Ajuste y diagnóstico de un modelo de regresión lineal simple

Ejemplo 1. Se tiene un negocio con una compañía que vende y repara computadoras. La compañía desea predecir el número de ingenieros para el servicio que requerirá en los siguientes años. Un elemento del procedimiento de predicción es en análisis de los tiempos de reparación de las computadoras. Estos tiempos dependen del número de componentes electrónicos en la computadora que deben ser reparados o reemplazados.

Unidades (x)	Minutos (y)
1	23
2	29
3	49
4	64
4	74
5	87
6	96
6	97
7	109
8	119
9	149
9	145
10	154
10	166

Tabla 1. Observaciones para las unidades reparadas (x) y tiempo de reparación en minutos (y).

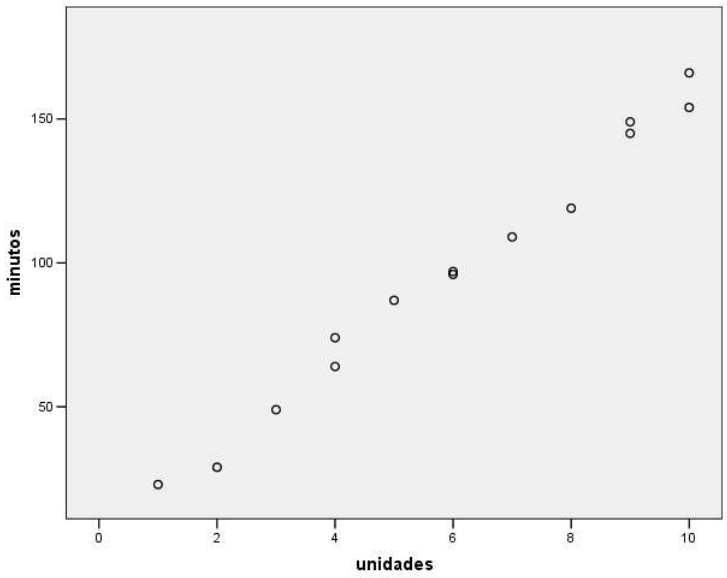

Figura 1. Gráfica de número de unidades vs. tiempos de reparación (en minutos).

En la Figura 1 puede verse que las variables tienen una relación lineal positiva (conforme aumenta el número de unidades a reparar aumenta el tiempo de reparación).

Otra medida que podría ayudarnos a saber qué tan relacionadas (linealmente) están las variables es el coeficiente de correlación. Como saben, los valores de este coeficiente se encuentran entre -1 y 1, mientras más cercano sea el valor a -1 o a 1 la relación lineal es más fuerte (en el sentido de que casi se dibuja una recta). Si es cercano a -1 su relación es inversa (una variable crece y la otra decrece) y si es cercano a 1 su relación es directa (una variable crece, la otra crece). En el presente ejemplo, el coeficiente de correlación es de 0.9934, lo que indica que los tiempos de reparación están fuertemente correlacionados con el número de unidades a reparar y que su relación es positiva. Esto se confirma observando la Figura 1.

A continuación se ajusta la recta de regresión, dado que en la Figura 1 se observa que hacerlo tiene sentido. Los estimadores de los parámetros, sus intervalos de confianza y pruebas de hipótesis se muestran en el siguiente cuadro:

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
(Constant)	4.162	3.355	1.240	.239	-3.148	11.472
UNIDADES	15.509	.505	30.712	.000	14.409	16.609

Tabla 2. Valores de los estimadores de los coeficientes de regresión ($\hat{\beta}_0$ y $\hat{\beta}_1$), prueba t e intervalos de confianza.

El valor del estimador de la pendiente ($\hat{\beta}_1$) es 15.509. Lo que indica que cuando los técnicos tienen que reparar un componente electrónico extra, se tardan, **en promedio**, 15.509 minutos más.

Si observamos el intervalo del 95% de confianza para este estimador, notamos que no contiene al cero. El valor del p-value (Sig.) para probar la hipótesis nula $H_0 : \beta_1 = 0$ es 0.000, por lo que rechazamos la hipótesis y podemos decir que el parámetro es **significativamente diferente de cero**.

El estimador $\hat{\beta}_0$ es 4.162. Este parámetro podría interpretarse como el tiempo promedio que los técnicos se tardarían en reparar 0 componentes electrónicos descompuestos de una computadora, lo cual, en este caso, suena raro. Al analizar el intervalo de confianza vemos que sí contiene al cero y el p-value para probar $H_0 : \beta_0 = 0$ es 0.239, por lo que no se rechaza esta hipótesis nula y se puede decir que el parámetro podría ser cero. Lo que indica que tal vez una regresión que pase por el origen (con $\beta_0 = 0$) ajustaría mejor a los datos (y, de hecho, en este ejemplo, al menos por interpretación, tendría más sentido).

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.994(a)	.987	.986	5.392	2.051

a Predictors: (Constant), unidades
b Dependent Variable: minutos

Tabla 3. Resumen del modelo.

Como se observa en la Tabla 3, el coeficiente de determinación $R^2 = 0.9874$ nos dice que el modelo explica el 98.74% de la variabilidad de los tiempos de reparación.

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	27419.509	1	27419.509	943.201	.000
Residual	348.848	12	29.071		
Total	27768.357	13			

a Predictors: (Constant), UNIDADES
b Dependent Variable: MINUTOS

Tabla 4. Tabla de Análisis de Varianza (ANOVA)

Con respecto a la tabla ANOVA, vemos que $\hat{\sigma}^2 = 29.071$, es decir, $\hat{\sigma} = 5.39$ (ver Tabla 3), la cual se considera una desviación estándar pequeña, considerando los valores de los tiempos de reparación. El p-value para la prueba F es 0.000, lo cual nos dice que la relación lineal entre X y Y es altamente significativa.

