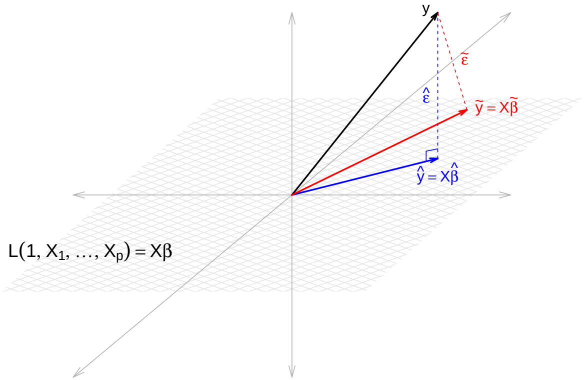
5 Regresión Lineal Múltiple



−ebz

Contenido

−ebz

Modelo de Regresión Lineal Múltiple

Modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

donde,

y_i	:	respuesta al i-ésimo nivel del regresor ($i = 1, \ldots, n$)
x_{ij}	:	j-ésimo regresor a su i-ésimo nivel ($j = 1, 2$)
β_0	:	ordenada al origen
β_j	:	coeficiente (tasa de cambio) del j-ésimo regresor
ϵ_i	:	error aleatorio

O bien, si $x_i = (1, x_{i1}, x_{i2})'$, y $\beta = (\beta_0, \beta_1, \beta_2)'$,

$$y_i = x_i'\beta + \epsilon_i, \qquad i = 1, \ldots$$

Equivalentemente,

$$y = X\beta + \epsilon$$

Supuestos:

$$\text{rango}(X) = 2 + 1 = 3 = q \; (\leq n)$$

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2) \text{ } \textit{i.i.d.} \quad \text{(Supuesto Esférico)}$$

–ebz

Mínimos Cuadrados

Suma de cuadrados

$$S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 = \sum_{i=1}^{n} (y_i - x_i' \beta)^2$$

Criterio de Mínimos Cuadrados:

$$\min_{\beta} S(\beta) \equiv \min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 \equiv \min_{\beta_0, \beta_1, \beta_2} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \right\}$$

$$\frac{\partial S(\beta)}{\partial \beta_j} = 0 \Longrightarrow \left\{ \begin{array}{rcl} 2\sum(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})(-1) & = & 0 \\ 2\sum(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})(-x_{i1}) & = & 0 \\ 2\sum(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})(-x_{i2}) & = & 0 \end{array} \right.$$

−ebz

Mínimos Cuadrados

Ecuaciones Normales

$$
\begin{array}{ccccccc}
n\beta_0 & + & \beta_1 \sum x_{i1} & + & \beta_2 \sum x_{i2} & = & \sum y_i \\
\beta_0 \sum x_{i1} & + & \beta_1 \sum x_{i1}^2 & + & \beta_2 \sum x_{i1} x_{i2} & = & \sum x_{i1} y_i \\
\beta_0 \sum x_{i2} & + & \beta_1 \sum x_{i1} x_{i2} & + & \beta_2 \sum x_{i2}^2 & = & \sum x_{i2} y_i
\end{array}
$$

cuya solución $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ son los *estimadores de mínimos cuadrados*.
El correspondiente estimador de la varianza σ^2 está dado por:

$$
s^2 = \frac{1}{n-3} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-3} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2
$$

Ejemplo: Desempeño de Supervisores[1]

La siguiente tabla muestra los resultados de un estudio de sicología industrial del desempeño de supervisores como función de distintas variables características (medidas entre 0 y 1) de los mismos. Las variables consideradas fueron:

variable	descripción
y	desempeño total del supervisor
x_1	manejo de quejas de los empleados
x_2	no permite privilegios especiales
x_3	oportunidad de aprender cosas nuevas
x_4	aumentos basados en desempeño
x_5	*muy crítico* en desempeño pobre
x_6	tasa de avance a mejores trabajos

