# Tarea 4

## EJERCICIO 1: MOLab (con α=0.05)

1. Para hallar la ecuación, coeficientes y varianza del modelo estimado se siguen los siguientes pasos en RStudio:

```
datos=read.table("tiemposCPU.txt", header=T)
attach(datos)

modelo=lm(CPU ~ size + format)
summary(modelo)
```

Obteniendo los siguientes resultados.

```
Call:
lm(formula = CPU ~ size + format)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36597 -0.14712 -0.01502  0.16922  0.60624

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.0077607  0.0900441  11.192 4.06e-16 ***
size         0.0028451  0.0001503  18.929  < 2e-16 ***
formatB      0.5417854  0.0674664   8.030 5.45e-11 ***
formatC     -0.2222004  0.0683042  -3.253  0.00191 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2186 on 58 degrees of freedom
Multiple R-squared:  0.8947,    Adjusted R-squared:  0.8893
F-statistic: 164.3 on 3 and 58 DF,  p-value: < 2.2e-16
```

Por tanto, la ecuación del modelo estimado quedaría:
**CPU=1.0077607 + 0.0028451*size + 0.5417854*Zb – 0.2222004*Zc + error**
Tomando como referencia la variable Za, variable cualitativa format de valor A.
Siendo los valores estimados: $\boldsymbol{\beta_0}$**=1.0077607**, $\boldsymbol{\beta_1}$**=0.0028451**, $\boldsymbol{\alpha_1}$**=0.5417854** y $\boldsymbol{\alpha_2}$**= - 0.2222004**.

El **coeficiente de determinación** del modelo será $\boldsymbol{R^2}$**=89,47%**. Los regresores explican un 89,47% de la variabilidad de CPU.
El **coeficiente de determinación corregido** será **88,93%**, y se utiliza para comparar con modelos con distinto número de regresores.
La desviación típica es 0.2186 por lo que la **varianza residual** resulta ser **0.04779**. Ya que la varianza es el cuadrado de la desviación.

2. Para el modelo anterior:
   - $\boldsymbol{\beta_0}$ **= 1.0077607.** Este valor tiene poca interpretación física, ya que es la ordenada en el origen tomando como referencia la variable Za, variable cualitativa format A. Según las alphas estimadas este valor se verá alterado según el respectivo formato, influyendo en la variable CPU. Es el valor que toma CPU cuando el resto de variables son nulas.

- $\beta_1$**=0.0028451.** En la ecuación del modelo este valor corresponde a la pendiente, valor que multiplica al valor dado de la variable size. Sabemos entonces que la pendiente es positiva. Esto significa que por cada incremento de una unidad de la variable size, la variable CPU aumentará en promedio un 0.0028451, siendo el resto de variables constantes.

Los intervalos de confianza para las betas anteriores han resultado ser:

```
> confint(modelo, level=0.95)
                  2.5 %       97.5 %
(Intercept)  0.827517885  1.188003586
size         0.002544218  0.003145946
formatB      0.406736637  0.676834089
formatC     -0.358926204 -0.085474608
```

- $\beta_0$**:** Tiene un intervalo de confianza de **[ 0.827517885 , 1.188003586 ]**.
- $\beta_1$: Tiene un intervalo de confianza de **[ 0.002544218 , 0.003145946 ]**.

3. Primero vamos a analizar los resultados del modelo anterior, tomando como referencia la letra A de la variable format.

El **formato B con A como referencia** tiene un **p-valor** de **5.45e-11**. Este valor es muy inferior para cualquier $\alpha$ por lo que **sí existen diferencias significativas** entre el tiempo computacional requerido por los formatos **A y B**.
El **formato C con A como referencia** tiene un **p-valor** de **0.00191**. Pese a que este valor no es tan pequeño como el anterior lo es lo suficiente para $\alpha$ =0.05 como para afirmar que **sí existen diferencias significativas** entre el tiempo computacional requerido por los formatos **A y C**.

Para estudiar la relación entre B – C y analizar su p-valor necesitamos tomar uno de estos dos como referencia. En mi caso, he tomado B como nueva referencia:

```
format2=relevel(format, ref='B')
modelo2=lm(CPU ~ size + format2)
summary(modelo2)
```

```
Call:
lm(formula = CPU ~ size + format2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36597 -0.14712 -0.01502  0.16922  0.60624

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.5495461  0.0902202   17.18  < 2e-16 ***
size         0.0028451  0.0001503   18.93  < 2e-16 ***
format2A    -0.5417854  0.0674664   -8.03 5.45e-11 ***
format2C    -0.7639858  0.0683044  -11.19 4.16e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2186 on 58 degrees of freedom
Multiple R-squared:  0.8947,    Adjusted R-squared:  0.8893
F-statistic: 164.3 on 3 and 58 DF,  p-value: < 2.2e-16
```

EL **formato C con B como referencia** tiene un **p-valor** de **4.16e-16**. Este valor es muy pequeño por lo que **sí existen diferencias significativas** entre el tiempo computacional requerido por los formatos **B y C**.

