
2001/02	6.88	10.65	1.1	–7.56	–0.70
2001/03	5.98	14.74	0.3	–16.62	–1.54
2002/03	5.25	17.50	0.3	–20.29	–1.88
2002/04	5.34	12.29	0.7	–12.75	–1.18
2003/01	5.44	12.45	2.5	–4.22	–0.39
2003/02	4.74	11.78	1.3	–9.95	–0.92
2003/03	4.07	10.96	1.1	–10.70	–0.99
2003/04	3.97	12.22	1.3	–11.18	–1.04
2004/01	4.32	12.68	3.7	0.14	0.01

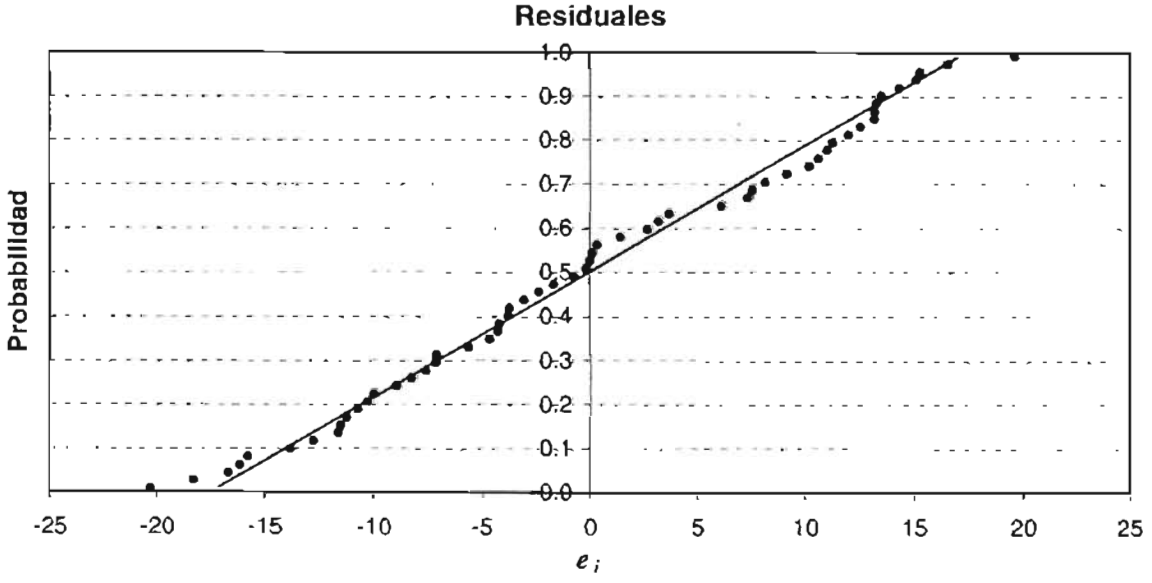


Figura 4.10

Tabla 4.10

Estadístico			Error estándar de la estimación	Suma de cuadrados de residuales
ρ	R^2	Durbin - Watson		
0.6	0.8383	1.2785	7.3565	2814.1477
0.7	0.7950	1.3975	7.0316	2571.0337
0.8	0.7373	1.4722	6.7726	2385.1498
0.9	0.6779	1.5169	6.5952	2261.8381
0.91	0.6730	1.5209	6.5847	2254.6197
0.92	0.6683	1.5249	6.5759	2248.6418
0.93	0.6640	1.5288	6.5692	2243.9960
0.94	0.6601	1.5328	6.5644	2240.7754
0.95	0.6566	1.5368	6.5619	2239.0725
0.955	0.6550	1.5387	6.5616	2238.8187
0.9555	0.6549	1.5389	6.5616	2238.8159
0.9556	**0.6548**	**1.5390**	**6.5616**	**2238.8158**
0.9557	0.6548	1.5390	6.5616	2238.8159
0.96	0.6535	1.5407	6.5618	2238.9782
0.97	0.6509	1.5445	6.5642	2240.5799

Asimismo, cabe señalar que la autocorrelación positiva sigue presente en todos los modelos, porque la estadística de Durbin–Watson es menor a $d_L = 1.44$; por consiguiente el siguiente paso, es eliminar ésta, para ello se utiliza el método de Hildreth–Lu. El anterior cuadro muestra las estimaciones del valor ρ, para encontrar el mínimo de $S(b_0, b_1, \rho)$, el cual se obtiene cuando $\rho = 0.9556$, porque con éste se tiene la menor suma de cuadrados del residual. Con este esquema, las variables transformadas se calculan como

$$y_t' = y_t - \hat{\rho} y_{t-1} = y_t - 0.9556 y_{t-1}$$

$$x_{t1}' = x_{t1} - \hat{\rho} x_{t-1\,1} = x_{t1} - 0.9556 x_{t-1\,1} \qquad \text{para } t = 2, 3, \ldots, 56$$

$$x_{t2}' = x_{t2} - \hat{\rho} x_{t-1\,2} = x_{t2} - 0.9556 x_{t-1\,2}$$

y en la tabla 4.11 de la siguiente página se pueden visualizar éstas.

Entonces, el modelo transformado a estimar con mínimos cuadrados ordinarios es:

$$y_t' = b_0' + b_1' x_{t1}' + b_2' x_{t2}' + a_t$$

con $x_{t1}' = x_{t1} - \hat{\rho} x_{t-1\,1}$, $x_{t2}' = x_{t2} - \hat{\rho} x_{t-1\,2}$, $b_0' = b_0\left(1 - \hat{\rho}\right)$, $b_1' = b_1$, $b_2' = b_2$ y $a_t = \varepsilon_t - \hat{\rho}\varepsilon_{t-1}$.

