Verificación de Calidad de Modelos en Regresión Lineal Software Estadístico de Regresión ERLA

Autores:
Juan Carlos Buenaño Cordero, Celia De La Cruz Cedeño
Coautor:
Gaudencio Zurita Herrera
Instituto de Ciencias Matemáticas
Escuela Superior Politécnica del Litoral
Km. 30.5 vía Perimetral, Edificio 32D, Guayaquil - Ecuador
jucabuen@espol.edu.ec, adela@espol.edu.ec, gzurita@espol.edu.ec

Resumen

En un análisis de Regresión Lineal, existen varios supuestos o premisas que deben ser considerados al momento de determinar la validez de un modelo, puesto que el no cumplimiento de alguno de estos supuestos podría conducirnos a modelos inestables, de ser así, un valor alto del estadístico R^2 o R^2 ajustado no garantiza que el modelo se ajuste bien a los datos. Entre los principales supuestos que se realizan están: La distribución del error normal con media 0 y varianza σ^2 constante, la no correlación de los errores y la relación no lineal entre las variables de explicación. Por otra parte los estadísticos de resumen como t, F o R^2, los Coeficientes de Regresión y la Media Cuadrática del Error son sensibles a la presencia de valores aberrantes o atípicos. En este artículo se revisan los métodos más usados, para verificar el cumplimiento de estos supuestos, detectar la presencia de valores aberrantes y puntos de influencia a través de su implementación y correspondiente validación en un software estadístico especializado en la técnica de Regresión Lineal llamado ERLA (Estadística de Regresión Lineal Avanzada), desarrollado por estudiantes del Instituto de Ciencias Matemáticas de la ESPOL.

Palabras Claves: *Regresión Lineal, Supuestos, Adecuación del Modelo, Implementación, Validación, Software, ERLA.*

Abstract

In Linear Regression Analysis, there are assumptions and premises that must be considered to determine the validity of a model. Since non-compliance with any of these assumptions could lead to unstable models, if so, a high value of R2 and adjusted R2 does not guarantee that the model fits the data well. Among the major assumptions of Linear Regression are: The normal error distribution with mean 0 and constant variance σ^2, non-correlation of the errors and the non-linear relationship between the variables of explanation. Moreover summary statistics such as t, F or R^2, Regression Coefficients and Mean Square of Error are sensitive to the presence of outliers or atypical values. This article reviews the methods used to verify compliance with these assumptions, the presence of outliers and leverage points through its implementation and validation by using a Linear Regression-oriented software named ERLA (Advanced Linear Regression Statistics), developed by students of the Institute of Mathematical Sciences at ESPOL.

Keywords*: Linear Regression, Assumptions, Model Adequacy, Deployment, Validation, Software, ERLA.*

1. Introducción

La Regresión Lineal es una de las técnicas estadísticas más poderosas y versátiles utilizada en diversas áreas, entre ellas la medicina y los negocios, ya que esta técnica permite explorar y cuantificar la relación entre una variable llamada variable de respuesta, explicada o pronosticada *Y*, y una o más variables predictoras o variables de explicación X_1, X_2,…, X_{p-1} siempre y cuando éstas sean cuantitativas. Debido a la importancia que demandan las conclusiones a las cuales se llega después de obtener un modelo, resulta imprescindible evaluar la calidad del mismo. Para ello existen numerosos métodos estadísticos basados principalmente en el análisis de residuales, recuérdese que:
Dado el modelo de Regresión Lineal:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{(p-1)} x_{(p-1)i} + \varepsilon_i$$
$$Para\ i=1,2,\ldots,n$$

Se define al error

$$\varepsilon_i = y_i - E[y_i] \text{ donde}$$
$$E[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{(p-1)i}$$

Y a los residuales como los estimadores del error:

$$e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i,$$
$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_{p-1} x_{(p-1)i}$$

Además se trabaja bajo los supuestos de que:
- La distribución del error ε_i es normal con media cero.
- La varianza de ε_i es constante.
- Los errores ε_i y ε_j para $i \neq j$ no están correlacionados.
- Las variables de explicación X_i y X_j para $i \neq j$ no están correlacionadas.

No sólo el no cumplimiento de los supuestos antes descritos, puede afectar la calidad del modelo de regresión lineal sino también la inclusión de valores aberrantes o puntos de influencia. A continuación se describen los métodos más usados para identificar este tipo de inadecuaciones. Además se explica brevemente cómo aplicar éstos métodos en el Software estadístico de Regresión ERLA.

