Taller 5 Estadísitca II

Autores

Giselle Tatiana Fernández López
José Luis Sánchez Escobar
Santiago Rendón Giraldo

Docente

Raúl Alberto Pérez Agamez

Asignatura

Estadística II

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Medellín
Noviembre de 2021

Índice

Índice de figuras

Índice de cuadros

1. Ejercicio 1

Se tiene del texto que se desea estimar un modelo de regresión multiple que explique el riesgo de infección en términos de todas las variables predictoras, este modelo tendrá la forma

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2);\ 1 \leq i \leq 80$$

1.1. Estimación del modelo, significancia e interpretación de coeficientes

Estimación del modelo

Cuadro 1: Resumen de los coeficientes

	Estimación	Error estándar	T_0	Valor P
(Intercept)	0.6025	0.7456	0.8081	0.4216
X1	0.3102	0.0897	3.4585	0.0009
X2	0.0420	0.0117	3.6029	0.0006
X3	-0.0003	0.0031	-0.1092	0.9133
X4	-0.0026	0.0040	-0.6505	0.5174
X5	0.0042	0.0023	1.8082	0.0746

De esta tabla se obtienen los estimadores para los diferentes parametros que utilizamos en el modelo.

$$\hat{y}_i = 0.6025 + 0.3102 x_{i1} + 0.0420 x_{i2} - 0.0003 x_{i3} - 0.0026 x_{i4} + 0.0042 x_{i5}$$

Significancia e interpretación de coeficientes Para esta prueba utilizaremos el siguiente juego de hipótesis

$$H_0 : \beta_j = 0\ vs\ H_1 : \beta_j \neq 0; j = 0, .., 5$$

De la tabla obtenemos los valores p para cada uno de los parámetros, y usando un $\alpha = 0.05$ llegamos a estas conclusiones:
β_0 tiene un valor-p muy por encima de 0.05, por tanto no se rechaza la hipotesis nula, dejando así que no es significativo; por otro lado no se puede interpretar
β_1 tiene un valor-p menor que 0.05, entonces es significativo, y se puede interpretar como el aumento en Y en promedio 0.310 unidades cuando la duración promedio de la estadía de todos los pacientes en el hospital aumenta una unidad, siempre que las otras variables de predicción se tengan constantes.
β_2 tiene un valor-p menor que 0.05, entonces es significativo y se puede interpretar como el aumento en Y en promedio 0.042 unidades cada que X_2 aumenta en una unidad, siempre que las otras variables de predicción se tengan constantes

β_3 tiene un valor-p por encima de 0.05, por tanto no se rechaza la hipotesis nula, dejando así que no es significativo; por otro lado no se puede interpretar
β_4 tiene un valor-p por encima de 0.05, por tanto no se rechaza la hipotesis nula, dejando así que no es significativo; por otro lado no se puede interpretar
β_5 tiene un valor-p por encima de 0.05, por tanto no se rechaza la hipotesis nula, dejando así que no es significativo; por otro lado no se puede interpretar

1.2. Significancia de la regresión

Para probar la significancia de la regresión estableceremos las siguientes hipótesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0, \ vs$$
$$H_1 : \ alg\acute{u}n \ \beta_j \neq 0, j = 1, 2, ..., 5$$

y utilizaremos la siguiente tabla ANOVA

Cuadro 2: Tabla ANOVA para el modelo

	Suma de cuadrados	gl	Cuadrado Medio	F_0	Valor P
Model	63.2061	5	12.641215	13.3516	2.82658e-09
Error	70.0628	74	0.946795		

Haciendo el análisis de la tabla ANOVA se concluye que el modelo de regresión sí es significativo, puesto que su valor-p es menor a 0.05(2.82658e-09). Rechazando así H_0, concluyendo que el riesgo de infección depende de al menos una de las variables predictoras del modelo.

