3,3

Trabajo 1

Equipo 28

Integrantes:
Lizeth Tatiana Oquendo Romero
Diana Sofía Ruiz Arteaga
Liliana Marcela Ruiz Piñeros
Daniel Vélez Vélez

Docente:
Francisco Javier Rodríguez Cortés

Estadística II

Un modelo RLM sobre la eficacia en el control de infecciones hospitalarias en EE.UU

UNIVERSIDAD NACIONAL DE COLOMBIA

Sede Medellín
29 de marzo de 2023

Índice

Índice de figuras

Índice de cuadros

1. Pregunta 1

15,5 pt

En un estudio sobre la eficacia en el control de infecciones hospitalarias en EE.UU, se recogió información de 113 hospitales. En este caso, se propondrá y analizará un modelo de Regresión Lineal Múltiple (RLM) en el que se pretende encontrar la relación que existe entre 5 variables regresoras y el riesgo de infección, la cual será la variable respuesta, con base en una muestra aleatoria de 65 hospitales.

Caption! Por qué ese formato de tabla? sólo usen un formato para todo el trabajo

Variable	Descripción
Y: Riesgo de infección	Probabilidad promedio estimada de adquirir infección en el hospital (en porcentaje).
X_1: **Duración de la estadía**	Duración promedio de la estadía de todos los pacientes en el hospital (en días).
X_2: **Rutina de cultivos**	Razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100.
X_3: **Número de camas**	Número promedio de camas en el hospital durante el periodo del estudio.
X_4: **Censo promedio diario**	Número promedio de pacientes en el hospital por día durante el periodo del estudio.
X_5: **Número de enfermeras**	Número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio.

El modelo RLM propuesto para analizar esta relación, en la cual hay 5 variables regresoras es:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$$

que tiene como supuestos lo siguiente:

$$\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \quad \forall_{i=} \ 1, 2, ..., 65$$

También se puede especificar el modelo en términos matriciales, así:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \quad \text{con} \quad \varepsilon \sim N_n(\underline{0}, \sigma^2 I)_{n\times n}$$

1.1. Modelo de regresión

2 pt

Al ajustar el modelo, se obtienen los siguientes coeficientes:

Cuadro 1: Tabla de valores coeficientes del modelo

→ formato de tabla?

	Valor del parámetro
$\beta 0$	1.104228166
$\beta 1$	0.183963744
$\beta 2$	-0.007392722
$\beta 3$	0.038605536
$\beta 4$	0.012098791
$\beta 5$	0.001189688

Por lo tanto, el modelo de regresión ajustado es:

Supuestos y error no va en ec. ajustada

$$\hat{Y}_i = 1.104228166 + 0.183963744\, x_{1i} - 0.007392722\, x_{2i} + 0.038605536\, x_{3i} + 0.012098791\, x_{4i} + 0.001189688\, x_{5i} + \cancel{\varepsilon_i}$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\quad 1 \leqslant i \leqslant 65$$

1.2. Significancia de la regresión

3pt

Ahora que el modelo se ha propuesto, se realizará una prueba de significancia general de la regresión con el objetivo de identificar si las variables predictoras están en capacidad de brindar información sobre la variable respuesta. Para ello, se plantea la siguiente prueba con base en la ANOVA, tabla de análisis de varianza.

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

$$H_0: \boxed{\cancel{\beta_0}} = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \rightarrow \text{No va } \beta_0$$

$$H_a: Algún\ \beta_j \neq 0, \quad \boxed{j = 1,2,\ldots,5} \rightarrow \text{Acá sí no ponen } \beta_0\text{?}$$

En esta prueba, se tiene la hipótesis nula de que todos los parámetros son igual a cero y la hipótesis alterna, que dicta que al menos uno de los parámetros sí es significativo.

Con el objetivo de tener un modelo que esté en capacidad de brindar información sobre el riesgo de infección, es de interés que se rechace la hipótesis nula a favor de la hipótesis alterna, indicando que al menos uno de los parámetros del modelo puede establecer una relación con la variable respuesta.

