Modelos de Regresión y Correlación IV. Aplicación de Pruebas de Significación Estadística

MODELS OF REGRESSION AND CORRELATION IV. APPLICATION OF STATISTICAL SIGNIFICANCE TESTS

MD Mauricio Salinas, Sonia Carlos C.
MD, MPH (c) Director Unidad de Epidemiología y Estadística. Fundación Científica y Tecnológica ACHS.
MS, Epidemióloga, Unidad de Epidemiología y Estadística. Fundación Científica y Tecnológica ACHS.

RESUMEN

Al hacer un análisis de regresión lineal se puede aplicar una prueba de significación estadística basado en el método de máxima verosimilitud o de mínimos cuadrados. En este artículo se explica el método de mínimos cuadrados, y cómo se aplica la prueba de significación estadística en este caso. Se aplica lo enseñado mediante un ejemplo.

(Salinas M. 2007. Modelos de Regresión y Correlación IV. Aplicación de Pruebas de Significación Estadística. Cienc Trab. Ene-Mar; 10 (27):34-37).

Descriptores: REGRESIÓN Y CORRELACIÓN, MÍNIMOS CUADRADOS, PRUEBAS DE SIGNIFICACIÓN ESTADÍSTICA.

ABSTRACT

When conducting a linear regression analysis, a statistical significance test based on the maximum verisimilitude or minimum square methods may be applied. This article explains the minimum square method and how the statistical significance test is applied in this case. Subject presented is explained by means of an example and the way of doing it using the EpiInfo software.

Descriptors: REGRESSION AND CORRELATION, MINIMUM SQUARES, STATISTICAL SIGNIFICANCE TESTS.

INTRODUCCIÓN

Continuando con la serie sobre regresión, en este artículo se explica los fundamentos de la aplicación y desarrollo de las pruebas de significación estadística en este caso particular. El fundamento de los modelos de regresión ha sido explicado en números anteriores, así como el de las pruebas de significación en general, por lo cual esas materias se asumen ya conocidas (Silva C, Salinas M. 2006; Salinas M. 2007).

Como breve repaso, en el modelo de regresión lineal simple hay tres parámetros que se deben estimar: los coeficientes de la recta de regresión, α y β; y la varianza de la distribución normal, σ^2. El cálculo de estimadores para estos parámetros puede hacerse por diferentes métodos, siendo los más utilizados el método de máxima verosimilitud y el método de mínimos cuadrados.

Correspondencia / *Correspondence*
Mauricio Salinas F.
Fundación Científica y Tecnológica ACHS
Vicuña Mackenna 210 piso 6, Providencia, Santiago
Tel.: (56-2) 685 38 84
e-mail: msalinasf@achs.cl
Recibido: 4 de enero de 2008 / Aceptado: 10 de marzo de 2008

A continuación se revisa el método de mínimos cuadrados para someter a prueba estadística el modelo de regresión.

Para entender la aplicación de estos modelos se utilizan ejemplos simples, basados principalmente en el modelo de regresión lineal. Los cálculos se pueden realizar en una planilla de cálculo como Excel o equivalente. Los valores estadísticos se pueden obtener utilizando el software EpiInfo de distribución gratuita (Software EpiInfo).

MODELO DE MÍNIMOS CUADRADOS

Existen muchos problemas donde un conjunto de datos asociados en pares indican que la regresión es lineal, donde desconocemos la distribución conjunta de las variables aleatorias; sin embargo, se requiere estimar los coeficientes de regresión α y β. Este tipo de problemas generalmente son desarrollados por un método de ajuste de curvas conocido como Método de Mínimos Cuadrados. Para entender el método de mínimos cuadrados se debe entender que el modelo de regresión se fundamenta en que se está modelando una respuesta de comportamiento lineal y se está cuantificando la variabilidad para cada punto del modelo respecto a la distribución de los valores reales (Silva C, Salinas M. 2006).

El método de los mínimos cuadrados nos permite encontrar la ecuación (la pendiente y la ordenada de origen) de la recta a partir de los datos experimentales.

