Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Segundo Semestre 2024

Curso : **Probabilidad y Estadística**
Sigla : **EYP1113**
Profesores : **Ricardo Aravena C., Cristian Capetillo C., Ingrid Guevara R., Bladimir Morales T., Ricardo Olea O. y Daniel Saavedra M.**

PAUTA INTERROGACIÓN 4

Pregunta 1

La duración del desempleo se refiere al tiempo que una persona permanece sin empleo antes de encontrar un nuevo trabajo. Esta variable es crucial para evaluar la salud del mercado laboral y la efectividad de las políticas de empleo. Los tiempos de desempleo pueden variar considerablemente entre individuos: mientras que algunos pueden encontrar trabajo rápidamente, otros pueden permanecer desempleados por periodos prolongados, generando así una distribución asimétrica. Dadas estas características, se propone un modelo de probabilidad para el tiempo que una persona permanece desempleada, cuya función de densidad está dada por:

$$f(x) = \frac{\beta\,\nu^{\beta\,k}}{\Gamma(k)} \cdot x^{\beta\,k-1} \cdot e^{-(\nu\,x)^{\beta}},$$

con $x \geq 0$, $k > 0$, $\nu > 0$ y $\beta > 0$.

Una propiedad de este modelo es que $X^{\beta} \sim$ Gamma $\left(k, \nu^{\beta}\right)$. A partir de una muestra aleatoria (independiente) de n tiempos de desempleo, obtenga el estimador de momentos para ν, suponiendo que β y k son parámetros conocidos. Calcule su error cuadrático medio aproximado de 1er orden y determine si el estimador es consistente o no.

Solución

Para obtener el estimador de momentos de ν, se puede resolver la ecuación $\mathsf{E}(X) = \bar{x}$ en función de ν, donde $\bar{x}$ es el promedio de la muestra aleatoria de n tiempos de desempleo. Sin embargo, dado que el parámetro β se supone conocido, usamos la propiedad dada del modelo para una estimación de momentos para ν resolviendo la ecuación

$$\mathsf{E}(X^{\beta}) = \overline{X^{\beta}} = \frac{1}{n}\sum_{i=1}^{n} X_i^{\beta}.$$

Como $X^{\beta} \sim$ Gamma(k, ν^{β}), por formulario se tiene que $\mathsf{E}(X^{\beta}) = \dfrac{k}{\nu^{\beta}}$.

Así,

$$\frac{k}{\nu^{\beta}} = \overline{X^{\beta}} \to \hat{\nu} = \left(\frac{k}{\overline{X^{\beta}}}\right)^{1/\beta}.$$

Calcular el error cuadrático medio aproximado de 1er orden corresponde a utilizar el método Delta para obtener el error cuadrático medio.

De formulario

$$Z = g(W) \approx g(\mu_W) + (W - \mu_W)g'(\mu_W).$$

Aplicando operadores $\mathsf{E}(\cdot)$ y $\mathsf{Var}(\cdot)$ se tiene:

$$\mathsf{E}(Z) = \mathsf{E}\big[g(\mu_W) + (W - \mu_W)g'(\mu_W)\big] = g(\mu_W) + 0 = g(\mu_W)$$

y

$$\mathsf{Var}(Z) = \left[g(\mu_W) + (W - \mu_W)g'(\mu_W)\right] = g'(\mu_W)^2\,\mathsf{Var}(W).$$

Si consideramos $W = \overline{X^\beta}$, entonces $Z = g(W) = (k/W)^{1/\beta}$, con

$$\begin{aligned}
\mu_W = \mathsf{E}(W) &= \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(X_i^\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{k}{\nu^\beta} = \frac{k}{\nu^\beta} \\
\mathsf{Var}(W) = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^\beta\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{Var}(X_i^\beta) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{k}{v^{2\beta}} = \frac{1}{n^2}\frac{kn}{\nu^{2\beta}} = \frac{k}{n\nu^{2\beta}}, \\
g'(W) &= \frac{d}{dW}\left(k^{1/\beta}W^{-1/\beta}\right) = -\frac{k^{1/\beta}}{\beta}W^{-(1+1/\beta)}.
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
\mathsf{E}(Z) = g(\mu_W) &= \left(\frac{k}{\mu_W}\right)^{1/\beta} = \left(\frac{k}{k/\nu^\beta}\right)^{1/\beta} = \nu, \\
\mathsf{Var}(Z) = g'(\mu_W)^2\,\mathsf{Var}(W) &= \left(-\frac{k^{1/\beta}}{\beta}\mu_W^{-(1+1/\beta)}\right)^2\frac{k}{n\nu^{2\beta}} = \frac{1}{n}\frac{k^{2/\beta+1}}{\beta^2\nu^{2\beta}}\left(\frac{k}{\nu^\beta}\right)^{-2-2/\beta} = \frac{1}{n}\frac{\nu^2}{k\beta^2}.
\end{aligned}$$

