UNIVERSIDAD NACIONAL DE COLOMBIA, SEDE MEDELLÍN- ESCUELA DE ESTADÍSTICA

PRIMER EXAMEN DE ESTADÍSTICA II, SEMESTRE 01 - 2021

EL EXAMEN CONSTA DE 15 PREGUNTAS DE SELECCIÓN MÚLTIPLE CON 4 OPCIONES DE RESPUESTA, DE LAS CUALES SOLO UNA ES CORRECTA

Tenga en cuenta el siguiente enunciado, para responder las preguntas 1 a 6.

Para determinar la relación que existe entre el tiempo (en días) de una reacción y la cantidad de células por milímetro (Cel.mL) de una variedad de cacao en una reacción química se hizo un experimento. Los datos se muestran en la siguiente tabla:

Tiempo (días)	9	12	15	18	21	24
Cel.mL	0.7917	1.5417	11.8750	2.0000	21.9375	32.4375
	0.2292	0.1041	2.6667	3.6458	8.2708	4.1458
	1.1458	0.6042	1.3125	24.2708	44.0625	8.2917

Considere el modelo Cel.mL vs Tiempo (Tablas 1 y 2 del Anexo) para las preguntas 1 y 2.

1. Al completar la tabla ANOVA incluida en la tabla 1 se obtiene como valor correcto:

   a. MSE = 1859.93 con 16 grados de libertad
   b. Los grados de libertad del SSE son 18
   c. SSR = 1859.93 con 1 grado de libertad
   d. SSE = 1859.93 con 16 grados de libertad

2. Se puede afirmar que:

   a. Los gráficos en la tabla 2, sugieren que el supuesto de varianza constante no se cumple
   b. $1 - R^2 = 0.01183$
   c. La prueba de significancia de la regresión concluye que el modelo no es significativo a un nivel $\alpha = 0.05$
   d. Las opciones a, b y c son falsas

Considere el modelo log(Cel.mL) vs. Log(Tiempo) (Tablas 3 y 4 del Anexo) para las preguntas 3 a 6.
(La abreviatura .log indica logaritmo natural)

3. De los resultados para el modelo 1: Cel.mL vs Tiempo y del modelo 2:log(Cel.mL) vs log(Tiempo), podemos afirmar que:

   a. El modelo 1 no parece cumplir e supuesto de homogeneidad de varianza de los errores.
   b. El modelo 1 tiene mejor R^2 que el modelo 2
   c. En el modelo 2 el supuesto de normalidad de los errores no se cumple
   d. A y b son correctas

4. De las siguientes opciones señale la correcta:

a. Al inicio de la reacción (tiempo = 0) se estima que el logaritmo natural de la cantidad de Cel.mL es -9.4080
b. 59.52% de la variabilidad total de la cantidad de Cel.mL la explica el tiempo
c. La variable logaritmo natural del tiempo explica el 59.52% de la variabilidad total del logaritmo natural de la cantidad de Cel.mL
d. Se estima que por cada día de aumento en el tiempo aumenta el promedio de la cantidad de Cel.mL en 3.8446

5. Sean t (0.025,16) = 2.12, t (0.025,17) = 2.11, t (0.05, 16) = 1.746. t (0.05, 17) = 1.74, algunos percentiles de la distribución t, y s.e ($\widehat{y_0^*}$) = 0.3281para X_0 = 20 dado. Del modelo de regresión Y*= log(Cel,mL) vs X* = log(Tiempo), se obtiene que:
   a. Un intervalo de confianza del 95% para la cantidad promedio de Cel.mL cuando el tiempo es 20, es (1.4138, 2.8050)
   b. Un intervalo de predicción del 95% para el logaritmo natural de la cantidad futura de Cel.mL cuando el tiempo es 20, es (0.6921, 98.1798)
   c. Un intervalo de predicción del 95% para la cantidad futura de Cel,mL resultante cuando el tiempo es 20, es (-0.3680, 4.5868)
   d. Un intervalo de confianza del 95% para la cantidad promedio de Cel.mL cuando el tiempo es 20, es (4.115, 16.5271)

6. Para una reacción que se deja 15 días es correcto:
   a. Se puede estimar que la cantidad promedio de Cel.mL es de 1.0034
   b. Como X_0 = 15 no está en el rango de la predictora, no se puede estimar la cantidad promedio de Cel.mL en 15 días
   c. Se puede estimar que el promedio del logaritmo natural de la cantidad de Cel.mL es de 1.0034
   d. Ninguna de las afirmaciones anteriores es correcta.

