Trabajo 1

Estudiantes

Jacobo Melendez Quintero
Equipo # 64

3,7

Docente

Javier Armando Lozano Rodriguez

Asignatura

Estadística II

Sede Medellín
05 de octubre de 2023

Índice

Índice de figuras

Índice de tablas

Universidad Nacional de Colombia. – Sede Medellín.

Escuela de Estadística. – Semestre 2023-2S.

Objetivo: Usar de manera eficiente las herramientas del análisis de regresión para resolver un problema práctico.

Problema: En un estudio a gran escala realizado en EE.UU sobre la eficacia en el control de infecciones hospitalarias se recogió información en 113 hospitales. A su equipo de trabajo le corresponde analizar una muestra aleatoria de 69 hospitales, la base de datos contiene las siguientes columnas (variables):

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

Preguntas a resolver.

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.

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.

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).

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.

Modelo teorico y supuestos 18pt

Punto 1. Modelo de Regresión Lineal Múltiple.

El objetivo del modelo de regresión lineal Múltiple solicitado es explicar la variable respuesta ***(Riesgo de infección)*** en función de las demás variables predictoras en la base de datos, si ajustamos el modelo con la base de datos actual se obtiene:

$$\widehat{Y}_i = -1.0857 + 0.2178X_{i1} + 0.0293X_{i2} + 0.05091X_{i3} + 0.0074X_{i4} + 0.0012X_{i5}, \quad i = 1, 2, \ldots, 69$$

(5 muestran donde ajustaron) 1pt

Significancia individual de los parámetros.

Se desea probar la significancia individual de los parámetros del modelo, es decir se desea probar las siguientes pruebas de hipótesis:

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

Se presenta la tabla de significancia individual con la información de los parámetros estimados, los estadísticos de prueba y los valores-p para cada prueba de hipótesis

Table 2: Tabla de significancia individual

	Estimación	Estadístico	Valor-P
(Intercept)	-1.0857134	-0.7378056	0.4633719
X1	0.2178569	3.0568068	0.0032801
X2	0.0293205	1.0643453	0.2912348
X3	0.0509184	3.4916858	0.0008828
X4	0.0074179	1.1324664	0.2617302
X5	0.0012208	1.8038867	0.0760302

6 pt

De la tabla anterior, de acuerdo con el valor-p y un nivel de significancia $\alpha = 0.05$, se puede concluir que los únicos parámetros significativos del modelo son β_1 y β_3, cuando los demás parámetros están presentes.

Interpretación de los parámetros.

Solo los psrámetros significativos aportan una interpretación relevante al modelo, por lo tanto solo β_1 y β_3 serán interpretados, ya que para que β_0 también sea interpretable el cero debería estar dentro de l rango de las variables, pero no es el caso con los datos actuales.

3 pt

En primer lugar, $\widehat{\beta_1}$ indica que por cada unidad de aumento en la duración promedio de la estadía el riesgo de infección aumenta en 0.2178569 unidades, mientras que por cada unidad que aumente el número promedio de camas en el hospital, el riesgo de infección aumentará en 0.0509184, esto mientras las demás predictoras permanecen fijas.

Significancia de la regresión.

Se desea probar si todos los parámetros del modelo son significativos simultáneamente, es decir:

$$\begin{array}{l} H_0 : \text{Ningún parámetro es significativo} \\ H_1 : \text{Al menos un parámetro es significativo} \end{array}$$

O equivalentemente:

$$\begin{array}{l} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \beta_i \neq 0, \text{para algún i} \in [1, 5] \end{array}$$

Se presenta la tabla ANOVA con la información del estadístico de prueba y el valor-p para la prueba mencionada anteriormente:

Table 3: Tabla ANOVA.

	Sum_of_Squares	DF	Mean_Square	F_Value	P_value
Model	45.8937	5	9.178743	10.5334	2.23107e-07
Error	54.8979	63	0.871395		

Dado que $V_p = 2.23107e^{-07}$ es considerablemente menor que un nivel de significancia $\alpha = 0.05$ se rechaza H_0 y se concluye que el modelo es significativo, en otras palabras, la probabilidad promedio estimada de adquirir infección en el hospital es explicada por al menos una de las variables predictoras.

