UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Hipótesis en el modelo de regresión lineal por Mínimos Cuadrados Ordinarios

Apellidos, nombre	Chirivella González, Vicente (vchirive@eio.upv.es)
Departamento	Estadística e Investigación Operativa Aplicadas y Calidad
Centro	Facultad de Administración y Dirección de Empresas

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

1 Resumen de las ideas clave

Para estimar los parámetros de un modelo de regresión lineal necesitamos realizar ciertos supuestos sobre el modelo y sobre la forma en que llevaremos a cabo dicha estimación. Aunque estos supuestos puedan estar claros, al menos desde un punto de vista estadístico, deberíamos entender, de forma sencilla, el motivo y el significado de los supuestos, ¿qué hay detrás de ellos?, ¿por qué los asumimos?, ¿podemos hacerlo?

2 Introducción

El método de los *mínimos cuadrados ordinarios* (*MCO*) es el método de estimación más habitual cuando se realiza el ajuste de un modelo de regresión lineal en los parámetros, aunque no es el único.

Atendiendo al nombre del método, parece claro que hay que minimizar el cuadrado de algo, y el añadido de ordinario sugiere que deben existir otros tipos de mínimos cuadrados. El "algo" en cuestión es el error, cuantificado como la diferencia entre el valor real de la observación y el valor previsto para ella. A esta diferencia se le llama residuo, y por tanto se minimiza la suma de cuadrados de los residuos con el objeto de estimar los parámetros del modelo. ¿Por qué hacemos esto? Por otra parte, tenemos el calificativo de ordinario, que debe entenderse como sinónimo de habitual. Este calificativo resume un cierto número de simplificaciones o hipótesis realizadas sobre el error, sobre las variables y sobre los parámetros del modelo. Debemos entender su significado para poder admitirlas como adecuadas respecto al problema que tratamos de resolver. ¿Se pueden asumir estas hipótesis en mi problema?

3 Objetivos

El objetivo de este artículo docente es que entendamos el motivo y la forma en que se ha minimizado el error del modelo de regresión por *MCO* y las implicaciones que tienen las hipótesis o simplificaciones sobre el error, que no son las únicas, pero si son las más relevantes de las realizadas.

4 Regresión lineal simple. Hipótesis.

En este apartado vamos a formular en primer lugar el modelo de regresión, identificando todas sus partes. A continuación constataremos que es imposible estimar los parámetros del modelo sólo con los datos disponibles, y que necesitamos más, lo que nos ofrece el método de los *MCO*. Veremos entonces lo que significan tanto el método como sus hipótesis.

4.1 Formulación del modelo

La expresión general de un modelo de regresión, para un total de *k* variables explicativas es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U$$

Ecuación 1: Modelo de regresión lineal

donde *Y* es la variable explicada, las X_j son las variables explicativas, y los parámetros β_j son unos parámetros que cuantifican la relación existente entre la variable explicada y cada variable explicativa. Hasta aquí, el modelo que uno puede encontrar en la teoría correspondiente para explicar a la variable. Por ejemplo, una teoría económica para explicar el consumo de un producto, o una ley que relaciona la intensidad y el voltaje en un transistor.

Ahora bien, debemos añadir al modelo el término *U*, una variable aleatoria que recoja la influencia sobre la variable explicada de otras variables explicativas que no hemos tenido en cuenta en el modelo (por el motivo que sea). El término *U* es necesario ya que la variable explicada (*Y*) es una variable aleatoria, pero al otro lado de la igualdad (en el modelo), no son aleatorios ni los parámetros β ni las variables explicativas *X*. Como debe existir una parte aleatoria en ese lado del modelo, añadimos este término *U*. Si despejamos *U* en la Ecuación 2, ¿qué nos queda? Pues la diferencia entre el valor real de la variable y el valor propuesto para la misma por el modelo, es decir, el error.

$$Y - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = U$$

Ecuación 2: El error en el modelo de regresión lineal

¿Qué distribución tiene este error? Pues lo habitual es asumir que este error es una variable aleatoria con distribución conjunta normal, de valor medio cero y de desviación típica desconocida.

