Trabajo 1

4,6

Estudiantes

Diana Sofía Coral Ibarra
Malcom Andrés Portela Muñoz
Vanessa Alejandra Mendoza Contreras
Sofía Ramírez Guzmán

Equipo 01

Docente

Julieth Veronica Guarin Escudero

Asignatura

Estadística II

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
5 de octubre de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

18 pt

Teniendo en cuenta la base de datos brindada, en la cual hay 5 variables regresoras dadas por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

El modelo contiene las siguientes variables:

- Y_i: Promedio de riesgo de infección.
- X_1: Duración de la estadía.
- X_2: Rutina de cultivos.
- X_3: Número de camas.
- X_4: Censo promedio diario.
- X_5: Número de enfermeras.

1.1. Modelo de regresión

Al ajustar el modelo, obtienen los siguientes coeficientes:

?

Cuadro 1: Tabla de valores coeficientes del modelo

2 pt

	Valor del parámetro
β_0	-1.2854
β_1	0.3184
β_2	0.0134
β_3	0.0406
β_4	0.0139
β_5	0.0003

Por lo tanto, el modelo de regresión ajustado es:

$$\hat{Y}_i = -1.2854 + 0.3184 X_{1i} + 0.0134 X_{2i} + 0.0406 X_{3i} + 0.0139 X_{4i} + 3 \times 10^{-4} X_{5i};\ 1 \leqslant i \leqslant 59$$

1.2. Significancia de la regresión

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

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \text{Algún } \beta_j \text{ distinto de 0 para j=1, 2,..., 5} \end{cases}$$

Cuyo estadístico de prueba es:

$$F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} f_{5\,53} \tag{1}$$

Ahora, se presenta la tabla Anova:

Cuadro 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	55.7241	5	11.144819	13.1847	2.36848e-08
Error	44.8000	53	0.845282		

De la tabla Anova, se observa un valor P demasiado pequeño, donde tomando un $\alpha = 0,05$ para compararlo, podemos rechazar la hipotesis nula en la que $\beta_j = 0$ con $0 \leqslant j \leqslant 5$, aceptando así la hipótesis alternativa en la que algún $\beta_j = 0$. Concluyendo así que la regresión es significativa. Además por medio del estadístico de prueba, el cual tiene un valor de $F_0 = 13,1847$, comparandoló con el valor que se obitene a partir de la distribución F, con sus respectivos grados de libertad tenemos que $f_{5,53} = 2.3894$, dando así un F_0 mucho mayor a $f_{5,53}$, por ende rechazamos la hipotesis nula como lo hicimos con el criterio anterior, confirmando así la aceptación de la hipotesis alternativa.

1.3. Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros, la cual permitirá determinar cuáles de ellos son significativos.

Cuadro 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	-1.2854	1.5696	-0.8189	0.4165
β_1	0.3184	0.1041	3.0587	0.0035
β_2	0.0134	0.0289	0.4632	0.6451
β_3	0.0406	0.0123	3.3048	0.0017
β_4	0.0139	0.0067	2.0758	0.0428
β_5	0.0003	0.0007	0.4713	0.6393

Teniendo en cuenta los P-valores de la tabla anterior, podemos concluir con un nivel de significancia $\alpha = 0.05$ que los parámetros β_1, β_3 y β_4 individualmente son significativos en el modelo, pues sus P-valores son menores a α.

1.4. Interpretación de los parámetros

Solo se interpretan los parámetros significativos, estos son:

$\hat{\beta_1} = 0.3184$**:** Indica que por cada unidad que aumente la duración de la estadía, el promedio del riesgo de infección aumente en 0.3184 unidades, cuando las demás predictoras se mantienen fijas.

$\hat{\beta_3} = 0.0406$**:** Indica que por cada unidad que aumente el número de camas, el promedio del riesgo de infección aumente en 0.0406 unidades, cuando las demás predictoras se mantienen fijas.

$\hat{\beta_4} = 0.0139$**:** Indica que por cada unidad que aumente el censo promedio diario, el promedio del riesgo de infección aumente en 0.0139 unidades, cuando las demás predictoras se mantienen fijas.

