Capítulo

8

Análisis de Regresión Lineal Múltiple.

8.1. Introducción

En el capítulo anterior se desarrollo el análisis de regresión cunado sobre la variable dependiente influye sólo una variable independiente. Por lo general, en la práctica este no es el caso. En este capítulo se extiende al caso donde hay más de una variable independiente, en cuyo caso se dice que se realiza un análisis de regresión lineal múltiple.

8.2. Modelo de Regresión Lineal Múltiple

En general se puede relacionar la variable respuesta y con k variables independientes $x_1, x_2, ..., x_k$, en ese caso el modelo está dado por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{8.1}$$

donde los coeficientes β_j, $j = 0, 1, ..., k$ son constantes desconocidas y son los parámetros del modelo. Cada β_j representa el cambio esperado en la respuesta y por

el cambio unitario en x_j cuando todas las demás variables independientes $x_i (i \neq j)$ se mantienen constantes. ϵ es un componente de error aleatorio.

En el caso de los modelos de regresión múltiple es preferible usar la notación matricial, pues dicha forma permite expresar el modelo en una forma más compacta y que con un poco de conocimiento del algebra matricial los resultados se simplifican considerablemente.

Forma Matricial: El modelo de Regresión Múltiple en su forma matricial es la siguiente:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{8.2}$$

donde

1. $\mathbf{y}$ es un vector $n \times 1$ observable;

2. $\mathbf{X}$ es una matriz $n \times p$ que contiene los valores de las variables independientes;

3. β es un vector $p \times 1$ de parámetros no observables;

4. ε es un vector $n \times 1$ de variables aleatorias no observables conocido como el vector de errores aleatorios.

Si se reescriben los vectores y las matrices de la ecuación 2.2 en detalle, se obtiene

$$\mathbf{y} = \begin{bmatrix} y_1 & 1 & x_{11} & x_{12} & \dots & x_{1k} \\ y_2 & 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ y_n & 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{8.3}$$

8.3. Ejemplo: Tiempo de Entrega

Este es un ejemplo tomado de Montgomery(2002):Un embotellador de bebidas gaseosas analiza las rutas de servicio de las máquinas expendedoras en su sistema de distribución. Le interesa predecir el tiempo necesario para que el representante de ruta atienda las máquinas expendedoras en una tienda. Esta actividad de servicio consiste en abastecer la máquina con productos embotellados, y algo de mantenimiento o limpieza. El ingeniero industrial responsable del estudio ha sugerido que las dos variables más importantes que afectan el tiempo de entrega y son la cantidad de cajas de producto abastecido, x_1, y la distancia caminada por el representante, x_2. El ingeniero ha reunido 25 observaciones de tiempo de entrega que se ven en la tabla 2.1. Se ajustará el modelo de regresión lineal simple siguiente

$$y = \beta_0 + \beta x_1 + \varepsilon$$

En este caso la matriz $\mathbf{X}$ y el vector $\mathbf{y}$ están dados por

Tabla 8.1: Datos de tiempo de entrega

Observación	y	x_1	x_2	Observación	y	x_1	x_2
1	16,68	7	560	14	19,75	6	462
2	11,5	3	220	15	24	9	448
3	12,03	3	340	16	29	10	776
4	14,88	4	80	17	15,35	6	200
5	13,75	6	150	18	19	7	132
6	18,11	7	330	19	9,5	3	36
7	8	2	110	20	35,1	17	770
8	17,83	7	210	21	17,9	10	140
9	79,24	30	1460	22	52,32	26	810
10	21,5	5	605	23	18,75	9	450
11	40,33	16	688	24	19,83	8	635
12	21	10	215	25	10,75	4	150
13	13,5	4	255				

$$\mathbf{X} = \begin{pmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ 1 & 4 & 80 \\ 1 & 6 & 150 \\ 1 & 7 & 330 \\ 1 & 2 & 110 \\ 1 & 7 & 210 \\ 1 & 30 & 1460 \\ 1 & 5 & 605 \\ 1 & 16 & 688 \\ 1 & 10 & 215 \\ 1 & 4 & 255 \\ 1 & 6 & 462 \\ 1 & 9 & 448 \\ 1 & 10 & 776 \\ 1 & 6 & 200 \\ 1 & 7 & 132 \\ 1 & 3 & 36 \\ 1 & 17 & 770 \\ 1 & 10 & 140 \\ 1 & 26 & 810 \\ 1 & 9 & 450 \\ 1 & 8 & 635 \\ 1 & 4 & 150 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 16{,}68 \\ 11{,}50 \\ 12{,}03 \\ 14{,}88 \\ 13{,}75 \\ 18{,}11 \\ 8{,}00 \\ 17{,}83 \\ 79{,}24 \\ 21{,}50 \\ 40{,}33 \\ 21{,}00 \\ 13{,}50 \\ 19{,}75 \\ 24{,}00 \\ 29{,}00 \\ 15{,}35 \\ 19{,}00 \\ 9{,}50 \\ 35{,}10 \\ 17{,}90 \\ 52{,}32 \\ 18{,}75 \\ 19{,}83 \\ 10{,}75 \end{pmatrix}$$

8.4. Estimación de los Parámetros del Modelo

8.4.1. Estimación de β.

El estimador de mínimos cuadrados de β, denotado por $\hat{\beta}$, es el valor de β que mínimiza

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Por lo tanto, lo que se debe hacer es derivar la expresión anterior y buscar el valor de β que la hace igual a cero. Antes de derivar note que la expresión anterior se puede escribir como

$$\begin{aligned} S(\beta) &= \mathbf{y}'\mathbf{y} - \beta'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta + \mathbf{X}'\beta'\beta\mathbf{X} \\ &= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \mathbf{X}'\beta'\beta\mathbf{X} \end{aligned}$$

Ahora si derivando e igualando a cero se obtiene

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

que se simplifica a

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \tag{8.4}$$

las cuales se conocen como las **ecuaciones normales de mínimos cuadrados.** Para hallar la expresión de $\hat{\beta}$ se premultiplica la ecuación anterior por la inversa de $\mathbf{X}'\mathbf{X}$ (que en este caso se asume que existe). Por lo tanto el estimador de β por mínimos cuadrados es

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{8.5}$$

Ejemplo 8.1 *Para el ejemplo se tiene que la matriz* $\mathbf{X}'\mathbf{X}$ *está dada por*

$$\begin{aligned}\mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix} \\ &= \begin{bmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{bmatrix}\end{aligned}$$

y el vector $\mathbf{X}'\mathbf{y}$ *es*

$$\begin{aligned}\mathbf{X}'\mathbf{y} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 16{,}68 \\ 11{,}50 \\ \vdots \\ 10{,}75 \end{bmatrix} \\ &= \begin{bmatrix} 559{,}60 \\ 7375{,}44 \\ 337072{,}00 \end{bmatrix}\end{aligned}$$

El estimador de β *por mínimos cuadrados es*

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

o sea

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{bmatrix}^{-1} \begin{bmatrix} 559{,}60 \\ 7375{,}44 \\ 337072{,}00 \end{bmatrix}$$

$$= \begin{bmatrix} 0{,}11321518 & -0{,}00444859 & -0{,}00008367 \\ -0{,}00444859 & 0{,}00274378 & -0{,}00004786 \\ -0{,}00008367 & -0{,}00004786 & 0{,}00000123 \end{bmatrix}^{-1} \begin{bmatrix} 559{,}60 \\ 7375{,}44 \\ 337072{,}00 \end{bmatrix}$$

$$= \begin{bmatrix} 2{,}34123115 \\ 1{,}61590712 \\ 0{,}01438483 \end{bmatrix}$$

El ajuste por mínimos cuadrados, con los coeficientes de regresión expresados con cinco decimales, es

$$\hat{y} = 2{,}34123 + 1{,}61591x_1 + 0{,}01438x_2$$

Procedimiento en R

La estimación de los parámetros se obtienen directamente usando la instrucción

```
> MRLM1<-lm(resp~x1+x2,data=Datos)
> MRLM1
```

con lo cual se obtiene

```
Call:
lm(formula = resp ~ x1 + x2, data = Datos)
```

```
Coefficients:
(Intercept)          x1          x2
    2.34123     1.61591     0.01438
```

en donde se tiene que $\hat{\beta}_0 = 2{,}34123$, $\hat{\beta}_1 = 1{,}61591$ y $\hat{\beta}_2 = 0{,}01438$.

