PEC 5

UOC

Introducción

El siguiente dataset contiene información sobre diversos factores que afectan al rendimiento de los estudiantes en los exámenes de la asignatura "Estadística y Probabilidad".

El fichero para realizar la PEC 5 es *"data_pac5"* y lo encontraréis en formato csv. Esta base de datos contiene información relacionada con las horas estudiadas, porcentaje de clases asistidas y otras variables.

- *Hours_Studied* : Número de horas dedicadas al estudio por semana.
- *Attendance* : Porcentaje de clases asistidas.
- *Parental_Involvement* : Nivel de involucración de los padres en la educación del estudiante (Bajo, Medio, Alto).
- *Access_to_Resources* : Disponibilidad de recursos educativos (Bajo, Medio, Alto).
- *Sleep_Hours* : Número promedio de horas de sueño por noche.
- *Previous_Scores* : Calificaciones de exámenes anteriores.
- *Motivation_Level* : Nivel de motivación del estudiante (Bajo, Medio, Alto)..
- *Tutoring_Sessions* : Número de sesiones de tutoría atendidas por mes.
- *Physical_Activity* : Número promedio de horas de actividad física por semana.
- *Exam_Score* : Calificación final del examen.

Os puede ser útil consultar el siguiente material:

- Módulos teóricos de Regresión lineal simple, múltiple y ANOVA.
- Actividades resueltas del Reto 5 (regresión lineal simple, múltiple y ANOVA).
- Modelos de regresión y análisis multivariante con R.

Hay que entregar la práctica en fichero pdf o html. Se recomienda generar el informe con Rmarkdown que genera automáticamente el html/pdf a entregar. Se puede utilizar el fichero .Rmd, que disponéis en la PEC, como plantilla para resolver los ejercicios.

Esta PEC debe realizarse de forma estrictamente individual, quedando del todo prohibido el uso de herramientas de IA. Cualquier indicio de copia será penalizado con un suspenso (D) por todas las partes implicadas y la posible evaluación negativa de la asignatura de forma íntegra.

```
##   Hours_Studied Attendance Parental_Involvement Access_to_Resources Sleep_Hours
## 1            23         84                  Low                High           7
## 2            19         64                  Low              Medium           8
## 3            24         98               Medium              Medium           7
## 4            29         89                  Low              Medium           8
## 5            19         92               Medium              Medium           6
## 6            19         88               Medium              Medium           8
##   Previous_Scores Motivation_Level Tutoring_Sessions Physical_Activity
## 1              73              Low                 0                 3
```

```
## 2              59              Low                 2                 4
## 3              91           Medium                 2                 4
## 4              98           Medium                 1                 4
## 5              65           Medium                 3                 4
## 6              89           Medium                 3                 3
##   Exam_Score
## 1         67
## 2         61
## 3         74
## 4         71
## 5         70
## 6         71
```

Pregunta 1 (resolver con R). (4 puntos)

Tenemos la siguiente salida por R:

```
lm(formula = Exam_Score ~ Tutoring_Sessions, data = datos_ej1)

Residuals:
    Min      1Q  Median      3Q     Max
-11.972  -2.458   0.000   2.028  32.028

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)              ?    0.07732  859.47   <2e-16 ***
Tutoring_Sessions  0.51411    0.04006       ?   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared:  0.02641,   Adjusted R-squared:  0.02625
```

a) Encuentre el intercepto. (1 punto).

Solución

```
beta1 <- 0.51411
beta0 <- mean(Exam_Score) - beta1 * mean(Tutoring_Sessions)
beta0
```

```
## [1] 66.46772
```

```
summary(lm(formula = Exam_Score ~ Tutoring_Sessions, data = datos_ej1))
```

```
##
## Call:
## lm(formula = Exam_Score ~ Tutoring_Sessions, data = datos_ej1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.991  -2.486   0.009   2.019  32.029
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       66.49648    0.07436  894.31   <2e-16 ***
## Tutoring_Sessions  0.49486    0.03842   12.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.843 on 6605 degrees of freedom
## Multiple R-squared:  0.0245, Adjusted R-squared:  0.02435
## F-statistic: 165.9 on 1 and 6605 DF,  p-value: < 2.2e-16
```

