CAPÍTULO 2

REGRESIÓN LINEAL MULTIPLE

2.1 Introducción

Es evidente que lo más económico y rápido para modelar el comportamiento de una variable Y es usar una sola variable predictora y usar un modelo lineal. Pero, algunas veces es bastante obvio de que el comportamiento de Y es imposible que sea explicada en gran medida por solo una variable. Por ejemplo, es imposible tratar de explicar el rendimiento de un estudiante en un examen, teniendo en cuenta solamente el número de horas que se preparó para ella. Claramente, el promedio académico del estudiante, la carga académica que lleva, el año de estudios, son tres de las muchas otras variables que pueden explicar su rendimiento. Tratar de explicar el comportamiento de Y con más de una variable predictora usando una funcional lineal es el objetivo de regresión lineal multiple.

Frecuentemente, uno no es muy familiar con las variables que están en juego y basa sus conclusiones solamente en cálculos obtenidos con los datos tomados. Es decir, si ocurre que el coeficiente de determinación R^2 sale bajo (digamos menor de un 30%) , considerando además que su valor no se ha visto afectado por datos anormales, entonces el modelo es pobre y para mejorarlo hay tres alternativas que frecuentemente se usan:

a) Transformar la variable predictora, o la variable de respuesta Y, o ambas y usar luego un modelo lineal.
b) Usar regresión polinómica con una variable predictora.
c) Conseguir más variables predictoras y usar una regresión lineal múltiple.

En el primer caso, se puede perder el tiempo tratando de encontrar la transformación más adecuada y se podria caer en sobre-ajuste (***"overfitting"***), es decir, encontrar un modelo demasiado optimista, que satisface demasiado la tendencia de los datos tomados pero que es pobre para hacer predicciones debido a que tiene una varianza grande.
En el segundo caso el ajuste es más rápido, pero es bien fácil caer en ***"overfitting"*** y, además se pueden crear muchos problemas de cálculo ya que pueden surgir problemas de colinealidad, es decir relación lineal entre los terminos del modelo polinómico.
El tercer caso es tal vez la alternativa más usada y conveniente. Tiene bastante analogía con el caso simple, pero requiere el uso de vectores y matrices.
En el siguiente ejemplo se mostrará el uso interactivo de las tres alternativas a través de seis modelos de regresión y servirá como un ejemplo de motivación para introducirnos en regresión lineal multiple.

Ejemplo 1: Considerar el conjunto de datos **millaje** donde la variable de respuesta esY= (MPG) millas promedio por galón de un auto, y las variables predictotras son; X_1=(VOL): Capacidad en volumen del carro, X_2=(HP): Potencia del Motor, X_3= (SP) :Velocidad Máxima y X_4=(WT): Peso del auto. Los datos fueron adaptados de la "Data and Story Library" (lib.stat.cmu.edu/DASL/) y están dsponibles en academic.uprm.edu/eacuna/millaje.txt.

Primero, se explorará las relaciones entre todas las parejas de variables, en particular la relación de Y con cada una de las variables predictoras. Esto se logra con una gráfica lllamada **plot matricial**, la cual está disponible en la mayoría de programas estadísticos. La function **pairs** de **R**

produce el plot matricial para las variables del ejemplo 1 tal como se muestra en la siguiente figura:

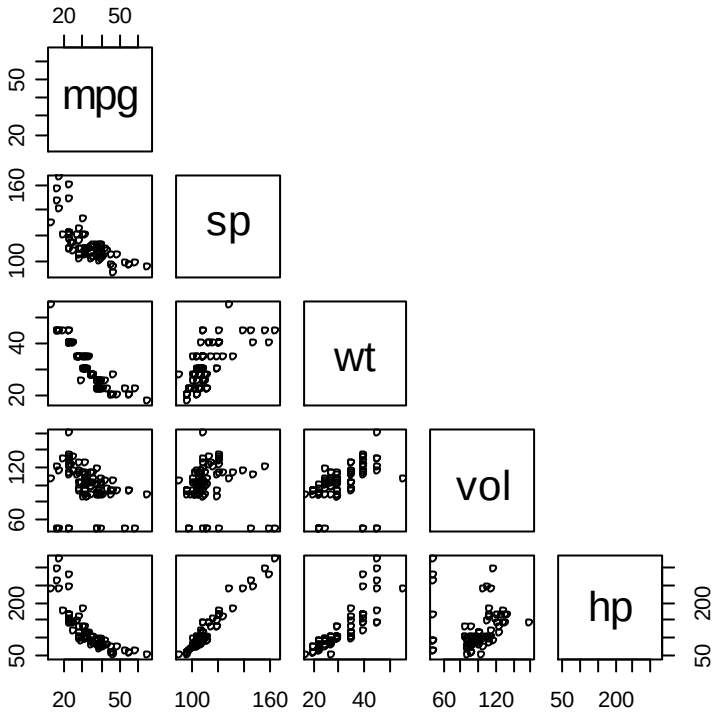

Figura 2.1. Plot matricial de las variables del conjutnode datos **millaje.**

Claramente se puede ver que la variable predictora (WT) es la que tiene mejor relación lineal con MPG y que VOL tiene una pobre relación lineal con MPG. En tanto que para HP y SP la relación lineal no es muy marcada.

Ahora , analizemos la relación entre HP y MPG. Un plot de estas variables se muestra en la figura 2.2. Si hacemos la regresión lineal entre las dos variables se obtiene

```
>  l1<-lm(mpg~hp)
> l1

Call:
lm(formula = mpg ~ hp)

Coefficients:
(Intercept)           hp
    50.0661      -0.1390

> summary(l1)$r.squared
[1] 0.6239
```

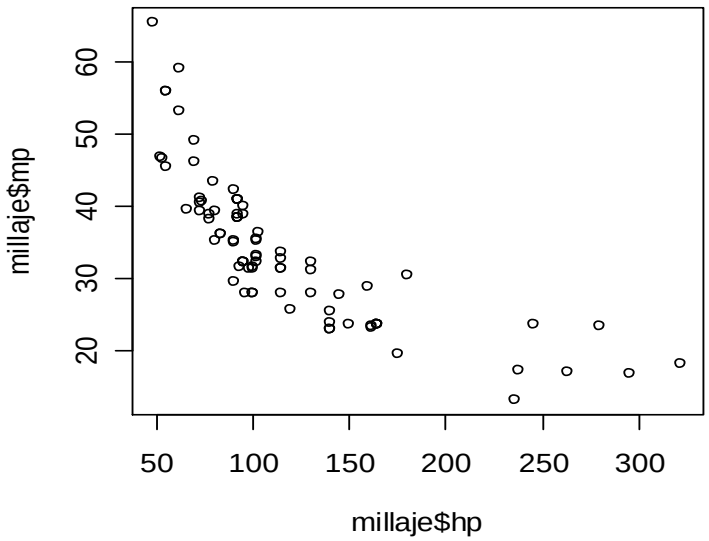

Figura 2.2. Plot de MPG versus HP.