Así, como $\mathsf{E}(\hat{\nu}) = \mathsf{E}(Z) = \nu$, el estimador es aproximadamente insesgado, y el Error Cuadrático Medio aproximado de 1er orden está dado por

$$\mathsf{ECM}(\hat{\nu}) = \mathsf{Sesgo}^2 + \mathsf{Var}(\hat{\nu}) = \mathsf{Var}(\hat{\nu}) = \frac{1}{n}\frac{\nu^2}{k\beta^2}.$$

Para determinar si el estimador de ν es consistente, se debe verificar si el límite del Error Cuadrático Medio para $n \to \infty$ es igual cero.

$$\lim_{n\to\infty}\mathsf{ECM}(\hat{\nu}) = \lim_{n\to\infty}\frac{1}{n}\frac{\nu^2}{k\beta^2} = \frac{\nu^2}{k\beta^2}\lim_{n\to\infty}\frac{1}{n} = 0.$$

Así, el estimador de momentos, $\hat{\nu}$, es consistente.

Asignación de Puntaje:

Alternativa 1:
Logro 1 [1 Ptos]: por uso de propiedad dada para plantear ecuación del estimador de momentos: $\mathrm{E}(X^\beta) = \overline{X^\beta}$.

- Si estudiante plantea y desarrolla $\mathrm{E}(X) = \overline{X}$, ir a Alternativa 2 de solución.

Logro 2 [1 Ptos]: por obtener que estimador de momentos es $\hat{\nu} = \left(\frac{k}{\overline{X^\beta}}\right)^{1/\beta}$.

Este resultado implica:

- **[0.5 Ptos]** por calcular $\mathrm{E}(X^\beta) = k/\nu^\beta$ (basta con citar formulario)
- **[0.5 Ptos]** por obtener el estimador tras resolver la ecuación de momentos.

Logro 3 [2 Ptos]: por plantear el método delta para obtener resultados aproximados de primer orden.

- **[0.5 Ptos]** por mencionar uso de método delta por $Z = g(w)$.
- **[0.5 Ptos]** por mencionar o plantear que se debe calcular $\mathrm{E}(Z) = g(\mu_W)$.
- **[1.0 Ptos]** por mencionar o plantear que se debe calcular $\mathrm{Var}(Z) = g'(\mu_W)^2\,\mathrm{Var}(W)$.

Logro 4: **[1 Ptos]** por obtener correctamente $\mathrm{E}(Z)$ **[0.4 Ptos]**, $\mathrm{Var}(Z)$ **[0.4 Ptos]**, y luego $\mathrm{ECM}(\hat{v})$ **[0.2 Ptos]**.

- No se asigna puntaje por mencionar que $\hat{\nu}$ es insesgado.

Logro 5: **[1 Ptos]** por verificar que $\hat{\nu}$ es un estimador consistente.

- **[0.5 Ptos]** por indicar cómo verificar consistencia de $\hat{\nu}$. Asignar este puntaje incluso si estudiante no tiene una expresión para $\hat{\nu}$.
- **[0.5 Ptos]** por verificar que $\lim_{n\to\infty} \mathrm{ECM}(\hat{\nu}) = 0$ y concluir correctamente.

Alternativa 2: estudiante decide calcular valor esperado por definición.

Logro 1: **[1 Ptos]** por plantear ecuación del estimador de momentos: $\mathrm{E}(X) = \overline{X}$ **[0.5 Ptos]** y por plantear la integral correcta para obtener $\mathrm{E}(X)$ **[0.5 Ptos]**.

Logro 2: **[1 Ptos]** por obtener que estimador de momentos es $\hat{\nu} = \dfrac{\Gamma(k+1/\beta)}{\overline{X}\,\Gamma(k)}$.

- **[0.5 Ptos]** por terminar de calcular correctamente $\mathrm{E}(X)$ para obtener que $\mathrm{E}(X) = \dfrac{\Gamma(k+1/\beta)}{\nu\,\Gamma(k)}$.
- **[0.5 Ptos]** por obtener $\hat{\nu}$ tras resolver la ecuación del estimador de momentos.

Logro 3 [2 Ptos]: equivalente a Alternativa 1.

