Regresión Lineal Multiple

DIMAS RAMIREZ LUIS DANIEL

7/12/2020

Regresión Lineal Múltiple

$$y_i = \hat{\beta} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + ...\hat{\beta}_k x_{ik} + \epsilon_i$$

Formato de vectores y matrices

$$Y = X\beta + \epsilon$$

La forma de calcular las betas es:

$$\hat{\beta} = (Xt X)^{-1} Xt Y$$

Modelo básico con 2 betas:

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.6.3
```

```
Base <- read_excel("Base de datos encuestas.xlsx")
names(Base)
```

```
##  [1] "MARCA TEMPORAL"
##  [2] "NOMBRE DE USUARIO"
##  [3] "¿CUAL ES TU EDAD?"
##  [4] "¿CUAL ES TU SEXO BIOLOGICO?"
##  [5] "¿QUE CARRERA ESTUDIAS?"
##  [6] "¿TE GUSTA TU CARRERA?"
##  [7] "¿EN DONDE HICISTE TU BACHILLERATO?"
##  [8] "¿CUAL FUE TU PROMEDIO DE BACHILLERATO?"
##  [9] "¿CUAL ES TU PROMEDIO ACTUAL EN LA CARRERA?"
## [10] "¿CUAL ES TU AVANCE DE CREDITOS?"
## [11] "¿CUAL ES TU NIVEL DE INGLES?"
## [12] "¿CUANTAS MATERIAS INSCRIBES EN PROMEDIO AL SEMESTRE?"
## [13] "¿CUANTAS MATERIAS HAS REPROBADO?"
## [14] "¿CUANTAS HORAS AL DIA, PASAS EN PROMEDIO EN LA FACULTAD?"
## [15] "¿REALIZAS ALGUNA ACTIVIDAD EXTRACURRICULAR?"
## [16] "¿CUANTAS HORAS EN PROMEDIO, LE DEDICAS A ESA ACTIVIDAD POR SEMANA?"
## [17] "¿CUANTAS HORAS AL DIA TE TOMA TRANSPORTARTE A LA ESCUELA?"
```

```
## [18] "¿QUE TIPO DE TRANSPORTE UTILIZAS PARA IR A LA ESCUELA?"
## [19] "¿ERES FORANEO?"
## [20] "¿CON QUIEN VIVES?"
## [21] "¿CUAL ES EL GRADO MAXIMO DE ESTUDIOS DE TU MADRE?"
## [22] "¿CUAL ES EL GRADO MAXIMO DE ESTUDIOS DE TU PADRE?"
## [23] "¿CUANTAS PERSONAS HABITAN CONTIGO?"
## [24] "¿CUENTAS CON HABITACION PROPIA?"
## [25] "¿ERES RESPONSABLE DE ALGUNA MASCOTA?"
## [26] "¿CUENTAS CON ALGUNA BECA?"
## [27] "¿ESTUDIAS Y TRABAJAS?"
## [28] "¿CUANTAS HORAS A LA SEMANA TRABAJAS?"
## [29] "¿CUAL ES EL INGRESO PROMEDIO MENSUAL DE TU FAMILIA?"
## [30] "27.\t¿CUANTO GASTAS SEMANALMENTE EN COSAS RELACIONADAS CON LA ESCUELA?"
## [31] "28.\t¿TE ENCUENTRAS EN UNA RELACION CON ALGUNA PERSONA?"
## [32] "29.\t¿TIENES HIJOS?"
## [33] "30.\t¿CUANTOS HIJOS TIENES?"
## [34] "31.\t¿TE HACES CARGO AL 100% DE TUS HIJOS?"
## [35] "32.\t¿CUANTAS HORAS DIARIAS DUERMES EN PROMEDIO?"
## [36] "33.\t¿CUENTAS CON LAPTOP PROPIA?"
## [37] "¿CUANTAS VECES AL MES CONSUMES ALCOHOL?"
## [38] "¿QUE TIPO DE BEBIDA ES LA QUE MAS CONSUMES?"
```

