EYP1113 - Probabilidad y Estadística

Capítulo 8: Análisis de Regresión (Otro Enfoque)

Ricardo Aravena C. Ricardo Olea O.

Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2014

Probability Concepts in Engineering

Alfredo H-S. Ang† and Wilson H. Tang‡

† University of Illinois at Urbana-Champaing and University of California, Irvine

‡ Hong Kong University of Science & Technology

Contenido I

El Modelo de Regresión Múltiple

Especificación del Modelo

La regresión lineal multiple permite obtener dos importantes resultados:

1. Una ecuación lineal estimada que predice la variable dependiente, Y, en función de K variables independientes observadas, x_j, donde $j = 1, \ldots, K$.

$$\hat{y}_i = b_0 + b_1\, x_{1\,i} + b_2\, x_{2\,i} + \cdots + b_K\, x_{K\,i}$$

donde $i = 1, \ldots, n$ observaciones.

El Modelo de Regresión Múltiple

Especificación del Modelo

2. La variación marginal de la variable dependiente, Y, provocada por las variaciones de las variables independientes, que se estima por medio de los coeficientes, b'_j. En la regresión multiple, estos coeficientes dependen de que otras variables se incluyan en el modelo.

El coeficiente b_j indica la variación de Y, dada una variación unitaria de x_j, descontando al mismo tiempo el efecto simultáneo de las demás variables independientes.

En algunos problemas, ambos resultados son igual de importantes. Sin embargo, normalmente predomina uno de ellos.

El Modelo de Regresión Múltiple

Desarrollo del Modelo

El modelo de regresión múltiple define la relación entre una variable dependiente o endógena, Y, y un conjunto de variables independientes o exógenas, x_j, donde $j = 1, \ldots, K$. Se supone que las $x_{j\,i}$ son números fijos; Y es una variable aleatoria definida para cada observación, i, donde $i = 1, \ldots, n$, y n es el número de observaciones.

El modelo se define de la forma siguiente:

$$y_i = \beta_0 + \beta_1\ x_{1\,i} + \beta_2\, x_{2\,i} + \cdots + \beta_K\, x_{K\,i} + \varepsilon_i$$

donde las β_j son coeficientes constantes y las ε son variables aleatorias de media 0 y varianza σ^2.

El Modelo de Regresión Múltiple

Gráficos Tridimensionales

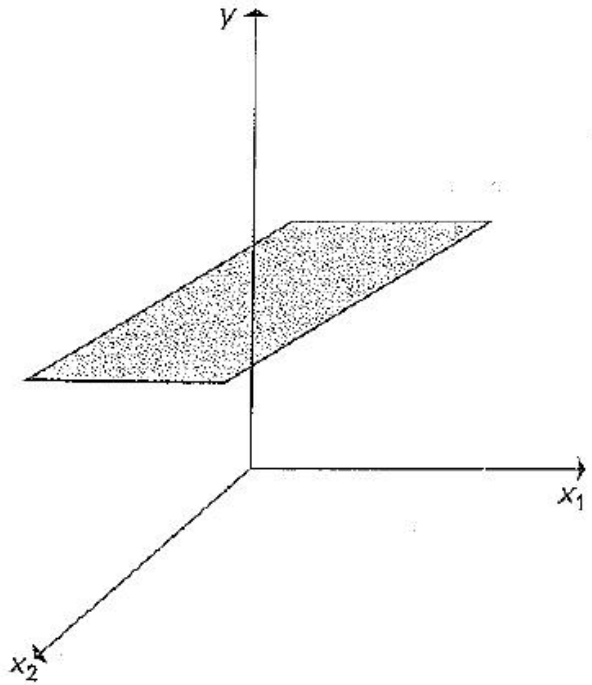

Figura 13.1. El plano es el valor esperado de Y en función de X_1 y X_2.