El R^2 =62.4% no está bajo, pero hay que tratar de mejorarlo, usando las alternativas a y b. En la gráfica se observa una curvatura, así que se podría ajustar una regresión cuadrática. Los resultados usando hp y hp2=hp^2 como variables predictoras son los siguientes:

```
> l2=lm ( mpg ~ hp + hp2)
> l2

Call:
lm(formula = mpg ~ hp + hp2)

Coefficients:
(Intercept)          hp         hp2
 71.2313424   -0.4598708    0.0009707

> summary(l2)$r.squared
[1] 0.8067
```

El R^2 resulta ser 80.7% lo que representa una gran mejora, pero hay un peligro de hacer predicciones porque al final la cuadrática tiende a subir, y se podría concluir que un auto con 400 HP podría tener un rendimiento de 42.59 millas por galón, similar al de un carro de 50 HP. Este es un ejemplo de un modelo sobreajustado ("overfitted"). Notar también el valor bien pequeño del coeficiente del término cuadrático, el cual podría causar problema en el cálculo de las predicciones.

Observando más detenidamente la gráfica de la figura 2.2 se puede ver que hay un comportamiento asintótico en la parte inferior, es decir, que después de cierto nivel de HP, el millaje tiende a estabilizarse. Esto sugiere que podriamos tratar un modelo hiperbólico de la

forma MPG=$\alpha+\beta\frac{1}{HP}$ para ajustar los datos. Considerando la predictora hp1=1/hp, se obtienen los siguientes resultados en R.

```
> l3=lm(mpg~hp1)
> l3

Call:
lm(formula = mpg ~ hp1)

Coefficients:
(Intercept)         hp1
      9.73     2373.11

> summary(l3)$r.squared
[1] 0.8429
```

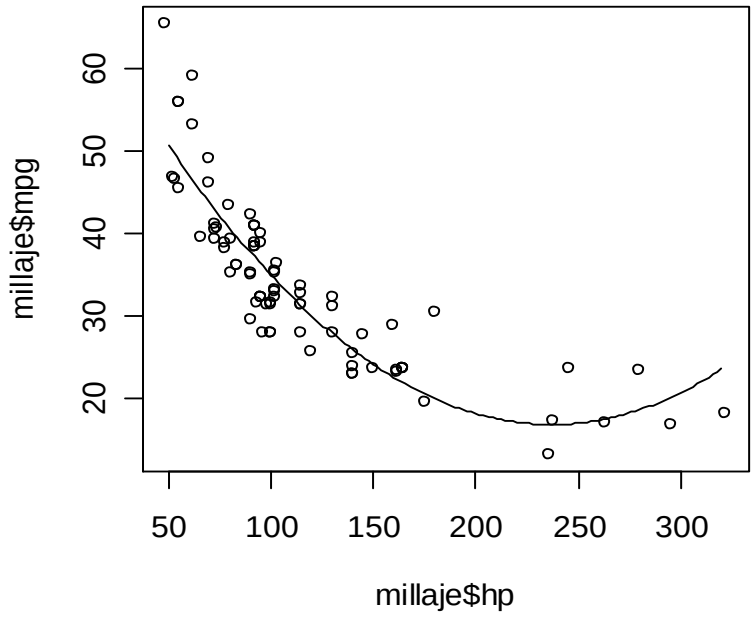

Figura 2.3. Regresión cuadrática de MPG versus HP

Notar que el R^2 =84.3% está bastante aceptable lo cual indica un buen ajuste del modelo. Así para un carro con 400 de HP su MPG será 15.66.

Alguién que no quiere perder el tiempo explorando relaciones polinómicas o haciendo transformación de variables, tratará de conseguir información acerca de otras variables

predictoras, con la esperanza de subir sustancialmente su R^2 pero usando solamente modelos lineales.
Del plot matricial que aparece en la figura 2.1 no hay ninguna duda de que la variable a considerar conjuntamente con HP sería WT. Haciendo uso de R se obtiene

```
> l4<-lm(mpg~hp+wt)
> l4

Call:
lm(formula = mpg ~ hp + wt)

Coefficients:
(Intercept)          hp          wt
   66.85500     -0.02097     -0.99037

> summary(l4)$r.squared
[1] 0.8235
```

El cual sería el segundo mejor modelo usando el criterio de R^2 ya que produce un valor de 82.4%. Si usamos el hecho de que la relación de Y con HP1 resulta ser bastante buena, podemos intentar ajustar un modelo lineal con HP1 y WT como las variables predictoras. Los resultados aparecen a continuación:

```
> l5<-lm(mpg~hp1+wt)
> l5

Call:
lm(formula = mpg ~ hp1 + wt)

Coefficients:
(Intercept)          hp1          wt
    36.5361    1387.1768      -0.5439

> summary(l5)$r.squared
[1] 0.8933
```

Este último sería el mejor modelo hasta ahora ya que su R^2=89.9 es el mayor de todos. Así se puede seguir explorando más modelos, pero teniendo cuidado de no caer en ***"overfitting"***.

Si ajustamos un modelo de regresión lineal multiple con las 4 variables predictoras disponibles se obtiene

```
> l6<-lm(mpg~vol+hp+sp+wt)
> l6

Call:
lm(formula = mpg ~ vol + hp + sp + wt)

Coefficients:
(Intercept)          vol          hp          sp          wt
  192.43775     -0.01565      0.39221     -1.29482     -1.85980

> summary(l6)$r.squared
```

[1] 0.8733

Si habría que decidir entre este último modelo y el anterior, habría que escoger el anterior porqué con solo dos variables predictoras se obtiene un mejor R^2.

2.2 El modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple con p variables predictoras y basado en n observaciones tomadas es de la forma

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} + e_i \quad (2.1)$$

para i=1,2,….n. Escribiendo el modelo para cada una de las observaciones, éste puede ser considerado como un sistema de ecuaciones lineales de la forma

$$y_1 = \beta_o + \beta_1 x_{11} + \beta_2 x_{12} + ...\beta_p x_{1p} + e_1$$
$$y_2 = \beta_o + \beta_1 x_{21} + \beta_2 x_{22} + ...\beta_p x_{2p} + e_2$$
……..
$$y_n = \beta_o + \beta_1 x_{n1} + \beta_2 x_{n2} + ...\beta_p x_{np} + e_n$$

que puede ser escrita en forma matricial como

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . & . \\ .. & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

O sea,

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e} \quad (2.2)$$

donde **Y** es un vector columna n dimensional**, X** es una matriz n x p', con p'=p+1, **β** es el vector de coeficientes de regresión a ser estimados, su dimensión es p' y **e** es un vector columna aleatorio de dimensión n

Por ahora, las únicas suposiciones que se requieren son que E(**e**)=**0** y que la matriz de varianza-covarianzas de los errores está dada por Var(**e**)=$\sigma^2\mathbf{I_n}$, donde $\mathbf{I_n}$ es la matriz identidad de orden n.

