4,4

Trabajo 1

Estudiantes:

Maria Camila Correa Toro

Simón Pedro Serna Cardona

Angie Pahola Tobar Calpa

Juan Manuel Teherán Machado

Grupo 35

Docente:

Francisco Javier Rodriguez Cortés

Estadística II

30 de marzo 2023

UNIVERSIDAD
NACIONAL
DE COLOMBIA

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 cuadrado.

Se plantea que los datos pueden seguir un modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i \quad i = 1, 2, \ldots, 50$$

$$\varepsilon_i \sim N(0, \sigma^2) \text{ idéntica e independientemente } \forall\ i = 1,\ldots, 50$$

Donde:

Y_i es el riesgo de infección

X_1 es la duración de la estadía

X_2 es la rutina de cultivos

X_3 es el número de camas

X_4 es el censo promedio diario

X_5 es el número de enfermeras

Al construir el modelo ajustado de la regresión, apoyados del software R, obtenemos los siguientes valores para los parámetros de la regresión.

	Valor estimado	Error Estándar	Estadístico T	Valor P
Intercepto	-0.336245991	1.7911618483	-0.18772507	0.851955226
β_1	0.318593025	0.0920597851	3.46071875	0.001209711
β_2	-0.003336017	0.0363059745	-0.09188616	0.927205565
β_3	0.064000612	0.0182651308	3.50397777	0.001066109
β_4	0.001687620	0.0086482408	0.19514022	0.846181498
β_5	0.001919818	0.0006749782	2.84426715	0.006730093

Con base en la tabla de parámetros estimados, se obtiene la ecuación de regresión ajustada:

$$\widehat{Y_i} = -0.336245991 + 0.318593025 X_{i1} - 0.003336017 X_{i2} + 0.064000612 X_{i3} + 0.001687620 X_{i4} + 0.001919818 X_{i5}$$

Significancia de los parámetros del modelo 6 pt

Para probar la significancia individual de un parámetro del modelo, se tienen las hipótesis:

$$H_0: \beta_j = 0 \text{ vs } H_1: \beta_j \neq 0$$
$$\text{para } i = 0, 1, \ldots, 5$$

A un nivel de significancia $\alpha = 0.05$ se concluye que los parámetros individuales $\beta_1, \beta_3 \, y \, \beta_5$ **son significativos** cada uno en presencia de los demás parámetros. Mientras que los parámetros $\beta_0, \beta_2 \, y \, \beta_4$ **no son significativos** individualmente.

Interpretación de los parámetros estimados 2,5 pt

Solo los parámetros significativos se pueden interpretar. $\widehat{\beta_1} = 0.318593025$ indica que por cada unidad promedio de días que aumente la estadía de todos los pacientes, la probabilidad de adquirir una infección en el hospital aumenta en 0.3185 unidades, cuando las demás variables predictoras se mantienen fijas . $\widehat{\beta_3} = 0.064000612$ indica que por cada unidad promedio de camas que aumente, la probabilidad de adquirir una infección en el hospital aumenta en 0.064 unidades, cuando las demás variables predictoras se mantienen fijas. $\widehat{\beta_5} = 0.001919818$ indica que por cada unidad promedio de enfermeras que aumente, la probabilidad de adquirir una infección en el hospital aumenta en 0.0019 unidades, cuando las demás variables predictoras se mantienen fijas.

Significancia de la regresión

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \text{ vs } H_1: \text{algún } \beta_j \neq 0, \text{ para } j = 1,\ldots,5$$

Tabla ANOVA					
	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico F	Valor P
Modelo	76.6161	5	15.323216	16.8458	$2.79971 * 10^{-9}$
Error	40.0231	44	0.909616		

De la tabla ANOVA se obtiene un Valor-P de $2.79971 * 10^{-9}$, asumiendo un nivel de significancia $\alpha = 0.05$, como el valor $P < 0.05$, se rechaza H_0 y se concluye que el modelo de regresión es significativo, lo cual quiere decir que, la probabilidad promedio estimada de adquirir infección en el hospital depende significativamente de al menos una de las variables.

5 pt

Coeficiente de determinación múltiple

Con la tabla ANOVA podemos calcular:

$$R^2 = \frac{SSR}{SST} = \frac{76.6161}{76.6161+40.0231} = 0.6568641$$

Lo cual quiere decir que el 65.68% de la variabilidad total de la probabilidad promedio de adquirir infección en el hospital es explicado por el modelo de RLM propuesto en el presente informe..