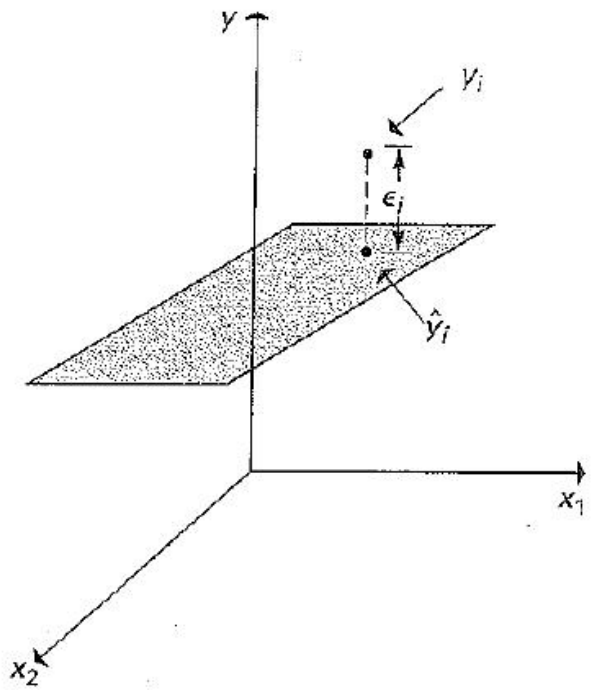

Figura 13.2. Comparación del valor observado y el esperado de Y en función de dos variables independiente

El Modelo de Regresión Múltiple

Datos Ejemplo

Ejemplo: Datos sobre ahorro y crédito inmobiliario.

La siguiente tabla proporciona el margen anual de beneficios (Y), ingresos anuales netos por dólar depositado (X_1), y el número de oficinas existentes ese año (X_2).

Tabla : Datos sobre ahorro y crédito inmobiliario

Año	Ingresos	Oficinas	Beneficio	Año	Ingresos	Oficinas	Beneficio
1	3.92	7298	0.75	14	3.78	6672	0.84
2	3.61	6855	0.71	15	3.82	6890	0.79
3	3.32	6636	0.66	16	3.97	7115	0.70
4	3.07	6506	0.61	17	4.07	7327	0.68
5	3.06	6450	0.70	18	4.25	7546	0.72
6	3.11	6402	0.72	19	4.41	7931	0.55
7	3.21	6368	0.77	20	4.49	8097	0.63
8	3.26	6340	0.74	21	4.70	8468	0.56
9	3.42	6349	0.90	22	4.58	8717	0.41
10	3.42	6352	0.82	23	4.69	8991	0.51
11	3.45	6361	0.75	24	4.71	9179	0.47
12	3.58	6369	0.77	25	4.78	9318	0.32
13	3.66	6546	0.78				

Estimación de Coeficientes

El modelo de regresión poblacional múltiple es

$$y_i = \beta_0 + \beta_1\ x_{1\,i} + \beta_2\, x_{2\,i} + \cdots + \beta_K\, x_{K\,i} + \varepsilon_i$$

y suponemos que se dispone de n conjuntos de observaciones. Se postulan los siguientes supuestos habituales para el modelo.

- Las $x_{j\,i}$ son o bien números fijos, o bien realizaciones de variables aleatorias, X_j, que son independientes de los términos del error, ε. En el segundo caso, la inferencia se realiza condicionada a los valores observados de las $x_{j\,i}$.
- El valor esperado de la variable aleatoria Y es una función lineal de kas variables independientes X_j.

Estimación de Coeficientes

- Los términos de error son variables aleatorias cuya media es cero y que tienen la misma varianza, σ^2. Este último supuesto se denomina homocedasticidad o varianza uniforme.

$$E(\varepsilon_i) = 0 \quad \text{y} \quad E(\varepsilon_i^2) = \sigma^2,$$

para $i = 1, \ldots, n$.

- Los términos de error aleatorio, ε_i, no están correlacionados entre sí, por lo que

$$E(\varepsilon_i \cdot \varepsilon_j) = 0 \quad \forall i \neq j$$

- No es posible hallar un conjunto de números que no sean iguales a cero, c_0, c_1,. . . , c_K, tal que

$$c_0 + c_1\, x_{1\,i} + \cdots + c_K\, x_{K\,i} = 0$$

Esta es la propiedad de la ausencia de relación lineal entre las X_j.

Estimación de Coeficientes

Métodos de Mínimos Cuadrados

Para una muestra de n observaciones $(x_{1\,i},\, x_{2\,i},\ldots,\, x_{K\,i},\, Y_i$, donde $i = 1,\ldots,n)$ medidas para u proceso cuyo modelo de regresión poblacional múltiple es

$$y_i = \beta_0 + \beta_1\, x_{1\,i} + \beta_2\, x_{2\,i} + \cdots + \beta_K\, x_{K\,i} + \varepsilon_i$$

Las estimaciones por mínimos cuadrados de los coeficientes $\beta_1,\, \beta_2,\ldots,\, \beta_K$ son los valores $b_0,\, b_1,\ldots,\, b_K$ para los que la suma de los cuadrados de las desviaciones

$$SCE = \sum_{i=1}^{n} (y_i - b_0 - b_1\, x_{1\,i} - b_2\, x_{2\,i} - \cdots - b_K\, x_{K\,i})^2$$

es la menor posible.

