UNIVERSIDAD
NACIONAL
DE COLOMBIA

3,7

TRABAJO CORTO 01

ESTADÍSTICA II

ESCUELA DE ESTADÍSTICA - FACULTAD DE CIENCIAS

EQUIPO 45

MALLERLY GALLEGO MORALES
C.C. 1.000.922.231

ALEJANDRA SERNA FLOREZ
C.C. 1.000.547.259

MIYANIS MANUELA LONDOÑO VÉLEZ
C.C. 1.000.086.192

VALENTINA SIERRA DURANGO
C.C. 1.000.292.082

MEDELLÍN

OCTUBRE DE 2023

Contents

Pregunta 1

18 pt

Teniendo en cuenta la base de datos asignada para el **equipo 45**, la cual cuenta con una muestra de tamaño $n = 54$ hospitales, se desea predecir el *riesgo promedio de infección en los hospitales de EE.UU*. Para ello, la base de datos brinda información de las siguientes variables para cada una de las muestras dadas (hospitales):

- **Y** (Variable respuesta): **Riesgo de infección** := probabilidad promedio estimada de adquirir infección en el hospital *(en porcentaje)*
- $\mathbf{X_1}$: **Duración de la estadía** := duración promedio de la estadía de todos los pacientes en el hospital *(en días)*.
- $\mathbf{X_2}$: **Rutina de cultivos** := razón del número de cultivos realizados en pacientes sin síntomas de infección hospitalaria, por cada 100.
- $\mathbf{X_3}$: **Número de camas** := número promedio de camas en el hospital durante el periodo del estudio.
- $\mathbf{X_4}$: **Censo promedio diario** := número promedio de pacientes en el hospital por día durante el periodo del estudio.
- $\mathbf{X_5}$: **Número de enfermeras** := número promedio de enfermeras, equivalentes a tiempo completo, durante el periodo del estudio.

En ese orden de ideas, el primer modelo propuesto, está dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i; \quad \text{con: } \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \quad 1 \leqslant i \leqslant 54$$

Modelo de regresión

Así, al ajustar el modelo, los coeficientes obtenidos son:

Table 1: Tabla de valores coeficientes del modelo

	Valor del parámetro
β_0	1.6795
β_1	0.1842
β_2	-0.0075
β_3	0.0449
β_4	0.0030
β_5	0.0011

2 pt

No va en ec. ajustada

Por lo tanto, el *modelo de regresión lineal múltiple ajustado*, es:

$$\hat{Y}_i = 1.6795 + 0.1842 X_{1i} - 0.0075 X_{2i} + 0.0449 X_{3i} + 0.003 X_{4i} + 0.0011 X_{5i} + \varepsilon_i; \quad \text{con: } \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \quad 1 \leqslant i \leqslant 54$$

Significancia de la regresión

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

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

Cuyo estadístico de prueba es:

$$F_0 = \frac{MST}{MSE} \overset{H_0}{\sim} f_{5,48} \tag{1}$$

Ahora, se presenta la tabla Anova:

Table 2: Tabla ANOVA para el modelo

	Sumas de cuadrados	g.l.	Cuadrado medio	F_0	P-valor
Regresión	30.2803	5	6.056070	6.84355	6.84273e-05
Error	42.4767	48	0.884931		

En la anterior Anova, se observa un *valor P* igual a $6.84273e-05$, el cual es muy pequeño y claramente menor al nivel de significancia con el que se trabaja en el ejercicio ($\alpha = 0.05$); por lo tanto, se rechaza la hipótesis nula H_0 para la no significancia de los parámetros y así, se concluye que al menos uno de los parámetros del modelo propuesto es significativo y por ende, el modelo también lo es.

Significancia de los parámetros

En el siguiente cuadro se presenta información de los parámetros, con la que se pretende determinar cuáles de ellos son significativos.

