1	2	3	Calificación

Introducción a la Estadística y Ciencia de Datos - Segundo cuatrimestre
Segundo examen parcial - 23/11/2023

Nombre y Apellido:
Cantidad Total de Hojas:
Tildar el acta de TP y Final de la materia en la que quiere figurar:

◯Estadística (MATE820015)
◯Introducción a la Estadística y Ciencia de Datos (LCDA210008)
◯Tópicos de Estadística (DOC8800931)

Por favor, numerar todas las hojas y colocar el nombre en ellas. Se aprueba con al menos 60 puntos, al menos 15 de ellos deben provenir del ejercico 1.

- Justificar todas las respuestas -

---

1. (42 puntos) Sea $X_1, X_2, ..., X_n$ una muestra aleatoria con función de densidad dada por

$$f(x,\theta) = 2\theta x e^{-\theta x^2} \; I_{(0,\infty)}(x) \quad \theta > 0 \tag{1}$$

a) (8 puntos) Deducir el test más potente de nivel α para $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, con $\theta_1 < \theta_0$.

b) (8 puntos) A partir de a) probar que la regla de decisión del test UMP de nivel α para testear

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta < \theta_0$$

se puede reescribir de la siguiente forma:

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & si \quad 2\theta_0 \sum_{i=1}^n X_i^2 \geq K \\ 0 & si \quad 2\theta_0 \sum_{i=1}^n X_i^2 < K \end{cases}$$

donde K es una constante que depende del nivel α del test. Asumiendo que $X_i^2 \sim \Gamma(1,\theta)$ y utilizando las propiedades de la distribución Gamma, hallar el valor de K.

c) (7 puntos) Hallar la función de potencia $\pi(\theta)$ y estudiar su monotonía como función de θ.

d) (7 puntos) Extender el test obtenido para chequear las hipótesis $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$ estudiando la función de potencia. Mostrar que tiene nivel α. ¿Es UMP para estas hipótesis?

e) (7 puntos) El tiempo de descarga de cierto río históricamente era una v.a. con densidad dada por (1) con $\theta = 1$. Después de una intervención en el cauce se sospecha que el parámetro θ ha cambiado a un valor menor. A fin de poner a prueba esta hipótesis,

se toman 10 mediciones del tiempo de descarga garantizándose que los registros resulten independientes, observándose $\sum_{i=1}^{10} x_i^2 = 31.1972$. ¿Son los datos significativos al 5 %? Plantee las hipótesis, el test que propone y establezca claramente su decisión.

f) (5 puntos) Halle el p-valor. ¿Serían significativos los datos al 1 %? ¿Y al 10 %? Justificar.

2. (43 puntos) Se analizó la información sobre 27 vinos borgoña añejos seleccionados vendidos en una serie de subastas en Londres en 1990-1991 [1]. Algunas de las variables medidas se detallan a continuación.

Variable	Descripción
`Price`	logaritmo del precio medio de mercado, relativo al precio de la cosecha de 1961.
`AGST`	temperatura promedio de la temporada de crecimiento (en grados Celsius).
`HarvestRain`	lluvia en la temporada de cosecha (en mm).

Se desea predecir `Price` usando las variables del dataset. Asumiendo normalidad y homogeneidad de varianzas, responda a las siguientes preguntas.

a) (5 puntos) A partir de los datos, se estimaron utilizando mínimos cuadrados los coeficientes del siguiente modelo, que llamaremos `mod1`,

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

donde indicamos con Y a la variable `Price`, $X_1 =$ `AGST`.

Se reporta abajo la salida parcial del comando `lm` tras ajustar el modelo con los datos. Escriba modelo que resulta ajustado e informe una estimación de σ y sus correspondientes grados de libertad.

```
Call: lm(formula = Price ~ AGST)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.5469     2.3641  -1.500 0.146052
AGST          0.6426     0.1434   4.483 0.000143
---

Residual standard error: 0.4819 on ???? degrees of freedom
Multiple R-squared:  0.4456,Adjusted R-squared:  0.4234
F-statistic: 20.09 on 1 and 25 DF,  p-value: 0.0001425
```

[1] Este dataset fue compilado y analizado por Ashenfelter, Ashmore, y Lalonde (1995)

b) (6 puntos) Testee la hipótesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Indique si debe agregar algún supuesto (especificándolo), qué estadístico utiliza (describiendo bien sus componentes) y dé su distribución bajo H_0. A partir de la salida dada, indique el valor observado del estadístico y el p-valor. Escriba la conclusión a nivel 0.05.

c) (3 puntos) ¿Cree que es necesario incluir la intercept? jutifique.

d) (3 puntos) Describa **brevemente**, según el modelo recién ajustado, si la mayor temperatura promedio en la temporada de crecimiento de las uvas redunda en un mayor precio de venta, o si la afirmación correcta es al revés.