Estimación de Coeficientes

Métodos de Mínimos Cuadrados

La ecuación resultante

$$\hat{y}_i = b_0 + b_1\ x_{1\,i} + b_2\, x_{2\,i} + \cdots + b_K\, x_{K\,i}$$

es la regresión múltiple de Y sobre X_1, X_2,. . . ,X_K.

Estimación de Coeficientes

Métodos de Mínimos Cuadrados

Ejemplo Consideremos el caso con dos variables de predicción

$$\hat{y}_i = b_0 + b_1\ x_{1\,i} + b_2\, x_{2\,i}$$

Los estimadores de los coeficientes pueden resolverse utilizando las formas siguientes:

$$b_1 = \frac{s_y(r_{x_1\,y} - r_{x_1\,x_2}\, r_{x_2\,y})}{s_{x_1}\,(1 - r^2_{x_1\,x_2})}$$

$$b_2 = \frac{s_y(r_{x_2\,y} - r_{x_1\,x_2}\, r_{x_1\,y})}{s_{x_2}\,(1 - r^2_{x_1\,x_2})}$$

$$b_0 = \overline{y} - b_1\,\overline{x}_1 - b_2\,\overline{x}_2$$

donde r es la correlación muestral y s es la desviación típica muestral.

Poder Explicativo de una Ecuación de Regresión Múltiple

Descomposición Suma de Cuadrados

Comenzamos con el modelo de regresión múltiple ajustado mediante mínimos cuadrados

$$y_i = b_0 + b_1\ x_{1\,i} + b_2\, x_{2\,i} + \cdots + b_K\, x_{K\,i} + e_i = \hat{y}_i + e_i$$

donde las b_j son las estimaciones por mínimos cuadrados de los coeficientes del modelo de regresión poblacional y las e son los residuos del modelos de regresión estimado.

Poder Explicativo de una Ecuación de Regresión Múltiple

Descomposición Suma de Cuadrados

La variabilidad del modelo puede dividirse en los componentes

$$STC = SCR + SCE$$

las que se definen de la siguiente manera

$$\begin{aligned} STC &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\ &= \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \end{aligned}$$

Poder Explicativo de una Ecuación de Regresión Múltiple

Descomposición Suma de Cuadrados

Esta descomposición puede interpretarse como

$$\text{Variabilidad Muestral Total} = \text{Variabilidad Explicada} + \text{Variabilidad No Explicada}$$

El coeficiente de determinación, R^2, de la regresión ajustada es la proporción de la variabilidad muestral total explicada por la regresión

$$R^2 = \frac{SCR}{STC} = 1 - \frac{SCE}{STC}$$

y se deduce que

$$0 \leq R^2 \leq 1$$

Poder Explicativo de una Ecuación de Regresión Múltiple

Estimación de la Varianza de los Errores

Dado el modelo de regresión poblacional múltiple

$$y_i = \beta_0 + \beta_1\ x_{1\,i} + \beta_2\, x_{2\,i} + \cdots + \beta_K\, x_{K\,i} + \varepsilon_i$$

y los supuestos habituales de la regresión, sea σ^2 la varianza común del término de error, ε_i. Entonces, una estimación insesgada de esta varianza es

$$s_e^2 = \frac{1}{n-K-1}\sum_{i=1}^{n} e_i^2 = \frac{SCE}{n-K-1}$$

donde K es el número de variables independientes en el modelo de regresión. La raíz cuadrada de la varianza, s_e, también se llama **error típico de la estimación**.

Poder Explicativo de una Ecuación de Regresión Múltiple

Coeficiente de Determinación Ajustado

El **coeficiente de determinación ajustado**, $\overline{R}^2$, se define de la forma siguiente:

$$\overline{R}^2 = 1 - \frac{SCE/(n-K-1)}{STC/(n-1)}$$

Utilizamos esta medida para tener en cuenta el hecho de que las variables independientes irrelevantes provocan una pequeña reducción de la suma de los cuadrados de los errores. Port lo tanto, el $\overline{R}^2$ ajustado permite comparara mejor los modelos de regresión múltiple que tienen diferentes números de variables independientes.