2. Normalidad

Entre los métodos que permiten verificar si la distribución del error es normal, se pueden mencionar los métodos de bondad de ajuste y el gráfico de probabilidad normal.

2.1. Gráfico de Probabilidad Normal

El Gráfico de Probabilidad Normal es un gráfico diseñado para que al graficarse la distribución normal acumulada se bosqueje una línea recta. Véase *Figura 1*

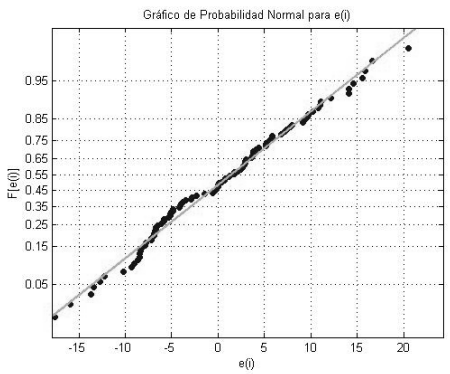


Figura 1. Gráfico de Probabilidad Normal

El eje vertical de ésta gráfica es construido a partir de la inversa acumulada de la normal estándar, razón por la que su valor varía entre 0 y 1. Véase *Figura 2*.

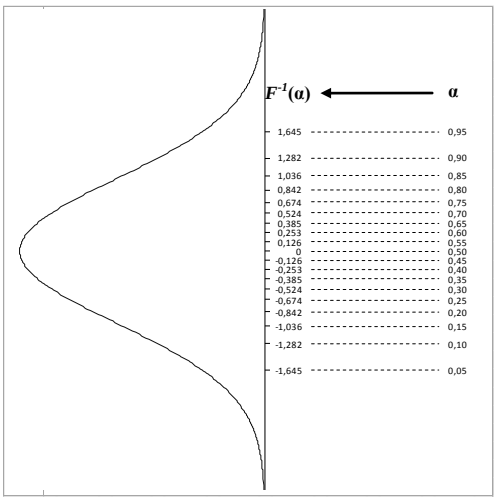


Figura 2. Escala de Probabilidad Normal

Sean $e_{[1]}, e_{[2]}, \ldots, e_{[n]}$ los residuales ordenados en forma ascendente. Si se grafica $e_{[i]}$ en función de la distribución acumulada $F(e_{[i]})=(i-0.5)/n$, para $i=1,2,\ldots,n$ en un gráfico de probabilidad normal, los puntos deberían aproximarse a una recta, si es que la distribución de probabilidad de los $e_{[i]}$ es normal.

3. Homocedasticidad - Varianza Constante

Una manera sencilla de verificar si la varianza del error ε_i es constante es realizando un gráfico de los residuales e_i contra los valores ajustados $\hat{y}_i$. Si la varianza es constante se esperaría que los errores fluctúen alrededor del eje horizontal, y que puedan ubicarse en una banda; caso contrario puede ser que la varianza no sea constante Véase *Figura* 3.

