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

Kevin García 1533173
Alejandro Vargas 1525953

21 de mayo de 2018

Introducción

En esta presentación 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.

Definiciones

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.

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 $PM_{2,5}$.

Modelo Ajustado

Teniendo en cuenta lo anterior, el modelo planteado es:

$$PM_{2,5} = \beta_0 + \beta_1 PM_{10}$$

El modelo ajustado, usando MCO es:

$$\hat{PM}_{2,5} = -4{,}29990 + 0{,}45776 PM_{10}$$

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.

Bondad del modelo

R^2	$R^2_{Ajustado}$	$\hat{\sigma}$	Estadistico F	Valor p
0.5755	0.5729	6.7707	225	$< 2{,}2e-16$

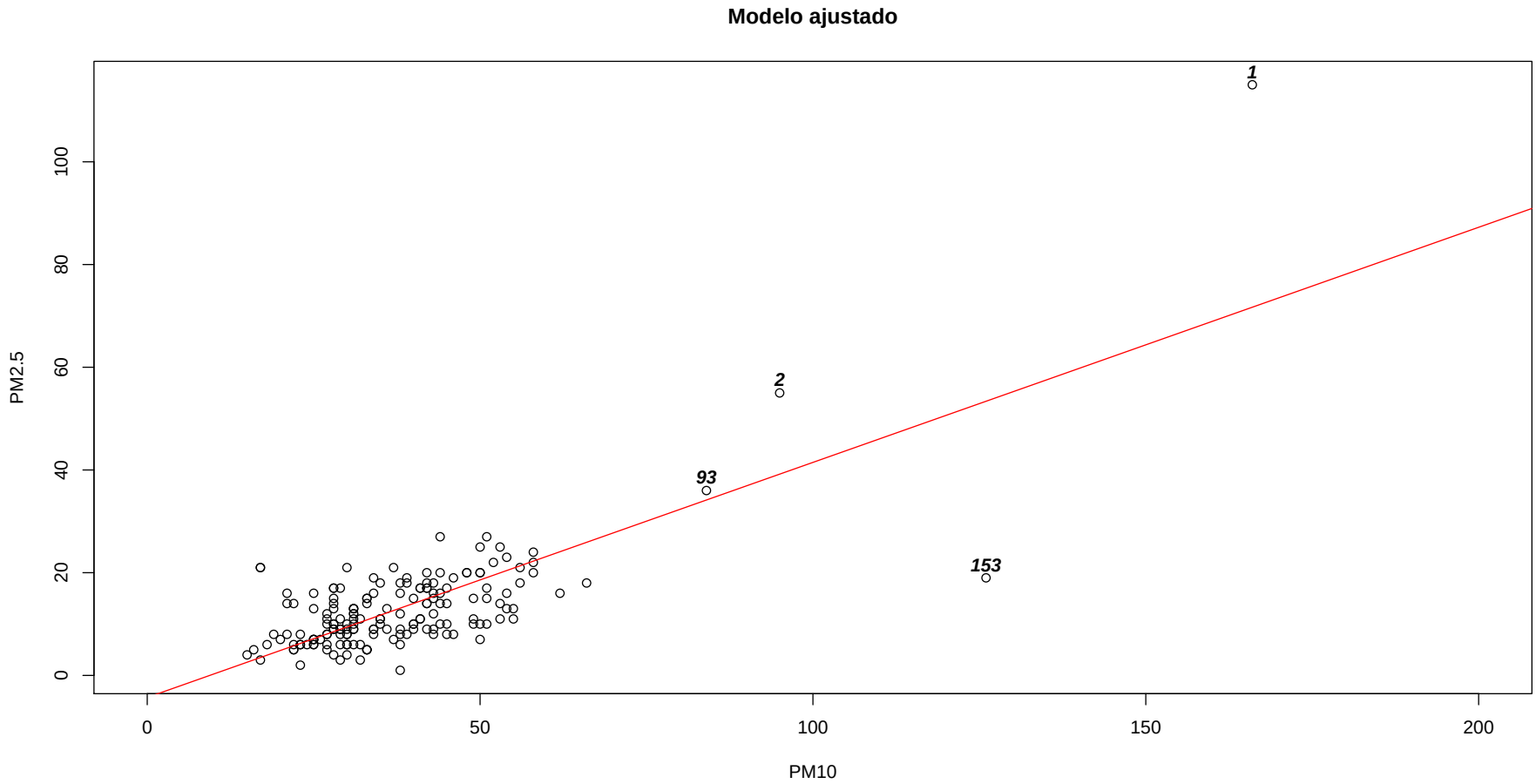
En esta tabla, podemos ver que tanto el R^2 como el $R^2_{Ajustado}$ son considerablemente altos, estos nos indican que cerca del 57 % de la variabilidad total de la variable $PM_{2,5}$ es explicada por la variable PM_{10}, además la desviación del modelo ajustado a los datos es de 6.7707, lo cuál se 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.

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.

