

PRÁCTICO GRUPAL:

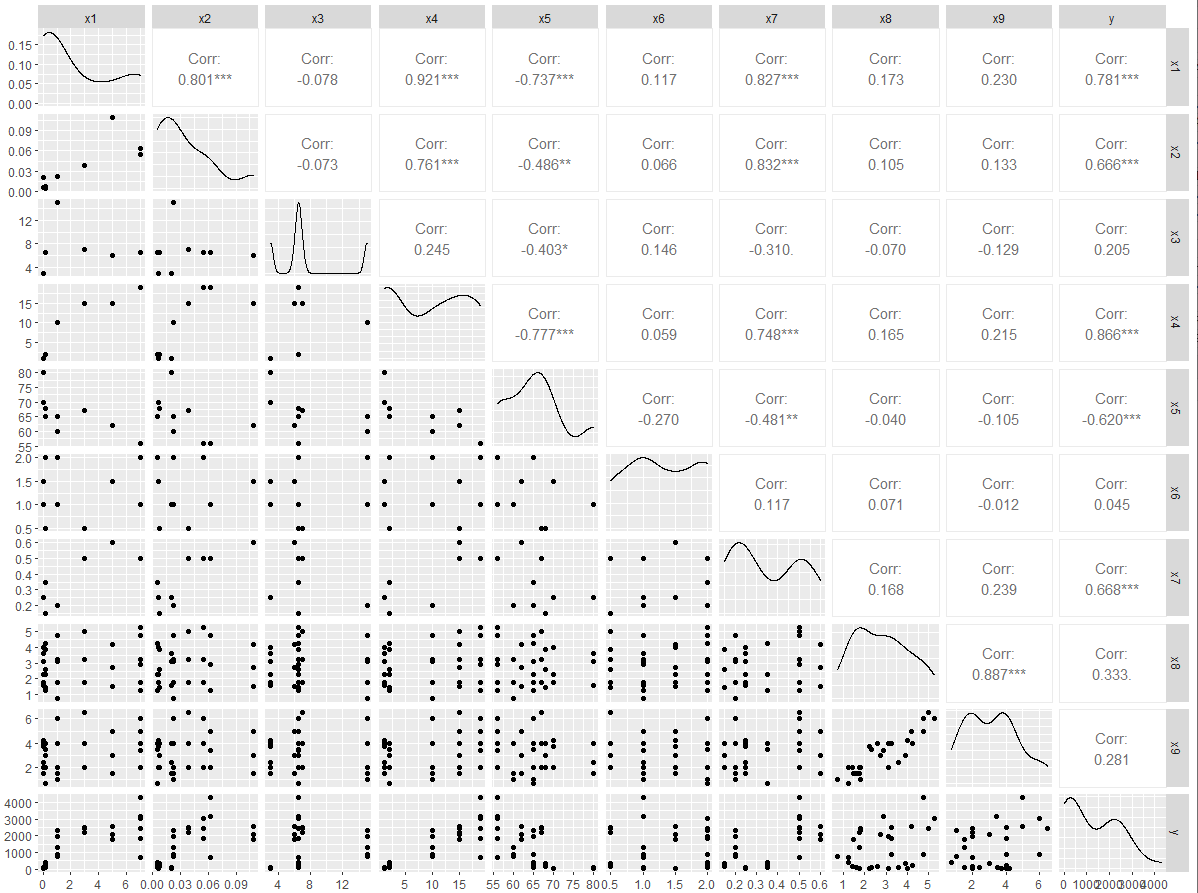
| INTEGRANTES | **D.N.I./LEGAJO** |
| --- | --- |
| Joaquín Bustamante | 670/17 |
| Israel Jaime | 576/11 |
| Hugo Cesar Galvan | 23785281 |

**AÑO 2021**

ÍTEM 1

Calculamos la matriz de correlación entre las variables x1, x2,..., x9, Y. Observando esta matriz se ve que las covariables que podrían ayudar a explicar la variable Y serían x1, x2, x4, x5 y x7, donde usamos como criterio para elegir que el coeficiente de correlación sea mayor o igual a 0.5.

