Jamie Brabston
Matt Caulfield
Mark Testa
December 2, 2008
MA 331A
We pledge our honor that we have abided by the Stevens Honor System.

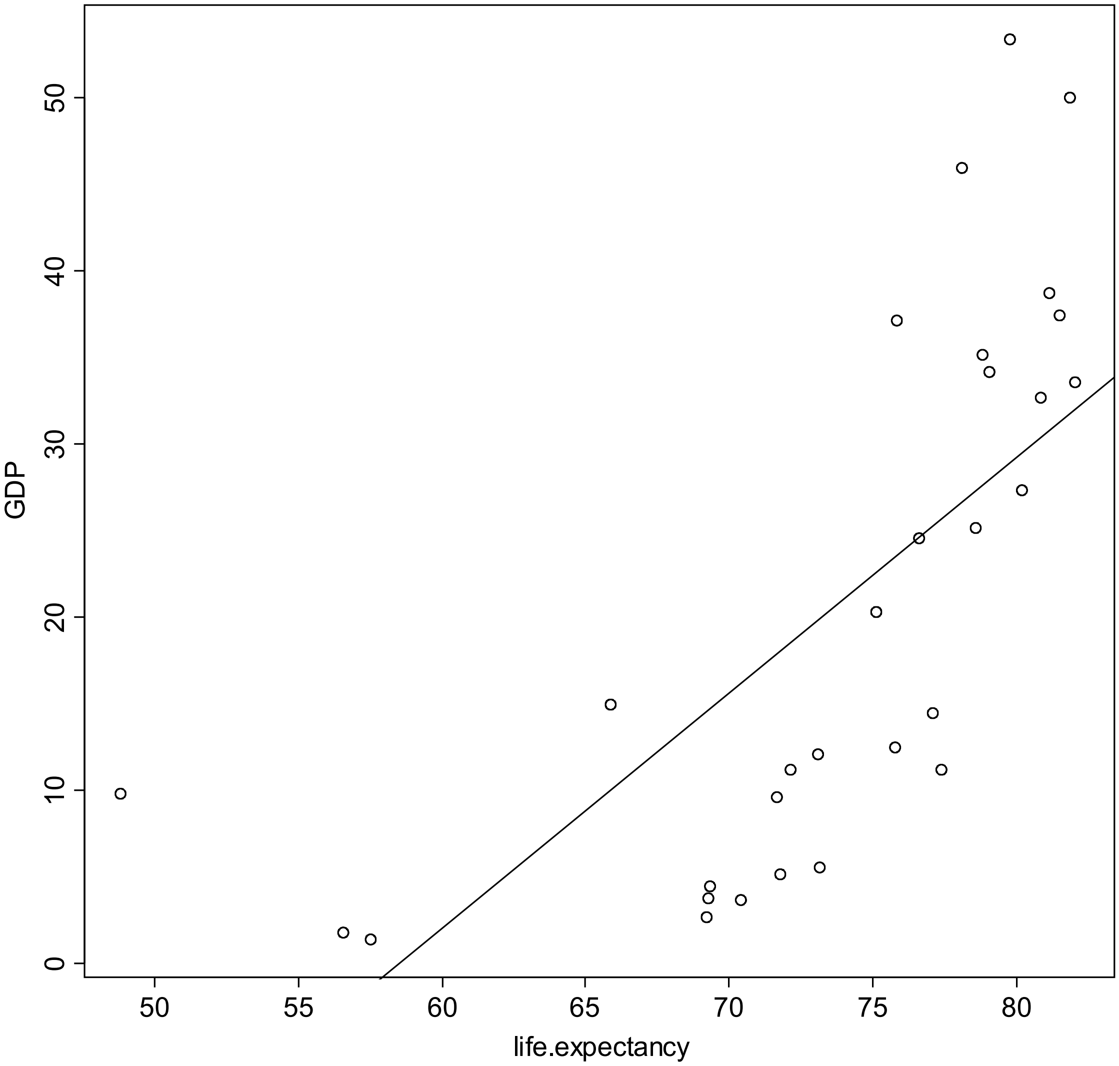
The Wealth of Nations

We did a study that attempted to identify some of the economic, governmental, and demographic factors that affect or are affected by how wealthy a country is. We are looking at thirty countries from around the world, selected for diversity in culture, location, and economic status. For each country, we have obtained from the CIA World Factbook economic data such as national debt, oil imports and exports, and unemployment rate, and demographic data such as life expectancy, population density, and literacy rate. We performed a multivariate regression analysis to determine the correlation between these and other variables, and GDP per capita. Once we had identified significant variables and their relationship to GDP, we attempted to identify the causal relationship between the variable and GDP: whether one causes the other, or if both are caused by a lurking variable. We hoped to identify relationships between economics, government, and demographics that are not immediately obvious.

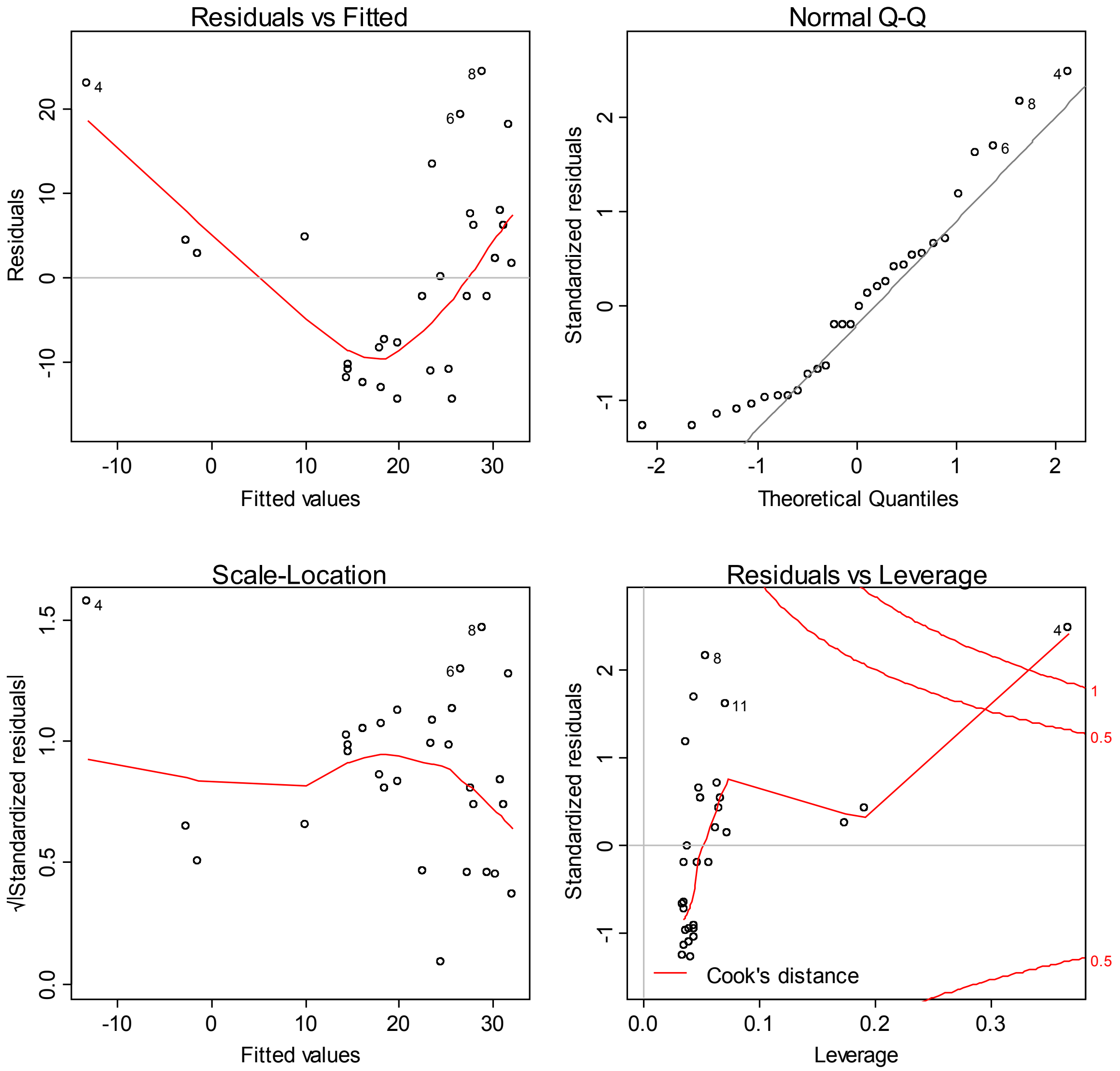
For this study, we made use of many of the multivariate regression techniques learned in this class. The first step of the study was be to identify which variables significantly affect GDP, and which do not, and to identify and remove any suspected outliers that could harm the analysis. We used stepwise and multiple regressions to find the best possible set of variables. We then did a partial simple linear regression using each variable individually, checking for normality, independence, and constant variance of the residuals. In the case of nonlinearity, we used a Box-Cox transform on the explanatory variable.

Regression of Each Variable Compared To GDP Per Capita

Life Expectancy



There is a perceptible positive correlation between life expectancy and GDP, which is not surprising. However, the data appears to be significantly nonlinear.



Residuals vs Fitted

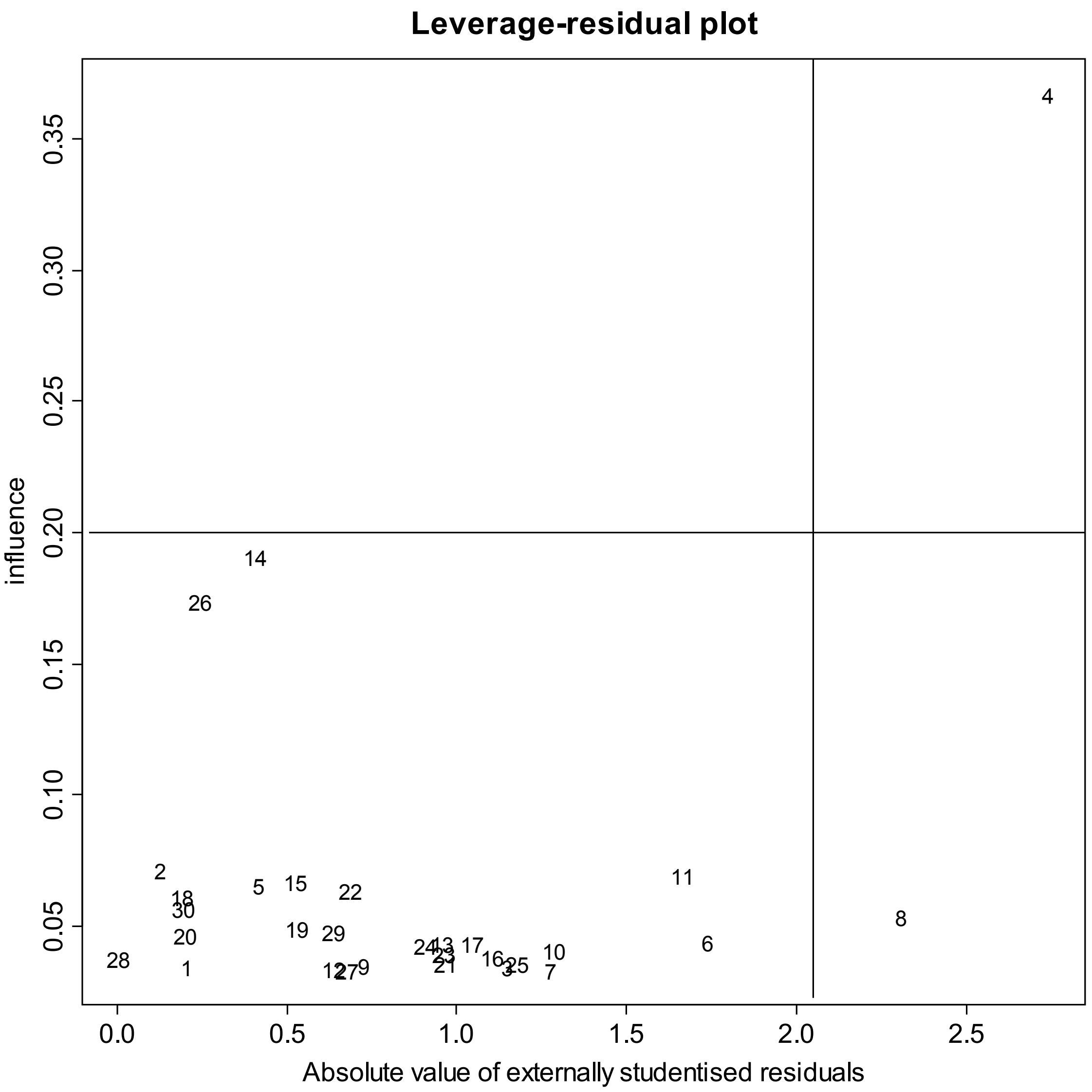
Normal Q-Q

Scale-Location

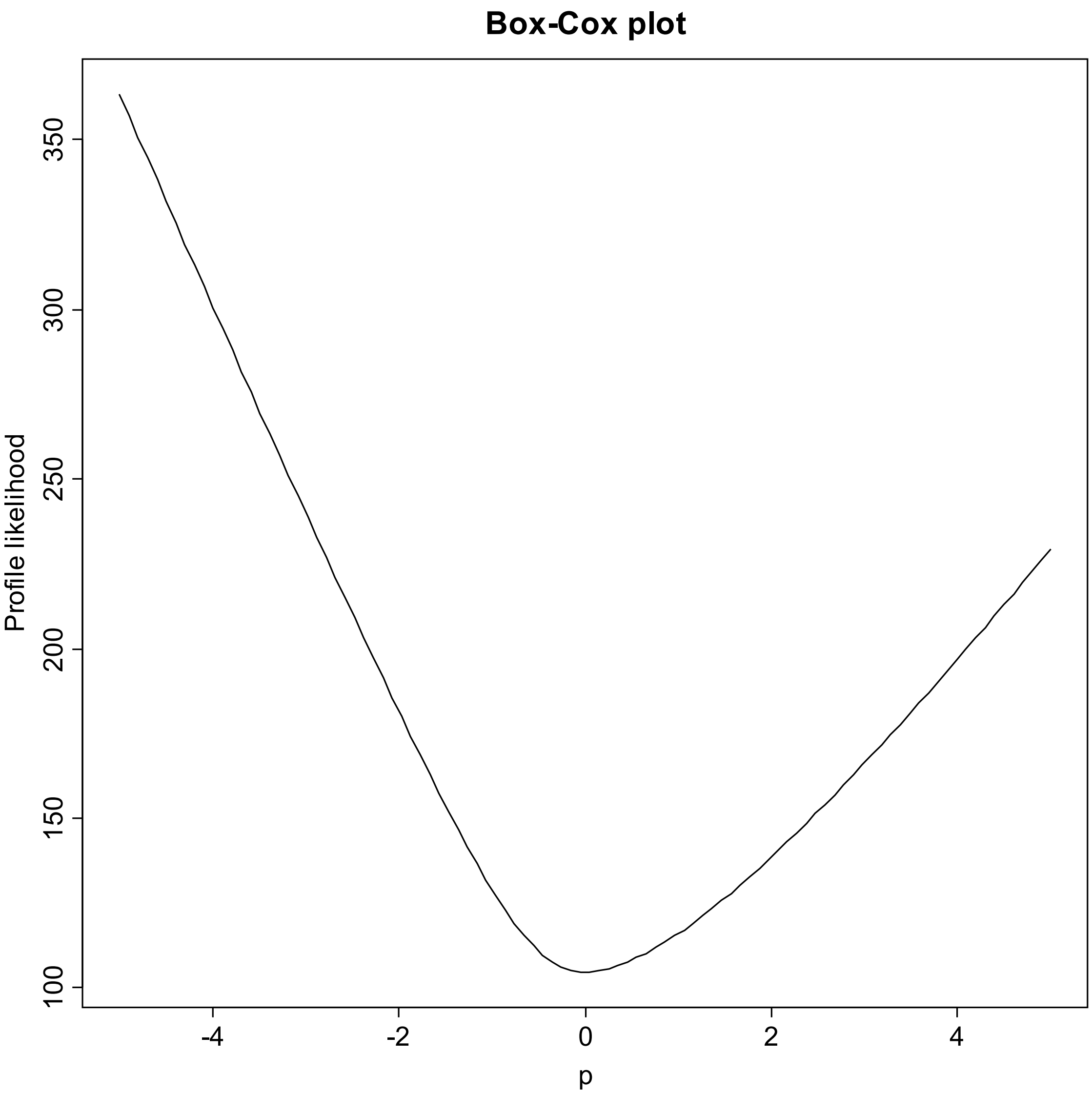
Residuals vs Leverage

The upper-left plot, which depicts the residuals as a function of the fitted values, would be a horizontal line if the data were linear. Obviously, this is not the case – the nonlinearity is a cause for concern.

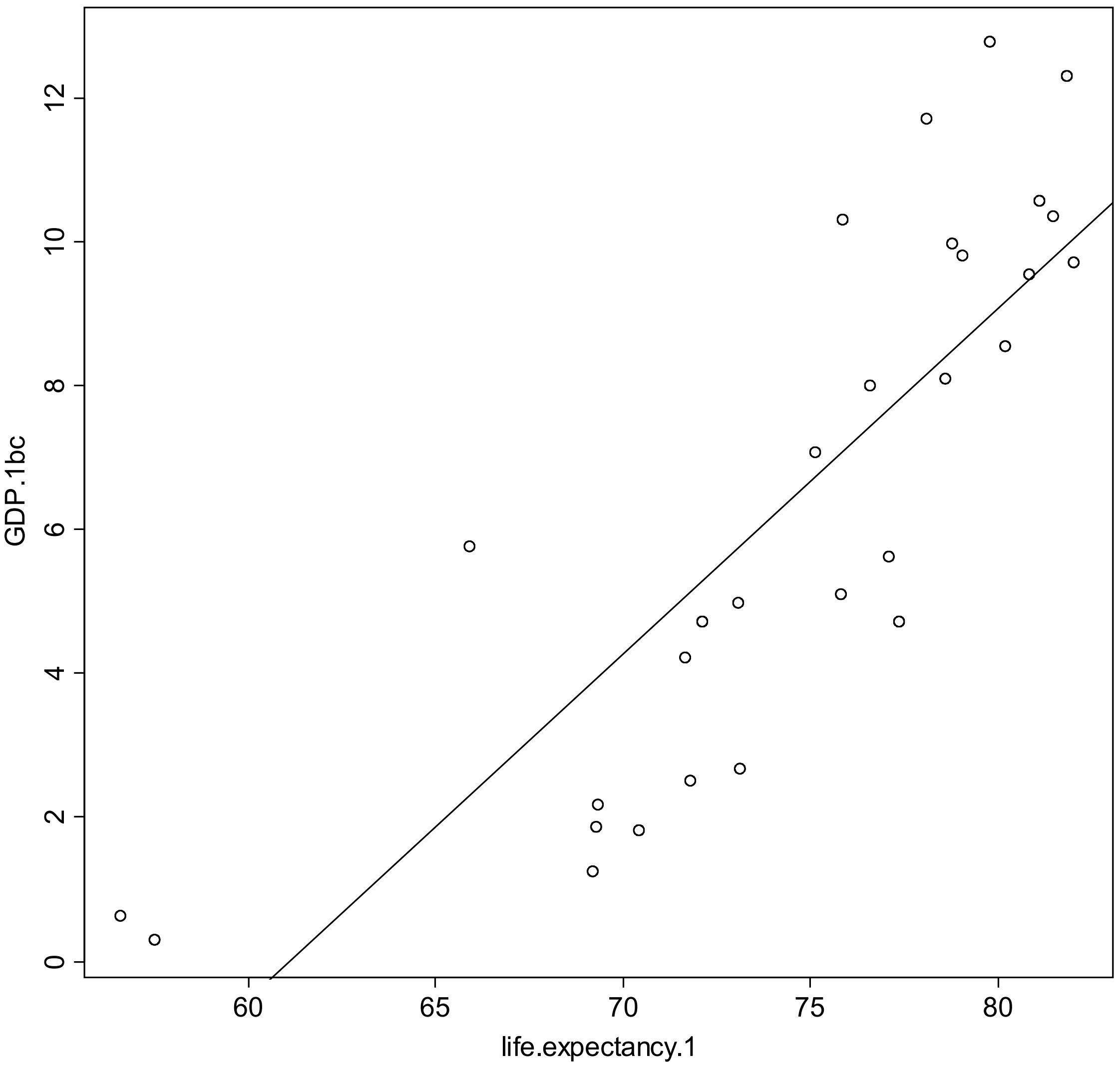
The upper-left plot is a Q-Q quantile plot of the residuals. There is some deviation from normality, but in the middle of the plot the residuals are normally distributed, which is encouraging.



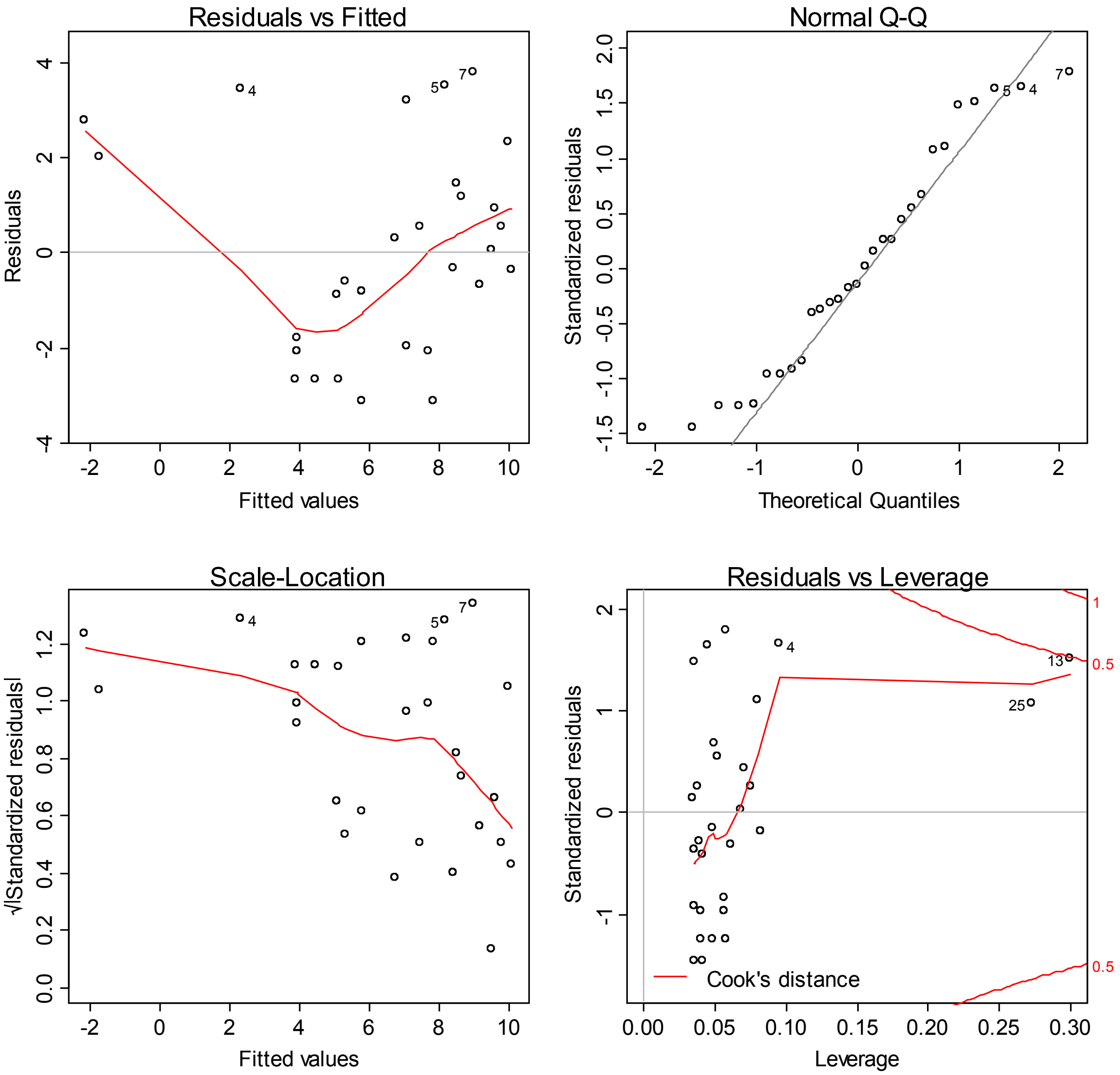
This leverage-residual plot sorts the data points according to suspected outliers and suspected non-outliers, on the right and left respectively, and influential points and non influential points on the top and bottom, respectively. We want to eliminate influential outliers, which means we will remove the point labeled (4) in the top right.



One way to fix the nonlinearity of the residuals is to perform a Box-Cox transform on the response variable. A Box-Cox plot essentially replaces the linear relationship with a power law relationship. The plot above is a measure of the goodness of fit as a function of p, the power in the power law. The minimum point on this curve gives the p which will give the best fit. In this case, the optimal p appears to be very close to 0.

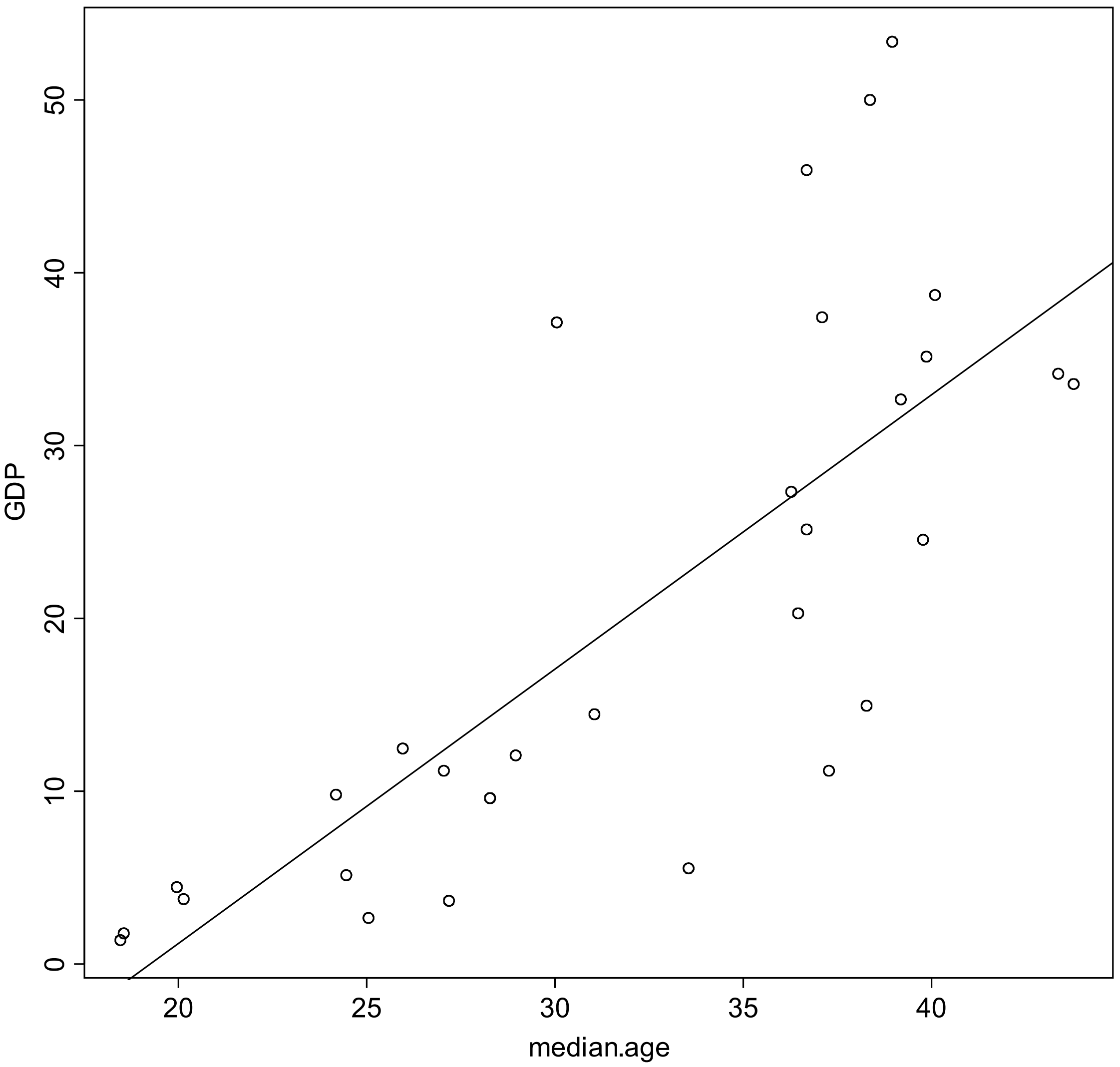


This is a plot of the Box-Cox transformed GDP as a function of life expectancy. It appears as if the nonlinearity has been slightly corrected.

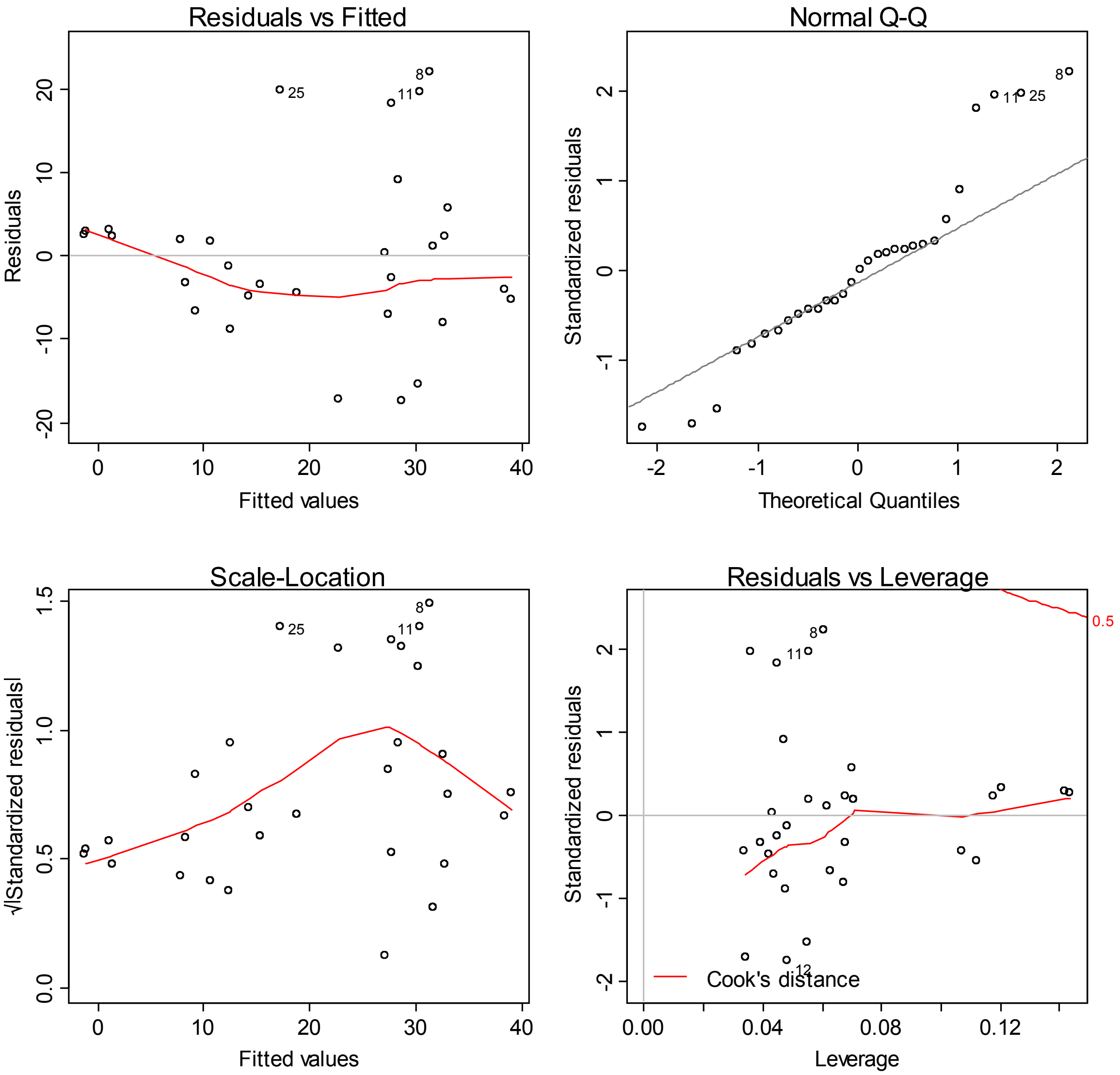


The Residuals vs. Fitted Values plot, however, still indicates significant nonlinearity. We can conclude that life expectancy has a positive relationship with GDP, but we cannot perform linear regression and expect to get good results. This means we cannot make precise predictions about the GDP of a test country based on its life expectancy.

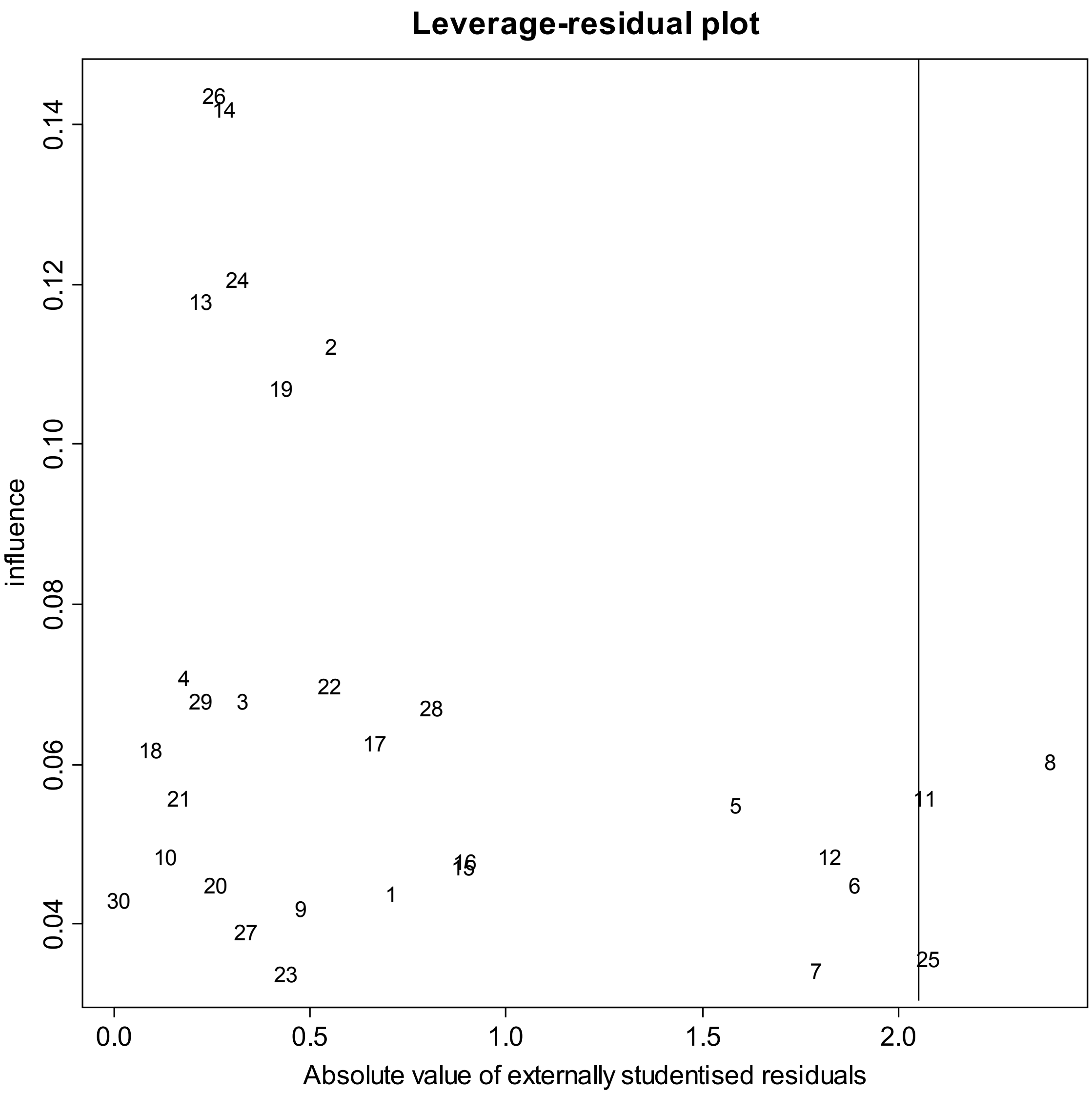
Median Age



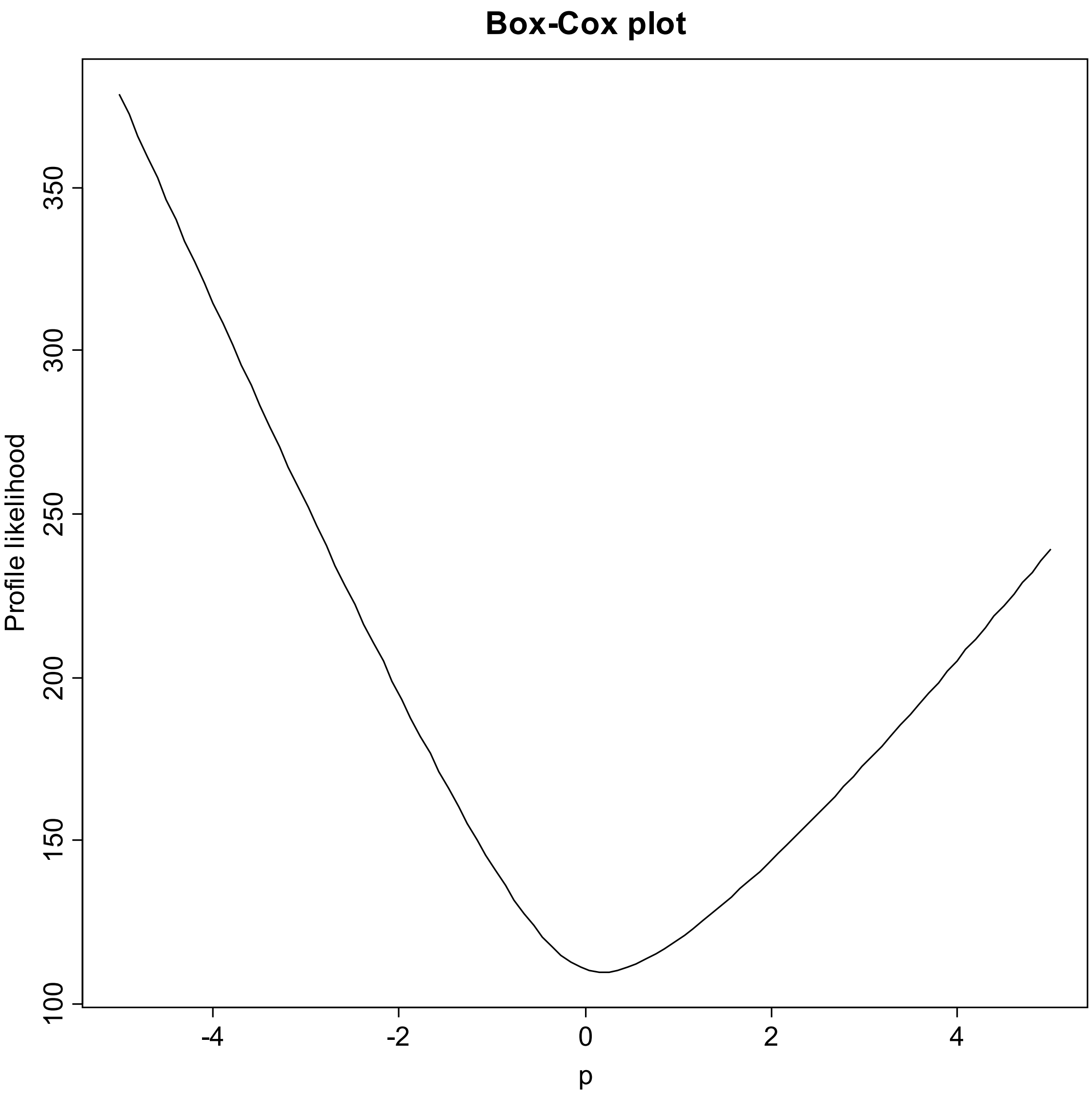
This is a simple linear regression plot of GDP as a function of median age.



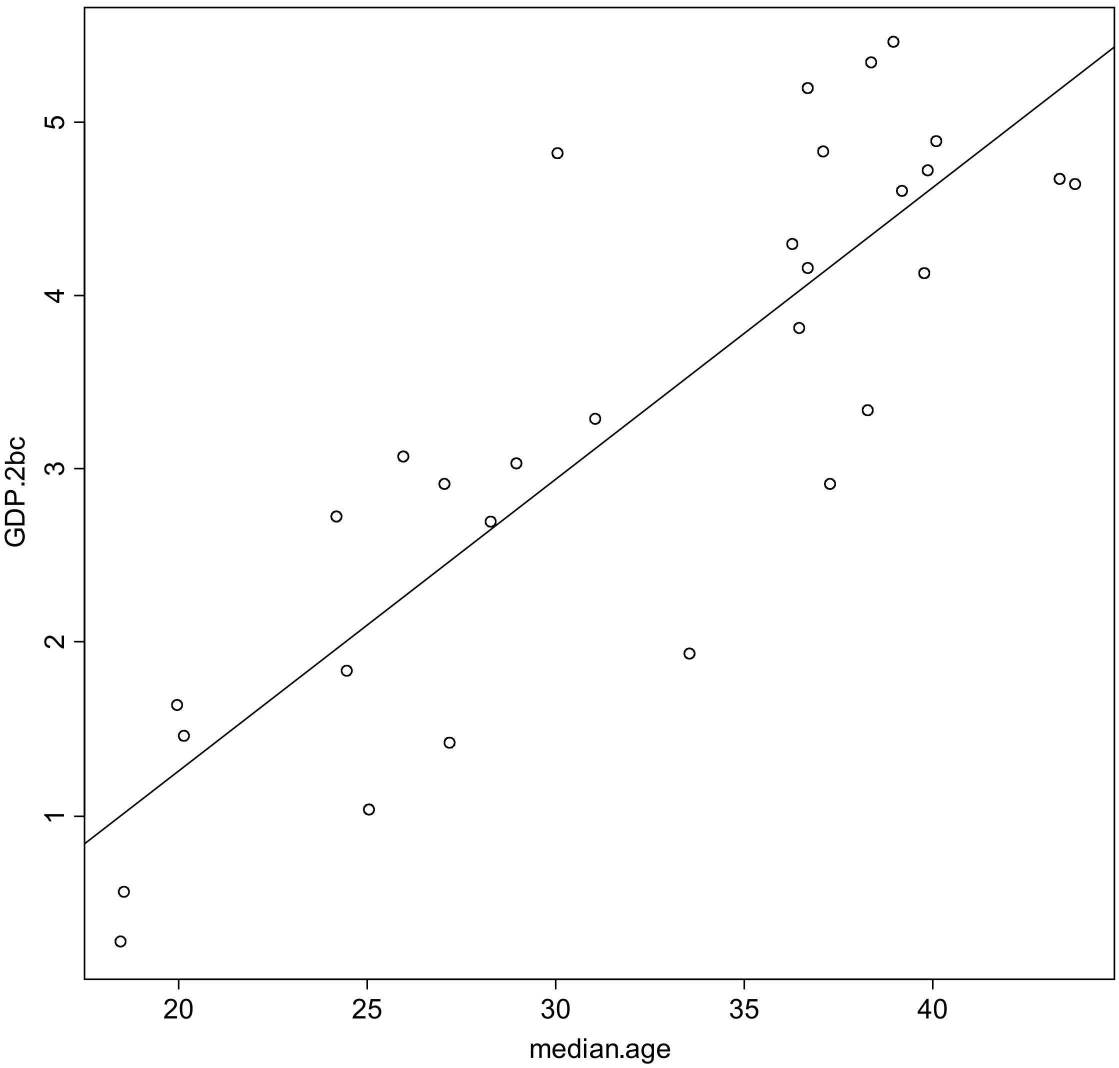
The residuals appear to be approximately linear, which is encouraging. However, their deviation from normality is significant, so we will try a Box-Cox transform to correct this.