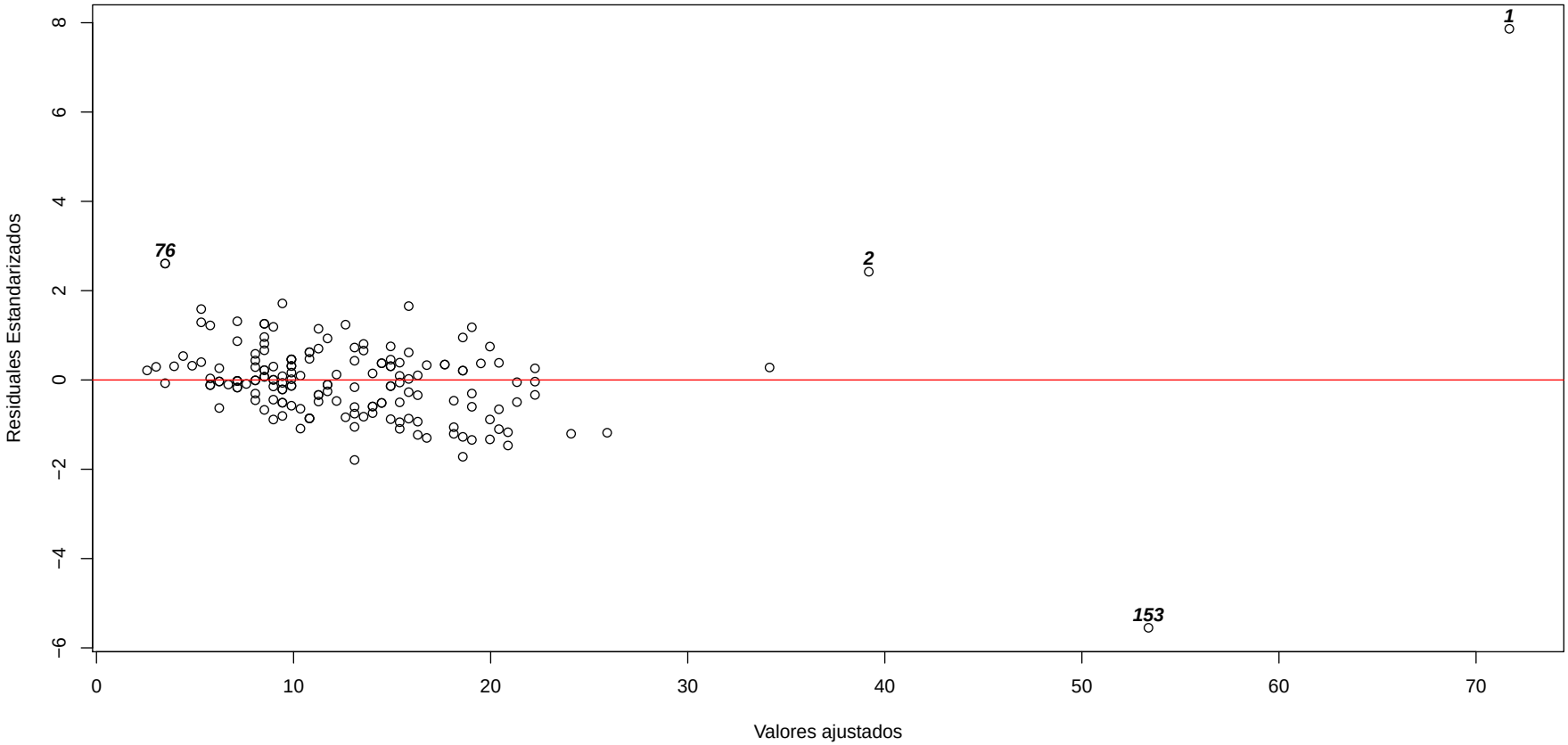
Pruebas de diagnostico

a) **Gráfica de dispersión:**



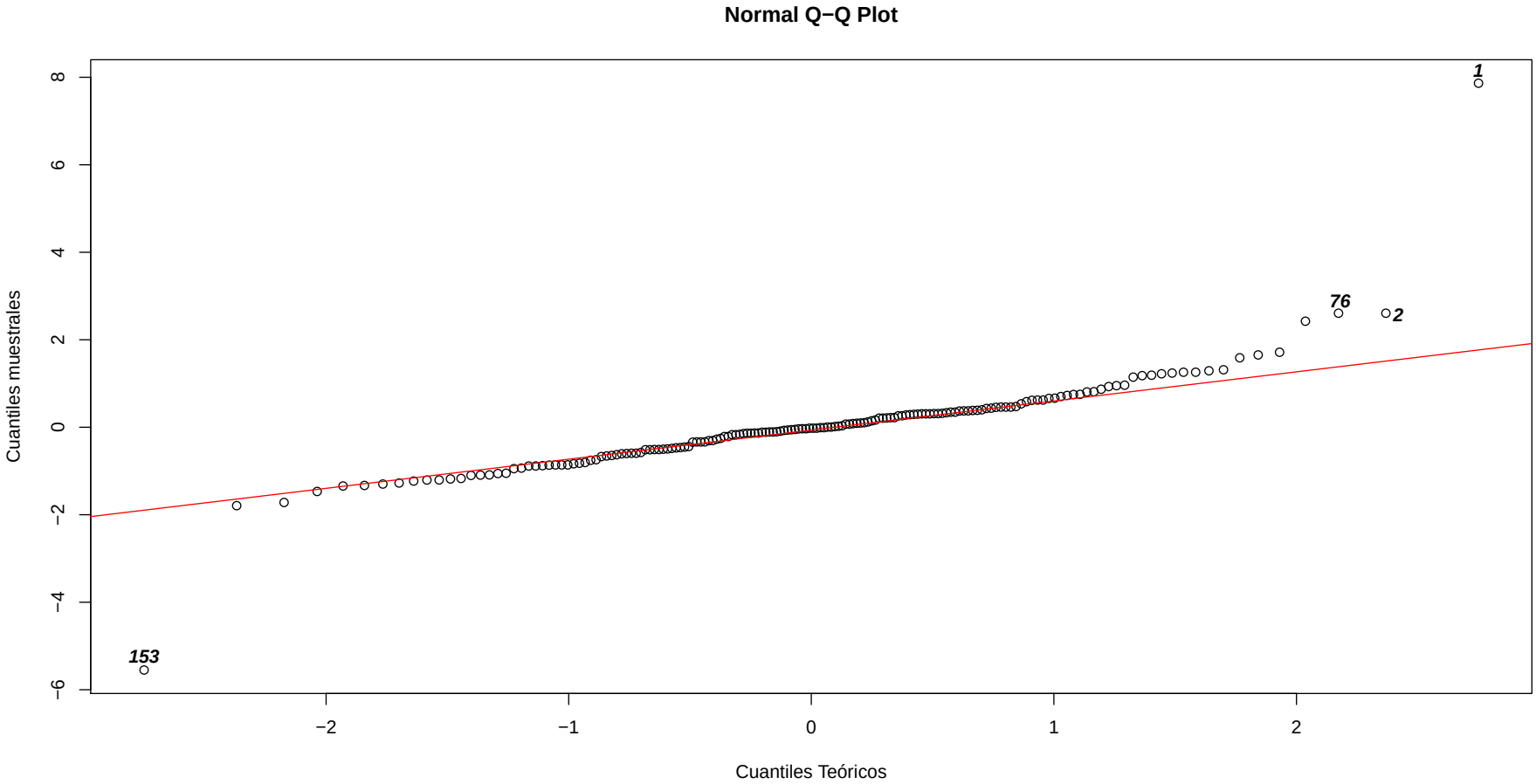
Pruebas de diagnostico

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



Pruebas de diagnostico

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



Pruebas de diagnostico

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.

Pruebas de diagnostico

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 $D_i(M; C) = \frac{(\hat{\vec{\beta}}_{(i)} - \hat{\vec{\beta}})^T M (\hat{\vec{\beta}}_{(i)} - \hat{\vec{\beta}})}{C}$; i=1,...,n, donde $M = X^T X$ y $C = p\hat{\sigma^2}$. El criterio es:

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

Pruebas de diagnostico

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.

Pruebas de diagnostico

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

- Si $|DFBETAS_{ji}| > 2/\sqrt{n} = 0{,}1543$, entonces debe examinarse la i-ésima observación.

Pruebas de diagnostico

h) **COVRATIO:** Esta medida sirve para expresar el papel de la i-ésima observación en la precisión de la estimación y se define como $\frac{|S^2_{(i)}(X_i^T X_{(i)})^{-1}|}{|S^2(X^T X)^{-1}|}$; 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 $COVRATIO_i > 1 + 3\frac{p}{n} = 1{,}0357$ o si $COVRATIO_i < 1 - 3\frac{p}{n} = 0{,}9643$, entonces la i-ésima observación debería ser considerada influyente.

Resultados medidas de influencia

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*

Se puede observar que todas las observaciones de la tabla, se podrían considerar influyentes por uno u otro criterio.

Modelo con intercepto 0

El modelo planteado para que pase por el origen es:

$$PM_{2,5} = \beta_1 PM_{10}$$

El modelo ajustado asumiendo que el intercepto es igual a cero es:

