Segundo Corte: Regresión lineal Múltiple

Andrés Martínez

2019

Índice

1. Regresión Lineal Múltiple

En un modelo lineal múltiple se asume qye Y es una función lineal de sus predictores mas un ruido blanco:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon \tag{1}$$

1.1. Supuestos

1. Todos los parámetros son lienales

2. La muestra de n observaciones sigue el modelo poblacional del primer supuesto
3. No existe perfecta colinearidad
4. EL error tiene un valor esperado de cero dado por variables independientes.
5. Homocedasticidad: El error tiene la misma varianza que los valores de las variables explicativas.

No hacemos suposiciones sobre las distribuciones (marginales o conjuntas) de X_i, pero suponemos que $E[\epsilon|X] = 0$, $V[\epsilon|X] = \sigma^2$, y ese ϵ no está correlacionado entre las mediciones. La forma matricial del modelo es:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{2}$$

Donde $\mathbf{X}$ incluye una columna inicial de 1.

Cuando se incluye se asume que este debe tener un comportamiento normal como se ve en la siguiente ecuación:

$$\epsilon \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}) \tag{3}$$

independiente de $\mathbf{X}$.

Los coeficientes se estiman de la siguiente forma:

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{4}$$

Bajo el supuesto de ruido blanco, este es el mismo resultado que la estimación por máxima verosimilitud

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{H}\mathbf{Y} \tag{5}$$

Donde $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

Que es simetrico e idempotente. Los residuales se obteienen así:

$$SSE = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y} \tag{6}$$

donde $\mathbf{I} - \mathbf{H}$ es simetrico e idempotente

1.2. Derivación de los Estimadores

$$\widehat{\beta}_{OLS} = argmin_{\beta\in\mathbb{R}^2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \tag{7}$$

$$\frac{\partial}{\partial\beta} = (\mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T(\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta)) \tag{8}$$

$$\frac{\partial}{\partial\beta} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\beta} = 0 \tag{9}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{10}$$

1.3. Propiedades de los Estimadores

Insesgamiento del estimador mínimo cuadrado

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{11}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}^T(\mathbf{X}\beta + \epsilon) \tag{12}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\beta \tag{13}$$

$$+(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\epsilon) \tag{14}$$

$$= \beta + (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\epsilon) \tag{15}$$

$$E[\hat{\beta}|x] = \beta \tag{16}$$

Varianza de los errores estandar: Suponiendo Homocedasticidad

$$V[\hat{\beta}|\mathbf{X}] = V[\beta + (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\epsilon)/\mathbf{X}] \tag{17}$$

$$V[\hat{\beta}|\mathbf{X}] = V[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\epsilon)/\mathbf{X}] \tag{18}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}V[\mathbf{X}^T\epsilon/\mathbf{X}](\mathbf{X}^T\mathbf{X})^{-1} \tag{19}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TV[\epsilon/\mathbf{X}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \tag{20}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \tag{21}$$

$$\hat{\beta} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \tag{22}$$

$$\hat{\beta} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{23}$$

Consistencia

- Ley de los grandes números
- Teorema Límite Central

Tenemos que $\hat{\beta} = \beta + (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\epsilon)$ y se desea ver si $\hat{\beta}$ es consistente. $\lim_{n\to p}(\hat{\beta} - \beta) = 0$

Eficiencia: Bajo los supuestos de linealidad, muestreo aleatorio, no colinealidad perfecta, media condicional cero de los errores y homocedasticidad, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ es el mejor estimador lineal insesgado.

Además $\hat{\beta}_1 = \beta_1 + (\mathbf{X}_1^T\mathbf{X}_1)^{-1}(\mathbf{X}_1^T\epsilon_2)$.

1.4. Colinealidad

Asumiendo que $(\mathbf{X}^T\mathbf{X})^{-1}$ existe y es invertible o no singular, existen un número equivalente de condiciones para una matriz que debe ser invertible:

- Su determinante no es cero
- Su rango de columnas es completo, lo que quiere decir que todas las columnas son lienalmente independientes.
- Su rango de filas es completo, lo que quiere decir que las filas son linealmente independientes.

Colinearidad perfecta:
Usando uno de los ejemplos del libro, que tiene la base de datos vote1, se hace una regresión entre los gastos de la campaña de A y B más los totales, con respecto a el voto de A.

```
> library(wooldridge)
> data("vote1")
> total=vote1$expendA +vote1$expendB
> model1=lm(vote1$voteA~ vote1$expendA+vote1$expendB+total)
> summary(model1)

Call:
lm(formula = vote1$voteA ~ vote1$expendA + vote1$expendB + total)

Residuals:
     Min       1Q   Median       3Q      Max
-27.5661  -8.6919   0.1641   9.0789  27.0131

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.618995   1.426147   34.79   <2e-16 ***
vote1$expendA  0.038331   0.003387   11.32   <2e-16 ***
vote1$expendB -0.036127   0.003107  -11.63   <2e-16 ***
```

```
total                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.58 on 170 degrees of freedom
Multiple R-squared:  0.5299,        Adjusted R-squared:  0.5244
F-statistic: 95.83 on 2 and 170 DF,  p-value: < 2.2e-16
```