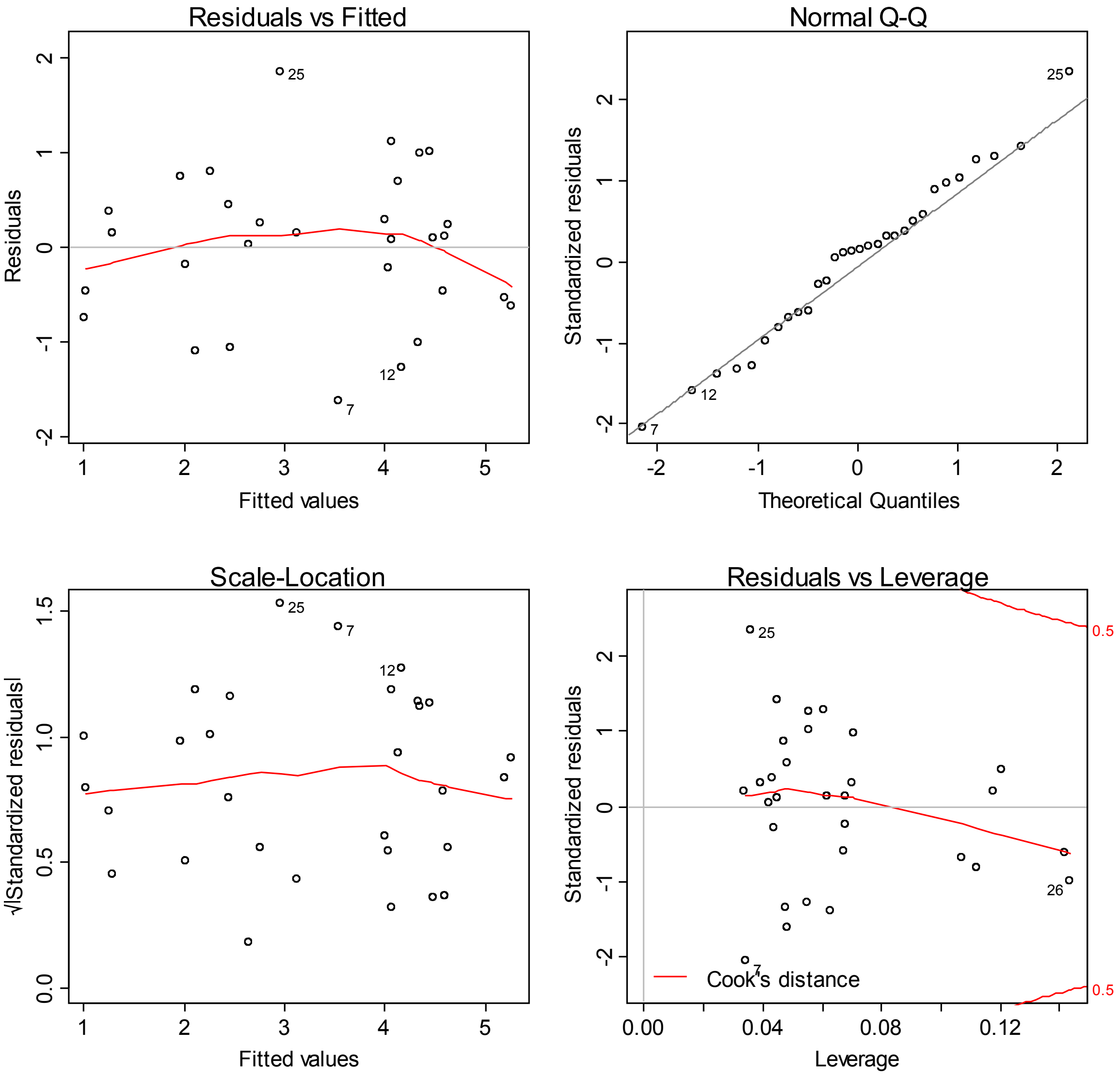
Before manipulating the data further, we checked for outliers. This plot indicates that there are no preeminently influential data points, so outliers are not a major concern. We did not remove any data points for this analysis.



This Box-Cox plot again gives a relatively small p.



This is the Box-Cox transformed scatter plot.

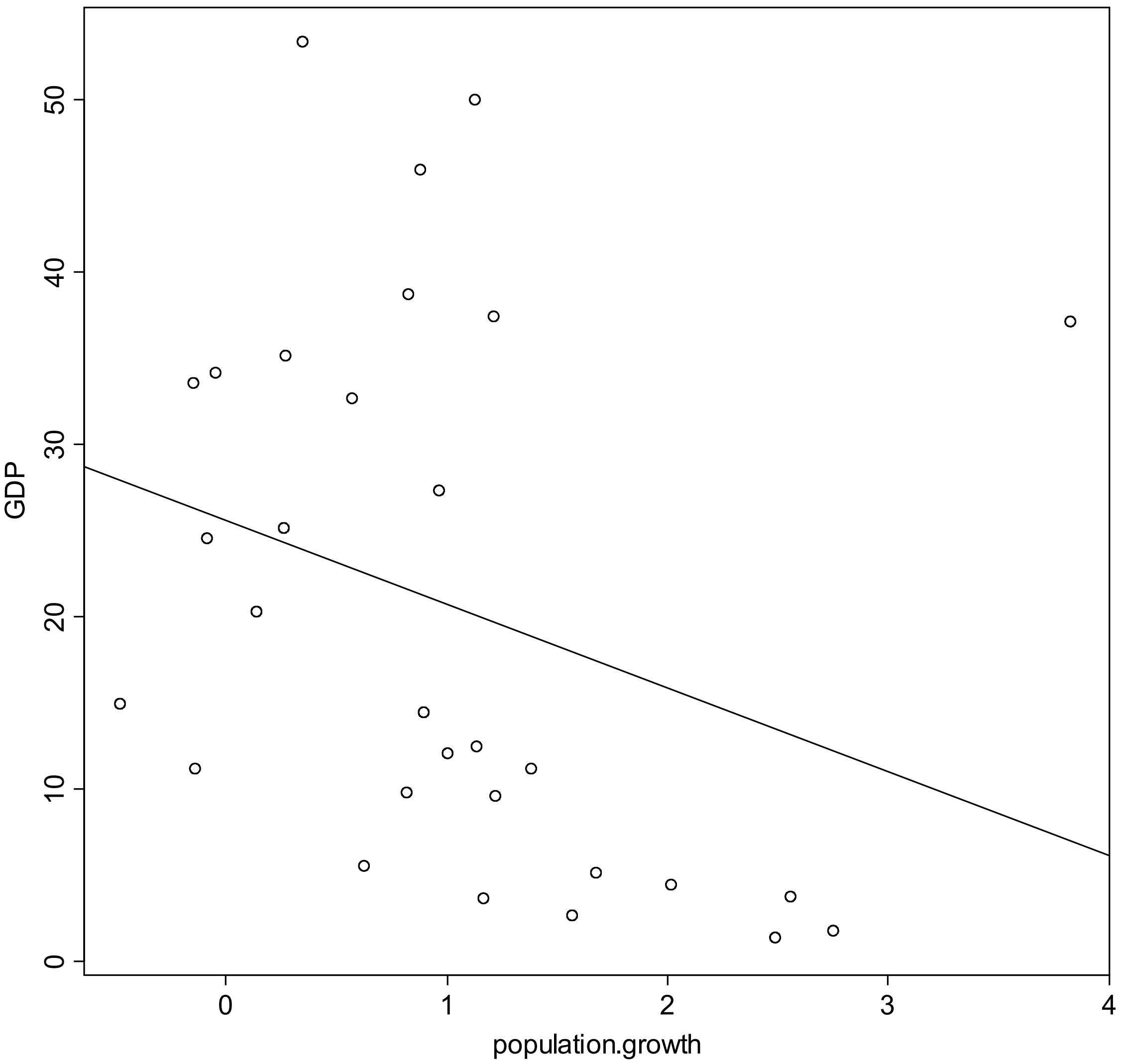


Linearity is not too badly affected, and normality is significantly better after the Box-Cox transform. We can conclude that an appropriate model is:

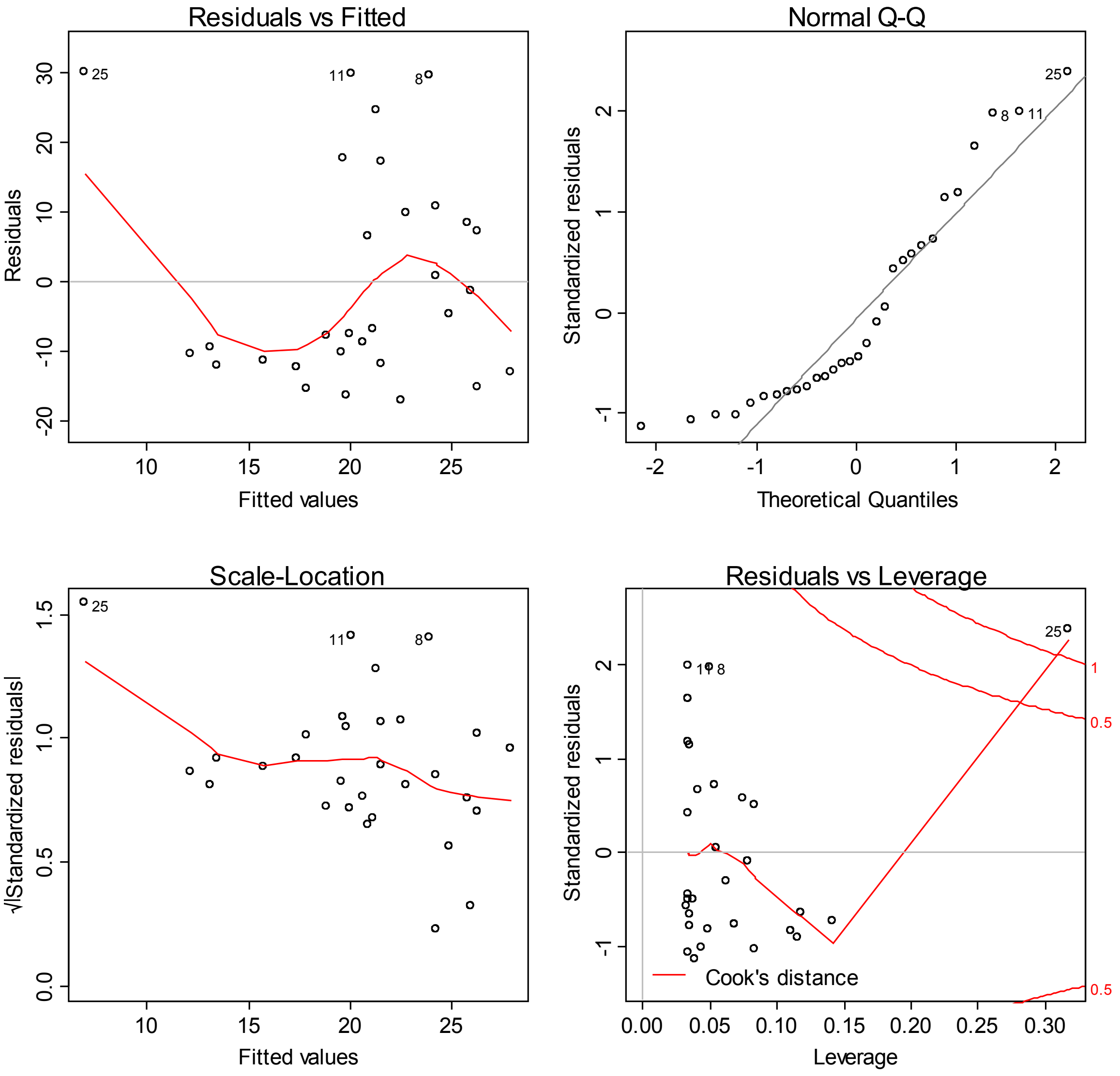
$((\text{GDP})^{0.15} – 1)/0.15 = -2.1 + 0.17(\text{Median Age})$

For this regression the R^2 is 0.72, and the p-value is so small as to pass any reasonable significance test.

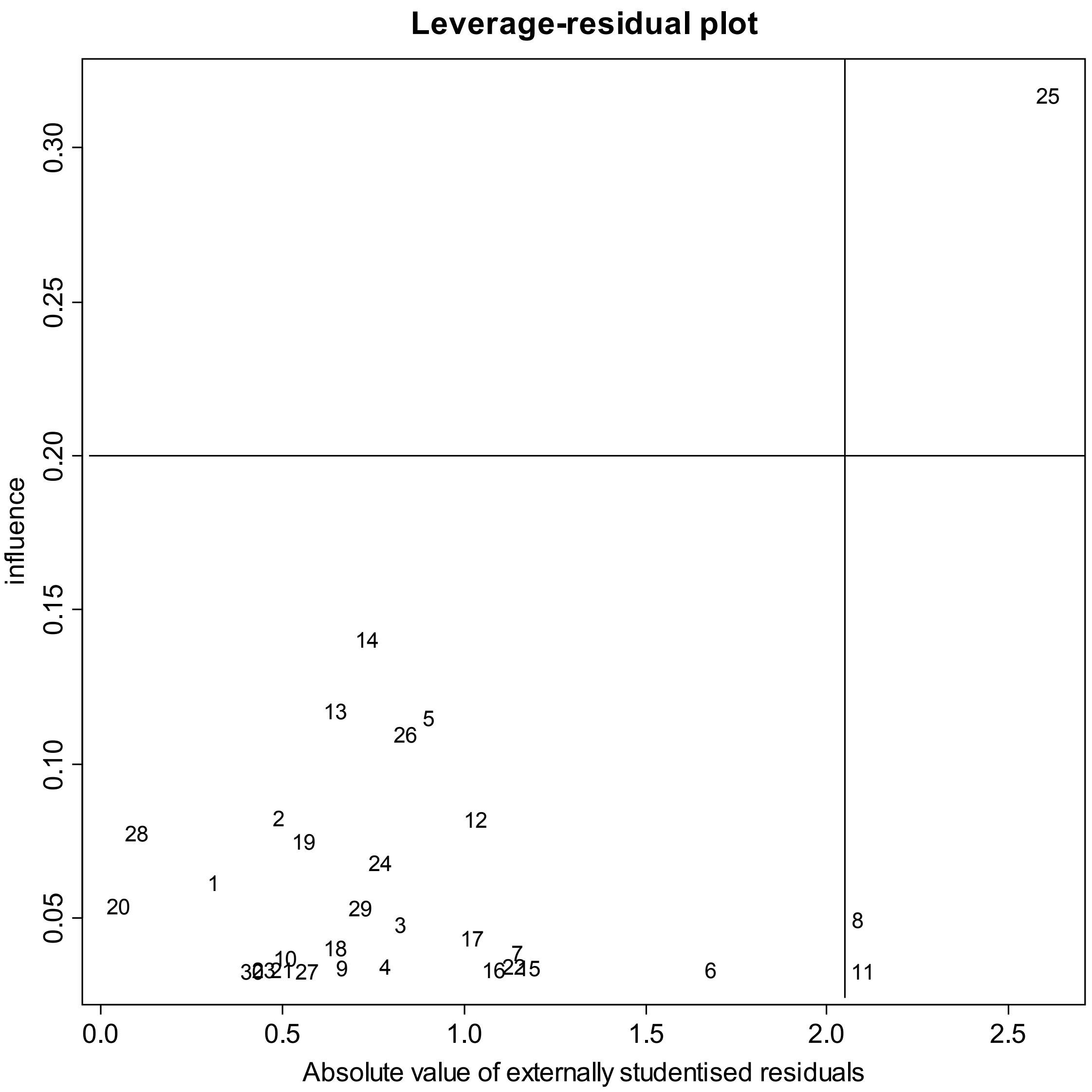
Population Growth



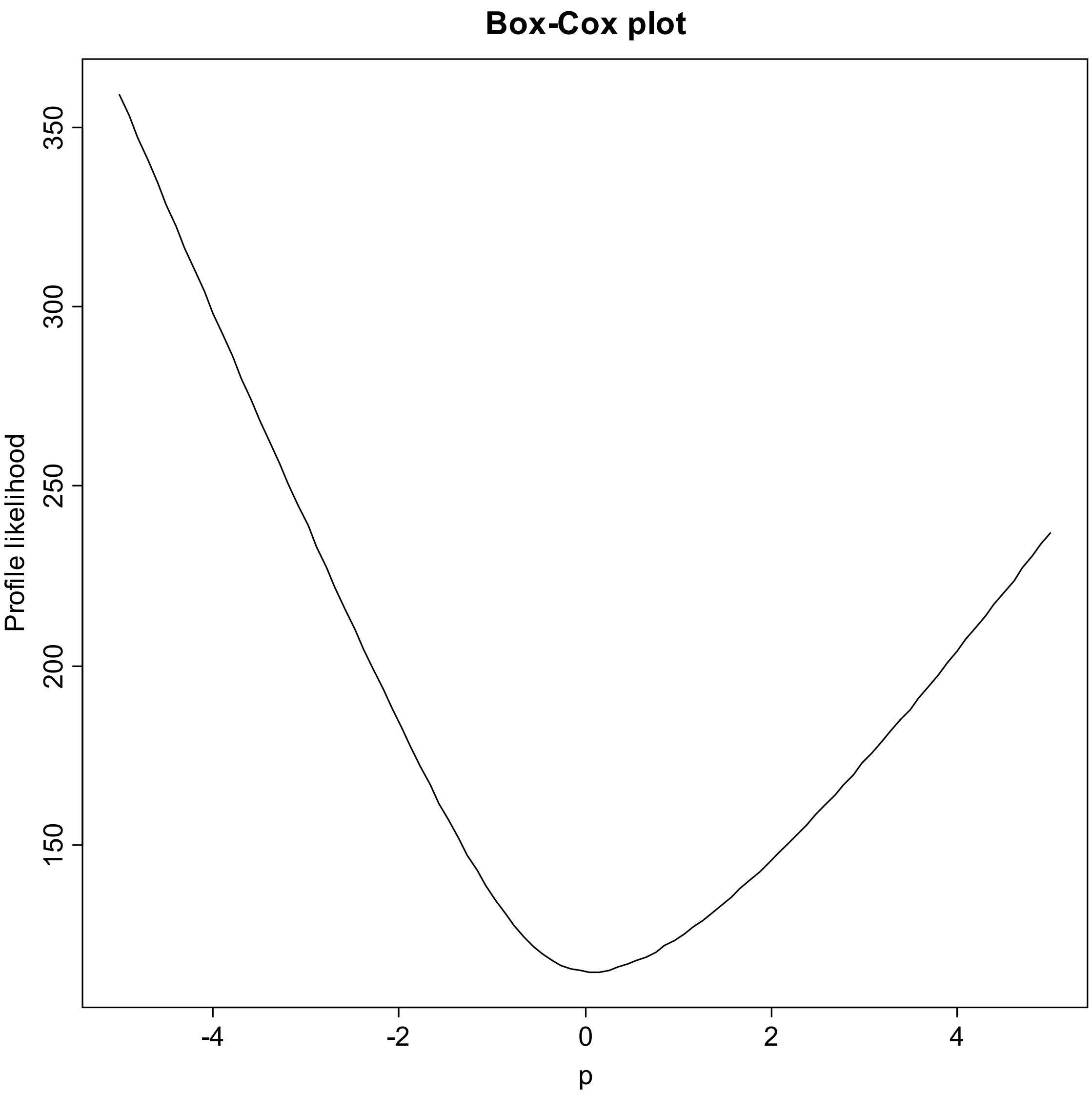
This is a plot of GDP as a function of population growth. It appears that there is a lot of noise in the data, and a few significant outliers.



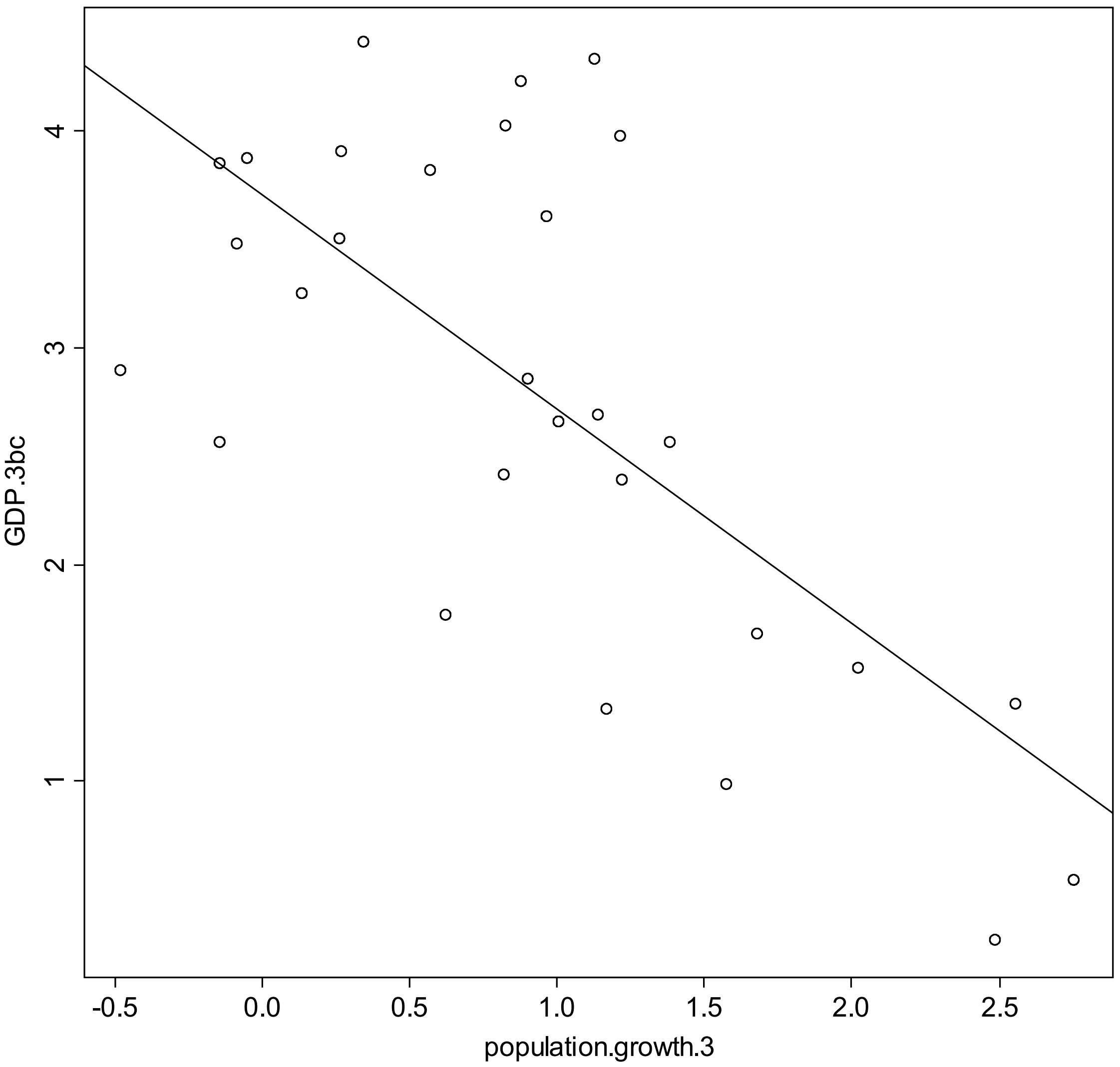
The Residuals vs. Fitted values plot indicates that the data is unacceptably nonlinear. A Box-Cox transform will be attempted to linearize the regression.



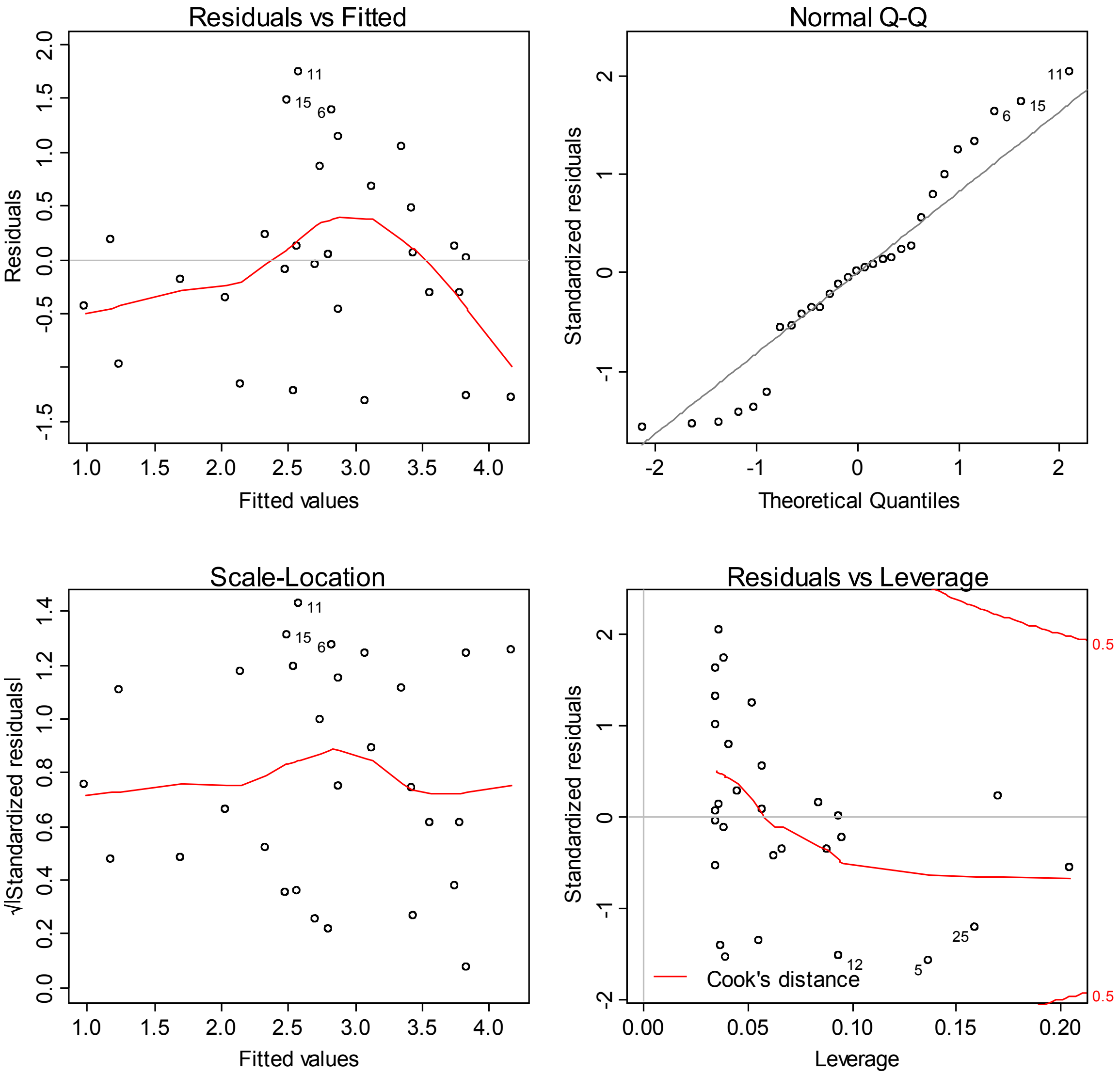
Again, we check for outliers. Point (25) is eliminated.



Box-Cox plot again gives a small value for the optimal p.

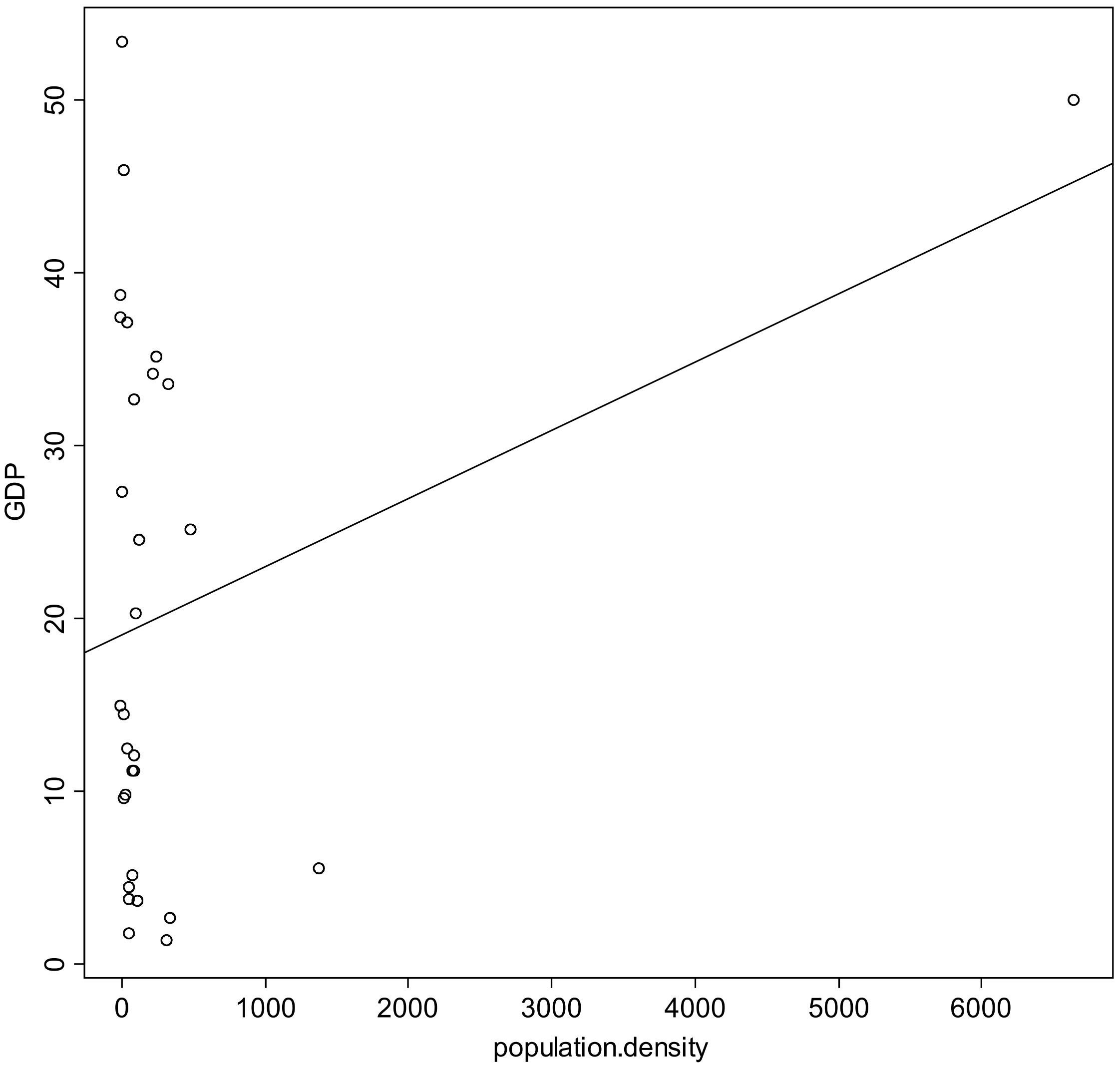


This is the Box-Cox transformed scatter plot, with no outlier. The linearity appears to be slightly improved.

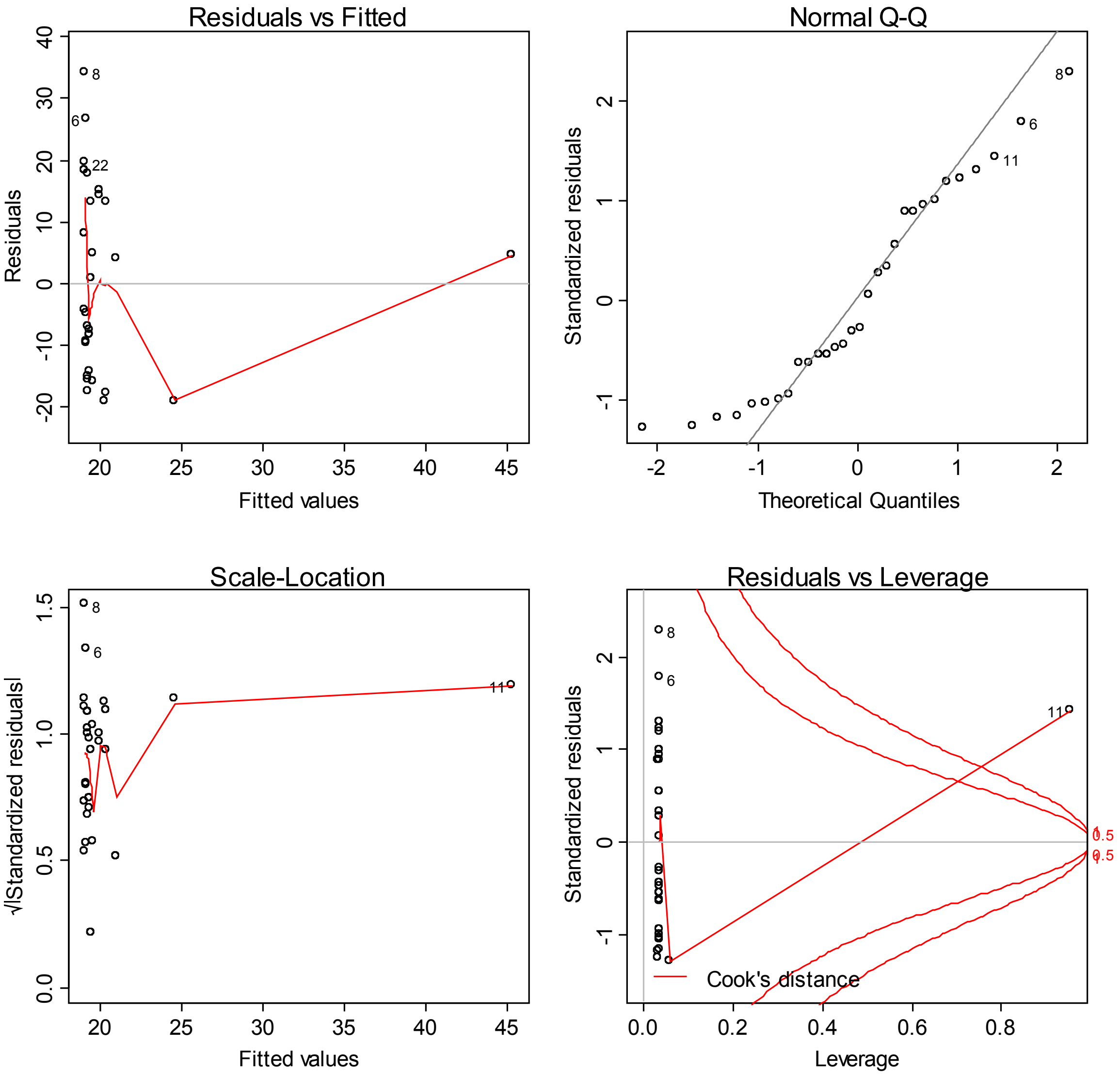


Significant nonlinearity remains, however. We can only conclude that there is a significant negative relationship between GDP and population growth, but we cannot formulate a specific model of that relationship.

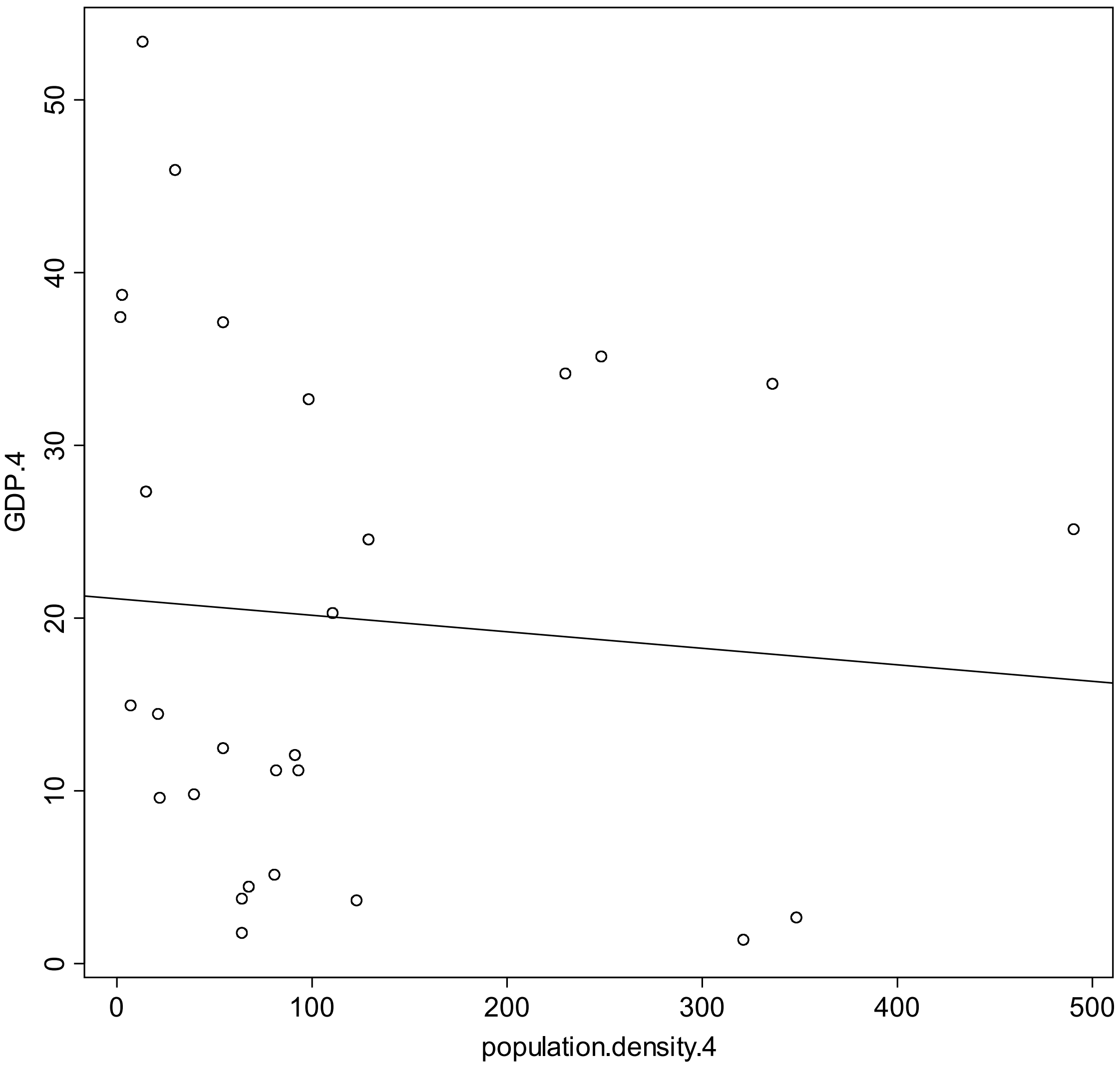
Population Density



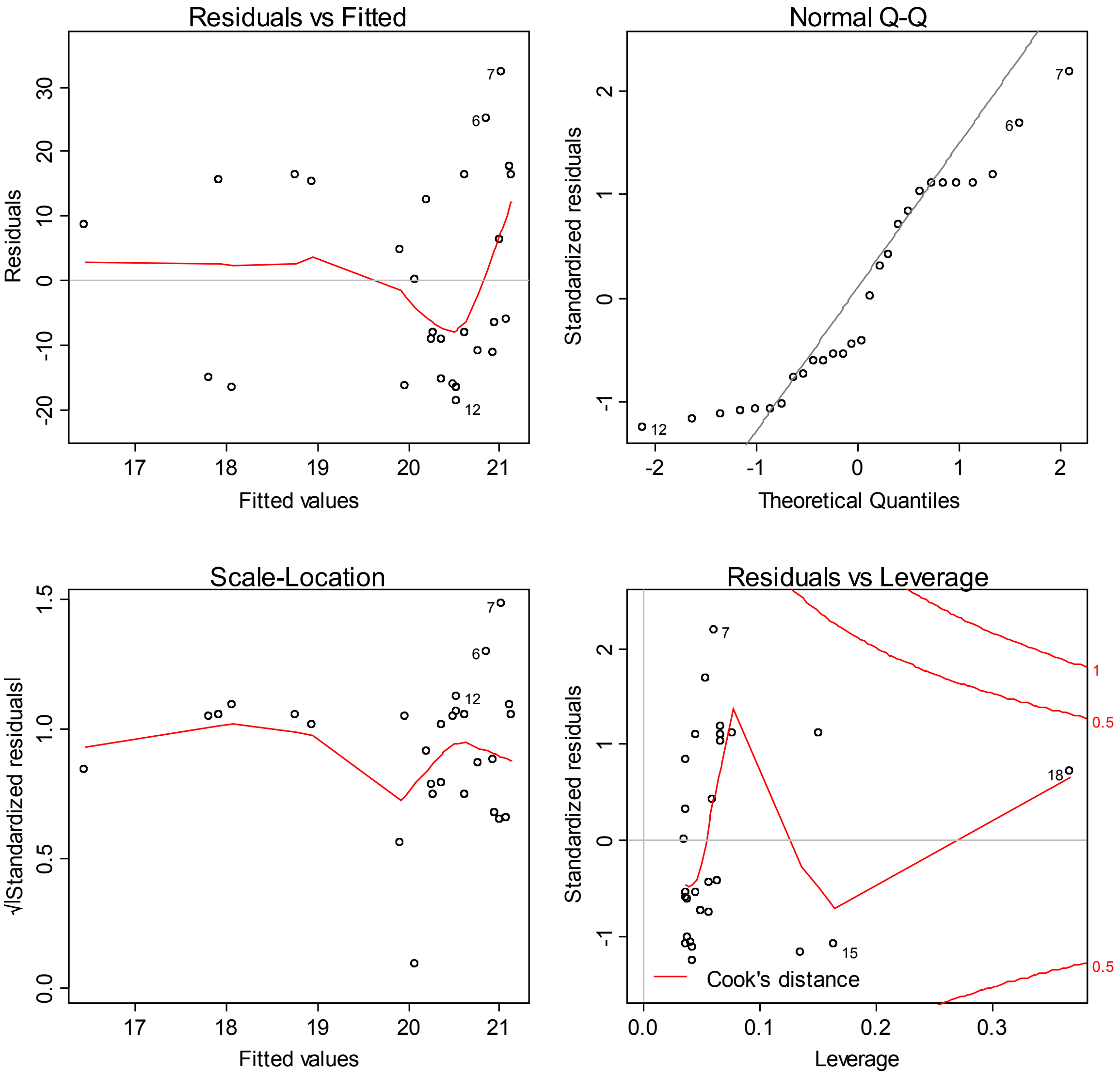
This is a plot of GDP as a function of population density. The extreme outlier in the upper right is Singapore, which is essentially a small island city. The data point to the right of 1000 corresponds to China. Both of these are suspected outliers, which could skew the analysis, so they are removed.