2.2.1 Estimación del vector de parámetros β por Cuadrados Mínimos

Al igual que en regresión lineal simple hay que minimizar la suma de cuadrados de los errores. La suma de cuadrados de los errores puede ser expresada vectorialmente de la siguiente manera

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (2.3)$$

donde el símbolo ' indica transpuesta del vector o matriz (es decir, la matriz que se obtiene intercambiando las fila por columnas en la matriz original). Haciendo operaciones con los vectores y matrices se obtiene

$$Q(\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \quad (2.4)$$

En la igualdad anterior se ha usado la propiedad **(AB)'=B'A'.** Derivando Q con respecto a β e igualando a cero se obtiene el sistema de ecuaciones normales;

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \quad (2.5)$$

de donde resolviendo para β se obtiene

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (2.6)$$

aqui $(\mathbf{X'X})^{-1}$ representa la matriz inversa de **(X'X).** Notar que **X'X** es simétrica, pues su transpuesta da la misma matriz.

En la regresión lineal simple, p=1 y el modelo puede ser escrito en forma matricial como

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ & \\ & \\ & \\ 1 & x_{n1} \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Manipulando las matrices se obtiene que

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_{11} & x_{21} & . & . & . & x_{n1} \end{bmatrix} \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ & \\ & \\ & \\ 1 & x_{n1} \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{bmatrix}$$

$$\mathbf{X'\,Y} = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_{11} & x_{21} & . & . & . & x_{n1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \end{bmatrix}$$

Luego las ecuaciones normales se reducen a:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \end{bmatrix}$$

Por comodidad podemos eliminar el segundo subindice de las x's ya que no afecta en nada. Como

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

se concluye que

$$\begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{nSxx} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \\ -\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + n\sum_{i=1}^{n} x_i y_i \end{bmatrix} \quad (2.7)$$

y haciendo manipuleo algebraíco se llega a las formulas para los estimadores del intercepto y de la pendiente que se vieron en la sección 1.2 del capítulo 1.

2.2.2 Propiedades del estimador $\hat{\boldsymbol{\beta}}$

En forma similar al caso simple, el estimador minimo cuadrático tiene las siguientes propiedades:

a) $\hat{\boldsymbol{\beta}}$ es insesgado, o sea $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. En efecto,

$$E(\hat{\boldsymbol{\beta}}) = E[\mathbf{(X'X)^{-1}X'Y}] = E[\mathbf{(X'X)^{-1}X'(X\beta+e)}] = E[\mathbf{(X'X)^{-1}(X'X)\,\beta+(X'X)^{-1}X'e}]$$
$$= \boldsymbol{\beta} + \mathbf{(X'X)^{-1}X'}E(\mathbf{e})$$

como $E(\mathbf{e})=0$, se llega a $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

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

Para probar esto debemos usar una propiedad de la matriz de varianza-covarianza de **Az** donde **A** es matriz y **z** vector columna. La propiedad dice que Var(**Az**)==**A**Var(**z**)**A' .**

Luego, $Var(\hat{\boldsymbol{\beta}}) = Var[(\mathbf{X'X})^{-1}\mathbf{X'Y}] = (\mathbf{X'X})^{-1}\mathbf{X'}Var(\mathbf{Y})((\mathbf{X'X})^{-1}\mathbf{X'})'$

Usando los hechos que $(\mathbf{X'})'=\mathbf{X}$ y que $[(\mathbf{X'X})^{-1}]'=(\mathbf{X'X})^{-1}$, por simetría de la matriz inversa de **X'X,** se obtiene que

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X'X})^{-1}\mathbf{X'}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X'X})^{-1} = \sigma^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1} = \sigma^2(\mathbf{X'X})^{-1}$$

c) Si no se asume normalidad el estimador minimo-cuadrático $\hat{\boldsymbol{\beta}}$ es el mejor estimador dentro de los estimadores lineales insesgados de $\boldsymbol{\beta}$**,** en el sentido que es el que tiene la varianza más pequeña. Este es llamado el **teorema de Gauss-Markov.**

d) Si se asume normalidad de los errores entonces $\hat{\boldsymbol{\beta}}$ es el mejor estimador entre todos los estimadores insesgados de $\boldsymbol{\beta}$

2.2.3 Estimación de la varianza σ^2

En un modelo de regresión lineal múltiple con p variables predictoras (con el intercepto habrían en total p+1 parámetros a estimar), se tiene que un estimado de la varianza de los errores es

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1} = \frac{\sum_{i=1}^{n}\hat{e}_i^2}{n-p-1} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-p-1} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-p-1} \qquad (2.8)$$

El numerador de la expresión representa la suma de cuadrados de los residuales y puede ser escrito como:

$$SSE = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{Y}-\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y})'(\mathbf{Y}-\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y}) = \mathbf{Y'}(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

donde $\mathbf{H}=\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ es conocida en regresión como la ***"Hat Matrix"*** (la matriz sombrero).
Notar que $\mathbf{H'}=\mathbf{H}$ y que $\mathbf{H}^2=\mathbf{HH}=\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\,\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}=\mathbf{H}$**.** En álgebra matricial se dice que H es idempotente. **H** tiene muy buenas propiedades una de ellas es que Traza(**H**)=rango(**H**)=p+1. Por otro lado, $(\mathbf{I-H})'(\mathbf{I-H})=(\mathbf{I-H})(\mathbf{I-H})=\mathbf{I-H-H}+\mathbf{H}^2=\mathbf{I-2H}+\mathbf{H}=\mathbf{I-H}$**.** Así que también **I-H** es también simétrica e idempotente.

En consecuencia, la varianza estimada de los errores puede ser escrita como:

$$\hat{\sigma}^2 = \frac{\mathbf{Y'(I - X(X'X)^{-1}X')Y}}{n-p-1} = \frac{\mathbf{Y'(I-H)Y}}{n-p-1} \qquad (2.9)$$

Más común es usar el símbolo s para la desviación estándar estimada de los errores y $s = \sqrt{\frac{SSE}{n-p-1}} = \sqrt{MSE}$

Propiedad: Sea **Y** un vector aleatorio n-dimensional tal que E(**Y**)= **μ** y VAR(**Y**)=**V** entonces

E(**Y'AY**)=Traza(**AV**)+**μ'Aμ** **(2.10)**

Usando la propiedad anterior con **μ=Xβ** y $\mathbf{V}=\sigma^2\mathbf{I_n}$ se puede mostrar que $E[s^2]=\sigma^2$.

