

Segunda entrega de ejercicios – Regresión Lineal Simple.

EJERCICIO 1

En el modelo lineal de la forma:

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

Donde la *matriz de diseño* X tiene n filas y $k + 1$ columnas (con $n < k + 1$), se define la matriz de proyección ortogonal:

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

Demuestre las siguientes propiedades:

1. $H^j = H \quad \forall j = 1, 2, 3, \dots$
2. Si H es idempotente, entonces $I_{n \times n} - H$ también lo es
3. $H(I_{n \times n} - H) = \mathbf{0}_{n \times n}$

1)

Idempotencia de la matriz de proyección H :

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

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

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

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

Lo mismo es cierto para cualquier cantidad de H hasta generalizarlo para $H^k \quad k \in \mathbb{N}$.

2)

$$(I_{n \times n} - H) = (I_{n \times n} - H)(I_{n \times n} - H)$$

$$(I_{n \times n} - H) = I_{n \times n}^2 - 2I_{n \times n}H + H^2$$

$$(I_{n \times n} - H) = I_{n \times n} - 2H + H \quad (\text{idempotencia de la matriz identidad})$$

$$(I_{n \times n} - H) = (I_{n \times n} - H)$$

3)

$$H(I_{n \times n} - H) = HI_{n \times n} - H^2$$

$$H(I_{n \times n} - H) = H - H \quad (\text{idempotencia de la matriz } H)$$

$$H(I_{n \times n} - H) = 0$$

EJERCICIO 7

En el archivo *pgc.txt* se encuentran las mediciones de 252 personas. Se desea construir una fórmula que sea capaz de estimar el porcentaje de grasa corporal en función de algunas mediciones (circunferencia abdominal, circunferencia de cintura, de cuello, entre otras).

1. Visualice la relación de cada una de las medidas corporales con respecto a la variable *porc_grasa* y estime las correlaciones.
2. Estime la matriz de correlación de las mediciones corporales.
3. Describa con sus palabras las conclusiones que haya podido obtener a partir de los dos puntos anteriores.
4. Realice un conjunto de modelos de *RLS* que expliquen la relación de cada medida corporal con respecto al porcentaje de grasa corporal. ¿Qué observa?
5. Ajuste la siguiente serie de modelo:

- $porc_grasa_i = \beta_0 + \beta_1 abdom_i + \epsilon_i$.
- $porc_grasa_i = \beta_0 + \beta_1 abdom_i + \beta_2 cintura_i + \epsilon_i$.
- $porc_grasa_i = \beta_0 + \beta_1 abdom_i + \beta_2 cintura_i + \beta_3 cuello_i + \epsilon_i$.
- $porc_grasa_i = \beta_0 + \beta_1 abdom_i + \beta_2 cintura_i + \beta_3 cuello_i + \beta_4 muñeca + \epsilon_i$.

A partir de las estimaciones de estos modelo:

- Compare la evolución del R^2 .
- Compare la evolución de la estimación de σ^2 .
- Compare la evolución del coeficiente asociado a la circunferencia abdominal.

```
## corplot 0.92 loaded
```

1. Visualice la relacion de cada una de las medidas corporales con respecto a la variable *porc_grasa* y estime las correlaciones

```
mod <- lm(porc_grasa ~ cuello + pecho + abdomen + cadera + muslo
          + rodilla + tobillo + biceps + brazo + muneca, data = datos )
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = porc_grasa ~ cuello + pecho + abdomen + cadera +
##      muslo + rodilla + tobillo + biceps + brazo + muneca, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.3159 -2.7435 -0.1584  2.8388 10.5150
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.228749   6.214309   1.163  0.24588
## cuello      -0.581947   0.208580  -2.790  0.00569 **
```

```
## pecho          -0.090847    0.085430   -1.063    0.28866
## abdomen         0.960229    0.071582   13.414    < 2e-16 ***
## cadera          -0.391355    0.112686   -3.473    0.00061 ***
## muslo           0.133708    0.124922    1.070    0.28554
## rodilla         -0.094055    0.212394   -0.443    0.65828
## tobillo         0.004222    0.203175    0.021    0.98344
## biceps          0.111196    0.159118    0.699    0.48533
## brazo           0.344536    0.185511    1.857    0.06450 .
## muneca          -1.353472    0.471410   -2.871    0.00445 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.071 on 241 degrees of freedom
## Multiple R-squared:  0.7351, Adjusted R-squared:  0.7241
## F-statistic: 66.87 on 10 and 241 DF,  p-value: < 2.2e-16
```