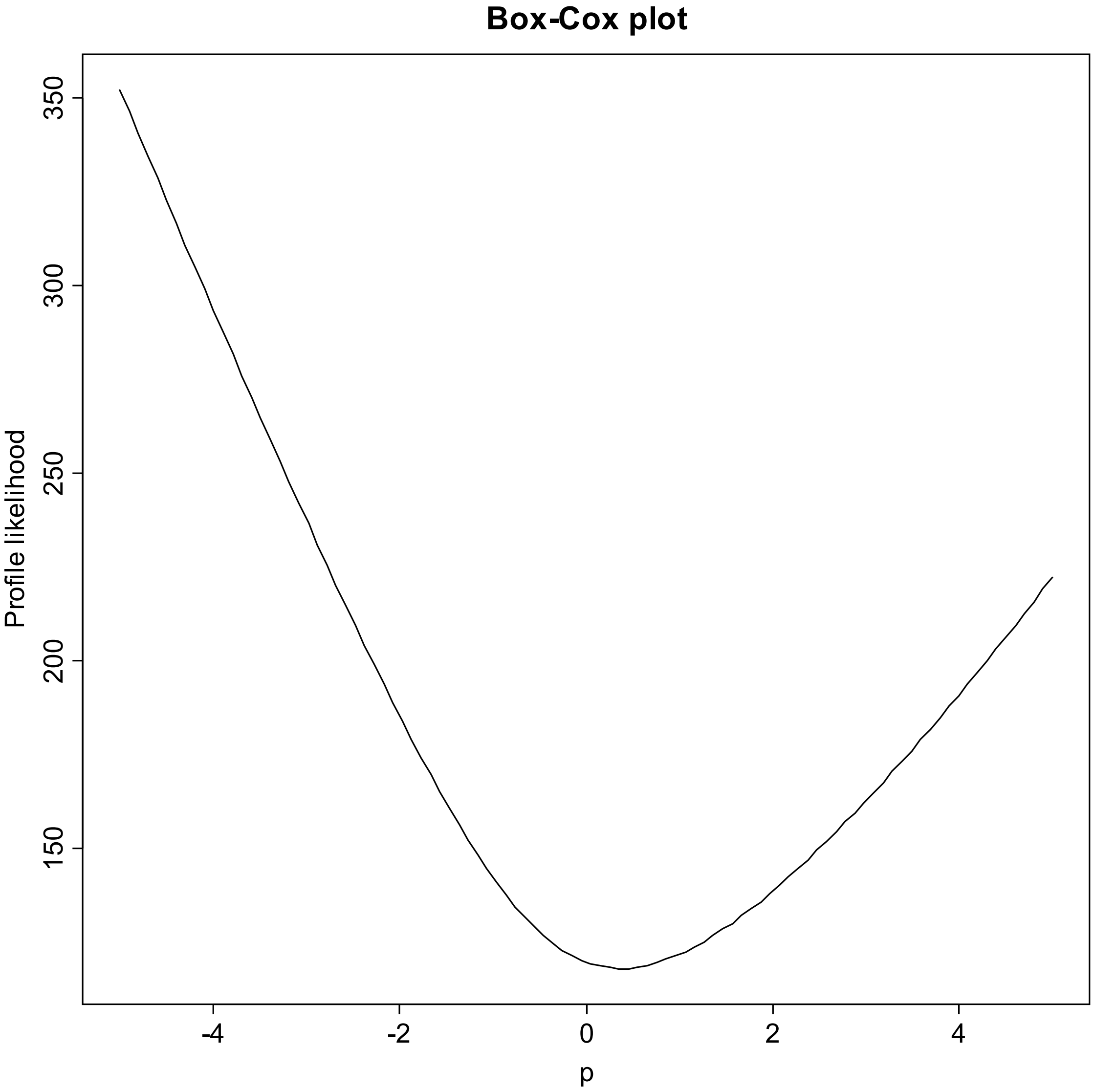
This is an analysis of the residuals. They are all negatively impacted by the presence of the extreme outliers.
=



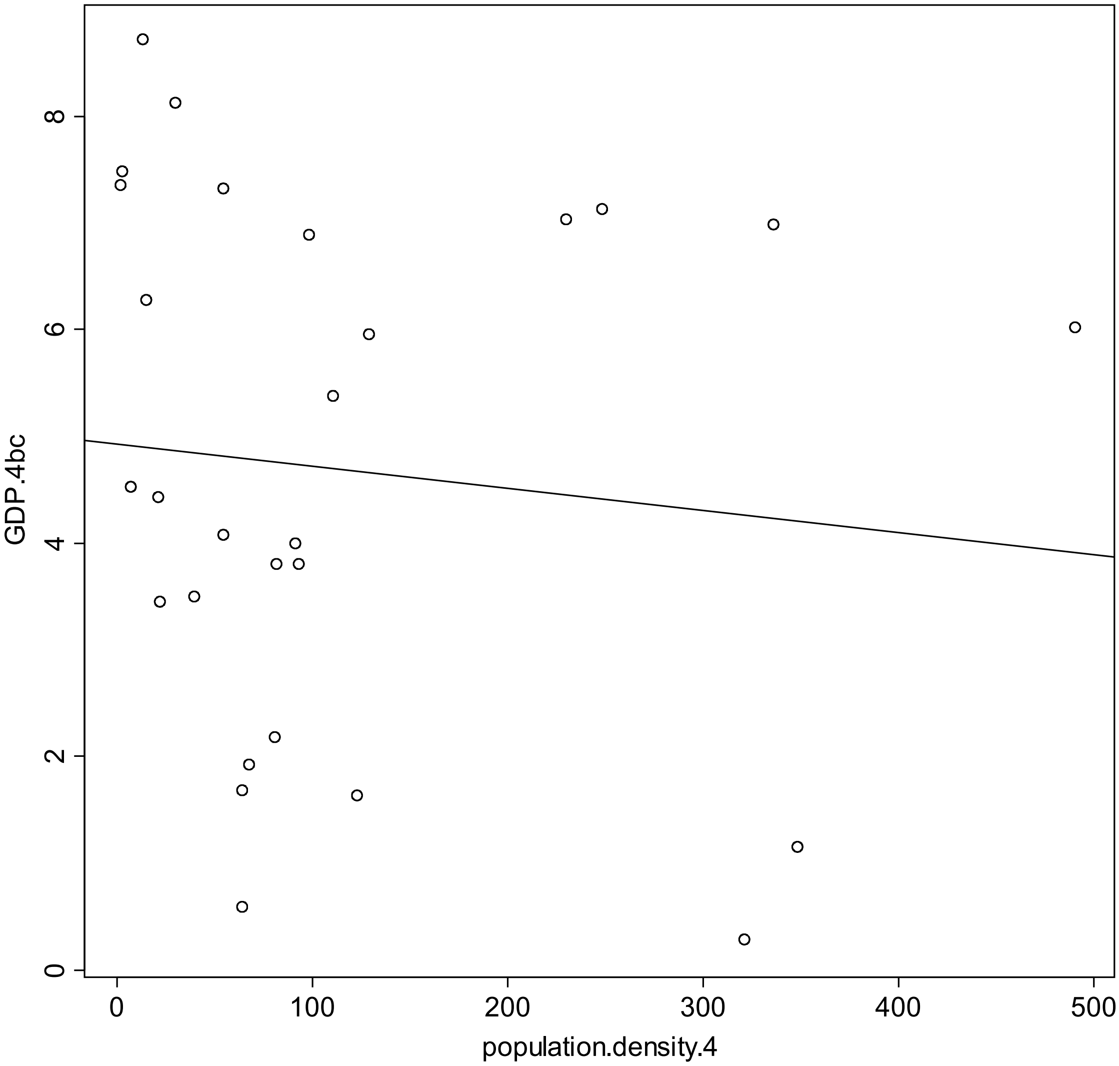
This is a scatter plot of GDP as a function of population density with the two outliers corresponding to China and Singapore removed. There appears to be a lot of noise around the regression line, which is perhaps not significantly far from horizontal (no relationship).



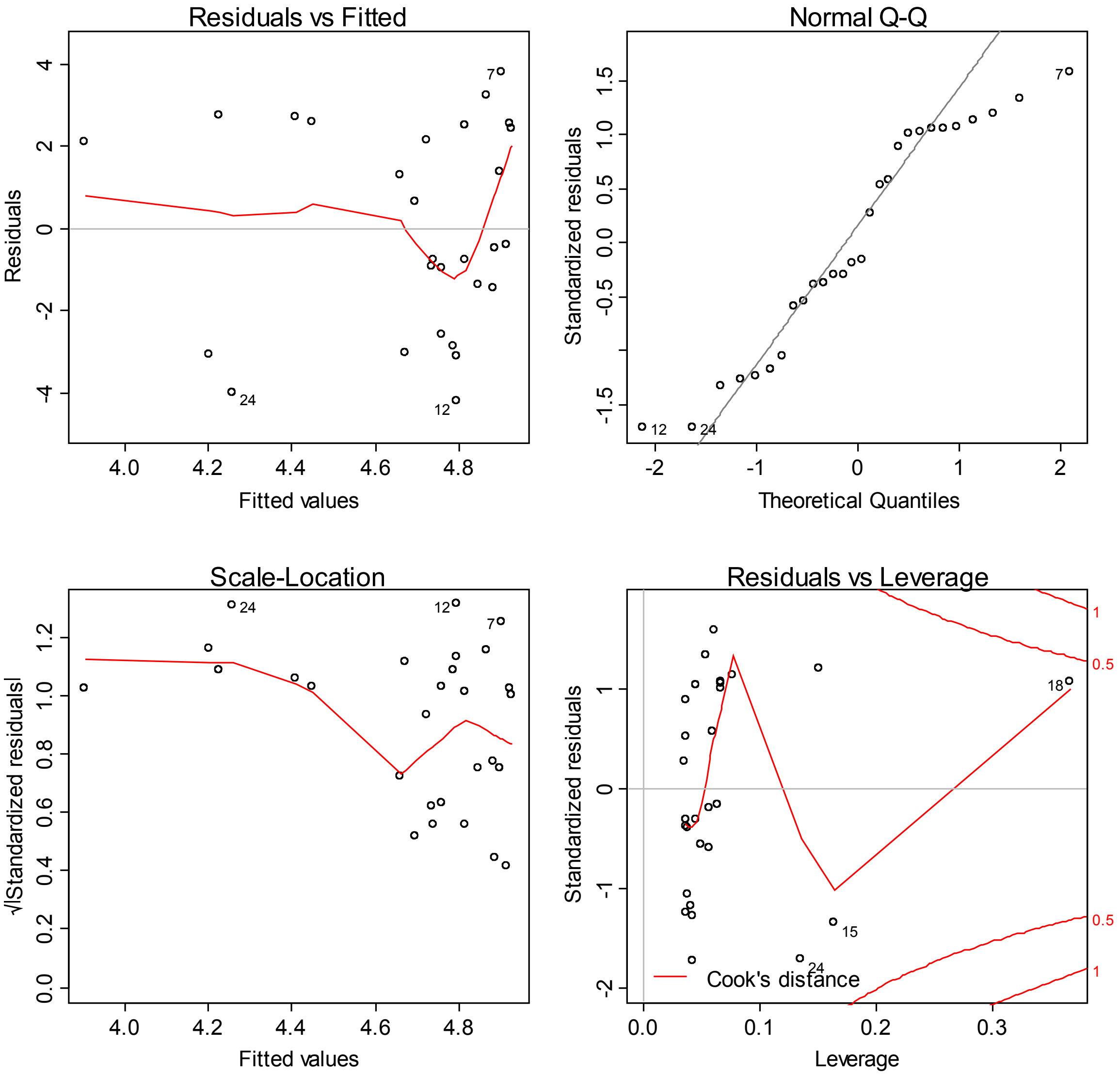
The residuals appear very nonlinear and nonnormal. Also, the p-value, which is 0.68, is very insignificant. A Box-Cox transform is attempted.



The optimal p-value is given by the minimum on this curve.

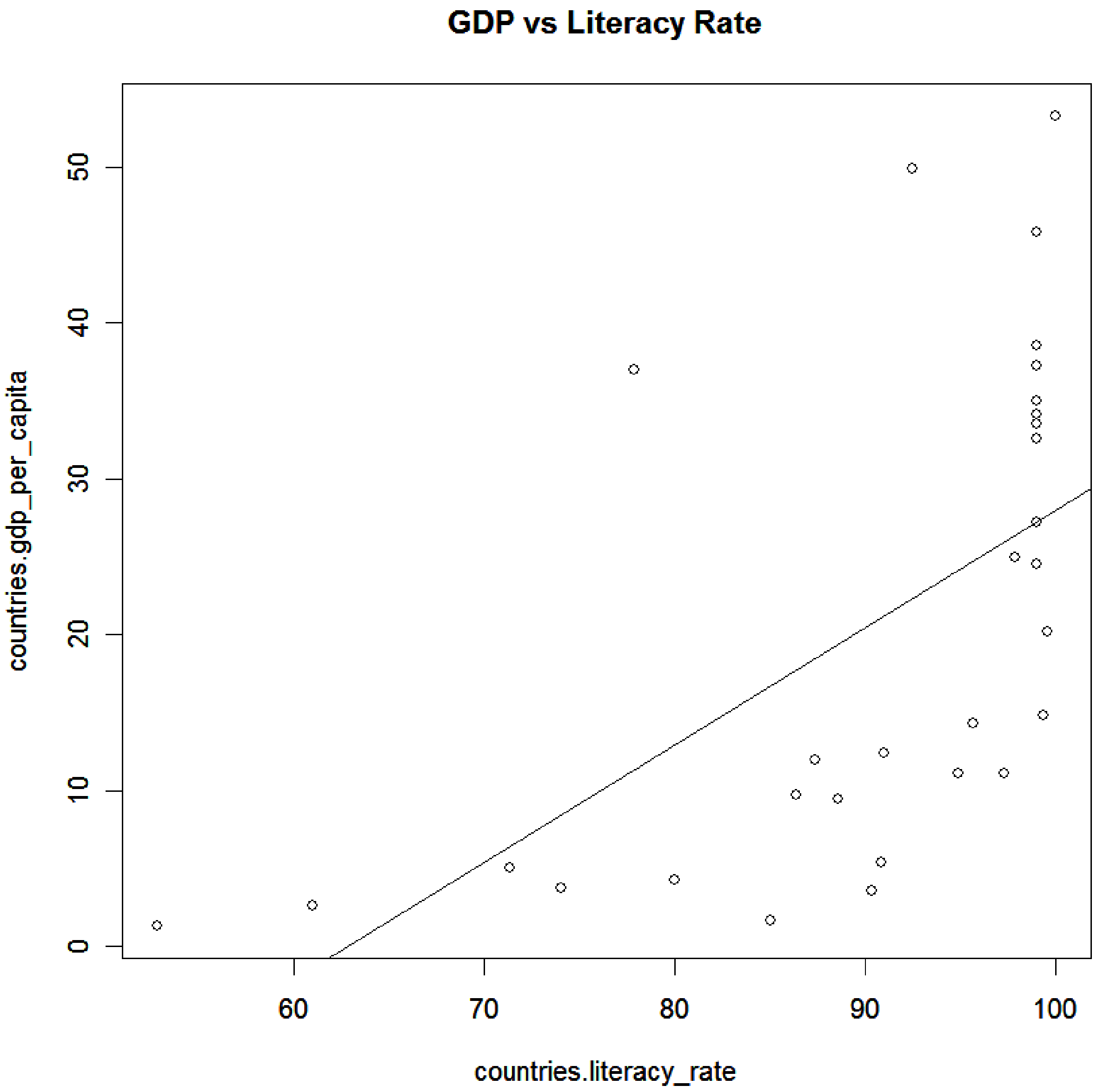


This is the Box-Cox transformed scatter plot.

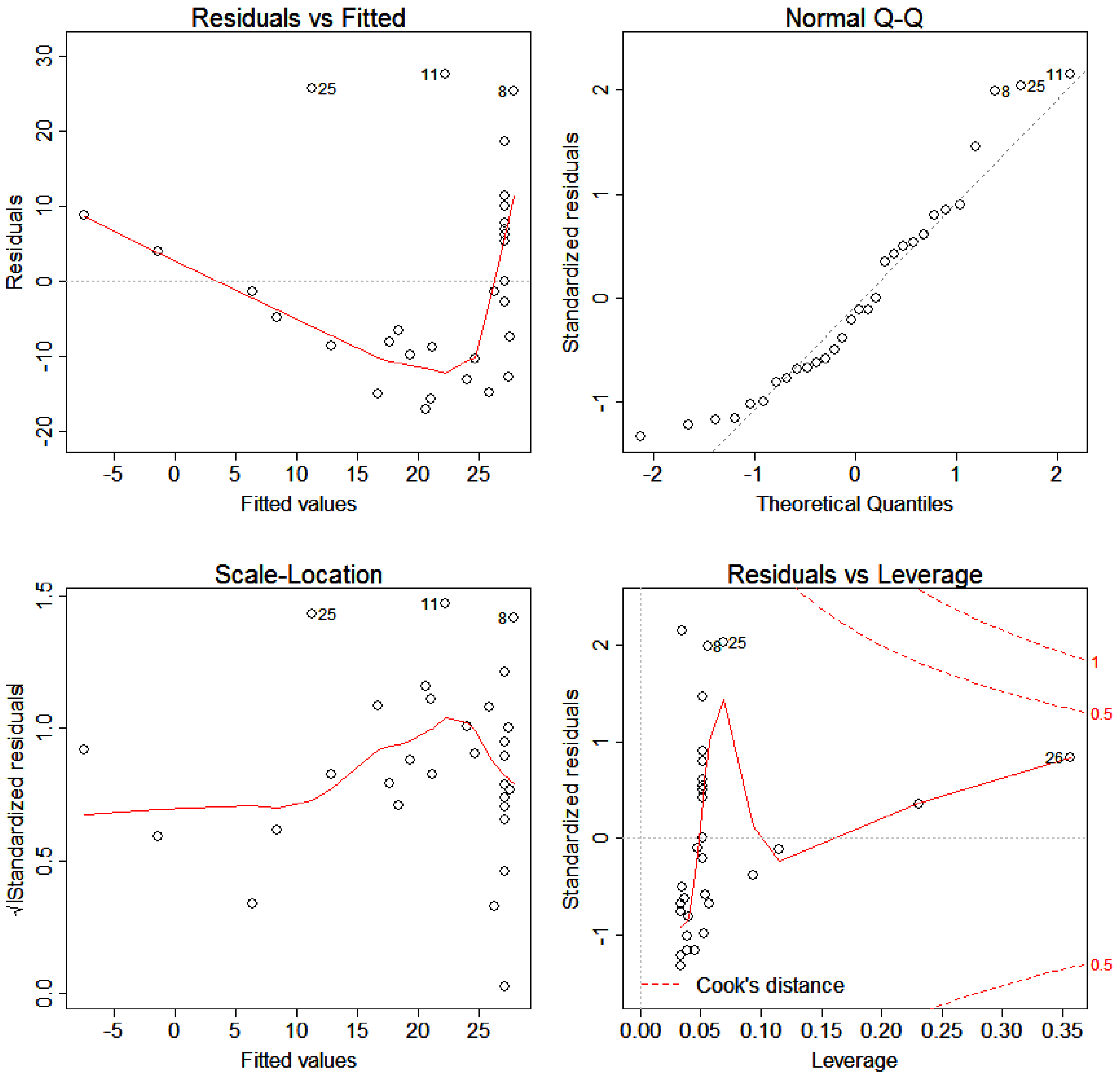


The Box-Cox plot does not reduce nonlinearity, and, because the p-value is given as 0.59, the relationship is not significant at all.

Literacy Rate



First, in the graph, you can already tell that the data needs desperately to be transformed. It looks almost as if it has a more exponential shape rather than linear. We now look at the residuals and Q-Q plot to see how our data looks there.

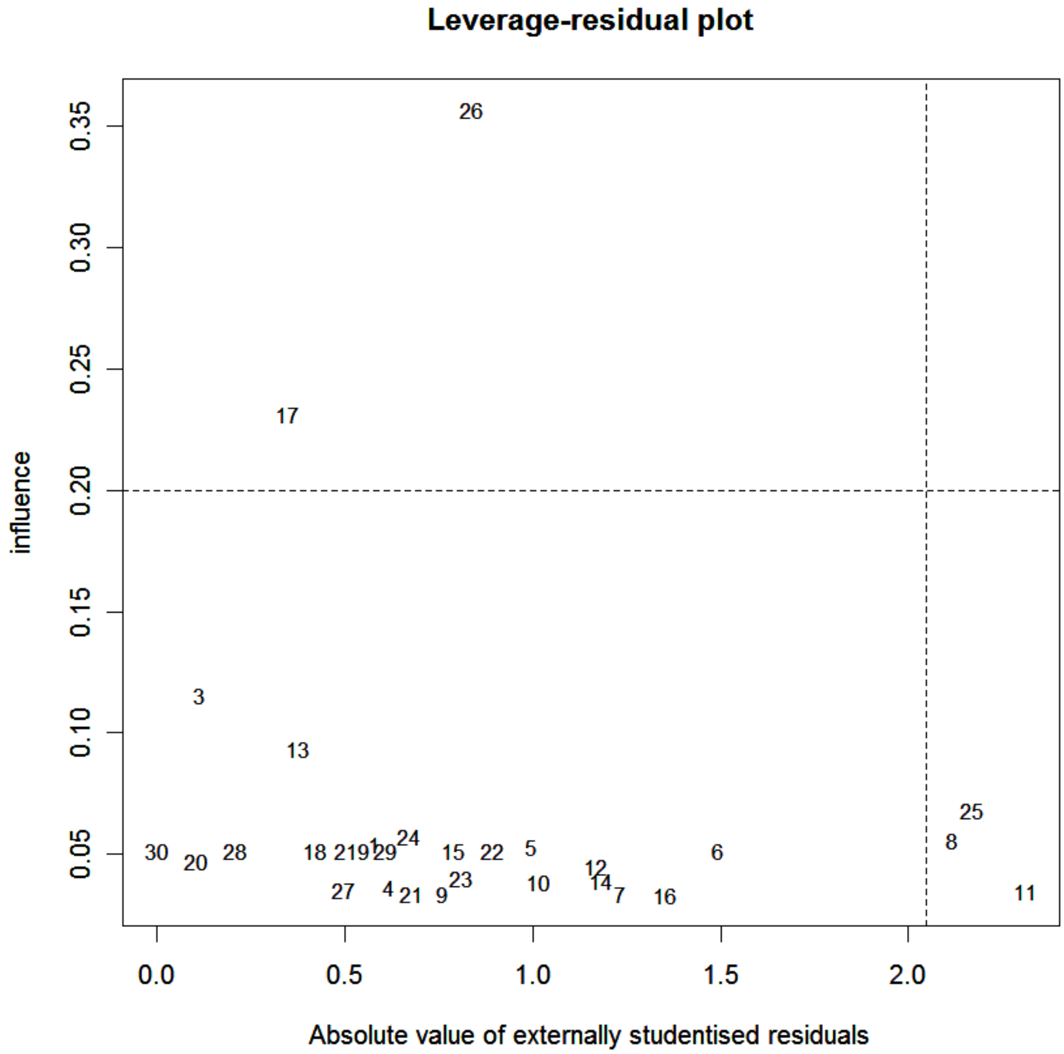


The residuals look awful, even though it looks like the data is normal from the Q-Q plot. When we look at the summary of the data:

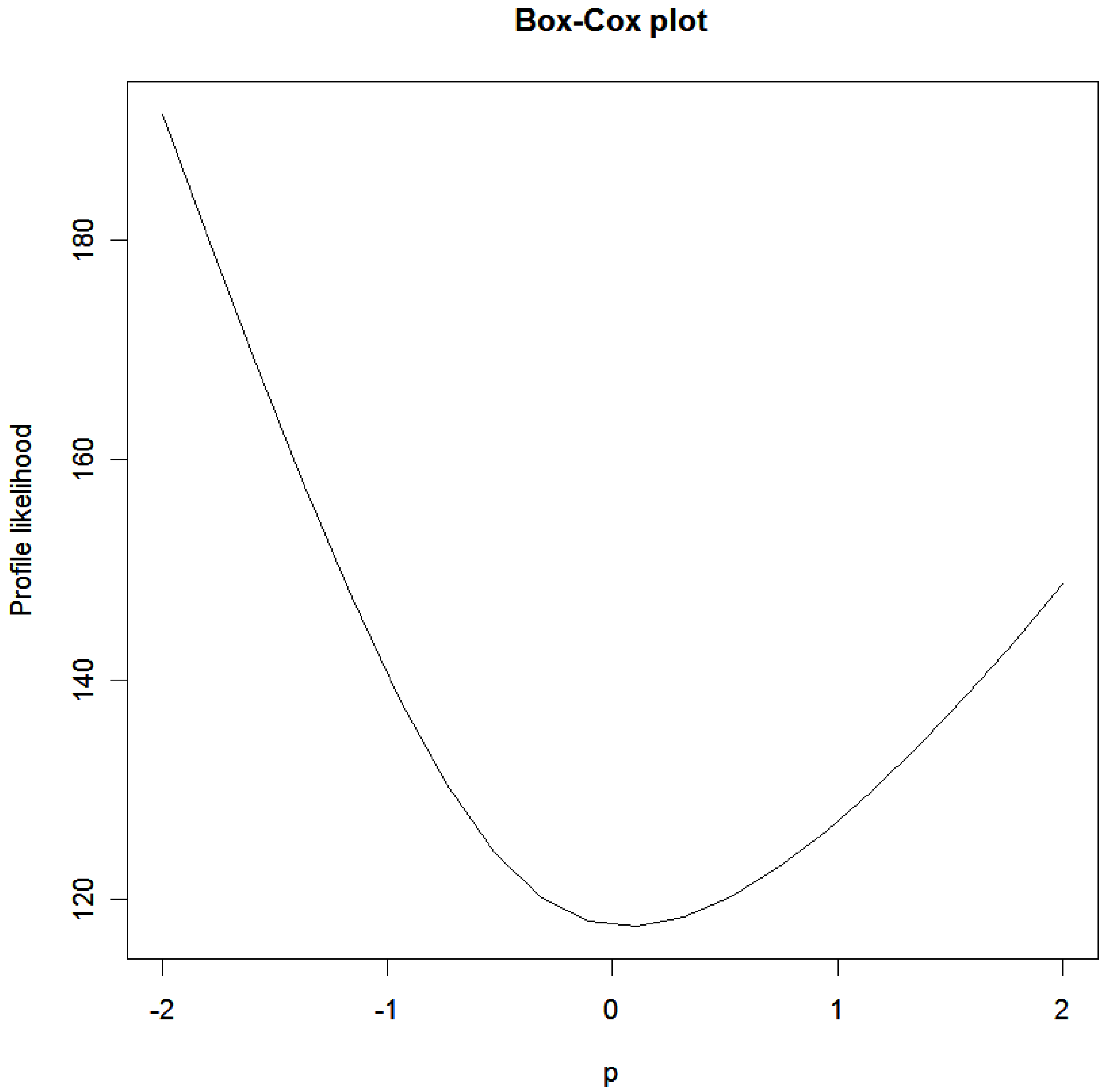
```
Call:
lm(formula = countries.gdp_per_capita ~ countries.literacy_rate)

Residuals:
    Min      1Q  Median      3Q     Max
-17.116  -9.587  -2.049   7.582  27.603

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -47.3684    18.1334  -2.612 0.014302 *
countries.literacy_rate   0.7531     0.1993   3.778 0.000759 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.07 on 28 degrees of freedom
Multiple R-Squared: 0.3377,     Adjusted R-squared: 0.314
```

F-statistic: 14.28 on 1 and 28 DF, p-value: 0.0007589

We have an R^2 of .3377. But we need to improve this model. First, we will look at the outliers and see if we can take anything out that might be skewing our data.



However, our Leverage- Residuals plot shows that we do not have any points that we can delete. So now we look to transform the data using the boxcox plot.



From this plot, we can transform the model using:

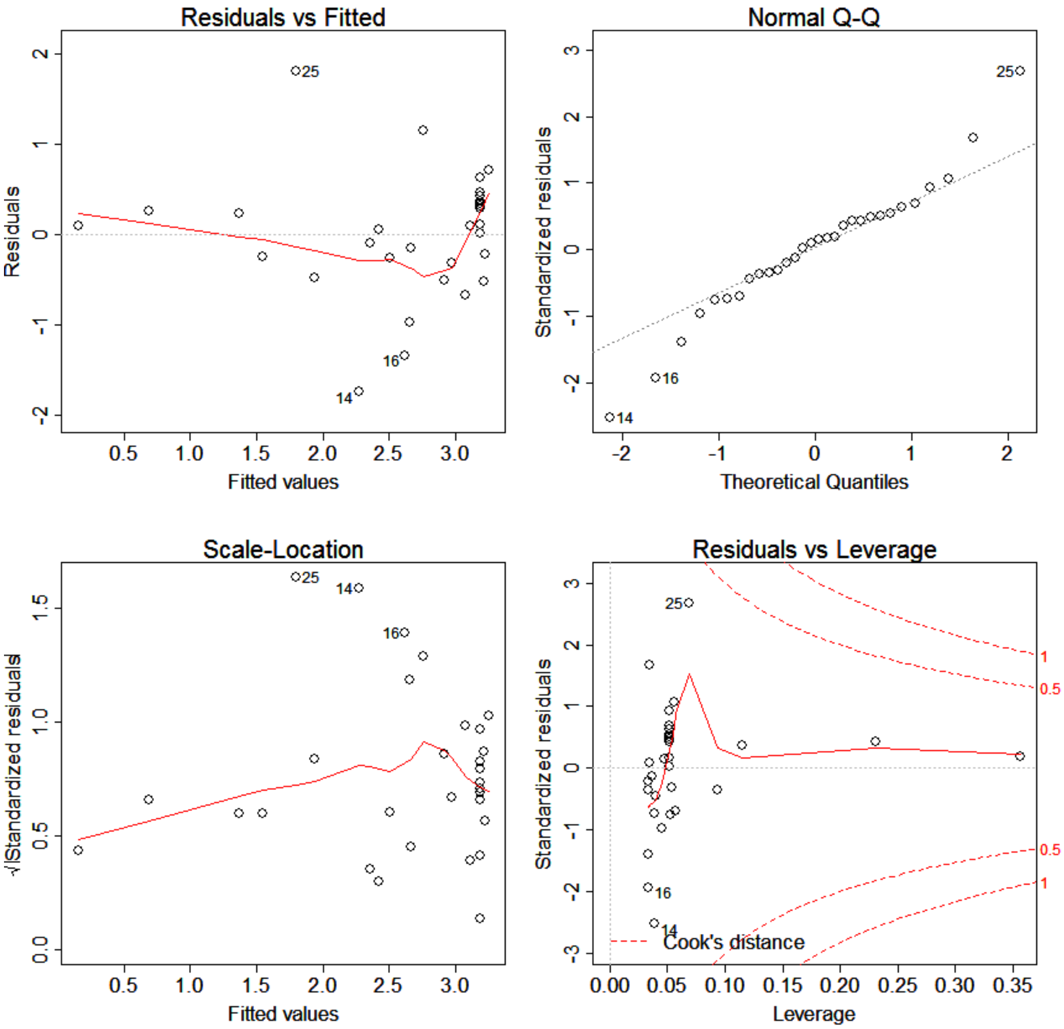
```
boxcox.lit_rate=lm(log(countries.gdp_per_capita) ~ countries.literacy_rate, data=countries)
```

This way we get a much better R^2, and the p value is significant:

```
Call:
lm(formula = log(countries.gdp_per_capita) ~ countries.literacy_rate,
    data = countries)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74177 -0.29434  0.08332  0.33898  1.81169

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -3.32024    0.97500  -3.405  0.00201 **
countries.literacy_rate  0.06572    0.01072   6.132 1.28e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

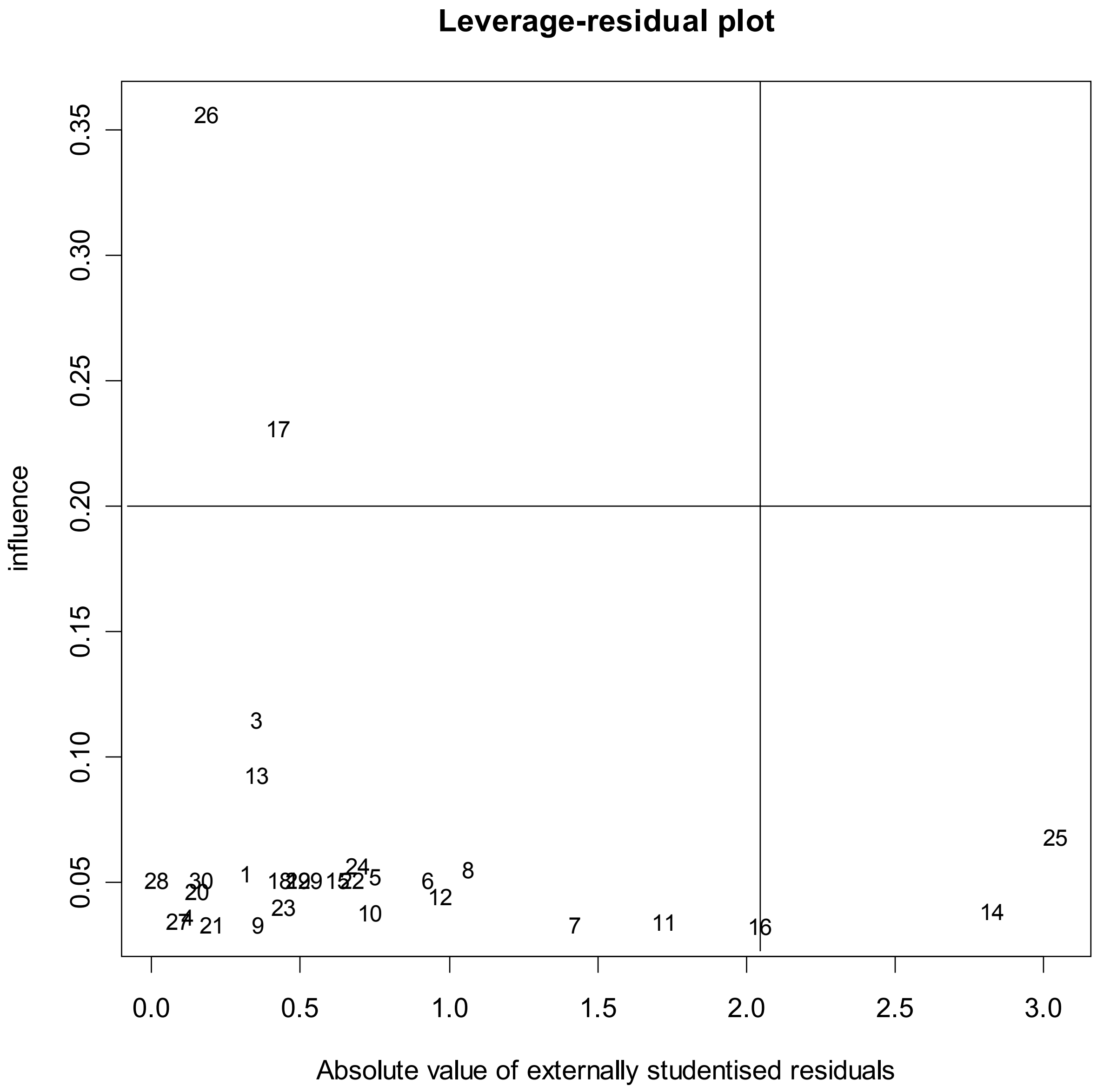
```
Residual standard error: 0.7026 on 28 degrees of freedom
Multiple R-Squared: 0.5732,     Adjusted R-squared: 0.5579
F-statistic:  37.6 on 1 and 28 DF,  p-value: 1.285e-06
```

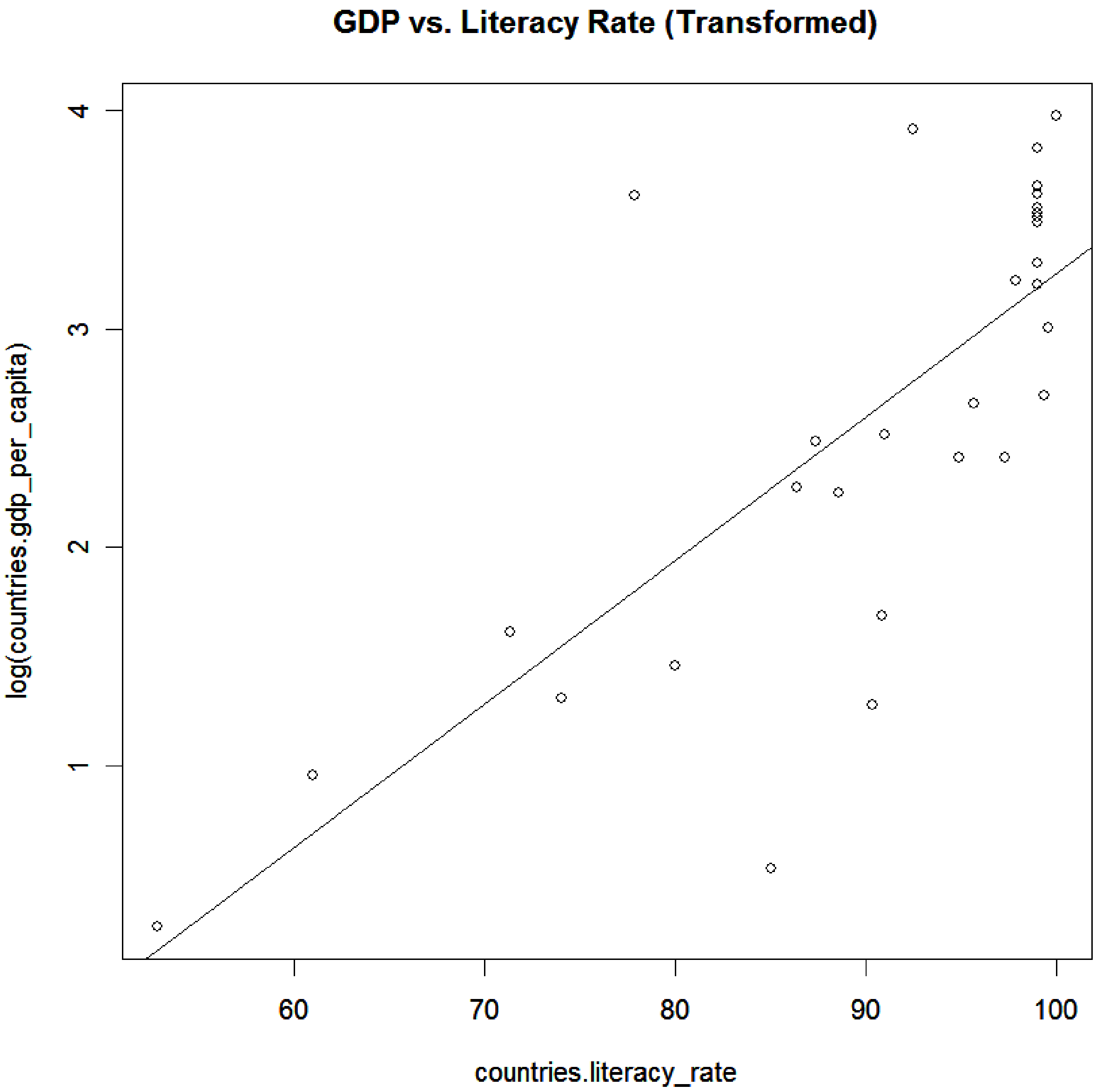
And better plots:



The residuals, although not perfect, look better than before, along with the Q-Q plot looking slightly more linear with the exception of the ends.

When we run the Leverage-Residuals plot again, we see that there are no outliers in this model either:


Leverage-residual plot
influence
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Absolute value of externally studentised residuals
26
17
3
13
25
28
30
20
1
18
2
29
15
22
24
5
6
8
12
27
4
21
9
23
10
7
11
16
14



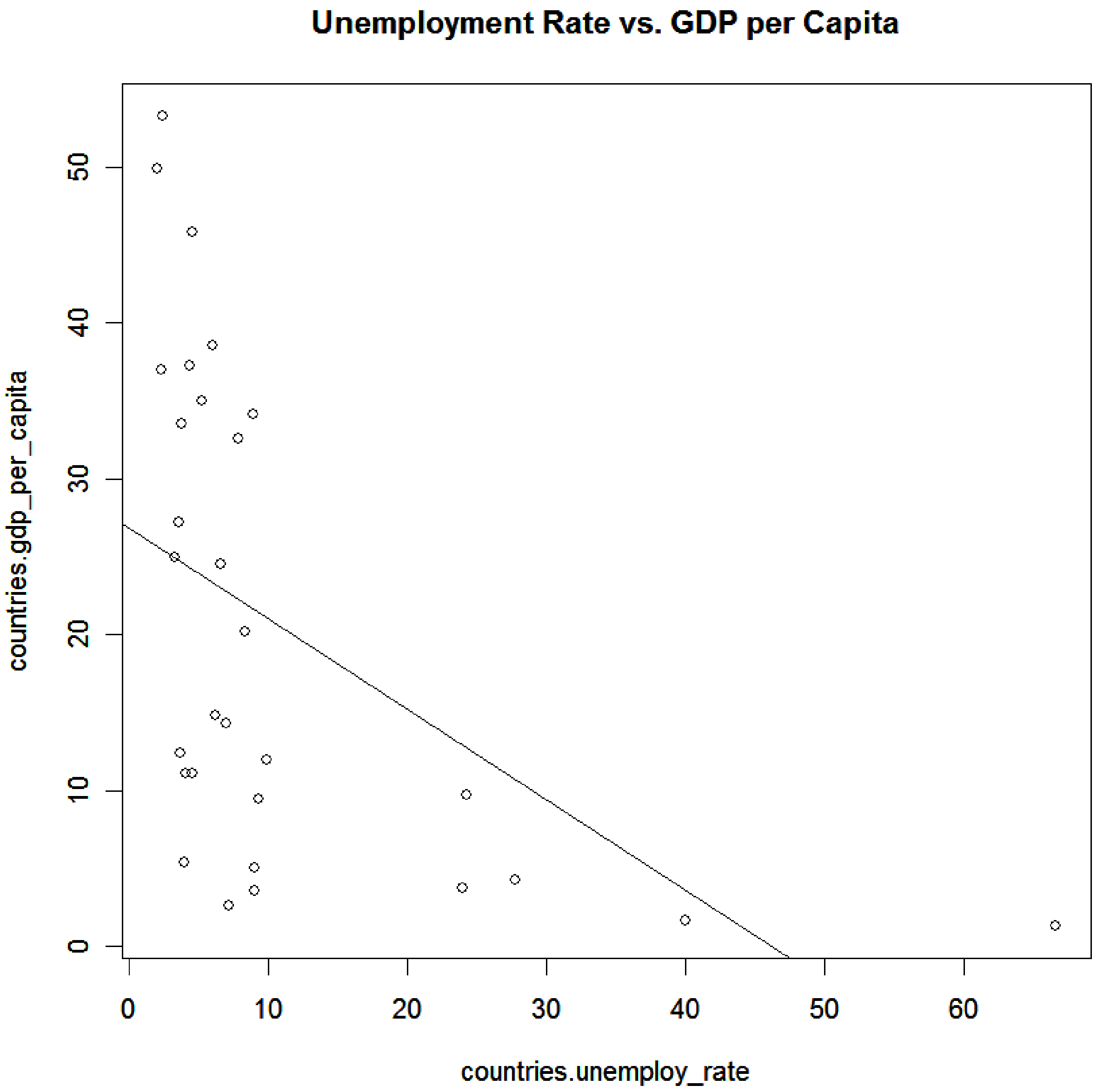
This plot also looks much better than before. Therefore, this model should be used rather than the original.

With the final model: GDP= -3.320 + .0657(literacy rate)

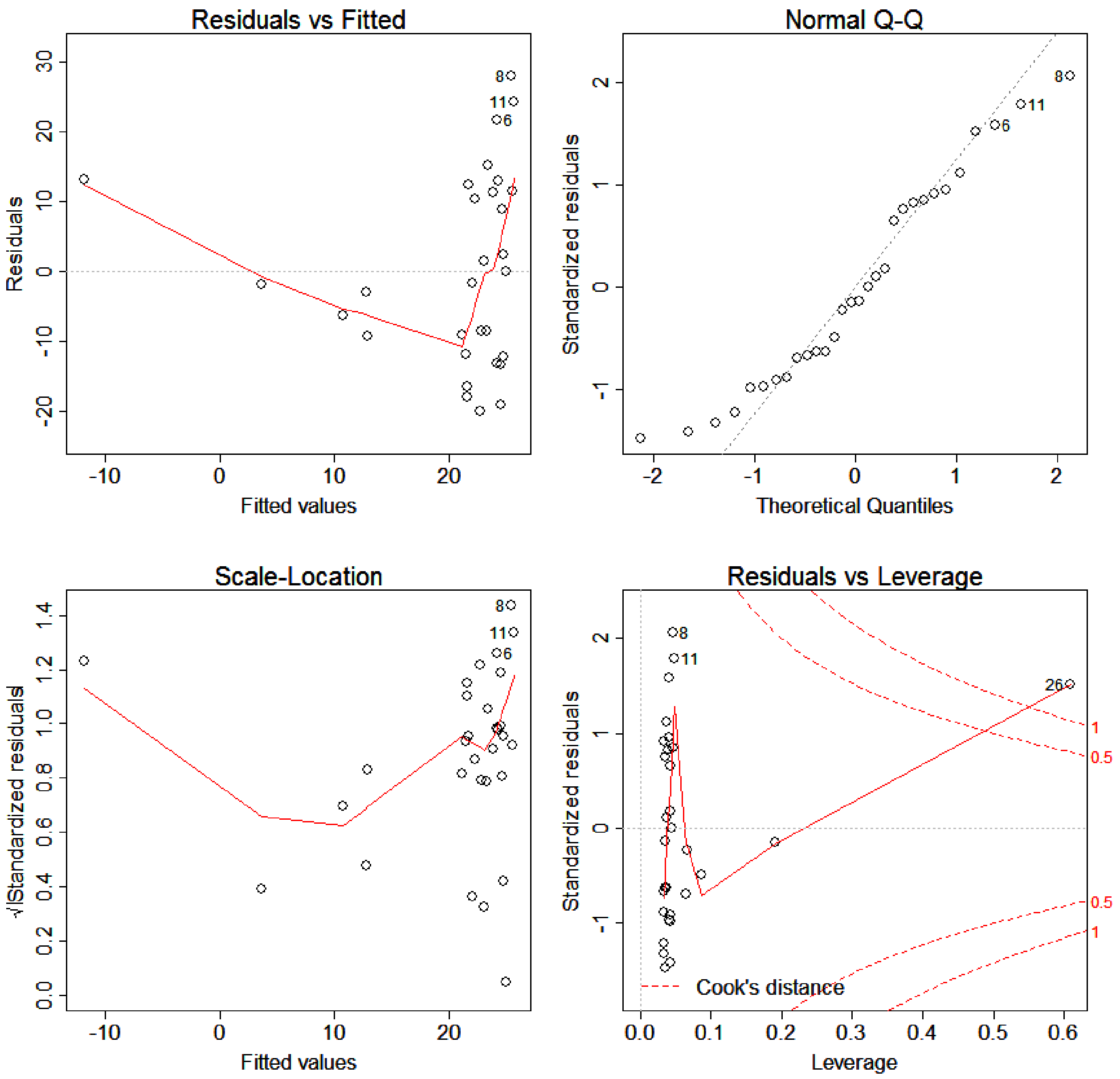
```
Call:
lm(formula = log(countries.gdp_per_capita) ~ countries.literacy_rate,     data = countries)