n	y	x1	x2	x3	x4	x5	x6	n	y	x1	x2	x3	x4	x5	x6
1	0.43	0.51	0.30	0.39	0.61	0.92	0.45	16	0.81	0.90	0.50	0.72	0.60	0.54	0.36
2	0.63	0.64	0.51	0.54	0.63	0.73	0.47	17	0.74	0.85	0.64	0.69	0.79	0.79	0.63
3	0.71	0.70	0.68	0.69	0.76	0.86	0.48	18	0.65	0.60	0.65	0.75	0.55	0.80	0.60
4	0.61	0.63	0.45	0.47	0.54	0.84	0.35	19	0.65	0.70	0.46	0.57	0.75	0.85	0.46
5	0.81	0.78	0.56	0.66	0.71	0.83	0.47	20	0.50	0.58	0.68	0.54	0.64	0.78	0.52
6	0.43	0.55	0.49	0.44	0.54	0.49	0.34	21	0.50	0.40	0.33	0.34	0.43	0.64	0.33
7	0.58	0.67	0.42	0.56	0.66	0.68	0.35	22	0.64	0.61	0.52	0.62	0.66	0.80	0.41
8	0.71	0.75	0.50	0.55	0.70	0.66	0.41	23	0.53	0.66	0.52	0.50	0.63	0.80	0.37
9	0.72	0.82	0.72	0.67	0.71	0.83	0.31	24	0.40	0.37	0.42	0.58	0.50	0.57	0.49
10	0.67	0.61	0.45	0.47	0.62	0.80	0.41	25	0.63	0.54	0.42	0.48	0.66	0.75	0.33
11	0.64	0.53	0.53	0.58	0.58	0.67	0.34	26	0.66	0.77	0.66	0.63	0.88	0.76	0.72
12	0.67	0.60	0.47	0.39	0.59	0.74	0.41	27	0.78	0.75	0.58	0.74	0.80	0.78	0.49
13	0.69	0.62	0.57	0.42	0.55	0.63	0.25	28	0.48	0.57	0.44	0.45	0.51	0.83	0.38
14	0.68	0.83	0.83	0.45	0.59	0.77	0.35	29	0.85	0.85	0.71	0.71	0.77	0.74	0.55
15	0.77	0.77	0.54	0.72	0.79	0.77	0.46	30	0.82	0.82	0.39	0.59	0.64	0.78	0.39

[1] Chatterjee, Hadi and Price (2000)

Ejemplo: Desempeño de Supervisores (cont.)

Considere el modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i, \quad i = 1, \ldots, 30$$

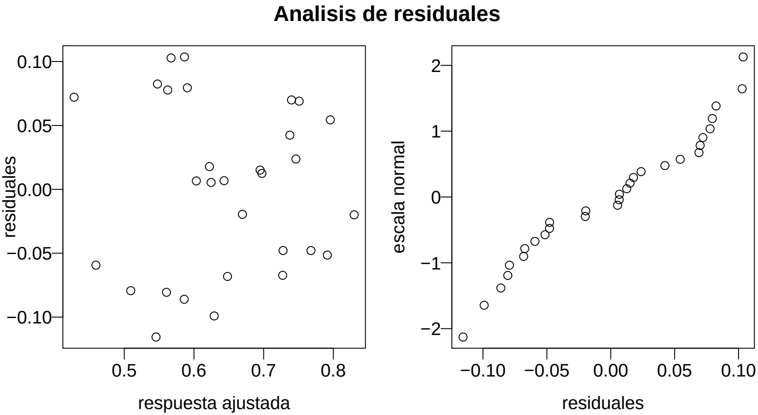
Modelo ajustado: $\hat{y} = 0.099 + 0.644x_1 + 0.211x_3$

$$s = 0.06817$$

n	y	$\hat{y}$	$\hat{e}$	n	y	$\hat{y}$	$\hat{e}$
1	0.43	0.51	−0.08	16	0.81	0.83	−0.02
2	0.63	0.62	0.01	17	0.74	0.79	−0.05
3	0.71	0.69	0.02	18	0.65	0.64	0.01
4	0.61	0.60	0.01	19	0.65	0.67	−0.02
5	0.81	0.74	0.07	20	0.50	0.59	−0.09
6	0.43	0.55	−0.12	21	0.50	0.43	0.07
7	0.58	0.65	−0.07	22	0.64	0.62	0.02
8	0.71	0.70	0.01	23	0.53	0.63	−0.10
9	0.72	0.77	−0.05	24	0.40	0.46	−0.06
10	0.67	0.59	0.08	25	0.63	0.55	0.08
11	0.64	0.56	0.08	26	0.66	0.73	−0.07
12	0.67	0.57	0.10	27	0.78	0.74	0.04
13	0.69	0.59	0.10	28	0.48	0.56	−0.08
14	0.68	0.73	−0.05	29	0.85	0.80	0.05
15	0.77	0.75	0.02	30	0.82	0.75	0.07

−ebz

Ejemplo: Desempeño de Supervisores (cont.)