Si tuviéramos que elegir una sería la de mayor correlación con Y. Es decir, x4, que tiene un coeficiente de 0.866.



ÍTEM 2

Realizamos una regresión lineal con intercept usando Y como variable dependiente y las demás variables como independientes.

El modelo propuesto es Y = a1.x1 + a2.x2 +...+ a9.x9 + e, donde suponemos que:

* Los errores “e” tienen una distribución normal
* E(e) = 0 para todos los errores
* Var(e) = sigma^2 para todos los errores

Calculamos el EMC

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 292.56 4428.62 0.066 0.9480

x1 -203.14 410.27 -0.495 0.6259

x2 1055.78 9833.70 0.107 0.9156

x3 -49.24 156.20 -0.315 0.7558

x4 209.76 162.05 1.294 0.2103

x5 -10.20 51.09 -0.200 0.8438

x6 -24.56 303.53 -0.081 0.9363

x7 142.78 3288.44 0.043 0.9658

x8 511.71 209.74 2.440 0.0241 \*

x9 -301.87 172.00 -1.755 0.0945 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 609.3 on 20 degrees of freedom

Multiple R-squared: 0.8214, Adjusted R-squared: 0.741

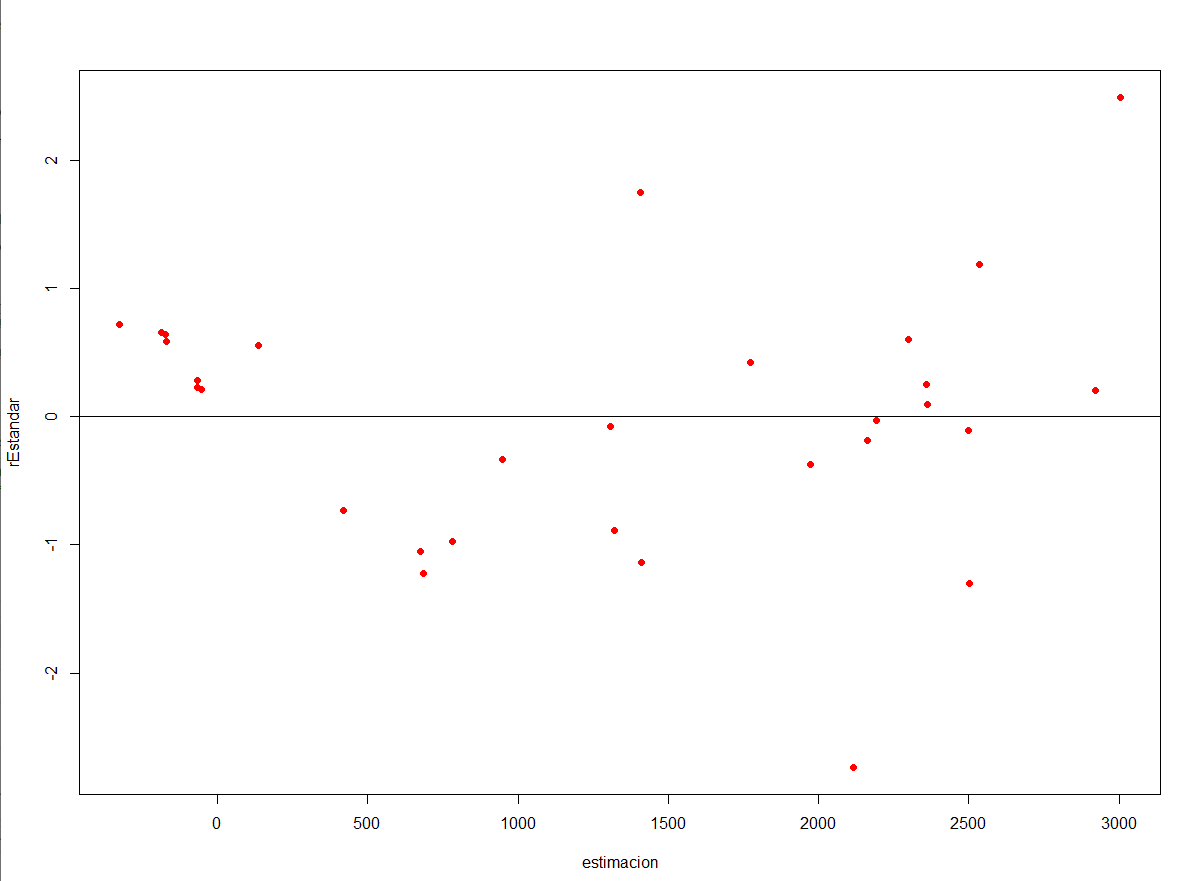
F-statistic: 10.22 on 9 and 20 DF, p-value: 9.744e-06