Tabla 4.11 Variables transformadas

Fecha	Inflación y_t'	M_1 x_{t1}'	PIB x_{t2}'
1987/02	−9.39	−16.68	−0.62
1987/03	−3.99	−16.26	−0.97
1987/04	−7.18	−10.42	−1.02
1988/02	6.20	2.31	−0.02
1988/03	45.19	29.70	0.56
1988/04	43.10	33.71	0.06
1989/01	41.96	35.23	−1.47
1989/02	9.55	16.40	−0.73
1989/03	2.21	0.49	−0.50
1989/04	−0.86	−1.44	0.58
1990/01	−3.76	−7.57	−0.01
1990/03	−3.23	−6.31	0.10
1992/04	15.31	33.01	1.06
1993/01	2.84	2.07	0.63
1993/02	1.35	2.17	1.18
1993/03	0.81	−2.80	0.18
1993/04	1.36	6.44	−0.02
1994/01	1.69	−2.78	−0.30
1994/02	0.64	7.66	−1.52
1994/03	0.47	3.36	−0.01
1994/04	0.13	2.95	−0.10
1995/01	−7.38	16.28	4.88
1995/02	−17.26	2.72	4.28
1995/03	−6.04	−5.06	0.74
1995/04	−4.87	−8.07	0.02
1996/01	2.70	−20.84	−6.30
1996/02	15.47	−18.05	−2.96
1996/03	5.01	0.68	−1.00
1996/04	3.65	3.65	−0.47
1997/01	3.77	1.51	0.71
1997/02	5.16	2.53	−1.61
1997/03	2.95	0.91	0.00
1997/04	2.74	6.47	0.30
1998/01	2.63	9.01	−0.37
1998/02	0.84	5.81	1.86
1998/03	0.21	5.48	0.45
1998/04	−1.16	3.27	1.02
1999/01	−0.22	0.72	2.99
1999/02	1.53	1.01	−0.58
1999/03	2.13	−0.55	−0.36
1999/04	3.38	−3.34	−0.33
2000/01	3.64	2.01	−3.37
2000/02	1.44	−0.38	0.33
2000/03	0.92	3.00	0.52
2000/04	0.51	5.67	0.89
2001/01	1.79	3.36	4.69
2001/02	0.89	4.60	0.95
2001/03	1.16	−3.43	0.81
2002/03	0.97	−1.98	0.01
2002/04	0.14	5.77	−0.37
2003/01	0.15	0.39	−1.69
2003/02	0.91	1.19	1.26
2003/03	0.85	1.31	0.25
2003/04	0.27	−0.72	−0.14
2004/01	−0.15	0.10	−2.23

Un ajuste con mínimos cuadrados a las variables transformadas produce el modelo

$$\hat{y}' = 1.2523 + 0.8098x_1' - 2.7528x_2'$$

Tabla 4.12

	B	Error estándar de B	t (52)	Significancia
Constante	1.2523	0.9076	1.3797	0.1736
M1	0.8098	0.0841	9.6281	0.0000
PIB	–2.7528	0.5317	–5.1775	0.0000
R	R^2	Error estándar de la estimación	F(2,52)	Durbin–Watson
0.8092	0.6548	6.5616	49.3278	1.5390

En la tabla anterior se muestran los estadísticos de resumen, en donde puede observarse que el estadístico de Durbin–Watson para el modelo transformado es d = 1.539, y al compararlo con los valores críticos para α = 1% y n = 60, d_L = 1.35 y d_U = 1.48, se concluye que los errores no están correlacionados. Asimismo, se puede apreciar que $\hat{b}_0$ tiene un nivel de significancia grande, lo que indica que debe eliminarse de la ecuación de regresión.

Se hace el nuevo ajuste, obteniéndose el siguiente modelo

$$\hat{y}' = 0.8356x_1' - 2.7851x_2'$$

En el cuadro siguiente se tienen los estadísticos resumen emitidos, se aprecia en éste que el poder explicativo de esta regresión es aproximadamente del 67%.

Tabla 4.13

	B	Error estándar de B	t (53)	Significancia
M1	0.8356	0.0827	10.1028	0.0000
PIB	–2.7851	0.5357	–5.1991	0.0000
R	R^2	Error estándar de la estimación	F(2,53)	Durbin–Watson
0.8184	0.6698	6.6173	53.7656	1.5226

Una vez que se ha eliminado la autocorrelación en los errores, se puede cambiar el modelo transformado obtenido anteriormente a las variables originales como sigue:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

donde $b_0 = \dfrac{b_0'}{1-\hat{\rho}}$, $b_1 = b_1'$, $b_2 = b_2'$.

Dado que el modelo anterior no tiene b_0, se puede concluir que el "mejor" ajuste de la ecuación de regresión múltiple para el conjunto de datos es:

$$\hat{y} = \mathbf{0.8356}x_1 - \mathbf{2.7851}x_2$$

Como puede apreciarse, los valores estimados tienen el signo que inicialmente se esperaba. Por cada aumento en la tasa del M_1, la inflación crece 0.8356, disminuye 2.7851 por cada incremento en el producto interno bruto, y el porcentaje de la variación total en la variable dependiente que se puede explicar por medio de los dos regresores (x_1, x_2) a través del modelo es del 67%.

En el anexo B.3, se encuentran los resultados emitidos por el paquete estadístico STATISTICA, concernientes al cálculo de esta última regresión.