Ejemplo 2: Hallar el modelo de regresión lineal multiple para explicar el rendimiento en millaje promedio por galón (MPG) de los vehículos de acuerdo a cuatro variables predictoras: VOL, HP, SP y WT e interpretar los valores estimados.

Las variables predictoras que fueron usadas antes en el ejemplo 1 están definidas como sigue:

VOL: Capacidad de la cabina en pies cúbicos
HP: Potencia del motor
SP: Velocidad máxima (mph)
WT: Peso del vehículo (100 lb)

El modelo de regresión que ya fue obtenido en el ejemplo 1 es el siguiente:

```
MPG = 192 - 0.0156 VOL + 0.392 HP - 1.29 SP - 1.86 WT
```

Interpretación de los coeficientes de regresión estimados:

$\hat{\beta}_1 =$ -0.0156 significa que el millaje promedio por galón baja en promedio en 0.0156 cuando el volumen interior del carro aumenta en un pie cubico, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_2 =$ 0.392 significa que el millaje promedio por galón aumenta en promedio en 0.392 cuando la potencia del motor aumenta en 1 HP, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_3 =$ -1.29 significa que el millaje promedio por galón baja en promedio en 1.29 cuando la velocidad máxima del carro aumenta en 1 milla por hora, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_4 =$ -1.86 significa que el millaje promedio por galón baja en 1.86 cuando el peso del vehículo aumenta en 100 libras, asumiendo que las otras variables permanecen fijas.

En general, un coeficiente de regresión representa el cambio promedio en la variable de respuesta cuando la variable predictora correspondiente se incrementa en una unidad adicional y asumiendo que las otras variables predictoras premanecen fijas.

2.3. Inferencia en Regresión lineal múltiple

En esta sección se harán pruebas de hipótesis e intervalos de confianza acerca de los coeficientes del modelo de regresión poblacional. También se calcularán intervalos de confianza de las predicciones que se hacen con el modelo.

De ahora en adelante vamos a suponer que **e~NI(0,σ^2I$_n$)** o equivalente que **Y~NI(Xβ, σ^2I$_n$).**

Al igual que en regresión lineal simple la variación total de Y se descompone en dos variaciones: una debido a la regresión y otra debido a causas no controlables. Es decir,

SST=SSR+ SSE

Por teoría de modelos lineales se puede determinar que las sumas de cuadrados que aparecen en el análisis de regresión son formas cuadráticas de la variable de respuesta Y. Por lo tanto, éstas se distribuyen como una Ji-cuadrado. Más especificamente, se pueden establecer los siguientes resultados:

i). $\frac{SST}{\sigma^2} \sim \chi^2_{(n-1)}$ Ji-cuadrado no central con n-1 grados de libertad. Los grados de libertad se pueden establecer de la fórmula de cálculo de SST, pues en ella se usan *n* datos, pero en ella aparece un valor estimado ($\bar{y}$) por lo tanto se pierde un grado de libertad.

ii). $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-p-1)}$ Ji-cuadrado con n-p-1 grados de libertad . Para calcular *SSE* se usan *n* datos pero hay presente un estimado $\hat{y}_i$ cuyo cálculo depende a su vez de p+1 estimaciones. Por lo tanto se pierden p+1 grados de libertad. También se puede escribir que $\frac{(n-p-1)s^2}{\sigma^2} \sim \chi^2_{(n-p-1)}$

iii). $\frac{SSR}{\sigma^2} \sim \chi^2_{(p)}$ Ji-cuadrado no central con p grados de libertad

2.3.1 Prueba de hipótesis acerca de un coeficiente de regresión individual

En este caso la hipótesis nula más importante es H_o: β_i=0 (i=1,2..,p), o sea la variable X_i no es importante en el modelo, versus la hipótesis alterna H_a:$\beta_i \neq 0$; la variable X_i si merece ser considerada en el modelo. La prueba estadística es la prueba de t, definida por

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{s\sqrt{C_{ii}}}$$

donde el error estándar de $\hat{\beta}_i$ se calcula por $se(\hat{\beta}_i) = s\sqrt{C_{ii}}$, C_{ii} es el i-ésimo elemento de la diagonal de **(X'X)$^{-1}$.** Esta t se distribuye como una t de Student con n-p-1 grados de libertad. R al igual que otros programas de computadoras, da el "P-value" de la prueba t. Para el ejemplo anterior se obtiene lo siguiente