y obtuvimos la siguiente información:

* Para el test: “ai.est IGUAL a 0 vs ai.est DISTINTO de 0” vemos que a un nivel de significación 0.05 sólo hay evidencia suficiente para afirmar que no es 0 la variable x8, y a un nivel de significación 0.1 hay evidencia suficiente de que no son nulas las variables x8 y x9.
* La significación de la regresión viene dada por el test: “todos los coeficientes del estimador son 0 vs alguno de los coeficientes del estimador es distinto de 0”. Dado que el p-valor de dicho test es 9.74x10^-6<<0.05 concluimos que hay evidencia suficiente para afirmar que la regresión es significativa.

ÍTEM 3

Para el modelo planteado en el ITEM 2 realizamos el gráfico de valores predichos vs residuos estandarizados

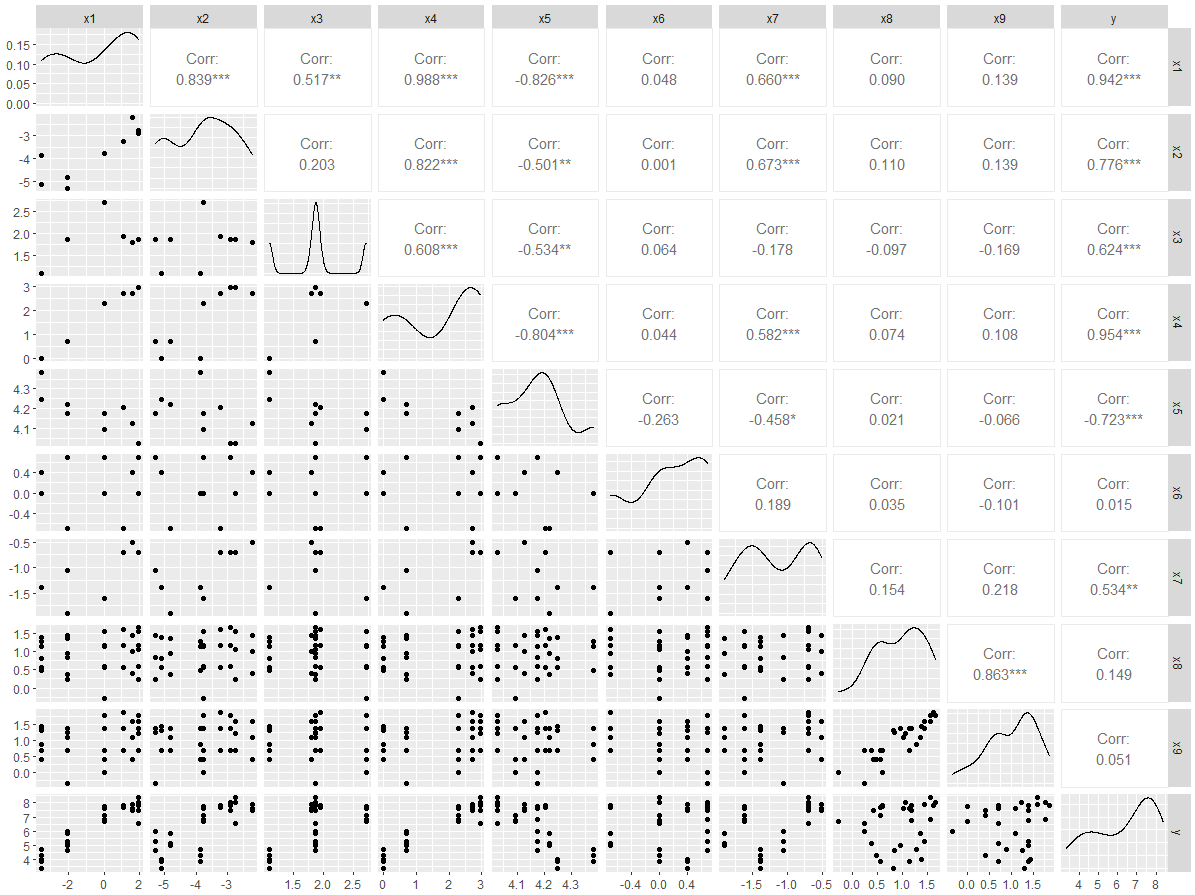


Como se puede ver, el gráfico no presenta ninguna estructura reconocible, sino que hay una nube de puntos distribuidos de forma aleatoria alrededor de la recta horizontal y = 0. Esto permite validar los supuestos de la regresión.

ÍTEM 4

Las correlaciones de los primeros cinco parámetros aumentan considerablemente con respecto al modelo del ÍTEM 2. El x7 bajó de 0.66 a 0.53.