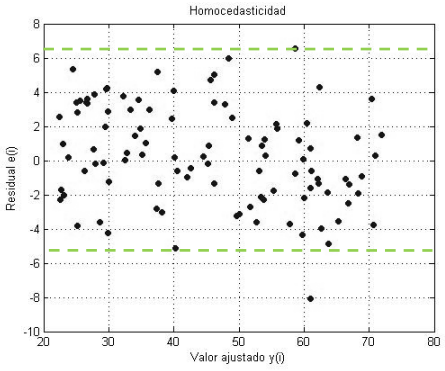


Figura 3. Residuales e_i vs. Valores ajustados $\hat{y}_i$

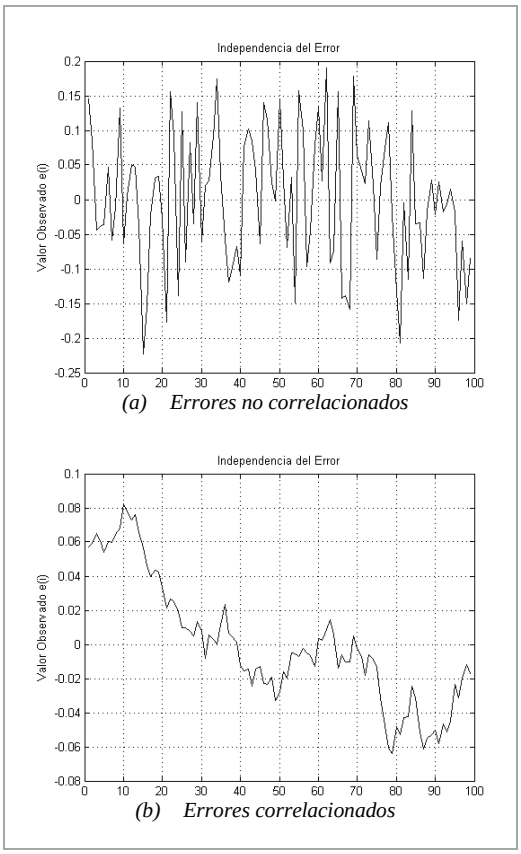
4. Errores no correlacionados

Otro de los supuestos que se realiza en un análisis de Regresión Lineal es que los errores no están

correlacionados, y; para verificar que este supuesto se cumpla se pueden usar dos métodos:

4.1. Gráfico de los residuales vs. Secuencia u orden

Al graficar los residuales de manera ordenada en el tiempo o espacio es posible detectar la presencia de correlación entre los errores. Si estos muestran algún tipo de patrón lineal o cíclico por ejemplo, los errores podrían estar correlacionados caso contrario no lo están. Véase *Figura 4*.



(a) Errores no correlacionados

(b) Errores correlacionados

Figura 4. Gráfico de residuales en el tiempo

4.2. Prueba Durbin-Watson

La prueba de Durbin-Watson es utilizada en Series de Tiempo para detectar Correlación Serial. Esta prueba se basa en la hipótesis de que los errores del modelo de regresión se generan en un proceso autorregresivo de primer orden, esto es:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i$$

donde δ_i es una variable aleatoria $N(0,\sigma^2)$ y ρ es el coeficiente de correlación. Ante esta situación Durbin y Watson plantearon la siguiente prueba unilateral:

$$H_o : \rho = 0 \quad \text{vs.} \quad H_1 : \rho > 0$$

y determinaron la región crítica de la prueba en base al estadístico:

$$d = \frac{\sum_{i=2}^{n}\left(e_i - e_{i-1}\right)^2}{\sum_{i=2}^{n} e_i^{\,2}}$$

Si la Hipótesis Nula de la prueba es verdadera, la distribución del estadístico *d* dependería de la matriz de diseño **X** y es desconocida. Sin embargo Durbin y Watson [1951] demostraron que *d* esta entre dos cotas d_L y d_U a través de las cuales se puede llegar a una conclusión respecto a la hipótesis nula planteada:

Si $d < d_L$, rechazar H_0
Si $d > d_U$, no rechazar H_0
Si $d_L \le d \le d_U$, la prueba no es concluyente.

Durbin y Watson tabularon los valores de los límites d_L y d_U para varios tamaños de muestra, diversas cantidades de regresores o variables de explicación y tres tasas de error tipo I (α=0.05, α=0.025, y α=0.01).

G. S. Maddala en el año de 1996 pudo probar que *d* es un valor comprendido entre 0 y 4. Véase *Figura 5*.

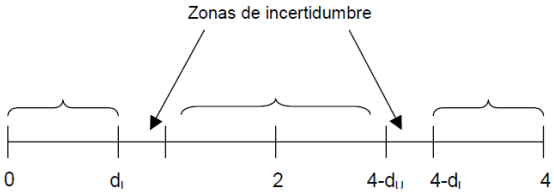


Figura 5. Región de Rechazo de la Prueba de Durbin-Watson

Así, si el valor del estadístico *d* es próximo a 2, ρ=0; si se aproxima a 4, ρ<0 y si se aproxima a 0 ρ>0.

5. Multicolinealidad - Variables de explicación no correlacionadas

Cuando existe una relación aproximadamente lineal entre las variables de explicación, es posible que los estimadores resultantes tengan varianzas muy grandes aunque siguen conservando la propiedad de insesgados, además se puede no rechazar la hipótesis nula de que un parámetro es cero, aun cuando la correspondiente variable sea relevante; y por último los coeficientes estimados serán muy sensibles a pequeños cambios en los datos.

Una forma de detectar multicolinealidad es calculando la matriz de correlación de las variables de explicación y ver qué pares de variables tienen correlación cercana a 1. Sin embargo existen métodos más formales como el Factor de Agrandamiento de la Varianza (FAV) y el Número de Condición.

