

Práctico 3

Inferencia

Modelos Lineales - 2024

EJERCICIO 1

Sea el modelo de regresión múltiple dado por:

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

Donde el vector Y es de $n \times 1$, la matriz X es de $n \times k + 1$, el vector β es de $k + 1 \times 1$ y el vector ϵ es de $n \times 1$.

El coeficiente de determinación está dado por la expresión:

$$R^2 = \frac{SCReg}{SCTot}$$

Y el estadístico F correspondiente a la significación global del modelo puede ser expresado como:

$$F = \frac{SCReg/k}{SCRes/(n - k - 1)}$$

Demuestre que:

$$F = \frac{R^2}{1 - R^2} \times \frac{n - k - 1}{k}$$

EJERCICIO 2

En el modelo de regresión múltiple descrito en el ejercicio 1, el estimador insesgado de la varianza de los errores es:

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n - k - 1}$$

Sabiendo que:

$$\frac{(n - k - 1) \hat{\sigma}^2}{\sigma^2} \sim \chi_{n-k-1}^2$$

Construya un intervalo de confianza al $(1 - \alpha) \%$ para σ^2 .

EJERCICIO 3

Dentro del paquete *faraway* encontrará un conjunto de datos llamado *cheddar*. En el mismo se encuentran 30 observaciones correspondientes a 30 hormas de queso cheddar y 4 variables. La primera de ellas un *score* otorgado por un conjunto de jueces y las restantes corresponden a concentraciones de ácido acético, ácido sulfhídrico y ácido láctico (estos 3 están expresados en escala logarítmica).

1. Ajuste un modelo donde se explique el sabor a partir de las 3 concentraciones.
2. Lleve a cabo la prueba de significación del modelo con un nivel de significación del 5 % e interprete el resultado.


3. Lleve a cabo las pruebas de significación de cada variable con un nivel de significación del 5 %.
4. Vuelva a ajustar el modelo removiendo la/las variables explicativas que no hayan resultado significativas en el punto anterior.
5. Determine qué porcentaje de la variabilidad del sabor es explicada por las variables de este último modelo.
6. Obtenga un intervalo de confianza para la respuesta media de una horma con valores promedio en las variables de este último modelo.

EJERCICIO 4

Un grupo de investigadores relevaron información en 498 pueblos para determinar qué asociación tienen algunos factores sociales respecto de la prevalencia (qué tan frecuente es una cierta enfermedad) de enfermedades cardíacas. En el set de datos *'cardiacos.txt'* se encuentran las variables *porc_EC*, *porc_fuma* y *porc_bicicleta* que representan el porcentaje de personas en cada pueblo que padecen enfermedades cardíacas, que fuman y que van a su trabajo en bicicleta.

A partir de estos datos se ajustó el siguiente modelo lineal:

$$porc_EC_i = \beta_0 + \beta_1 porc_fuma_i + \beta_2 porc_bici_i + \epsilon_i$$

A continuación se presenta el *summary* posterior al ajuste de este modelo en .

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	14.984	?	186.99	< 0.001
porc_bici	-0.200	0.0013	?	< ?
porc_fuma	?	0.0035	50.39	< 0.001

Residual standard error: 0.654 on ? degrees of freedom

Multiple R-squared: 0.9796

F-statistic: 11895.24 on 2 and ? DF, p-value: < 0.001

A partir de esta salida, se pide:

- Determine si el modelo es significativo en su conjunto.
- Determine cuántos grados de libertad tiene el estimador de la varianza de los errores.
- Complete los valores marcados con ? en la tabla.
- Determine qué variables explicativas son significativas.
- Interprete los coeficientes asociados a las dos variables explicativas.
- Interprete el valor del R^2 .

EJERCICIO 5

La salida que se presenta a continuación fue construída en base al set de datos *teengamb* del paquete *faraway*. El conjunto de datos comprende 47 observaciones correspondientes a adolescentes del Reino Unido y 5 variables. Se ajustó un modelo lineal para tratar de predecir el gasto en apuestas (en libras por año) en función del sexo (0=Hombre, 1=Mujer), su ingreso (en libras por semana), el nivel socioeconómico de los padres y un indicador de fluidez verbal.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.56	17.197	1.312	0.197
sexo	-22.12	8.211	-2.694	0.010
NSE	0.05	0.281	0.186	0.853
ingreso	4.96	1.025	4.839	<0.001
f_verbal	-2.96	2.172	-1.362	0.180

SCT = 45689.49

SCRes = 21623.8

1. A partir de los datos provistos, calcule el estadístico F , sus grados de libertad y lleve a cabo la prueba de significación del modelo explicitando su conclusión.
2. Determine cuál/cuales de las variables explicativas es/son significativa/s considerando un nivel de significación del 5 %.
3. Empleando la información provista calcule un intervalo de confianza al 90 % para el coeficiente asociado a la variable sexo.
4. Interprete el intervalo calculado en el punto anterior.
5. Realice la predicción del gasto anual en apuestas para un adolescente de sexo masculino con un NSE de 60, un ingreso de 3 libras a la semana y un *score* de fluidez verbal de 8.
6. Para la predicción realizada en el punto anterior, calcule el intervalo de confianza para la respuesta media.

EJERCICIO 6

En un estudio sobre la incidencia que puede tener la comprensión lectora y la capacidad intelectual sobre el rendimiento en lenguaje, se obtuvieron datos sobre 100 estudiantes tomados al azar de un curso de educación inicial.

El modelo propuesto es:

$$Lenguaje_i = \beta_0 + \beta_1 Comp_lect_i + \beta_2 Cap_int_i + \epsilon_i$$

Luego de ajustar el modelo se obtuvieron las siguientes estimaciones:

$\hat{\beta}_0$	-0,754
$\hat{\beta}_1$	0,474
$\hat{\beta}_2$	0,584
$\hat{\sigma}^2$	0,949

Adicionalmente, se sabe que:

$$(X'X)^{-1} = \begin{pmatrix} 0,00106 & -0,00087 \\ -0,00087 & 0,00096 \end{pmatrix}$$

A partir de esta información:

1. Realice la predicción para un niño con una comprensión lectora de 7 y una capacidad intelectual de 9.
2. Calcule el desvío standard para la predicción anterior (S_{FO}).
3. A partir de los datos anteriores, construya el intervalo de predicción al 95 %. Interpretelo.