Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Estadística Aplicada II
Cali, mayo de 2018

Tarea 5: Diagnósticos en el Análisis de Regresión,Regresión a Través del Origen y Regresión Robusta

Kevin Steven García[a], Alejandro Vargas[b]

1. Introducción

En este trabajo se tratara de encontrar un modelo lineal adecuado, para predecir la cantidad de $PM_{2.5}$ a partir de la cantidad de PM_{10}, para ello se ajustara un modelo simple por MCO y se llevaran a cabo pruebas de diagnostico para detectar e identificar puntos con altos residuales, atípicos, influyentes o de alto leverage, y se tratará de mejorar el modelo llevando a cabo un modelo ajustado que pase por el origen, al cuál se le aplicaran nuevamente las pruebas de diagnostico y se le evaluaran el cumplimiento de los supuestos.Finalmente, se aplicara regresión robusta y se compararan todos los modelos, eligiendo el mejor o el más adecuado.

1.1 Definiciones

Para entender un poco el objetivo del modelo, se van a definir las dos variables en cuestión:
El material particulado respirable presente en la atmósfera de nuestras ciudades en forma sólida o líquida (polvo, cenizas, hollín, partículas metálicas, cemento y polen, entre otras) se puede dividir, según su tamaño, en dos grupos principales. A las de diámetro aerodinámico igual o inferior a los 10 μm o 10 micrómetros (1 μm corresponde a la milésima parte de un milímetro) se las denomina PM_{10} (partículas respirables) y a la fracción respirable más pequeña, $PM_{2.5}$ (partículas finas). Estas últimas están constituidas por aquellas partículas de diámetro aerodinámico inferior o igual a los 2.5 micrómetros, es decir, son 100 veces más delgadas que un cabello humano, además son más peligrosas, ya que, al ser inhaladas, pueden alcanzar las zonas periféricas de los bronquiolos y alterar el intercambio pulmonar de gases.

2. Punto 1: Modelo Ajustado

Para ajustar un modelo lineal para estas partículas, utilizamos el paquete Zeileis & Hothorn (2002) y decidimos que la variable independiente o regresora fuera el PM_{10} y la variable dependiente fuera el $PM_{2.5}$, se hizo esta selección, ya que las partículas con un diámetro menor o igual a 10 micrómetros son más fáciles de medir que las de diámetro menor o igual a 2.5 micrómetros, y la idea es que se ahorren costos y esfuerzo humano adicional, midiendo solo las partículas PM_{10} (que son un poco más grandes) y a partir de ellas logren obtener una buena estimación para las partículas $PM_{2.5}$.
Teniendo en cuenta lo anterior, el modelo planteado es:

$$PM_{2.5} = \beta_0 + \beta_1 PM_{10}$$

El modelo ajustado, usando MCO es:

$$\hat{PM}_{2.5} = -4.29990 + 0.45776 PM_{10}$$

3. Punto 2: Bondad del modelo

Para este modelo ajustado, tenemos los siguientes resultados que nos ayudaran a juzgar que tan bueno es:

	Estimación	Desviación Estandár	Estadistico T	P valor
β_0	-4.29990	1.27362	-3.376	0.000916
β_1	0.45776	0.03051	15.001	<2e-16

En la tabla anterior se puede apreciar que ambas estimaciones de los coeficientes son significativas a muy bajos niveles de α,ya que el p valor para ambos es casi cero, además las desviaciones estándar de ambos son muy bajas, esto nos indica que están muy cerca del verdadero valor del parámetro, ó en otras palabras, tienen una buena precisión.

R^2	$R^2_{Ajustado}$	$\hat{\sigma}$	Estadistico F	Valor p	AIC	BIC
0.5755	0.5729	6.7707	225	<2.2e-16	1123.39	1132.762

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co

En esta tabla, podemos ver que tanto el R^2 como el $R^2_{Ajustado}$ son considerablemente altos, estos nos indican que cerca del 57 % de la variabilidad total de la variable $PM_{2.5}$ es explicada por la variable PM_{10}, además la desviación del modelo ajustado a los datos es de 6.7707, lo cuál se podría considerar un buen valor ya que no es tan alto, y finalmente, vemos que el valor p de la prueba de significancia es casi cero, esto nos dice que $\beta_1 \neq 0$ o en otras palabras, que el modelo es significativo.