El intercepto es 66.46772 para la salida del enuncioado, sin embargo en el dataset es 66.49648, lo que indica que hay un error en la salida del enunciado, o no son los mismos datasets, ya que el intercepto es el valor de la variable dependiente cuando la variable independiente es 0, en este caso tendriamos en el enunciado que cuando un alumno no da clases, tiene una nota de 66.46772 vs 66.49648 en el dataset, lo que es muy similar, pero no es lo mismo.

b) Encuentre el t valor para la hipótesis nula H0 : $\beta 1 = 1$ y explique la conclusión (nivel significación de 0.05). Considere Z $\sim$ N(0,1). (1.5 puntos).

Solución

```
# Usando la salida del enunciado
se_beta1 <- 0.04006
t <- (beta1 - 1) / se_beta1
t
```

```
## [1] -12.12906
```

```
# Usando el dataset
se_beta1 <- 0.03842
t <- (beta1 - 1) / se_beta1
t
```

```
## [1] -12.6468
```

c) ¿Cuál es el valor de $\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}$? Interprete el resultado. (1.5 puntos).

Solución

$R^2 = 0.02641$ es el coeficiente de determinación, que indica que el 2.641% de la variabilidad de la variable dependiente (Exam_Score) es explicada por la variable independiente (Tutoring_Sessions). Esto significa que el modelo no es muy bueno para predecir la variable dependiente, un buen modelo puede explicar al menos un 70-80% de la variabilidad.

Pregunta 2 (resolver con R). (3 puntos)

En este ejercicio se pide realizar un ANOVA para estudiar si existen diferencias entre las medias de las notas del examen (variable *Exam_Score*) según las diferentes categorías de la variable *Access_to_Resources*.

a) Formula la hipótesis nula y alternativa. (1 punto).

Solución: La hipotesis nula H0 es que las medias de Exam_Score son iguales para todas las categorías de Access_to_Resources. La hipótesis alternativa H1 es que al menos una de las medias es diferente.