Coefficients:
        (Intercept)  countries.literacy_rate
          -3.32024                0.06572
```

Overall, we can assume that literacy rate is a relatively significant factor by itself in predicting the GDP.

Unemployment Rate



First, in the graph, you can already tell that the data needs desperately to be transformed. It looks almost as if it has a more exponential shape rather than linear. We now look at the residuals and Q-Q plot to see how our data looks there.

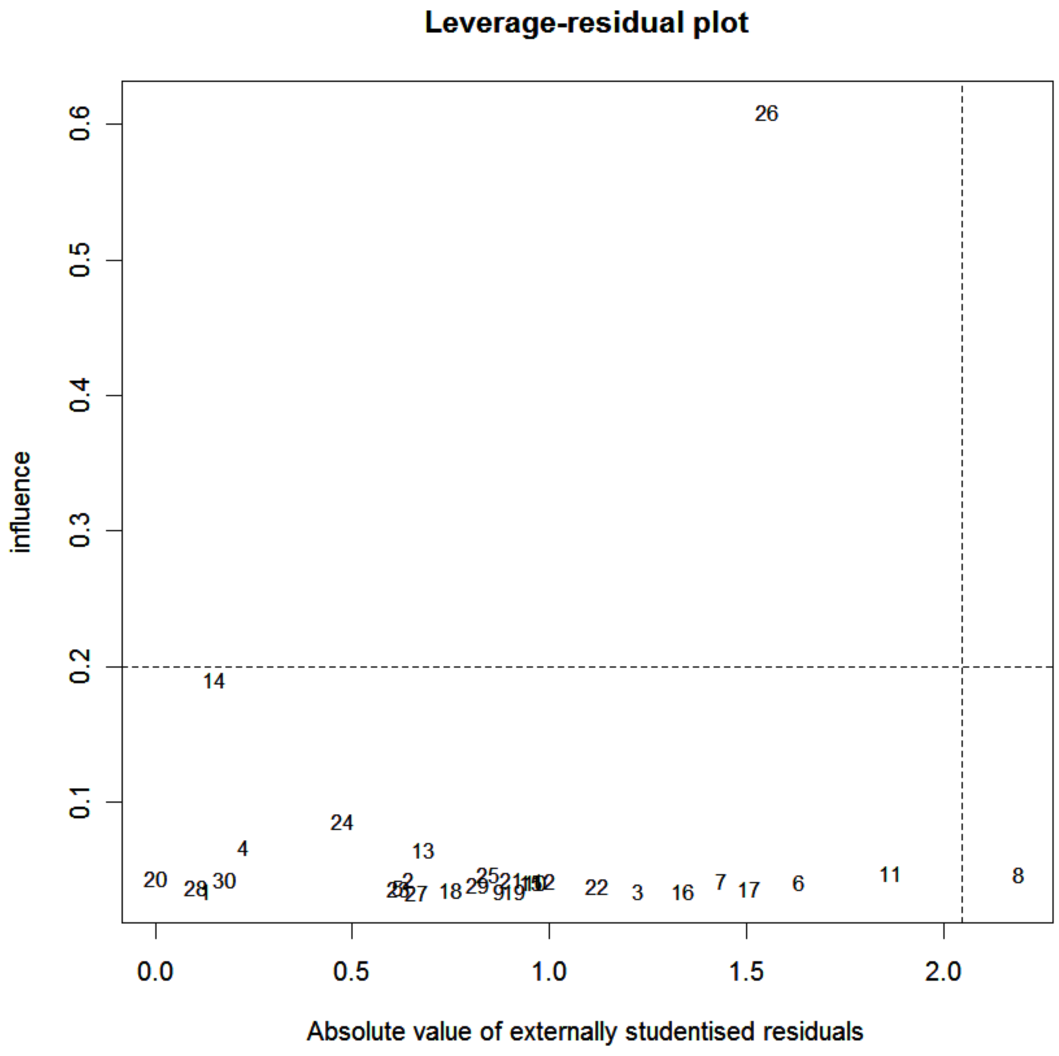


The residuals look awful, and the Q-Q plot doesn't look great either, definitely signaling that we need a transformation. When we look at the summary of the data:

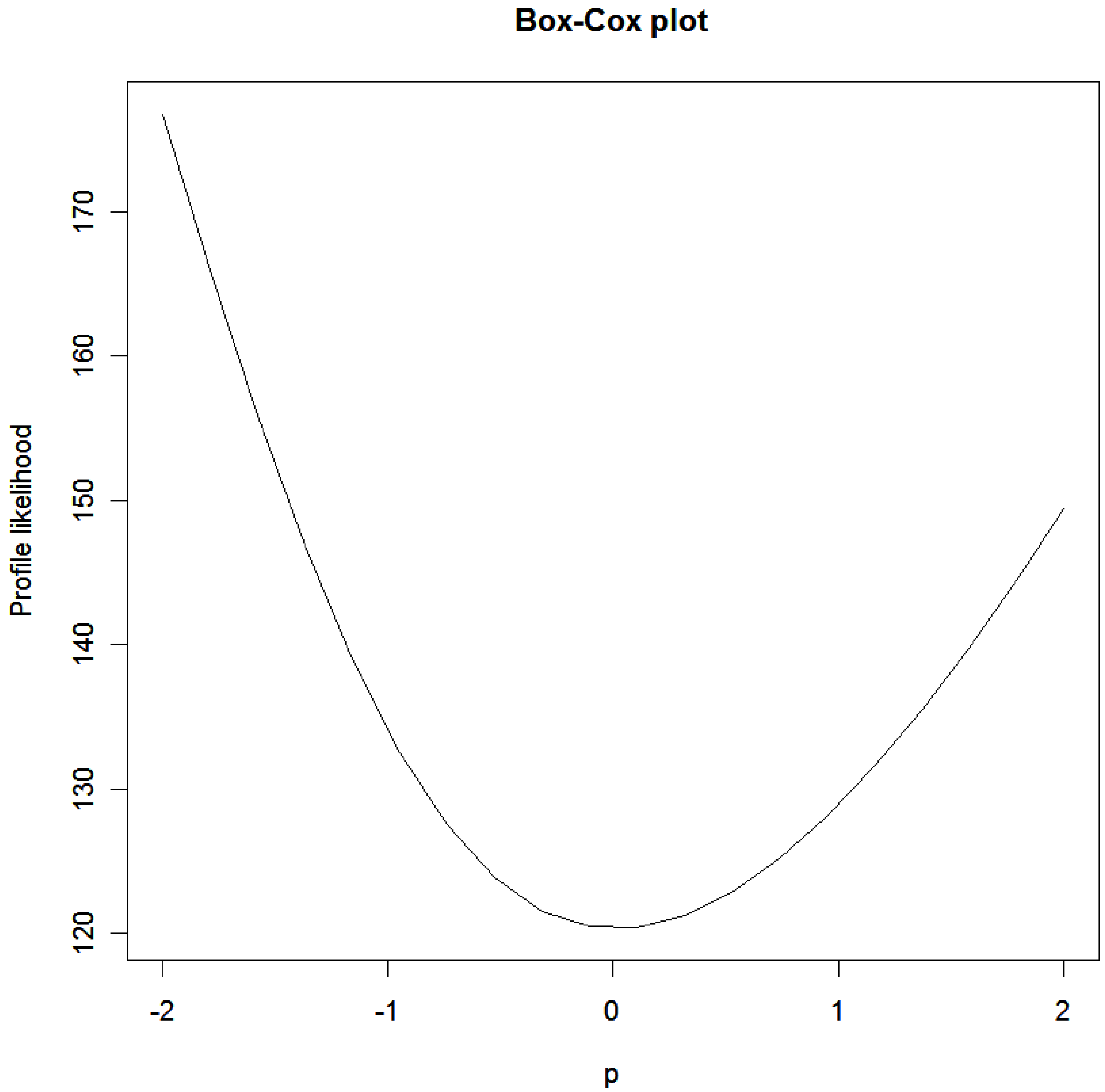
```
Call:
lm(formula = countries.gdp_per_capita ~ countries.unemploy_rate)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.098 -11.289  -1.859  11.431  27.867 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              26.8868     3.2657   8.233 5.83e-09 ***
countries.unemploy_rate  -0.5817     0.1891  -3.076  0.00465 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.88 on 28 degrees of freedom
```

Multiple R-Squared: 0.2526, Adjusted R-squared: 0.2259
F-statistic: 9.465 on 1 and 28 DF, p-value: 0.004645

We have an R^2 of .2526, and it looks as though the unemployment rate is significant in this model. However, we need to improve this model. First, we will look at the outliers and see if we can take anything out that might be skewing our data:



However, our Leverage- Residuals plot shows that we do not have any points that we can delete. So now we look to transform the data using the boxcox plot.



From this plot, we can transform the model using:

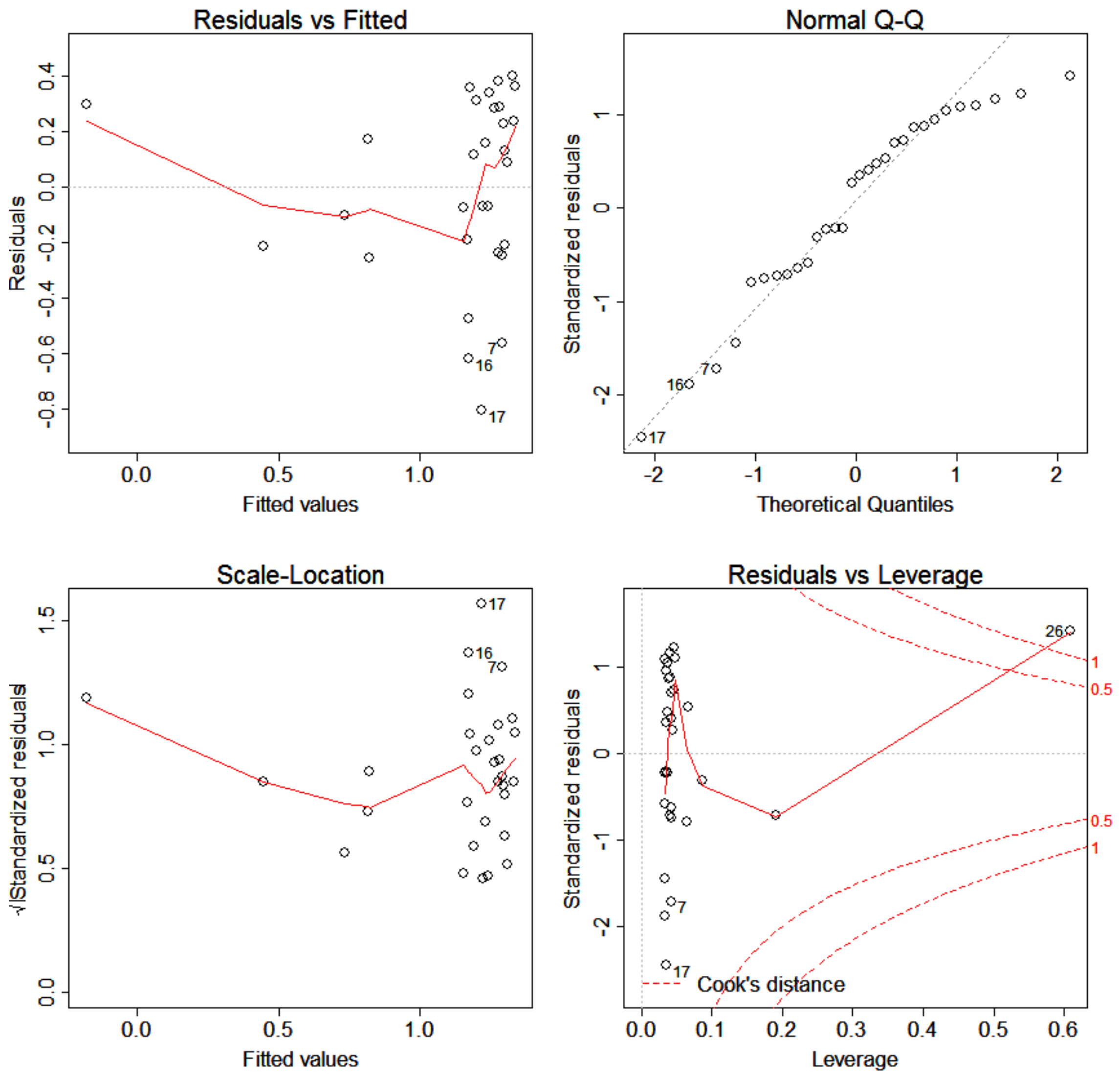
```
boxcox.unemp_rate=lm(log(countries.gdp_per_capita, 10) ~ countries.unemploy_rate, data=countries)
```

This way we get a much better R^2:

```
Call:
lm(formula = log(countries.gdp_per_capita, 10) ~ countries.unemploy_rate,
    data = countries)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8039 -0.2140  0.1010  0.2853  0.3972

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.388410   0.078522   17.68  < 2e-16 ***
countries.unemploy_rate -0.023549   0.004546   -5.18 1.70e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

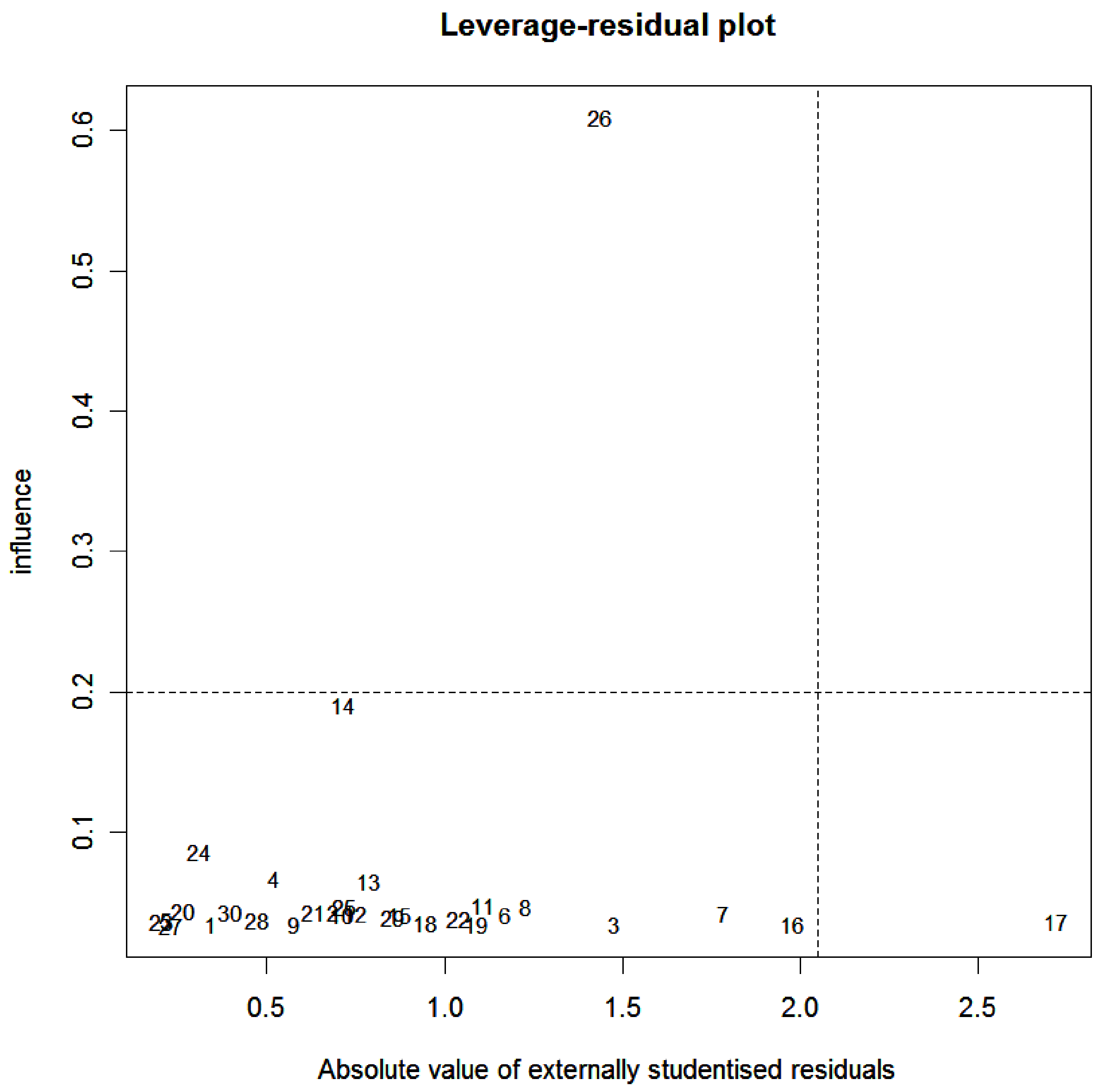
```
Residual standard error: 0.3337 on 28 degrees of freedom
Multiple R-Squared: 0.4893,     Adjusted R-squared: 0.4711
F-statistic: 26.83 on 1 and 28 DF,  p-value: 1.697e-05
```

Check the plots:

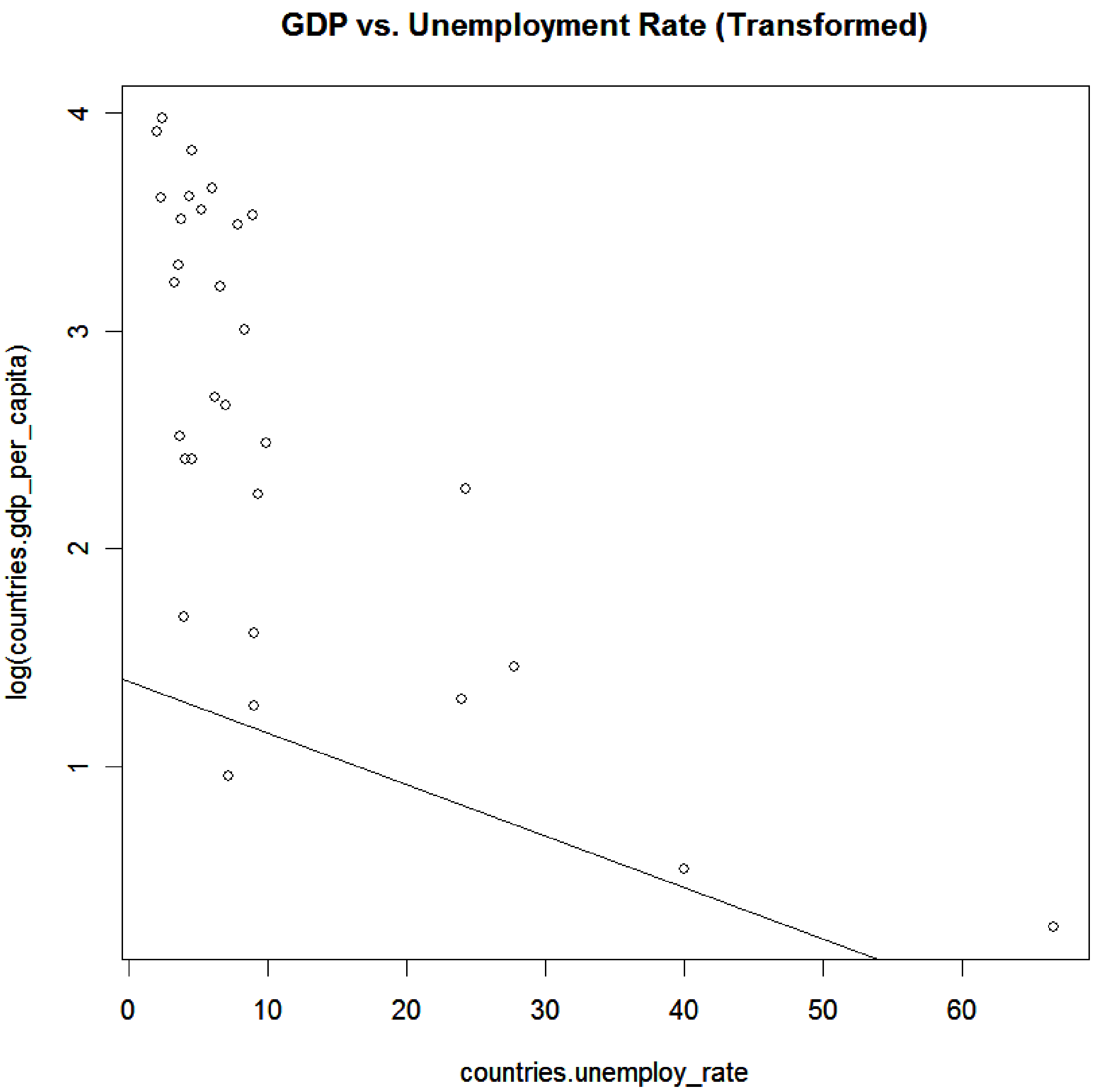


These plots still look really bad.

When we run the Leverage-Residuals plot again, we see that there are no outliers in this model either:



Therefore, our final graph is:

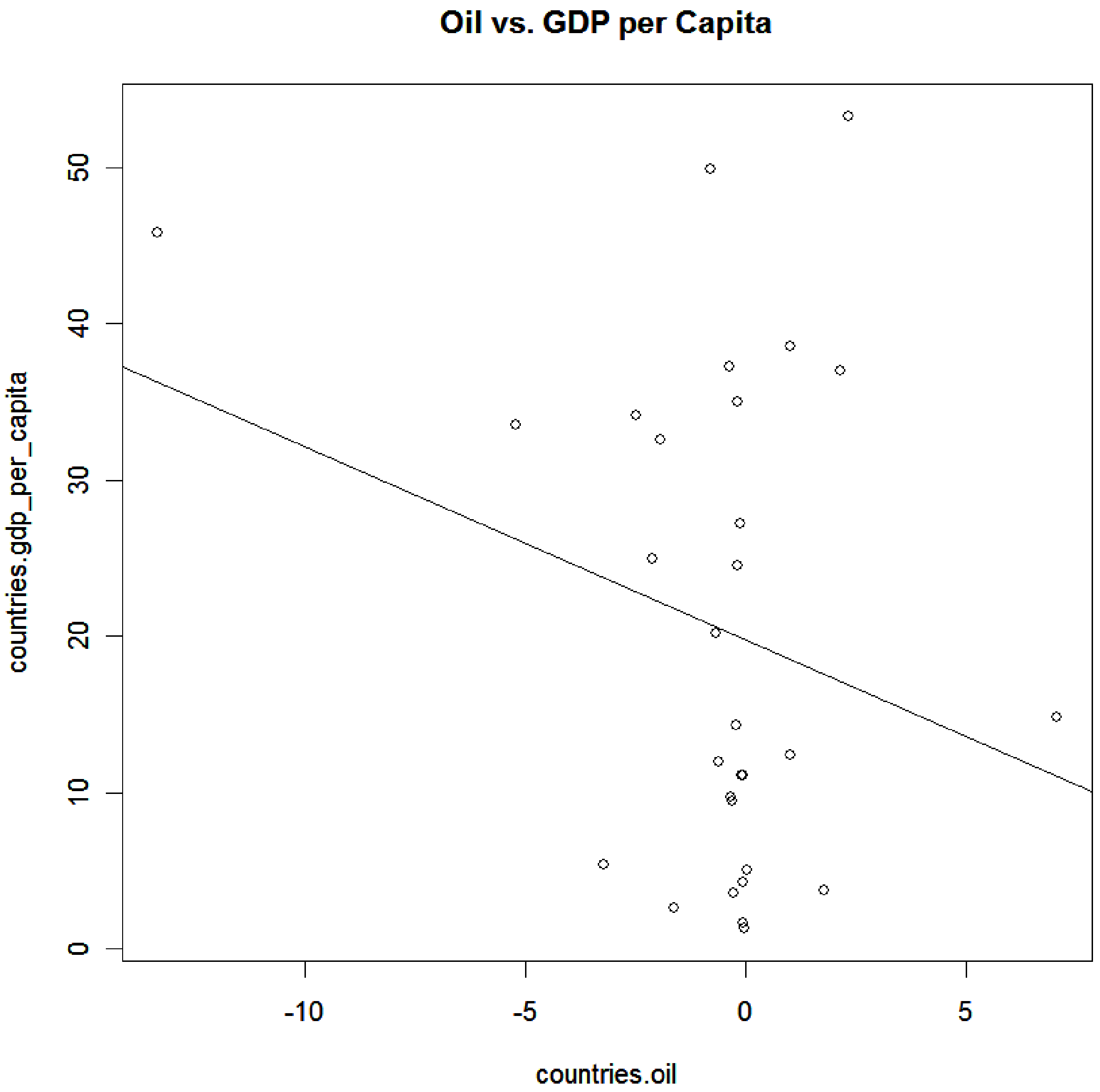


If we try to get a final model: GDP= 1.388 -.0236(unemployment rate)

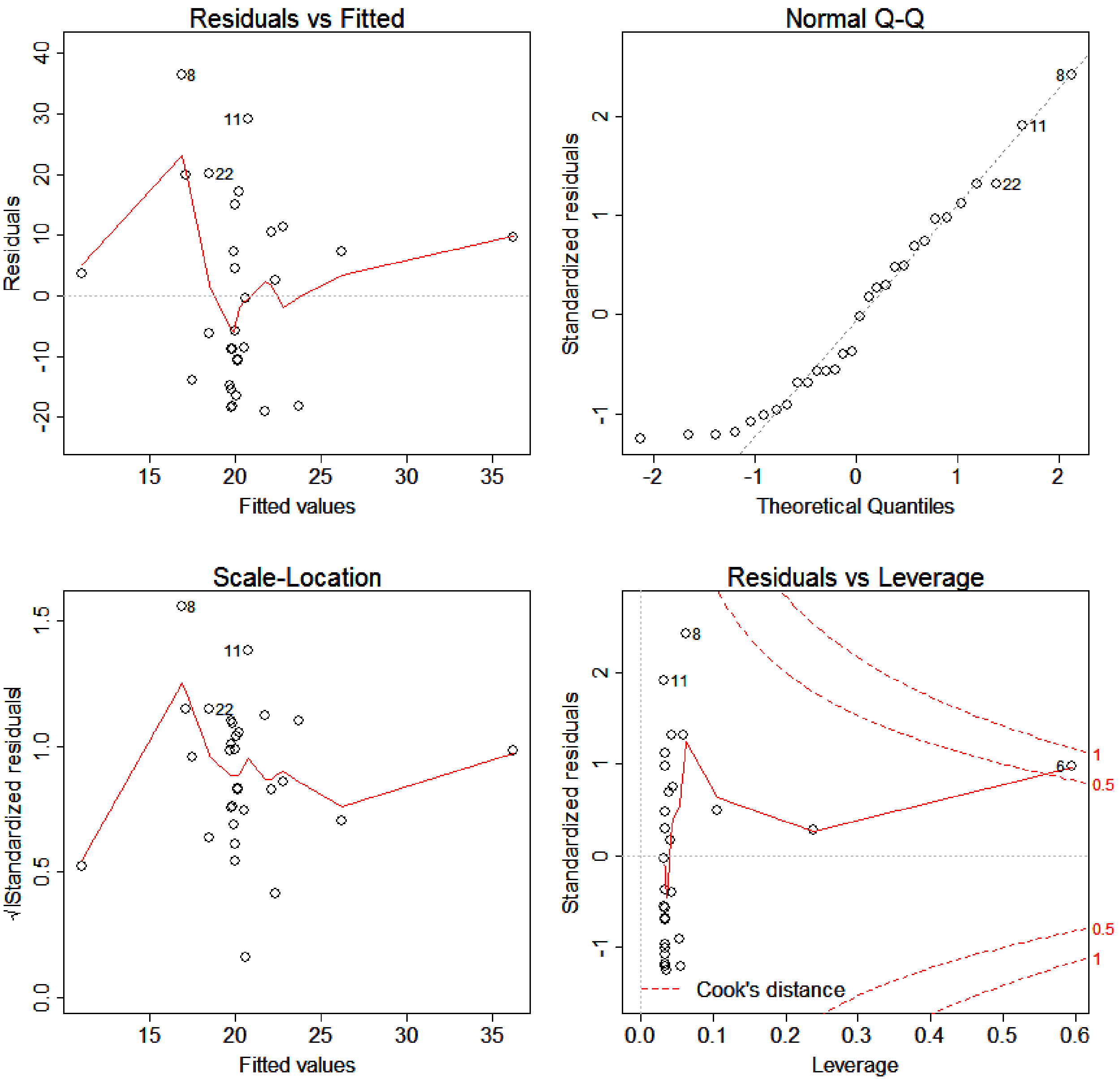
```
Call:
lm(formula = log(countries.gdp_per_capita, 10) ~ countries.unemploy_rate,     data = countries)

Coefficients:
           (Intercept)  countries.unemploy_rate
              1.38841                -0.02355
```

Overall, this model is simply unacceptable. We cannot use this model because the data is not normal or equally variant.

Oil



First, in the graph, you can already tell that the data needs desperately to be transformed. This data is definitely not linear. We now look at the residuals and Q-Q plot to see how our data looks there.

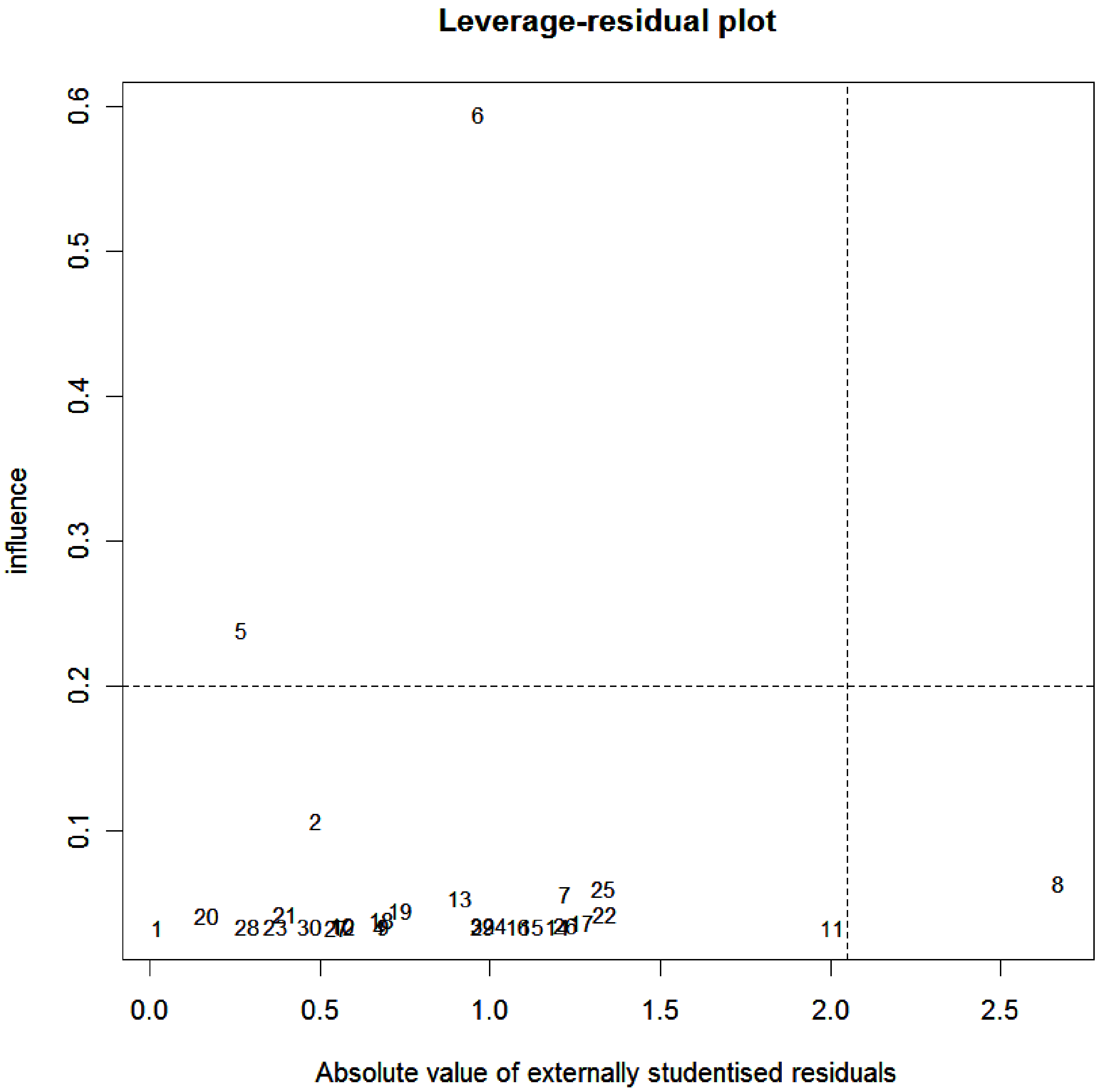


The residuals look really bad, but the Q-Q plot does not look so bad. We definitely need a transformation. When we look at the summary of the data:

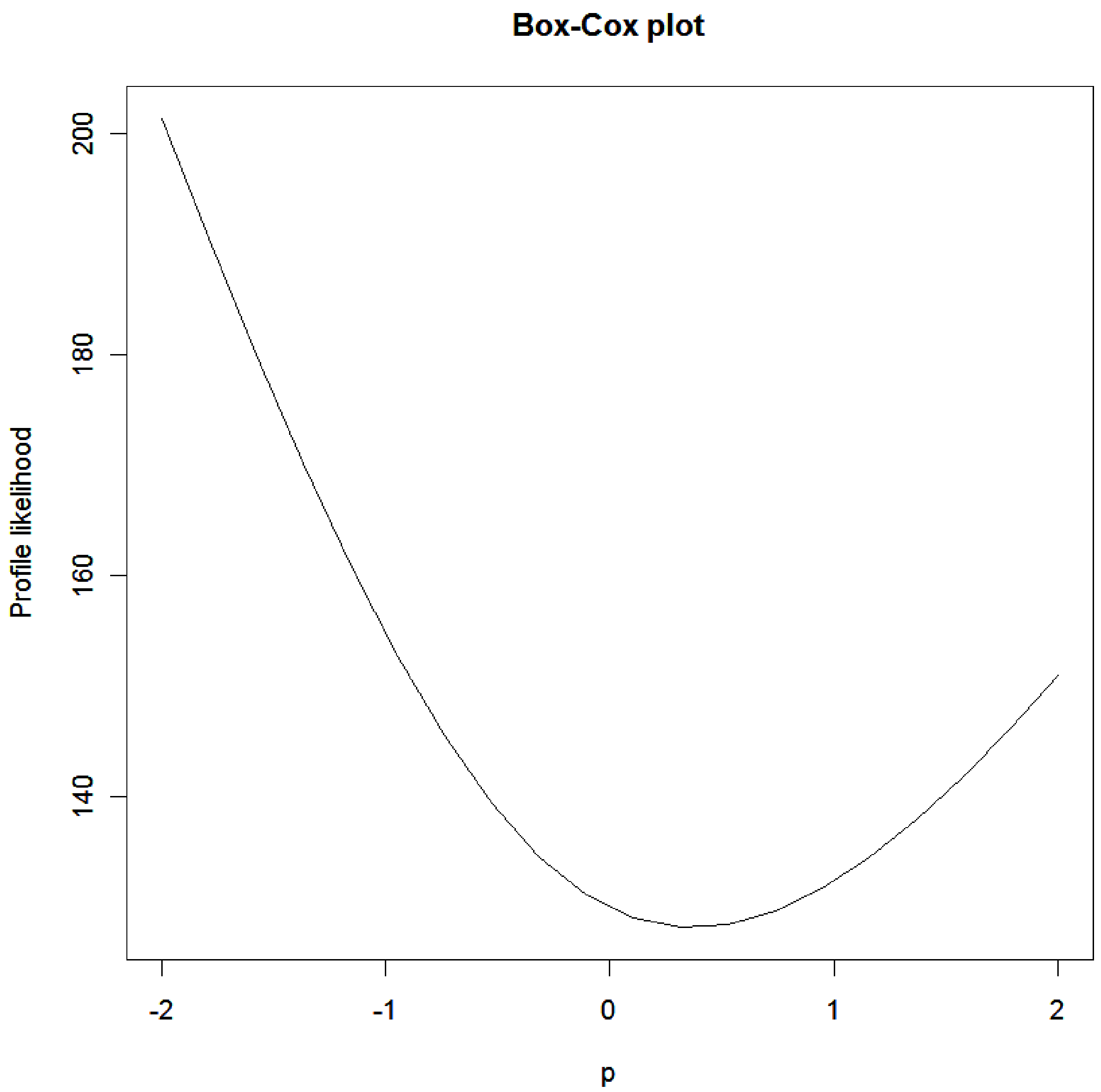
```
Call:
lm(formula = countries.gdp_per_capita ~ countries.oil)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.182 -13.055  -3.073  10.243  36.398 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    19.7754     2.8992   6.821 2.07e-07 ***
countries.oil  -1.2323     0.9175  -1.343     0.19    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 15.56 on 28 degrees of freedom
```

Multiple R-Squared: 0.06053,   Adjusted R-squared: 0.02698
F-statistic: 1.804 on 1 and 28 DF, p-value: 0.19

We have an R^2 of .06053, and it looks as though the oil is NOT a significant factor. However, we need to improve this model. First, we will look at the outliers and see if we can take anything out that might be skewing our data:



However, our Leverage- Residuals plot shows that we do not have any points that we can delete. So now we look to transform the data using the boxcox plot.