5.1. Factor de Agrandamiento de la Varianza (FAV)

Si consideramos el modelo de regresión lineal múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{(p-1)} X_{(p-1)} + \varepsilon$$

para *i=1,2,...,n*. Entonces se puede probar que la varianza del *j-ésimo* coeficiente de regresión estimado es:

$$s_{\hat{\beta}_j} = \frac{\hat{\sigma}_2}{n(1-R_j^2)s_j^2}$$

para *j=1,2,...,p-1* y donde R_j^2 es el coeficiente de determinación obtenido al hacer la regresión de X_j sobre el resto de las variables de explicación del modelo, y s_j^2 es la varianza muestral de la variable X_j.

Si la correlación entre las variables de explicación fuera nula, la fórmula para estimar la varianza del j-ésimo coeficiente de regresión se reduciría a:

$$s_{\hat{\beta}_j} = \frac{\hat{\sigma}_2}{ns_j^2}$$

El FAV es la razón entre la varianza observada y la que habría sido en caso de que no estuviera correlacionada con el resto de las variables de explicación del modelo:

$$FAV = \frac{1}{1-R_j^2}$$

Es decir que el FAV mide cuanto crece la varianza del j-ésimo coeficiente de regresión como consecuencia de que las variables estén altamente correlacionadas. Una variable de explicación con un FAV entre 5 y 10 puede causar multicolinealidad.

5.2. Número de condición

El número de condición es la razón entre la raíz característica más grande $(\lambda_{\max})$ y la raíz característica más pequeña $(\lambda_{\min})$ de la matriz $\mathbf{X^T X}$, siendo $\mathbf{X}$ la matriz de diseño sin la columna de unos:

$$k(\mathbf{X}) = \frac{(\lambda_{\max})}{(\lambda_{\min})}$$

Recuérdese que la matriz $\mathbf{X^T X}$ es una matriz cuadrada y simétrica. El problema de la multicolinealidad es grave cuando el número de condición toma un valor mayor que 1000.

Entre las soluciones que pueden darse a la multicolinealidad están:

1. Eliminar del modelo las variables que tienen una correlación muy alta.
2. Incrementar el tamaño de la muestra
3. Regresión Ridge
4. Componentes principales
5. Mínimos Cuadrados Parciales

6. Valores aberrantes o atípicos

En todo análisis estadístico resulta importante detectar la presencia de valores aberrantes o atípicos, ya que éstos pueden afectar drásticamente a los estimadores, por ello existen varios criterios para su identificación basados en el análisis de residuales.

El Gráfico de los Residuales e_i en función de los valores ajustados $\hat{y}_i$ y el Gráfico de Probabilidad Normal son también útiles para detectar valores atípicos potenciales.

6.1. Residuales

Ya se habían definido antes los residuales como:

$$e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$$

Además, se puede probar que:

$$E(e_i) = 0$$

$$Var(e_i) = MCE$$

Se consideran valores atípicos potenciales, los residuales cuyo valor absoluto es mayor a tres desviaciones estándar respecto de la media. Se recomienda además analizar los residuales que se detallan a continuación

6.1.1. Residuales Estandarizados. Ya que la varianza aproximada del error se estima con la MCE, los residuales estandarizados serán:

$$d_i = \frac{e_i}{\sqrt{MCE}}$$

Los residuales estandarizados tienen media cero y varianza aproximadamente unitaria. Un residual estandarizado mayor que 3 indica que la observación *i-ésima* es un valor atípico potencial.

6.1.2. Residuales Estudentizados. Sea H, la conocida Matriz Hat definida como:

$$\mathbf{H = X(X^T X)^{-1} X^T}$$

y h_{ij} sus elementos; además sea **e** el vector de residuales, se puede probar que:

$$Var(\mathbf{e}) = \sigma^2 \left(\mathbf{I - H}\right)$$

Esto quiere decir que:

$$Var(e_i) = \sigma^2(1-h_{ii}) \text{ y } Cov(e_i, e_j) = -\sigma^2(h_{ij})$$

Por tanto se definen los residuales estudentizados dividiendo el *i-ésimo* residual entre su desviación estándar "exacta":

$$r_i = \frac{e_i}{\sqrt{MCE(1-h_{ii})}}$$

Los residuales estandarizados y estudentizados aportan con frecuencia información equivalente. En conjuntos grandes de datos los residuales estandarizados no serán muy diferentes de los estudentizados. Un residual estudentizado r_i mayor que 3 indica la presencia de un valor atípico potencial.

