|  |
| --- |
| Statistics 333 Term Project |
| **Factors Predicting the GDP of a Nation** |
| **A Multiple Regression Analysis** |

|  |
| --- |
| **Team 9**  **12/4/2015** |

**INTRODUCTION**

Gross domestic product, or GDP, is a fundamental indicator of a society’s well-being. Few other variables depict the standard of living for a citizen than GDP in its per capita form. As such, it is imperative in stately analysis to understand what factors best predict GDP. Such details allow authorities to control these factors in an aim to ameliorate their country’s GDP per capita (or prevent it from deteriorating).

To promote comprehension of what variables best explain gross domestic product per capita, we assembled a dataset from both the official databases of the United Nations, and the Annual Register. The 1992 data, whilst not contemporary, still provides meaningful insight into how the many facets of a country’s population (e.g. life expectancy, infant mortality, etc.) co-exist with wealth. In finding an explanation of GDP, we would allow greater control in terms of how a nation considers regulating its wealth.

**DATA SET**

Our dataset consists of 10 predictor variables against the response variable of GDP. The dataset provides data for 97 countries in the world, divided into 6 groups. Five of these groups are location-based, with the last consisting of the developed/super power nations. We found our data set from two major online sources the American Statistical Association website ([www.amstat.org](http://www.amstat.org/)) and the United Nations website ([www.un.org](http://www.un.org/)). The data in our data set corresponds to the year 1992.

We had a few missing data points in the AMSAT dataset, which was our primary dataset. We filled these from the UN website. The AMSTAT dataset comprised of 6 predictor variables and the response variable. Data for the remaining 4 predictor variables were obtained from the UN website. Since both of our sources are very credible sources, we were pretty confident with the accuracy of data. However we cross checked them with each other and also with a few other sources. For the UN variables, we collected 1992 data where possible and 1990-95 data when only a range was available (i.e. in the case of dependency).

**VARIABLES**

The following are our **predictor variables**:

BIR – Live birth rate per 1,000 of population

DEA – Death rate per 1,000 of population

INF– Infant deaths per 1,000 of population under 1 year old

LEM – Life expectancy at birth for males

LEF – Life expectancy at birth for females

POP – Total population of both sexes combined (in thousands)

MIG – Net migration rate per 1,000 of population

AGE – Median age of population

DEP – Total dependency ratio (ratio of population 0-19 & 70+ per 100 population 20-69)

GRO– Country Group, Factored via six levels:

1 = Eastern Europe

2 = South America and Mexico

3 = Western Europe, North America, Japan, Australia, New Zealand

4 = Middle East

5 = Asia

6 = Africa

The **response variable** is: GDP – Gross Domestic Product per capita in 1992 U.S dollars

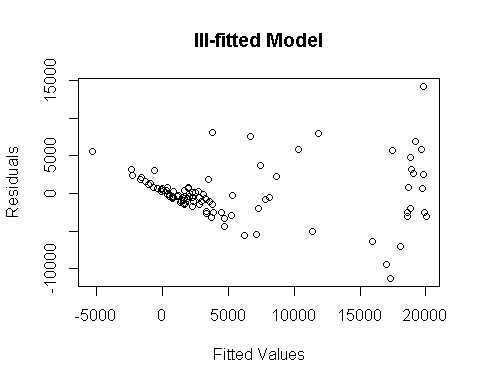
**MODEL SELECTION AND ANALYSIS:**

**Method of Model-Choosing:**

Our models are initially built off of backwards elimination with a p-value cutoff of 0.1. We intend to capture models with high adjusted R^2 and low AIC, but always choose marginally worse criterion scores in exchange for less variables and more efficient models. Therefore, models created by step functions often produced better AIC, but had insignificant variables.

**Creating a model with the base variables**

As a first step we tried to build a model with only the base predictor variables and no interaction terms. We arrived at the model **gdp ~ bir + lem + gro + dep** which granted an AIC of 1877 and an adjusted R^2 of 0.7867 (both values could be improved with additional variables, but only minimally so). The step function included the age variable as well which provided a small decrease in AIC and increase in adjusted R^2. Nevertheless, neither model suffices as our residual plot features clear indicators of unequal variance and a lack of model fit.