Para que ello ocurra, se debe establecer el estadístico de prueba, así como la condición de rechazo de la hipótesis nula:

$$\text{Estadístico de prueba} \rightarrow F_0 = \frac{MSR}{MSE} \sim F_{5,59}$$

$$\text{Se rechaza } H_0 \text{ si} \rightarrow F_0 > f_{\alpha,\boxed{k,n-p}}, \text{ donde se asume } \alpha = 0.05, \text{ o si } V_P < \alpha$$

reemplazar

Cuadro 2: Tabla ANOVA para el modelo

	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Valor F	Valor P
Modelo	48.2966	5	9.659321	11.4625	9.39699 ×10^ (-8)
Error	49.7188	59	0.842691		

✓

Es un lenguaje

A partir de la tabla ANOVA que se obtiene ~~a partir del software~~ R, se observa un valor P = $9.39699 \times 10^{-8} < 0.05$, por lo que se rechaza la hipótesis nula en la que $\beta_j = 0$ con $1 \leqslant j \leqslant 5$, aceptando la hipótesis alternativa en la que algún $\beta_j \neq 0$, por lo tanto, la regresión es significativa. ✓

1.3. Significancia de los parámetros

6 pt

Luego de comprobar que la regresión es significativa, es de interés evaluar cuáles parámetros aportan información al modelo y cuáles no son significativos. Para esto, se establece la prueba de hipótesis de los parámetros individuales como sigue:

$$\boldsymbol{H_0}: \boldsymbol{\beta_j} = \mathbf{0} \quad , \quad \boxed{\boldsymbol{j} = \mathbf{1, 2, \ldots, 5.}}$$
$$\boldsymbol{H_a}: \boldsymbol{\beta_j} \neq \mathbf{0}$$

¿y β_0?

Rechazar la hipótesis nula en la prueba de cada parámetro implica que dicho parámetro β_j es significativo. Para definir el resultado de esta prueba, se emplea el siguiente estadístico de prueba, con su respectiva condición de rechazo:

$$\text{Estadístico de prueba} \rightarrow \quad \boldsymbol{T_j} = \frac{\widehat{\boldsymbol{\beta_j}}}{\boldsymbol{se}(\widehat{\boldsymbol{\beta_j}})}$$

$$\text{Se rechaza } H_0 \text{ si} \rightarrow \quad |\boldsymbol{T_j}| > \boldsymbol{t}_{\frac{\alpha}{2},\boldsymbol{n-p}}, \text{ donde se asume } \propto = 0.05, \text{ o si } \boldsymbol{V_P} < \boldsymbol{\alpha}$$

A partir de la tabla resumen de los coeficientes del modelo, se obtienen los valores estimados de los parámetros, así como el valor p, para evaluar la significancia de cada parámetro individual:

Cuadro 3: Resumen de los coeficientes

Formato de tabla

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	1.104228166	1.4936374011	0.7392880	0.462663419
B_1	0.183963744	0.0710759169	2.5882711	0.012126365
B_2	-0.007392722	0.0279555735	-0.2644454	0.792358356
B_3	0.038605536	0.0122034281	3.1634992	0.002463915
B_4	0.012098791	0.0067991182	1.7794648	0.080314197
B_5	0.001189688	0.0006390146	1.8617542	0.067620820

Al contrastar, se tiene que los valores P de los parámetros β_1 **y** β_3 son los únicos que cumplen la condición de significancia menor a $\alpha = 0.05$:

$$0.012126365 < 0.05 \quad \text{y} \quad 0.002463915 < 0.05$$

Por lo tanto, los parámetros que tienen significancia en la regresión son β_1 y β_3, es decir, que existe una relación positiva entre el riesgo de contraer una infección hospitalaria en EE.UU y la duración de la estadía en un hospital y el número de camas en el hospital.

1.4. Interpretación de los parámetros 1,5 pt

Cuadro 4: Valores mínimos y máximos de cada variable

	Y	x_1	x_2	x_3	x_4	x_5
min	1.3	6.70	38.8	1.9	39.6	70
max	7.8	19.56	65.9	60.5	133.5	835

El parámetro $\hat{\beta}_0$ no tiene interpretación pues ninguna variable predictora incluye al "0" en su rango de valores, lo cual se concluye a partir del cuadro 4.