Logro 4 [1 Ptos]: equivalente a Alternativa 1.

Logro 5 [1 Ptos]: equivalente a Alternativa 1.

+ 1 Punto Base

Pregunta 2

Suponga que el número de reclamos diarios de afiliados de *Isapres* recibido por cada oficina de la Superintendencia de Salud, relacionados con los montos de las devoluciones, sigue una distribución de Poisson(λ). Debido que en el país hay más de un centenar de oficinas distribuidas a lo largo de todo el territorio, usted selecciona una muestra aleatoria de 35 oficinas y registra el número de reclamos recibidos el lunes 02 de diciembre, los que suman 95 reclamos. Calcule un intervalo de confianza aproximado para la probabilidad de que una oficina reciba más de un reclamo en un día cualquiera, a partir del estimador máximo verosímil de λ.

Solución

Sea $X \sim$ Poisson(λ) la variable aleatoria que representa el número de reclamos diarios de afiliados recibidos por una oficina.

Se solicita un intervalo de confianza aproximado para $g(\lambda)$ dado por

$$g(\lambda) = \mathsf{P}(X > 1) = 1 - \mathsf{P}(X \leq 1) = 1 - \mathsf{P}(X = 0) - \mathsf{P}(X = 1) = 1 - e^{-\lambda}(1 + \lambda). \tag{1}$$

Para construir el intervalo de confianza aproximado se utilizan las propiedades de normalidad asintótica del estimador máximo verosímil de $g(\lambda)$.

Según formulario, el EMV de $g(\hat{\lambda})$ es $g(\hat{\lambda})$, donde $\hat{\lambda}$ es el EMV de λ.

Se puede verificar que el EMV de λ es $\hat{\lambda} = \overline{X}$, donde $\overline{X}$ el número promedio de reclamos entre las 35 oficinas muestreadas.

Así, el EMV de $g(\lambda)$ es $g(\hat{\lambda}) = 1 - e^{-\overline{X}}\,(1 + \overline{X})$.

Como $g(\hat{\lambda})$ es un EMV, se tiene que

$$\frac{g(\hat{\lambda}) - g(\lambda)}{\sqrt{\mathsf{Var}\left(g(\hat{\lambda})\right)}} \dot{\sim} \mathsf{Normal}(0, 1), \quad \text{cuando } n \to \infty, \tag{2}$$

por lo tanto, podemos considerar que $g(\hat{\lambda})$ distribuye aproximadamente $\mathsf{Normal}\left(g(\lambda), \sqrt{\mathsf{Var}\left(g(\hat{\lambda})\right)}\right)$.

De formulario se tiene que

$$\mathsf{Var}\left(g(\hat{\lambda})\right) = \frac{[g'(\lambda)]^2}{I_n(\lambda)}.$$

Para obtener $I(\lambda)$ realizamos lo siguiente:

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!} \to \log L(\lambda) = -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \log\left(\prod_{i=1}^{n} X_i!\right)$$

$$\frac{\partial}{\partial\lambda} \log L(\lambda) = -n + \frac{\sum_{i=1}^{n} X_i}{\lambda}, \quad \frac{\partial^2}{\partial\lambda^2} \log L(\lambda) = -\frac{\sum_{i=1}^{n} X_i}{\lambda^2}$$

$$\to I_n(\lambda) = -\mathsf{E}\left(\frac{\sum_{i=1}^{n} X_i}{\lambda^2}\right) = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}.$$

Además

$$g'(\lambda) = \frac{d}{d\lambda}\left[1 - e^{-\lambda}\,(1 + \lambda)\right] = \lambda\, e^{-\lambda}$$

Por lo tanto

$$\text{Var}\left(g(\hat{\lambda})\right) = \frac{[g'(\lambda)]^2}{I_n(\lambda)} = \frac{\lambda^2\, e^{-2\lambda}}{n/\lambda} = \frac{\lambda^3\, e^{-2\lambda}}{n},$$

y su estimación está dada por

$$\widehat{\text{Var}\left(g(\hat{\lambda})\right)} = \frac{\hat{\lambda}^3\, e^{-2\hat{\lambda}}}{n} = \frac{(\overline{X})^3\, e^{-2\overline{X}}}{n}$$