$$\hat{PM}_{2,5} = 0{,}3638 PM_{10}$$

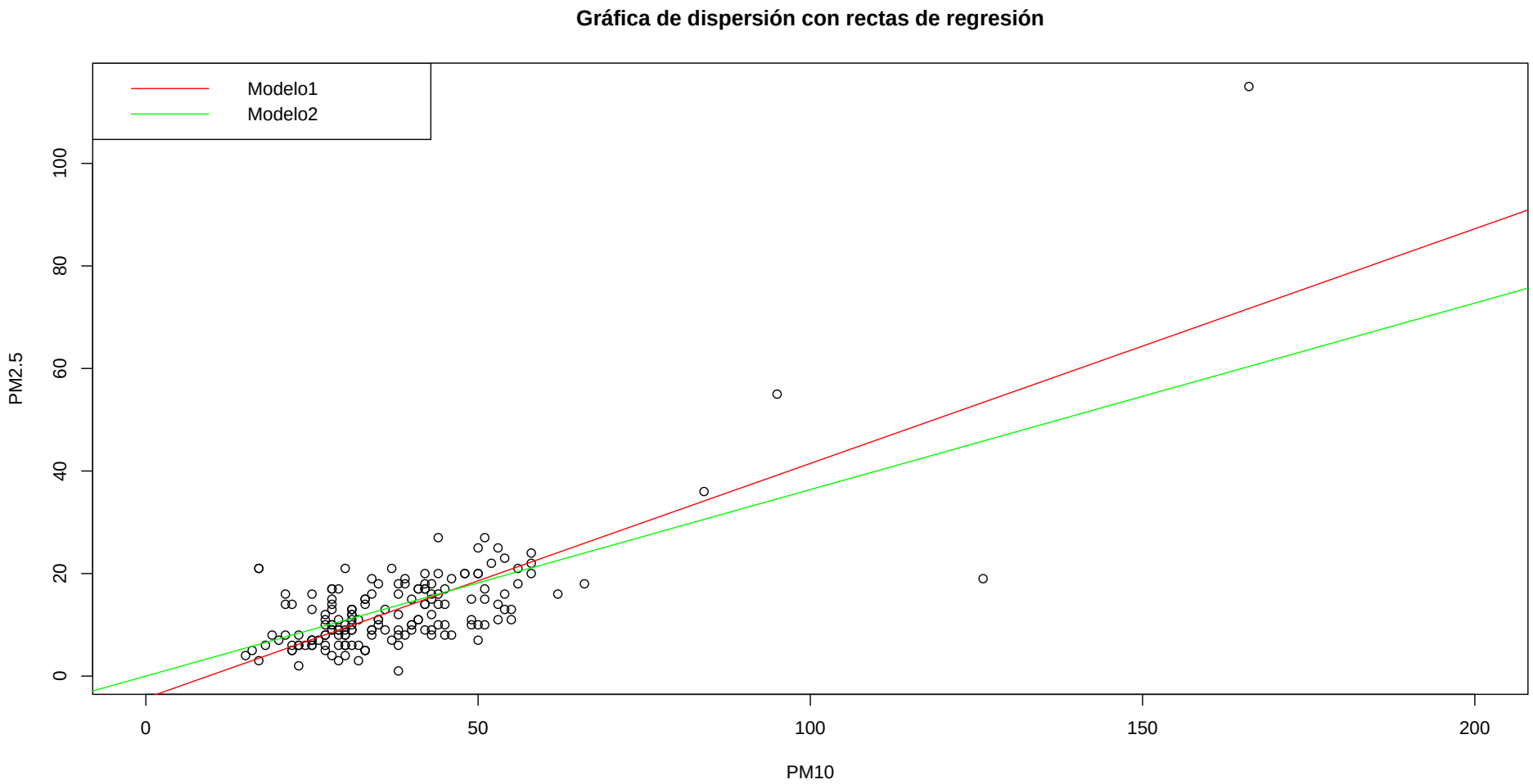
Bondad del modelo

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_1	0.3638	0.0129	28.2	$< 2e-16$

R^2	$R^2_{Ajustado}$	$\hat{\sigma}$	Estadistico F	Valor p
0.8265	0.8254	6.978	795.4	$< 2{,}2e-16$

Comparación de modelos



Comparación de modelos

Para comparar los modelos tenemos la siguiente tabla comparativa:

Modelos	$R^2_{Ajustado}$	$\hat{\sigma}$	Estadistico F	Valor P	AIC	BIC	$SD(\beta_1)$
Modelo 1	0.5729	6.7707	225	$< 2{,}2e-16$	1123.39	1132.762	0.03051
Modelo 2	0.8254	6.978	795.4	$< 2{,}2e-16$	1132.546	1138.794	0.0129

Comparación modelos

Seleccionamos como mejor modelo, al segundo, en el cuál se asumió que el intercepto es igual a 0. Se hizo esa elección ya que viendo el R^2 y el $R^2_{Ajustado}$, hubo un aumentó bastante grande en ambos (cerca del 0.25) lo cuál es muy bueno, ya que nos dice que la variabilidad del $PM_{2,5}$ explicada por el PM_{10} aumentó en un 25 % asumiendo que el intercepto es 0. Además, si nos fijamos en la desviación estándar de la estimación, vemos que se hace más pequeña, lo cuál quiere decir que la precisión de la estimación aumentó y, por último, la prueba de significancia también se rechazo, es decir, $\beta \neq 0$, por lo tanto, el modelo es significativo. Hubo un aumento en la desviación estándar de 0.2, lo cuál no nos pareció una razón de peso para no elegirlo.