Por último, con este modelo se pronosticará la tasa de inflación para el segundo y el tercer trimestre de 2004. Los datos reales que se obtienen del Banco de México y del INEGI, se presentan en la tabla 4.14.

Si x'_0 = [15.10, 3.8], entonces una estimación puntual de la observación futura y_0 en este punto es aproximadamente de

$$\hat{y}_0 = 0.8356(15.10) - 2.7851(3.8) = 2.0342$$

Además, un intervalo de predicción del 95% para esta misma observación futura es:

$$y_0 \in (-2.0824,\ 6.1498)$$

Tabla 4.14

Fecha	Inflación	M_1	PIB
2004/02	4.29	15.10	3.8
2004/03	4.79	14.83	4.0

es decir, el pronóstico de la tasa de inflación estimada para el segundo trimestre de 2004, con una tasa de crecimiento monetario M_1 de 15.10 y una tasa de variación del PIB igual a 3.8, sería de entre **–2.0824** y **6.1498** unidades; que como se puede apreciar en el cuadro anterior, efectivamente estuvo en este intervalo (4.29).

Si ahora x'_0 = [14.83, 4.0], entonces una estimación puntual de la observación futura y_0 es aproximadamente de

$$\hat{y}_0 = 0.8356(14.83) - 2.7851(4.0) = 1.2515$$

De igual manera, un intervalo de predicción del 95% para esta misma observación futura es:

$$y_0 \in (-3.0289,\ 5.5311)$$

es decir, el pronóstico de la tasa de inflación estimada para el tercer trimestre de 2004, con una tasa de crecimiento monetario de 14.83 y una tasa de variación del PIB de 4.0, sería de entre **–3.0289** y **5.5311** unidades; que según la tabla 4.14, efectivamente su valor cae en este intervalo (4.79).

CONCLUSIÓN

El empleo de métodos de predicción para conocer los comportamientos económicos y sociales, entre otros, se ha vuelto una actividad cotidiana y fundamental para la toma de decisiones en cualquier campo. Por ejemplo, para el caso del gerente su quehacer básico es la toma de decisiones y por lo tanto debe elaborar estimaciones de lo que sucederá en el futuro; asimismo, debe prever escenarios que le permitan anticiparse a las posibles eventualidades que le indicarán la conveniencia o inconveniencia de una alternativa. En particular para analizar decisiones de inversión es necesario hacer predicciones de muy diversas variables: precios, tasas de interés, volúmenes de venta o de producción, etc., por lo tanto, es necesario que el analista conozca, por lo menos la existencia de ciertas técnicas que le ayuden en esta tarea.

Para elaborar pronósticos se cuenta con diversos métodos, en los cuales se hace uso de la información histórica, ya sea para predecir el comportamiento futuro o para suponer que el comportamiento histórico se seguirá manteniendo y sobre esta base hacer las estimaciones.

Dado que en el proceso de toma de decisiones se involucra el comportamiento humano, por ejemplo, a través de los dictámenes de los individuos a quienes está dirigida un determinado producto o servicio; las resoluciones del mercado están compuestas por muchísimas medidas individuales, imposibles de predecir con exactitud, o bien, existen factores externos imposibles de controlar (como el medio ambiente); cualquier estimación que se lleve a cabo implicará un grado de error inevitable. Es decir, se debe tener presente que no existe ningún método de pronóstico infalible; lo que hacen estos procedimientos es construir el "mejor modelo", pero el proceso de observación y registro de los datos que se tienen de un fenómeno, impiden la exactitud plena en los resultados, por lo cual el elemento aleatorio o de error siempre estará presente y será impredecible.

En el presente trabajo se describen sólo los métodos de predicción utilizando regresión lineal que consiste básicamente en modelar la relación funcional entre variables independientes, de manera que conociendo alguna o algunas de ellas se pueda predecir el valor de otra. La mayor fuerza del análisis de regresión lineal se refiere a que es un método que permite determinar (estimar) prácticamente cualquier tipo de relación lineal que pudiera existir entre una variable dependiente y uno o más regresores.

Por supuesto, existen algunas desventajas del uso de la regresión múltiple. Una se refiere a qué requiere valores (estimados o reales) de las variables independientes antes de que se pueda hacer un pronóstico. En este contexto, habría que preguntarse por la calidad de los datos.

Por principio, en México no es fácil realizar una encuesta y no existe una cultura para responderla; asimismo, los organismos gubernamentales que podrían tener información estadística "confiable", no la tienen completa, o la existente difiere entre ellos, aunque su fuente sea la misma. Sumado a esto, los datos pueden cambiar de un día a otro, sin notificar al usuario de la causa. Todo lo anterior trae como consecuencia que se dude de la calidad de los datos y por consiguiente se tenga incertidumbre, tanto en el modelo desarrollado, como en el pronóstico elaborado.

Otra desventaja potencial es la tendencia a pensar que siempre que exista un alto valor de R^2, la ecuación de regresión es automáticamente buena, si así fuera el caso, se deben de satisfacer los supuestos de la regresión (que no siempre se cumplen) y disponer de datos suficientes (al menos 30 observaciones). Por lo anterior, se debe de hacer un análisis de los residuales, ver si no se están violando los supuestos, entre otros, para así determinar si el modelo es bueno o no.