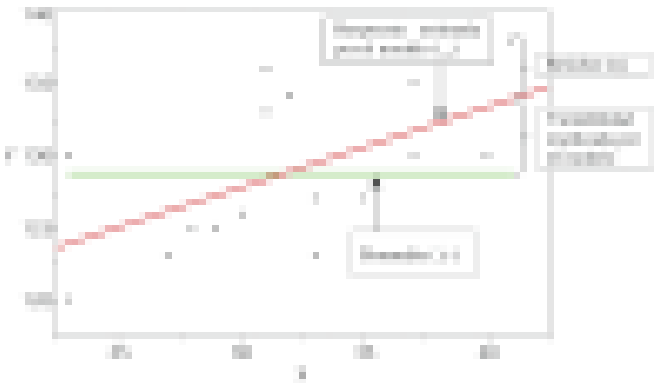
La forma de cuantificar la variabilidad se basa en calcular los cuadrados de las diferencias entre el valor real de la variable

dependiente *y*, el valor predicho de ésta y el promedio de la muestra, para cada valor de *X*.
Como se explicó en un capítulo anterior (Silva C, Salinas M. 2006), dado un conjunto de datos provenientes de una muestra aleatoria y utilizando un modelo de regresión lineal simple, observamos los siguientes valores (Figura 1).

y_i: Valor de la respuesta *y* para x_i,
$\hat{y}_i$: Valor de la respuesta *y*, estimado de acuerdo al modelo, para x_i
$\bar{y}$: Valor promedio de la respuesta y a través de las n observaciones de la muestra aleatoria utilizada
x_i: Valor de x para la observación i - ésima

Se pueden establecer las siguientes relaciones matemáticas entre estos valores:
$\Sigma (y - \bar{y})^2$: Es la suma de las diferencias entre cada valor de *y* y la media. Representa la variabilidad total de la respuesta *y*, sin prestar atención a la relación que ella pueda tener con X.
$\Sigma (\hat{y} - \bar{y})^2$: Es la suma de las diferencia entre cada valor estimado de Y por el modelo escogido y la media. Es la variabilidad explicada por el modelo de regresión.
$\Sigma (y - \hat{y})^2$: Es la suma de las diferencias entre cada valor real de *y* y su estimado por el modelo. Es la variabilidad de *y* no explicada por el modelo y se le llama Error Residual o residuo.

Figura 1.
Gráfica de regresión lineal simple.



Lo que se está haciendo al sumar estas diferencias es cuantificar a qué distancia queda cada punto real del estimado por el modelo. Mientras más cerca estén los puntos reales del modelo, mejor será el modelo, y menor será el residuo. La manera de cuantificarlo es hacer la diferencia de cada punto, elevarlo al cuadrado y sumar éstas. Debe elevarse al cuadrado ya que, si no, la suma de las diferencias respecto al promedio siempre daría cero.
Si se coloca los valores mencionados en una tabla se puede observar las siguientes relaciones: (Tabla 1)

Tabla 1.
Relación entre suma y media de cuadrados para un modelo de regresión lineal simple.

	Suma de cuadrados (SC)	g.l.*	Media de cuadrados (MC)
Modelo	$\Sigma (\hat{y} - \bar{y})^2$	1	$\Sigma (\hat{y} - \bar{y})^2 / 1$
Error	$\Sigma (y - \hat{y})^2$	n - 2	$\Sigma (y - \hat{y})^2 / n - 2$
Total	$\Sigma (y - \bar{y})^2$	n - 1	$\Sigma (y - \bar{y})^2 / n - 1$

