Trabajo Corto 01 – Estadística II

3,65

Grupo #09

Andrés Vélez Vélez

Diego Flórez Múnera

Juan David Garro Arboleda

Juan Daniel Cardona Ruiz

Estadística II

Julieth Verónica Guarín Escudero

Universidad Nacional de Colombia

Facultad de Ciencias

Escuela de estadística

Medellín

Marzo de 2023

Pregunta 1 19 pt

Significancia de la regresión:

Se parte por plantear el modelo de regresión lineal múltiple (RLM), en el cual se identifica una variable respuesta (Y), cinco variables predictoras $(X_1, \dots, X_5)$ y seis parámetros $(\beta_0, \dots, \beta_5)$.

$i = 1, 2, \dots$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i \text{ con } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \text{ Donde:}$$

$Y = Riesgo\ de\ infección$ $X_3 = Numero\ de\ camas$

$X_1 =$ Duración de la estadía $X_4 = Censo\ promedio\ diario$

$X_2 = Rutina\ de\ cultivos$ $X_5 = Numero\ de\ enfermeras$

Para garantizar la validez de los análisis futuros, se parte por comprobar la significancia de la regresión, razón por la cual se procede con el análisis de la varianza para un nivel de significancia $\alpha = 0{,}05$ de. Así, se inicia planteando las pruebas de hipótesis:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \qquad vs \qquad H_1: para\ algún\ \beta_j \neq 0\ ; j = 1, \dots, 5$$

Y se toma por estadístico de prueba a:

$$F_0 = \frac{MSR}{MSE} \sim f_{k,n-p}$$

¿n-p? ¿k?

En donde MSR corresponde a la suma de cuadrados medios descrito por el cociente entre la suma de cuadrados de la regresión (SSR) y sus respectivos grados de libertad (GL), y a su vez el MSE representa la razón entre la suma de cuadrados del error (SSE) y sus correspondientes grados de libertad (GL).

Posteriormente, los valores hallados por el estadístico F se contrastarán con los respectivos criterios de rechazo, los cuales, de cumplirse, probarán la significancia de la regresión. Estos criterios son:

$$F_0 > f_{\alpha,k,n-p} \quad \text{y} \quad VP < \alpha$$

Ambos criterios deben llevar necesariamente a la misma conclusión. Es importante destacar que, para la regresión presente, al no indicarse un nivel de significancia específico se trabajará con un Alpha de 0.05. Además, el término $f_{\alpha,k,n-p}$ corresponde al cuantil de la distribución f indicada por Alpha, y los grados de libertad de ambas sumas de cuadrados, obteniendo: $f_{\alpha,k,n-p} = 2.370977$.

¿k? ¿n-p?

La información anteriormente descrita se encuentra consignada en la tabla a continuación:

	Suma de cuadrados	*Grados de libertad*	*Suma de cuadrados medios*	*Estadístico F*	*Valores P*
Modelo	43.4086	5	8.681720	10.147	4.787×10^{-7}
Error	50.4800	59	0.855594		

¿Ecuación ajustada? Modelo ajustado antes de cualquier significancia opt

Con la información suministrada se plantea el estadístico de prueba y se somete a los criterios de rechazo:

$$10.147 > 2.370977$$

$$4.787 \times 10^{-7} < 0.05$$

Como puede observarse, ambos criterios indican que, con una significancia de 0.05, la hipótesis nula debe rechazarse, indicando que ninguno de los parámetros toma el valor de cero, dotando a la regresión de significancia.

Significancia de los parámetros individuales:

Una vez probada la significancia de la regresión, se continúa verificando esta característica para los parámetros individuales de la regresión, es decir, para $\beta_0,\ \beta_1, \beta_2, \beta_3,\ \beta_4\ y\ \beta_5$.