Dado que esta técnica de pronóstico son una herramienta necesaria para la planeación macro y microeconómica, se han considerado algunos de los detalles de la aplicación de la regresión múltiple en la práctica. En estos ejemplos, se puede observar primeramente que

el "modelo perfecto" de regresión lineal no existe, el error aleatorio siempre está presente; que no es fácil de encontrar la mejor ecuación de regresión lineal de los datos, se necesita de paciencia para analizar varios modelos hasta encontrar el "mejor ajuste"; pero además de una sensibilidad muy especial al fenómeno en análisis, que sólo se puede conseguir convirtiéndose en un experto del tema; que los supuestos no siempre se cumplen y se violaran algunos, por lo cual es necesario cambiar de método o bien hacer una transformación de los datos; y por último, el más importante, el pronóstico puede fallar o acertar.

ANEXO A

MÉTODO DE MÍNIMOS CUADRADOS ORDINARIOS (MCO)

El objetivo del método de los mínimos cuadrados ordinarios es encontrar una recta que se ajuste de una manera adecuada a la nube de puntos definida por todos los pares de valores muestrales (x_i, y_i).

Los residuos e_i se definen como la distancia que hay entre el valor observado y_i y el correspondiente valor estimado $\hat{y}_i$, que será la imagen en el eje de abscisas del valor x_i.

A.1. AJUSTE DE UNA LÍNEA RECTA

El procedimiento que existe para construir una línea que se ajuste a un determinado número de valores observados se denomina *método de mínimos cuadrados ordinarios*, de acuerdo con esta propuesta, la recta se debe construir considerando que la suma de los cuadrados de las desviaciones verticales de todos los puntos respecto a la recta sea mínima. En seguida se describe el método.

Supóngase que hay n pares de observaciones, por ejemplo (x_1, y_1), (x_2, y_2), . . . , (x_n, y_n). Estos datos se emplearan para estimar los parámetros desconocidos b_0 y b_1 en la ecuación

$$y = b_0 + b_1 x + \varepsilon$$

donde ε es un error aleatorio con media cero y varianza σ^2. Los $\{\varepsilon\}$ se supone también que son variables aleatorias no correlacionadas. Así, el modelo muestral de regresión se puede escribir

$$y_i = b_0 + b_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

y el criterio de mínimos cuadrados ordinarios es

$$S(b_0,\ b_1)=\sum_{i=1}^{n}(e_i)^2=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$$

Los estimadores, por MCO, de b_0 y b_1, que se designarán por $\hat{b}_0$ y $\hat{b}_1$, deben de satisfacer

$$\left.\frac{\partial S}{\partial b_0}\right|_{\hat{b}_0,\ \hat{b}_1}=-2\sum_{i=1}^{n}\left(y_i-\hat{b}_0-\hat{b}_1x_i\right)=0$$

y

$$\left.\frac{\partial S}{\partial b_1}\right|_{\hat{b}_0,\ \hat{b}_1}=-2\sum_{i=1}^{n}\left(y_i-\hat{b}_0-\hat{b}_1x_i\right)x_i=0$$

La simplificación de estas dos ecuaciones produce

$$n\hat{b}_0+\hat{b}_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i$$

$$\hat{b}_0\sum_{i=1}^{n}x_i+\hat{b}_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}y_ix_i$$

Las anteriores igualdades se denominan **ecuaciones normales de mínimos cuadrados**. Su solución es la siguiente:

$$\hat{b}_0=\bar{y}-\hat{b}_1\bar{x}$$

y

$$\hat{b}_1=\frac{\sum_{i=1}^{n}(x_iy_i)-\frac{\left(\sum_{i=1}^{n}y_i\right)\left(\sum_{i=1}^{n}x_i\right)}{n}}{\sum_{i=1}^{n}\left(x_i^2\right)-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}=\frac{\sum_{i=1}^{n}(x_iy_i)-n\bar{x}\,\bar{y}}{\sum_{i=1}^{n}\left(x_i^2\right)-n\bar{x}^2}$$

en donde:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{y} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

son los promedios de y_i y x_i, respectivamente. Por consiguiente, $\hat{b}_0$ y $\hat{b}_1$, en las ecuaciones anteriores, son los **estimadores por mínimos cuadrados** de la ordenada al origen y la pendiente, respectivamente. El modelo ajustado de regresión lineal simple es, entonces,

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

A.2. PROPIEDADES DE LOS ESTIMADORES DE REGRESIÓN LINEAL SIMPLE

Los estimadores por cuadrados mínimos tienen algunas propiedades estadísticas importantes, que son útiles en evaluar la suficiencia del modelo. Primero, obsérvese que $\hat{b}_0$ y $\hat{b}_1$ son variables aleatorias, puesto que son justamente combinaciones lineales de las observaciones y_i, y éstas lo son.

Los estimadores $\hat{b}_0$ y $\hat{b}_1$ son *insesgados* de los parámetros b_0 y b_1 del modelo. Para demostrarlo con $\hat{b}_1$, considérese

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

porque $\sum_{i=1}^{n} \left(x_i y_i\right) - n\bar{x}\,\bar{y} = \sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)$ y $\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$

