EX.20250903

Ejercicios elaborados con fines educativos, inspirados en los contenidos evaluados en el exámen del 03/09/2025 de Modelado Estadístico de Datos de la UNED (convocatoria Sep-2025).

Este documento no es una copia ni una transcripción del examen oficial, sino una redacción propia de ejercicios conceptualmente equivalentes.

Se puede resolver con el apoyo de cualquier tipo de material escrito y de una calculadora programable.

EX.20250903.1

Enunciado

(1 punto) En el contexto del modelo de regresión lineal, sean H la matriz *hat*, $M = I - H$ y C la matriz asociada a la ortogonalización. Determina si es correcta la siguiente afirmación:

$$CM = I.$$

a) Verdadero.

b) Falso.

Razona adecuadamente tu respuesta.

Solución

Respuesta: b) Falso.

Recordemos que en el modelo lineal

$$\hat{y} = Hy, \qquad H = X(X^TX)^{-1}X^T, \qquad M = I - H,$$

y que la matriz de ortogonalización se define como

$$C = (X^TX)^{-1}X^T.$$

La afirmación a comprobar es $CM = I$. Operamos:

$$CM = C(I - H) = C - CH.$$

Ahora bien,

$$CH = (X^TX)^{-1}X^T\,X(X^TX)^{-1}X^T = (X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = (X^TX)^{-1}X^T = C.$$

Sustituyendo:

$$CM = C - CH = C - C = 0.$$

Por tanto, CM no es la identidad, sino la matriz nula, y la proposición es **falsa**.

EX.20250903.2

Enunciado

(2 puntos) Con los resultados muestrales $a_1 = 14$ de $n_1 = 50$ observaciones en el primer grupo y $a_2 = 14$ de $n_2 = 50$ observaciones en el segundo grupo, plantea y lleva a cabo un contraste bilateral para evaluar si existe diferencia entre las proporciones poblacionales de ambos grupos:

$$H_0 : \ \pi_1 - \pi_2 = 0$$

$$H_1 : \ \pi_1 - \pi_2 \neq 0.$$

Solución

A partir de los datos muestrales se calculan las proporciones observadas en cada grupo:

$$\hat{\pi}_1 = \frac{14}{50} = 0.28, \qquad \hat{\pi}_2 = \frac{14}{50} = 0.28.$$

Se quiere contrastar

$$H_0 : \ \pi_1 - \pi_2 = 0 \qquad \text{frente a} \qquad H_1 : \ \pi_1 - \pi_2 \neq 0.$$

Bajo H_0 se utiliza la aproximación normal del contraste de dos proporciones con proporción combinada:

$$\hat{\pi} = \frac{a_1 + a_2}{n_1 + n_2} = \frac{28}{100} = 0.28.$$

El error estándar bajo H_0 es:

$$EE = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.28 \cdot 0.72 \left(\frac{1}{50} + \frac{1}{50}\right)} = \sqrt{0.28 \cdot 0.72 \cdot 0.04}.$$

El estadístico de contraste queda:

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{EE} = \frac{0.28 - 0.28}{EE} = 0.$$

Para un contraste bilateral, el p-valor es:

$$p = 2(1 - \Phi(|Z|)) = 2(1 - \Phi(0)) = 1.$$

Como $p = 1 > \alpha$ (por ejemplo, $\alpha = 0.05$), no se rechaza H_0.
Por tanto, con estos datos no hay evidencia estadística de diferencia entre las proporciones poblacionales de ambos grupos.

EX.20250903.3

Enunciado

(2 puntos) Se dispone de un conjunto de datos correspondiente a 26 deportistas, con el fin de estudiar si el peso corporal en kilogramos (*rta*) puede explicarse a partir de la medida de la cintura en centímetros (*exp1*), el número de kilómetros de entrenamiento realizados (*exp2*) y la modalidad de entrenamiento seguida (*exp3* = 1: *Body building*, *exp3* = 2: *Fitness*).

