Revista Colombiana de Estadística
Volumen 24 (2001) N° 1, páginas 33 *a* 43

COMPARACIÓN DE TRES MÉTODOS DE REGRESIÓN LINEAL USANDO PROCEDIMIENTOS DE SIMULACIÓN

JUAN C. TORRES C.*

Resumen

Cuando desea ajustarse un modelo lineal a un conjunto de datos, el método de regresión usualmente más empleado es el de mínimos cuadrados. Este método es óptimo si la distribución de los residuos es gaussiana. Existen casos en donde el supuesto de normalidad en los residuales no se cumple y se hace necesario el uso de métodos alternativos de regresión, como la regresión vía mínimas desviaciones absolutas (LAD) o la regresión no paramétrica basada en rangos, los cuales no requieren de supuestos distribucionales sobre los residuos y permiten obtener una mejor estimación de los parámetros del modelo de regresión.
Palabras clave: Modelo lineal; análisis de regresión; simulación.

Abstract

When it's necessary to fit a lineal model to a data set, least squares regression method is usually used. This method is optimum if the residuals distribution is normal. When the assumption of residuals normality doesn't comply it's necessary to use alternative regression methods, as Least absolute deviations (LAD) or Non parametric regression based on ranks, which don't need the assumption about the residuals distribution and allow a better estimation of regression model parameters.
Key words: Linear model, regression analysis, simulation.

*Estadístico egresado de la Universidad Nacional de Colombia y Consultor del Departamento Administrativo Nacional de Estadística DANE. Correo electrónico: jctorresc@dane.gov.co

1. Introducción

El objetivo de este artículo es determinar cuál de los tres métodos de regresión: Mínimo cuadrática, No paramétrica basada en rangos o Mínima desviación absoluta, presenta un mejor ajuste cuando la distribución de los errores aleatorios tiene forma diferente de la gaussiana en elongación y cuando hay presencia de observaciones atípicas u "outliers". Para cumplir con el objetivo, se calcularon estimaciones de los coeficientes del modelo de regresión y sus medidas de ajuste para los tres métodos de regresión mediante procedimientos de simulación realizados con el paquete estadístico SAS.

2. Regresión Mínimo Cuadrática

Para el modelo dado por la expresión $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, el método construye el estimador $\hat{\boldsymbol{\beta}} = \boldsymbol{(X'X)^{-1}X'Y}$, que minimiza la suma de cuadrados de los errores. Para un conjunto de datos observados, cuando la expresión $f(\boldsymbol{b}) = \sum (Y_i - \boldsymbol{x_i b})^2$ se hace mínima, el vector de valores $\boldsymbol{b}$ se conoce como la estimación mínimo cuadrática de $\boldsymbol{\beta}$. En la función anterior $\boldsymbol{x_i}$ representa la i-ésima fila de la matriz $\boldsymbol{X}$. Se estima la varianza de la población de los errores con $s^2 = \sum e_i^2/(n-p-1)$, en donde los e_i son los errores obtenidos con los datos observados, n es el número de observaciones realizadas y p es el número de variables explicativas en el modelo.

3. Regresión no Paramétrica Basada en Rangos

El método se basa en el modelo lineal $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. No existen fórmulas explícitas para el estimador $\widehat{\beta}$. Sin embargo, mediante un algoritmo iterativo es posible obtener el vector de estimación de $\boldsymbol{\beta}$. Para un conjunto de datos observados, el vector $\beta = [b_0\ \boldsymbol{b}]$ se conoce como la estimación no paramétrica de $\boldsymbol{\beta}$, en donde el vector $\boldsymbol{b}$ de tamaño $1 \times p$, minimiza la función

$$g(\boldsymbol{b}) = \sum \Big[rango(y_i - \boldsymbol{x_i b}) - \frac{n+1}{2} \Big] (y_i - \boldsymbol{x_i b})$$

en donde $\boldsymbol{b} = [b_1, b_2, \cdots, b_p]'$ y $\boldsymbol{x_i} = [x_{i1}, x_{i2}, \cdots x_{ip}]$ y la estimación no paramétrica de b_0 se obtiene como la mediana de las diferencias de $y_i - \boldsymbol{x_i b}$. Ya que no existen fórmulas explícitas para calcular los coeficientes estimados, Birkes y Dodge (1993) describen un algoritmo para estimarlos. De igual manera, los mismos autores describen un algoritmo para estimar la desviación estándar de

los errores, cuya notación usual es τ. Además afirman que un estimador para la desviación estándar de los errores es $\tilde{s} = \tau/\ 1.023$ si la distribución de los errores es normal.