Usando el criterio de usar las que tengan valor de correlación mayor o igual a 0.5 consideramos significativos x1, x2, x3, x4, x5 y x7.



ÍTEM 5

Replanteamos el modelo del ÍTEM 2, como Ln(Y) = a1Ln(x1) + a2Ln(x2) +...+ a9Ln(x9) + e

y los mismos supuestos respecto a los errores. Y los llamamos Ln.model

Summary del modelo

Coefficients:

Estimate Std. Error t value Pr(>|t|)

Intercept -0.93045 5.53556 -0.168 0.8682

x1 0.59299 0.24455 2.425 0.0249 \*

x2 -0.11077 0.13516 -0.820 0.4222

x3 0.28705 0.26897 1.067 0.2986

x4 0.34341 0.38318 0.896 0.3808

x5 1.44111 1.29147 1.116 0.2777

x6 -0.27121 0.12069 -2.247 0.0361 \*

x7 0.06234 0.25853 0.241 0.8119

x8 1.63613 0.18395 8.894 2.18e-08 \*\*\*

x9 -1.39735 0.16082 -8.689 3.18e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2148 on 20 degrees of freedom

Multiple R-squared: 0.9871, Adjusted R-squared: 0.9813

F-statistic: 169.8 on 9 and 20 DF, p-value: < 2.2e-16

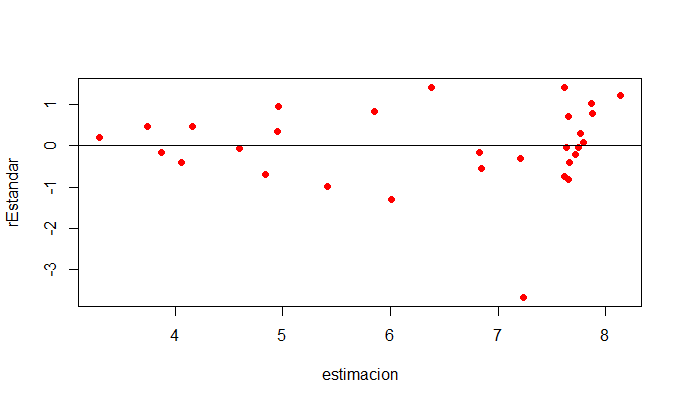
Pasamos de tener dos variables significativas a tener cuatro, es más, dos con un nivel de significación 0.001.

