FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

APRENDIZAJE ESTADÍSTICO

2º CUATRIMESTRE 2020

Ejercicio - Regresión Lineal

AUTOR

Giampieri Mutti, Leonardo - #102 358
<lgiampieri@fi.uba.ar>

DOCENTE

García, Jemina María

Índice

1. Introducción **2**

2. Matriz de correlación **3**

3. Modelo Naïve **5**

4. Modelo sin intercept **11**

5. Modelo significativo **14**

6. Selección de modelos **18**

7. Conclusión **21**

8. Anexo **22**

1. Introducción

En el siguiente trabajo se pide relacionar los datos de distintas muestras de composiciones de cemento para analizar la relación que hay entre los componentes y el calor generado al fraguar las mismas. A continuación vemos el set de datos.

	x1	x2	x3	x4	x5	y
1	6.00	7.00	26.00	60.00	2.50	85.50
2	15.00	1.00	29.00	52.00	2.30	76.00
3	8.00	11.00	56.00	20.00	5.00	110.40
4	8.00	11.00	31.00	47.00	2.40	90.60
5	6.00	7.00	52.00	33.00	2.40	103.50
6	9.00	11.00	55.00	22.00	2.40	109.80
7	17.00	3.00	71.00	6.00	2.10	108.00
8	22.00	1.00	31.00	44.00	2.20	71.60
9	18.00	2.00	54.00	22.00	2.30	97.00
10	4.00	21.00	47.00	26.00	2.50	122.70
11	23.00	1.00	40.00	34.00	2.20	83.10
12	9.00	11.00	66.00	12.00	2.60	115.40
13	8.00	10.00	68.00	12.00	2.40	116.30
14	18.00	1.00	17.00	61.00	2.10	62.60

Las variables explicativas $\mathbf{x_i}$ indican el porcentaje de peso de cada componente en la muestra, $\mathbf{Y}$ indica la cantidad de calor generado al fraguar la misma. Buscamos ajustar un modelo mediante el método de regresión lineal, de la forma:

$$Y = X\beta + \varepsilon$$

En caso de ser la matriz de diseño de rango completo, vamos a poder encontrar una única solución para el vector β, mediante las ecuaciones normales.

2. Matriz de correlación

A continuación, analizo el grado de correlacion lineal que hay entre todas las variables del modelo, incluido **Y**. Para esto, represento la matriz de correlación en un gráfico tipo heatmap, sacando la diagonal y la mitad superior de la matriz, ya que es simétrica.

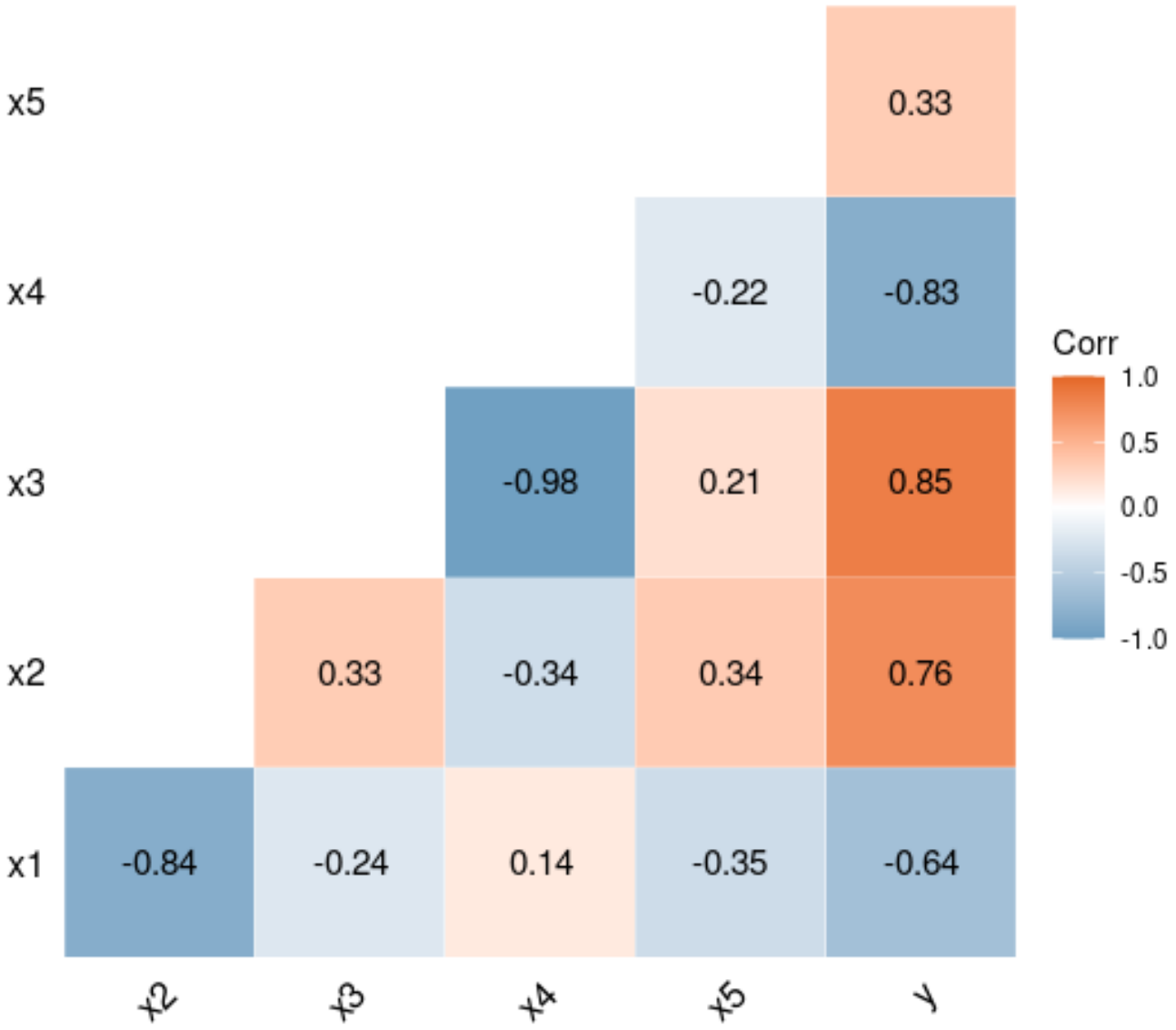

Figura 2.1

Notamos que, de tener que elegir a través de la matriz de correlación cuáles son aquellas variables que podrían contribuir significativamente para explicar a **Y**, tomaríamos a $\mathbf{X_3}$, ya que en módulo es la que mayor correlación tiene, seguida en orden de $\mathbf{X_4}$, $\mathbf{X_2}$, $\mathbf{X_1}$ y $\mathbf{X_5}$. Es importante notar también que hay una alta correlación entre $\mathbf{X_3}$ y $\mathbf{X_4}$, y lo mismo sucede con $\mathbf{X_1}$ y $\mathbf{X_2}$, lo que puede dar indicios de que en los modelos en donde se incluyan todas las variables, se introduzca multicolinealidad. Esto puede llegar a ser algo no deseado, ya que

las varianzas de los estimadores que se obtengan por el método de mínimos cuadrados no escalan, es decir, serían muy altas. Esto también repercute sobre todo el análisis estadístico que se haga, incluidos los p-valores, los intervalos de confianza y predicción, etc, de modo que puede que el análisis de qué variables son realmente significativas a la regresión se vuelva un poco más complicado.