> summary(l6)

```
Call:
lm(formula = mpg ~ vol + hp + sp + wt)

Residuals:
   Min     1Q Median     3Q    Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
vol          -0.01565    0.02283  -0.685    0.495
hp            0.39221    0.08141   4.818 7.13e-06 ***
sp           -1.29482    0.24477  -5.290 1.11e-06 ***
wt           -1.85980    0.21336  -8.717 4.22e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los "P-values" de la prueba de t sugieren que la variable ***VOL*** no contribuye al modelo, pues se acepta la hipótesis nula de que dicho coeficiente es cero. Las otras tres variables *WT, HP y SP* si parecen ser importantes en el modelo ya que los "P-values" de la prueba t correpondientes son menores que .05.

2.3.2 Prueba de Hipótesis de que todos los coeficientes de regresión son ceros.

En este caso la hipótesis nula es Ho: $\beta_1=\beta_2=\ldots=\beta_p=0$, o sea que el modelo no sirve, versus la hipótesis alterna H_a: Al menos uno de los coeficientes es distinto de cero, o sea al menos una de las variables del modelo puede ser usada para explicar la variación de Y.
En la sección 2.2.3 se mencionó que $E(s^2)=\sigma^2$. La suma de cuadrados del error tiene n-p-1 g.l. Nuevamente usando esperado de una formas cuadrática se puede mostrar que

$$E(SSR)=E[\mathbf{Y'(H-11'/n)Y}]=p\sigma^2+ \boldsymbol{\beta}\mathbf{'X'(H-11'/n)X}\boldsymbol{\beta} \qquad \mathbf{(2.11)}$$

Donde **1** es un vector columna de n unos Si la hipótesis nula se cumpliera entonces $E(MSR)= \sigma^2$
La suma de cuadrados de Regresión tiene p grados de libertad que es igual al número de variables predictoras en el modelo.
Se puede mostrar que si la hipótesis nula es cierta entonces :

$$F = \frac{\dfrac{SSR}{p}}{\dfrac{SSE}{n-p-1}} = \frac{MSR}{MSE}$$

se distribuye como una F con p grados de libertad en el numerador y n-p-1 g.l en el denominador.

La prueba de F se obtiene al hacer la tabla del análisis de varianza para la regresión múltiple, la cual se muestra a continuación:

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados Medios	F
Regresión	SSR	P	MSR=SSR/p	F=MSR/MSE
Error	SSE	n-p-1	MSE=SSE/n-p-1	
Total	SST	n-1		

Para el ejemplo 1, usando todas las variables predictoras, se tiene,

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-Squared: 0.8733, Adjusted R-squared: 0.8667
F-statistic: 132.7 on 4 and 77 DF, p-value: < 2.2e-16

Notar que la desviación estimada del error es s=3.653= $\sqrt{MSE} = \sqrt{13.3}$. El "P-value" de la Prueba de F es 0.0000, lo cual lleva a la conclusión de que al menos una de las variables predictoras presentes en el modelo es importante para predecir MPG.

El coeficiente de Determinación R^2 tiene la misma interpretación que en regresión lineal simple y se calcula por $R^2 = \frac{SSR}{SST}$.

El R^2=87.3%, lo que quiere decir que hay un ajuste bastante bueno asumiendo que no hay datos contaminados en el conjunto de datos. El 87.3% de la variación del millaje promedio por galón es explicada por su relación lineal con VOL, HP, SP y WT. El R-Sq(adj) llamado el R^2 ajustado será definido más adelante en el capitulo de selección de variables.

La suma de cuadrados de regresión puede ser particionada en tantas partes como variables predictoras existen en el modelo. Esto es llamado un particionamiento secuencial de la suma de cuadrados de regresión y sirve para determinar la contribución de cada una de las variables predictoras al comportamiento de Y. Formalmente,

$SSR(\beta_1,\beta_2,,\ldots\beta_p/\beta_0)=SSR(\beta_1/\beta_0)+SSR((\beta_2,/\beta_1,\beta_0)+\ldots\ldots\ldots\ldots..SSR(\beta_p/\beta_{p-1},,\ldots\beta_1,\beta_0)$

Aquí $SSR(\beta_k/\beta_{k-1},,\ldots\beta_1,\beta_0)$ significa el incremento en la suma de cudrados de regresión cuando la variable X_k es incluida en el modelo, el cual ya contiene las variables predictivas $X_1,\ldots X_{k-1}$
La función **anova** de **R** produce estas sumas parciales. Para el ejemplo anterior se obtiene lo siguiente:

```
> l6

Call:
lm(formula = mpg ~ vol + hp + sp + wt)

Coefficients:
(Intercept)         vol          hp          sp          wt
  192.43775    -0.01565     0.39221    -1.29482    -1.85980

> anova(l6)
Analysis of Variance Table
```

```
Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
vol        1 1101.6  1101.6  82.563 8.172e-14 ***
hp         1 4731.1  4731.1 354.584 < 2.2e-16 ***
sp         1  233.6   233.6  17.509 7.515e-05 ***
wt         1 1013.8  1013.8  75.979 4.221e-13 ***
Residuals 77 1027.4    13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La Suma de Cuadrados de Regresión es 7080.1=1101.6+4731.1+233.6+1013.8. 233.6 significa que la suma de cuadrado de regresión aumenta en 233.6 cuando la variable SP es añadida al modelo, después que las variables VOL y HP ya están incluidas. El problema ahora es tratar de establecer pruebas para determinar si una variable predictora o un subconjunto de ellas efectivamente debe estar o no en el modelo.

Las sumas de cudrados de regresión secuenciales varia si se cambia el orden de las anteriores predictoras al momento de ajustar el modelo. Asi, si elegimos el orden WT, HP, SP y al final VOL se obtiene el siguiente resultado.

```
> l6<-lm(mpg~wt+hp+sp+vol)
> anova(l6)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq  F value    Pr(>F)
wt         1 6641.5  6641.5 497.7630 < 2.2e-16 ***
hp         1   35.4    35.4   2.6516    0.1075
sp         1  397.0   397.0  29.7522 5.739e-07 ***
vol        1    6.3     6.3   0.4698    0.4951
Residuals 77 1027.4    13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notar que las variables HP y VOL no son significativas. Es claro que la variable VOL es la menos importante de las variables predictoras.

2.3.3 Prueba de hipótesis para un subconjunto de coeficientes de regresión

Algunas veces estamos interesados en probar si algunos coeficientes del modelo de regresión son iguales a 0 simultáneamente. Por ejemplo, si el modelo tiene *p* variables predictoras, quisieramos probar si los k primeros coeficientes son ceros, o sea H_o: $\beta_1=\ldots..=\beta_k=0$. De rechazarse la hipótesis nula implicaría que las k primeras variables predictoras pueden ser excluidas del modelo.
Al modelo en donde se consideran todas las p variables se le llama el ***modelo completo*** y al modelo que queda asumiendo que la hipótesis nula es cierta se le llama el ***modelo reducido*.**
Es decir, que el modelo reducido sería

$$Y=\beta_{k+1}X_{k+1}+\beta_{k+2}X_{k+2}+\ldots..+\beta_pX_p+e \qquad (2.12)$$

Para probar si la hipótesis nula es cierta se usa una prueba de F que es llamada F-parcial. La prueba de F parcial se calcula por

$$Fp = \frac{\frac{SSR(C) - SSR(R)}{k}}{\frac{SSE(C)}{n-p-1}} = \frac{\frac{SSR(C) - SSR(R)}{k}}{MSE(C)}$$

donde SSR(C) y MSE(C) representan la suma de cuadrados de regresión y el cuadrado medio del error del modelo completo respectivamente, y SSR(R) es la suma de cuadrados de regresión del modelo reducido.

SSR(C)=SSR($\beta_1,\beta_2,\ldots\beta_p/\beta_o$) y
SSR(R) =SSR($\beta_{k+1},\beta_{k+2},\ldots,\beta_p/\beta_o$)

SSR(C) –SSR(R)=SSR($\beta_1,\beta_2,\ldots\beta_k/\beta_{k+1},\beta_{k+2},\ldots\beta_p$)

Esta última diferencia representa el incremento en la suma de cuadrados de regresión cuando $X_1,\ldots X_k$ son añadidas al modelo en el cual ya están presentes $X_{k+1},\ldots X_p$ y la constante

Si F_p es mayor que $F_{1-\alpha}$ usando k grados de libertad para el numerador y n-p-1 para el denominador entonces se rechaza H_o, en caso contrario se acepta.

R no tiene una opción que haga directamente la prueba de F parcial. Hay que calcular los dos modelos de regresión y usar las sumas de cuadrados de regresión de ambos modelos para calcular la prueba de F parcial.

Ejemplo 3: En el ejemplo 1, probar que las variables VOL y HP no son importantes y pueden ser excluidas del modelo

Haciendo el análisis de regresión sin incluir VOL y HP se obtiene:

```
> l2

Call:
lm(formula = mpg ~ sp + wt, data = millaje)

Coefficients:
(Intercept)           sp           wt
   75.64938     -0.09816     -0.99738

> anova(l2)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
sp         1 3842.6  3842.6  219.49 < 2.2e-16 ***
wt         1 2881.8  2881.8  164.61 < 2.2e-16 ***
Residuals 79 1383.0    17.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Luego SSR(R)=6724.4 y SSR(C)-SSR(R)=7080.1-6724.4=355.7.

```
> #Suma de cuadrados de regresion del modelo completo
> a=sum(anova(l1)$Sum[-(p+1)])
> #Suma de cuadrados de regresion del modelo reducido
> b=sum(anova(l2)$Sum[-(k+1)])
> #Cuadrado Medio del error del modelo completo
> c=anova(l1)$Mean[p+1]
> #Calculo del F parcial
> fp<-((a-b)/2)/c
> fp
[1] 13.32720
```

Luego la F-parcial será

$$F_p = \frac{\frac{355.63}{2}}{13.34} = \frac{177.81}{13.34} = 13.33$$

Usando un nivel de significación del 5%, Hay que comparar Fp con F(.95,2,77).

```
> #Hallando el percentil de la F con alpha=.05
> qf(.95,k,n-p-1)
[1] 3.115366
```

Luego Fp> F(.95,2,77)=3.11 por lo tanto se rechaza la prueba y se concluye que VOL y HP no pueden ser eliminadas simúltanemanente, al menos una de ellas es importante.

2.3.4 Intervalos de Confianza y de Predicción en Regresión Lineal Múltiple.

Supongamos que se desea predecir el valor medio de la variable de respuesta Y para una combinación predeterminada de las variables predictoras $X_1,\ldots\ldots X_p$,.Consideremos el vector $\mathbf{x'}_o = (1, x_{1,0},\ldots.x_{p,0})$ donde $x_{1,0},\ldots.x_{p,0}$ son los valores observados de $X_1,\ldots..X_p$ respectivamente.

El valor predicho para el valor medio de la variable de respuesta Y será $\hat{y}_o = \mathbf{x'_o}\,\hat{\boldsymbol{\beta}}$. De donde, $Var(\hat{y}_o) = \mathbf{x'_o}\,\mathbf{Var}((\hat{\boldsymbol{\beta}})\mathbf{x_o} = \sigma^2 \mathbf{x'_o}\,(\mathbf{X'X})^{-1}\mathbf{x_o}$. En consecuencia asumiendo que los errores están normalmente distribuidos se tiene que un intervalo del 100(1-α)% para el valor medio de Y dado que $\mathbf{x}=\mathbf{x_o}$ es de la forma

$$\hat{y}_o \pm t_{(\alpha/2,n-p-1)} s\sqrt{\mathbf{x'_o}\,(\mathbf{X'X})^{-1}\mathbf{x_o}} \qquad (2.13)$$

También usando la misma derivación que se hizo en el caso de regresión lineal simple se llega a establecer que un intervalo de confianza (más conocido como intervalo de predicción) del 100(1-α)% para el valor individual de Y dado $\mathbf{x}=\mathbf{x_o}$ es de la forma

$$\hat{y}_o \pm t_{(\alpha/2,n-p-1)} s\sqrt{1+\mathbf{x'_o}\,(\mathbf{X'X})^{-1}\mathbf{x_o}} \qquad (2.14)$$

Ejemplo 4: Usando el conjunto de datos **millaje,** hallar un intervalo de confianza del 95% para el millaje promedio por galón de todos los vehículos con capacidad interior de 90 pies cúbicos, un

HP de 50 una velocidad máxima de 1200 millas por galón y un peso de 20,000 libras. Hallar un intervalo de predicción para el millaje de un carro con las mismas características anteriores.

Usando R se obtiene

```
>   # hallando el intervalo de confianza del 95% para el valor medio
>   sp<-100
>   wt<-20
>   vol<-90
>   hp<-50
>   nuevo<-as.data.frame(cbind(sp,wt,vol,hp))
>   nuevo
  sp wt vol hp
1 100 20  90 50
>   predict.lm(l1,nuevo,se.fit=T,interval=c("confidence"),level=.95)
$fit
       fit     lwr     upr
[1,] 43.9624 42.41585 45.50894

$se.fit
[1] 0.7766682

$df
[1] 77

$residual.scale
[1] 3.652755

>  #Hallando el ntervalo de prediccion del 99% para los mismos datos
> predict.lm(l1,nuevo,se.fit=T,interval=c("prediction"),level=.99)
$fit
       fit     lwr     upr
[1,] 43.9624 34.09908 53.82571

$se.fit
[1] 0.7766682

$df
[1] 77

$residual.scale
[1] 3.652755
```

Hay un 95% de confianza de que el millaje promedio de todos los carros con las características dadas caiga entre 42.41 y 45.50 millas por galón. Hay un 99% de confianza de que el rendimiento de millas por galón de un carro con las características mencionadas caiga entre 34.09 y 53.82

2.3.5 La prueba de Falta de Ajuste

Es una prueba que se usa para determinar si la forma del modelo que se está considerando es adecuada. Es decir, si la regresión debe o no incluir términos potencias o interacciones entre las

variables predictoras. En el caso de regresión simple la prueba requiere que haya por lo menos un valor de la variable predictora con varias observaciones de y. En regresión múltiple se debe suponer que hay m combinaciones distintas de las n observaciones de las p variables predictoras y que por cada una de esas combinaciones hay n_i (i=1,…m) observaciones de la variable de respuesta, es decir, $\sum_{i=1}^{m} n_i = n$.

La Suma de Cuadrados del Error se particiona de la siguiente manera

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i-\hat{y}_i)^2 \qquad (2.15)$$

donde $\hat{y}_i$ es el valor predicho por el modelo de regresión para la i-ésima combinación de las variables predictoras , mientras que $\bar{y}_i$ es el valor promedio de la variable predictora para la i-ésima combinación.

La primera suma de cuadrados del lado derecho es llamado la **Suma de Cuadrados del Error Puro (SSPE)** y tiene n-m grados de libertad. Si no hubiera varios valores de la variable de respuesta por cada combinación de las predictoras esta suma sería cero. Se puede demostrar que el valor esperado del cuadrado medio del error puro es igual a la varianza poblacional σ^2 , sea o no sea el modelo de regresión adecuado.