2. Estime la matriz de correlacion de las mediciones corporales.

```
#Matriz de correlaciones
```

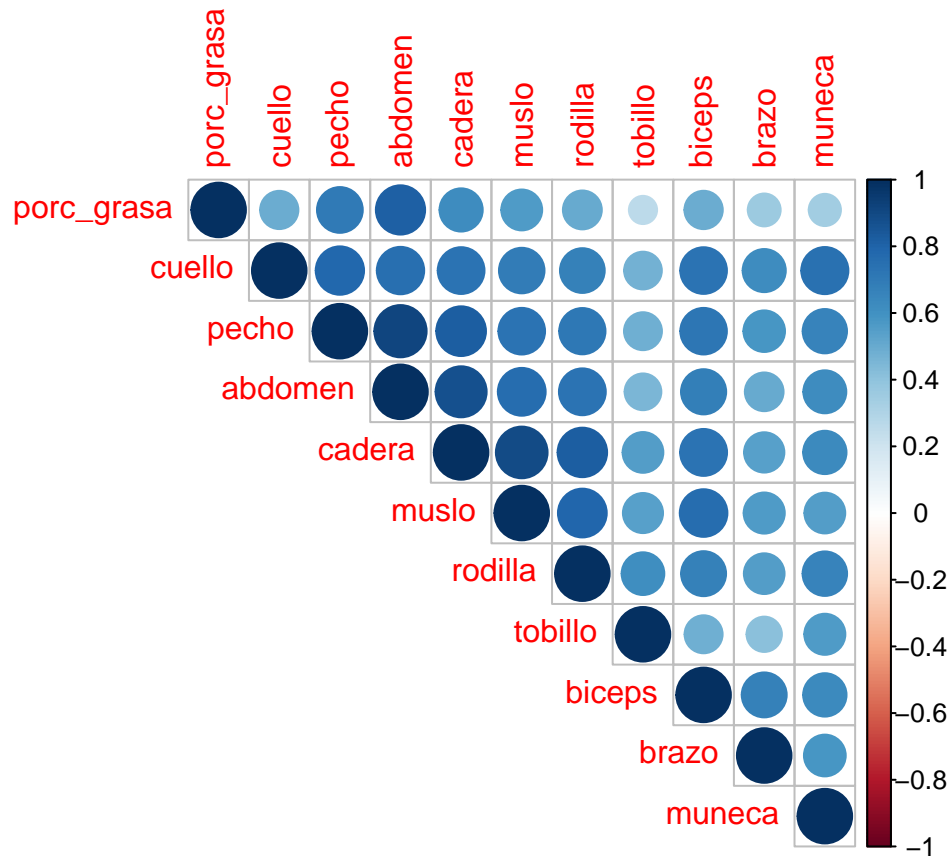
```
analisis <- datos[,-(c(2,3,4))]
```

```
correlacion <- round(cor(analisis),3)
```

```
correlacion
```

```
##          porc_grasa  cuello pecho abdomen cadera muslo rodilla tobillo biceps
## porc_grasa      1.000  0.491 0.703   0.814  0.626 0.561   0.508  0.267  0.493
## cuello          0.491  1.000 0.785   0.754  0.735 0.696   0.672  0.478  0.731
## pecho           0.703  0.785 1.000   0.916  0.829 0.730   0.719  0.483  0.728
## abdomen         0.814  0.754 0.916   1.000  0.874 0.767   0.737  0.453  0.685
## cadera          0.626  0.735 0.829   0.874  1.000 0.896   0.823  0.558  0.739
## muslo           0.561  0.696 0.730   0.767  0.896 1.000   0.799  0.540  0.761
## rodilla         0.508  0.672 0.719   0.737  0.823 0.799   1.000  0.612  0.679
## tobillo         0.267  0.478 0.483   0.453  0.558 0.540   0.612  1.000  0.485
## biceps          0.493  0.731 0.728   0.685  0.739 0.761   0.679  0.485  1.000
## brazo           0.363  0.624 0.580   0.503  0.545 0.567   0.556  0.419  0.678
## muneca          0.348  0.745 0.660   0.620  0.630 0.559   0.665  0.566  0.632
##
##          brazo muneca
## porc_grasa 0.363  0.348
## cuello     0.624  0.745
## pecho      0.580  0.660
## abdomen    0.503  0.620
## cadera     0.545  0.630
## muslo      0.567  0.559
## rodilla    0.556  0.665
## tobillo    0.419  0.566
## biceps     0.678  0.632
## brazo      1.000  0.586
## muneca     0.586  1.000
```

```
corrplot(correlacion, method="circle", type="upper", pch.col = 10)
```



3. Describa con sus palabras las conclusiones que haya podido obtener a partir de los dos puntos anteriores.

Las variables están corelacionadas de manera positiva, lo cual tiene sentido al menos desde una noción más biológica donde si una parte del cuerpo es de un tamaño la otra guardará cierta proporción, así mismo se observa que es el tamaño del tobillo la que menos incide en las otras variables.

