Apuntes Matemáticas II. BQ y BIO: Tema 7 Regresión.

15 de mayo de 2012

Introducción

Consideremos los pares de observaciones de dos variables:

$$\{(x_i, y_i)|i = 1, 2, \ldots, n\}$$

- La variable y es la variable dependiente o de respuesta.
- La variable x es la variable de control o independiente o de regresión.
- El problema que se intenta resolver es encontrar la mejor relación funcional que explique la variable y conocido el valor de la variable x: Y/x. En nuestro caso esta función será una recta.

Ejemplo

Ejemplo: Consideremos los datos siguientes donde x representa los meses e y representa el crecimiento de un determinado tipo de planta en mm.

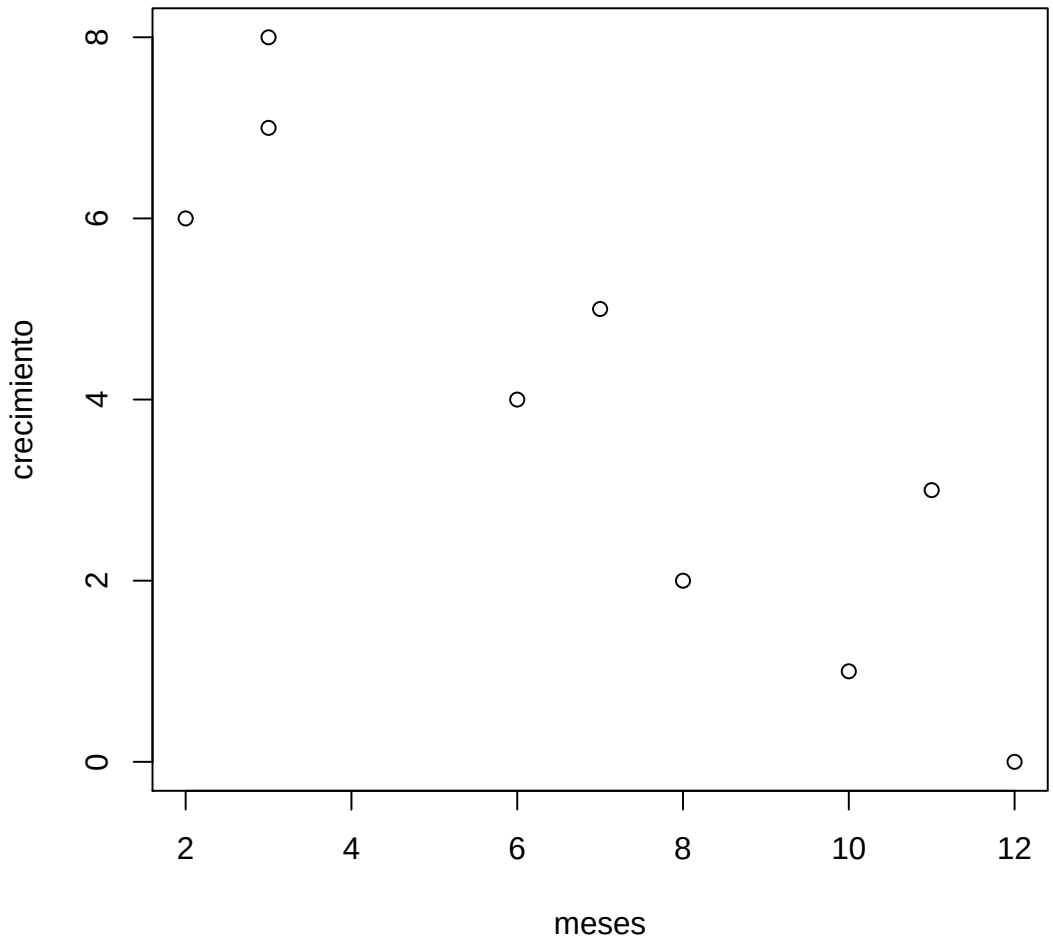
Meses	Crecimiento
12	0
10	1
8	2
11	3
6	4
7	5
2	6
3	7
3	8

El código R para introducir los datos y hacer un gráfico es:

```
> meses=c(12,10,8,11,6,7,2,3,3)
> crecimiento=c(0,1,2,3,4,5,6,7,8)
> plot(meses,crecimiento)
```

Ejemplo

```
> plot(meses,crecimiento)
```



El modelo de Regresión lineal simple

En realidad, en un análisis más riguroso, el modelo de regresión lineal es el siguiente:

$$\mu_{Y/x} = \beta_0 + \beta_1 x,$$

donde $\mu_{Y/x}$ es el valor esperado que toma la variable y cuando la variable de control vale x, mientras que β_0 (término independiente) y β_1 (pendiente) son dos parámetros a determinar. Dada una muestra calcularemos las estimaciones b_0 y b_1 de β_0 y de β_1 respectivamente. Notemos que para muestras diferentes, las estimaciones serán diferentes. Una vez obtenidas las estimaciones podemos calcular la recta de regresión estimada, que es:
$\hat{y} = b_0 + b_1 x.$

Regresión lineal simple por Mínimos cuadrados

- Existen diversas maneras de calcular las estimaciones de los coeficientes de una regresión lineal: Regresión ortogonal, métodos robustos, regresión mínimo cuadrática o de mínimos cuadrados,... Nosotros optaremos por el método más habitual que es el de mínimos cuadrados (m.c.).
- Modelo

$$Y_i = \beta_0 + \beta_1 x_i + E_i,$$

  Donde E_i es una nueva variable llamada error o residuo.
- Una vez planteado el modelo y dada una muestra el modelo se debe ajustar a los datos de ésta:

$$y_i = \beta_0 + \beta_1 x_i + E_i, \text{ para } i = 1, 2, \ldots n.$$

Regresión lineal simple por Mínimos cuadrados

- Cuando ajustamos por las estimaciones b_0 y b_1 obtenemos la recta de regresión ajustada

$$\hat{y} = b_0 + b_1 x.$$

- Podemos calcular para cada par de observaciones:

$$y_i = b_0 + b_1 x_i + e_i, \qquad \hat{y}_i = b_0 + b_1 x_i \text{ para } i = 1, 2, \ldots n.$$

- Entonces el error o residuo de la i-ésima observación, $i = 1, 2, \ldots, n$, es

$$e_i = y_i - \hat{y}_i.$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes

Cálculo de b_0 y b_1 por M.C.