18,39% ó 0,183963744

El parámetro $\hat{\beta}_1$ indica un aumento promedio de ~~0.183963744%~~ en la probabilidad de infectarse en hospitales de EE.UU, cuando se da un aumento de un día en la estadía de todos los pacientes en el hospital, dejando las demás variables regresoras constantes.

; sólo el de estudio

3,8605536%

El parámetro $\hat{\beta}_3$ indica un aumento promedio de ~~0.038605536~~ % en la probabilidad de infectarse en hospitales de EE.UU, cuando se da un aumento de una cama en los hospitales, dejando las demás variables regresoras constantes.

El resto de los parámetros no tienen interpretación pues no son significativos para la regresión, según la prueba de hipótesis anterior.

1.5. Coeficiente de determinación múltiple R^2 3pt

El modelo tiene un coeficiente de determinación múltiple $R^2 = 0.4927$, lo que significa que aproximadamente el 49.*27* % de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

Ahora, se procederá a calcular el coeficiente R^2 y el coeficiente de determinación múltiple R^2 *ajustado*, haciendo uso de los valores arrojados por la tabla ANOVA del software R:

$$R^2 = \frac{SSR}{SST} = \frac{48.2966}{48.2966 + 49.7188} = 0.49274501761$$

$$R^2_{adj} = 1 - \frac{(n-1)\,MSE}{SST} = 1 - \frac{64 \times 0.842691}{48.2966 + 49.7188} = 0.4497576503$$

El R^2_{adj} brinda una medida de bondad de ajuste sobre parámetros no significativos, es decir, penaliza la regresión cuando se añaden variables que no son significativas para el modelo. Al analizar la brecha entre el R^2 y el R^2_{adj}, se concluye que dicha diferencia ocurre por la presencia de estas variables "vacías" en el modelo ajustado y, por lo tanto, que se podría prescindir de aquellas que no sean significativas.

2. Pregunta 2

4,5 pt

2.1. Planteamiento pruebas de hipótesis y modelo reducido

Como solicitado en el punto 2, se procederá a evaluar la significancia simultánea del subconjunto de tres variables con los valores p más grandes. Para la muestra aleatoria de 65 hospitales en la que se basa este análisis, las tres variables con los valores p más grandes son $x_2,\ x_4\ y\ x_5$

La prueba de hipótesis asociada a la validación de la significancia de este subconjunto es como sigue:

$$H_0: \beta_2 = \beta_4 = \beta_5 = 0$$

$$H_a: Algún\ \beta_j \neq 0, \quad j = 2,4,5.$$

Cuadro 5: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	49.7188	X1 X2 X3 X4 X5
Modelo reducido	54.264	X1 X3

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon;\ \varepsilon_i \overset{iid}{\sim} N\,(0, \sigma^2);\ 1 \leqslant i \leqslant 65$$

2.2. Estadístico de prueba y conclusión

Con el objetivo de evaluar la significancia o no de esta prueba de hipótesis, se establecerá el estadístico de prueba, así como la condición de rechazo. La hipótesis consiste en probar la significancia de los parámetros indicados, dada la presencia de los demás parámetros en la regresión. Esto, a través de las sumas de cuadrados extras, es decir, de la reducción marginal en el SSE, definida mediante la diferencia entre el SSE del modelo reducido y el SSE del modelo completo. A su vez, esta suma de cuadrados tiene tantos grados de libertad como la cantidad de β_j del subconjunto de parámetros a probar, que, para este caso, al estar probando 3 β_j, se tienen 3 g.l. Este procedimiento está definido a continuación:

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

$F_0 \sim F_{3,59}$

3pt

$$F_0 = \frac{\frac{[54.264 - 49.719]}{3}}{0.842691} = 1.797891121$$

La hipótesis nula se rechazará si $F_0 > f_{\propto,k,n-p}$, es decir, si $1.797891121 > f_{0.05,5,59}$

Reemplazar

A partir del cálculo en R: qf(0.05,5,59,lower.tail=FALSE)
~~[1]~~ 2.370977

1,5 pt

No pongan salidas

¡¿Qué?!

Por lo tanto, $\boxed{1.797891121 > 2.370977}$, es decir, se acepta la hipótesis nula (al no cumplirse la condición de rechazo), y se puede prescindir de estas 3 variables en el modelo ya que, al no ser significativas, ninguna de las variables probadas está aportando información al modelo.

