Tarea 5 Estadística Multivariada

Code

Hairo Ulises Miranda Belmonte

15 de Marzo del 2019

EJERCICIO 1

1. Uso de la regresó lineal simple.

(a) Utilice el conjunto de datos /states.rds/.

Code

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

Code

Code

(b) Ajusta un modelo que prediga la energía consumida per capita (energy) con respecto al porcentaje de residentes que viven en áreas metropolitanas (metro). Reporta lo siguiente:

- Examina / gráfica los datos antes de aplicar el modelo

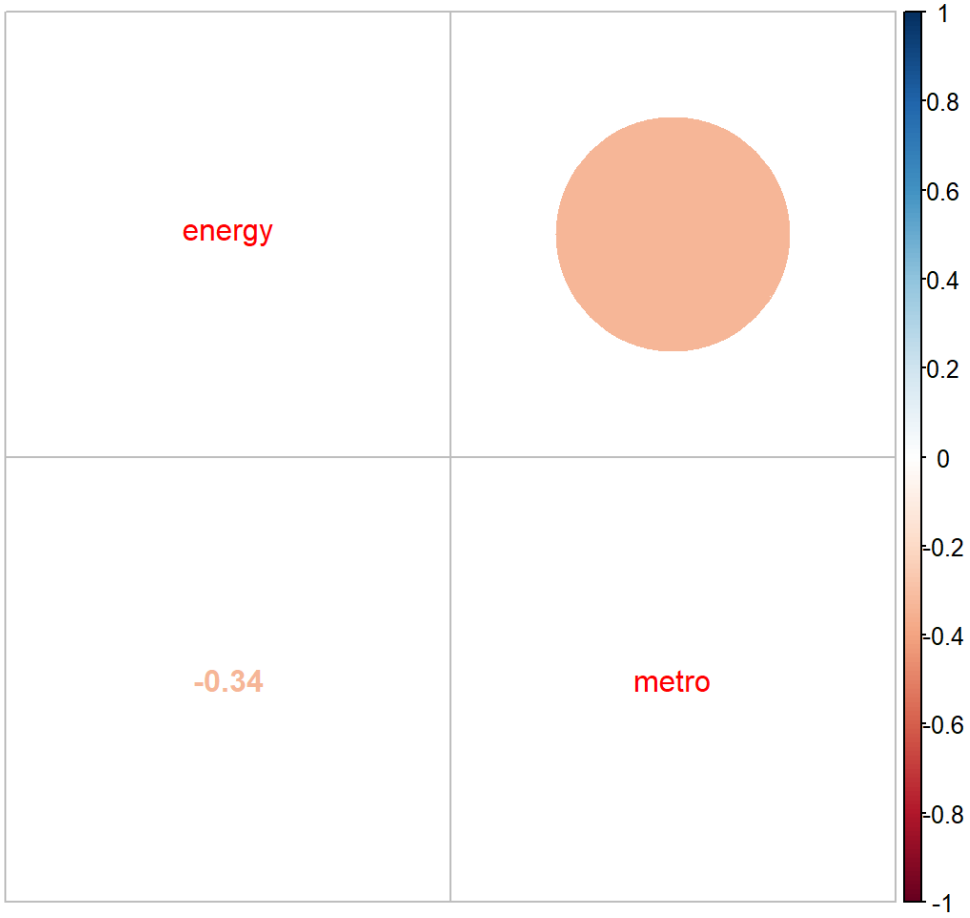
Code

Previo a realizar el análisis gráfico retiramos los valores NA a las variables *energy* y *metro*

Code

Se realiza un gráfico de correlación para observar la relación lineal entre las dos variables de interes.

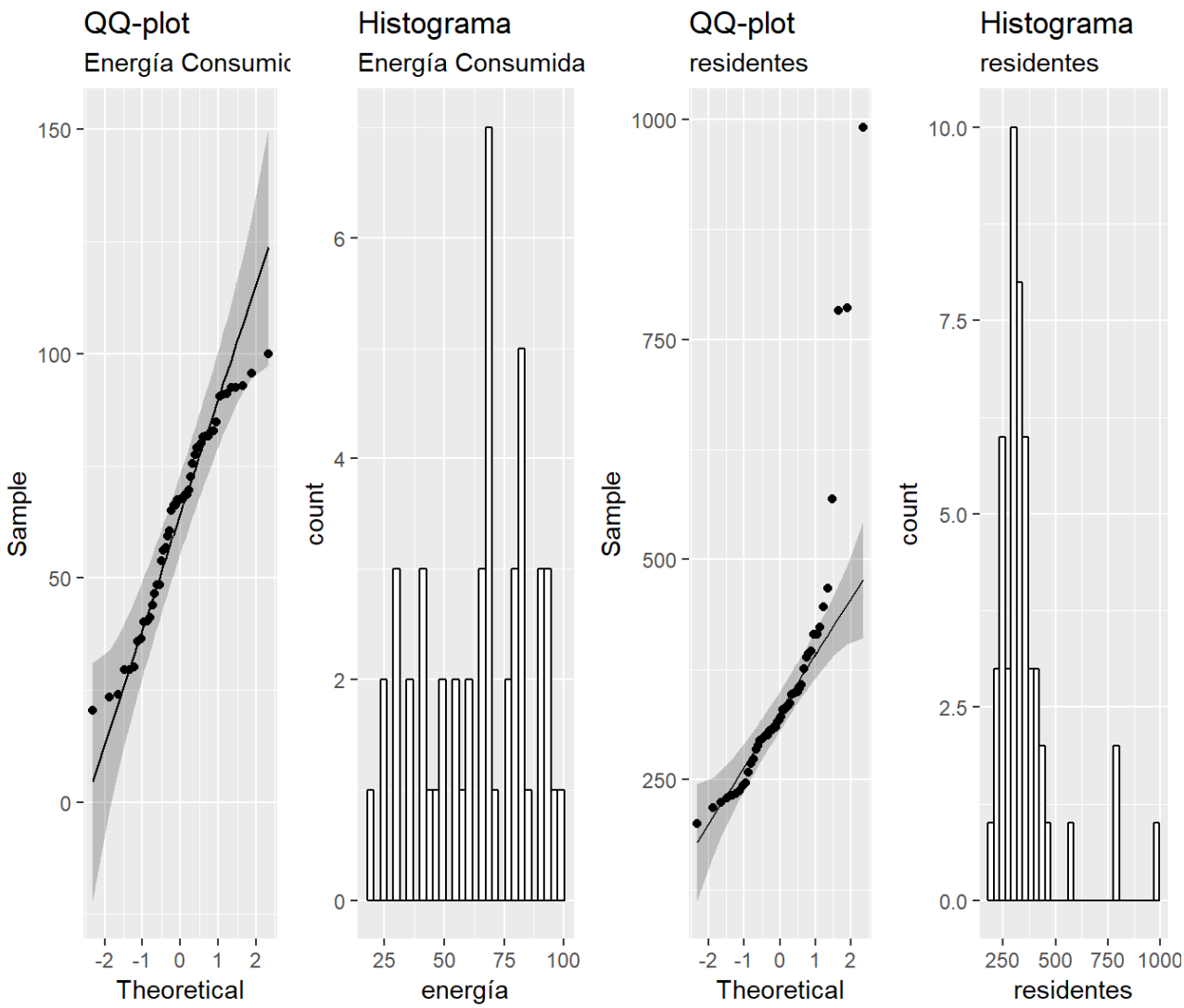
Code



Existe una relación inversa entre la variable energía percapita respecto al porcentaje de residentes que viven en áreas metropolitanas.

Se presentan gráficos exploratorios previo a realizar el modelo.

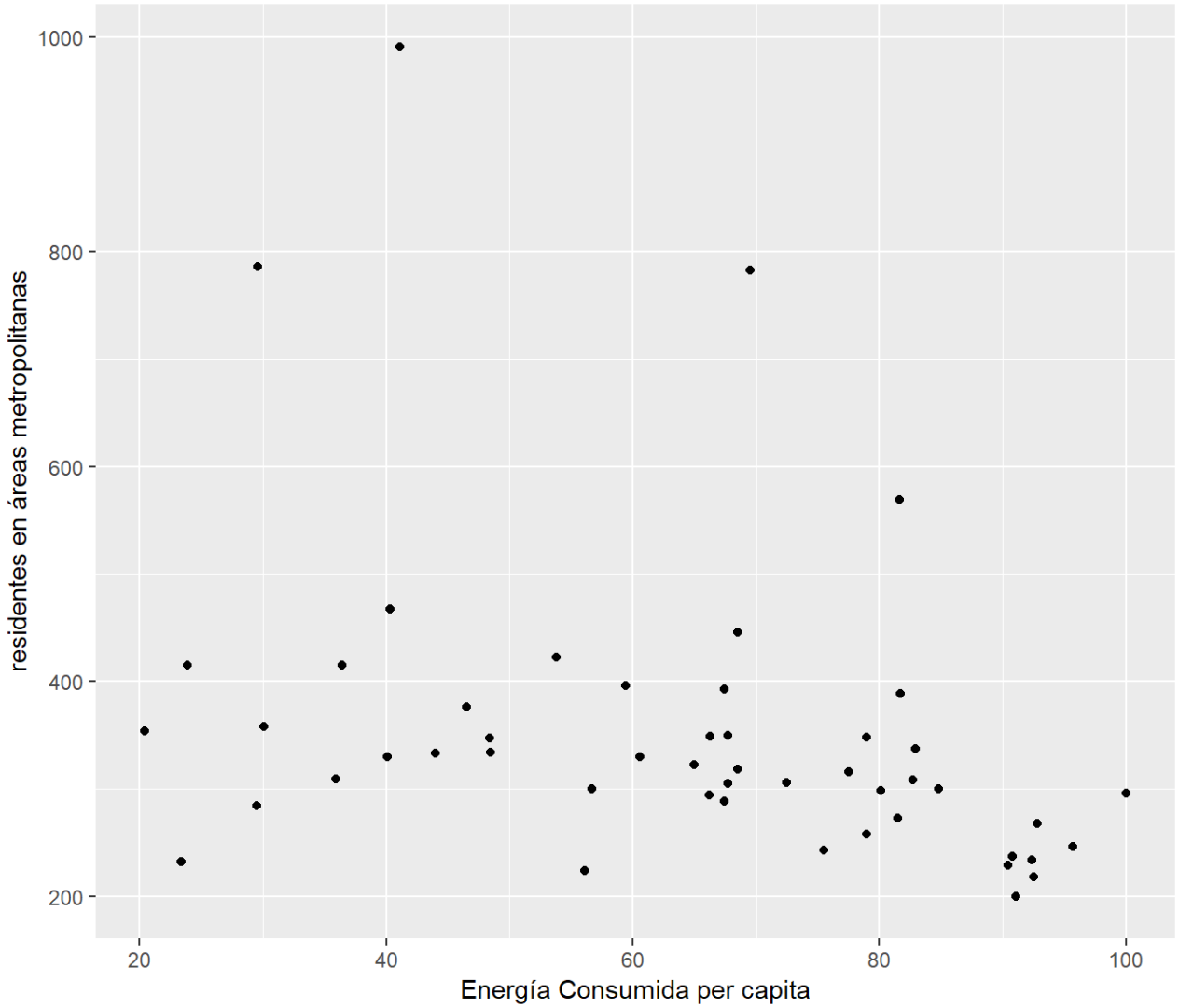
Code



Se tiene cuatro gráficos, los primeros dos (de izquierda a derecha), se observa el QQ-plot de la energía consumida percapita, se observa una cola pesada. La variable el número de residentes cuenta con colas ligeramente pesadas; no obstante, con su respectivo histograma se observa que datos atipicos afectan la forma de la distrbución de la variable. De esta manera, las variables no parecen distribuirce de forma normal, presentando datos en los extremos que afecten su comportamiento.

A continuación, se presenta un gráfico de dispersión de la energía percapita respecto al número de residentes, con el cual se puede observar algunos valores atipicos.

Code



ii. Imprime e interpreta el modelo

Realizamos el siguiente modelo

$$Energy = \beta_0 + \beta_1 metto$$

Code

```
##
## Call:
## lm(formula = states2$energy ~ states2$metro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -215.51  -64.54  -30.87   18.71  583.97
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   501.0292    61.8136   8.105 1.53e-10 ***
## states2$metro  -2.2871     0.9139  -2.503   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 140.2 on 48 degrees of freedom
## Multiple R-squared:  0.1154, Adjusted R-squared:  0.097
## F-statistic: 6.263 on 1 and 48 DF,  p-value: 0.01578
```

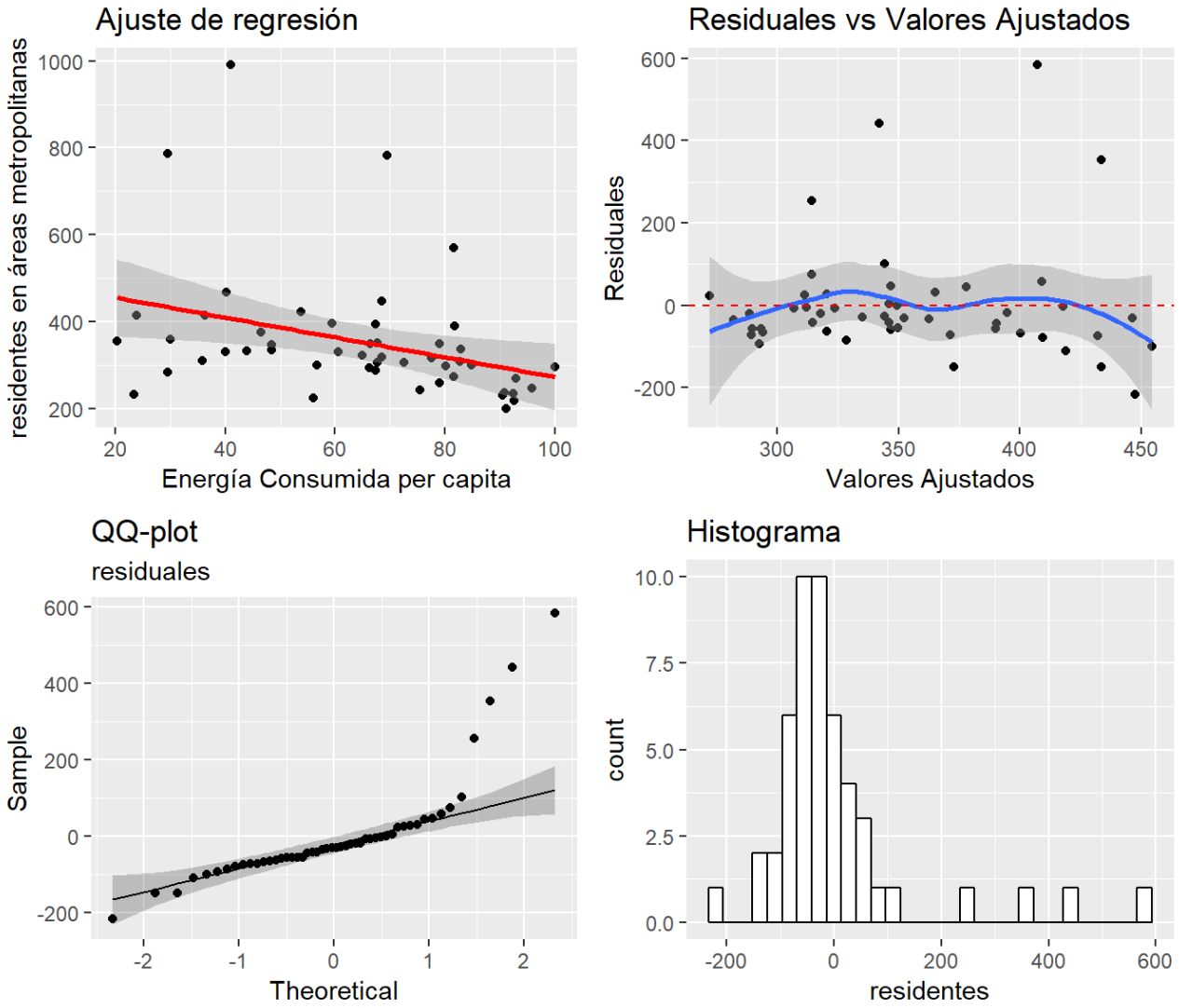
En el modelo se observa que tanto el intercepto y la variable de residentes en área metropolitana, son significativos al 99% y 100% de significancia. La variable *metro* presenta una relación inversa, lo cual índica que a un incremento de una unida en el número de residentes en el área metropolinada, reducirá -2.2871 la energía consumida percapita. El R^2 es del 0.1154, indicando que la covariable no explica en su totalidad la variabilidad de la energía percapita.

** iii. Gráfica el modelo para buscar desviaciones de los supuestos de modelado**

Code

Realizamos el anáisis a los residuales del modelo. El grpafico superior izquierdo muestra el ajuste de la regresión a las observaciones, se observan un par de observaciones que no son ajustados por la recta de regresión. El gráfico superior derecho se presentan los residuales respecto a los valores ajustados, en el cual claramente se observa como los residuales siguen la tendencia del ajuste de la regresión, indicando correlación etre los rsiduos y las estimaciones ($\epsilon' \hat{y} \neq 0$).Por otro lado, asumimos que los residuales on ruido blanco; sin embargo, en las gráficas inferiores se observan que los residuales no son normales.

Code



Se utiliza la prueba de Shapiro y Watstos de normalidad sobre los residuales. En base al críterio del p-valor, se tiene evidencia suficiente para rechazar la hipótesis de normalidad en los residuales.

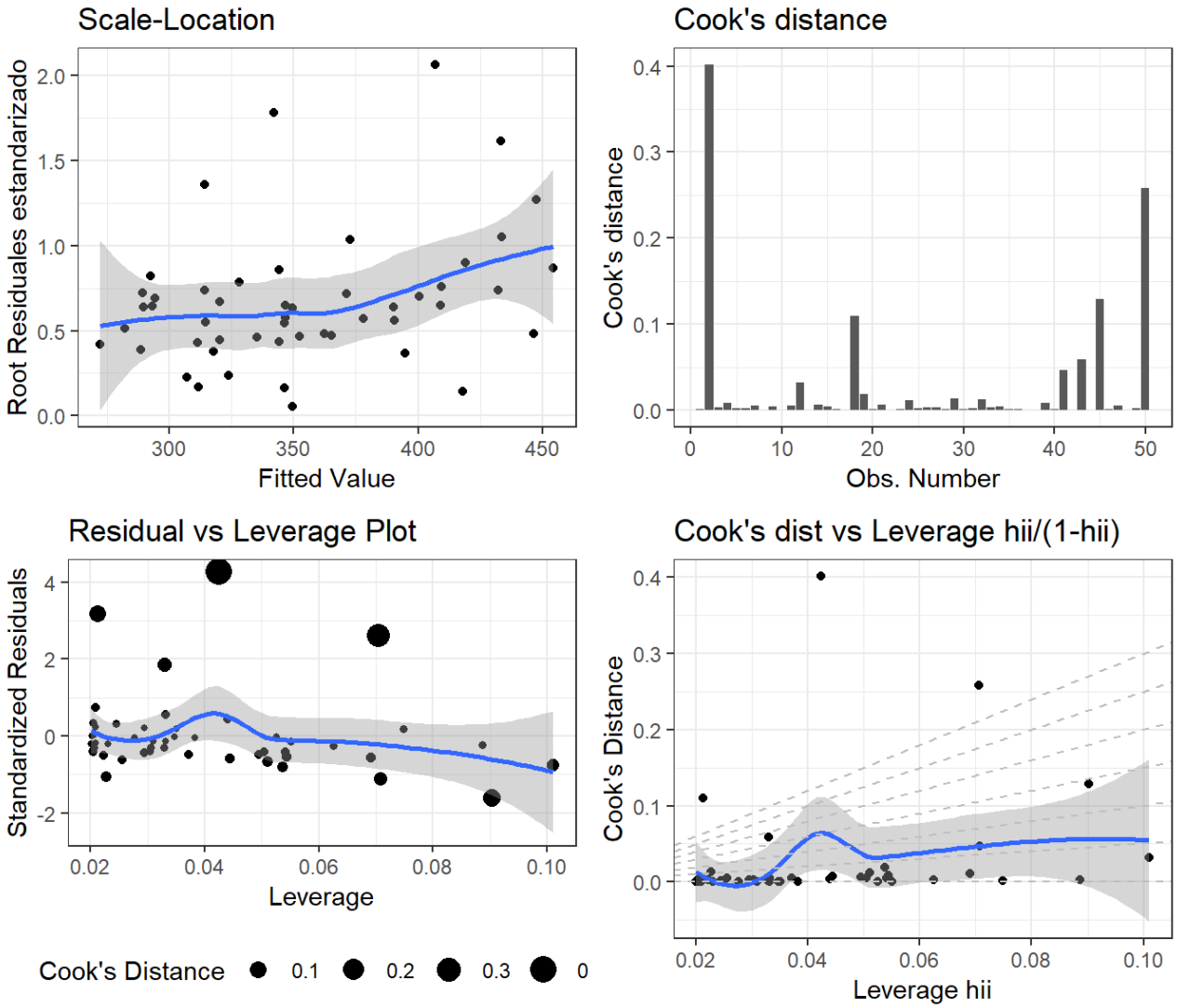
$$H_0 : normalidad$$

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  m1$residuals
## W = 0.71279, p-value = 1.414e-08
```

Ahora, observamos si los datos atipicos influyen en los resultados.

Code



Utilizando la desviación estandar de los residuales (scale-location) se observa la siperción de los residuales sobre los valores ajustados del modelo, se observa que la dispersión de los residuales no es la misma para todo el rango de los valores ajustados, mostrando una linea de ajuste cada vez menos horizontal, sugeriendo que el modelo presenta efectos de heterocedasticidad en los residuos. Por otra parte, la distancia de Cook's nos indica la influencia de los datos sobre los resultados del modelo, la figura superior derecha, muestra con barras, la existencia de 5 observaciones que pueden estar influyendo en los resultados del modelo; asimismo, con el grafico inferior derecho, la distancia de Cook's y el laverage hill, indican observaciones atipicas en las covariables.

Se procede a retirar los valores atipicos para actualizar el modelo. se retiran las observaciones $18, 50, 2, 43$ del modelo.

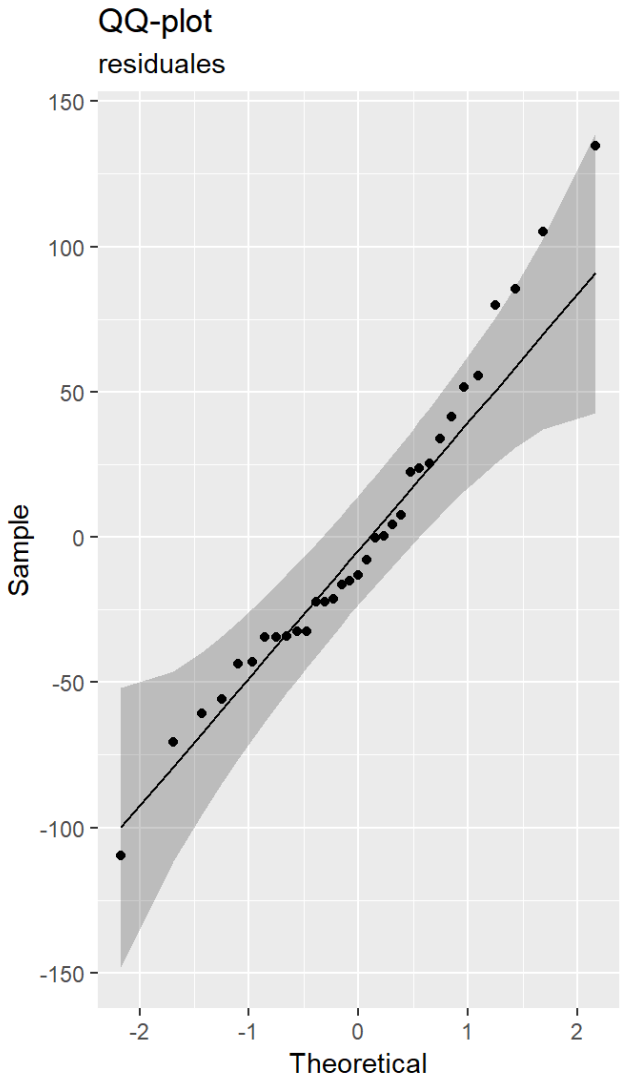
Code

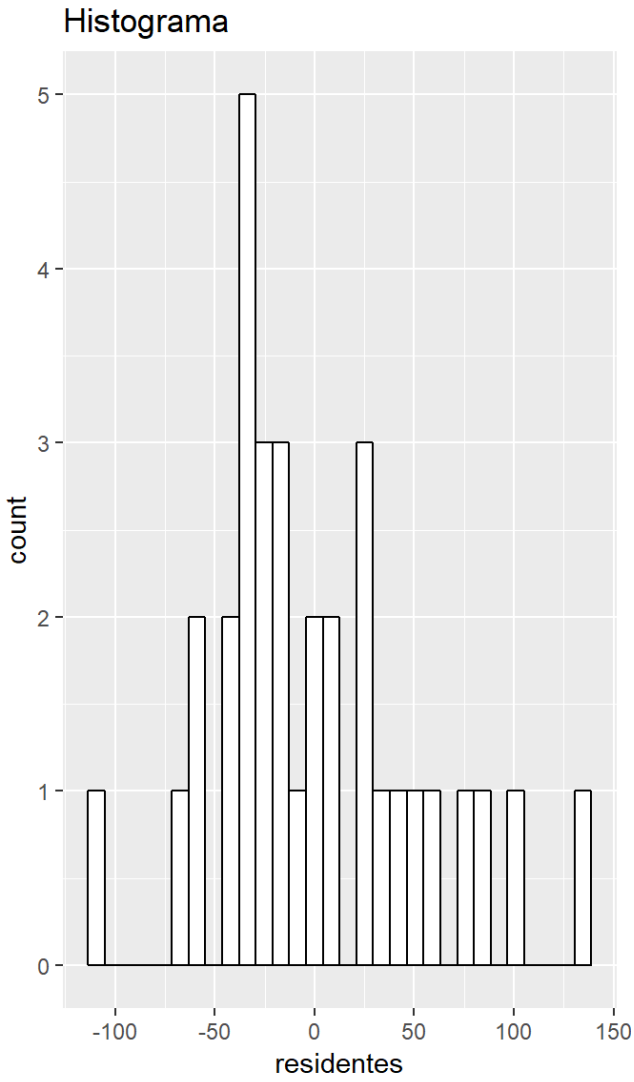
```
##
## Call:
## lm(formula = states2$energy ~ states2$metro, subset = (1:35)[-c(18,
##     50, 2, 43)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -109.58  -34.24  -13.23   25.18  134.70
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   434.3571    29.2423  14.854 1.21e-15 ***
## states2$metro  -1.7964     0.4236  -4.241 0.000186 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.45 on 31 degrees of freedom
## Multiple R-squared:  0.3672, Adjusted R-squared:  0.3467
## F-statistic: 17.99 on 1 and 31 DF,  p-value: 0.0001862
```

Los coeficientes del intercepto y los residentes en el área metropolitana son significativos, casi al 100%. El R^2 mejora ligeramente, ajustando mejor el modelo sin observaciones atipicasrespecto al modelo anterior. De esta forma, un cambio de una unidad en el número de residentes en el área metropolitana, reducirá en -1.7964 la energía percapita. El $F-statistic$ indica que el modelo en general explica bien.

Evaluamos normalidad en los residuos del nuevo modelo. Se observa que los residuales al retirar los datos atipicos se parecen más a observaciones provenientes de distribuciones normales.

Code





Code