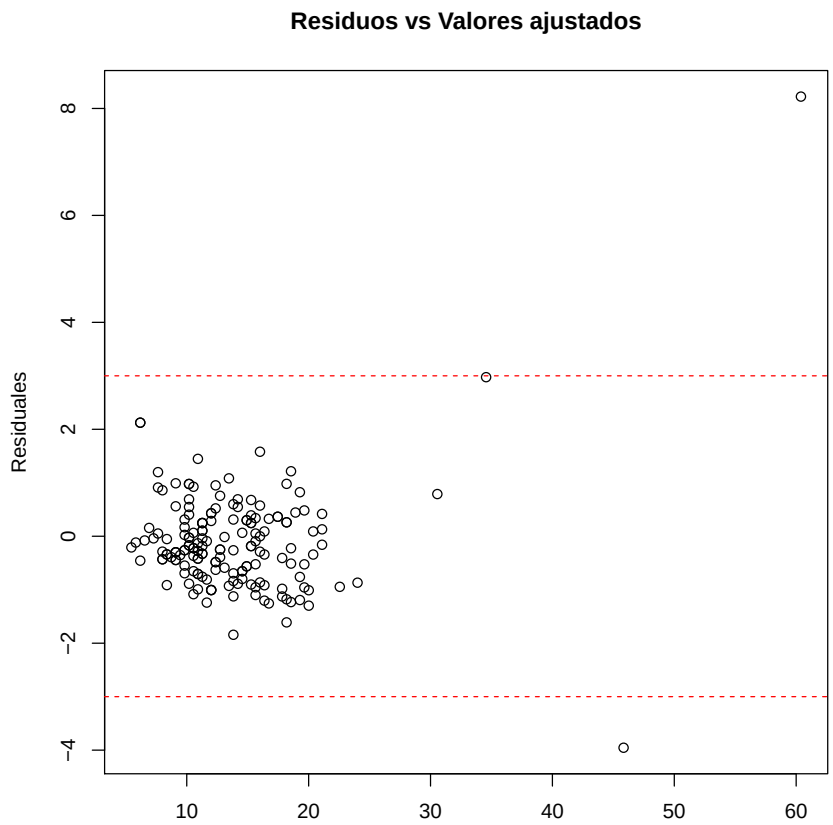
Pruebas de diagnostico modelo 2

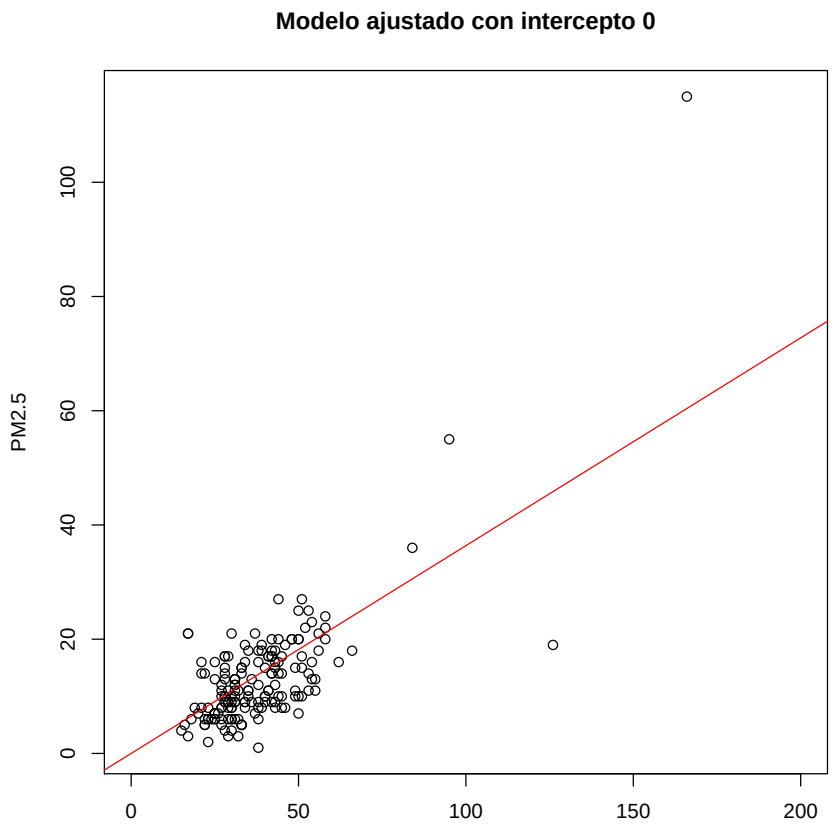
Los resultados de las medidas de influencia y las pruebas de diagnostico para este modelo se resumen en la siguiente tabla:

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*

Validación de supuestos, modelo 2

a) Correcta especificación (Linealidad): Para evaluar la correcta especificación del modelo, simplemente nos basamos en las siguientes dos gráficas.