4. Regresión Vía Mínima Desviación Absoluta

El método se basa en el modelo lineal $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. No existen fórmulas explícitas para el estimador $\widehat{\beta}$. Sin embargo, mediante un algoritmo iterativo es posible obtener el vector de estimación de $\boldsymbol{\beta}$ el cual es escogido de tal forma que la suma de los valores absolutos de los errores, $\sum |e|$, sea la más pequeña posible. Para un conjunto de datos observados, cuando la expresión $f(\boldsymbol{b}) = \sum |y_i - \boldsymbol{x_i b}|$ se hace mínima, el vector de estimación $\boldsymbol{b}$ obtenido, es conocido como la estimación mínima desviación absoluta de $\boldsymbol{\beta}$. En la función anterior $\boldsymbol{x_i}$ representa la i-ésima fila de la matriz $\boldsymbol{X}$. Como no existen fórmulas explícitas para calcular los coeficientes estimados, Birkes y Dodge (1.993) describen un algoritmo para estimar los coeficientes de regresión. La desviación estándar de los errores se estima con $\tilde{s} = 1{,}483 * MAD$, donde MAD es la mediana de las desviaciones absolutas con respecto a la mediana de los residuales, calculado con los residuos diferentes de cero.

5. Distribuciones g y h de Tukey

La familia de distribuciones g y h comprende una considerable variedad de distribuciones con características especiales en cuanto a asimetría y elongación, por lo cual resulta de gran utilidad cuando se desea simular datos que provengan de distribuciones con formas diferentes a la distribución normal.

5.1. Asimetría

Si Z es una variable normal estándar y g es una constante real, la variable aleatoria $Y_g(Z)$, definida como $Y_g(Z) = (e^{gz} - 1)g^{-1}$ se dice que tiene la distribución g de Tukey para un valor dado de g. El parámetro g controla la magnitud y la dirección de la asimetría.

5.2. Elongación

Si Z es una variable aleatoria normal estándar y h es una constante real, la variable aleatoria $Y_h(Z)$ dada por $Y_h(Z) = Ze^{hZ^2/2}$ se dice que tiene distribución h de Tukey para un valor dado de h. El parámetro h controla la cantidad de la elongación de la distribución. Las distribuciones de la familia h de Tukey son simétricas y su valor esperado es cero.

6. Proceso de Simulación

Para el cálculo de las estimaciones del ajuste con errores aleatorios provenientes de la familia de distribuciones h de Tukey, se efectuaron 1000 simulaciones de modelos lineales con una y dos variables independientes, con tamaños de muestra 20 y 50, y usando cada uno de los tres métodos de regresión. Se utilizó el mismo procedimiento para errores provenientes de una distribución normal contaminada con un porcentaje de "outliers" dado. Todas las simulaciones fueron llevadas a cabo con el paquete estadístico SAS.

En el caso de la regresión simple, el modelo simulado fue $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \boldsymbol{e}$ donde $\beta_0 = 4$ y $\beta_1 = 5$ y la variable explicativa X_1 se dejó fija en cada simulación y fue generada de una distribución uniforme $U(2, 5)$ para los tamaños de muestra 20 y 50. Para la regresión múltiple, el modelo simulado fue $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boldsymbol{e}$ donde $\beta_0 = 3$, $\beta_1 = 4$, y $\beta_2 = 5$. Las variables explicativas X_1 y X_2 se fijaron para cada simulación y fueron generadas de distribuciones $U(1, 4)$ y $U(2, 5)$ respectivamente, para los tamaños de muestra 20 y 50. Para la simulación de errores aleatorios con formas diferentes a la distribución normal se utilizaron las siguientes distribuciones:

Tabla 1: Distribuciones h de Tukey
Simuladas con su varianza teórica

h	$V(e)$
0	1
0.2	2.1517
0.4	11.1803
0.6	Infinita
0.8	Infinita
1	Infinita

En la simulación de errores aleatorios con distribución normal y presencia de "outliers", se utilizó el modelo contaminado $f(x) = (1-\epsilon)\phi(x) + \epsilon\phi(x/k)$ donde $\phi(x)$ es la función de distribución normal estándar y k es igual a 3. Los valores de contaminación ϵ que se utilizaron fueron 0.05, 0.10, 0.15 y 0.20, lo que equivale a contaminar la distribución $N(0,1)$ con los porcentajes de "outliers" $5\,\%, 10\,\%, 15\,\%$ y $20\,\%$, generados de una distribución $N(0,3)$, para cada uno de los modelos descritos anteriormente.