4. **Punto 3: Pruebas de diagnostico**
Para detectar los puntos con altos residuales, atípicos, influyentes o de alto leverage, se van a realizar las gráficas comunes para tener idea sobre cuales parejas de datos podrían ser tales puntos (gráfica de dispersión con recta ajustada, gráfica de residuales estandarizados contra los Y ajustados y gráfica q-qplot), y además, se utilizara el criterio de la matriz hat y el residual studentizado, y las medidas de influencia D de Cook, DFFITS, DFBETAS y COVRATIO.

a) **Gráfica de dispersión:**

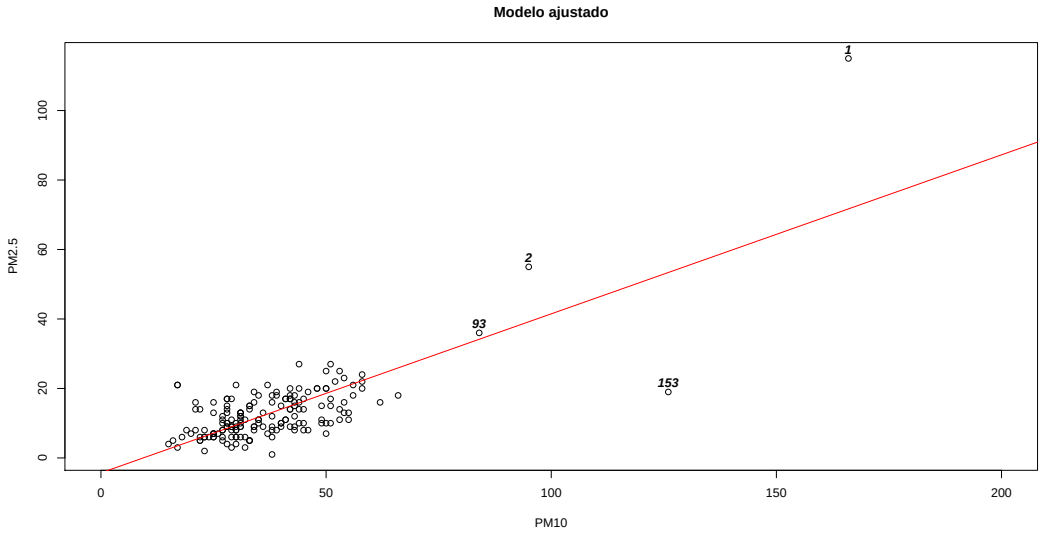

FIGURA 1: Gráfico de dispersión con recta ajustada y valores candidatos a atipicos

b) **Gráfica de $\hat{d}$ vs $\hat{y}$:**

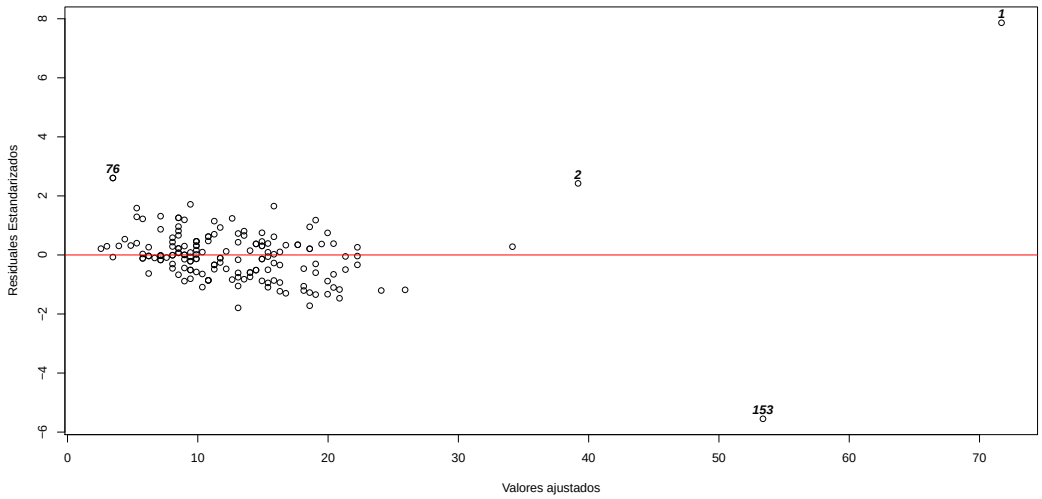

FIGURA 2: Gráfico de residuales estandarizados vs los valores ajustados

En las dos gráficas anteriores y en la siguiente podemos ver como las observaciones 1,2,76,93 y 153 sobresalen del resto, por lo cuál son las primeras sospechosas a ser puntos atípicos y van a ser los puntos en los cuales nos vamos a centrar en las medidas de influencia.