Finalmente, por propiedades de la distribución normal, un intervalo de confianza aproximado a una confianza del $100(1-\alpha)\,\%$ está dado por

$$< g(\lambda) >_{1-\alpha} \in \left(g(\hat{\lambda}) \pm k_{1-\alpha/2}\sqrt{\widehat{\text{Var}\left(g(\hat{\lambda})\right)}}\right) = \left(1 - e^{-\overline{X}}\,(1+\overline{X}) \pm k_{1-\alpha/2}\sqrt{\frac{(\overline{X})^3\, e^{-2\overline{X}}}{n}}\right)$$

Por enunciado, $\overline{X} = \dfrac{95}{35} = 2.714$, por lo que a un 95 % de confianza (mencionado en aclaración de la evaluación),

$$< g(\lambda) >_{95\,\%} \in \left(1 - e^{-2.714}\,(1+2.714) \pm 1.96\sqrt{\frac{2.714^3\, e^{-2\cdot 2.714}}{35}}\right) = (0.656,\, 0.852) \approx (0.66,\, 0.85).$$

Asignación de Puntaje:

Logro 1 [1.5 Ptos]: por indicar que el EMV de $g(\lambda)$ es $g(\hat{\lambda}$ con $\hat{\lambda})$ el EMV de λ.

- **[0.5 Ptos]** por identificar cantidad de interés $g(\lambda) = 1 - e^{-\lambda}\,(1+\lambda)$
- **[0.5 Ptos]** por obtener EMV de λ dado por $\hat{\lambda} = \overline{X}$.
- **[0.5 Ptos]** por indicar que EMV de $g(\lambda)$ es $g(\hat{\lambda})$.

Logro 2 [2 Ptos]: por plantear propiedad de normalidad asintótica de EMV dada por ecuación (2) o por indicar que $g(\hat{\lambda})$ distribuye aproximadamente Normal$\left(g(\lambda), \sqrt{\text{Var}\left(g(\hat{\lambda})\right)}\right)$.

- **[0.5 Ptos]** por indicar $g(\lambda)$ en propiedad de normalidad asintótica.
- **[1.0 Ptos]** por indicar que $\text{Var}(g(\hat{\lambda})) = [g'(\lambda)]^2/I_n(\lambda)$.
- **[0.5 Ptos]** por normalidad asintótica.

Logro 3 [1.5 Ptos]: por obtener que la varianza del EMV es $\text{Var}(g(\hat{\lambda})) = \lambda^3\, e^{-2\lambda}/n$.

- **[0.5 Ptos]** por $g'(\lambda) = \lambda\, e^{-\lambda}$.
- **[1.0 Ptos]** por $I_n(\lambda) = n/\lambda$

Logro 4 [1 Ptos]: por obtener el intervalo de confianza aproximado.

- **[0.8 Ptos]** por plantear fórmula de intervalo de confianza bajo normalidad: $g(\hat{\lambda}) \pm k_{1-\alpha/2}\sqrt{\widehat{\text{Var}(g(\hat{\lambda})}}$.
- **[0.2 Ptos]** por obtener que $< g(\lambda) >_{95\,\%} \in (0.574,\, 0.934) \approx (0.57,\, 0.93)$.

+ 1 Punto Base

Pregunta 3

En la actualidad, las devoluciones de los excesos cobrados por las *Isapres* son un tema ampliamente discutido. Usted está interesado en demostrar que el monto promedio de devolución obtenido por los hombres es mayor que el de las mujeres. Para ello, realiza una encuesta a 140 afiliados de distintas *Isapres*, obteniendo únicamente 27 respuestas con los montos de las devoluciones, en UF, que recibirán. La Tabla 1 presenta un resumen según género (asuma que la devolución en UF sigue una distribución normal).

Responda las siguientes preguntas indicando: los parámetros involucrados, hipótesis nula y alternativa, valor del estadístico del test, valor-*p* y decisión, justificando esta última. Redacte la conclusión en el contexto del problema. Considere un nivel de significancia del $10\,\%$. En caso de no ser posible calcular un valor-*p* exacto a partir de las tablas de percentiles, proporcione un intervalo donde se encuentre este valor-p.

Característica	**Hombres**	**Mujeres**
Tamaño de muestra	12	15
Devolución media (en UF)	59	50
Desv. estándar (en UF)	8	12

Tabla 1: Resumen de las devoluciones en UF según género.

(a) **[3.0 Ptos.]** ¿Es posible afirmar que los montos, en UF, de las devoluciones en hombres y mujeres presentan la misma variabilidad?

(b) **[3.0 Ptos.]** ¿Es posible afirmar que los montos, en UF, de las devoluciones en hombres es mayor que en las mujeres?

Solución

(a) Sean H y M las variables aleatorias que representan los montos de las devoluciones de los hombres y las mujeres, respectivamente.