Las variables que son significativamente distintas de cero son x1, x6, x8 y x9, y por lo tanto eliminaríamos, guiándonos por este modelo, las covariables x2, x3, x4, x5 y x7.

El modelo es significativo ya que tenemos un p-value: 2.2e-16. Es decir, usando un test con nivel 0.05 rechazaríamos la hipótesis de que todas las beta-i sean distintas de cero.

ITEM 6

Para este nuevo modelo Ln.model hacemos el gráfico de predichos vs residuos estandarizados



La distribución de los errores estandarizados se concentran más alrededor del 0 en una franja más fina en comparación con el modelo del ÍTEM 2. Es decir, está dentro de las 2 desviaciones estándar, sosteniendo el supuesto de normalidad. Además, la distribución denota aleatoriedad, sin existencia de algún patrón.

ÍTEM 7

Al modelo del ítem 2 le aplicamos el procedimiento stepwise en la versión forward

* Usando como criterio de bondad el R^2 ajustado

ECM según modelo

[1] 0.7405769

[2] 0.7709796

[3] 0.7904453

[4] 0.7847644

[5] 0.7773775

[6] 0.7742061

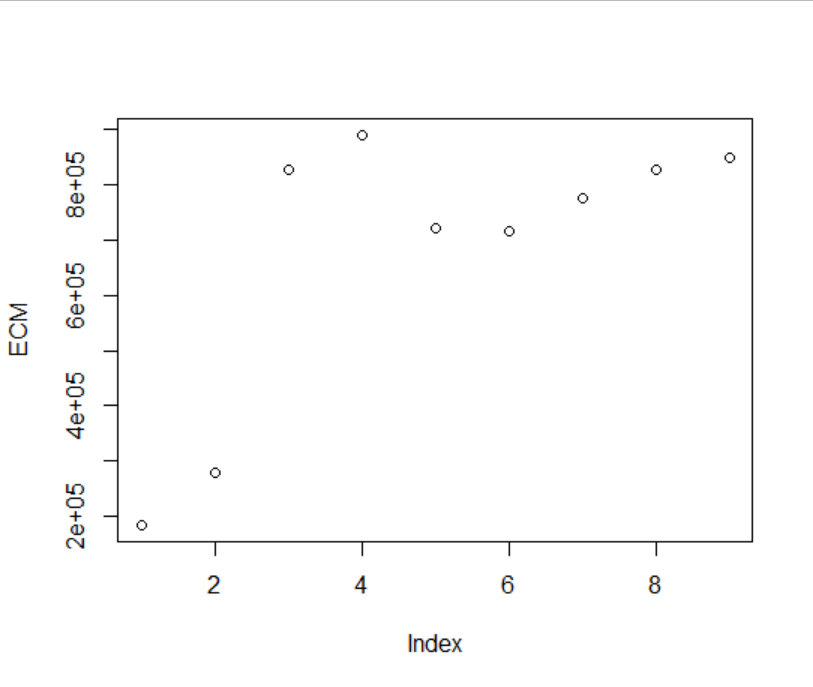
[7] 0.7644606

[8] 0.7533035

[9] 0.7409931

Observamos en el summary el retorno de regsubsets de cada modelo y vemos que el que tiene tres parámetros tiene el mayor R^2 ajustado (0.79) y toma el x4, x8 y x9. Esto nos indica que las variables mencionadas explican ‘y’ aproximadamente en un 80%.

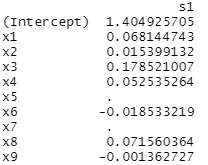
* Usando como criterio el Error cuadrático medio de validación cruzada



Observamos en el gráfico que el primer modelo es el de menor ECM = 184180 y consta de una intercept y la covariable x4.

ITEM 8

Haciendo LASSO con convalidación cruzada y utilizando la misma división de datos de entrenamiento y testeo se dejan de lado los parámetros x5 y x7 y el modelo tiene un ECM de 0.2.