* g.l. Grados de libertad

El concepto de g.l. es muy abstracto y de difícil definición, pero tiene relación con el número de categorías posibles y tiene una forma específica de estimarse según la prueba o modelo que se esté utilizando. Por ejemplo, en la regresión lineal, el total de grados de libertad es n-1; los g.l. del modelo corresponden al número de variables independientes que participan en el modelo y la diferencia da los grados de libertad para el residuo. Se sugiere consultar la bibliografía para ahondar en este concepto.
La media de cuadrados corresponde a la suma de cuadrados dividida por el número de grados de libertad respectivo. La media de cuadrados representa la diferencia promedio entre cada punto y su referencia, sea ésta la media o el *y* estimado.
Se ha demostrado que la relación entre la media de cuadrados del modelo y la media del error se distribuye probabilísticamente con distribución F. Brevemente, la distribución F es una distribución de probabilidades, como la distribución normal, por ejemplo, pero su curva tiene otra forma (Taucher E. 1997).
Finalmente, se divide el cuadrado medio del modelo sobre el cuadrado medio del residuo, es decir:

$$F = \frac{\text{MC modelo}}{\text{MC residuo}}$$

El valor obtenido corresponde al valor de la estadística F y su probabilidad asociada. Estos valores se encuentran en tablas en libros de estadística o pueden obtenerse en algunas páginas Web (Lowry, Richard. VassarStats).
A continuación se analizará un ejemplo para entender mejor los conceptos. Para aquellos que deseen ir haciendo los cálculos, se sugiere ir copiando los datos en una planilla de computador.

APLICANDO LOS CONCEPTOS

En una empresa dada los trabajadores trabajan expuestos a una

Tabla 2.
Valores de concentración en sangre y orina en trabajadores expuestos a la sustancia T1.

	Concentración plasmática (mg/dl)	Concentración orina (ug/dl)
1	0,4	12,0
2	0,6	13,0
3	0,7	12,3
4	1,1	12,5
5	1,1	12,5
6	1,4	12,6
7	1,7	13,6
8	1,8	13,3
9	2,0	13,4
10	2,1	12,7
11	2,6	12,3
12	2,6	12,7
13	2,7	13,0
14	2,8	13,5
15	3,0	13,0
16	3,2	13,8
17	3,2	14,1
18	3,4	14,5
19	3,5	15,0
20	3,8	13,7
21	4,1	14,8
22	4,5	13,9
23	5,0	16,0
24	4,9	16,0
25	5,3	15,9

sustancia química potencialmente tóxica denominada T1. Para poder monitorizar el nivel de exposición de estos trabajadores, se desea evaluar la correlación entre la concentración en sangre de la sustancia y su valor en orina. Para ello se obtiene una muestra aleatoria de 25 individuos con los valores mostrados en la Tabla 2.
Si se grafica estos datos (Figura 2) podemos ver que existe una relación entre las mediciones que parece tener carácter lineal. De tal manera que la regresión lineal parece una buena aproximación para modelar la relación entre estas variables. Se ha verificado los supuestos del modelo de regresión lineal y estos se cumplen, así que el análisis y conclusiones son válidos (Silva C, Salinas M. 2006).
Entonces se decide utilizar un modelo de regresión lineal y someter a prueba estadística la hipótesis de que la concentración urinaria es un buen predictor de la concentración plasmática de la sustancia en estudio. Se define un nivel de significación de 0,05 y se plantean las dos hipótesis correspondientes:
H_0. No existe asociación entre la concentración plasmática y urinaria.
H_1. Existe asociación entre la concentración plasmática y urinaria.
Lo primero que se hace es aplicar el modelo de regresión lineal y estimar los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ como se ha explicado anteriormente (Silva C, Salinas M. 2006).
Se obtiene lo siguiente:

$\hat{\beta}_0 = -11,28$
$\hat{\beta}_1 = 1,03$

La ecuación del modelo lineal queda entonces:

$y = 1,03 * x + -11,28$

Con la fórmula anterior se debe calcular el valor de y para cada x, es decir, el valor de T1 en sangre que el modelo predice basado en la concentración urinaria. Se obtienen los valores mostrados en la Tabla 3.
A continuación se procede con la suma de cuadrados, según las fórmulas anteriormente enseñadas.

La variabilidad total es:

$$\sum (y - 2,7)^2 = 48,2$$

La variabilidad explicada por el modelo es:

$$\sum (\hat{y} - 2,7)^2 = 35,5$$

La diferencia entre ambos valores, da la variabilidad del residuo. También se puede calcular con la fórmula $\sum (y - \hat{y})^2$. El valor es 12,7.