Si se quiere conocer el valor de uno de los estimadores en particular se usa la instrucción

```
> objetolm$coef[j+1]
```

Con lo cual se obtiene el valor estimado del parámetro j. Por ejemplo, si quiere conocer el valor de β_1 se coloca la instrucción

```
> MRLM1$coef[2]
```

Otra manera de obtener las estimaciones usando R es usando las siguientes instrucciones

- Creación de la matriz $\mathbf{X}$ y el vector $\mathbf{y}$

```
> X<-matrix(c(idv,x1,x2),nrow=25,ncol=3)
> y<-matrix(c(resp),nrow=25,ncol=1)
```

- Se calculan las estimaciones usando la ecuación ** por medio de las siguientes instrucciones

```
> beta<-solve((t(X)%*%X))%*%t(X)%*%y
> beta
```

Si se desea conocer el valor de alguno de los β_j se usa la siguiente instrucción

```
> beta[j]
```

8.4.2. Estimación de σ^2.

Al igual que en el caso de la regresión lineal simple, el estimador de σ^2 se puede obtener a partir de la suma de cuadrados de los residuales:

$$\begin{aligned} SC_{Res} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} r_i^2 \\ &= \mathbf{r}'\mathbf{r} \end{aligned}$$

Sustituyendo $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\beta}$ se obtiene

$$\begin{aligned} SC_{Res} &= (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\beta} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\beta}'\mathbf{X}'\mathbf{y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \end{aligned}$$

como $\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$, la última ecuación se transforma en

$$SC_{Res} = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} \tag{8.6}$$

la cual tiene $n - p$ grados de libertad (pues hay que p parámetros en el modelo de regresión múltiple). Por lo tanto el cuadrado medio del residual es

$$CM_{Res} = \frac{SC_{Res}}{n - p} \tag{8.7}$$

cuyo valor esperado es σ^2. Por lo tanto un estimador insesgados de σ^2, denotado por $\hat{\sigma}^2$ es

$$\hat{\sigma}^2 = M_{Res} \tag{8.8}$$

Ejemplo 8.2 *Se estimará la varianza del error, σ^2, para el ajuste del modelo de regresión múltiple a los datos de tiempo de entrega de bebidas gaseosas en el ejemplo ***. Ya que*

$$\mathbf{y}'\mathbf{y} = 18310,6290$$

y

$$\begin{aligned} \hat{\beta}'\mathbf{X}'\mathbf{y} &= \begin{bmatrix} 2,34123115 & 1,61590721 & 0,01438483 \end{bmatrix} \begin{bmatrix} 559{,}60 \\ 7375{,}44 \\ 337072{,}00 \end{bmatrix} \\ &= 18076,90304 \end{aligned}$$

la suma de cuadrados de residuales es

$$\begin{aligned} SC_{Res} &= \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} \\ &= 18310,6290 - 18076{,}9030 = 233,7260 \end{aligned}$$

Por consiguiente, el estimado de σ^2 es el cuadrado medio de residuales

$$\hat{\sigma}^2 = \frac{SC_{Res}}{GL_{Res}} = \frac{223,7260}{25-3} = 10,6239$$

Procedimiento en R

La estimación de $\sqrt{\sigma^2}$ se obtiene como uno de los resultados arrojados por la instrucción

```
> summary(objetolm)
```

donde objetolm es un objeto de la instrucción lm(). Otra manera de obtener la estimación de σ^2 es usando las siguientes instrucciones

```
> varest<-(t(y)%*%y-t(beta)%*%t(X)%*%y)/(nrow(y)-nrow(beta))
> varest
```

8.4.3. Propiedades de los estimadores.

1. **Son estimadores insesgados**. En la sección anterior se probo que $\hat{\sigma}^2$ es un estimador insesgados de σ^2. Por lo tanto sólo falta probar con $\hat{\beta}$.

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{X}\beta+\varepsilon) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta \end{aligned}$$

2. $Cov(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

3. $\hat{\beta}$ y $\hat{\sigma}^2$ son independientes.

4. Si se supone que los errores son normales se tiene que $\hat{\beta}$ también se distribuye normal y que una función de $\hat{\sigma}^2$ se distribuye chi cuadrado. Además $\hat{\beta}$ y $\hat{\sigma}^2$ son los estimadores de máxima verosimilitud.

8.5. Prueba de hipótesis en la Regresión Lineal Múltiple

Nota: Esta sección es tomada del libro Introducción al análisis de regresión lineal de Montgomery, Pecky Vining.

Una vez estimados los parámetros del modelo, surgen de inmediato dos preguntas:

1. ¿Cuál es la adecuación general del modelo?

2. ¿Cuáles regresores específicos parecen importantes?.

Hay varios procedimientos de prueba de hipótesis que demuestran su utilidad para contestar estas preguntas. Las pruebas formales requieren que los errores aleatorios sean independientes y tengan una distribución normal con promedio 0 y varianza constante (σ^2).

8.5.1. Prueba de la significancia de la regresión

La prueba de la significancia de la regresión es para determinar si hay una relación lineal entre la respuesta y cualquiera de las variables regresoras $x_1, x_2, ..., x_k$. Este procedimiento suele considerarse como una prueba general o global de la adecuación del modelo. Las hipótesis pertinentes son:

$$H_0 \quad : \quad \beta_0 = \beta_1 = ... = \beta_k = 0$$

$$H_1 \quad : \quad \beta_j \neq 0 \text{ Para al menos una } j$$

El rechazo de la hipótesis nula implica que al menos uno de los regresores $x_1, x_2, ..., x_k$ contribuye al modelo significativamente. El procedimiento de prueba es una generalización del **análisis de varianza** que se usó en la regresión lineal simple. La **suma de cuadrados total** SC_T se divide en una **suma de cuadrados debida a la regresión**, SC_R, y a una suma de cuadrados de residuales, SC_{Res}. Donde,

$$\begin{aligned} SC_T &= \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\ SC_R &= \hat{\beta}'\mathbf{X}\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\ SC_{Res} &= SC_T - SC_R \end{aligned}$$

Bajo la hipótesis nula cierta, se puede demostrar que SC_R/σ^2 tiene una distribución χ^2_k, donde k es el numero de variables independientes. También SC_{Res}/σ^2 tiene una distribución χ^2_{n-k-1} y que además SC_{Res} y SC_R son independientes. Por lo tanto, de acuerdo con la definición de un estadístico F se tiene que

$$F_0 = \frac{SC_R/k}{SC_{Res}/n-k-1} = \frac{CM_R}{CM_{Res}}$$

tiene una distribución $F_{k,n-k-1}$. Donde $CM_R = SC_R/k$ es el cuadrado medio de la regresión y $CM_{Res} = SC_{Res}/n-k-1$ es el cuadrado medios de los residuales, cuyos valores esperados son respectivamente

$$\begin{aligned} E(CM_R) &= \sigma^2 + \frac{\beta^{*'}\mathbf{X}_c'\mathbf{X}_c\beta^*}{k\sigma^2} \\ E(CM_{Res}) &= \sigma^2 \end{aligned}$$