e) (3 puntos) Se agrega ahora la variable `HarvestRain` al modelo `(mod 1)` del ítem $2a$. Escriba el modelo propuesto de forma análoga a como está escrito el `(mod 1)` del ítem $2a$, definiendo las variables apropiadas que incluye. ¿Cuántos coeficientes involucra el nuevo modelo? Referiremos a este modelo como `(mod 2)`. Se reporta abajo la salida parcial del comando `lm`:

```
Call: lm(formula = Price ~ AGST + HarvestRain)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.7210031  1.8290026  -1.488 0.149853
AGST         0.6298525  0.1103314   5.709    7e-06 ***
HarvestRain -0.0042511  0.0009954  -4.271 0.000265 ***
---

Residual standard error: 0.3707 on 24 degrees of freedom
Multiple R-squared:  0.685,Adjusted R-squared:  0.6587
F-statistic: 26.09 on 2 and 24 DF,  p-value: 9.548e-07
```

f) (7 puntos) A partir de la salida del ítem anterior, realice un test de nivel 0.05 que le permita decidir si incorporar o no la variable `HaervestRain` al `(mod 1)`. Escriba las hipótesis que testea. Informe el valor del estadístico observado, el $p-$valor correspondiente a este test y su conclusión. Escriba dicha conclusión en los términos del problema (o sea, más allá de escribir "rechazo H_0 o no la rechazo").

g) (2 puntos) Escriba el valor predicho de `Price` de acuerdo al ajuste realizado con `(mod 2)` para un vino con `AGST`=16 y`HarvestRain` = 200. ¿Cuánto da la predicción?

h) (7 puntos) Calcule un intervalo de predicción de nivel 0.95 para `Price` para un vino con `AGST`=16 y`HarvestRain` = 200.

Sugerencia: Elija de los siguientes productos el/los que considere más adecuado de acuerdo al propósito planteado y la salida reportada. Se denota $\mathbb{X}$ a la matriz de diseño.

$\boldsymbol{x}_o = (1, 0, 0)'$ $\quad \boldsymbol{x}_o'(\mathbb{X}'\mathbb{X})^{-1}\boldsymbol{x}_o = 24.34192$

$\boldsymbol{x}_o = (0,1,0)'$ $\quad \boldsymbol{x}_o'(\mathbb{X}'\mathbb{X})^{-1}\boldsymbol{x}_o = 0.08857778$

$\boldsymbol{x}_o = (0,0,1)'$ $\quad \boldsymbol{x}_o'(\mathbb{X}'\mathbb{X})^{-1}\boldsymbol{x}_o = 0.000007$

$\boldsymbol{x}_o = (1,200,16)'$ $\quad \boldsymbol{x}_o'(\mathbb{X}'\mathbb{X})^{-1}\boldsymbol{x}_o = 2982.469$

$\boldsymbol{x}_o = (1,16,200)'$ $\quad \boldsymbol{x}_o'(\mathbb{X}'\mathbb{X})^{-1}\boldsymbol{x}_o = 0.07807224$

$\boldsymbol{x}_o = (16,200,1)'$ $\quad \boldsymbol{x}_o'(\mathbb{X}'\mathbb{X})^{-1}\boldsymbol{x}_o = 413.3303$

i) (7 puntos) Basándose en el ajuste que se hizo del `(mod 2)`, realice un test de nivel simultáneo 0.05 que le permita decidir si las variables `AGST` y `HarvestRain` son significativas en el este modelo. Escriba las hipótesis que testea, el estadístico que utiliza (describiendo bien sus componentes) y su distribución bajo H_0. Informe el valor del estadístico observado, el $p-$valor correspondiente a este test y su conclusión detalladamente.

3. **Teórico** (15 puntos) Supongamos que los vectores $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^t \in \mathbb{R}^p, 1 \leq i \leq n$, son determinísticos y que para cierto $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t$ desconocido, las respuestas y_i cumplen:

$$\begin{aligned} y_i &= \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n \\ \mathbb{E}(\varepsilon_i) &= 0 \\ \mathbb{V}(\varepsilon_i) &= \sigma^2 \\ \mathbb{C}ov(\varepsilon_i, \varepsilon_j) &= 0 \quad i \neq j \end{aligned}$$

Asumamos que la matriz de diseño $\mathbb{X} \in \mathbb{R}^{n \times p}$ con filas $\mathbf{x_i^t}$ tiene rango completo p.

Sea $\widehat{\boldsymbol{\theta}}$ el estimador de mínimos cuadrados de $\boldsymbol{\theta}$.

a) (3 puntos) Describa el problema usando notación matricial definiendo los elementos que sean necesarios. Escriba el estimador de mínimos cuadrados $\widehat{\boldsymbol{\theta}}$ en términos de la matriz $\mathbb{X}$.

b) (6 puntos) Probar que $\widehat{\boldsymbol{\theta}}$ es un estimador insesgado del vector de parámetros $\boldsymbol{\theta}$.

c) (6 puntos) Probar que la matriz de covarianza de $\widehat{\boldsymbol{\theta}}$ satisface $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}} = \sigma^2(\mathbb{X}^t\mathbb{X})^{-1}$.