Por otra parte, el modelo del punto uno, explica en un 73,5% el porcentaje de grasa corporal, siendo cadera, cuello, abdomen, y muñeca las que más información arrojan sobre el modelo en sí.

4. Realice un conjunto de modelos de RLS que expliquen la relacion de cada medida corporal con respecto al porcentaje de grasa corporal. ¿Que observa?

```
mod1 <- lm(porc_grasa ~ cuello , data = datos )
mod2 <- lm(porc_grasa ~ pecho , data = datos )
mod3 <- lm(porc_grasa ~ abdomen , data = datos )
mod4 <- lm(porc_grasa ~ cadera , data = datos )
mod5<- lm(porc_grasa ~ muslo , data = datos )
mod6<- lm(porc_grasa ~ rodilla , data = datos )
```

```

mod7<- lm(porc_grasa ~ tobillo , data = datos )
mod8<- lm(porc_grasa ~ biceps , data = datos )
mod9<- lm(porc_grasa ~ brazo , data = datos )
mod10<- lm(porc_grasa ~ muneca , data = datos )

algo <- rbind(coef(mod1),coef(mod2),coef(mod3),coef(mod4),coef(mod5),coef(mod6)
             ,coef(mod7),coef(mod8),coef(mod9),coef(mod10))

algo <- algo[,2]

RLS_porcGrasa <- as.data.frame(cbind(c("Cuello","Pecho","Abdomen","Cadera","Muslo","Rodilla","Tobillo",
colnames(RLS_porcGrasa)<-c("Variable","Beta")

class(RLS_porcGrasa)

```

```
## [1] "data.frame"
```

```
RLS_porcGrasa <- RLS_porcGrasa[order(RLS_porcGrasa$Beta,decreasing = T),]
```

```
RLS_porcGrasa
```

```
##      Variable      Beta
## 10  Muñeca  2.88563598657192
## 6   Rodilla  1.63187973004883
## 1   Cuello   1.5670899630347
## 9   Brazo   1.39343956604193
## 8   Biceps  1.26483448827388
## 7   Tobillo  1.22001365754821
## 5   Muslo   0.828661701987687
## 4   Cadera  0.676950137461153
## 2   Pecho   0.646222311718054
## 3   Abdomen 0.584890527012418
```

```
#abdomen_grafico <- ggplot(data = datos, mapping = aes(x=porc_grasa,y=abdomen))+geom_point()+ylab("Centim
```

```
#muneca_grafico <- ggplot(data = datos, mapping = aes(x=porc_grasa,y=muneca))+geom_point()+ylab("Centim
```

```
RLS_porcGrasa
```

```
##      Variable      Beta
## 10  Muñeca  2.88563598657192
## 6   Rodilla  1.63187973004883
## 1   Cuello   1.5670899630347
## 9   Brazo   1.39343956604193
## 8   Biceps  1.26483448827388
```

```
## 7 Tobillo 1.22001365754821
## 5 Muslo 0.828661701987687
## 4 Cadera 0.676950137461153
## 2 Pecho 0.646222311718054
## 3 Abdomen 0.584890527012418
```

```
summary(mod3)
```

```
##
## Call:
## lm(formula = porc_grasa ~ abdomen, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6257  -3.4672   0.0111   3.1415  11.9754
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.19661     2.46229  -14.29  <2e-16 ***
## abdomen      0.58489     0.02643   22.13  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.514 on 250 degrees of freedom
## Multiple R-squared:  0.6621, Adjusted R-squared:  0.6608
## F-statistic: 489.9 on 1 and 250 DF, p-value: < 2.2e-16
```

```
#abdomen_grafico
```

```
summary(mod10)
```

```
##
## Call:
## lm(formula = porc_grasa ~ muneca, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8183  -5.5840   0.1231   5.0703  25.6703
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.6660     8.9870  -3.746 0.000223 ***
## muneca       2.8856     0.4923   5.861 1.45e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.282 on 250 degrees of freedom
## Multiple R-squared:  0.1208, Adjusted R-squared:  0.1173
## F-statistic: 34.35 on 1 and 250 DF, p-value: 1.446e-08
```

```
#muneca_grafico
```

Según este resumen de datos, cada variable tiene una relación positiva con el porcentaje de grasa. Pondremos particular atención en las variables abdomen y muñeca.