El resultado evidencia que cuando existe colinealidad perfecta, no es posible obtener el coeficiente para ese valor, porque este está directamente relacionado con los gastos en A y B. Esto se abordará más adelante cuando se profundice más en multicolinealidad.

```
> votop=1.2*vote1$voteA+rnorm(1)
> model2=lm(vote1$voteA~ vote1$expendA+vote1$expendB+votop)
> summary(model2)

Call:
lm(formula = vote1$voteA ~ vote1$expendA + vote1$expendB + votop)

Residuals:
       Min         1Q     Median         3Q        Max
-9.242e-14 -2.264e-15 -1.960e-16  1.816e-15  7.407e-14

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)
(Intercept)   7.302e-01  3.800e-15 1.922e+14  < 2e-16 ***
vote1$expendA 5.349e-18  4.248e-18 1.259e+00 0.209708
vote1$expendB 1.526e-17  3.943e-18 3.869e+00 0.000156 ***
votop         8.333e-01  6.054e-17 1.376e+16  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.096e-14 on 169 degrees of freedom
Multiple R-squared:      1,        Adjusted R-squared:      1
F-statistic: 1.344e+32 on 3 and 169 DF,  p-value: < 2.2e-16
```

Si no existe correlación perfecta, pero hay un alto grado de colinealidad como en el modelo 2, puede haber un problema de especificación probocando que los errores de los coeficientes aumenten. Para evitar la colinealidad entre las variables, se debe tener claro desde antes la relación entre las variables, pero también se deben probar los modelos de tal forma que los errores en los coeficientes sean mínimos y estas variables sean significativas.

1.5. Goodness-of-fit

El objetivo más importante de estos modelos es obtener los errores, pues permiten establecer la eficiencia del modelo y las posibles falencias de especifi-

cación que este pueda tener. Para esto definimos rapidamente los tres errores que siempre se buscan en una regresión lineal múltiple o simple:

$$SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \mathbf{Y}^T(\mathbf{I} - \mathbf{I}_{n\times n})\mathbf{Y}. \tag{24}$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}. \tag{25}$$

$$SSR = \sum_{i=1}^{n}(\hat{y} - \overline{y})^2 = \mathbf{Y}^T(\mathbf{H} - \mathbf{I}_{n\times n})\mathbf{Y}. \tag{26}$$

Con estos resultados se pouede obtener el coeficiente de determinación múltiple R^2.

$$R^2 = \frac{SSR}{SCT} = 1 - \frac{SSE}{SST}. \tag{27}$$

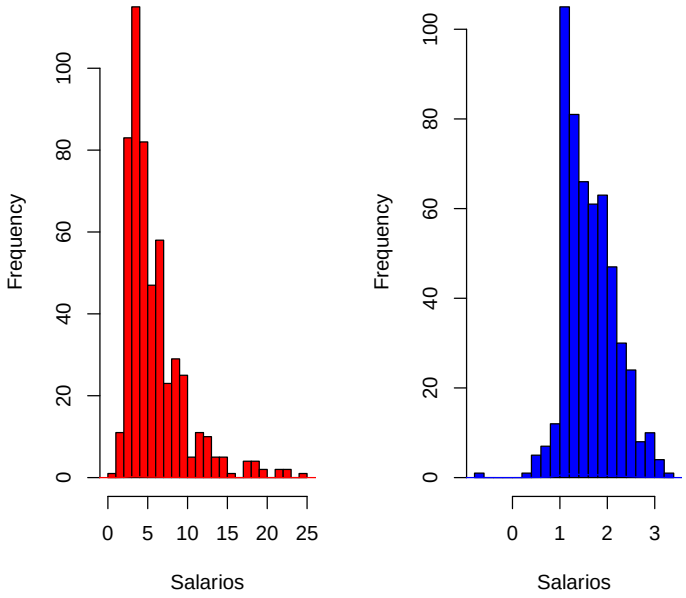
1.6. Propiedades de los estimadores de mínimos cuadrados

- $E[\hat{\beta}_i] = \beta_i$
- $V[\hat{\beta}_i] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- Un estimado insesgado de σ^2 es $S^2 = SSE/[n - (k + 1)]$, donde $SSE = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$
- Cada β_i está distribuido normalmente

1.7. Ejemplo

Usando el ejemplo 3.2 del libro de wooldrige, se hace la regresión del logaritmo de los salarios con respecto a la educación, la experiencia y el tiempo empleado. El proposito de obtener el logaritmo de los salarios es para generar una distribución simétrica y mejorar la escala.

```
> data("wage1")
> par(mfrow=c(1,2))
> hist(as.numeric(wage1$wage),breaks=20,main="",xlab="Salarios", col="red")
> lines(density(wage1$wage), col = 10)
> hist(as.numeric(wage1$lwage),breaks=20,main="",xlab="Salarios", col="blue")
> lines(density(wage1$lwage), col = 4)
>
>
```



La idea es trabajar con variables que tengan una distribución simétrica y al generar el logaritmo, los salarios adquieren una mejor forma como se ve en la gráfica anterior.