Poder Explicativo de una Ecuación de Regresión Múltiple

Coeficiente de Correlación Multiple

El **coeficiente de correlación múltiple** es la correlación entre el valor predicho y el valor observado de la variable dependiente.

$$R = r(\hat{y},\, y) = \sqrt{R^2}$$

y es igual a la raíz cuadrada del coeficiente múltiple de determinación. Utilizamos R como otra medida de la fuerza de la relación entre variable dependiente y las variables independientes. Por lo tanto, es comparable a la correlación entre Y y X en la regresión simple.

Poder Explicativo de una Ecuación de Regresión Múltiple

Salida R

```
> summary(lm(Y~X1+X2))

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.085090 -0.039102 -0.003341  0.030236  0.105692

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.564e+00  7.940e-02  19.705 1.82e-15 ***
X1           2.372e-01  5.556e-02   4.269 0.000313 ***
X2          -2.491e-04  3.205e-05  -7.772 9.51e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0533 on 22 degrees of freedom
Multiple R-squared: 0.8653,     Adjusted R-squared: 0.8531
F-statistic: 70.66 on 2 and 22 DF,  p-value: 2.65e-10
```

Poder Explicativo de una Ecuación de Regresión Múltiple

Salida Excel

Resumen								
Estadísticas de la regresión								
Coeficiente de correlación múltiple	0,930212915							
Coeficiente de determinación R^2	0,865296068							
R^2 ajustado	0,853050256							
Error típico	0,053302217							
Observaciones	25							
ANÁLISIS DE VARIANZA								
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
Regresión	2	0,40151122	0,20075561	70,66057082	2,64962E-10			
Residuos	22	0,06250478	0,002841126					
Total	24	0,464016						
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	1,564496771	0,079395981	19,70498685	1,81733E-15	1,399839584	1,729153958	1,399839584	1,729153958
Ingresos	0,237197475	0,055559366	4,269261695	0,000312567	0,121974402	0,352420548	0,121974402	0,352420548
Oficinas	-0,000249079	3,20485E-05	-7,771949195	9,50879E-08	-0,000315544	-0,000182615	-0,000315544	-0,000182615

Intervalos de Confianza y Contraste de Hipótesis

Base para Inferencia

Sea el modelo de regresión poblacional

$$y_i = \beta_0 + \beta_1\ x_{1\,i} + \beta_2\, x_{2\,i} + \cdots + \beta_K\, x_{K\,i} + \varepsilon_i$$

Sean b_0, $b_1, \ldots,$ b_K las estimaciones por mínimos cuadrados de los parámetros poblacionales y s_{b_0}, $s_{b_1}, \ldots,$ s_{b_K} las desviaciones típicas de los estimadores de mínimos cuadrados. Entonces, si se cumplen los supuestos habituales de la regresión y si los términos de error, ε_i, siguen una distribución normal,

$$t_{b_j} = \frac{b_j - \beta_j}{s_{b_j}}, \quad j = 1, 2, \ldots, K$$

sigue una distribución t-student con $(n - K - 1)$ grados de libertad.

Intervalos de Confianza y Contraste de Hipótesis

Intervalos de Confianza

Si lo errores de la regresión poblacional, ε_i, siguen una distribución normal y se cumplen los supuestos habituales de la regresión, los intervalos de confianza bilaterales al $100(1-\alpha)\%$ de los coeficientes de regresión, β_j, son

$$b_j - t_{n-K-1,\,1-\alpha/2} < \beta_j < b_j + t_{n-K-1,\,1-\alpha/2}$$

donde $t_{n-K-1,\,1-\alpha/2}$ es el número para el que

$$P(t_{n-K-1} > t_{n-K-1,\,1-\alpha/2}) = \frac{\alpha}{2}$$

y la variable aleatoria t_{n-K-1} sigue una distribución t-student con $(n-K-1)$ grados de libertad.