1.5. Coeficiente de determinación múltiple R^2

El modelo tiene un coeficiente de determinación múltiple

$$R^2 = \frac{SSR}{SST} = \frac{55.7241}{55.7241 + 44.8000} = 0,5543 \tag{2}$$

lo que significa que aproximadamente el 55, 43 % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

2. Pregunta 2

5 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Las covariable con el P-valor más pequeño en el modelo fueron X_1,X_3 y X_4, por lo tanto a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

Cuadro 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	44.800	X1 X2 X3 X4 X5
Modelo reducido	95.369	X2 X5

Luego un modelo reducido para la prueba de significancia del subconjunto es:

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_5 X_{5i} + \varepsilon;\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

2.2. Estadístico de prueba y conclusión

3 pt

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_2, \beta_5) - SSE(\beta_0, \cdots, \beta_5))/3}{MSE(\beta_0, \cdots, \beta_5)} \overset{H_0}{\sim} f_{3,53} \\ &= \frac{(95.369 - 44.800)/3}{0.845282} \\ &= 19,942 \end{aligned} \qquad (3)$$

2 pt

Ahora, comparando el $F_0 = 19,942$ con $f_{0.95,3,53} = 2.7791$, se puede ver que F_0 es mayor a $f_{0.95,3,53}$, por tanto se rechazaría la hipotesis nula, tomando así que los β_1, β_3 y β_4 son significativos para el modelo, por lo cual no es posible descartar estas variables. De igual manera en el punto anterior ya habiamos hecho la prueba de significancia por cada parámetro, aquí podemos confirmar los resultados anteriores.

3. Pregunta 3

5 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

Se hacen las siguientes preguntas:

- ¿El efecto de la duracion de la estadia sobre el promedio de riesgo de infeccion es igual a 4 veces el efecto del numero de camas sobre el promedio de riesgo de infeccion?
- ¿El efecto que tiene el censo promedio diario sobre el promedio de riesgo de infeccion es igual a 2 veces el efecto que tiene el numero de camas sobre el promedio de riesgo de infeccion?

por consiguiente se plantea la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = 4\beta_3;\ \beta_4 = 2\beta_3 \\ \mathrm{H}_1 : \text{Alguna de las igualdades no se cumple} \end{cases}$$

reescribiendo matricialmente:

$$\begin{cases} \mathrm{H}_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ \mathrm{H}_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Con $\mathbf{L}$ dada por

$$L = \begin{bmatrix} 0 & 1 & 0 & -4 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 & 0 \end{bmatrix}$$

El modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_3 X^*_{3i} + \beta_5 X_{5i} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2);\ 1 \leqslant i \leqslant 59$$

Donde $X^*_{3i} = 4X_{1i} + X_{3i} + 2X_{4i}$

3 pt

3.2. Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - 44.8000)/2}{0.845282} \overset{H_0}{\sim} f_{2,53}$$

2 pt

4. Pregunta 4

17,5

4.1. Supuestos del modelo

4.1.1. Normalidad de los residuales

Para la validación de este supuesto, se planteará la siguiente prueba de hipótesis la cual se realizará mediante el test de shapiro-wilk, acompañada de un gráfico cuantil-cuantil:

$$\begin{cases} H_0 : \varepsilon_i \sim \text{Normal} \\ H_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$