Lo anterior tiene sentido si se recuerda que el subconjunto probado estaba compuesto por las variables con el valor p más alto, de las cuales ninguno estaba por debajo a 0.05 (la significancia mínima).

Entonces, se puede afirmar que la rutina de cultivos, el censo promedio diario y el número de enfermeras no son variables que aportan información con el objetivo de determinar qué variables se relacionan con el riesgo de infección en ~~los hospitales de EE.UU~~.

No leyeron bien el ejercicio

3. Pregunta 3

4 pt

3.1. Prueba de hipótesis y prueba de hipótesis matricial

En este caso, se plantea la pregunta de si la duración promedio de la estadía de todos los pacientes en el hospital (en días), es igual al número promedio de enfermeras, (equivalentes a tiempo completo), a la vez que el número promedio de camas en el hospital es igual al número promedio de pacientes en el hospital:

Mal formuladas

$$H_0: \beta_1 - \beta_5 = 0 \quad \text{a la vez que} \quad \beta_3 - \beta_4 = 0$$
$$H_a: \beta_1 - \beta_5 \neq 0 \quad \text{o} \quad \beta_3 - \beta_4 \neq 0$$

De forma matricial, la hipótesis nula es la siguiente:

$$H_0: \quad L\beta = 0 \quad \rightarrow \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{matrix} 0 \\ 0 \end{matrix}$$

1,5 pt

Quién es específicamente L

Note que la matriz $\boldsymbol{L}$ tiene $r = 2$ filas linealmente independientes (observe que una fila no puede escribirse como un múltiplo escalar de la otra).

El modelo reducido en este caso es:

$$\begin{aligned} Y &= \beta_0 + \beta_1(X_1+X_5) + \beta_2 X_2 + \beta_3(X_3+X_4) + \boxed{\varepsilon} \rightarrow \text{y los supuestos} \\ &= \beta_0 + \beta_1 X_{1,5} + \beta_2 X_2 + \beta_3 X_{3,4} + \varepsilon \end{aligned}$$

0,5 pt

donde $X_{1,5} = X_1 + X_5$, y $X_{3,4} = X_3 + X_4$

3.2. Estadístico de prueba

2 pt

Finalmente, la expresión para el estadístico de prueba es:

$$F_0 = \frac{MSH}{MSE} = \frac{SSH/2}{MSE} = \boxed{\frac{[SSE(RM)* - SSE(FM)]/2}{MSE}} = \frac{[SSE(RM)* - 49.719]/2}{0.842691} \sim F_{2,59}$$

Se rechaza la hipótesis nula si $F_0 > f_{\propto,r,n-p}$, por lo tanto, solo resta establecer el valor en (*) que es SSE(RM), el cual no se puede obtener de la tabla de todas las regresiones posibles, ya que ésta no admite sumas de variables entre sus opciones.

4. Pregunta 4

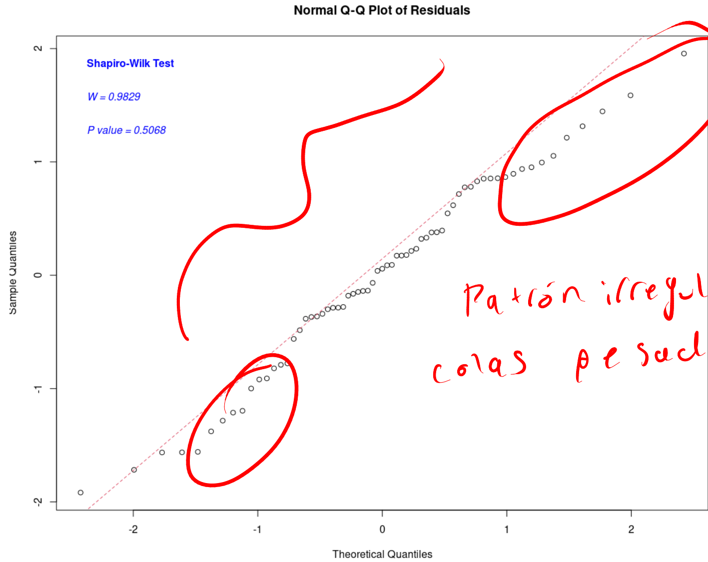
9 pt

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

4.1. Supuestos del modelo

Es bien sabido que un modelo RLM debe cumplir ciertos supuestos en los errores del mismo: ser independientes unos de otros, tener media cero, distribuirse normales y tener varianza constante.