6.1.3. Residuales PRESS. Los residuales PRESS o residuales de predicción se definen como la diferencia entre el valor observado y_i para $i = 1, 2, ..., n$ y el valor estimado de esta observación basado en todas las observaciones excepto esta *i-ésima*:

$$e_{[i]} = y_i - \hat{y}_{[i]}$$

Es decir, se elimina la *i-ésima* observación y se ajusta el modelo de regresión a las *n-1* observaciones restantes, para estimar y_i.

Se puede probar que existe una relación entre los residuales PRESS y los residuales usuales:

$$e_{[i]} = \frac{e_i}{1 - h_{ii}}$$

Una gran diferencia entre el residual ordinario y el residual PRESS indica un valor atípico potencial.

7. Puntos de Influencia

Los Puntos de Influencia o valores influyentes son aquellos que tienen un impacto notable sobre los coeficientes del modelo, por ello la importancia de localizarlos.

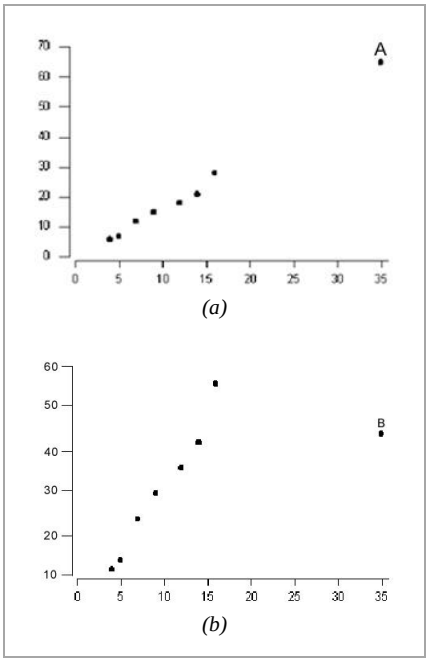


Figura 6. Puntos de Influencia

La mayoría de los textos llaman "valores aberrantes" a un valor alejado solamente en la dirección vertical y Punto de influencia a una observación alejada en la dirección horizontal. Véase *Figura 6*. El punto A no afecta las estimaciones de los coeficientes de regresión, mientras el punto B si tiene un impacto notable en la estimación de estos coeficientes puesto que atrae a la recta de regresión en su dirección.

A continuación se presentan dos métodos para detectar puntos de influencia:

7.1. Apalancamiento

Sea H, la antes mencionada Matriz Hat definida como:

$$\mathbf{H = X(X^T X)^{-1} X^T}$$

La varianza del vector de estimaciones puede escribirse como:

$$Var(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{H}$$

Los elementos h_{ij} de la matriz **H** son una medida de lugar o ubicación del i-ésimo punto en el espacio de x, por lo tanto son vistos como la cantidad de "balanceo" o "apalancamiento" de la i-ésima observación yi sobre el i-ésimo valor ajustado $\hat{y}_i$. Por esta razón, los valores grandes en la diagonal de la matriz **H** indican observaciones que son potencialmente influyentes, esto es; valores de $h_{ii} > 2p/n$ lo cual no aplica para casos donde $2p/n > 1$.

7.2. Distancia de Cook

La distancia de Cook mide el cambio que ocurriría en el vector de coeficientes estimado de regresión si la i-ésima observación fuera omitida. Esta distancia se define como:

$$CD_i^2 = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]})}{pMCE}$$

Dónde:

- $\hat{\boldsymbol{\beta}}$ es el vector de coeficientes estimado con el modelo completo
- $\hat{\boldsymbol{\beta}}_{[i]}$ es el vector de coeficientes estimado sin la *i-ésima* observación
- **X** es la matriz de Diseño
- MCE es el estimador de σ^2
- *p* es el número de parámetros en el modelo

Sea P_i el *i-ésimo* punto para *i=1,2,…,n* de *p* coordenadas. Dado el siguiente contraste de Hipótesis:

H_0 : P_i no es un punto de influencia

vs.

$H_1: P_i$ es un punto de influencia

Con (1-α)100% de confianza se rechaza H_0 a favor de H_1 si el estadístico CD_i^2 es mayor que $F_{(\alpha,p,n-p)}$.