DIAGNÓSTICO

El diagnóstico se realiza para probar si se cumplen los supuestos del modelo de regresión lineal y que son los siguientes:

- La relación entre la variable respuesta y y las variables explicativas es lineal, al menos aproximadamente.
- El término del error ε tiene media cero.
- El término del error ε tiene varianza constante.
- Los errores no están correlacionados.
- Los errores se distribuyen normal.

Residuos

Los residuos se definen como

$$e_i = y_i - \hat{y}_i, \quad i = 1,2,\ldots,n$$

donde y_i es la i-ésima observación y $\hat{y}_i$ es el correspondiente valor ajustado.

Los residuos pueden considerarse como la desviación entre los valores observados y el ajuste; son una medida de la variabilidad de la variable respuesta no explicada por el modelo de regresión. También pueden considerarse como los valores observados de los errores del modelo y por esto, el análisis de los residuos es una manera efectiva de descubrir insuficiencias en el modelo.

Los residuos tienen media cero y su varianza promedio se aproxima por:

$$\frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SCE}{n-2} = MCE$$

Muchas veces es útil trabajar con los residuos reescalados. Estos residuos reescalados son útiles para detectar valores extremos o "aberrantes" (valores "raros").

Residuos estandarizados

$$d_i = \frac{e_i}{\sqrt{MSC}}, \quad i = 1,2,\ldots,n$$

Los residuos estandarizados tienen media cero y varianza aproximadamente uno. Un residuo estandarizado grande $(d_i > 3)$ potencialmente indica que el valor es "aberrante" o extremo.

Residuos estudentizados

$$r_i = \frac{e_i}{\sqrt{MSC\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{XX}}\right)\right]}} = \frac{e_i}{\sqrt{MSC[1-h_{ii}]}}, \quad i = 1,2,\ldots,n$$

En muestras grandes estos residuos no difieren mucho de los residuos estandarizados. Tienen distribución aproximadamente normal estándar y son más útiles para detectar observaciones "aberrantes" o influyentes.

Paréntesis. Cómo obtener los residuos estandarizados y estudentizados.

En SPSS al correr el análisis de regresión: Analyze – Regression – Linear... en la opción SAVE se seleccionan los tipos de residuos que se desea obtener.

En S-PLUS (o R) es un poco más complicado. Se tiene que realizar algo de programación, para esto, escriban las siguientes instrucciones en la "commands window".

1. `> compuaj_ls.fit(tcompus[,1],tcompus[,2])` Donde "compuaj" es el nombre que le quiero poner al ajuste del modelo. Si ustedes escriben `> compuaj` les dará la salida de todo lo que contiene el ajuste (coeficientes, residuos y más cosas). tcompus[,1] quiere decir que la variable independiente (x) es la columna 1(unidades) de mi conjunto de datos que se llama tcompus, y tcompus[,2] es la columna 2 de mi conjunto de datos, es decir, y (tiempo).
2. `> ajuste_ls.diag(compuaj)` Con esta instrucción se obtienen los residuos del modelo ajustado, que en este caso llamé "compuaj". Si escriben `> ajuste` les dirá qué tantos ajustes realizó, se obtienen residuos, distancias de cook, y otras cosas. Si se fijan hay un vector que se llama "$stud.res", ese es el que contiene a los residuos estudentizados. El vector que se llama "$std.res" contiene a los residuos estandarizados. Para obtenerlos en un vector aparte utilizamos la siguiente instrucción:
3. `> resstu_ajuste$stud.res` Aquí le están llamando "resstu" al vector que contiene a los residuos estudentizados.
4. `> resstan_ajuste$std.res` Aquí le están llamando "resstan" al vector que contiene a los residuos estandarizados.

Si quieren pegar estos vectores a su base de datos que tiene los datos originales lo pueden hacer con la siguiente instrucción:

5. `> tcompus_cbind(tcompus,resstu,resstan)` La instrucción "cbind" pega vectores en columnas, así que lo que harán con esta instrucción es pegarle a "tcompus" los vectores de residuos estudentizados y estandarizados.

De aquí ya pueden realizar las gráficas utilizando esas nuevas columnas (variables).

Gráficas de residuos (análisis cualitativo)

- *Gráfica de probabilidad Normal*

El que la distribución de los errores esté alejada de la normalidad puede traer serios problemas, ya que las estadísticas t o F y los intervalos de confianza y de predicción dependen del supuesto de normalidad; por lo cual, es importante probar si este supuesto se cumple. Una opción es realizar una gráfica de probabilidad normal (Q-Q plot).

Si los errores tienen distribución normal, los puntos caerán, aproximadamente, en una línea recta. Pueden existir algunas desviaciones, sobre todo en la cola, como las que se muestran en la Figura 2.