```
##
##  Shapiro-Wilk normality test
##
## data:  m1$residuals
## W = 0.71279, p-value = 1.414e-08
```

Code

```
##
##  Jarque-Bera test for normality
##
## data:  m1$residuals
## JB = 153.57, p-value = 5e-04
```

Con la prueba de Shapiro y Watson, no se tiene evidencia suficiente para rechazar la hipótesis nula, por lo tanto los residuales son normales.

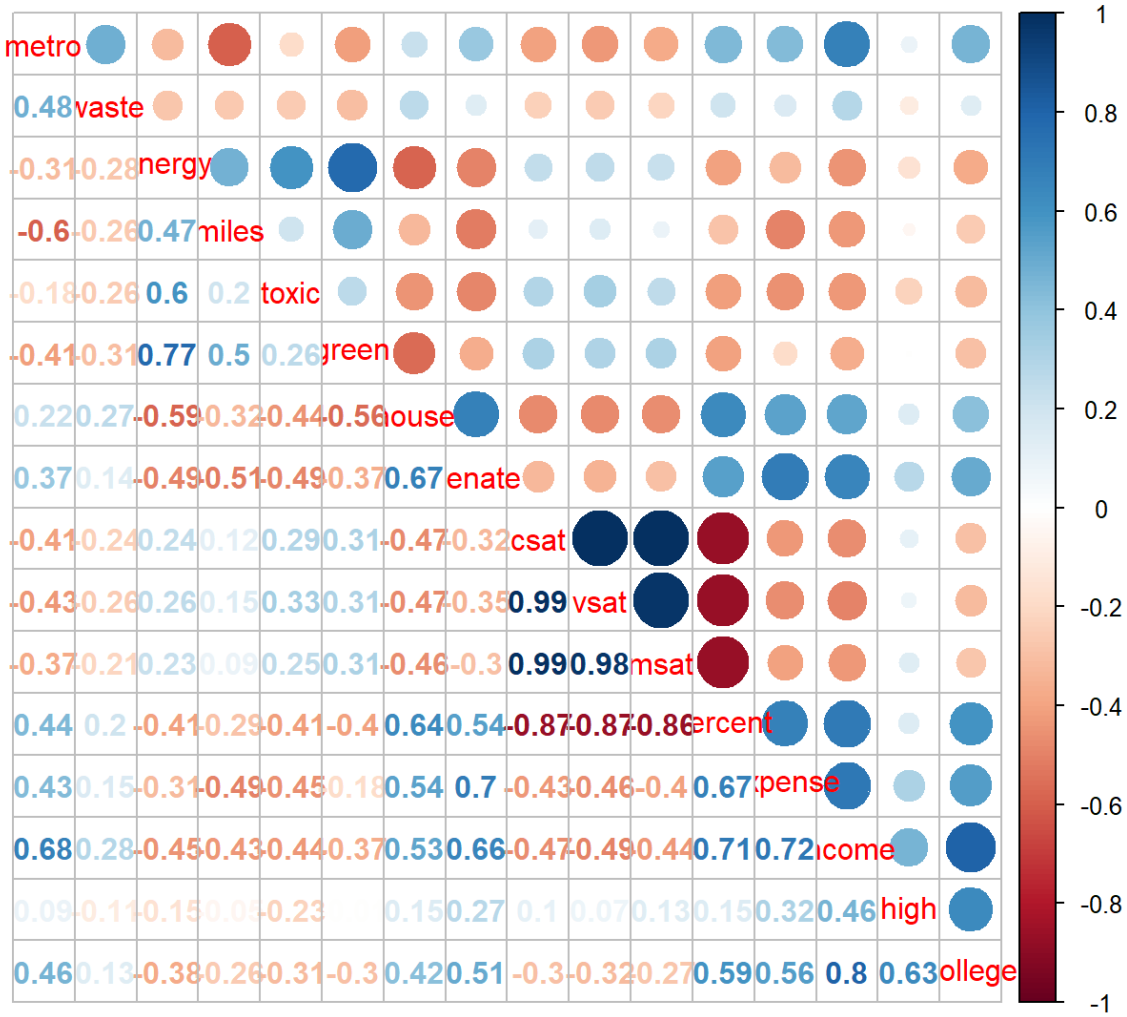
Code

```
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.96415, p-value = 0.3372
```

(c) Selecciona uno o más predictores adicionales para agregar al modelo y repita los pasos anteriores.¿Este modelo significativamente mejor que el modelo con la variable / metro / solo comoúnico predictor?

Como en el caso anterior retiramos los valores NA de la base. Realizamos un gráfico de correlación para seleccionar variables con mayor relación lineal respecto a la energía percapita.

Code



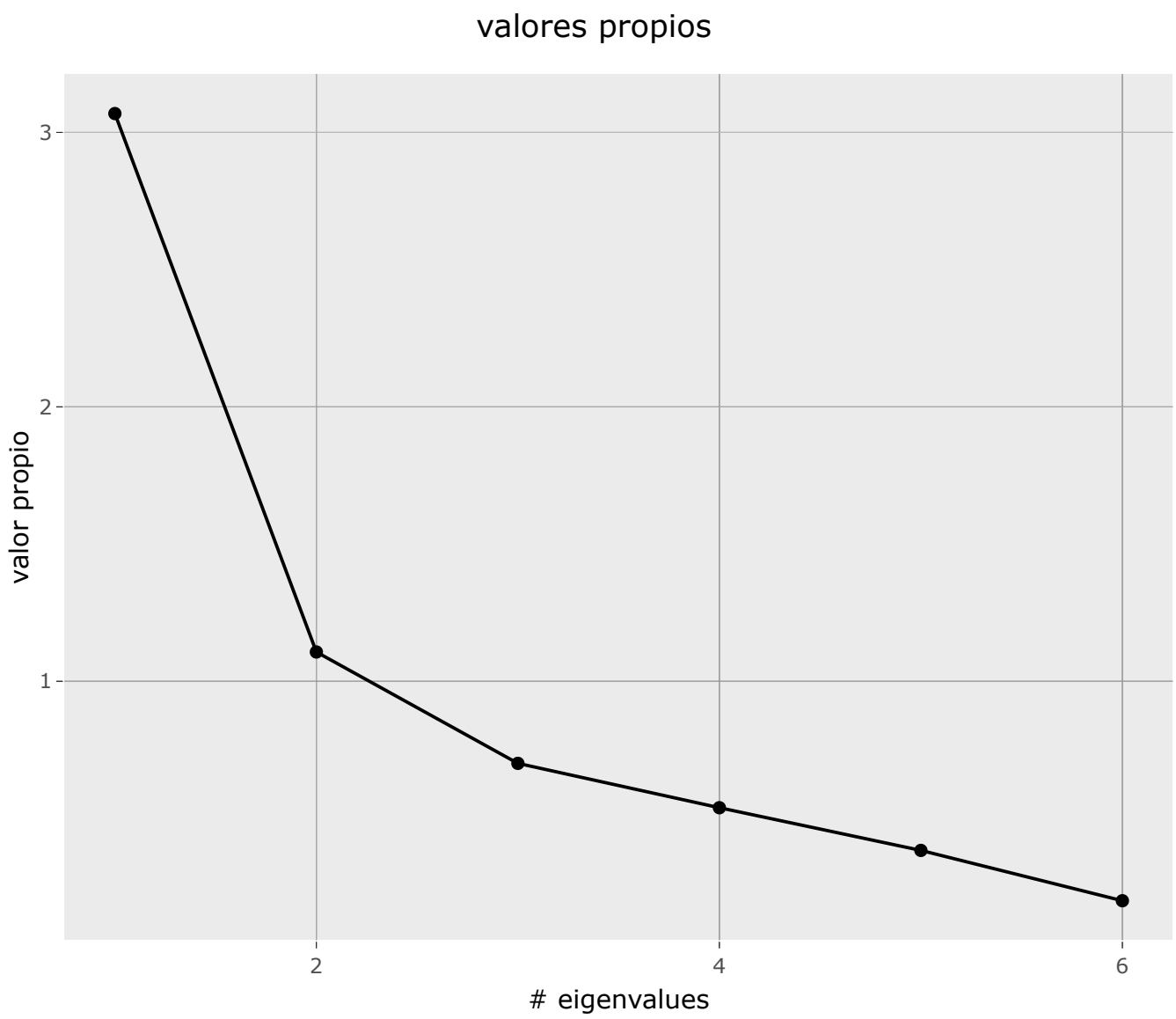
Code

La energia presenta correlación respecto a la variables toxic, hause, metro, miles, green y sense; algunas relación positiva, y otra a manera negativa.

Code

Por medio de los valores propios de la matriz de correlación de las observaciones, se bucan indicios de multicolinealidad en los datos.

Code



Code

$\lambda_{max}/\lambda_{min} = 15.305834989065$, indicando algo de multicolinealidad; i.e, covariables se afectan entre si.

Realizamos el modelo

$$energy = \beta_0 + \beta_1 metro + \beta_2 miles + \beta 3 toxic + \beta 4 green + \beta 5 house + \beta 5 senate + \epsilon$$

Code

```
##
## Call:
## lm(formula = states3$energy ~ ., data = states3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -174.417  -23.517   -9.184   24.061  177.718
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.0094   121.8444   0.369    0.714
## metro         0.5874     0.5014   1.171    0.248
## miles        12.6413    10.5538   1.198    0.238
## toxic         2.5384     0.5363   4.733 2.63e-05 ***
## green         4.4411     0.7263   6.115 2.97e-07 ***
## house        -0.1096     0.7178  -0.153    0.879
## senate       -0.1122     0.5241  -0.214    0.832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.27 on 41 degrees of freedom
## Multiple R-squared:  0.776,  Adjusted R-squared:  0.7432
## F-statistic: 23.67 on 6 and 41 DF,  p-value: 7.182e-12
```

Solamente el coeficeinte de la variable toxic y green son significativas casi al 100, la R^2 ajustada y sin ajustar, indican buen ajuste en el modelo; no obstate, muchas variables tienen significacia estadísittca.

Planteamos el siguiente modelo:

$$energy = \beta_0 + \beta_1 metro + \beta 2house + \epsilon$$

Code

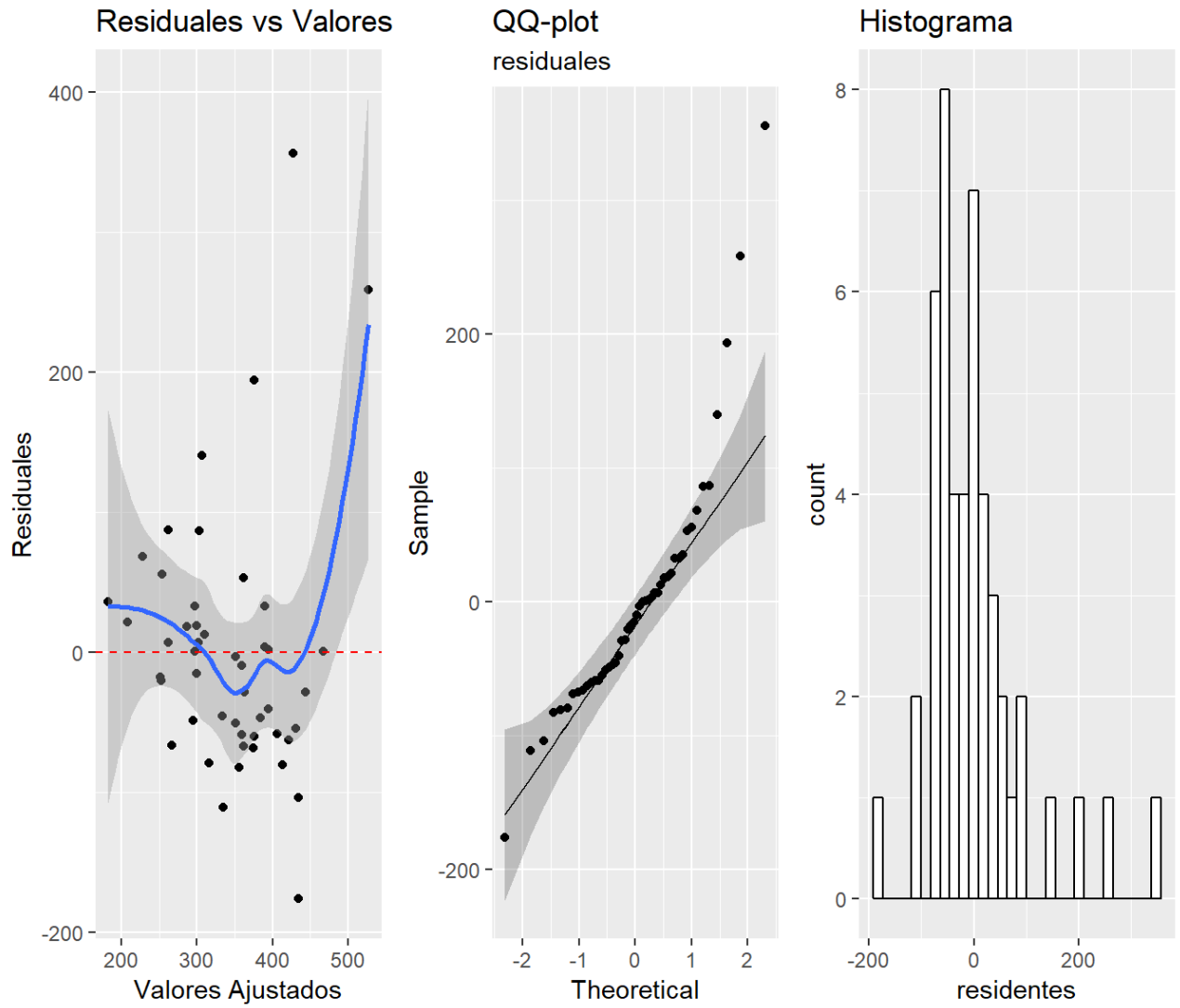
```
##
## Call:
## lm(formula = states3$energy ~ +states3$metro + states3$house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -176.04  -58.73  -12.54   24.07  355.81
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   557.3483    48.5702  11.475 5.88e-15 ***
## states3$metro  -1.0172     0.6389  -1.592    0.118
## states3$house  -3.3032     0.7321  -4.512 4.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.24 on 45 degrees of freedom
## Multiple R-squared:  0.3783, Adjusted R-squared:  0.3507
## F-statistic: 13.69 on 2 and 45 DF,  p-value: 2.264e-05
```

Se tiene que solo el coeficiente de *house* es significativa a casi el 100, y el coeficiente de metro se aproxima a ser sifnificativa al 90; a su vez, el valor del R^2 ajustado, y sin ajustar, disminuye.

*iii. Gráfica el modelo para buscar desviaciones de los supuestos de modelado**

Ahora, le realizamos pruebas a los residuales del modelo anterior para detectar violaciones en el supuesto

Code



Se observa, correlación en los residuales y a su vez, la dispersión de ellos en la primera grafica nos indican efectos de heterocedasticidad en el modelo; por otro lado, vemos que los residuales parecen normales en el centro de la distribución , pero no en las colas.

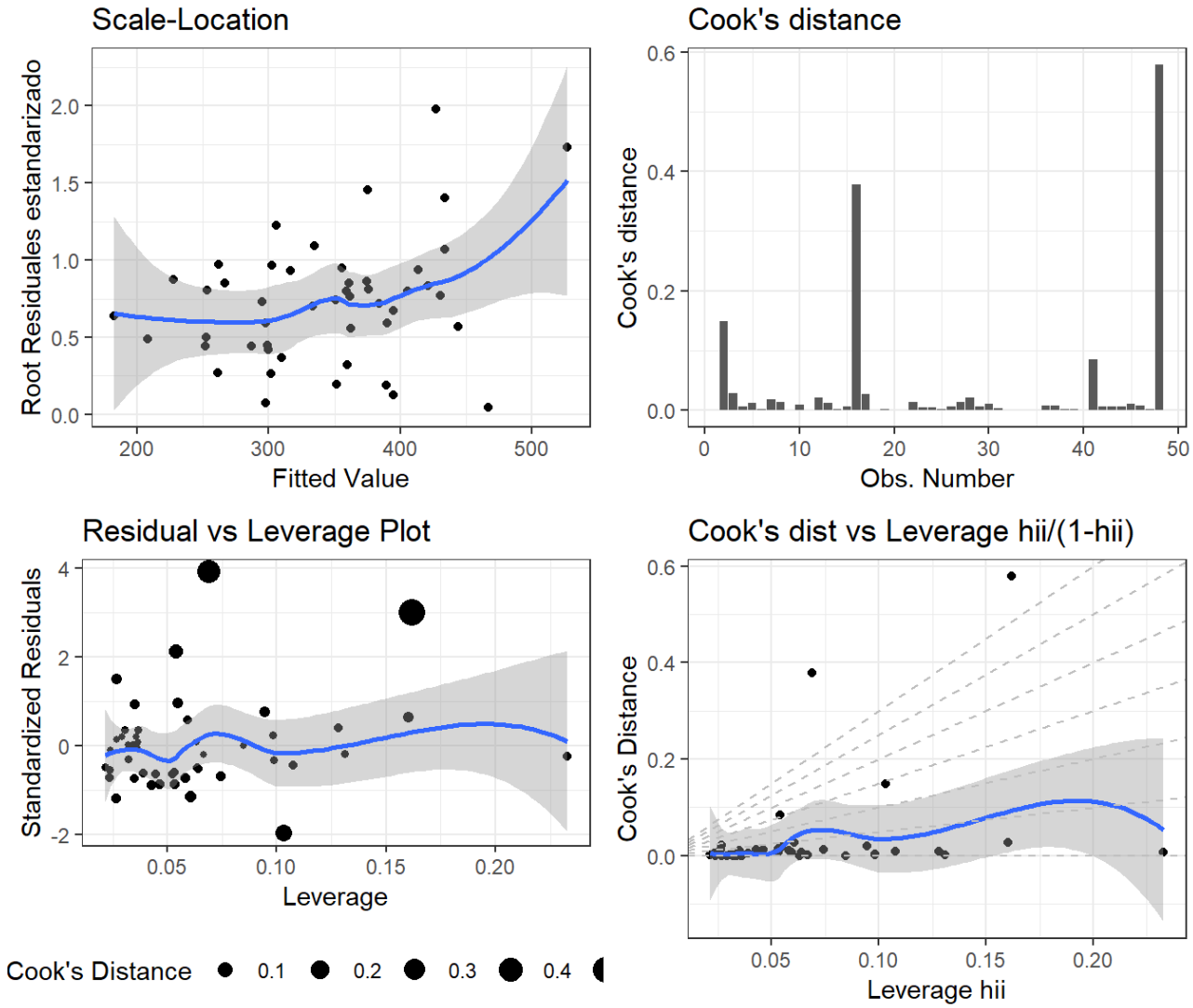
Aplicamos el test de normalidad sobre los residuos del modelo, podemos decir, que se tiene evidencia para rechazar la hipotesis nula de normalidad en los residuales.

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  m3$residuals
## W = 0.85591, p-value = 3.132e-05
```

Ahora, realizamos gráficos para detectar valores atipicos que afecten los resultados en el modelo.

Code



Es claro que utilizando los estadísticos de Cook´s y los valores leverege, se detecta la existencia de un grupo de observaciones atipicas en las covariables que ocasiones una mala especificación en el modelo. Se retiran las observaciones $16, 48, 41, 12, 32$

Se procede a retirar los valores atipicos. Planteamo de nuevo el modelo:

$$energy = \beta_0 + \beta_1 metro + \beta 2house + \epsilon$$

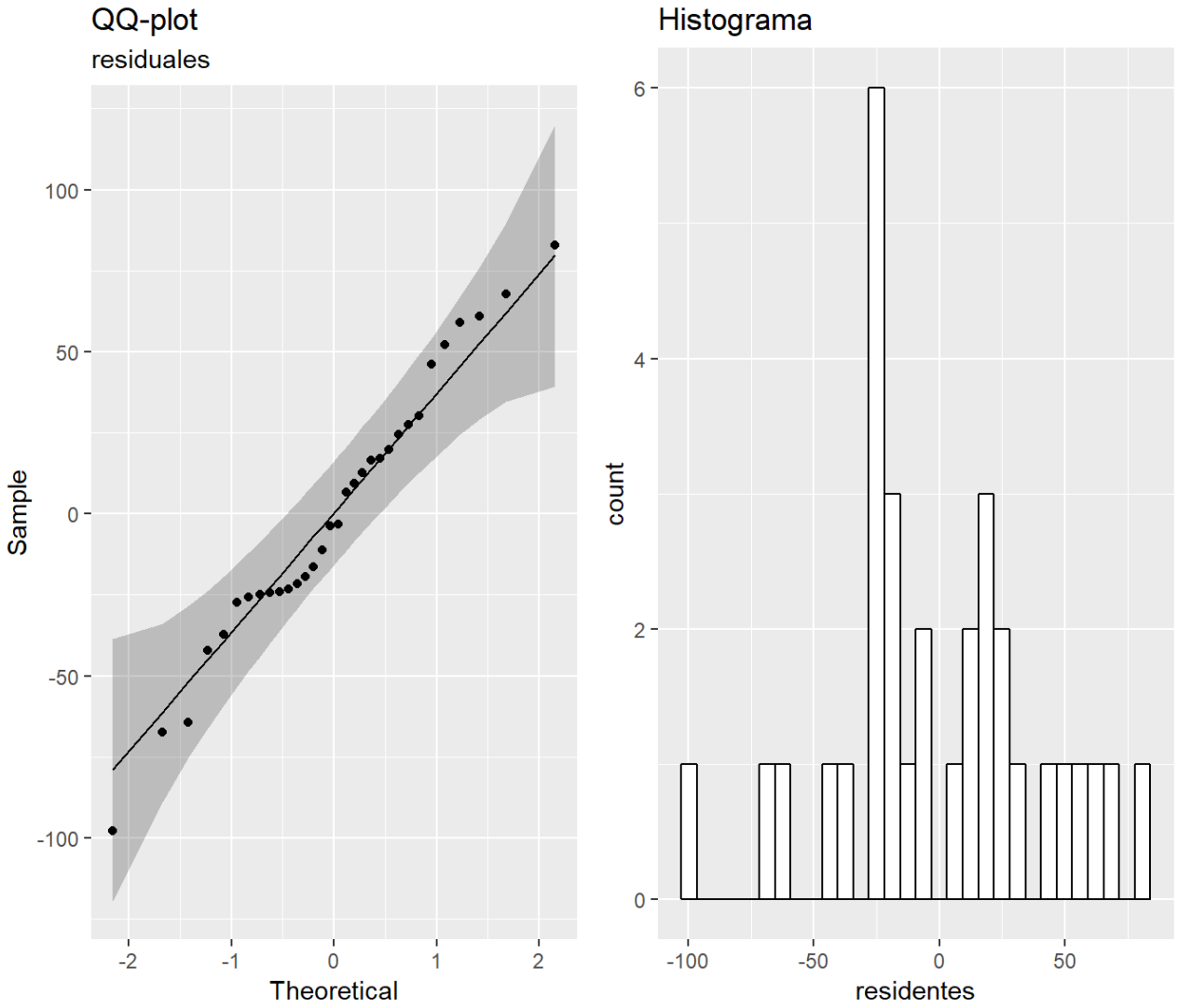
Code

```
##
## Call:
## lm(formula = states3$energy ~ +states3$metro + states3$house,
##     subset = (1:35)[-c(16, 48, 41, 12, 32)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.790 -24.469  -3.523  25.249  83.006
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   445.6271    27.4541  16.232 4.31e-16 ***
## states3$metro  -1.3688     0.3752  -3.648  0.00103 **
## states3$house  -0.9409     0.4915  -1.914  0.06548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.94 on 29 degrees of freedom
## Multiple R-squared:  0.4706, Adjusted R-squared:  0.4341
## F-statistic: 12.89 on 2 and 29 DF,  p-value: 9.872e-05
```

Retirando valores atipicos, se observa que los coeficientes del intercepto, metro y house son significativas, al 100%, 99% y 95%. Con un R^2 de 0.4706, lo cual es un ajuste ligeramente bueno. El valor F indica que el modelo en global es adecuado. De esta manera, un incremento de una unidad en la variable *house*, genera un cambio de -0.9409 en la energía consumida percapita; a su vez, un incremento de una unidad en el número de residentes en el área metropolitana, hace que la energia consumida cambie en -0.9409.

Realizamos el analisis de normalidad en los residuales del modelo.

Code



Se observa que los residuales se parecen ligeramente a una normal; sin embargo, utilizamos la prueba de Shapiro y Watson para argumentar al respecto.

Bajo el crpiterio del p-valor, no se tiene evidencia suficiente para rechazar la hipótesis nula; por lo tanto los residuales del modelo se distribuyen de forma normal.

