PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE MATEMÁTICAS
Departamento de Estadística
EYP1113 – Probabilidades y Estadística
Segundo semestre 2024

Ayudantes: Michael Ramón: maramon@uc.cl
Juan Merino: jcamilo.merino@uc.cl
Diego Pérez: dtperez1@uc.cl
Christoph Bürger: christoph.buerger@uc.cl
Oscar Guerrero: oscar.guerrero@uc.cl

Ayudantía 11

Test de Hipótesis, Regresión Lineal

Problema 1 *Prueba de Proporción y Pruebas de Hipótesis Bajo Normalidad*

Un equipo de psicólogos lleva a cabo una investigación respecto a la atención y comprensión de un tópico específico dentro de una clase en un curso universitario. El equipo busca verificar un conjunto de hipótesis, las cuales son:

- *Hipótesis 1*: En cursos masivos, menos de dos tercios de los alumnos logran comprender un tópico específico.
- Frente a una consulta específica del tópico, los alumnos que lograron una buena comprensión responderán acertadamente con:
  - (a) Un tiempo medio menor a 20 segundos.
  - (b) Una desviación estándar mayor a 10 segundos.

Asuma que el tiempo se modela mediante una distribución Normal. Para obtener una muestra se considera un curso masivo de 124 alumnos, quienes son consultados al final respecto a la comprensión del tópico específico. Del total, solo 72 consideran haber comprendido el tópico. Posterior a la clase, los alumnos que indican haber comprendido el tópico son sometidos a una evaluación con la aplicación Kahoot!. Esta aplicación permite medir el acierto y tiempo (en segundos) utilizado para una o más consultas. El resultado para una consulta específica se muestra a continuación:

	Respuesta	
	Correcta	Incorrecta
N	28	44
mean	16	24
sd	12	18

Indique y desarrolle las respectivas pruebas para verificar o refutar las hipótesis planteadas. Utilice un nivel de significancia del 5 %.

Solución:

Hipótesis 1:

Debido que la hipótesis 1 pide analizar una fracción de una cierta población, entonces el test a utilizar es test de proporción. Las hipótesis a probar son:

$$H_0 : p = 2/3 \qquad \text{vs} \qquad H_a : p = 2/3$$

El estadístico de prueba de este test es:

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0\,(1-p_0)}{n}}} \overset{\cdot}{\sim} \text{Normal}(0,1)$$

donde:

$$\hat{p} = \overline{X}_{124} = \frac{72}{124} \qquad p_0 = \frac{2}{3} \qquad n = 124$$

Evaluando es estadístico de prueba se obtiene:

$$Z_0 = -2{,}03$$

El valor-p es:

$$\begin{aligned} \text{valor-p} &= P(Z < Z_0) \\ &= \Phi(-2{,}03) \\ &= 0{,}0212 \end{aligned}$$

Debido a que valor-p $< \alpha$, entonces, se rechaza H_0 a favor de H_a, es decir, existe evidencia para afirmar que menos de $^2\!/_3$ de los alumnos logran la comprensión del tópico.

Hipótesis 2.a:

Se pide probar un tiempo medio, por lo tanto, el test a utilizar es test de media con varianza desconocida, debido a que no se otorga el valor de σ. Las hipótesis a probar son:

$$H_0 : \mu = 20 \qquad \text{vs} \qquad H_a : \mu < 20$$

El estadístico de prueba de este test es:

$$T_0 = \frac{\hat{\mu} - \mu_0}{S/\sqrt{n}} \sim \text{t-Studnet}(n-1)$$

donde:

$$\hat{\mu} = \texttt{mean(Correcta)} = 16 \qquad S = \texttt{sd(Correcta)} = 12 \qquad n = 28$$

evaluando Z_0 se obtiene:

$$Z_0 = -1{,}76$$

El valor-p se obtiene de la siguiente forma:

$$\text{valor-p} = P(T < T_0)$$

Debido a la tabla t-Student es para percentiles, entonces se debe aproximar esta probabilidad, el intervalo d aproximación es:

$$2{,}50\,\% < \text{valor-p} < 5\,\%$$

Debido a que valor-p $< \alpha$, se rechaza H_0 a favor de H_a, es decir, existe suficiente evidencia para afirmar que el tiempo medio es menor a 20 segundos.

