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 solucionar esto llevando a cabo un modelo ajustado que pase por el origen y aplicando regresión robusta. Finalmente, para estos dos nuevos modelos, se evaluaran nuevamente las pruebas de diagnostico y se concluirá sobre el mejor modelo, teniendo en cuenta los resultados.

1.0.1. 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, 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:

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

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

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
0.5755	0.5729	6.7707	225	<2.2e-16

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.

- Gráfica de dispersión:
- Gráfica de $\hat{d}$ vs $\hat{y}$:
- q-qplot de los residuales :
- 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 $h_{ii} > \frac{2p}{n} = 2\bar{h}$ (como $\sum h_{ii} = p \rightarrow \bar{h} = \frac{p}{n}$) y el residual Studentizado es grande, entonces la i-ésima observación puede considerarse influyente.
  - Si solamente $h_{ii} > \frac{2p}{n}$, entonces la i-ésima observación es solo de balanceo.
- 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 $M = X^T X$ y $C = p\hat{\sigma^2}$. El criterio es:

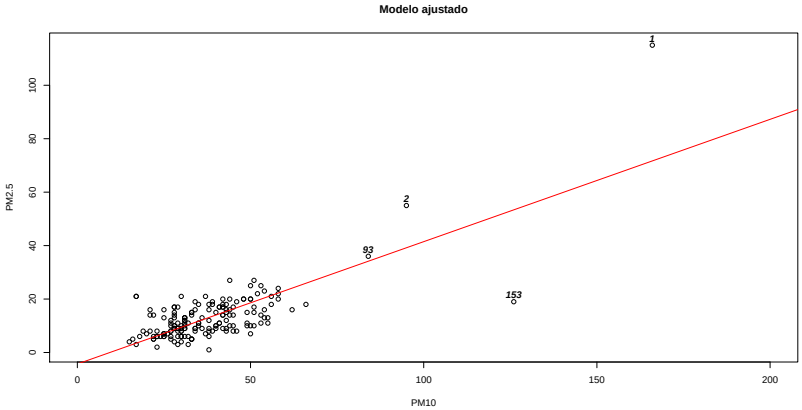

FIGURA 1: Gráfico de dispersión con recta ajustada y valores candidatos a atipicos

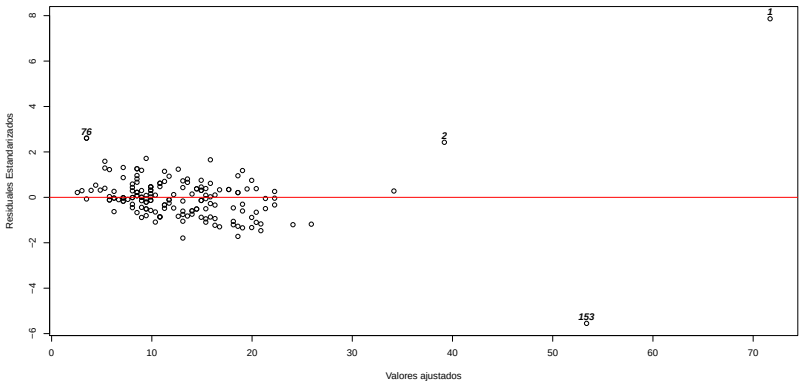

FIGURA 2: Gráfico de residuales estandarizados vs los valores ajustados

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

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

- 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 $DFBETAS_{ji} = \frac{\hat{\beta}_j - \hat{\beta_{j(i)}}}{\sqrt{S^2_{(i)} C_{jj}}}$; $C = (X^T X)^{-1}$. El criterio es:

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

- COVRATIO: Esta medida sirve para expresar el papel de la i-ésima observación en la precisión de la estimación y se define como $\frac{|S^2_{(i)}(X_i^T X_{(i)})^{-1}|}{|S^2(X^T X)^{-1}|}$; 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 $COVRATIO_i > 1 + 3\frac{p}{n}$ o si $COVRATIO_i < 1 - 3\frac{p}{n}$, entonces la i-ésima observación debería ser considerada influyente.

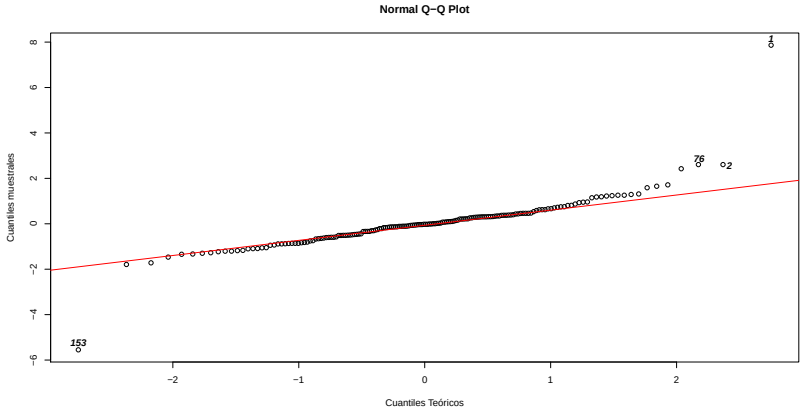

FIGURA 3: q-qplot de los residuales

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:

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:

$$\hat{PM_{2.5}} = 0.3638 PM_{10}$$

? ? ?