Si se divide esos valores por los respectivos g.l. se obtienen los cuadrados medios. Esta información se resume en la Tabla 4.

Figura 2.
Gráfico de dispersión entre concentración plasmática y urinaria.

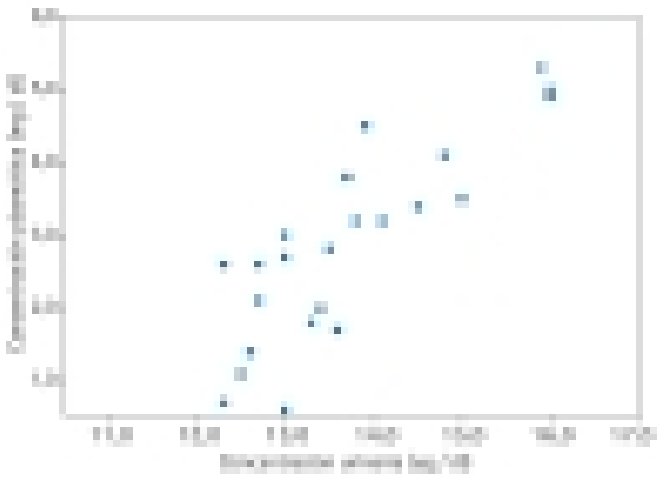


Tabla 3.
Valores medidos de concentración urinaria y valores estimados de concentración plasmática.

	Concentración orina (ug/dl)	Concentración plasmática (mg/dl) estimada (modelo)
1	12,0	1,05
2	13,0	2,08
3	12,3	1,36
4	12,5	1,57
5	12,5	1,57
6	12,6	1,67
7	13,6	2,70
8	13,3	2,39
9	13,4	2,49
10	12,7	1,77
11	12,3	1,36
12	12,7	1,77
13	13,0	2,08
14	13,5	2,59
15	13,0	2,08
16	13,8	2,90
17	14,1	3,21
18	14,5	3,62
19	15,0	4,14
20	13,7	2,80
21	14,8	3,93
22	13,9	3,00
23	16,0	5,16
24	16,0	5,16
25	15,9	5,06

Tabla 4.
Suma y media de cuadrados para el ejemplo.

	g.l.	Suma de cuadrados	Media de cuadrados
Modelo	1	35,5	35,52
Residuo (error)	23	12,7	0,55
Total	24	48,2	

Por último, se calcula el valor de la estadística F:

$$F = \frac{\text{MC modelo}}{\text{MC residuo}}$$

$$F = \frac{35,52}{0,55} = 64,6$$

El valor de probabilidad asociado al valor F 64,6 es < 0,0001. Como el valor encontrado es menor al nivel de significación establecido, se rechaza la hipótesis H_0.

COMENTARIO FINAL

El método de mínimos cuadrados es probablemente la forma más utilizada para someter a prueba de hipótesis los modelos de regresión en general. Es un método subjetivo, una estimación probabilística y con distribución de probabilidad conocida. Su aplicación en modelos de regresión múltiple o logística tiene algunas variaciones algebraicas, pero el fundamento es el mismo. Los cálculos mostrados acá son realizados rápidamente por diversos softwares estadísticos y lo que se ha intentado mostrar es su racionalidad y fundamentación.

REFERENCIAS

Silva C, Salinas M. 2006. Modelos de regresión y correlación. Cienc Trab 8(22): 185 – 189.

Salinas M. 2007. Pruebas de significación estadística. Cienc Trab 9(26): 200 – 203.

Software EpiInfo. http://www.cdc.gov/EpiInfo/. Accesado el 2 de febrero de 2008.

Lowry, Richard. VassarStats. Web site for statistical computation. http://faculty.vassar.edu/lowry/tabs.html#f. Accesado el 2 de febrero de 2008.

Taucher E. 1997 Bioestadística. Santiago: Editorial Unversitaria.