- Los valores de b_0 y b_1 buscados son los que minimizan el error cuadrático:

$$SSE = \sum_{i=1}^{n} e_i^2.$$

- Estos valores serán los estimadores de β_0 y β_1 por el método de mínimos cuadrados.

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes

- En primer lugar tenemos que:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2.$$

- Calculando las derivadas parciales respecto a b_0 y a b_1, e igualando a cero:

$$\frac{\partial SSE}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0,$$

$$\frac{\partial SSE}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0.$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes. Ecuaciones normales

- Las ecuaciones anteriores reciben el nombre de ecuaciones normales:

$$\left.\begin{aligned} nb_0 + \sum_{i=1}^{n} x_i b_1 &= \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i b_0 + \sum_{i=1}^{n} x_i^2 b_1 &= \sum_{i=1}^{n} x_i y_i \end{aligned}\right\}$$

- Las soluciones de estas ecuaciones son:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \qquad b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}.$$

Ejemplo

En el ejemplo anterior, tenemos:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{9 \cdot 175 - 62 \cdot 36}{9 \cdot 536 - 62^2} = -0.6704,$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \frac{36 - (-0.6704) \cdot 62}{9} = 8.6184.$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes. Definición de los momentos de primer y segundo orden.

- Definimos las medias y varianzas de las variabes x e y como:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n},$$

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2,$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n} = \frac{\sum_{i=1}^{n} y_i^2}{n} - \overline{y}^2,$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes. Definición de los momentos de primer y segundo orden.

- Definimos la covarianza entre las variabes x e y como:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{x} \cdot \overline{y}.$$

Ejemplo anterior

- Los momentos de primer orden son para el ejemplo anterior:

$$\overline{x} = \frac{62}{9} = 6.89, \quad \overline{y} = \frac{36}{9} = 4.$$

- Los momentos de segundo orden son:

$$s_x^2 = \frac{536}{9} - 6.89^2 = 12.09877, s_y^2 = \frac{204}{9} - 4^2 = 6.67,$$

$$s_{xy} = \frac{175}{9} - 6.89 \cdot 4 = -8.111111.$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes en función de los momentos de primer y segundo orden

- Los coeficientes de la recta de regresión son en función de las medias, varianzas y covarianza entre las variables x e y:

$$b_1 = \frac{s_{xy}}{s_x^2}, \quad b_0 = \overline{y} - b_1 \overline{x}.$$

- Ejemplo anterior:

$$b_1 = \frac{-8.11}{12.09877} = -0.6704,$$

$$b_0 = 4 - (-0.6704) \cdot 6.89 = 8.6184.$$

Propiedades de los estimadores

- La recta de regresión pasa por el vector de medias $(\overline{x}, \overline{y})$, es decir:

$$b_0 + b_1\overline{x} = \overline{y}$$

- La media de los valores estimados es igual a la media de los observados

$$\overline{\hat{y}} = \frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \overline{y}$$

Ejemplo anterior

- Comprobemos que la recta de regresión pasa por el vector de medias que será $(6.89, 4)$:

$$\hat{y} = b_0 + b_1 x, \text{ si } x = 6.89, \text{ queda}$$
$$8.6184 - 0.6704 \cdot 6.89 \approx 4.$$

- Veamos que la media de los valores estimados es la misma que los valores observados; o sea, 4. Los valores estimados son:

$$0.57347,\ 1.91429,\ 3.25510,\ 1.24388,\ 4.59591,$$
$$3.92551,\ 7.27755,\ 6.60714,\ 6.60714$$

Si hallamos la media, vale efectivamente 4.

Consideraciones sobre el modelo de regresión lineal

- Se supone que los errores del modelo E_i tienen una distribución normal de media 0 y desviación típica σ.
- Los errores de la estimación por mínimos cuadrados tienen media 0. Efectivamente,