```
P_A <- Base$`¿CUAL ES TU PROMEDIO ACTUAL EN LA CARRERA?` #y
P_B <- Base$`¿CUAL FUE TU PROMEDIO DE BACHILLERATO?` #x

P_A
```

```
##   [1] 8.70 8.90 9.20 8.90 8.00 7.30 8.00 8.90 7.40 8.50 8.30 8.17 8.80 8.60 8.40
##  [16] 9.20 8.60 8.04 8.85 9.10 9.20 8.75 8.00 7.61 9.50 8.60 7.00 7.30 8.50 7.99
##  [31] 8.89 9.00 9.00 9.00 8.50 9.40 7.80 8.60 8.20 8.00 7.80 8.90 7.00 8.00 8.50
##  [46] 8.70 9.64 7.60 8.65 9.00 9.36 9.10 8.00 8.50 8.20 9.60 8.70 8.60 8.90 8.70
##  [61] 9.30 8.00 8.51 9.00 8.00 8.50 8.57 8.96 9.27 8.60 7.00 6.50 8.94 7.00 8.91
##  [76] 7.20 9.30 8.80 7.28 8.30 7.99 8.40 8.10 9.00 7.90 8.10 8.50 8.80 8.50 8.80
##  [91] 8.40 9.70 8.90 8.00 8.44 8.50 7.91 9.50 7.85 8.10 9.50 8.30 9.50 9.60 9.20
## [106] 8.50 9.80 8.90 8.20 8.27 8.38 8.10 7.20 8.30 8.03 8.20 8.30 7.72 8.50 8.10
## [121] 8.40 7.60 8.80 8.90 9.60 8.68 8.30 8.10 7.90 8.40 8.30 8.30 9.70 9.60 8.90
## [136] 8.70 9.00 9.10 9.13 7.90 9.48 8.05 8.83 8.86 8.50 8.89 7.60 7.90 8.50 8.50
## [151] 9.10 9.00 8.70 7.44 9.40 9.40 8.88 8.24 9.61 8.48 9.41 8.96 8.00 8.60 8.85
## [166] 8.75 8.54 8.50 8.70 8.32 9.00 7.51 8.32 9.10 9.30 8.90 9.50 8.70 9.30 9.13
## [181] 8.50 8.50 9.46 7.80 8.82 8.26 8.30 8.32 9.30 8.60 9.00 8.59 8.60 8.20 8.60
## [196] 8.60 8.45 8.00 8.90 8.10 8.80 9.00 8.00 6.40 8.00 8.70 9.00 8.90 8.60 8.23
## [211] 8.23 8.23 8.56 8.14 9.20 9.30 8.00 8.53 8.87 7.79 8.50 8.30 8.73 7.90 9.70
## [226] 9.76 8.12 8.60 8.90 7.80 9.86 9.70 9.70 8.93 8.80 8.68 8.30 8.80 9.00 9.10
## [241] 9.02 8.17 8.50 9.21 8.00 9.00 8.80 8.80 8.80 8.60 9.51 8.35 8.40 8.20 7.80
## [256] 8.50 9.20 9.30 8.60 9.50 9.02 8.41 7.80 8.70 8.90 8.54 8.68 8.00 8.60 9.95
## [271] 9.30 7.79 7.90 7.50 8.00 8.10 9.24 8.10 8.90 8.00 8.34 8.50 8.30 8.40 7.90
## [286] 7.80 8.00 7.70 7.56 7.60 8.30 8.15 8.00 9.23 9.10 9.03 8.40 9.22 7.90 7.00
## [301] 7.00 7.40 8.30 8.25 8.80 8.20 6.50 8.20 9.20 9.46 9.00 9.70 9.21 9.00 7.80
## [316] 8.00 7.60 8.00 8.60 8.80 8.60 7.83 9.01 7.90 7.80 8.32 7.70 8.40 8.30 7.90
## [331] 8.74 8.60 9.00 8.20 8.70 8.60 9.00 9.10 8.10 9.20 7.90 9.20 9.50 9.20 8.95
## [346] 7.70 8.70 7.00 8.00 7.20 8.50 7.80 8.30 8.30 9.01 7.00 7.40 7.40 8.60 8.97
## [361] 9.03 8.80 8.00 7.80 7.80 8.10
```

```r
M_A <- rep(NA, times=366*2)#matriz equis ampliada
dim(M_A) <- c(366,2)
M_A[,1] <- 1
M_A[,2] <- P_B
```

Calculo de las betas:

```r
betas<- solve(t(M_A)%*%M_A)%*%t(M_A)%*%P_A
betas
```

```
##           [,1]
## [1,] 4.3215036
## [2,] 0.4770765
```

```r
# t= transpuesta
#para multiplicar las matrices se tiene que utilizar % * %
# solve se utilizar para sacar la inversa

lm(P_A~P_B)
```

```
##
## Call:
## lm(formula = P_A ~ P_B)
##
## Coefficients:
## (Intercept)          P_B
##      4.3215       0.4771
```

Modelo múltiple:

```r
#Filtrar datos
#0 y 1 hacen la Función indicadora

lt<- Base$`33.  ¿CUENTAS CON LAPTOP PROPIA?`
lt[lt=="SI"] <- 1
lt[lt=="NO"] <- 0
lt<- as.numeric(lt)
```

Modelo incluyendo si el estudiante tiene o no tiene laptop, análisis estadística para verificar que la variable es significativa

```r
# as.factor funciona para denotar al 0 y 1 como presencia y ausencia de dato
ml_m1 <- lm(P_A~P_B+as.factor(lt))
```

$$C_A = 4.322 + 0.477P_B - 0.000518LT$$

Notamos que la variable de si tiene o no tiene laptop es casi cero

Checando el análisis estadístico de laptop

```r
summary(ml_m1)
```

```
##
## Call:
## lm(formula = P_A ~ P_B + as.factor(lt))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83352 -0.37672 -0.00373  0.33596  1.33890
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.322e+00  4.194e-01  10.305   <2e-16 ***
## P_B             4.771e-01  4.867e-02   9.802   <2e-16 ***
## as.factor(lt)1 -8.193e-05  9.280e-02  -0.001    0.999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5658 on 363 degrees of freedom
## Multiple R-squared:  0.2161, Adjusted R-squared:  0.2117
## F-statistic: 50.02 on 2 and 363 DF,  p-value: < 2.2e-16
```

Del resumen notamos que lt es una varaiable que no es representativa en el modelo debido a que nuestra signifacancia es del 5% por que $\alpha/2 = 2.5\%$ y la probabilidad mostrada para lt es de 0.99 por lo tanto se debe rechazar esa variable del modelo ya que no es representativa. Adicional a esto, notamos que R no muestra ningún asterisco que nos de indicio para tomarlo en cuenta.