Siendo $\beta^* = (\beta_1, \beta_2, ..., \beta_k)'$ y $\mathbf{X}_c$ es la matriz "centrada" del modelo, definida por

$$\begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ldots & \vdots \\ x_{i1} - \bar{x}_1 & x_{i2} - \bar{x}_2 & \cdots & x_{ik} - \bar{x}_k \\ \vdots & \vdots & \ldots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

El procedimiento de prueba se resume normalmente en una **tabla de análisis de varianza**, como la tabla ****

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_0
Regresión	SC_R	k	CM_R	$\frac{CM_R}{CM_{Res}}$
Residuales	SC_{Res}	$n-k-1$	CM_{Res}	
Total	SC_T	$n-1$		

Ejemplo 8.3 (Datos del tiempo de entrega) *Se probará la significancia de la re-*

*gresión con los datos del tiempo de entrega del ejemplo ****. Note que*

$$\begin{aligned} SC_T &= \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\ &= 18310,6290 - \frac{(559,60)^2}{25} = 5784,5426 \\ SC_R &= \hat{\beta}'\mathbf{X}\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\ &= 18076,9030 - \frac{(559,60)^2}{25} = 5550,8166 \\ SC_{Res} &= SC_T - SC_R \\ &= 5784,5426 - 5550,8166 = 233,7260 \end{aligned}$$

*El análisis de varianza se muestra en la tabla ***. Para probar* $H_0 : \beta_1 = \beta_2 = 0$*, se calcula el estadístico*

$$F_0 = \frac{CM_R}{CM_{Res}} = \frac{2775,4083}{10,6239} = 261,24$$

Como el valor de F_0 *es mayor al valor tabulado,* $F_{\alpha;k;n-k-1} = F_{0,05;2;22} = 3{,}44$*, entonces se rechaza* H_0*, lo cual implica que el tiempo de entrega depende del volumen de entrega y/o de la distancia. Sin embargo eso no implica necesariamente que la relación que se encontró sea adecuada para predecir el tiempo de entrega en función del volumen y de la distancia. Se requieren más pruebas de adecuación del modelo.*

Como hacerlo en R

Para obtener la tabla de análisis de varianza como la expresada anteriormente es necesario calcular cada uno de sus elementos, para ellos se usan las siguientes instrucciones

- Sumas de cuadrados

```
> SCT<-sum((data\$Y-mean(data\$Y))^2)
> SCR<-sum((objetolm\$fitted-mean(data\$Y))^2)
> SCRes<-sum(objetolm\$residuals^2)
```

Para el ejemplo *** las instrucciones son

```
> SCT<-sum((Datos$resp-mean(Datos$resp))^2)
> SCR<-sum((MRL1$fitted-mean(Datos$resp))^2)
> SCRes<-sum(MRL1$residuals^2)$
```

con lo cual se obtienen los siguientes resultados

```
> 5784.543
> 5550.811
> 233.7317
```

Los cuales son parecidos a los obtenidos haciendo lo cálculos, la diferencia se debe a errores de redondeo.

- Grados de libertad

```
> n<-nrow(cbind(Y))
> GLT<- n-1
```

```
> GLRes<- df.residuals(objetolm())
> GLR<- GLT-GLRes
```

Para el ejemplo *** las instrucciones son

```
> n<-nrow(cbind(resp))
> GLT<- n-1
> GLRes<- df.residual(MRL1)
> GLR<- GLT-GLRes
```

con lo cual se obtienen los siguientes resultados

```
> 24
> 22
> 2
```

- Cuadrados Medios

```
> CMR<-SCR/GLR
> CMRes<-SCRes/GLRes
```

obteniéndose en el ejemplo

```
> 2775.405
> 10.62417
```

- F calculado

```
> Fo<-CMR/CMRes
```

lo cual para el ejemplo se obtiene

```
> 261.2351
```

- Valor P

```
> pv<-1 - pf(F0, GLR,GLRes)
```

que para el ejemplo es

```
> 4.440892e-16
```

los cuales coinciden con los resultados mostrados en la tabla de análisis de varianza (tabla ***).

R^2 y R^2 ajustada

Otras dos maneras de evaluar la adecuación general del modelo son los estadísticos R^2 y R^2 ajustada; esta última se representa por R^2_{Adj}. El R^2 mide la variabilidad de la variable respuesta que es explicada por el modelo, esta dada por

$$R^2 = \frac{SCR}{SCT}$$

La desventaja del R^2 es que por lo general dicha cantidad aumenta cuando se agrega un regresor al modelo, independientemente del valor de la contribución de esa variable. En consecuencia es díficil juzgar si un aumento de R^2 dice en realidad algo importante. Algunas personas que trabajan con modelo de regresión prefieren usar el estadístico R^2_{Adj}, que se define como sigue:

$$R^2_{Adj} = 1 - \frac{SC_R/GL_R}{SC_T/GL_T}$$

En vista de que $\frac{SC_R/GL_R}{SC_T/GL_T}$ es el cuadrado medio de los residuales y SC_T/GL_T es constante, independientemente de cuántas variables hay en el modelo, R^2_{Adj} sólo aumentará al agregar una variable al modelo si esa adición reduce el cuadrado medio residual. En R estos valores son obtenidos al usar la función **summary()**.

8.5.2. Pruebas sobre coeficientes individuales de regresión

Una vez determinado que al menos uno de los regresores es importante, la pregunta lógica es ¿cuál(es) sirve(n) de ellos?. Si se agrega una variable a un modelo de regresión, la suma de cuadrados de la regresión aumenta, y la suma de cuadrados residuales disminuye. Se debe decidir si el aumento de la suma de cuadrados de la regresión es suficiente para garantizar el uso del regresor adicional en el modelo. La adición de un regresor también aumenta la varianza del valor ajustado$\hat{y}$, por lo que se debe tener cuidado de incluir sólo regresores que tenga valor para explicar la respuesta. Además, si se agrega un regresor no importante se puede aumentar el cuadrado medio de residuales, y con eso se disminuya la utilidad del modelo.

Las hipótesis para probar la significancia de cualquier coeficiente individual de regresión, por ejemplo β_j, son

$$H_0 \quad : \quad \beta_j = 0$$

$$H_1 \quad : \quad \beta_j \neq 0$$

Si no se rechaza H_0, quiere decir que se puede eliminar el regresor x_j del modelo. El estadístico de prueba para esta hipótesis es

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\sqrt{var(\hat{\beta}_j)}} \tag{8.9}$$

donde C_{ij} es el j-ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$ que corresponde a $\hat{\beta}_j$. Se rechaza H_0 si $|t_0| > t_{\alpha/2,n-k-1}$. Nótese que ésta es en realidad una prueba parcial o marginal, porque el coeficiente de regresión $\hat{\beta}_j$ depende de todas las demás variables regresoras x_i $(i \neq j)$, que hay en el modelo. Así, se trata de una prueba de la contribución de x_j dados los demás regresores del modelo.

Ejemplo 8.4 *Para ilustrar el procedimiento se usarán los datos de tiempos de entrega del ejemplo ***. Se supone que se desea evaluar la importancia de la variable regresora DISTANCE (distancia,x_2) dado que el regresor CASES (cajas,x_1) está en el modelo. Las hipótesis son*

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_1 &: \beta_2 \neq 0 \end{aligned}$$

El elemento de la diagonal principal de $(\mathbf{X}'\mathbf{X})^{-1}$ que corresponde a β_2 es $C_{22} =$ 0,00000123, por lo que el estadístico de la ecuación 2.9 es

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{0,01438}{\sqrt{(10,6239)(0,00000123)}} = 3,98$$

En vista de que $t_{0,025;22} = 2,074$, se rechaza H_0, y la conclusión es que el regresor DISTANCE, o x_2, contribuye en forma significativa al modelo, dado que CASES, o x_1,

ya está también en el modelo.