c) **q-qplot de los residuales :**

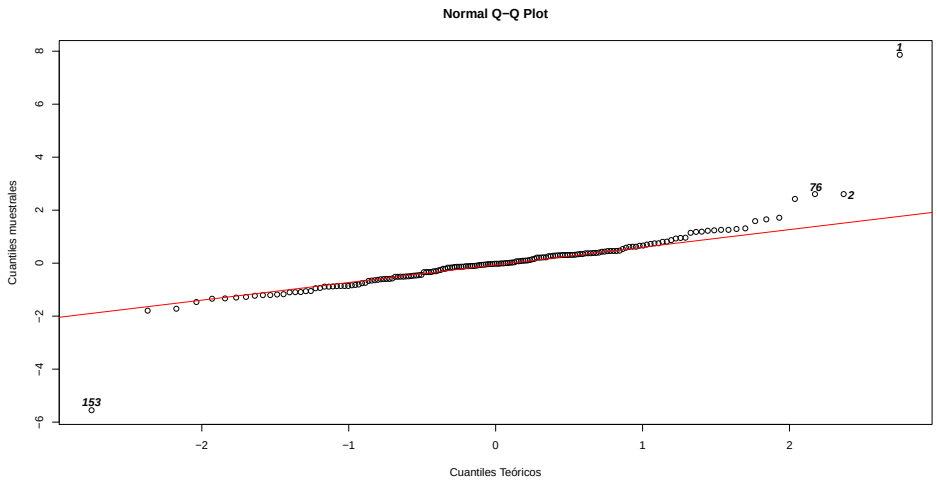

FIGURA 3: q-qplot de los residuales

d) **Matriz HAT y residual Studentizado:** La matriz HAT definida como $H = x(x'x)^{-1}x^T$ y más específicamente el i-ésimo valor de su diagonal $h_{ii} = x_i'(X'X)^{-1}x_i$ y el residual Studentizado, definido como $\hat{r_i} = \frac{\hat{e_i}}{\sqrt{1-h_{ii}}\hat{\sigma}}$ sirven para ver si los puntos pueden ser influyentes o solo de balanceo, considerando los siguientes criterios:

- Si $h_{ii} > \frac{2p}{n} = 2\bar{h} = 0.02381$ (como $\sum h_{ii} = p \rightarrow \bar{h} = \frac{p}{n}$) y el residual Studentizado es grande (> 3), entonces la i-ésima observación puede considerarse influyente.
- Si solamente $h_{ii} > \frac{2p}{n} = 0.02381$, entonces la i-ésima observación es solo de balanceo.

e) **Distancia de Cook (D de Cook):** La distancia de Cook mide cómo cambia el vector de estimadores $\hat{\beta}$ cuando se elimina cada observación, por lo cuál es una medida útil para considerar como influyente una observación. La D de Cook se define como $D_i(M;C) = \frac{(\hat{\beta_{(i)}}-\hat{\beta})^T M(\hat{\beta_{(i)}}-\hat{\beta})}{C}$; i=1,...,n, donde $M = X^TX$ y $C = p\hat{\sigma^2}$. El criterio es:

- Si $D_i \geq 1$, entonces la i-ésima observación puede considerarse como influyente.

f) **DFFITS:** $DFFITS_i$ es la cantidad de desviaciones estándar que cambia el valor ajustado $\hat{y_i}$ si se elimina la observación i. Este criterio no define exactamente si el punto es influyente o de balanceo, solo sugiere que se debe examinar la observación. Se define como $DFFITS_i = \frac{\hat{y_i}-\hat{y_{(i)}}}{\sqrt{S^2_{(i)}h_{ii}}}$ El criterio es:

- Si $|DFFITS_i| > 2\sqrt{\frac{p}{n}} = 0.2182$, sugiere que se debe investigar la influencia de la i-ésima observación.

g) **DFBETAS:** Los $DFBETAS_{j,i}$ indica cuánto cambia el coeficiente de regresión $\hat{\beta_j}$, en unidades de desviación estándar, si se omitiera la i-ésima observación. Esta se define como $DFBETAS_{ji} = \frac{\hat{\beta_j}-\hat{\beta_{j(i)}}}{\sqrt{S^2_{(i)}C_{jj}}}$; $C = (X^TX)^{-1}$. El criterio es:

- Si $|DFBETAS_{ji}| > 2/\sqrt{n} = 0.1543$, entonces debe examinarse la i-ésima observación.

h) **COVRATIO:** Esta medida sirve para expresar el papel de la i-ésima observación en la precisión de la estimación y se define como $\frac{|S^2_{(i)}(X_i^TX_{(i)})^{-1}|}{|S^2(X^TX)^{-1}|}$; i=1,2,...,n. El criterio es:

- Si $COVRATIO_i > 1$, entonces la i-ésima observación mejora la precisión de la estimación.
- Si $COVRATIO_i < 1$, entonces la inclusión de la i-ésima observación disminuye la precisión de la estimación.
- Si $COVRATIO_i > 1 + 3\frac{p}{n} = 1.0357$ o si $COVRATIO_i < 1 - 3\frac{p}{n} = 0.9643$, entonces la i-ésima observación debería ser considerada influyente.

4.1 Resultados pruebas diagnostico

Como se tienen 168 resultados de cada una de las medidas de influencia no se pueden mostrar todas, en la siguiente tabla se mostraran las medidas de los puntos más importantes, las que cumplan con el criterio de influencia se les asignara un asterisco (*) al lado del valor, para así detectar con que criterios es influyente o de balanceo cada punto. **Tanto en este modelo como en el que pasa por origen (ver tabla del punto 5) las observaciones 1 y 153 son claramente influyentes por todos los criterios y el resto resultan influyentes por el COVRATIO, excepto la observación 2, que es solo de balanceo en el primer modelo**.

Observación	HAT	$\hat{r_i}$	D_i	$DFFITS_i$	$DFBETAS_{(0),i}$	$DFBETAS_{1,j}$	$COVRATIO_i$
1	0.3384*	9.8987499*	15.8*	7.07945*	-6.014452*	7.02*	0.602*
2	0.07179*	2.4606893	0.227	0.684351*	-0.516887*	0.655*	1.015
76	0.01497	2.6539008	0.0516	0.327122*	0.316146*	-0.254*	0.945*
79	0.01497	2.6539008	0.0516	0.327122*	0.316146*	-0.254*	0.945*
93	0.04881*	0.2790688	0.00201	0.063217	-0.044971	0.0592	1.063*
153	0.16301	-6.1311115*	3*	-2.705776*	2.210180*	-2.66*	0.802*

5. **Punto 4: Modelo por el origen**
   El modelo ajustado asumiendo que el intercepto es igual a cero es: $\hat{PM}_{2.5} = 0.3638 PM_{10}$
   Para este modelo ajustado, tenemos los siguientes resultados que nos ayudaran a juzgar que tan bueno es:

	Estimación	Desviación Estandár	Estadistico T	P valor
β_1	0.3638	0.0129	28.2	<2e-16

R^2	$R^2_{Ajustado}$	$\hat{\sigma}$	Estadistico F	Valor p	AIC	BIC
0.8265	0.8254	6.978	795.4	<2.2e-16	1132.546	1138.794

6. **Punto 5: Pruebas de diagnostico**
   Los resultados de las medidas de influencia y las pruebas de diagnostico para este modelo se resumen en la siguiente tabla:

Observación	HAT	$\hat{r_i}$	D_i	$DFFITS_i$	$DFBETAS_{1,j}$	$COVRATIO_i$
1	0.094157*	10.625441*	7.03*	3.43*	3.43*	0.661*
2	0.03084*	3.048048*	0.282	0.544*	0.544*	0.983
76	0.000987	2.146917	0.00446	0.0675	0.0675	0.980*
79	0.000987	2.146917	0.00446	0.0675	0.0675	0.980*
93	0.024110*	0.788304	0.0154	0.124	0.124	1.027*
153	0.054247*	-4.141674*	0.897	-0.992*	-0.992*	0.964*

6. **Punto 6: Elección del modelo**
   Seleccionamos como mejor modelo, al primero, en el cuál se ajusto con su intercepto, ya que los criterios de selección AIC y BIC dieron menores para este, además, la desviación del modelo a los datos es menor. Nos podría engañar en la selección el R^2 y el $R^2_{Ajustado}$, ya que estos aumentan casi en un 25 % lo cuál nos podría tentar a elegirlo como el mejor,este aumento ocurre ya que, al no haber intercepto, casi toda la variabilidad de Y ($PM_{2.5}$) es atribuible a X(PM_{10}) eso explica el R^2 de 82 %. Además, en el modelo sin intercepto $E(\hat{e})$ no es precisamente cero, en este caso esta cerca de -1, lo cuál es inadecuado.