Intervalos de Confianza y Contraste de Hipótesis

Contraste de Hipótesis

Si los errores de la regresión, ε_i, siguen una distribución normal y se cumplen los supuestos habituales del análisis de regresión, los siguientes contrastes de hipótesis tienen el nivel de significación α:

- Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \beta_j = \beta_j^* \quad o \quad H_0 : \beta_j \leq \beta_j^*$$

frente a la hipótesis alternativa

$$H_1 : \beta_j > \beta_j^*$$

Se rechaza H_0 si

$$\frac{b_j - \beta_j^*}{s_{b_j}} > t_{n-k-1,\,1-\alpha}$$

Intervalos de Confianza y Contraste de Hipótesis

Contraste de Hipótesis

- Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \beta_j = \beta_j^* \quad o \quad H_0 : \beta_j \geq \beta_j^*$$

frente a la hipótesis alternativa

$$H_1 : \beta_j < \beta_j^*$$

Se rechaza H_0 si

$$\frac{b_j - \beta_j^*}{s_{b_j}} < t_{n-k-1,\,\alpha} = -t_{n-k-1,\,1-\alpha}$$

Intervalos de Confianza y Contraste de Hipótesis

Contraste de Hipótesis

- Para contrastar la hipótesis nula

$$H_0 : \beta_j = \beta_j^*$$

frente a la hipótesis alternativa

$$H_1 : \beta_j \neq \beta_j^*$$

Se rechaza H_0 si

$$\left| \frac{b_j - \beta_j^*}{s_{b_j}} \right| > t_{n-k-1,\, 1-\alpha/2}$$

Contraste de Coeficientes de Regresión

Contraste de todos los Coeficientes

Consideremos el modelo de regresión múltiple

$$y_i = \beta_0 + \beta_1\ x_{1\,i} + \beta_2\, x_{2\,i} + \cdots + \beta_K\, x_{K\,i} + \varepsilon_i$$

Para contrastar la hipótesis nula

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$$

Frente a la hipótesis alternativa

$$H_1 :\ \text{Al menos un } \beta_j \neq 0$$

Contraste de Coeficientes de Regresión

Contraste de todos los Coeficientes

A un nivel de significación α, utilizamos la regla de decisión:

$$\text{Rechazar } H_0 \text{ si } \quad \frac{CMR}{s_e^2} > F_{K,\,n-K-1,\,1-\alpha}$$

donde $F_{K,\,n-K-1,\,\alpha}$ es el valor crítico de F (ver Tabla 9) para el que

$$P(F_{K,\,n-K-1} > F_{K,\,n-K-1,\,1-\alpha}) = \alpha$$

La variable aleatoria calculada $F_{K,\,n-K-1}$ sigue una distribución Fisher con K grados de libertad en el numerados y $(n-K-1)$ grados de libertad en el denominador.

Contraste de Coeficientes de Regresión

Contraste de un conjunto de Coeficientes

Dado un modelo de regresión con la descomposición de las variables independientes en los subconjuntos X y Z,

$$y_i = \beta_0 + \beta_1\ x_{1\,i} + \beta_2\, x_{2\,i} + \cdots + \beta_K\, x_{K\,i} + \alpha_1\, z_{1\,i} + \cdots + \alpha_r\, z_{r\,i} + \varepsilon_i$$

Para contrastar la hipótesis nula

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

de que los parámetros de regresión de un subconjunto son simultáneamente iguales a cero, frente a la hipótesis alternativa

$$H_1 :\ \text{Al menos un } \alpha_j \neq 0, \quad j = 1, \ldots, r$$

Contraste de Coeficientes de Regresión

Contraste de un conjunto de Coeficientes

Comparamos la suma de los cuadrados de los errores del modelo completos con la suma de los cuadrados de los errores del modelo restringido.

Se rechaza H_0 si

$$\frac{(SCE(r) - SCE)/r}{s_e^2} > F_{r,\, n-K-1,\, 1-\alpha}$$

con r el número de variables eliminadas.

Predicción

Dado que se cumple el modelo de regresión poblacional

$$y_i = \beta_0 + \beta_1\ x_{1\,i} + \beta_2\, x_{2\,i} + \cdots + \beta_K\, x_{K\,i} + \varepsilon_i, \quad i = 1, \ldots, n$$

y que los supuestos habituales del análisis de regresión son válidos, sean $b_0, b_1, \ldots, b_K$ las estimaciones por mínimos cuadrados de los coeficientes del modelo, β_j, siendo $j = 1, \ldots, K$, basados en los puntos de datos $x_{1\,i}, x_{2\,i}, \ldots, x_{K\,i}$, $(i = 1, \ldots, n)$. En tal caso, dada una nueva observación de un punto de datos $x_{1,\,n+1}, x_{2,\,n+1}, \ldots, x_{K,\,n+1}$, la mejor predicción lineal insesgada de $\hat{y}_{n+1}$ es

$$\hat{y}_{n+1} = b_0 + b_1\ x_{1,\,n+1} + b_2\ x_{2,\,n+1} + \cdots + b_K\ x_{K,\,n+1}$$