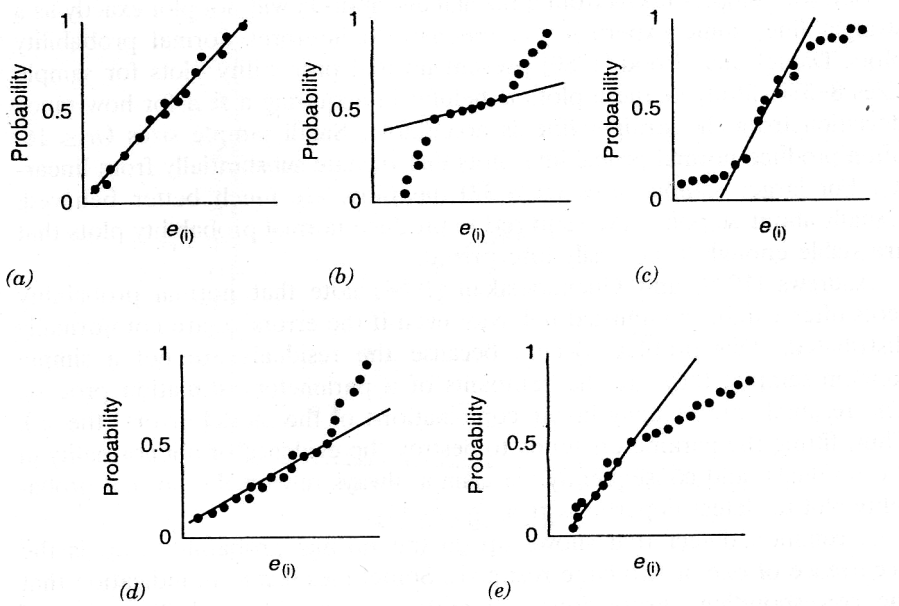

Figura 2. Gráficas de probabilidad normal. (a) ideal; (b) distribución con colas ligeras; (c) distribución con colas pesadas; (d) sesgo positivo; (e) sesgo negativo.

En la Figura 2a puede verse como sería la gráfica de probabilidad normal "ideal". Nótese que los puntos caen aproximadamente sobre la recta. Las gráficas b-e representan algunos problemas típicos. La gráfica (b) muestra puntos que las colas de la distribución son más ligeras que como lo serían con una distribución normal. Por el contrario, la gráfica (c) muestra cuál es el comportamiento típico de una distribución con colas más pesadas que la normal. Las gráficas (d) y (e) exhiben patrones asociados con sesgos positivo y negativo, respectivamente. NOTA: en las gráficas se supone que los residuos se grafican en el eje x, si se graficaran en el eje y, la interpretación es al contrario.

En realidad este tipo de gráficas no son tan fáciles de interpretar, sin embargo, nos pueden dar alguna idea de si se cumple el supuesto de normalidad. Lo mejor es utilizar también una prueba no paramétrica de normalidad para los errores.

- *Gráfica de residuos contra los valores ajustados* $\hat{y}_i$

Una gráfica de los residuos e_i (o los residuos reescalados d_i o r_i) contra los correspondientes valores ajustados $\hat{y}_i$ es útil para detectar algunas insuficiencias en el modelo. Si la gráfica no tiene ningún patrón, significa que no hay defectos obvios en el modelo y que podemos considerar que los errores tienen varianza constante. Si existe algún patrón en los datos esto puede indicar que la varianza no es constante, que no hay linealidad o que tal vez se necesiten otras variables explicativas en el modelo (ver Figura 3). En cualquiera de estos casos una solución podría ser transformar alguna de las variables.

- *Gráfica de residuos contra la variable independiente*

Realizar estas gráficas nos va ayudar también a ver si la supuesta relación lineal entre las variables no es correcta y se necesita realizar alguna transformación.

En el caso de regresión lineal simple, no es necesario realizar esta gráfica, ya que los valores ajustados son combinaciones lineales de los valores de la variable independiente, así que las gráficas sólo diferirán en la escala de la abscisa.

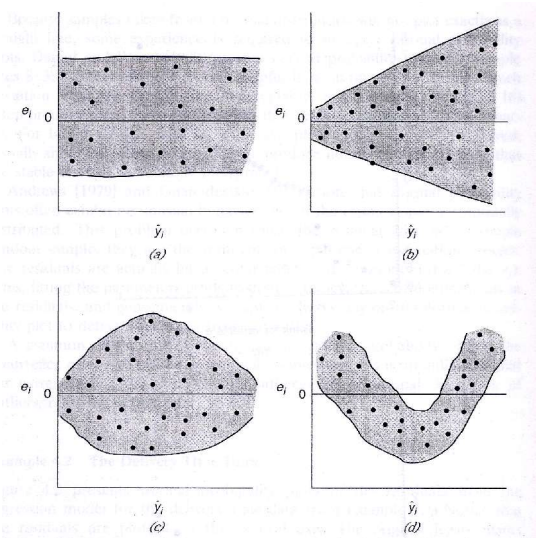

Figura 3. Patrones de gráficas de residuos. (a) satisfactorio; (b) embudo; (c) doble arco (double bow); (d) no lineal.

- *Gráfica de índice*

Graficar los residuos en el orden en que se recolectaron los datos, si este se conoce, puede ser útil, ya que se puede detectar si existen patrones en los errores, es decir, si estos están correlacionados de alguna manera. Lo ideal es que no se encuentre ningún patrón. Esta gráfica sólo tiene sentido si, por diseño del estudio, se tienen observaciones repetidas sobre el mismo individuo (pueden ser tomadas a través del tiempo).

Ejemplo.

A continuación se realizará el diagnóstico del modelo, probando los supuestos.

- ***Supuesto de normalidad***

Este supuesto se puede probar tanto gráfica como cuantitativamente (pruebas no paramétricas). Se realizarán las gráficas de los residuos estudentizados, aunque en realidad se pueden hacer con cualquier tipo de residuo.

- *Estadísticas descriptivas de los residuos estudentizados*

Se desea comparar ciertas estadísticas descriptivas de los residuos estudentizados con las de una N(0,1). Una normal estándar tiene

Sesgo = 0. El *sesgo* mide la simetría de la distribución de una variable. Un sesgo > 0 sugiere que hay más residuos positivos que negativos.

Kurtosis = 3. La *kurtosis* refleja el peso de las colas de la distribución en relación al valor central. Generalmente, aunque no siempre, la kurtosis se resta al 3 para medir distancia con respecto al valor "normal". Un valor de kurtosis <3 es evidencia de colas más pesadas que la normal y un valor >3 es evidencia de colas más ligeras.