3. Modelo Naïve

Empezamos ajustando un modelo lineal utilizando todas las variables, incluyendo intercept, de la forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i, \forall i \in [1, 14]$$

El ajuste de este modelo produce los siguientes resultados:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	73.61	105.97	0.69	0.51
x1	-0.45	1.13	-0.40	0.70
x2	1.30	1.07	1.22	0.26
x3	0.56	1.06	0.53	0.61
x4	-0.17	1.05	-0.16	0.88
x5	-0.39	1.52	-0.25	0.81

Los p-valores estan calculados con un nivel de significación 0.05. Notamos que en la columna 'Estimate' tenemos los estimadores de mínimos cuadrados, el vector $\hat{\beta}$. La columna 'Std. Error' nos da una idea de la varianza de la estimación para cada parámetro, y por lo que podemos ver es inmensa para la intercept. A partir de esta tabla podemos testear las siguientes hipótesis:

$$H_0 : \beta_i = 0 \quad vs \quad H_1 : \beta_i \neq 0$$

Esto es, queremos ver si X_i es una variable significativa para la regresión o no. El estadístico para cada uno de estos test, que se encuentra calculado para cada parámetro en la columna 't value', es tal que $t \sim t_{8;0,975}$. Podemos ver para cada beta si cae en la zona de rechazo, es decir, si podemos afirmar que la hipótesis de que $\beta_i = 0$ es verdadera.

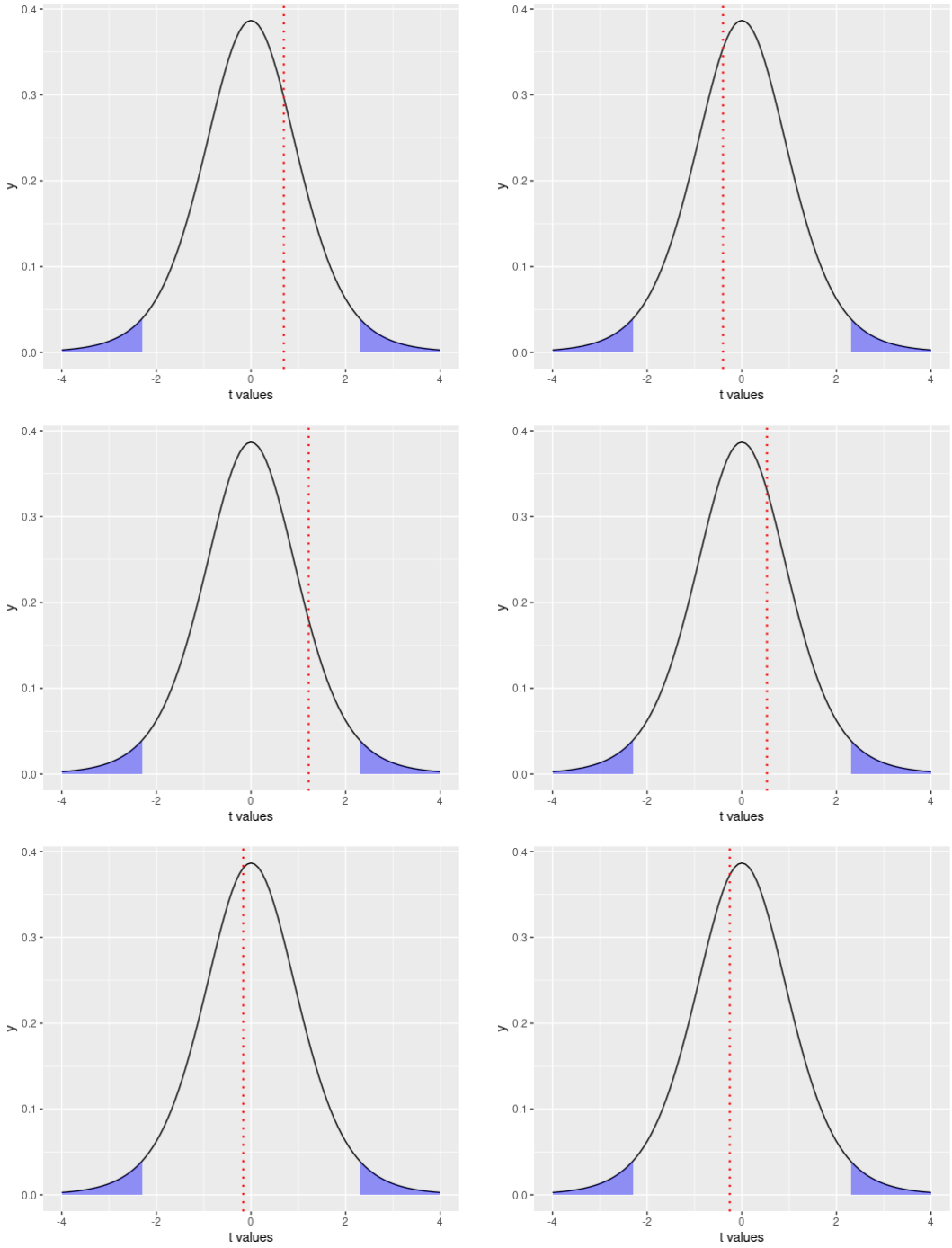

Figura 3.1

De izquierda a derecha vemos, para cada parámetro, la zona de rechazo marcada en azul. La línea punteada vertical representa el valor del estadístico del test. Vemos que ningún valor cae en la zona de rechazo. Se puede notar también porqué los p-valores son tan altos. A medida que mas se acerca la línea vertical al eje y, mayor es la probabilidad de que exista una muestra

más extrema bajo H_0. Mientras más cerca de cero el valor del estadístico, mayor el p-valor. Como ningún p-valor es menor que 0.05, no hay evidencia suficiente como para rechazar la hipótesis nula de que cualquiera de estas variables sea no significativa. Pero sin embargo la regresión sí es significativa.

	F-Statistic
value	122.18
p-value	2.48e-07

El test que se hace en este caso es el siguiente:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_5 = 0 \quad vs \quad H_1 : Algun \quad \beta_i \neq 0, i = 1, ..., 5$$