A continuación se muestra el código en R utilizado para el análisis, junto con los resultados obtenidos en cada una de las etapas del procedimiento. Se pide describir el proceso llevado a cabo y realizar una interpretación detallada del modelo final y de los contrastes realizados.

```
datos <- read.table("c_ccd.txt", header = TRUE)
str(datos)
datos$exp3_bb <- as.integer(datos$exp3 == 1)
modelo_inicial <- lm(
  rta ~ exp1 + exp2 + exp3_bb + exp1:exp3_bb,
  data = datos
)
modelo_final <- step(modelo_inicial, direction = "both", trace = TRUE)
summary(modelo_final)
shapiro.test(residuals(modelo_final))
```

Solución

La relación entre el peso en kg (*rta*) y las variables explicativas se ha estudiado mediante un **modelo de regresión lineal**.
El tipo de entrenamiento ($exp3$), al ser una variable cualitativa con dos categorías, se ha incorporado al modelo mediante una **variable dummy** definida como

$$exp3d1 = \mathbf{1}(exp3 = 1),$$

donde $exp3 = 1$ corresponde a *Body building* y la categoría de referencia es *Fitness* ($exp3 = 2$).

Se parte de un modelo inicial que incluye la cintura ($exp1$), los km de entrenamiento ($exp2$), el tipo de entrenamiento ($exp3d1$) y la interacción entre $exp1$ y $exp3d1$.
El procedimiento de selección automática por **AIC** elimina primero la interacción y posteriormente la variable $exp2$, ya que su inclusión no mejora el ajuste del modelo. De este modo, el modelo final queda dado por

$$rta = \beta_0 + \beta_1\, exp1 + \beta_2\, exp3d1 + \varepsilon.$$

En el modelo seleccionado, todos los coeficientes resultan estadísticamente significativos. El coeficiente asociado a la cintura es positivo ($\hat{\beta}_1 \approx 0.60$), lo que

indica que, manteniendo constante el tipo de entrenamiento, **un aumento de 1 cm en la cintura implica un incremento medio de aproximadamente 0.6 kg en el peso**.
Por otra parte, el coeficiente de la variable dummy del entrenamiento ($\hat{\beta}_2 \approx 4.16$) muestra que, a igualdad de cintura, los deportistas que practican *Body building* pesan de media **unos 4 kg más** que los que practican *Fitness*. El intercepto es significativo, aunque su interpretación directa no es relevante desde el punto de vista práctico.

La calidad del ajuste es muy elevada, con un coeficiente de determinación

$$R^2 \approx 0.95,$$

lo que indica que el modelo explica la mayor parte de la variabilidad observada en el peso. El contraste global del modelo también es significativo, por lo que se concluye que el conjunto de variables explicativas incluidas resulta relevante.

Finalmente, el contraste de **Shapiro–Wilk** aplicado a los residuos no rechaza la hipótesis de normalidad (p-valor alto), lo que respalda la validez de los supuestos del modelo lineal. En conjunto, el modelo final es parsimonioso, presenta un excelente ajuste y permite concluir que el peso depende fundamentalmente de la cintura y del tipo de entrenamiento.

EX.20250903.4

Enunciado

(2 puntos) Se ha ajustado un modelo de regresión logística para estudiar la probabilidad de supervivencia de los pasajeros del Titanic en función de distintas variables explicativas. A continuación se muestra el código empleado en R y el resultado obtenido tras el ajuste del modelo.

Se pide interpretar el resultado proporcionado, comentando el efecto de las variables incluidas, su significación estadística y la información global del modelo.

```
library(titanic)
datos <- titanic_train
datos <- as.data.frame(datos)
modelo_log <- glm(
  formula = Survived ~ Pclass + Sex + Age + SibSp,
  family  = binomial(link = "logit"),
  data    = datos
)
summary(modelo_log)
```

```
Llamada:
glm(formula = Survived ~ Pclass + Sex + Age + SibSp,
    family = binomial, data = df)
```

Residuales de la desviación:

```
     Min        1Q    Median        3Q       Max
-2.7714  -0.6445  -0.3836   0.6276   2.4585

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.600846   0.543441  10.306  < 2e-16 ***
Pclass      -1.317398   0.140900  -9.350  < 2e-16 ***
Sexmale     -2.623483   0.214524 -12.229  < 2e-16 ***
Age         -0.044385   0.008155  -5.442 5.26e-08 ***
SibSp       -0.376119   0.121080  -3.106 0.00189 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Otra información:

```
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 636.72  on 709  degrees of freedom