Table 3: Resumen de los coeficientes

	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	T_{0j}	P-valor
β_0	1.6795	1.7226	0.9750	0.3345
β_1	0.1842	0.0777	2.3691	0.0219
β_2	-0.0075	0.0302	-0.2470	0.8060
β_3	0.0449	0.0154	2.9250	0.0052
β_4	0.0030	0.0076	0.3962	0.6937
β_5	0.0011	0.0008	1.4088	0.1653

Los P-valores presentes en la tabla permiten concluir que con un nivel de significancia 5%, los parámetros β_1 y β_3 son significativos, pues sus P-valores son *menores* a $\alpha = 0.05$.

Interpretación de los parámetros

$\hat{\beta}_{\mathbf{0}} = \mathbf{1.6795}$, en este caso, el intercepto no puede ser interpretado puesto que el vector $\mathbf{0}$ no se encuentra dentro del rango experimental.

$\hat{\beta}_{\mathbf{0}} = \mathbf{0.1842}$, indica que por cada unidad que aumenta la duración de la estadía, la probabilidad de riesgo de infección aumenta 0.1842 unidades, cuando las demás variables predictoras permanecen fijas.

$\hat{\beta}_{\mathbf{0}} = \mathbf{0.0449}$, indica que por cada unidad que aumenta el número de camas, la probabilidad de riesgo de infección aumenta 0.0449 unidades, cuando las demás variables predictoras permanecen fijas.

Coeficiente de determinación múltiple R^2

El coeficiente de determinación se calcula usando $R^2 = \frac{SSR}{SST}$, lo cual puede obtenerse de la tabla ANOVA $R^2 = \frac{42.4767}{30.2803+42.4767}$ y cuyo resultado es 0.583816. De tal forma que, modelo tiene un coeficiente de determinación múltiple $R^2 = 0.583816$, lo que significa que aproximadamente el 58.3% de la variabilidad total observada en la respuesta es explicada por el modelo de regresión propuesto en el presente informe.

De igual manera, es posible calcular R^2_{adj} como una medida de bondad de ajuste, donde $R^2_{adj} = 1 - \frac{(n-1)MSE}{SST} = 1 - \frac{(54-1)0.884931}{30.2803+42.4767} = 0.35537$. Así, se alcanza a notar que $R^2_{adj} < R^2$, por lo que se puede decir que hay variables redundantes o innecesarias presentes en el modelo que deben ser removidas para que dichos valores sean más próximos.

Pregunta 2

Planteamiento pruebas de hipótesis y modelo reducido

Los parámetros cuyos valores P fueron más bajos en el modelo fueron $X_1,\ X_3$, debido a ello, a través de la tabla de todas las regresiones posibles se pretende hacer la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_3 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \text{ distinto de } 0 \text{ para } j = 1, 3 \end{cases}$$

Table 4: Resumen tabla de todas las regresiones

	SSE	Covariables en el modelo
Modelo completo	42.477	X1 X2 X3 X4 X5
Modelo reducido	44.726	X1 X3

Lugeo, un modelo reducido para la prueba de significancia del subconjunto, está dado:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i; \quad \text{con} \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2); \quad 1 \leq i \leq 54$$

Estadístico de prueba y conclusión

Se construye el estadístico de prueba como:

$$\begin{aligned} F_0 &= \frac{(SSE(\beta_0, \beta_1, \beta_3) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5))/3}{MSE(\beta_0, \ \cdots, \ \beta_5)} \overset{H_0}{\sim} f_{3,48} \\ &= \frac{(44.726 - 42.477)/3}{26.776/48} \\ &= 0.8471408 \end{aligned} \tag{2}$$

```
## [1] "F0 es:  0.8471408"
```

2 pt

Ahora, comparando a un nivel de significancia $\alpha = 0.05$, el F_0 con $f_{0.95,3,48}$, se puede ver que $F_0 = 0.847 < f_{0.95,1,48} = 2.7981$, es decir, no se rechaza H_0 y se concluye que la eficacia en el control de infecciones hospitalarias depende de las variables "Duración de la estadía" y "Número de camas".