Se utilizarán estas medidas descriptivas y otras como media, mediana, varianza, etc.

En este ejemplo se tienen las siguientes medidas descriptivas para los residuos estudentizados:

Case Summaries

Studentized Residual

N	Mean	Median	Kurtosis	Skewness	Minimum	Maximum	Std. Deviation	Variance
14	.0032599	-.1374791	-1.000	-.079	-1.81141	1.53110	1.04251926	1.087

Tabla 5. Estadísticas descriptivas de los residuos estudentizados.

Comparando estas estadísticas con las de una normal estándar puede pensarse que los datos no se distribuyen normal.

- *Histograma*

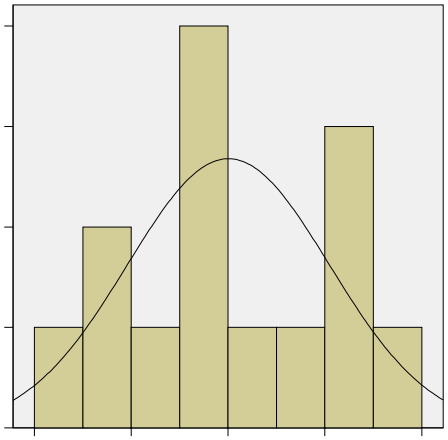

Figura 4. Histograma de los residuos estudentizados.

Como puede observarse en el histograma, los residuos no parecen tener una distribución normal.

- *Gráfica de probabilidad normal (Q-Q plot)*

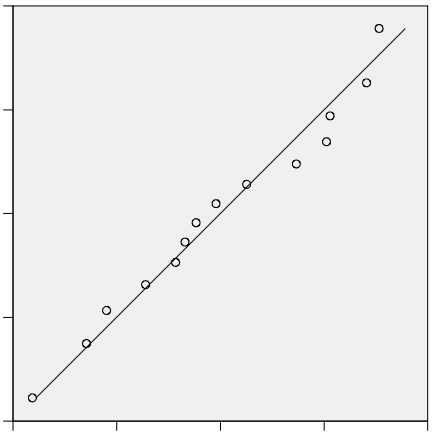

Figura 5. Gráfica de probabilidad normal de los residuos estudentizados.

Si los residuos se distribuyen normal, la mayoría de los puntos deberían de caer sobre la línea punteada (la diagonal). Como puede observarse en esta gráfica, los residuos en el ejemplo no parecen alejarse demasiado de la distribución Normal.

- *Gráfica de caja (box-plot)*

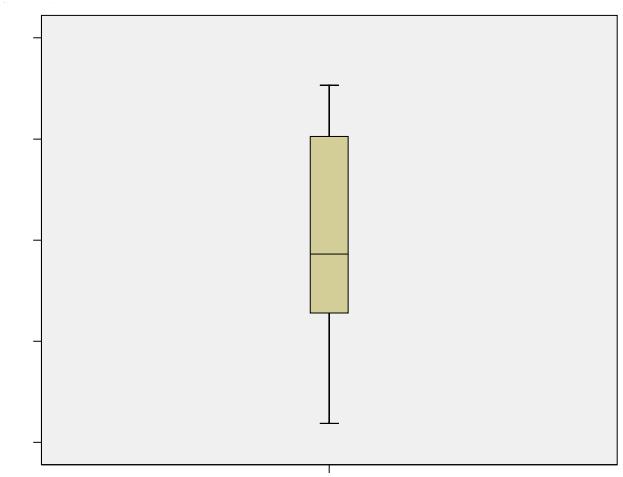

Figura 6. Gráfica de caja de los residuos estudentizados.

Con esta gráfica podemos observar qué tan simétrica es la distribución de los datos. En este caso no parece ser muy simétrica, aparentemente la distribución no es Normal.

- *Pruebas Kolmogorov-Smirnov(Lilliefors) y Shapiro-Wilk*

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Studentized Residual	.122	14	.200*	.963	14	.764

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tabla 6. Resultados de las pruebas Kolmogorov-Smirnov (Lilliefors) y Shapiro-Wilk para normalidad.

El estadístico de prueba para Lilliefors es 0.122, y una cota inferior para el p-value es 0.20. Por lo tanto no se rechaza la hipótesis de normalidad. Lo cual nos indica que las desviaciones de la normalidad que se ven en las gráficas no son tan graves. Por lo tanto, podemos decir que los residuos se distribuyen Normal.

Lo mismo se puede concluir de la prueba Shapiro-Wilk, ya que el p-value es de 0.764.

- ***Supuesto de homoscedasticidad (varianza constante).***

Este supuesto se probará utilizando las gráficas de residuos estudentizados.

- *Gráfica de residuos contra los valores ajustados* $\hat{y}_i$

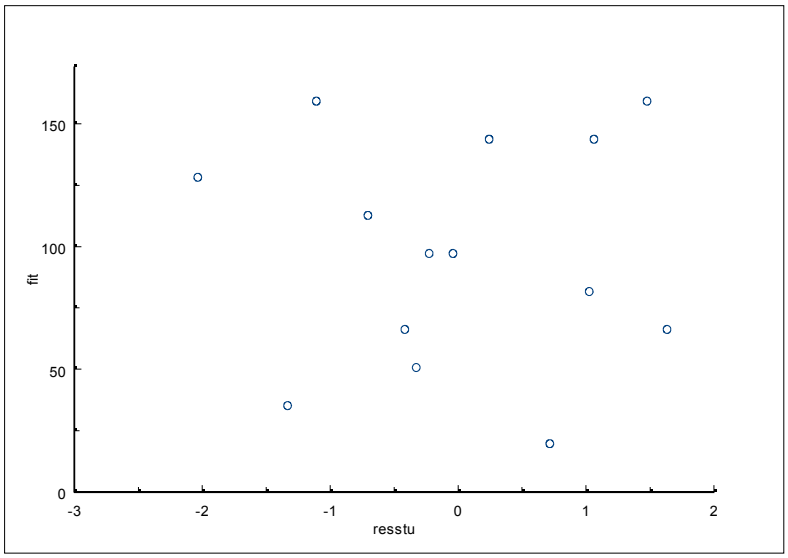

Figura 7. Gráfica de residuos estudentizados contra los valores ajustados.