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0.$$

En conclusión

$$\overline{e} = \frac{\sum_{i=1}^{n} e_i}{n} = 0,$$

y

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n} = \frac{SSE}{n}.$$

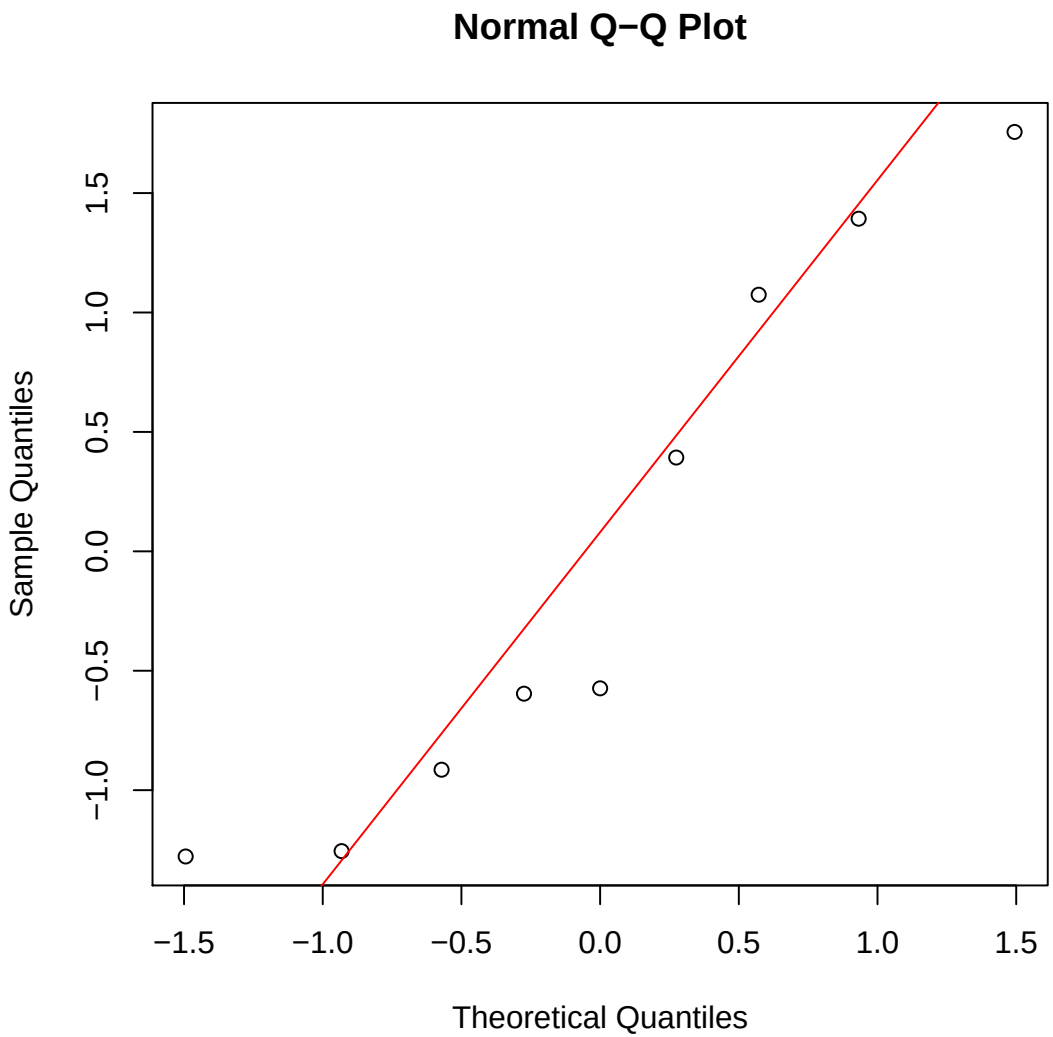
Ejemplo anterior

Para comprobar la normalidad se realiza un QQ-plot. Los errores son los siguientes:

$$-0.57347,\ -0.91429,\ -1.25510,\ 1.75612,\ -0.59591,\ 1.07449,\ -1.27755,\ 0.39286,\ 1.3928571$$

El código para el qq-plot es

```
> errores=c(-0.57347, -0.91429,-1.25510,1.75612,-0.59591,
+  1.07449,  -1.27755, 0.39286,1.3928571)
> qqnorm(errores)
> qqline(errores,col="red")
```



A medida de que los puntos se aproximan a la recta (diagonal del primer cuadrante) más se aproximan a estar normalmente distribuidos.

Definición de las sumas de cuadrados

- Llamaremos suma de cuadrados de los residuales o del error a

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- Llamaremos suma de cuadrados de totales a

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

- Llamaremos suma de cuadrados de la regresión a

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2.$$

Relación entre las sumas de cuadrados

- En una regresión lineal por el método de mínimos cuadrados se tiene que:

$$SST = SSR + SSE.$$

- La expresión anterior es equivalente a

$$S_y^2 = S_{\hat{y}}^2 + S_e^2.$$

Ejemplo anterior

- Las sumas de cuadrados son:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 11.06020,$$
$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = 48.9398,$$
$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 = 60.$$

Se puede comprobar que se cumple la igualdad $SST = SSE + SSR$.

El coeficiente de determinación R^2 y la estimación de la varianza.

- Se define como
$$R^2 = \frac{SSR}{SST}.$$
- En el caso de regresión lineal m.c. se cumple que: $R^2 = \frac{S^2_{\hat{y}}}{S^2_y}$, $R^2 = 1 - \frac{SSE}{SST}$, $R^2 = 1 - \frac{S^2_e}{S^2_y}$.
- $R^2 = r^2_{xy}$, donde $r_{xy} = \frac{s_{xy}}{s_x s_y}$.
- Por lo tanto R^2 es la proporción de varianza de la variable y que queda explicada por la regresión lineal.
- Una estimación insesgada de σ^2 (la varianza del error E) en m.c. es
$$S^2 = \frac{SSE}{n-2}.$$

Ejemplo anterior

- El coeficiente R^2 valdrá: $R^2 = \frac{48.9398}{60} = 0.8157$. Por tanto, se explica el $81.57\,\%$ de la varianza del crecimiento de la planta.
- Estimación insesgada de la varianza:

$$S^2 = \frac{SSE}{n-2} = \frac{11.06020}{7} = 1.5800.$$

Intervalos de confianza

- Suponemos de que los residuos siguen una ley normal.
- Intervalo de confianza al nivel $(1-\alpha)100\,\%$ para el parámetro β_1: $(\mu_{Y/x} = \beta_0 + \beta_1 x)$

$$b_1 - \frac{t_{n-2,1-\alpha/2} S}{\sqrt{n S_x^2}} < \beta_1 < b_1 + \frac{t_{n-2,1-\alpha/2} S}{\sqrt{n S_x^2}}$$