7. La función de regresión que relaciona Y, el puntaje de una estudiante en una prueba después de asistir a un curso nivelatorio (Y en puntos) y X, el correspondiente puntaje del estudiante en una prueba antes del nivelatorio (X en puntos) es E [Y|X = x] = 32 + 0.90x, donde X es una variable fija con valores entre 60 y 97 puntos. De las siguientes afirmaciones señala cuál es la correcta:
   a. Se puede predecir que un estudiante con 90 puntos antes del curso obtendrá en la prueba después del curso 113 puntos
   b. Se puede concluir que el promedio en la prueba después del curso para los estudiantes que sacaron 50 puntos en la prueba antes del curso nivelatorio, es 77 puntos
   c. Se puede concluir que para los estudiantes que sacaron cero en el examen antes del curso nivelatorio, el promedio en la prueba después del curso es 32 puntos
   d. Las opciones a, b y c son correctas

8. Si un investigador está interesado en probar la hipótesis de falta de ajuste o linealidad del modelo, donde SSE = SSLOF + SSPE, con las siglas LOF para falta de ajuste y PE para error puro, y F_0 es el estadístico de prueba a ser utilizado. ¿Cuál de los siguientes estadísticos le recomendaría?

a. $F0 = \dfrac{SSLOF / (m - 2)}{SSPE / (n - m)}$

b. $F0 = \dfrac{SSLOF / m}{SSE / (n - 2)}$

c. $F0 = \dfrac{SSLOF / (m - 2)}{SSE / (n - 2)}$

d. $F0 = \dfrac{SSLOF / m}{SSPE / (n - m)}$

9. Relacionado con el coeficiente de determinación R2, se puede afirmar lo siguiente:
   a. El R^2 nos da la proporción de variabilidad total de la variable Y que es explicada por el error aleatorio
   b. Un R^2 grande (cercano a 1) nos garantiza que el modelo es lineal
   c. Un R^2 pequeño (cercano a 0) nos dice que no existe ninguna relación entre las dos variables consideradas
   d. 1 - R^2 nos da la proporción de variabilidad total de la variable Y que es explicada por el modelo considerado

10. De la figura 1 que incluye dos gráficas propias del análisis de un modelo de RLS, se puede afirmar que:

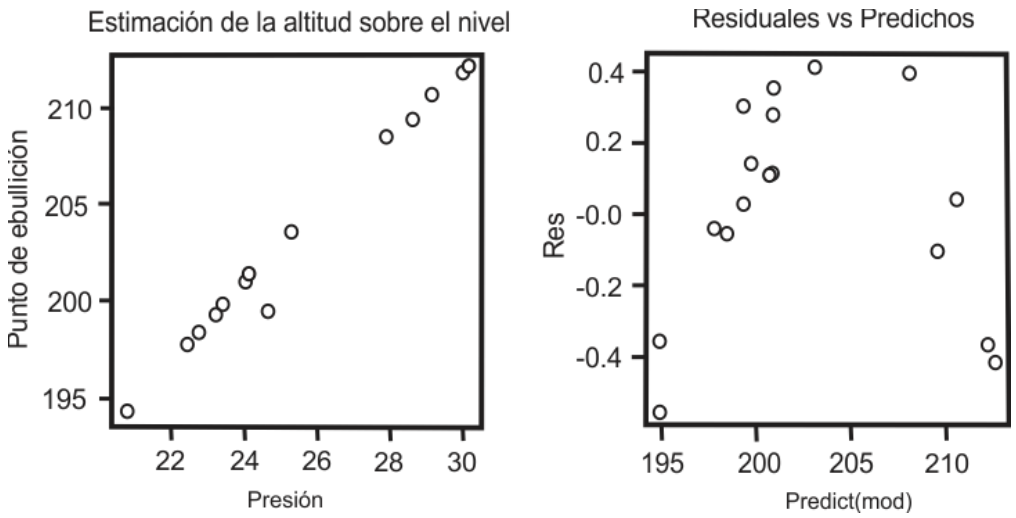

Figura 1: Gráfico de dispersión Y vs. X y Gráfico de residuales ei vs. Predichos $\widehat{y_i}$

   a. Los errores provienen de una distribución normal
   b. El modelo lineal no parece apropiado
   c. La varianza de los errores no parece constante
   d. La varianza de los errores disminuye a medida que aumentan los valores predichos

11. Para definir un intervalo de confianza del 90% para la respuesta media E [Y|x_0], dado un valor apropiado X = x_0, señale cuál de las siguientes expresiones es correcta: (Tenga en cuenta que $\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$, $t_{\nu,\, gl}$ es el valor de la distribución con gl grados de libertad que deja una probabilidad a derecha de ν, y $S_{XX} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$

   a. $\widehat{y_0} \pm t_{(0.05,n-2)} \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}\right)}$

   b. $\widehat{y_0} \pm t_{(0.025,n-2)} \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}\right)}$

   c. $\widehat{y_0} \pm t_{(0.025,n-2)} \sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}\right)}$

   d. $\widehat{y_0} \pm t_{(0.05,n-2)} \sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}\right)}$

12. En relación a un valor futuro $y_0 = Y| x_0$, y a la respuesta media E [Y| x_0] para un punto apropiado y fijo X = x_0. Se puede afirmar que:
    a. A un nivel de confianza (1- α) 100%, los intervalos de predicción para los valores futuros son más estrechos que los intervalos de confianza para la respuesta media.
    b. Los valores futuros son estimados de forma insesgada usando la ecuación de regresión ajustada evaluada en el valor X = x_0
    c. La estimación por intervalo de la respuesta media es más precisa que la correspondiente estimación por intervalo de un valor futuro de la respuesta.
    d. La estimación puntual de la respuesta media y de un valor futuro de la respuesta son diferentes.