From this plot, we can transform the model using:

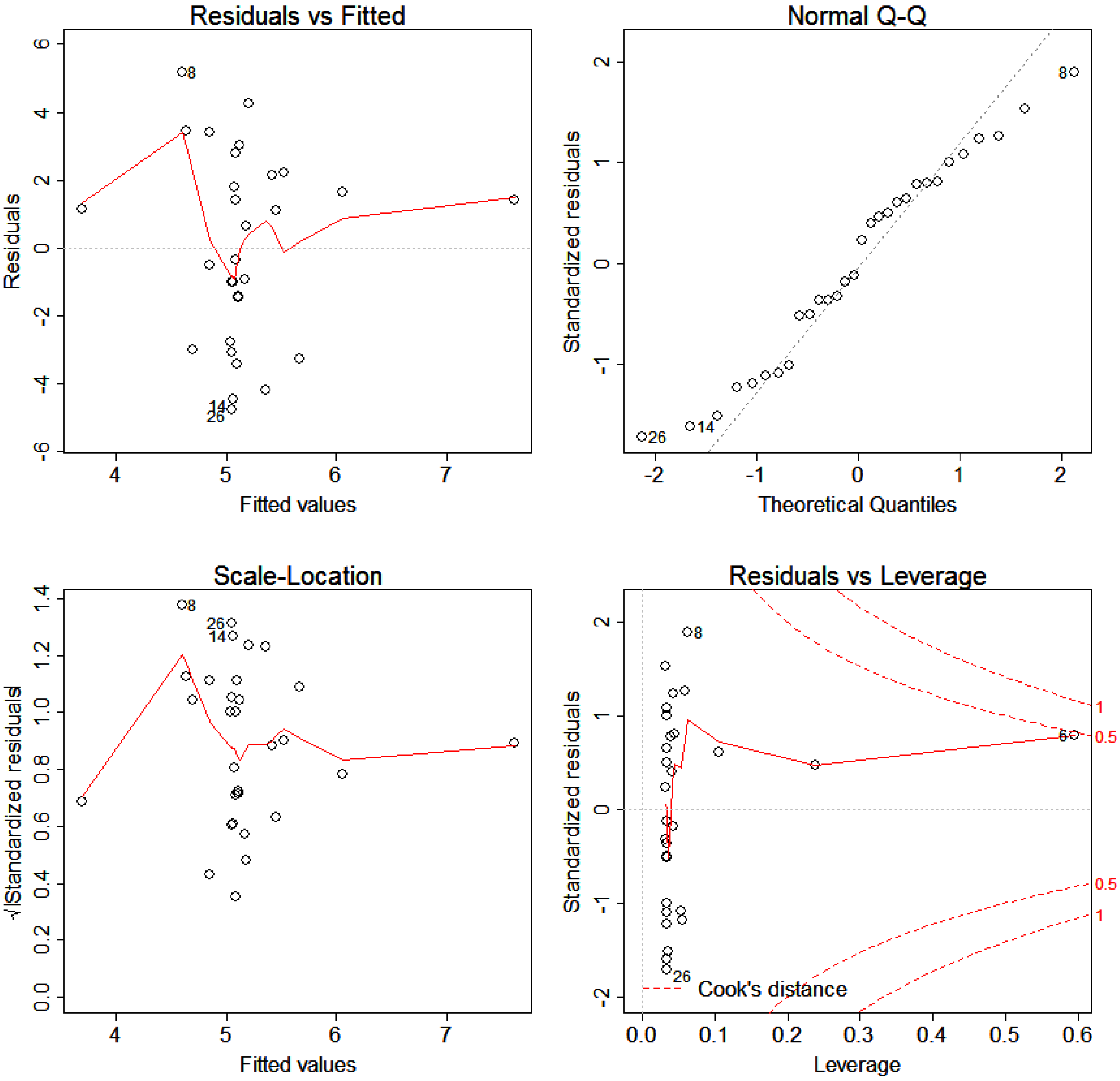
```
boxcox.unemp_rate=lm(((countries.gdp_per_capita^(0.4)) - 1)/(0.4) ~ countries.oil, data=countries)
```

This way we get an even worse R^2:

```
Call:
lm(formula = ((countries.gdp_per_capita^(0.4)) - 1)/(0.4) ~ countries.oil)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.7744 -2.4515  0.1482  2.0668  5.1643 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     5.0487     0.5268   9.583 2.44e-10 ***
countries.oil  -0.1925     0.1667  -1.155    0.258    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

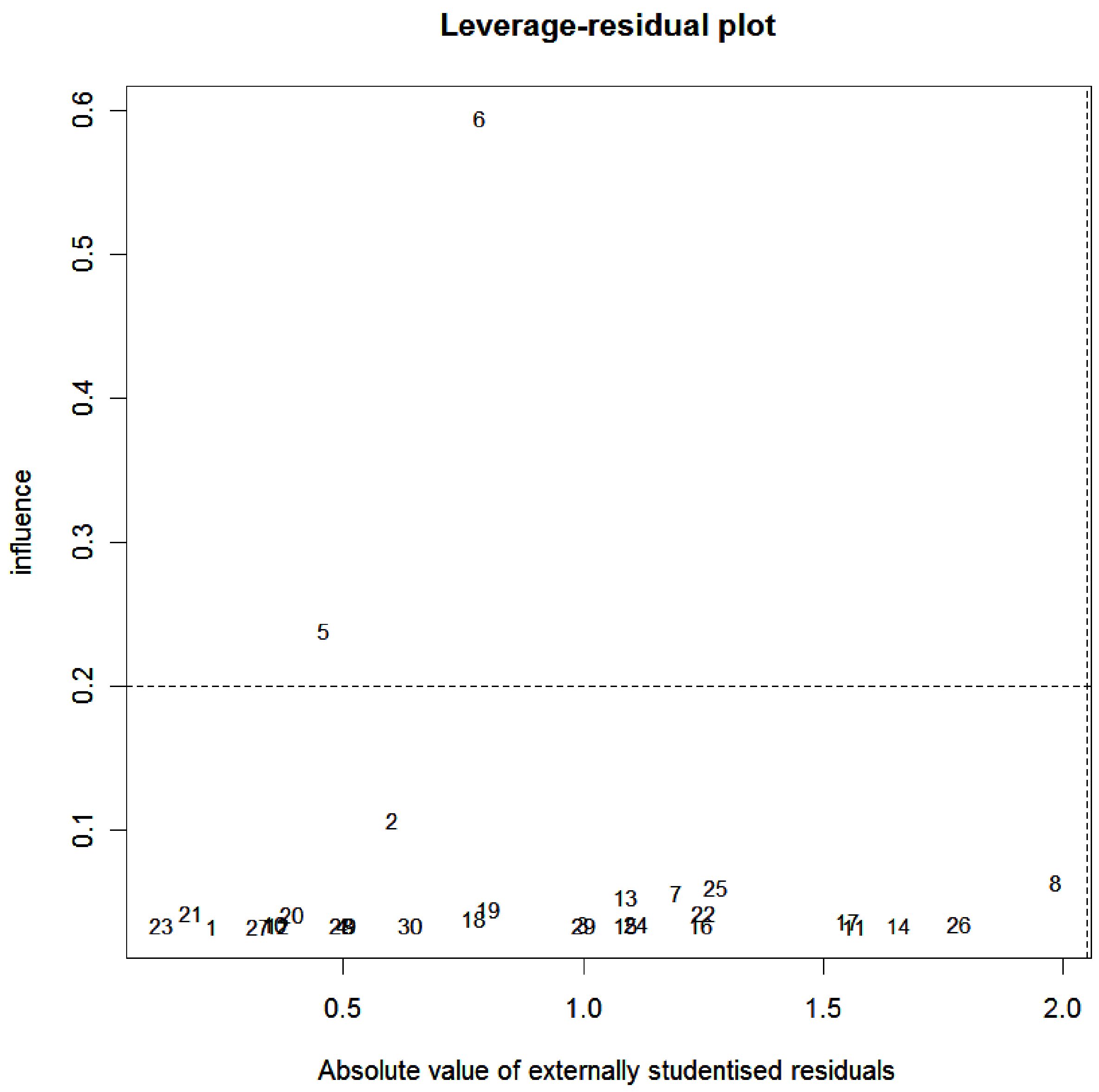
```
Residual standard error: 2.828 on 28 degrees of freedom
Multiple R-Squared: 0.04547,    Adjusted R-squared: 0.01138
F-statistic: 1.334 on 1 and 28 DF,  p-value: 0.2579
```

If we check the plots:

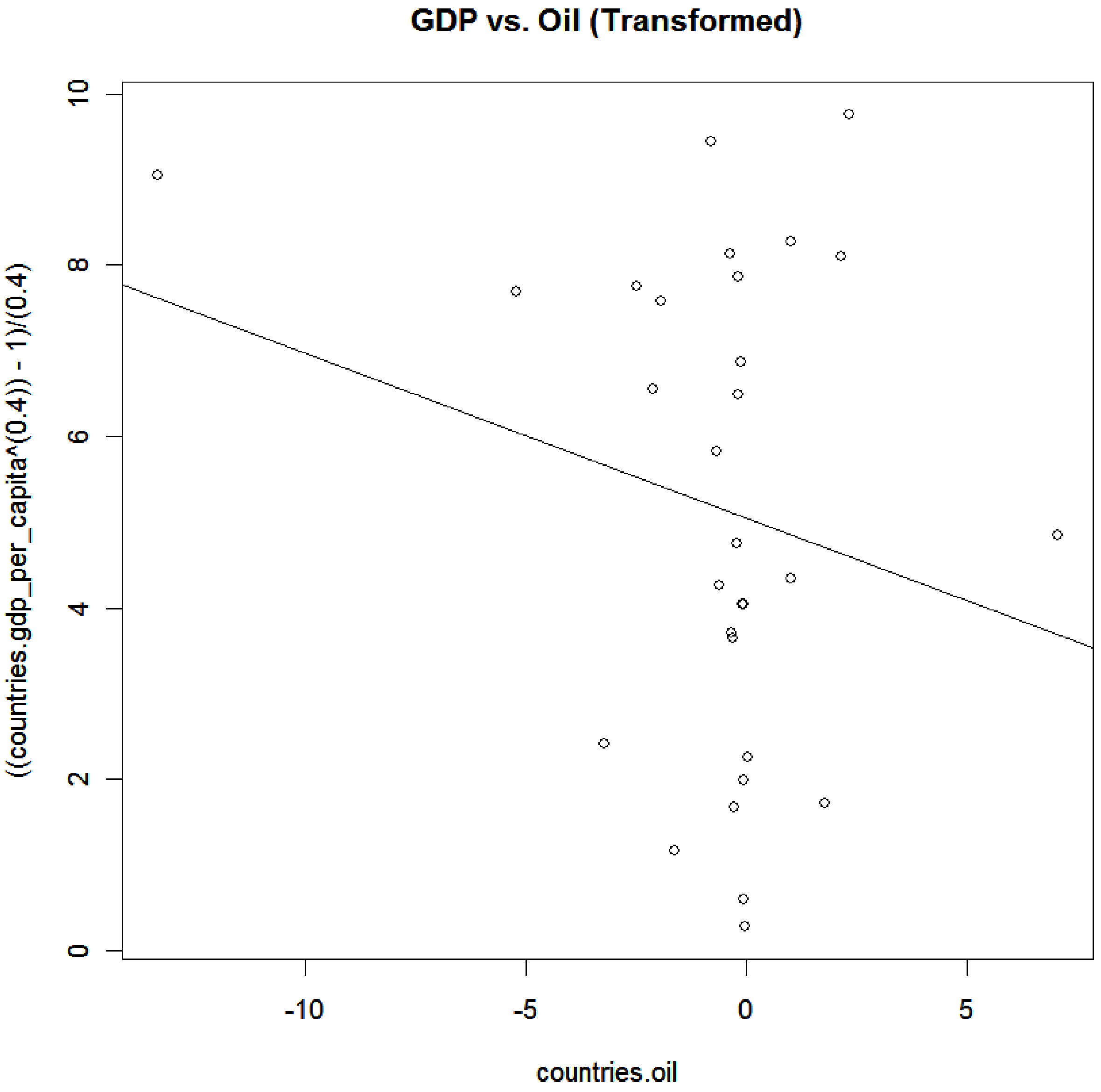


These plots still look pretty bad.

When we run the Leverage-Residuals plot again, we see that there are no outliers in this model either:



Therefore, out final transformed graph is:

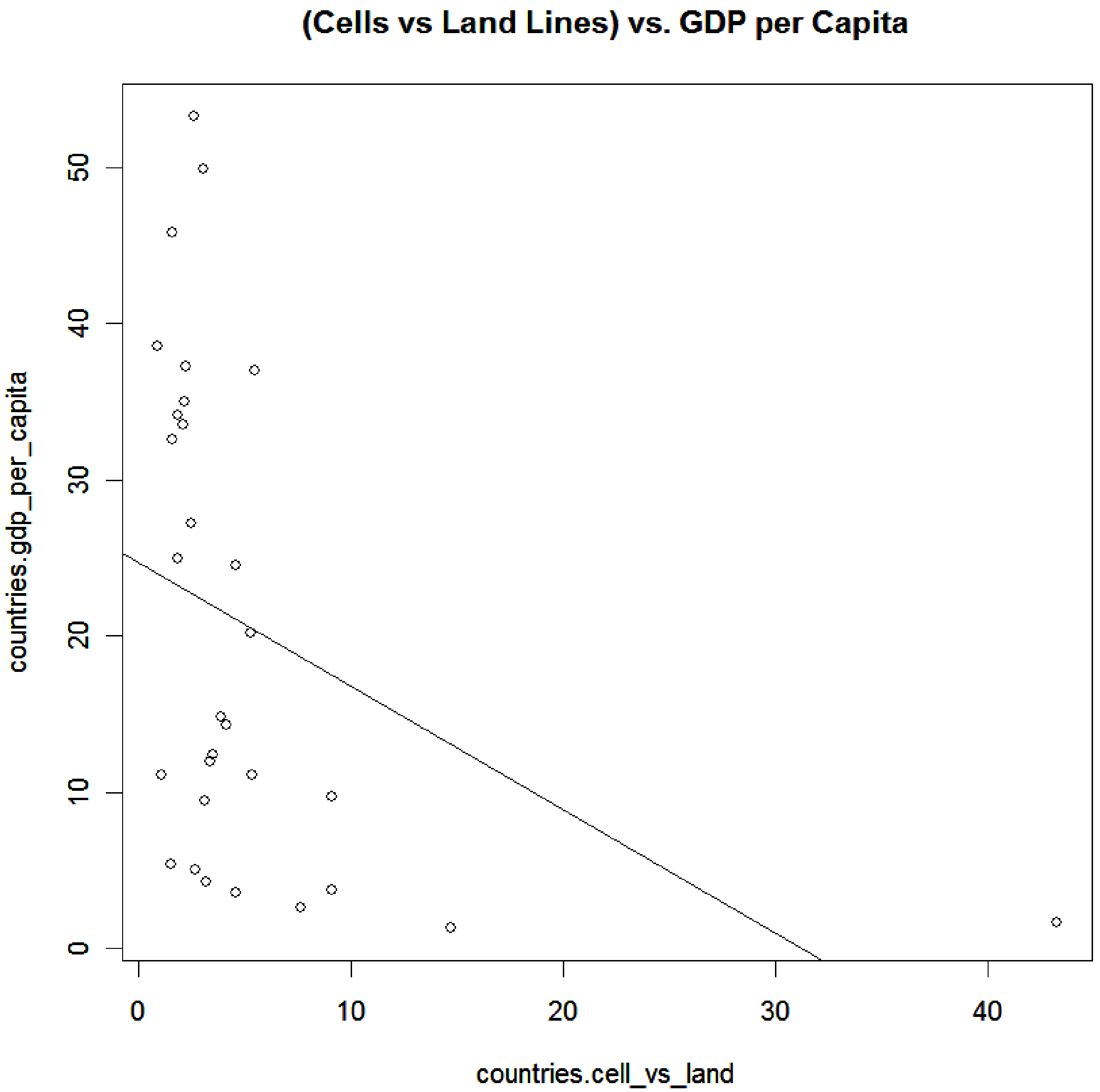


With the final model (using original data): GDP= 19.775 -1.232(oil)

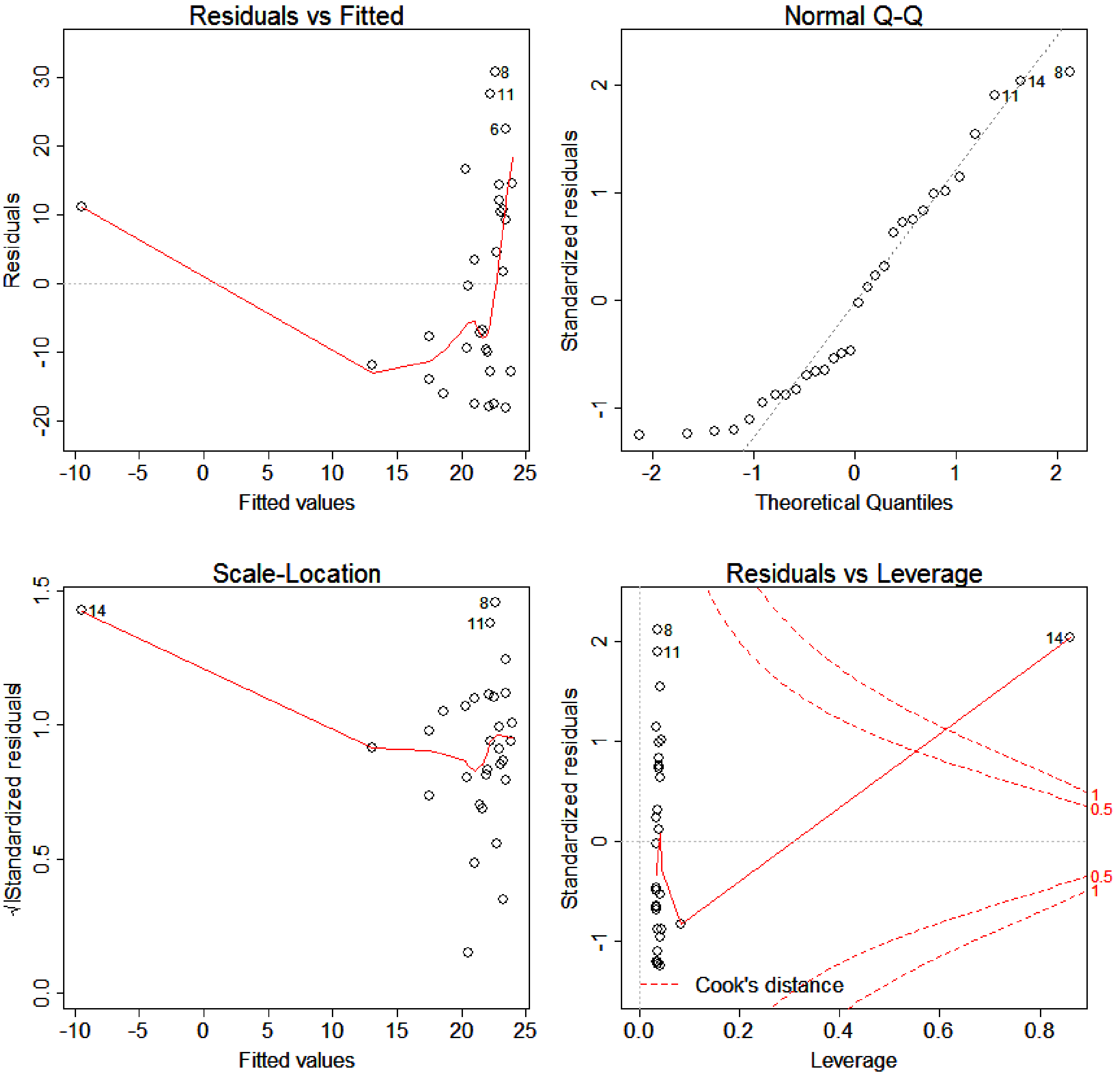
```
Call:
lm(formula = countries.gdp_per_capita ~ countries.oil)

Coefficients:
 (Intercept)  countries.oil
      19.775        -1.232
```

Overall, we once again cannot use this model to predict GDP. The model is not linear, and the data does not have equal variance, even after the transformation, which just made things worse.

Cells vs Landlines



First, in the graph, you can already tell that the data, in its current state, is not linear. It looks like the point to the right is really affecting the plot. We now look at the residuals and Q-Q plot to see how our data looks there.

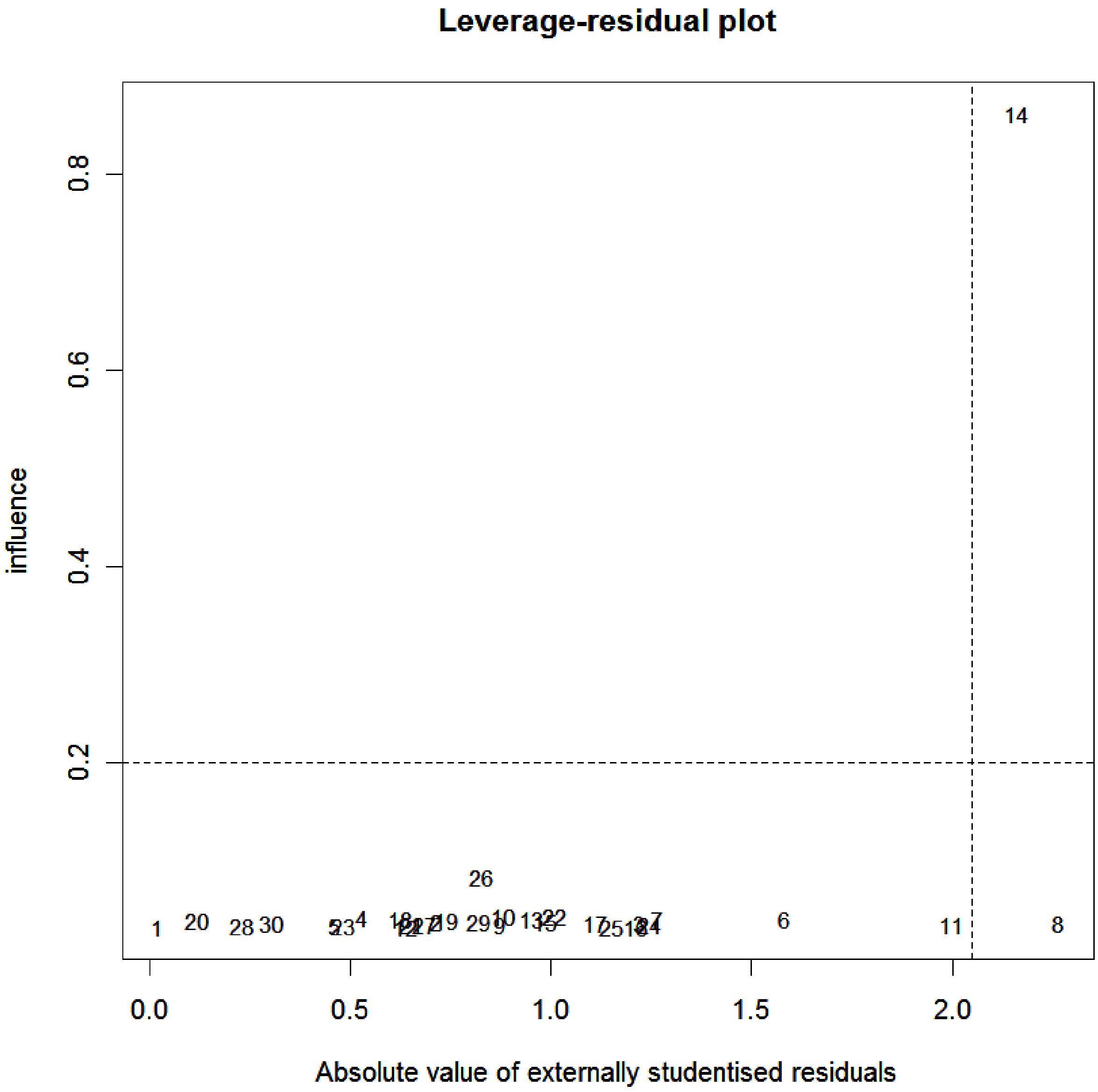


The residuals look awful, and the Q-Q plot does not look much better. When we look at the summary of the data:

```
Call:
lm(formula = countries.gdp_per_capita ~ countries.cell_vs_land)

Residuals:
    Min      1Q  Median      3Q     Max
-18.118 -12.526  -3.583  11.114  30.665

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           24.7038     3.2789   7.534 3.31e-08 ***
countries.cell_vs_land  -0.7919     0.3545  -2.234   0.0337 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.79 on 28 degrees of freedom
Multiple R-Squared: 0.1513,     Adjusted R-squared: 0.1209
```

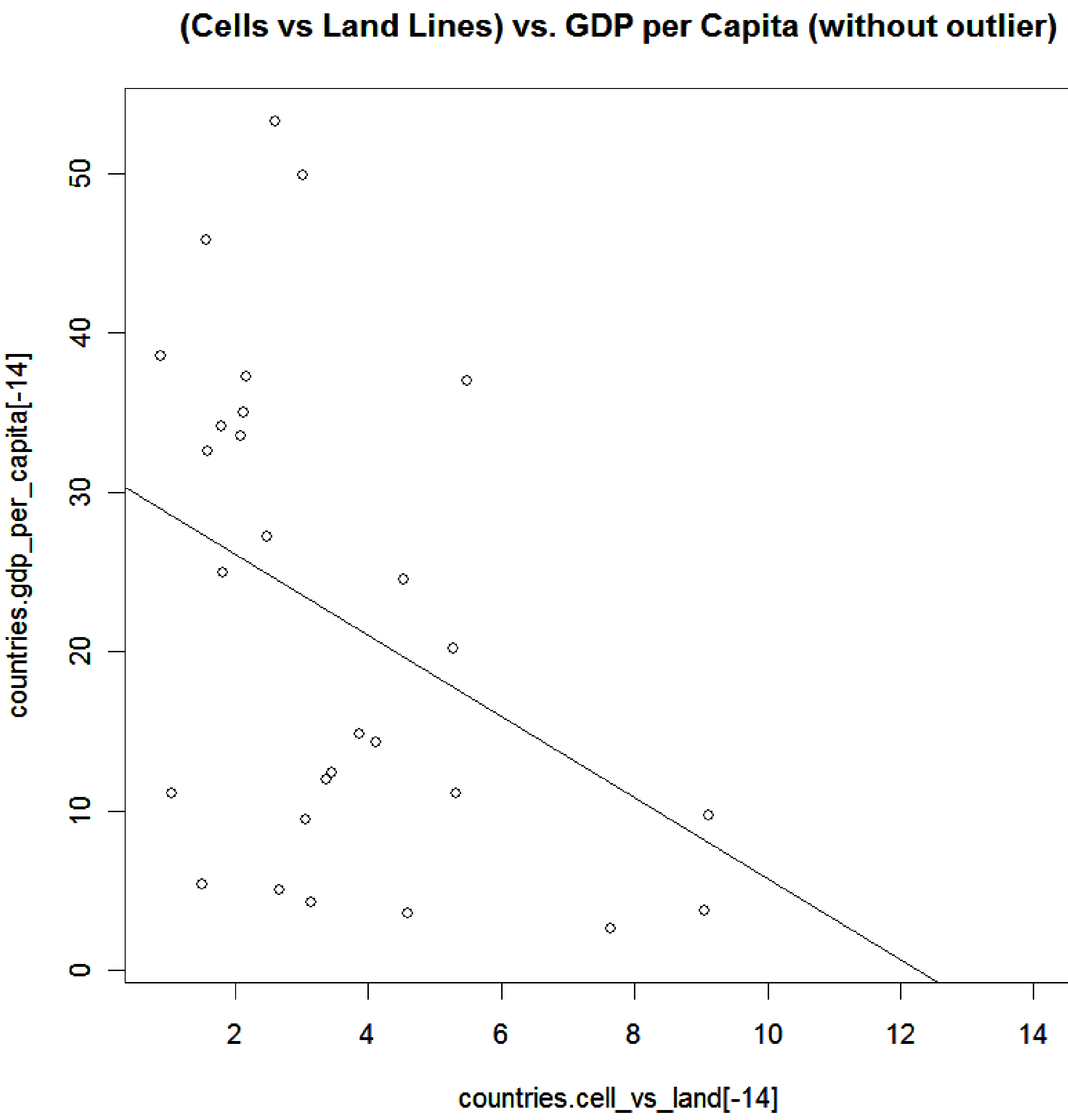
F-statistic: 4.99 on 1 and 28 DF, p-value: 0.03366

We have an R^2 of .1513. We need to improve this model. First, we will look at the outliers and see if we can take anything out that might be skewing our data.

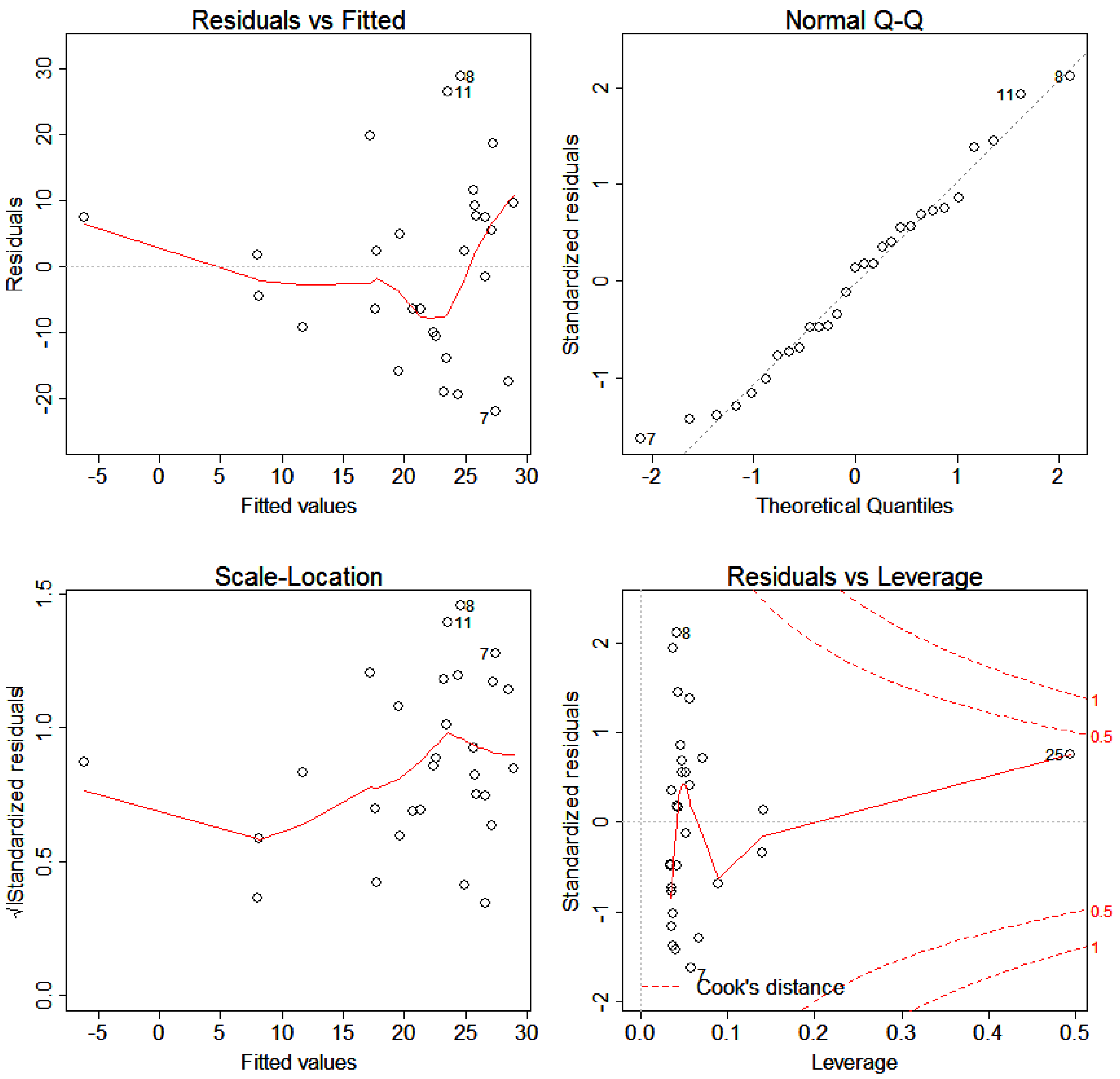


Our Leverage- Residuals plot shows that we do have a point that is considered an influential outlier, and therefore we need to get rid of it.

Our new plot looks much better with that outlier deleted:



When we check out the other plots:

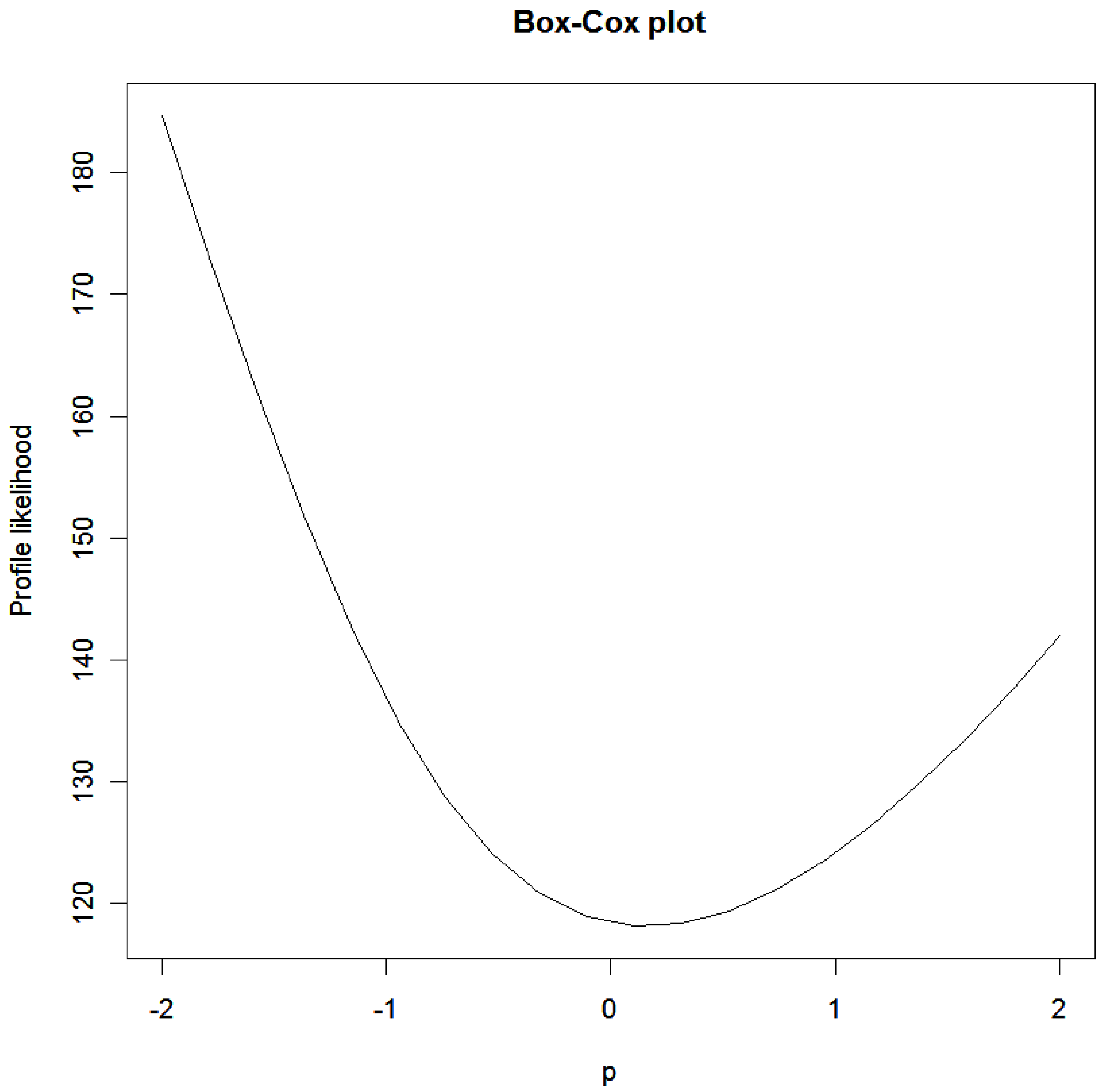


The residuals, although better, still look pretty bad. The normal Q-Q plot looks pretty good. Lets look at the summary:

```
Call:
lm(formula = countries.gdp_per_capita[-14] ~ countries.cell_vs_land[-14])

Residuals:
    Min      1Q  Median      3Q     Max 
-22.016 -10.027   1.695   7.606  28.724 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  31.2351     4.3172   7.235 8.81e-08 ***
countries.cell_vs_land[-14]  -2.5494     0.8789  -2.901  0.00732 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 13.91 on 27 degrees of freedom
Multiple R-Squared: 0.2376,     Adjusted R-squared: 0.2094 
```

```
F-statistic: 8.414 on 1 and 27 DF,  p-value: 0.007318
```

Although our R^2 is improved, we should try to see if we can change it more with a Box-Cox plot:



From this plot, we can transform the model using:

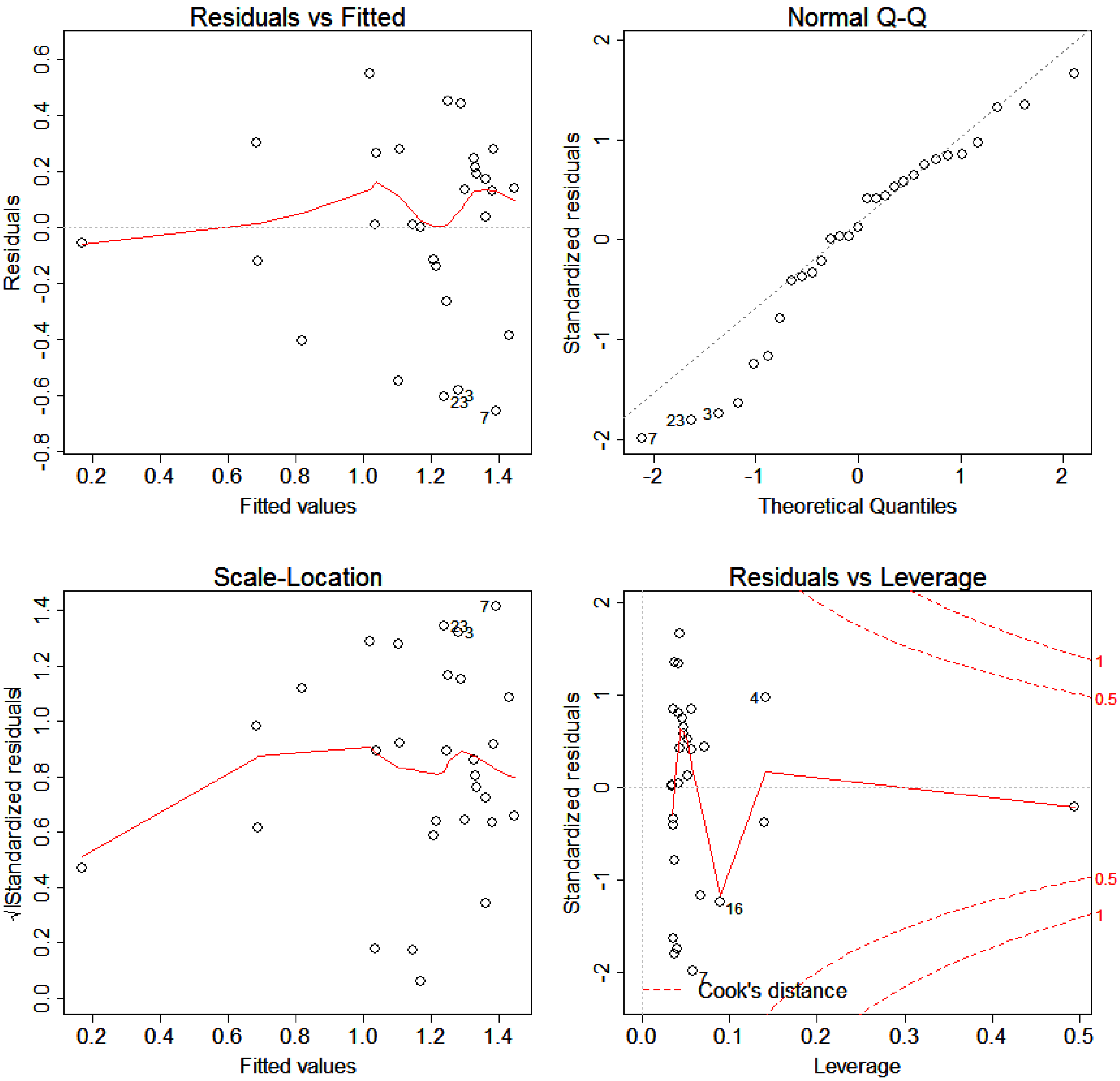
```
boxcox.cell_vs_land2=lm(log(countries.gdp_per_capita[-14], 10) ~ countries.cell_vs_land[-14],
data=countries)
```

This way we get a better R^2, and the p value is significant:

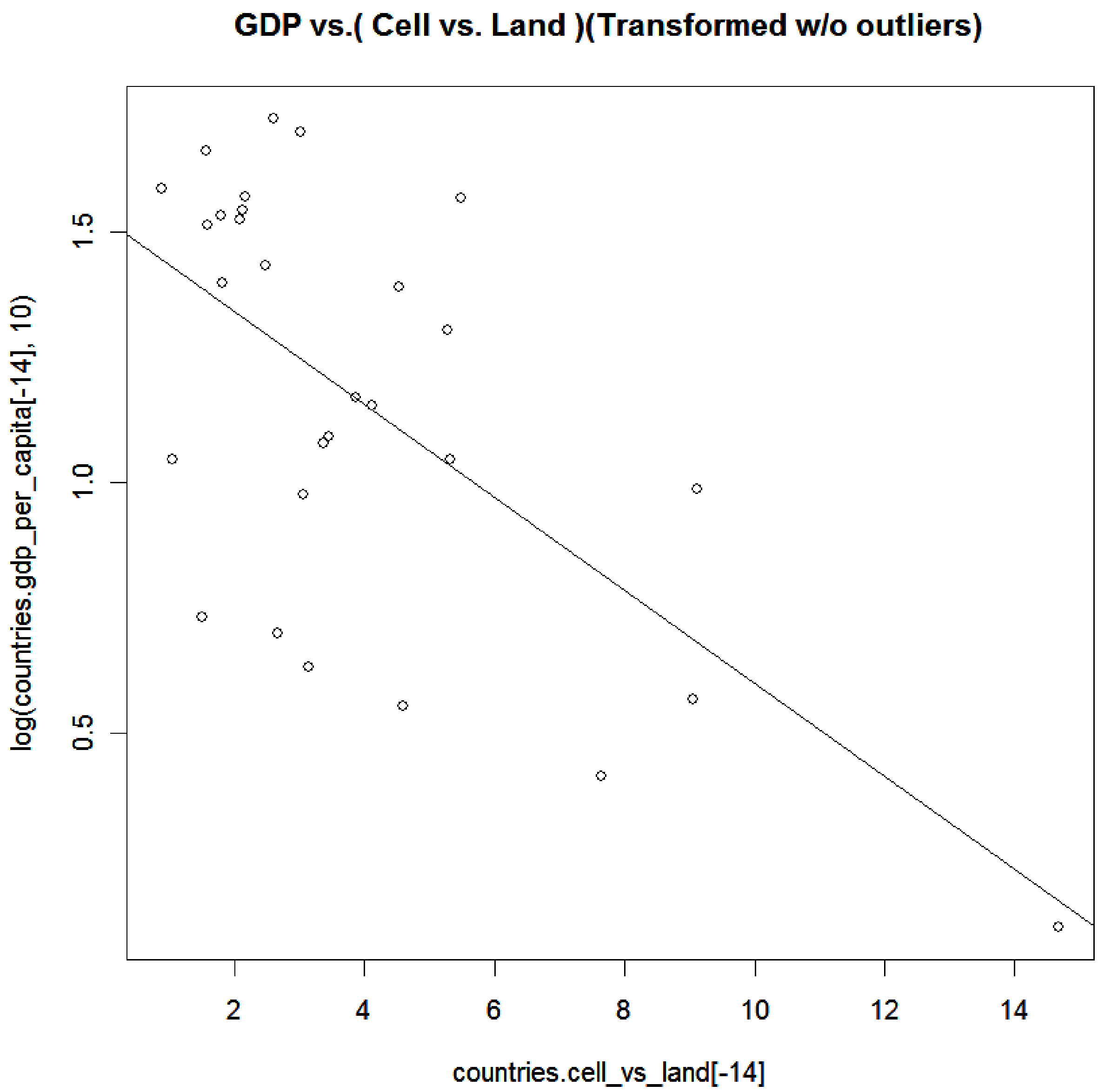
```
Call:
lm(formula = log(countries.gdp_per_capita[-14], 10) ~ countries.cell_vs_land[-14],
    data = countries)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65675 -0.13668  0.03866  0.24564  0.54881

Coefficients:
```

```
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.52811    0.10538  14.502 2.91e-14 ***
countries.cell_vs_land[-14] -0.09276    0.02145  -4.324 0.000187 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3395 on 27 degrees of freedom
Multiple R-Squared: 0.4092,     Adjusted R-squared: 0.3873
F-statistic:  18.7 on 1 and 27 DF,  p-value: 0.0001867
```

And better plots:



The residuals, although not perfect, look better than before. The Q-Q plot looks less linear however, which means the data is not as normal, which could cause problems in our final model.



This final plot also looks better than before. Therefore, this model should be used rather than the original.

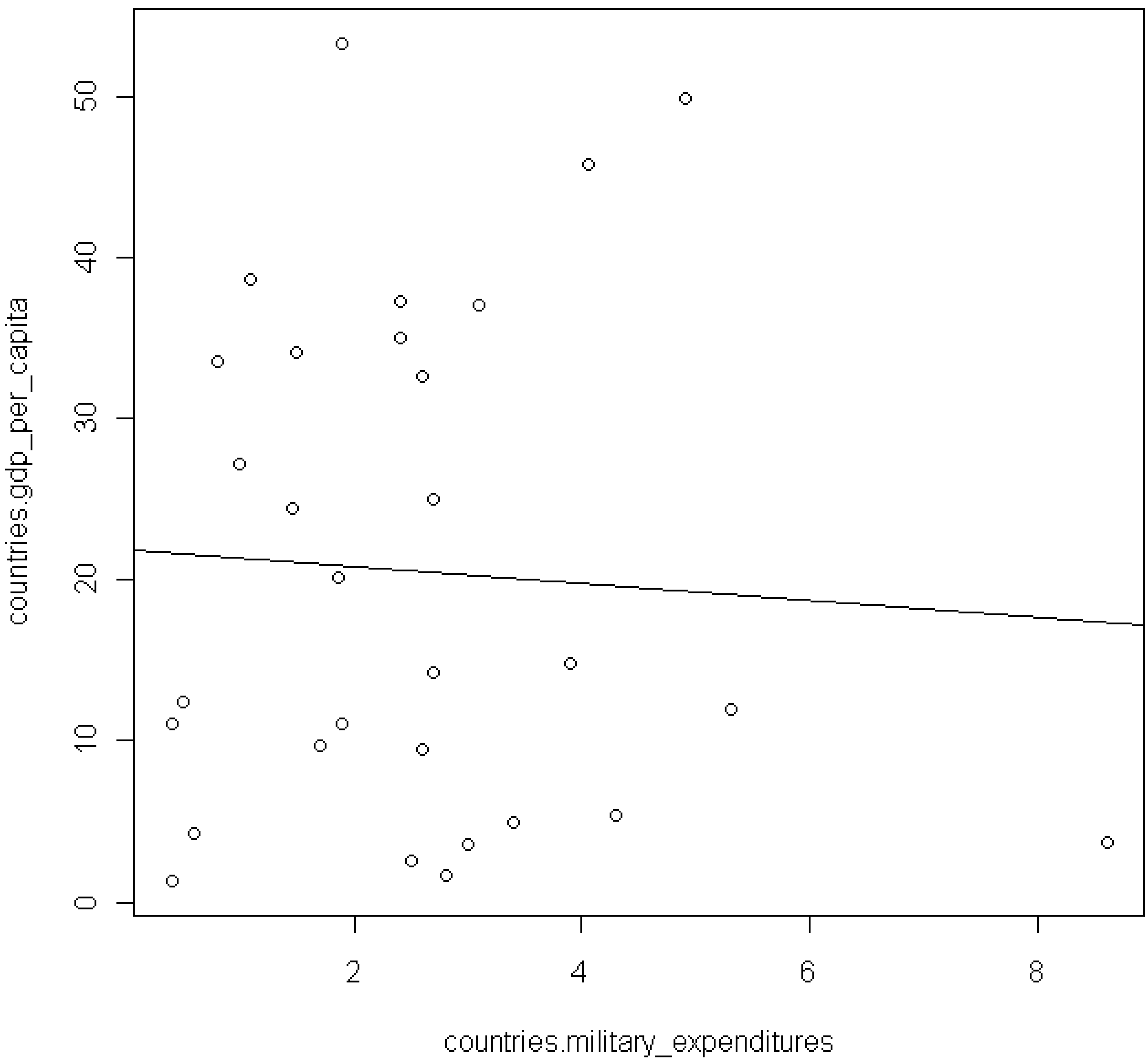
With the final model: GDP= 1.52811 - .0928(cells vs landlines)

```
Call:
lm(formula = log(countries.gdp_per_capita[-14], 10) ~ countries.cell_vs_land[-14],     data = countries)