−ebz

Modelo de Regresión Lineal Múltiple

Modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \qquad i = 1, \ldots, n$$

donde,

y_i	:	respuesta al nivel i-ésimo nivel del regresor ($i = 1, \ldots, n$)
x_{ij}	:	j-ésimo regresor a su i-ésimo nivel ($j = 1, \ldots, p$)
β_0	:	ordenada al origen
β_j	:	coeficiente (tasa de cambio) del j-ésimo regresor

O bien, si $x_i = (1, x_{i1}, \ldots, x_{ip})'$ y $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$,

$$y_i = x_i'\beta + \epsilon_i, \quad i = 1, \ldots, n$$

$$y = X\beta + \epsilon$$

Supuestos

$$\text{rango}(X) = p + 1 = q\ (\leq n)$$

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2)\ \textit{i.i.d.}$$

Se pide que la matriz X sea de rango completo. Esto es, que tenga sus q columnas linealmente independientes.

Regresión lineal múltiple

En general, un modelo de *regresión lineal múltiple* es una aproximación a la variable de respuesta y como tal es un *modelo empírico*.
Ejemplos:

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\
y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \\
y &= \alpha + \beta P + \gamma T + \delta L + \epsilon \\
y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon
\end{aligned}
$$

−ebz

Matricialmente, si

$$y_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}; \quad \epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}; \quad \beta_{q\times 1} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}; \quad X_{n\times q} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

entonces el modelo de *regresión lineal múltiple* se puede expresar como

$$y = X\beta + \epsilon$$

con los supuestos:

$$\text{rango}(X) = p + 1 = q \ (\leq n)$$

$$\epsilon \sim \mathrm{N}_n(0, \sigma^2 I_n)$$

−ebz

Mínimos Cuadrados

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \sum (y_i - x_i' \beta)^2 \\ &= ||y - X\beta||^2 = (y - X\beta)'(y - X\beta) = y'y - 2y'X\beta + \beta' X' X \beta \end{aligned}$$

Problema de Mínimos Cuadrados:

$$\min_{\beta} S(\beta) \equiv \min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 \equiv \min_{\beta_0,\ldots,\beta_p} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 \right\}$$

$$\frac{\partial S}{\partial \beta_j} = 0 \Longrightarrow \left\{ \begin{array}{rcl} 2\sum (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^1 (-1) & = & 0 \\ \vdots & & \vdots \\ 2\sum (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p c_{ip})^1 (-x_{ip}) & = & 0 \end{array} \right.$$

que da lugar a las *ecuaciones normales* (q ecuaciones y q incógnitas) :

Ecuaciones normales

Ecuaciones normales

$$
\begin{array}{ccccccccc}
n\beta_0 & + & \beta_1 \sum x_{i1} & + & \cdots & + & \beta_p \sum x_{ip} & = & \sum y_i \\
\beta_0 \sum x_{i1} & + & \beta_1 \sum x_{i1}^2 & + & \cdots & + & \beta_p \sum x_{i1} x_{ip} & = & \sum x_{i1} y_i \\
\vdots & & \vdots & & \ddots & & \vdots & & \vdots \\
\beta_0 \sum x_{ip} & + & \beta_1 \sum x_{i1} x_{ip} & + & \cdots & + & \beta_p \sum x_{ip}^2 & = & \sum x_{ip} y_i
\end{array}
$$

cuya solución $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ son los *estimadores de mínimos cuadrados (*EMC*)*.

Mínimos cuadrados y ecuaciones normales

Matricialmente el problema es

$$\min_{\beta} S(\beta) \equiv \min_{\beta} \left\{ y'y - 2y'X\beta + \beta'X'X\beta \right\}$$

$$0 = \frac{\partial S}{\partial \beta} = -2X'y + 2X'X\beta \Longrightarrow X'X\beta = X'y$$

Ecuaciones normales

$$X'X\beta = X'y$$

Esto es,

$$\begin{bmatrix} n & \sum x_{i1} & \cdots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \cdots & \sum x_{i1}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ip} & \sum x_{i1}x_{ip} & \cdots & \sum x_{ip}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{ip}y_i \end{bmatrix}$$

Puesto que rango$(X) = q$, se tiene que rango$(X'X) = q$ y por lo tanto $X'X$ es invertible. Entonces,

Estimadores de mínimos cuadrados (EMC)

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

−ebz

El valor de la respuesta y_u ajustado al nivel $x_u = (1, x_{u1}, \ldots, x_{up})'$ del vector de regresores está dado por:

$$\hat{y}_u = x_u'\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_{u1} + \cdots + \hat{\beta}_p x_{up}$$