Para esto se realizará un procedimiento análogo al anterior, donde se cuenta con una prueba de hipótesis, un estadístico de prueba y un criterio de rechazo. Asimismo, se trabaja con un nivel de significancia de 0.05 y se halla un cuantil:

$$f_{\frac{\alpha}{2},n-p} = 2.000995$$

$$H_0: \beta_j = 0\ ;\ j = 1, \dots, 5 \qquad vs \qquad H_1: \beta_j \neq 0\ ; j = 1, \dots, 5$$

$$T_{j,o} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim f_{\alpha,n-p}$$

$$|T_{j,0}| > f_{\frac{\alpha}{2},n-p}$$

La información necesaria para llevar a cabo el proceso anteriormente descrito se consigna a continuación:

	Valor estándar	*Error estándar*	*Estadístico de prueba T*	*Valor P*
Intercepto	−0.460802065	1.4303940434	−0.3221504	0.74847745
$\hat{\beta}_1$	0.109963306	0.0840846514	1.3077691	0.19602454
$\hat{\beta}_2$	0.021857473	0.0266305541	0.8207668	0.41508312
$\hat{\beta}_3$	0.040440059	0.0181032136	2.2338608	0.02929576
$\hat{\beta}_4$	0.018181483	0.0072449560	2.5095367	0.01485364
$\hat{\beta}_5$	0.001492129	0.0007293195	2.0459196	0.04522992

De la información suministrada anteriormente se pueden comparar los valores de los respectivos estadísticos de prueba con el cuantil hallado, permitiendo determinar así para cada caso su significancia:

$$2.2338608 > 2.000995 \qquad |-0.3221504| < 2.000995$$

$2.5095367 > 2.000995$ $\quad$ $1.3077691 < 2.000995$

$2.0459196 > 2.000995$ $\quad$ $0.8207668 < 2.000995$

Se tiene entonces que los parámetros β_3, β_4 y β_5 son significativos individualmente para la regresión, pues al cumplir el criterio, se rechaza la hipótesis nula, descartando la posibilidad de que alguno de estos tome el valor de cero. Por otra parte, los parámetros β_0, β_1 y β_2 no son significativos en la regresión, pues no cumplen el criterio de rechazo.

Interpretación de los parámetros individuales

De entrada, puede afirmarse que β_0 no es interpretable pues este es un punto de extrapolación, es decir, $X_i = 0 \notin [Xmin, Xmax]$ para cada una de las covariables X_i ; $i = 1, \dots, 5$. Además, tampoco es significativo.

	y	X_1	X_2	X_3	X_4	X_5
$ValorMin$	1.3	6.7	38.8	2.2	40.4	29
$ValorMax$	7.6	17.94	65.9	36.7	133.5	835

El análisis de $\hat{\beta}_3$ indica que el aumento unitario del promedio de camas en el hospital representa un aumento promedio de 0.040440059% en la probabilidad de contagio, relación que previamente se probó significativa individualmente para la regresión. Todo esto, bajo el supuesto de que todo lo demás se mantiene constante.

El aumento unitario en el número promedio de pacientes por día, dado por $\hat{\beta}_4$, indica una tasa de crecimiento del riesgo promedio de infección en un 0.018181483%. Al igual que $\hat{\beta}_3$, este parámetro probó su importancia al cumplir con la prueba de significancia individual y al suponer constantes las demás variables.

Finalmente, $\hat{\beta}_5$, que también parte asumiendo a las demás variables como constantes, propone que, ante un aumento de una unidad en el número promedio de enfermeras, la probabilidad promedio de contagio aumenta en un 0.001492129%. Estableciéndose así la última relación significativa individual de esta regresión.

En conclusión, se tiene que de plantearse regresiones lineales simples entre cada una de las variables regresoras y la variable respuesta Y, únicamente cumplirían la prueba de significancia individual las variables acompañadas por los parámetros β_3, β_4 y β_5, pues son las únicas que al suponer todo lo demás constante cumplen con la prueba de hipótesis planteada.

Cálculo e interpretación del coeficiente de determinación múltiple R^2

Al tratarse de un modelo RLM se hace evidente que debe recurrirse a un R^2_{adj}, con el fin de evitar la inflación artificial del R^2 al contemplar múltiples variables regresoras.

$$R^2_{adj} = 1 - \frac{(n-1)MSE}{SST} = 1 - \frac{64(0.855594)}{93.8886} = 0.4167$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{50.4800}{93.8886} = 0.462341 = 46.23\%$$

Si se desea aumentar el valor del R^2_{adj} entonces deben incluirse en la regresión nuevas variables con MSE bajos, es decir, que contribuyan significativamente a la regresión y marginalmente al error. El R^2 en este caso nos dice que el 46.23% de la variabilidad total de los datos de la variable respuesta es explicada por la regresión, mientras que el R^2_{adj} nos habla de un 0.4167. → ¿y qué significa el R^2_{adj}?