7. Criterios de Comparación.

Los criterios de comparación utilizados fueron los siguientes:

- Error cuadrático medio de los estimadores de los coeficientes del modelo (ECM)
- Error absoluto medio de los estimadores de los coeficientes del modelo (MAE)
- Promedios de los coeficientes de determinación estimados.

Cuando se simulan los errores de una distribución normal con un porcentaje de "outliers", se compararon los promedios de las cantidades de "outliers" detectados que fueron generados para cada método de regresión, usando el método de los residuales estandarizados. Según Birkes y Dodge (1993), un residual estandarizado mayor a 2.5 en valor absoluto, puede catalogarse como *"outlier"*.

8. Resultados de las Simulaciones

Para el análisis de los resultados, se compararon el error cuadrático medio (ECM) y el valor absoluto medio (MAE) de los estimadores de los parámetros obtenidos con cada uno de los tres métodos de regresión: mínimos cuadrados, no paramétrica basada en rangos y LAD. Se exponen solamente los resultados para regresión simple ya que los correspondientes a regresión múltiple con dos variables son similares. Así mismo, se presentan los resultados para el error cuadrático medio (ECM) de los estimadores, ya que los resultados para el error absoluto medio (MAE) son similares (Ver Pulido y Torres, 1997). En cada una de las tablas presentadas los datos en letra negrilla son los menores valores de las medidas de comparación, los que poseen letra cursiva son los que siguen

en el orden de menor a mayor y los datos en letra normal corresponden a los mayores valores.

Tabla 2: Resultados de Regresión simple $n = 20$

Distribución		β_0			β_1		
		MC	NP	LAD	MC	NP	LAD
normal		**1.1716**	*1.2739*	1.8169	**0.0448**	*0.0472*	0.0688
h	0.2	2.8079	**1.8567**	*2.2316*	0.1047	**0.0708**	*0.0869*
	0.4	12.2780	**2.4435**	*2.6841*	0.3529	**0.0972**	*0.1051*
	0.6	60.3540	*3.3256*	**3.2183**	2.6535	*0.1332*	**0.1269**
	0.8	828.1600	*4.2417*	**3.2424**	27.5080	*0.1738*	**0.1318**
	1	5432.562	*5.4006*	**3.6525**	174.5601	*0.2148*	**0.1440**
% outliers	5	*1.7075*	**1.5531**	2.1669	*0.0652*	**0.0588**	0.0827
	10	*2.1565*	**1.8126**	2.4468	*0.0835*	**0.0687**	0.0938
	15	2.7169	**1.9623**	*2.4529*	0.1058	**0.0755**	*0.0965*
	20	3.4506	**2.5211**	*3.0297*	0.1345	**0.0963**	*0.1164*

Tabla 3: Resultados de regresión simple $n = 50$

Distribución		β_0			β_1		
		MC	NP	LAD	MC	NP	LAD
normal		**0.4407**	*0.483*	0.7151	**0.0154**	*0.0166*	0.0248
h	0.2	1.0208	**0.676**	*0.7780*	0.0356	**0.0235**	*0.0271*
	0.4	3.4151	**0.7573**	*0.7885*	0.1207	**0.0260**	*0.0271*
	0.6	17.2182	*0.9823*	**0.8817**	0.7438	*0.0341*	**0.0306**
	0.8	236.1500	*1.1016*	**0.8631**	7.7681	*0.0382*	**0.0295**
	1	1546.770	*1.3799*	**0.9745**	49.4998	*0.0485*	**0.0338**
% outliers	5	*0.5962*	**0.5575**	0.7769	*0.0207*	**0.0187**	0.0267
	10	*0.8799*	**0.6575**	0.8821	*0.0307*	**0.0226**	0.0309
	15	0.9968	**0.6949**	*0.9055*	0.0342	**0.0238**	*0.0316*
	20	1.1307	**0.7779**	*1.0061*	0.0391	**0.0271**	*0.0359*

8.1. Resultados de la Estimación de Parámetros

Los resultados se presentan en las Tablas 2 y 3. En ellas puede observarse que el ECM de los estimadores de los parámetros tiene un comportamiento similar para los tamaños de muestra 20 y 50.

8.1.1. Vector de errores con distribución normal

Observando las tablas 2 y 3, para los tamaños de muestra 20 y 50, el menor ECM de los estimadores de los parámetros se obtuvo con la regresión mínimo cuadrática y sigue en orden ascendente la regresión no paramétrica basada en rangos.