4. (A)
   Para esta primera predicción se realiza en Rstudio lo siguiente:

```
> prediccion1 = data.frame( size=200, format = 'B')
> predict(modelo, prediccion1, interval = 'prediction')
       fit      lwr      upr
1 2.118562 1.661083 2.576042
```

   Esta **primera predicción** toma un valor de **2.118562**.
   Teniendo cuidado, sabemos que nos piden el **intervalo de una nueva observación**, no el de medias. Siendo este **[ 1.661083 , 2.576042]**.

   (B)
   Para la segunda predicción realizamos y obtenemos:

```
> prediccion2 = data.frame( size=100, format = 'A')
> predict(modelo, prediccion2, interval = 'confidence')
       fit      lwr      upr
1 1.292269 1.136724 1.447814
```

   Obtenemos un valor para la **predicción** de **1.292269**.
   Calculando esta vez el **intervalo de las medias**. Siendo **[ 1.136724 , 1.447814 ]**.

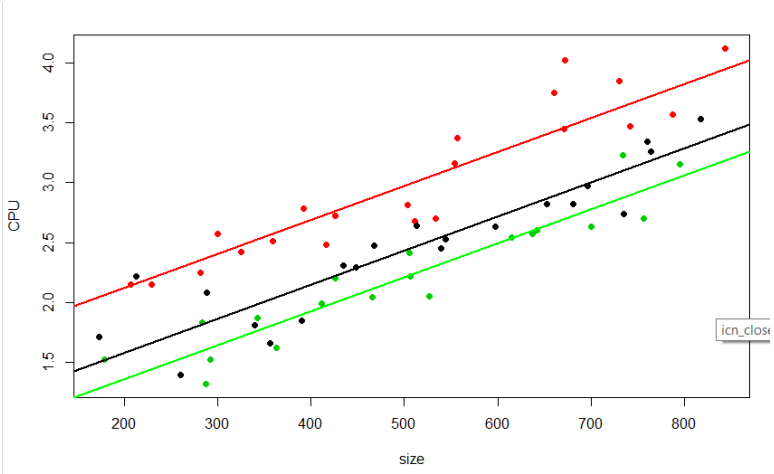
   Como suele ser normal, el intervalo de la observación tiene mayor amplitud que el de las medias.

5. Realizando en Rstudio:

```
plot(CPU~size, pch=19, data=datos, col=format, ylabel='CPU', xlabel='size')

abline(c(1.0077607,0.0028451), col='black', lwd=2)
abline(c(1.0077607 - 0.2222004,0.0028451), col='green', lwd=2)
abline(c(1.0077607 + 0.5417854,0.0028451), col='red', lwd=2)
```

   Se obtiene el siguiente gráfico.



Las ecuaciones de las rectas correspondientes son:

- Formato A ( Negro) : **CPU=1.0077607 + 0.0028451*size + error**
- Formato B ( Rojo): **CPU= (1.0077607 + 0.5417854) + 0.0028451*size + error**
  **CPU= (1.5495461) + 0.0028451*size + error**
- Formato C (Verde): **CPU= (1.0077607 – 0.2222004) + 0.0028451*size + error**
  **CPU= (0.7855603) + 0.0028451*size + error**

## EJERCICIO 2: Bebés (con α=0.05)

1. Para estimar el modelo primero tenemos que realizar los siguientes pasos en RStudio, donde se pasan las variables smoke y parity (que vienen expresada con datos numéricos pese a ser variables cualitativas), a partir de comparaciones, a variables completamente cualitativas.

```
datos=read.csv("babies2.csv", header=T)
attach(datos)

zpar1 = parity == 1
znopar1 = parity == 2
zfum = smoke == 1
znofum = smoke == -1

modelo=lm(bwt ~ gestation + age + height + weight + zpar1 + zfum)
summary(modelo)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1626.05  -288.67    -3.28   269.72  1455.67

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2247.3015   402.6569  -5.581 2.97e-08 ***
gestation      12.5887     0.8249  15.260  < 2e-16 ***
age            -0.2717     2.4333  -0.112  0.91113
height         12.6891     2.2555   5.626 2.31e-08 ***
weight          3.1076     1.5728   1.976  0.04841 *
zpar1TRUE     -94.7987    32.0090  -2.962  0.00312 **
zfumTRUE     -237.9990    27.0414  -8.801  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 448.8 on 1167 degrees of freedom
Multiple R-squared:  0.2579,    Adjusted R-squared:  0.2541
F-statistic: 67.59 on 6 and 1167 DF,  p-value: < 2.2e-16
```