8. Software estadístico de Regresión ERLA

ERLA es un software estadístico especializado en la técnica de regresión lineal, desarrollado por estudiantes del Instituto de Ciencia Matemáticas mediante el uso del MCR (MATLAB Component Runtime) y VisualBasic.NET. A continuación se presenta cómo obtener un modelo de regresión lineal en ERLA, y cómo evaluar la calidad del mismo utilizando los métodos antes mencionados.

Figura 7. Inicio ERLA

8.1 Regresión Lineal en ERLA

Para explicar cómo se realiza un análisis de Regresión Lineal en ERLA, se ha considerado, guardando la correspondiente confidencialidad que la ética estadística exige, una base de datos correspondiente a un estudio realizado en la Escuela Superior Politécnica del Litoral por el Centro de Estudios e Investigaciones Estadísticas, llamado "Imagen de la ESPOL en Guayaquil". Este estudio presenta un total de 12 proposiciones calificadas en una escala del 1 al 10. Al obtener la matriz de correlación de las proposiciones, se encontró que P9 (*Identifico a los estudiantes de la ESPOL por su responsabilidad*) y P10 (*Identifico a los estudiantes de la ESPOL por su honestidad*) están altamente correlacionadas. Para obtener el modelo de regresión que explique a P10 en términos de P9 se sigue la secuencia:

1. Barra de menues➤ Análisis de datos➤ Regresión➤Regresión lineal

2. Seleccione la variable a ser explicada y las variables de explicación en el cuadro de diálogo *Regresión Lineal* (*Véase Figura 8*), luego seleccione los indicadores para evaluar la calidad del modelo a través del botón **Opcione**s. Véase *Figura 9*.

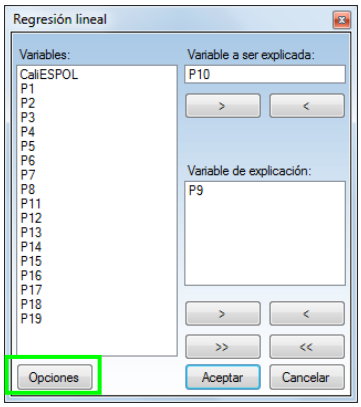


Figura 8. Cuadro de Diálogo Regresión Lineal ERLA

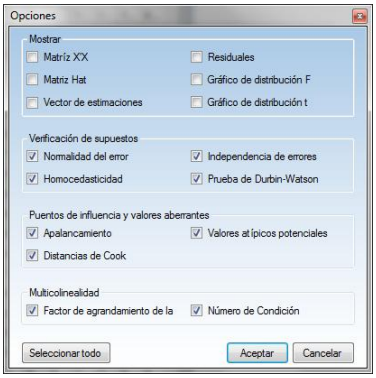


Figura 9. Cuadro de Diálogo Opciones de Regresión Lineal ERLA

En el cuadro de diálogo *Opciones* se seleccionan todos los ítems de "Verificación de supuestos", "Puntos de influencia" y "Valores aberrantes" y "Multicolinealidad". Los resultados se muestran en la *Figura 10*.

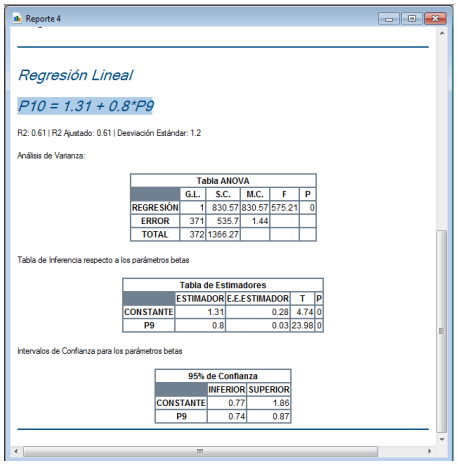


Figura 10. Modelo de Regresión Lineal obtenido en ERLA

Estos resultados son:

1. El modelo.
2. La potencia de explicación del modelo (R^2 y R^2 Ajustado).
3. La desviación estándar (s).
4. La tabla ANOVA con todos sus valores característicos: Fuentes de variación (1era columna), grados de libertad (G.L.), sumas y medias cuadráticas (S.C. y M.C., respectivamente), el estadístico de prueba F (F) y el valor p (P).
5. La tabla de inferencia respecto a los parámetros betas. El valor del estimador (ESTIMADOR), el error estándar del estimador (E. E. ESTIMADOR), el estadístico de prueba t (T) y el valor p (P).
6. Los intervalos de confianza para los parámetros betas utilizando un nivel de confianza del 95%. Se puede distinguir en los resultados de la tabla el límite inferior (INFERIOR) y el límite superior (SUPERIOR).