El vector de la respuesta ajustada $\hat{y}$ es

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_1'\hat{\beta} \\ \vdots \\ x_n'\hat{\beta} \end{bmatrix} = X\hat{\beta} = X(X'X)^{-1}X'y$$

$$\hat{y} = [X(X'X)^{-1}X']y = Hy$$

donde $H = X(X'X)^{-1}X'$ es la *matriz gorro* o *matriz sombrero* (*"hat matrix"*). Nuevamente, los *residuales* $\hat{e}$ están dados por

$$\begin{aligned} \hat{e} &= y - \hat{y} = y - X\hat{\beta} \\ &= y - Hy = (I - H)y \\ &= My \end{aligned}$$

con $M = (I - H) = (I - X(X'X)^{-1}X')$.

−ebz

Ejemplo: Desempeño de Supervisores (cont.)

Considere el modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i, \quad i = 1, \ldots, 30$$

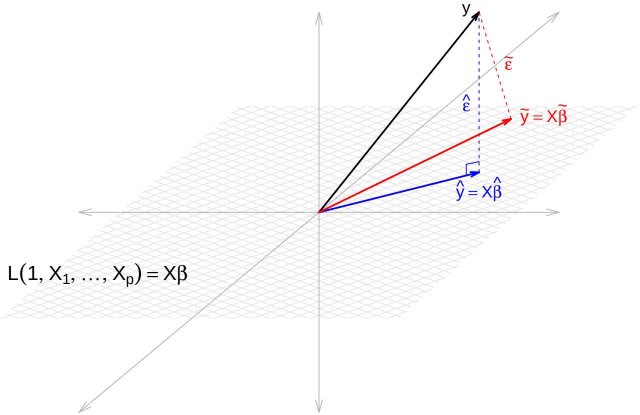
$$X'X = \begin{bmatrix} 30.00 & 19.98 & 16.91 \\ 19.98 & 13.82 & 11.53 \\ 16.91 & 11.53 & 9.93 \end{bmatrix}; \; X'y = \begin{bmatrix} 19.39 \\ 13.30 \\ 11.19 \end{bmatrix}$$

$$\begin{bmatrix} 30.00 & 19.98 & 16.91 \\ 19.98 & 13.82 & 11.53 \\ 16.91 & 11.53 & 9.93 \end{bmatrix}^{-1} \begin{bmatrix} 19.39 \\ 13.30 \\ 11.19 \end{bmatrix} = \begin{bmatrix} 0.099 \\ 0.644 \\ 0.211 \end{bmatrix} = \hat{\beta}$$

Modelo ajustado:

$$\hat{y} = 0.099 + 0.644 x_1 + 0.211 x_3$$

Representación Geométrica de Mínimos Cuadrados



Interpretación geométrica: $\hat{y}$ es la proyección ortogonal sobre el plano generado por los regresores.

Ejemplo: Servicio de televisión por cable [2]

	Variable	Descripción
1	Colonia	Colonia a la que pertenece el hogar de la zona residencial
2	Manzana	Número de manzana a la que pertenece el hogar
3	Adultos	Número de adultos por hogar
4	Niños	Número de niños menores de 12 años por hogar
5	Teles	Número de televisores por hogar
6	Tipo	Tipo de televisor que posee: blanco y negro (B), color (C), ambos (A)
7	TVtot	Suma del número de horas frente al televisor en la semana de todos los miembros de la familia
8	Renta	Cantidad máxima de renta que el jefe del hogar estaría dispuesto a pagar al mes por servicio de TV por cable
9	Valor	Valor catastral del hogar (m$). La respuesta se usa para dar idea aproximada del ingreso familiar

obs.	colonia	manzana	adultos	niños	teles	renta	tvtot	tipo	valor
1	2	20	3	2	2	50	68	B	79928
2	2	25	3	3	1	65	82	B	94415
3	2	20	1	2	1	45	40	A	120896
4	2	8	2	2	2	35	56	A	132867
5	2	25	1	2	0	0	0	N	141901
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
36	1	2	2	0	2	60	20	A	332699
37	1	2	3	0	3	70	28	C	336290
38	1	9	3	0	5	85	28	C	355641
39	1	9	2	0	3	70	20	C	357972
40	1	4	3	0	4	80	28	C	370325

[2] Aguirre et al. 2006.

Ejemplo: Servicio de televisión por cable (cont.)