Entonces su valor esperado es

$$E\left(\hat{b}_1\right) = E\left[\frac{\sum_{i=1}^{n} y_i\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right] = \frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} E\left[\sum_{i=1}^{n} y_i\left(x_i - \bar{x}\right)\right]$$

$$= \frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} E\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(b_0 + b_1 x_i + \varepsilon_i\right)\right]$$

$$= \frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\left\{E\left[b_0\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\right] + E\left[b_1\sum_{i=1}^{n} x_i\left(x_i - \bar{x}\right)\right] + E\left[\sum_{i=1}^{n}\varepsilon_i\left(x_i - \bar{x}\right)\right]\right\}$$

$$= \frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} b_1\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = b_1$$

puesto que $\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$, y por la suposición de que $E(\varepsilon_i) = 0$. De tal manera, se concluye que efectivamente $\hat{b}_1$ es un estimador insesgado de la pendiente verdadera b_1. Considérese ahora la varianza de este parámetro, supóngase que $V(\varepsilon_i) = \sigma^2$ se concluye que $V(y_i) = \sigma^2$, y

$$V\left(\hat{b}_1\right) = V\left[\frac{\sum_{i=1}^{n} y_i\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right] = \frac{1}{\left\{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right\}^2} V\left[\sum_{i=1}^{n} y_i\left(x_i - \bar{x}\right)\right]$$

Las variables aleatorias $\{y_i\}$ no están correlacionadas debido a que $\{\varepsilon_i\}$ no lo están. Por tanto, la varianza de la suma en la última ecuación es justo la suma de las varianzas, y la correspondiente a cada término en la suma, dígase $V\left[y_i\left(x_i - \bar{x}\right)\right]$, es $\sigma^2\left(x_i - \bar{x}\right)^2$. En consecuencia,

$$V\left(\hat{b}_1\right) = \frac{1}{\left\{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right\}^2}\sigma^2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Usando un planteamiento similar, se puede demostrar que

$$E\left(\hat{b}_0\right)=b_0 \quad \text{y} \quad V\left(\hat{b}_0\right)=\sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right]$$

La covarianza de $\hat{b}_0$ y $\hat{b}_1$ no es cero; efectivamente, $Cov\left(\hat{b}_0, \hat{b}_1\right)=-\sigma^2\dfrac{\bar{x}}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$.

Obsérvese que $\hat{b}_0$ es un estimador neutral de b_0.

A.3. AJUSTE DE UNA FUNCIÓN LINEAL DE VARIAS VARIABLES

El problema ahora es explicar y en términos de $x_1, x_2, \ldots, x_k$, mediante la relación

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k+\varepsilon \tag{1}$$

De la misma manera, este método puede utilizarse para estimar los coeficientes de regresión en la ecuación anterior. Para ello, supóngase que se disponen $n > k$ observaciones, y déjese que x_{ij} denote la observación i–ésima o el nivel de la variable j. Se considerará además, que el término del error ε en el modelo tiene $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$, y que las $\{\varepsilon_i\}$ son variables aleatorias no correlacionadas. Los datos se presentan en la tabla A.1.

Tabla A.1 Datos para la regresión lineal múltiple

Observación	Respuesta	Regresores			
i	y	x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Se puede escribir el modelo muestral de regresión, correspondiente a la ecuación anterior, en términos de las observaciones como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$y_i = \beta_0 + \sum_{j=1}^{k} \left(\beta_j x_{ij}\right) + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

La función de mínimos cuadrados es

$$S(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{k} \left(b_j x_{ij}\right) \right)^2$$

La función S se minimizará con respecto a b_0, b_1, b_2, ..., b_k. Los estimadores de mínimos cuadrados de éstos, denominados por $\hat{b}_0$, $\hat{b}_1$, ..., $\hat{b}_k$, deben satisfacer

$$\left.\frac{\partial S}{\partial b_0}\right|_{\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{b}_0 - \sum_{j=1}^{k} \left(\hat{b}_j x_{ij}\right) \right) = 0$$

y

$$\left.\frac{\partial S}{\partial b_j}\right|_{\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{b}_0 - \sum_{j=1}^{k} \left(\hat{b}_j x_{ij}\right) \right) x_{ij} = 0, \qquad j = 1, 2, \ldots, k$$

Al simplificar las anteriores igualdades, se obtienen las **ecuaciones normales de mínimos cuadrados**

$$\begin{array}{ccccccccc}
n\hat{b}_0 & + & \hat{b}_1 \sum_{i=1}^{n} (x_{i1}) & + & \hat{b}_2 \sum_{i=1}^{n} (x_{i2}) & + \cdots + & \hat{b}_k \sum_{i=1}^{n} (x_{ik}) & = & \sum_{i=1}^{n} (y_i) \\
\hat{b}_0 \sum_{i=1}^{n} (x_{i1}) & + & \hat{b}_1 \sum_{i=1}^{n} (x_{i1}^2) & + & \hat{b}_2 \sum_{i=1}^{n} (x_{i1} x_{i2}) & + \cdots + & \hat{b}_k \sum_{i=1}^{n} (x_{i1} x_{ik}) & = & \sum_{i=1}^{n} (x_{i1} y_i) \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\hat{b}_0 \sum_{i=1}^{n} (x_{ik}) & + & \hat{b}_1 \sum_{i=1}^{n} (x_{ik} x_{i1}) & + & \hat{b}_2 \sum_{i=1}^{n} (x_{ik} x_{i2}) & + \cdots + & \hat{b}_k \sum_{i=1}^{n} (x_{i\,k}^2) & = & \sum_{i=1}^{n} (x_{ik} y_i)
\end{array}$$

Nótese que hay $p = k + 1$ ecuaciones normales, para cada uno de los coeficientes de regresión desconocidos. La solución para las ecuaciones normales serán los **estimadores de mínimos cuadrados** de los coeficientes de regresión $\hat{b}_0$, $\hat{b}_1$, ..., $\hat{b}_k$.