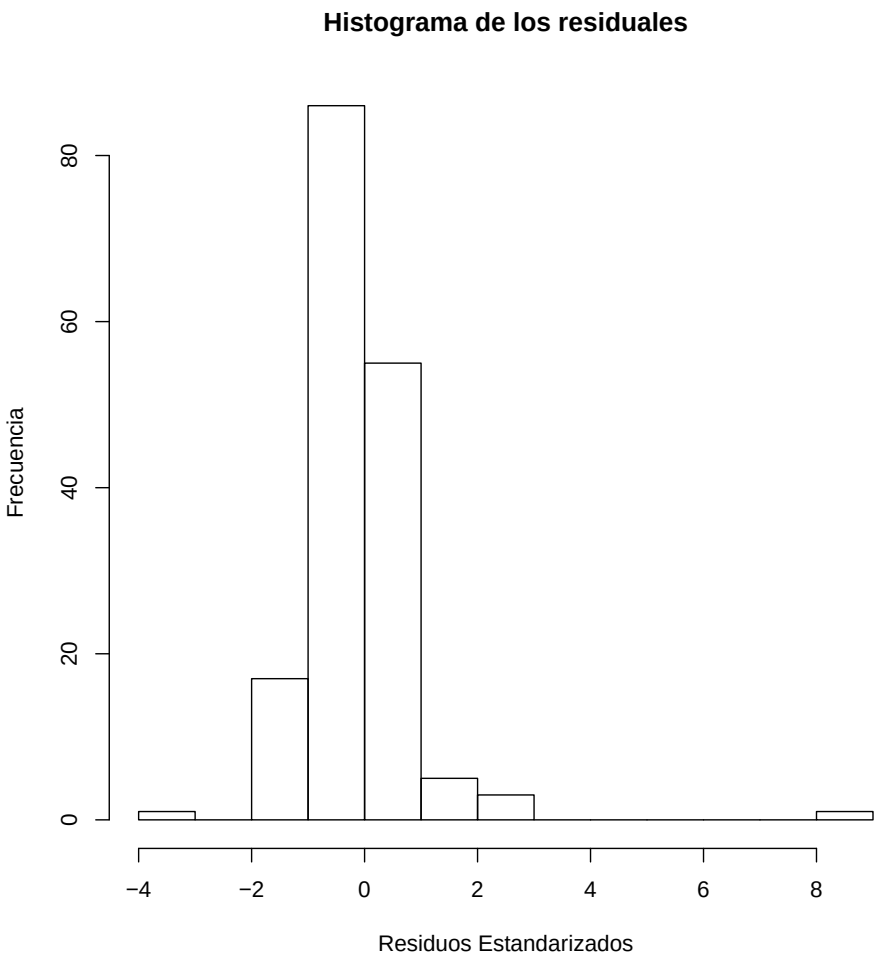


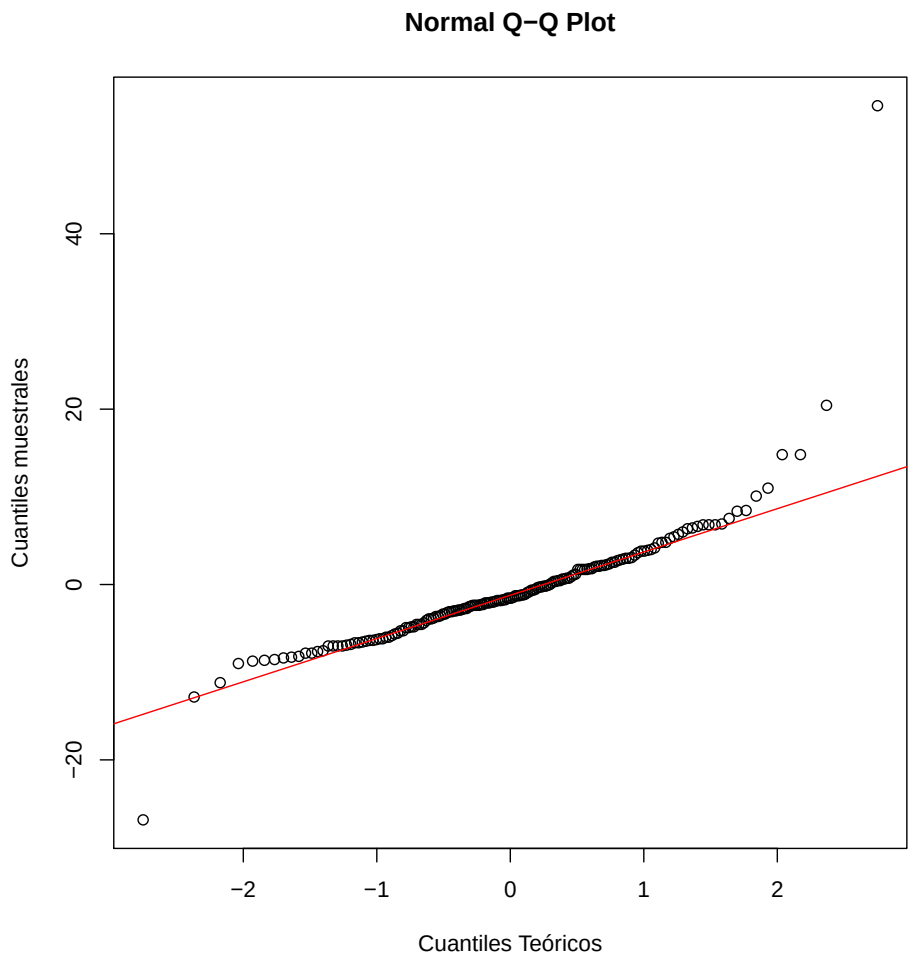


Validación de supuestos, modelo 2

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 estadistico de $W = 0{,}77708$ y un p valor= 1.032e-14, por tanto se rechaza H_0 y se concluye que la distribución de los residuales no es normal.