Según enunciado, $H \sim \text{Normal}(\mu_H, \sigma_H)$ y $M \sim \text{Normal}(\mu_M, \sigma_M)$.

Se pide llevar a cabo una prueba de hipótesis para evaluar si la varianza entre las distribuciones de H y M son iguales o no.

Así los parámetros involucrados son σ_H^2 y σ_M^2.

La hipótesis nula y alternativa a evaluar son $H_0 : \sigma_H^2 = \sigma_M^2 \quad \text{vs} \quad H_a : \sigma_H^2 \neq \sigma_M^2$.

Para evaluar la hipótesis nula, se realizará la prueba F de homogeneidad de varianzas, la cual, según formulario, es llevada a cabo con el estadístico de prueba F dado por

$$F = \frac{S_H^2/\sigma_H^2}{S_M^2/\sigma_M^2} \sim F(n-1,\, m-1) \quad \text{o} \quad F = \frac{S_M^2/\sigma_M^2}{S_H^2/\sigma_H^2} \sim F(m-1,\, n-1)$$

donde n y m son los tamaños muestrales de hombres y mujeres, respectivamente, y S_H^2 y S_M^2 son las varianzas muestrales de los hombres y las mujeres.

Bajo H_0, $\sigma_H^2 = \sigma_M^2$, y según la Tabla 1, $n = 12$, $m = 15$, $S_H^2 = 8^2 = 64$ y $S_M^2 = 12^2 = 144$.

Así, por simplicidad utilizaremos

$$F_0 = \frac{S_M^2}{S_H^2} = \frac{144}{64} = 2.25 \rightarrow \text{valor-p} = 2 \cdot P(F_{m-1,n-1} > 2.25) > 2 \cdot P(F_{m-1,n-1} > 2.74) = 2 \cdot 0.05 = 10\,\%$$

Por lo tanto no se rechaza la hipótesis que la variabilidad en los montos de las devoluciones son iguales.

(b) Considerando las mismas variables aleatorias especificadas en (a), se llevará a cabo una prueba de hipótesis para evaluar si los montos medios de las devoluciones en hombres es mayor que en las mujeres.

Así, los parámetros involucrados son μ_H y μ_M, ya que interesa determinar si las medias de las distribuciones es igual o no.

La hipótesis nula y alternativa a evaluar es

$$H_0 : \mu_H = \mu_M \quad \text{vs} \quad H_a : \mu_H > \mu_M.$$

Para evaluar la hipótesis nula, se realiza una prueba t de comparación de medias bajo el caso σ_H^2 y σ_M^2 desconocidos e iguales.

Según formulario, esta se evalúa con el estadístico de prueba t dado por

$$T = \frac{(\bar{H} - \bar{M}) - (\mu_H - \mu_M)}{S_p\sqrt{1/n + 1/m}} \sim t\text{-Student}(\nu = n + m - 2),$$

con

$$S_p = \sqrt{\frac{(n-1)\,S_H^2 + (m-2)\,S_M^2}{n - m - 2}} = 10.43.$$

Bajo H_0, $\mu_H - \mu_M = 0$, y según la Tabla 1, $\bar{H} = 59$, y $\bar{M} = 50$. Así,

$$T_0 = \frac{59 - 50}{10.43\,\sqrt{64/12 + 144/15}} = 2.228 \rightarrow 1\,\% < \text{valor-p} < 2.5\,\%$$

.
Por lo que existe evidencia estadística suficiente para rechazar la hipótesis nula, así que es posible afirmar que los montos, en UF, de las devoluciones en hombres es mayor que en las mujeres.

Asignación de Puntaje:

(a) [3 Ptos]

Logro 1 [1.5 Ptos]: por indicar elementos básicos para realizar prueba de hipótesis, correspondientes a

- **[0.5 Ptos]** los parámetros involucrados en la prueba de hipótesis, σ_H^2 y σ_M^2,
- **[0.5 Ptos]** la hipótesis a evaluar, y
- **[0.5 Ptos]** por indicar el estadístico de prueba a utilizar para evaluar la hipotesis.

Logro 2 [1 Ptos]: por evaluar el estadístico de prueba y calcular valor-p.

- **[0.2 Ptos]** por indicar $\sigma_H^2/\sigma_M^2 = 1$.
- **[0.4 Ptos]** por obtener $F_0 = 2.25$ o $F_0 = 0.4444$.
- **[0.4 Ptos]** por indicar un intervalo para el valor-p. Debe indicarse por qué no puede obtenerse de forma exacta. Asignar **[0.2 Ptos]** si sólo da intervalo correcto sin indicar por qué no el valor-p exacto.