El estadístico del test en este caso es $F \sim F_{5,8,0,95}$

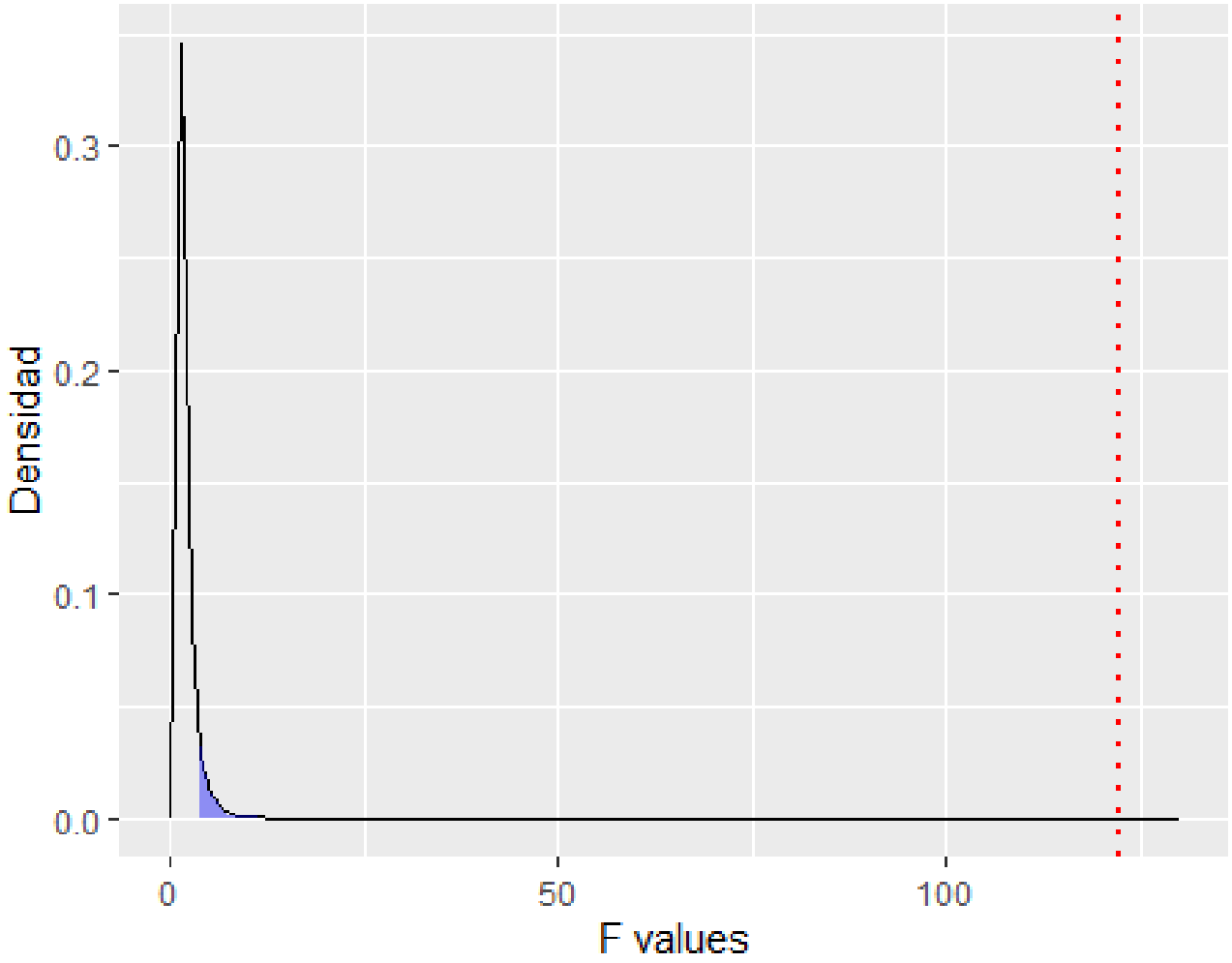

Figura 3.2

Notamos que rechazamos ahora la hipótesis de que no haya ninguna variable significativa, ya que el p-valor es órdenes de magnitud menor que el umbral de rechazo. Por lo tanto, la

regresión es significativa. Las contradicciones que se observan recaen sobre los p-valores y las correlaciones. Testeando individualmente por cada parámetro, no pudimos encontrar alguno que sea significativamente distinto de cero, pero sin embargo el test de la regresión indica que se puede explicar a $\mathbf{Y}$ en función de alguna de las otras variables. Además, la alta correlación que encontramos entre $\mathbf{X_3}$ e $\mathbf{Y}$ no la vemos reflejada en el resultado de los p-valores. Puede entonces que se introduzca ruido al considerar un modelo con todas las variables. Para verificar que los parámetros no son significativamente distintos de cero, calculamos los intervalos de confianza simples para cada uno.

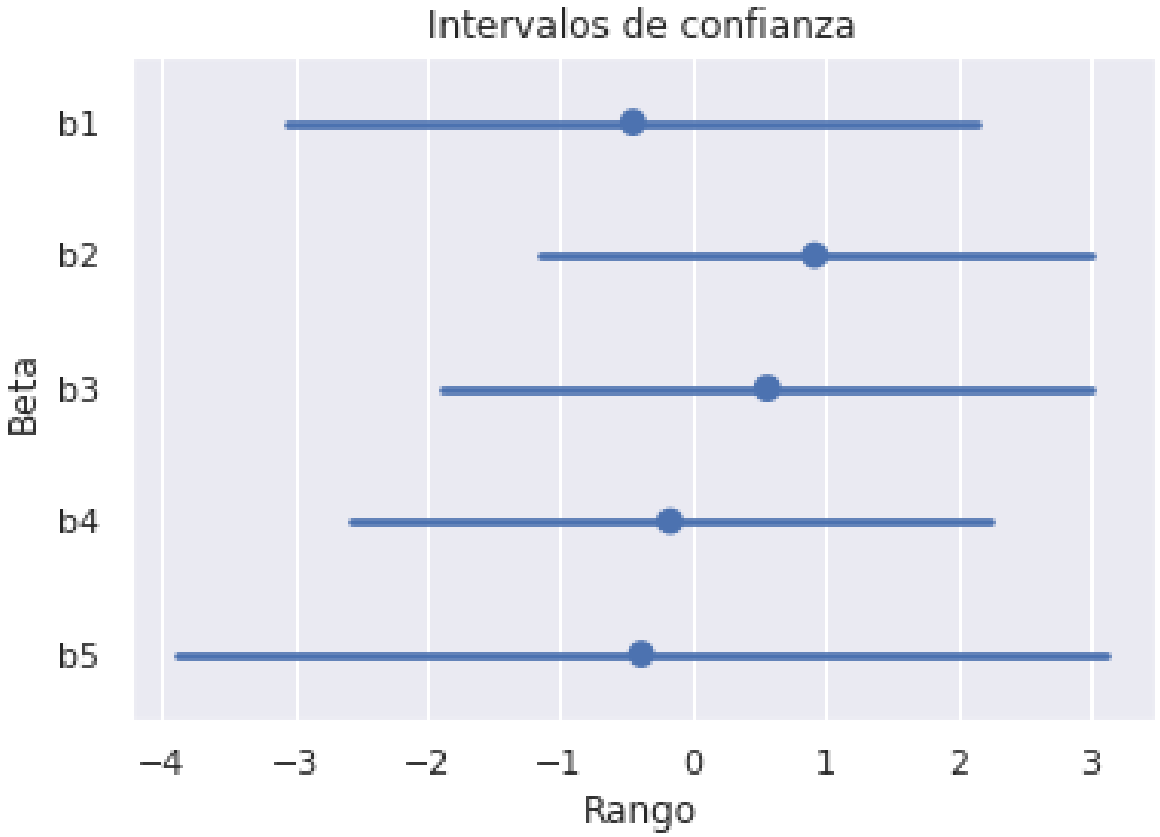

Figura 3.3

Estos intervalos se calcularon a partir del pivote para una distribución $t \sim t_8$, de nivel 0.95 cada uno.

Vale la pena hacer un nuevo modelo para ver que variables entran en la regresión, ya que no sabemos cuáles son significativas para la misma.

Queremos analizar si este modelo es válido, es decir, si se cumplen los supuestos del modelo lineal. No podemos visualizar ya que el gráfico sería en $\mathbb{R}^6$, pero si podemos analizar los residuos.

En primer lugar vemos el supuesto de esperanza nula de los errores.