2. Use la tabla de todas las regresiones posibles, para probar la significancia simultánea del subconjunto de tres variables con los valores p más grandes del punto anterior. Según el resultado de la prueba es posible descartar del modelo las variables del subconjunto? Explique su respuesta.

De los datos obtenidos anteriormente, podemos ver que las variables con valores P más grandes fueron: X_2, X_4 y X_5. Así, planteamos las siguientes hipótesis:

$$H_0: \beta_2 = \beta_4 = \beta_5 = 0 \text{ vs } H_1: \exists \beta_j \neq 0, \ para\ algún\ j = 2, 4, 5.$$

Así, para probar nuestra hipótesis, tenemos que construir un modelo de regresión dada H_0, el cual llamaremos Modelo Reducido (MR), que tiene la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i; \varepsilon_i \sim N(0, \sigma^2)\ idéntica\ e\ independientemente\ \forall\ i = 1,\ldots,50$$

Para poner a prueba H_0, utilizaremos el estadístico F dado por:

$$F_0 = \frac{SSR(X_2, X_4, X_5|X_1, X_3)/gl(SSR(X_2, X_4, X_5|X_1, X_3))}{MSE} = \frac{[SSE(MR) - SSE(MC)]/(GL(MR) - GL(MF))}{MSE}$$

Donde MC se refiere al Modelo Completo, MR es el Modelo Reducido, MSE es la suma de cuadrados medios del error, correspondiente al Modelo Completo, y GL a los Grados de Libertad de la suma de cuadrados correspondiente.
Utilizando R, podemos obtener la siguiente información para el modelo de regresión ajustado con las variables no nulas bajo H_0, sean estas X_1 y X_3.

#Variables Regresoras	R Cuadrado	Suma de Cuadrados del Error	Variables
2	0.593	47.448	X_1, X_3

Vemos que el número de parámetros para el MR es p = 3, y SSE(MR) = 47.448, así:

$$F_0 = \frac{[69.817 - 40.023]/(50 - 3 - (50 - 6))}{0.909616} \approx 10.91816, donde\ F_0 \sim F(3, 44)$$

Donde 3 corresponde a los grados de libertad del numerador y 44 a los grados de libertad del MSE.

Así, calculando el valor P de F_0, se tiene que:

$$P(F(3, 44) > F_0) = 1.459349 * 10^{-5}$$

Asumiendo un valor $\alpha = 0.05$, vemos que el valor P es muy pequeño respecto a α, por tanto se rechaza H_0 y se concluye que al menos uno de los parámetros entre X_2, X_4 y X_5 es significativo, por tanto, ninguno puede ser descartado en virtud de la prueba realizada.

3. Plantee una pregunta donde su solución implique el uso exclusivo de una prueba de hipótesis lineal general de la forma H0 : Lβ = 0 (solo se puede usar este procedimiento y no SSextra). Especifique claramente la matriz L, el modelo reducido y la expresión para el estadístico de prueba (no hay que calcularlo).

Queremos saber según el modelo ajustado, si $\beta_1 y\, \beta_2$ son equivalentes y a vez, si también lo son $\beta_3\, y\, \beta_4$.

Para ello planteamos la siguiente prueba de hipótesis.

$$H_0: \beta_1 = \beta_2,\ \beta_3 = \beta_4 \textbf{ vs } H_1: Algunas\ de\ las\ desigualdades\ no\ se\ cumplen$$

Reescribiendo matricialmente:

$$H_0: L\bar{\beta} = 0 \textbf{ vs } H_1: L\bar{\beta} \neq 0$$

donde $\bar{\beta} = [\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4\ \beta_5]$ y $L = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}$

Así, podemos plantear un modelo teórico, bajo H_0 de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_{i1,2} + \beta_3 X_{i3,4} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)\ idéntica\ e\ independientemente\ \forall\ i = 1,\ldots, 50$$

$$X_{i1,2} = X_{i1} + X_{i2},\ X_{i3,4} = X_{i3} + X_{i4}$$

Este modelo lo llamaremos Modelo Reducido dado H_0, o para abreviar, MR.

$$F_0 = \frac{MSH}{MSE} = \frac{SSH/2}{MSE} = \frac{[\,SSE(MR) - SSE(MC)]/2}{MSE(MC)} = \frac{[SSE(MR) - 40.0231]/2}{0.909616}$$

$$Donde\ F_0 \sim F_{2,44}$$

SSH hace referencia a la suma de cuadrados del modelo dado H_0 cierto, MC hace referencia al modelo completo y el MSE corresponde también al MC, para el cual ya se ha ajustado un modelo y declarado el valor de sus sumas de cuadrados anteriormente, junto con sus grados de libertad.