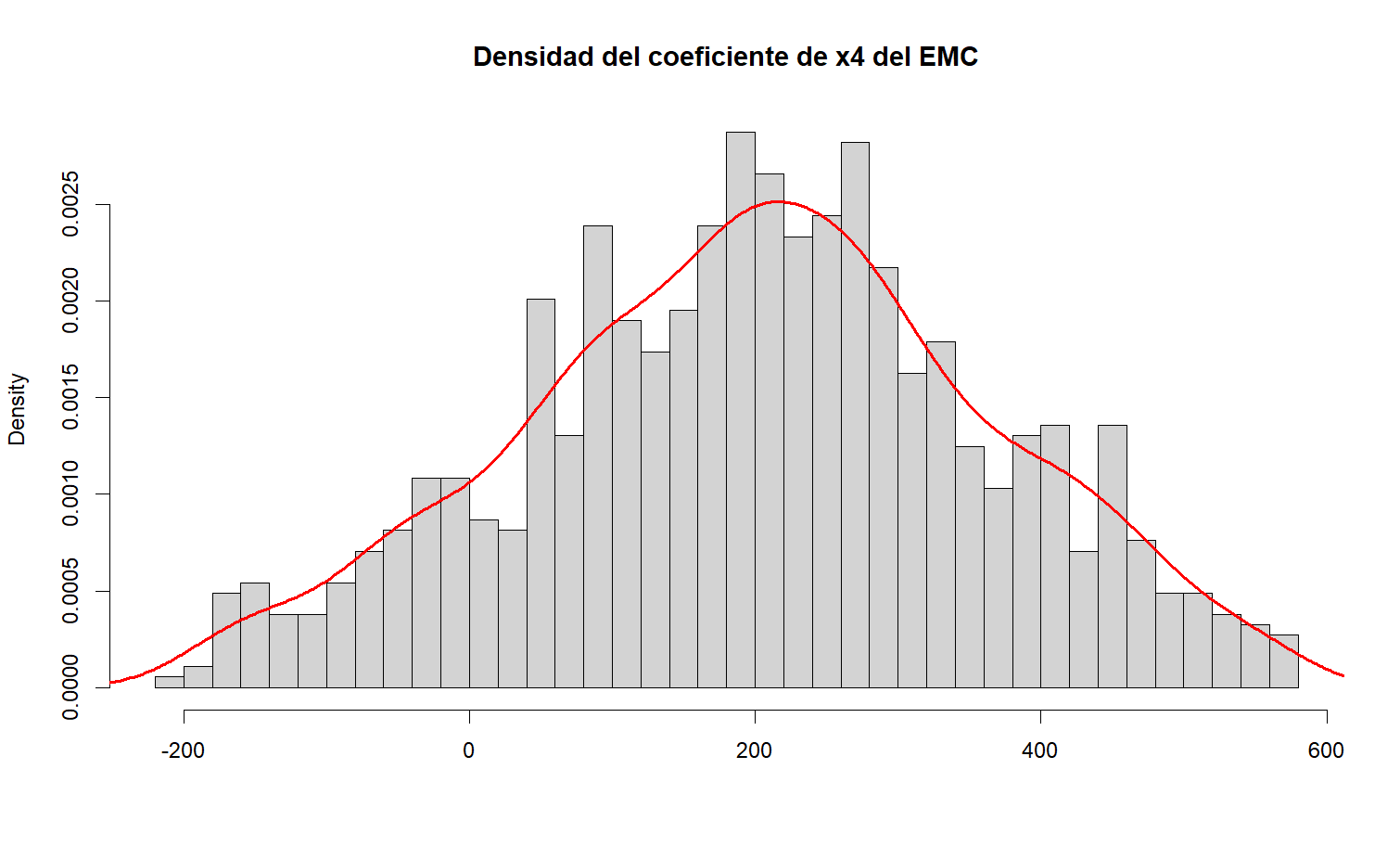
Consideramos el modelo del ítem 7 a) por tener un R^2 ajustado interesante y pocas variables (tres), sin embargo nos quedamos con el modelo del ítem 8 usando el criterio del ECM, dado que en este caso resulta muy pequeño, lo que nos dice que el modelo tiene un alto grado de predicción, la contra es que tiene 7 covariables.

