Trabajo 1

3, 3

Estudiantes

Brian Alexander Guerrero Bohórquez
Cristobal Henao Rueda
Deisy Johana España Jamioy

Equipo 50

Docente

Carlos Mario Lopera

Asignatura

Estadística II

Sede Medellín
5 de octubre de 2023

Primer trabajo Estadística 2 - Equipo 50

Se tiene una base de datos con 64 registros correspondientes a una muestra aleatoria de 113 hospitales para estudiar la eficacia en el control de infecciones en EE.UU, las variables a estudiar son:

Variable	***Descripción***
Y: Riesgo de infección.	Probabilidad promedio estimada de adquirir infección en el hospital en porcentaje
X1: Duración de la estadía.	Duración promedia de la estadía de todos los pacientes en el hospital en días
X2: Rutina de cultivos.	Razón del número de cultivos realizados en pacientes sin síntomas de infección, por cada 100
X3: Número de camas.	Número promedio de camas en el hospital durante el periodo del estudio
X4: Censo promedio diario.	Número promedio de pacientes en el hospital por día durante el periodo del estudio
X5: Número de enfermeras.	Número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio

1. Estime un modelo de regresión lineal múltiple que explique el riesgo de infección en términos de las variables restantes (actuando como predictoras) Analice la significancia de la regresión y de los parámetros individuales. Interprete los parámetros estimados. Calcule e interprete el coeficiente de determinación múltiple R^2.

Se quiere plantear un modelo de regresión lineal múltiple de la forma:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i, \quad i = 1, 2, \ldots, 64,$$

después de plantear el modelo RLM con ayuda de la función lm() tenemos que las estimaciones de los parámetros son:

Table 2: Tabla de parámetros estimados

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.2506067	1.6153946	-0.7741803	0.4419686
X1	0.1842449	0.1104669	1.6678744	0.1007315
X2	0.0250173	0.0297570	0.8407209	0.4039570
X3	0.0645838	0.0168107	3.8418312	0.0003051
X4	0.0123891	0.0077146	1.6059358	0.1137195
X5	0.0016635	0.0007512	2.2145135	0.0307368

Teniendo en cuenta esta información, se plantea la ecuación de regresión ajustada que está dada por:

$$\widehat{Y}_i = -1.250607 + 0.184245 X_{i1} + 0.025017 X_{i2} + 0.064584 X_{i3} + 0.012389 X_{i4} + 0.001664 X_{i5}, \quad i = 1, 2, \ldots, 64$$

Significancia de los parámetros.

Para estudiar la significancia del modelo planteamos el siguiente juego de hipótesis:

$$\begin{array}{l} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{array} \quad \text{para} \quad j = 0, 1, \ldots, 5.$$

Para ello, tomamos los valores de la última columna de la tabla anterior correspondiente al valor-p para la prueda de la significancia individual de cada parámetro y un nivel de significancia $\alpha = 0.05$. Así podemos concluir que β_3 y β_3 son significativamente distintos de cero cuando los demás parámetros están presentes, por el contrario, β_1, β_0, β_2 y β_4 no lo son.

Interpretación de los parámetros:

6 pt

- $\widehat{\beta}_0$ = -1.2506067. Para que β_0 sea interpretable, cero debe estar entre el rango de las observaciones. En este caso como ningúna de las variables toma el valor de cero, β_0 no es interpretable
- $\widehat{\beta}_1 = 0.1842449$, No significativa
- $\widehat{\beta}_2 = 0.0250173$, No significativa
- $\widehat{\beta}_3 = 0.0645838$, Indica que por cada unidad que aumente el número de camas (X_3) el riesgo de infección aumenta en 0.0645838 unidades, cuando las demás predictoras se mantienen fijas.
- $\widehat{\beta}_4 = 0.0123891$, No significativa
- $\widehat{\beta}_5 = 0.0016635$, Indica que por cada unidad que aumente el número de enfermeras (X_5) el riesgo de infección aumenta en 0.0016635 unidades, cuando las demás predictoras se mantienen fijas.

Significancia de la regresión

Para estudiar la significancia de la regresión se plantea el siguiente juego de hipótesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_5 = 0, \quad \text{vs.}$$
$$H_1 : \exists \beta_j \neq 0, j = 1, \ldots, 5.$$

5 pt

Veamos la tabla ANOVA del modelo:

Table 3: Tabla ANOVA.