Coeficiente de determinación R^2

El coeficiente de determinación R^2 es un estadístico que nos indica la proporcion de variabilidad explicada por un modelo, definido como

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{45.8937}{45.8937 + 54.8979} = 0.4553$$

Por lo tanto, el modelo de regresión lineal múltiple planteado explica solo el 45.53% aproximadamente de la variabilidad total presente en el riesgo de infección.

Punto 2. Significancia de un subconjunto del modelo

Se propone probar la significancia del subconjunto de parámetros conformado por los parámetros con los tres valores-p más pequeños, de la ***Tabla 1.*** sabemos que los el subconjunto estará conformado por β_1, β_3 y β_5. Entonces la prueba de hipótesis correspondiente es:

$$H_0: \ \beta_1 = \beta_3 = \beta_5 = 0 \quad \text{vs.} \quad H_1: \ \beta_i \neq 0, \ \text{para algún i} = 1, 3, 5.$$

El estadístico de prueba para esta prueba de hipótesis es:

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

Para calcular el $\text{SSE}(\beta_0, \beta_2, \beta_4)$ tomemos la información de la tabla de todas las regresiones posibles, las dos filas de interés son:

Table 4: Muestra de todas las regresiones posibles.

	GL	R^2	R^2_{adj}	SSE	Cp	Variables
14	2	0.171	0.146	83.537	32.866	X2 X4
31	5	0.455	0.412	54.898	6.000	X1 X2 X3 X4 X5

De esta forma el estadístico de prueba es :

$$F_0 = \frac{(83.537 - 54.898)/3}{0.871395} = 10.95523$$

Como $F_0 = 10.95523 > f_{0.05,3,63} = 2.7505411$, entonces se rechaza H_0 y se concluye que el conjunto de predictoras es significativo.

Prueba de hipótesis lineal general.

Es de interés en el estudio comparar el efecto de algunas variables sobre la respuesta, las directrices desean saber si los efectos de el número promedio de camas y enfermeras tienen el mismo efecto y si la rutina de cultivos presenta diferencias al censo promedio diario, si vemos el juego de hipótesis de forma matricial tenemos

$$H_o : \begin{cases} \beta_2 - \beta_4 = 0 \\ \beta_3 - \beta_5 = 0 \end{cases}$$

Viéndolo como una prueba de hipótesis lineal general se tendría:

$$H_0 = \mathbf{L}\beta = 0 \quad \Longrightarrow \quad H_0 = \begin{bmatrix} 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Así el modelo reducido es :

$$Y = \beta_0 + \beta_2(X_2 + X_4) + \beta_3(X_3 + X_5) + \varepsilon$$

Para esta prueba en particular, el estadístico es de la forma:

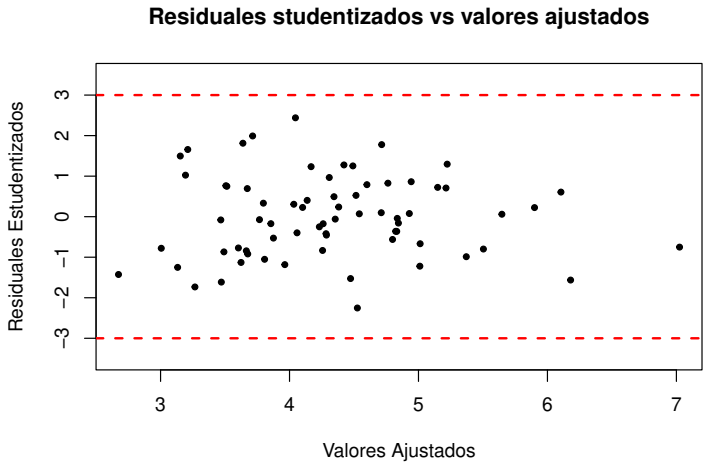
$$F_0 = \frac{\frac{SSE(RM) - SSE(FM)}{GL_{RM} - GL_{FM}}}{MSE_{FM}}$$

Punto 4.1 Datos atípicos, influenciales y de balanceo.