Pregunta 2 9,5pt

Significancia simultánea de un subconjunto de variables

El subconjunto de variables elegidas para realizar esta prueba corresponde a aquellas cuyos estimadores tienen asociados el valor p más alto. Así, se identifican a X_1, X_2, y X_5 como las variables que cumplen con el requisito solicitado. Cabe anotar que si bien β_0 presenta el mayor valor p hallado en la anterior tabla, este no cuenta con una variable asociada, lo cual inmediatamente lo excluye de ser elegible.

Una vez identificadas las variables que desean ser estudiadas se recurre una vez más al planteamiento de una prueba de hipótesis, de un estadístico de prueba y la evaluación de un criterio de rechazo. Esto, apoyándose en la suma de cuadrados extra (SSextra), la cual permite determinar la significancia para el conjunto de variables seleccionadas simultáneamente.

Buscando desarrollar el proceso anteriormente descrito se desarrolla la suma de cuadrados extra desde la perspectiva de los SSE, apoyándose en la tabla de todas las regresiones posibles.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
K	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3
R^2	0.248	0.247	0.237	0.176	0.013	0.385	0.371	0.369	0.353	0.326	0.286	0.271	0.261	0.238	0.180	0.432	0.418
R^2_{adj}	0.236	0.235	0.225	0.163	−0.002	0.366	0.351	0.349	0.332	0.304	0.263	0.247	0.238	0.213	0.154	0.404	0.390
SSE	70.588	70.731	71.600	77.333	92.645	57.704	59.059	59.216	60.786	63.278	67.049	68.460	69.346	71.569	76.970	53.296	54.603
CP	21.502	21.669	22.685	29.385	47.281	8.443	10.027	10.211	12.046	14.957	19.366	21.015	22.050	24.649	30.961	5.291	6.818
X_i en el modelo	X_3	X_4	X_1	X_5	X_2	X_4, X_5	X_1, X_3	X_1, X_4	X_3, X_5	X_3, X_4	X_2, X_3	X_1, X_5	X_2, X_4	X_1, X_2	X_2, X_5	X_3, X_4, X_5	X_1, X_3, X_4

18	19	20	21	22	23	24	25	26	27	28	29	30	31
3	3	3	3	3	3	3	3	4	4	4	4	4	5
0.416	0.398	0.391	0.377	0.373	0.370	0.356	0.271	0.456	0.447	0.424	0.417	0.405	0.462
0.387	0.369	0.361	0.347	0.342	0.339	0.324	0.235	0.420	0.410	0.386	0.378	0.365	0.417
54.827	56.502	57.190	58.462	58.872	59.176	60.481	68.444	51.056	51.943	54.061	54.750	55.868	50.480
7.080	9.038	9.843	11.330	11.809	12.164	13.689	22.995	4.674	5.710	8.186	8.990	10.298	6.000
X_1, X_4, X_5	X_1, X_3, X_5	X_2, X_4, X_5	X_1, X_2, X_3	$X_2 X_3, X_5$	X_1, X_2, X_4	X_2, X_3, X_4	X_1, X_2, X_5	X_1, X_3, X_4, X_5	X_2, X_3, X_4, X_5	X_1, X_2, X_3, X_4	X_1, X_2, X_4, X_5	X_1, X_2, X_3, X_5	Todas

¿Para qué toda esa tabla y tantos datos?

Surgiendo así el planteamiento de la siguiente SSextra:

$$SSR(\beta_1, \beta_2, \beta_5 | \beta_0, \beta_3, \beta_4) = SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$$ ✓

→ ¿Filas = datos y columnas = datos, la hicieron al revés.