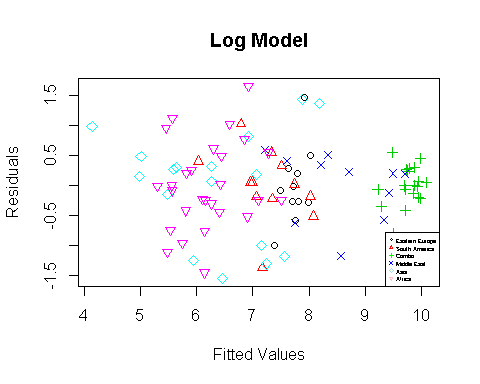
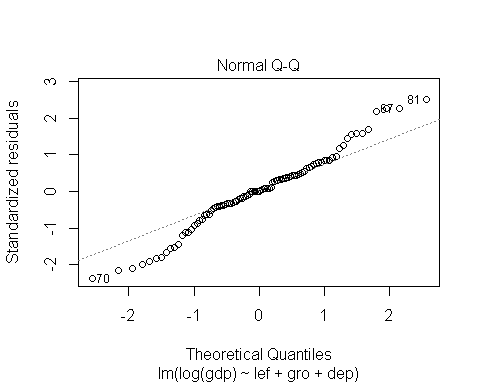
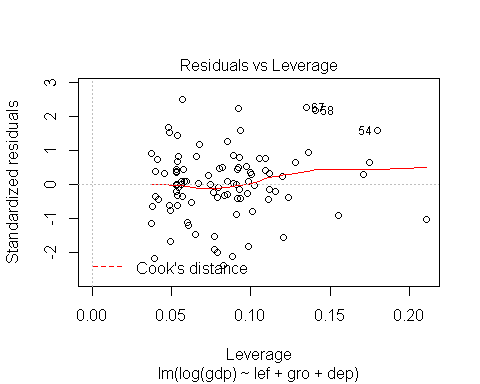
**Transforming GDP**

With the goal of producing a more reasonable residual plot, we took to transforming the GDP variable. A log-based transformation proved to hold the most merit. Again using backwards elimination at a cutoff of 0.1, we ended up with the model:

**log(gdp) ~ lef + gro + dep.** It has an AIC of 209.9305 and an adj-R^2 of 0.831.

The residual plot is much nicer, but features a right end with lower variance than the rest of the plot. Unequal variance is a worrisome fault in our assumptions, but notwithstanding that caveat, we believe in the model’s robustness. The summary table for this model is as follows:

## Call:  
## lm(formula = log(gdp) ~ lef + gro + dep, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.54470 -0.27259 0.00888 0.32990 1.65535   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.047842 1.311171 3.087 0.002694 \*\*   
## lef 0.064662 0.014146 4.571 1.56e-05 \*\*\*  
## gro2 0.560189 0.316839 1.768 0.080479 .   
## gro3 1.794455 0.261095 6.873 8.31e-10 \*\*\*  
## gro4 2.026960 0.333589 6.076 2.99e-08 \*\*\*  
## gro5 -0.023556 0.301106 -0.078 0.937819   
## gro6 1.015495 0.369024 2.752 0.007181 \*\*   
## dep -0.016768 0.004791 -3.500 0.000728 \*\*\*  
## Residual standard error: 0.6794 on 89 degrees of freedom  
## Multiple R-squared: 0.8432, Adjusted R-squared: 0.8308   
## F-statistic: 68.36 on 7 and 89 DF, p-value: < 2.2e-16