Llegando a obtener el modelo:

**bwt = -2247.3015 +12.5887*gestation – 0.2717*age + 12.6891*height + 3.1076*weight – 94.7987*Zpar1 – 237.9990*Zfum + error**

Donde Zpar1 hace referencia cuando la variable cualitativa parity toma el valor 1, cuando es el primer parto de la mujer.

Y Zfum hace referencia cuando la variable cualitativa smoke toma el valor 1, cuando la madre ha fumado durante el embarazo.

El coeficiente de determinación resultante es de 25.79%.

La estimación de las variables ha resultado ser:

-$\boldsymbol{\beta_0}$ **= -2247.3015.** Carece de sentido físico puesto que es la ordenada en el origen del modelo. Dependiendo de los valores que tomen las variables cualitativas y sus correspondientes $\alpha$ este valor se verá afectado. Es significativo puesto que su p-valor es muy pequeño.
-$\boldsymbol{\beta_1}$ **= + 12.5887.** Explica, a igualdad del resto de variables, la relación entre bwt y el regresor gestation. Su valor es significativo debido a su p-valor menor a 2e-16.
-$\boldsymbol{\beta_2}$ **= - 0.2717.** Siendo el resto de los regresores constantes este valor expresa la relación entre bwt y age. Debido a que su p-valor es 0.91113 su valor puede considerarse no significativo.
-$\boldsymbol{\beta_3}$ **= + 12.6891.** Expresa el cambio que sufre bwt al variar height, siendo el resto de las variables constantes. Es significativo debido a su p-valor.
-$\boldsymbol{\beta_4}$ **= + 3.1076.** Este valor muestra como afecta la variable weight a bwt, mientras el resto no varían. No tiene un p-valor de los más pequeños pero lo suficiente como para que sea significativa.
-$\boldsymbol{\alpha_{par1}}$ **= - 94.7987.** Este valor explica cómo los bebés que provienen de un primer parto pesan 94.7987 gramos menos, a igualdad de variables, y respecto de una madre no primeriza. Debido a su p-valor, es significativo su valor.
-$\boldsymbol{\alpha_{fum}}$ **= -237.9990.** Este valor explica como los bebés de madres fumadoras pesan 237.9990 gramos menos, a igualdad de variables, y respecto de una madre no fumadora. Es significativo debido a su p-valor tan pequeño.

Para calcular los intervalos de los parámetros mencionados se realiza en RStudio:

```
> confint(modelo, level=0.95)
                    2.5 %       97.5 %
(Intercept) -3037.3138732 -1457.289048
gestation      10.9701323    14.207179
age            -5.0458232     4.502509
height          8.2637408    17.114492
weight          0.0217786     6.193333
zpar1TRUE    -157.6004580   -31.997042
zfumTRUE     -291.0542589  -184.943770
```

- $\boldsymbol{\beta_0}$**:** Tiene un intervalo de confianza de **[ - 3037.3138732 , - 1457.289048 ]**.
- $\boldsymbol{\beta_1}$: Tiene un intervalo de confianza de **[ 10.9701323 , 14.207179 ]**.
- $\boldsymbol{\beta_2}$**:** Tiene un intervalo de confianza de **[ - 5.0458232 , 4.502509 ]**.
- $\boldsymbol{\beta_3}$: Tiene un intervalo de confianza de **[ 8.2637408 , 17.114492]**.
- $\boldsymbol{\beta_4}$**:** Tiene un intervalo de confianza de **[ 0.0217786 , 6.193333]**.
- $\boldsymbol{\alpha_{par1}}$: Tiene un intervalo de confianza de **[ - 157.6004580 ,- 31.997042]**.
- $\boldsymbol{\alpha_{fum}}$: Tiene un intervalo de confianza de **[ - 291.0542589 , - 184.943770]**.