2,5 p +

$$SSR(\beta_1, \beta_2, \beta_5 | \beta_0, \beta_3, \beta_4) = 63.278 - 50.480$$

$$SSR(\beta_1, \beta_2, \beta_5 | \beta_0, \beta_3, \beta_4) = 12.798$$ ✓

Continuando este proceso se hallan los GL asociados a la SSextra:

$$SSR(\beta_1, \beta_2, \beta_5 | \beta_0, \beta_3, \beta_4) = SSE(\beta_0, \beta_3, \beta_4) - SSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$$

De hecho siempre va primero la PH1 como estadístico y después ahí sí ecuaciones.

$$\textbf{GL } SSR(\beta_1, \beta_2, \beta_5 | \beta_0, \beta_3, \beta_4) = (n-3) - (n-6)$$

$$\textbf{GL } SSR(\beta_1, \beta_2, \beta_5 | \beta_0, \beta_3, \beta_4) = 3$$ ✓

Una vez que se dispone de toda la información necesaria se procede con el planteamiento de la prueba de hipótesis descrita anteriormente para las variables de interés, asumiendo un nivel de significancia de $\alpha = 0.05$:

$$H_0: \beta_1 = \beta_2 = \beta_5 = 0 \qquad vs \qquad H_1: Algún\ \beta_j \neq 0\ ; j = 1, 2, 5$$

Para continuar con el estadístico de prueba:

$$F_0 = \frac{\frac{SSR(\beta_1, \beta_2, \beta_5 | \beta_0, \beta_3, \beta_4)}{\textbf{GL } SSR(\beta_1, \beta_2, \beta_5 / \beta_0, \beta_3, \beta_4)}}{MSE(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)} \sim f_{k,n-p} = \frac{\frac{12.798}{3}}{0.855594} = 4.9860$$ ✓

Y finalizar con el criterio de rechazo:

$$F_0 > f_{\alpha,k,n-p}$$ ✓

Donde

$$f_{0.05,3,59} = 2.760767$$ ✓

Estableciendo que:

$$4.9860 > 2.760767$$ ✓

Esto quiere decir que se cumple el criterio de rechazo y, por lo tanto, se hace posible afirmar que el conjunto de los parámetros $\beta_1, \beta_2\ y\ \beta_5$ son significativos para la regresión, es decir, alguno de estos parámetros toma un valor diferente de cero. Lo anterior permite concluir que no es correcto descartar del modelo a las variables X_1, X_2, y X_5, asociadas a los parámetros $\beta_1, \beta_2\ y\ \beta_5$. Esto ya que la prueba de hipótesis ha demostrado la relevancia de al menos uno de dichos parámetros al interior del modelo de regresión. ✓

2p +

Pregunta 3

Planteamiento de la pregunta:

A partir de la información suministrada y obtenida en el trabajo, se desea plantear una pregunta que permita conocer simultáneamente si el comportamiento de cuatro variables predictoras seleccionadas puede ser clasificada en dos grupos. Para este fin se planteó la siguiente pregunta:

¿Presenta el aumento unitario de la duración promedio de estadía de los pacientes en el hospital y el número promedio de camas el mismo impacto sobre el riesgo de infección promedio; a la vez que esta última variable presenta un mismo comportamiento ante cambios de una unidad en el censo promedio diario y el número de enfermeras?

Para aproximarse al anterior interrogante se recurre al planteamiento de una prueba de hipótesis lineal general a un nivel de significancia de $\alpha = 0.05$ y un estadístico de prueba:

$$H_0: \beta_1 = \beta_3, \beta_4 = \beta_5 \ \ vs \ \ H_1: \beta_1 \neq \beta_3, \beta_4 \neq \beta_5$$

Se reescribe la prueba de hipótesis como un sistema de ecuaciones:

$$H_0: \begin{cases} \beta_1 - \beta_3 = 0 \\ \beta_4 - \beta_5 = 0 \end{cases} \ vs \ H_1: \begin{cases} \beta_1 - \beta_3 \neq 0 \\ \beta_4 - \beta_5 \neq 0 \end{cases}$$

Finalmente, esta es expresada de forma matricial:

$$H_0: L\beta = 0 \ \ vs \ \ H_1: L\beta \neq 0$$

Donde L es una matriz de tamaño mxp, donde m = 2 corresponde al número de ecuaciones y p = 6 al número de parámetros del modelo completo. Asimismo, 0 representa un vector nulo 2x1:

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

;