Logro 3 [0.5 Ptos]: por indicar la decisión tomada en base al contexto de acuerdo a los resultados obtenidos.

(b) [3 Ptos]

Logro 4 [1.5 Ptos]: por indicar elementos básicos para realizar prueba de hipótesis, correspondientes a

- **[0.5 Ptos]** los parámetros involucrados en la prueba de hipótesis, μ_H y μ_M,
- **[0.5 Ptos]** la hipótesis a evaluar, y
- **[0.5 Ptos]** por indicar el estadístico de prueba a utilizar para evaluar la hipotesis.

Logro 5 [1 Ptos]: por evaluar el estadístico de prueba y calcular valor-p.

- **[0.2 Ptos]** por indicar $\mu_H - \mu_M = 0$.
- **[0.2 Ptos]** por obtener $T_0 = 2.228$.
- **[0.2 Ptos]** por obtener $\nu = n + m - 2 = 25$.
- **[0.4 Ptos]** por indicar un intervalo para el valor-p. Debe indicarse por qué no puede obtenerse de forma exacta. Asignar **[0.2 Ptos]** si sólo da intervalo correcto sin indicar por qué no el valor-p exacto.

Logro 6 [0.5 Ptos]: por indicar la decisión tomada en base al contexto de acuerdo a los resultados obtenidos.

Importante: si estudiante no resuelve (a), desconociendo así la decisión de homogeneidad de varianzas, y decide realizar (b), podrá obtener logros 4, 5, y 6 si especificó en su respuesta que la elección del estadístico de prueba es dependiente de si las varianzas son iguales o no. Si estudiante resuelve (b) bajo ambos casos, puede recibir logros 4, 5, y 6.

+ 1 Punto Base

Pregunta 4

El consumo de gas residencial, en kg, está fuertemente influenciado por las condiciones ambientales extremas. Factores como la temperatura mínima, la humedad, y la cobertura de nubes son claves para entender cómo los hogares adaptan su consumo energético. Para realizar este análisis usted se pone en contacto con una distribuidoras de combustible ubicadas en región metropolitana solicitando la demanda de gas residencial que se observo en 40 días que usted previamente escogió al azar entre los años 2013 y 2023. Posteriormente empalmo a cada día las variables ambientales antes mencionadas a partir de la información proveniente de la estación de monitoreo más cercana a la distribuidora disponible en `https://sinca.mma.gob.cl`, las cuales se describen a continuación:

- **residentialgasdemand**: Demanda de gas residencial diaria, medida en kilogramos (kg).
- **cloudcov_max**: Cobertura máxima de nubes observada durante el día (proporción entre 0 y 1).
- **dayhumidity_max**: Humedad relativa máxima observada durante el día (proporción entre 0 y 1).
- **temp_min**: Temperatura mínima registrada durante el día (en grados Celsius).
- **visibility_min**: Visibilidad mínima registrada durante el día (en kilómetros).

(a) **[4.0 Ptos.]** Para los siguientes cuatro modelos de regresión simple, complete la información faltante:

```
summary(residentialgasdemand)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.       sd.
   1456   13025   18998   20018   26717   43807  9621.529
```

```
------------------------------------------------------------
Modelo 1: lm(residentialgasdemand ~ temp_min)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29258.7     1219.3   24.00  < 2e-16
temp_min     -1775.9     XXXX.X   XX.XX 2.73e-12

Residual standard error: 5085
Multiple R-squared:  0.7278, Adjusted R-squared:  0.7207
F-statistic: 101.6 on 1 and XXXX DF,  p-value: 2.73e-12
------------------------------------------------------------
```

```
------------------------------------------------------------
Modelo 2: lm(residentialgasdemand ~ dayhumidity_max)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)       -28942       6016  -4.811 2.39e-05
dayhumidity_max    66726       8102   8.236 5.59e-10

Residual standard error: 5841
Multiple R-squared:  0.641, Adjusted R-squared:  0.6315
F-statistic: XX.XX on 1 and XXXX DF,  p-value: 5.595e-10
------------------------------------------------------------
```

```
------------------------------------------------------------
Modelo 3: lm(residentialgasdemand ~ visibility_min)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)    27416.4     1744.6  15.715  < 2e-16
visibility_min -2361.0      421.3  -5.604 1.99e-06

Residual standard error: XXXX
Multiple R-squared:  0.4525, Adjusted R-squared:  0.4381
F-statistic:  31.4 on 1 and XXXX DF,  p-value: 1.991e-06
------------------------------------------------------------
```

```
------------------------------------------------------------
Modelo 4: lm(residentialgasdemand ~ cloudcov_max)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -10214      12017  -0.850   0.4007
cloudcov_max   31992      12627   2.534   X.XXXX

Residual standard error: 9016
Multiple R-squared:  X.XXXX, Adjusted R-squared:  X.XXXX
F-statistic: 6.419 on 1 and XXXX DF,  p-value: X.XXXX
------------------------------------------------------------
```