La segunda suma de cuadrados que también puede ser escrita como $\sum_{i=1}^{m} n_i(\bar{y}_i-\hat{y}_i)^2$ es llamada la **Suma de Cuadrados de Falta de Ajuste (SSLOF)** y tiene m-p-1 grados de libertad. Si el modelo especificado es correcto entonces el valor esperado del cuadrado medio de Falta de Ajuste es igual a σ^2. Si le faltan términos al modelo (por ejemplo: potencias, productos de variables, etc.) entonces el cuadrado medio del la falta de ajuste sobreestimará a σ^2.

En resumen, la hipótesis nula será Ho: El modelo es adecuado (no hay falta de ajuste) versus Ha: el modelo no es adecuado y la prueba estadística es una prueba de F dada por

$$F = \frac{SSLOF/(m-p-1)}{SSPE/(n-m)} = \frac{MSLOF}{MSPE}$$

que se distribuye como una F(m-p-1,n-m) si la hipótesis nula es cierta. La hipótesis nula es rechazada si el valor de la prueba estadstica es mayor que F(1-α,m-p-1, n-m).

R no tiene una función para calcular directamente la prueba de bondad de ajuste. Hay que introducir una variable adicional que identifique los valores de y correspondiente al mismo valor de la variable predictora.

Ejemplo 5: Usando el conjunto de datos **millaje,** hacer una prueba de Falta de Ajuste si se considera la variable de respuesta MPG y la variable predictora HP.

Usando el laboratorio 8 del apéndice del texto se obtiene los siguientes resultados

```
> millajelf=millaje[,c(1,5)]
> table(millajelf$hp)

 49  52  53  55  62  66  70  73  74  78  80  81  84  90  92  93  95  96  98 100
```

```
  1    1   1  3   2    1   2  3    1  2    1   2   2   4   7   1   5   1   1   4
102 103 115 120 130 140 145 150 160 162 165 175 180 236 238 245 263 280 295 322
  5    1   5    1   3   4   1   1   1   2   4   1   1   1   1   1   1   1   1   1

># Hay m=40 valores distintos de la predictora
>.# anadiendo una columna  var3 que identifica a que grupo pertenece cada
># observación
> millajelf[1:10,]
   mpg hp var3
1  65.4 49    1
2  56.0 55    4
3  55.9 55    4
4  49.0 70    7
5  46.5 53    3
6  46.2 70    7
7  45.4 55    4
8  59.2 62    5
9  53.3 62    5
10 43.4 80   11
………………
………
> #haciendo la regresion lineal simple
> l1=lm(mpg~hp,data=millajelf)
> l1

Call:
lm(formula = mpg ~ hp, data = millajelf)

Coefficients:
(Intercept)           hp
    50.0661      -0.1390

> anova(l1)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
hp         1 5058.0  5058.0  132.69 < 2.2e-16 ***
Residuals 80 3049.4    38.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Haciendo el analisis de varianza de claificacion simple de mpg
># entre los 40 grupos diferentes
>l2=lm(mpg~factor(var3),data=millajelf)
> anova(l2)
Analysis of Variance Table

Response: mpg
             Df Sum Sq Mean Sq F value    Pr(>F)
factor(var3) 39 7794.4   199.9  26.809 < 2.2e-16 ***
Residuals    42  313.1     7.5
```

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

> # Haciendo el anova  para comparar  los dos modelos .   Se extrae la suma de cuadrados del
>#  error Puro y la suma de cuadrados de falta de Ajuste.
>#
>anova(l1,l2)
Analysis of Variance Table

Model 1: mpg ~ hp
Model 2: mpg ~ factor(var3)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     80 3049.44
2     42  313.11 38   2736.33   9.6592  1.703e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La suma de cuadrados de Eror Puro es 313.11 y la suma de cuadrados de Falta de ajuste es 2736.33. Si usamos el "P-value" de la prueba F de Falta de Ajuste se concluye que se rechaza la hipótesis nula, es decir, hay suficiente evidencia para concluir que faltan términos en el modelo.

MINITAB y otros programas estadísticos dan el "P-value" de esta prueba. MINITAB además da una prueba de Falta de Ajuste que no requiere que hayan varios valores de la variable de respuesta para cada combinación. La prueba de Falta de ajuste que da MINITAB es más informativa aún, dice que hay una posible curvatura en HP (ver el plot de la Figura 2.2), que hay outliers en la dirección de la variable predictora y finalmente da una prueba de ajuste global.

Consideremos ahora una prueba de ajuste usando todas las variable predictoras

```
> millajep=millaje[,2:5]
> dim(unique(millajep))
[1] 70  4

>#  Hay m=70 valores distintos de la predictora
>.# anadiendo una columna  var4 que identifica a que grupo pertenece cada
>#  observación
> millajelf=cbind(millaje,var4)
> millajelf[1:10,]
   mpg  sp   wt vol hp var4
1  65.4  96 17.5  89 49    1
2  56.0  97 20.0  92 55    2
3  55.9  97 20.0  92 55    2
4  49.0 105 20.0  92 70    3
5  46.5  96 20.0  92 53    4
6  46.2 105 20.0  89 70    5
7  45.4  97 20.0  92 55    2
8  59.2  98 22.5  50 62    6
9  53.3  98 22.5  50 62    6
10 43.4 107 22.5  94 80    7
>#  Haciendo  ;la regresion lineal multiple
> l2=lm(mpg~sp+wt+vol+hp,data=millajelf)
```

```
> anova(l2)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq  F value    Pr(>F)
sp         1 3842.6  3842.6 287.9944 < 2.2e-16 ***
wt         1 2881.8  2881.8 215.9879 < 2.2e-16 ***
vol        1   46.0    46.0   3.4451   0.06727 .
hp         1  309.7   309.7  23.2093 7.131e-06 ***
Residuals 77 1027.4    13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Haciendo el analisis de varianza de claificacion simple de mpg
># entre los 70 grupos diferentes

> l3=lm(mpg~factor(var4),data=millajelf)
> anova(l3)
Analysis of Variance Table

Response: mpg
             Df Sum Sq Mean Sq F value    Pr(>F)
factor(var4) 69 7990.3   115.8  11.859 2.130e-05 ***
Residuals    12  117.2     9.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # Haciendo el anova  para comparar  los dos modelos .   Se extrae la suma de cuadrados del
>#  error Puro y la suma de cuadrados de falta de Ajuste.
>#
> anova(l2,l3)
Analysis of Variance Table

Model 1: mpg ~ sp + wt + vol + hp
Model 2: mpg ~ factor(var4)
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     77 1027.38
2     12  117.18 65    910.20 1.434 0.2518
```

La clásica prueba de Falta de Ajuste acepta la hipótesis nula, es decir, que no hay suficiente evidencia para concluir que haya Falta de Ajuste. Sin embargo, la prueba de Falta de ajuste de MINITAB es mas infomativa y concluye que hay interacción entre las variables predictoras HP y SP, que hay que transformar WT y además hay outliers.