1.3. Cálculo de R^2

Para el cálculo de R^2 utilizaremos $R^2 = \frac{SSR}{SST}$ y obtendremos un valor de 0.4743, implica que un 47.43 % de la variabilidad de la variable de respuesta es explicada por la regresión

2. Ejercicio 2

Las variables que tuvieron mayor valor P son:X_3, X_4, X_5
Para probar la significancia de este subconjunto de parametros, se utilizará el siguiente juego de hipotesis:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0, vs H_1 : alg\acute{u}n \ \beta_j \neq 0 \ , j = 3, 4, 5$$

Para esto definiremos lo siguiente

$$MF = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

$$MR = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

De los cuales se obtiene $SSR(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2) = SSE(\beta_0, \beta_1, \beta_2) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ cuyas Sumas cuadráticas de errores se obtienen de la tabla de todas las regresiones posibles dando un resultado de

$$SSR(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2) = 74.534 - 70.063 = 4.471$$

Por otro lado, se tiene que $MSR = \frac{SSR(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2)}{G.L.} = \frac{4.471}{3} = 1.490$ para así hallar el estadístico de prueba que usaremos, que es:

$$F_0 = \frac{MSR(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2)}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)}$$

de la muestra se tiene que

$$MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = \frac{SSE}{n - p} = \frac{70.063}{80 - 6} = 0.9468$$

entonces se tiene que: $F_0 = \frac{4.471}{0.9468} = 4.722$ y se rechazará H_0 si $F_0 > f_{0.05,3,70}$; $f_{0.05,3,70} = 0.1167 < F_0$ por tanto se rechaza H_0 y se concluye que al menos uno de estos parametros es signifactivo para la regresión

3. Ejercicio 3: Prueba de hipótesis lineal general

Dado el modelo completo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Veamos si se puede reducir el modelo basado en una prueba de hipótesis.

3.1. Prueba de hipótesis lineal general

Sean las hipótesis:

$$\begin{cases} H_0: & \beta_2 - \beta_4 = 0, \quad \beta_3 - \beta_5 = 0 \\ H_1: & \beta_2 - \beta_4 \neq 0, \quad \beta_3 - \beta_5 \neq 0 \end{cases}$$

Podemos ver que la hipótesis nula H_0 tiene dos ecuaciones, por lo que la matriz $\mathbf{L}$ tendrá la forma 2×6. Reescribamos entonces la hipótesis nula H_0 como un sistema de ecuaciones 2×2:

$$H_0 : \begin{cases} \beta_2 - \beta_4 = 0 \\ \beta_3 - \beta_5 = 0 \end{cases}$$

En forma matricial se puede expresar como:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

De la formulación anterior podemos identificar la matriz $\mathbf{L}$ y observar que sus filas son linealmente independientes.

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

3.1.1. Modelo reducido

Entonces el **modelo reducido** para esta prueba de hipótesis lineal general, es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(X_2 + X_4) + \beta_3(X_3 + X_5) + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_{2,4} + \beta_3 X_{3,5} + \varepsilon, \text{ donde } X_{2,4} = X_2 + X_4, \text{ y } X_{3,5} = X_3 + X_5$$

A este modelo se halla asociado la suma de cuadrados del error $SSE(X1, X_{2,4}, X_{3,5})$ con $n - 4$ grados de libertad, y la suma de cuadrados de regresión $SSR(X1, X_{2,4}, X_{3,5})$ con 3 grados de libertad. Para probar la hipótesis dada es necesario comparar el modelo reducido (RM) vs. el modelo completo (FM) en términos de la razón del cuadrado medio de la diferencia de las sumas de cuadrados de los errores. Para esto, definamos primero:

$$\begin{aligned} SSH &= SSE(RM) - SSE(FM) \\ &= SSR(FM) - SSR(RM) \\ &= SSE(X1, X_{2,4}, X_{3,5}) - SSE(X1, X2, X3, X4, X5) \\ &= SSR(X1, X2, X3, X4, X5) - SSR(X1, X_{2,4}, X_{3,5}) \end{aligned}$$

Con esto, definimos a MSH como:

$$MSH = \frac{SSH}{r}$$

donde r es el número de filas linealmente independientes de $\mathbf{L}$. Entonces el estadístico de prueba F_0 está dado por:

$$F_0 = \frac{MSH}{MSE} \sim F_{2,74} \text{ bajo } H_0$$
$$= 9.2913$$

con lo que se obtiene un Valor P de 0.0003, por lo que se rechaza la hipótesis nula H_0.

4. Ejercicio 4

Para la validación de los supuestos de los errores, asumimos que se cumple que los errores del modelo tienen media cero y que son independientes.