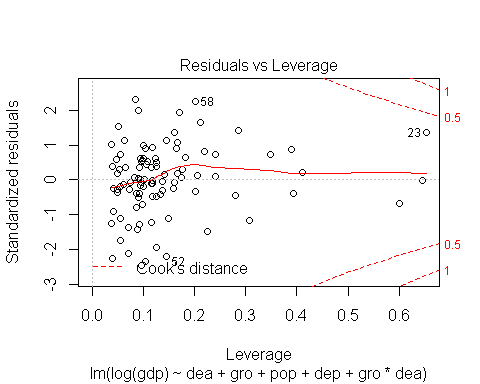
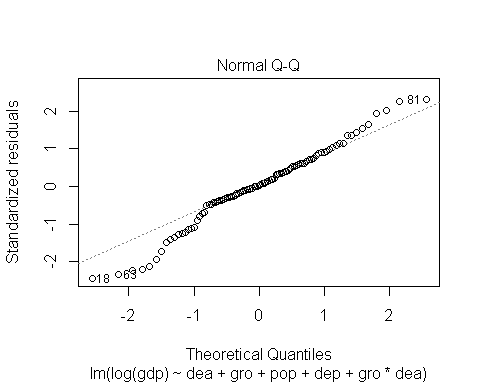
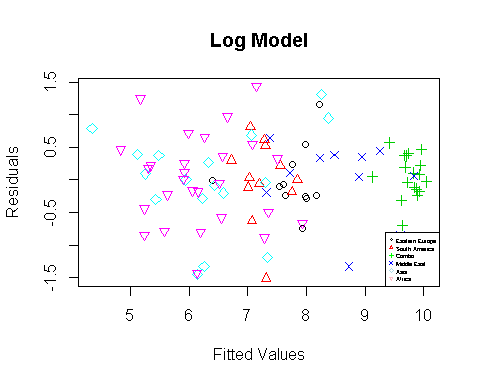
VIF assures us that multicollinearity is a non-factor with values being under the 10 cutoff.

## lef 5.041262## gro 4.215896 ## dep 5.473953

**Including Interactions**

In an aim to increase the effectiveness of our model, we added interaction variables. Logically, the easily observable interactions come from the effects each variable has and how that differs from country group to country group. Thus, we added interactions between group and every other variable. From there, we used backwards elimination. Many of the interactions had decent significance, but hardly improved our model criteria. Resultantly, only death rate by country group remained. As we can see in the table below, death rate, the interaction term and population size replace life expectancy of women.

## Call:  
## lm(formula = log(gdp) ~ dea + gro + pop + dep + gro \* dea, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.50923 -0.24043 0.01313 0.37715 1.43415   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.891e+00 1.125e+00 7.014 5.74e-10 \*\*\*  
## dea 1.583e-01 9.939e-02 1.592 0.11513   
## gro2 1.770e+00 1.132e+00 1.563 0.12181   
## gro3 3.399e+00 1.453e+00 2.339 0.02173 \*   
## gro4 3.627e+00 1.208e+00 3.001 0.00355 \*\*   
## gro5 2.458e+00 1.141e+00 2.154 0.03413 \*   
## gro6 3.813e+00 1.152e+00 3.311 0.00138 \*\*   
## pop -1.099e-06 4.733e-07 -2.321 0.02271 \*   
## dep -2.595e-02 4.167e-03 -6.229 1.85e-08 \*\*\*  
## dea:gro2 -1.167e-01 1.062e-01 -1.099 0.27491   
## dea:gro3 -1.287e-01 1.433e-01 -0.898 0.37187   
## dea:gro4 -1.590e-01 1.320e-01 -1.205 0.23177   
## dea:gro5 -2.639e-01 1.072e-01 -2.462 0.01589 \*   
## dea:gro6 -2.828e-01 1.033e-01 -2.738 0.00756 \*\*   
##   
## Residual standard error: 0.652 on 83 degrees of freedom  
## Multiple R-squared: 0.8653, Adjusted R-squared: 0.8442   
## F-statistic: 41.02 on 13 and 83 DF, p-value: < 2.2e-16



It’s AIC and adjusted R^2 are nearly the same as the non-interaction model, which makes this model suffer as it has more variables. However, this model has better plots with a less pronounced fan shape in the residual plot and a better normal plot. Making a judgement between the two is a matter of efficiency versus soundness of model. The second model with interactions is better, marginally, in every way, but at the expense of simplicity--again marginally.

Multicollinearity is difficult to decipher for models with interactions as interaction terms are obviously related to their respective counterpart predictor variables.