Como puede notarse en la Figura 7, los residuos no tienen ningún patrón en especial, por lo que se puede decir que la varianza es constante. Con respecto al supuesto de media cero de los errores, éste se puede comprobar por medio de esta gráfica y con la media obtenida en la Tabla 4.

- ***Supuesto de linealidad.***

- *Gráfica de residuos contra la variable independiente*

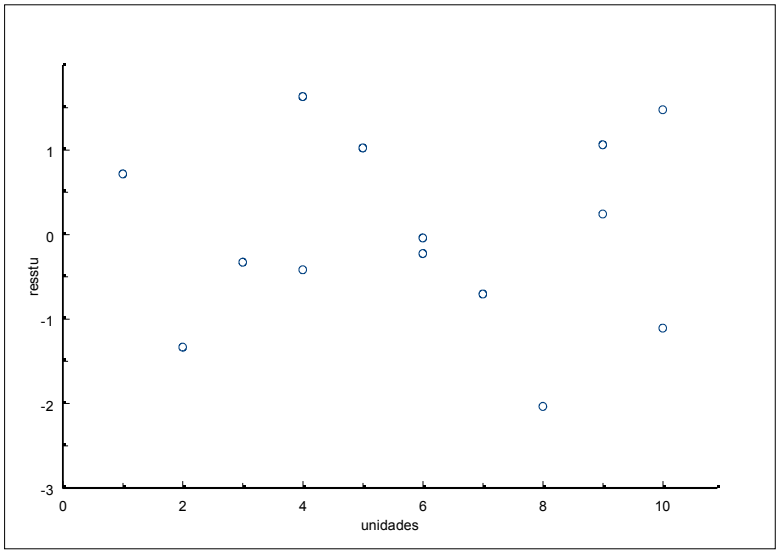

Figura 8. Gráfica de la variable independiente contra los residuos estudentizados.

En esta gráfica no se nota ningún patrón, por lo que se acepta el supuesto de varianza constante. Aunque en este caso no tiene mucho sentido hacerla, dado que sólo se tiene una variable explicativa.

- ***Supuesto de residuos no correlacionados.***

- *Gráfica de índice.*

No tiene sentido, dado que las observaciones se consideran independientes por ser tiempos de reparación de componentes electrónicos de diferentes computadoras.

- *Prueba Durbin-Watson.*

Para probar que los residuos no están autocorrelacionados se utiliza la prueba *Durbin-Watson*. Esta prueba utiliza los residuos de una regresión lineal por mínimos cuadrados. Sea e_i el residuo para la i-ésima observación. Entonces, la estadística de prueba, denotada como D, se calcula como sigue:

$$D = \frac{\sum_{i>2}\left(e_i - e_{i-1}\right)}{\sum_{i>2} e_i^2}$$

La distribución muestral de esta estadística es algo inusual. El rango de la distribución es de 0 a 4, y bajo la hipótesis nula de no autocorrelación, la media de la distribución es cercana a 2. Valores cercanos a "0" o a "4" indican problemas de autocorrelación entre los residuos. Esta prueba sólo es útil si la etiqueta i significa un orden en la muestra.

En este ejemplo esta prueba no tiene sentido. Sin embargo, se obtuvo el valor de la estadística (2.051) y puede observarse en la Tabla 3.

Observaciones "aberrantes" e influyentes (outliers).

Una observación *aberrante* (outlier o valor extremo) es aquella que es marcadamente diferente del resto de las observaciones del conjunto de datos, es decir, que es atípica. Una observación puede ser aberrante con respecto a la variable respuesta y/o las variables independientes. Específicamente, una observación extrema en la variable respuesta es llamada "outlier", mientras que valores extremos en las variables independientes se dice que tienen una "palanca" (leverage) grande y frecuentemente se les llama *puntos palanca*.

Los residuos que son considerablemente grandes, en valor absoluto, con respecto a los otros (pueden estar alejados de la media en 3 o 4 desviaciones estándar), indican observaciones aberrantes. Son puntos que son atípicos. Dependiendo de su localización en el espacio X, este tipo de observaciones pueden tener desde efectos moderados hasta graves sobre el modelo de regresión lineal. Las gráficas de residuos contra $\hat{y}_i$ y las gráficas de probabilidad normal son útiles para identificar "outliers". Examinar residuos reescalados, como los estudentizados es una manera excelente para identificar observaciones atípicas potenciales.

A una observación que provoca que los estimadores de la regresión sean sustancialmente diferentes de lo que serían si se eliminara a tal observación del conjunto de datos se le llama *observación influyente*. Las observaciones que son aberrantes (outliers) o tienen una palanca grande no necesariamente son influyentes, mientras que las observaciones influyentes son aberrantes y tienen palanca grande.

Se debe investigar cuidadosamente la razón por la cual existen observaciones atípicas. Algunas veces son valores que están "mal" que surgen debido a errores de dedo, errores de medición, etc. En este caso el valor debe corregirse o eliminarse del conjunto de datos. La razón para eliminarlos es que son valores que están mal y pueden afectar nuestro ajuste.