Pregunta 3

4 pt

Prueba de hipótesis y prueba de hipótesis matricial

Se plantea el siguiente juego de hipótesis:

$$\begin{cases} H_0 : \beta_1 = \beta_5; \ \beta_2 = \beta_4 \\ H_1 : \beta_1 \neq \beta_5; \text{ó } \beta_2 \neq \beta_4 \end{cases}$$

Lo cual es equivalente a tener el juego de hipótesis:

$$\begin{cases} H_0 : \beta_1 - \beta_5 = 0; \ \beta_2 - \beta_4 = 0 \\ H_1 : \beta_1 - \beta_5 \neq 0; \text{ó } \beta_2 - \beta_4 \neq 0 \end{cases}$$

Reescribiendo matricialmente:

$$\begin{cases} H_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ H_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Con $\mathbf{L}$ dada por

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

2 pt

El modelo reducido está dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_3 X_{3i} + \varepsilon_i$$

supuestos 0 pt

Donde: $X_{1i}^* = X_{1i} + X_{5i}$ y $X_{2i}^* = X_{2i} + X_{4i}$

Estadístico de prueba

El estadístico de prueba F_0 está dado por:

$$F_0 = \frac{(SSE(MR) - SSE(MF))/r}{MSE(MF)} \overset{H_0}{\sim} f_{0.95,\ 48}$$

Con los datos:

$$F_0 = \frac{(57.052 - 42.477)/2}{0.8849} \overset{H_0}{\sim} f_{0.95,\ 2,\ 48}$$