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	63.4148	5	12.68296	12.4881	3.0654e-08
Error	58.9052	58	1.01561		

Como $V_p = 3.0654e^{-08} < \alpha = 0.05$ se rechaza H_0 y se concluye que el modelo es significativo, es decir que al menos una predictora del modelos es significativamente distinta de cero.

Coeficiente de determinación R^2

3 pt

Tenemos que $\mathbf{R^2} = \frac{\text{SSR}}{\text{SST}} = \mathbf{0.5184}$, por lo tanto el modelo de RLM anterior, explica el 53.12% de la variabilidad total en el riesgo de infección.

2. Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más pequeños del punto anterior. Según el resultado de la prueba este subconjunto de parámetros son todos significativos? Explique su respuesta.

5 pt

Si tomamos las tres variables con el valor-p más pequeno, basados en la tabla anterior, probaremos la significancia simultánea del subconjunto de las variables β_1, β_3 y β_5.

Se quiere probar que:

$$H_0 : \ \beta_1 = \beta_3 = \beta_5 = 0 \quad \text{vs.} \quad H_1 : \ \exists \beta_j \neq 0, \ j = 1, 3, 5.$$

Observando la tabla de todas las regresiones posibles, tenemos como filas de interés:

Table 4: Muestra de todas las regresiones posibles.

	GL	R^2	$R^2{}_{adj}$	SSE	Cp	Variables
14	2	0.227	0.201	94.581	35.128	X2 X4
16	3	0.490	0.465	62.334	5.376	X1 X3 X5
31	5	0.518	0.477	58.905	6.000	X1 X2 X3 X4 X5

Que contienen la información del modelo reducido, su complemento y el modelo completo, Para validar la hipótesis tenemos que el estadístico de prueba F_0 es:

$$F_0 = \frac{\text{MSextra}}{\text{MSE}} = \frac{[\text{SSE}(\beta_0, \beta_2, \beta_4) - \text{SSE}(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)]/3}{\text{MSE}}$$
$$= \frac{[94.581 - 58.905]/3}{1.01561} = 11.7092191$$

Como $F_0 = 11.70922 > f_{0.05,3,58} = 2.7635518$, entonces se rechaza H_0 y se concluye que el conjunto de predictoras individualmente son significativas por lo tanto no se podrían excluir del modelo.

3. Plantee una pregunta donde su solución implique el uso exclusivo de una prueba de hipótesis lineal general de la forma $H_0 : L\beta = 0$ (solo se puede usar este procedimiento y no SSextra). Especifique claramente la matriz $\mathbf{L}$, el modelo reducido y la expresión para el estadístico de prueba (no hay que calcularlo).

Supongamos que se quiere probar si el Número promedio de pacientes en el hospital y el numero promedio de camas son iguales, esto para estudiar la capacidad del hospital, simultáneamente se quiere saber si la razón de cultivos realizados en pacientes sin síntomas de infección es significativa:

$$H_o : \begin{cases} \beta_3 - \beta_4 = 0 \\ \beta_2 = 0 \end{cases}$$

si se analiza de la forma $H_o : \mathbf{L}\underline{\beta} = 0$ se tiene

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

Así el modelo nulo es

$$Y = \beta_0 + \beta_3(X_3 + X_4) + \varepsilon$$

Luego, se tiene que el estadístico de prueba está dado por:

$$F_0 = \frac{\frac{SSE(RM) - SSE(FM)}{62-58}}{MSE} = \frac{\frac{SSE(RM) - SSE(FM)}{4}}{1.01561} = \frac{SSE(RM) - SSE(FM)}{4.06244}$$

4. Realice una validación de los supuestos en los errores y examine si hay valores atípicos, de balanceo e influenciales. Qué puede decir acerca de la validez de éste modelo?. Argumente su respuesta.

7 pt

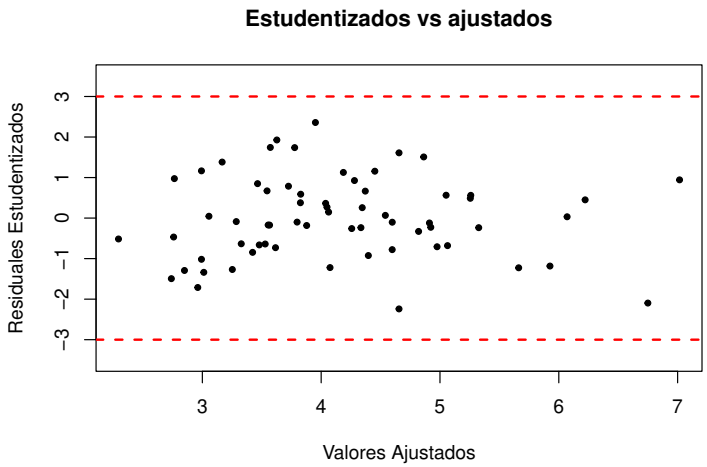
4.1 Análisis de observaciones extremas.