Para el caso de abdomen, el incremento de un punto porcentual de grasa corporal implica el aumento de medio centímetro de abdomen, mientras que para la medida de muñeca representa un incremento de 2.8 centímetros. Por otra parte, explicar el porcentaje de grasa corporal mediante la medida de abdomen es más confiable que por el de la muñeca, ya que entre otras cosas, el coeficiente R2 es de 66.2% para el primero y de 12.1% para el segundo

5. Ajuste la siguiente serie de modelo:

- $porc_grasa_i = \beta_0 + \beta_1 abdomen_i + \epsilon_i.$
- $porc_grasa_i = \beta_0 + \beta_1 abdomen_i + \beta_2 cintura_i + \epsilon_i.$
- $porc_grasa_i = \beta_0 + \beta_1 abdomen_i + \beta_2 cintura_i + \beta_3 cuello_i + \epsilon_i.$
- $porc_grasa_i = \beta_0 + \beta_1 abdomen_i + \beta_2 cintura_i + \beta_3 cuello_i + \beta_4 muñeca + \epsilon_i.$

A partir de las estimaciones de estos modelo:

Compare la evolucion del R2.

Compare la evolucion de la estimacion de sigma2.

Compare la evolucion del coefiente asociado a la circunferencia abdominal.

```
mod11<- lm(porc_grasa ~ abdomen , data = datos )
```

```
summary(mod11)
```

```
##
## Call:
## lm(formula = porc_grasa ~ abdomen, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6257  -3.4672   0.0111   3.1415  11.9754
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.19661     2.46229  -14.29  <2e-16 ***
## abdomen      0.58489     0.02643   22.13  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.514 on 250 degrees of freedom
## Multiple R-squared:  0.6621, Adjusted R-squared:  0.6608
## F-statistic: 489.9 on 1 and 250 DF,  p-value: < 2.2e-16
```

```
mod12<- lm(porc_grasa ~ abdomen + cadera , data = datos )
```

```
summary(mod12)
```



```
##
## Call:
## lm(formula = porc_grasa ~ abdomen + cadera, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.532  -3.153  -0.256   2.953  11.746
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.09863    4.30711  -3.970 9.41e-05 ***
## abdomen      0.81259     0.05194  15.644 < 2e-16 ***
## cadera      -0.39210     0.07818  -5.015 1.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.311 on 249 degrees of freedom
## Multiple R-squared:  0.6931, Adjusted R-squared:  0.6907
## F-statistic: 281.2 on 2 and 249 DF,  p-value: < 2.2e-16
```

```
mod13<- lm(porc_grasa ~ abdomen + cadera + cuello , data = datos )
summary(mod13)
```

```
##
## Call:
## lm(formula = porc_grasa ~ abdomen + cadera + cuello, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.9972 -2.9498 -0.1737  2.8267 12.4451
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.33037    5.07977  -0.852   0.395
## abdomen      0.89166     0.05330  16.728 < 2e-16 ***
## cadera      -0.31127     0.07771  -4.006 8.18e-05 ***
## cuello      -0.74128     0.16939  -4.376 1.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.162 on 248 degrees of freedom
## Multiple R-squared:  0.7151, Adjusted R-squared:  0.7117
## F-statistic: 207.5 on 3 and 248 DF,  p-value: < 2.2e-16
```

```
mod14<- lm(porc_grasa ~ abdomen + cadera + cuello + muneca , data = datos )
summary(mod14)
```

```
##
## Call:
## lm(formula = porc_grasa ~ abdomen + cadera + cuello + muneca,
##      data = datos)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3584  -2.7101  -0.2303   2.9092  10.9033
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.65504     5.79616   0.803 0.422675
## abdomen       0.89202     0.05243  17.014 < 2e-16 ***
## cadera       -0.28141     0.07705  -3.652 0.000317 ***
## cuello       -0.43807     0.19388  -2.260 0.024722 *
## muneca       -1.29023     0.42188  -3.058 0.002471 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.094 on 247 degrees of freedom
## Multiple R-squared:  0.7255, Adjusted R-squared:  0.7211
## F-statistic: 163.2 on 4 and 247 DF,  p-value: < 2.2e-16
```

```
comparación <- as.data.frame(cbind(c(0.6621,0.6931,0.7151,0.7255),
                                   c(4.514,4.311,4.162,4.094),
                                   c(0.58489,0.81259,0.89166,0.89202)))

colnames(comparación)<-c("R-squared","Residual standard error","Beta_abdomen")

rownames(comparación)<-c("Abdomen","Abdomen/Cadera","Abdomen/Cadera/Cuello","Abdomen/Cadera/Cuello/Muñeca")

comparación
```

```
##
##              R-squared Residual standard error Beta_abdomen
## Abdomen              0.6621              4.514      0.58489
## Abdomen/Cadera       0.6931              4.311      0.81259
## Abdomen/Cadera/Cuello 0.7151              4.162      0.89166
## Abdomen/Cadera/Cuello/Muñeca 0.7255              4.094      0.89202
```