2 pt

Pregunta 4

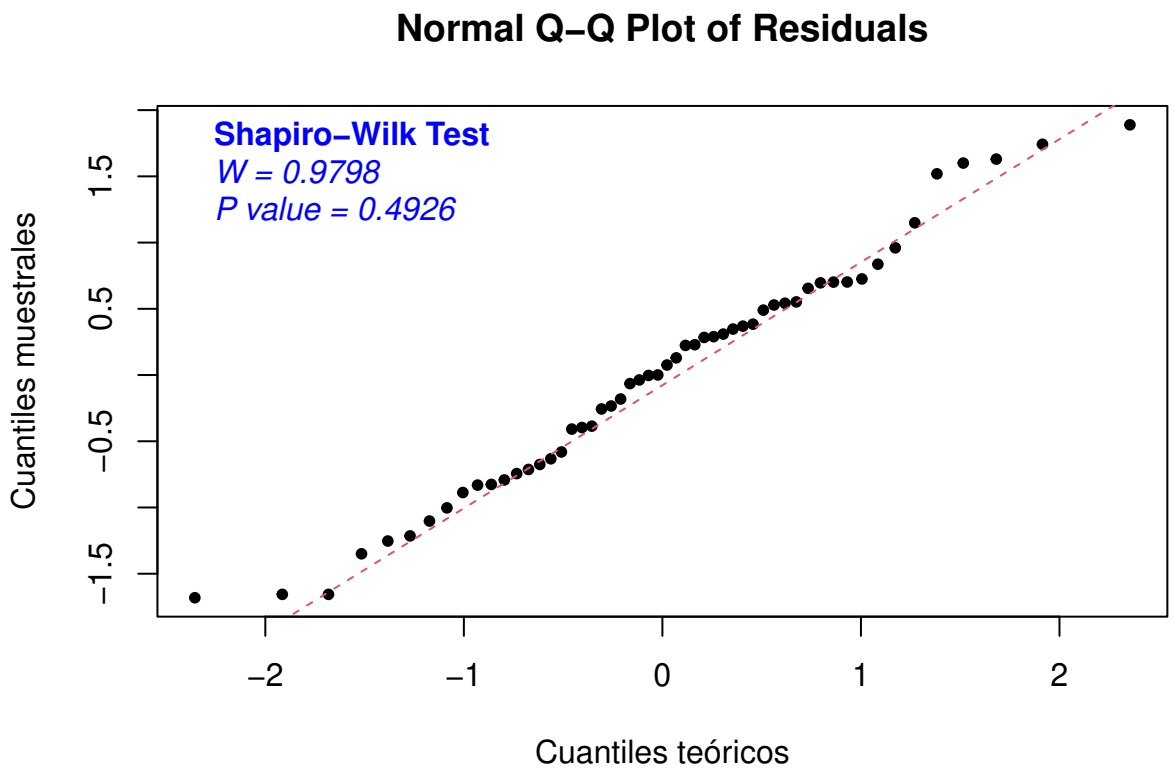
13pt

Supuestos del modelo

Normalidad de los residuales

Para la validación del supuesto de normalidad, se plantea la siguiente prueba de hipótesis que se realizará utilizando el test de *Shapiro-Wilk* y acompañada de un gráfico cuantil-cuantil:

$$\begin{cases} \mathrm{H}_0 : \varepsilon_i \sim \text{Normal} \\ \mathrm{H}_1 : \varepsilon_i \nsim \text{Normal} \end{cases}$$



2pt

Figure 1: Gráfico cuantil-cuantil y normalidad de residuales

Nótese que se obtuvo, con el Test de Shapiro-Wilk un valor P igual a 0.4926 y considerando que se está trabajando con un nivel de significancia $\alpha = 0.05$, no se rechaza la hipótesis nula pues el valor P es evidentemente mayor; es decir, que hay evidencia suficientemente significativa para decir que los errrores del modelo se *distribuyen normal* con media μ y varianza σ^2.