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  m4$residuals
## W = 0.9803, p-value = 0.8082
```

Construimos intervalos de confianza a los coeficientes de la regresión

Code

	parametros	ICInferior	ICUpper
Intercepto	445.6270645	389.477	501.777

	parametros	ICInferior	ICUpper
metro	-1.3688021	-2.136	-0.601
house	-0.9409379	-1.946	0.064

Para constastar los resultado realizamos un análisis de varianza con la prueba anova. Se observa que los factores que se eligen son significativos para la variable respuesta.

Code

```
## Analysis of Variance Table
##
## Response: states3$energy
##               Df Sum Sq Mean Sq F value    Pr(>F)
## states3$metro  1  62448   62448  7.0318   0.01102 *
## states3$house  1 180772  180772 20.3552 4.583e-05 ***
## Residuals     45 399639    8881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

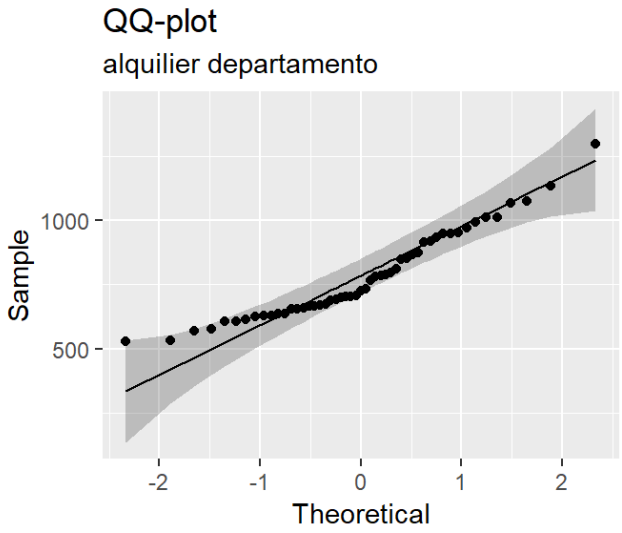
EJERCICIO 2

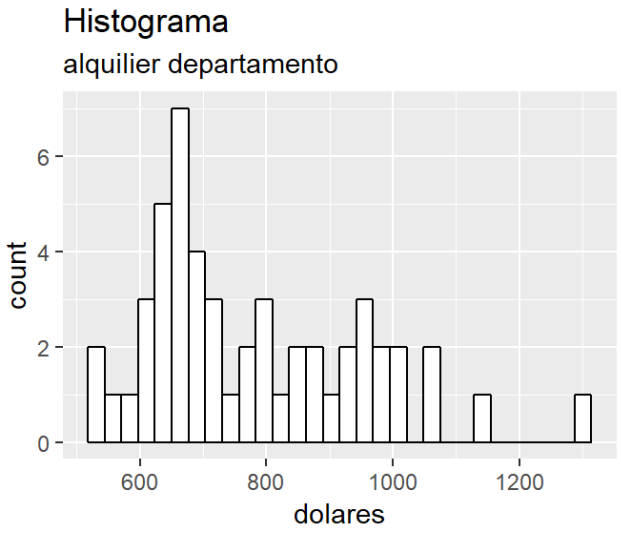
*2. Los datos del archivo costoliving.txt enumeran algunas estadÃ□???sticas del costo de vida para cada uno de los 50 estados de los USA. Los tres costos son: alquileres de apartamentos (rent) costo de casas (house) el índice de costo de vida. (cost of live). resto de variables (income) and (pop)

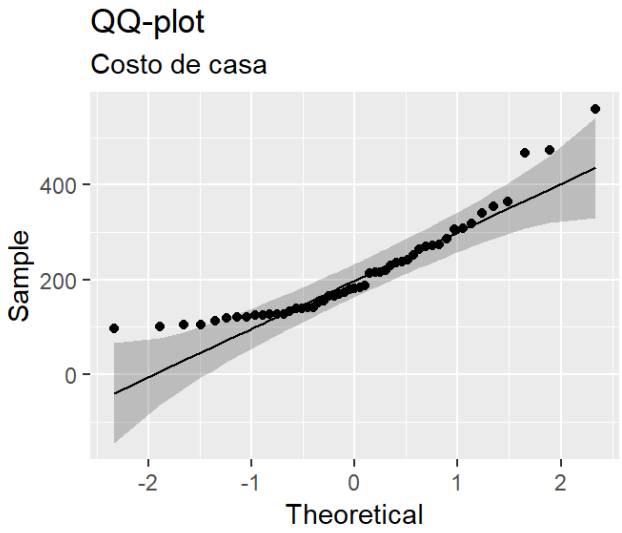
Code

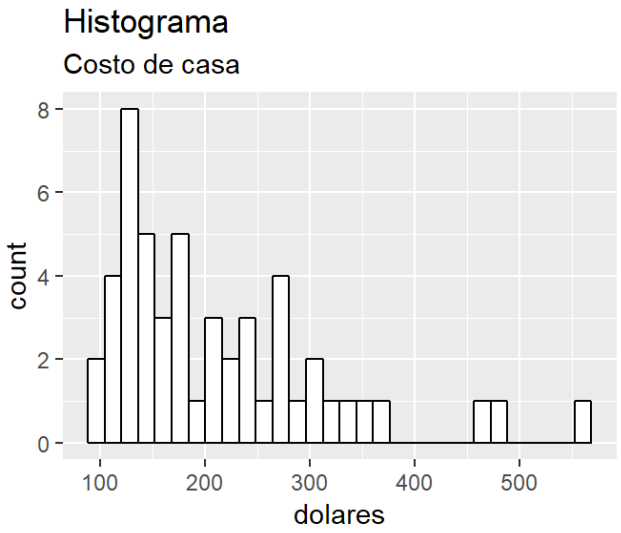
Realizamos un analisis gráfico de las variables.

Code



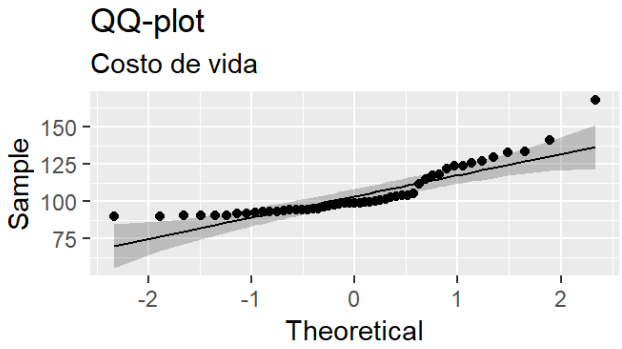


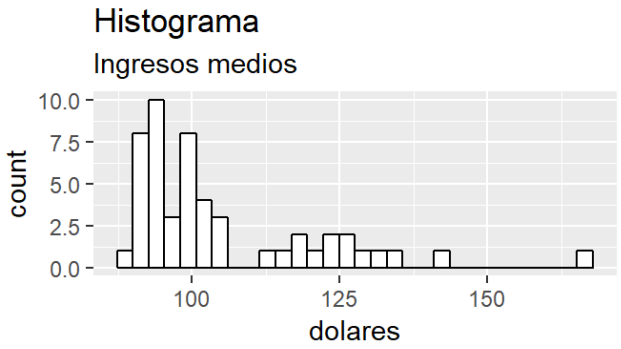


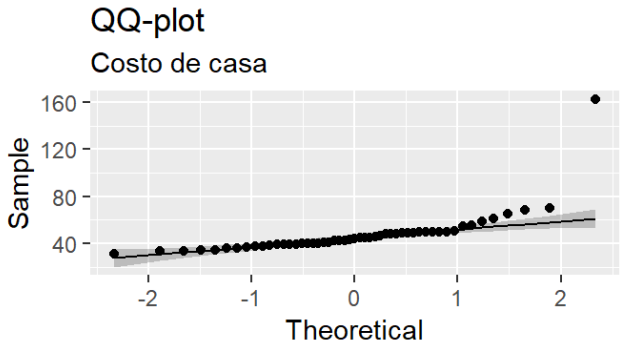


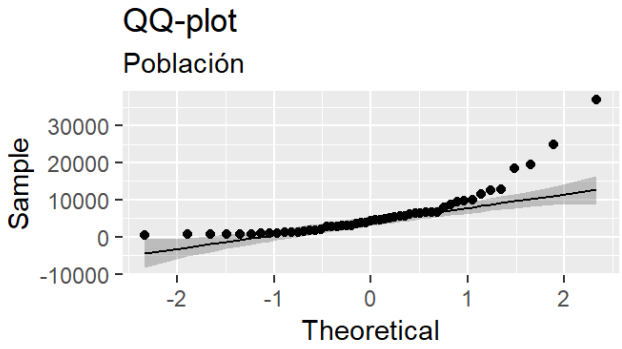
Podemos obsservar que las variables del costo de casa y alquiler de departamento presentan colas pesadas, las cueles pueden ser generadas por datos atipicos.

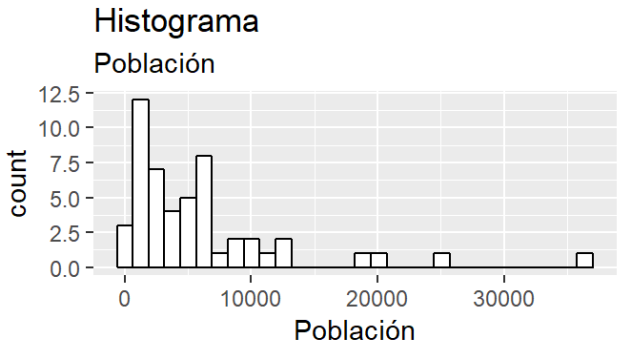
Code







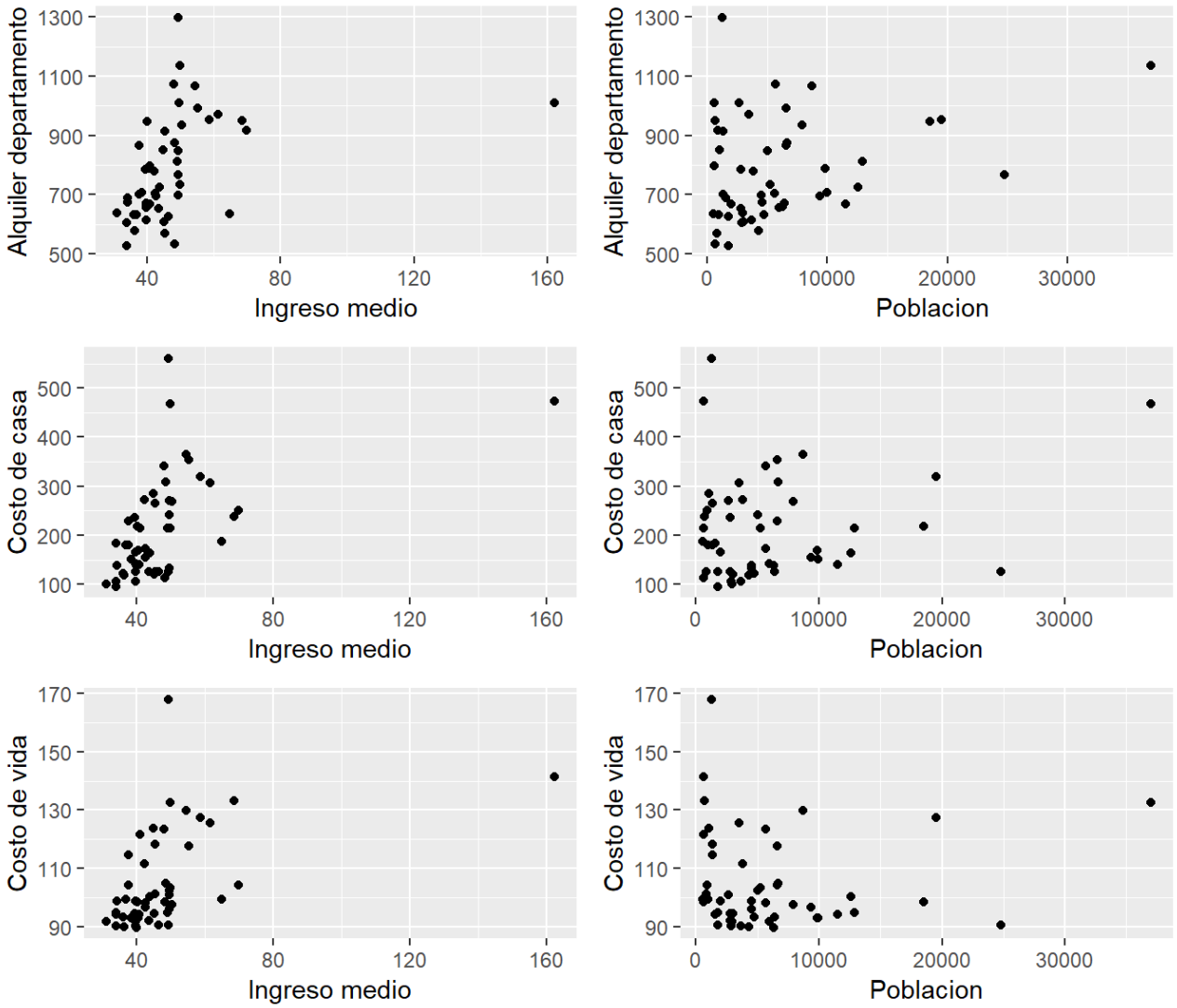




El resto de las variables presentan un comportamiento similar a las anteriores, colas pesadas, y algunas observaciones atipicas.

A continuación presentamos gráficos de disperción de las variables respuestas respectos las covariables a utilizar en el modelo.

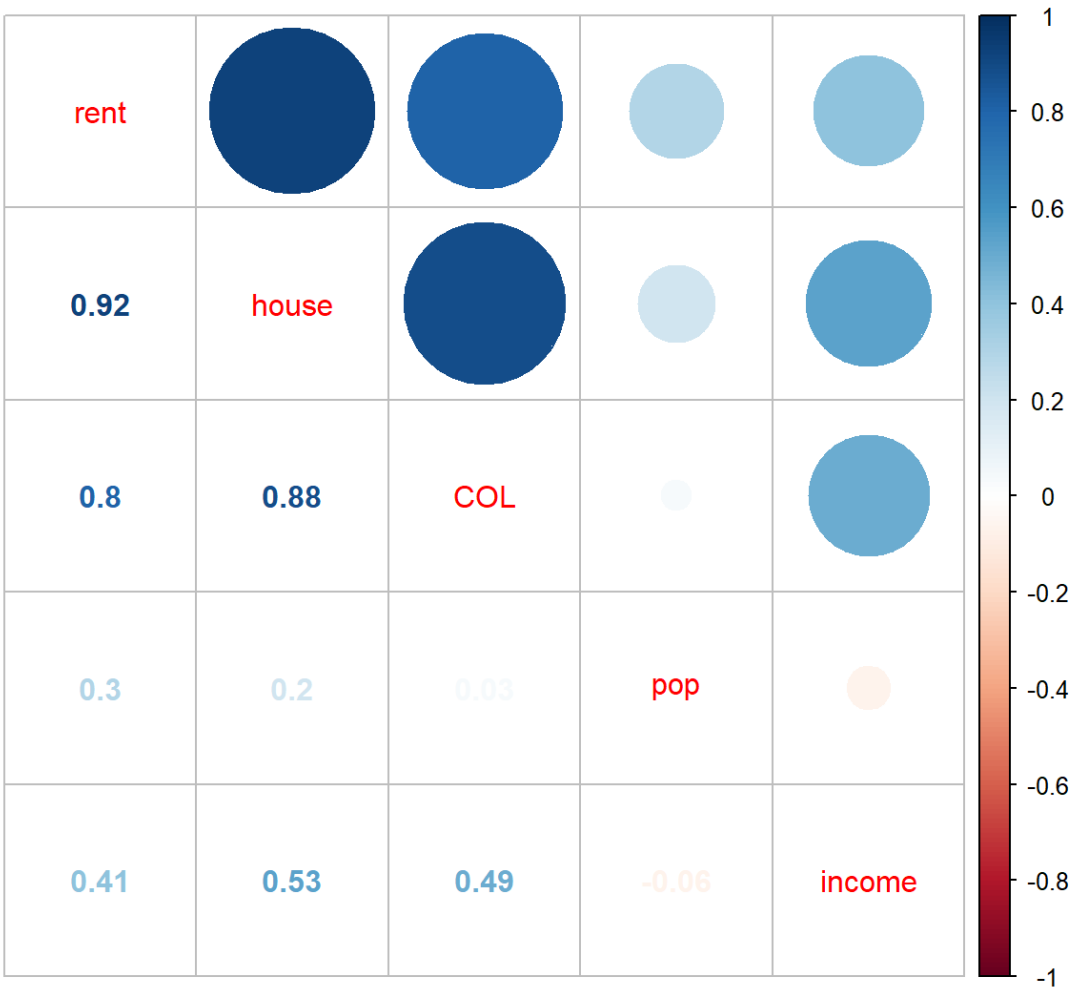
Code



A primera vista se presentan observaciones en las covariables atipicas.

Se realiza un gráfico de correlación para ver el grado de relación lineal entre las variables respuestas y covariables.

Code



Se observa una relación negativa entre las variables, con excepción a la población de cada estado.

(a) Realiza una regresíon lineal multivariada para explicar estas tres métricas en terminos de las poblaciones estatales e ingresos medios. Ã□Â¾Son Ã□Âºtiles estas variables independientes para explicar conjuntamente las variables de costo?

Code

```
## Response rent :
##
## Call:
## lm(formula = rent ~ costolive$income + costolive$pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -206.10  -95.66  -49.76   89.30  548.41
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.440e+02  6.211e+01   8.759 1.61e-11 ***
## costolive$income 3.954e+00  1.142e+00   3.462  0.00114 **
## costolive$pop    8.236e-03  3.122e-03   2.638  0.01121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149.4 on 48 degrees of freedom
## Multiple R-squared:  0.2711, Adjusted R-squared:  0.2407
## F-statistic: 8.924 on 2 and 48 DF,  p-value: 0.0005065
##
##
## Response house :
##
## Call:
## lm(formula = house ~ costolive$income + costolive$pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158.30  -58.89  -22.76   43.04  359.10
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      46.721108  35.721748   1.308   0.1971
## costolive$income  3.037709   0.656892   4.624 2.86e-05 ***
## costolive$pop     0.003567   0.001796   1.987   0.0527 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.92 on 48 degrees of freedom
## Multiple R-squared:  0.3357, Adjusted R-squared:  0.308
## F-statistic: 12.13 on 2 and 48 DF,  p-value: 5.453e-05
##
##
## Response COL :
##
```

```
## Call:
## lm(formula = COL ~ costolive$income + costolive$pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.964  -9.202  -4.365   6.258  62.750
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.340e+01  6.022e+00  13.850  < 2e-16 ***
## costolive$income 4.351e-01  1.107e-01   3.929 0.000273 ***
## costolive$pop    1.526e-04  3.027e-04   0.504 0.616599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.48 on 48 degrees of freedom
## Multiple R-squared:  0.2441, Adjusted R-squared:  0.2126
## F-statistic: 7.751 on 2 and 48 DF,  p-value: 0.00121
```

Tenemos el modelo multivariado en la parte de arriba, para las tres respuestas. La respuestade la variable costo de renta presenta a sus tres coeficientes significativos a diferentes niveles de significancia; de esta manera incrementos de una unidad en el ingreso promedio aumentan en los costos de la renta, asimismo, incrementos de una unidad en el número de habitantes generan incrementos en el costo de renta.

El modelo de la respuesta costo de habitación cuenta con los coeficientes de las covariables significativos a distintos niveles, pero no el coeficiente de la constante.

En la respuesta del coosto de vida, el coeficiente del ingreso medio es significativo, ya que incrementos de ua unidad del ingreso medio genera cambios del del $4.351e-01$, los cuales son pequeños pero significativos.

Los ajustes del modelo por medio del R^2 no son los mejores, ya que reporta niveles menores al $.5$.

Ahora realizamos la prueba de normalidad a los residuales de las respuestas.

Code

```
##           [,1]                           [,2]
## statistic 0.907725                       0.8615029
## p.value   0.0007694821                   2.738788e-05
## method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
## data.name "modelo$residuals[, i]"       "modelo$residuals[, i]"
##           [,3]
## statistic 0.7862688
## p.value   3.51343e-07
## method    "Shapiro-Wilk normality test"
## data.name "modelo$residuals[, i]"
```

En conclusión se tiene evidencia suficiente para rechazar la hipótesis nula de normalidad en los residuales bajo el críterio del p-valor.

Vemos que los residuales no estén correlacionados con los ajustes de las respuestas y que la suma de residuales sea cero.

Code

	rent	house	COL
rent	0	0	0
house	0	0	0
COL	0	0	0

Suma de residuales igual a cero

Code

	.
rent	0
house	0
COL	0

Ahora, por medio del análisis de varianza utilizamos la prueba MANOVA para rectificar el resultado de la regresión

Code

```
##  Response rent :
##                                                 Df  Sum Sq Mean Sq
## as.matrix(cbind(costolive$income, costolive$pop))  2  398346  199173
## Residuals                                         48 1071246   22318
##                                                 F value    Pr(>F)
## as.matrix(cbind(costolive$income, costolive$pop))  8.9245 0.0005065 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response house :
##                                                 Df Sum Sq Mean Sq
## as.matrix(cbind(costolive$income, costolive$pop))  2 179076   89538
## Residuals                                         48 354358    7382
##                                                 F value    Pr(>F)
## as.matrix(cbind(costolive$income, costolive$pop))  12.129 5.453e-05 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response COL :
##                                                 Df  Sum Sq Mean Sq
## as.matrix(cbind(costolive$income, costolive$pop))  2  3252.3  1626.2
## Residuals                                         48 10070.3   209.8
##                                                 F value  Pr(>F)
## as.matrix(cbind(costolive$income, costolive$pop))  7.7511 0.00121 **
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

Nos indica que vajo el críterio del p-valor los factores son significativos para cada respuesta.

Se retiran valores atipicos para mejorar el modelo. Quitamos aquellos valores atipicos que se presenten para cada respuesta los cuales son: $5, 12, 20, 32, 8$ Ajustamos de nuevo el modelo y tenemos lo siguiente:

Code

```
## Response rent :
##
## Call:
## lm(formula = rent ~ costolive$income + costolive$pop, subset = (1:35)[-
c(5,
##     12, 20, 32, 8)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -211.27  -68.90  -29.46   48.70  282.79
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.976e+02  1.038e+02   2.868 0.007922 **
## costolive$income 9.149e+00  2.203e+00   4.153 0.000295 ***
## costolive$pop    9.617e-03  4.372e-03   2.200 0.036578 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114.3 on 27 degrees of freedom
## Multiple R-squared:  0.449,  Adjusted R-squared:  0.4082
## F-statistic:    11 on 2 and 27 DF,  p-value: 0.00032
##
##
## Response house :
##
## Call:
## lm(formula = house ~ costolive$income + costolive$pop, subset = (1:35)
[-c(5,
##     12, 20, 32, 8)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72.69 -32.96 -19.18  37.18 139.14
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -22.878115  48.691592  -0.470 0.642230
## costolive$income   4.264146   1.033518   4.126 0.000317 ***
## costolive$pop      0.003585   0.002051   1.747 0.091927 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.63 on 27 degrees of freedom
## Multiple R-squared:  0.4257, Adjusted R-squared:  0.3831
## F-statistic: 10.01 on 2 and 27 DF,  p-value: 0.0005607
```

```
##
##
## Response COL :
##
## Call:
## lm(formula = COL ~ costolive$income + costolive$pop, subset = (1:35)[-c
(5,
##     12, 20, 32, 8)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.564  -5.454  -2.664   4.123  19.357
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.415e+01  8.754e+00   7.328 6.98e-08 ***
## costolive$income 8.235e-01  1.858e-01   4.432  0.00014 ***
## costolive$pop    1.049e-04  3.688e-04   0.284  0.77829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.641 on 27 degrees of freedom
## Multiple R-squared:  0.422,  Adjusted R-squared:  0.3792
## F-statistic: 9.855 on 2 and 27 DF,  p-value: 0.0006114
```

Primero, se observa que el R^2 ajustado, y sin ajustar, mejora, indicando un mejor ajuste para las tres respuestas. Segundo, para la primera respuesta, que es el costo de la rentas, los coeficientes del intercepto, ingreso medio, y población, son significativas a diferentes nveles, ambos afectan de forma positiva al costro de renta.

En el caso de la respuesta, costo de casa, el intercepto no es significativo, pero si las covariables, por lo que un incremento en el ingreso medio genera un incremento en el costo de las casas aproximado de 4.264146, y un incremento en una unidad de la población genera u aumento del costo de casas del 0.003585.

En la respuesta del costo de vida, solo el ingreso medio y el intercepto son significativos; de esta manera, un incremento de una unidad en el ingreso medio, genera un cambio en el índice del costo de vida del $8.235e-01$ aproximadamente.

Ahora evaluamos normalidad en los residuales.

Code

```
##           [,1]                             [,2]
## statistic 0.9389446                        0.9338786
## p.value   0.0851983                        0.06229999
## method    "Shapiro-Wilk normality test"    "Shapiro-Wilk normality test"
## data.name "modelo_outlier$residuals[, i]" "modelo_outlier$residuals[,
i]"
##           [,3]
## statistic 0.9347035
## p.value   0.06554826
## method    "Shapiro-Wilk normality test"
## data.name "modelo_outlier$residuals[, i]"
```

Como se observa ahora no se tiene evidencia suficiente para rechazar la hipótesis nula, por lo tanto los residuales se distribuyen de forma normal.

Constuimos intervalos de confianza

Code

	2.5 %	97.5 %
rent:intercepto	84.692	510.568
rent:income	4.629	13.668
rent:pop	0.001	0.019
house:intercepto	-122.785	77.029
house:income	2.144	6.385
house:pop	-0.001	0.008
COL:intercepto	46.190	82.114
COL:income	0.442	1.205
COL:pop	-0.001	0.001

En los tres modelos la variable del número de población estados no era muy significativa. Para esto, se realiza la prueba de verosimilitud para bajo la hipótesis nula $H_0 = \beta_2 = 0$, que el coeficiente de la población no es necesario para explicar a las respuestas “en conjunto”.

Code

$$\Lambda = \frac{|E|}{|E+H|} = 0.03529033$$

$$F_{3,46} = 4.238306$$

$$\Lambda \leq F_{3,46}$$

Por lo tanto, no se tiene evidencia suficiente para rechazar la hipótesis nula; entonces, la variable de población no explica a la respuestas en su conjunto.

(b) Ajusta tres modelos de regresión lineal de manera separada y veri□ca la utilidad de las variables independientes en cada uno ellos. Compara los resultados con los obtenidos en el inciso (a)

MODELO 1:

$$rent = \beta_0 + \beta_1 income + \beta_2 pop + \epsilon$$

Code

```
##
## Call:
## lm(formula = costolive$rent ~ costolive$income + costolive$pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -206.10  -95.66  -49.76   89.30  548.41
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.440e+02  6.211e+01   8.759 1.61e-11 ***
## costolive$income 3.954e+00  1.142e+00   3.462  0.00114 **
## costolive$pop    8.236e-03  3.122e-03   2.638  0.01121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149.4 on 48 degrees of freedom
## Multiple R-squared:  0.2711, Adjusted R-squared:  0.2407
## F-statistic: 8.924 on 2 and 48 DF,  p-value: 0.0005065
```

Observamos significancia estadística para los tres coeficientes, pero un $\R^2$ ajustado, y sin ajustar bajo.

MODELO 2:

$$house = \beta_0 + \beta_1 income + \beta_2 pop + \epsilon$$

Code