Para observar mejor el cambio, primero se hace la regresión con el logaritmo de los salarios, luego se hace con los salarios y se compara el resultado de cada uno de los coeficientes y el coeficiente de correlación

```
> modelo=lm(wage1$lwage~ wage1$educ+wage1$exper+wage1$tenure)
> summary(modelo)

Call:
lm(formula = wage1$lwage ~ wage1$educ + wage1$exper + wage1$tenure)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05802 -0.29645 -0.03265  0.28788  1.42809

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.284360   0.104190   2.729  0.00656 **
wage1$educ   0.092029   0.007330  12.555  < 2e-16 ***
wage1$exper  0.004121   0.001723   2.391  0.01714 *
wage1$tenure 0.022067   0.003094   7.133 3.29e-12 ***
```

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

Residual standard error: 0.4409 on 522 degrees of freedom
Multiple R-squared:  0.316,        Adjusted R-squared:  0.3121
F-statistic: 80.39 on 3 and 522 DF,  p-value: < 2.2e-16

>
>
```

El segundo modelo es

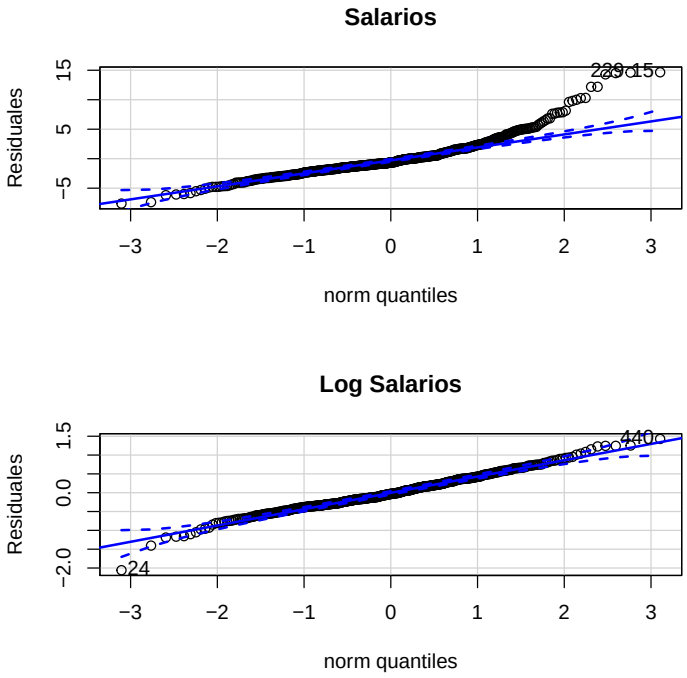
```
> modelo1=lm(wage1$wage~ wage1$educ+wage1$exper+wage1$tenure)
> summary(modelo1)

Call:
lm(formula = wage1$wage ~ wage1$educ + wage1$exper + wage1$tenure)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6068 -1.7747 -0.6279  1.1969 14.6536

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.87273    0.72896  -3.941 9.22e-05 ***
wage1$educ    0.59897    0.05128  11.679  < 2e-16 ***
wage1$exper   0.02234    0.01206   1.853   0.0645 .
wage1$tenure  0.16927    0.02164   7.820 2.93e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.084 on 522 degrees of freedom
Multiple R-squared:  0.3064,        Adjusted R-squared:  0.3024
F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16

>
```

Como se puede ver el error en los coeficientes de cada variable aumenta, si bien el coeficiente de correlación aun no es óptimo para generar la regresión, si se puede ver una mejora cuando se trabaja con el logaritmo de los salarios.

Para continuar con el análisis se hace un gráfico qq plot de los residuales con el objetivo de observar cuál de los dos cumple con el suspuesto de normalidad.

```
> library(car)
> par(mfrow=c(2,1))
> qqPlot(modelo1$residuals,main = "Salarios",ylab="Residuales")
```