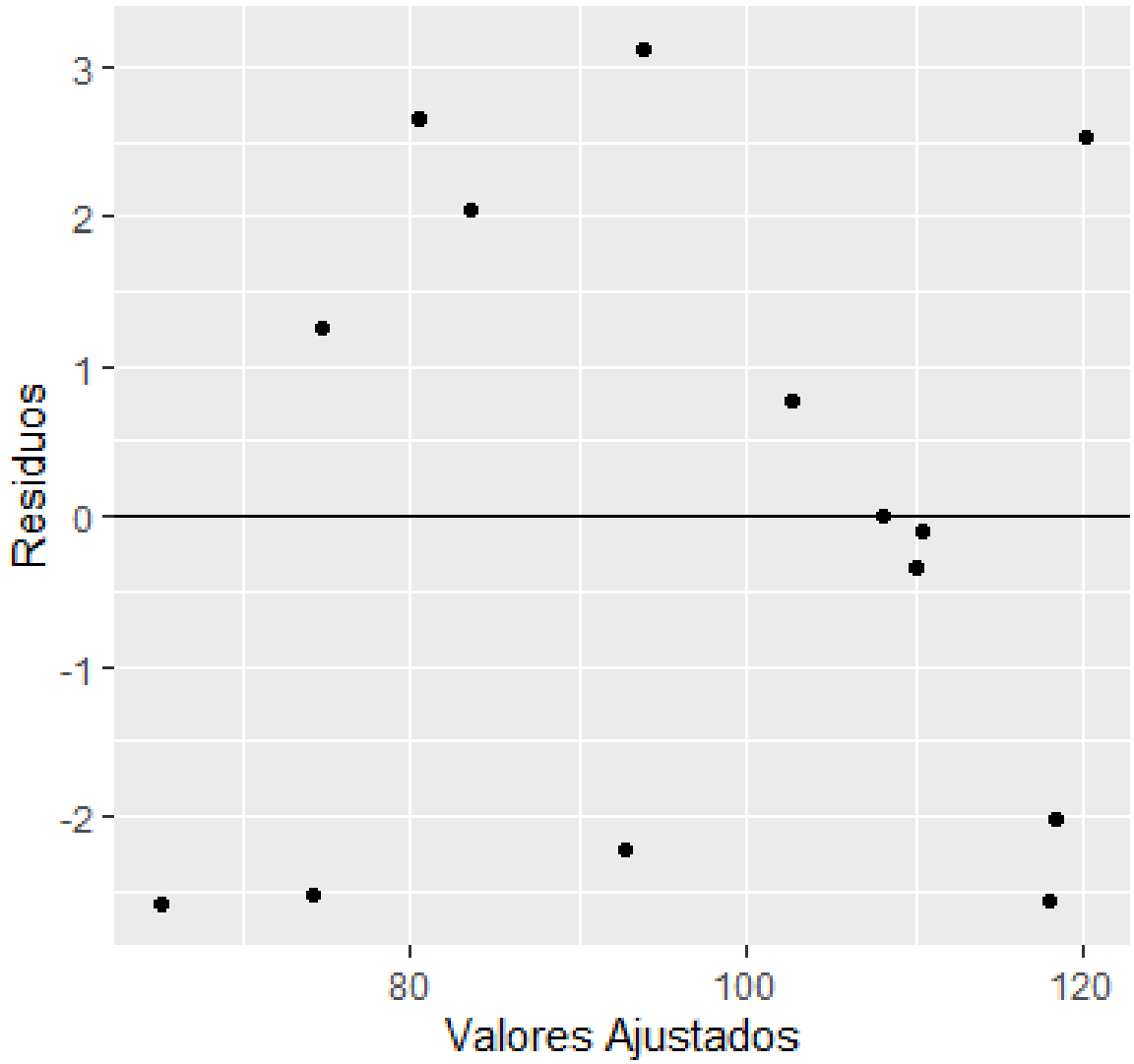

Figura 3.4

Si bien son pocos los puntos, no hay ninguna estructura aparente, de modo que se asume que los residuos tienen esperanza nula.

Ahora checkeamos la homocedasticidad. Queremos ver si la varianza de los errores es constante para todas las observaciones. Para esto, realizamos un qqplot.

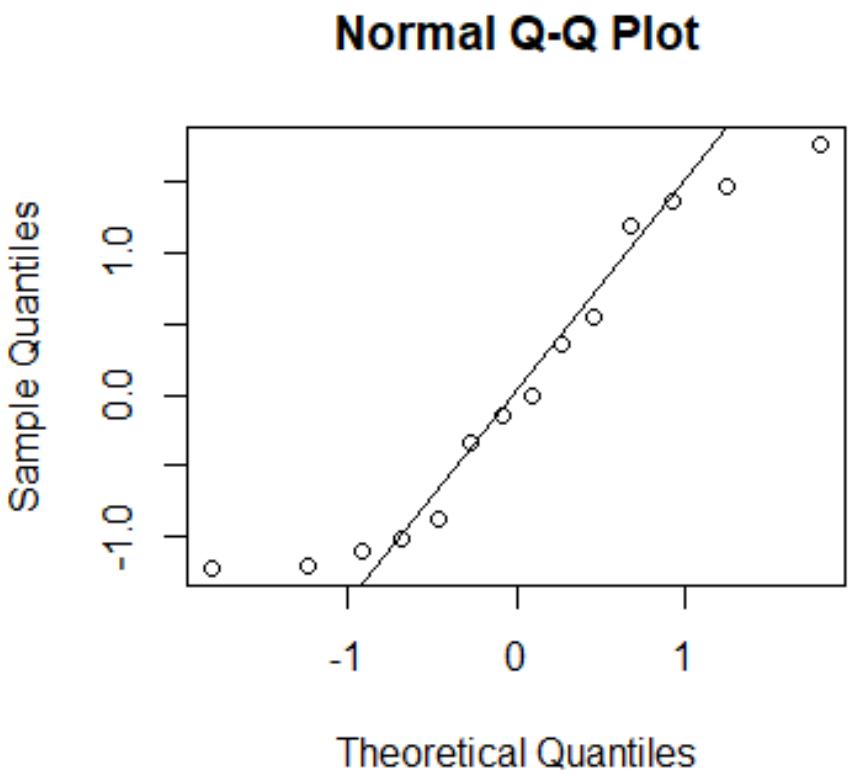

Figura 3.5

Nuevamente, son pocos los puntos para hacer un análisis exhaustivo, pero se nota la tendencia hacia "light tails". Esto indica que la distribución de los errores no es estrictamente normal, de modo que podemos decir que no se cumple el supuesto de normalidad, de modo que podemos descartar éste modelo para buscar uno nuevo.

4. Modelo sin intercept

Una de las razones por las cuáles se podría tener en cuenta este modelo es que, al ser las variables explicativas composicionales, la suma de las mismas resulta en una columna que es (casi) combinación lineal de la columna necesaria para el intercept en la matriz de diseño, la columna de unos.