Es más cómodo resolver las ecuaciones normales si ellas se expresan en notación de matriz. Eso permite presentar en forma compacta al modelo, los datos y los resultados. En notación matricial la ecuación (1), en términos de las observaciones, puede escribirse como

$$\boldsymbol{y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$$

donde

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \qquad \text{y} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

En general, $\boldsymbol{y}$ es un vector ($n \times 1$) de las observaciones, $\boldsymbol{X}$ es una matriz ($n \times p$) de los niveles de las variables independientes, $\boldsymbol{B}$ es un vector ($p \times 1$) de los coeficientes de regresión, y $\boldsymbol{\varepsilon}$ es un vector ($n \times 1$) de los errores aleatorios.

Se desea encontrar el vector $\hat{\boldsymbol{B}}$ de los estimadores de mínimos cuadrados que minimice

$$S\left(\hat{\boldsymbol{B}}\right) = \sum_{i=1}^{n} \left(\varepsilon_i\right)^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \left(\boldsymbol{y} - \boldsymbol{XB}\right)'\left(\boldsymbol{y} - \boldsymbol{XB}\right)$$

Nótese que $S(\boldsymbol{B})$ puede expresarse como

$$\begin{aligned} S\left(\boldsymbol{B}\right) &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{B}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{XB} + \boldsymbol{B}'\boldsymbol{X}'\boldsymbol{XB} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{B}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{B}'\boldsymbol{X}'\boldsymbol{XB} \end{aligned}$$

puesto que $\boldsymbol{B'X'y}$ es una matriz de (1×1), es decir, un escalar, y su transpuesta $(\boldsymbol{B'X'y})' = \boldsymbol{y'XB}$ es el mismo escalar. Los estimadores de mínimos cuadrados deben de satisfacer

$$\left.\frac{\partial S}{\partial B}\right|_{\hat{B}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{B}} = 0$$

que se simplifica a

$$\boldsymbol{X'X}\hat{\boldsymbol{B}} = \boldsymbol{X'y}$$

Las ecuaciones anteriores son las *ecuaciones normales de mínimos cuadrados*. Ellas son idénticas a las primeras. Para resolver estas igualdades, se multiplica ambos lados por la inversa de $\boldsymbol{X'X}$, de tal modo, el *estimador de mínimos cuadrados* de $\boldsymbol{B}$ es

$$\hat{\boldsymbol{B}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

siempre y cuando exista la matriz inversa $(\boldsymbol{X'X})^{-1}$. Lo anterior se cumple si los regresores son linealmente independientes, esto es, si ninguna columna de la matriz $\boldsymbol{X}$ es una combinación lineal de las demás columnas.

Es fácil ver que la forma matricial de las ecuaciones normales es idéntica a la forma escalar. Así la forma completa es

$$\begin{bmatrix} n & \sum_{i=1}^{n}(x_{i1}) & \sum_{i=1}^{n}(x_{i2}) & \cdots & \sum_{i=1}^{n}(x_{ik}) \\ \sum_{i=1}^{n}(x_{i1}) & \sum_{i=1}^{n}(x_{i1}^2) & \sum_{i=1}^{n}(x_{i1}x_{i2}) & \cdots & \sum_{i=1}^{n}(x_{i1}x_{ik}) \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n}(x_{ik}) & \sum_{i=1}^{n}(x_{ik}x_{i1}) & \sum_{i=1}^{n}(x_{ik}x_{i2}) & \cdots & \sum_{i=1}^{n}(x_{ik}^2) \end{bmatrix} \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}(y_i) \\ \sum_{i=1}^{n}(x_{i1}y_i) \\ \vdots \\ \sum_{i=1}^{n}(x_{ik}y_i) \end{bmatrix}$$

Si se efectúa la multiplicación matricial indicada, resultará la forma escalar de las ecuaciones normales. En esta forma es fácil ver que $\boldsymbol{X'X}$ es una matriz simétrica ($p \times p$) y $\boldsymbol{X'y}$ es un vector columna ($p \times 1$). Adviértase la estructura especial de $\boldsymbol{X'X}$. Los componentes de la diagonal de $\boldsymbol{X'X}$ son las sumas de cuadrados de los elementos en las columnas de $\boldsymbol{X}$, y los términos fuera de la diagonal son las sumas de los productos cruzados de las columnas de $\boldsymbol{X}$. Asimismo, nótese que los integrantes de $\boldsymbol{X'y}$ son las sumas de los productos cruzados de las columnas de $\boldsymbol{X}$ y las observaciones $\{y_i\}$.

El modelo ajustado de regresión que corresponde a los niveles de las variables independientes es $\hat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}$. En notación escalar, el vector de valores ajustados $\hat{y}_i$ es

$$\hat{y}_i = \hat{b}_0 + \sum_{i=1}^{n} \hat{b}_j x_{ij} \qquad i = 1, 2, \ldots, n$$

A.4. PROPIEDADES DE LOS ESTIMADORES DE REGRESIÓN LINEAL MÚLTIPLE

Las propiedades estadísticas del estimador de mínimos cuadrados se demuestran con facilidad. Se examinará primero el sesgo:

$$E\left(\widehat{\boldsymbol{B}}\right) = E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\right] = E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{B}+\varepsilon\right)\right]$$
$$= E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}+\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\varepsilon\right] = \boldsymbol{B}$$

porque $E(\varepsilon) = \mathbf{0}$ y $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \mathbf{1}$. Entonces $\widehat{\boldsymbol{B}}$ es un *estimador insesgado* de $\boldsymbol{B}$.