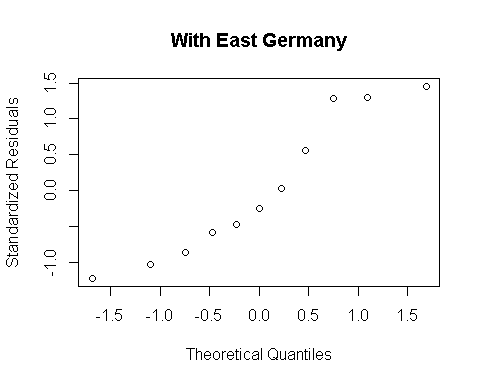
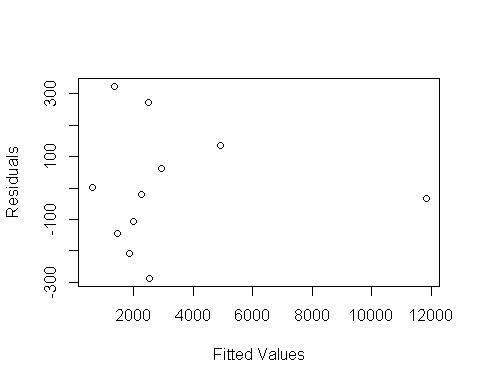
As we can see from both of our complete models, the group that clumps to the right and has the smallest variance comes from our combo group. Accompanied with this is the clear indication that groups are significantly related to their respective countries' GDP per capita values. Our factored group variable is extremely significant in both interactions and as a base variable. The next step is to isolate the groups and evaluate models that best fit each.

**CREATING INDIVIDUAL MODELS FOR EACH GROUP**

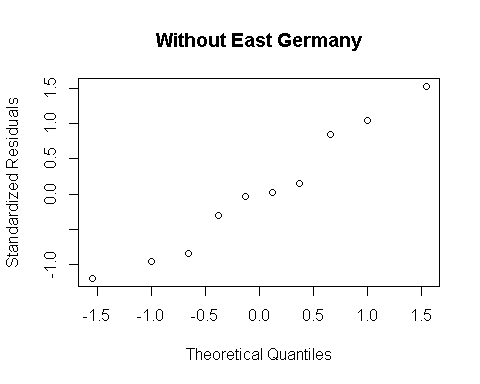
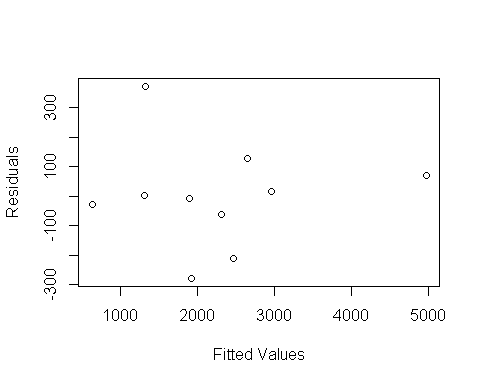
**Evaluating Eastern Europe**

We arrive at the model **gdp ~ dea + lem + pop + age + dep** which features a few of the base variables and a markedly high adjusted R^2 value of 0.9925 and an AIC of 159.6315. The step function creates the same model with infant mortality and female life expectancy still present, rendering an AIC of 157. However, both of those variables had p-values higher than .1 and thus create a slightly better, but notably less efficient model. So we stick with our backwards elimination-bred structure.

What is notable here however is the effect of East Germany on the model. It carries a large GDP per capita in comparison to the rest of Eastern Europe, and therefore is very influential on the dataset. The normal plot has a slight err to it. Residuals are spread evenly in the clumped area of the countries with smaller GDP. East Germany however is fitted rather precisely as is observed from its small residual.



To test the 'contribution' that the former East Germany has on the model, we test the same formula without it. Removing East Germany and running the same model lowers the adjusted R^2 to a still unreasonable 0.9506 and lowers AIC to 144.9833, so the model, and its components, remain sound. The residual plot has less skew, and the normal plot has a less distinct curve.