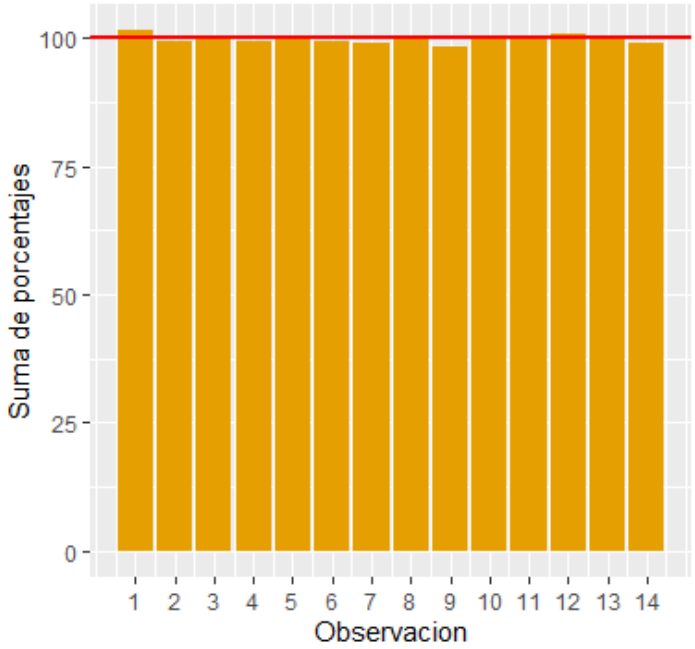

Figura 4.1

Aca explicamos el casi. Los datos no suman exactamente 100 % del peso del cemento, ya sea por errores de redondeo o por otra causa, por ende, la matriz de diseño sigue siendo de rango completo, pero sin embargo está mal condicionada, $\mathbf{X^T X}$ resulta ser una matriz cercana a singular. La justificación del parecido entre los totales es que, como las variables son porcentajes, cada una representa su aporte a la composición en el cemento, de modo que la suma debería dar el total, en porcentaje, de la masa de la muestra. De todas maneras, mas allá de tener o no intercept, de conservar todas las variables para ajustar un nuevo modelo, seguiríamos teniendo multicolinealidad, ya que como vimos en la matriz de correlación, $\mathbf{X_3}$ y $\mathbf{X_4}$ están fuertemente correlacionados, ídem $\mathbf{X_1}$ y $\mathbf{X_2}$. Además, el intercept captura todo aquello que el modelo no puede explicar. Droppear el intercept se podría interpretar de la siguiente manera: al no tener composición de ningún material, no se debería generar ningún calor. Esto es cierto a nivel físico, pero lo que implica es que nuestro modelo debería ser perfectamente lineal desde donde interpolamos hasta el origen, lo cual no sabemos, ya que no tenemos datos cercanos al origen. De hecho, tampoco tenemos el origen. Recordamos que el modelo lineal tampoco es excelente para extrapolar nuestros datos. Al extrapolar hacia el origen, lograríamos meter ruido en nuestros estimadores.

Ajustamos un nuevo modelo sin usar intercept.

De la tabla vemos que las variables más significativas son $\mathbf{X_2}$, $\mathbf{X_3}$ y $\mathbf{X_4}$, ya que tienen un

	Estimate	Std. Error	t value	Pr(>|t|)
x1	0.33	0.17	1.91	0.0883
x2	2.03	0.21	9.83	4.14e-06
x3	1.30	0.06	21.65	4.52e-09
x4	0.56	0.05	11.07	1.53e-06
x5	0.35	1.05	0.34	0.745

p-valor menor a la significación del test. Esto es diferente al modelo anterior, en donde no se tenía una claridad sobre que variables contribuian a explicar al modelo, y acá se ve claro. Se ve también que los valores estimados cambiaron significativamente respecto a los del previo modelo.

La regresión es significativa.

	F-Statistic
value	3932.64
p-value	9.69e-15

Estudiamos ahora los residuos del modelo. Primero, estudiamos si los residuos son homecedásticos.

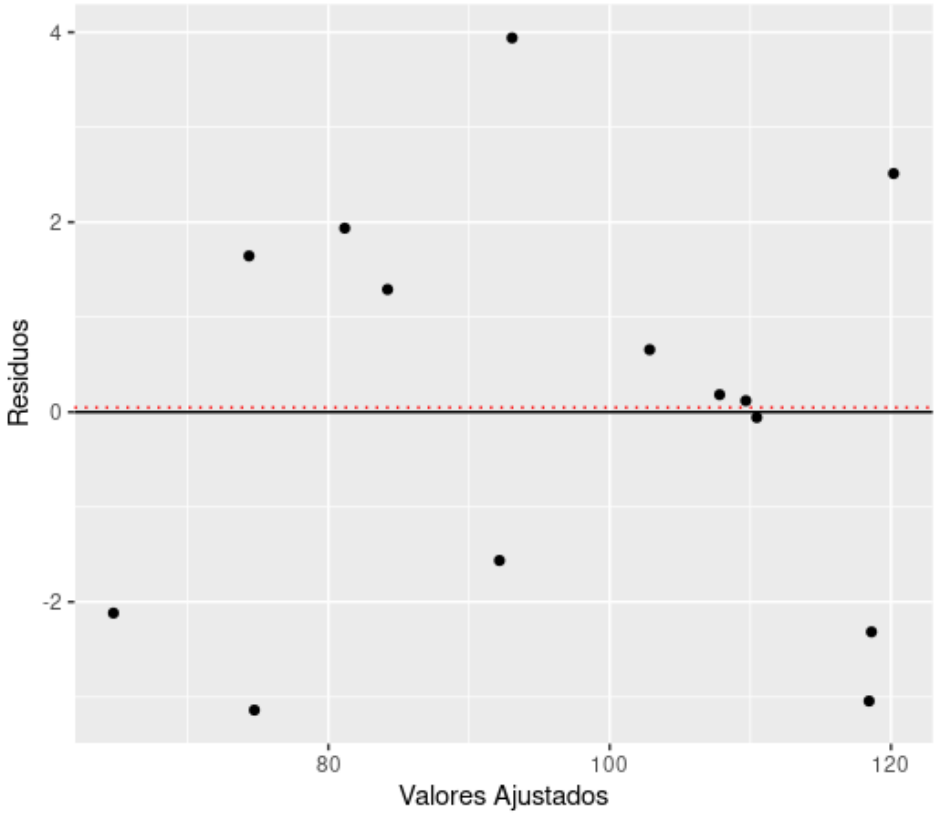

Figura 4.2

Notamos que si bien no hay una estructura aparente, la linea punteada que se ve representa la suma de los residuos. Es significativamente distinta de cero, ya que corta al eje y en 0,0478. Esto muestra que lo que no puede explicar la intercept, lo compensan los residuos, por lo tanto, si bien la regresión es significativa, no se justifica droppear la intercept.

Vemos ahora si se cumple el supuesto de normalidad de los residuos.

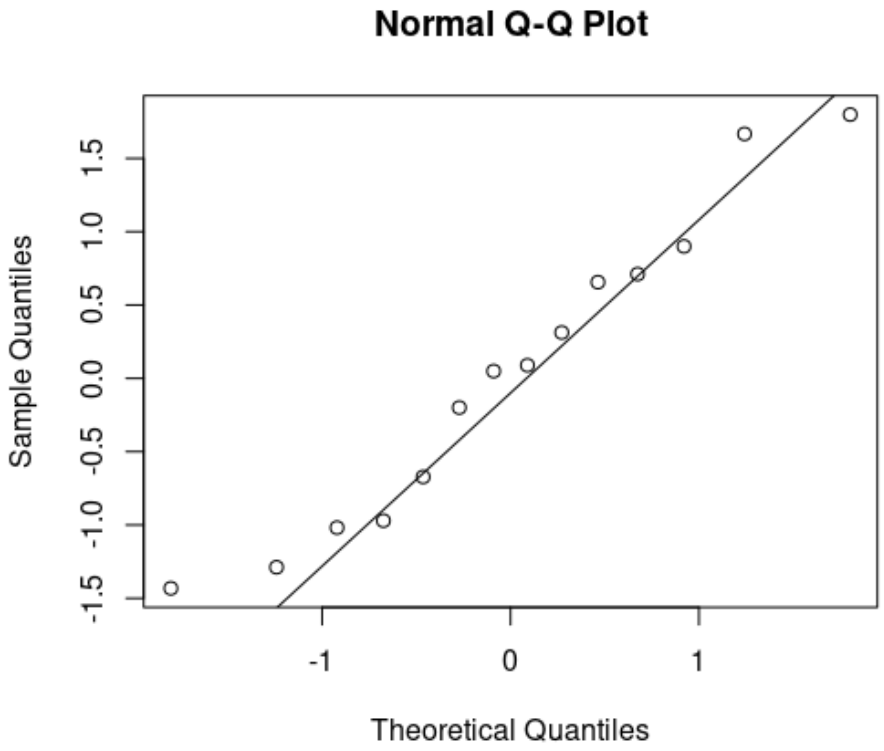

Figura 4.3

No parece haber ni colas en los extremos ni curvatura en el medio, por lo cual podemos decir que se cumple el supuesto de normalidad.