Hipótesis 2.b:

Se pide probar una desviación estándar de tiempo, por lo tanto, el test a utilizar es test de varianza con media desconocida. Las hipótesis a probar son:

$$H_0 : \sigma = 10 \qquad \text{vs} \qquad H_a : \sigma > 20$$

El estadístico de prueba de este test es:

$$C_0 = \frac{(n-1)\,S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

donde $\sigma_0 = 10$, evaluando C_0se obtiene:

$$C_0 = 38{,}88$$

El valor-p se obtiene de la siguiente forma:

$$\text{valor-p} = P(C > C_0)$$

Debido a que la tabla chi-cuadrado es de percentiles, entonces se debe aproximar esta probabilidad, el intervalo de aproximación es:

$$5\,\% < \text{valor-p} < 10\,\%$$

Debido a que valor-p $> \alpha$, entonces no se rechaza H_0 a favor de H_a, es decir, existe suficiente evidencia para rechazar la afirmación de que hay la desviación estándar del tiempo es mayor a 10 segundos.

Problema 2 *Inferencia Estadística*

Frente al tema de género que se ha instalado en la discusión nacional, un investigador busca determinar si existen diferencias basales en el desempeño según género. Específicamente, su hipótesis es que frente a situaciones de estrés (por ejemplo, desarrollo de una evaluación) las mujeres tienen un mayor control. Con el fin de verificar o refutar su hipótesis lleva a cabo un cuasi-experimento que consisten en someter a una situación estresantes a dos grupos - Hombres y Mujeres - seleccionados al azar dentro de los alumnos y alumnas del curso, registrando la presión arterial durante la evaluación. Los resultados son:

Resultados	Hombres	Mujeres	Ambos
Num Casos	15	17	32,00
Promedio	85	92	88,00
Mediana	80	85	84,00
Desv. Estándar	8	13	10,00
Promedio LN	-	-	4,48
Desv. Estándar LN	-	-	0,12

Asumiendo que la presión arterial diastólica se comporta de acuerdo a una distribución Normal y teniendo claro que un alza de presión implica una pérdida de control frente a situaciones estresantes, ¿es válida la afirmación del experto para un nivel de significancia de 10 %? Debe plantear hipótesis, indicar el test a utilizar, especificar y validar supuestos (cuando sea necesario) y sus conclusiones deben estar basadas en el valor-p.

Solución:

Un mayor control ante situaciones de estrés indica un bajo valor de presión arterial, por lo que las hipótesis a analizar son en base al promedio de la presión arterial entre hombres y mujeres:

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

ya que las varianzas poblacionales no se conocen, se pueden utilizar dos test, comparación de medias con σ_H y σ_M desconocidos pero iguales o diferentes, para saber cual utilizar es necesario realizar primero un test de varianzas bajo las siguientes hipótesis:

$$H_0 : \sigma_H = \sigma_M \quad \text{vs} \quad H_a : \sigma_H \neq \sigma_M$$

Bajo H_0, el estadístico de prueba a utilizar es:

$$F_0 = \frac{S_H^2}{S_M^2} \sim \text{Fisher}(n_H - 1, n_M - 1)$$

con $S_H^2 = 8^2$, $S_M^2 = 13^2$, $n_H = 15$ y $n_M = 17$, por lo que el estadístico tiene un valor de:

$$F_0 = 0{,}3786982 \sim \text{Fisher}(14, 16)$$

El criterio de rechazo de H_0 es el siguiente:

- $F_0 > F_{1-\alpha/2}(14, 16)$
- $F_0 < F_{\alpha/2}(14, 16)$

Ambos criterios implican que valor-p $< 10\,\%$, para concluir se debe obtener por percentiles $F_{0,05}(14, 16)$ y $F_{0,95}(14, 16)$, de la tabla Fisher se puede obtener el segundo valor fijando $\text{df}_1 = 14$ y $\text{df}_2 = 16$:

$$F_{0,95}(14, 16) = 2{,}37$$

Mediante la siguiente relación se obtiene el segundo percentil:

$$F_{0,05}(14, 16) = \frac{1}{F_{0,95}(16, 14)}$$

de tabla se tiene que $F_{0,95}(16, 14) = 2{,}44$, por lo que

$$F_{0,05}(14, 16) = \frac{1}{2{,}44} = 0{,}409836$$

Comparando se tiene que:

$$F_0 < F_{0,05}(14, 16) \longrightarrow \text{valor-p} < 10\,\%$$

en base a esto se concluye que existe evidencia para rechazar H_0, es decir, las varianzas poblaciones se pueden considerar diferentes con un 10 % de significancia.