No hacen análisis gráfico y es más importante aún en esta prueba

Varianza constante

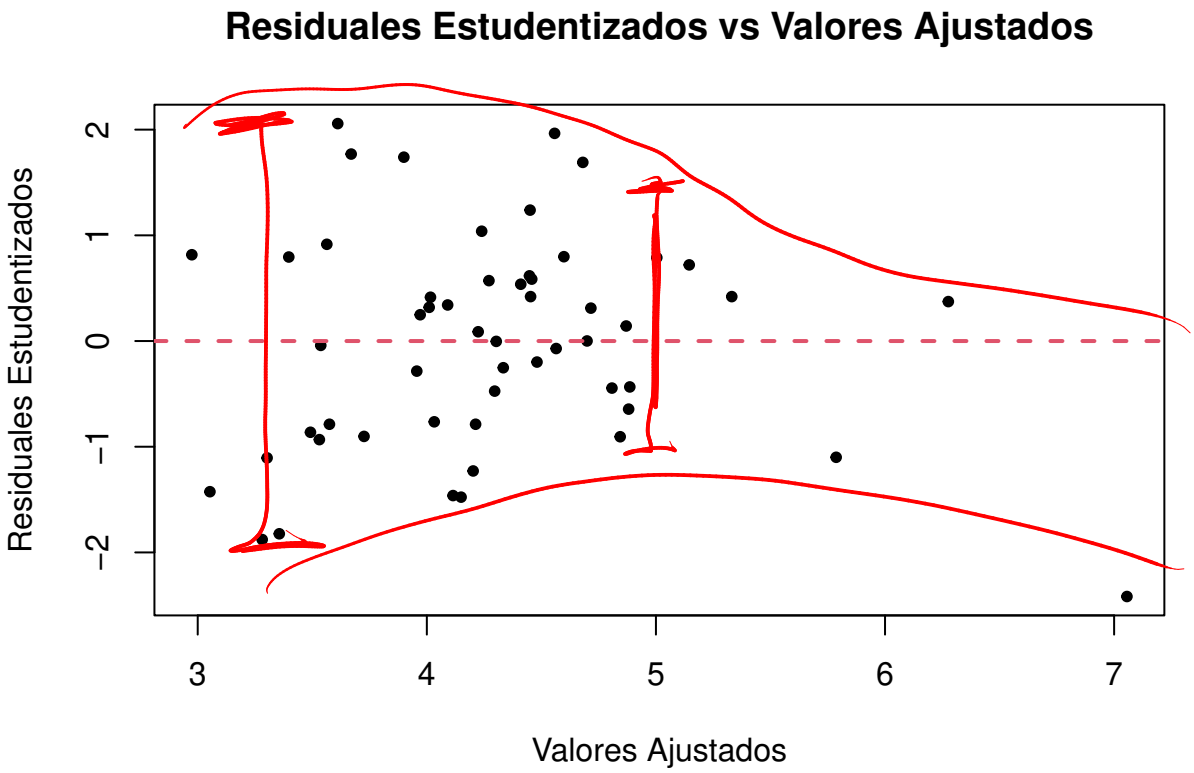


Figure 2: Gráfico residuales estudentizados vs valores ajustados

Ahora, para el análisis de la varianza constante, en el gráfico de residuales estudentizados vs valores ajustados se puede observar que no hay patrones en los que la varianza aumente, decrezca o de un compotamiento que muestre algún tipo de tendencia marcada; por lo tanto, se puede decir que el supuesto de varianza constante se cumple. Por otro lado, dado que los puntos están al rededor del cero, lo que implica que puede considerarse la media como igual a cero.