```
##
## Call:
## lm(formula = costolive$house ~ costolive$income + costolive$pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158.30  -58.89  -22.76   43.04  359.10
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      46.721108  35.721748   1.308   0.1971
## costolive$income  3.037709   0.656892   4.624 2.86e-05 ***
## costolive$pop     0.003567   0.001796   1.987   0.0527 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.92 on 48 degrees of freedom
## Multiple R-squared:  0.3357, Adjusted R-squared:  0.308
## F-statistic: 12.13 on 2 and 48 DF,  p-value: 5.453e-05
```

El coeficiente del intercepto no es significativo, el del ingreso medio es significativo y posotivo, pero el de la población es significativo al 95%. El ajuste es bajo ya que el R^2 es menor al $.5$.

MODELO 3:

$$COL = \beta_0 + \beta_1 income + \beta_2 pop + \epsilon$$

Code

```
##
## Call:
## lm(formula = costolive$COL ~ costolive$income + costolive$pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.964  -9.202  -4.365   6.258  62.750
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.340e+01  6.022e+00  13.850  < 2e-16 ***
## costolive$income 4.351e-01  1.107e-01   3.929 0.000273 ***
## costolive$pop    1.526e-04  3.027e-04   0.504 0.616599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.48 on 48 degrees of freedom
## Multiple R-squared:  0.2441, Adjusted R-squared:  0.2126
## F-statistic: 7.751 on 2 and 48 DF,  p-value: 0.00121
```

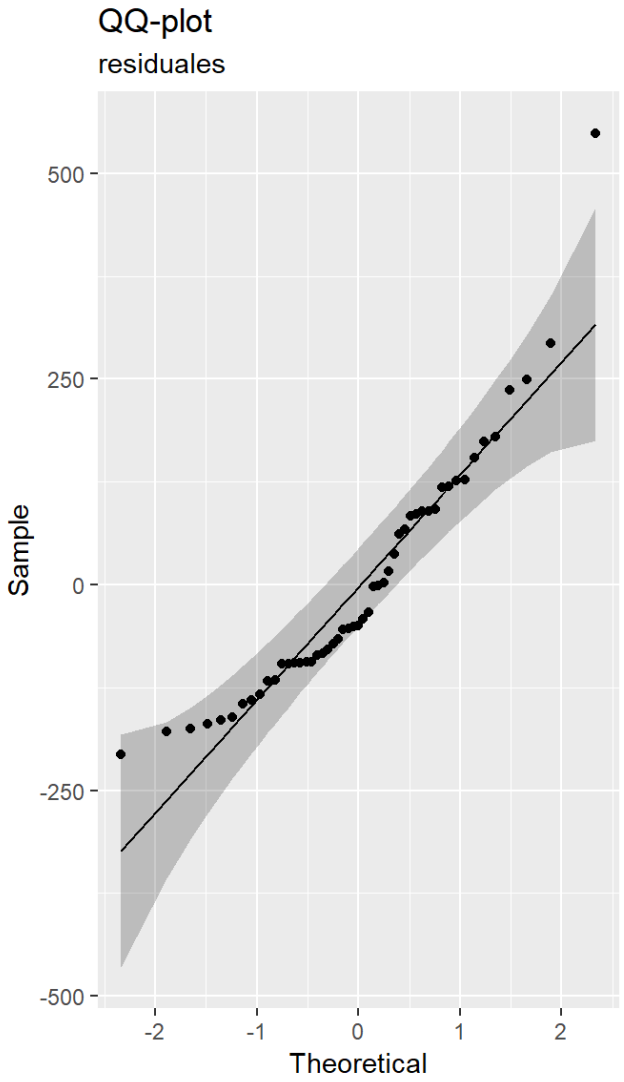
En este modelo, solo el coeficiente y el ingreso medio son significativos, por lo tento incrementos en el ingreso medio de una unidad generan en promedio un aumneto del índice de costo de vida del $4.351e-01$.

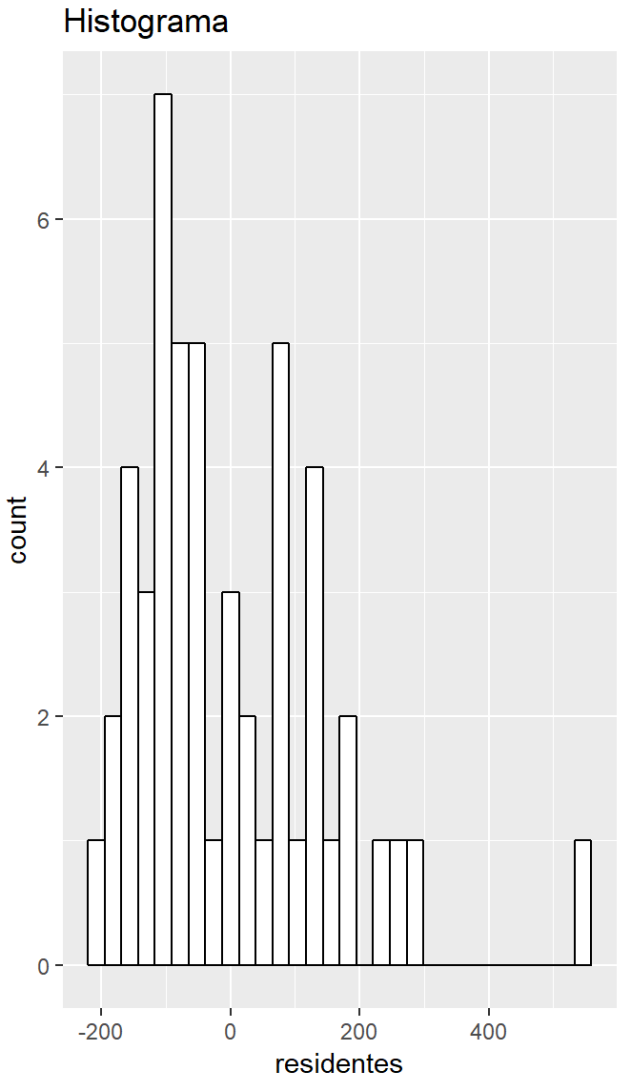
Ahora, tratamos de mejorar el modelo uno mediante la detección de ouliers, y analisando sus residuales.

MODELO 1 MODELO 1:

$$rent = \beta_0 + \beta_1 income + \beta_2 pop + \epsilon$$

Code





Onservamos que la distribución de los residuales tiene colas pesadas, y alguno que otro valor atipico. Aplicando test de normalidad a los residuales obtenemos lo siguiente:

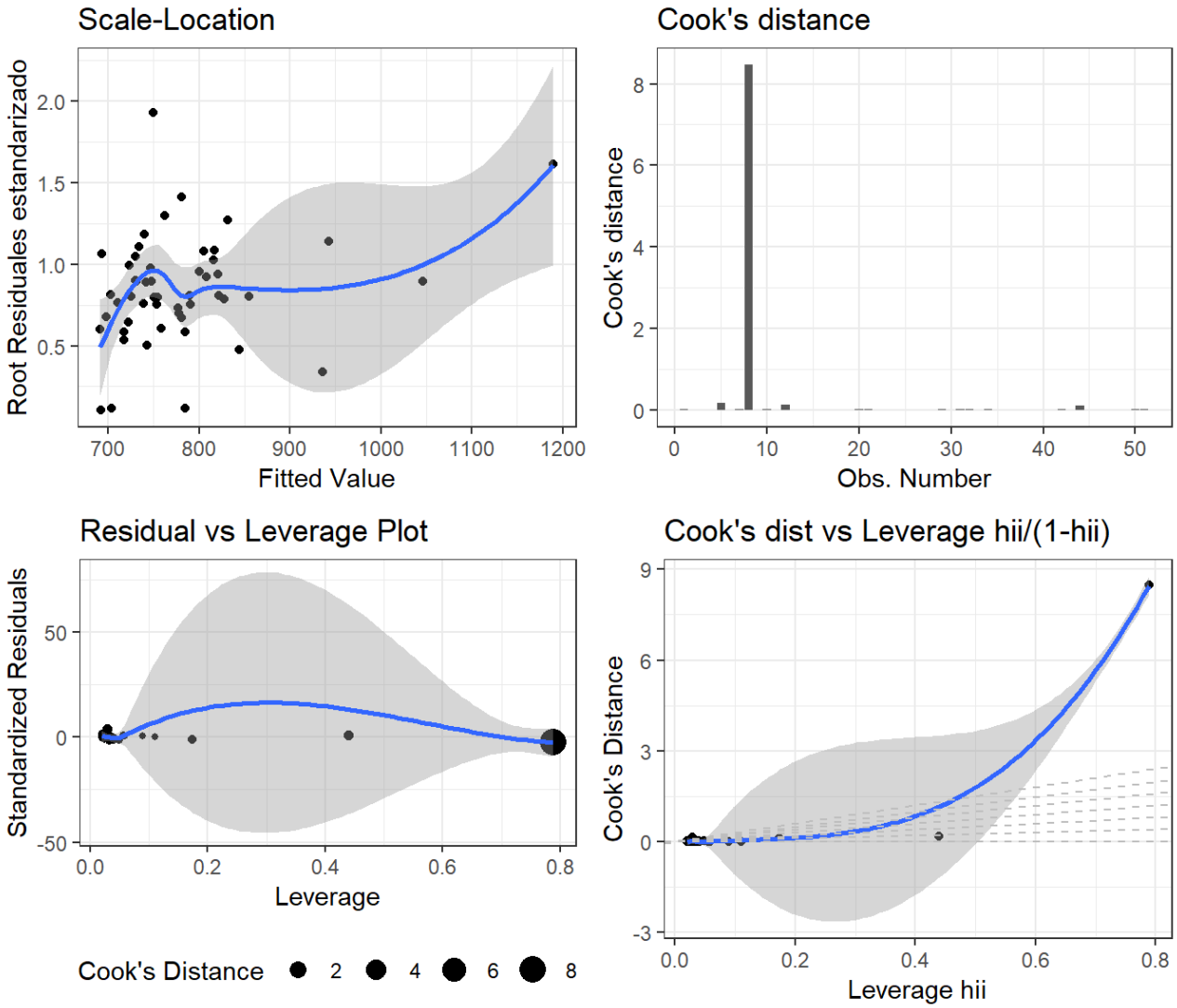
Code

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo1$residuals
## W = 0.90773, p-value = 0.0007695
```

Bajo el crpiterio del p-valor los residuales no son normales.

Ahora detectamos outliers que influyan en el resultado del modelo.

Code



Con la distancia de cook's se detectan valores atipicos que afecten en los resultados del modelo; con el levereg nos menciona que existe algun valor en alguna covariable que es extremo (atipico); a su vez, con el gráfico superior izquierdo, vemos como los residuales siguen el patron de los valores ajustados, con una dispersión que incrementa en el rango de los valores ajustados en el modelo1, por lo tanto da indicios de heterocedasticidad.

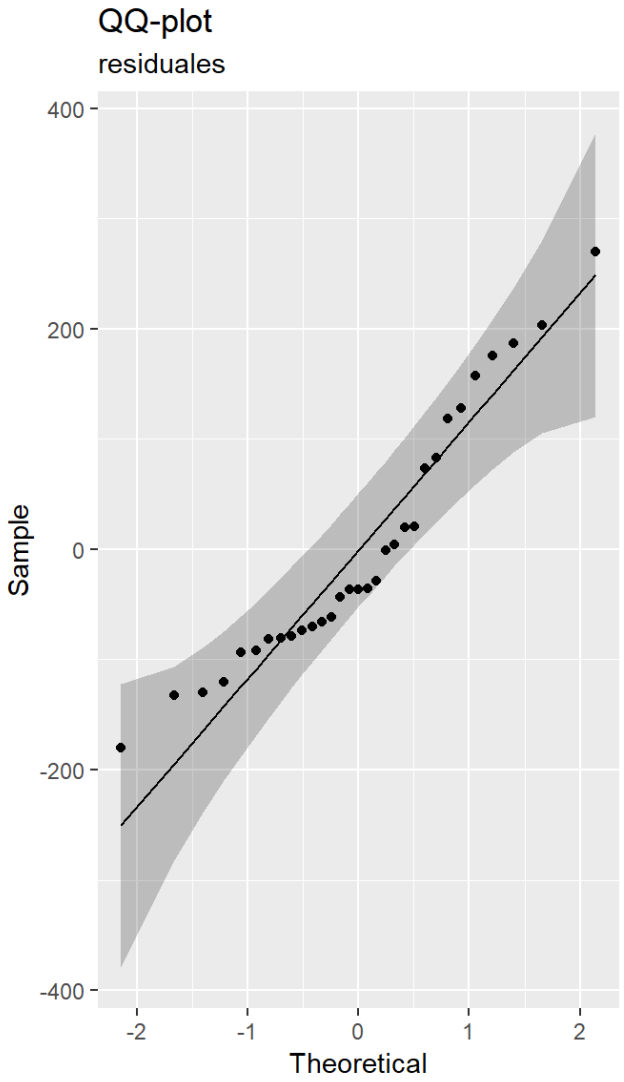
Retiramos los valores qtipicos de las observaciones 5, 12, 21, 32, 44, y actualizamos el modelo 1

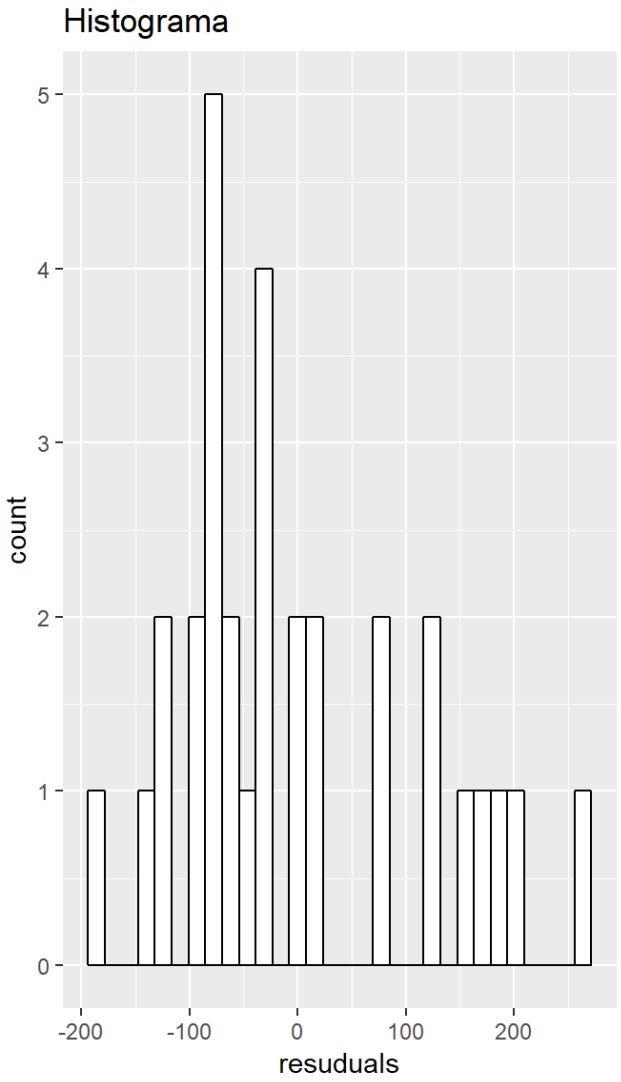
Code

```
##
## Call:
## lm(formula = costolive$rent ~ costolive$income + costolive$pop,
##     subset = (1:35)[-c(5, 12, 21, 32, 44)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -180.18  -79.51  -35.99   77.86  270.31
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.254e+02  5.881e+01   8.933 1.09e-09 ***
## costolive$income 3.771e+00  9.453e-01   3.989 0.000433 ***
## costolive$pop    1.064e-02  4.505e-03   2.361 0.025412 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 118.2 on 28 degrees of freedom
## Multiple R-squared:  0.4044, Adjusted R-squared:  0.3619
## F-statistic: 9.507 on 2 and 28 DF,  p-value: 0.0007064
```

Obtenemos significancia en los tres coeficientes del modelo, todos con una relación positiva, y un R^2 cercana a .5, a no ajustada, indicando una mejora en el ajuste del modelo.

Code





Observamos la gráfica sin los outliers, la distribución de los residuales mejora en las colas, pero no tanto en el centro. Utilizamos el test de normalidad.

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo1_outlier$residuals
## W = 0.92363, p-value = 0.0295
```

se tiene un estadístico W cerano a uno, pero un p-valor que no rechaza la hpótesis nula pero solo al 97 aproximadamente. Realizamos intervalos de ocnfianza para los coeficientes estimados en el modelo uno.

Code

	parametros	ICInferior	ICUpper
Intercepto	525.3558761	404.882	645.829

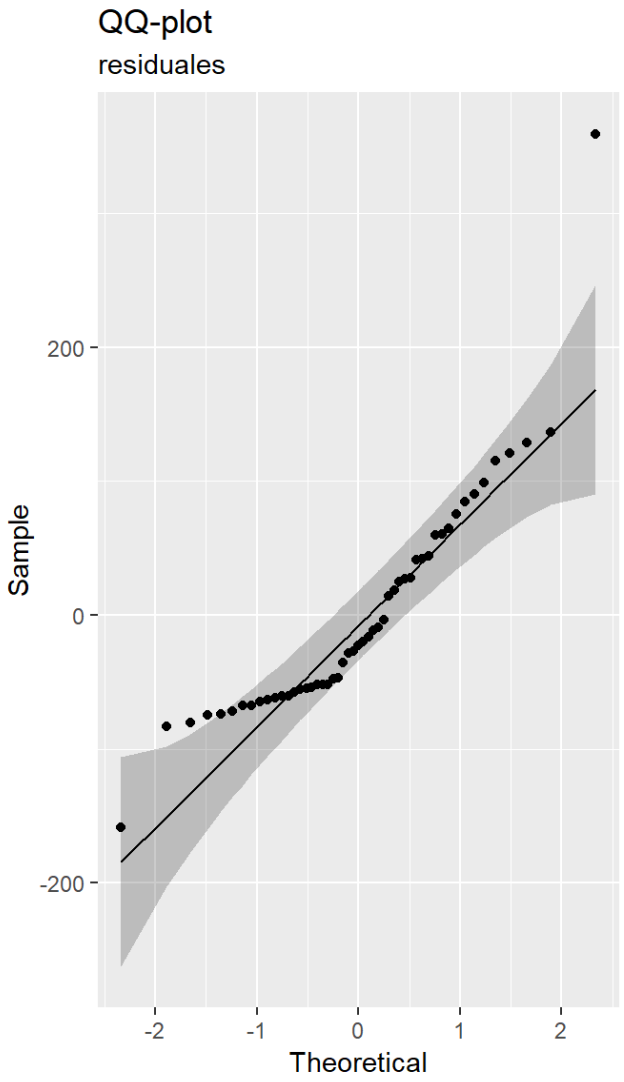
	parametros	ICInferior	ICUpper
income	3.7709144	1.835	5.707
pop	0.0106363	0.001	0.020

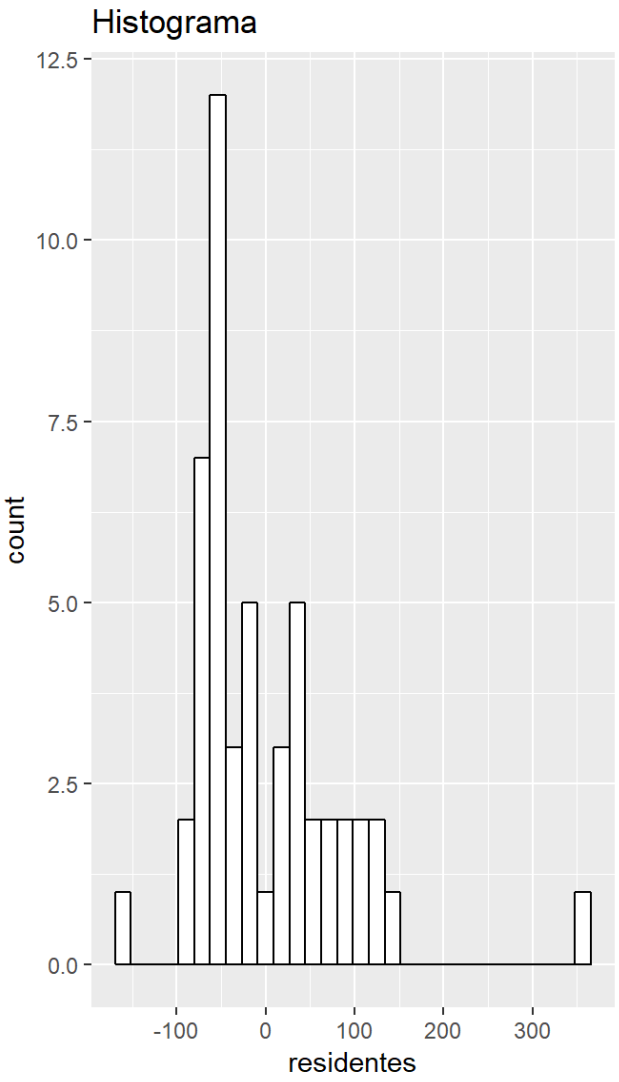
MODELO 2 MODELO 2:

$$house = \beta_0 + \beta_1 income + \beta_2 pop + \epsilon$$

Se realiza el mismo análisis al modelo 2.

Code





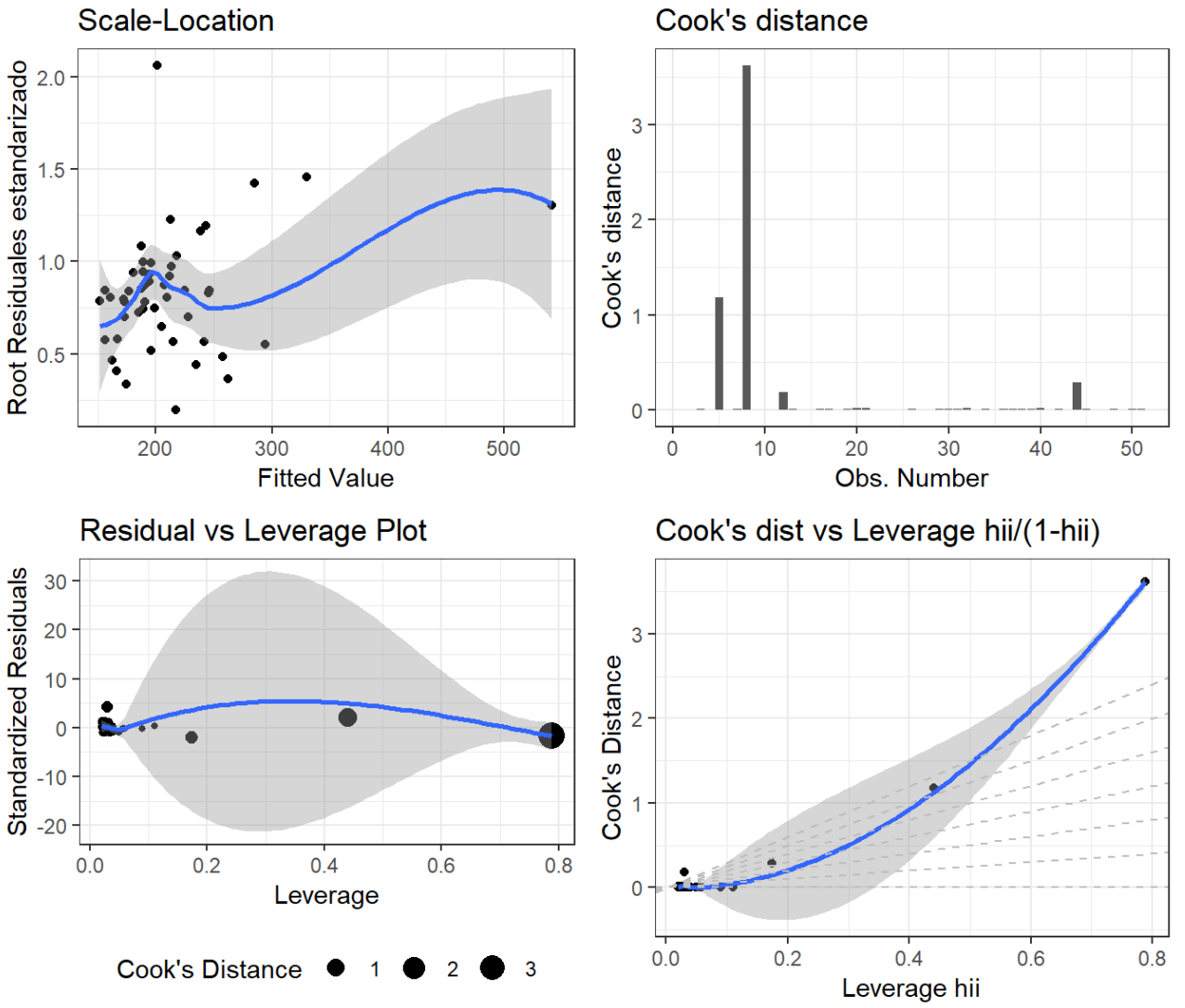
Los residuales parecen no distribuirse normales, pero se detecta mucho valor atipico. Bajo el test de normalidad se rechaza la hipótesis nula.

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo1$residuals
## W = 0.90773, p-value = 0.0007695
```

Se observa a detectar valores atipicos que afecten el resultado del modelo

Code



Observamoscuatro valores atipocos en la distancia de cooks, que superan el valor uno, los cuales afectan el resultado del modelo; por otro lado, en el gráfico superior izquierdo, se observa a los residuales que crecen conforme el rango de los valores ajustados aumenta, de esta manera, se dice que la variable no es constante.