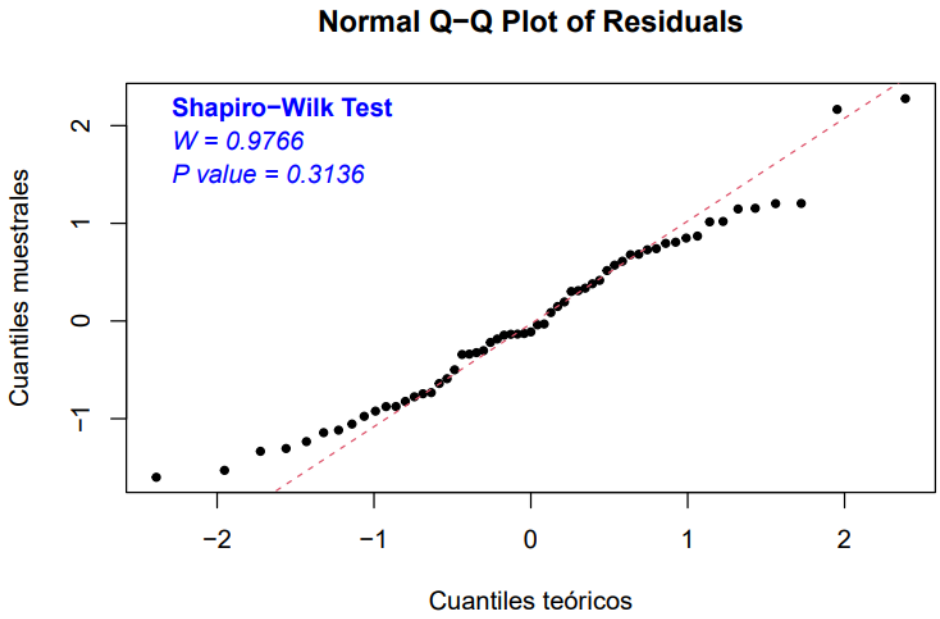

Figura 1: Gráfico cuantil-cuantil y normalidad de residuales

Al ser el P-valor aproximadamente igual a 0.3136 y teniendo en cuenta que el nivel de significancia es $\alpha = 0.05$, el P-valor es mucho mayor y por lo tanto, no se rechazaría la hipótesis nula, es decir que los residuales distribuyen normal, sin embargo la gráfica de comparación de cuantiles permite ver colas más pesadas y patrones irregulares, Por lo tanto se concluye que el supuesto de normalidad se cumple pero se deja constancia de irregularidades en el ajuste a la recta normal, que se podría ver afectado por datos extremos que analizaremos más adelante.

Ahora se validará si la varianza cumple con el supuesto de ser constante.

4.1.2. Varianza constante

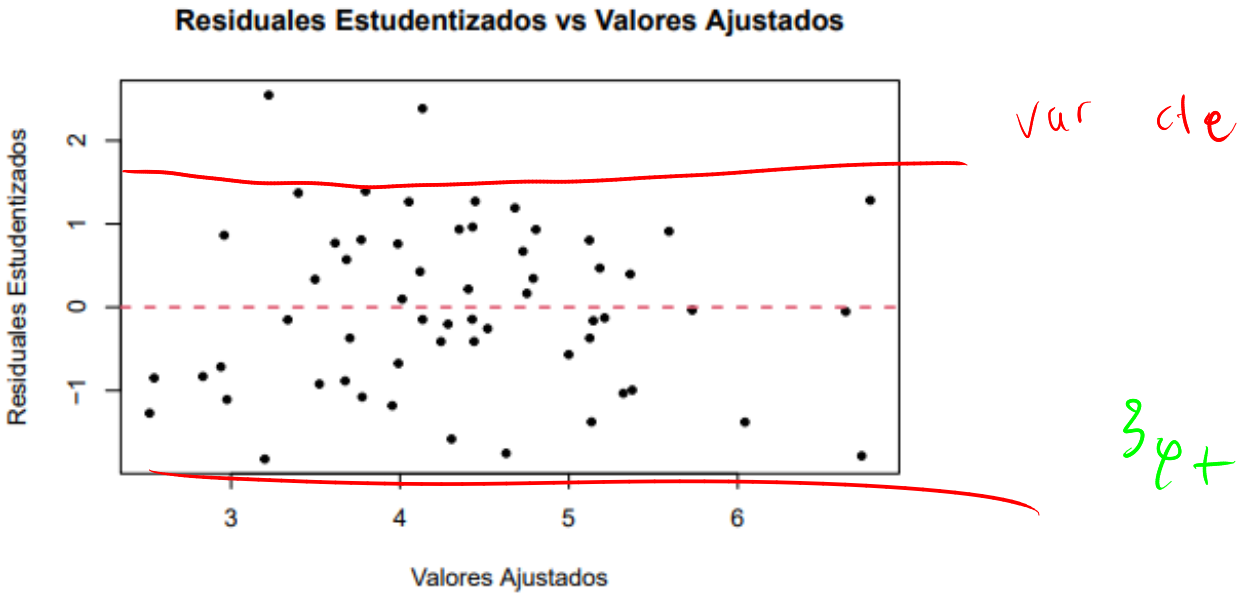

Figure 2: Gráfico residuales estudentizados vs valores ajustados

En el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente, decrezca ni un comportamiento que permita descartar una varianza constante, al no haber evidencia suficiente en contra de este supuesto se acepta el supuesto de varianza constante.