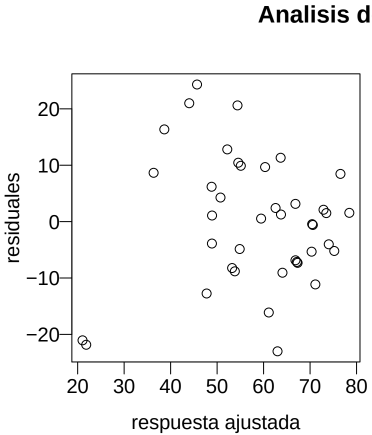
```
Response: renta
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.80566   10.32633   0.950 0.348838
ninos          -4.91432    2.73477  -1.797 0.080973
adultos         2.64006    2.44211   1.081 0.287065
tvtot           0.45053    0.11445   3.936 0.000375
I(valor/1000)   0.12989    0.03141   4.135 0.000211

Residual standard error: 11.99 on 35 degrees of freedom
Multiple R-squared: 0.5916,     Adjusted R-squared: 0.545
F-statistic: 12.68 on 4 and 35 DF,  p-value: 1.772e-06


Analysis of Variance Table

               Df    Sum Sq Mean Sq F value     Pr(>F)
SCReg           4    7293.4  1823.4 12.6788 1.768e-06
SCRes          35    5034.2   143.8
SCTot          39   12327.6   316.1
```

Ejemplo: Servicio de televisión por cable (cont.)



¿Tendencia? ¿Datos atípicos?

Propiedades de los EMC

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X'X)^{-1}X'y] = \mathbb{E}[(X'X)^{-1}X'(X\beta + \epsilon)] \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mathbb{E}[\epsilon] \\
&= \beta \\
\mathrm{cov}(\hat{\beta}) &= \mathrm{cov}((X'X)^{-1}X'y) \\
&= (X'X)^{-1}X'\mathrm{cov}(y)X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1} \\
&= \sigma^2 C
\end{aligned}
$$

con $C = (X'X)^{-1}$. Así,

$$
\begin{aligned}
\mathrm{var}(\hat{\beta}_j) &= \sigma^2 C_{jj} \\
\mathrm{cov}(\hat{\beta}_j, \hat{\beta}_k) &= \sigma^2 C_{jk}
\end{aligned}
$$

Propiedades de los EMC

Por otro lado,

$$\begin{aligned}\sum \hat{e}_i^2 &= ||\hat{e}||^2 = \hat{e}'\hat{e} \\ &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - y'X\hat{\beta}\end{aligned}$$

pues por las ecuaciones normales $X'X\beta = X'y$. Entonces,

$$S(\hat{\beta}) = ||\hat{e}||^2 = y'y - \hat{\beta}'X'y = (y - X\hat{\beta})'y$$

Se puede mostrar que

$$s^2 = \frac{\text{SC}_{\text{Res}}}{n-q} = \text{CM}_{\text{Res}}$$

es un estimador insesgado de σ^2, aunque dependiente del modelo.

Intervalos de Confianza para los coeficientes β_j

El hecho que $y = X\beta + \epsilon$, con $\epsilon \sim \mathrm{N}(0, \sigma^2 I)$ implica que $y \sim \mathrm{N}_n(X\beta, \sigma^2 I)$, y siendo $\hat{\beta} = (X'X)^{-1}X'y$, $\hat{\beta}$ es una transformación lineal de un vector normal multivariado. Entonces,

$$\hat{\beta} \sim \mathrm{N}_q\left(\beta, \sigma^2 (X'X)^{-1}\right)$$

y en particular, $\mathrm{var}(\hat{\beta}_j) = \sigma^2 C_{jj}$, y $\mathrm{ee}(\hat{\beta}_j) = \sigma\sqrt{C_{jj}}$, donde $C_{jj} = (X'X)^{-1}_{jj}$.

