Primera Revisión
Jueves 12 de mayo

Modelos Lineales - 2022

EJERCICIO 1

En el modelo de regresión lineal múltiple donde intervienen k variables explicativas, las definiciones del R^2 y el $R^2_{ajustado}$ están dadas por las siguiente fórmulas.

$$\begin{aligned} R^2 &= 1 - \frac{SCRes}{SCTot} \\ R^2_{aj} &= 1 - \frac{SCRes/(n-k-1)}{SCTot/(n-1)} \end{aligned}$$

Manipule estas fórmulas para encontrar una expresión que le permita expresar el $R^2_{ajustado}$ en función del R^2.

EJERCICIO 2

Una empresa fabrica un dispositivo electrónico para ser utilizado en un rango de temperatura muy amplio. La empresa sabe que el aumento de temperatura acorta el tiempo de vida del dispositivo, por lo que se realiza un estudio en el que se determina el **tiempo de vida** en función de la **temperatura**. Se encuentran los siguientes datos:

- $n = 9$.
- $\sum_i x_i = 450$.
- $\sum_i y_i = 1691$.
- $\sum_i x_i y_i = 52670$.
- $\sum_i x_i^2 = 28500$.
- $\sum_i y_i^2 = 489857$.

Siendo X la variable "Temperatura" y Y la variable "Tiempo de vida". A partir de estos datos:

1. Plantee el modelo de regresión lineal simple en el que se pretende explicar el tiempo de vida en función de la temperatura.
2. Obtenga el valor de la correlación entre X e Y.
3. Obtenga el valor del R^2.
4. Estime los componentes del vector β.
5. Obtenga la predicción del tiempo de vida para un dispositivo que funciona a 65 grados.

EJERCICIO 3

Una empresa fabricante de autos y camiones está interesada en conocer el rendimiento de sus vehículos (medido en kilómetros por litro) en función de algunas características del motor (el volumen de combostible que desplaza en cada ciclo y si tiene turbo o no). Adicionalmente, le interesa conocer si el rendimiento es igual para autos y camiones.
Con este objetivo, se obtuvo información sobre un conjunto de vehículos y se planteó un modelo de regresión linal múltiple.
A continuación se presenta le *summary* del modelo de *RLM* correspondiente a esta situación.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.8875     0.5547  28.640  <0.0001 ***
volumen_despl -1.5477     0.1627  -9.512  <0.0001 ***
turbo(Si)     -1.3374     0.3607  -3.708 0.000582 ***
tipo(Camion)  -1.2143     0.3720  -3.265 0.002125 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.2 on 44 degrees of freedom
Multiple R-squared:  0.743,     Adjusted R-squared:  0.7255
F-statistic:  42.4 on 3 and 44 DF,  p-value: <0.0001
```

A partir de esta información responda los siguientes puntos.

1. ¿Con cuantas observaciones se contó para estimar el modelo?
2. ¿Qué porcentaje de la variabilidad del rendimiento pudo ser explicada por el modelo?
3. Determine la significación global del modelo justificando sus afirmaciones.
4. Considerando un nivel de significación del 5 %, ¿es significativo el aporte de la variable “*tipo*”?
5. ¿Podría contestar la pregunta de los fabricantes respecto a la comparación entre el rendimiento entre autos y camiones?
6. Asumiendo que la variable “*volumen desp*” indica la cantidad de litros desplazados por el motor del vehículo en cada ciclo de combustión, interprete el valor estimado del parámetro asociado a esta variable.
7. Realice la predicción correspondiente a un automóvil cuyo motor desplaza un volumen de 2 litros y que adicionalmente tiene turbo.

EJERCICIO 4

Sea el modelo $y \sim N\left(X\beta, \sigma^2 I_{n\times n}\right)$. Siendo X una matriz de rango completo k y β y σ^2 parámetros desconocidos. A partir la expresión:

$$\hat{\beta}_j = \frac{y^{'} M_{(j)} x_j}{x_j^{'} M_{(j)} x_j}$$

- Interpretar qué son $x_j^{'} M_{(j)}$ y $y^{'} M_{(j)}$. Con $M_{(j)} = I_{n\times n} - X_j \left(X_j^{'} X_j\right)^{-1} X_j^{'}$, siendo X_j la matriz X desprovista de la columna x_j.
- Indique a qué modelo de RLS corresponde la estimación $\hat{\beta}_j$.

EJERCICIO 5

Un grupo de científicos llevó a cabo un estudio para determinar los factores asociados a un mejor desempeño en una carrera de 10 kilómetros. Para ello, obtuvieron datos correspondientes a 50 participantes *amateur* referentes a su, sexo, edad, iíndice de masa corporal, años de experiencia corriendo y si si entrena en grupo o solo.
A partir de esta información se llevó a cabo un procedimiento de selección de variables y a continuación se presentan los resultados.

```
Backward elimination, alpha-to-remove: ???

Full model: tiempo ~ sexo_F + edad + entrena_grupo + imc + experiencia

              Step    RSS    AIC  R2pred     Cp  Pr(>F)
entrena_grupo    1 159.30 67.937 0.80724 4.0167  0.8977
experiencia      2 161.89 66.745 0.81010 2.7341  0.3963
sexo_F           3 171.35 67.584 0.80859 3.3471  0.0880

Call:
lm(formula = tiempo ~ edad + imc, data = datos)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6416     2.7898   3.456  0.00117 **
edad          0.1747     0.0341   5.124  < 0.001 ***
imc           1.3740     0.0921  14.919  < 0.001 ***
```

A partir de esta información.

- ¿Qué procedimiento se implementó?
- ¿Qué nivel de significación se utilizó como criterio para incluir/eliminar variables?
- Describa lo que sucedió en cada paso del algoritmo.
- Describa el modelo final.