- Intervalo de confianza al nivel $(1-\alpha)100\,\%$ para el parámetro β_0:

$$b_0 - \frac{t_{n-2,1-\alpha/2} S \sqrt{\sum_{i=1}^n x_i^2}}{n s_x} < \beta_0 < b_0 + \frac{t_{n-2,1-\alpha/2} S \sqrt{\sum_{i=1}^n x_i^2}}{n s_x}$$

Intervalos de confianza

- Intervalo de confianza al nivel $(1-\alpha)100\,\%$ para la respuesta media μ_{Y/x_0}:

$$\hat{y}_0 - t_{n-2,1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{nS_x^2}} < \mu_{Y/x_0} <$$

$$\hat{y}_0 + t_{n-2,1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{nS_x^2}}$$

- Intervalo de confianza al nivel $(1-\alpha)100\,\%$ para el valor de y_0 cuando $x = x_0$:

$$\hat{y}_0 - t_{n-2,1-\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}} < y_0 <$$

$$\hat{y}_0 + t_{n-2,1-\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}}$$

Ejemplo anterior

- Intervalo de confianza al nivel $95\,\%$ para la respuesta media μ_{Y/x_0}:

$$\hat{y}_0 - t_{n-2,1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{n s_x^2}} < \mu_{Y/x_0} <$$
$$\hat{y}_0 + t_{n-2,1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{n s_x^2}}$$
$$\hat{y}_0 - t_{7,0.975} \sqrt{\frac{1.58}{9} + \frac{1.58\cdot(x_0-6.89)^2}{9\cdot 12.09877}} < \mu_{Y/x_0} <$$
$$\hat{y}_0 + t_{7,0.975} \sqrt{\frac{1.58}{9} + \frac{1.58\cdot(x_0-6.89)^2}{9\cdot 12.09877}}$$
$$\hat{y}_0 - 2.36 \cdot \sqrt{0.1756 + \frac{(x_0-6.89)^2}{68.917}} < \mu_{Y/x_0} <$$
$$\hat{y}_0 + 2.36 \cdot \sqrt{0.1756 + \frac{(x_0-6.89)^2}{68.917}}$$

Si tomamos $x_0 = 11$ meses, el intervalo anterior vale:
$-0.287 < \mu_{Y/x_0} < 2.775.$

ANOVA en la recta de regresión lineal

Muy brevemente el Análisis de la Varianza (ANalisys Of VAriance) consiste en contrastar si la media de una variable en k poblaciones independientes, con distribución normal de igual varianza, son iguales contra que al menos dos son distintas.

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \ldots = \mu_k \\ H_1: & \text{no todas las medias son iguales} \end{cases}$$

En el caso de la regresión lineal, usaremos la técnica anterior para contrastar si las medias de los grupos que conforman las variables son iguales o no (el grupo k está formado por los valores cuya media vale μ_{Y/x_k}). En caso afirmativo, decir que las medias son iguales es equivalente a afirmar que $\beta_1 = 0$ y, por lo tanto, el modelo de regresión lineal no es bueno. Por tanto, para que el modelo sea bueno, hemos de rechazar la hipótesis nula en el contraste ANOVA

ANOVA en la recta de regresión lineal

- Test a realizar:

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

- Tabla a calcular:

Fuente de variación	Suma de cuadrados	g. l.	Cuadrados medios	F
Regresión	SSR	1	SSR	SSR/S^2
Error	SSE	$n-2$	$S^2 = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

ANOVA en la recta de regresión lineal

- Ahora rechazamos la hipótesis nula al nivel de significación α si $f > f_{1-\alpha,1,n-2}$ donde $f_{1-\alpha,1,n-2}$ es el valor de una distribución F de con grados de libertad 1 y $n-2$.
- Esta prueba, en el caso de regresión lineal simple tiene un efecto a otra parecida en la que se contrasta con una t de student.

Ejemplo anterior

- Tabla ANOVA:

Fuente de variación	Suma de cuadrados	g. l.	Cuadrados medios	F
Regresión	48.9398	1	48.9398	30.974
Error	11.0602	7	$S^2 = 1.5800$	
Total	60	8		

- Tomando $\alpha = 0.05$. El valor $f_{0.95,1,7}$ vale 5.59. Como $f = 30.974 > 5.59$, rechazamos la hipótesis nula y concluimos que $\beta_1 \neq 0$. Por tanto, nuestro modelo es adecuado según este análisis.