Ahora bien, sean $q = p + 1$ y $\nu = n - q$ los grados de libertad de los residuales. Se sigue entonces que $\frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu$, independientemente de $\hat{\beta}$ y se tiene que

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{C_{jj}}} \sim t_\nu$$

Por lo tanto, un intervalo del $100(1-\alpha)$ % de confianza para β_j estaría dado por

$$\hat{\beta}_j \pm t_{(1-\alpha/2,\nu)} s\sqrt{C_{jj}}$$

Una región de *confianza conjunta* del $100(1-\alpha)$ % para el vector de parámetros $\beta = (\beta_0, \ldots, \beta_p)'$, de acuerdo a Bonferroni, estaría dada por el producto de los intervalos del $100(1-\alpha/q)$ % de confianza marginales. Esto es,

$$\hat{\beta}_j \pm t_{(1-\alpha/2q,\nu)} s\sqrt{C_{jj}}, \quad j = 0, \ldots, p;\ (q = p + 1)$$

−ebz

Intervalos de confianza y de predicción para la respuesta $\hat{y}(x)$

Considere el regresor al nivel $x_0 = (1, x_{10}, \ldots, x_{p0})'$. La respuesta media ajustada correspondiente es $\hat{y}_0 = \hat{y}(x_0) = x_0'\hat{\beta}$, con

$$\text{var}(\hat{y}_0) = \text{var}(x_0'\hat{\beta}) = x_0'\text{var}(\hat{\beta})x_0 = \sigma^2 x_0'(X'X)^{-1}x_0$$

Por lo tanto, un intervalo del $100(1-\alpha)\,\%$ de confianza para la respuesta media $\hat{y}$ al nivel x_0 esta dado por

$$\hat{y}(x_0) \pm t_{(1-\alpha/2,\nu)} s \sqrt{x_0'(X'X)^{-1}x_0}$$

Similarmente, para el mismo nivel x_0, el correspondiente intervalo de *predicción* para una respuesta nueva $\hat{y}_0 + \epsilon$ es

$$\hat{y}(x_0) \pm t_{(1-\alpha/2,\nu)} s \sqrt{1 + x_0'(X'X)^{-1}x_0}$$

puesto que

$$\text{var}(\hat{y}(x_0) + \epsilon) = \text{var}(x_0'\hat{\beta} + \epsilon) = \sigma^2 x_0'(X'X)^{-1}x_0 + \sigma^2 = \sigma^2\left(x_0'(X'X)^{-1}x_0 + 1\right)$$

por independencia de la nueva observación.

Prueba de Hipótesis

Considere el modelo de regresión

$$y = X\beta + \epsilon = X_1 B_1 + X_2 B_2 + \epsilon \tag{MC}$$

donde

$$B_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_r \end{bmatrix}_{r\times 1} \; ; \; B_2 = \begin{bmatrix} \beta_{r+1} \\ \vdots \\ \beta_p \end{bmatrix}_{q-r\times 1} \; ; \; \beta = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}_{q\times 1}$$

y suponga que se desea contrastar $\mathrm{H}_0 : B_2 = 0$ *vs.* $\mathrm{H}_1 : B_2 \neq 0$

Entonces para el *modelo completo* (MC)

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ \mathrm{SC}_{\mathrm{Reg}}(B_1, B_2) &= \hat{\beta}'X'y \\ \mathrm{CM}_{\mathrm{Res}} &= \left(y'y - \hat{\beta}'X'y\right)/(n-q) \end{aligned}$$

Por otro lado, considere el *modelo reducido* (MR)

$$y = X_1 B_1 + \epsilon \tag{MR}$$

luego

$$\begin{aligned} \hat{B}_1 &= (X_1'X_1)^{-1}X_1'y \\ \mathrm{SC}_{\mathrm{Reg}} &= \hat{B}_1'X_1'y \\ \mathrm{CM}_{\mathrm{Res}} &= \left(y'y - \hat{B}_1'X_1'y\right)/(n-r) \end{aligned}$$

–ebz

Entonces, la *suma de cuadrados extra* debida a los regresores X_2 dado que los regresores X_1 están en el modelo sería

$$
\begin{aligned}
\mathrm{SC}_{\mathrm{Reg}}(B_2|B_1) &= \mathrm{SC}_{\mathrm{Reg}}(B_1, B_2) - \mathrm{SC}_{\mathrm{Reg}}(B_1) \\
&= [\mathrm{SC}_{\mathrm{Tot}} - \mathrm{SC}_{\mathrm{Res}}(\beta_1, \beta_2)] - [\mathrm{SC}_{\mathrm{Tot}} - \mathrm{SC}_{\mathrm{Res}}(\beta_1)] \\
&= \mathrm{SC}_{\mathrm{Res}}(\beta_1) - \mathrm{SC}_{\mathrm{Res}}(\beta_1, \beta_2) \\
&= \mathrm{SC}_{\mathrm{Res}}(\mathrm{MR}) - \mathrm{SC}_{\mathrm{Res}}(\mathrm{MC})
\end{aligned}
$$