Por otro lado, se pueden tener valores atípicos que son observaciones "raras" pero perfectamente plausibles. Eliminar estos puntos para "mejorar el ajuste" puede ser peligroso, ya que esta acción puede dar al usuario la falsa sensación de precisión en la estimación o predicción. Ocasionalmente puede encontrarse que los valores atípicos son más importantes que el resto de los datos porque pueden controlar muchas de las propiedades del modelo (*valores influyentes*).

Ejemplo 2. Este ejemplo consiste de 10 observaciones generadas de un modelo de regresión lineal simple. Los valores de las observaciones de la variable independiente x tienen valores de 1 al 10. El modelo es

$$y = 3 + 1.5x + \varepsilon$$

Donde ε es una variable aleatoria Normal con media cero, desviación estándar 3 y sin observaciones atípicas. Los datos se muestran en la Tabla 6.

Obs	x	y
1	1	6.2814
2	2	5.1908
3	3	8.6543
4	4	14.3411
5	5	13.8374
6	6	11.1229
7	7	16.5987
8	8	19.1997
9	9	20.0782
10	10	19.7193

Tabla 7. Ejemplo para ilustrar valores extremos

Los resultados del análisis de regresión ajustado se muestran a continuación:

```
       *** Linear Model ***

Call: lm(formula = y ~ x, data = inventado, na.action = na.exclude)
Residuals:
    Min     1Q Median    3Q   Max
 -3.234 -1.248 0.5007 1.041 3.402

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 4.1044 1.4139     2.9030  0.0198
       x 1.7087 0.2279     7.4989  0.0001

Residual standard error: 2.07 on 8 degrees of freedom
Multiple R-Squared: 0.8755
F-statistic: 56.23 on 1 and 8 degrees of freedom, the p-value is
0.00006935
```

```
Analysis of Variance Table

Response: y

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
        x  1  240.8762 240.8762 56.23292 0.0000693522
  Residuals  8   34.2684   4.2835
```

Tabla 8. Resultados del análisis de regresión para el Ejemplo 2.

Los resultados son razonables. La regresión es significativa, los valores de los estimadores están dentro de una desviación estándar de los valores reales.

Para propósitos de ilustración, se considerarán 2 escenarios diferentes:

Escenario 1. Cuando $x = 5$ y y se incrementa en 10 unidades.
Escenario 2. Cuando $x = 10$ y y se incrementa en 10 unidades.

Nótese que en cada escenario el valor observado de la variable dependiente se ha incrementado en más de 3 desviaciones estándar, esto es un resultado que no esperamos que ocurra y por lo que el valor se considera extremo (outlier). La diferencia entre los dos escenarios es la localización del valor extremo: El primero ocurre en la mitad del rango de los valores de x, mientras que el segundo ocurre al final. La Tabla 8 muestra un resumen de los resultados para la regresión para los datos originales y los dos escenarios para observaciones extremas.

Escenario	$\hat{\beta}_0$ error std.	$\hat{\beta}_1$ error std.	MSE
Datos originales	4.104 1.414	1.709 0.228	4.284
Escenario 1	5.438 2.936	1.648 0.473	18.469
Escenario 2	2.104 2.024	2.254 0.326	8.784

Tabla 9. Resumen de las regresiones.

Los resultados muestran un sesgo en las estimaciones debido a los valores extremos, pero los sesgos son bastante diferentes para los dos escenarios. Para el escenario 1, $\hat{\beta}_1$ ha cambiado poco, $\hat{\beta}_0$ se ha incrementado algo, y el MCE y, consecuentemente, los errores estándar de ambos coeficientes son más grandes. Para el escenario 2, $\hat{\beta}_1$ ha cambiado dramáticamente, mientras que los cambios en los valores de $\hat{\beta}_0$ y MCE no son tan marcados. Estos resultados pueden verse en la Figura 8, que muestra las gráficas de los valores observados y la recta ajustada. Estas gráficas muestran que el efecto más visible en la recta estimada es debido al outlier que se encuentra al final de los valores de x.

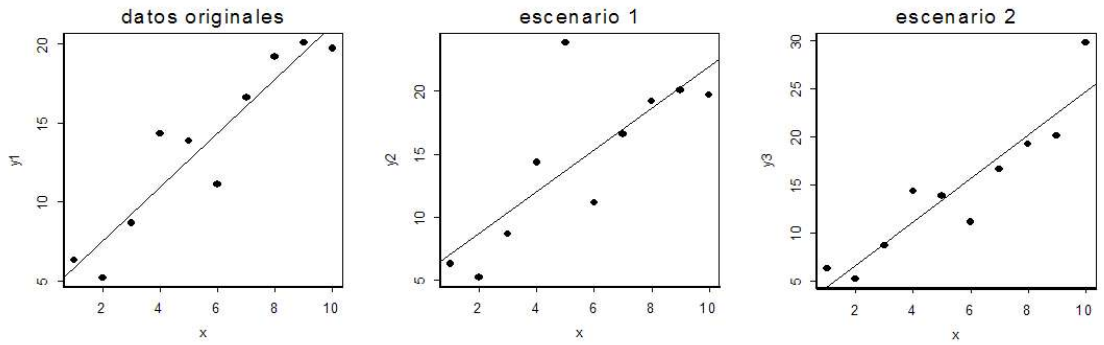

Figura 9. Efectos de los valores extremos en la regresión.

En la gráfica del Escenario 1 puede notarse que al tener un valor extremo a la mitad del rango de los valores de x, la recta incrementa un poco la ordenada al origen, pero mantiene la pendiente casi igual. Sin embargo, en la gráfica del Escenario 2, puede notarse que el valor extremo al final del rango de las x actúa como "palanca" y hace que cambien tanto la ordenada al origen como la pendiente. Por lo tanto, la localización del outlier afecta la naturaleza del sesgo en los estimadores de los parámetros. Esta combinación de un valor extremo con un valor grande de la palanca define a una observación *influyente*.