Código R para regresión lineal simple

```
> regresion=lm(crecimiento~meses)
> summary(regresion)

Call:
lm(formula = crecimiento ~ meses)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2775 -0.9143 -0.5735  1.0745  1.7561

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.6184     0.9296   9.271 3.52e-05 ***
meses        -0.6704     0.1205  -5.565 0.000846 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Código R para regresión lineal simple

El siguiente código muestra como se pueden dibujar algunos gráficos básicos en regresión.

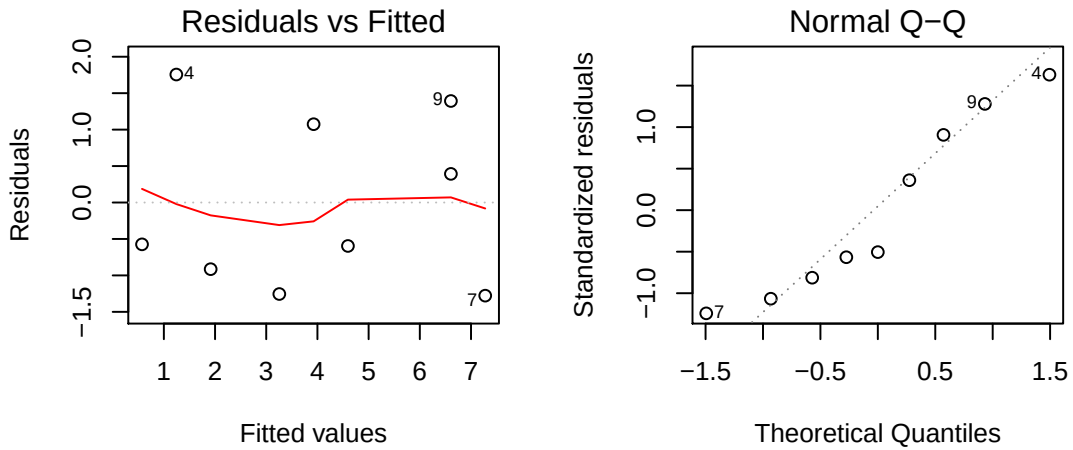
```
> par(mfrow=c(1,2))
> plot(regresion,which=c(1,2))
> par(mfrow=c(1,1))
```

Solo utilizaremos los dos primeros, para lo que utilizaremos el parámetro `which=c(1,2)`.

Código R para regresión lineal simple

- El gráfico de valores estimados (*fitted values*) contra errores (*residuals*). Si el ajuste es bueno deberían distribuirse lo mś cerca de cero y sin mostrar ninguna tendencia específica.
- El gráfico de los cuantiles de los errores contra los cuantiles de una la normal; que es un `qq-plot`. Deberían ajustarse lo más posible a la diagonal y sin mostrar ninguna tendencia.
- Las etiquetas numéricas indican (por defecto) el índice de los tres individuos que más extremos.

Código R para regresión lineal simple



Código R para regresión lineal simple

```
> summary(aov(regresion))

            Df Sum Sq Mean Sq F value   Pr(>F)
meses        1  48.94   48.94   30.97 0.000846 ***
Residuals    7  11.06    1.58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Introducción

- Tenemos k variables independientes $x_1, \ldots, x_k$ y una variable dependiente y.
- Postulamos el modelo de regresión lineal como:

$$\mu_{Y,x_1,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k.$$

Los parámetros β_i son desconocidos y se pueden estimar a partir de una muestra:

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) | i = 1, 2, \ldots, n\}$$

de la que se exige que $n > k$, es decir el número de observaciones sea mayor que el número de variables.

Introducción

- El modelo es el siguiente: Consideramos una conjunto de k variables aleatorias $X_1, X_2, \ldots, X_k$. Suponemos que existen variables aleatorias respuestas $Y_1, \ldots, Y_k$ cuya relación con las anteriores es:

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + E_i,$$

  donde E_i son variables aleatorias que representan el error aleatorio del modelo asociado a la respuesta Y_i.
- El problema es estimar los parámetros β_i a partir de una muestra de datos que representan una muestra aleatoria simple de las variables X_i y de la variable Y de tamaño n:

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) | i = 1, 2, \ldots, n\}.$$

Introducción

- Llamaremos y_i al valor obtenido de la variable Y_i usando las estimaciones b_i de los parámetros β_i:

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + e_i \text{ para } i = 1, 2, \ldots, n$$

  donde e_i será la estimación de la variable error residual E_i asociado a la respuesta Y_i.

- Llamaremos

$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}.$$

  Entonces $e_i = y_i - \hat{y}_i$.

Introducción

- Definimos los vectores siguientes:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}, \ \hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}, \ \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

- Definimos la matriz siguiente:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

Introducción

- Podemos escribir el modelo de regresión múltiple matricialmente como:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{e}, \\ \hat{\mathbf{y}} &= \mathbf{X}\mathbf{b}, \\ \mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}}. \end{aligned}$$

Cálculo de los coeficientes b_i usando el método de mínimos cuadrados

- Definimos el error cuadrático SSE como:

$$\begin{aligned} SSE = & \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ = & \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2. \end{aligned}$$

- Los estimadores por el método de mínimos cuadrados serán los valores $b_0, b_1, \ldots, b_k$ que minimicen SSE.
- Para resolver este problema calculamos las derivadas parciales de SSE respecto a cada b_i para $i = 1, 2, \ldots, n$ y se obtiene el un sistema de ecuaciones que recibe el nombre de ecuaciones normales.

Cálculo de los coeficientes b_i usando el método de mínimos cuadrados

$$\left.\begin{aligned}
nb_0 + b_1 \sum_{i=1}^{n} x_{i1} + \ldots + b_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + & \\
b_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i \\
\ldots \quad \ldots & \\
b_0 \sum_{i=1}^{n} x_{ik} + b_1 \sum_{i=1}^{n} x_{ik}x_{i1} + b_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \ldots + & \\
b_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i
\end{aligned}\right\}$$

Cálculo de los coeficientes b_i usando el método de mínimos cuadrados

- El sistema anterior se puede expresar en forma matricial de la forma siguiente:

$$\left(\mathbf{X}^\top \mathbf{X}\right) \cdot \mathbf{b} = \mathbf{X}^\top \cdot \mathbf{y}.$$

- La solución buscada del sistema anterior será:

$$\mathbf{b} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \cdot \left(\mathbf{X}^\top \mathbf{y}\right).$$