ÍTEM 9

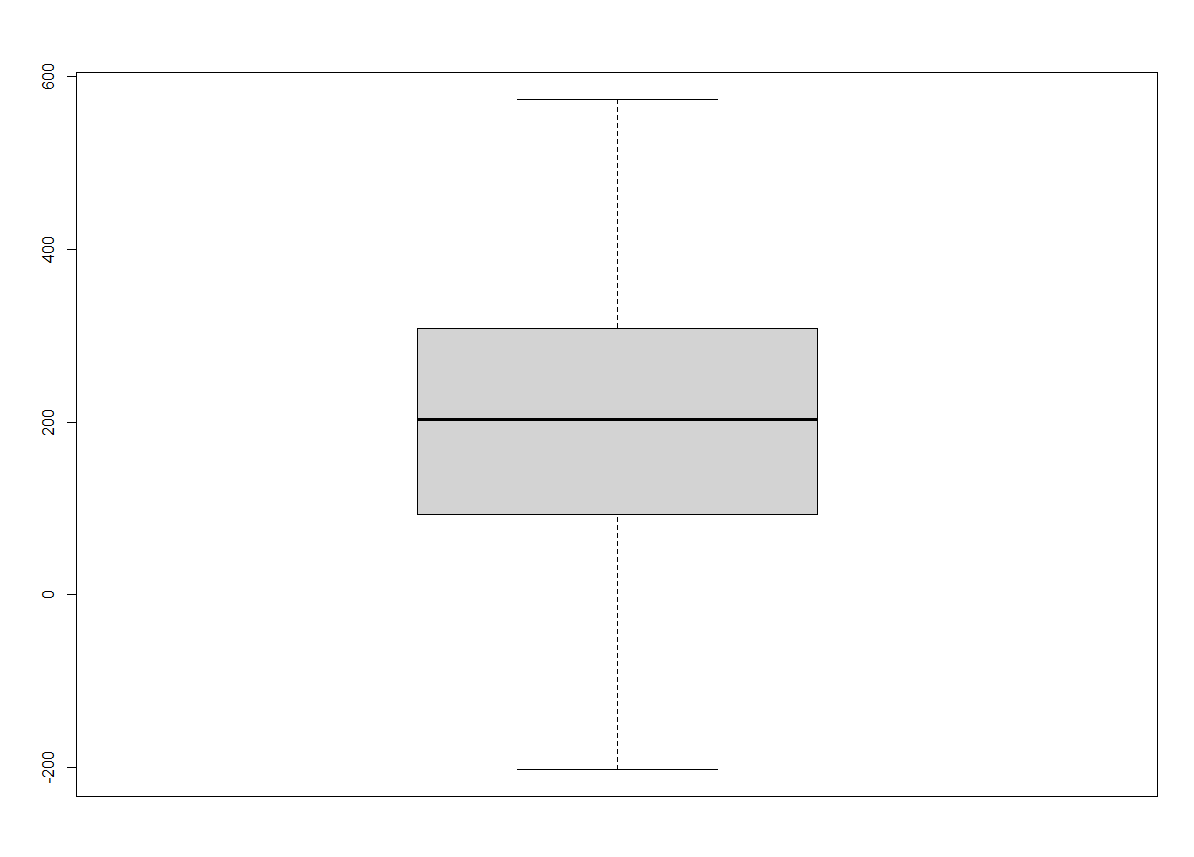
Usando el método Bootstrap no paramétrico a partir de la muestra de 30 observaciones de las variables x1, x2,..., x9, Y conseguimos otras 1000 muestras y para cada una de ellas ajustamos el modelo de regresión lineal del ÍTEM 2. De esta manera construimos 1000 observaciones de los estimadores de cuadrados mínimos.

En particular podemos hacer un análisis de la densidad de los mismos. En este caso, nos proponemos verificar la “normalidad” de los estimadores. Por ejemplo del coeficiente correspondiente a la covariable x4.