Por lo tanto, puede presentarse la prueba de hipótesis matricial de forma explícita:

$$H_0: \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \ vs \ H_1: \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Ahora, puede plantearse un Modelo Reducido (MR) y un Modelo Completo (MC), los cuales se exponen a continuación:

$$\text{MR:} \quad Y = \beta_0 + \beta_1(X_1 + X_3) + \beta_2 X_2 + \beta_4(X_4 + X_5) + \varepsilon_i$$
$$MC: \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i$$

Donde se define:

$$X_{1,3} = X_1 + X_3 \quad , \quad X_{4,5} = X_4 + X_5$$ ✓

Sintetizando el modelo reducido:

$$\text{MR:} \quad Y = \beta_0 + \beta_1(X_{1,3}) + \beta_2 X_2 + \beta_4(X_{4,5})$$ ✓

Una vez definidos claramente la prueba de hipótesis y el MR de la regresión, se procede con la elaboración del estadístico de prueba. Para ello, será necesario definir previamente la suma de cuadrados debida a la hipótesis (SSH) y la suma de cuadrados medios (MSE) para el modelo completo (MC):

Se parte por el planteamiento de la SSH y sus grados de libertad:

$$SSH = SSE(MR) - SSE(MC)$$

Los GL de la SSH corresponden al número de filas linealmente independientes de la matriz L:

$$GL\ SSH = 2$$

Una vez que se dispone de esta información puede hallarse la suma de cuadrados medios debido a la hipótesis (MSH):

$$MSH = \frac{SSH}{GL\ SSH} = \frac{SSH}{2}$$

Recurriendo al previamente utilizado MSE del MC para conocer la totalidad de los términos del estadístico de prueba:

$$MSE = \frac{SSE(MC)}{GL\ MC}$$

Finalmente, se presenta un estadístico de prueba de la forma:

Reemplazar lo que conocen

$$F_0 = \frac{MSH}{MSE(MC)} \sim f_{k,n-p}$$

¿cuánto da!

¿K? → Nope, es r=2

0,5 pt

Pregunta 4 15 pt

<u>Validación de los supuestos del modelo sobre los errores:</u>

<u>Supuesto de medio cero:</u>
Se asume cierto.

<u>Supuesto de independencia:</u>
Se asume cierto.

Supuesto de normalidad:

Para probar este supuesto se cuenta con dos herramientas fundamentales: las pruebas de normalidad como la prueba Shapiro-Wilk y la gráfica de los residuales. Esta última compara los cuantiles teóricos de una distribución normal con los cuantiles observados en el modelo en cuestión. Se realizarán ambas pruebas para mayor exhaustividad, pero partiendo de la certeza que el criterio gráfico tiene la última palabra.

Se plantea entonces la prueba de hipótesis para aceptar o rechazar el supuesto de normalidad:

$$H_0 = \varepsilon_i \sim Normal \;\; vs \;\; H_1 = \varepsilon_i \nsim Normal$$

Se plantea el criterio de rechazo:

$$VP < \alpha$$

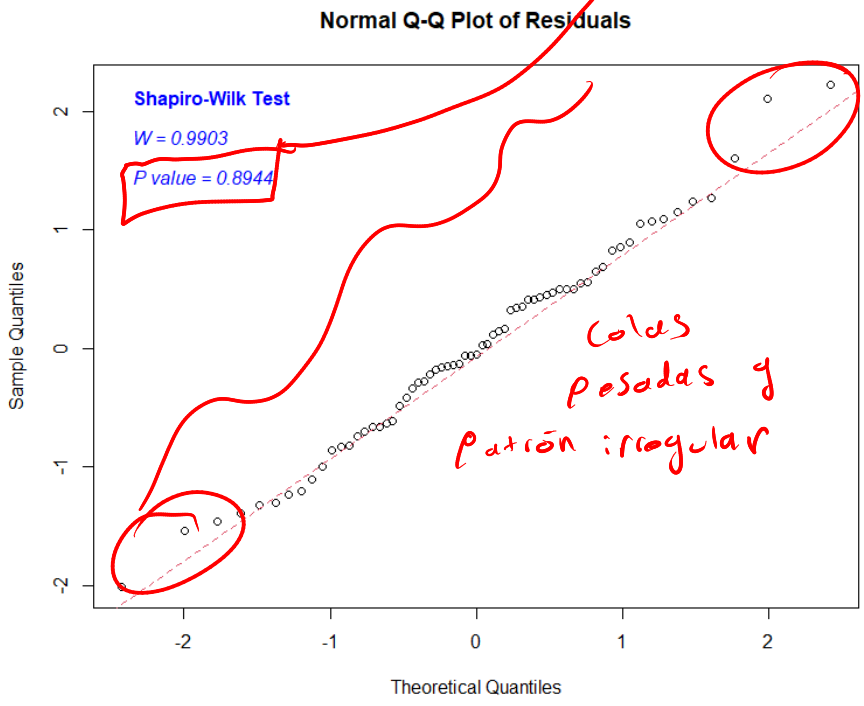
Y finalmente se ejecuta esta prueba con ayuda de los valores de la siguiente tabla:

	Valor P	*Nivel de significancia* (α)
Prueba Shapiro − Wilk	0.9903	0.05

$0.9903 > 0.05$

Por lo tanto, el criterio de rechazo no se cumple, aceptando la hipótesis nula y estableciendo que los errores cuentan con una distribución normal.

Ahora se presentará la gráfica de los residuales con el fin de someter el supuesto de normalidad al rigor de una segunda prueba:



Tal como puede advertirse, los cuantiles teóricos de la distribución se acercan bastante en coincidir con la muestra de cuantiles de la regresión, esto es, se sitúa con bastante precisión sobre la línea punteada roja que siguen las distribuciones normales.

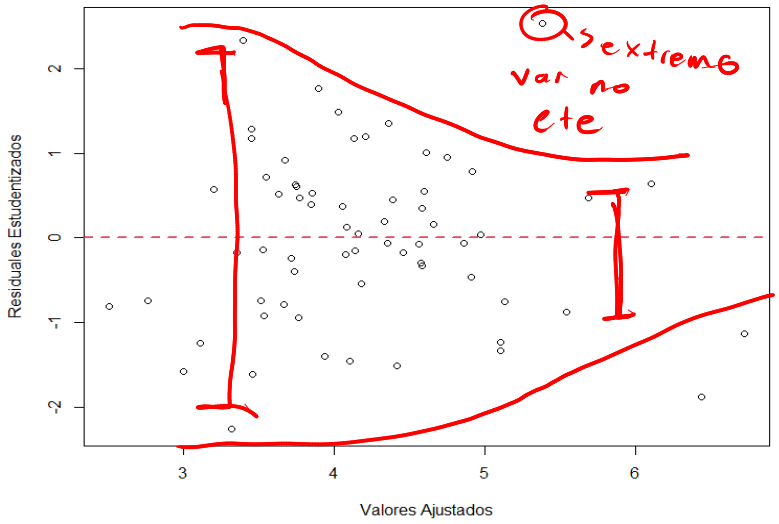
Ahora, puede establecerse indudablemente que la hipótesis nula debe aceptarse. Es decir, los errores presentan una distribución normal, cumpliendo con el supuesto del modelo.

Supuesto de varianza constante: 1,5 pt

Para aceptar o rechazar este supuesto se recurre al criterio gráfico, puntualmente a la gráfica de residuales estudentizados vs valores ajustados. Lo que se busca al analizar esta gráfica es que las observaciones más lejanas de la dispersión de puntos presenten aproximadamente la misma distancia entre ellas a lo largo de toda la nube de puntos.

Se plantea para este fin, a siguiente prueba de hipótesis y la gráfica:

$$H_0: V[\varepsilon_i] = \sigma^2 \ \ vs \ \ H_1: V[\varepsilon_i] \neq \sigma^2$$



Tras analizar la gráfica se afirma que no existe evidencia que permita afirmar que la varianza no es constante. Esto, pues no se evidencia explícitamente la forma de parlante característica, así como tampoco puede afirmarse que los puntos que harían pensar que la varianza es constante sean observaciones atípicas, como se demostrará más adelante.

Observaciones atípicas, puntos de balanceo y observaciones influenciales:

Para determinar la presencia de estas observaciones extremas se hace necesario recurrir a una serie de criterios que contrastados con los datos de interés permitirá realizar una identificación precisa de los mismos.