4.1. Normalidad

Se verifica la normalidad del modelo con el test de Shapiro Wilk:

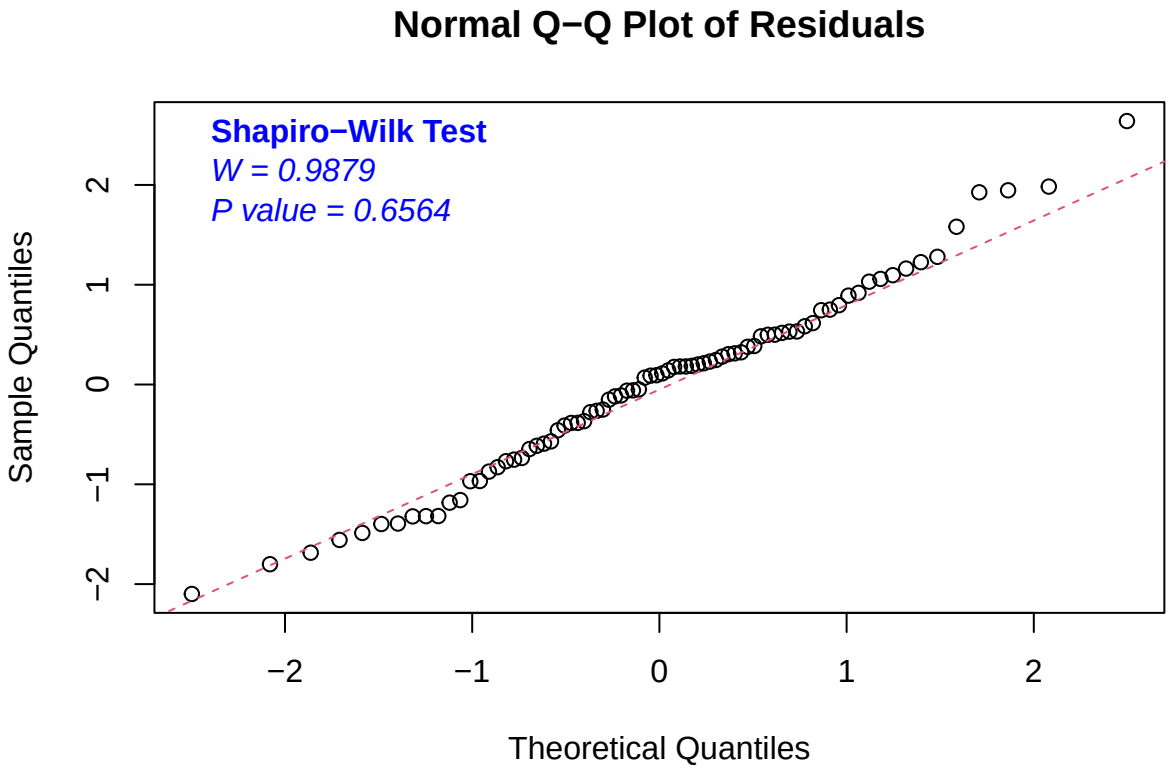

Figura 1: Análisis de normalidad

En esta gráfica observamos que en la escala normal los datos se ajustan por una línea recta, además, la Prueba de Shapiro-Wilk nos arroja un valor p de 0.6564, por lo que no se rechaza la hipótesis de que los residuales vienen de una población normal.

4.2. Varianza constante y valores atípicos

Ahora se evalúa el supuesto de varianza constante con un gráfico de residuales vs valores ajustados de la respuesta:

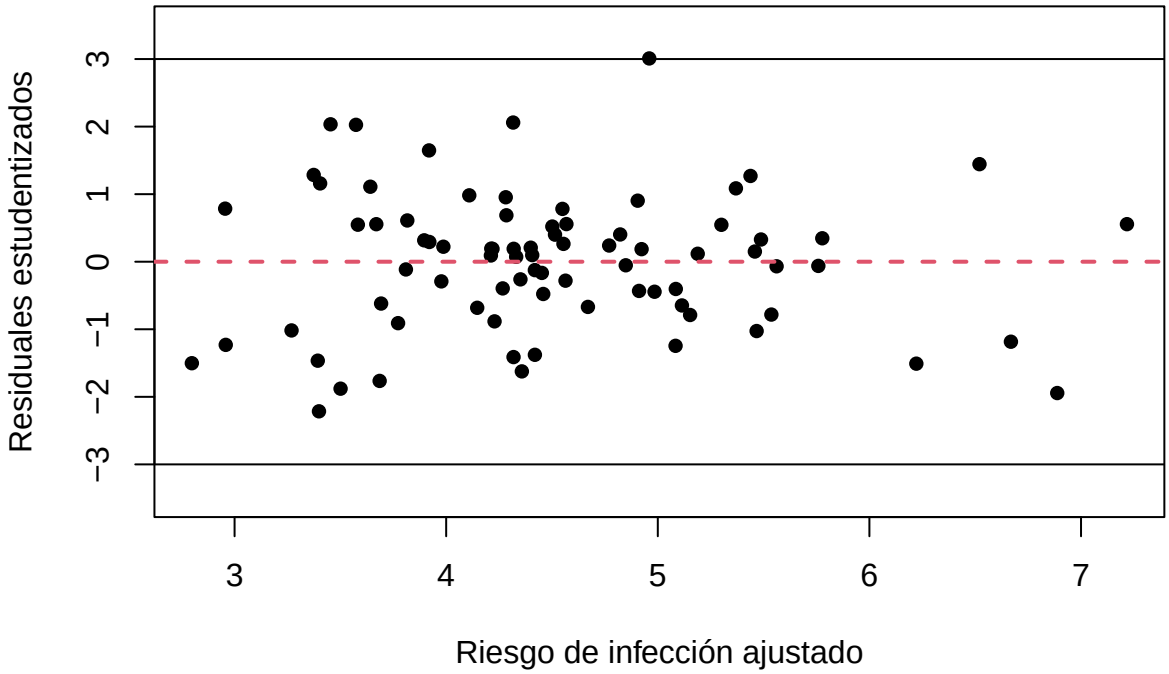

Figura 2: Análisis de variabilidad

Del anterior gráfico se puede observar que los residuales del modelo no cumplen con el supuesto de varianza constante. Por otro lado, esta gráfica indica que existe una observación atípica, el residual estudentizado es tal que su valor absoluto es mayor que 3.

4.3. Tabla de valores para el diagnóstico de valores extremos

4.4. Análisis de balanceo

Para identificar los puntos de balanceo tenemos en cuenta que una observación es un punto de balanceo si hii $>$ 2p/n. En este caso, tenemos que hii $> 2(6/80) = 0.15$.