```
[1]  15 229
> qqPlot(modelo$residuals,main = "Log Salarios",ylab="Residuales")
[1]  24 440
```



Al observar los residuales de ambos modelos, es mucho más claro que transformar la variable de salarios genera un mejor proceso de la regresión.

1.8. Inferencia del Modelo

Para generar la inferencia de cada uno de los estimadores, primero se hará el proceso de forma individual, y luego de forma global con la prueba F. En el test de significancia individual, se debe tener cuidado con las muestras pequeñas, dado que se puede rechazar la hipotesis nula H_0 cuando esta es cierta cayendo en error tipo uno. Para esto es mejor realizar una hipotesis a dos colas. De forma matricial el estadístico T se obtiene de la siguiente forma:

Usando la prueba t se obtiene que

$$T = \frac{\hat{\beta}_i - \beta_i}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}}} \tag{28}$$

Donde β_i corresponde al coeficiente que se está analizando. El error estandar es $S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}}$.

El intervalo para el coeficiente se obtiene con:

$$\hat{\beta}_i \pm T_{\alpha/2} S \sqrt{(\mathbf{X}^T\mathbf{X})^{-1}} \tag{29}$$

Recurde que para que el coeficiente sea significativo este no puede pasar por cero.

1.9. Revisando adentro de las matrices

Para entender mejor que sucede en el modelo, recrearemos las matrices para un modelo de regresión lineal simple.

Primero revisamos $(\mathbf{X}^T\mathbf{X})$

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

Como se puede ver en la matriz, en la posición $(1,1)$ se encuentra el tamaño de la muestra, en la posición $(2,2)$ se encuentra la suma de x al cuadrado y en las demás la suma de x. Solo cuando esta matriz tiene un determinante, se puede obtener la matriz inversa de $(\mathbf{X}^T\mathbf{X})$.

Con la matriz inversa se obtienen los calculos iniciales para encontrar las varianzas y covarianzas de cada uno de los coeficientes.

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{\sum x_i^2}{nS_{xx}} & -\frac{\overline{x}}{S_{xx}} \\ -\frac{\overline{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{pmatrix}$$

Como se vió en el primer corte, la diagonal de la matriz multiplicada por S^2 sirve para obtener la varianza de cada coeficiente, mientras que los demás valores deberían tender a cero ya que representan el primer cálculo de la covarianza, por lo tanto si existiera una relación entre las variables, se incumpliria el supuesto de independencia entre las variables generando colinearidad.

De acuerdo a lo anterior, $V(\hat{\beta}_0) = \frac{\sum x_i^2}{nS_{xx}} S^2$ y $V(\hat{\beta}_1) = \frac{1}{S_{xx}} S^2$ y con estos resultados se puede obtener el estadístico T para validar la significancia individual.

1.10. Significancia Global

Para encontrar la significancia global, se plantea la hipotesis $H_0 : \beta_{g+1} = \beta_{g+1} = ... = \beta_k = 0$. La idea es encontrar un conjunto de parámetros que expliquen de forma conjunta la variable dependiente.

Para valorar el modelo se crea un modelo reducido para contrastarlo con el modelo completo.

$$modeloR : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_g x_g + \epsilon \tag{30}$$

$$modeloC : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_g x_g + .. + \beta_k x_k + \epsilon \tag{31}$$

De esta forma se obtienen la suma de los errores SSE_r y SSE_c de cada modelo. Si el modelo completo aporta más información, SSE_c debería ser menor al SSE_r . Por lo tanto entre mayor sea la diferencia $(SSE_r - SSE_c)$ más fuerte es la evidencia de que el modelo completo tiene más información y por lo tanto soporta la hipótesis alternativa a la planteada al principio de la sección.

Dado que estamos trabajando con la varianza y estamos haciendo una comparación de varianzas es conveniente trabajar con la distribución Chi quadrado y con la distribución F.

Para cada suma de cuadrados se tiene las siguientes pruebas:

$$\chi_3^2 = \frac{SSE_r}{\sigma^2} \tag{32}$$

$$\chi_2^2 = \frac{SSE_c}{\sigma^2} \tag{33}$$

$$\chi_1^2 = \frac{SSE_r - SSE_c}{\sigma^2} \tag{34}$$

Recuerde que la función χ^2 trabaja con grados de libertad. Por lo tanto los grados de libertad para cada muestra son $(n - [g + 1])$, $(n - [k + 1])$ y $(k - g)$ respectivamente.

Comparando las dos distribuciones se crea la prueba F y se obtiene el estadístico para determinar se se debe on no rechazar la hipótesis nula,

$$F = \frac{\chi_1^2/(k-g)}{\chi_2^2/(n-[k+1])} \tag{35}$$

1.10.1. Ejemplo

Realice una regresión lineal usando las matrices con dos variables independientes. Donde $X_1 = -2, -1, 0, 1, 2$, $X_2 = 4, 1, 0, 1, 4$ y $Y = 0, 0, 1, 1, 3$.

```
> Y=as.vector(c(0,0,1,1,3))
> X=matrix(c(1,1,1,1,1,-2,-1,0,1,2,4,1,0,1,4),nrow=5,ncol=3)
> X
```

```
     [,1] [,2] [,3]
[1,]    1   -2    4
[2,]    1   -1    1
[3,]    1    0    0
[4,]    1    1    1
[5,]    1    2    4
```

```
> Beta=solve(t(X)%*%X)%*%t(X)%*%Y
> Beta
```

```
          [,1]
[1,] 0.5714286
[2,] 0.7000000
[3,] 0.2142857
```

```
> SSE=t(Y-X%*%Beta)%*%(Y-X%*%Beta)
> SSE
```

```
          [,1]
[1,] 0.4571429
```

```
> S2=SSE/(5-3)
> S2
```

```
          [,1]
[1,] 0.2285714
```

```
> VB=solve(t(X)%*%X)
> VB
```

```
           [,1] [,2]        [,3]
[1,]  0.4857143  0.0 -0.14285714
[2,]  0.0000000  0.1  0.00000000
[3,] -0.1428571  0.0  0.07142857
```