Buscamos ahora otro modelo que contenga a las variables significativas que encontramos en este modelo.

5. Modelo significativo

Para ajustar el siguiente modelo, utilizamos la intercept y las variables que en el ajuste anterior resultaron significativas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.0783	14.6424	2.19	0.0533
x2	1.7207	0.1218	14.12	6.23e-08
x3	0.9752	0.1915	5.09	0.00047
x4	0.2388	0.1818	1.31	0.218

Observamos que ahora no hay evidencia suficiente para decir que $\mathbf{X_4}$ es significativa para la regresión. Mientras que si lo son $\mathbf{X_2}$ y $\mathbf{X_3}$. ¿Qué sucede?. Esto sucede debido a la alta correlación entre $\mathbf{X_3}$ y $\mathbf{X_4}$ que vimos en el primer heatmap. De nuevo notamos como escala la varianza del intercept, ya que trata de amortiguar el problema que trae la multicolinealidad.

Analizamos la relación entre $\mathbf{X_3}$ y $\mathbf{X_4}$.

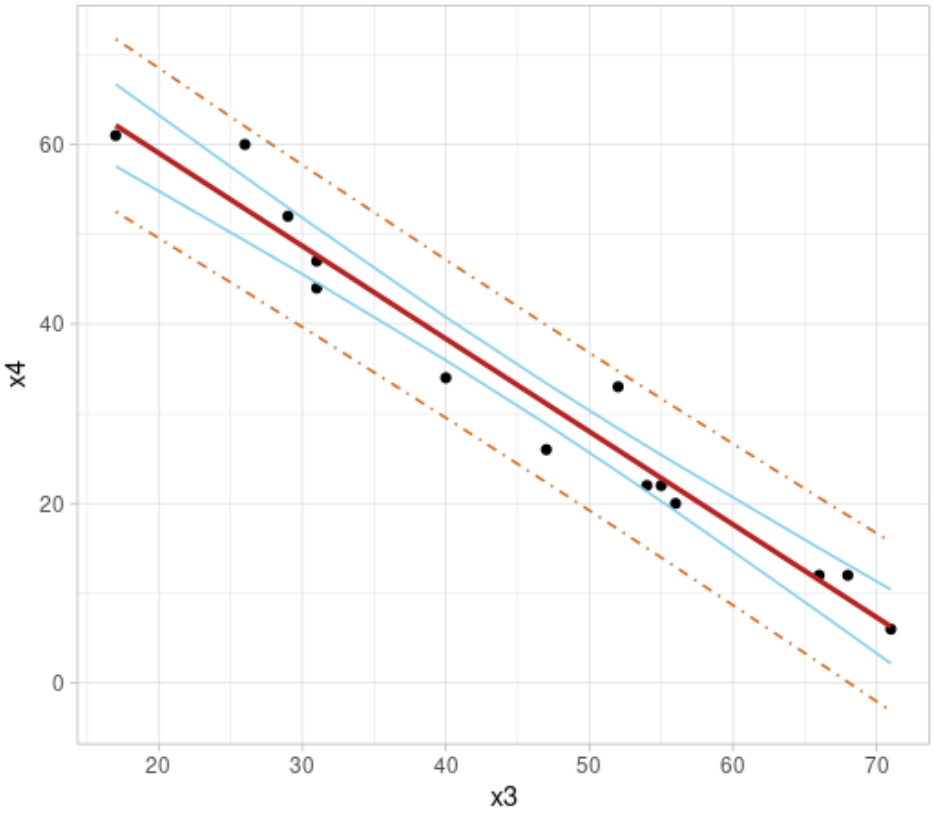

Figura 5.1

Notamos que las proporciones siguen una relación inversa, a medida que aumenta el porcentaje de una disminuye el porcentaje de la otra, aparentemente de manera lineal. La regresión es significativa, con un $p - value = 1{,}373e - 09$.

Este mismo análisis lo podemos realizar con $\mathbf{X_1}$ y $\mathbf{X_2}$.

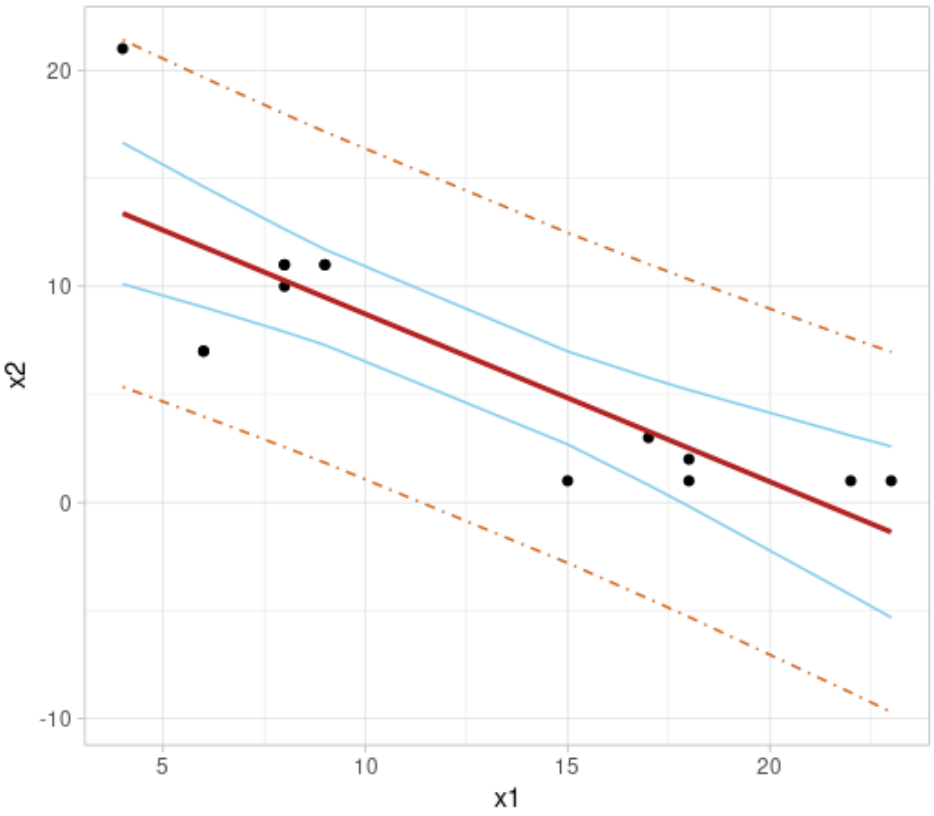

Figura 5.2

La regresión es significativa con un $p - value = 0{,}0001879$.