Nuestro objetivo es estimar el valor de los parámetros β y de la desviación típica del error, aquello que no conocemos del modelo. Para ello debemos apoyarnos en los datos de que disponemos, los *n* valores observados de las variables. Si sustituimos en el modelo tenemos un sistema de *n* ecuaciones con *n+k+1* incógnitas, las *k+1* incógnitas que son los parámetros β y las *n* incógnitas que son el error *U*, que nos permitirán estimar luego la desviación típica del error. Como tenemos más incógnitas que ecuaciones, no hay una única solución.

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + U_1 \\ Y_2 &= \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \cdots + \beta_k X_{k2} + U_2 \\ &\cdots \\ Y_n &= \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_k X_{kn} + U_n \end{aligned}$$

Ecuación 3: Sistema de ecuaciones al aplicar los n datos disponibles

Es necesario conseguir más ecuaciones para resolver el problema, un total de *k+1* ecuaciones más. Tendremos que centrar la atención sobre el error *U* para hallarlas.

Dado que *U* es un error, sería interesante que el error fuera lo más pequeño posible. Pero como el error es una variable aleatoria, "lo más pequeño posible" se traduce en que su valor medio sea cero, y en que su varianza sea (de nuevo) lo más pequeña posible. De la minimización de esa varianza del error se obtendrán las *k+1* ecuaciones que nos faltan, y de aquí viene el nombre de método de los mínimos cuadrados ordinarios.

4.2 Hipótesis

Para poder estimar los parámetros del modelo y la varianza del error debemos establecer algunas hipótesis simplificadoras respecto a la perturbación, respecto a las variables explicativas y la explicada, y respecto a los parámetros β del modelo. De no hacerlo así, obtener las estimaciones deseadas se convertiría en una tarea bastante complicada, si no imposible. Nos centramos en las hipótesis sobre el error, por ser las más importantes:

1. Los errores son variables aleatorias de media nula.
2. Todos los errores tienen la misma varianza.
3. Todos los errores están incorrelacionados entre sí.
4. El error tiene una distribución conjunta normal. Junto a la hipótesis anterior se concluye que los errores son independientes entre sí.
5. El error no depende de las variables explicativas X_i.

Pero, ¿qué significado tienen estas hipótesis establecidas? Vayamos por partes.

1- Los errores son variables aleatorias de media nula.

Esto quiere decir que no existe error sistemático, es decir, que no predecimos la variable explicada siempre por encima (o por debajo) de su valor real. En ocasiones lo haremos por encima, en ocasiones lo haremos por debajo, y en promedio no estaremos añadiendo ni quitando nada. Si, bueno... lo ideal sería que el error fuese cero, pero como no puede ser constante (el cero es constante), ya que es una variable aleatoria, pues hay que conformarse con esto.

El cumplimiento de esta condición está garantizado por el método de los MCO, así que no debemos preocuparnos por ella.

2- Todos los errores tienen la misma varianza.

Esto quiere decir que la importancia del error es siempre la misma. El error no, el error particular e individual será diferente en cada observación. Hablamos aquí de su importancia en conjunto, de su orden de magnitud. El error estará "acotado" y se espera que sus valores no superen un cierto límite. Evidentemente, desearemos que ese límite sea lo más pequeño posible.

Sin embargo esta simplificación puede ser excesiva en algunos casos, pues hay variables que por su propia naturaleza hacen que la varianza del error no sea constante. ¿Por ejemplo? Pues podría ser el Gasto en I+D realizado en las empresas cuando se explica mediante sus INGRESOS, tal y como aparece en el Gráfico 1.

Si tenemos un grupo de empresas que tienen ingresos pequeños, la diferencia de gasto en I+D entre ellas será también pequeño.