```
> VB0=S2*VB[1,1]
> VB0
```

```
          [,1]
[1,] 0.1110204
```

```
> VB1=S2*VB[2,2]
> VB1
```

```
           [,1]
[1,] 0.02285714
```

```
> VB2=S2*VB[3,3]
> VB2
```

```
           [,1]
[1,] 0.01632653

>

> T0=Beta[1]/sqrt(VB0)
> T0

         [,1]
[1,] 1.714986

> T1=Beta[2]/sqrt(VB1)
> T1

         [,1]
[1,] 4.630065

> T2=Beta[3]/sqrt(VB2)
> T2

         [,1]
[1,] 1.677051

> ## Buscar el test d edos colas por ser una muestra pequeña
> pb0=2*pt(-abs(T0),df=3, lower.tail = TRUE)
> pb0

          [,1]
[1,] 0.1848574

> pb1=2*pt(-abs(T1),df=3)
> pb1

           [,1]
[1,] 0.01897554

> pb2=2*pt(-abs(T2),df=3)
> pb2

          [,1]
[1,] 0.1921256

> # Modelo reducido
> XR=X[,1]
> Beta=solve(t(XR)%*%XR)%*%t(XR)%*%Y
> Beta

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

```
> SSER=t(Y-XR%*%Beta)%*%(Y-XR%*%Beta)
> SSER

     [,1]
[1,]    6

>

> F=((SSER-SSE)/(2-0))/(SSE/(5-3))
> F

       [,1]
[1,] 12.125

> FAOV <- aov(Y~X[,2]+X[,3])
> FAOV

Call:
   aov(formula = Y ~ X[, 2] + X[, 3])

Terms:
                  X[, 2]   X[, 3] Residuals
Sum of Squares  4.900000 0.642857  0.457143
Deg. of Freedom        1        1         2

Residual standard error: 0.4780914
Estimated effects may be unbalanced

>
```

Coeficente de Correlación

```
> R2=1-SSE/(t(Y)%*%Y)
> R2

          [,1]
[1,] 0.9584416

> modelm=lm(Y~X[,2]+X[,3])
> summary(modelm)

Call:
lm(formula = Y ~ X[, 2] + X[, 3])

Residuals:
       1        2        3        4        5
-0.02857 -0.08571  0.42857 -0.48571  0.17143
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5714     0.3332   1.715   0.2285
X[, 2]        0.7000     0.1512   4.630   0.0436 *
X[, 3]        0.2143     0.1278   1.677   0.2355
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4781 on 2 degrees of freedom
Multiple R-squared:  0.9238,        Adjusted R-squared:  0.8476
F-statistic: 12.13 on 2 and 2 DF,  p-value: 0.07619
```

```
> FAOV<- anova(modelm)
> FAOV

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X[, 2]     1 4.9000  4.9000 21.4375 0.04362 *
X[, 3]     1 0.6429  0.6429  2.8125 0.23553
Residuals  2 0.4571  0.2286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1.11. Predicción del modelo

La predicción de un modelo varia de acuerdo a la transformación de las variables que se usaron para generar el modelo estimado. No obstante para determinar el valor de la predicción se obtienen los intervalos de confianza que permite establecer el valor esperado y la variación con respecto a la media de la variable que se desea estimar.

1.12. Transformación de Variables

En esta sección vamos a trabajar con diferentes regresiones que tiene como factor comun, la transformación de una o más variables.

EL primer ejemplo se encuentra en el capítulo 6 del libro del curso wooldrige, en donde se evalua el peso de los recien nacidos en relación con los cigarrillos que una madre fuma durante el tiempo de gestación.

```
> library(wooldridge)
> data("bwght")
> modelo3=lm(bwght$bwght ~ bwght$cigs+bwght$faminc)
> summary(modelo3)
```

```
Call:
lm(formula = bwght$bwght ~ bwght$cigs + bwght$faminc)

Residuals:
    Min      1Q  Median      3Q     Max
-96.061 -11.543   0.638  13.126 150.083

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  116.97413    1.04898 111.512  < 2e-16 ***
bwght$cigs    -0.46341    0.09158  -5.060 4.75e-07 ***
bwght$faminc   0.09276    0.02919   3.178  0.00151 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.06 on 1385 degrees of freedom
Multiple R-squared:  0.0298,        Adjusted R-squared:  0.0284
F-statistic: 21.27 on 2 and 1385 DF,  p-value: 7.942e-10
```

```
> hist(as.numeric(bwght$bwght),breaks=20,main="",xlab="Salarios", col="red")
```

El siguiente ejemplo busca generar el modelo en libras en vez de onzas, y para esto se divide cada una de las variables en 16, notese que son todas las variables las que se dividen para mantener la medida.