Por otra parte los resultados que pueden obtenerse para evaluar la calidad del modelo:

8.1.1. Normalidad del error. ERLA muestra el Gráfico de probabilidad normal de los residuales:

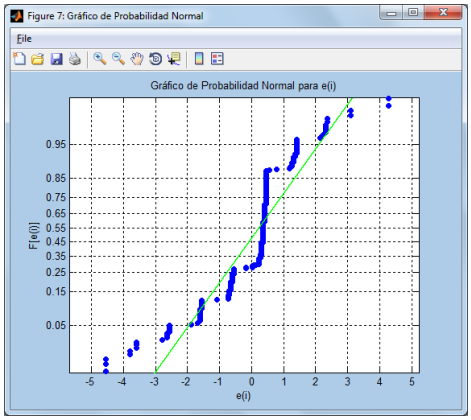


Figura 11. Gráfico de Probabilidad Normal

8.1.2. Homocedasticidad. ERLA muestra el Gráfico de Residuales vs. Valores Ajustados. Véase *Figura 12*.

8.1.3. Errores no correlacionados. ERLA muestra el Gráfico de los residuales en vs. secuencia/orden. Véase *Figura 13*. Además presenta el estadístico de Durbin Watson para una prueba de dos colas con su respectivo valor *p*. Véase *Figura 14*.

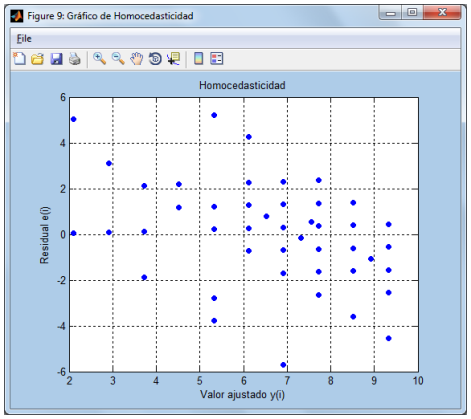


Figura 12. Gráfico de Residuales vs. Valores Ajustados

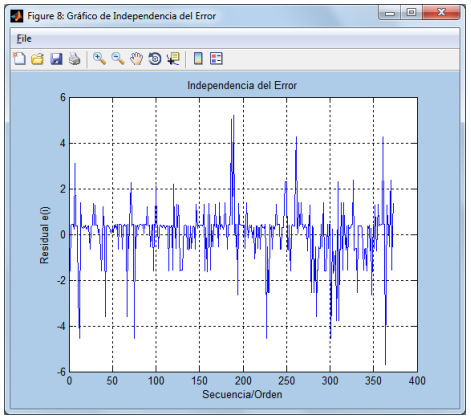


Figura 13. Gráfico de los Residuales vs. Secuencia/Orden

Prueba de Durbin-Watson

Estadístico D-W: 1.75

Valor p: 0.01

Figura 14. Prueba de Durbin-Watson

8.1.4. Multicolinealidad. ERLA presenta el Factor de agrandamiento de la varianza (FAV) para cada una de las variables de explicación incluidas en el modelo, y también el Número de Condición. Véase *Figuras 15 y 16*.

Factor de Agrandamiento de la Varianza

Factor de Agrandamiento de la Varianza		
	R2	FAV
P13	0.58	2.39
P14	0.58	2.39

Figura 15. Factor de Agrandamiento de la Varianza

Número de Condición

Número de condición = 66.38

Figura 16. Número de Condición

8.1.5. Valores Aberrantes. ERLA presenta los valores aberrantes potenciales en una tabla con sus correspondientes residuales. Véase *Figura 17*.