La propiedad de varianza de $\widehat{\boldsymbol{B}}$ se expresa con la matriz de covarianza

$$Cov\left(\widehat{\boldsymbol{B}}\right) = E\left\{\left[\widehat{\boldsymbol{B}} - E\left(\widehat{\boldsymbol{B}}\right)\right]\left[\widehat{\boldsymbol{B}} - E\left(\widehat{\boldsymbol{B}}\right)\right]'\right\}$$

que es una matriz simétrica de ($p \times p$), cuyo j–ésimo elemento diagonal es la varianza de $\hat{b}_j$ y cuyo (ij)–ésimo componente fuera de ésta, es la covarianza entre $\hat{b}_i$ y $\hat{b}_j$.

La covarianza de la matriz de $\widehat{\boldsymbol{B}}$ es

$$Cov\left(\widehat{\boldsymbol{B}}\right) = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Por ende, si se hace $\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, la varianza de $\hat{b}_j$ es $\sigma^2 C_{jj}$, y la covarianza entre $\hat{b}_i$ y $\hat{b}_j$ es $\sigma^2 C_{ij}$.

ANEXO B

RESULTADOS EMITIDOS POR EL PROGRAMA STATISTICA DE LOS CASOS PRÁCTICOS

B.1. LEY DE OKUN EN MÉXICO

Tabla B.1 Resumen del modelo

R	R^2	R^2 ajustada	Error estándar de la estimación	Estadística Durbin-Watson
0.5240	0.2745	0.2631	0.9489	0.2440

Predictores: (Constante), PIB

Variable dependiente: DESEMPLEO

Tabla B.2 ANOVA

	Suma de cuadrados	Grados de libertad	Cuadrado medio	F(1, 64)	Significancia
Regresión	21.7989	1	21.7979	24.2105	0.0000
Residual	57.6223	64	0.9003		
Total	79.4202	65			

Predictores: (Constante), PIB

Variable dependiente: DESEMPLEO

Tabla B.3 Coeficientes

	Coeficientes ordinarios B	Error estándar	Coeficientes estandarizados Beta	t (64)	Significancia
(Constante)	0.0394	0.1632		24.1232	0.0000
PIB	–0.1977	0.0402	–0.5239	–4.9204	0.0000

Variable dependiente: DESEMPLEO

Tabla B.4 Residuales

	Mínimo	Máximo	Media	Desviación estándar	*n*
Valor pronosticado	2.4537	5.1619	3.3756	0.5791	66
Valor pronosticado estandarizado	−1.5919	3.0845	0.0000	1.0000	66
Error estándar del valor pronosticado	0.1168	0.3814	0.1560	0.0548	66
Valor pronosticado ajustado	2.3938	4.9811	3.3650	0.5551	66
Residual	−1.6107	2.4335	7.54×10^{-16}	0.9415	66
Residual estandarizado	−1.6975	2.5650	0.0000	0.9920	66
Residual PRESS	−1.7214	2.7092	0.0106	0.9875	66
Distancia Mahalanobis	0.0004	9.5144	0.9848	1.8180	66
Distancia de Cook	0.0000	0.6194	0.0255	0.0800	66

B.2. INFLACIÓN EN MÉXICO

Tabla B.5 Resumen del modelo con 69 datos

Estadísticos	Valor
R	0.798299
R^2	0.637282
R^2 ajustada	0.626291
F(2, 66)	57.979766
Significancia	0.000000
Error estándar de la estimación	23.576377
Durbin - Watson	0.286119

Predictores: (Constante), PIB, M_1

Variable dependiente: INFLACIÓN

Tabla B.6 ANOVA

	Suma de cuadrados	Grados de libertad	Cuadrado medio	F(2, 66)	Signifi-cancia
Regresión	64455.589561	2	32227.794781	57.979767	0.000000
Residual	36685.805830	66	555.845543		
Total	101141.395392	68			

Predictores: (Constante), PIB, M_1

Variable dependiente: INFLACIÓN

Tabla B.7 Coeficientes

	BETA	Error estándar de BETA	B	Error estándar de B	t (66)	Signifi-cancia
Constante			10.977133	4.801593	2.286144	0.025463
M_1	0.768681	0.075038	0.797358	0.077837	10.243950	0.000000
PIB	–0.365091	0.075038	–4.889548	1.004954	–4.865445	0.000007

Tabla B.8 Correlación de los coeficientes

		PIB	M_1
Correlaciones	PIB	1.000000	–0.154791
	M_1	–0.154791	1.000000
Covarianzas	PIB	1.009932	–0.012108
	M_1	–0.012108	0.006059

Variable dependiente: INFLACIÓN

Tabla B.9 Residuales

	Mínimo	Máximo	Media	Desviación estándar	*n*
Valor pronosticado	–7.3555	109.0990	29.3516	30.7880	69
Valor pronosticado estandarizado	–1.1923	2.5902	0.0000	1.0000	69
Error estándar del valor pronosticado	2.8673	9.5287	4.6409	1.6332	69
Valor pronosticado ajustado	–8.2141	103.9280	29.1341	30.4892	69
Residual	–73.6062	68.8089	-1.157×10^{-14}	23.2265	69

	Mínimo	Máximo	Media	Desviación estándar	*n*
Residual estandarizado	–3.1222	2.9186	0.0000	0.9850	69
Residual PRESS	–81.6102	78.2438	0.2176	25.1818	69
Distancia Mahalanobis	0.0203	10.1222	1.9710	2.3280	69
Distancia de Cook	0.0000	0.4427	0.0293	0.0800	69