```
> y=bwght$bwght/16
> x1=bwght$cigs/16
> x2=bwght$faminc/16
> modelo4=lm(y~ x1 + x2)
> summary(modelo4)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0038 -0.7215  0.0399  0.8204  9.3802

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.31088    0.06556 111.512  < 2e-16 ***
x1          -0.46341    0.09158  -5.060 4.75e-07 ***
x2           0.09276    0.02919   3.178  0.00151 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.254 on 1385 degrees of freedom
```

```
Multiple R-squared:  0.0298,        Adjusted R-squared:  0.0284
F-statistic: 21.27 on 2 and 1385 DF,  p-value: 7.942e-10
```

Si bien el intercepto cambia, se puede ver a primera vista que los coeficientes se mantienen iguales. Por lo tanto, si las condiciones se mantienen iguales para todas las variables, se tendrán los mismos coeficientes.

El siguiente ejemplo usa los datos de la base de precios de casas 2 del libro de wooldrige, en donde se buscar la relación entre el logarítmo del precio con respecto a la contaminación y el número de habitaciones.

```
> data(hprice2)
> modelo5=lm(hprice2$lprice~ hprice2$lnox + hprice2$rooms)
> summary(modelo5)

Call:
lm(formula = hprice2$lprice ~ hprice2$lnox + hprice2$rooms)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06485 -0.12331  0.00782  0.14471  1.38770

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.23374    0.18774   49.18   <2e-16 ***
hprice2$lnox  -0.71767    0.06634  -10.82   <2e-16 ***
hprice2$rooms  0.30592    0.01902   16.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.286 on 503 degrees of freedom
Multiple R-squared:  0.5137,        Adjusted R-squared:  0.5118
F-statistic: 265.7 on 2 and 503 DF,  p-value: < 2.2e-16
```

En este ejemplo se hace la misma regresión pero en este caso se usan directamente los precios. Es interesante ver como mejora el r cuadrado, pero también aumentan los errores de los coeficientes.

```
> modelo6=lm(hprice2$price~ hprice2$nox + hprice2$rooms)
> summary(modelo6)

Call:
lm(formula = hprice2$price ~ hprice2$nox + hprice2$rooms)

Residuals:
   Min     1Q Median     3Q    Max
-17952  -3303   -600   2672  39669
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -18423.4     3346.7  -5.505 5.89e-08 ***
hprice2$nox    -1884.7      253.6  -7.432 4.62e-13 ***
hprice2$rooms   8178.6      418.1  19.562  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6291 on 503 degrees of freedom
Multiple R-squared:  0.5352,        Adjusted R-squared:  0.5333
F-statistic: 289.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

Finalmente en este ejemplo se incluyen más variables ya que se busca evitar caer en el error de omitir variables significativas. En comparación con los demás modelos, se puede ver que las nuevas variables son significativas y también el r cuadrado mejora considerablemente.

```
> modelo7=lm(hprice2$lprice~ hprice2$lnox + hprice2$rooms+hprice2$dist+hprice2$crime+hprice2
> summary(modelo7)

Call:
lm(formula = hprice2$lprice ~ hprice2$lnox + hprice2$rooms +
    hprice2$dist + hprice2$crime + hprice2$stratio + hprice2$lproptax)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85915 -0.12314 -0.01216  0.10615  1.32918

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      10.985234   0.299231  36.712  < 2e-16 ***
hprice2$lnox     -0.701876   0.107804  -6.511 1.83e-10 ***
hprice2$rooms     0.242073   0.017006  14.235  < 2e-16 ***
hprice2$dist     -0.040705   0.008835  -4.607 5.18e-06 ***
hprice2$crime    -0.012730   0.001521  -8.369 5.86e-16 ***
hprice2$stratio  -0.040195   0.005787  -6.945 1.18e-11 ***
hprice2$lproptax -0.073268   0.042150  -1.738   0.0828 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2419 on 499 degrees of freedom
Multiple R-squared:  0.6549,        Adjusted R-squared:  0.6508
F-statistic: 157.8 on 6 and 499 DF,  p-value: < 2.2e-16

>
```

1.13. Estandarización de las variables

Cuando se trabaja con variables estandarizadas, se evalua cada una de las variables de acuerdo a su escala en términos de una distribución normal estandar.

$$Z_y = \frac{y - \overline{y}}{sd(y)} \tag{36}$$

$$Z_x = \frac{x - \overline{x}}{sd(x)} \tag{37}$$

La interpretación de estos resultados se mide a través de la variación de las desviaciones estandar. Es decir el cambio de la variable independiente cuando varia la desviación estnadar.