(177 observations deleted due to missingness)
AIC: 646.72
Number of Fisher Scoring iterations: 5
```

Solución

Se ha estimado un modelo de regresión logística binaria con el objetivo de analizar cómo distintas características de los pasajeros influyen en la probabilidad de supervivencia ($Survived = 1$). Dado el tipo de modelo empleado, los parámetros estimados describen cambios en los log-odds de supervivencia asociados a cada variable explicativa, manteniendo constantes las demás.

El término independiente recoge el valor del logit de la probabilidad de supervivencia para el perfil de referencia: mujer, en la categoría base de la clase del pasajero, edad nula y sin hermanos ni cónyuge a bordo. Aunque esta situación carece de un significado realista, el intercepto es un componente necesario del modelo y permite obtener las probabilidades predichas para perfiles concretos.

La variable correspondiente a la clase del pasajero presenta un coeficiente negativo y altamente significativo. Esto implica que, al pasar de clases más altas a clases más bajas, los log-odds de supervivencia disminuyen de forma acusada. En consecuencia, pertenecer a una clase inferior se asocia con una probabilidad considerablemente menor de sobrevivir.

El coeficiente asociado al sexo indica que ser hombre reduce de manera muy notable los log-odds de supervivencia respecto a la categoría de referencia, que es la mujer. Este efecto es estadísticamente muy significativo y confirma que, a

igualdad del resto de variables, los hombres tenían una probabilidad de supervivencia sensiblemente inferior.

La edad también muestra un efecto negativo y significativo. Cada incremento de un año en la edad del pasajero implica una reducción en los log-odds de supervivencia, lo que se traduce en una menor probabilidad de sobrevivir conforme aumenta la edad, manteniendo constantes las demás características.

Asimismo, el número de hermanos o cónyuge a bordo presenta un coeficiente negativo y significativo. Esto sugiere que viajar acompañado por más familiares cercanos reduce la probabilidad de supervivencia, una vez controlados los efectos del sexo, la clase y la edad.

Desde el punto de vista del ajuste global, la inclusión de las variables explicativas produce una reducción importante de la deviance respecto al modelo nulo, lo que indica que el modelo explica una parte sustancial de la variabilidad observada en la respuesta. El valor del AIC permite además comparar este modelo con otros posibles, y el proceso de estimación converge correctamente. Cabe señalar que parte de las observaciones han sido excluidas debido a la presencia de valores perdidos.

En conjunto, los resultados muestran que todas las variables consideradas influyen de forma significativa en la supervivencia. Las mayores probabilidades de sobrevivir se asocian con ser mujer, pertenecer a clases más altas, tener menor edad y viajar con un número reducido de familiares cercanos.

EX.20250903.5

Enunciado

¿En qué consisten las variables *dummy* y para qué se utilizan en el análisis estadístico y en los modelos de regresión?

Solución

Las variables dummy son variables artificiales de tipo binario que toman valores 0 o 1 y se utilizan para codificar variables cualitativas o categóricas. Cada dummy indica la pertenencia o no de una observación a una determinada categoría.

Su utilidad principal es permitir la inclusión de variables categóricas en modelos estadísticos como la regresión lineal o la regresión logística, que requieren variables numéricas. Además, facilitan la interpretación de los coeficientes del modelo, ya que estos miden el efecto de cada categoría en la variable respuesta en comparación con una categoría de referencia.

EX.20250903.6

Enunciado

Sea θ un parámetro que representa una proporción, con $\theta \in (0, 1)$.
Desde el punto de vista de la inferencia bayesiana, ¿qué familias de distribuciones se utilizan habitualmente como distribuciones *a priori* para modelar la incertidumbre inicial sobre θ?

Solución

Cuando se desea modelizar una proporción o probabilidad $\theta \in (0, 1)$ dentro de un enfoque bayesiano, es habitual emplear distribuciones definidas en el intervalo unitario. Las más utilizadas pertenecen a la familia Beta, debido a su flexibilidad y a su conjugación con la distribución binomial.

Entre las elecciones más habituales destacan:

- Una distribución Beta impropia, que corresponde a asumir ausencia total de información previa sobre θ.
- La prior de Jeffreys, que es invariante frente a reparametrizaciones y asigna mayor peso a los valores cercanos a 0 y 1.
- La distribución uniforme en $(0, 1)$, que representa una creencia previa neutral al otorgar la misma probabilidad a todos los valores posibles de θ.

Estas distribuciones permiten incorporar distintos grados de información previa en el análisis bayesiano de proporciones.