Grados de libertad:

$$
q_2 = (n-r) - (n-q) = q - r
$$

Entonces, para contrastar $\mathrm{H}_0 : B_2 = 0$ *vs.* $\mathrm{H}_1 : B_2 \neq 0$, se utiliza el estadístico

$$
\begin{aligned}
F &= \frac{\mathrm{SC}_{\mathrm{Reg}}(B_2|B_1)/q_2}{\mathrm{SC}_{\mathrm{Res}}(\mathrm{MC})/(n-q)} \\
&= \frac{\left[\left(y'y - \hat{\beta}'X'y\right) - \left(y'y - \hat{B}_1'X_1'y\right)\right]/(q-r)}{\mathrm{SC}_{\mathrm{Res}}(\beta)/(n-q)} \\
&= \frac{[\mathrm{SC}_{\mathrm{Res}}(B_1) - \mathrm{SC}_{\mathrm{Res}}(B1, B2)]/(q-r)}{\mathrm{SC}_{\mathrm{Res}}(B_1, B_2)/(n-q)} \\
&= \frac{\mathrm{CM}_{\mathrm{Extra}}}{\mathrm{CM}_{\mathrm{Res}}(MC)} \sim F_{q_2, n-q}
\end{aligned}
$$

−ebz

Ejemplo: Servicio de televisión por cable (cont.)

```
Response: renta
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.80566   10.32633   0.950 0.348838
ninos          -4.91432    2.73477  -1.797 0.080973
adultos         2.64006    2.44211   1.081 0.287065
tvtot           0.45053    0.11445   3.936 0.000375
I(valor/1000)   0.12989    0.03141   4.135 0.000211

Residual standard error: 11.99 on 35 degrees of freedom
Multiple R-squared: 0.5916,     Adjusted R-squared: 0.545
F-statistic: 12.68 on 4 and 35 DF,  p-value: 1.772e-06


Analysis of Variance Table
              Df Sum Sq Mean Sq F value    Pr(>F)
ninos          1  793.1   793.1  5.5138 0.0246456
adultos        1 2198.8  2198.8 15.2873 0.0004048
tvtot          1 1841.7  1841.7 12.8041 0.0010366
I(valor/1000)  1 2459.8  2459.8 17.1014 0.0002106
Residuals     35 5034.2   143.8
```

–ebz

Ejemplo: Servicio de televisión por cable (cont.)

```
Response: renta
           Df Sum Sq Mean Sq F value    Pr(>F)
ninos       1    793  793.08  5.5138 0.0246456
adultos     1   2199 2198.82 15.2873 0.0004048
tvtot       1   1842 1841.66 12.8041 0.0010366
vc          1   2460 2459.76 17.1014 0.0002106
Residuals  35   5034  143.83
TotCorr    39  12328
Correction  1 135722
Total      40 148050
```

```
Suma de Cuadrados secuencial (Tipo I)
Model 1: renta ~ 1
Model 2: renta ~ ninos
Model 3: renta ~ ninos + adultos
Model 4: renta ~ ninos + adultos + tvtot
Model 5: renta ~ ninos + adultos + tvtot + vc
  Res.Df     RSS Df Sum of Sq       F    Pr(>F)
1     39 12327.5
2     38 11534.4  1    793.08  5.5138 0.0246456
3     37  9335.6  1   2198.82 15.2873 0.0004048
4     36  7493.9  1   1841.66 12.8041 0.0010366
5     35  5034.2  1   2459.76 17.1014 0.0002106
```

−ebz

Igualdad de parámetros

Suponga que se tiene el siguiente *modelo completo*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \tag{MC}$$

y se desea probar la siguiente *hipótesis compuesta*

$$\mathrm{H}_0 : \beta_1 = \beta_2 \text{ y } \beta_3 = \beta_4 \quad \textit{vs.} \quad \mathrm{H}_1 : \beta_1 \neq \beta_2 \text{ o } \beta_3 \neq \beta_4$$

Entonces bajo H_0, el modelo se puede reexpresar como el siguiente *modelo reducido*

$$y = \beta_0 + \beta_1(x_1 + x_2) + \beta_3(x_3 + x_4) + \epsilon \tag{MR}$$

y se contrasta H_0 con el estadístico de prueba

$$F = \frac{\left[\mathrm{SC}_{\mathrm{Res}}(\mathrm{MR}) - \mathrm{SC}_{\mathrm{Res}}(\mathrm{MC})\right]/(5-3)}{\mathrm{SC}_{\mathrm{Res}}(\mathrm{MC})/(n-5)} \sim F_{2,n-5}$$

Si $F > F_{1-\alpha/2}$, se rechaza H_0 con una significancia α y se ha de trabajar con el modelo completo MC.