Para el estudio de estos datos, se tienen en cuenta los métodos y restricciones vistas en clase.

Datos Atípicos.

Como diagnóstico para identificar datos atípicos tenemos como parámetro los residuales estudentizados como se ve en la sigueinte gráfica

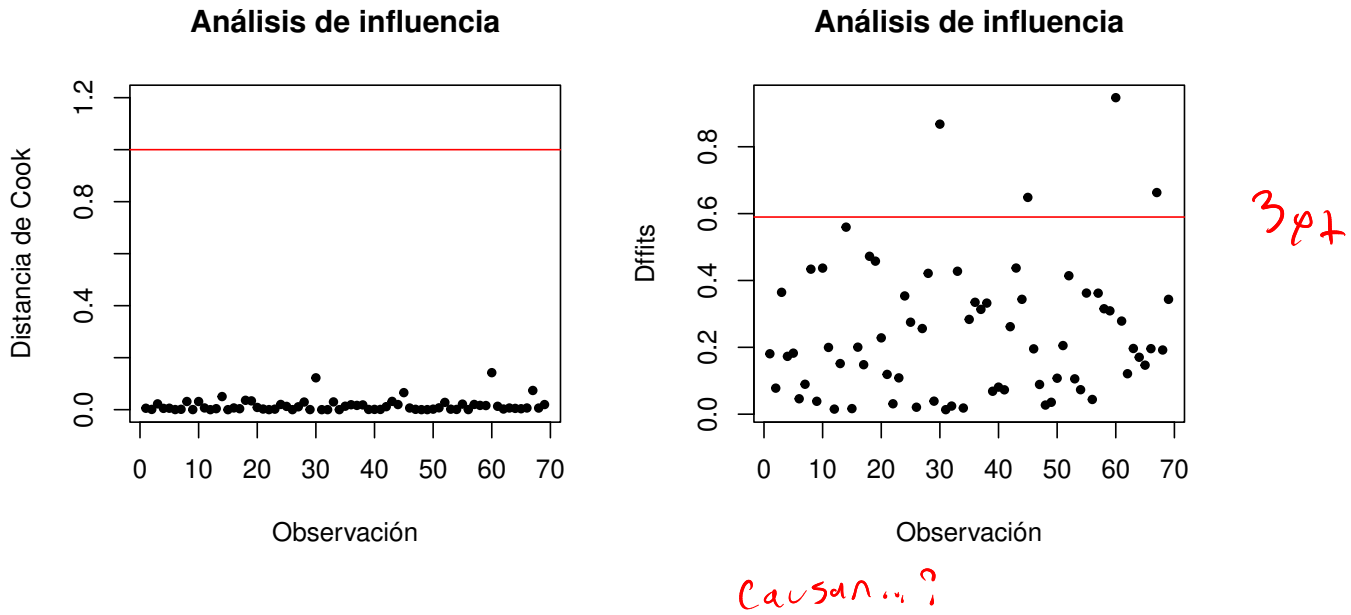


3 pt

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

Observaciones Influenciales.

Se puden detectar observaciones influenciales por varios métodos, se implementa el criterio de la distancia de Cook ($D_i > 1$) y el criterio del diagnóstico DFFITS ($|\mathbf{DFFITS}_i| > 2\sqrt{\frac{6}{69}} = 0.5897678$):



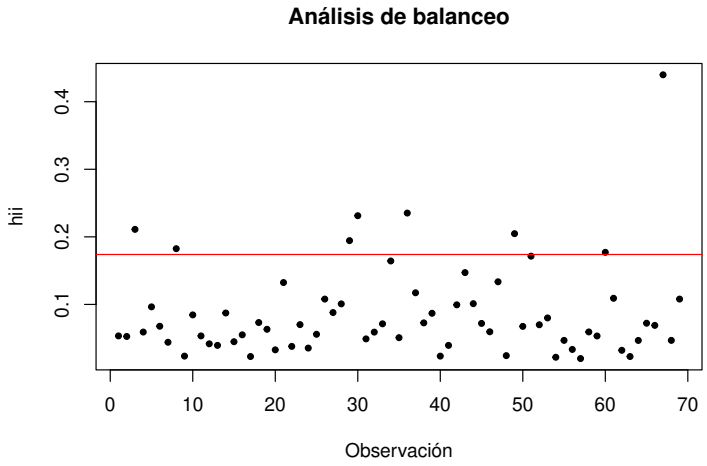
3 pt

Causan...?