Multi-collinearity proves a non-factor with E. Germany, but without it, the values increase dramatically:

## dea lem pop age dep

w/E.Germany 1.928555 1.518601 2.101974 2.590852 2.227930

w/out 54.36517 4.20541 7.97133 101.4144 45.759654

So we remove highly mutlicollinear variables and end up with a model of

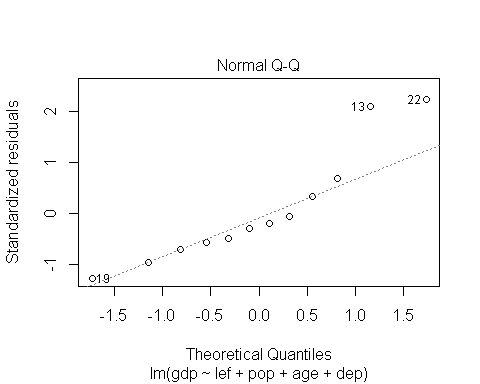
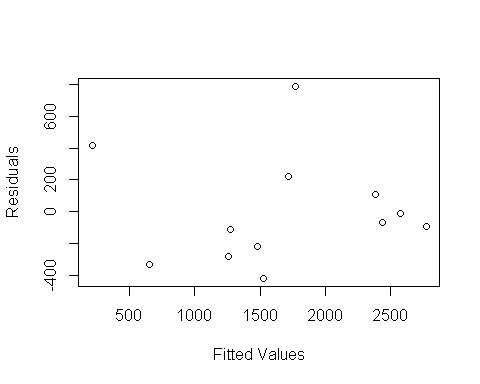
**Gdp ~ lem + pop + dep**, which has more realistic criterion values of 0.778 Adj R^2 and 160 AIC. It

has no multicollinearity issues and features much the same plots as above, but with higher residual levels.

**Studying South America and Mexico**

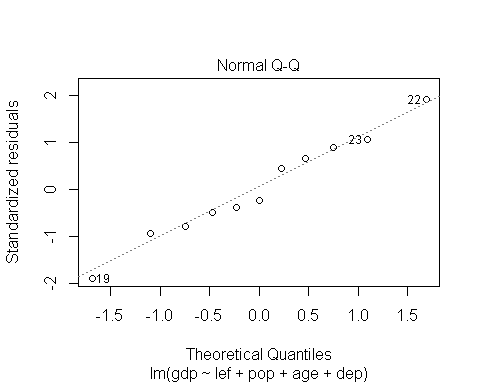
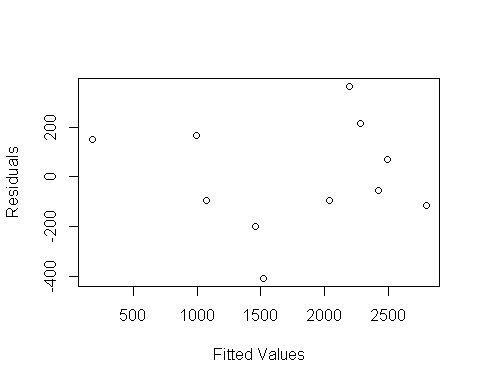
Our favored model through backwards elimination is **gdp ~ lef + pop + age + dep.**

This has an adjusted R^2 of 0.7459 and an AIC of 185.0144. Again, step allows for lower AIC but at the expense of retaining inefficient variables. The following plots suggest an inconsistency in the model. Although small in observations, the residual plot is slightly concerning in that 8 of the 12 points are negative. The normal plot is offset at the rightmost points. Most telling is the Cook's distance of Bolivia (point 13), which is well over 1 given a large GDP relative to its comparatively underwhelming predictor variable stats.



Removing Bolivia yields an improved model with an adjusted R^2 of 0.8824 and an AIC of 160.6462. The predictor variables remain significant. We test to see if Bolivia was an outlier with predict and a significance level of 1-0.05/12.

The interval obtained is [-4071.533, 1424.029], Bolivia’s wealth is 630. This illustrates that Bolivia fits with the new model, the issue is with its leverage more so than its GDP disparity. Nevertheless, our model without Bolivia has an improved residual plot and is fitted normally.



Based on these data, we'd consider the model without Bolivia to be superior, but note that there is something to consider regarding an additional factor, unrelated to population, that has affected Bolivia's GDP per capita.

Multi-collinearity proves a non-factor.