4.2. Verificación de las observaciones

Se va a realizar un análisis de observaciones extremas dividido en tres partes, datos atípicos, observaciones influenciales y puntos de balanceo.

4.2.1. Datos atípicos

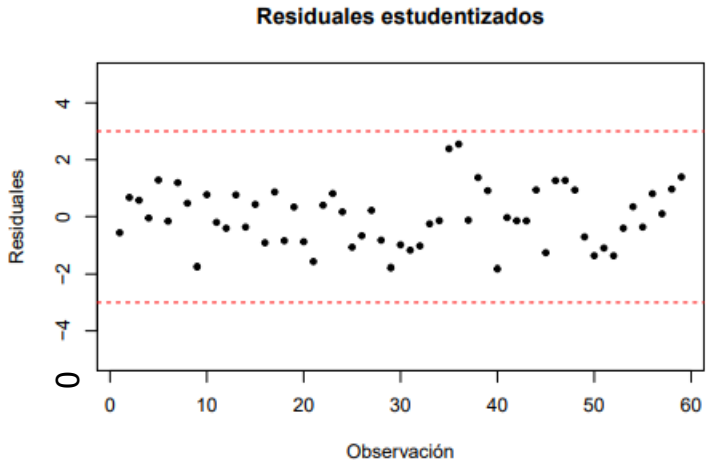

Figure 4: Identificación de datos atípicos

Como se puede observar en la gráfica de residuales, no hay datos atípicos en el conjunto de datos pues ningúnresidual estudentizado sobrepasa el criterio de $|r_{estud}| > 3$.

4.2.2. Puntos de balanceo

se puede apreciar que existen 6 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 0.202338$, los cuales se presentan en la siguiente tabla.

Figura 4: Identificación de puntos de balanceo

	res.stud	Cooks.D	hii.value	Dffits
4	-0.0525	0.0002	0.3348	-0.0369
5	1.2841	0.0970	0.2608	0.7675
13	0.7602	0.0292	0.2324	0.4166
25	-1.0780	0.0540	0.2179	-0.5699
29	-1.7850	0.2726	0.3392	-1.3066
55	-0.3706	0.0059	0.2058	-0.1871

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor

$h_{ii} = 2\frac{p}{n} = 0.202338$

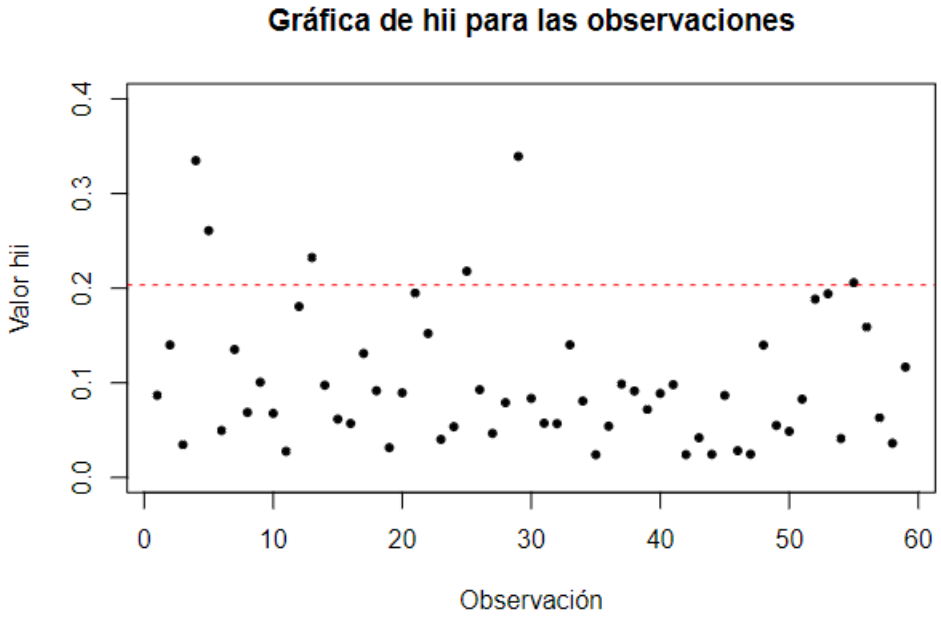

Figure 3: Identificación de puntos de balanceo

Así, las observaciones 4, 5, 13, 25, 29 y 55 se consideran puntos de balanceo.