Validación de supuestos, modelo 2





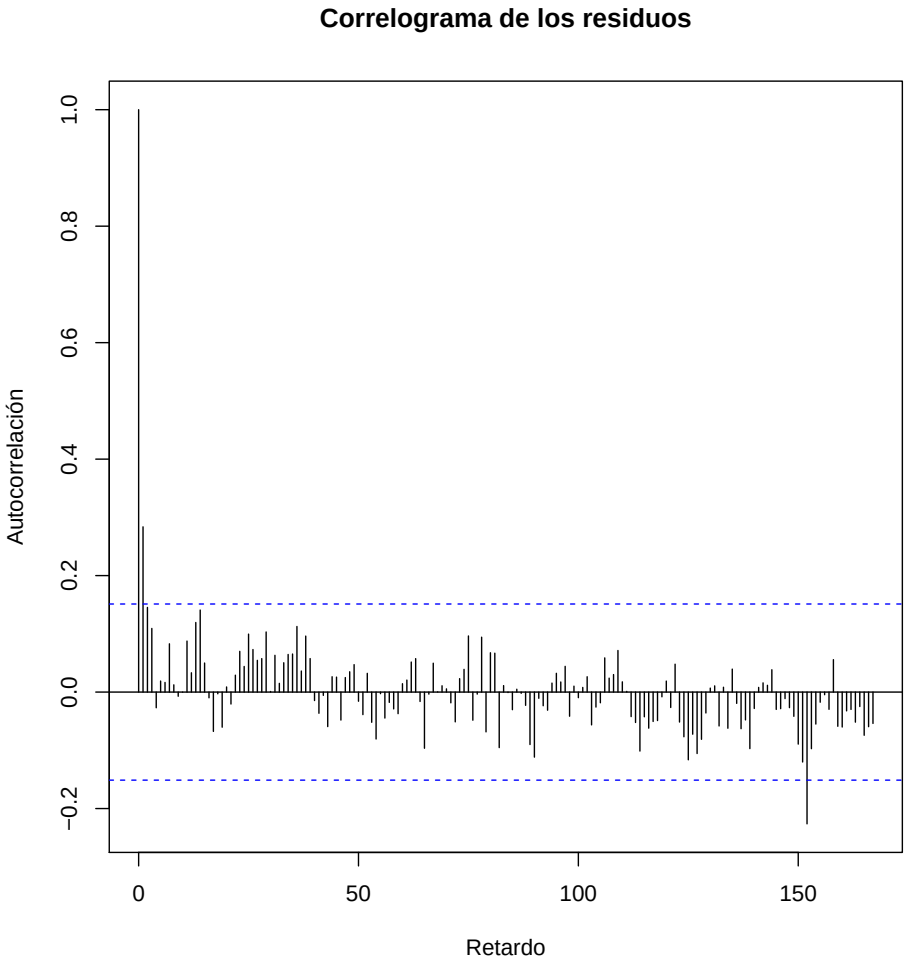
Validación de supuestos, modelo 2

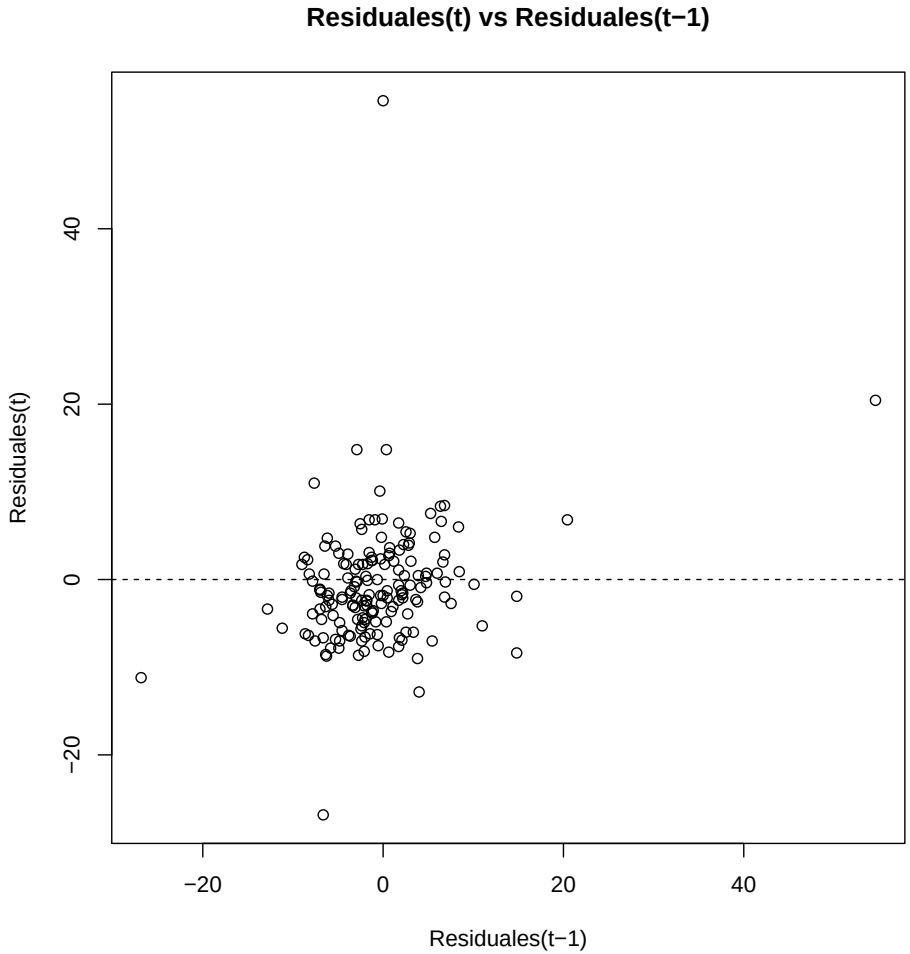
c) Homocedasticidad: Para probar la homocedasticidad, se analiza la misma gráfica realizada en el literal a (los residuales vs los $\hat{y}$ donde se ve que las varianzas son muy parecidas, por lo cual se sospecha que se cumple este supuesto, sin embargo se realizó también la prueba Goldfeld-Quandt, se eligió esta, ya que todas las otras pruebas conocidas (Levene, Barlett, Cochran, Breusch-Pagan, etc) algunas requieren repeticiones en X y otros supuestos, y otras requieren o necesitan el intercepto (es el caso de la prueba de Breusch-Pagan) y este modelo no tiene intercepto. La prueba Goldfeld-Quandt nos arrojo un estadístico de prueba $GQ = 0{,}44682$ y un p valor de 0.9998. Por lo cuál no se rechaza la hipotesis nula,y concluimos que los residuos del modelo no dan evidencia para decir que tienen heterogeneidad de varianza.

Validación de supuestos, modelo 2

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 rachas para darnos cuenta si estos son totalmente aleatorios o si tienen cierta correlación o dependencia. La prueba de rachas nos arrojo un estadístico aproximado a la normal estándar de -2.3369 y un p valor de 0.01945, entonces, para un nivel de significancia de 0.05, se rechaza H_0 y se concluye que los residuos no son totalmente aleatorios, es decir, hay cierta correlación entre ellos.