- ***Detección de valores extremos y observaciones influyentes.***

- *Gráficas de residuos*

Ahora que se ha visto el efecto de los valores extremos en la regresión, se necesitan métodos para detectarlos. Anteriormente ya se mencionó que una manera de hacerlo es utilizando gráficas de residuos.

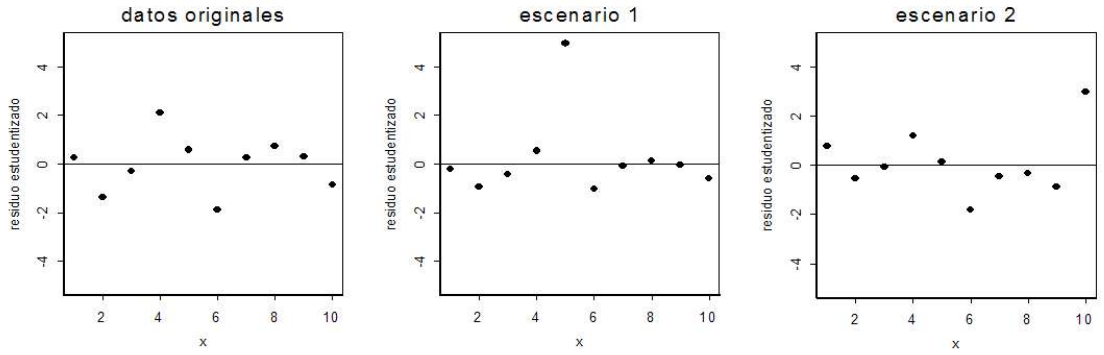

Figura 10. Gráficas de la variable independiente vs. los residuos estudentizados para los diferentes escenarios.

Como puede notarse en las gráficas, en el escenario 1 el valor extremo sobresale de los demás, sin embargo, en el escenario 2, el valor extremo no es muy diferente, en magnitud, a los demás residuos. Por esta razón se necesita de otras medidas que ayuden a identificar si una observación es influyente o no.

- *Estadísticas para medir la "palanca"*

Como ya se vio, el tamaño de la "palanca" es un elemento importante para estudiar el efecto de los valores extremos. La medida más común para medir la palanca son los elementos de la diagonal de la matriz H, h_i, misma que se definirá más adelante. Estos elementos son medidas estándar de la distancia ente los valores de x para la i-ésima observación y la media de

los valores de x para todas las observaciones. Una observación con un valor grande de h_i puede considerarse un valor extremo para la variable independiente. Obviamente, una observación puede tener una palanca grande sin ejercer una gran influencia en el modelo. De hecho, una interpretación de una observación con una palanca grande es que es una observación que podría causar problemas. En cualquier caso, es útil identificar observaciones con una palanca grande y luego determinar la influencia que éstas tienen en el modelo.

Ahora sólo falta especificar qué valores de h_i implican un alto grado de "palanca" (leverage). Puede mostrarse que

$$\sum h_i = (k+1)$$

donde k es el número de variables independientes ($(k+1)$ es el número de parámetros cuando se incluye $\hat{\beta}_0$ en el modelo). Entonces, el valor promedio para las h_i es

$$\bar{h} = \frac{k+1}{n}$$

Como una regla de dedo un valor que exceda al doble del promedio, esto es, $h_i \geq 2\bar{h}$, es considerado un indicador de un alto grado de palanca, aunque esta regla es algo arbitraria. Sin embargo, hay que tener cuidado con esta regla, puede haber situaciones en las cuales $2\bar{h} > 1$, en estos casos la regla no aplica.

Para una variable independiente

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2},$$

que, de hecho, es una medida de la magnitud relativa de la distancia al cuadrado de x a $\bar{x}$ para la i-ésima observación.

En el ejemplo 2, los valores de la palanca para $x = 5$ y $x = 10$, son respectivamente, son 0.103 y 0.345. El valor de $2\bar{h} = 0.4$, por lo tanto, no se considera que estas observaciones tengan palanca grande.

Nota. En SPSS, al realizar la regresión, en el botón SAVE... se puede pedir el "leverage", sin embargo, éstos valores son sólo el segundo término de la suma en h_i, así que hay que tener cuidado.

En S-Plus los valores de la palanca se obtienen igual que los residuos estandarizados (pasos 1 y 2 iguales), pero con la instrucción > resstu_ajuste$hat.

- *Medidas de influencia: Distancias de Cook.*

Esta estadística es igual a

$$D = \frac{r_i^2}{k+1} \frac{h_i}{(1-h_i)}$$

Los puntos donde $D_i > 1$ se consideran influyentes. Esta distancia es una *medida de supresión*, esto es, mide la influencia de la i-ésima observación removiéndola del modelo.

Nota. En SPSS, al realizar la regresión, en el botón SAVE... se pueden pedir las distancias de Cook ".

En S-Plus los valores de la palanca se obtienen igual que los residuos estandarizados (pasos 1 y 2 iguales), pero con la instrucción > resstu_ajuste$cooks.

Datos originales	Escenario 1	Escenario 2
0.0206	0.0141	0.1697
0.2791	0.1495	0.0506
0.0100	0.0209	0.0007
0.2257	0.0242	0.0986
0.0211	0.3578	0.0016
0.1563	0.0613	0.1480
0.0055	0.0006	0.0157
0.0614	0.0023	0.0130
0.0182	0.0004	0.1341
0.2040	0.1056	**1.1812**

Tabla 10. Distancias de Cook para los diferentes escenarios del ejemplo 2.