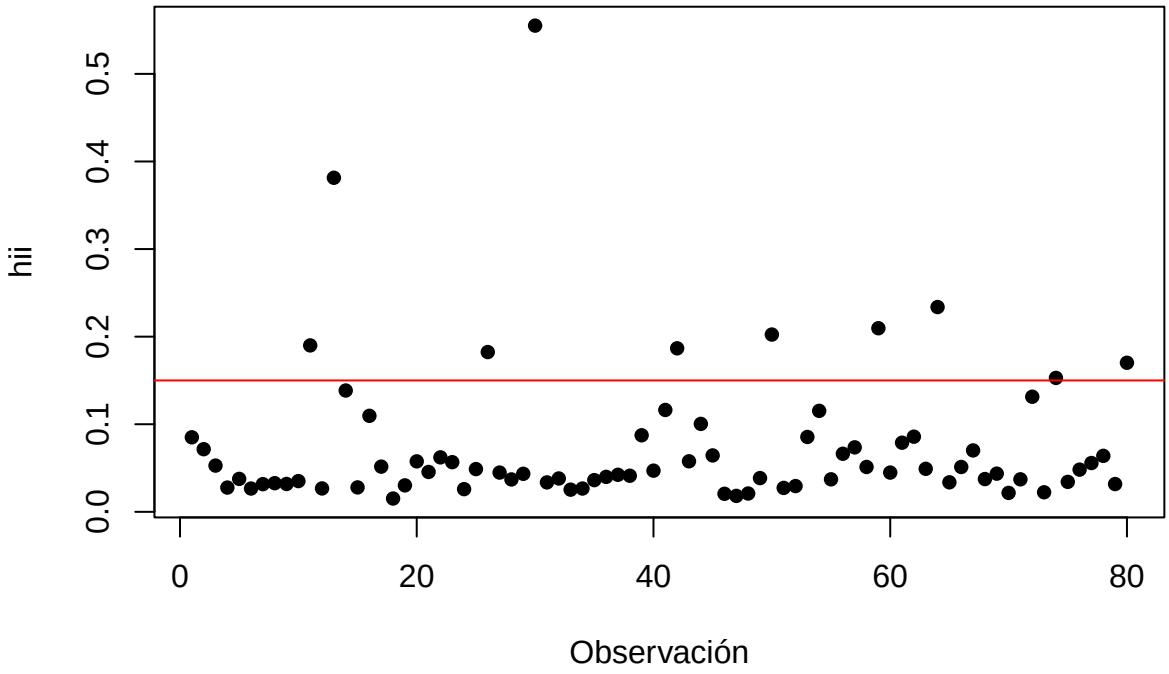

Figura 3: Análisis de balanceo

Analizamos la columna de hii.value de valores de la diagonal de la matriz H y nos damos cuenta de que las observaciones 11,26,30,42,50,59,64,74 y 80 son puntos de balanceo.

Las observaciones con hii grandes probablemente serán influenciales.

4.5. Análisis de influencia

Una observación es influencial cuando la distancia de Cook es mayor a 1 y cuando el valor absoluto del diagnóstico DFFITS es mayor a $2\sqrt{\frac{6}{80}}$.

En nuestro caso el DFFITS debe ser mayor a 0.5477.

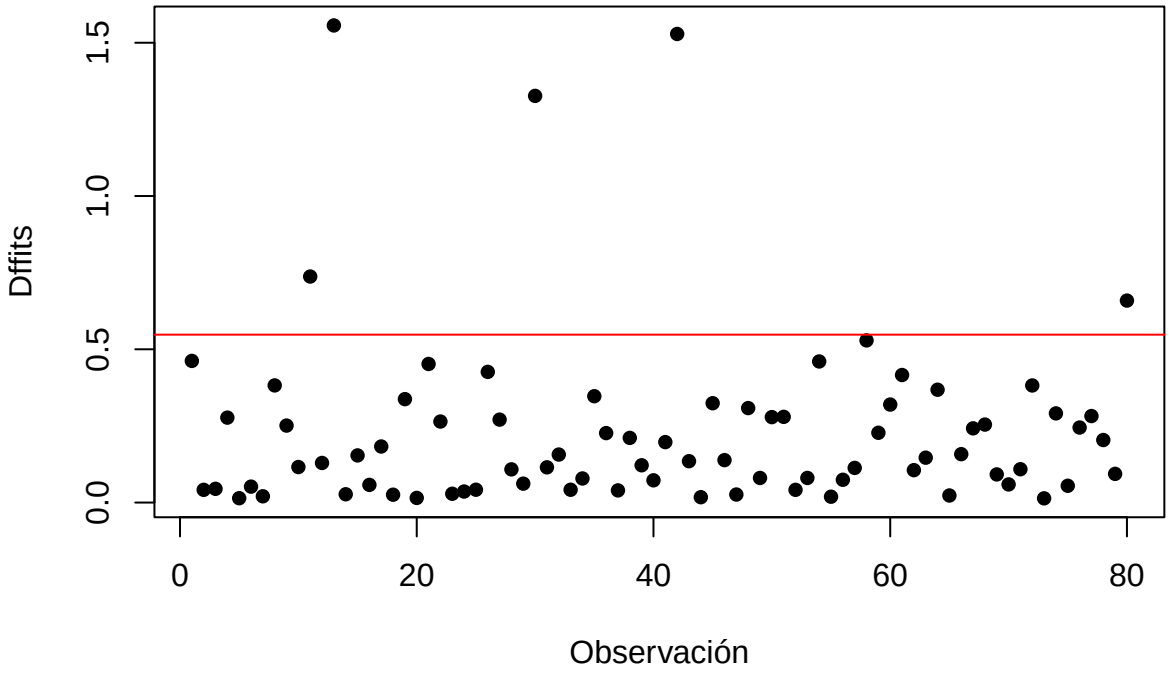

Figura 4: Análisis de influencia

Según la gráfica de DFFITS se encuentran 5 observaciones influenciales.