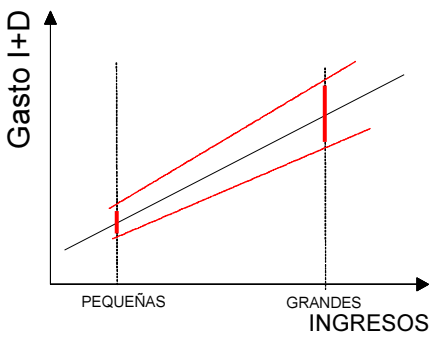

Gráfico 1: Relación entre Gasto en I+D e Ingresos de las empresas

Dado que no ingresan mucho dinero, tampoco habrá mucho dinero que invertir, y el gasto de las empresas que tienen mucho interés en realizar I+D no será muy diferente de las que tienen poco interés en ello. Sin embargo, si las empresas tienen ingresos elevados, las diferencias de gastos entre ellas serán mucho más grandes. Si hay interés, una empresa podrá invertir mucho más dinero que si no lo hay, y si no hay interés, la inversión puede ser muy, muy pequeña. La variabilidad (varianza) del gasto en I+D depende de los ingresos, es mucho más pequeña en las empresas de ingresos bajos que en las de ingresos altos. Esta es la naturaleza de la relación.

Hay una ventaja adicional en este supuesto, y es que es más sencillo estimar algo que es constante, que estimar algo que no lo es. Las expresiones de los estimadores serán así considerablemente más sencillas.

A esta característica de varianza no constante se le llama heterocedasticidad, no es inusual, y hace que se incumpla esta hipótesis sobre el error.

3- Todos los errores están incorrelacionados entre sí.

Esto quiere decir que el error cometido con una observación, un año, una comunidad autónoma, no influirá en el error que se cometerá con otra observación, otro año u otra comunidad autónoma. Si el error es positivo y grande con una comunidad autónoma de España, este valor no tiene que forzar a que el error de otra comunidad sea también positivo, o negativo, o grande, o pequeño. No tiene nada que ver, y así debe reflejarse en el modelo.

Claro, que hay variables que por su propia naturaleza presentan relación con ellas mismas. ¿Por ejemplo? Pues el consumo trimestral de helado. El consumo de helado de un trimestre está relacionado con el del trimestre anterior. Si en el trimestre anterior se consumió mucho helado, cabe esperar que el siguiente trimestre también se consuma mucho. Vale, uno puede argumentar que cada vez se consume más helado porque cada vez hay más personas para consumir, y porque la gente tiene cada vez más dinero. Muy bien. ¿Alguna variable explicativa más? Es que si nos dejamos alguna variable explicativa, su efecto de "cada vez más" va a aparecer en el error. Recordemos: el error recoge el efecto de las variables explicativas no consideradas en el modelo. Basta con que nos dejemos una variable explicativa para que haya un "cada vez más" en el error, y por ello relación del error con su pasado.

A esta característica se le llama autocorrelación. Tampoco es una propiedad inusual, y es otra de las hipótesis que se suele incumplir. De hecho, cuando existe autocorrelación, es mejor dejar los *modelos de regresión* para utilizar *modelos ARIMA* o de F*unción de Transferencia*, creados específicamente para cuando existe autocorrelación.

4- El error tiene una distribución conjunta normal. Los errores son independientes.

Esta simplificación es necesaria para más tarde poder realizar pruebas de hipótesis con distribuciones conocidas (distribuciones *t* y *F*), y tener expresiones sencillas de los estimadores puntuales y de los intervalos de confianza de los parámetros y de la desviación típica del error.

Recordemos que la distribución normal es la distribución más habitual que uno puede encontrar para una variable aleatoria, así que asumir normalidad no es una condición muy exigente.

5- El error no depende de las variables explicativas *X*

Esto quiere decir que el modelo está bien planteado, que la relación entre la variable explicada y la variable explicativa (cada una de ellas) es correcta y completa. Si la relación es por tramos, lineal, cuadrática, exponencial,... lo que sea, el modelo está conforme a esa relación real existente.

Esta es una hipótesis que se incumple con mucha frecuencia. Las teorías para explicar a una variable suelen proponer más de un modelo para explicarla, modelos que difieren en las variables explicativas que lo componen, o en la relación existente entre la variable explicada y la explicativa. Y si no existe el modelo, entonces debemos proponerlo de forma racional. Pero por muy racional que sea la propuesta, va a ser complicado que la propongamos bien a la primera.