En cuanto a los supuestos de independencia y media cero, serán asumidos por teoría de acuerdo con la metodología empleada en la asignatura.



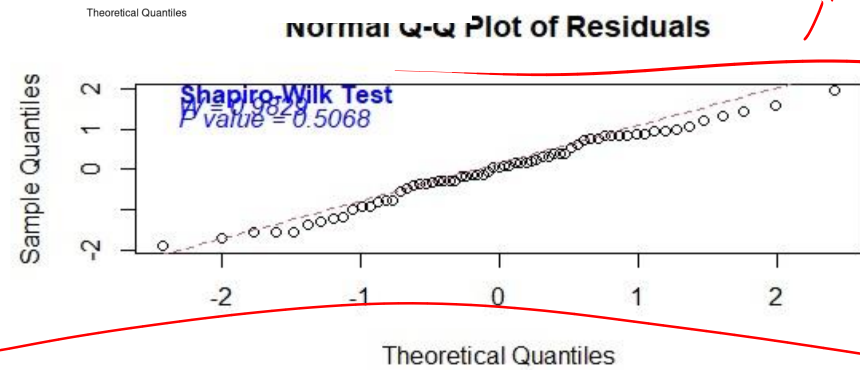
Patrón irregular
colas pesadas

muy distinto

1,5 pt

ales

Hagan bien ese copy-paste.



muy mal análisis gráfico, tal vez debido a pegar mal la gráfica

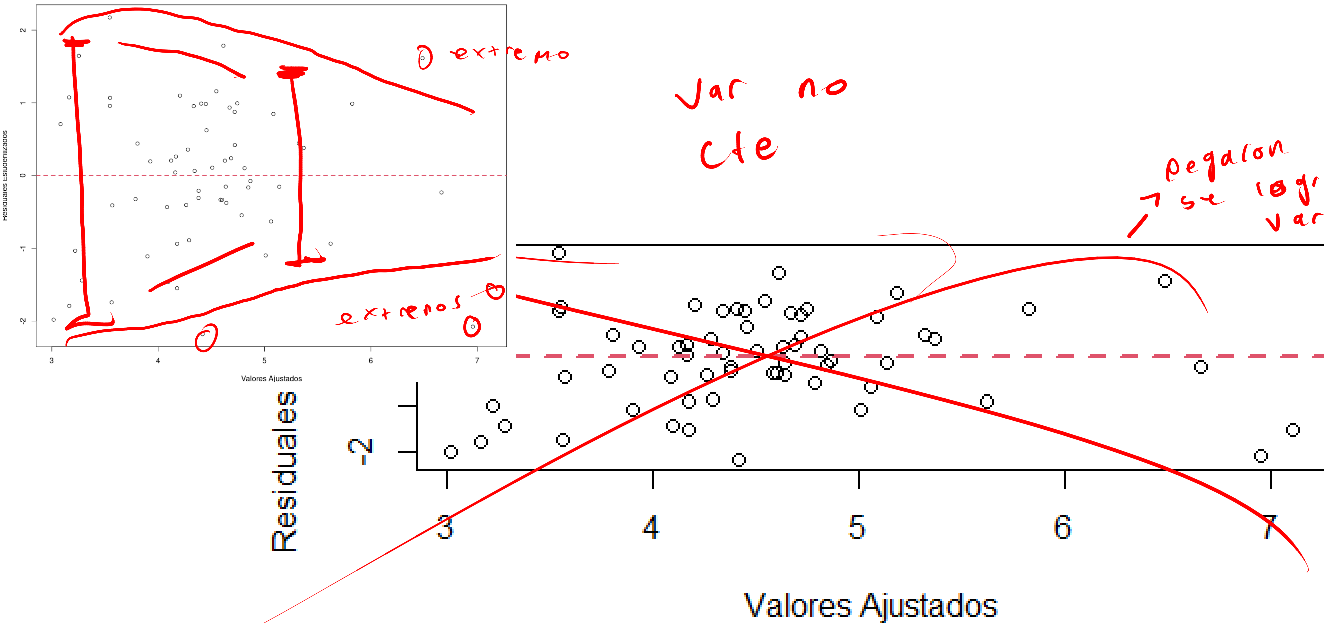
La gráfica Q-Q plot expuesta permite conocer que los errores se distribuyen normales, pues los puntos se mantienen cerca de la línea de normalidad. Esto puede ser corroborado por la prueba ~~de hipótesis~~ Shapiro Wilk, que tiene el siguiente juego de hipótesis:

$$H_0\text{: los errores se distribuyen normales}$$
$$H_a\text{: los errores no se distribuyen normales}$$