Como hacerlo en R

La prueba de hipótesis referidas a coeficientes individuales se obtiene con la instrucción

```
> summary(objetolm())
```

En el ejemplo sería

```
> summary(MRL1)
```

Con lo cual se obtienen diversos resultados (como se explico antes) entre los cuales se encuentran los correspondientes a los parámetros del modelo, y se muestran a continuación

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
x1          1.615907   0.170735   9.464 3.25e-09 ***
x2          0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En dichos resultados se observa, por ejemplo que $\hat{\beta}_2 = 0,014385$, $\sqrt{Var\hat{\beta}_2} = 0{,}003613$, $t_0 = 3,981$ y el valor de P es 0,000631. Los cuales coinciden con los valores obtenidos anteriormente.

Otra alternativa de realizar las pruebas sobre los coeficientes individuales

También se puede determinar directamente la contribución de la suma de cuadrados de un regresor en la regresión, por ejemplo de x_j, dado que otros regresores x_i $(i \neq j)$, están ya en el modelo; para eso se usa el **método de suma extra de cuadrados**. Con este procedimiento también se puede investigar la contribución de un subconjunto de las variables regresoras para el modelo.

Considérese el modelo de regresión con k regresores

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

donde $\mathbf{y}$ es un vector $n \times 1$, $\mathbf{X}$ es una matriz $n \times p$,β es un vector $p \times 1$, ε es un vector $n \times 1$ y $p = k + 1$. Se desea determinar si algún subconjunto de $r < k$ regresores contribuyen en forma significativa al modelo de regresión. Se a seccionado como sigue el vector de los coeficientes de regresión:

$$\beta = \begin{bmatrix} \underline{\beta_1} \\ \beta_2 \end{bmatrix}$$

donde β_1 es un vector $(p-r) \times 1$ y β_2 es un vector $r \times 1$. Se desean probar las siguientes hipótesis

$$H_0 \quad : \quad \beta_2 = 0$$

$$H_1 \quad : \quad \beta_2 \neq 0$$

Este modelo se puede escribir como sigue:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon$$

en el que la matriz $\mathbf{X}_1$ de $n \times (p-r)$ representa a las columnas de $\mathbf{X}$ asociadas con β_1 y la matriz $\mathbf{X}_2$ de $n \times r$ representa a las columnas de $\mathbf{X}$ asociadas con β_2. A éste se le llama el modelo completo.

Para el modelo completo, se sabe que $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$. La suma de cuadrados de regresión para este modelo es

$$SC_R(\beta) = \hat{\beta}'\mathbf{X}'\mathbf{y} \quad (p \text{ grados de libertad})$$

y

$$CM_{Res} = \frac{\mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}}{n-p}$$

Para determinar la contribución de los términos de β_2 a la regreseión se ajusta el modelo suponiendo que es cierta la hipótesis nula $H_0 : \beta_2 = \emptyset$. Este es conocido como el **modelo reducido** y está dado por

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \varepsilon \tag{8.10}$$

El estimador de β_1 por mínimos cuadrados en el modelo reducido es $\hat{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{y}$. La suma de cuadrados de la regresión es

$$SC_R(\beta_1) = \hat{\beta}_1'\mathbf{X}_1'\mathbf{y} \quad (p-r \text{ grados de libertad}) \tag{8.11}$$

La suma de cuadrados de la regresión debida a β_2 dado que β_1 ya está en el modelo es

$$SC_R(\beta_2|\beta_1) = SC_R(\beta) - SC_R(\beta_1) \tag{8.12}$$

con $p-(p-r)=r$ grados de libertad. Esta suma de cuadrados se llama **suma extra de cuadrados debida a** β_2, porque mide el aumenta de la suma de cuadrados de la regresión debida a agregar los regresores $x_{k-r+1}, x_{k-r+2}, ..., x_k$ a un modelo que ya contiene $x_1, x_2, ..., x_{k-r}$. Ahora, $SC_R(\beta_2|\beta_1)$ es independiente del CM_{res}, y se puede probar la hipótesis nula $\beta_2 = \emptyset$ mediante el estadístico

$$F_0 = \frac{SC_R(\beta_2|\beta_1)/r}{CM_{Res}} \tag{8.13}$$

Si $\beta_2 \neq \emptyset$, entonces F_0 sigue una distribución F no central, con parámetro de no centralidad igual a

$$\lambda = \frac{1}{\sigma^2}\beta_2'\mathbf{X}_2'\left[\mathbb{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\right]\mathbf{X}_2\beta_2$$

Este resultado es muy importante. Si hay multicolinealidad en los datos, hay casos en os que β_2 es definitivamente distinto de cero, pero esta prueba en realidad casi no tiene potencia (capacidad para indicar esta diferencia) porque hay una relación casi colineal entre $\mathbf{X}_1$ y $\mathbf{X}_2$. En este caso, λ es casi cero aún cuando β_2 sea realmente importante. Esta relcaión también hace destacar que la máxima potencia de la prueba se alcanza cuando $\mathbf{X}_1$ y $\mathbf{X}_2$ son ortogonales entres sí. Por ortogonales se entiende que $\mathbf{X}_2'\mathbf{X}_1 = \emptyset$. Si $F_0 > F_{\alpha,r,n-p}$, se rechaza H_0 y se concluye que al menos uno de los parámetros en β_2 es distinto de cero, y en consecuencia que al menos uno de los regresores $x_{k-r+1}, x_{k-r+2}, ..., x_k$ en X_2 contribuyen en forma significativa al modelo de regresión.

Algunos autores llaman la prueba 2.13 **prueba parcial F**, o **prueba F parcial**, porque mide la contribución de los regresores en xv_2, dado que los demás regresores en $\mathbf{X}_1$ ya están el el modelo. Para ilustrar la utilidad de este procedimiento, considérese el modelo

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \varepsilon$$

Las sumas de cuadrados

$$SC_R(\beta_1|\beta_0, \beta_2, \beta_3)$$

$$SC_R(\beta_2|\beta_0, \beta_1, \beta_3)$$

y

$$SC_R(\beta_3|\beta_0, \beta_1, \beta_2)$$

son sumas de cuadrados de un grado de libertad que miden la contribución de cada regresor $x_j, j = 1, 2, 3$, al modelo, dado que todos los demás regresores ya estaban en él. Esto es, evalúa la ventaja de agregar x_j a un modelo que no incluía a este regresor. En general, se puede determinar

$$SC_R(\beta_j|\beta_0, \beta_1, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_k), \quad 1 \leq j \leq k$$

que es el aumento de la suma de cuadrados de regresión, debido a agregar x_j a un modelo que ya contiene $x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_k$. Hay quienes creen de utilidad imaginar que esto mide la **contribución de $x - j$ como si fuera la última variable agregada al modelo.**

Se puede demostrar que la prueba F parcial sobre una variable única x_j equivale a la prueba t en 2.9. Sin embargo, la prueba F parcial es un procedimiento más general,

porque se puede medir el efecto de conjuntos de variables. Esta prueba se usa en la formación de modelos, es decir, en la búsqueda del mejor conjunto de regresores que se deben usar en el modelo.