## lef pop age dep

## 3.149656 1.215085 5.004212 4.563514

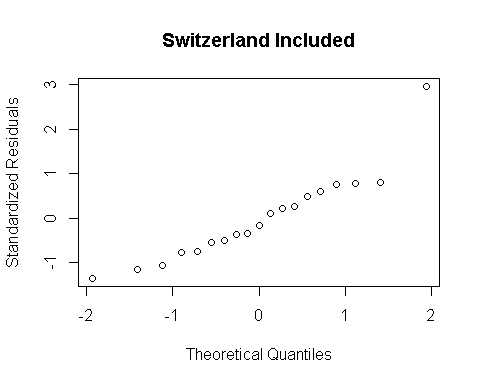
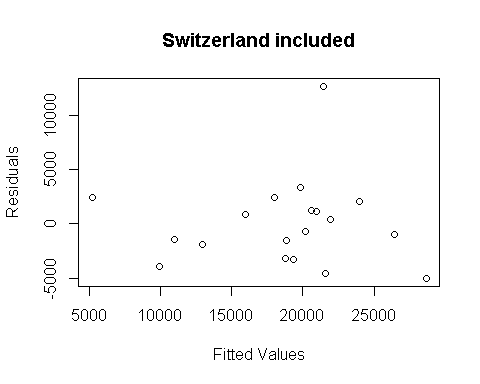
**Contemplating group 3**

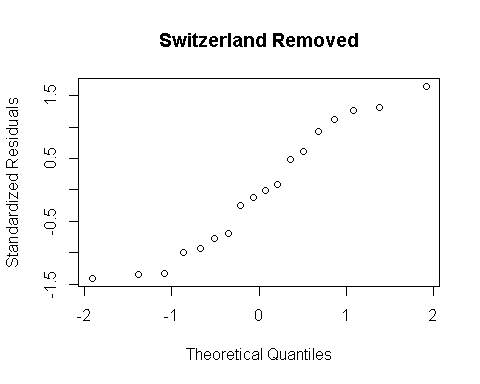
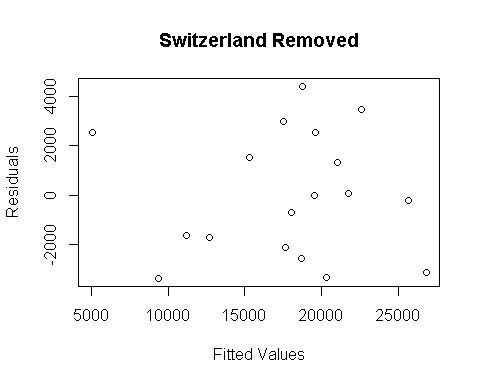
Our best model through backwards elimination is below:

What is immediately evident through the residual plots of our models is Switzerland's potential as an outlier. Our variables at hand do not consider details such as banking revenue, or some other outlet Switzerland has that leads it to such high GDP per capita levels. In terms of these population-based variables, Switzerland is not elite.

To prove its outlier status, we remove it from the best model involving Switzerland **GDP~ Birth + InfantMortality + Population + Median Age**, and run a Bonferroni monitored prediction interval with significance of 1-.05/19. The interval obtained is [12556.4, 27610.93].

Switzerland's response GDP (34064) is clearly influenced in a great way by other variables and is not confidently predicted by this model. The removal of Switzerland increases the predictive ability of the model and the residual shape (AIC drops from 379 to 343 and ajdR^2 rises from 0.58 to 0.776).