Se rechaza la hipótesis nula si $V_P < \alpha$. Para este caso, se tiene que $0.5068 < 0.05$, por lo tanto, no se cumple la condición de rechazo y se acepta la hipótesis nula, que los errores efectivamente se distribuyen normales.

4.1.2. Varianza constante

0 pt



En el gráfico de residuales estudentizados vs valores ajustados, observamos que no hay una secuencia que nos diga que la varianza no es constante, por consiguiente, al no tener suficiente información para descartar este supuesto lo aceptamos como verdadero. Lo anterior, se puede corroborar al observar que, a lo largo de todo el rango, la distancia de los residuales respecto a la media es aproximadamente igual. X

4.2. Verificación de las observaciones

4.2.1. Datos atípicos

1,5 pt ¿Dónde está? Deben colocar todo lo que mencionan

Se considera que una observación es atípica cuando su residual estudentizado ri, es tal que: |ri| > 3. Por lo que, de acuerdo con la tabla de diagnóstico que arrojó el programa, se puede ver que, en todos los datos, el valor absoluto del residual estudentizado da menor que tres, lo que indica que no hay ningún dato atípico.

A ver la gráfica? siempre es bueno mostrar.

4.2.2. Puntos de balanceo

2 pt

Para determinar si hay puntos de balanceo, los cuales son observaciones que están alejados de la mayoría de las predictoras, se analizan los elementos de la diagonal principal de la matriz H (h_{ii}), si estos cumplen que $h_{ii} > 2\frac{P}{n}$, ó $h_{ii} > 0.184615$ son puntos de balanceo. En la tabla se pudo ver que los que obtuvieron un $h_{ii} > 0.184615$ fueron los datos 4, 9, 28, 29, 32, 51, por lo tanto, son puntos de balanceo. como se puede ver en la siguiente tabla:

¿Cuál? Al menos una gráfica.

Cuadro 6: Resumen tabla de todas las regresiones

DATOS	h_{ii}
4	0.2040
9	0.2716
28	0.3321
29	0.4096

y esta partición

32	0.2126
51	0.2681

¿Qué causan?

4.2.3. Puntos influenciales

Para determinar los puntos influenciales, los cuales son observaciones que halan al modelo en su dirección y que tienen un impacto considerable en los coeficientes de regresión ajustados, se utilizan dos medidas, las cuales son la distancia de Cook y el diagnóstico DFFITS.

1pt ¿cuál? ¿cómo va a saber el lector que lo que dicen es cierto?

- Se dice que la observación i será influencial si Di > 1.
- Una observación será influencial si |DFFITSi| > $\boxed{2}$

Para la distancia de Cook se verifica que un valor es influencial si $D_i > 1$; y en la tabla no hay ningún valor que sobrepase a uno. Ahora bien, se realiza el diagnóstico de DFFITS, el cual dice que un dato es influencial si $|DFFITS| > 2\sqrt{\frac{p}{n}}$, ó |DFFITS| > 0.6076, para lo cual se encontraron que los datos 4, 7, 9, 32, 60 son valores influenciales, como se puede ver en la siguiente tabla.

DATOS	|DFFITS|
4	0.7467
7	0.7083
9	0.9563
32	0.8505
60	0.6713

1pt

¿Qué causan?

4.3. Conclusión

2pt

En conclusión, se puede decir que este modelo es válido para hacer ~~un análisis de correlación~~ entre el riesgo de infección en ~~hospitales de EE.UU~~ y las variables regresoras significativas, en tanto que cumple con los supuestos de los errores, como fue mostrado previamente en este punto.

Al menos fueron congruentes con el error