Se procede a detectar y retiras los valores atipico. Se retiran las observaciones $5,8,12,44, y re-ajustamos el modelo

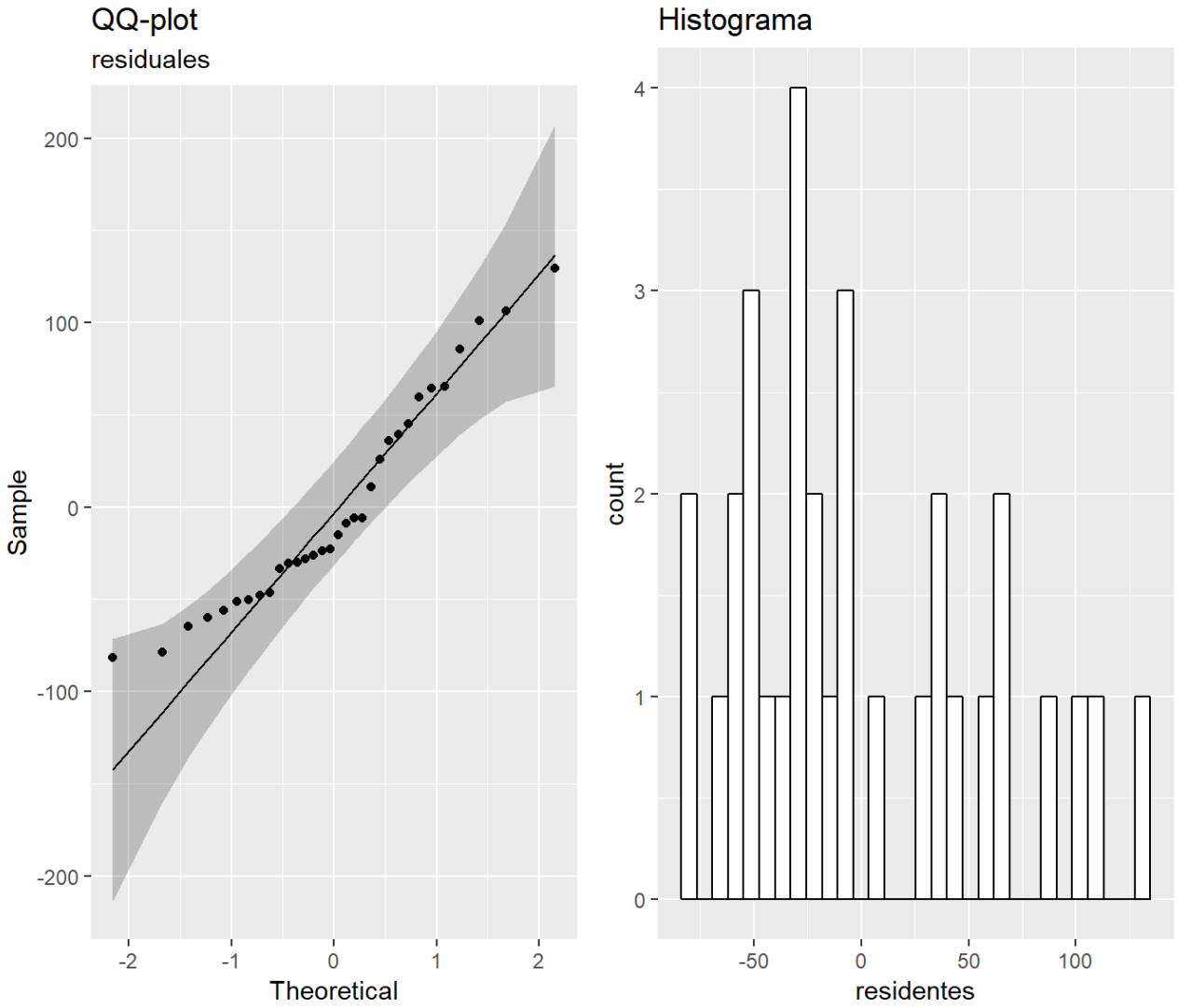
Code

```
##
## Call:
## lm(formula = costolive$house ~ costolive$income + costolive$pop,
##     subset = (1:35)[-c(5, 8, 12, 44)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81.64 -46.76 -18.89  40.66 129.55
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -54.587389  52.646033  -1.037   0.3084
## costolive$income   5.037262   1.108031   4.546 8.94e-05 ***
## costolive$pop      0.004326   0.002254   1.919   0.0649 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.46 on 29 degrees of freedom
## Multiple R-squared:  0.4613, Adjusted R-squared:  0.4242
## F-statistic: 12.42 on 2 and 29 DF,  p-value: 0.0001271
```

El coeficiente R^2 incrementa a casi punto cinco, el intercepto no es significativo, pero el ingreso y la población lo son al, 99% y 95% respectivamente.

Evaluado normalidad en los residuales:

Code



Parece que los residuales se aproximan a la distribución normal. Se utiliza el tes de normalidad y se tiene:

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo2_outlier$residuals
## W = 0.9329, p-value = 0.04719
```

Se tiene evidencia suficiente para rechazar la hipótesis nuela al 95, por lo tanto los residuales son normales.

Se realizan intervalos de confianza a los coeficientes del modelo 2

Code

	parametros	ICInferior	ICUpper
Intercepto	-54.5873888	-162.261	53.086

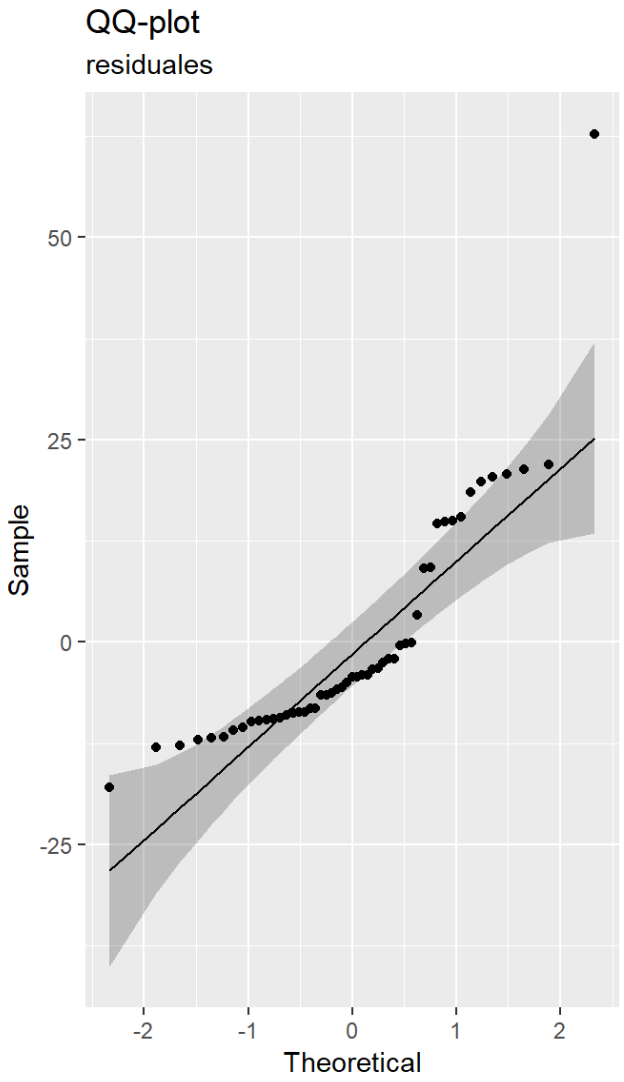
	parametros	ICInferior	ICUpper
income	5.0372624	2.771	7.303
pop	0.0043257	0.000	0.009

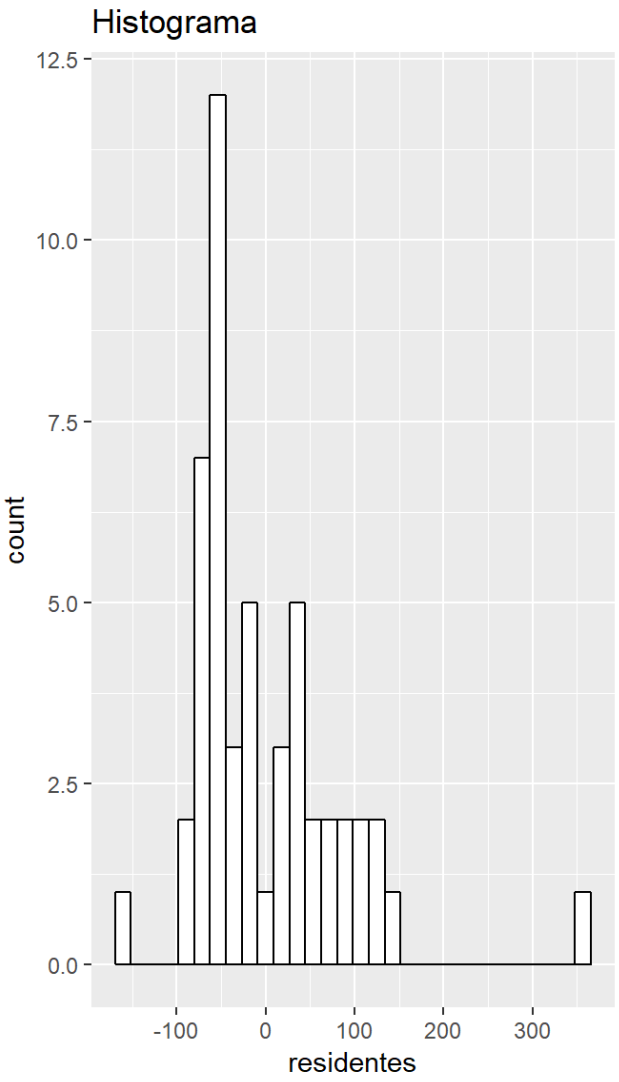
MODELO 3 MODELO 3:

$$COL = \beta_0 + \beta_1 income + \beta_2 pop + \epsilon$$

Aplicamos al tercer modelo lo anterior, mejorar el modelo retirando valores atipicos.

Code





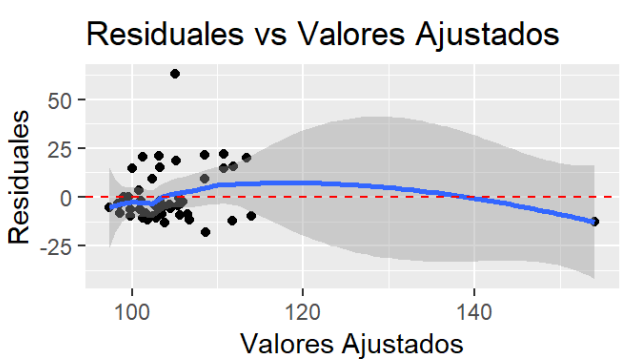
Observamos que os residuales del modelo tres no son normales, y presentan valores atipicos. El test de normalidad rechaza la hipótesis nula de normalidad en los residuales.

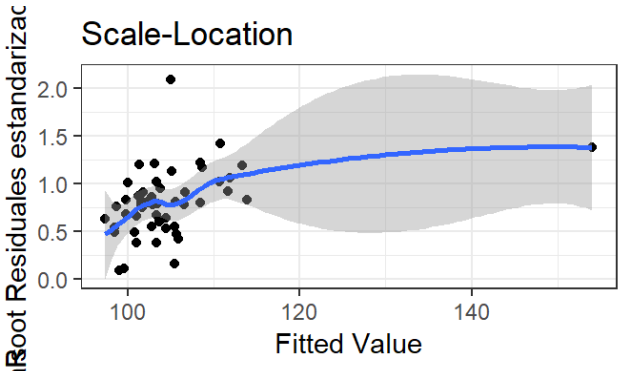
Code

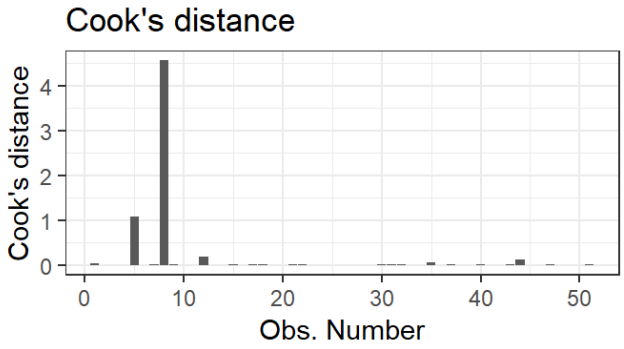
```
##
##  Shapiro-Wilk normality test
##
## data:  modelo3$residuals
## W = 0.78627, p-value = 3.513e-07
```

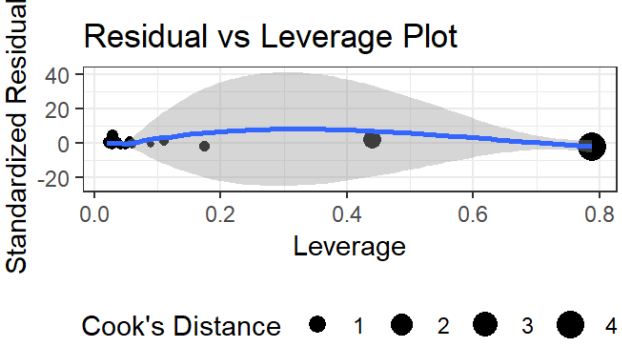
Se observa si los valores atipicos afectan a los resultados del modelo.

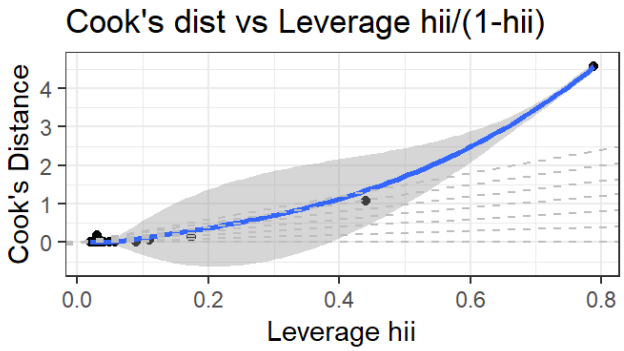
Code











El gráfico superior izaquiedo, presentan residuales muy dispersos, pero no de forma constante; el superior derecho, argumenta lo anterior, ya que los resduales tienden a ajustarce a la linea del ajuste de los valores del modelo; por otro lado, la distacia de cooks, detecta cinco valores que pueden ocacionar que el modelo no se desempeñe de manera correcta.

Presentamo el modelo sin los datos atipicos.

Code

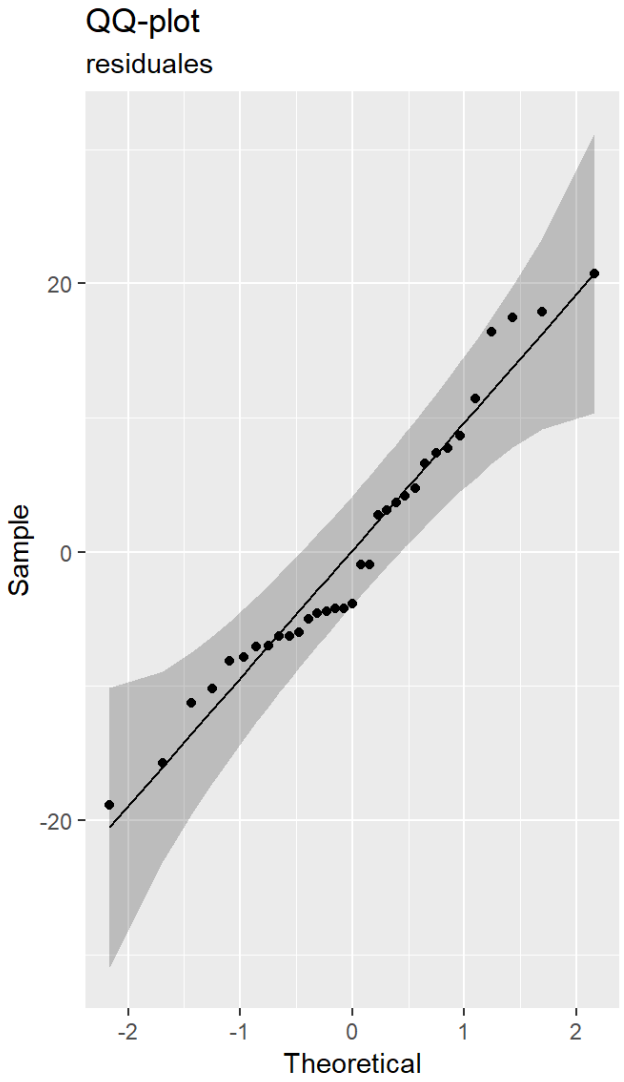
```
##
## Call:
## lm(formula = costolive$COL ~ costolive$income + costolive$pop,
##     subset = (1:35)[-c(8, 12)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.860  -6.301  -3.834   6.559  20.743
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.857e+01  8.777e+00   6.673 2.16e-07 ***
## costolive$income 9.161e-01  1.879e-01   4.877 3.30e-05 ***
## costolive$pop    5.574e-04  2.482e-04   2.246   0.0322 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.09 on 30 degrees of freedom
## Multiple R-squared:  0.5062, Adjusted R-squared:  0.4732
## F-statistic: 15.37 on 2 and 30 DF,  p-value: 2.534e-05
```

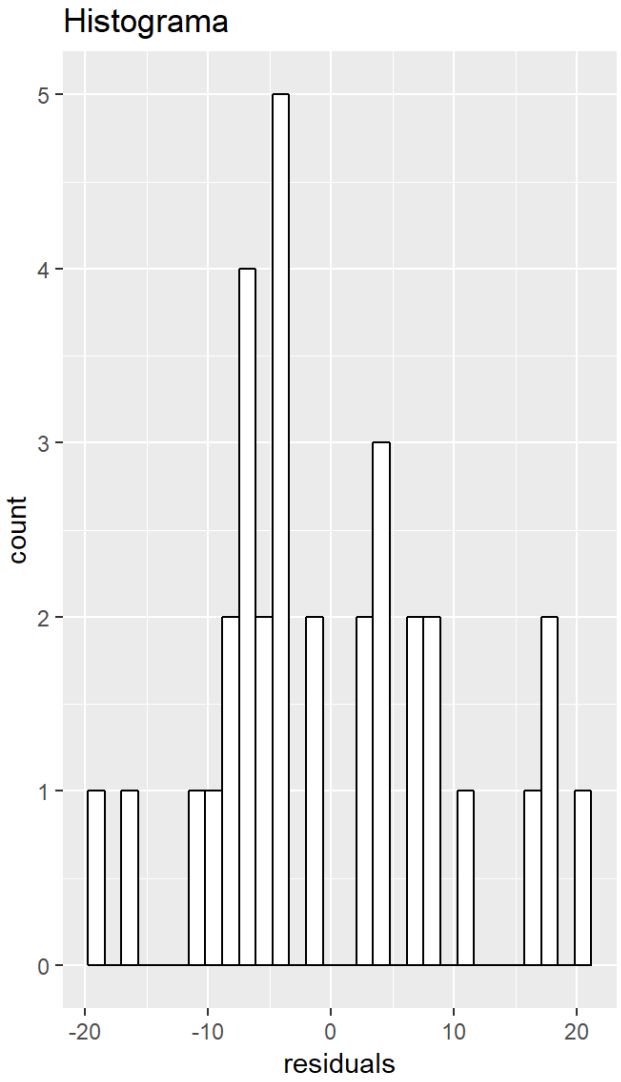
La mejora en el modelo es notoria, los coeficientes del intercepto, ingreso promedio, y población, son significativos a distintos niveles. El valor del R^2 no ajustada en del 0.05, y el ajustado se aproxima a ese valor; a su vez, el estadístico F, menciona que a manera global el modelo estpa bien especificado.

De esta manera, cambios en una unidad del ingreso promedio y del número de población afectan de manera positiva al índice de costo de vida en la magnitud de los coeficientes que se observan arriba.

Revisamos normalidad en el modelo 3.

Code





Sin los valores atipicos os reisduales se aproximan a la normal en las colas, no obstante aún se ve que existen valores que hacen que los residuales no se comporten exactamente con observaciones normales. Por otro lado, si utilizamos el test de normalidad nos indica que no se tiene evidencia suficiente para rechazar la hipótesis nula, por lo cual, los residuales son se distribuyen normales.

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo3_outlier$residuals
## W = 0.9588, p-value = 0.239
```

Ahora construimos intervalos de confianza a los coeficentes del modelo 3.

Code

	parametros	ICInferior	ICUpper

	parametros	ICInferior	ICUpper
Intercepto	58.5743850	40.648	76.500
income	0.9161397	0.532	1.300
pop	0.0005574	0.000	0.001

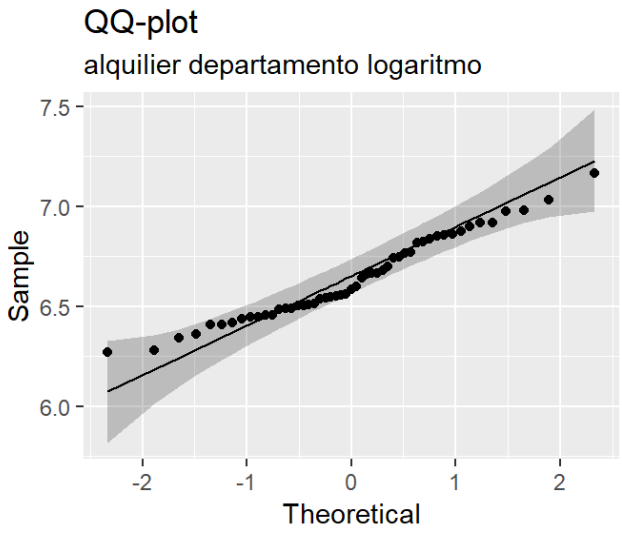
MODELOS INDIVIDUALES CON LOGARITMO

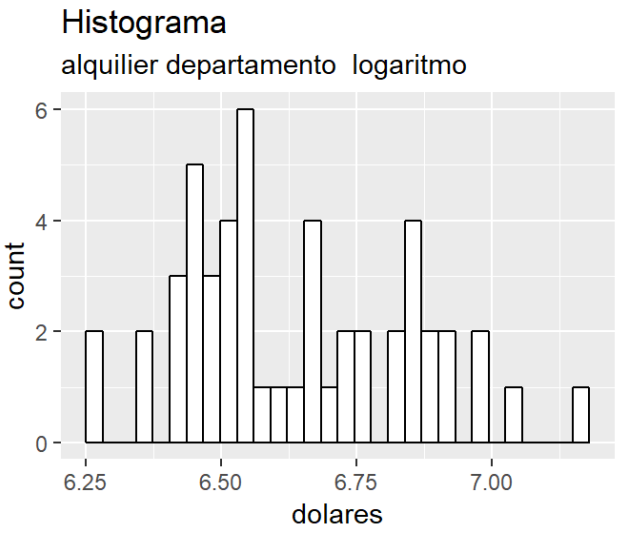
A manera de ejercicio, aplicamos logaritmo a los modelos individuales, para ver que ta su mejora; no se pretende presentar todos los resultados anteriores, sin embargo, se realizaron y quedan a su disposución en caso de necesitarlos. Nota, no se aplico logaritmo a la población.

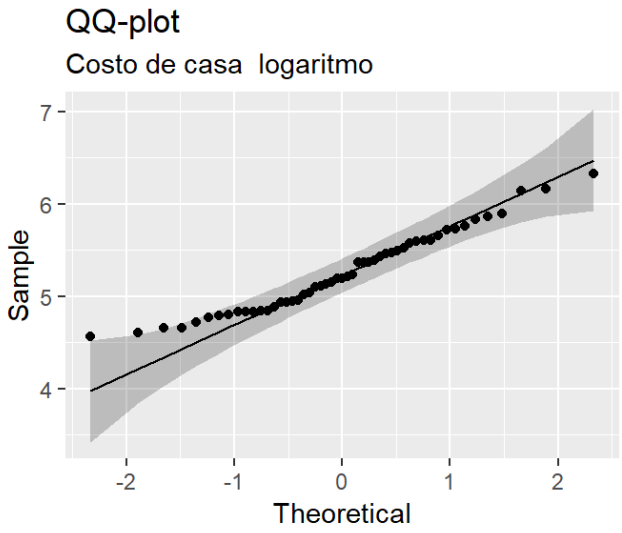
Code

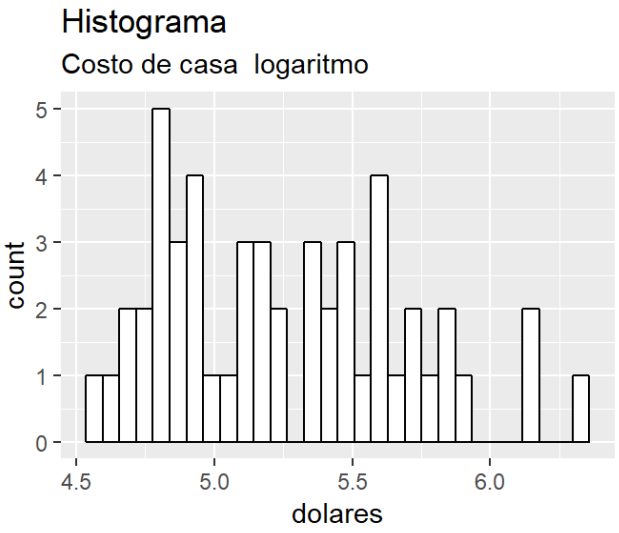
Se presenta descriptivos de la observaciones ahora con logaritmos, sus distribuciones se aproximan a la normal.