Ejemplo

Se postula que la estatura de un niño recién nacido (y) tiene una relación con su edad en días x_1, su estatura al nacer en cm. (x_2), su peso en Kg. al nacer (x_3) y el aumento en tanto por ciento de su peso actual con respecto a su peso al nacer (x_4). Se pudo obtener una pequeña muestra con $n = 9$ niños cuyos resultados fueron:

y	x_1	x_2	x_3	x_4
57.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67	88	49	5.52	36.5
53.5	67	43	3.21	27.2
62.7	80	48	4.32	27.7
56.2	74	48	2.31	28.3
68.5	94	53	4.3	30.3
69.2	102	58	3.71	28.7

Ejemplo

La matriz $\mathbf{X}$ es:

$$\mathbf{X} = \begin{pmatrix} 1 & 78 & 48.2 & 2.75 & 29.5 \\ 1 & 69 & 45.5 & 2.15 & 26.3 \\ 1 & 77 & 46.3 & 4.41 & 32.2 \\ 1 & 88 & 49 & 5.52 & 36.5 \\ 1 & 67 & 43 & 3.21 & 27.2 \\ 1 & 80 & 48 & 4.32 & 27.7 \\ 1 & 74 & 48 & 2.31 & 28.3 \\ 1 & 94 & 53 & 4.3 & 30.3 \\ 1 & 102 & 58 & 3.71 & 28.7 \end{pmatrix}$$

Ejemplo

El vector $\mathbf{y}$ es:

$$\mathbf{y} = \begin{pmatrix} 57.5 \\ 52.8 \\ 61.3 \\ 67 \\ 53.5 \\ 62.7 \\ 56.2 \\ 68.5 \\ 69.2 \end{pmatrix}$$

Ejemplo

El producto $\mathbf{X}^\top\mathbf{X}$ es el siguiente:

$$\begin{pmatrix} 9 & 729 & 439 & 32.68 & 266.7 \\ 729 & 60123 & 35947.2 & 2702.41 & 21715.3 \\ 439 & 35947.2 & 21568.18 & 1604.388 & 13026.01 \\ 66.07 & 6108.19 & 3541.008 & 128.66 & 1948.561 \\ 266.7 & 21715.3 & 13026.01 & 990.27 & 7980.83 \end{pmatrix}$$

El producto $\mathbf{X}^\top\mathbf{y}$ es el siguiente:

$$\mathbf{X}^\top\mathbf{y} = \begin{pmatrix} 548.7 \\ 45001 \\ 26946.89 \\ 2035.52 \\ 16348.29 \end{pmatrix}$$

Ejemplo

Resolviendo el sistema $(\mathbf{X}^\top \mathbf{X}) \cdot \mathbf{b} = \mathbf{X}^\top \mathbf{y}$ obtenemos como solución:

$$\mathbf{b} = \begin{pmatrix} 7.1475 \\ 0.1001 \\ 0.7264 \\ 3.0758 \\ -0.03 \end{pmatrix}$$

La recta de regresión estimada es :

$$\hat{y} = 7.1475 + 0.1001x_1 + 0.7264x_2 + 3.0758x_3 - 0.03x_4.$$

Propiedades de la recta de regresión

La recta de regresión ajustada pasa por el vector de medias. O sea, si llamamos $\overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$, para $i = 1, \ldots, k$, $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, se verifica que:

$$\overline{y} = b_0 + b_1\overline{x}_1 + \cdots + b_k\overline{x}_k.$$

La suma de los errores e_i es 0: $\sum_{i=1}^{n} e_i = 0$ y por lo tanto su media también: $\overline{e} = 0$. La media de los valores estimados coincide con la media de los valores de la muestra: $\overline{\hat{y}} = \overline{y}$.

Ejemplo

Veamos que la recta de regresión pasa por el vector de medias. Éste vale:

$$\overline{x}_0 = 1,\ \overline{x}_1 = 81,\ \overline{x}_2 = 48.778,\ \overline{x}_3 = 3.631,\ \overline{x}_4 = 29.633.$$

La media del vector $\mathbf{y}$ vale: $\overline{y} = 60.967$.
Se cumple:

$$\begin{aligned} 60.967 \approx \quad & 7.1475 + 0.1001 \cdot 81 + 0.7264 \cdot 48.778 \\ & +3.0758 \cdot 3.6311 - 0.03 \cdot 29.633. \end{aligned}$$

Ejemplo

Veamos que la media de los errores es nula. Los valores de los vectores $\hat{\mathbf{y}}$ y $\mathbf{e}$ son:

$$\hat{\mathbf{y}} = \mathbf{X} \cdot b = \begin{pmatrix} 57.541 \\ 52.929 \\ 61.085 \\ 67.432 \\ 54.146 \\ 62.479 \\ 55.678 \\ 67.372 \\ 70.039 \end{pmatrix}, \ \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -0.0405 \\ -0.1290 \\ 0.2150 \\ -0.4324 \\ -0.6461 \\ 0.2214 \\ 0.5225 \\ 1.1277 \\ -0.8385 \end{pmatrix}$$

Puede comprobarse que la media del vector $\mathbf{e}$ es 0 y que $\overline{\hat{y}} = \overline{y} = 60.967$.