Ejemplo: Producción de un reactor [3]

Se han realizado una serie de ensayos para investigar la influencia de algunos factores *físicos* y *químicos* en la producción de cierto reactor. Las variables consideradas se muestran a continuación:

	variable	tipo
y	producción	respuesta
x_1	voltaje (volts)	físico
x_2	tiempo (min)	físico
x_3	agitación (rpm)	químico
x_4	relación entre reactivos (%)	químico

n	x1	x2	x3	x4	y	n	x1	x2	x3	x4	y
1	150	75	260	0.35	70	9	150	75	260	0.65	60
2	250	75	260	0.35	60	10	250	75	260	0.65	49
3	150	105	260	0.35	89	11	150	105	260	0.65	88
4	250	105	260	0.35	81	12	250	105	260	0.65	82
5	150	75	340	0.35	69	13	150	75	340	0.65	60
6	250	75	340	0.35	62	14	250	75	340	0.65	52
7	150	105	340	0.35	88	15	150	105	340	0.65	86
8	250	105	340	0.35	81	16	250	105	340	0.65	79

[3] Box, Hunter and Hunter (1978)

Ejemplo: Producción de un reactor (cont.)

Modelo Completo:

Incluye variables *físicas* y *químicas*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	256.00	256.00	28.231	<2e-16
x2	1	2304.00	2304.00	254.075	<2e-16
x3	1	0.25	0.25	0.028	0.871
x4	1	121.00	121.00	13.343	0.004
Resid	11	99.75	9.07		

Modelo Reducido:

Incluye solamente variables *físicas*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	256	256	15.059	0.002
x2	1	2304	2304	135.529	<2e-16
Resid	13	221	17		

Suma extra de cuadrados:

$$\begin{aligned} F &= \frac{\left[SC_{\text{Res}}(MR) - SC_{\text{Res}}(MC)\right]/(5-3)}{SC_{\text{Res}}(MC)/(n-5)} \\ &= \frac{[221.00 - 99.75]/(5-3)}{99.75/(16-5)} = \frac{121.25/2}{99.75/11} \\ &= 6.685 \ (p = 0.013) \end{aligned}$$

```
Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + x3 + x4

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     13 221.00
2     11  99.75  2    121.25 6.6855 0.01259
```

Conclusión: Los datos muestran ($p = 0.013$) que las variables *químicas* también son importantes en la explicación del producción del reactor.

−ebz

Condiciones y Teorema Gauss-Markov

Propiedades deseables de los estimadores se pueden mostrar a partir de *Condiciones de Gauss–Markov*:

$$\left.\begin{array}{rcl} \mathbb{E}[\epsilon_i] & = & 0 \\ \mathrm{var}(\epsilon_i) & = & \sigma^2 \\ \mathrm{cov}(\epsilon_i, \epsilon_j) & = & 0 \end{array}\right\} \quad \begin{array}{rcl} \mathbb{E}[\epsilon] & = & 0 \\ \mathrm{cov}(\epsilon) & = & \sigma^2 I_n \end{array}$$

En muchas de las aplicaciones de la regresión lineal múltiple, estamos interesados en estimaciones de funciones lineales de β. E. g., $\ell'\beta$, o bien, $L\beta$, donde ℓ y L son vector y matriz respectivamente. Por ejemplo,

$$\hat{y}_i = x_i'\hat{\beta}, \quad \hat{y} = X\hat{\beta}, \quad \text{o incluso} \quad \hat{\beta} = I\hat{\beta}$$

Teorema Gauss-Markov

Considere el modelo $y = X\beta + \epsilon$ y sea $\hat{\beta} = (X'X)^{-1}X'y$. Bajo las condiciones de Gauss-Markov, el estimador de mínimos cuadrados, $\ell'\hat{\beta}$, de la función estimable $\ell'\beta$, tiene varianza mínima entre todos los estimadores lineales insesgados de $\ell'\beta$. $\ell'\hat{\beta}$ es BLUE, *"best linear unbiased estimator"*.

Esto es,

$$\mathrm{var}(\ell'\hat{\beta}) \leq \mathrm{var}(d'y), \quad \forall\ d'y \ \ni\ \mathbb{E}[d'y] = \ell'\beta$$

para todo estimador lineal $d'y$ que sea estimador insesgado de $\ell'\beta$.

−ebz