

Análisis de regresión

Capítulo 2: Análisis de regresión lineal múltiple

Mg. Sc. J. Eduardo Gamboa U.



Presentación

Hemos visto que es posible explicar la relación de dependencia entre un par de variables mediante una línea de regresión. En aplicaciones reales, la variable respuesta suele estar influenciada por más de una variable independiente, por lo que el modelo general adopta la forma:

$$y = f(x_1, x_2, \dots, x_p, \varepsilon)$$

Ejemplo aplicado

Se busca estudiar el efecto lineal de la edad (años), el sexo (femenino / masculino) y los años de educación sobre el sueldo mensual (en miles de soles). Adicionalmente, se incorpora una variable ficticia X_4 , generada al azar, para evaluar su comportamiento en el modelo. Se dispone de una muestra de $n = 24$ observaciones.

Una primera etapa del análisis consiste en explorar la relación entre la variable respuesta y cada variable explicativa mediante diagramas de dispersión.

Modelo de regresión lineal múltiple

Dadas p variables independientes X_1, \dots, X_p y una variable respuesta Y , el modelo de regresión lineal múltiple se define como:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_p X_{p,i} + \varepsilon_i, \quad i = 1, \dots, n$$

donde:

- ▶ β_0 es el intercepto
- ▶ β_1, \dots, β_p son los coeficientes de regresión
- ▶ ε_i es el término de error aleatorio, con

$$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2$$

Si las variables X_j son aleatorias, sus observaciones deben ser al menos independientes (Montgomery, Peck y Vining, 2012).

En notación matricial:

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

Nuestro objetivo es estimar $\boldsymbol{\beta}$ y σ^2 a partir de la muestra.

```
library(readxl)
(datos = read_excel('U2_datos_1.xlsx'))
```

```
# A tibble: 24 x 5
```

	Sueldo	Educacion	Sexo	Edad	X4
	<dbl>	<dbl>	<chr>	<dbl>	<dbl>
1	4.22	12	M	31	10
2	5.89	18	M	23	-3
3	9.88	20	M	50	-6
4	2.35	7.5	F	19	-4
5	7	20	F	39	-2
6	1.25	9	M	18	7
7	6.78	3	M	39	-7
8	5.19	15	F	32	8
9	8.16	21	F	35	5
10	6.1	18	F	34	7

```
# i 14 more rows
```

Inferencia

Estimación puntual

El estimador de mínimos cuadrados ordinarios del vector de coeficientes es:

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

Además:

$$E(\hat{\beta}) = \beta \quad \text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

El estimador de la varianza del error es:

$$\hat{\sigma}^2 = \frac{SC_{Error}}{n - k}$$

donde $k = p + 1$ es el número de coeficientes estimados.

```
library(dplyr)
y = datos$Sueldo
X = model.matrix(Sueldo~Educacion+Sexo+Edad+X4,data=datos)
solve(t(X)%*%X)%*%t(X)%*%y
```

```
          [,1]
(Intercept) -0.73053615
Educacion    0.15695057
SexoM        0.82211261
Edad         0.10462774
X4           -0.04537505
```



```
modelo = lm(Sueldo ~ ., datos)
(beta = coef(modelo))
```

(Intercept)	Educacion	SexoM	Edad	X4
-0.73053615	0.15695057	0.82211261	0.10462774	-0.04537505

```
(sigma = summary(modelo)$sigma)
```

```
[1] 1.13476
```

Estimación intervalar

Si g_{ii} es el elemento i -ésimo de la diagonal de $(X'X)^{-1}$, entonces:

$$IC(\beta_i) = \hat{\beta}_i \pm t_{1-\alpha/2, n-k} \hat{\sigma} \sqrt{g_{ii}}$$

```
G      = solve(t(X)%*%X)
g      = G |> diag()
n      = datos |> nrow()
k      = beta |> length()
valt   = qt(0.975, 24-5)
beta - valt*sigma*sqrt(g)
```

(Intercept)	Educacion	SexoM	Edad	X4
-2.58809073	0.02638407	-0.18289392	0.04174457	-0.12999986

```
beta + valt*sigma*sqrt(g)
```

(Intercept)	Educacion	SexoM	Edad	X4
1.12701843	0.28751706	1.82711914	0.16751091	0.03924976

```
modelo |> confint()
```

	2.5 %	97.5 %
(Intercept)	-2.58809073	1.12701843
Educacion	0.02638407	0.28751706
SexoM	-0.18289392	1.82711914
Edad	0.04174457	0.16751091
X4	-0.12999986	0.03924976

Prueba de hipótesis global

Se contrasta:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

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

$$GL_{Total} = n - 1 \quad GL_{Reg} = k - 1 \quad GL_{Error} = n - k$$

$$SC_{Total} = (\mathbf{y} - \bar{y}\mathbf{1})'(\mathbf{y} - \bar{y}\mathbf{1}) \quad \bar{y} = \frac{1}{n}\mathbf{1}'\mathbf{y}$$

$$SC_{Reg} = (\hat{\mathbf{y}} - \bar{y}\mathbf{1})'(\hat{\mathbf{y}} - \bar{y}\mathbf{1})$$

$$SC_{Error} = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

$$CM_{Reg} = \frac{SC_{Reg}}{GL_{Reg}} \quad CM_{Error} = \frac{SC_{Error}}{GL_{Error}}$$