Valores atípicos potenciales

Las siguientes observaciones son consideradas valores atípicos potenciales:

Valores atípicos potenciales				
Observaciones	Residuales	Residuales Estandarizados	Residuales Estudentizados	Residuales Press
12	-4.32	-3.6	-3.67	-4.5
75	-4.32	-3.6	-3.67	-4.5
187	3.89	3.23	3.35	4.18
189	4.68	3.9	4	4.93
227	-4.32	-3.6	-3.67	-4.5
261	3.88	3.23	3.29	4.02
301	-4.32	-3.6	-3.67	-4.5
307	-4.32	-3.59	-3.68	-4.52
310	-4.32	-3.59	-3.68	-4.52
361	3.88	3.23	3.29	4.02
364	-5.92	-4.93	-5.11	-6.36

Figura 17. Valores Aberrantes

8.1.5. Puntos de Influencia. Para detectar puntos influyentes, ERLA presenta el vector de apalancamientos junto al vector que contiene las distancias de Cook como nuevas variables en la ventana de datos. Véase *Figura 18*.

C:\Users\...\Desktop\MATERIA DE GRADUACION\Ejemplos\IMAGEN_ESPOL.csv

	APALANCAMIENTO	RESIDUALES	RES_ESTANDARIZADOS	RES_ESTUDENTIZADOS	RES_PRESS	DISTANCIA_DE_COOK
1	0.00	-1.52	-1.27	-1.27	-1.53	0.00
2	0.00	0.28	0.23	0.23	0.28	0.00
3	0.01	0.68	0.56	0.57	0.68	0.00
4	0.01	0.68	0.56	0.57	0.68	0.00
5	0.01	0.68	0.56	0.57	0.68	0.00
6	0.01	-0.32	-0.26	-0.27	-0.32	0.00
7	0.03	2.08	1.73	1.77	2.17	0.05
8	0.00	0.28	0.23	0.23	0.28	0.00
9	0.00	0.28	0.23	0.23	0.28	0.00

Figura 18. Puntos de Influencia

9. Conclusiones y recomendaciones

9.1. Conclusiones

1. El no cumplimiento de los supuestos en un análisis de regresión lineal hace que los estimadores de los coeficientes del modelo dejen de ser eficientes, los intervalos de confianza y las pruebas de hipótesis basadas en las distribuciones t y F dejan de ser confiables. El modelo se vuelve inestable, en el sentido de que muestras diferentes pueden conducir a modelos diferentes.

3. La presencia de valores aberrantes y puntos de influencia en un modelo de regresión lineal pueden disminuir la potencia de explicación del modelo.

3. Para el caso del ejemplo, está claro que el modelo no cumple el supuesto de normalidad del error, esto definitivamente hace que el modelo no sea del todo confiable.

9.2. Recomendaciones

1. Siempre debe verificarse el cumplimiento de los supuestos o premisas bajo los cuales se trabaja en un análisis de regresión, puesto que la calidad del modelo encontrado puede verse afectada y las conclusiones finales pueden ser erradas.

2. Se debe tener mucho cuidado si quiere eliminar valores aberrantes y puntos de influencia, ya que estos no siempre provienen de un error de medición o digitación y en estos casos debe considerarse el uso de técnicas robustas de estimación que no sean tan sensibles a puntos influyentes como lo son los mínimos cuadrados.

10. Referencias Bibliográficas

[1] MONTGOMERY, D. (2002), "Introducción al Análisis de Regresión Lineal", Editorial Continental, México-México.

[2] SEBER, A. & LEE, A. (2003), "Linear Regression Analysis", (2da Edición), Editorial Wiley, New York – U.S.A.

[3] GUJARATI, D. (2004), "Econometría Básica", (4ta Edición), Editorial Mc Graw Hill, México-México.

[4] MORILLAS, A. & DÍAZ, B. (2007), "El Problema de los Outliers Multivariantes en el Análisis de Sectores Clave y Cluster Industrial", Universidad de Málaga, España.

[5] ZURITA, G. (2010), "Probabilidad y Estadística: Fundamentos y Aplicaciones", (2da Edición), Talleres Gráficos ESPOL, Guayaquil-Ecuador.

[6] ACUÑA FERNÁNDEZ, E. "Diagnósticos de Regresión", Universidad de Puerto Rico, obtenido en agosto de 2010 desde http://math.uprm.edu/~edgar/cap3sl.ppt

[7] ACUÑA FERNÁNDEZ, E. "Multicolinealidad", Universidad de Puerto Rico, obtenido en agosto de 2010 desde http://math.uprm.edu/~edgar/cap7sl.ppt

[8] RAMIREZ, D. "Autocorrelación", obtenido en septiembre de 2010 desde fhttp://webdelprofesor.ula.ve/economia/dramirez/MICRO/FORMATO_PDF/Materialeconometria/Autocorrelacion.pdf.