El punto es que si utilizamos estas variables en conjunto, seguro vamos a introducir colinealidad al modelo.

Analizamos los residuos de este modelo.

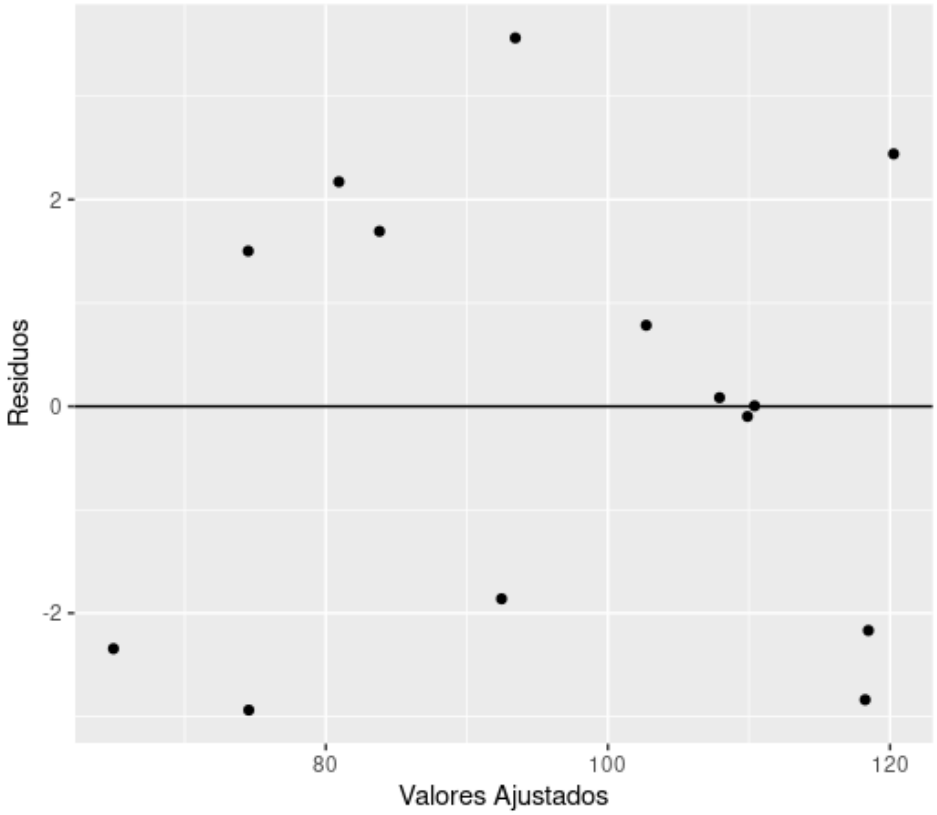

Figura 5.3

Nuevamente, ninguna estructura se presenta, por lo que se concluye que la varianza de los residuos es constante.

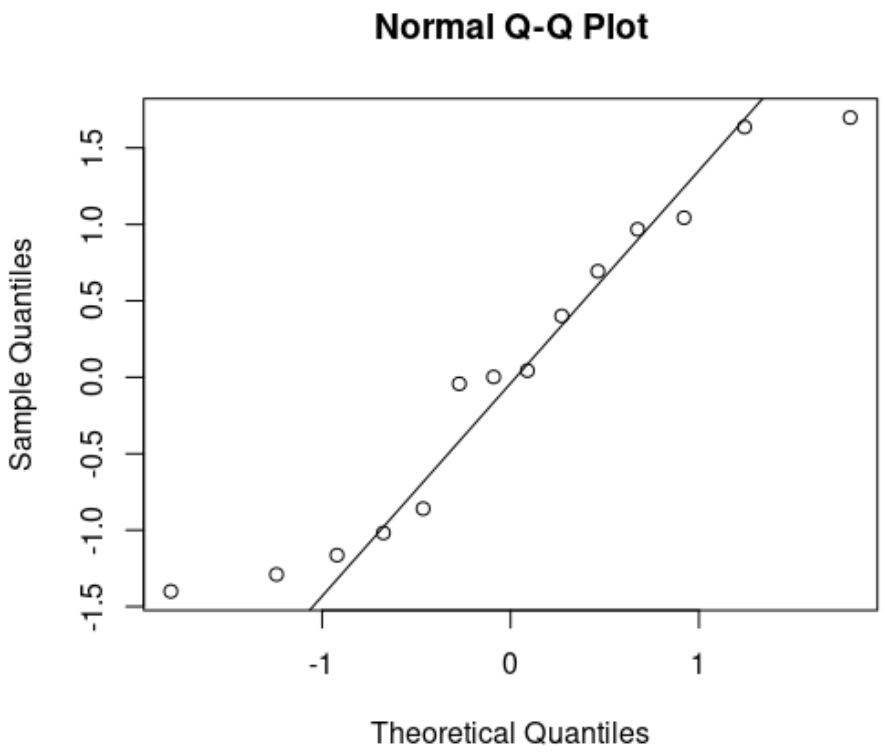

Figura 5.4

Acá notamos que hay una leve tendencia hacia las light tails, símil al primer modelo. Decimos entonces que los residuos no tienen distribución normal.

6. Selección de modelos

Vimos que sacar la intersección generalmente es una mala idea, sobre todo cuando no podemos visualizar el hiperplano generado por el método de cuadrados mínimos, por lo cual descartamos el modelo sin intercept. De elegir entre los dos restantes analizados, habría que analizar si el poder predictivo de ambos es similar. En ese caso, elegiríamos aquel con menor cantidad de variables. Como vimos, las columnas con alta correlación no estan perfectamente correlacionadas, de modo que no podemos expresar totalmente una variable en función de la otra, pero sin embargo, al incluirlas todas se vuelven los estimadores más variables, lo cual no es deseado. A modo de decidir cual es el modelo ideal, utilizamos la librería **leaps** con una búsqueda exhaustiva (ya que son pocas observaciones), utilizando el criterio de Cp de Mallow. Sabemos que si el calculo del Cp se aproxima a la cantidad de variables utilizadas, entonces puedo expresar buena parte de la variabilidad que tendría utilizando todas las variables, pero utilizando menor cantidad.