Sumas de cuadrados en la regresión

- Llamaremos suma de cuadrados de los residuales o del error a

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- Llamaremos suma de cuadrados de totales a

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

- Llamaremos suma de cuadrados de la regresión a

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2.$$

- Se verifica:

$$SST = SSR + SSE,$$

  o equivalentemente:

$$s_y^2 = s_{\hat{y}}^2 + s_e^2.$$

Ejemplo

- La suma de cuadrados del error vale en el ejemplo anterior:

$$SSE = (57.5 - 57.5405)^2 + \cdots + (69.2 - 70.0385)^2 = 2.9656.$$

- La suma de cuadrados totales vale:

$$SST = (57.5 - 60.967)^2 + \cdots + (69.2 - 60.967)^2 = 321.24.$$

- La suma de cuadrados de la regresión es:

$$\begin{aligned} SSR = & \quad (57.5405 - 60.967)^2 + \cdots + (70.0385 - 60.967)^2 \\ = & \quad 318.274. \end{aligned}$$

- Puede observarse que se cumple:

$$SST = SSE + SSR, \quad 321.24 = 2.9656 + 318.274.$$

Definición del coeficiente de determinación

- Definimos el coeficiente de determinación como

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

o también

$$R^2 = \frac{S^2_{\hat{y}}}{S^2_y} = 1 - \frac{S^2_e}{S^2_y}.$$

R^2 se interpreta como la proporción de varianza de la variable y que es explicada por el modelo de regresión múltiple.

- Definimos el coeficiente de determinación ajustado como:

$$R^2a = R^2 - \frac{k(1-R^2)}{n-k-1}.$$

- Definimos el coeficiente de correlación múltiple r de la variable y respecto de las variables $x_1, \ldots, x_k$ como $r = \sqrt{R^2}$.

Ejemplo

- El coeficiente de determinación vale en nuestro ejemplo:

$$R^2 = \frac{318.274}{321.24} \approx 0.9908.$$

- El coeficiente de determinación ajustado vale:

$$R^2a = 0.9908 - \frac{4 \cdot (1 - 0.9908)}{9 - 4 - 1} = 0.9815.$$

- El coeficiente de correlación múltiple r de la variable y respecto de las variables $x_1,\ x_2,\ x_3,\ x_4$ vale: $r = \sqrt{0.9908} = 0.9954$.

Consideraciones sobre el modelos de regresión múltiple

- Suponemos que las variables aleatorias error E_i son independientes e idénticamente distribuidas según una normal de media 0 y varianza σ^2.
- Bajo el supuesto anterior, los estimadores $b_0, \ldots, b_k$ de $\beta_0, \ldots, \beta_k$ son insesgados. O sea, $E(b_i) = \beta_i,\ i = 0, \ldots, k$.
- La matriz $(X^\top X)^{-1}\sigma^2$ es la matriz de covarianzas de $\beta_0, \ldots, \beta_k$.
- Un estimador insesgado de σ^2 es

$$S^2 = \frac{SSE}{n-k-1}.$$

Ejemplo

- Una estimación de la varianza σ^2 será:

$$S^2 = \frac{2.9656}{9-4-1} = 0.7414.$$

- Una estimación de la matriz de covarianzas de $\beta_0, \ldots, \beta_4$: $(X^\top X)^{-1} S^2 =$

$$\begin{pmatrix} 270.919 & 5.325 & -12.521 & -13.743 & -1.4 \\ 5.325 & 0.115 & -0.266 & -0.326 & -0.0176 \\ -12.521 & -0.266 & 0.618 & 0.742 & 0.0416 \\ -13.743 & -0.326 & 0.742 & 1.122 & -0.00598 \\ -1.4 & -0.0176 & 0.0416 & -0.00598 & 0.0277 \end{pmatrix}$$

ANOVA en regresión lineal múltiple

- El contraste ANOVA en la regresión lineal múltiple nos permite contrastar la adecuación del modelo. Se trata de contrastar:

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: & \text{hay alguna } \beta_i \neq 0 \end{cases}$$

- Si aceptamos H_0 estamos diciendo que la estimación dada por la regresión es constante. Por tanto el modelo no sería adecuado.
- En la tabla siguiente aparece los pasos necesarios para realizar el contraste. Como puede observarse, se usa como estadístico de contraste el cociente $\frac{MSR}{MSE}$ que, suponiendo normalidad, sigue una distribución F de Snédecor de $k, n-k-1$ grados de libertat:

ANOVA en regresión lineal múltiple

F.V.	S.C.	g.l.	C.M.	f
Regresión	SSR	k	$MSR = \frac{SSR}{k}$	$f = \frac{MSR}{MSE}$
Error	SSE	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$	
Total	SST	$n-1$		

F.V.: Fuente de variación.

S.C.: Suma de cuadrados.

g.l.: grados de libertad.

C.M.: Cuadrados medios.

Rechazaremos la hipótesis nula $H_0 : \beta_1 = \ldots = \beta_k = 0$ al nivel de significación α si $f > f_{1-\alpha,k,n-k-1}$ donde $f_{1-\alpha,k,n-k-1}$ es el cuantil de una distribución F de con grados de libertad k y $n-k-1$.

Ejemplo

- La tabla ANOVA es en nuestro ejemplo:

F.V.	S.C.	g.l.	C.M.	f
Regresión	318.274	4	$MSR = 79.569$	$f = 107.323$
Error	2.9656	4	$MSE = 0.7414$	
Total	321.24	8		

- Cogiendo $\alpha = 0.05$, el cuantil es $f_{1-0.05,4,4}$ vale 6.388. Como $f = 107.323 > f_{0.95,4,4} = 6.388$, rechazamos la hipótesis nula y concluimos que el modelo es adecuado según este análisis.

ANOVA en regresión lineal múltiple

- El modelo de regresión lineal con este conjunto de x puede no ser el único que se puede utilizar. Es posible que con algunas transformaciones de las x mejore el valor de f.
- El modelo podría ser más eficaz si se incluyen otras variables o podría continuar siendo casi igual de eficaz si se eliminan algunas (principio de parsimonia).