Coefficients:
        (Intercept)  countries.cell_vs_land[-14]
           1.52811                  -0.09276
```

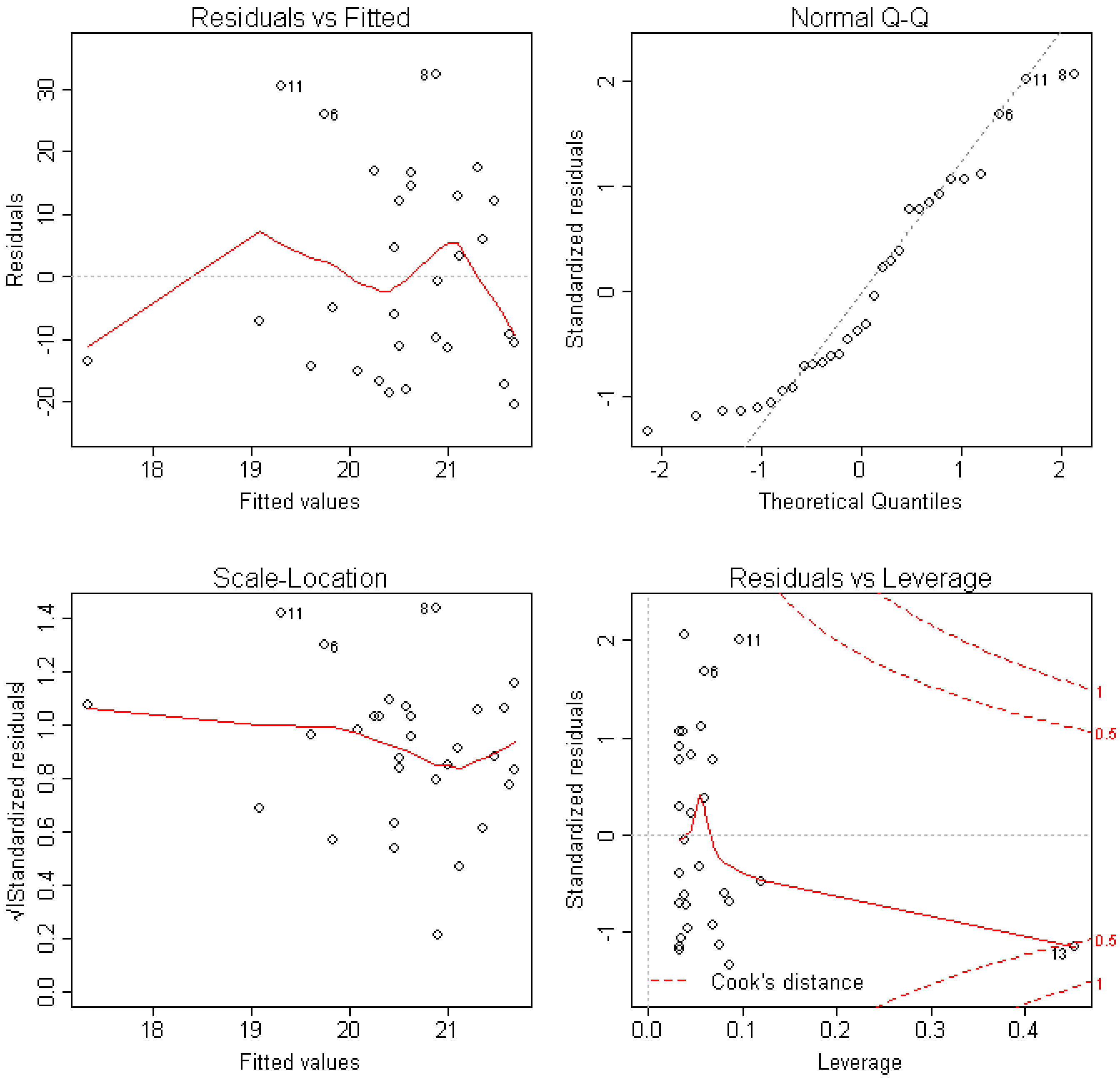
Overall, we pretty much cannot use this model even after the transformation. The Q-Q plot looks pretty curved at some points at the end.

Military Expenditures

```
> plot(countries.military_expenditures, countries.gdp_per_capita)
> model_mil_expend = lm(countries.gdp_per_capita ~ countries.military_expenditures)
> abline(model_mil_expend, lty = 1)
```



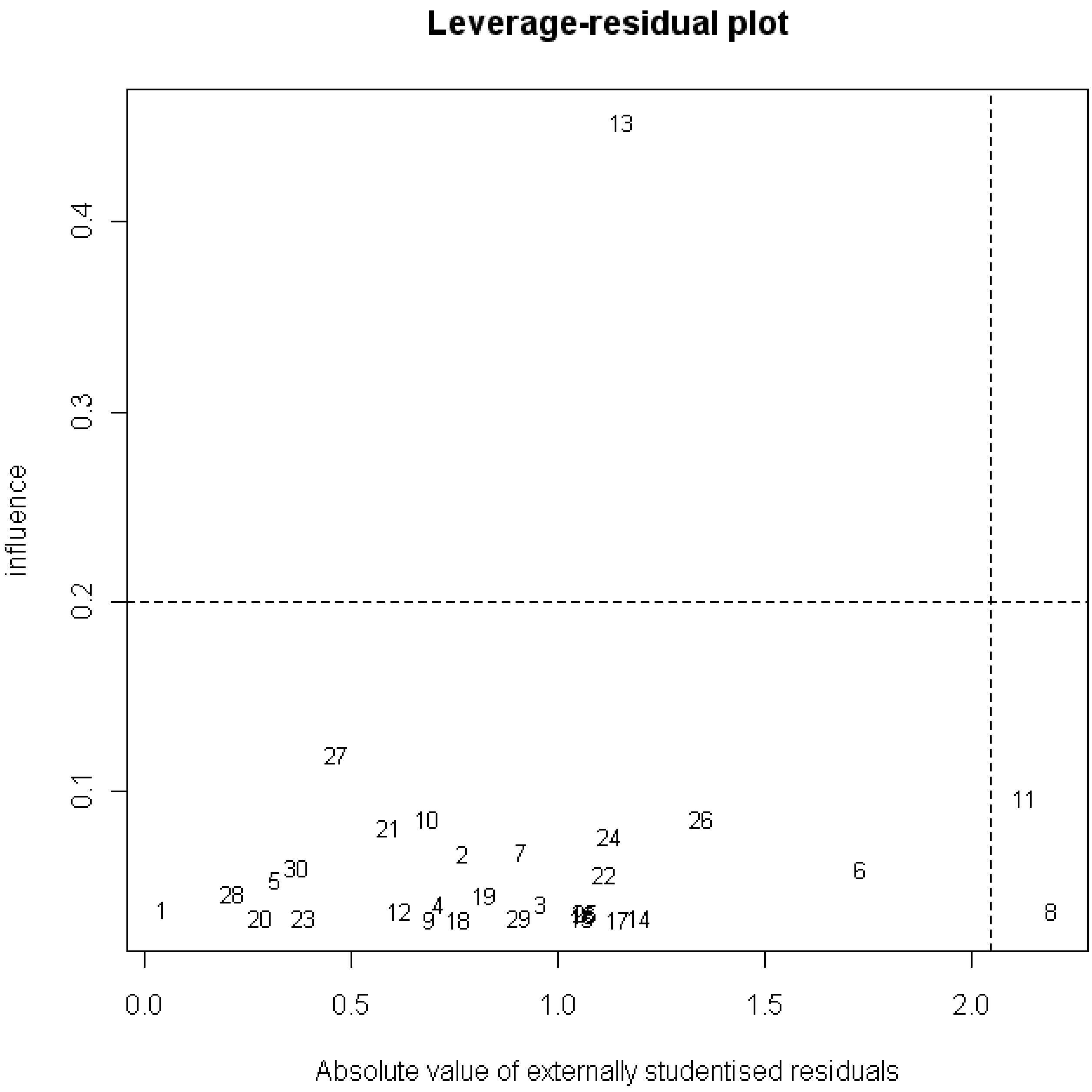
```
> par(mfrow=c(2,2),mex=0.6)
> plot(model_mil_expend)
> par(mfrow=c(1,1),mex=1)
```



This is a very bad model for military expenditures. It doesn't pass the conditions for regression; the data isn't linear, residuals are not random, the Q-Q plot is curved, and there are a lot of outliers.

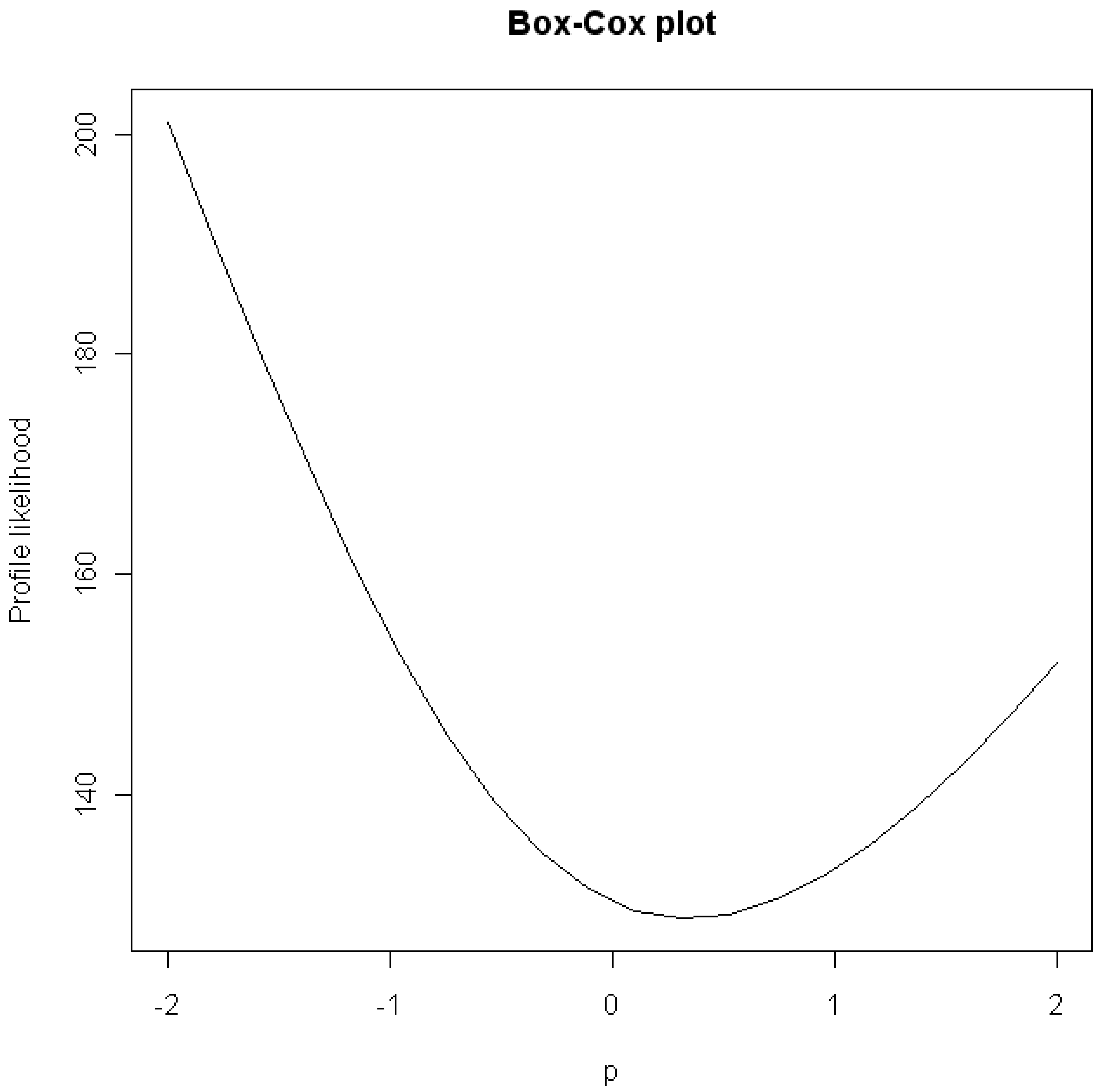
Looking at the leverage-residual plot, there are no points to remove from this data:

```
> lrplot(model_mil_expend)
```

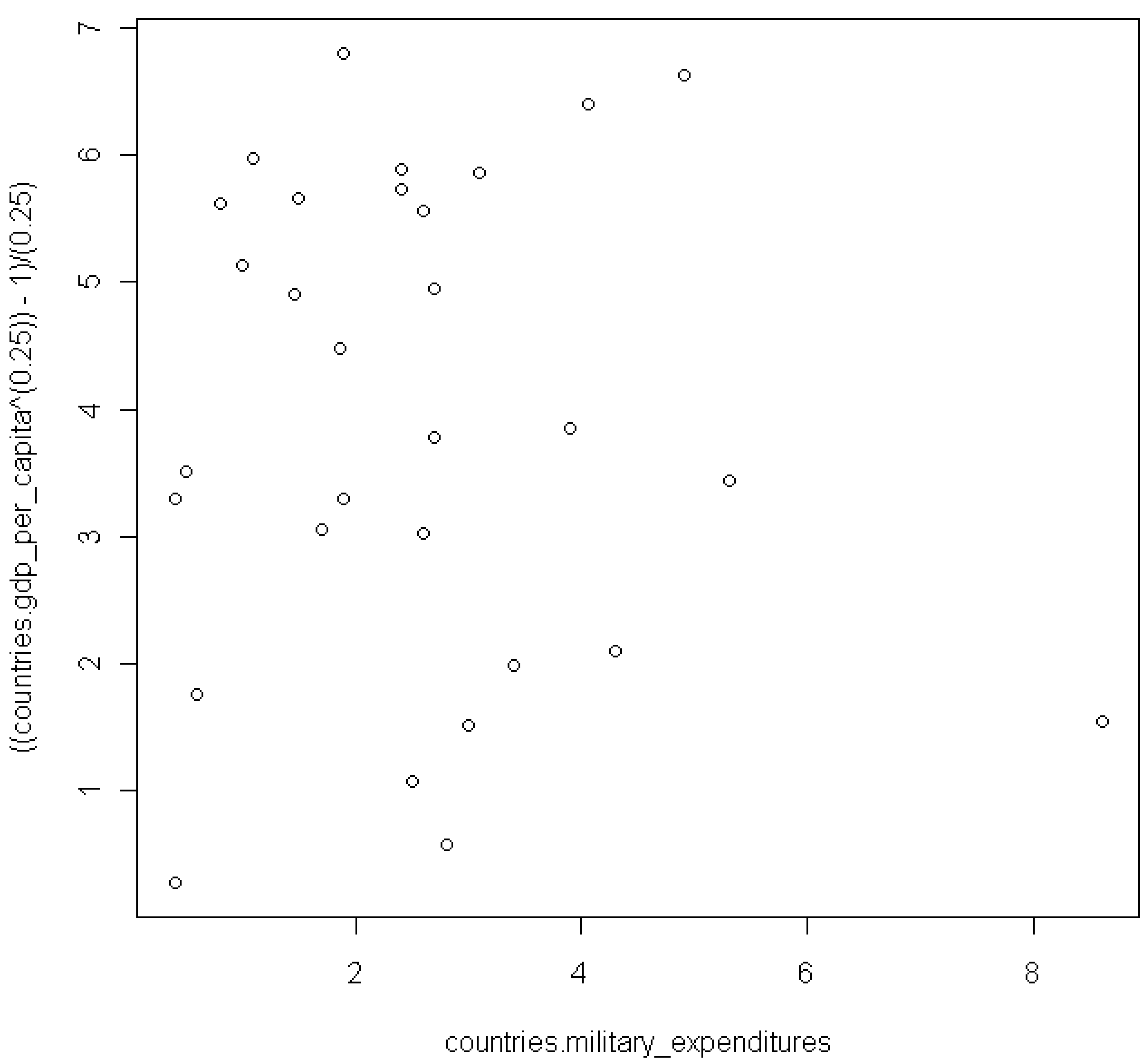


Trying a boxcox transformation to improve the model:

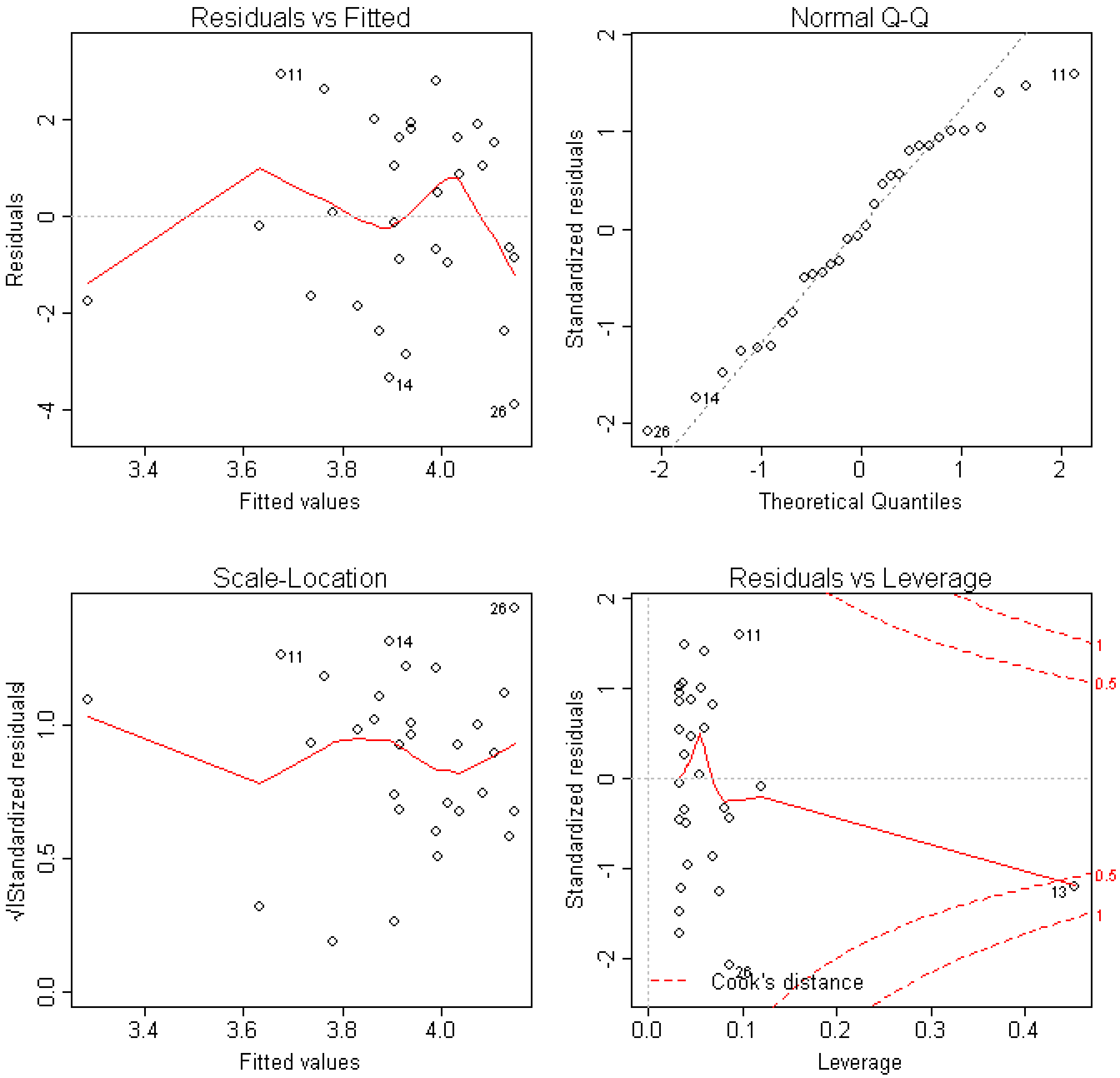
```
> boxcoxplot(countries.gdp_per_capita ~ countries.military_expenditures, data=factors.df)
```



```
> plot(countries.military_expenditures, ((countries.gdp_per_capita^(.25))-1)/(.25))
```



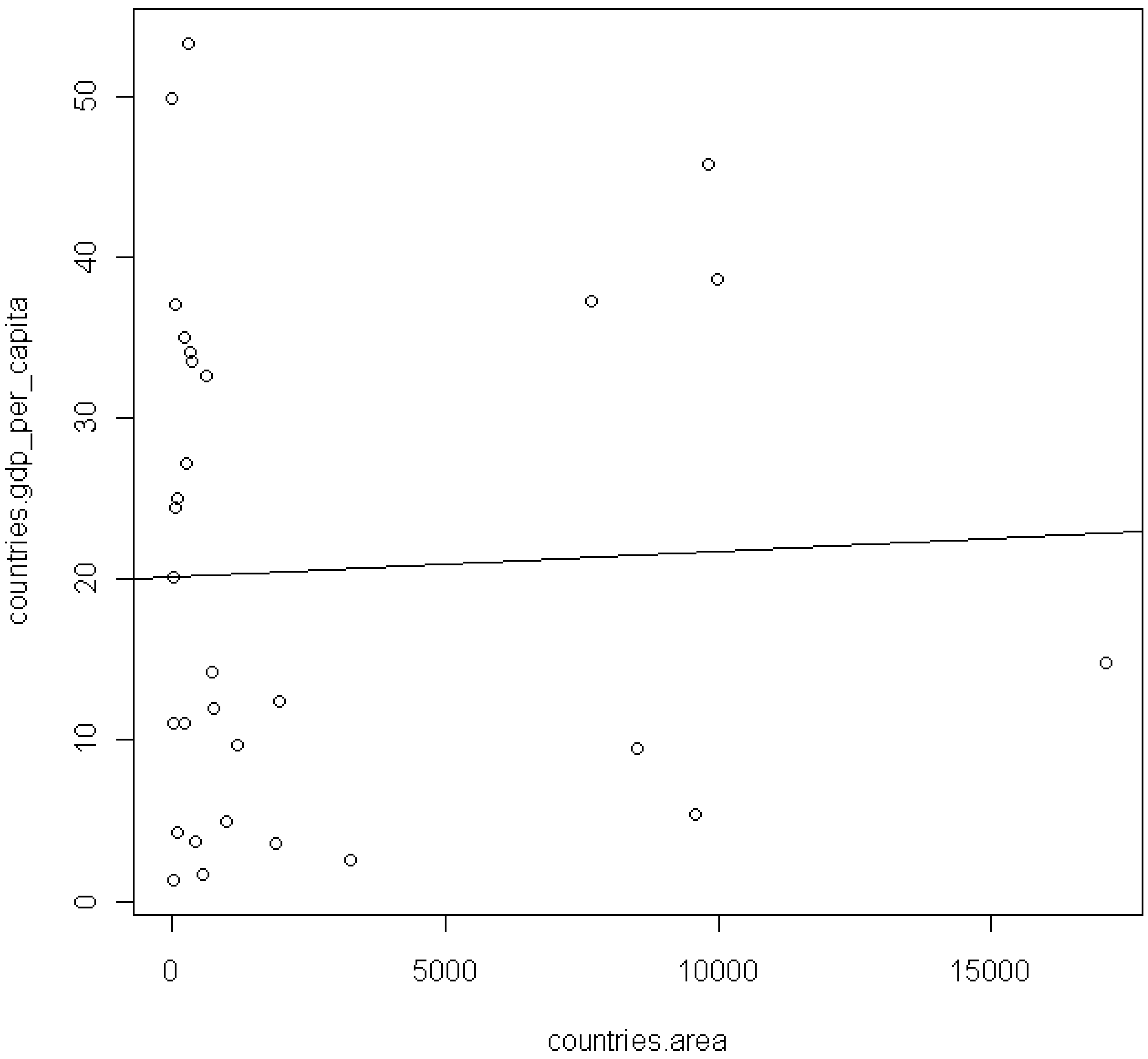
```
> par(mfrow=c(2,2),mex=0.6)
> plot(lm(((countries.gdp_per_capita^(.25))-1)/(.25) ~ countries.military_expenditures))
> par(mfrow=c(1,1),mex=1)
```



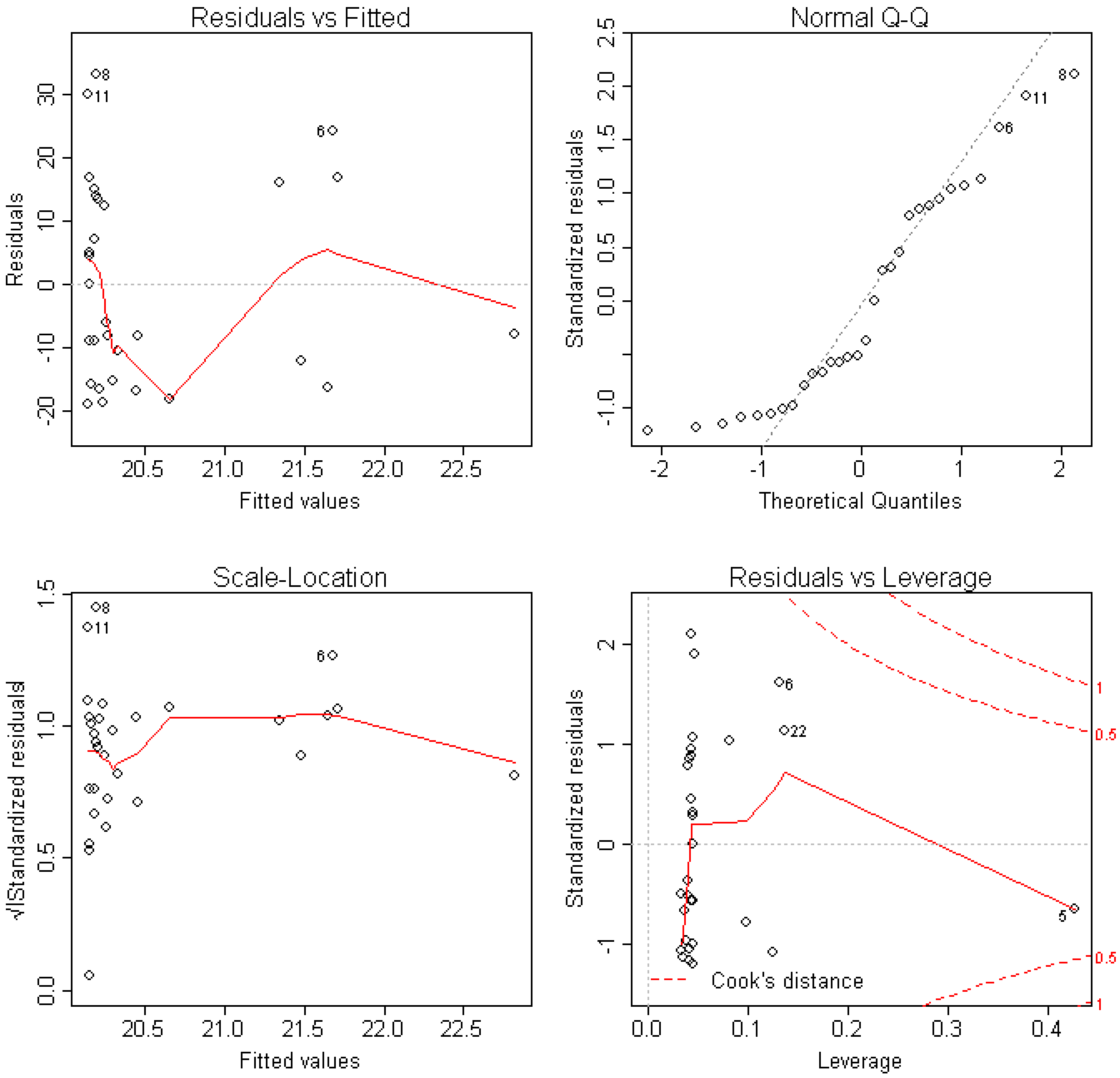
Looking at the new plot after the boxcox transformation, the data is still not linear and we still can't do regression on it.

Area

```
> plot(countries.area, countries.gdp_per_capita)
> model_area = lm(countries.gdp_per_capita ~ countries.area)
> abline(model_area, lty = 1)
```



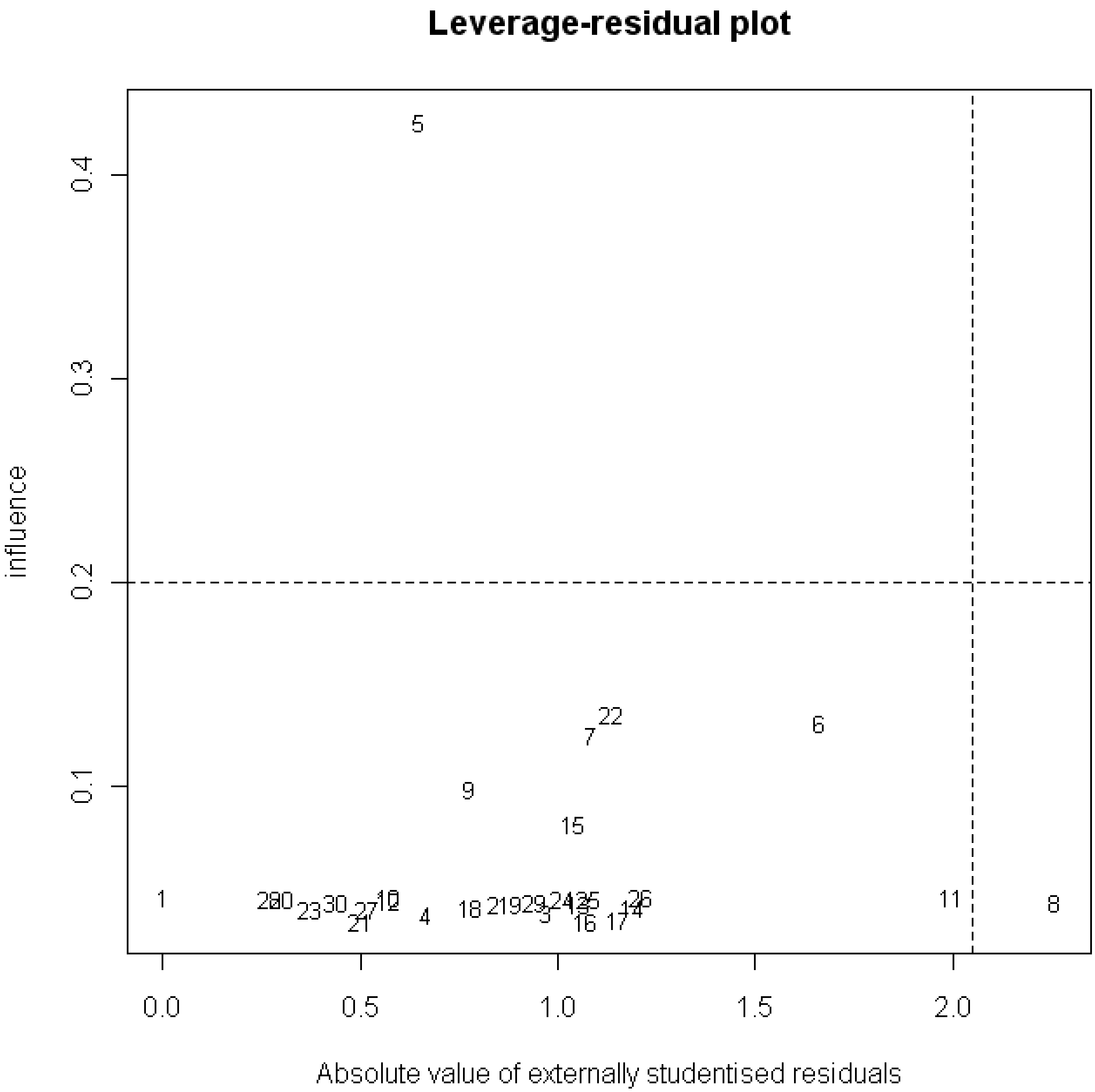
```
> par(mfrow=c(2,2),mex=0.6)
> plot(model_area)
> par(mfrow=c(1,1),mex=1)
```



This is a very bad model that doesn't pass the conditions for regression. The data is not linear, residuals are not random, the Q-Q plot is not linear, and there a lot of outliers.

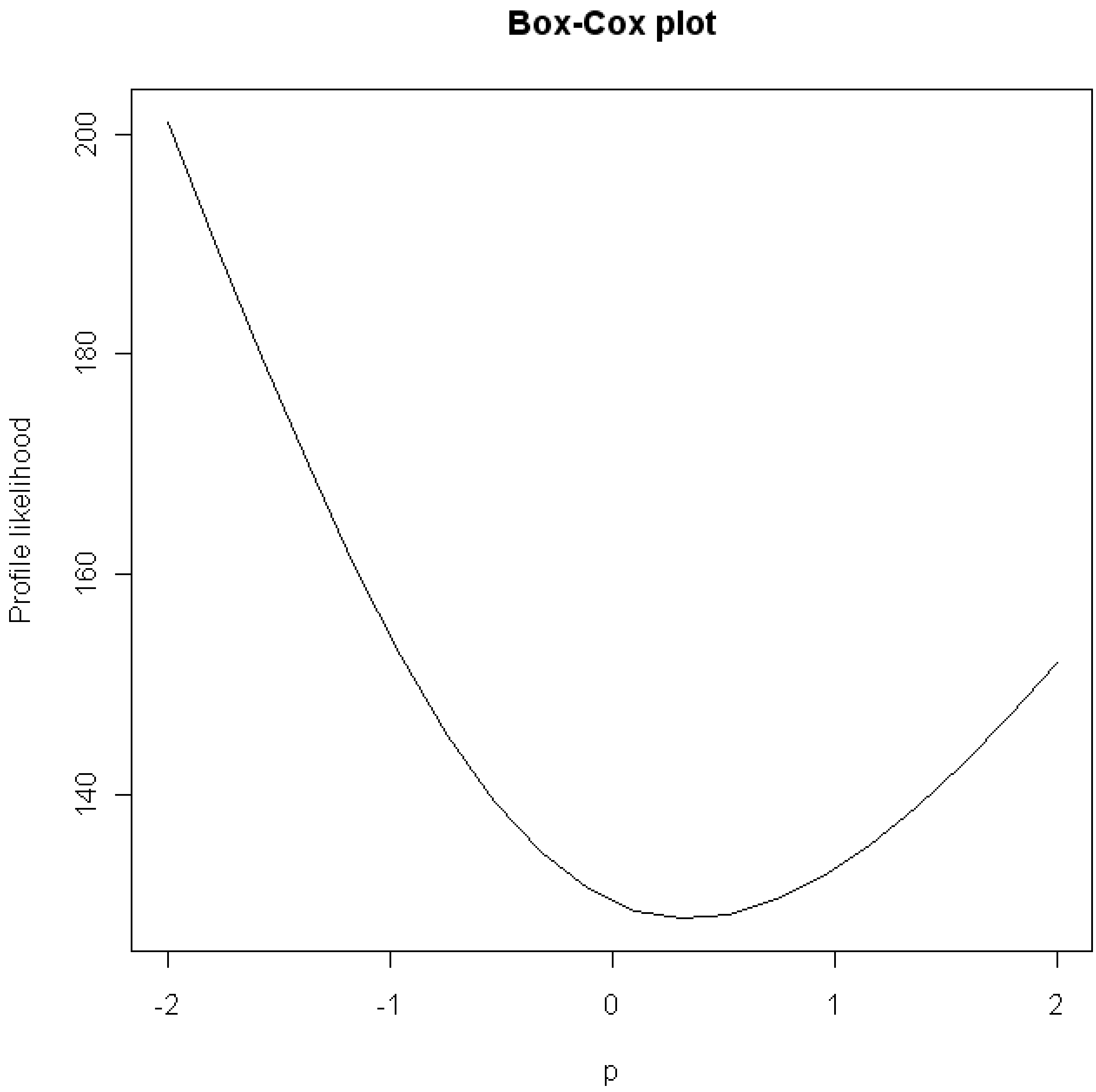
Looking at the leverage-residual plot, there are no points to remove from this data:

```
> lrplot(model_area)
```

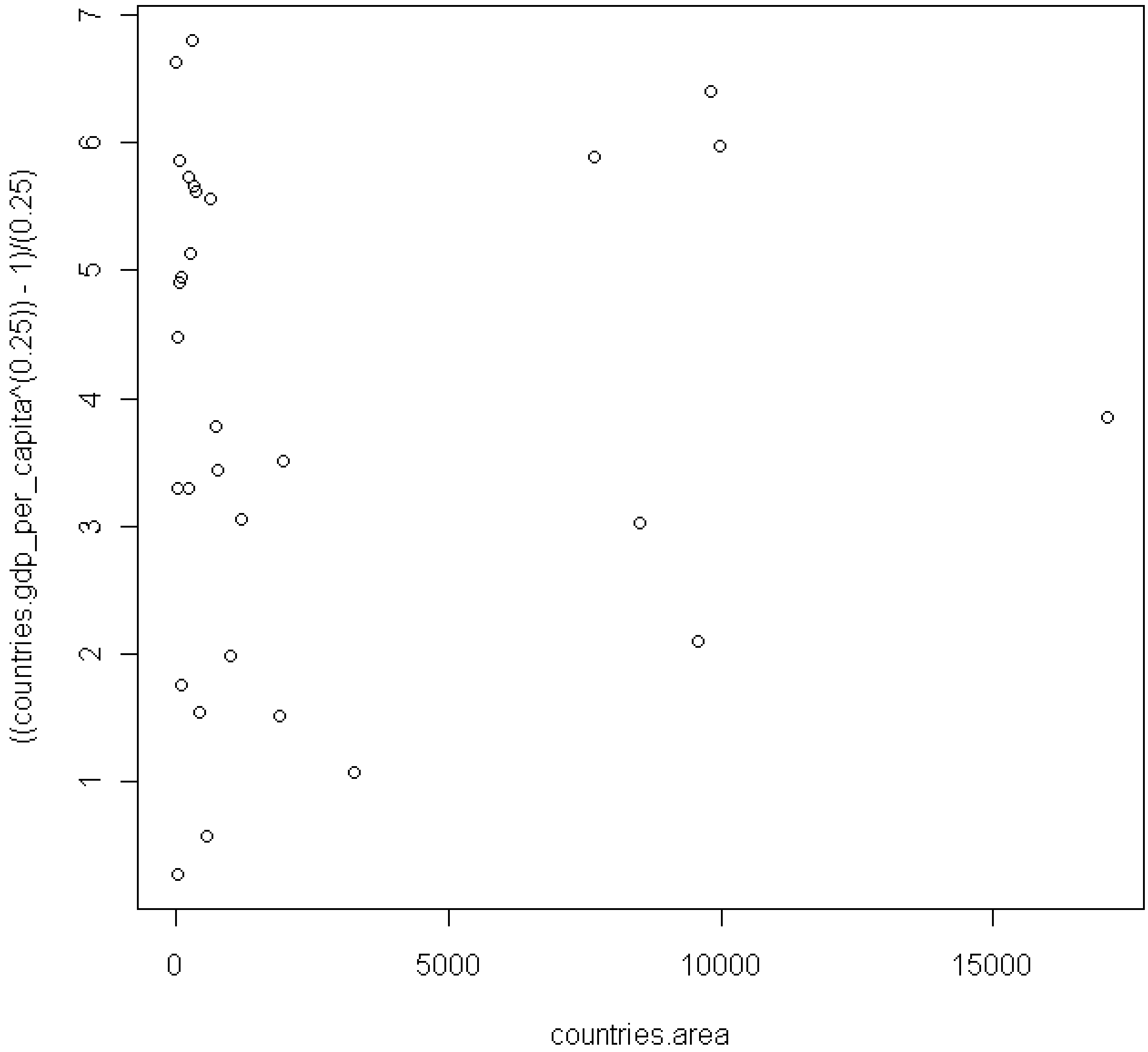


Trying a boxcox transformation to improve the model:

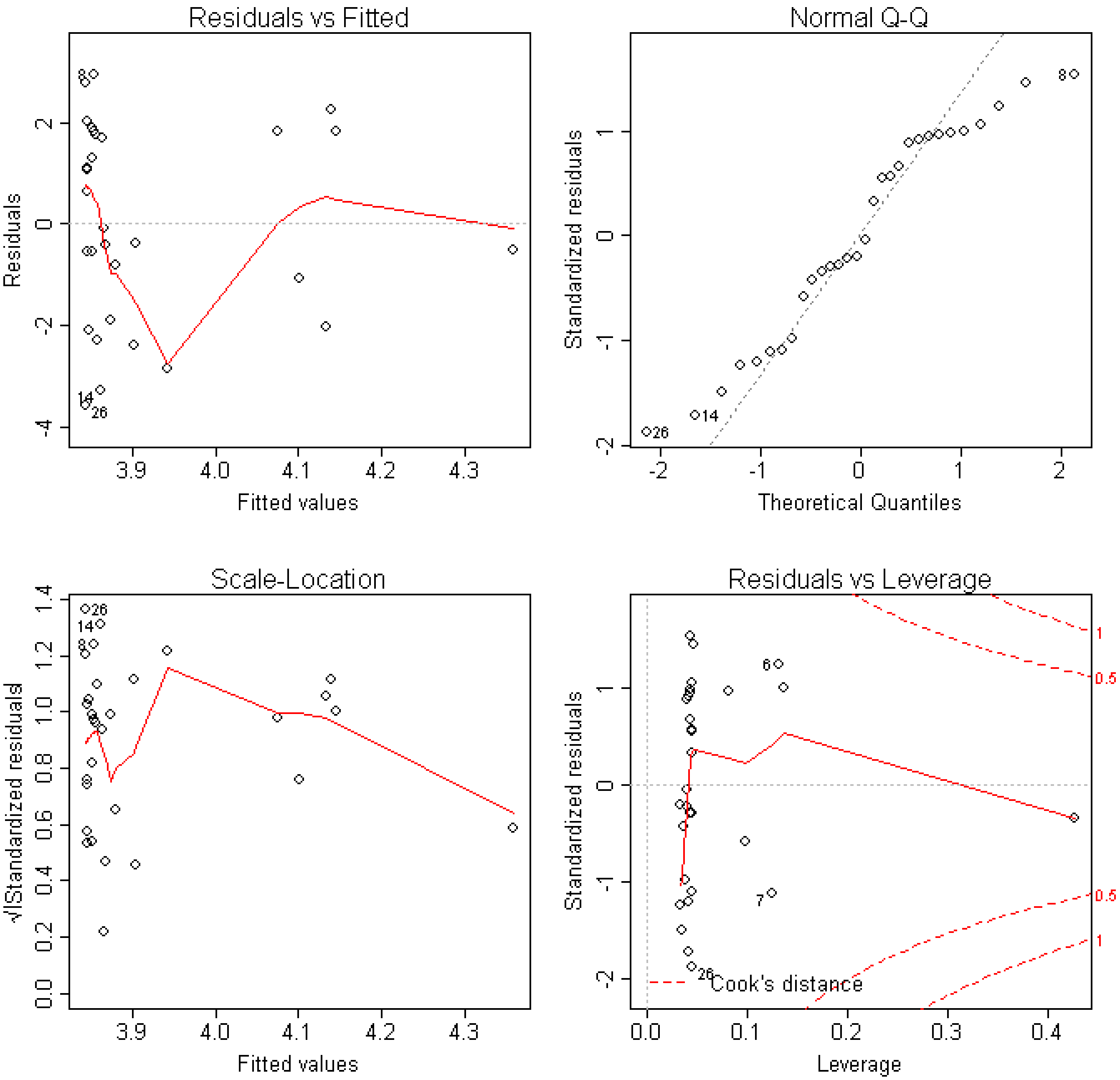
```
> boxcoxplot(countries.gdp_per_capita ~ countries.area, data=factors.df)
```