Code

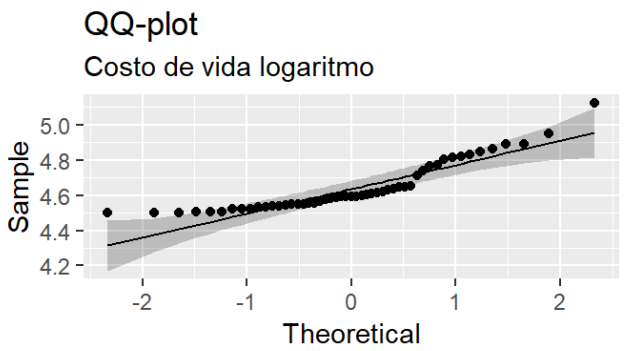


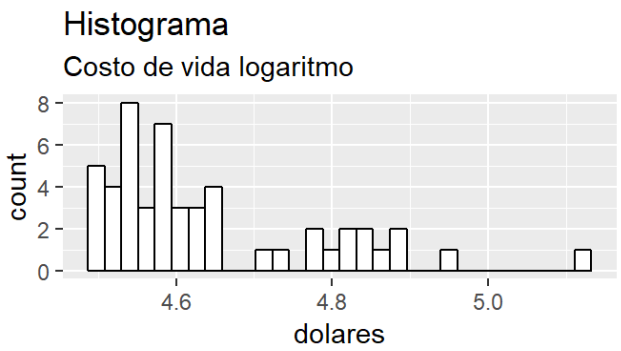


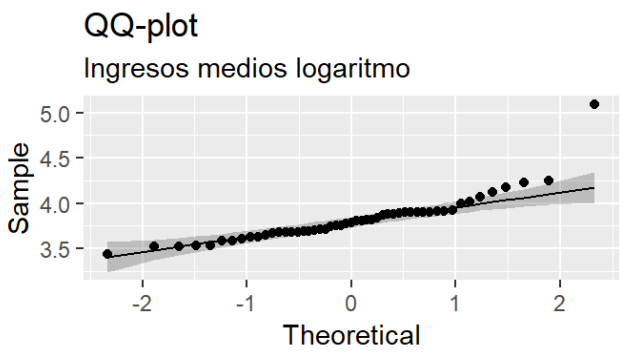


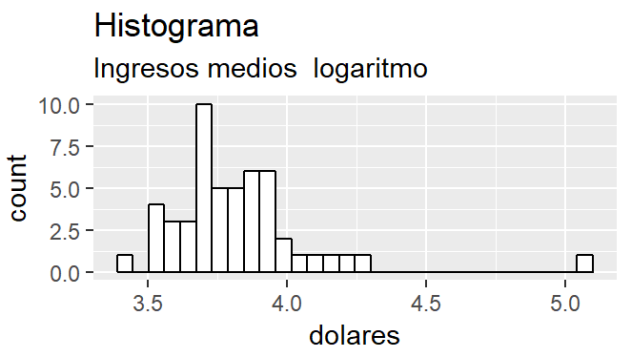


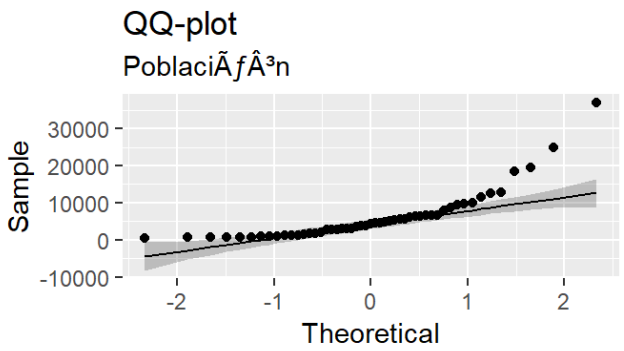
Code

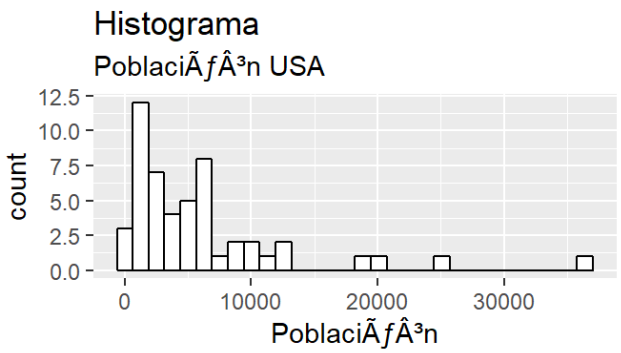






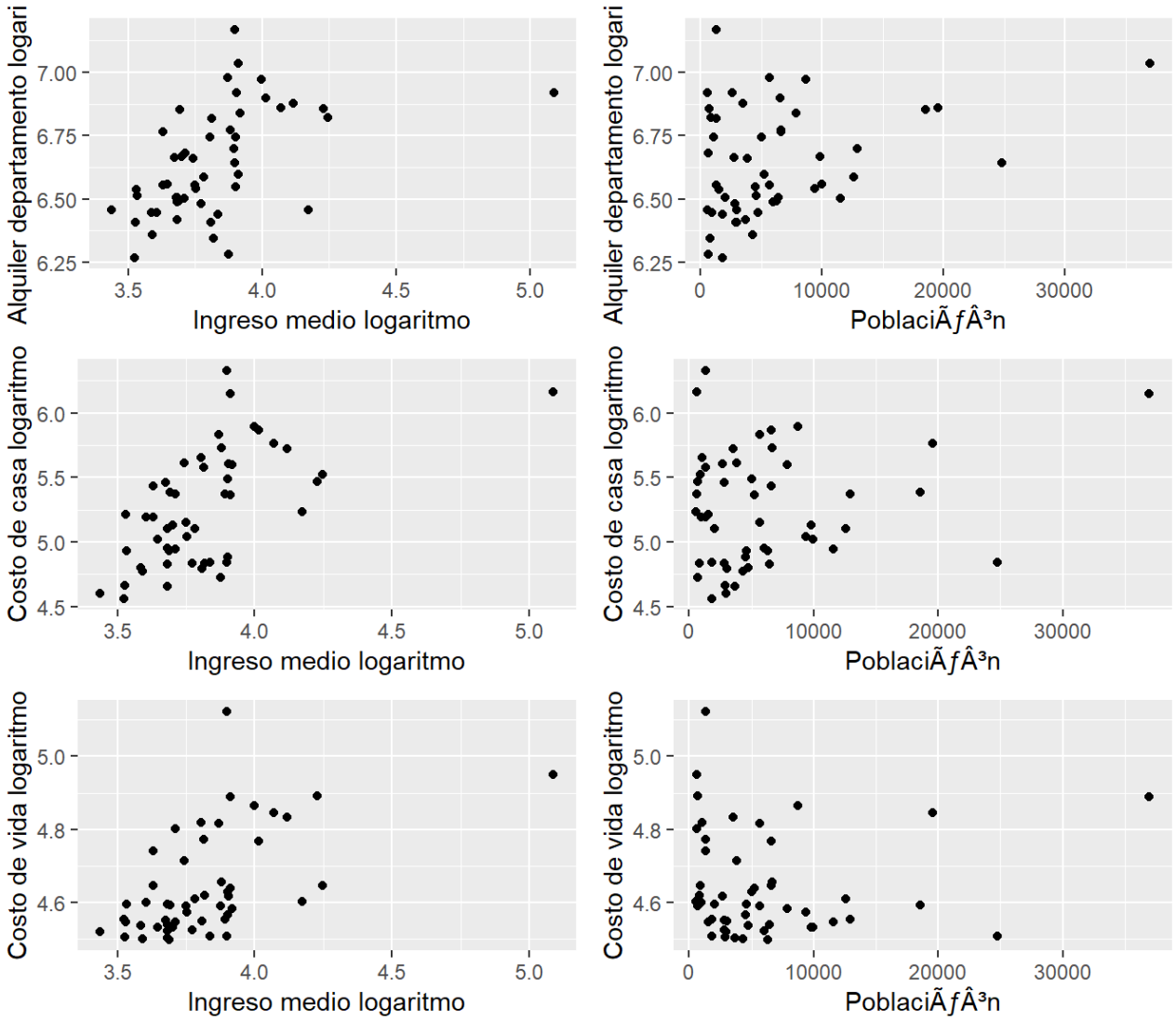






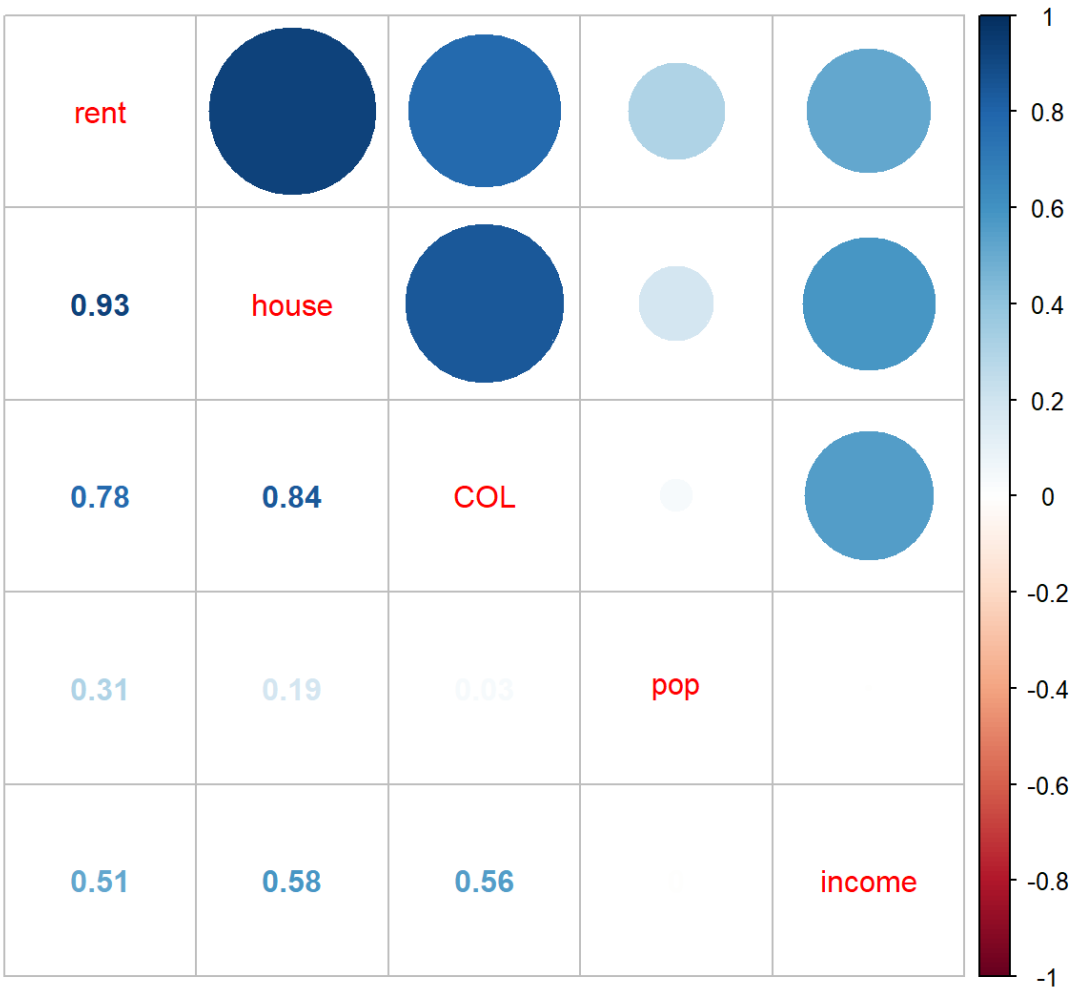
La dispersión entre variables respuestas y covariables mejora, en cuestion que se observa relación en casi todos los casos.

Code



La corelación entre variables es mayor respetco a las variables repsuestas y covariales.

Code



MODELO 1:

$$log(rent) = \beta_0 + \beta_1 log(income) + \beta_2 pop + \epsilon$$

Code

```
##
## Call:
## lm(formula = costolive$rent ~ costolive$income + costolive$pop)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33103 -0.10349 -0.03799  0.10989  0.54188
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.976e+00  3.621e-01  13.741  < 2e-16 ***
## costolive$income 4.203e-01  9.445e-02   4.450 5.09e-05 ***
## costolive$pop    9.594e-06  3.582e-06   2.678   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1717 on 48 degrees of freedom
## Multiple R-squared:  0.3596, Adjusted R-squared:  0.3329
## F-statistic: 13.48 on 2 and 48 DF,  p-value: 2.263e-05
```

El modelo tiene coeficientes significativos. La interpretación es que un incremento porcentual en lo logartmo del ingreso medio afecta a costo de la renta en $4.203e-01$. El R^2 mejora ligeramente.

MODELO 2:

$$log(house) = \beta_0 + \beta_1 log(income) + \beta_2 pop + \epsilon$$

Code

```
##
## Call:
## lm(formula = costolive$house ~ costolive$income + costolive$pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7252 -0.2712 -0.0472  0.2907  1.0478
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.365e+00  7.468e-01   1.827   0.0739 .
## costolive$income 1.001e+00  1.948e-01   5.136 5.06e-06 ***
## costolive$pop    1.217e-05  7.387e-06   1.648   0.1059
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3542 on 48 degrees of freedom
## Multiple R-squared:  0.3773, Adjusted R-squared:  0.3513
## F-statistic: 14.54 on 2 and 48 DF,  p-value: 1.157e-05
```

El modelo solo tiene coeficientes de ingreso medio significativo. La interpretación es que un incremento porcentual en lo logartmo del ingreso medio afecta a costo de la renta en $1.001e+00$. El R^2 mejora ligeramente.

MODELO 3:

$$log(COL) = \beta_0 + \beta_1 log(income) + \beta_2 pop + \epsilon$$

Code

```
##
## Call:
## lm(formula = costolive$COL ~ costolive$income + costolive$pop)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17412 -0.07765 -0.03291  0.06247  0.45860
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.467e+00  2.526e-01  13.729  < 2e-16 ***
## costolive$income 3.068e-01  6.587e-02   4.657 2.56e-05 ***
## costolive$pop    7.267e-07  2.498e-06   0.291    0.772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1198 on 48 degrees of freedom
## Multiple R-squared:  0.3121, Adjusted R-squared:  0.2834
## F-statistic: 10.89 on 2 and 48 DF,  p-value: 0.0001262
```

El modelo tiene coeficientes significativos el intercepto e ingreso medio. La interpretación es que un incremento porcentual en lo logartmo del ingreso medio afecta a costo de la renta en $3.068e-01 $. El R^2 mejora ligeramente.

Detectamos y quitamos outliers para mejorar el modelo, ya que sabemos que estos valores atipicos están afectando la calidad del modelo.

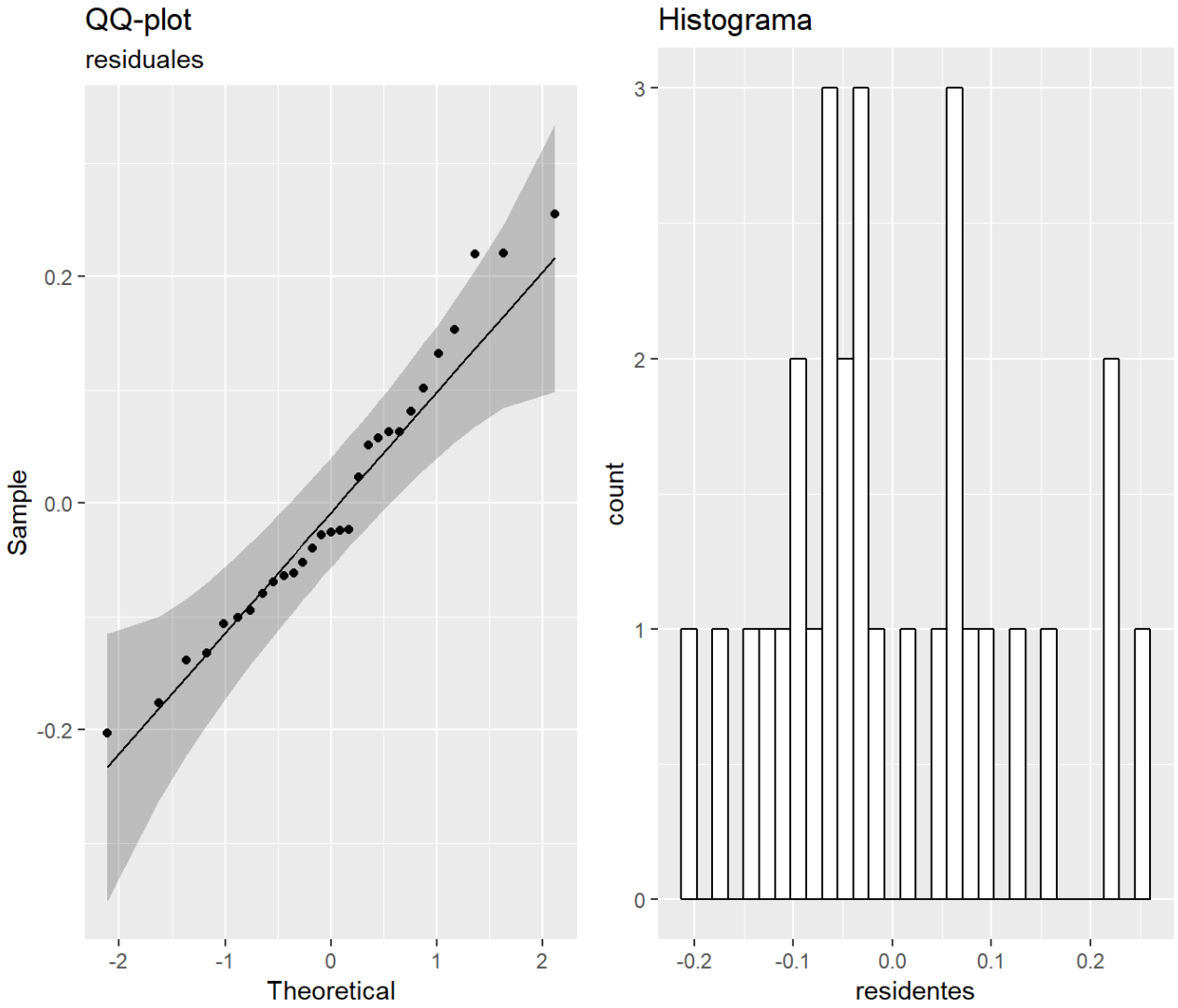
MODELO1 SIN OUTLIERS

Code

```
##
## Call:
## lm(formula = costolive$rent ~ costolive$income + costolive$pop,
##     subset = (1:35)[-c(5, 8, 12, 21, 29, 32, 44, 50, 51)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20340 -0.08008 -0.02599  0.06332  0.25485
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.369e+00  4.284e-01  10.198 1.41e-10 ***
## costolive$income 5.809e-01  1.131e-01   5.135 2.35e-05 ***
## costolive$pop    1.003e-05  4.823e-06   2.080   0.0476 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.124 on 26 degrees of freedom
## Multiple R-squared:  0.5504, Adjusted R-squared:  0.5158
## F-statistic: 15.92 on 2 and 26 DF,  p-value: 3.065e-05
```

Se observa significancia en sus coeficientes, y un buen ajuste con el R^2

Code



Los residuales se aproximas a distribuirse normales y la prueba de normalidad indica que son normales.

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo1_outlier$residuals
## W = 0.96054, p-value = 0.339
```

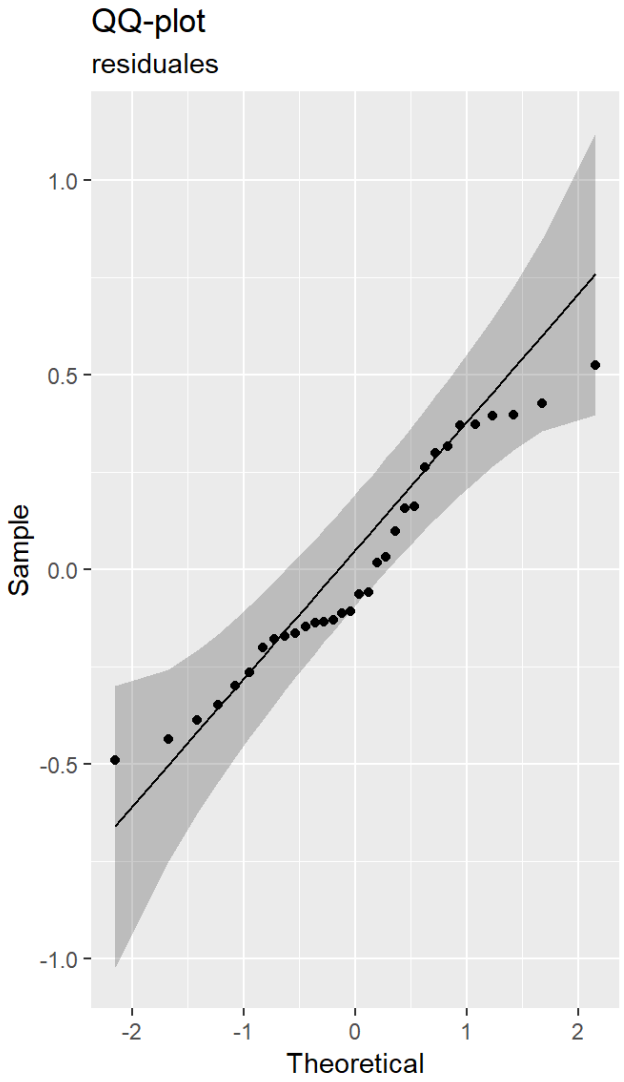
MODELO2 SIN OUTLIERS

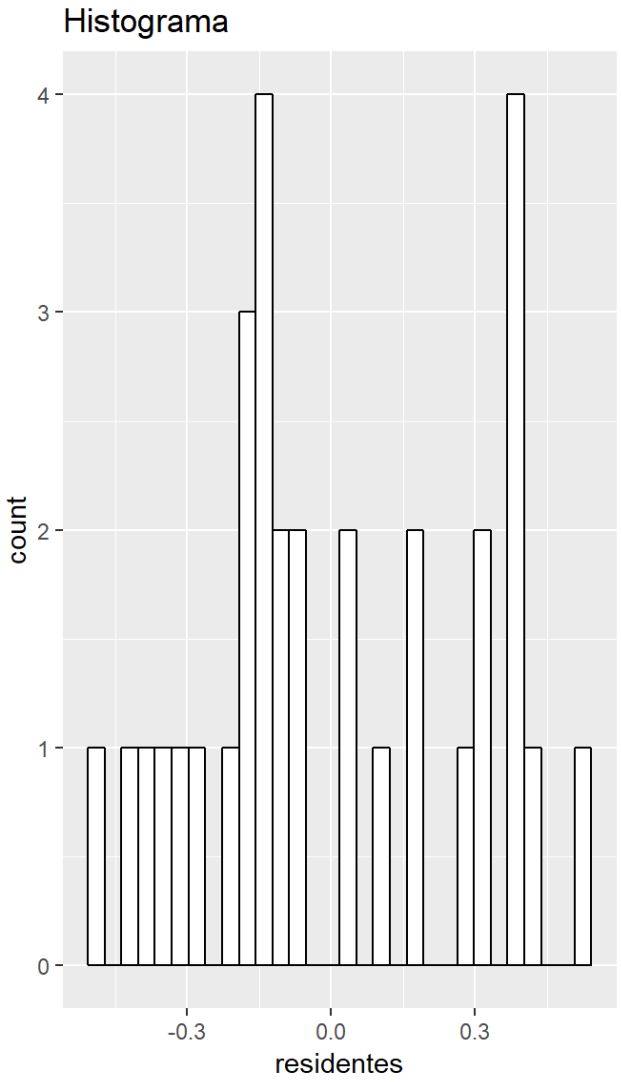
Code

```
##
## Call:
## lm(formula = costolive$house ~ costolive$income + costolive$pop,
##     subset = (1:35)[-c(5, 8, 12, 44, 51)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49037 -0.17284 -0.08553  0.27197  0.52416
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.498e-01  9.883e-01   0.556    0.582
## costolive$income 1.200e+00  2.603e-01   4.609 7.51e-05 ***
## costolive$pop    2.023e-05  1.107e-05   1.827    0.078 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2917 on 29 degrees of freedom
## Multiple R-squared:  0.4692, Adjusted R-squared:  0.4326
## F-statistic: 12.82 on 2 and 29 DF,  p-value: 0.0001027
```

Se observa significancia en sus coeficientes, y un buen ajuste con el R^2

Code





Los residuales mejoran su distribución a una normal. Bajo el test de normalidad no rechazamos la hipótesis nula de normalidad, por lo tanto los residuales distrinuyen normales.

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo2_outlier$residuals
## W = 0.94765, p-value = 0.1235
```

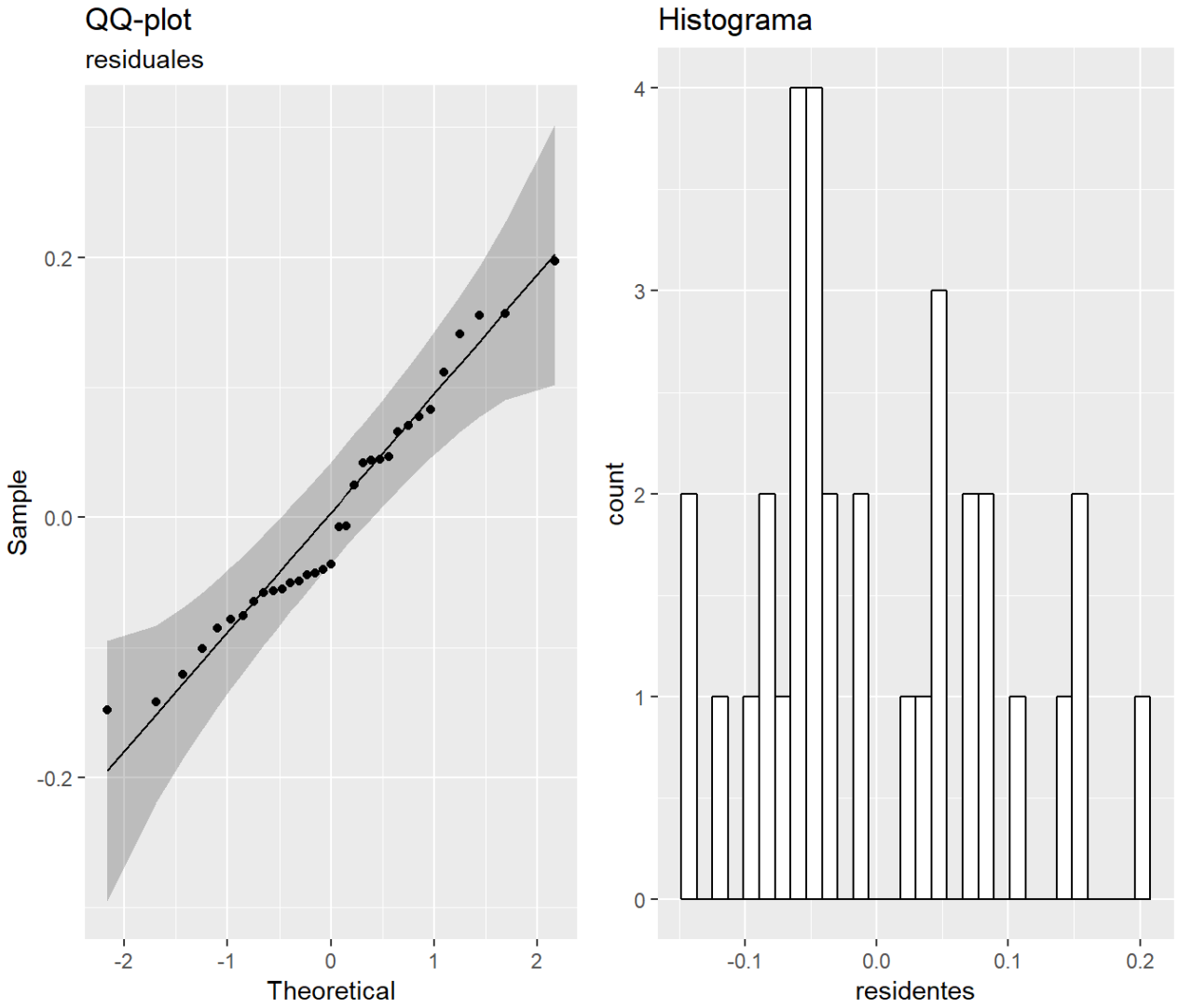
MODELO3 SIN OUTLIERS

Code

```
##
## Call:
## lm(formula = costolive$COL ~ costolive$income + costolive$pop,
##     subset = (1:35)[-c(8, 12)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14812 -0.05807 -0.03564  0.06554  0.19666
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.093e+00  3.158e-01   9.793 7.43e-11 ***
## costolive$income 3.984e-01  8.331e-02   4.782 4.31e-05 ***
## costolive$pop    4.505e-06  2.305e-06   1.954   0.0601 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09339 on 30 degrees of freedom
## Multiple R-squared:  0.4928, Adjusted R-squared:  0.459
## F-statistic: 14.58 on 2 and 30 DF,  p-value: 3.777e-05
```

Se observa significancia en sus coeficientes, y un buen ajuste con el R^2

Code



Los residuales mejoran su distribución a una normal. Bajo el test de normalidad no rechazamos la hipótesis nula de normalidad, por lo tanto los residuales distrinuyen normales.