Así, se presenta el siguiente fragmento de la tabla diagnóstico, donde se halla la totalidad de observaciones extremas al interior del modelo, así como otros pocos que no lo son:

Sólo pongan lo que interesa en un reporte

falta dato 33

Observación	*Residuales Estudentizados*	*Distancia de Cook*	*Valor h_{ii}*	*Diferencia en el ajuste (DFFITS)*
1	1.1747	0.0146	0.0599	0.2974
2	1.1787	0.0083	0.0346	0.2240
3	−1.2363	0.0783	0.2351	−0.6886
4	0.7167	0.0030	0.0342	0.1342
5	−1.8759	0.1620	0.2164	−1.0080
6	−1.2414	0.0195	0.0707	−0.3439
7	2.5309	0.1184	0.0999	0.8852
9	−0.7586	0.0229	0.1927	−0.3692
21	−0.0598	0.0001	0.1271	−0.0226
22	−0.7452	0.0241	0.2069	−0.3792
23	0.6284	0.0056	0.0788	0.1828
48	1.4846	0.0596	0.1395	0.6042
49	0.6381	0.0278	0.2907	0.4065
51	−1.1291	0.1297	0.3791	−0.8843
65	−0.0678	0.0001	0.1548	0.0288

¿Los tachados para qué los colocaron?

Observaciones atípicas: 3pt

Las observaciones atípicas pueden ser definidas como aquellos datos que están alejados de los otros en la variable Y. Para identificar estos valores atípicos, se emplea el criterio:

$$|r_i| > 3$$

Tal como puede expresarse en la tabla de diagnóstico, ningún residual estudentizado cumple con el criterio, por lo tanto, no puede afirmarse que existen observaciones atípicas.

¿Gráfica?

Puntos de balanceo 3pt

Estos puntos son entendidos como los datos que se encuentran alejados de los demás en la variable x. El criterio que permite identificar estos puntos recurre a comparar los valores de la diagonal principal (h_{ii}) de la matriz Hat, el cual enuncia:

$$h_{ii} > 2\left(\frac{p}{n}\right)$$

Donde:

$$2 \cdot \frac{p}{n} = 2 \cdot \frac{6}{65} = 0.1846$$

Estableciendo:

$$h_{ii} > 0.1846$$

Al someter los valores consignados en la tabla de diagnóstico al criterio se establece que las observaciones 3, 5, 9, 22, 49 y 51 son puntos de balanceo. ✓

Hagan gráfica

Observaciones influenciales:

Estas observaciones, de presentarse en el modelo, afectan los valores ajustados de los parámetros. Para identificarlas, se cuenta con la prueba de la distancia de Cook (D_i), y el requisito de la diferencia en el ajuste (DFFITS).

Se inicia entonces por presentar el criterio de la distancia de Cook:

$$D_i > 1$$

Tras contrastar la totalidad de los valores provistos en la tabla se concluye que no existe un solo dato que según la prueba de la distancia de Cook sea una observación influencial.

Se finaliza entonces con la presentación del criterio DFFITS:

$$|DFFITS| > 2 \cdot \sqrt{\frac{P}{N}}$$

Donde:

$$2 \cdot \sqrt{\frac{P}{N}} = 2 \cdot \sqrt{\frac{6}{65}} = 0.6076$$

Por lo tanto:

$$|DFFITS| > 0.6076$$

Al someter los datos provistos en la tabla con el criterio identificado anteriormente se encuentra que los datos 3, 5, 7, 33 y 51 pueden definirse como observaciones influenciales.

Para concluir, puede establecerse con la ayuda de los resultados obtenidos para cada una de las pruebas desarrolladas a lo largo de este punto, que el modelo es válido. Es importante aclarar que esta conclusión se alcanzó gracias a las pruebas realizadas sobre los supuestos de la regresión, puesto que si bien las pruebas de las observaciones extremas sirven para realizar un análisis más exhaustivo, la validez está determinada por los supuestos. Esta afirmación se realiza tras encontrar que la distribución cumple con el supuesto distribucional de normalidad y que, dada la información suministrada para la varianza, no puede afirmarse que este supuesto se viole. Así, se expone como conclusión que el modelo cuenta con validez, a la vez que se reconoce que deben llevarse a cabo una serie de pruebas más amplias y exhaustivas que respalden esta afirmación con mayor rigor.