Usando la función scale se puede estandarizar directamente cada una de las variables.Por lo tanto, la nueva regresión se puede leer de la siguiente forma:

$$z_y = b_1 z_{x1} + b_2 z_{x2} + u \tag{38}$$

Para representar este ejemplo, se utilizara la base de datos de los precios de las casas del libro de wooldrige

```
> modelo8=lm(scale(hprice2$lprice)~ scale(hprice2$lnox) + scale(hprice2$rooms)+scale(hprice2
> summary(modelo8)

Call:
lm(formula = scale(hprice2$lprice) ~ scale(hprice2$lnox) + scale(hprice2$rooms) +
    scale(hprice2$dist) + scale(hprice2$crime) + scale(hprice2$stratio))

Residuals:
    Min      1Q  Median      3Q     Max
-2.1330 -0.3111 -0.0158  0.2512  3.1827

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            -8.753e-16  2.632e-02   0.000        1
scale(hprice2$lnox)    -3.881e-01  4.712e-02  -8.237 1.56e-15 ***
scale(hprice2$rooms)    4.155e-01  2.925e-02  14.204  < 2e-16 ***
scale(hprice2$dist)    -2.188e-01  4.524e-02  -4.836 1.77e-06 ***
scale(hprice2$crime)   -2.867e-01  2.995e-02  -9.572  < 2e-16 ***
scale(hprice2$stratio) -2.295e-01  2.913e-02  -7.878 2.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5922 on 500 degrees of freedom
Multiple R-squared:  0.6528,        Adjusted R-squared:  0.6494
F-statistic:   188 on 5 and 500 DF,  p-value: < 2.2e-16
```

1.14. Modelos con funciones cuadraticas

Continuando con las transformaciones, se pueden encontrar también aquellas en donde una variable se eleva al cuadrado para capturar los efectos marginales. Usando el ejemplo de los salarios, se puede ver un efecto marginal negativo de la experiencie en el crecimiento de los salarios.

```
> data("wage1")
> modelo8=lm(wage1$wage~ wage1$educ+wage1$exper+wage1$expersq)
> summary(modelo8)

Call:
lm(formula = wage1$wage ~ wage1$educ + wage1$exper + wage1$expersq)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0692 -2.0837 -0.5417  1.2860 15.1363

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.964890   0.752153  -5.271 1.99e-07 ***
wage1$educ     0.595343   0.053025  11.228  < 2e-16 ***
wage1$exper    0.268287   0.036897   7.271 1.31e-12 ***
wage1$expersq -0.004612   0.000822  -5.611 3.26e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.166 on 522 degrees of freedom
Multiple R-squared:  0.2692,        Adjusted R-squared:  0.265
F-statistic: 64.11 on 3 and 522 DF,  p-value: < 2.2e-16

>
```

Realice un modelo en donde usando las variables de esta base de datos se obtenga un r cuadrado significativo

```
> #modelo9=lm(wage1$wage~ )
>
> #summary(modelo9)
```

1.15. Predicción de Variables

Usando el ejemplo de los logarítmos del salario nos concetramos en determinar el valor $\hat{y}$ de acuerdo al modelo propuesto anteriormente.

```
> library(wooldridge)
> data("ceosal2")
```

```
> modelo10=lm(ceosal2$lsalary~ceosal2$lsales+ceosal2$lmktval+ceosal2$ceoten)
>  pred=predict(modelo10,interval = "prediction")
> matplot(pred,type="l")

> data("hprice2")
> modelo11<-lm(hprice2$lprice~hprice2$nox+hprice2$dist+hprice2$rooms+I(hprice2$rooms^2)+hpri
> pred1<-predict(modelo11,interval="confidence")
> matplot(pred1,type="l")
> X<-data.frame(rooms=3,nox=5.5498,dist=3.7958,stratio=18.4593)
> pred2<-predict(modelo11,X,interval = "prediction")
> matplot(pred2,type="l")
```

Tarea, generar el verdadero valor de la variable dependiente usando la varianza estimada del modelo.

```
> data("gpa2")
> modelo12<-lm(colgpa~sat+hsperc+hsize+I(hsize^2),data=gpa2)
> cvalues<-data.frame(sat=c(1200,900,1400),hsperc=c(30,20,5),hsize=c(5,3,1))
> pre=predict(modelo12,cvalues,interval = "prediction")
> plot(pre[,1] ,type="l")
> lines(pre[,2],col="red")
> lines(pre[,3],col="green")
```

1.16. Variables Binarias y Cambios estructurales

El uso de variables cualitativas en los modelos ayuda a disminuir problemas como por ejemplo eliminación de variables irrelevantes en los modelos. Por ejemplo, en algunos modelos económicos, la variable dependiente que puede ser el ingreso, depende no solo de variables cuantitativas com el tiempo de estudio, sino también del genero o la raza. Si bien han cambiando algunos de estos aspectos en algunos estudios, estos todavía son relevantes y merecen ser estudiados en algunas ocasiones.

En cuanto al cambio estructural, este se puede dar en dos vias, la primera en una serie temporal que no es evaluada en este curso pero que se presenta mucho en series que han sufrido algún cambio adicional en su estrcutura por algún choque. En series de corte tranvsersal son variables que se diferencian de otras y que por su nueva condición se debe tener en cuenta ese cambio dentro del modelo.

En cuanto al valor que se le otorga a las variables cualitativas, este puede variar de acuerdo al tipo de información que busca almacenar. Existen variables que son nominales en donde los valores asignables están en uno o cero, mientras que las variables ordinales cambian de acuerdo al orden jearquico de la observación.