Al incumplimiento de esta hipótesis se le denomina error de formulación.

4.3 Estimación de parámetros

El método de los *mínimos cuadrados ordinarios* consiste en la obtención de un hiperplano de forma que se minimice la suma de los cuadrados de las distancias entre cada una de las observaciones de la variable y dicho hiperplano (residuos).

Vayamos por partes con las explicaciones. Nuestro interés es encontrar una recta (hiperplano si tienes tres o más variables explicativas) que pase lo más cerca posible de todos los puntos. Dicho así parece sencillo, pero... ¿cómo cuantificamos ese "lo más cerca posible"? Hay muchas maneras de hacerlo. Podemos medir la distancia vertical desde la observación a la recta, podemos medir la distancia en horizontal entre la observación y la recta, podemos trazar la perpendicular a la recta por cada observación y medir la distancia en esa dirección, podemos medir la distancia en vertical pero añadirle un peso para que ciertas observaciones sean más relevantes que otras en el ajuste,... muchas formas. En el método de los *MCO* se escoge medir la distancia en vertical desde el punto a la recta, como aparece en el Gráfico 2. A esa distancia se le denomina, como ya hemos dicho, residuo.

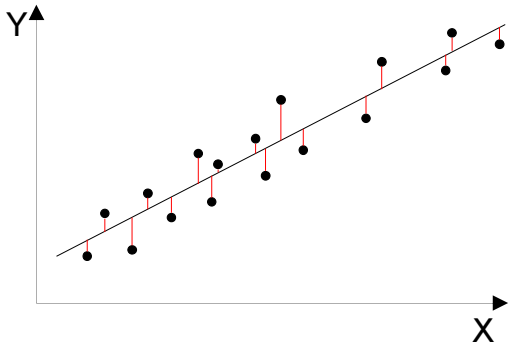

Gráfico 2: Recta ajustada y error cometido (residuo)

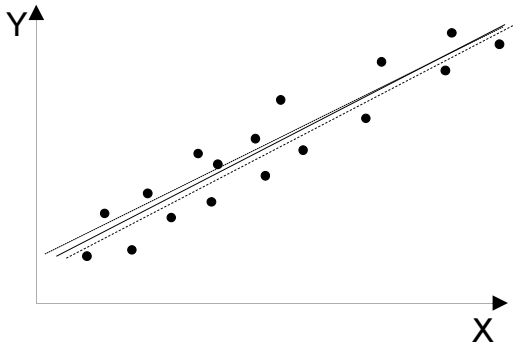

Gráfico 3: Posibles rectas para explicar la relación entre variables. ¿Cuál es mejor?

Si modifico ligeramente la posición de la recta, Gráfico 3, tenemos unas nuevas distancias (residuos), con lo que nos planteamos la pregunta de ¿cuál de ellas pasa más cerca de todos los puntos y es mejor recta? El ser humano tiene ciertas limitaciones en sus capacidades, y comparar el conjunto de, pongamos, veinte

residuos de un ajuste con los otros veinte de otro resulta muy complicado. ¿La solución? Acumular de alguna forma esos veinte números en uno solo. Así es mucho más sencillo. La forma más habitual de hacerlo es sumarlos. ¿Es buena solución? Pues no mucho, porque hay residuos negativos y positivos que pueden cancelarse mutuamente, dando como resultado un número acumulado muy pequeño cuando proviene en realidad de distancias muy grandes. ¿Y si se prescinde del signo? Pues si, pero vamos a optar por elevar el residuo al cuadrado, mejor que prescindir del signo. Al elevar al cuadrado perdemos el signo, si, pero además convertimos un número grande en otro mucho más grande aún, así que penalizamos aquellos puntos muy alejados de la recta. Mejor cuanto más cerca. Sumamos los residuos al cuadrado (*SCR*), y este número nos ofrece una forma de medir "lo cerca que pasa la recta de los puntos". Un valor de *SCR* grande significa que todos los puntos quedan muy lejos de la recta, mientras que un valor pequeño significa que quedan muy cerca. Cada recta tiene su número, y podemos elegir una de ellas como la mejor de todas por tener ese número pequeño, por pasar lo más cerca posible de las observaciones.