Verificación de las observaciones

Datos atípicos

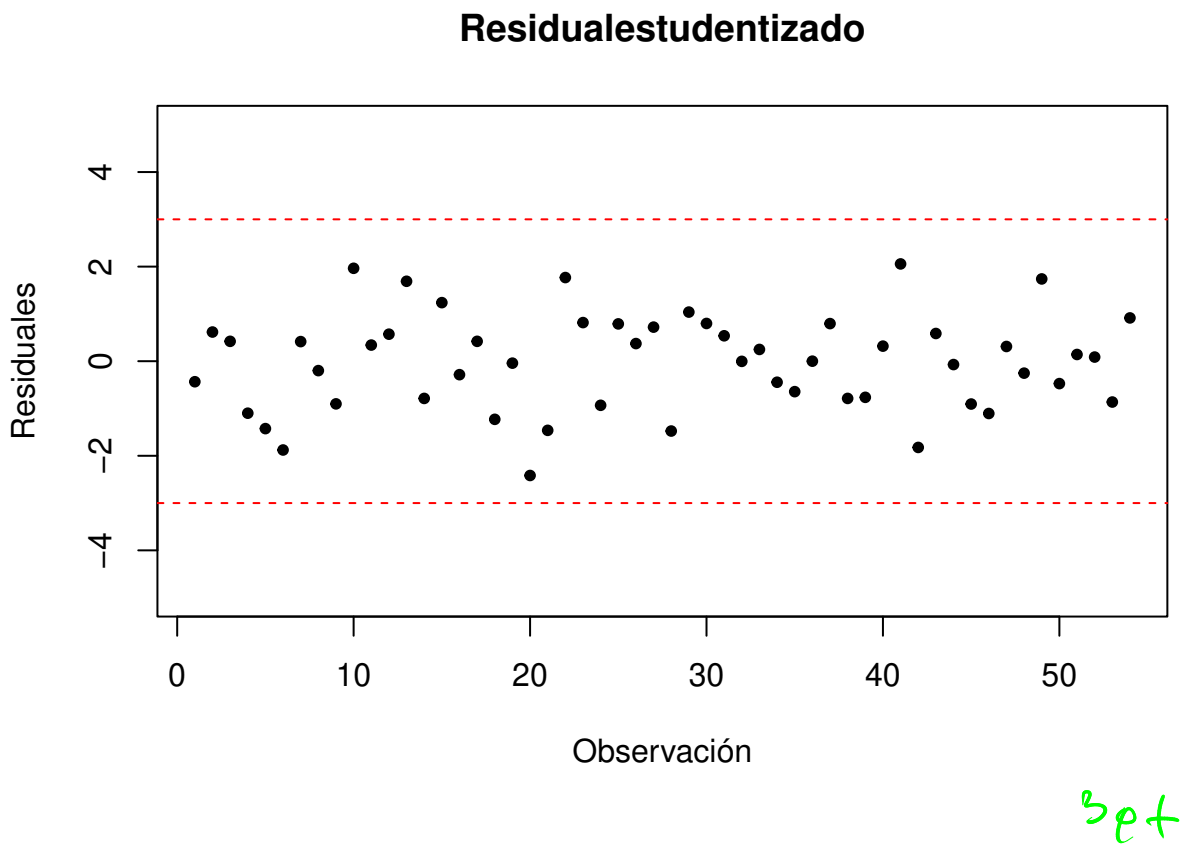


Figure 3: Identificación de datos atípicos

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

Puntos de balanceo

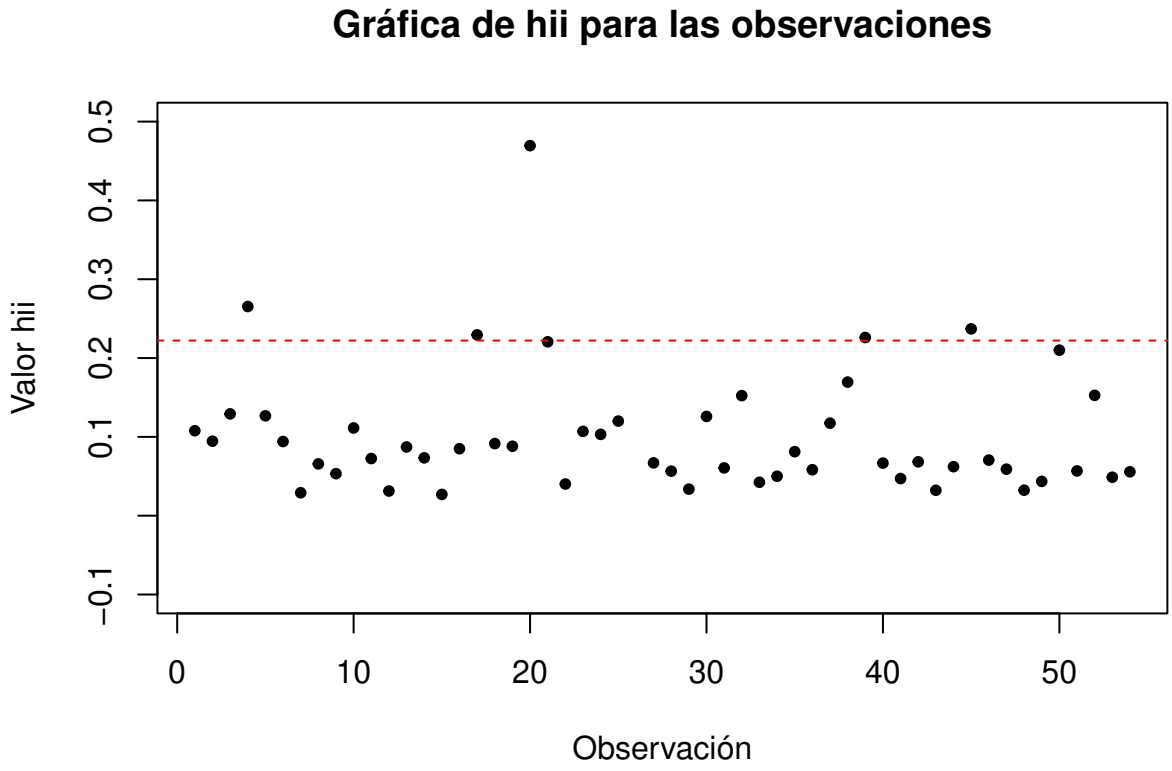


Figure 4: Identificación de puntos de balanceo

```
##    res.stud Cooks.D hii.value  Dffits
## 4   -1.0998  0.0728    0.2654 -0.6624
## 17   0.4203  0.0088    0.2294  0.2273
## 20  -2.4169  0.8616    0.4695 -2.4007
## 26   0.3730  0.0339    0.5937  0.4468
## 39  -0.7642  0.0284    0.2261 -0.4112
## 45  -0.9055  0.0424    0.2370 -0.5037
```

2 pt

Al observar la gráfica de observaciones vs valores h_{ii}, donde la línea punteada roja representa el valor $h_{ii} = 2\frac{p}{n} = 0.2222$, se puede apreciar que existen 6 datos del conjunto que son puntos de balanceo según el criterio bajo el cual $h_{ii} > 0.2222$, los cuales son los presentados en la tabla previa.