Como puede verse en la Tabla 10, para el modelo que utiliza los datos originales y para el escenario 1 no hay observaciones influyentes, sin embargo, para el Escenario 2, la observación correspondiente a $x = 10$ sí se considera influyente.

- *Otras medidas de influencia: DFFITS y DFBETAS*

Estas medidas también son de *supresión*. La primera de éstas es una estadística que indica qué tanto cambia el coeficiente $\hat{\beta}_j$, tomando como unidad la desviación estándar, si se elimina a la *i*-ésima observación. Un valor grande (en valor absoluto) de las $DFBETAS_{j,i}$ indica que la observación *i* tiene una influencia considerable en el *j*-ésimo coeficiente de regresión. Nótese que hay un vector de *DFBETAS* para cada coeficiente. Se sugiere que cualquier observación para la cual $\left|DFBETAS_{j,i}\right| > \frac{2}{\sqrt{n}}$ necesita de atención.

También es importante investigar la influencia de la supresión de la i-ésima observación en el valor ajustado. Esto lleva a una segunda medida de diagnóstico llamada $DFFITS_i$, que es el número de desviaciones estándar que el valor ajustado $\hat{y}_i$ *cambia* si se elimina la observación *i*. Se sugiere que cualquier observación para la cual $\left|DFFITS_i\right| > 2\sqrt{(k+1)/n}$ requiere de atención.

Nota. En SPSS, al correr la regresión, en el botón SAVE... se pueden pedir las DFBETAS y DFFITS *estandarizadas*. Esas son las que se ocupan.

En Splus, de la misma manera que se obtienen los residuos estandarizados y distancias de cook se obtienen los DFFITS. > resstu_ajuste$dfits. En este programa no se pueden encontrar las DFBETAS, al menos no con alguna instrucción.

Datos originales			Escenario 1			Escenario 2		
dfit	dfbeta0	dfbeta1	dfit	dfbeta0	dfbeta1	dfit	dfbeta0	dfbeta1
0.19097	0.19025	-0.16097	-0.1578	-0.1572	0.1330	0.56823	0.56609	-0.47898
-0.78674	-0.77012	0.60817	-0.5432	-0.5317	0.4199	-0.30355	-0.29713	0.23465
-0.13326	-0.12409	0.08749	-0.1937	-0.1803	0.1272	-0.03415	-0.03180	0.02242
0.80269	**0.65873**	-0.37157	0.2100	0.1723	-0.0972	0.45572	0.37399	-0.21096
0.19695	0.11976	-0.03378	**1.6823**	**1.0229**	-0.2885	0.05228	0.03179	-0.00897
-0.64391	-0.19577	-0.11043	-0.3517	-0.1069	-0.0603	-0.61821	-0.18796	-0.10602
0.09899	0	0.04582	-0.0335	0.0000	-0.0155	-0.16805	0	-0.07779
0.34017	-0.07918	0.22333	0.0640	-0.0149	0.0420	-0.15175	0.03532	-0.09963
0.17971	-0.07036	0.13892	-0.0278	0.0109	-0.0215	-0.51096	0.20006	-0.39498
-0.62867	0.31315	-0.52993	-0.4411	0.2197	-0.3718	**2.16624**	**-1.07904**	**1.82598**

Tabla 11. DFFITS y DFBETAS para los diferentes escenarios.

En la Tabla 11 se muestran las DFFITS y DFBETAS para los 3 diferentes escenarios del ejemplo 2. Como se vio anteriormente, el punto de corte para las DFBETAS es $\frac{2}{\sqrt{10}} = 0.6325$. Por lo tanto, para el modelo con los datos originales, la observación 4 es influyente para $\hat{\beta}_0$, es decir, si se elimina del los datos el valor de la ordenada al origen cambia. Para el modelo del escenario 2, la observación 5 es influyente para $\hat{\beta}_0$. Y para el modelo del escenario 3, la observación 10 es influyente tanto para $\hat{\beta}_0$ como para $\hat{\beta}_1$. Esto es algo que ya se había notado al calcular los parámetros para los 3 escenarios.

El punto de corte para las DFFITS es $2\sqrt{(k+1)/n} = 0.8944$. Por lo tanto, para el escenario 1, la observación 5 es influyente para $\hat{y}_5$, es decir, si se elimina esta observación el valor ajustado cambia. Para el escenario 2, la observación 10 es influyente para su valor ajustado.

En conclusión, para el escenario 1, la observación 5 es influyente tanto para su valor ajustado como para la ordenada al origen, y para el escenario 2, la observación 10 es influyente para su valor ajustado y para los dos estimadores de los parámetros.

- *Tratamiento de observaciones influyentes.*

La pregunta ahora es si las observaciones influyentes deben o no eliminarse del modelo. Como ya se mencionó al hablar de valores extremos, si la observación es atípica debido a errores de dedo, medición, etc., es decir, que por alguna razón la observación es inválida, entonces eliminarla del modelo es apropiado. Sin embargo, si el análisis revela que un punto influyente es una observación válida, no hay justificación para eliminarla. Claro que siempre se pueden eliminar estas observaciones con la finalidad de explorar el ajuste del modelo sin ellas.

BIBLIOGRAFIA

- Chatterjee, S & Price, B (1991) *Regression Analysis by Example*. John Wiley & sons. 1a. ed.
- Draper, Norman & Smith, Harry (1998) *Applied Regression Analysis*. John Wiley & sons. 3a. ed.
- Freund, R & Wilson, W (1998) *Regression Analysis. Statistical Modeling of a response variable*. Academic Press. 1a. ed.
- Montgomery, D, et. al (2001) *Introduction to linear regression analysis.* John Wiley & sons.
- Kleinbaum, D. et. al (1998) Applied Regression Analysis and other multivariate methods. Duxbury. 3a. ed.