Los grados de libertad para el numerador corresponden a la cantidad de ecuaciones linealmente independientes que tenemos en la hipótesis nula.

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.

Supuestos del modelo

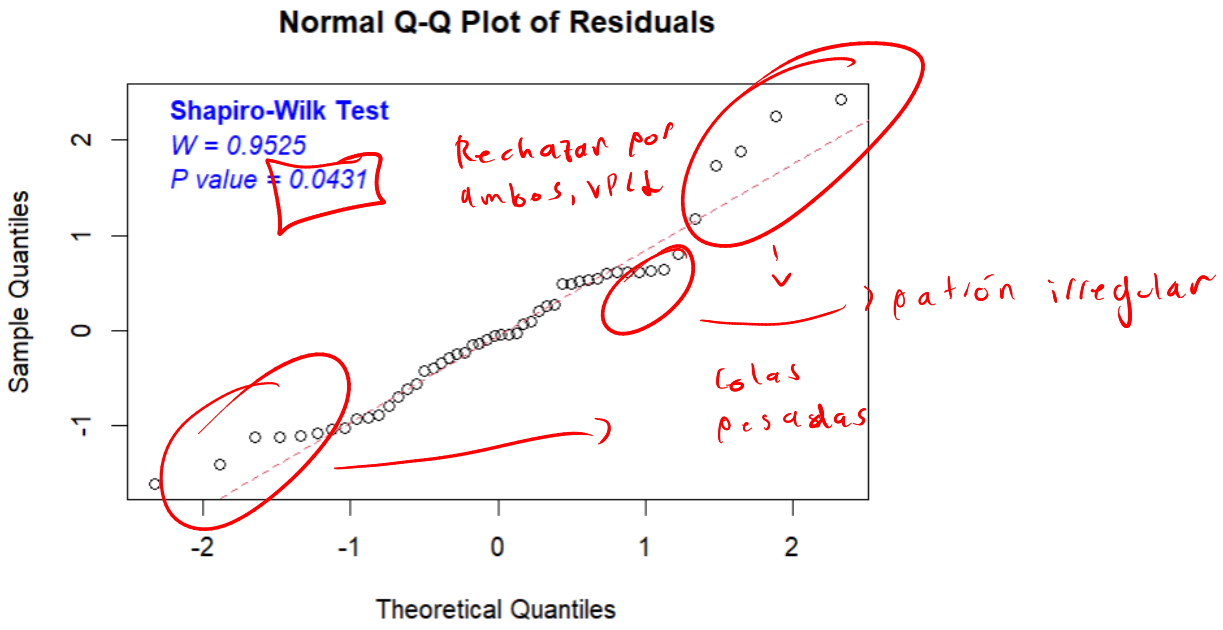
-Normalidad de los residuales:

Para probar este supuesto, partimos del siguiente par de hipótesis:

$$H_0: \varepsilon_i \sim N(0, \sigma^2)\ idéntica\ e\ independientemente\ \ \textbf{vs}\ H_1: \varepsilon_i\ no\ se\ distribuye\ normal$$

$$\forall\ i = 1,\ldots, 50$$

Apoyados del software estadístico R, generamos el siguiente gráfico y efectuamos el Test de Shapiro-Wilk sobre los residuales del modelo ajustado.