Validación de supuestos, modelo 2





Regresión 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.
Para aplicar regresión robusta existen varios métodos, entre lo más utilizados están:

- Método M (máxima verosimilitud): 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.
- Método LTS (Cuadrados menos recortados): Intenta minimizar la suma de residuos cuadrados.

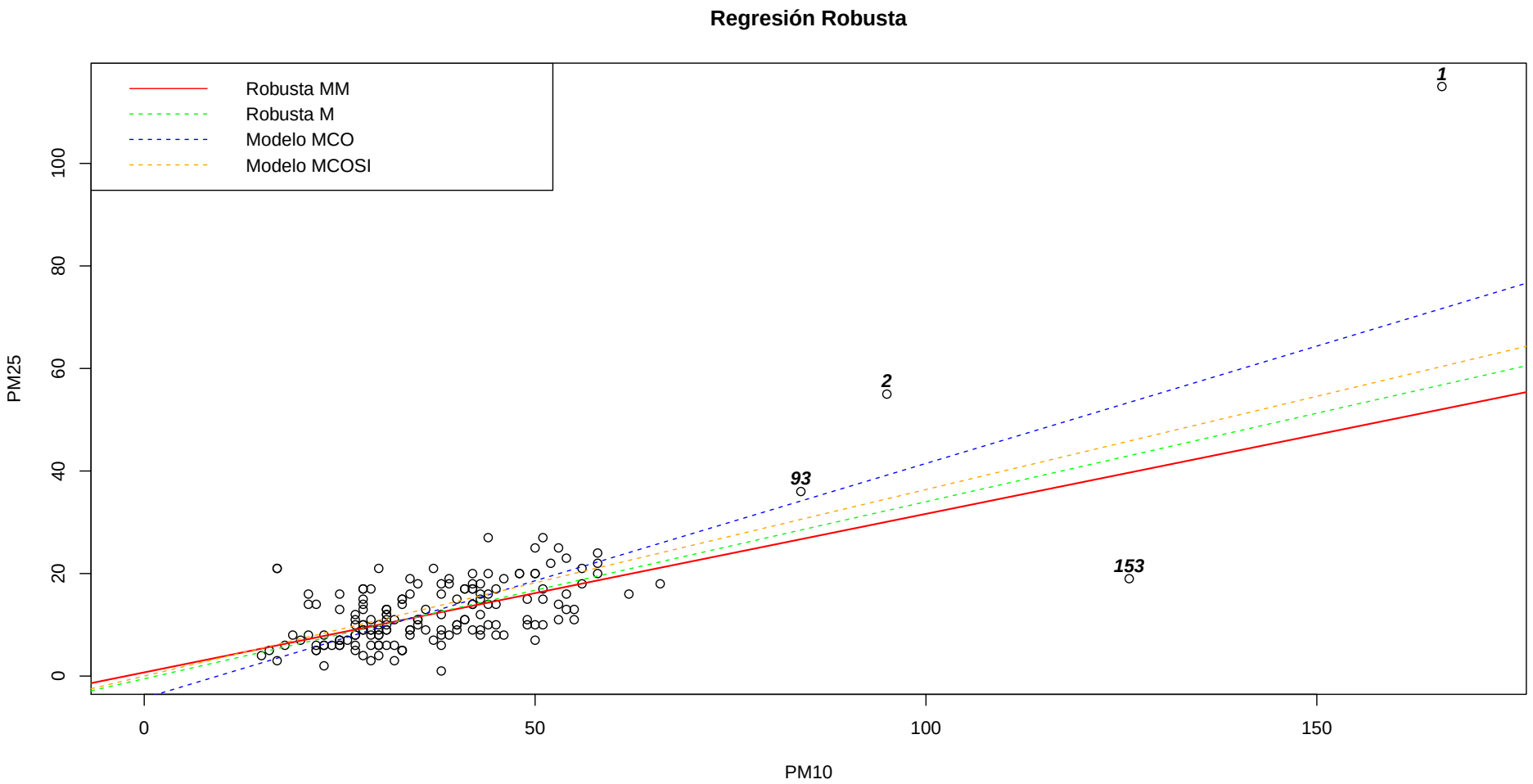
Regresión Robusta

- Método S: 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.
- Método MM: 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, el modelo ajustado usando este método fue:

$$\hat{PM}_{2,5} = 0{,}7196 + 0{,}3092 PM_{10}$$

Regresión Robusta



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	0.7196	0.9259	0.7771	0.4382
β_1	0.3092	0.0222	13.9372	0

$\hat{\sigma}$	Estadistico F	Valor p
4.824	194.2455	0

Comparación de modelos

Para comparar los modelos tenemos la siguiente tabla comparativa:

Modelos	$R^2_{Ajustado}$	$\hat{\sigma}$	Estadistico F	Valor P	AIC	BIC	$SD(\beta_1)$
Modelo 1	0.5729	6.7707	225	$< 2{,}2e-16$	1123.39	1132.762	0.03051
Modelo 2	0.8254	6.978	795.4	$< 2{,}2e-16$	1132.546	1138.794	0.0129
Modelo Robusto		4.824	194.2455	0	1147.121	1156.493	0.0222

Podemos concluir que el mejor modelo es el ( ), porque ( )