7. **Validación de supuestos del modelo con intercepto**

a) Correcta especificación (Linealidad): Para evaluar la correcta especificación del modelo, simplemente nos basamos en dos gráficas, la primera, es la gráfica de los residuos contra los valores ajustados, y la segunda, es la gráfica de dispersión común y corriente con la recta ajustada.En ambas gráficas se ve que el modelo esta correctamente especificado.

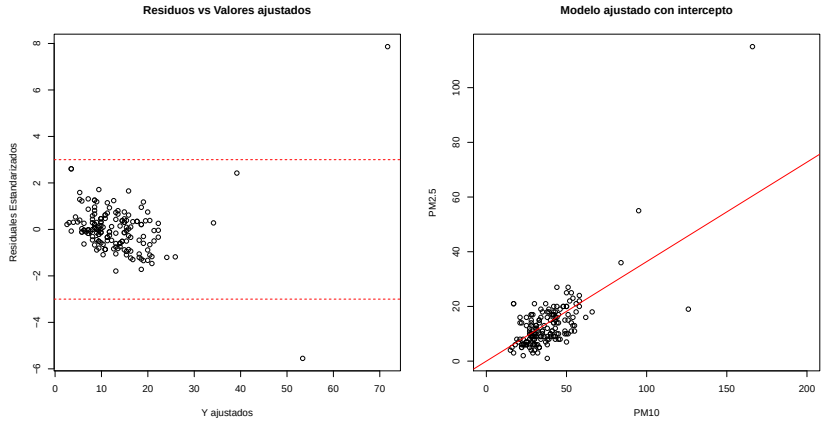

Figura 4: Gráficos para detectar incorrecta especificación

b) Normalidad de los errores: Para verificar este supuesto, se realizaron nuevamente dos gráficas, la primera, un histograma de los residuales para ver una posible forma de distribución de estos, y la segunda, el Q-Qplot, que muestra que tanto se asemeja la distribución de los residuales a la distribución normal. En ambas gráficas se evidencia que la distribución de los residuales no es normal, se puede ver en ambas que la cola derecha de la distribución de los residuales es mucho más pesada que en la distribución normal. Además aplicamos la prueba Shapiro-Wilk para confirmar nuestra creencia, esta prueba nos arrojo un estadístico de $W = 0.85121$ y un p valor= 8.796e-12, por tanto se rechaza H_0 y se concluye que la distribución de los residuales no es normal.