Tabla B.10 Resumen del modelo con 56 datos transformados

Estadísticos	Valor
R	0.818441
R^2	0.669846
R^2 ajustada	0.657387
F(2, 66)	53.765573
Significancia	0.000000
Error estándar de la estimación	6.617257
Durbin – Watson	1.522605

Predictores: (Constante), PIB, M1

Variable dependiente: INFLACIÓN

Tabla B.11 ANOVA

	Suma de cuadrados	Grados de libertad	Cuadrado medio	F(2, 66)	Signifi–cancia
Regresión	4708.583754	2	2354.291877	53.765572	0.000000
Residual	2320.768903	53	43.788093		
Total	7029.352658	55			

Predictores: (Constante), PIB, M1

Variable dependiente: INFLACIÓN

Tabla B.12 Coeficientes

	BETA	Error estándar de BETA	B	Error estándar de B	t (53)	Signifi–cancia.
M_1	0.834093	0.082560	0.835566	0.082706	10.102822	0.000000
PIB	–0.429243	0.082560	–2.785085	0.535682	–5.199137	0.000003

BIBLIOGRAFÍA

BANCO DE MÉXICO

http://www.banxico.org.mx

México

Diciembre de 2004

CEBRIÁN López, Inmaculada.

La ley de Okun.

http//:www2.uah.es/ icebrian/cap10_1.pdf.

Universidad de Alcalá, España

Diciembre de 2003

COLE, Julio H.

Nociones de regresión lineal

http://www.economia.ufm.edu.gt/Catedraticos/jhcole/Nociones.doc

Guatemala

Enero de 2004

CORTÉS, Fernando

Regresión logística en la investigación social: potencialidades y limitaciones

http://www.rau.edu.uy/fcs/soc/revista_13/cortes13.htm

Uruguay

Enero de 2004

CUBILLO Arias, Eilyn, Ana Cecilia VALVERDE Kikut y Jorge MADRIGAL Badillo

Estimación de la ley de Okun para Costa Rica

http://www.bccr.fi.cr/udie/Documentos/DIE-03-2002-NTESTIMACION%20LA%20LEY%20DE%20OKUN.pdf

Costa Rica

Febrero de 2004

CUEVAS Salgado, Yolanda

Pronóstico de divisas. México. Tesis maestría. Facultad de Ingeniería, UNAM. 1997. 119 p.

DI PIETRO, Sergio R.

Desempleo 2001: una odisea nacional

http://www.vaneduc.edu.ar/uai/comuni/pulso/numero-07/pi07_07.htm

Argentina

Febrero de 2004

ELÍAS Mata, Nereo

Modelos de regresión en pronóstico. México. México. Tesis maestría. Facultad de Ingeniería, UNAM. 1992. 134 p.

ENCICLOPEDIA MULTIMEDIA VIRTUAL DE ECONOMÍA (EMVI)

http://www.eumed.net.cursecon/medir/index.htm

Marzo de 2004

GARAVITO, Cecilia

La ley de Okun en el Perú: 1970 – 2000

http://www.pucp.edu.pe/economia/pdf/DDD212.pdf

Perú

Febrero de 2004

GESELL, Silvio.

El dinero tal y como es.

http://www.systemfehler.de/es/parte1/16.htm

Alemania

Febrero de 2004

HILLIER, Frederick S. y Gerarld J. LIEBERMAN

Investigación de Operaciones. Séptima edición. México. McGraw – Hill. 2002. 1223 p.

HINES, William W. y Douglas C. MONTGOMERY

Probabilidad y estadística para ingeniería y administración. Tercera edición. México. Compañía Editorial Continental. 1993. 834 p.

HOEL, Paul Gerhard. y Raymond James JESSEN

Estadística básica para negocios y economía. México. Compañía Editorial Continental. 1980. 452 p.

INSTITUTO NACIONAL DE ESTADÍSTICA, GEOGRAFÍA E INFORMÁTICA (INEGI)

http://www.inegi.gob.mx

México

Diciembre de 2004

INSTITUTO TECNOLÓGICO DE LA PAZ

El dinero y la teoría monetaria

http://www.itlp.edu.mx/publica/tutoriales/economia2/tema22.htm

México

Abril de 2004

KAZMIER, Leonard J.

Estadística aplicada a la administración y a la economía. Tercera edición. México. McGraw – Hill. 2001. 416 p.

LEMOIS, Félix A.

Estimación de la ley de Okun para Puerto Rico

http://www.jpops01.jp.gobierno.pr:7778/pls/portal/url/ITEM/B423B085DF794407B7C5B7924C836904

Puerto Rico

Enero de 2004

MAKRIDAKIS, Sypros y Steven C.WHEELWRIGTH

Manual de técnicas de pronósticos. México. Quinta Edición. Editorial Limusa. 1989. 470 p.

MAKRIDAKIS, Sypros y Steven C. WHEELWRIGTH

Métodos de pronósticos y aplicaciones. Primera edición. México. Editorial Limusa. 2000. 482 p.

MASON, Robert Deward y Lind A. DOUGLAS

Estadística para administración y economía. México. Alfaomega Grupo Editor. 1992. 911 p.

MONTGOMERY, Douglas C., Elizabeth A. PECK y G. Geoffrey VINING

Introducción al análisis de regresión lineal. Tercera edición. México. Compañía Editorial Continental. 2004. 588 p.

NETER, J., M. H. KUTHER, C. J. NACHTSHEIM y W. WASSERMAN

Applied Linear Statistical Models. Cuarta edición. Richard D. Irwin, Homewood, III, 1996.