En el primer ejemplo vamos a tratar el salario por hora de acuerdo a las diferencias por genero en donde las mujeres toman el valor de uno y los hombres

el valor de cero. Esta variable se trabajará en conjunto coin las demás que se han usado para determinar el salario por hora.

```
> data("wage1")
> modelo13=lm(wage~female+educ+exper+tenure,data=wage1)
> summary(modelo13)

Call:
lm(formula = wage ~ female + educ + exper + tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7675 -1.8080 -0.4229  1.0467 14.0075

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.56794    0.72455  -2.164   0.0309 *
female      -1.81085    0.26483  -6.838 2.26e-11 ***
educ         0.57150    0.04934  11.584  < 2e-16 ***
exper        0.02540    0.01157   2.195   0.0286 *
tenure       0.14101    0.02116   6.663 6.83e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.958 on 521 degrees of freedom
Multiple R-squared:  0.3635,        Adjusted R-squared:  0.3587
F-statistic:  74.4 on 4 and 521 DF,  p-value: < 2.2e-16

>

> modelo14=lm(wage~educ+exper+tenure,data=wage1)
> summary(modelo14)

Call:
lm(formula = wage ~ educ + exper + tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6068 -1.7747 -0.6279  1.1969 14.6536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
educ         0.59897    0.05128  11.679  < 2e-16 ***
exper        0.02234    0.01206   1.853   0.0645 .
tenure       0.16927    0.02164   7.820 2.93e-14 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.084 on 522 degrees of freedom
Multiple R-squared:  0.3064,        Adjusted R-squared:  0.3024
F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

Como se observa en los resultados de cada una de las regresiones, el modelo 13 presenta la variable dummy de genero que no solo es significativa, sino que también tiene un coeficiente de correlación más alto mejorando el resultado del modelo.

A diferencia del ejemplo anterior, el siguiente combina dos variables categóricas en donde el efecto que se busca es no solo mirar el genero, sino también el efecto de que sea una persona casada.

```
> modelo15=lm(wage~married*female+educ+exper+tenure+I(exper^2)+I(tenure^2),data=wage1)
> summary(modelo15)

Call:
lm(formula = wage ~ married * female + educ + exper + tenure +
    I(exper^2) + I(tenure^2), data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1941 -1.6748 -0.4404  1.1067 13.0270

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.8482401  0.7178692  -3.968 8.28e-05 ***
married         1.3007620  0.3973564   3.274  0.00113 **
female         -0.3938605  0.4001190  -0.984  0.32540
educ            0.5229128  0.0480534  10.882  < 2e-16 ***
exper           0.1831766  0.0376334   4.867 1.50e-06 ***
tenure          0.1941521  0.0485380   4.000 7.26e-05 ***
I(exper^2)     -0.0037014  0.0007926  -4.670 3.85e-06 ***
I(tenure^2)    -0.0026140  0.0016599  -1.575  0.11591
married:female -2.3074103  0.5151465  -4.479 9.23e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.823 on 517 degrees of freedom
Multiple R-squared:  0.4246,        Adjusted R-squared:  0.4157
F-statistic: 47.68 on 8 and 517 DF,  p-value: < 2.2e-16
```

Con respecto al modelo 14 y al modelo 13, se puede ver que la variable dummy de genero dejó de ser significativa, dando paso a la variable estado civil y a la variable estado civil unida con genero. Es interesante ver como el hecho

de que una mujer esté casada en este modelo produsca un efecto negativo en el salario pagado por hora.

1.16.1. Variables Lógicas

El uso de variables lógicas en los modelos es otra alternativa para incluir variables binarias. Estos valores pueden ser transformados en cero y uno con la función en R as.numeric donde TRUE es igua a 1 y FALSE es igual a cero.

```
> wage1$female<-as.logical(wage1$female)
> table(wage1$female)

FALSE  TRUE
  274   252

> modelo16<-lm(wage~female+educ+exper+tenure,data=wage1)
> summary(modelo16)

Call:
lm(formula = wage ~ female + educ + exper + tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7675 -1.8080 -0.4229  1.0467 14.0075

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.56794    0.72455  -2.164   0.0309 *
femaleTRUE  -1.81085    0.26483  -6.838 2.26e-11 ***
educ         0.57150    0.04934  11.584  < 2e-16 ***
exper        0.02540    0.01157   2.195   0.0286 *
tenure       0.14101    0.02116   6.663 6.83e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.958 on 521 degrees of freedom
Multiple R-squared:  0.3635,        Adjusted R-squared:  0.3587
F-statistic:  74.4 on 4 and 521 DF,  p-value: < 2.2e-16
```

1.16.2. Factores

Los factores en R son elementos que se dejan transformar en una lista y por lo general muestran niveles, por ejemplo en una lista pueden haber continentes, países y ciudades en donde cada región es un factor. Si estas variables están generadas como factores permitirá manejar cierto tipode gráficos y funciones de forma más eficiente.