Por el criterio de la distancia de Cook ($D_i > 1$) no se identifican datos influenciales, pero se tiene que por Dffits las observaciones 30, 45, 60 y 67 son influenciales.

Puntos de balanceo.

Los puntos de balanceo son detectados mediante el análisis de los elementos de la diagonal principal de la matriz H, de acuerdo con el Diagnóstico DFBETAS ($h_{ii} > 2\left(\frac{6}{69}\right) = 0.173913$)



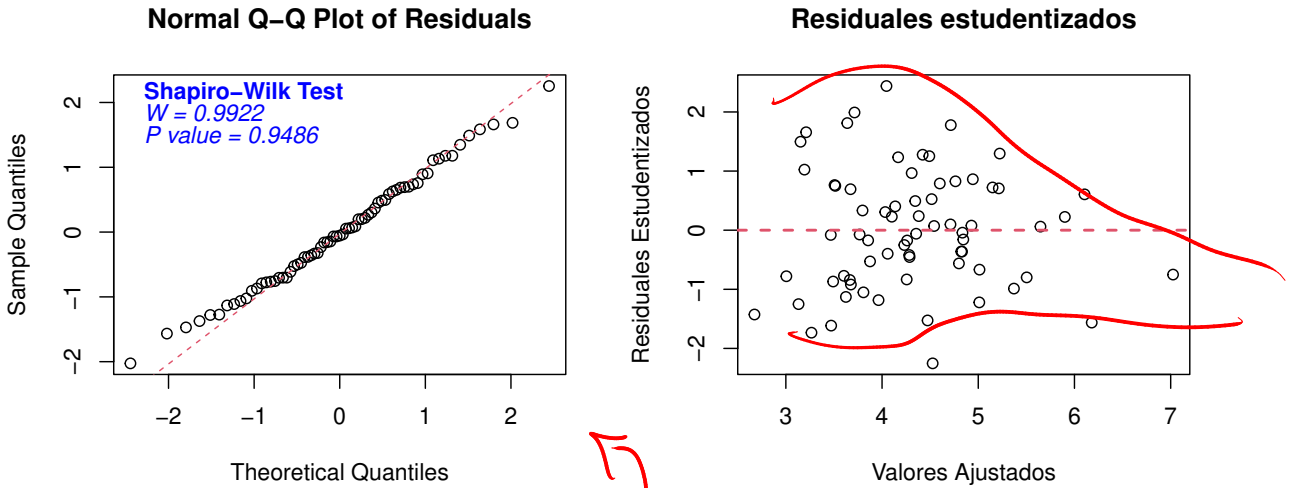
Según este criterio las observaciones 3, 8, 29, 30, 36, 49, 60 y 67 se consideran puntos de balanceo.

4.2 Supuestos del modelo.

Dentro de los supuestos del modelo, se desea validar el supuesto de normalidad y varianza constante. Para la normalidad se realizará un análisis numérico apoyados en la prueba de shapiro-wilk para normalidad donde se prueba:

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

Adicionalmente, para ambos supuestos se realizará una prueba gráfica:



En primer lugar, para la normalidad se puede ver que como $V_p > \alpha = 0.05$ se rechaza H_0 y se concluye que el supuesto de normalidad se cumple. Para el supuesto de varianza constante veamos el gráfico de residuales estudentizados vs valores ajustados en donde no se evidencian patrones de crecimiento, decrecimiento o agrupaciones en los residuales, por lo tanto se concluye que el supuesto de varianza constante también se

cumple. Como ambos supuestos se cumplen se puede decir que el modelo es apto para realizar estimaciones sobre la variable respuesta, no obstante, como se vio anteriormente el coeficiennte de determinación R^2 es muy bajo y hay precensia de observaciones extremas que podrían influir en la calidad de las estimaciones.

Válido o no?

2pt