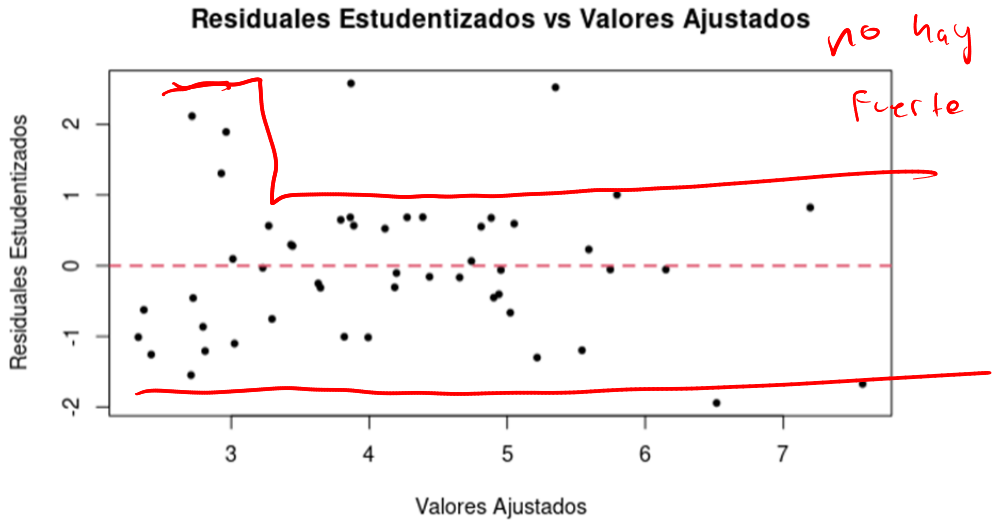


Vemos que se tiene un valor p aproximadamente a 0.0431 y con asumiendo nivel de significancia del α=0.05, se rechaza la hipótesis nula, por lo que se concluye que el supuesto

de distribución de los datos no es normal. Más importante que ésta prueba analítica, en el gráfico Cuantil-Cuantil se puede observar la falta de ajuste de los residuales, colas pesadas y patrones irregulares.

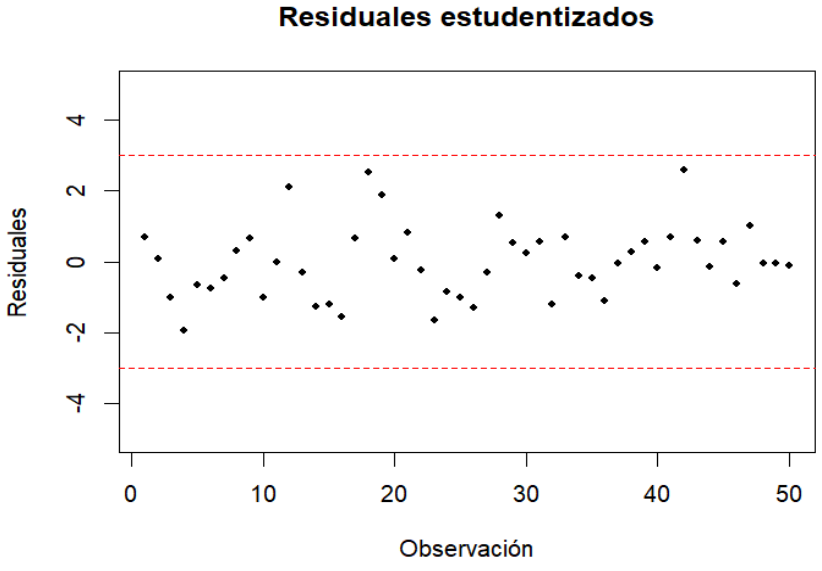
-Varianza constante.

Generando un gráfico de residuales estudentizados respecto a los valores ajustados, tenemos lo siguiente:



Se observa de la gráfica estudentizados vs valores ajustados, no se tiene suficiente información que nos permita concluir una varianza constante o no constante, ya que, no se evidencia una dispersión creciente o decreciente de los puntos a medida que aumente la variable independiente, pero si se observa media 0. Finalmente con esta información se acepta el supuesto de varianza constante.

Verificación de las observaciones
-Datos atípicos.



Se observa que el conjunto de datos del modelo de regresión lineal múltiple, no se evidencia algún dato atípico, ya que, ninguna de las observaciones se encuentran más arriba de 3 o más abajo que -3. Se concluye que no existen datos atípicos en las observaciones realizadas.