8.1.2. Vector de error con distribución h de Tukey

Se observa en la Figura 1, realizada para el ECM de β_1, regresión simple y tamaño de muestra 50, que cuando el parámetro h de la distribución de los errores es igual 0.2, 0.4, 0.6, 0.8 y 1, el ECM de los estimadores calculados con la regresión de mínimos cuadrados aumenta aceleradamente, mientras que con la regresión LAD y la regresión no paramétrica basada en rangos es más estable.

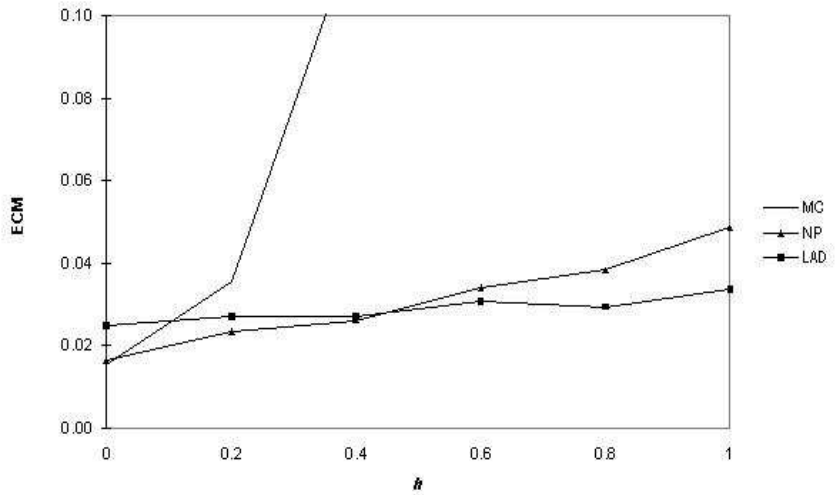

Figura 1. Error cuadrático medio. Estimador de β_1.
Errores familia h de Tukey

En la figura también se observa que para distribuciones con parámetros h igual a 0.2 y 0.4 el menor ECM de los parámetros estimados fue obtenido con la regresión no paramétrica basada en rangos y le sigue la regresión LAD. Para distribuciones h igual a 0.6, 0.8 y 1 el menor ECM de los parámetros estimados fue obtenido con la regresión LAD y le sigue la regresión no paramétrica basada en rangos.

8.1.3. Distribución de errores normal con un porcentaje de "outliers"

Como se observa tanto en la Figura 2, realizada para β_1, con el modelo simple y tamaño de muestra 50, así como en la Tabla 2, el menor ECM de los estimadores de los parámetros para los porcentajes de "outliers" simulados, fue obtenido con la regresión no paramétrica basada en rangos.

En orden ascendente siguen las estimaciones obtenidas con la regresión mínimo cuadrática, con un 5 % y 10 % de "outliers" en la distribución de los errores, y cuando se tiene un 15 % y 20 % de "outliers" en la distribución de los errores, le sigue el ECM de los estimadores calculados vía regresión LAD.

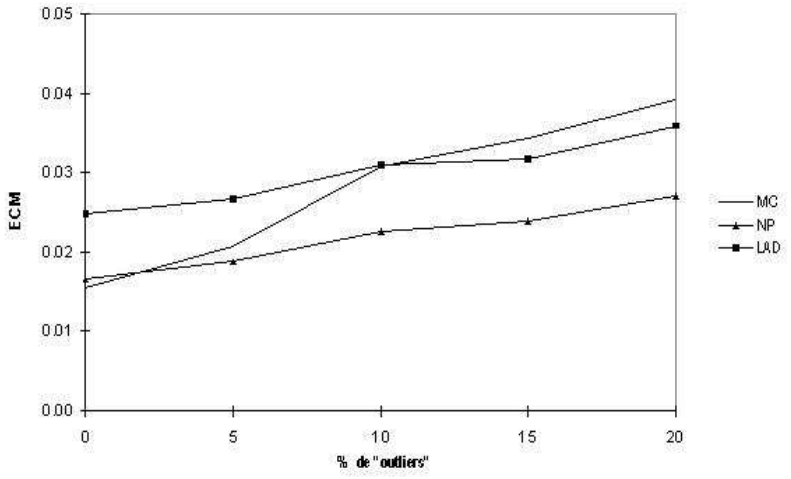

Figura 2. Error cuadrático Medio para β_1.
Errores con porcentaje de "outliers"

8.2. Resultados de Generación y Detección de Outliers

Como se puede observar en la tabla 4, el método que más " outliers" detectó en promedio, usando el método de los residuales estandarizados fue la regresión LAD y le sigue la regresión no paramétrica basada en rangos, para los diferentes porcentajes de "outliers" generados.