Causan...?

Puntos influenciales

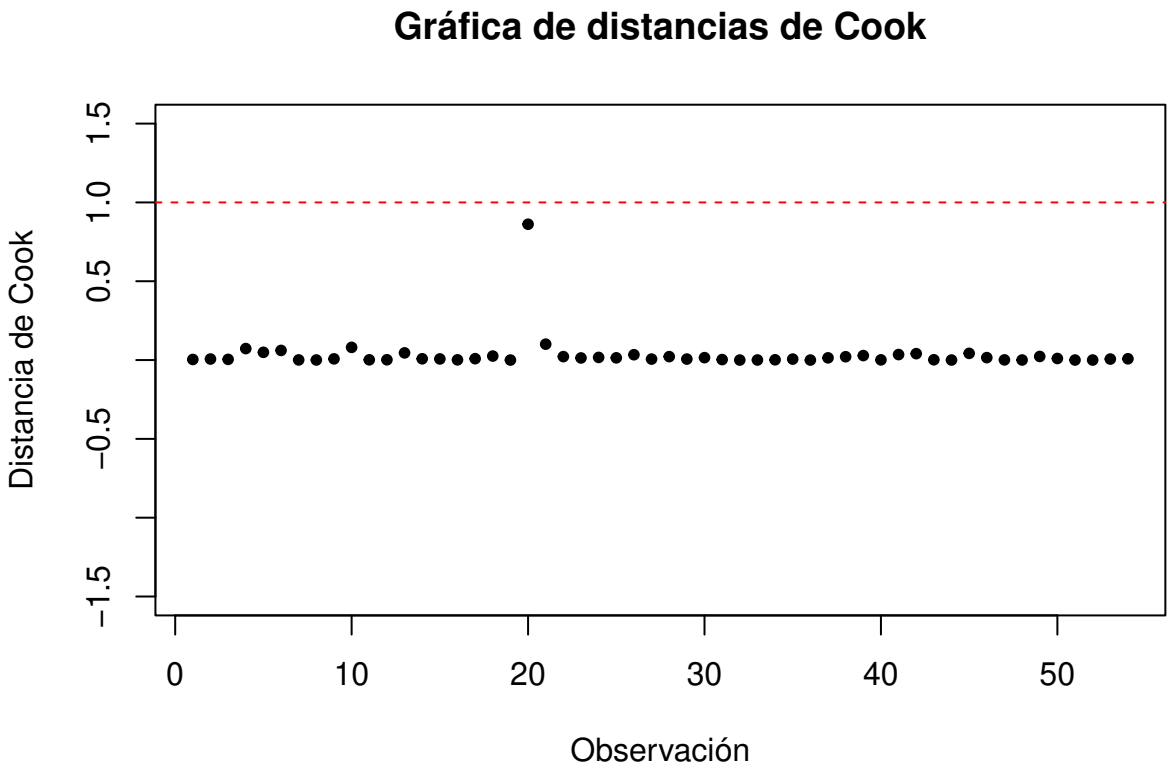


Figure 5: Criterio distancias de Cook para puntos influenciales

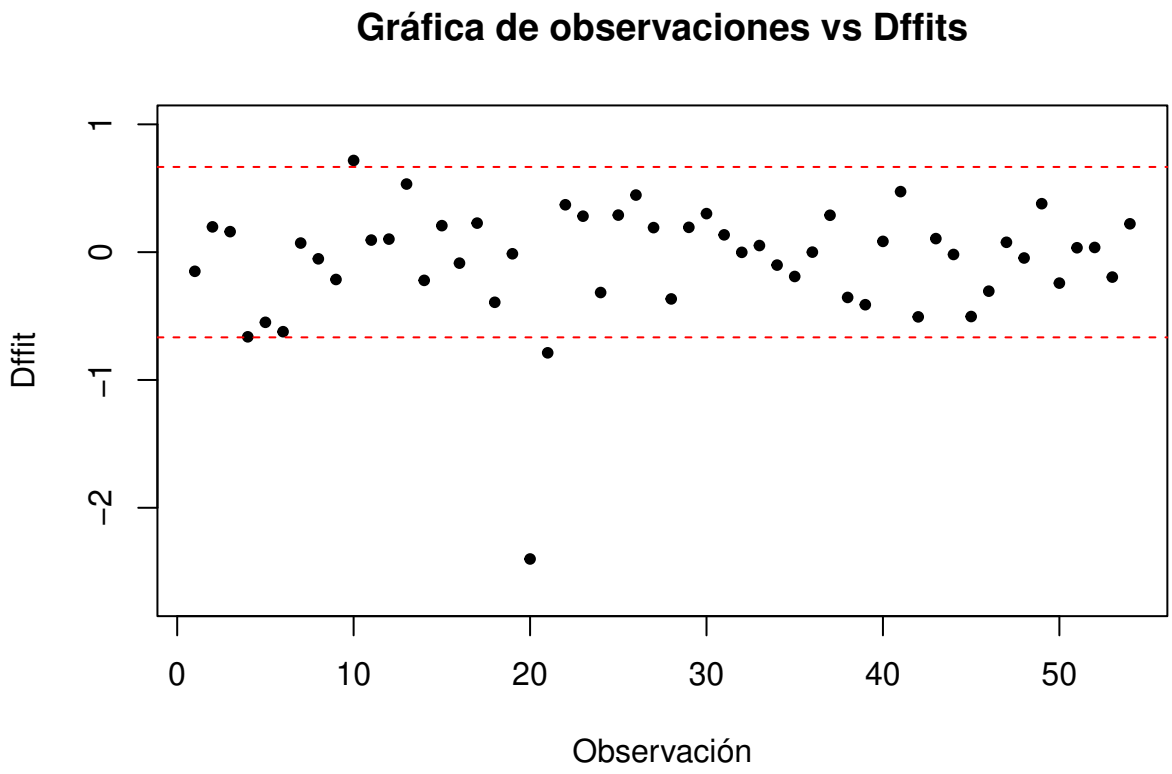


Figure 6: Criterio Dffits para puntos influenciales

```
##    res.stud Cooks.D hii.value  Dffits
## 10   1.9647  0.0805    0.1113  0.7173
## 20  -2.4169  0.8616    0.4695 -2.4007
## 21  -1.4622  0.1009    0.2206 -0.7876
```

Causan...? 3pt

Considerando el criterio de *Dffits*, las observaciones 10, 20, 21 son puntos influenciales dado que cumplen con la ecuación $|D_{ffit}| > 2\sqrt{\frac{p}{n}} = 0.6667$ lacual determina que son influenciales.

Luego, es importante mencionar también que, con el criterio de *distancias de Cook*, en el cual para cualquier punto cuya $D_i > 1$, es un punto influencial, ninguno de los datos cumple con serlo.

¿Qué define que sea aceptable?

Conclusión

1pt

El modelo "full" aquí planteado tiene buena estimación predecir el riesgo promedio de infección en los hospitales de EE.UU pues no solo cumplió con todos los supuestos (normalidad, varianza constante, media cero), sino que además arrojó un coeficiente de determinación aceptable (58.38%). También debe considerarse que se obtuvo como significativo en la prueba de hipótesis. No obstante, teniendo en cuenta de que se evidenció que hay variables no significativas, el modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i$ (obtenido en el ítem 2), es un buen modelo a considerar pero es necesario verificar todo lo anterior.

Válido o no?