b) Realiza un análisis ANOVA y explica las conclusiones del mismo (para este ejercicio se debe tener en cuenta un nivel de significancia del 5%). (1 punto).

```
anova <- aov(Exam_Score ~ Access_to_Resources,datos_ej1)
summary(anova)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## Access_to_Resources    2   2882  1441.0      98 <2e-16 ***
## Residuals           6604  97104    14.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor de p es 0.0000…0002, lo que significa que es menor que el nivel de significancia ($p < 0.05$). Por lo tanto, rechazamos la hipótesis nula y concluimos que existen diferencias significativas entre las medias de las calificaciones del examen según el acceso a recursos.

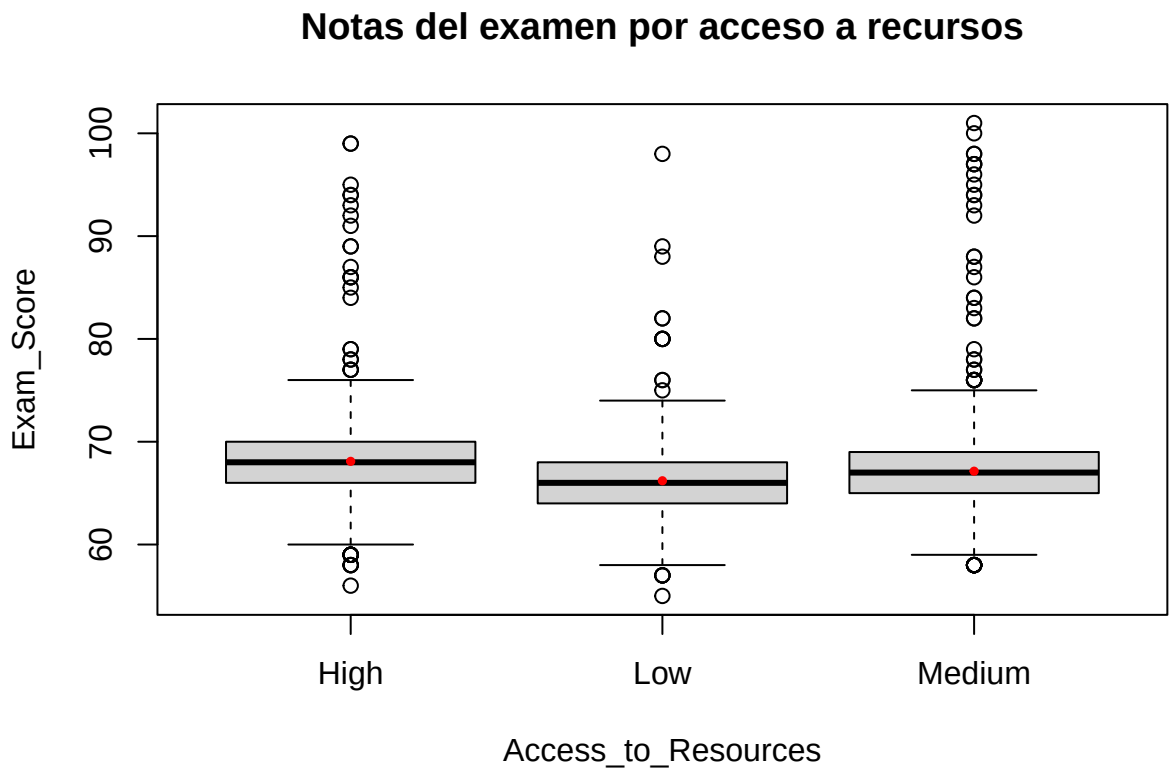
En el análisis, tenemos un total de 6605 observaciones, y los 6604 grados de libertad residuales explican la variabilidad no explicada por el modelo. El valor de F = 98 indica diferencias grandes entre las medias de las calificaciones según el acceso a recursos, en comparación con la variabilidad dentro de los grupos (1441/14.7).

La variable Access_to_Resources tiene 2 grados de libertad, lo que corresponde a 3 grupos (Low, Medium y High).

c) Haga un gráfico boxplot (o diagrama de cajas) de la variable *Exam_Score* para cada grupo de la variable *Access_to_Resources* y comente los resultados comparándolos con el apartado b). (1 punto).

Solución:

```r
boxplot(Exam_Score ~ Access_to_Resources, data = datos_ej1,
        main = "Notas del examen por acceso a recursos",
        xlab = "Access_to_Resources",
        ylab = "Exam_Score",
        col = "lightgray")
media <- tapply(datos_ej1$Exam_Score, datos_ej1$Access_to_Resources, mean, na.rm = TRUE)
points(1:length(media), media, col = "red", pch = 19, cex = 0.5)
```



Segun ANOVA hay diferencias significativas entre las medias de las notas del examen segun el acceso a recursos. En el boxplot se observa un solapamiento de las cajas y bigotes, que el grupo de acceso alto tiene una media y mediana mayor que los otros dos grupos, y con muchos outliers en la parte alta. El grupo de acceso a recursos medium tiene una media y mediana menor que el grupo de acceso alto, pero mayor que el grupo de acceso bajo, y continua teniendo una cantidad importante de outliers en la parte alta. Por ultimo, el grupo Low, tiene pocos outliers y una media y mediana menor que los otros dos grupos, es decir, la media y mediana de las notas del examen es menor en el grupo de acceso bajo, afectando menos para el calculo de la media y mediana los outliers ya que no tiene. Ademas si bien la mediana y media casi coinciden en los 3 grupos, la media es un poco mayor, cuando la media es mayor que la mediana significa que hay outliers que elevan la media, aunque como he dicho es un poco, no es muy significativo. Los bigotes y las cajas siguen la misma tendencia que las medias y medianas, es decir, el grupo de acceso alto tiene una caja y bigotes mas altos, el grupo de acceso medio tiene una caja y bigotes mas bajos que el grupo de acceso alto, pero mas altos que el grupo de acceso bajo, y el grupo de acceso bajo tiene una caja y bigotes mas bajos que los otros dos grupos. Por lo tanto, quizas lo que yo veo es, que el grupo de acceso bajo tiene una media y mediana menor, pero mas estable, mientras que los otros dos grupos tienen medias mayores, pero con mas variabilidad y dispersion en las notas. Eso se podria explicar con que hay alumnos outliers que elevan la media en los grupos de acceso alto y medio, pero no tanto en el grupo de acceso bajo. En resumen, hay

diferencias significativas, los boxplot se solapan, tienen medias/medianas diferentes, low es el grupo menos disperso, pero a nivel practico su efecto no es tan significativo.

Pregunta 3 (resolver con R). (3 puntos).

Se quiere hacer una estimación de las notas (variable *Exam_Score*) mediante un modelo de regresión lineal múltiple con las siguiente variables:

- Hours_Studied
- Sleep_Hours
- Previous_Scores

a) Escribe la ecuación de regresión lineal múltiple para explicar la variable *Exam_Score* utilizando las variables *Hours_Studied*, *Sleep_Hours* y *Previous_Scores*. (1 punto).

Solución:

```
modelo_multiple <- lm(Exam_Score ~ Hours_Studied + Sleep_Hours + Previous_Scores, data = datos_ej1)
summary(modelo_multiple)
```

```
##
## Call:
## lm(formula = Exam_Score ~ Hours_Studied + Sleep_Hours + Previous_Scores,
##     data = datos_ej1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.399 -2.220 -0.174  1.983 33.275
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     58.527785   0.332082  176.25   <2e-16 ***
## Hours_Studied    0.286780   0.007035   40.76   <2e-16 ***
## Sleep_Hours     -0.048516   0.028705   -1.69    0.091 .
## Previous_Scores  0.044230   0.002927   15.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.424 on 6603 degrees of freedom
## Multiple R-squared:  0.2257, Adjusted R-squared:  0.2253
## F-statistic: 641.5 on 3 and 6603 DF,  p-value: < 2.2e-16
```

Exam_Score = _0 + _1 Hours_Studied + _2 Sleep_Hours + _3 Previous_Scores

b) Analiza el modelo estimado en conjunto (significación con un nivel del 5% y coeficiente de determinación). (1 punto).

Solución:

H0 : $\beta_1 = \beta_2 = \beta_3 = 0$ al 5% de significancia H1 : Al menos uno de los β es diferente de 0 al 5% de significancia

El modelo estimado es: Exam_Score = _0 + _1 Hours_Studied + _2 Sleep_Hours + _3 Previous_Scores Pues bien, yo veo un valor de F = 641.5 y un p-value de 2.2e-16, lo que significa que el modelo es en su conjunto significativo, ademas podemos rechazar la H0 Solamente podemos explicar el $R^2 = 0.2257$ (o 22,57%) de la variabilidad de la variable dependiente (Exam_Score) con las tres variables independientes (Hours_Studied, Sleep_Hours y Previous_Scores). Esto significa que el modelo no es bueno para predecir la variable dependiente, un buen modelo puede explicar al menos un 70-80% de la variabilidad. Sin embargo

eso no quiere decir que no tenga relacion estadisticamente significativa, ya que el valor de F es muy alto y el p-value muy bajo, lo que significa que el modelo es significativo, teniendo relacion estadustica significativa, en su conjunto.

Por ultimo _0, _1 (Hours_Studied) y _3 (Previous_Scores) tienen p-values muy bajos, lo que significa que son muy significativos al 5%. En cambio, _2 (Sleep_Hours) tiene un p-value de 0.09, con lo cual no es significativo al 5% pero sí podría serlo al 10%. Por tanto, existe cierta evidencia de que las horas de sueño influyen en la calificación, aunque no podemos confirmarlo con un nivel de significancia del 5%.

c) Estima la calificación del examen (variable *Exam_Score*) de una persona que ha estudiado 17 horas (variable *Hours_Studied*), que ha dormido 5 horas (variable *Sleep_Hours*) y que en el anterior examen saco un 76.5 de nota (variable *Previous_Scores*). (1 punto).

Solución:

```
nueva_persona <- data.frame(Hours_Studied = 17, Sleep_Hours = 5, Previous_Scores = 76.5)
predict(modelo_multiple,  nueva_persona)
```

```
##        1
## 66.54406
```