2. Para analizar la influencia de una madre fumadora durante el embarazo debemos tener en cuenta que el modelo anterior toma como referencia a una no fumadora. A continuación, tenemos que ver si la variable Zfum es significativa. Para ello proponemos el siguiente contraste: Ho: $\alpha_{fum}$ = 0 y H1: $\alpha_{fum}$ =! 0. Volviendo al modelo anterior sabemos que para el estimador de $\alpha_{fum}$

obtenemos un p-valor menor que 2e-16, lo que significa que rechazaríamos la hipótesis inicial. Por lo que esta variable de tipo cualitativa **sí es significativa**. Y por tanto $\boldsymbol{\alpha_{fum}}$ **= -237.9990** se puede aceptar. Este valor indica que el peso de los bebés de madres que han fumado durante el parto es 237.9990 gramos menor que los de las madres que no han fumado, las de referencia en el modelo. Todo esto a igualdad del resto de variables.

3. Para analizar la influencia en el peso de los bebés que provienen de un primer parto volveremos al modelo estimado anterior. Este modelo toma como referencia las madres cuyo parto no ha sido el primero. Teniendo en cuenta esto planteamos el siguiente contraste: Ho: $\alpha_{par1}$ = 0 y H1: $\alpha_{par1}$ =! 0. Revisando los resultados del modelo estimado vemos que para el parámetro $\alpha_{par1}$ se tiene un p-valor de 0.00312, que es un valor lo suficientemente pequeño para rechazar la hipótesis inicial y aceptar el estimador, $\boldsymbol{\alpha_{par1}}$ **= - 94.7987.** Esto significa que los **bebés que nacen como "primer parto" pesan 94.7987 gramos menos que los que nacen de una madre no primeriza**, las de referencia del modelo. Todo esto a igualdad del resto de variables.

4. Para los resultados de este apartado se han utilizado las siguientes ordenes en RStudio:

```
prediccion=data.frame(gestation=284, age=24, height=168, weight=53, zpar1=T, zfum=F)
predict(modelo, prediccion, interval='confidence')
predict(modelo, prediccion, interval='prediction')
```

Para los valores dados por el enunciado de las respectivas variables se obtiene una **predicción del peso de 3523.03 gramos.**

```
> prediccion=data.frame(gestation=284, age=24, height=168, weight=53, zpar1=T, zfum=F)
> predict(modelo, prediccion, interval='confidence')
       fit      lwr      upr
1 3523.03 3461.24 3584.82
> predict(modelo, prediccion, interval='prediction')
       fit      lwr      upr
1 3523.03 2640.362 4405.699
```

Los intervalos de predicción resultantes son los siguientes:

- Para la **media**: **[ 3461.24 , 3584.82 ].**
- Para una **nueva observación**: **[ 2640.362 , 4405.699 ].**

Como es común el intervalo de la media es menor que respecto al de la nueva observación.

5.

1. En el modelo anterior la variable **age** resulta ser **no significativa** por tener un **p-valor** de **0.91113**. Este valor es demasiado grande para cualquier α. Por ello realizamos un nuevo modelo eliminado esta variable de él.

```
modelo1=lm(bwt ~ gestation + height + weight + zpar1 + zfum)
summary(modelo1)


Residuals:
     Min       1Q   Median       3Q      Max
-1629.19  -288.29    -2.81   270.94  1452.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2256.4249   394.1101  -5.725 1.31e-08 ***
gestation      12.5916     0.8242  15.278  < 2e-16 ***
height         12.7001     2.2524   5.638 2.15e-08 ***
weight          3.0853     1.5595   1.978  0.04811 *
zpar1TRUE     -93.5976    30.1338  -3.106  0.00194 **
zfumTRUE     -237.7918    26.9663  -8.818  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 448.6 on 1168 degrees of freedom
Multiple R-squared:  0.2579,    Adjusted R-squared:  0.2547
F-statistic: 81.17 on 5 and 1168 DF,  p-value: < 2.2e-16
```

En este nuevo modelo la variable weight tiene un p-valor en el límite, pero si probamos otro nuevo modelo eliminándola comprobamos que el coeficiente de determinación se ve alterado para peor por poco. Por lo que es correcto seguir trabajando con la variable weight como significativa.

Llegando a obtener el modelo:
**bwt = -2256.4249 + 12.5916*gestation + 12.7001*height + 3.0853*weight – 93.5976*Zpar1 – 237.7918*Zfum + error**