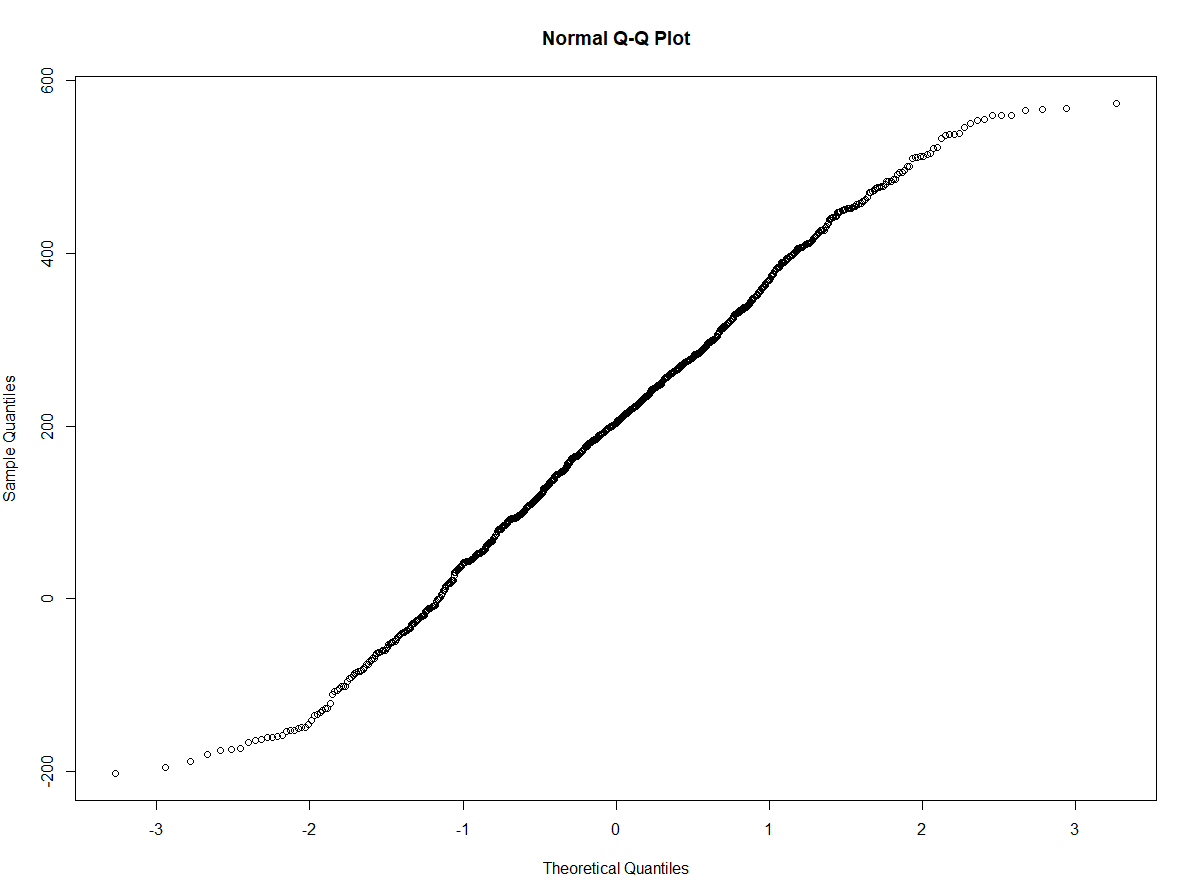
* En primer lugar, con el histograma y el gráfico de la densidad se observa una forma aproximadamente acampanada como la esperada en una distribución normal.



* Por otro lado, mirando el boxplot se ve claramente la distribución simétrica, habiendo dejado afuera los outliers.



* Finalmente, haciendo el QQplot se observa una buena correlación con la distribución normal, ya que salvo los extremos donde hay poca densidad de datos se ve una recta.



En base a los datos anteriores validamos la normalidad del estimador.