Ejemplo 8.5 *En los datos de tiempo de entrega de gaseosas del ejemplo ***, supóngase que se trata de investigar la contribución de la variables distancia* (x_2) *al modelo. Las hipótesis correspondientes son*

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_1 &: \beta_2 \neq 0 \end{aligned}$$

Para probar estas hipótesis se necesita la suma de cuadrados debida a β_2*, que es*

$$\begin{aligned} SC_R(\beta_2|\beta_1,\beta_0) &= SC_R(\beta_1,\beta_2,\beta_0) - SC_R(\beta_1,\beta_0) \\ &= SC_R(\beta_1,\beta_2|\beta_0) - SC_R(\beta_1|\beta_0) \end{aligned}$$

*De acuerdo con el ejemplo ***,*

$$SC_R(\beta_1,\beta_2|\beta_0) = \hat{\beta}'\mathbf{X}\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 5550,8166$$

con 2 grados de libertad. El modelo reducido $y = \beta_0 + \beta_1 x_1 + \varepsilon$ *se ajusto en el ejemplo ***, y se obtuvo* $\hat{y} = 3,3201 + 2,1762x_1$*. La suma de cuadrados de regresión para este modelo es*

$$SC_R(\beta_1|\beta_0) = \hat{\beta}_1'\mathbf{X}_1\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 5382,4077$$

con 1 grado de libertad. Por consiguiente,

$$SC_R(\beta_2|\beta_1,\beta_0) = 5550,8166 - 5382,4088 = 168,4078$$

Es un aumento de la suma de cuadrados de la regresión, que se debe agregar x_2 al modelo que ya contenía a x_1. Para probar $H_0 : \beta_2 = 0$ se forma el estadístico de prueba

$$F_0 = \frac{SC_R(\beta_2|\beta_1,\beta_0)/1}{CM_{Res}} = \frac{168,4078/1}{10,6239} = 15,85$$

Obsérvese que el CM_{Res} del modelo completo, que contiene a x_1 y x_2, se usa en el denominador del estadístico. Como $F_{0,05;1;22} = 4,30$, se rechaza $H_0 : \beta_2 = 0$ y se concluye que la distancia (x_2) contribuye al modelo en forma significativa.
Como esta prueba F parcial implica a una sola variable, equivale a la prueba t.

Como hacerlo en R

Los rsultados de las pruebas F parciales para cada variable regresora se obtienen directamente de la tabla de análisis de varianza al usar la instrucción

```
> anova(objetolm())
```

Para el ejemplo anterior, al usar la instrucción **anova(MRL1)** se obtiene la tabla **??** En dichos resultados se observa que en la fila correspondiente a la variable x_2 se

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	5382.41	5382.41	506.62	0.0000
x2	1	168.40	168.40	15.85	0.0006
Residuals	22	233.73	10.62		

Tabla 8.2: Anlisis de Varianza

encuentran la suma de cuadrados correspondiente a la agregación de dicha variable

al modelo, el estadístico de prueba y el valor de P los cuales permiten evaluar la significancia del coeficiente β_2.

8.5.3. Prueba de la hipótesis lineal general

Se pueden probar muchas hipótesis acerca de los coeficientes de regresión, si se usa un método unificado. El método de suma extra de cuadrados es un caso especial de este procedimiento. En el procedimiento más general, la suma de cuadrados con la que se calcula la hipótesis es como la diferencia de dos sumas de cuadrados de residuales. A continuación se describirá el procedimiento. Para conocer demostraciones y descripciones más detalladas, consúltese Graybill[1976], Searle[1971] o Seber[1977]. Supóngase que la hipótesis nula de interés se expresa en la forma $H_0 : \mathbf{H}\beta = 0$, donde Hv es una matriz de constantes $q \times p$, tal que sólo r de las q ecuaciones de $\mathbf{H}\beta$ son independientes (es decir $\mathbf{H}$ es de rango r). El modelo completo es $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, siendo $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, y la suma de cuadrados de residuales, para este modelo es

$$SC_{Res}(MC) = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} \quad (n - p \text{ grados de libertad})$$

Para obtener el modelo reducido, se usan las r ecuaciones independientes en $H_0 : \mathbf{H}\beta = 0$ para calcular los r coeficientes de regresión en el modelo completo, en función de los $p - r$ coeficientes restantes de regresión. Esto conduce al modelo reducido $\mathbf{y} = \mathbf{Z}\gamma + \varepsilon$, por ejemplo, donde $\mathbf{Z}$ es una matriz $n \times (p-r)$ y γ es un vector $(p-r) \times 1$, de coeficientes desconocidos de regresión. El estimado de γ es

$$\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

y la suma de cuadrados de residuales, para este modelo es

$$SC_{Res}(MR) = \mathbf{y}'\mathbf{y} - \hat{\gamma}'\mathbf{Z}'\mathbf{y} \quad (n - p + r \text{ grados de libertad})$$

El modelo reducido contiene menos parámetros que el modelo completo, así que $SC_{Res}(MR) \geq SC_{Res}(MC)$. para probar la hipótesis $H_0 : \mathbf{H}\beta = 0$ se emplea la diferencia de sumas de cuadrados de residuales

$$SC_H = SC_{Res}(MR) - SC_{Res}(MC) \tag{8.14}$$

con $n-p+r-(n-p) = r$ grados de libertad. En ella, SC_H se llama suma de cuadrados debida a la hipótesis $H_0 : \mathbf{H}\beta = 0$. El estadístico de prueba para esta hipótesis es

$$F_0 = \frac{SC_H/r}{SC_{Res}(MC)/(n-p)} \tag{8.15}$$

Se rechaza $H_0 : \mathbf{H}\beta = 0$ si $F_0 > F_{\alpha;r;n-p}$.

Prueba de igualdad de coeficientes de regresión

Para probar la igualdad de los coeficientes de regresión se puede usar el método de la hipótesis lineal general. Por ejemplo suponga el siguiente modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Para el modelo completo, SC_{Res} tiene $n-p = n-4$ grados de libertad. Se desea probar $H_0 : \beta_1 = \beta_3$. Esta hipótesis se puede enunciar como $H_0 : \mathbf{H}\beta = 0$, siendo

$$\mathbf{H} = [0, 1, 0, -1]$$

un vector 1×4. Hay sólo una ecuación en $H_0 : \mathbf{H}\beta = 0$, que es $\beta_1 - \beta_3 = 0$. Si se sustituye esta ecuación en el modelo completo, se obtiene el modelo reducido

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_3 + \varepsilon \\ &= \beta_0 + \beta_1(x_1 + x_3) + \beta_2 x_2 + \varepsilon \\ &= \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \varepsilon \end{aligned}$$

donde $\gamma_0 = \beta_0, \gamma_1 = \beta_1 (= \beta_3), z_1 = x_1 + x_3, \gamma_2 = \beta_2$ y $z_2 = x_2$. Al ajustar el modelo reducido se calcularía la $SC_{Res}(MR)$ con $n - 4 + 1 = n - 3$ grados de libertad. La suma de cuadrados debida a la hipótesis $SC_H = SC_{Res}(MR) - SC_{Res}(MC)$ tiene $n - 3 - (n - 4) = 1$ grado de libertad. El cociente F (ecuación 2.15) es

$$F_0 = \frac{SC_H/1}{SC_{Res}(MC)/(n-4)}$$

Notése que esta hipótesis también se podría probar con el estadístico t:

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_3}{\sqrt{var(\hat{\beta}_1 - \hat{\beta}_3)}} = \frac{\hat{\beta}_1 - \hat{\beta}_3}{\sqrt{\hat{\sigma}^2(C_{11} + C_{33} - 2C_{13})}}$$

con $n - 4$ grados de libertad.