Continuando con el test de comparación de medias, como σ_H y σ_M son desconocidas pero diferentes, entonces se utiliza el siguiente estadístico de prueba:

$$T_0 = \frac{\overline{H} - \overline{M}}{\sqrt{\dfrac{S_H^2}{n_H} + \dfrac{S_M^2}{n_M}}} \sim \text{t-Student}(\nu)$$

con $\nu = \dfrac{\left(\dfrac{S_H^2}{n_H} + \dfrac{S_M^2}{n_M}\right)^2}{\dfrac{(S_H^2/n_H)^2}{n_H - 1} + \dfrac{(S_M^2/n_M)^2}{n_M - 1}}$, reemplazando con los datos se obtiene:

$$T_0 = -1{,}85709 \sim \text{t-Student}(105)$$

El valor-p a calcular es:

$$\text{Valor-p} = P(T > T_0) = 1 - P(T \leq T_0)$$

En la tabla se fija $\nu = \infty$ y se debe buscar dos valores donde se encuentre T_0, ya que este es negativo no se encuentra en la tabla, pero se puede utilizar la siguiente relación:

$$t_p(\nu) = -t_{1-p}(\nu)$$

entonces, $-T_0$ se encuentra entre $t_{0,95}(105) = 1{,}645$ y $t_{0,975}(105) = 1{,}960$, por lo que:

$$t_{0,95}(105) = 1{,}645 < 1{,}85709 < t_{0,975}(105) = 1{,}960$$

entonces:

$$t_{0,025}(105) = -1{,}960 < -1{,}85709 < t_{0,05}(105) = -1{,}645$$

$$0{,}025 < P(T \leq T_0) < 0{,}05$$

$$0{,}95 < P(T > T_0) < 0{,}975$$

Como valor-p $> 10\,\%$, por lo tanto no se rechaza H_0 y no se apoya la afirmación del experto.

Problema 3 *Regresión Lineal*

Desde el explorador solar del ministerio de energía se descargó información de la radiación solar que ha afectado el campus San Joaquín UC al mediodía entre 2004 y 2016, y a partir de una muestra aleatoria se construyeron algunos modelos y análisis estadísticos. Entre las variables analizadas están la radiación global (glb) en $^{W}/_{m^2}$, temperatura (temp) a una altura de 2 metros en grados Celsius, la velocidad del viento (vel) en $^{m}/_{s}$ y si existía en ese momento presencia (1: si, 0: no) de nubosidad (cloud).

A continuación se presenta un resumen para la variable `glb`, los coeficientes de determinación R^2 para siete modelos d regresión lineal para predecir `glb` y los valores-p de la prueba KS para la normalidad de los residuos de estos modelos:

```
-----------------------------------------
     n   mean     sd median   min     max
-----------------------------------------
glb 28 714.61 395.78 973.71 47.17 1079.47
-----------------------------------------

lm(glb ~ regresores):
-------------------------------------------
modelo       regresores R-squared   p.value
-------------------------------------------
     1             temp    0.5492  0.982598
     2              vel    0.2556  0.745679
     3            cloud    0.8676  0.275897
     4        temp, vel    0.5819  0.981228
     5       cloud, vel    0.8917  0.973311
     6      cloud, temp    0.9249  0.846099
     7 cloud, temp, vel    0.9283  0.642811
-------------------------------------------
```

a) Complete la información faltante del modelo 2:

```
lm(formula = glb ~ vel, data = data_aux)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    229.7      175.1   1.312  0.20095
vel            384.4      128.6   X.XXX  0.00606
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: XXXX on 26 degrees of freedom
Multiple R-squared:  0.2556,Adjusted R-squared:  X.XXX
F-statistic: X.XXX on 1 and 26 DF,  p-value: X.XXXXX
```

b) Compare el modelo 7 con el mejor modelo simple. ¿El aporte conjunto de las dos variables que se agregan al mejor modelo simple es significativo? Utilice un nivel de significancia del 5 %.

c) ¿Cuál de los modelos cumple el supuesto de Normalidad de los residuos de mejor manera? Justifique su respuesta.

Solución:

Solución a) Primero definamos las variables a utilizar asociadas al modelo 2:

- X: Velocidad del viento `vel`.
- Y: Radiación global `glb`.

El `t value` para la pendiente del modelo es:

$$\texttt{t value} = \frac{\hat{\beta}_1}{\sqrt{\widehat{\text{Var}(\hat{\beta}_1)}}} = \frac{\texttt{Estimate(vel)}}{\texttt{Std. Error(vel)}} = \frac{384{,}4}{128{,}6} = 2{,}989$$

En regresión lineal simple (y solo en regresión lineal simple) sucede que el valor-p asociado al `t value` de la pendiente es igual al valor-p asociado al `F-statistic`, por lo tanto:

$$\texttt{p-value} = \texttt{Pr(>|t|)} = 0{,}00606$$

de la salida de R se tienen los siguientes datos:

$$\texttt{on 26 degrees of freedom} \longrightarrow 28 - 2 = 26 \longrightarrow n = 28$$

$$\texttt{sd(glb)} = S_Y = 385{,}78$$
$$R^2 = \texttt{R-squared(modelo 2)} = 0{,}2556$$