13. En relación a la estimación por mínimo cuadrados del intercepto β_0 y de la pendiente β_1 de un modelo de RLS, diga cuál de las siguientes afirmaciones es falsa:

    a. El método de estimación por mínimos cuadrados requiere supuestos distribucionales
    b. $\widehat{\beta_0}$ Es el estimador por mínimos cuadrados del intercepto del modelo RLS
    c. Los estimadores mínimo cuadrático para β_0 y β_1 resultan como solución a las ecuaciones normales de mínimos cuadrados
    d. Algunas de las afirmaciones es falsa

14. Dado el modelo de regresión lineal simple $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ¿Cuál de los siguientes enunciados No es un supuesto?
    a. ε_i ~ Normal, para todo i = 1, 2,…, n
    b. Los ε_i's son v.a.s' dependientes, i = 1, 2,…, n
    c. V [ε_i] = σ^2, para todo i = 1, 2, …, n
    d. E [Y| x_i] = $\beta_0 + \beta_1 x_i$

15. En el modelo exponencial $Y_i = \beta_0 e^{\{x_i \beta_1 + \varepsilon_i\}}$, aplique la transformación $Y_i^* = \log(Yi)$, y para el modelo de RLS resultante diga cuales son los supuestos que debe cumplir: (Nota: log indica logaritmo natural)
    a. $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$
    b. $\log(\varepsilon_i) \overset{i.i.d}{\sim} N(0, \sigma^2)$
    c. $\log(\varepsilon_i) \overset{i.i.d}{\sim} N(\beta_0^* + \beta_1^* x_i, \sigma^2)$
    d. $\varepsilon_i \overset{i.i.d}{\sim} N(\beta_0^* + \beta_1^* x_i, \sigma^2)$

Anexo 1

Tabla 1. Modelo Cel.ml vs. Tiempo

Analysis of Variance Table

Response: Concentración

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
Tiempo			937.33		0.01183
Residuals					
Total		2797.26			

Solution

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
Tiempo	**1**	**937.33**	937.33	**8.06**	0.01183
Residuals	**16**	**1859.93**	**116.246**		
Total	**17**	2797.26			

Anexo 2

Coefficients: (Parámetros Estimados)

	Estimate	Std. Error	T value	Pr (> | t |)
(Intercept)	-13.832	8.570	-1.614	0.1260
Tiempo	1.409	0.496	2.840	0.0118

Tabla 2. Análisis gráfico del modelo Cel.ml vs. Tiempo

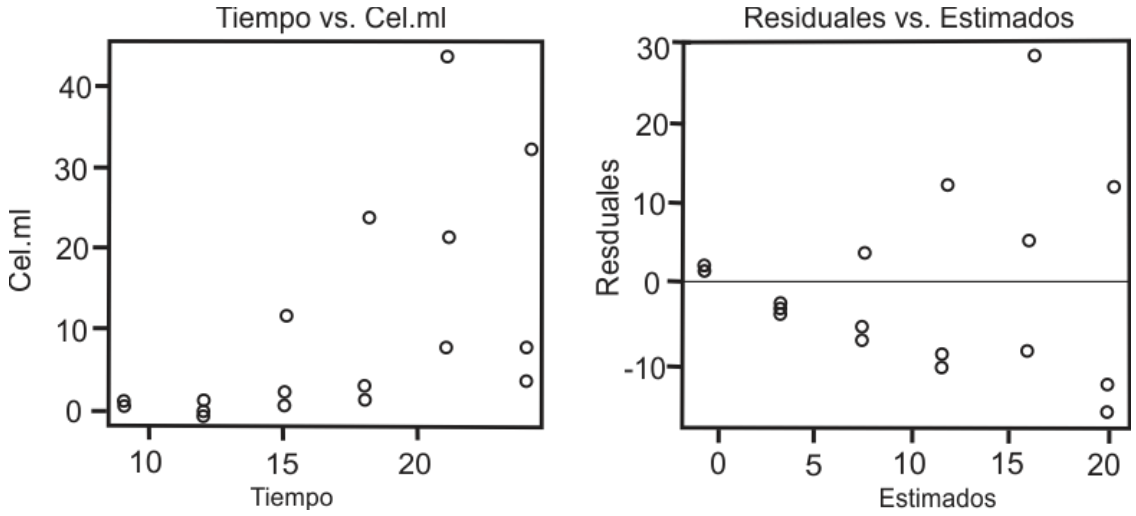

Tabla 3. Modelo log(Cel.ml) vs. log(Tiempo)

Response: log(Concentración)

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
log(Tiempo)	1	29.589	29.5894	23.524	0.0001773
Residuals	16	20.126	1.2579		

Coefficients: (Parámetros Estimados)

	Estimate	Std. Error	T value	Pr (> | t |)
(Intercept)	-9.4080	2.1963	-4.284	0.000570
log(Tiempo)	3.8446	0.7927	4.850	0.000177

Tabla 4. Análisis Gráfico del Modelo log(Cel.ml) vs. log(Tiempo)