Intervalos de confianza

- Un intervalo de confianza al nivel $(1-\alpha)100\,\%$ para la respuesta media $\mu_{Y/x_{10},x_{20},\ldots,x_{k0}}$ es

$$\hat{y}_0 - t_{1-\alpha/2,n-k-1} S\sqrt{x_0^\top (X^\top X)^{-1} x_0} <$$

$$\mu_{Y/x_{10},x_{20},\ldots,x_{k0}} < \hat{y}_0 + t_{1-\alpha/2,n-k-1} S\sqrt{x_0^\top (X^\top X)^{-1} x_0}$$

  donde $t_{1-\alpha/2,n-k-1}$ es el cuantil $1-\alpha/2$ de una t de student con $n-k-1$ grados de libertad, $x_0 = (1, x_{10}, x_{20}, \ldots, x_{k0})^\top$ y $\hat{y}_0 = b_0 + b_1 x_{10} + \cdots + b_k x_{k0}$.

- La cantidad $S\sqrt{x_0^\top (X^\top X)^{-1} x_0}$ recibe el nombre de error estándar de predicción.

Ejemplo

- Para $\alpha = 0.05$, hallemos un intervalo de confianza para la respuesta media $\mu_{Y/x_{10},x_{20},x_{30},x_{40}}$, para
  $x_{10} = 69,\ x_{20} = 45.5,\ x_{30} = 2.15,\ x_{40} = 26.3.$
- El cuantil $t_{1-0.025,4}$ vale 2.776. El valor $\hat{y}_0$ valdrá:

$$\begin{aligned} \hat{y}_0 = & \quad b_0 + \sum_{i=1}^{4} b_i x_{i0} \\ = & \quad 7.1475 + 0.1001 \cdot 69 + 0.7264 \cdot 45.5 + 3.0758 \cdot 2.15 \\ & \quad -0.03 \cdot 26.3 = 52.929. \end{aligned}$$

- El valor de $x_0^\top (X^\top X)^{-1} x_0$ es:

$$(1, 69, 45.5, 2.15, 26.3) \cdot (X^\top X)^{-1} \cdot \begin{pmatrix} 1 \\ 69 \\ 45.5 \\ 2.15 \\ 26.3 \end{pmatrix} = 0.3615.$$

Ejemplo

El intervalo de confianza es:

$$52.929 - 2.776\sqrt{0.7414 \cdot 0.3615} < \mu_{Y/x_{10},x_{20},x_{30},x_{40}} < 52.929 + 2.776 \cdot \sqrt{0.7414 \cdot 0.3615}$$

operando

$$51.492 < \mu_{Y/x_{10},x_{20},x_{30},x_{40}} < 54.366.$$

Intervalos de confianza

- Un intervalo de confianza al nivel $(1-\alpha)100\,\%$ para una predicción individual y_0 para los valores de la variables dependientes $x_{10}, x_{20}, \ldots, x_{k0}$ es

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

$$\hat{y}_0 + t_{1-\alpha/2,n-k-1} S \sqrt{1 + x_0^\top (X^\top X)^{-1} x_0},$$

donde $t_{1-\alpha/2,n-k-1}$ es el cuantil de una t de student con $n-k-1$ grados de libertad, y $x_0 = (1, x_{10}, x_{20}, \ldots, x_{k0})^\top$.

Intervalos de confianza

- Un intervalo de confianza al nivel $(1-\alpha)100\,\%$ para el parámetro β_k es:

$$b_k + t_{1-\alpha/2,n-k-1} s_{\beta_k} < \beta_k < b_k + t_{1-\alpha/2,n-k-1} s_{\beta_k},$$

donde s_{β_k} es la raiz cuadrada del elemento k-ésimo de la diagonal de la matriz $(X^\top X)^{-1} S^2$: $s_{\beta_k} = \sqrt{((X^\top X)^{-1} S^2)_{kk}}$.

Ejemplo

- Para $\alpha = 0.05$, hallemos un intervalo de confianza para el parámetro β_2.
- El cuantil $t_{1-0.025,4}$ vale 2.776. Los valores de la diagonal de la matriz $(X^\top X)^{-1} S^2$ son:

$$270.919, \quad 0.1154, \quad 0.6176, \quad 1.1219, \quad 0.02775.$$

  El valor que nos interesa es para el parámetro β_2: 0.726.
- El intervalo será:

$$\begin{aligned} 0.726 - 2.776 \cdot \sqrt{0.6176} < \quad & \beta_2 \\ < 0.726 + 2.776 \cdot \sqrt{0.6176}, & \\ -1.4556 < \quad & \beta_2 < 2.908. \end{aligned}$$

El problema de la selección del modelo. Colinealidad

- Dado un problema regresión lineal múltiple podemos ajustar todos los submodelos lineales posibles

$$\begin{aligned} Y = & \quad \beta_0 + \beta_1 x_1, \\ Y = & \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2, \\ & \quad \dots \\ Y = & \quad \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \end{aligned}$$

  ya que es posible que entre variables x_i exista una fuerte relación lineal y entonces *sobren del modelo*. Este problema recibe el nombre de colinealidad.
- Existen también otro tipo de problemas que se pueden resolver usando la técnica de la regresión lineal múltiple. Véase como ejemplo el problema de la regresión polinomial:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k.$$

El problema de la selección del modelo. Colinealidad

Una solución para ver qué modelo lineal es el más simple y adecuado es recurrir a los llamados métodos secuenciales de selección del modelo lineal, como son los siguientes:

- Regresión paso paso (Stepwise).
- Selección hacia adelante(Forward).
- Selección hacia atrás (Backward).