```
> plot(countries.area, ((countries.gdp_per_capita^(.25))-1)/(.25))
```



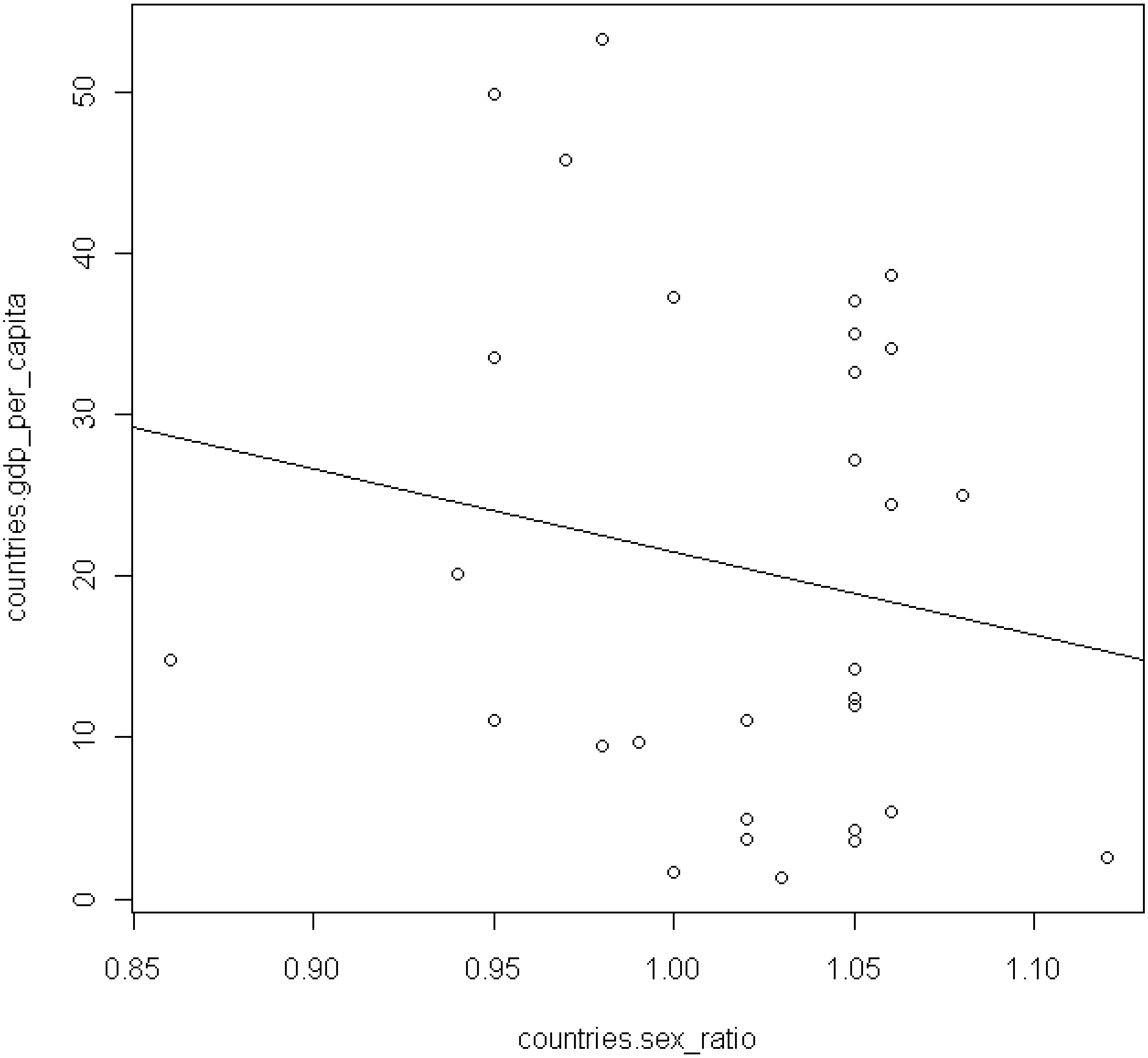
```
> par(mfrow=c(2,2),mex=0.6)
> plot(boxcox.area)
> par(mfrow=c(1,1),mex=1)
```



Looking at the new plot after the boxcox transformation, the data is still not linear and we still can’t do regression on it.

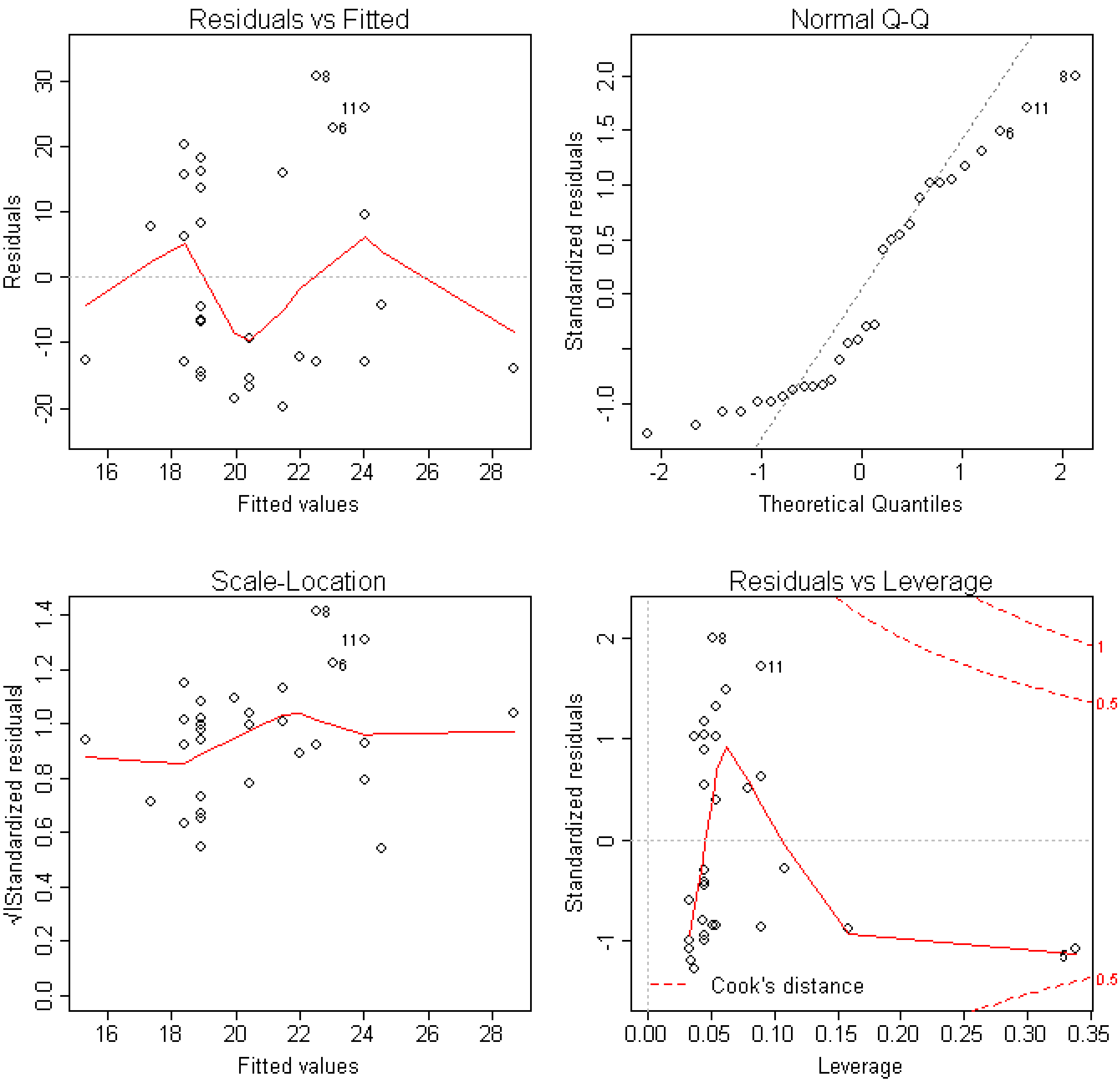
Sex ratio

```
> plot(countries.sex_ratio, countries.gdp_per_capita)
> model_sex_ratio = lm(countries.gdp_per_capita ~ countries.sex_ratio)
> abline(model_sex_ratio, lty = 1)
```



```
> summary(model_sex_ratio)

Call:
lm(formula = countries.gdp_per_capita ~ countries.sex_ratio)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.790 -13.015  -5.576  15.183  30.785 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)            72.76      56.24   1.294    0.206
```

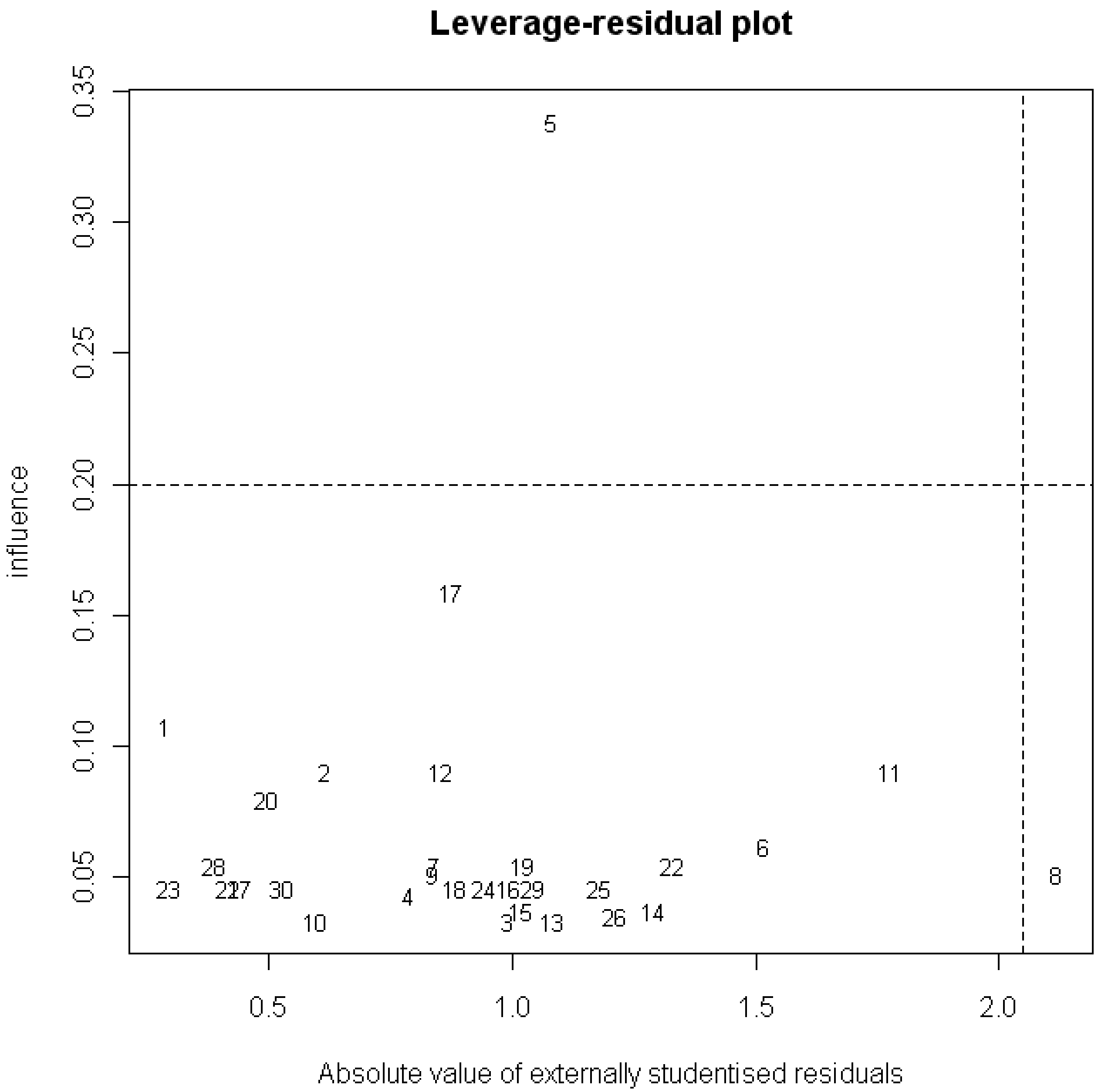
```
countries.sex_ratio   -51.27      55.15  -0.930    0.360

Residual standard error: 15.81 on 28 degrees of freedom
Multiple R-squared: 0.02994,    Adjusted R-squared: -0.004702
F-statistic: 0.8643 on 1 and 28 DF,  p-value: 0.3605

> par(mfrow=c(2,2),mex=0.6)
> plot(model_sex_ratio)
> par(mfrow=c(1,1),mex=1)
```



This is a very bad model that doesn’t pass the conditions for regression. The data is not linear, residuals are not random, the Q-Q plot is not linear, and there a lot of outliers.

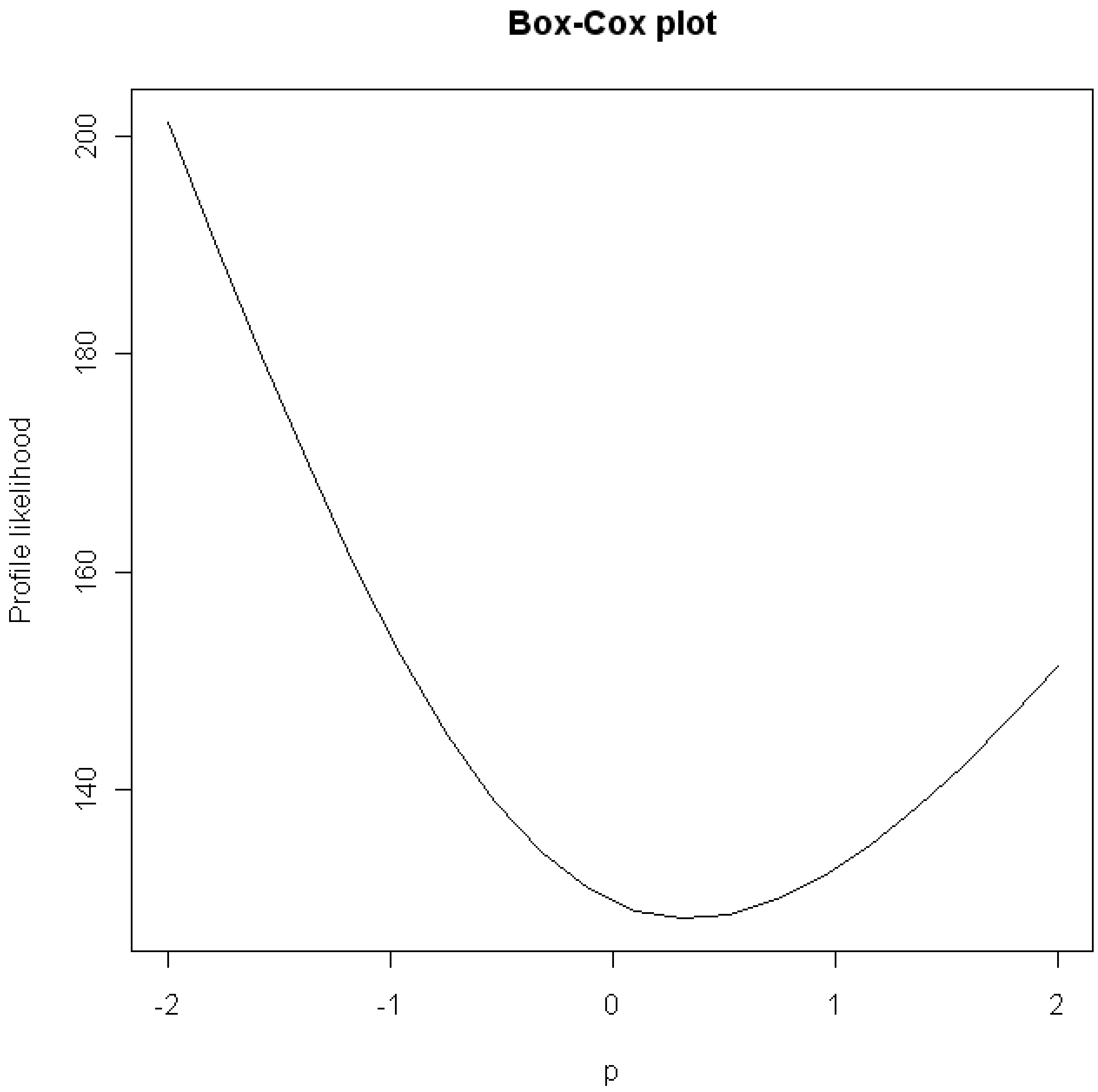
Looking at the leverage-residual plot, there are no points to remove from this data:

```
> lrplot(model_sex_ratio)
```

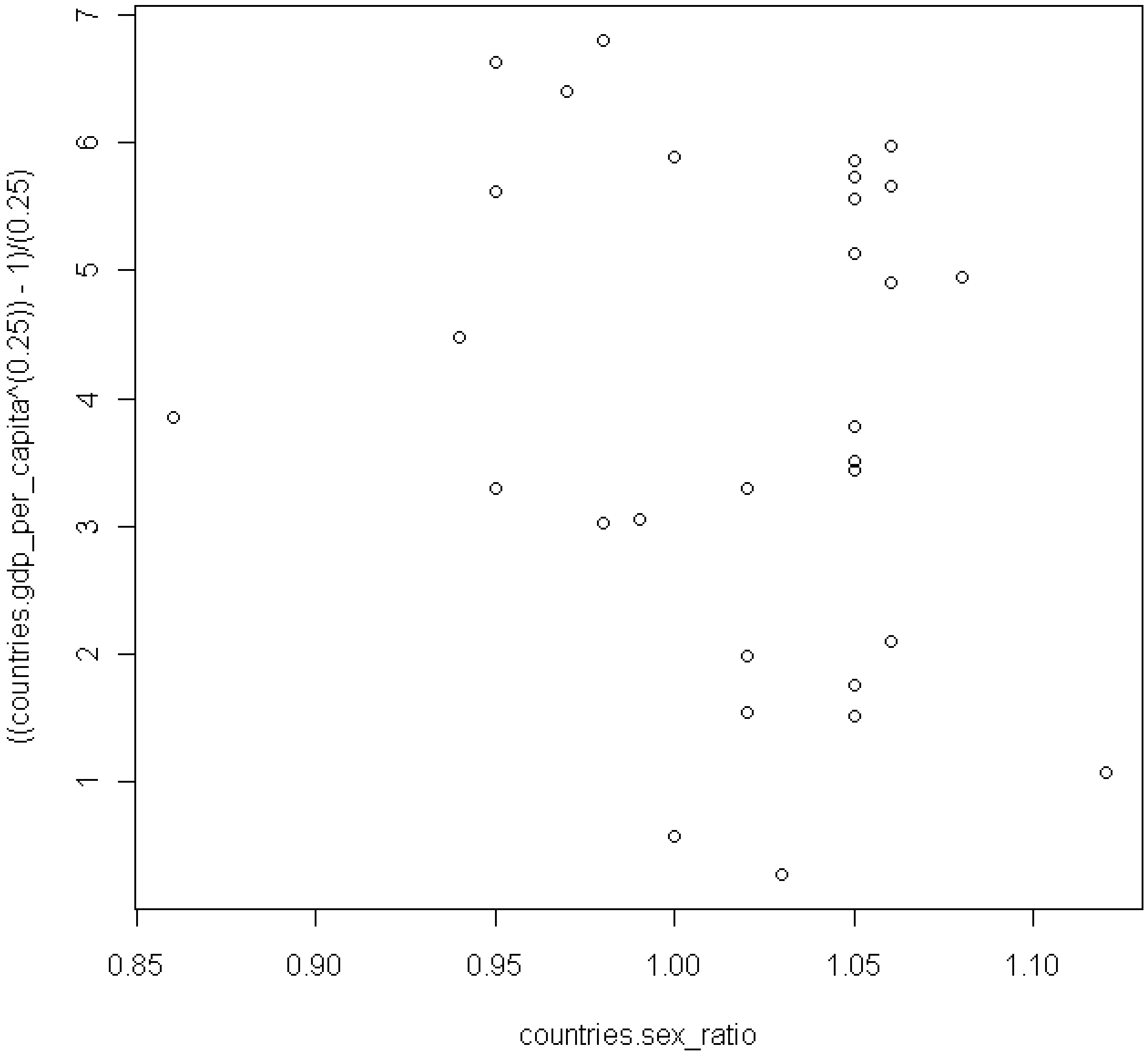


Trying a boxcox transformation to improve the model:

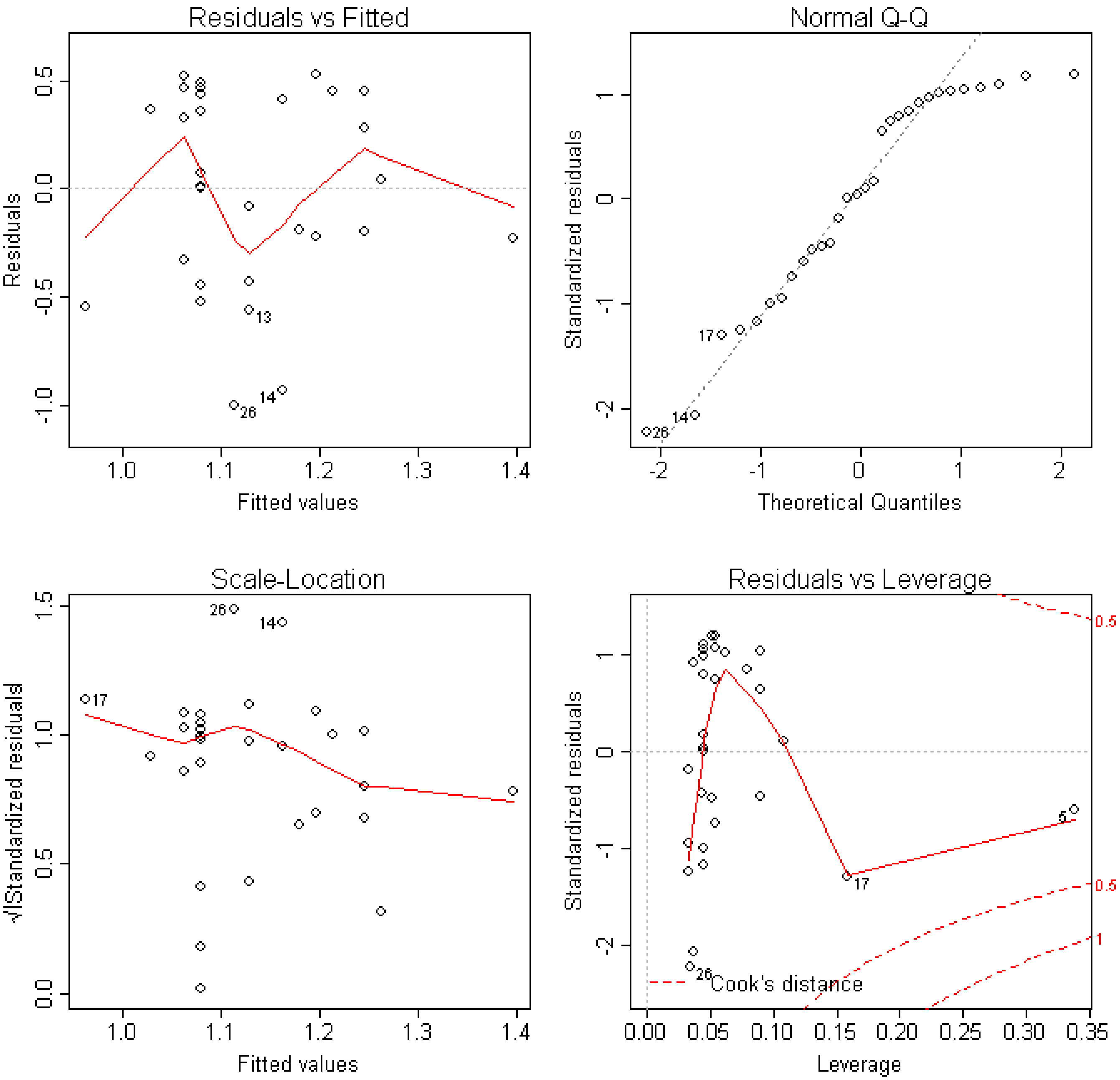
```
> boxcoxplot(countries.gdp_per_capita ~ countries.sex_ratio, data=factors.df)
```



```
> plot(countries.sex_ratio, ((countries.gdp_per_capita^(.25))-1)/(.25))
```



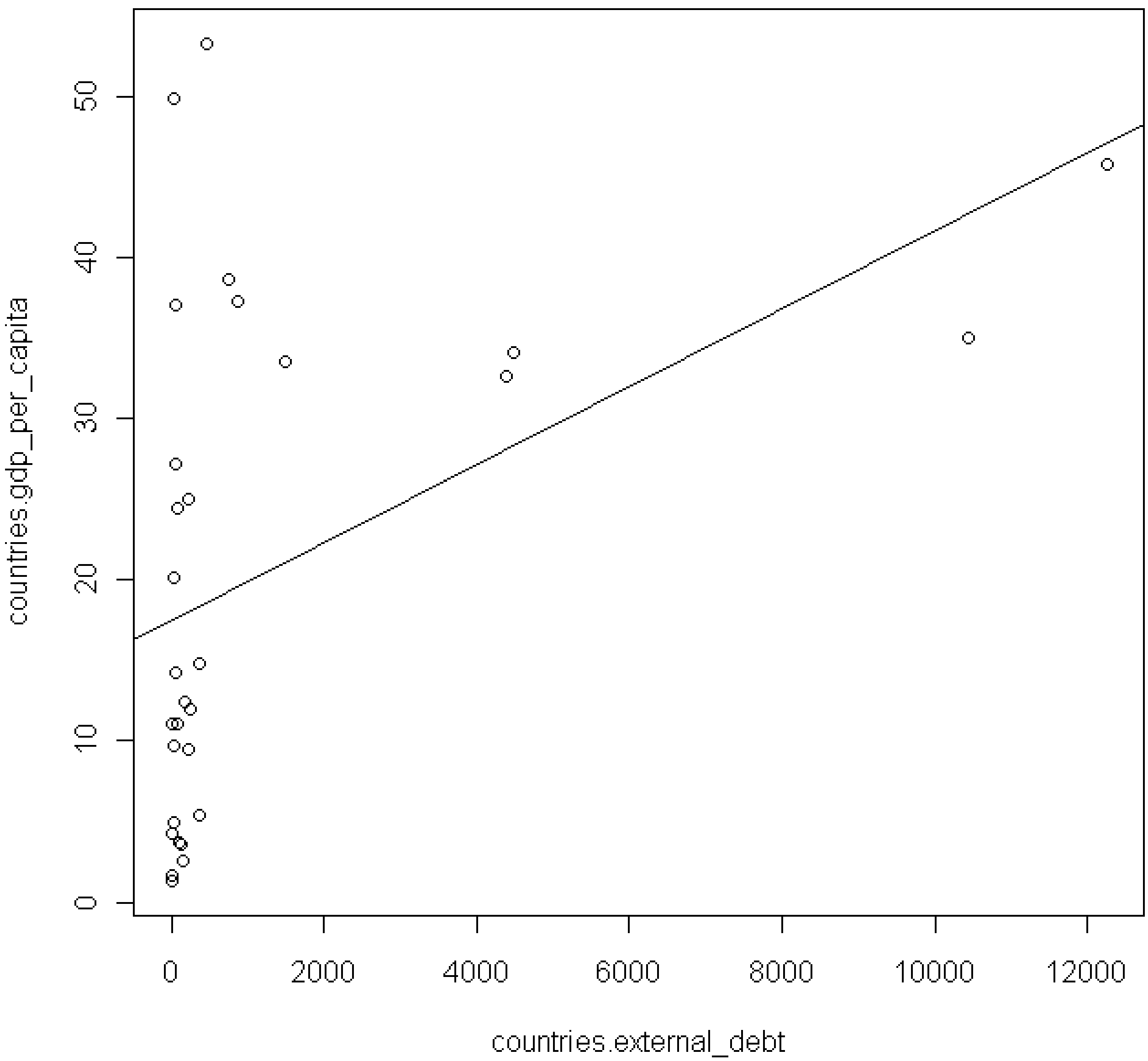
```
> par(mfrow=c(2,2),mex=0.6)
> plot(boxcox.sex_ratio)
> par(mfrow=c(1,1),mex=1)
```



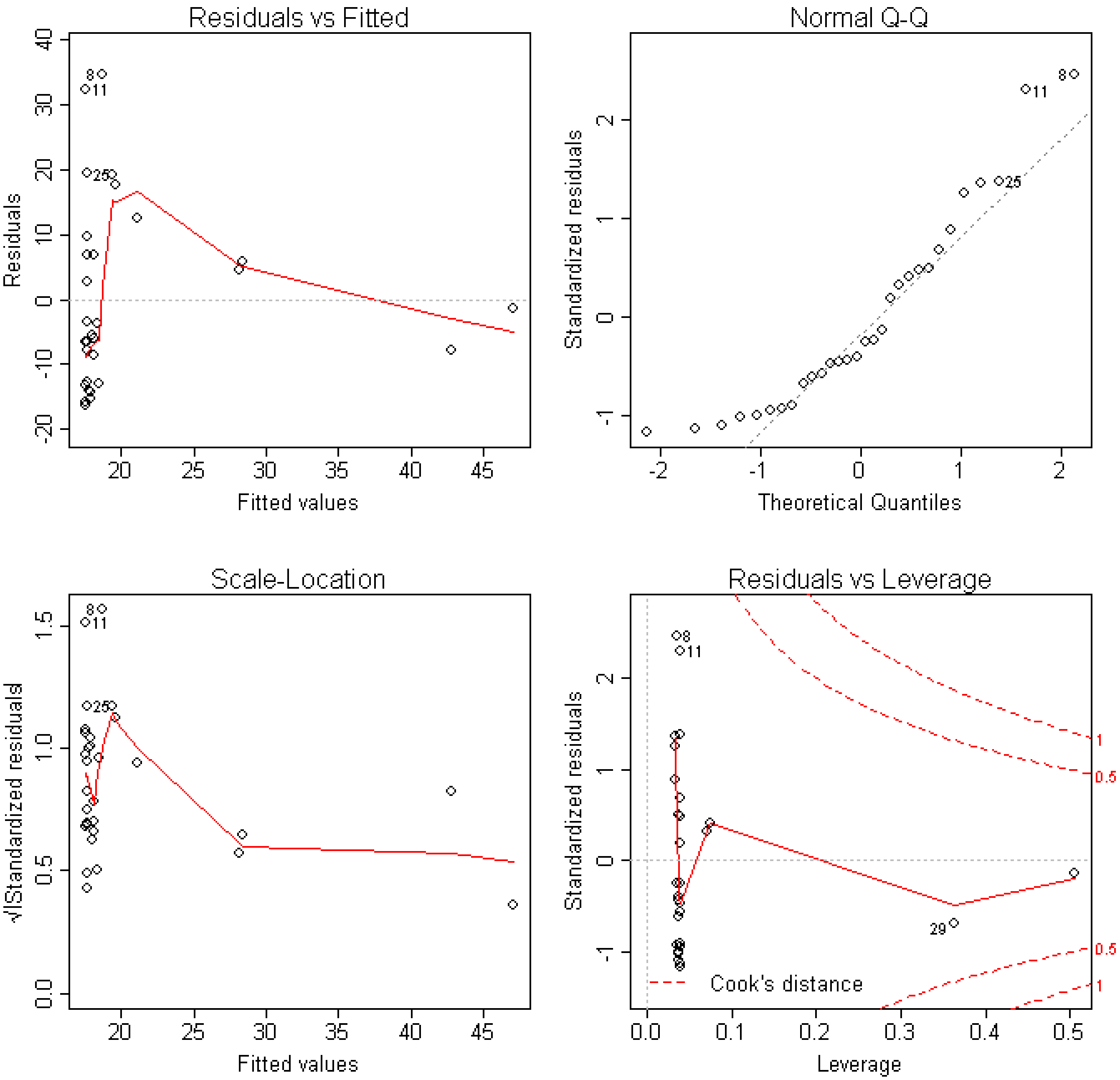
Looking at the new plot after the boxcox transformation, the data is still not linear and we still can’t do regression on it.

External debt

```
> plot(countries.external_debt, countries.gdp_per_capita)
> model_external_debt = lm(countries.gdp_per_capita ~ countries.external_debt)
> abline(model_external_debt, lty = 1)
```



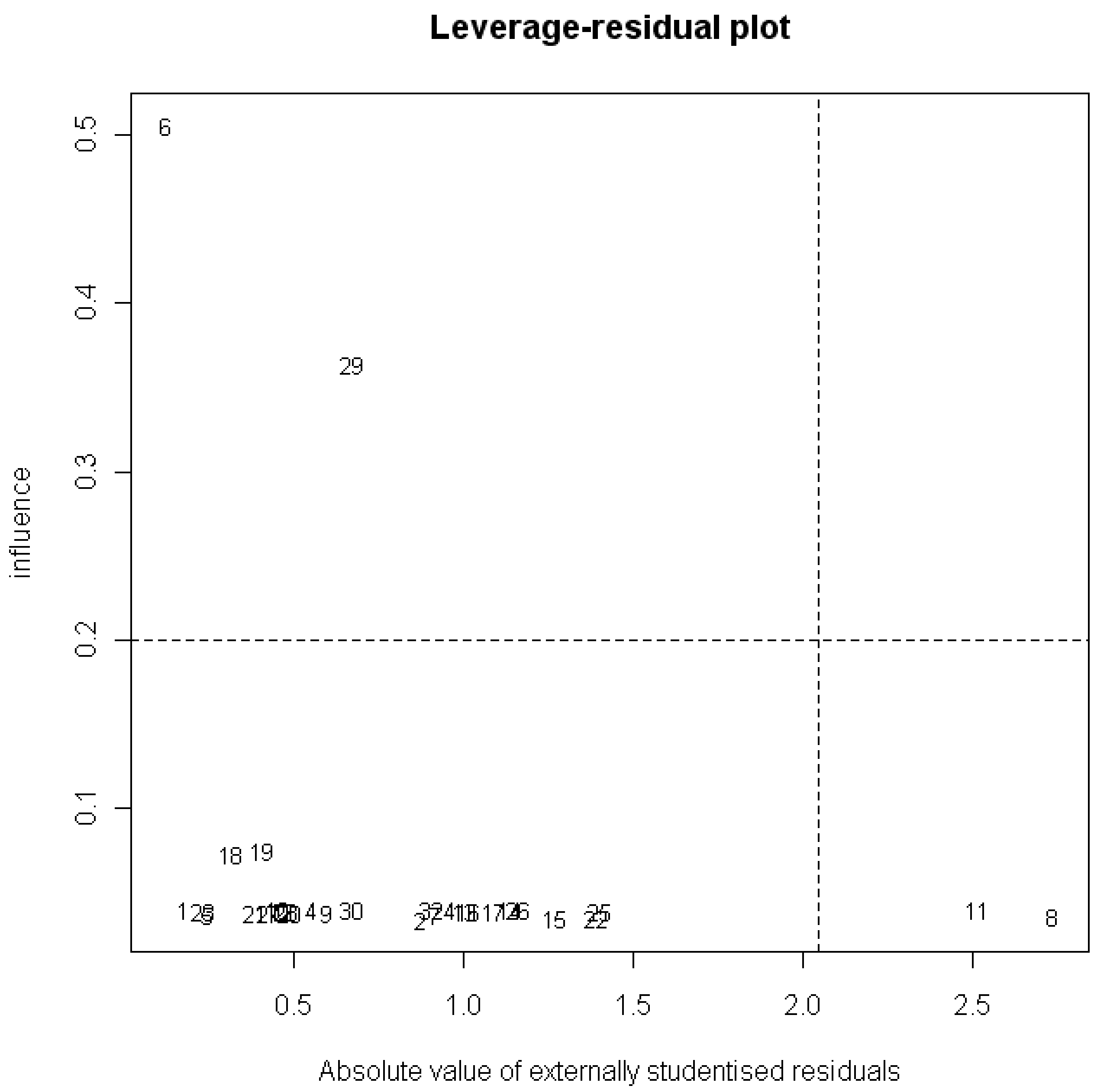
```
> par(mfrow=c(2,2),mex=0.6)
> plot(model_external_debt)
> par(mfrow=c(1,1),mex=1)
```



This is a very bad model that doesn't pass the conditions for regression. The data is not linear, residuals are not random, the Q-Q plot is not linear, and there a lot of outliers.

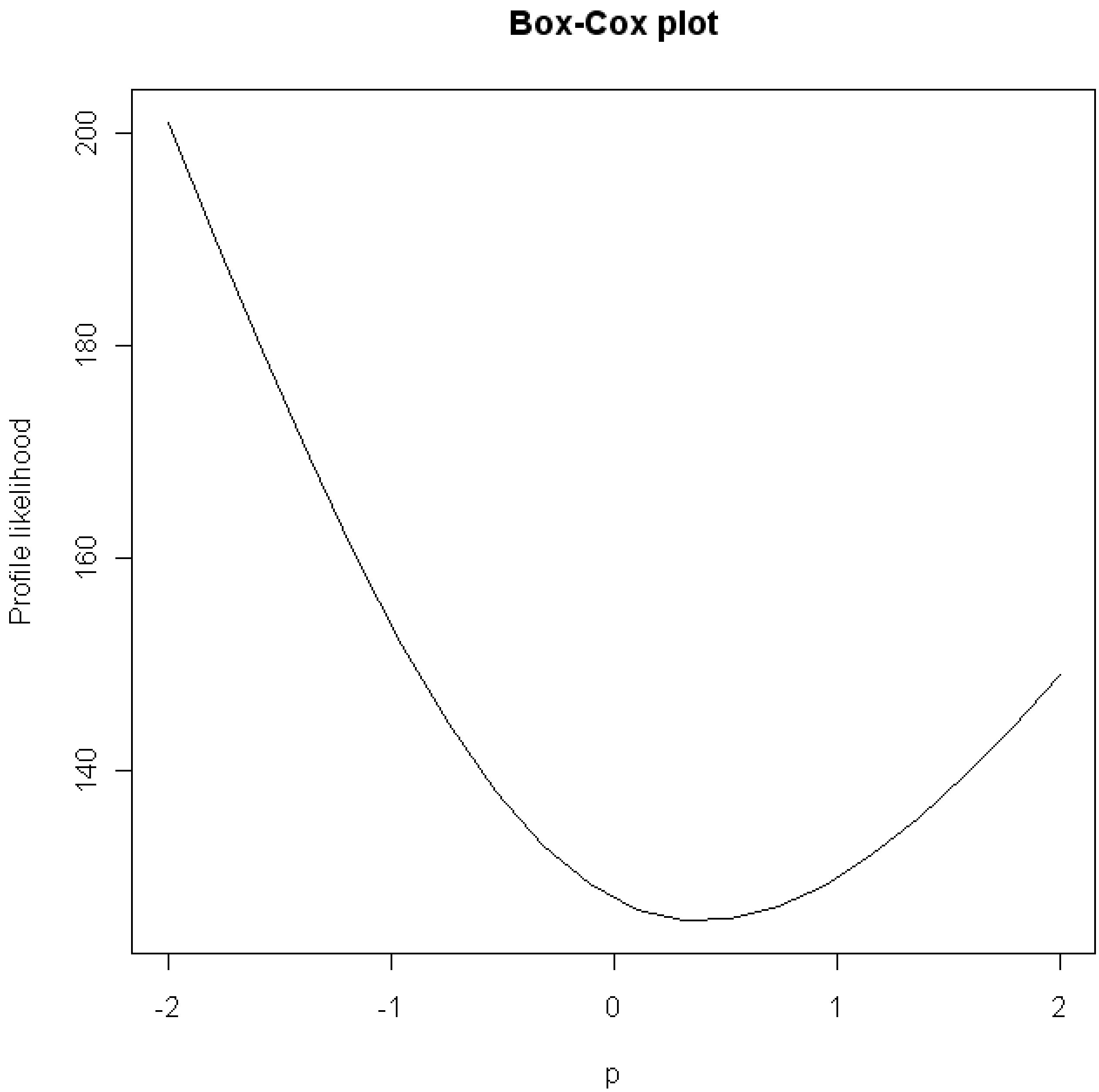
Looking at the leverage-residual plot, there are no points to remove from this data:

```
> lrplot(model_external_debt)
```

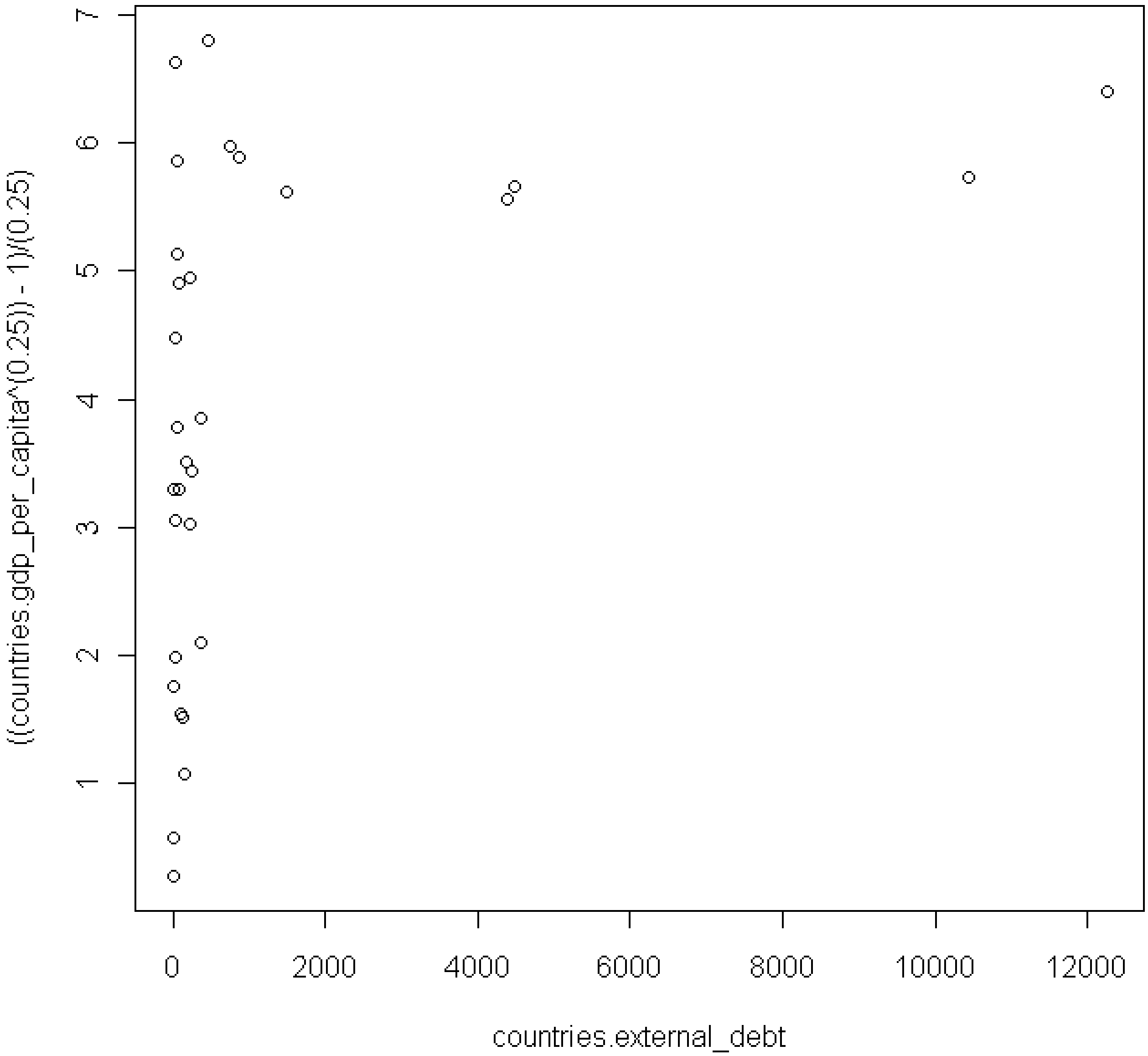


Trying a boxcox transformation to improve the model:

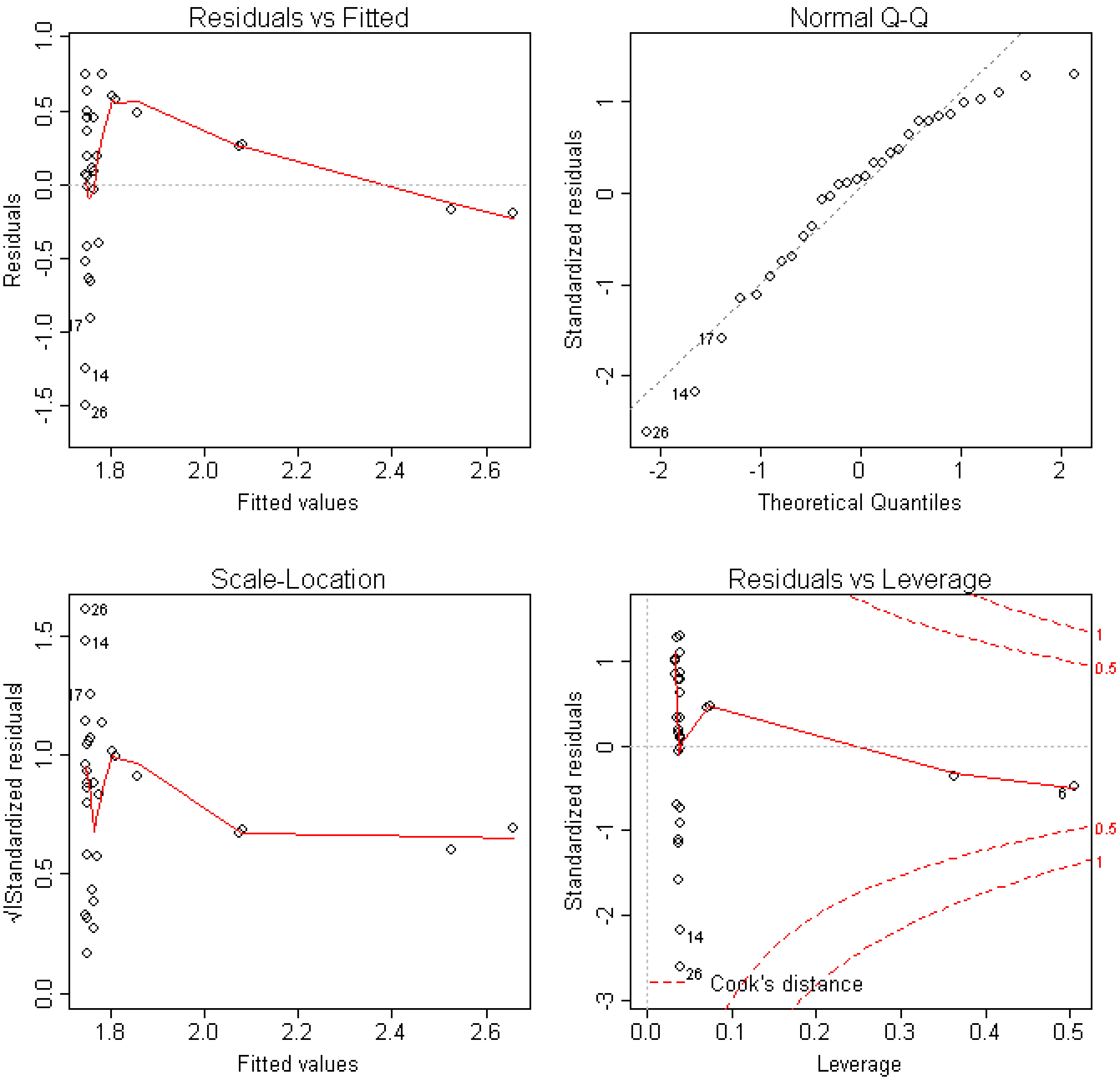
```
> boxcoxplot(countries.gdp_per_capita ~ countries.external_debt, data=factors.df)
```



```
> plot(countries.external_debt, ((countries.gdp_per_capita^(.25))-1)/(.25))
```



```
> par(mfrow=c(2,2),mex=0.6)
> plot(boxcox.external_debt)
> par(mfrow=c(1,1),mex=1)
```



Looking at the new plot after the boxcox transformation, the data is still not linear and we still can't do regression on it.

Variance Inflation Factors

Variance Inflation Factors, and simple regression on life expectancy, median age, population growth, and population density.