¿Son todos los modelos significativos al 1 %?, ¿cuál de los modelo presenta un mayor porcentaje de variabilidad explicada?

(b) **[2.0 Ptos.]** Uno de los supuestos del modelo de regresión simple, es que el error del modelo distribuye Normal$(0, \sigma)$. Para verificar este supuesto, usted realiza una prueba de bondad de ajuste χ^2 sobre los residuos del mejor modelo ajustado en el item (a). El estadístico de prueba χ^2 es igual a 0.2492 al considerar los siguientes intervalos: `(-Inf,-3000]`, `(-3000,0]`, `(0,+3000]`, `(+3000, +Inf]`. Calcule el valor-p de esta prueba estadística. ¿Existe evidencia al 5 % de significancia para rechazar este supuesto?

Solución

(a) [4 Ptos]

Modelo 1:

- Para obtener `t value` de `temp_min` se utiliza la propiedad de que en una regresión lineal simple el estadístico F de significancia global es el cuadrado del `t value`. Así. `t-value`$= \pm\sqrt{101.6} = \pm 10.07$. Como el signo de $\hat{\beta}$ es negativo, debe ser que `t-value` $= -10.07$.

- Para obtener `Std. Error` de `temp_min` se utiliza definición de `t value`: `t value`$= \hat{\beta}_1 / s_{\hat{\beta}_1}$, de lo que se obtiene que `Std.Error` $= s_{\hat{\beta}_1} = \hat{\beta}_1 / (\texttt{t value}) = -1775.9/-10.07 = 176.355$.
- Para obtener el segundo grado de libertad de la distribución F se extrae del formulario que el estadístico de prueba de la prueba F de significa global sigue una distribución $F(1, n-2)$. Así, el valor faltante es $n - 2 = 40 - 2 = 38$.

Modelo 2:

- Para obtener `F-statistic` se utiliza la propiedad de que en una regresión lineal simple el estadístico F de significancia global es el cuadrado del `t value`. Así. `F-statistic`$= 8.236^2 = 67.831$.
- Para obtener el segundo grado de libertad de la distribución F se extrae del formulario que el estadístico de prueba de la prueba F de significa global sigue una distribución $F(1, n-2)$. Así, el valor faltante es $n - 2 = 40 - 2 = 38$.

Modelo 3:

- Para obtener el segundo grado de libertad de la distribución F se extrae del formulario que el estadístico de prueba de la prueba F de significa global sigue una distribución $F(1, n-2)$. Así, el valor faltante es $n - 2 = 40 - 2 = 38$.
- `Residual standard error` es $\sqrt{s^2_{Y|x}}$. Del formulario se tiene que $R^2 = 1 - \frac{(n-2)}{(n-1)} \frac{s^2_{Y|x}}{S^2_Y}$, donde S^2_Y es la varianza muestral. Del `summary` entregado en la primera línea, se tiene la desviación estándar muestral de Y, $\sqrt{S^2_Y}$, por lo que $S^2_Y = 9621.529^2 = 92573820$. Además, se da $R^2 =$ `Multiple R-squared` $= 0.4525$. Así,

$$s_{Y|x} = \texttt{Residual standard error} = \sqrt{\frac{S^2_Y (n-1)(1-R^2)}{n-2}} = 7212.348$$

Modelo 4:

- Para calcular el valor-p (`Pr(>|t|)`) asociado a `cloudcov_max` se debe calcular `p value` $= 2\,\text{Pr}(T_{38} > |2.534|)$. Como no se dispone de percentiles t-student con 38 grados de libertad, se utiliza el caso $\nu = \infty$, el cual es equivalente a utilizar la distribución normal estándar. Así, se obtiene $\text{Pr}(T_{38} > |2.534|) = \text{Pr}(Z > 2.534) = 1 - 0.9946 = 0.0054$. Así, `p value` $= 0.0108$.
- `Multiple R-squared` es el R^2, el cual por formulario es igual a $R^2 = 1 - \frac{(n-2)}{(n-1)} \frac{s^2_{Y|x}}{S^2_Y}$. De esta expresión, $S^2_{Y|x} = (\texttt{Residual standard error})^2 = (9016)^2$, y S^2_Y fue calculado para el modelo 3, dado por $S^2_Y = 92573820$. Así,

$$R^2 = 1 - \frac{38}{39} \cdot \frac{(9016)^2}{92573820} = 0.144$$

- `Adjusted R-squared` está dado por en formulario por:

$$\texttt{Ajusted R-squared} = r^2 = 1 - \frac{n-1}{n-2}\frac{SCE}{SCT} = 1 - \frac{n-1}{n-2}(1-R^2) = 0.121.$$

- Para obtener el segundo grado de libertad de la distribución F se extrae del formulario que el estadístico de prueba de la prueba F de significa global sigue una distribución $F(1, n-2)$. Así, el valor faltante es $n - 2 = 40 - 2 = 38$.
- El `p-value` de la prueba F es equivalente al valor-p de la prueba de significancia de `cloudcov_max`, calculado previamente como 0.0108.

Para responder si todos los modelos son significativos al 1 %, se analiza el valor-p de la prueba F bajo cada modelo. Modelos 1, 2, y 3 son significativos al 1 % por tener valores-p menores a 0.01, mientras que el modelo 4 no lo es porque el valor-p = $0.0108 > 0.01$.

El porcentaje de variabilidad explicada está dado por el valor del coeficiente de determinación R^2. El modelo que más variabilidad explica es el modelo 1, con un $R^2 = 0.727$.
(b) [2 Ptos]

El estadístico de prueba de esta prueba de bondad de ajuste sigue una distribución χ^2 con $k-1-\nu = 4-1-1 = 2$ grados de libertad. $k = 4$ debido a que se consideran cuatro intervalos, y $\nu = 1$ debido a que se desea verificar el supuesto bajo una distribución con un solo parámetro, Normal$(0, \sigma)$.

Así, el valor-p se calcula como

$$\text{valor-p} = P(C > 0.2482) = 1 - P(C \leq 0.2482)$$

Como no se dispone de valores exactos de la distribución χ^2_2 para calcular este valor-p, se entrega un intervalo de dónde se encuentra el valor-p.

Como $c_{0.1}(2) = 0.211$ y $c_{0.2}(2) = 0.445$, entonces $P(C \leq 0.2482) \in (0.10,\ 0.20)$, por lo que

$$\text{valor-p} \in (0.80,\ 0.90).$$

A un 5 % de significancia, se tiene que el valor-p es mayor a 0.05, por lo que no existe evidencia significativa para rechazar el supuesto de normalidad.

Asignación de Puntaje:

(a) [4 Ptos]

Logro 1 [3 Ptos]: por completar toda la información faltante.

- **[0.4 Ptos]** por indicar correctamente cada uno de los siguientes: 1) `Std. Error`, 2) `t value`, 3) `F-statistic`, 4) `Residual standard error`, 5) `Multiple R-squared`, 6) `df2` de prueba F, y 7) `p-value` de prueba F.
- **[0.2 Ptos]** por `Adjusted R-squared`.

Logro 2 [0.5 Ptos]: por evaluar la significancia global de todos los modelos con prueba F o con prueba t (sólo si especifica equivalencia). Descontar **[0.1 Ptos]** por cada evaluación incorrecta. Si tiene sólo una evaluación correcta, asignar **[0.2 Ptos]**.

Logro 3 [0.5 Ptos]: por comparar coeficientes R^2 entre todos los modelos, e indicar que Modelo 1 es el que explica el mayor porcentaje de variabilidad.

- Si no calcula R^2 del modelo 4, el puntaje máximo de este ítem es **[0.3 Ptos]**.
- Adimitir puntaje máximo de **[0.5 Ptos]** si no calcula R^2 pero usa el `Residual standard error` del Modelo 4 para complementar el análisis y sacar la conclusión correcta.

(b) [2 Ptos]

Logro 4 [0.5 Ptos]: por obtener que los grados de libertad del estadístico χ^2 es iguales a $\nu = 2$.

Logro 5 [0.5 Ptos]: por obtener un intervalo donde se encuentra el valor-p de la prueba de bondad de ajuste. Intervalo = (0.80, 0.90).

Logro 6 [1 Ptos]: por concluir respecto al resultado de la prueba de hipótesis (no se rechaza normalidad).

+ 1 Punto Base
Tiempo: 120 minutos