8.6. Intervalos de Confianza en Regresión Múltiple

Los intervalos de confianza de los coeficientes de regresión individuales, y los intervalos de confianza de la respuesta media, para niveles específicos de los regresores, juegan el mismo papel importante que en la regresión lineal simple. En esta sección se desarrollan los intervalos de confianza, uno por uno, para estos casos. También se presentarán en forma breve los intervalos de confianza simultáneos para los coeficientes de regresión.

8.6.1. Intervalos de confianza de los coeficientes de regresión

Para construir intervalos de confianza de los coeficientes de regresión β_j, se continuará suponiendo que los errores ε_i están distribuidos normal e independientemente, con media cero y varianza σ^2. En consecuencia, las observaciones y_i están distribuidas en forma normal e independientemente, con media $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ y varianza σ^2. Como el estimador $\hat{\beta}$ por mínimos cuadrados es una combinación lineal de las observaciones, también está distribuido normalmente, con media β y matriz de covarianza $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Esto implica que la distribución marginal de cualquier coeficiente de regresión $\hat{\beta}_j$ es normal, con media β_j y varianza $\sigma^2 C_{jj}$, donde C_{jj} es el $j-$ésimo elemento diagonal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. En consecuencia, cada una de los estadísticos

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \quad j = 0, 1, 2, ..., k \tag{8.16}$$

se distribuye t-student con $n - p$ grados de libertad, donde $\hat{\sigma}^2$ es el estimador de la varianza.

De acuerdo con el resultado de la ecuación 2.16 se puede definir un intervalo de confianza

de $100(1-\alpha)$ por ciento para el coeficiente de regresión $\beta_j, j = 0, 1, ..., k$, como sigue

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \tag{8.17}$$

Ejemplo 8.6 *Se calculará un intervalo de confianza del 95 % para el parámetro* β_1 *en el ejemplo ***. La estimación puntual de* β_1 *es* $\hat{\beta}_1 = 10,6239$ *(de acuerdo con el ejemplo ***). Se aplica la ecuación 2.17 y se ve que*

$$\hat{\beta}_1 - t_{0,025;22}\sqrt{\hat{\sigma}^2 C_{11}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0,025;22}\sqrt{\hat{\sigma}^2 C_{11}}$$

$$1,61591 - (2,074)\sqrt{(10,6239)(0,00274378)} \leq \beta_1 \leq 1,61591 + (2,074)\sqrt{(10,6239)(0,00274378)}$$

$$1,61591 - (2,074)(0,17073) \leq \beta_1 \leq 1,61591 + (2,074)(0,17073)$$

y el intervalo de confianza de 95 % para β_1 *es*

$$1,26181 \leq \beta_1 \leq 1,97001$$

8.6.2. Intervalo de confianza de la respuesta media

Se puede establecer un intervalo de confianza para la respuesta media en determinado punto, como $x_{01}, x_{02}, ..., x_{0k}$. Defínase el vector $\mathbf{x}_0$ como sigue

$$\mathbf{x}_0 = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{pmatrix}$$

El valor ajustado en este punto es

$$\hat{\mathbf{y}}_0 = \mathbf{x}_0'\hat{\beta} \tag{8.18}$$

Es un estimador insesgado de $E(y|\mathbf{x}_0)$, porque $E(\hat{y}_0) = \mathbf{x}_0'\beta = E(y|\mathbf{x}_0)$, la varianza de $\hat{y}_0$ es

$$Var(\hat{y}_0) = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \tag{8.19}$$

Por consiguiente, un intervalo de confianza de $100(1-\alpha)$ por ciento de la respuesta media en el punto $x_{01}, x_{02}, ..., x_{0k}$ es

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{Var(\hat{y}_0)} \leq E(y|\mathbf{x}_0) \leq \hat{y}_0 - t_{\alpha/2,n-p}\sqrt{Var(\hat{y}_0)} \tag{8.20}$$

Ejemplo 8.7 *El embotellador de gaseosas del ejemplo *** quiere establecer un intervalo de confianza de 95 % para el tiempo medio*

8.6.3. Intervalos de confianza simultáneos para coeficientes de regresión

Se han descrito los procedimientos para establecer diversos tipos de intervalos de confianza y de predicción para el modelo de regresión lineal. Se ha hecho notar que éstos son intervalos de uno por uno, estop es, son los tipos usuales de intervalo de confianza o de predicción, en donde el coeficiente de confianza $1 - \alpha$ indica la proporción de estimaciones correctas que resulta cuando se seleccionan muestras aleatorias repetidas. En algunos problemas se necesita construir varios intervalos de confianza o de predicción con los mismos datos de la muestra. En esos casos, el analista suele interesarse en la especificación de un coeficiente de confianza que se aplique en forma simultánea, o al mismo tiempo, a todo el conjunto de estimaciones por intervalo. Un conjunto de intervalos de confianza o predicción que son todos ciertos en forma simultánea, con $1 - \alpha$ de probabilidad, se llama conjunto de intervalos **simultáneos** o **conjuntos de confianza** o **de predicción**.

Por ejemplo, se tiene un modelo de regresión lineal simple. Suponga que el analista desea sacar inferencias acerca de la ordenada al origen β_0 y la pendiente β_1, una posibilidad sería establecer intervalos de confianza, por ejemplo de 95 %, para ambos parámetros, sin embargo, si esos estimados son independientes, la probabilidad de que ambas afirmaciones sean correctas es $(0,95)^2 = 0,9025$. Así, no se tiene un nivel de confianza de 95 % asociado con ambas afirmaciones. Además, como los intervalos se establecen usando el mismo conjunto de datos muestrales, no son independientes. Esto introduce mayor complicación en la determinación del nivel de confianza para el conjunto de afirmaciones.

Es relativamente fácil definir una región de confianza conjunta para los parámetros β

del modelo de regresión múltiple. Se puede demostrar que

$$\frac{(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)}{pCM_{Res}} \sim F_{p,n-p}$$

y eso implica que

$$P\left[\frac{(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)}{pCM_{Res}} \leq F_{\alpha,p,n-p}\right] = 1-\alpha$$

En consecuencia, una región de confianza conjunta de $100(1-\alpha)$ por ciento, para todos los parámetros en β es

$$\frac{(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)}{pCM_{Res}} \leq F_{\alpha,p,n-p} \tag{8.21}$$

Esta desigualdad describe una región de forma elíptica.

8.7. Otras Funciones de R

Para realizar las pruebas de hipótesis y encontrar los intervalos de confianza que no se obtienen directamente a partir de la instrucción lm(), se usan operaciones básicas de matrices y el uso de las formulas antes descritas. A continuación se muestran algunas de las instrucciones usadas.

8.7.1. Definición de una matriz en R

Recuerde que una matriz $A_{m\times n}$ es un arreglo rectangular de n filas y m columnas, es decir

$$A = \begin{pmatrix} a_{11} & a_{11} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

En R una matriz se define usando la función matrix(), cuya sintaxis es

matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE)

donde

- data: es un vector de datos
- nrow: es el número de filas deseadas
- ncol: es el número de columnas deseadas
- byrow: es una variable lógica. Si es "FALSE" (por defecto) la matriz es llenada por columnas, en caso contrario es llenada por filas.