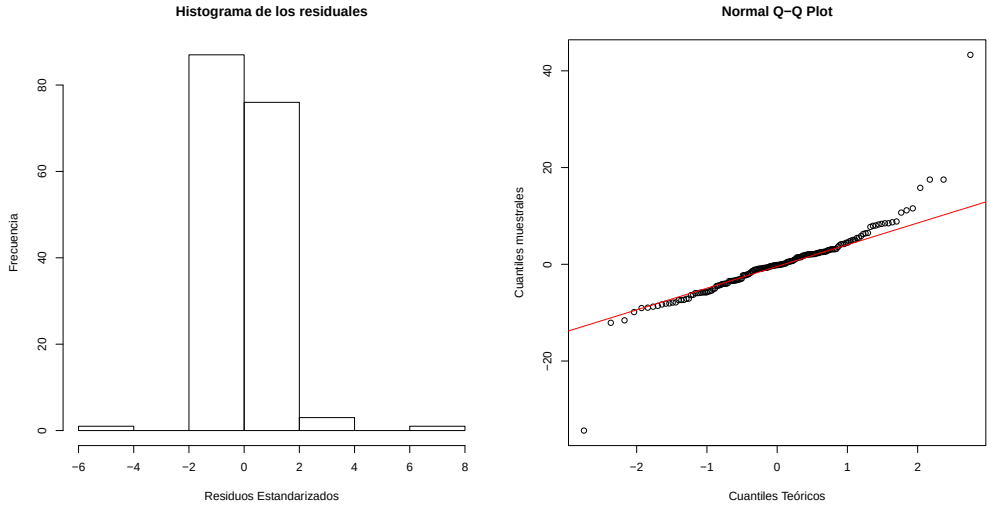

FIGURA 5: Gráficos para evaluar la normalidad en los errores

c) Homocedasticidad: Para probar la homocedasticidad, se analiza la misma gráfica realizada en el literal a (los residuales vs los $\hat{y}$), donde no se ve claramente el incumplimiento de este supuesto, por lo cual se realizaron las pruebas Goldfeld-Quandt y Breusch-Pagan, se eligieron estas, ya que todas las otras pruebas más conocidas y utilizadas (Levene, Barlett, Cochran, etc) algunas requieren repeticiones en X y otros supuestos. La prueba Goldfeld-Quandt nos arrojo un estadístico de prueba $GQ = 0.6068$ y un p valor de 0.02487, y la prueba Breusch-Pagan nos arrojó un estadístico de prueba $BP = 83.079$ y un p valor asociado de aproximadamente 0 (<2.2e-16). Por lo cuál en ambas pruebas se rechaza la hipótesis nula,y concluimos que los residuos del modelo no cumplen con el supuesto de homocedasticidad de varianza.

d) Independencia en los errores: Sabemos que no se puede probar a ciencia cierta la independencia en los errores, sin embargo, para tener cierta idea sobre este supuesto, analizamos los residuales gráficamente (correlograma) y le aplicamos la prueba de Durbin Watson para darnos cuenta si estos tienen cierta correlación o dependencia. La prueba de Durbin-Watson nos arrojo un estadístico $DW = 1.1683$ y un p valor de 2.183e-08, entonces, para un nivel de significancia de 0.05, se rechaza H_0 y se concluye que la correlación entre los errores es diferente de cero.

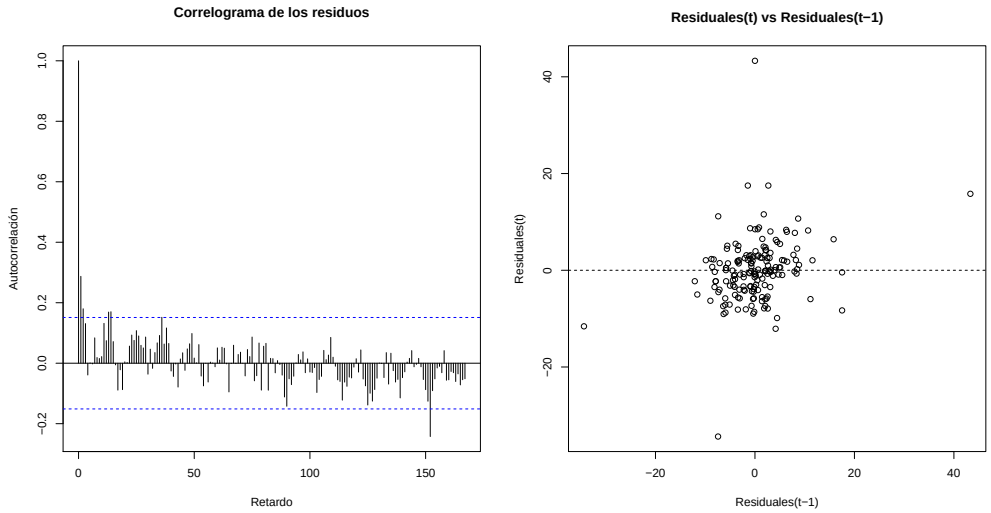

FIGURA 6: Gráficos para evaluar la independencia en los errores

8. Regresión Robusta

Las estimaciones con los mínimos cuadrados son altamente sensible a los valores atípicos. Si bien no existe una definición exacta de un valor atípico o de una observación atípica, los valores atípicos son observaciones que no siguen el patrón de las otras observaciones. Esto no es normalmente un problema si el valor atípico es simplemente una observación extrema extraída de la cola de una distribución normal, pero si los resultados atípicos de error de medición no normal o alguna otra violación de supuestos ordinarios de mínimos cuadrados estándar, entonces se compromete la validez de los resultados de la regresión si se utiliza una técnica de regresión no robusta. Un proceso de regresión robusta es aquel que amortigua el efecto de las observaciones que serían muy influyentes si se usaran los mínimos cuadrados.