El estadístico de prueba es:

$$F_{calc} = \frac{CM_{Reg}}{CM_{Error}}$$

Se rechaza H_0 si:

$$F_{calc} > F_{1-\alpha, k-1, n-k}$$

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1 : \text{Al menos un } \beta_j \neq 0, j = 1, 2, 3, 4$$

Grados de libertad

$$(GL_{Total} = n - 1)$$

$$[1] \quad 23$$

$$(GL_{Reg} = k - 1)$$

$$[1] \quad 4$$

$$(GL_{Error} = n - k)$$

$$[1] \quad 19$$

Sumas de cuadrados

```
uno = rep(1,n)
ybarra = as.numeric(1/n*(uno)%*%y)
yhat = predict(modelo)
(SCTotal = t(y - ybarra*uno)%*%(y - ybarra*uno))
```

```
[,1]
```

```
[1,] 103.0379
```

```
(SCReg = t(yhat - ybarra*uno)%*%(yhat - ybarra*uno))
```

```
[,1]
```

```
[1,] 78.57194
```

```
(SCError = t(y - yhat)%*%(y-yhat))
```

```
[,1]
```

```
[1,] 24.46592
```


Cuadrados medios

```
(CMReg = SCReg / GLReg)
```

```
[,1]
```

```
[1,] 19.64299
```

```
(CMError = SError / GLError)
```

```
[,1]
```

```
[1,] 1.28768
```

Estadístico de prueba

```
(Fcalc = CMReg / CMError)
```

```
      [,1]  
[1,] 15.25455
```

Valor crítico:

```
qf(0.95,4,19)
```

```
[1] 2.895107
```

Pvalor:

```
pf(Fcalc,4,19, lower.tail = F)
```

```
      [,1]  
[1,] 9.638667e-06
```

```
lm(y~X) |> aov() |> summary()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	4	78.57	19.643	15.26	9.64e-06 ***
Residuals	19	24.47	1.288		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimación de la media de la respuesta

Para un vector de valores explicativos \mathbf{x} :

$$\hat{\mu} = \hat{y} = \mathbf{x}'\hat{\beta}$$

Intervalo de confianza:

$$IC(\mu|\mathbf{x}) = \hat{\mu} \pm t_{1-\alpha/2, n-k} \sqrt{\hat{\sigma}^2 \mathbf{x}'(X'X)^{-1}\mathbf{x}}$$

Estimación del sueldo medio para una mujer de 35 años de edad, con 12 años de educación, y tomando $X_4 = 1$.

Puntual:

```
x0 = c(1, 12, 0, 35, 1)
yest = x0%*%beta
yest
```

[,1]

[1,] 4.769467

Intervalar al 95% de confianza:

```
me = qt(0.975, n-k)*sqrt(sigma^2*t(x0)%*%solve(t(X)%*%X)%*%x0)
(LI = yest - me)
```

[,1]

[1,] 3.944129

```
(LS = yest + me)
```

[,1]

[1,] 5.594804

```
predict(modelo,  
  data.frame(Educacion = 12,  
             Sexo = "F",  
             Edad = 35,  
             X4 = 1),  
  interval = "confidence")
```

	fit	lwr	upr
1	4.769467	3.944129	5.594804

```
predict(modelo,  
  data.frame(Educacion = c(12,7),  
    Sexo = c("F","M"),  
    Edad = c(35,28),  
    X4 = c(1,-2)),  
  interval = "confidence")
```

	fit	lwr	upr
1	4.769467	3.944129	5.594804
2	4.210557	3.177470	5.243645

Predicción de un nuevo valor

El predictor puntual es:

$$\hat{y}_0 = \mathbf{x}'\hat{\beta}$$

Intervalo de predicción:

$$IP(y|\mathbf{x}) = \hat{y}_0 \pm t_{1-\alpha/2, n-k} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}'(X'X)^{-1}\mathbf{x})}$$

Intervalar al 95% de confianza:

```
me = qt(0.975, n-k)*sqrt(sigma^2*(1+t(x0)%*%solve(t(X)%*%X)%*%x0))  
(LI = yest - me)
```

```
[,1]
```

```
[1,] 2.255071
```

```
(LS = yest + me)
```

```
[,1]
```

```
[1,] 7.283862
```

```
predict(modelo,  
  data.frame(Educacion = 12,  
             Sexo = "F",  
             Edad = 35,  
             X4 = 1),  
  interval = "prediction")
```

	fit	lwr	upr
1	4.769467	2.255071	7.283862

Residuales

El vector de residuales es:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (I - H)\mathbf{y}$$

donde:

$$H = X(X'X)^{-1}X'$$

El análisis de residuales permite verificar normalidad, homocedasticidad e independencia.

Bibliografía

- ▶ Montgomery, D., Peck, E., Vining, G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
- ▶ Weisberg, S. (2014). *Applied Linear Regression*. Wiley.