Tabla 4: Resultados, Generación y Detección de Outliers
Regresión Simple (Promedio)

		Generados	Detectados		
n	%		MC	NP	LAD
20	5	0.966	**0.223**	*0.266*	0.337
	10	1.981	**0.336**	*0.462*	0.583
	15	3.055	**0.440**	*0.637*	0.897
	20	3.994	**0.545**	*0.755*	1.112
50	5	2.473	**0.736**	*0.838*	0.910
	10	5.146	**1.212**	*1.596*	1.754
	15	7.603	**1.410**	*2.152*	2.444
	20	9.864	**1.589**	*2.552*	2.950

8.3. Resultados de la Medida de Ajuste R^2

En la Tabla 5 se puede observar que los valores de R^2 son mayores con la regresión mínimo cuadrática y sigue la regresión no paramétrica basada en rangos, para las diferentes distribuciones y los tamaños de muestra usados en la simulación.

9. Conclusiones

- Cuando la distribución de los errores es simétrica, de colas pesadas, generada de las distribuciones h de Tukey, se obtienen mejores estimaciones de los parámetros del modelo con los métodos de regresión no paramétrica basada en rangos y LAD que con la regresión mínimo cuadrática. Si la distribución de los errores es simétrica de colas muy pesadas, las mejores estimaciones se obtienen con la regresión LAD.

Tabla 5: Valores de R^2 para Regresión Simple (Promedio)

		20			50		
Distribución		MC	NP	LAD	MC	NP	LAD
normal	0	0.968	*0.967*	**0.966**	0.970	*0.970*	**0.970**
	0.2	0.935	*0.932*	**0.931**	0.939	*0.938*	**0.938**
	0.4	0.856	*0.847*	**0.845**	0.849	*0.845*	**0.845**
h	0.6	0.728	*0.707*	**0.705**	0.677	*0.667*	**0.666**
	0.8	0.584	*0.575*	**0.543**	0.483	*0.466*	**0.4650**
	1	0.482	*0.436*	**0.432**	0.328	*0.303*	**0.302**
	5	0.957	*0.955*	**0.954**	0.959	*0.958*	**0.958**
%	10	0.944	*0.942*	**0.941**	0.947	*0.946*	**0.946**
outliers	15	0.931	*0.928*	**0.926**	0.936	*0.935*	**0.935**
	20	0.919	*0.916*	**0.914**	0.927	*0.926*	**0.925**

- Se puede observar la robustez de los métodos de regresión LAD y no paramétrica basada en rangos, frente a la regresión mínimo cuadrática, cuando la distribución de los errores presenta colas pesadas.

- Cuando la distribución de los errores es normal, con un porcentaje de "outliers" mayor o igual al 5 %, se obtienen mejores estimaciones de los parámetros con la regresión no paramétrica y le sigue la regresión LAD.

- El empleo del coeficiente de determinación usual no es adecuado para comparar el ajuste de modelos con criterios de estimación diferentes, ya que para todas las distribuciones simuladas, éste fue mayor para la regresión mínimo cuadrática a pesar de que no siempre tuvo la mejor estimación de los parámetros del modelo de regresión.

- En cuanto a la detección de "outliers", cuando la distribución de los errores es normal, el método que detectó más en promedio usando el criterio de los residuales estandarizados, fue la regresión LAD, seguido de la regresión no paramétrica basada en rangos.

- En general, se puede observar que ninguno de los tres métodos de regresión en estudio fue uniformemente mejor para la estimación de los parámetros. Si la distribución de los errores aleatorios simétrica de colas muy pesadas, es mejor estimar los parámetros con la regresión LAD,

mientras que en el caso de que los errores tengan distribución normal con altos porcentajes de "outliers" es preferible estimar los parámetros usando la regresión no paramétrica basada en rangos.

Agradecimientos El autor expresa sus agradecimientos al Doctor Jorge Martínez Collantes, por sus valiosos aportes en la elaboración de este documento. Este artículo se basó en el trabajo de grado del mismo nombre realizado por Elsa Leonor Pulido y Juan Carlos Torres en 1.997, para optar al título de Estadístico en la Universidad Nacional de Colombia.

Referencias

[1] BIRKES, D. y DODGE, Y. (1993). *Alternative methods of regression.* John Wiley & Sons, Inc.

[2] PULIDO, E. y TORRES, J.(1.997). *Análisis comparativo de tres métodos de regresión: mínimos cuadrados, no paramétrica basada en rangos y mínima desviación absoluta, usando métodos de simulación.* Tesis de Grado. Universidad Nacional de Colombia.