Hay varios métodos para realizar regresión robusta, entre los más utilizados están el método M (máxima verosimilitud) que le asigna una función de peso a los residuales, de forma que entre mayor sea el valor del residual menos peso le pone, por ende este método le da menos importancia a valores con residuales grandes; el método LTS (Cuadrados menos recortados)que intenta minimizar la suma de residuos cuadrados; el método S que encuentra una línea (plano o hiperplano) que minimiza una estimación robusta de la escala (de la cual el método obtiene la S en su nombre) de los residuos y finalmente el método MM que intenta retener la solidez y la resistencia de la estimación S, mientras que gana la eficiencia de la estimación M. El método procede al encontrar una estimación S altamente robusta y resistente que minimiza una M-estimación de la escala de los residuos.

Seleccionamos el método MM para la estimación de los coeficientes por ser el más robusto y el de mejor comportamiento,utilizando el paquete Venables & Ripley (2002), el modelo ajustado usando este método fue: $\hat{PM_{2.5}} = 0.7196 + 0.3092 PM_{10}$

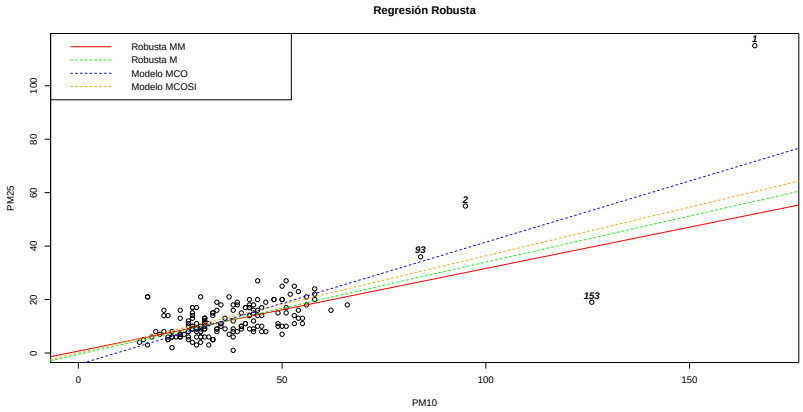

Figura 7: Diagrama de dispersión con todos los modelos ajustados

Para este modelo ajustado, tenemos los siguientes resultados que nos ayudaran a juzgar que tan bueno es:

	Estimación	Desviación Estandár	Estadistico T	P valor
β_0	0.7196	0.9259	0.7771	0.4382
β_1	0.3092	0.0222	13.9372	0

$\hat{\sigma}$	AIC	BIC
4.824	1147.121	1156.493

Podemos concluir que el mejor modelo es el último (Regresión Robusta), ya que, el criterio de comparación entre un modelo MCO y un modelo Robusto es la varianza del modelo, el cambio de los coeficientes de regresión y las desviaciones de esos coeficientes. Podemos ver que la desviación del modelo disminuyó de 6.7707 a 4.824,los coeficientes estimados cambiaron considerablemente y las desviaciones de las estimaciones también disminuyeron considerablemente. En este caso, no se toma muy en cuenta el AIC y el BIC, ya que estos criterios no se pueden usar para comparar modelos obtenidos por métodos diferentes (MCO y Regresión Robusta). Además, es muy razonable elegir el modelo ajustado por regresión robusta, ya que los supuestos del modelo 1 no se cumplen, por lo cuál las estimaciones por MCO y en general la teoría de los Mínimos Cuadrados Ordinarios es inadecuada y no se debe usar al no cumplirse sus supuestos.

Referencias

Douglas C. Montgomery, Elizabeth A. Peck, G. G. V. (2002), *Introducción al análisis de regresión lineal*, tercera edn, Compañía Editorial Continental.

Faraway, J. J. (2005), *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Vol. 1 edn., Chapman y Hall / CRC.

Norman R. Draper, H. S. (1998), *Applied Regression Analysis, Third Edition (Wiley Series in Probability and Statistics)*, third edn. *http://gen.lib.rus.ec/book/index.php?md5=903983CE6C8FBCBB1F81265002FE576B

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Relación entre $PM_{2.5}$ y PM_{10} en la ciudad de Bogotá (2005), *Revista de Ingeniería* .

Tusell, F. (2011), *Análisis de Regresión.Introducción Teórica y Práctica basada en R.*

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York. ISBN 0-387-95457-0.
*http://www.stats.ox.ac.uk/pub/MASS4

Zeileis, A. & Hothorn, T. (2002), 'Diagnostic checking in regression relationships', *R News* **2**(3), 7–10.
*https://CRAN.R-project.org/doc/Rnews/