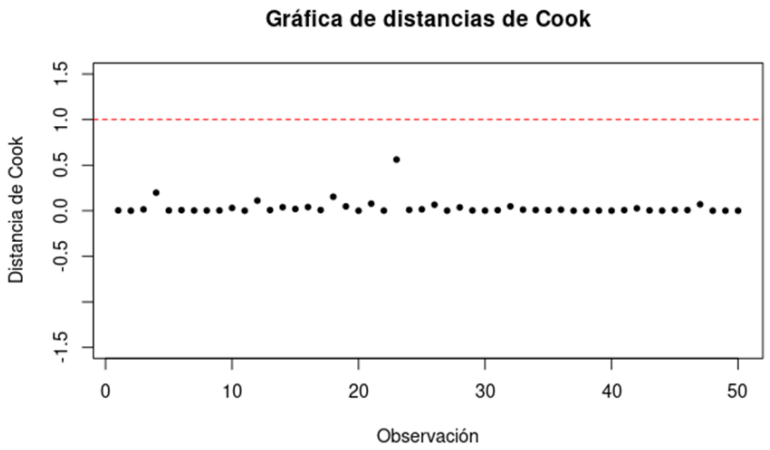
-Puntos de balanceo

Se observa, primero teniendo en cuenta que la línea roja que está discontinua es representada como el valor $2\frac{p}{n} = 0.24$, siendo p el número de parámetros del modelo y n el número de observaciones. Con base a esto, se evidenciaron 4 puntos de balanceo, esto debido a que su valor h_{ii} es mayor que 0.24, lo que indica un alejamiento importante del centro del espacio definido por las variables predictoras. Estos puntos son presentados en la siguiente tabla:

	Residual Estudentizado	Distancia de Cook	Valor h_{ii}	Dffits
13	-0.3126	0.0075	0.3154	-0.2100
21	0.8231	0.0779	0.4083	0.6812
23	-1.6737	0.5612	0.5459	-1.8747
47	1.0040	0.0700	0.2941	0.6481

-Puntos de influencia

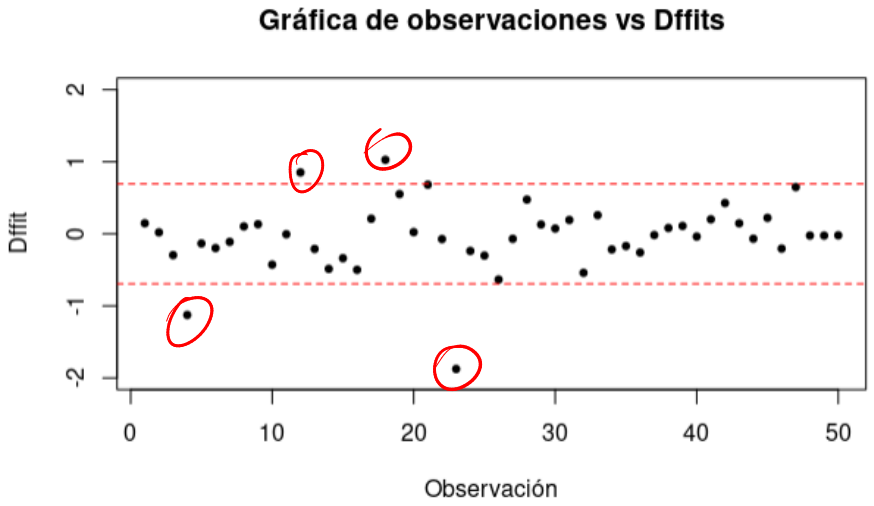
Utilizando el criterio de la distancia de Cook para verificar si existen puntos influenciales dentro de los datos usados para ajustar el modelo, presentamos la siguiente gráfica:



2 pt

Se observa que no hay presencia de ningún punto de influencia porque ninguno de los valores de distancia de Cook asociados a los datos supera al valor de 1.

Continuando con el análisis, ahora usamos el criterio de los Dffits, tenemos la siguiente gráfica:



Donde las barras rojas delimitan los valores $\pm 2\sqrt{\frac{p}{n}} \approx \pm 0.6928$, que nos indica los valores frontera tras los cuales un dato es candidato a ser punto de influencia. Así, analizando dato a dato respecto a este valor, encontramos los siguientes:

	Residual Estudentizado	Distancia de Cook	Valor h_{ii}	Dffits
4	-1.9438	0.1976	0.2388	-1.1258
12	2.1174	0.1110	0.1293	0.8513
18	2.5241	0.1534	0.1263	1.0257
23	-1.6737	0.5212	0.5459	-1.8747

1 pt

Así, podemos ver que estos 4 puntos son candidatos a ser puntos influenciales por el criterio de los Dffits, además; la observación 23 también fue destacada como un posible punto de balanceo.

¿Qué causan? 3 pt

Conclusión

En conclusión la validez del modelo no se cumple, debido a que los errores del modelo no tiene distribución normal, justificado esto por la prueba de Shapiro-Wilk y el criterio gráfico. También se observaron 4 puntos de balanceo y 4 datos de influencia, los cuales afectan la precisión de los parámetros del modelo ajustado. Esto sería muy importante dado el caso que el modelo cumpliera todos los supuestos para hacer inferencia sobre los datos, pero puesto que no es el caso, la proposición de medidas al respecto de estas observaciones se hace innecesaria.

✓

Muy bien :3