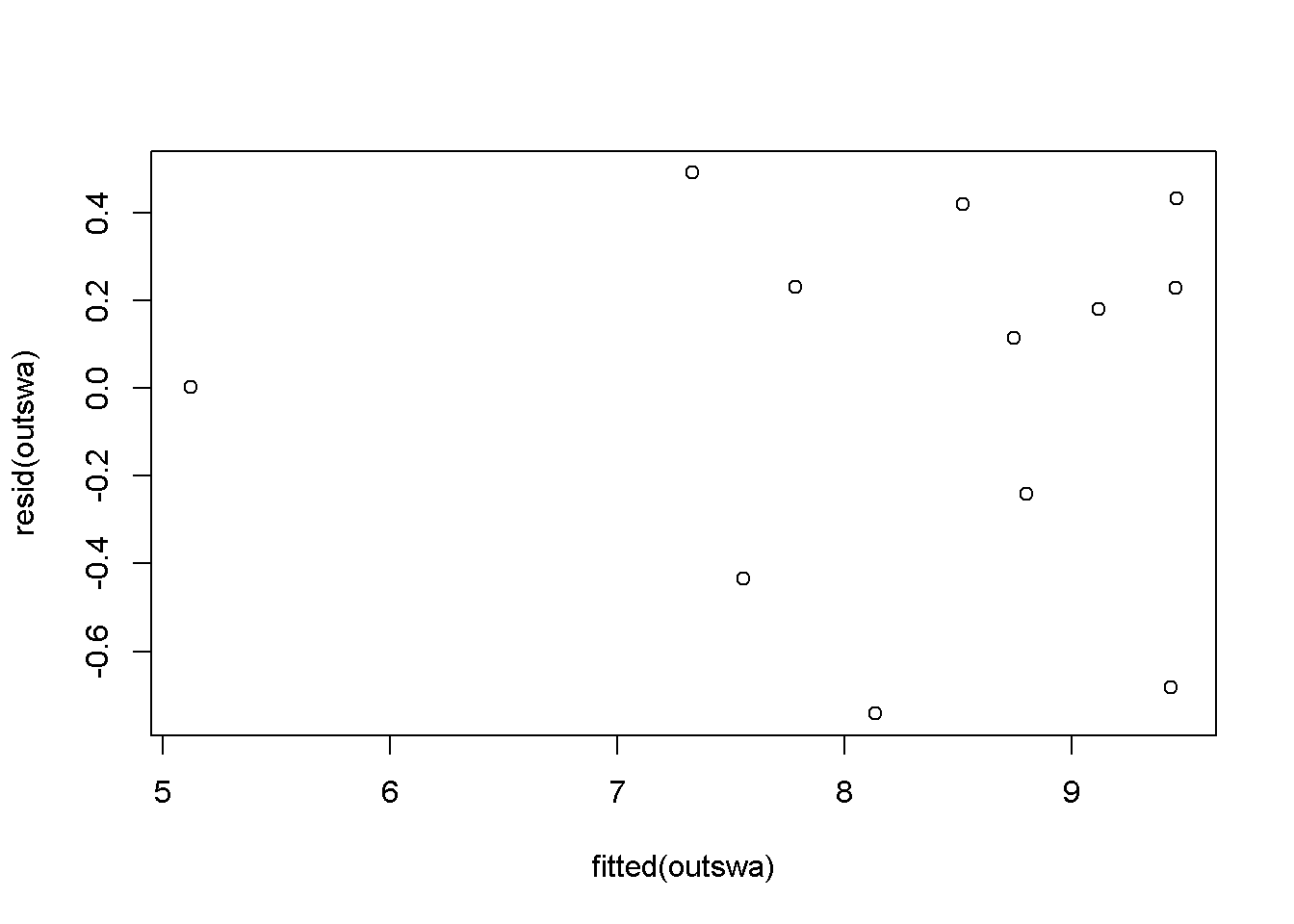


VIF values showed no multicollinearity issues.

**Summarizing Middle East**

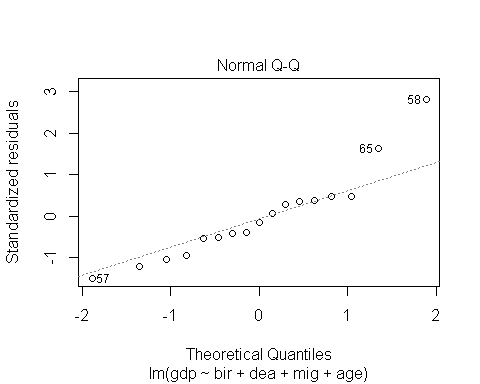
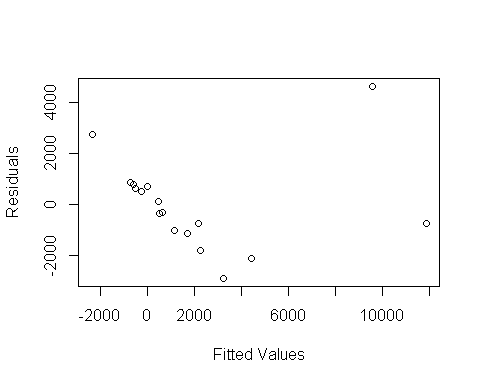
The Middle East was a difficult model to produce; the countries vary so significantly in predictor variables and gross domestic product per capita that every model had its issues. Our best result was to log transform the GDP values with a formula of **Log(gdp) ~ bir + lem + dep**

This granted an adjusted R^2 of 0.85, an AIC of 22 and no VIF values near or above 10. None of the cook distances are above 0.5 so we have no reason to remove any points. Given the difficulty of the region’s variables, our best residual plot still has noticeable issues in its lopsidedness and high values.