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo3_outlier$residuals
## W = 0.95606, p-value = 0.1995
```

En conclusión los modelos independientes se ajustan mejor que un modelo en conjunto, ya que existen covariables que en conjunto no afectan a las repsuesta, pero de forma individual sí. Asimismo, los modelos individuales con el ingreso medio en ogaritmos tiene mejor ajuste en base al R^2

EJERCICIO 3

3. Muchos inversionistas estÃ¡n buscando dividendos que se pagarÃ¡n de los bene□cios futuros. Los datos del archivos ash hi tech.txt enumeran una serie de caractericas sobre su situaciÃ³n □nanciera, hasta septiembre del 2010, de varias empresas de tecnologs e informaci. Las variables resultado a explicar son los dividendos actuales y futuros (current y 60% payout).

Code

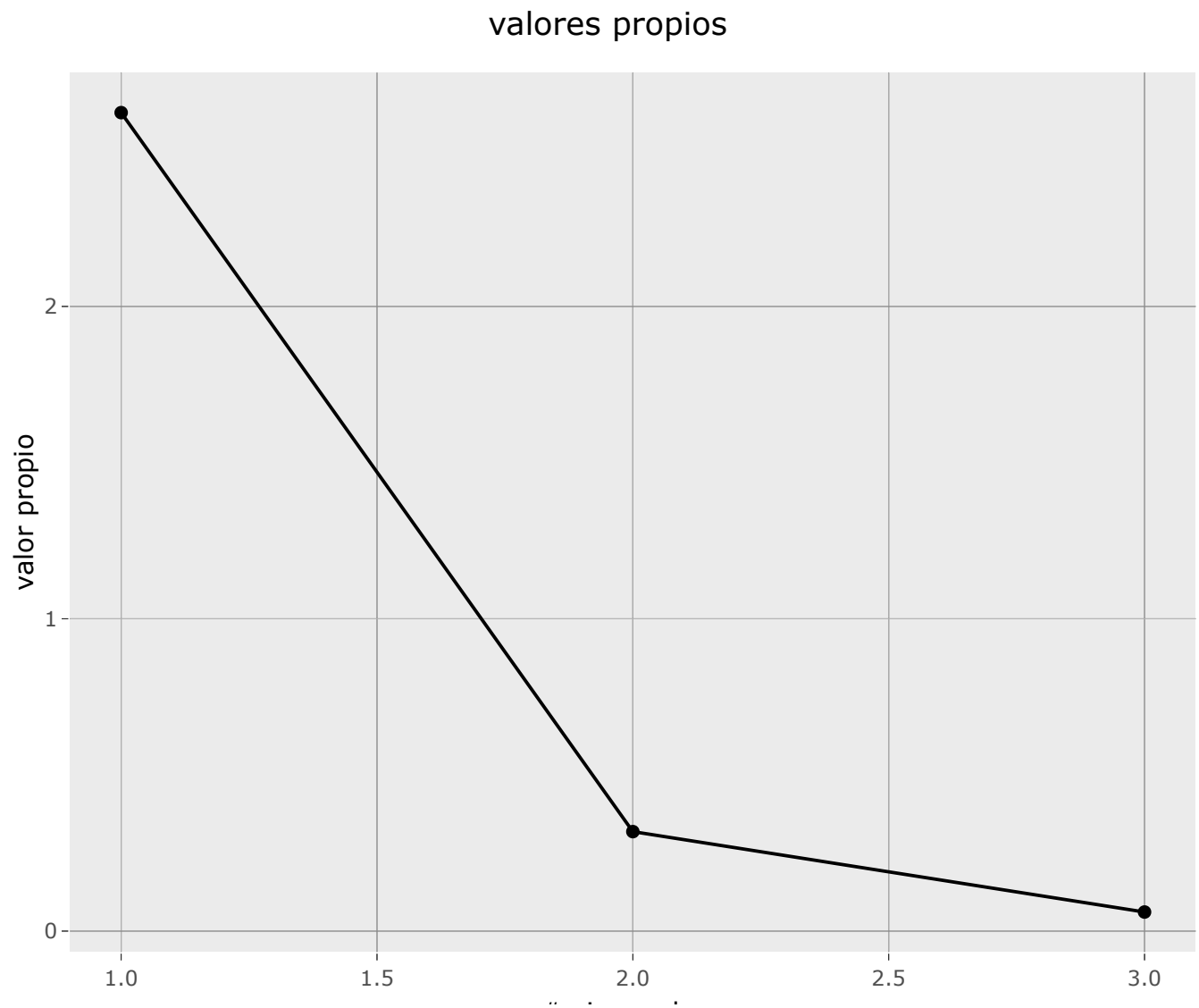
(a) Desarrolla un modelo de regresiÃ³n multivariada partir de la capitalizaciÃ³n de mercado (market cap), efectivo neto (net cash) y □ujo de efectivo (cash □ow) y analiza el efecto que tienen conjuntamente respecto a los dividendos.

Utilizando las respuestas dividendos actuales y futuros, con las covariables ya mencionadas, utilizamos la matriz de covarianza de las covariables y vemos si existe mutlicolinealidad.

Code

Vemos el screplot:

Code



```
# eigenvalues
```

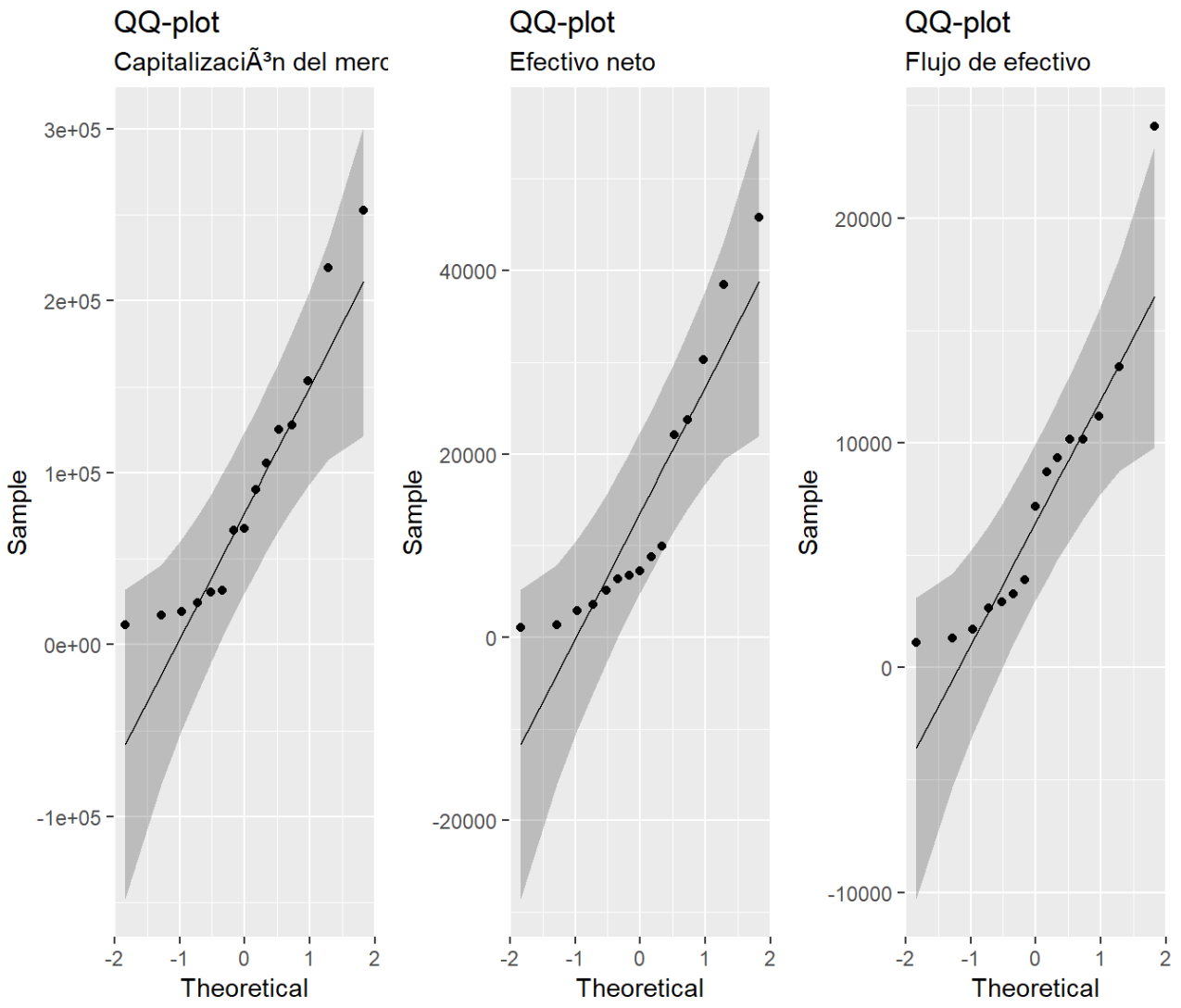
$$\lambda_{max}/\lambda_{min} = 42.89294$$

El cual indica colinealidad entre las variables, que claro esta, porque muchas tienen que ver con el ingreso total del mercado de capitales.

Code

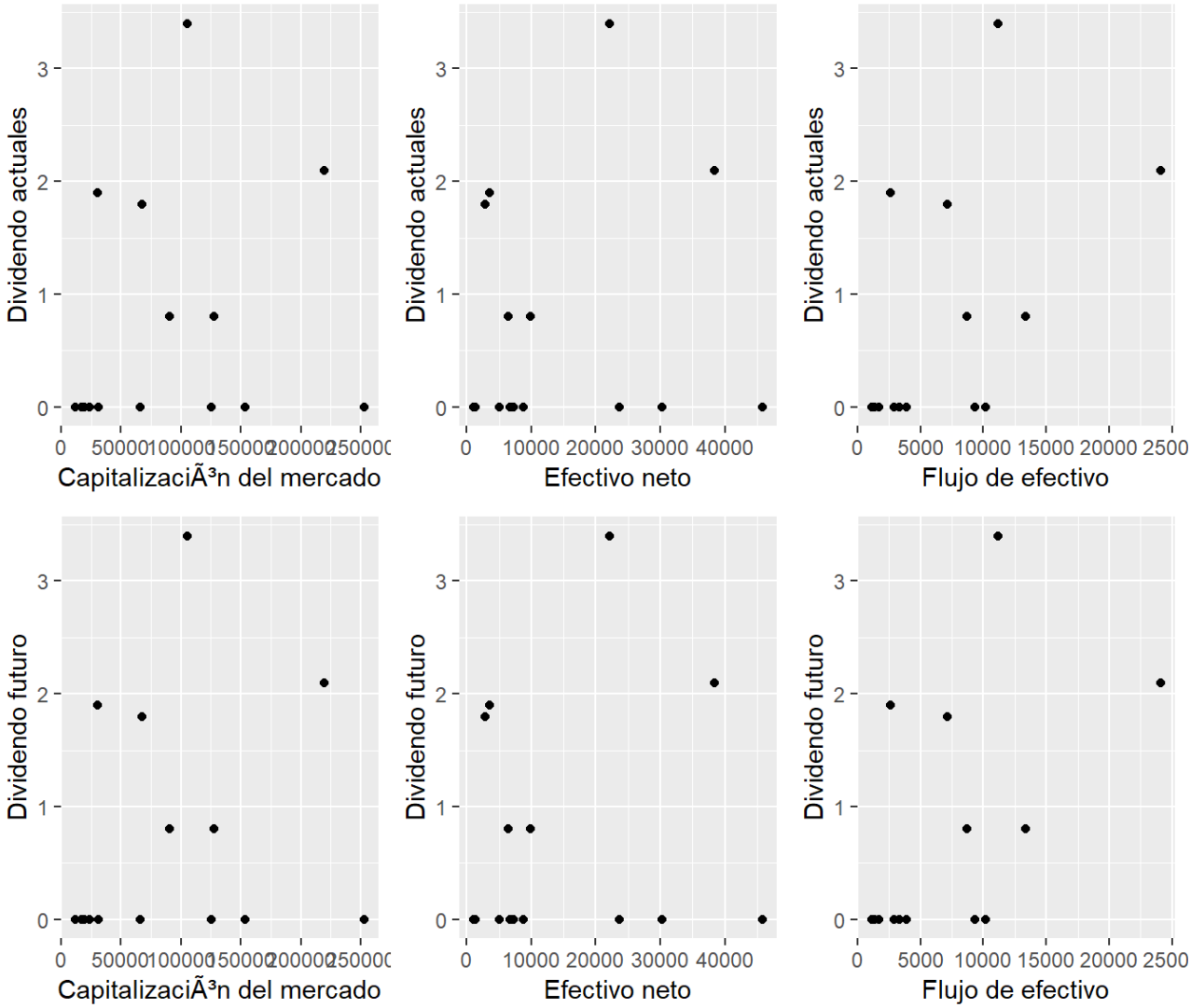
```
## [1] 42.89294
```

Code



La observaciones se encunetran dentro del intevalo de confianza pero no parece normales, cuentan con colas pesadad.

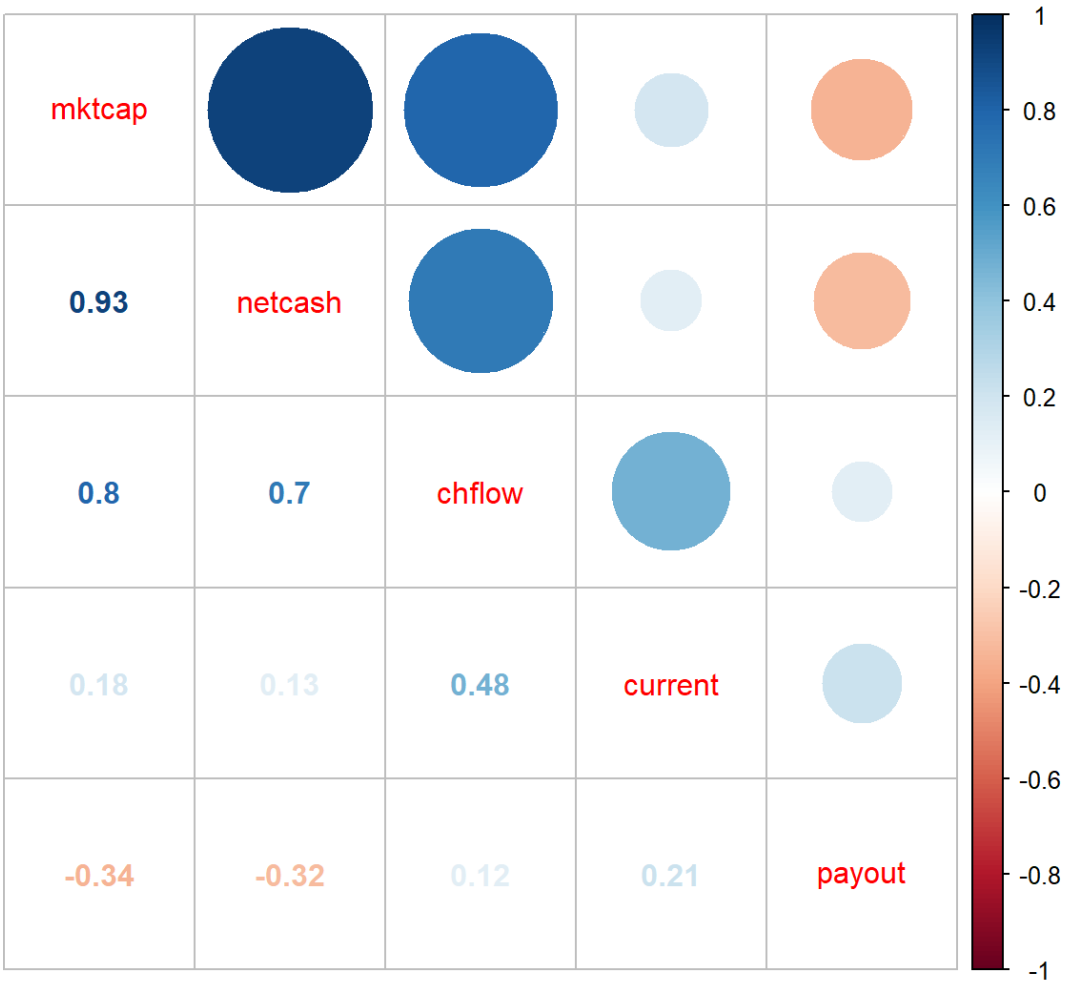
Code



Son pocos datos, la distrribución entre respuesta y covariables se ven como en el gráfico de arriba.

A continuación presentamos el gráfico de correlación de las observaciones, las covariables tiene relación lineal negativa respecto a las respuestas.

Code


mktcap
0.93
netcash
0.8
0.7
chflow
0.18
0.13
0.48
current
-0.34
-0.32
0.12
0.21
payout
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Code

```
## Response current :
##
## Call:
## lm(formula = current ~ X$mktcap + X$netcash + X$chflow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9046 -0.4708 -0.2848  0.1002  2.2458
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.293e-01  4.280e-01   0.536   0.6027
## X$mktcap    -6.406e-06  1.116e-05  -0.574   0.5776
## X$netcash   -6.747e-06  4.980e-05  -0.135   0.8947
## X$chflow     1.567e-04  7.228e-05   2.168   0.0529 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9958 on 11 degrees of freedom
## Multiple R-squared:  0.3318, Adjusted R-squared:  0.1496
## F-statistic: 1.821 on 3 and 11 DF,  p-value: 0.2016
##
##
## Response payout :
##
## Call:
## lm(formula = payout ~ X$mktcap + X$netcash + X$chflow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7960 -0.7866 -0.3777  0.8996  2.9271
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.902e+00  7.101e-01   8.312 4.53e-06 ***
## X$mktcap    -4.337e-05  1.852e-05  -2.341  0.03909 *
## X$netcash    4.209e-05  8.262e-05   0.509  0.62052
## X$chflow     3.962e-04  1.199e-04   3.304  0.00703 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.652 on 11 degrees of freedom
## Multiple R-squared:  0.5574, Adjusted R-squared:  0.4367
## F-statistic: 4.617 on 3 and 11 DF,  p-value: 0.0252
```

Corremos el modelo y en la primera respuesta que son los dividendos actuales, no se tiene signidicancia en los coeficientes y un R^2 muy bajo.

Code

```
##           [,1]                           [,2]
## statistic 0.813379                       0.9088529
## p.value   0.005487337                    0.1300349
## method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
## data.name "modelo2$residuals[, i]"      "modelo2$residuals[, i]"
```

Probando normalida en los residuales tenemos en forma separada existe normalidad solo para los residuales de la segunda respuesta.

Observamos que los residuales no esten correlacionados con el ajuste del modelo conjunto.

Code

	rent	house	COL
rent	0	0	0
house	0	0	0
COL	0	0	0

Que la suma de residuales sea cero.

Code

	.
rent	0
house	0
COL	0

Ahora para contrastar lo anterior, la regresión multivariada, aplicamos la preuba MANOVA

Code

```
##  Response current :
##              Df  Sum Sq Mean Sq F value Pr(>F)
## as.matrix(X)  3  5.4168 1.80559   1.821 0.2016
## Residuals    11 10.9072 0.99157
##
##  Response payout :
##              Df Sum Sq Mean Sq F value Pr(>F)
## as.matrix(X)  3 37.810 12.6035  4.6171 0.0252 *
## Residuals    11 30.027  2.7297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que solo en la primera respues, dividendos actuales, los factores no son significativos; sin embargo, para la segunda respuesta por lo son ligeramente.

Construimos los intervalos de confianza

Code

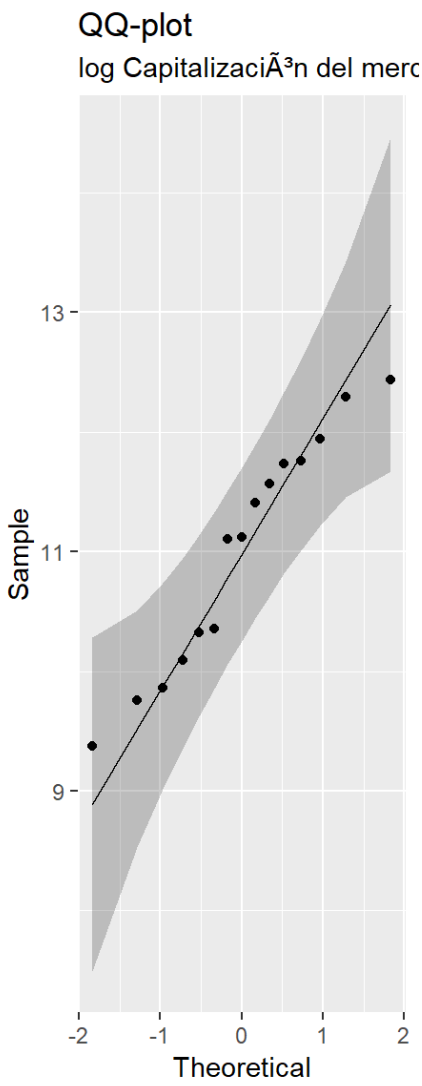
	2.5 %	97.5 %
rent:intercepto	-0.713	1.171
current:mktcap	0.000	0.000
current:netcash	0.000	0.000
current:chflow	0.000	0.000
payout:intercepto	4.339	7.465
payout:mktcap	0.000	0.000
payout:netcash	0.000	0.000
payout:chflow	0.000	0.001

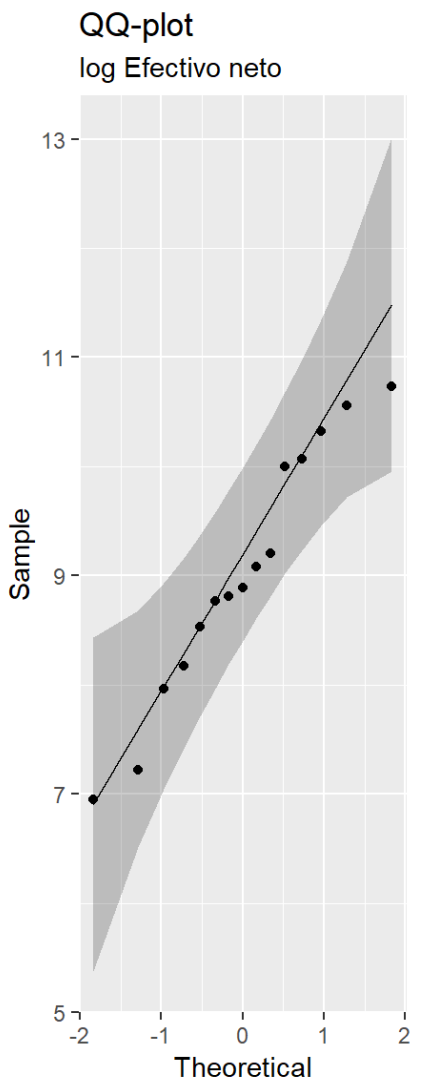
Ahora le aplicamos logaritmo al modelo multivariado con el fin de ver la mejora.

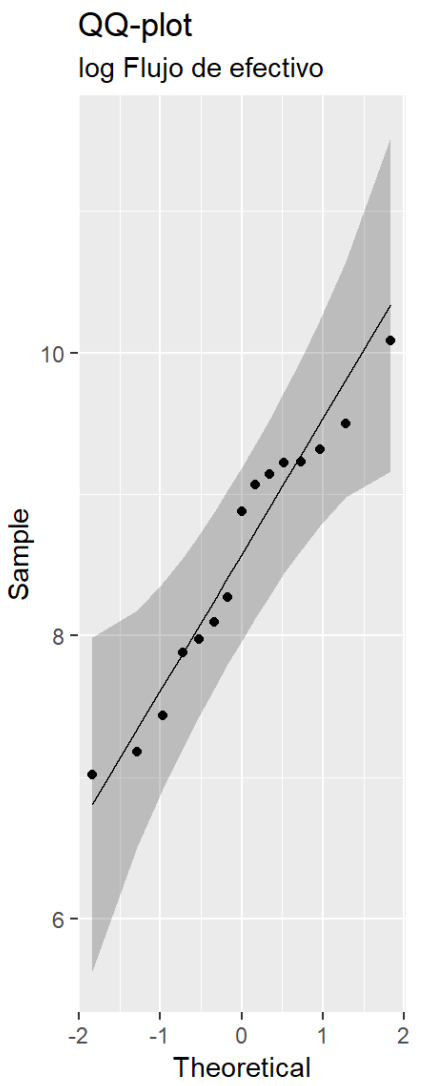
Code

Observamos gráficos para evaliar normalidad en la distribución de las variables

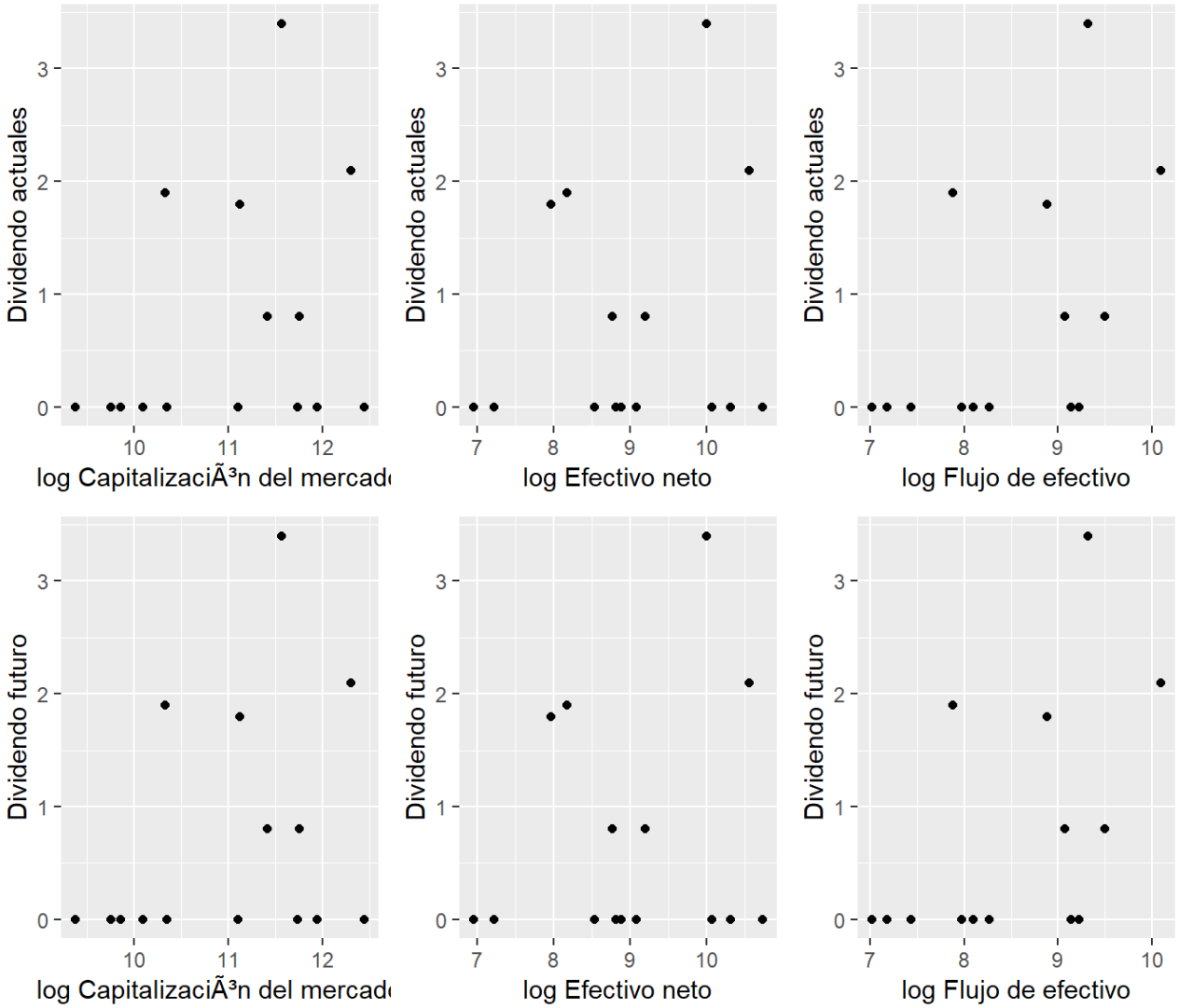
Code


QQ-plot
log CapitalizaciÃ³n del merc
Sample
Theoretical
13
11
9
-2
-1
0
1
2


QQ-plot
log Efectivo neto
Sample
Theoretical
13
11
9
7
5
-2
-1
0
1
2


QQ-plot
log Flujo de efectivo
Sample
Theoretical
10
8
6
-2
-1
0
1
2

Code



Realizamos el modelo con logaritmos. Nota: no se quitan outliers por que se tienen muy pocas observaciones.