Para realizar el análisis de las observaciones extremas se presenta la tabla de diagnósticos con las Distancias de Cook (Cooks.D), los Valores de la diagonal de la matriz H (hii) y los valores Dffits.

Table 5: Tabla de diagnósticos. (n=10)

base.Y	yhat	se.yhat	residuals	res.stud	Cooks.D	hii	Dffits
4.6	4.3456	0.1894	0.2544	0.2570	0.0004	0.0353	0.0488
4.9	6.7496	0.4870	-1.8496	-2.0963	0.2231	0.2335	-1.1931
4.4	4.0378	0.1552	0.3622	0.3637	0.0005	0.0237	0.0563
6.2	4.6554	0.3078	1.5446	1.6096	0.0444	0.0933	0.5236
4.7	4.9226	0.2288	-0.2226	-0.2268	0.0005	0.0516	-0.0524
1.3	2.9615	0.2720	-1.6615	-1.7122	0.0384	0.0729	-0.4883
1.3	2.7391	0.2970	-1.4391	-1.4943	0.0354	0.0868	-0.4659
5.5	3.7770	0.1826	1.7230	1.7385	0.0171	0.0328	0.3260
4.3	4.0492	0.3851	0.2508	0.2693	0.0021	0.1460	0.1105
4.2	3.8253	0.1856	0.3747	0.3783	0.0008	0.0339	0.0703

Datos Atípicos.

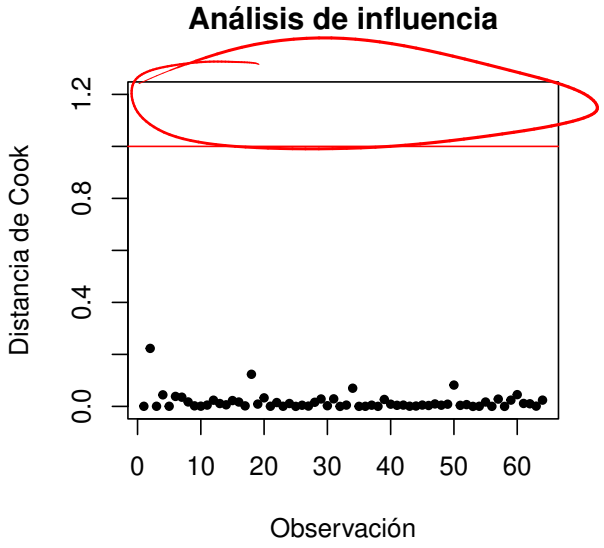


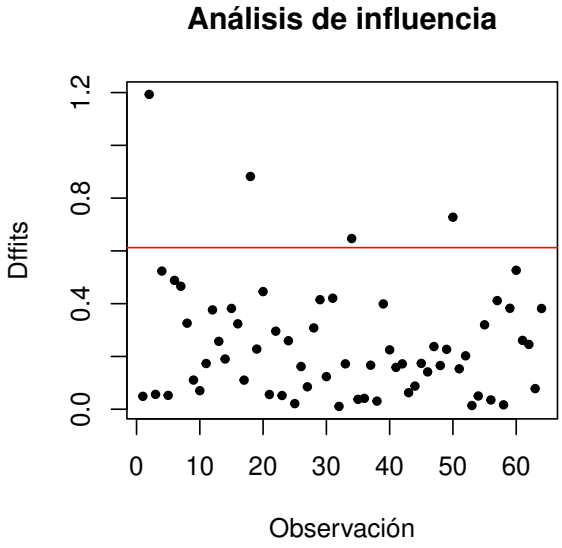
3 pt

Un valor se considera atípico si $-3 < r_i < 3$, por lo tanto según el gráfico no hay observaciones atípicas.

Observaciones Influenciales.