Si se quiere definir una matriz $A_{3\times 3}$ se usa la siguiente instruccción

```
> A<-matrix(c(a11,a12,a13,a21,a22,a23,a31,a32,a33),
  nrow=3,ncol=3,byrow=TRUE)
```

con lo cual se obtiene

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{11} & a_{11} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Ejemplo 8.8 *Para construir la matriz*

$$\mathbf{X} = \begin{pmatrix} 16 & 8 & 12 & -4 \\ 8 & 5 & 11 & -4 \\ 12 & 11 & 70 & -31 \\ -4 & -4 & -31 & 63 \end{pmatrix}$$

se usa la siguiente instrucción

```
> X<-matrix(c(16,8,12,-4,8,5,11,-4,12,11,70,-31,-4,-4,-31,63),
  nrow=4,ncol=4,byrow=TRUE)
```

8.7.2. Operaciones de matrices en R

En la siguiente tabla se muestran las operaciones básicas entre matrices que necesarias para los cálculos en un modelo lineal general

Tabla 8.3: Operaciones bsicas sobre matrices

Operación	Operador	Ejemplo
Suma	+	$\boldsymbol{A} + \boldsymbol{B}$
Resta	-	$\boldsymbol{A} - \boldsymbol{B}$
Multiplicación	$\%*\%$	$\boldsymbol{A} \%*\% \boldsymbol{B}$

Ejemplo 8.9 *Sean las matrices* $\boldsymbol{A}$ *y* $\boldsymbol{B}$ *dadas a continuación*

$$\boldsymbol{A} = \begin{pmatrix} 16 & 8 & 12 & -4 \\ 8 & 5 & 11 & -4 \\ 12 & 11 & 70 & -31 \\ -4 & -4 & -31 & 63 \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} 6 & 4 & 2 & -4 \\ 8 & 5 & 1 & -4 \\ 2 & 1 & 7 & -3 \\ -4 & -4 & -1 & 3 \end{pmatrix}$$

Se esta interesado en hallar $\boldsymbol{A} + \boldsymbol{B}$, $\boldsymbol{A} - \boldsymbol{B}$ *y* $\boldsymbol{A} \% * \% \boldsymbol{B}$.

Para crear las matrices $\boldsymbol{A}$ *y* $\boldsymbol{B}$ *se usan las siguientes instrucciones*

```
> A<-matrix(c(16,8,12,-4,8,5,11,-4,12,11,70,-31,-4,-4,-31,63),
    nrow=4,ncol=4,byrow=TRUE)
> B<-matrix(c(6,4,2,-4,8,5,1,-4,2,1,7,-3,-4,-4,-1,3),nrow=4,ncol=4,
    byrow=TRUE)
```

luego,

- *Para la suma se usa la siguiente instrucción*

  ```
  > A+B
  ```

  Con lo que se obtiene

  ```
       [,1] [,2] [,3] [,4]
  [1,]   22   12   14   -8
  [2,]   16   10   12   -8
  [3,]   14   12   77  -34
  [4,]   -8   -8  -32   66
  ```

- *Para la resta se usa la siguiente instrucción*

```
> A-B
```

Con lo que se obtiene

```
     [,1] [,2] [,3] [,4]
[1,]   10    4   10    0
[2,]    0    0   10    0
[3,]   10   10   63  -28
[4,]    0    0  -30   60
```

- *Para la multiplicación se usa la siguiente instrucción*

```
> A%*%B
```

Con lo que se obtiene

```
     [,1] [,2] [,3] [,4]
[1,]  200  132  128 -144
[2,]  126   84  102  -97
[3,]  424  297  556 -395
[4,] -370 -319 -292  314
```

8.7.3. Operaciones de matrices en R

Al igual que en el caso de las operaciones a continuación se muestran sólo las funciones necesarias en el modelo lineal general

Tabla 8.4: Funciones bsicas sobre matrices

Función	**Operador**	**Ejemplo**
Traspuesta	t()	$t(\boldsymbol{A})$
Inversa	solve()	$solve(\boldsymbol{A})$

Ejemplo 8.10 *Para la matriz* $\boldsymbol{A}$ *definida en el ejemplo anterior, se tiene que*

- *Para hallar la traspuesta de* $\boldsymbol{A}$ $(\boldsymbol{A}')$ *se usa la siguiente instrucción*

  ```
  > t(A)
  ```

  obteniendose

  ```
       [,1] [,2] [,3] [,4]
  [1,]   16    8   12   -4
  [2,]    8    5   11   -4
  [3,]   12   11   70  -31
  [4,]   -4   -4  -31   63
  ```

- *Para hallar la inversa de* $\boldsymbol{A}$ $(\boldsymbol{A}^{-1})$ *se usa la siguiente instrucción*

  ```
  > solve(A)
  ```

  obteniéndose

  ```
                [,1]         [,2]        [,3]         [,4]
  [1,]  0.397888322 -0.74433107  0.04988662  0.002551020
  [2,] -0.744331066  1.69954649 -0.14399093 -0.010204082
  [3,]  0.049886621 -0.14399093  0.03287982  0.010204082
  [4,]  0.002551020 -0.01020408  0.01020408  0.020408163
  ```

8.7.4. Valores tabulados y P valor

Para obtener los valores tabulados y el P valor de la distribución t-Student se usan las siguientes instrucciones

```
> qt(probabilidad, grados de libertad, lambda, lower.tail = TRUE)
> pt(valor de t, grados de libertad, lambda, lower.tail = TRUE)
```

lower.tail = TRUE en caso de que las probabilidades son $P[X <= x]$, de lo contrario, $P[X > x]$.

8.8. Ejercicios

1. Para los datos de la Liga Nacional de Fútbol:

   a) Ajustar un modelo de regresión lineal múltiple que relacione la cantidad de juegos ganados con las yardas por aire del equipo (x_2), el porcentaje de jugadas por tierra (x_7) y las yardas por tierra del contrario (x_8).

   b) Formar la tabla de análisis de varianza y probar la significancia de la regresión.

   c) Calcular el estadístico t para probar las hipótesis $H_0 : \beta_2 = 0, H_0 : \beta_7 = 0$ y $H_0 : \beta_8 = 0$. ¿Qué conclusiones se pueden sacar acerca del papel de las variables x_2, x_7 y x_8 en el modelo?.

   d) Calcular R^2 y R^2_{Adj} para este modelo.

   e) Con la prueba F parcial, determinar la contribución de x_7 al modelo. ¿Cómo se relaciona el estadístico F parcial con la prueba t calculada en el inciso c.?

   f) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

   g) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

h) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

i) Calcular un intervalo de confianza de 95 % para β_7 y un intervalo de confianza de 95 % para la cantidad media de juegos ganados por un equipo cuando $x_2 = 2300, x_7 = 56$ y $x_8 = 2100$.

j) ajustar un modelo a esos datos, usando solo x_7 y x_8 como regresores y probar la significancia de la regresión.

k) Calcular R^2 y R^2_{Adj}. ¿Compararlos con los resultados del modelo anterior.

l) Calcular un intervalo de confianza de 95 % para β_7. También, un intervalo de confianza de 95 % para la cantidad media de juegos ganados por un equipo cuando $x_7 = 56$ y $x_8 = 2100$. Comparar la longitudes de esos intervalos de confianza con las longitudes de los correspondientes al modelo anterior.

m) ¿Qué conclusiones se pueden sacar de este problema, acerca de las consecuencias de omitir un regresor importante de un modelo?