Code

```
## Response current :
##
## Call:
## lm(formula = current ~ X$mktcap + X$netcash + X$chflow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9391 -0.6565 -0.2328  0.4266  2.2477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.1850     3.2570  -0.671    0.516
## X$mktcap     -0.4935     0.8027  -0.615    0.551
## X$netcash    -0.2256     0.4284  -0.527    0.609
## X$chflow      1.2126     0.7165   1.693    0.119
##
## Residual standard error: 1.025 on 11 degrees of freedom
## Multiple R-squared:  0.2927, Adjusted R-squared:  0.09978
## F-statistic: 1.517 on 3 and 11 DF,  p-value: 0.2646
##
##
## Response payout :
##
## Call:
## lm(formula = payout ~ X$mktcap + X$netcash + X$chflow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3862 -0.3311 -0.2036  0.3634  0.8635
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.57727    1.56002  11.908 1.26e-07 ***
## X$mktcap    -5.35360    0.38447 -13.925 2.49e-08 ***
## X$netcash    0.03467    0.20520   0.169    0.869
## X$chflow     5.33105    0.34316  15.535 7.88e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4907 on 11 degrees of freedom
## Multiple R-squared:  0.961,  Adjusted R-squared:  0.9503
## F-statistic: 90.24 on 3 and 11 DF,  p-value: 4.978e-08
```

El modelo de la respuesta 1, dividendos corrietes, no mejora, presenta coeficientes no significativos. En cambio, la respuesta 2, dividendos futuros, mejora notablemente, ocn coeficientes muy significatvos, pero no el efectivo neto. El mercado de capitales afecta deforma negativa al dividendo futuro, con camios de una unidad porcentual afectaría en $-5.35360 $ los dividendos futuros, e incrementos de una unidad porcentual en los flujos de efectivo, implica un cabio en las unidades de los dividendos futuros de $5.33105 $ aproximadamente. El valor del R^2 en el modelo individual mejora, pero en conjunto la primera respuesta no es adecuada.

Code

```
##           [,1]                            [,2]
## statistic 0.8750351                       0.8159974
## p.value   0.04003093                      0.005942202
## method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
## data.name "modelo2_log$residuals[, i]"  "modelo2_log$residuals[, i]"
```

Se tiene que los residuales no son normales e base al tes de normalidad para los residuales seprados.

Se realiza análisis de varianza con MANOVA como segunda añternativa para identificar si los fatores son significativos para las respuestas.

Code

```
##  Response current :
##              Df  Sum Sq Mean Sq F value Pr(>F)
## as.matrix(X)  3  4.7778  1.5926  1.5173 0.2646
## Residuals    11 11.5462  1.0496
##
##  Response payout :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## as.matrix(X)  3 65.188 21.7295  90.236 4.978e-08 ***
## Residuals    11  2.649  0.2408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vemos que el modelo con logaritmos en los factores solo es adecuado para la segunda respueta, pero no para la primera.

Construimos intervalose de confianza para el modelos en logaritmos.

Code

	2.5 %	**97.5 %**
rent:intercepto	-9.354	4.984

	2.5 %	97.5 %
current:mktcap	-2.260	1.273
current:netcash	-1.169	0.717
current:chflow	-0.364	2.790
payout:intercepto	15.144	22.011
payout:mktcap	-6.200	-4.507
payout:netcash	-0.417	0.486
payout:chflow	4.576	6.086

Con el modelo en logaritomos lo coeficientes para la respuesta de los dividendos actuales no eran significativos, y en todos los casos el flujo neto no presenta significancia en los modelos, esto se puede deber a multicolinealidad o que en realidad no explique nada en las respuestas.

Aplicamos la prueba de razón de verosimilitud bajo la hipótesis de $H_0 : \beta_2 = 0$; esto es que el flujo neto explica a las respuestas en su conjunto.

Code

```
## [1] 0.0759562
```

Code

$$\Lambda = \frac{|E|}{|E+H|} = 0.9162667$$

$$F_{3,46} = 4.102821$$

$$\Lambda \leq F_{3,46}$$

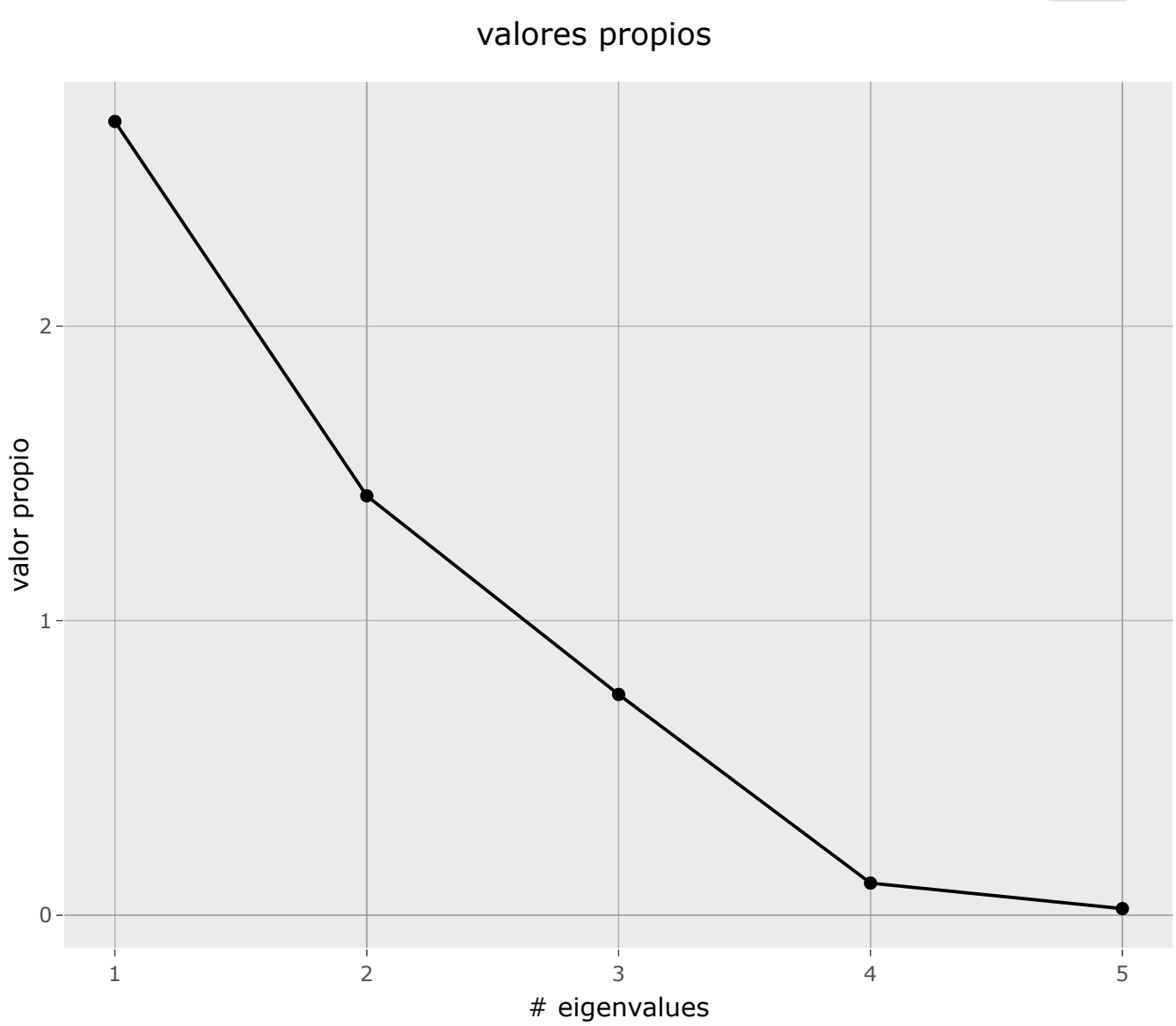
Podeos concluir que no se rechaza la hipótesis nula y, por lo tanto, el flujo neto no estaría presentando efecto en la respuestas “en conjunto”.

(b) También verifica el uso de otras variables explicativas basadas en funciones no lineales tales como la proporción entre el flujo de efectivo y la capitalización

Code

Vemos si existe multicolinealidad añadiendo a las covariables los porcentajesde capitalización y de flujos en efectivo.

Code



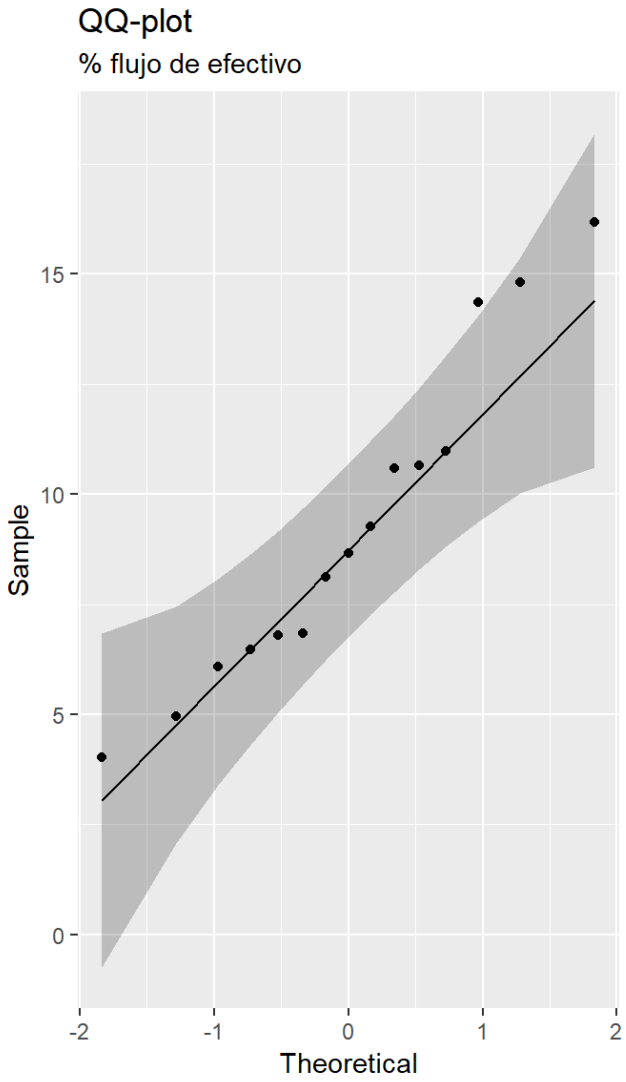
Detecta mucha colinealidad, esto debido a que los porcentajes se construyen con algunas covariables.

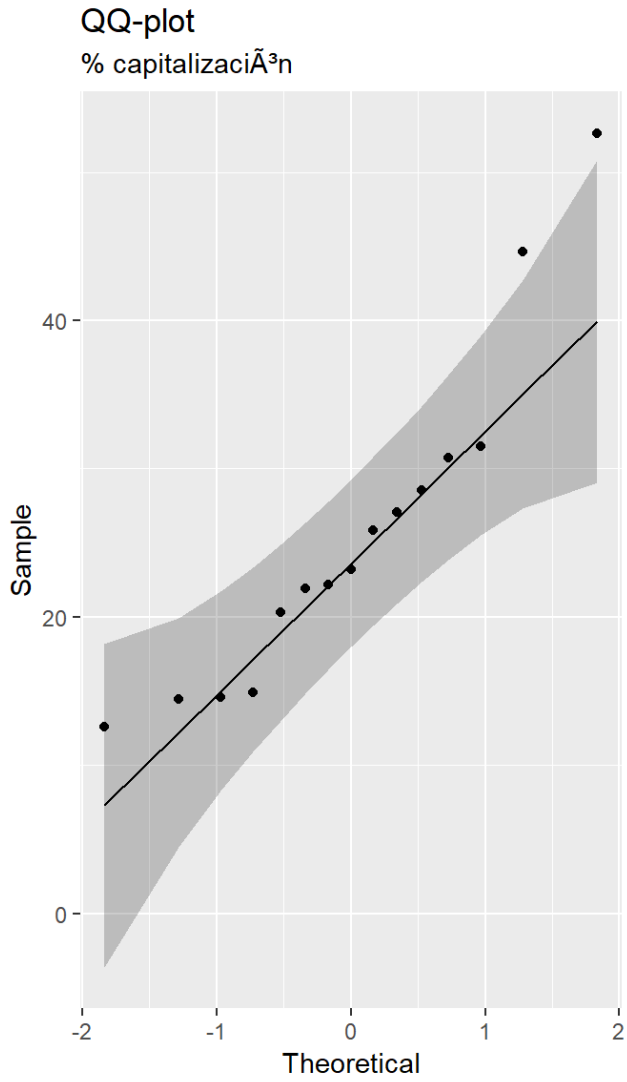
Code

```
## [1] 119.5909
```

Se utilizan las variables de dinero en logaritmo, las de porcentaje no. Precentamos las variables de los porcentajes

Code

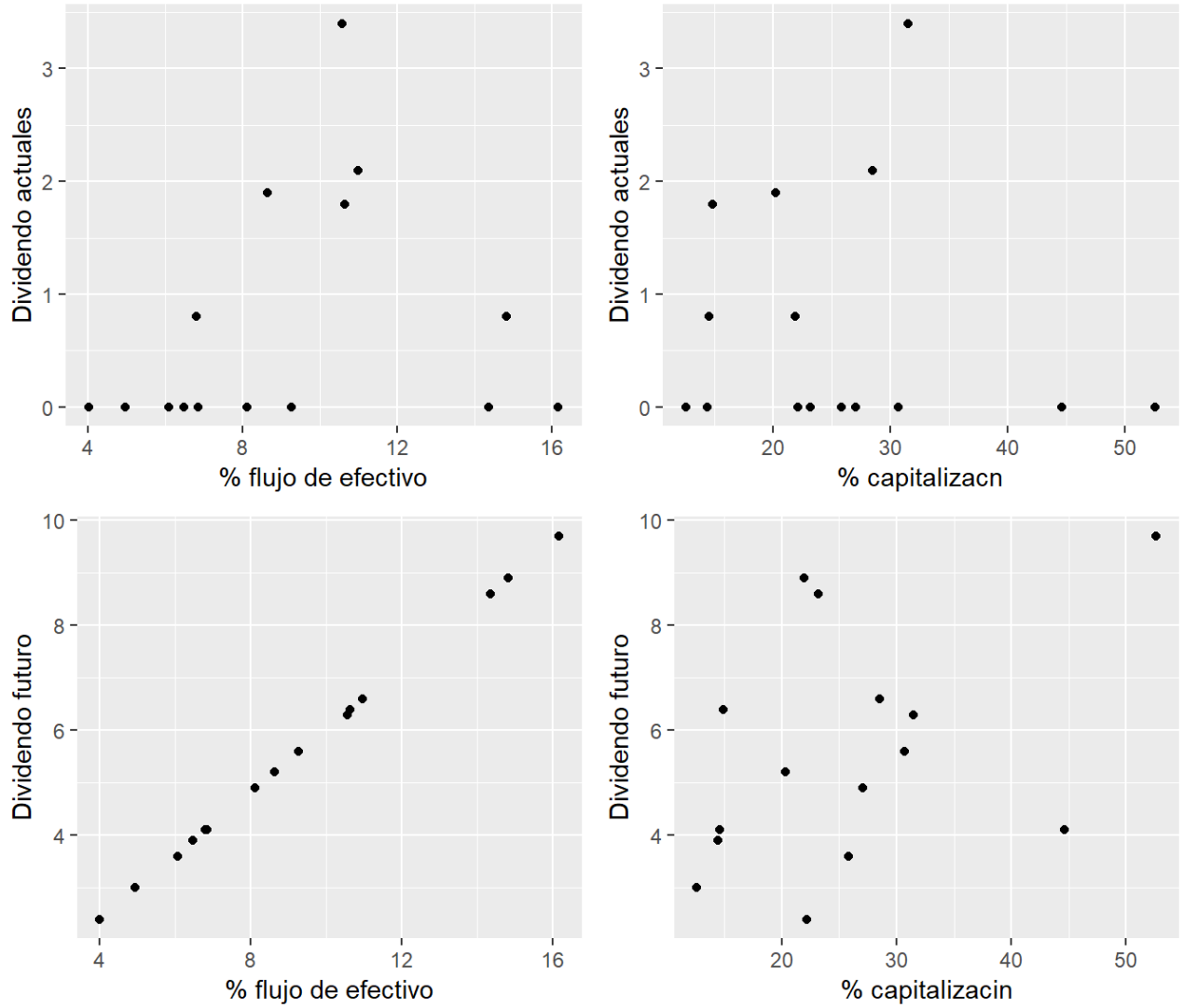




Se observan colas pesada.

Vemos la dispersión de los porcentajers respecto a las varriables respuestas.

Code



Aplicamos el modelo que considera a todas las variables

$$Y_i = Z_{ij}\beta_j$$

con β que incluye al mercado de capitales, el flujo neto, el flujo de efectivo y, el porcentaje de flujo de efectivo y capitalización,

Code

```
## Response current :
##
## Call:
## lm(formula = current ~ X$mktcap + X$netcash + X$chflow + X$prflow +
##     X$prpcap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36758 -0.40906 -0.00359  0.20118  1.88882
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.204723  12.204450   1.410   0.1922
## X$mktcap    -6.097395   3.475977  -1.754   0.1133
## X$netcash   -0.204983   1.530501  -0.134   0.8964
## X$chflow     6.814597   3.282266   2.076   0.0677 .
## X$prflow    -0.632132   0.391692  -1.614   0.1410
## X$prpcap     0.001863   0.097180   0.019   0.9851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9769 on 9 degrees of freedom
## Multiple R-squared:  0.4739, Adjusted R-squared:  0.1816
## F-statistic: 1.621 on 5 and 9 DF,  p-value: 0.249
##
##
## Response payout :
##
## Call:
## lm(formula = payout ~ X$mktcap + X$netcash + X$chflow + X$prflow +
##     X$prpcap)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044694 -0.008455  0.001263  0.014692  0.036507
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1656669  0.3638747   0.455    0.660
## X$mktcap    -0.0336504  0.1036360  -0.325    0.753
## X$netcash   -0.0159892  0.0456318  -0.350    0.734
## X$chflow     0.0465973  0.0978605   0.476    0.645
## X$prflow     0.5940110  0.0116783  50.865  2.2e-12 ***
## X$prpcap     0.0002939  0.0028974   0.101    0.921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.02913 on 9 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9998
## F-statistic: 1.599e+04 on 5 and 9 DF,  p-value: < 2.2e-16
```

Los ajustes en el modelo son malos, asimismo, se juega con varias combinaciones.

En este modelo solo utilizamos los porcentajes como covariables, ya que utilizar las otras presentan mayor colinealidad, ya que estos porcentajes se construyen con esas variables.

Code

```
## Response current :
##
## Call:
## lm(formula = current ~ X$prflow + X$prpcap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2349 -0.6115 -0.4291  0.3963  2.6837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.42695    0.90184   0.473    0.644
## X$prflow     0.09047    0.08922   1.014    0.331
## X$prpcap    -0.02120    0.02927  -0.724    0.483
##
## Residual standard error: 1.115 on 12 degrees of freedom
## Multiple R-squared:  0.08672,    Adjusted R-squared:  -0.06549
## F-statistic: 0.5698 on 2 and 12 DF,  p-value: 0.5802
##
##
## Response payout :
##
## Call:
## lm(formula = payout ~ X$prflow + X$prpcap)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.048002 -0.006511  0.004183  0.011712  0.039617
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0141361  0.0210955   0.670    0.515
## X$prflow     0.6005080  0.0020869 287.750   <2e-16 ***
## X$prpcap    -0.0006665  0.0006846  -0.974    0.350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02607 on 12 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 4.989e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

El modelo de la respuesta uno no tiene coeficientes significativos e indica un mal ajuste. El modelo de la repsuesta dos, solocuenta con el porcentaje de flujos de efectivos como significativo, y con un R^2 alto.

En conclusión, las covariables mercado de capital, flujo de efectivo, el capital neto, porcentaje de flujo neto, y porcentaje de capitalización, no explican adecuadamente a los dividendo actuales y futuros, esto por la alta multicolinealidad entre las covariables, que hace que se afecten entre si, haciendo que no sean significativos. A su vez, realizar regresión con logaritmos y con respuestas separadas, puede que mejore un poco en la explicaciónd de algunas variables a cada repsuesta.

Alternativas Podemos realizar PCA a la matriz de corralación de las observaciones, y utilizar las proyecciónes para construir una base ortogonal y tener covariables ortogonales, capaces de mejorar la explicación del modelo.

Utilizamos PCA con tres componentes ya que son los que explican aproximadamente el 80% de la varianza.

Code

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3      Comp.4
## Standard deviation     1.651321 1.2051961 0.8977436 0.107147980
## Proportion of Variance 0.545372 0.2904995 0.1611887 0.002296138
## Cumulative Proportion  0.545372 0.8358715 0.9970603 0.999356411
##                             Comp.5
## Standard deviation     0.0567269428
## Proportion of Variance 0.0006435892
## Cumulative Proportion  1.0000000000
```

Code

Realizamos el modelo con las proyeciones de la componete una y dos.

Code

```
## Response current :
##
## Call:
## lm(formula = current ~ X_scores[, 1] + X_scores[, 2] + X_scores[,
##     3])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9044 -0.6991 -0.3586  0.5062  2.3366
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.7200     0.2734   2.633   0.0233 *
## X_scores[, 1]   0.1864     0.1656   1.126   0.2843
## X_scores[, 2]  -0.1158     0.2269  -0.510   0.6198
## X_scores[, 3]   0.4338     0.3046   1.424   0.1821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.059 on 11 degrees of freedom
## Multiple R-squared:  0.2443, Adjusted R-squared:  0.03821
## F-statistic: 1.185 on 3 and 11 DF,  p-value: 0.3599
##
##
## Response payout :
##
## Call:
## lm(formula = payout ~ X_scores[, 1] + X_scores[, 2] + X_scores[,
##     3])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18974 -0.03550  0.02198  0.03510  0.10993
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.55333    0.02272  244.39  < 2e-16 ***
## X_scores[, 1] -0.39374    0.01376  -28.61 1.12e-11 ***
## X_scores[, 2] -1.33548    0.01885  -70.83 5.52e-16 ***
## X_scores[, 3]  1.36585    0.02531   53.96 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08801 on 11 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9984
## F-statistic:  2916 on 3 and 11 DF,  p-value: 3.124e-16
```

Observamos que la respuesta uno tiene el intecepto significativos; no obstante, los coeficiendes de la componente uno y dos no los son, su ajuste no es el adecuado, pero esto ya se explica dado a que muchas de las observacioenes del dividendo actual tienen valores cero. Por otra parte la respuesta de los dividendos actuales mejora de forma significativa, con sus coeficientes significativos y una $R2$ ceracana a uno.

Probamos normalidad en los residuales. Concluyendo que a maner aindividual los residuales son normales.

Code

```
##           [,1]                              [,2]
## statistic 0.8519828                         0.90586
## p.value   0.01852766                        0.1169986
## method    "Shapiro-Wilk normality test"     "Shapiro-Wilk normality test"
## data.name "modelo3_scores$residuals[, i]" "modelo3_scores$residuals[,
i]"
```

Contrastamos la regresión con MANOVA para ver si los factores en este modelo son sifnificativos.

Code

```
##  Response current :
##                              Df Sum Sq Mean Sq F value Pr(>F)
## as.matrix(X_scores[, (1:3)])  3  3.988  1.3293  1.1854 0.3599
## Residuals                    11 12.336  1.1215
##
##  Response payout :
##                              Df Sum Sq Mean Sq F value    Pr(>F)
## as.matrix(X_scores[, (1:3)])  3 67.752 22.5840  2915.8 3.124e-16 ***
## Residuals                    11  0.085  0.0077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Concluimos que los scores son significativos solo para la segunda respuesta pero no para el modelo en conjunto.

La dificultad en PCR, es la interpretación de los coeficientes, para esto realizamos la rotación de los coeficientes con el vector propio de los tres primeros componentes.

$$\beta_{ols} = V\beta pcr$$

Code

	current	**payout**

	current	payout
mktcap	0.1490122	-0.1991428
netcash	-0.0323161	-0.4326074
chflow	0.2950863	0.5612501
prflow	0.3162275	1.7877323
prpcap	-0.1612193	0.2574486

En conclusión los modelos presentan multicolinealidad en las covariables que hacen que el modelo se desempeñe mal, esto utilizando las series a niveles y en logaritmos. Por otro lado, las variable de los dividendos contiene muchos valores ceros, los cuales no se retiran por se pocas observaciones, y por se lo que reportan las compañias, pero generando problemas, lo que se puede hacer es una regresión al origen, pero afecta a la otra respuesta; es por eso que estas respuestas se modelarían meojor de forma separada. Por último, otra alternativa fue PCR, en cual trajó modelos mejores a los anteriores; no obstante los dividendos actuales caudan ruido en la regresión.