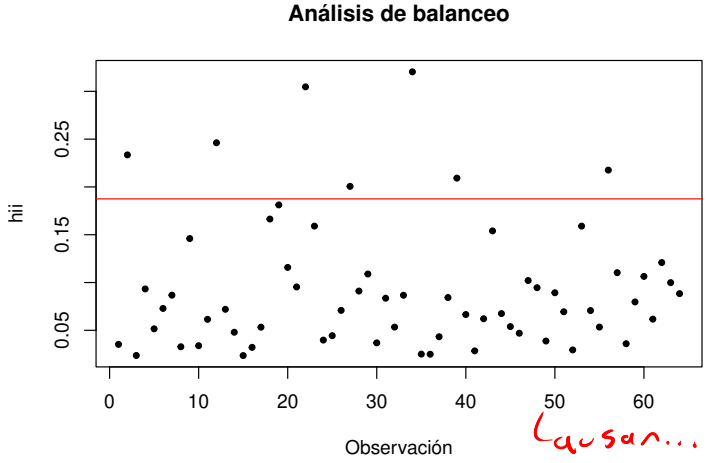
Analicemos la existencia o no de observaciones influenciales por el criterio de la distancia de Cook ($D_i > 1$) y el criterio del diagnóstico DFFITS ($|\mathbf{DFFITS}_i| > 2\sqrt{(\frac{6}{64})} = 0.6123724$):





En primer lugar podemos observar que por el criterio de la distancia de Cook ($D_i > 1$) no se observan datos influenciales, por otro lado, tenemos que por Dffits las observaciones 2, 18, 34 y 50 son influenciales.

Puntos de balanceo.



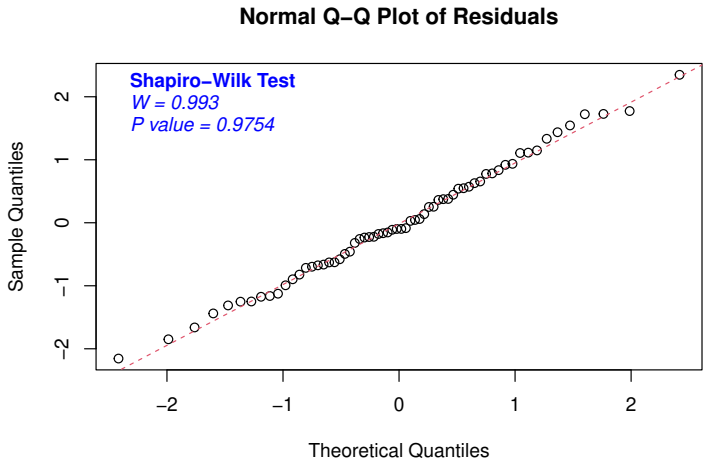
De acuerdo con el Diagnóstico DFBETAS ($h_{ii} > 2 \times \frac{6}{64} = 0.1875$) se sabe que las observaciones 2, 12, 22, 27, 34, 39, y 56 se consideran puntos de balanceo.

4.2 Supuestos del modelo

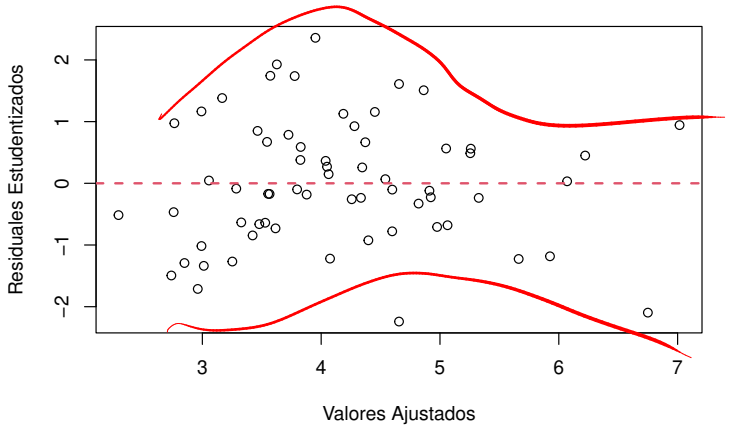
Los supuestos del modelo a validar son normalidad y varianza constante de los errores.

Apoyados en la prueba de Shapiro-Wilk y el gráfico QQplot se desea evaluar:

$$H_0: \varepsilon_i \sim \text{Normal} \text{ vs. } H_1: \varepsilon_i \not\sim \text{Normal}$$



Como $V_p > \alpha = 0.05$ podemos concluir que el supuesto de normalidad se cumple. Para el supuesto de varianza constante veamos el gráfico de residuales estudentisados vs valores ajustados



Con esto se quiere probar:

$$H_0: \ V[\varepsilon_i] = \sigma^2 \quad \text{vs.} \quad H_1: \ V[\varepsilon_i] \neq \sigma^2$$

A primera vista se puede observar que el supuesto de varianza constante en los residuales se cumple, puesto que no se observa un patrón fuerte en la distribución de los mismos.