EJERCICIOS

1. Propiedades de la matriz HAT $\mathbf{H=X(X'X)^{-1}X'}$

a) La traza de una matriz es igual a la suma de los elementos que están en su diagonal. Probar que Traza(H)=p' con p'=p+1, donde p es el número de variables predictoras.

b) Probar que $\frac{1}{n} \leq h_{ii} \leq 1$, donde h_{ii} es el i-ésimo elemento de la diagonal de H.

c) Probar que $\mathbf{H1_n=1_n}$ aqui $\mathbf{1_n}$ es un vector columna cuyo elementos son todos unos

2.Usar el conjunto de datos **Fuel** con variable de respuesta Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas. Los datos están disponible en la página de internet del texto

a) Hallar la variable que tiene correlación más slta con la variable de respuesta
b) Hacer un plot matricial para ver si no hay outliers y determinar si el coeficiente de correlación es confiable.
c) Hacer una regresión lineal de Y versus la variable determinada en los pasos a y b y tratar otros modelos: cuadrático, exponencial, logaritmico para mejorar el R^2, si es posible

d) Hallar un Intervalo de Confianza del 99% para el valor medio y el valor Predicho de Y, escogiendo un valor adecuado de la variables predictora usando el modelo lineal . Trazar las bandas de confiianza. Comentar sus resultados.

e) Hallar el modelo de regresion múltiple considerando todas las variable predictoras e interpretar los coeficientes de regresión.

f) Interpretar el coeficiente de Determinación R^2.

g) Probar que todos los coeficientes del modelo de regresión son ceros. Comentar el resultado.

h) Probar que cada uno de los coeficientes del modelo de regresión es cero. Comentar el resultado.

3. Sea $\hat{\beta} = (X'X)^{-1}X'Y$. Probar que

$$(Y - X\beta)'(Y - X\beta) = (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) + Y'(I - H)Y$$

y en consecuencia $\hat{\beta}$ es el estimador mínimo cuadrático.

4. Uso de la factorizacion QR en Regresión Supongamos que tenemos una matriz ortogonal **Q** de orden nx p' (es decir, **Q'Q=I**) y una matriz triangular superior **R** tal que **QR=X**)

a) Probar que **R'R=X'X**

b) Escribir el estimador minimocuadrático $\widehat{\boldsymbol{\beta}}$ en terminos de Q, R y Y. Cual sería la ventaja de usar esta fórmula con respecto a la fórmula original.?

c) Expresar $\widehat{\mathbf{Y}}$ y $\widehat{\mathbf{e}}$ en terminos de Y y Q.

5. Usar el conjunto de datos **Highway**, con variable de respuesta es TASA y todas las otras como variables predictoras para responder las siguientes preguntas. Los datos están en la página de internet del texto.

a)Hallar la variable que tiene correlación más slta con la variable de respuesta

b)Hacer un plot matricial para ver si no hay outliers y determinar si el coeficiente de correlación es confiable

c) Hacer una regresión lineal de Y versus la variable determinada en los pasos a y b y tratar otros modelos: cuadrático, exponencial, logaritrmico para mejorar el R^2, si es posible

d) Hallar el modelo de regresion múltiple considerando todas las variable predictoras e interpretar los coeficientes de regresión.

e) Interpretar el coeficiente de Determinación R^2.

f) Probar que todos los coeficientes del modelo de regresión son ceros. Comentar el resultado.

g) Probar que cada uno de los coeficientes del modelo de regresión es cero. Comentar el resultado.

h) Hallar las dos variables que están menos correlacionadas con la variable de respuesta y probar la hipótesis de que ambas variables deben ser excluidas simúltaneamente del modelo.

i) Hallar un Intervalo de Confianza para el valor medio de Y y el valor Predicho del 99% para Y, escogiendo valores adecuados de las variables predictoras. Comentar sus resultados.

j) Hacer un análisis de falta de ajuste usando como variable predictora, aquella obtenida en a).

6. Considerando un modelo de regresión lineal multiple probar que $E[s^2]=\sigma^2$ donde s^2 es la varianza estimada del error definida en la ecuación 2.9.

7. Verificar la identidad de la ecuación 2.11

8. **Efecto de subajuste**. Supongamos que se ajusta el modelo $\mathbf{Y=X\beta+e}$ donde $\mathbf{X}$ es una matriz n x r cuando en realidad el modelo verdadero incluye s adicionales variables predictoras contenidas en la matriz $\mathbf{Z}$. Es decir, que el verdadero modelo es $\mathbf{Y=X\beta+Z\gamma+ e}$. Mostrar que en general el estimador mínimos cuadrático $\hat{\boldsymbol{\beta}}$ usando el modelo reducido es sesgado. Asímismo mostrar que el estimador de la varianza es sesgado. Bajo que condiciones ambos estimadores serían insesgado?

9. Supongamos que se ha obtenido la siguiente regresión usando una muestra de 75 observaciones

$$Y= -5.16 + .325X_1 + 5.55X_2 + .3X_3 + .01X_4 + 8.75X_5 - .97X_6$$

a) Interpretar cualquiera de los coeficientes de las variables predictoras

b) Hallar el valor de la prueba estadística de F si el coeficiente de determinación $R^2=.95$

c) Explicar detalladamente como se probaría la hipótesis $Ho:\beta_1=\beta_3=\beta_4=\beta_6$

10. **Efecto de sobreajuste.** Supongamos que realmente el modelo $Y=X\beta+e$ (1) ajusta a nuestro conjunto de datos. Reescribamos el modelo anterior por

$$Y=X_1\beta_1+X_2\beta_2+e \quad , \text{donde } X=(X_1 \mid X_2)$$

X_1 es de orden n x k y X_2 es de orden n x (p-k), n es el número de observaciones y p es el número de parametros del modelo, es decir el número de variables predictoras mas el intercepto. Consideremos que en lugar del modelo (1) se usa el siguiente modelo para ajustar los datos

$$Y=X_1\beta_1+e \quad (2)$$

a) Hallar el esperado del estimador mínimo cuadrático de β_1 usando el modelo (2), pero considerando que realmente (1) es el que se cumple.

b) Hallar el esperado del estimador minimocuadrático de la varianza estimada usando el modelo (2).