```
> diag(solve(cor(countries)))
      life.expectancy            median.age     population.growth 
             4.971212              6.594239              4.129733 
   population.density         literacy.rate     unemployment.rate 
             1.353014              4.379077              3.645132 
        oil.p.minus.c             cell.land military.expenditures 
             1.781451              2.652049              1.367757 
            sex.ratio         external.debt 
             1.540358              2.141105 
```

Multicollinearity is a concern if any of the variance inflation factors is sufficiently high; in practice, approximately 10. None of the VIF's above are close to 10, so it seems that there are no Variance Inflation Factors to worry about.

Multiple Regression

Our multiple regression ended up producing this summary:

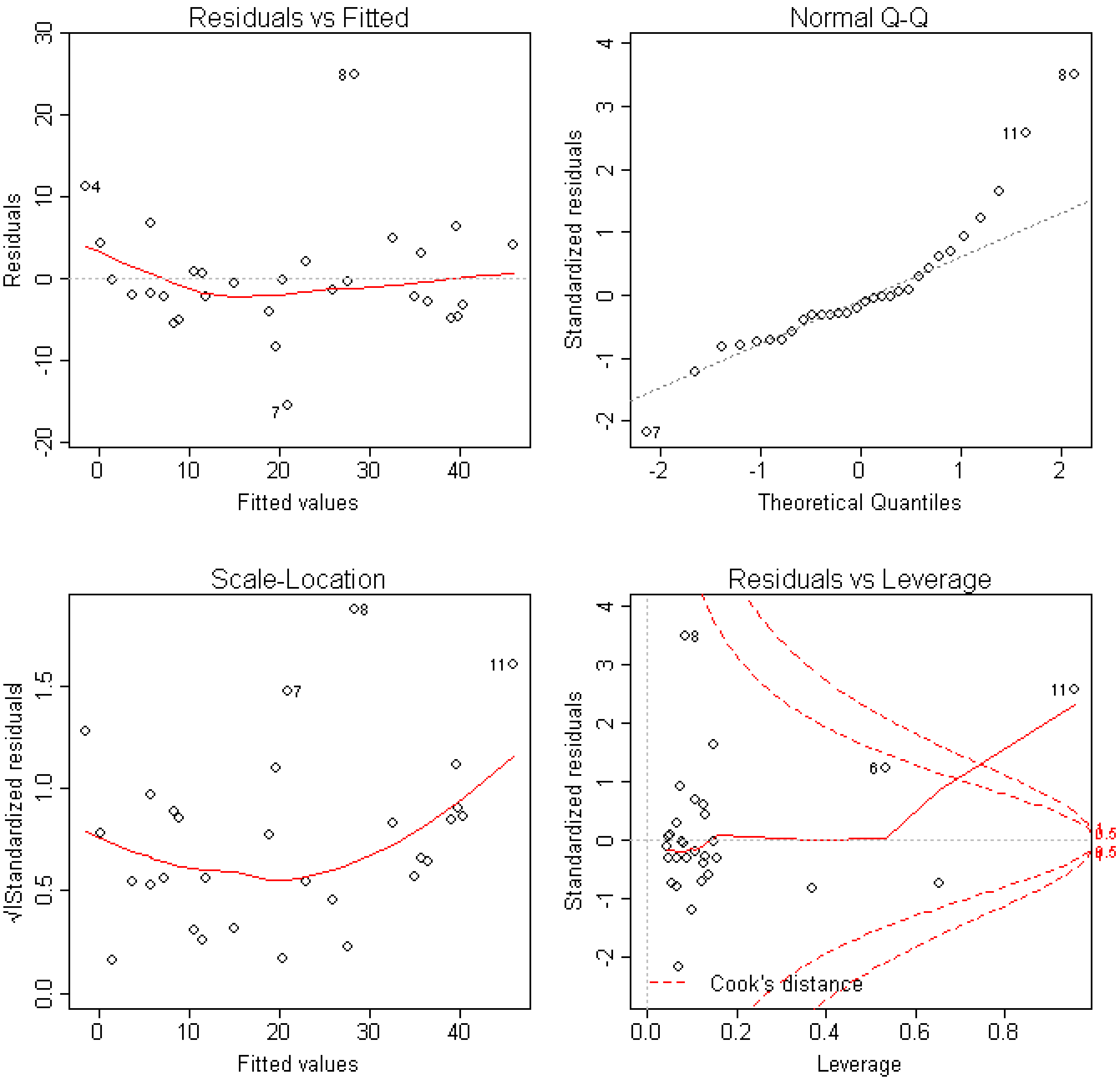
```
Call:
lm(formula = countries.gdp_per_capita ~ countries.life_expectancy +
    countries.median_age + countries.pop_growth + countries.pop_density +
    countries.literacy_rate + countries.unemploy_rate + countries.oil +
    countries.cell_vs_land + countries.military_expenditures +
    countries.area + countries.sex_ratio + countries.external_debt)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8218 -3.5091 -0.8347  3.8812 22.8029

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -4.386e+01  4.928e+01  -0.890 0.385822
countries.life_expectancy      -3.390e-02  4.297e-01  -0.079 0.938033
countries.median_age            2.041e+00  5.072e-01   4.025 0.000879 ***
countries.pop_growth            1.136e+01  3.131e+00   3.629 0.002076 **
countries.pop_density           2.090e-03  1.462e-03   1.429 0.171018
countries.literacy_rate         2.812e-01  2.578e-01   1.091 0.290604
countries.unemploy_rate         8.290e-02  2.119e-01   0.391 0.700500
countries.oil                   1.221e-01  6.354e-01   0.192 0.849948
countries.cell_vs_land         -2.820e-01  3.193e-01  -0.883 0.389378
countries.military_expenditures -7.948e-01  1.026e+00  -0.775 0.449042
countries.area                 -3.203e-04  4.121e-04  -0.777 0.447769
countries.sex_ratio            -3.404e+01  3.635e+01  -0.937 0.362085
countries.external_debt         1.162e-03  7.414e-04   1.568 0.135341
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.074 on 17 degrees of freedom
Multiple R-Squared: 0.8465,     Adjusted R-squared: 0.7381
F-statistic:  7.81 on 12 and 17 DF,  p-value: 9.356e-05
```

Although we get quite a high R^2, this model cannot really be relied on because we have too many variables and too few observations for it to be accurate. This model pretty much fits the data rather than a trend that we can use accurately to predict GDP for any country.

Stepwise Regression

Using stepwise regression, we get the following linear model:

```
> factors.df = data.frame(countries.life_expectancy, countries.median_age,
countries.pop_growth, countries.pop_density, countries.literacy_rate,
countries.unemploy_rate, countries.oil, countries.cell_vs_land,
countries.military_expenditures, countries.area, countries.sex_ratio,
countries.external_debt)
> null.model= lm(countries.gdp_per_capita ~ 1, data = factors.df)
> full.model= lm (countries.gdp_per_capita ~ countries.life_expectancy +
countries.median_age + countries.pop_growth + countries.pop_density +
countries.literacy_rate + countries.unemploy_rate + countries.oil +
countries.cell_vs_land + countries.military_expenditures + countries.area +
countries.sex_ratio + countries.external_debt, data = factors.df)
> full.model.formula= countries.gdp_per_capita ~ countries.life_expectancy +
countries.median_age + countries.pop_growth + countries.pop_density +
countries.literacy_rate + countries.unemploy_rate + countries.oil +
countries.cell_vs_land + countries.military_expenditures + countries.area +
countries.sex_ratio + countries.external_debt
> step(null.model, full.model.formula, direction="forward", trace = 0)

Call:
lm(formula = countries.gdp_per_capita ~ countries.median_age +     countries.pop_growth +
countries.external_debt + countries.pop_density,     data = factors.df)

Coefficients:
            (Intercept)     countries.median_age     countries.pop_growth
countries.external_debt    countries.pop_density
             -6.499e+01                2.296e+00                9.385e+00
9.723e-04                1.808e-03

> best_model = lm(countries.gdp_per_capita ~ countries.median_age + countries.pop_growth
+ countries.external_debt + countries.pop_density, data = factors.df)
> summary(best_model)

Call:
lm(formula = countries.gdp_per_capita ~ countries.median_age +
    countries.pop_growth + countries.external_debt + countries.pop_density,
    data = factors.df)

Residuals:
    Min      1Q  Median      3Q     Max
-15.530  -3.153  -1.072   2.759  24.973

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -6.499e+01  1.126e+01  -5.769 5.17e-06 ***
countries.median_age     2.296e+00  3.006e-01   7.640 5.39e-08 ***
countries.pop_growth     9.385e+00  2.199e+00   4.268 0.000248 ***
countries.external_debt  9.723e-04  5.113e-04   1.902 0.068802 .
countries.pop_density    1.808e-03  1.192e-03   1.517 0.141765
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.449 on 25 degrees of freedom
Multiple R-squared: 0.8078,     Adjusted R-squared: 0.7771
F-statistic: 26.27 on 4 and 25 DF,  p-value: 1.234e-08
> par(mfrow=c(2,2),mex=0.6)
> plot(best_model)
> par(mfrow=c(1,1),mex=1)
```

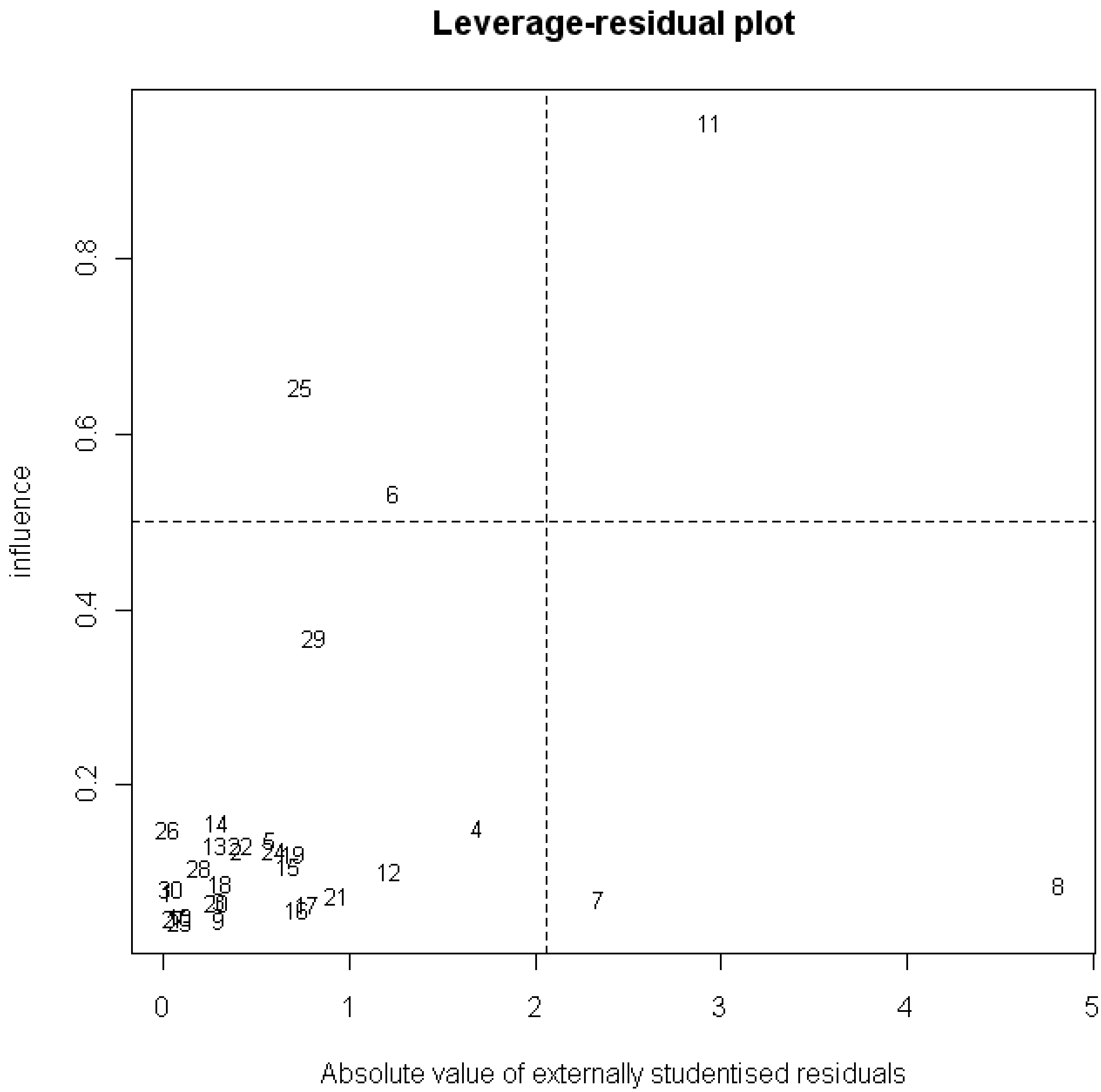


GDP = -6.499e+01 + 2.296(median age) + 9.385(population growth) + 9.723e-04(external debt) + 1.808e-03(population density)

The residuals for the stepwise model are a little curved, but besides that they are independent and have equal variance. There are a couple of possible outliers in the residual plot. The Q-Q plot is fairly linear but with a few outliers. The R^2 is strong for this model, 80.78% of the variability in GDP per capita is accounted for by the linear association with median age, population growth, external debt, and population density.

Looking at the leverage-residual plot, there is an influential outlier that should be removed.

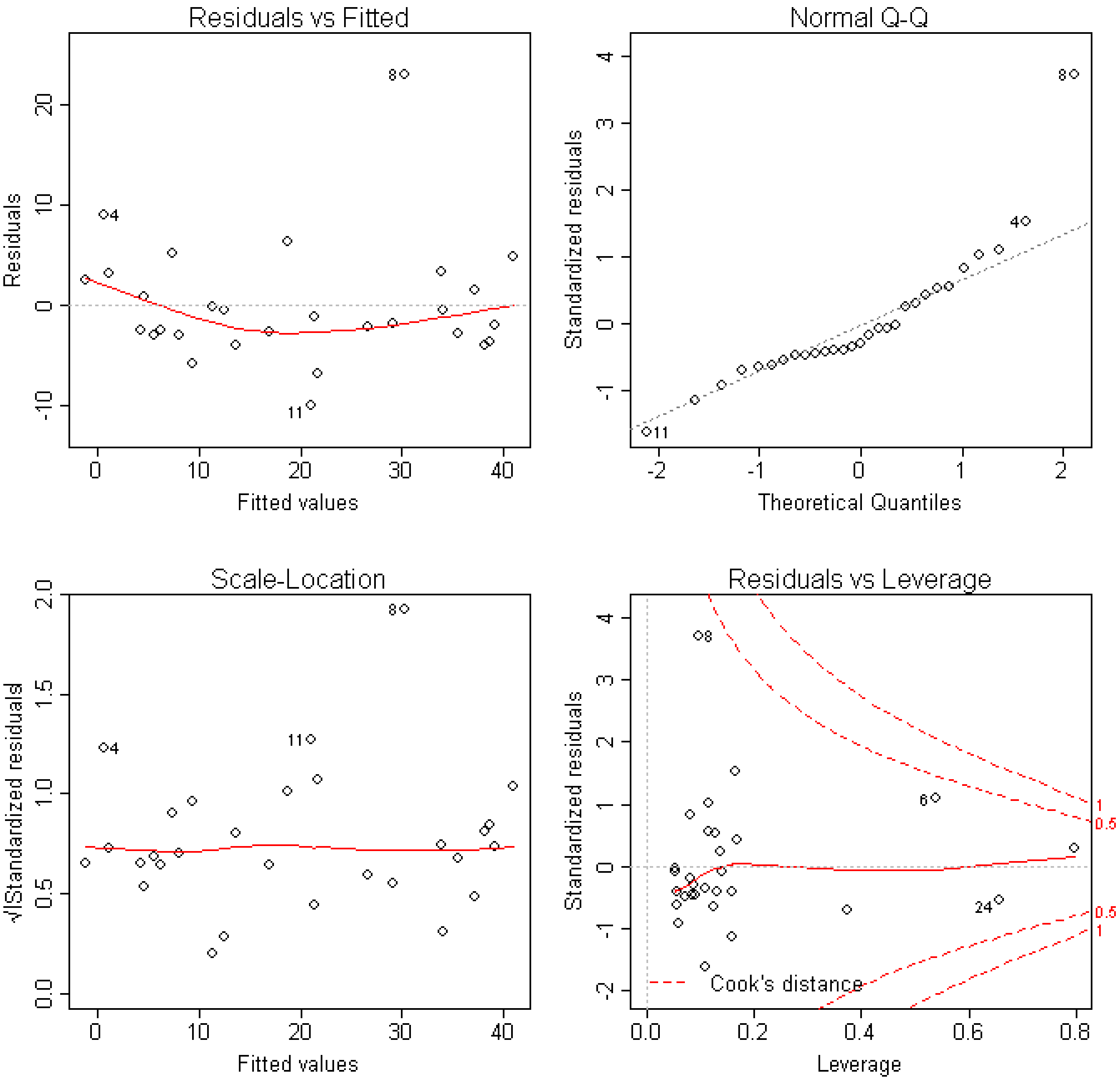
```
> lrplot(best_model)
```



Data point 11 refers to Singapore. Singapore has a very high population density, almost five times higher than the next highest country. All the other variables for Singapore that are in the model are also high. The reason why this country is an influential outlier is because it is a small country, but it has a lot of people that are financial well off.

Removing Singapore from the data set gives a stronger linear association.

```
> countries = countries[-11,]
> countries.country = countries$country
> countries.life_expectancy = countries$life.expectancy..years.
> countries.median_age = countries$median.age..years.
> countries.pop_growth = countries$population.growth....
> countries.pop_density = countries$population.density..population...country.area.
> countries.literacy_rate = countries$literacy.rate....
> countries.unemploy_rate = countries$unemployment.rate....
> countries.oil = countries$oil.production...oil.consumption..mill..bbl.day.
> countries.cell_vs_land = countries$cell.phone.vs.land.line.ratio..cell...land.
> countries.military_expenditures = countries$military.expenditures....of.GDP.
> countries.area = countries$country.area..thousands.sq.km.
> countries.sex_ratio = countries$sex.ratio..male.female.
> countries.external_debt = countries$external.debt..billion.USD.
> countries.gdp_per_capita = countries$GDP.per.capita..thousands.USD.
> factors.df = data.frame(countries.life_expectancy, countries.median_age,
countries.pop_growth, countries.pop_density, countries.literacy_rate,
countries.unemploy_rate, countries.oil, countries.cell_vs_land,
countries.military_expenditures, countries.area, countries.sex_ratio,
countries.external_debt)
> best_model_remove_outlier = lm(countries.gdp_per_capita ~ countries.median_age +
countries.pop_growth + countries.external_debt + countries.pop_density, data = factors.df)
> summary(best_model_remove_outlier)

Call:
lm(formula = countries.gdp_per_capita ~ countries.median_age +
    countries.pop_growth + countries.external_debt + countries.pop_density,
    data = factors.df)

Residuals:
   Min     1Q Median     3Q    Max
-9.959 -2.947 -1.890  2.512 23.039

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.975e+01  1.002e+01  -5.966 3.71e-06 ***
countries.median_age     2.224e+00  2.641e-01   8.423 1.25e-08 ***
countries.pop_growth     8.467e+00  1.949e+00   4.344  0.00022 ***
countries.external_debt  9.790e-04  4.473e-04   2.189  0.03859 *
countries.pop_density   -1.158e-02  4.669e-03  -2.481  0.02049 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.517 on 24 degrees of freedom
Multiple R-squared: 0.8389,     Adjusted R-squared: 0.812
F-statistic: 31.24 on 4 and 24 DF,  p-value: 3.382e-09

> par(mfrow=c(2,2),mex=0.6)
> plot(best_model_remove_outlier)
> par(mfrow=c(1,1),mex=1)
```

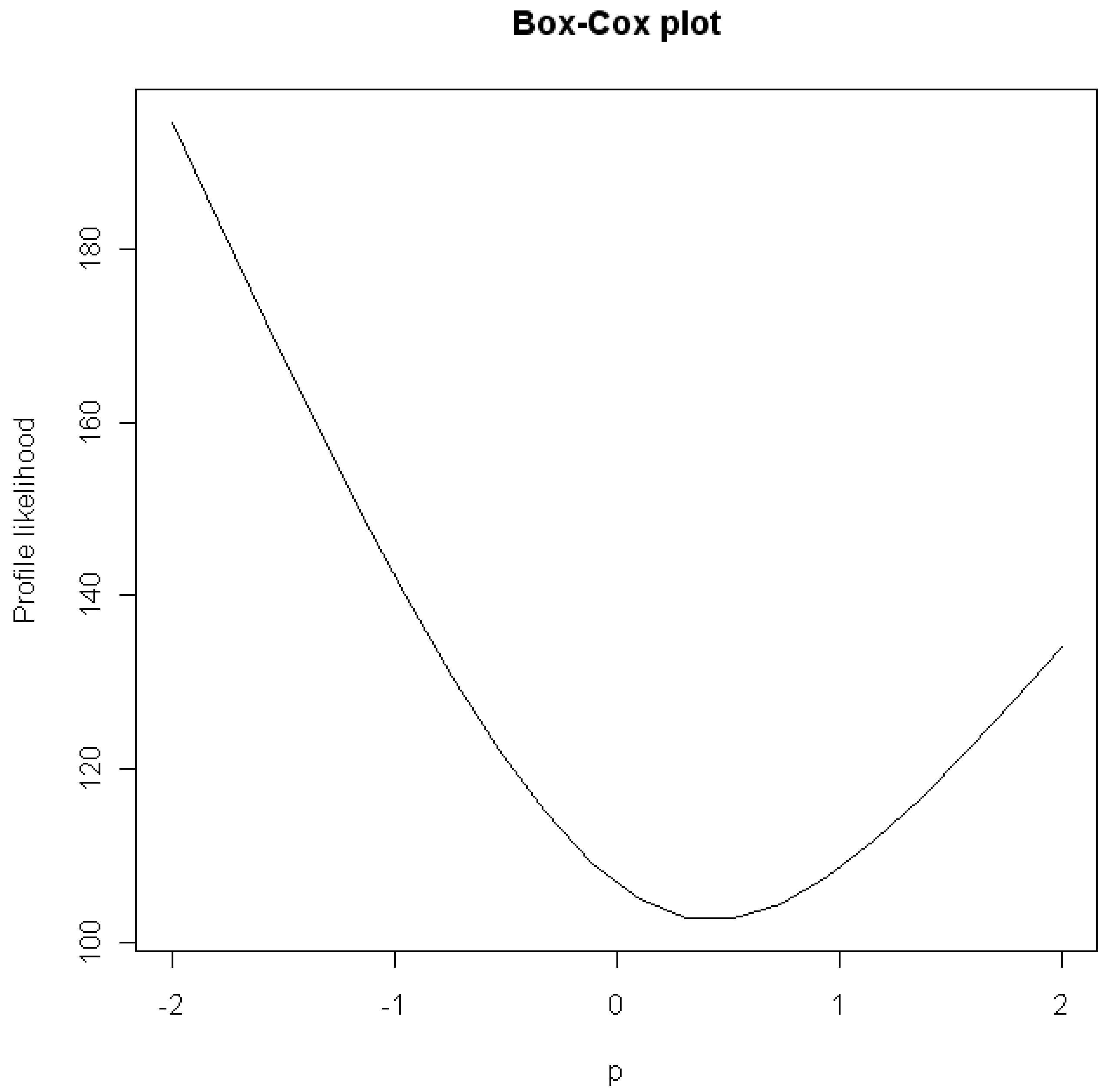


GDP = -6.277e+01 + 2.257(median age) + 8.885(population growth) + 9.274e-04(external debt) + 2.232e-03(population density)

The residual plot looks slightly more random than the last model, but the Q-Q plot looks much more linear. The R^2 is a little stronger for this model, 83.89% of the variability in GDP per capita is accounted for by the linear association with median age, population growth, external debt, and population density.

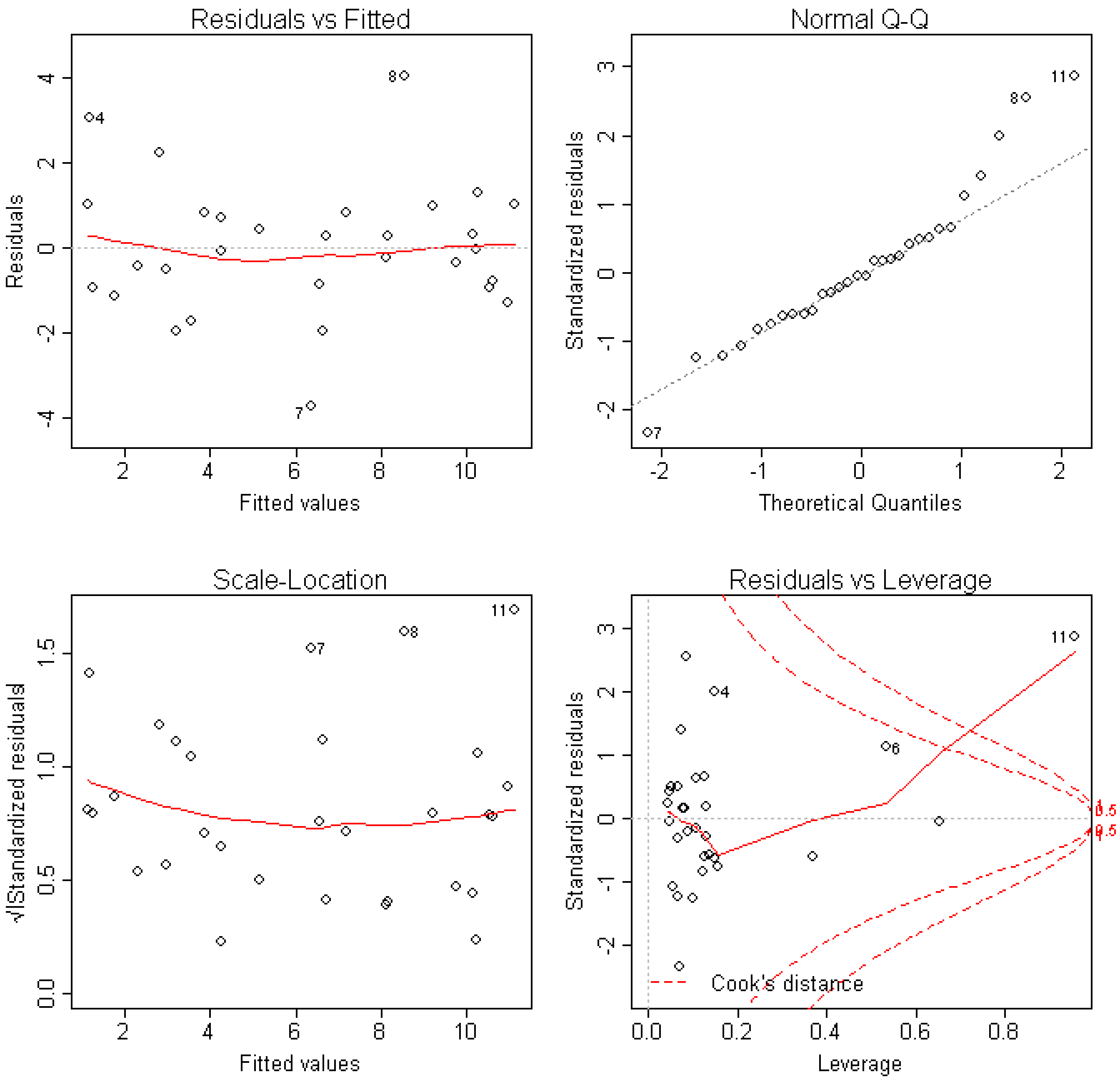
Since the residual plot in out first stepwise model was a little curved, we will try a boxcox transformation to make the date more linear.

```
> boxcoxplot(countries.gdp_per_capita ~ countries.median_age + countries.pop_growth + 
countries.external_debt + countries.pop_density, data=factors.df)
```



```
> boxcox.best_model=lm(((countries.gdp_per_capita^(.5))-1)/(.5) ~ countries.median_age +
countries.pop_growth + countries.external_debt + countries.pop_density)
> summary(boxcox.best_model)

Call:
lm(formula = ((countries.gdp_per_capita^(0.5)) - 1)/(0.5) ~ countries.median_age +
    countries.pop_growth + countries.external_debt + countries.pop_density)

Residuals:
     Min       1Q   Median       3Q      Max
-3.72869 -0.93466 -0.06819  0.80971  4.04569

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.388e+01  2.514e+00  -5.521 9.73e-06 ***
countries.median_age     5.560e-01  6.707e-02   8.290 1.22e-08 ***
countries.pop_growth     1.915e+00  4.907e-01   3.901 0.000638 ***
countries.external_debt  1.665e-04  1.141e-04   1.459 0.156983
countries.pop_density    2.228e-04  2.659e-04   0.838 0.409978
```

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

Residual standard error: 1.662 on 25 degrees of freedom
Multiple R-squared: 0.828,      Adjusted R-squared: 0.8004
F-statistic: 30.08 on 4 and 25 DF,  p-value: 3.165e-09
> par(mfrow=c(2,2),mex=0.6)
> plot(boxcox.best_model)
> par(mfrow=c(1,1),mex=1)
```

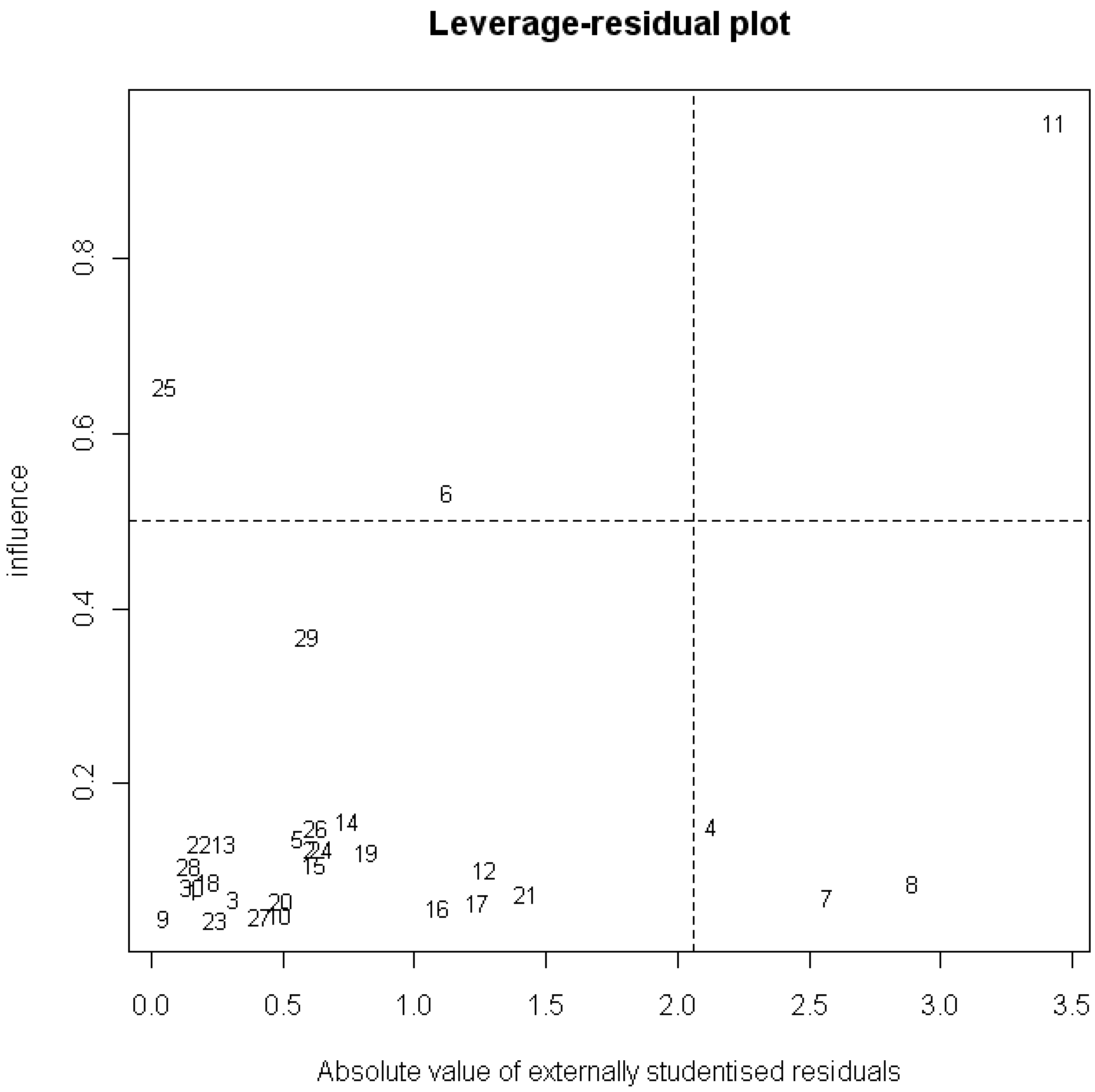


(GDP$^{0.5}$-1) / (0.5) = -1.388e+01 + 5.560e-01(median age) + 1.915(population growth) + 1.665e-04(external debt) + 2.228e-04(population density)

The residual plot for the boxcox transformed stepwise model is much more random than it was originally. There may be a few outliers. The Q-Q plot is more linear than the original stepwise model, but looks like there are some possible outliers. The R^2 is a little stronger than the original stepwise model, 82.8% of the variability in GDP per capita is accounted for by the linear association with median age, population growth, external debt, and population density.

Looking at the leverage-residual plot, there is an influential outlier that should be removed.

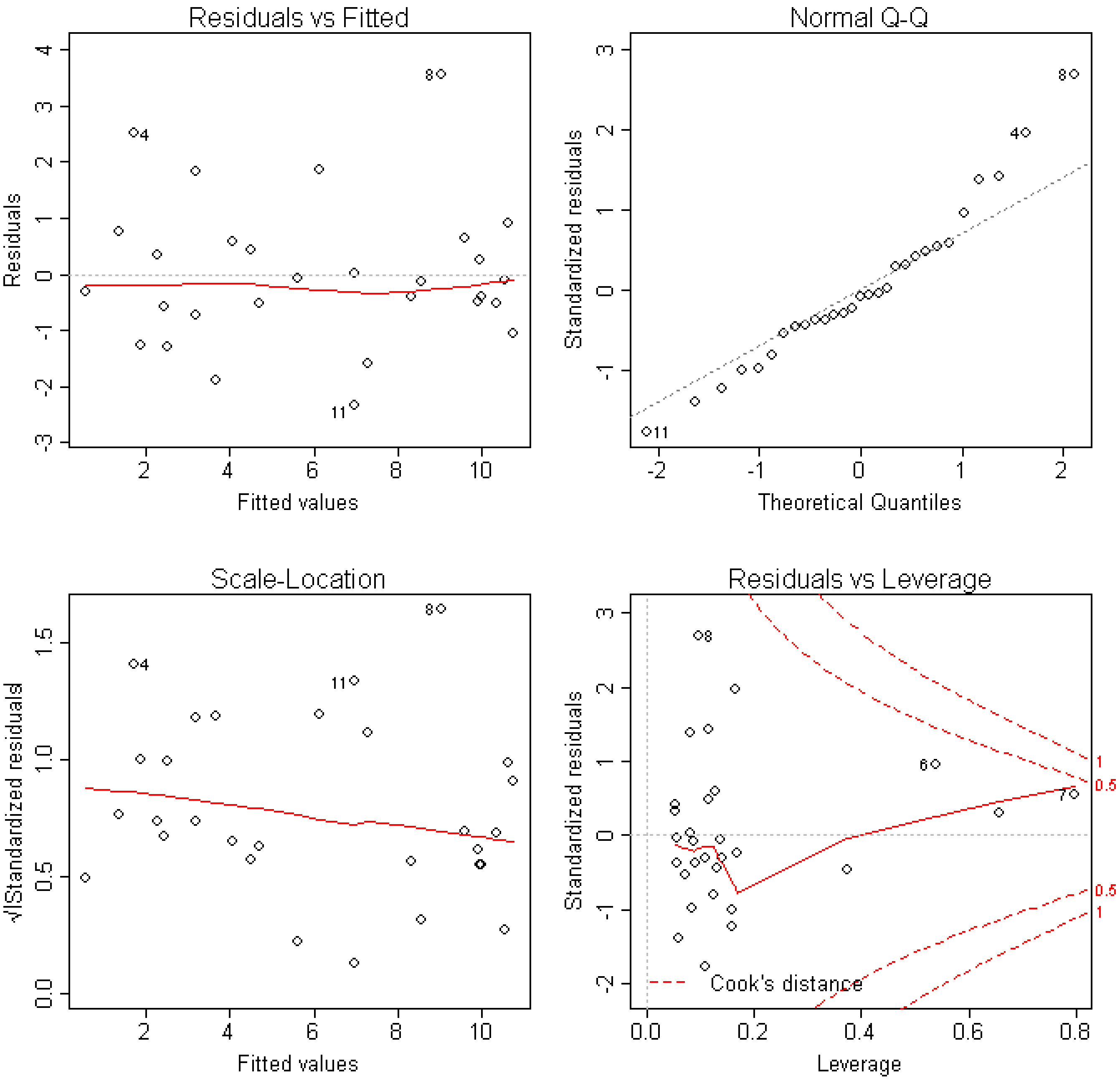
```
> lrplot(boxcox.best_model)
```



Data point 11 refers to Singapore. This is the same influential outlier that we removed from the original stepwise model.

The following model removes Singapore from the boxcox stepwise model:

```
> countries = countries[-11,]
> countries.country = countries$country
> countries.life_expectancy = countries$life.expectancy..years.
> countries.median_age = countries$median.age..years.
> countries.pop_growth = countries$population.growth....
> countries.pop_density = countries$population.density..population...country.area.
> countries.literacy_rate = countries$literacy.rate....
> countries.unemploy_rate = countries$unemployment.rate....
> countries.oil = countries$oil.production...oil.consumption..mill..bbl.day.
> countries.cell_vs_land = countries$cell.phone.vs.land.line.ratio..cell...land.
> countries.military_expenditures = countries$military.expenditures....of.GDP.
> countries.area = countries$country.area..thousands.sq.km.
> countries.sex_ratio = countries$sex.ratio..male.female.
> countries.external_debt = countries$external.debt..billion.USD.
> countries.gdp_per_capita = countries$GDP.per.capita..thousands.USD.
> factors.df = data.frame(countries.life_expectancy, countries.median_age,
countries.pop_growth, countries.pop_density, countries.literacy_rate,
countries.unemploy_rate, countries.oil, countries.cell_vs_land,
countries.military_expenditures, countries.area, countries.sex_ratio,
countries.external_debt)
> boxcox.best_model_remove_outliers=lm(((countries.gdp_per_capita^(.5))-1)/(.5) ~
countries.median_age + countries.pop_growth + countries.external_debt +
countries.pop_density)
> summary(boxcox.best_model)

Call:
lm(formula = ((countries.gdp_per_capita^(0.5)) - 1)/(0.5) ~ countries.median_age +
    countries.pop_growth + countries.external_debt + countries.pop_density)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3258 -0.5837 -0.1258  0.5691  3.5650

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.258e+01  2.136e+00  -5.889 4.48e-06 ***
countries.median_age     5.382e-01  5.631e-02   9.557 1.18e-09 ***
countries.pop_growth     1.686e+00  4.156e-01   4.057 0.000456 ***
countries.external_debt  1.682e-04  9.539e-05   1.763 0.090648 .
countries.pop_density   -3.106e-03  9.955e-04  -3.120 0.004656 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.39 on 24 degrees of freedom
Multiple R-squared: 0.8735,     Adjusted R-squared: 0.8524
F-statistic: 41.42 on 4 and 24 DF,  p-value: 1.932e-10

> par(mfrow=c(2,2),mex=0.6)
> plot(boxcox.best_model)
> par(mfrow=c(1,1),mex=1)
```



$(GDP^{0.5}-1) / (0.5) = -1.258e{+}01 + 5.382e{-}01(\text{median age}) + 1.686(\text{population growth}) + 1.682e{-}04(\text{external debt}) – 3.106e{-}03(\text{population density})$

The residual plot looks random, and the Q-Q plot looks fairly linear. The R^2 is stronger, 87.35% of the variability in GDP per capita is accounted for by the linear association with median age, population growth, external debt, and population density.

Conclusion

Our case study was very successful. We were able to create a model that predicts GDP per capita. Our final model:

$$((\text{predicted GDP})^{0.5}-1) / (0.5) = -1.258e+01 + 5.382e-01(\text{median age}) + 1.686(\text{population growth}) + 1.682e-04(\text{external debt}) - 3.106e-03(\text{population density})$$

Although we are very happy with our final model, we did have a lot of problems with the single regressions. Much of our data did not work out very well, which made it difficult to get any really answers about the individual variables compared to the GDP. We also had difficulty with our multiple regression, with all twelve variables, because we have too many variables and not enough observations. Both of these problems might have been resolved if we were able to get many, many observations, but because the data is difficult to pull together, we would never have enough time to do this.

To test our final model, we picked a random country, Greece, to predict its GDP:

Media age (years) = 41.5

Population growth (%) = 0.146

External debt (billion USD) = 86.72

Population density (population / country area) = 81.27

$$\begin{aligned}((\text{predicted GDP})^{0.5}-1) / (0.5) &= -1.258e+01 + 5.382e-01(41.5) + 1.686(0.146)+ 1.682e-04(86.72) - 3.106e-03(81.27)\\ &= -12.58 + 22.3353 + 0.246156 + 0.014586304 -0.25242462\\ &= 9.763617684\end{aligned}$$

predicted GDP = 34.60

The actual GDP per capita for Greece is 30.6. So for Greece our model was only off by 13.1%.