2. Véase los datos de rendimiento de gasolina.

a) Ajustar un modelo de regresión lineal múltiple que relacione el rendimiento de la gasolina y, en millas por galón, la cilindradada del motor (x_1), y la cantidad de gargantas del carburador, (x_6).

b) Formar la tabla de análisis de varianza y probar la significancia de la regresión.

c) Calcular R^2 y R^2_{Adj} para este modelo. Compararlas con las R^2 y R^2_{Adj} para el modelo de regresión lineal simple, que relaciona las millas con la cilindrada.

d) Determinar un intervalo de confianza para β_1.

e) Determinar un intervalo de confianza de 95 % para el rendimiento promedio de la gasolina, cuando $x_1 = 225\ pulg^3$ y $x_6 = 2\ gargantas$.

f) Determinar un intervalo de predicción de 95 % para una nueva observación de rendimiento de gasolina, cuando $x_1 = 225\ pulg^3$ y $x_6 = 2\ gargantas$.

g) Considere el modelo de regresión lineal simple, que relaciona las millas con la cilindrada.Contnstruya un intervalo de confianza de 95 % para el rendimiento promedio de la gasolina y un intervalo de predicción para el rendimiento, cuando $x_1 = 225\ pulg^3$. Compara las longitudes de estos intervalos con los intervalos obtenidos en los dos incisos anteriores. ¿Tiene ventajas agregar x_6 al modelo.

h) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

i) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

j) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

3. Véase los datos sobre precios de viviendas

a) Ajustar un modelo de regresión lineal múltiple que relacione el precio de venta con los nueve regresores.

b) Probar la significancia de la regresión.¿Qué conclusiones se pueden sacar?

c) Usar pruebas t para evaluar la contribución de cada regresor al modelo.

d) Calcular R^2 y R^2_{Adj} para este modelo.

e) ¿Cuál es la contribución del tamaño del lote y el espacio vital para el modelo, dado que se incluyeron todos los demás regresores?.

f) En este modelo, ¿la colinealidad es un problema potencial?.

g) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

h) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

i) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

4. Para los datos sobre la eficiencia de un proceso químico, en función de varias variables controlables del proceso se pide

a) Ajustar un modelo de regresión lineal múltiple que relacione el CO_2 del producto (y) con el solvente total (x_6) y el consumo de hidrógeno (x_7).

b) Probar la significancia de la regresión.

c) Calcular R^2 y R^2_{Adj} para este modelo.

d) Usar pruebas t para evaluar la contribución de x_6 y x_7 al modelo.

e) Establecer intervalos de confianza de 95 % para β_6 y β_7.

f) Volver a ajustar el modelo sólo con x_6 como regresor. Probar la significancia de la regresión y calcular R^2 y R^2_{Adj}. Comentar los resultados. Con base en estos estadísticos, ¿es satisfactorio el modelo?.

g) Establecer un intervalo de confianza de 95 % para β_6, con el modelo que se ajustó en el inciso d. ¿Se deduce algo importante acerca de la contribución de x_7 al modelo?.

h) Comparar los valores de CM_{Res} obtenidos con los dos modelos que se ajustaron (partes a y e). ¿Cómo cambio el CM_{Res} al quitar $x-7$ del modelo? ¿Indica lo anterior algo importante acerca de la contribución de x_7 al modelo?.

i) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

j) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

k) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

a) En los datos se muestra la concentración de $NbOCL_3$ en un reactor de tubo de flujo, en función de varias variables controlables.

b) Ajustar un modelo de regresión lineal múltiple que relacione la concentración de $NbOCL_3$ (y) con la $COCL_(2)$ (x_1) y la fracción mol (x_4).

c) Probar la significancia de la regresión.

d) Calcular R^2 y R^2_{Adj} para este modelo.

e) Usar pruebas t para evaluar la contribución de x_6 y x_7 al modelo.

f) Con pruebas t, determinar la contribución de x_1 y x_4 al modelo. ¿Son necesarios los dos regresores?

g) En este problema, ¿es la colinealidad un problema potencial?

h) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

i) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

j) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

5. Se cree que la calidad del vino Pinot Noir se relaciona con sus propiedades de claridad, aroma, cuerpo, sabor y fuerza. Se registraron los datos de 38 vinos.

a) Ajustar un modelo de regresión lineal múltiple que relacione la calidad del vino con esos regresores.

b) Probar la significancia de la regresión. ¿A qué conclusiones se puede llegar?

c) Use pruebas t para evaluar la contribución de cada regresor al modelo. Comentar los resultados.

d) Calcular R^2 y R^2_{Adj}. Comparar esos valores con R^2 y R^2_{Adj} para el modelo de regresión lineal que relacione la calidad del vino con su aroma y sabor. Comentar los resultados.

e) Determinar un intervalo de confianza de 95 % para el coeficiente de regresión del sabor, para los dos modelos de la parte d. Comentar las diferencias encontradas.

f) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

g) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

h) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

6. Un ingeniero hizo un experimento para determinar la presión, temperatura y flujo de $C =_2$, la humedad y el tamaño de partícula de los cacahuates sobre el rendimiento total de aceite por lote de cacahuates.

a) Ajustar un modelo de regresión lineal múltiple que relacione el rendimiento con esos regresores.

b) Probar la significancia de la regresión. ¿A qué conclusiones se puede llegar?

c) Hacer pruebas t para evaluar la contribución de cada regresor al modelo. Comentar los resultados.

d) Calcular R^2 y R^2_{Adj}. Comparar esos valores con R^2 y R^2_{Adj} para el modelo de regresión lineal que relacione el rendimiento con la temperatura y el tamaño de partícula. Comentar los resultados.

e) Establecer un intervalo de confianza de 95 % para el coeficiente de regresión de la temperatura, para los dos modelos de la parte d. Comentar las diferencias encontradas.

f) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

g) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

h) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

7. Un ingeniero químico estudió el efecto de la cantidad de surfactante y el tiempo sobre la formación de catrato. Los catratos se usan como medio de conservación en frío.

   a) Ajustar un modelo de regresión lineal múltiple que relacione la formación de catrato con esos regresores.

   b) Probar la significancia de la regresión. ¿A qué conclusiones se puede llegar?

   c) Hacer pruebas t para evaluar la contribución de cada regresor al modelo. Comentar los resultados.

   d) Calcular R^2 y R^2_{Adj}. Comparar esos valores con R^2 y R^2_{Adj} para el modelo de regresión lineal que relacione la formación de catrato con el tiempo. Comentar los resultados.

   e) Establecer un intervalo de confianza de 95 % para el coeficiente de regresión del tiempo, para los dos modelos de la parte d. Comentar las diferencias encontradas.

   f) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

   g) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

   h) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el

regresor?.

8. Un ingeniero estudió el efecto de cuatro variables de un factor adimensional con el que se describen las caídas de presión en una columna de burbujeo de platos perforados. Los catratos se usan como medio de conservación en frío.

   a) Ajustar un modelo de regresión lineal múltiple que relacione ese número adimensional con los cuatro regresores.

   b) Probar la significancia de la regresión. ¿A qué conclusiones se puede llegar?

   c) Hacer pruebas t para evaluar la contribución de cada regresor al modelo. Comentar los resultados.

   d) Calcular R^2 y R^2_{Adj}. Comparar esos valores con R^2 y R^2_{Adj} para el modelo de regresión lineal que relacione el número adimensional con x_2 y x_3. Comentar los resultados.

   e) Determinar un intervalo de confianza de 99 % para el coeficiente de regresión de x_2, para los dos modelos de la parte d. Comentar las diferencias encontradas.

   f) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

   g) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

   h) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

9. La viscosidad cinemática de cierto sistema de solventes depende de la relación entre los dos solventes y la temperatura.

   a) Ajustar un modelo de regresión lineal múltiple que relacione la viscosidad con los dos regresores.

   b) Probar la significancia de la regresión. ¿A qué conclusiones se puede llegar?

   c) Hacer pruebas t para evaluar la contribución de cada regresor al modelo. Comentar los resultados.

   d) Calcular R^2 y R^2_{Adj}. Comparar esos valores con R^2 y R^2_{Adj} para el modelo de regresión lineal que relacione la viscosidad sólo con la temperatura. Comentar los resultados.

   e) Establecer un intervalo de confianza de 99 % para el coeficiente de regresión de la temperatura, para los dos modelos de la parte d. Comentar las diferencias encontradas.

   f) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

   g) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

   h) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.