De estos datos se puede obtener la suma de cuadrados totales SCT y suma de cuadrados del error SCE:

$$\text{SCT} = (n-1)S_Y^2 = 4229329$$
$$\text{SCE} = (1-R^2)\text{SCT} = 3148312$$

Entonces, el valor de `Residual standard error` es:

$$\texttt{Residual standard error} = S_{Y|X} = \sqrt{\frac{\text{SCE}}{n-2}} = 347{,}9784$$

El valor de `Adjusted R-squared` es:

$$\texttt{Adjusted R-squared} = r^2 = 1 - \frac{S_{Y|X}^2}{S_Y^2} = 0{,}226969$$

Finalmente, el valor de `F-statistic` se obtiene de:

$$\texttt{F-statistic} = \frac{\text{SCR}/1}{\text{SCE}/(n-2)} = \frac{(\text{SCT}-\text{SCE})/1}{\text{SCE}/(n-2)} = 8{,}927458$$

Alternativamente, existe una ecuación que relaciona R^2 y r^2, esta es:

$$R^2 = 1 - (1-r^2)\frac{n-2}{n-1}$$

Dado que $R^2 = 0{,}2556$, despejando r^2 de la ecuación anterior se obtiene el mismo resultado:

$$r^2 = 0{,}226969$$

Una vez obtenido r^2 y conociendo que $S_Y = 385{,}78$, mediante la definición de r^2 se puede obtener $S_{Y|X}$:

$$r^2 = 1 - \frac{S_{Y|X}^2}{S_Y^2} \longrightarrow S_{Y|X} = 347{,}9784$$

Nuevamente, sólo en regresión lineal simple, el valor de `F-statistic` se puede obtener elevando al cuadrado el `t value` de la pendiente:

$$\texttt{F-statistic} = \texttt{t value(vel)}^2 = 8{,}927458$$

Cómo última opción, una formula alternativa de `F-statistic` es:

$$\texttt{F-statistic} = (n-1)\frac{S_Y^2}{S_{Y|X}^2} - (n-2) = 8{,}927458$$

Solución b) Para conocer el mejor modelo de regresión lineal simple se debe determinar cual tiene el mayor valor de R^2, en base a esto, el mejor modelo simple es el número 3. Para comprara modelos de regresión lineal es necesario obtener el valor del estadístico F, que es:

$$F_0 = \frac{(\text{SCE}_1 - \text{SCE}_2)/r}{\text{SCE}_2/(n-(k+r)-1)} \sim \text{Fisher}(r, n-(k+r)-1)$$

donde SCE_i es la suma de cuadrados del error del modelo correspondiente. Para obtener SCE se utiliza la definición de R^2:

$$R^2 = 1 - \frac{\text{SCE}}{\text{SCT}} \longrightarrow \text{SCE} = (1-R^2)\text{SCT}$$

donde la suma de cuadrado totales, al solo depender de Y, es igual para todos los modelos, por lo tanto, el SCE para el modelo 3 es:

$$\text{SCE}_{\text{modelo 3}} = (1-0{,}8676)\,\text{SCT} = 559963{,}1$$

El SCE del modelo 7 es:

$$\text{SCE}_{\text{modelo 7}} = (1-0{,}9283)\,\text{SCT} = 303242{,}9$$

Se debe cumplir además que $\text{SCE}_1 > \text{SCE}_2$, entonces:

$$\text{SCE}_1 = 559963{,}1 \qquad\qquad \text{SCE}_2 = 303242{,}9$$

Se tienen los siguientes valores:

- r: Cantidad de variables no compartidas entre el modelo 3 y 7, $r = 2$
- k: Cantidad de variables en común entre el modelo 3 y 7, $k = 1$
- n: Cantidad de datos, $n = 28$

Reemplazando en el estadístico F se obtiene:

$$F_0 = 10{,}159 \sim \text{Fisher}(2, 24)$$

El valor crítico es:

$$F_C = F_{1-\alpha}(r, n-(k+r)-1)$$

dado que $\alpha = 5\,\%$, entonces:

$$F_C = F_{1-0{,}05}(2, 24) = 3{,}40$$

Se tiene que:

$$F_0 > F_C$$

entonces, se concluye que las variables `vel` y `temp`, que se agregan a la variable básica `cloud`, tienen un aporte conjunto significativo.

Solución c) Para determinar cual de los modelos tiene el mejor ajuste a una distribución Normal de los residuos, para esto se busca cual tiene el mayor valor-p de la prueba KS. Se puede observar que el modelo 1 presenta el mayor valor-p, por lo que se puede concluir que tiene el mejor ajuste Normal en los residuos.