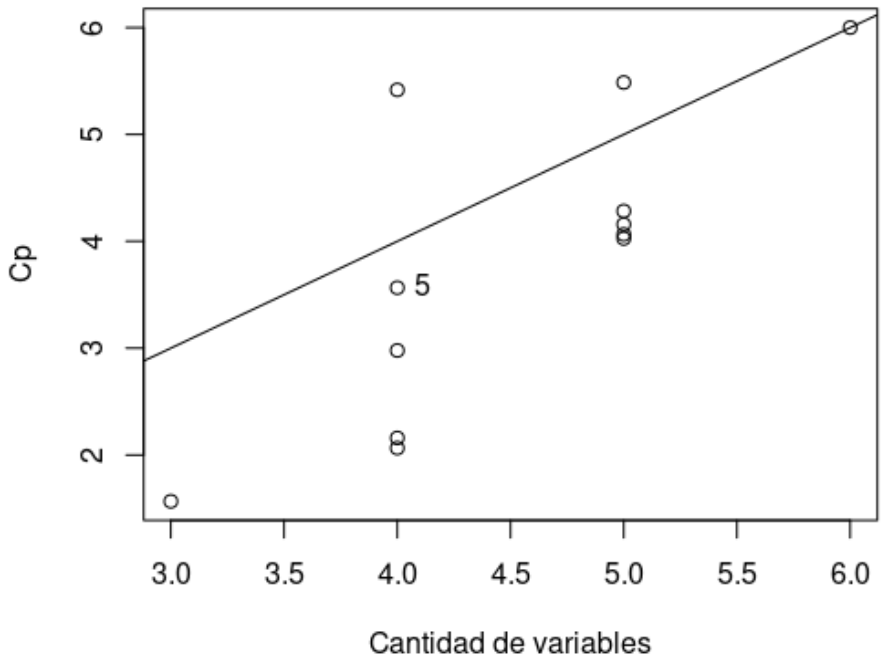

Figura 6.1

La cantidad de variables incluye a la intercept. Filtramos todos aquellos modelos que tengan Cp mayor a 6, ya que no tenemos más de 6 variables explicativas en el modelo. El punto titulado con un '5' es el modelo que decidimos quedarnos, ya que es el que más se aproxima al valor de $p = 4$ ($Cp = 3{,}565100$). Este modelo hace uso del intercept y $\mathbf{X_2}$, $\mathbf{X_3}$ y $\mathbf{X_5}$. Vemos que las variables que se incluyen son las que maximizan la correlación con $\mathbf{Y}$, y

que además no se incluyen las variables que inflan la varianza, $\mathbf{X_1}$ y $\mathbf{X_4}$, ya que su variabilidad puede ser explicada a través de $\mathbf{X_2}$ y $\mathbf{X_3}$, respectivamente.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	51.1986	3.0489	16.79	1.18e-08
x2	1.7083	0.1375	12.42	2.11e-07
x3	0.7300	0.0457	15.98	1.90e-08
x5	-0.0299	1.0791	-0.03	0.9785

La regresión es significativa.

	F-Statistic
value	212.34
p-value	2.372e-09

Analizamos los residuos del modelo.

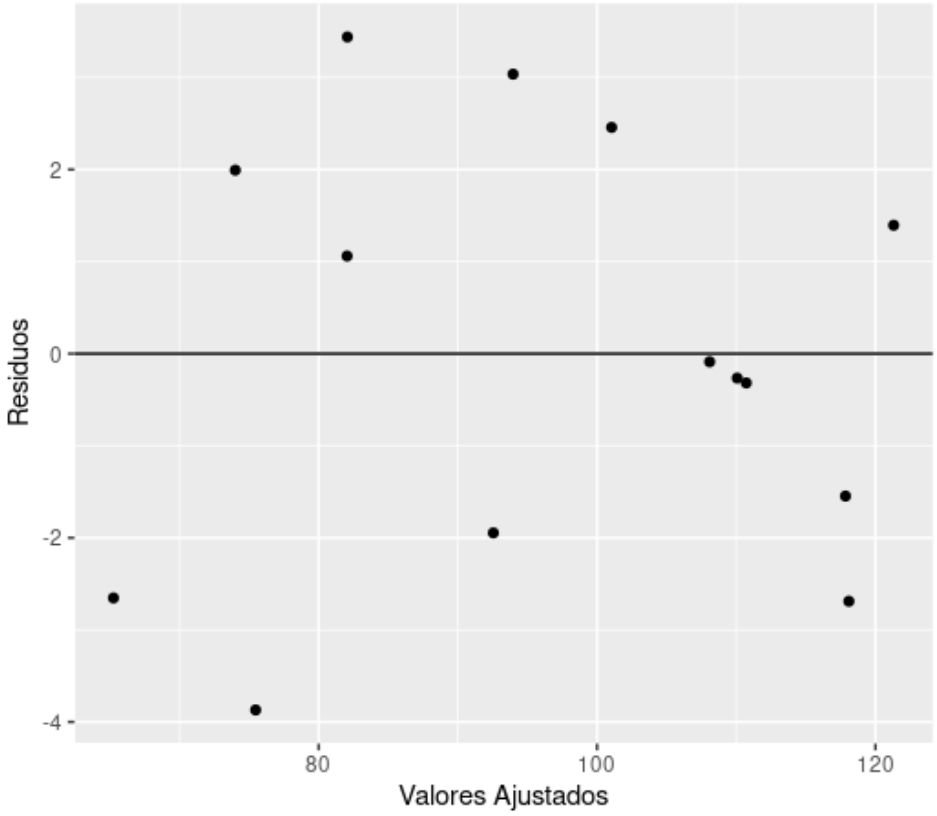

Figura 6.2

No vemos ninguna estructura. Los residuos son homocedásticos.

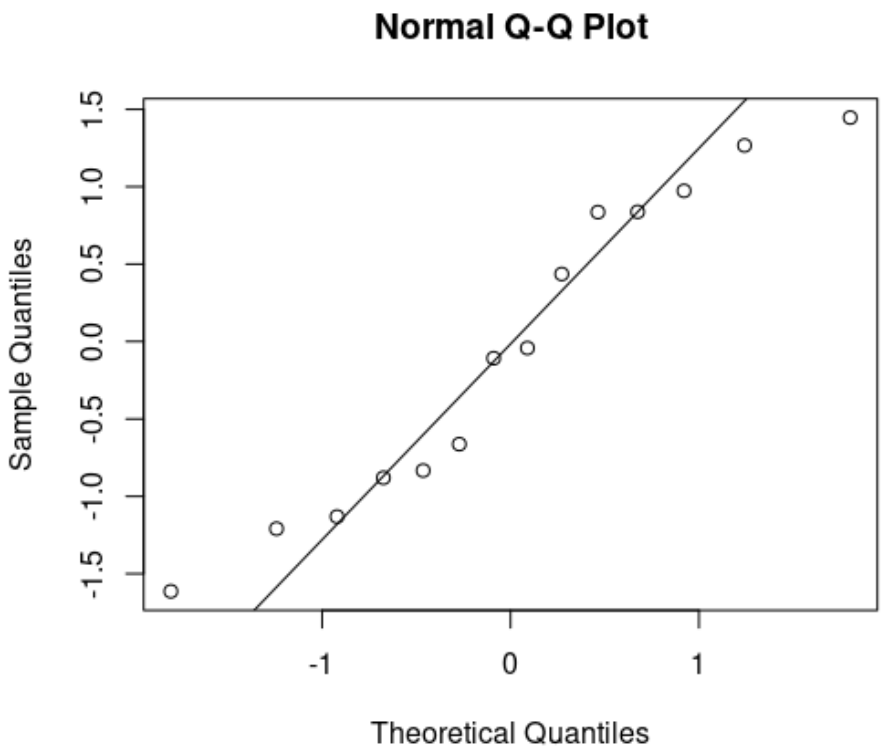

Figura 6.3

Se nota que 2 puntos que se alejan significativamente de la recta, pero sin embargo el resto se ajusta, de modo que consideramos que los residuos se distribuyen de manera normal.

7. Conclusión

En este trabajo se vió cómo impacta en la varianza de los estimadores y en el cálculo estadístico en general la presencia de colinealidad en los datos de entrada. También vimos que remover la intercept de un modelo por lo general no es una buena idea, ya que el mismo pierde expresividad, y que no necesariamente todas las variables aportan significativamente a explicar la variabilidad de la respuesta. Esto último lleva a comparar distintos modelos, combinando las variables de entrada y utilizando alguna medida, el Cp de Mallow, para poder seleccionar un buen modelo.

8. Anexo

Dejo el repositorio en donde se encuentran el codigo y los plots utilizados.
https://github.com/leogm99/Aprendizaje_Estadistico_tp1