1,5 pt

¿Qué causan?

4.2.3. Puntos influenciales

Observemos en primer lugar, el método de la distancia de Cook para identificar puntos influenciales, este método considera una observación como influencial si $D_i > 1$, donde D_1 representa la distancia de Cook para cada observación.

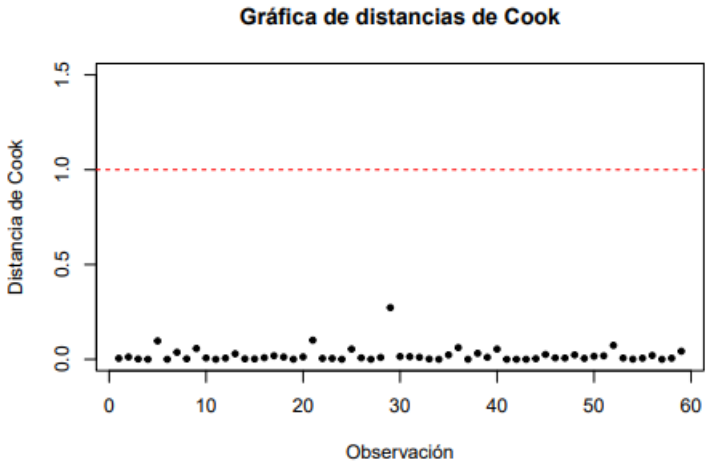

Figure 5: Criterio distancias de Cook para puntos influenciales

Por este método no se identifican observaciones influenciales, pro no significa que no existan, veamos tambiénel método el criterio Dffits

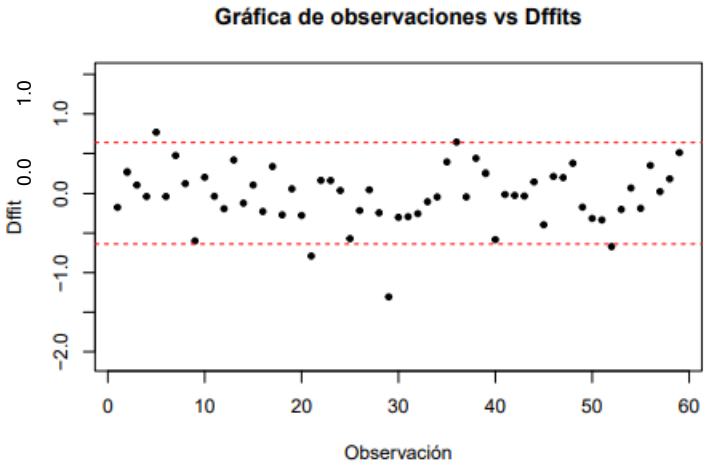

Figure 6: Criterio Dffits para puntos influenciales

Según el criterio de Dffits, Una observación se considera influencial si $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.6377928$ como se puede ver a continuación:

Table 2: Observaciones Influenciales

	res.stud	Cooks.D	hii.value	Dffits
5	1.2841	0.0970	0.2608	0.7675
21	-1.5822	0.1009	0.1947	-0.7896
29	-1.7850	0.2726	0.3392	-1.3066
36	2.5465	0.0618	0.0541	0.6439
52	-1.3798	0.0737	0.1884	-0.6706

3 pt

De esta forma, las observaciones 5, 21, 29, 36 y 52 son consideradas influenciales.

¿Qué causan?

4.3. Conclusión

El modelo de regresión lineal múltiple planteado cumple con los supuestos de normalidad y varianza con- stante, por lo que es valido para hacer estimaciones y predicciones, sin embargo se podría considerar un mejor modelo. Ya que el anterior solo explica aproximadamente el 55.43% de la variabilidad total de la probabilidad del riesgo de infección. Sería interesante ver si al eliminar las observaciones influenciales y puntos de balanceo el modelo tienen cambios importantes y analizar la posibilidad de eliminar las variables no significativas o aumentar labase de datos

✓

5 pt