La suma de cuadrados de residuo depende de las estimaciones de los parámetros β, por lo que podemos hacer que sea lo más pequeña posible:

$$SCR = \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 = \sum_{j=1}^{n} (Y_j - (b_0 + b_1 X_{1j} + b_2 X_{2j} + \cdots + b_k X_{kj}))^2$$

Ecuación 4: Suma de cuadrados de residuos

donde los *b* que aparen en la Ecuación 4 son las estimaciones de los parámetros β. Al derivar parcialmente respecto a cada *b* e igualar a cero (condición de mínimo), se obtiene el resto de las ecuaciones necesarias (*k+1*) para estimar los parámetros del modelo y finalmente la expresión de los estimadores *b*. No entramos en el detalle de los cálculos, pero el estimador de los parámetros se obtiene con un simple producto matricial, y la estimación de la varianza del error ni siquiera es una matriz:

$$b = (X'X)^{-1}X'Y$$

Ecuación 5: Estimador de los parámetros del modelo

$$\hat{\sigma}^2 = \frac{SCR}{n-k-1}$$

Ecuación 6: Estimador de la varianza del error

donde *Y* es el vector de la variable a explicar, *X* es la matriz de datos, como ya se ha visto anteriormente. Esta expresión tan simple es el motivo por el cual las calculadoras, o el propio *Excel*, pueden hacer ajustes de regresión.

Para que el sistema de ecuaciones tenga solución única (y no sea indeterminado) el producto de matrices *X'X* debe ser invertible, y para ello debe cumplirse que:

1. el número de datos sea mayor que el de los parámetros a estimar, *n>k+1*
2. no deben existir relaciones exactas entre las variables explicativas *X*

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

5 Cierre

Hemos visto las expresiones de los estimadores de los parámetros y de la varianza del error para un modelo de regresión ajustado mediante el método de los mínimos cuadrados ordinarios.

Estas expresiones se han obtenido tras realizar ciertos supuestos sobre el término de error, sobre las variables explicativas y sobre los parámetros del modelo. Aunque los supuestos son razonables, son arbitrarios y tiene unas claras limitaciones.

La obtención del mejor hiperplano de regresión se basa en minimizar las distancias medidas verticalmente entre las observaciones y el hiperplano. Esta es sólo una de las muchas formas en que se pueden medir las distancias. Además, sumar las distancias al cuadrado también es otro supuesto razonable, pero arbitrario, de acumular las distancias para poder ser entendidas y manejadas. Si estos supuestos nos parecen bien, tenemos un método de estimación. El método de los mínimos cuadrados ordinarios.

Pero no sólo hay supuestos para el método, también hay supuestos para uno de los componentes del modelo, el término de error o perturbación. Frente a supuestos razonables, y sobre los que uno no debe preocuparse demasiado, como que el valor medio del error sea cero y que tenga distribución normal, hay otros que si exigen atención. Debemos preocuparnos de que no existan en nuestro modelo problemas de heterocedasticidad y de autocorrelación, y asegurarnos de que el modelo esté correctamente formulado.

Hemos visto que tanto la heterocedasticidad como la autocorrelación son propiedades naturales de ciertas variables, y que no es extraño encontrarlas en nuestros problemas. Por otra parte, disponer de la relación real entre variable explicada y explicativas puede ser imposible, y este método supone que la conocemos.

Habría que tener cuidado y estar atentos al incumplimiento de las simplificaciones en nuestro problema. Es muy tentador utilizar un programa de ordenador, que nos aparezcan unos resultados, y que los demos como válidos cuando en realidad no lo son, y no se ha utilizado la herramienta adecuada para el análisis.

6 Bibliografía

D. Gujarati : "Basic Econometrics", Ed. McGrawHill - 4ª edition, páginas 335-560.

D. Peña: "Estadística: Modelos y Métodos. (Vol.2) Modelos lineales y Series temporales", Ed. Alianza Universidad-Textos, páginas 307-548.