Donde Zpar1 hace referencia cuando la variable cualitativa parity toma el valor 1, cuando es el primer parto de la mujer.
Y Zfum hace referencia cuando la variable cualitativa smoke toma el valor 1, cuando la madre ha fumado durante el embarazo.
Ambas como en el modelo anterior.
El coeficiente de determinación resultante es de 25.79%, igual al del modelo anterior.
La estimación de las variables ha resultado ser:
-$\beta_0$ **= -2256.4249.** Carece de sentido físico puesto que es la ordenada en el origen del modelo. Dependiendo de los valores que tomen las variables cualitativas y sus correspondientes $\alpha$ este valor se verá afectado. Es significativo puesto que su p-valor es muy pequeño.
-$\beta_1$ **= + 12.5916.** Explica, a igualdad del resto de variables, la relación entre bwt y el regresor gestation. Su valor es significativo debido a su p-valor menor a 2e-16.
-$\beta_3$ **= + 12.7001.** Expresa el cambio que sufre bwt al variar height, siendo el resto de las variables constantes. Es significativo debido a su p-valor.
-$\beta_4$ **= + 3.0853.** Este valor muestra cómo afecta la variable weight a bwt, mientras el resto no varían. No tiene un p-valor de los más pequeños pero lo suficiente como para que sea significativa.
-$\alpha_{par1}$ **= - 93.5976.** Este valor explica como los bebés que provienen de un primer parto pesan 93.5976 gramos menos, a igualdad de variables, y respecto de una madre no primeriza. Debido a su p-valor, es significativo su valor.

-$\boldsymbol{\alpha_{fum}}$ **= -237.7918.** Este valor explica como los bebés de madres fumadoras pesan 237.7918 gramos menos, a igualdad de variables, y respecto de una madre no fumadora. Es significativo debido a su p-valor tan pequeño.

Para calcular los intervalos de los parámetros mencionados se realiza en RStudio:

```
> confint(modelo1, level=0.95)
                    2.5 %       97.5 %
(Intercept) -3.029668e+03 -1483.182046
gestation    1.097465e+01    14.208628
height       8.280876e+00    17.119399
weight       2.565733e-02     6.145025
zpar1TRUE   -1.527200e+02   -34.475137
zfumTRUE    -2.906995e+02  -184.884104
```

- $\boldsymbol{\beta_0}$**:** Tiene un intervalo de confianza de **[ - 3029.668 , - 1483.182046 ]**.
- $\boldsymbol{\beta_1}$: Tiene un intervalo de confianza de **[ 10.97465 , 14.208628 ]**.
- $\boldsymbol{\beta_3}$: Tiene un intervalo de confianza de **[ 8.280876 , 17.119399 ]**.
- $\boldsymbol{\beta_4}$**:** Tiene un intervalo de confianza de **[ 0.02565733 , 6.145025 ]**.
- $\boldsymbol{\alpha_{par1}}$: Tiene un intervalo de confianza de **[ - 152.7200 ,- 34.475137 ]**.
- $\boldsymbol{\alpha_{fum}}$: Tiene un intervalo de confianza de **[ - 290.6995 , - 184.884104 ]**.

2. Para analizar la influencia de una madre fumadora durante el embarazo debemos tener en cuenta que el modelo anterior toma como referencia a una no fumadora. A continuación, tenemos que ver si la variable Zfum es significativa. Para ello proponemos el siguiente contraste: Ho: $\alpha_{\text{fum}}$ = 0 y H1: $\alpha_{\text{fum}}$ =! 0. Volviendo al modelo anterior sabemos que para el estimador de $\alpha_{\text{fum}}$ obtenemos un p-valor menor que 2e-16, lo que significa que rechazaríamos la hipótesis inicial. Por lo que esta variable de tipo cualitativa **sí es significativa**. Y por tanto $\boldsymbol{\alpha_{fum}}$ **= -237.7918** se puede aceptar. Este valor indica que el peso de los bebés de madres que han fumado durante el parto es 237.7918 gramos menor que los de las madres que no han fumado, las de referencia en el modelo. Todo esto a igualdad del resto de variables. Respecto al modelo 1 la diferencia es inferior en unidades de la magnitud de miligramos.

3. Para analizar la influencia en el peso de los bebés que provienen de un primer parto volveremos al modelo estimado anterior. Este modelo toma como referencia las madres cuyo parto no ha sido el primero. Teniendo en cuenta esto planteamos el siguiente contraste: Ho: $\alpha_{\text{par1}}$ = 0 y H1: $\alpha_{\text{par1}}$ =! 0. Revisando los resultados del modelo estimado vemos que para el parámetro $\alpha_{\text{par1}}$ se tiene un p-valor de 0.00194, que es un valor lo suficientemente pequeño para rechazar la hipótesis inicial y aceptar el estimador, $\boldsymbol{\alpha_{par1}}$ **= - 93.5976.** Esto significa que los **bebés que nacen como "primer parto" pesan 93.5976 gramos menos que los que nacen de una madre no primeriza**, las de referencia del modelo. Todo esto a igualdad del resto de variables. La diferencia respecto al modelo del primer apartado es de poco más de un gramo.