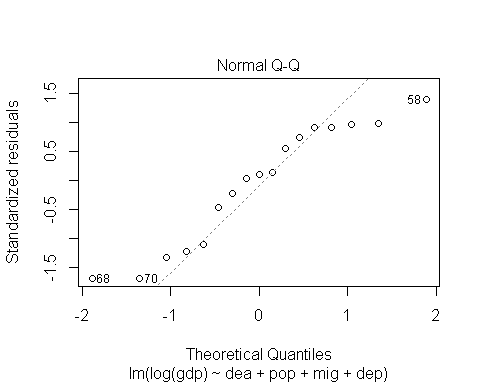
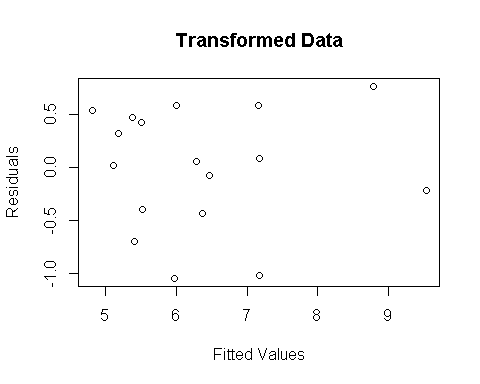


**Analyzing Asia**

Our best, through backwards elimination, normal Asia model re-introduces our original problem of an incorrect approach. The data needs to be transformed in some way as every normal model features the same skew featured below.



Resultantly, we return to the log transformation of GDP, which again produces an improved model and residual plot. The model is **log(gdp)~dea+pop+mig+dep** with a 0.7855 adj R^2 and AIC of 39.



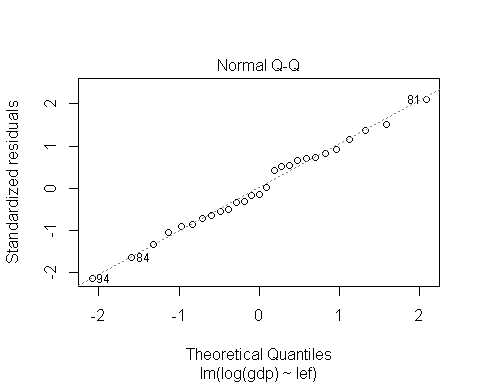
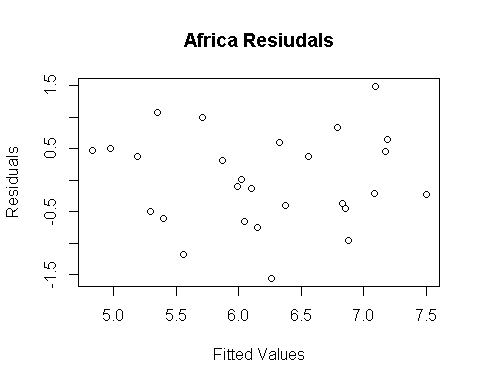
**Analyzing Africa**

Using a normal analysis, we get a rather ineffective model with a low R^2 and high AIC. On top of that, the variables feature pronounced multicollinearity.

VIF values:

## lef pop age dep   
## 1.545762 1.051320 18.238049 16.629151

So we return to a log model with **log(gdp)~lef.** Africa had a lot of variety and thus was best predicted by just the one variable. It has a comparably poor adjusted R^2 and AIC of 0.4914 and 64 respectively. The plots from this model, however, are the most consistent with our assumptions of normalcy, model fit and homogeneous variance.



**CONCLUSION**

The conclusions from these data illustrate the complexity of explaining Gross Domestic Product per capita. In a world-wide attempt to predict individual wealth, we most prefer a log-transformed model with interactions. Breaking down into specific country groups or creating a categorical variable based on country groups is imperative in GDP prediction. Our groups typically lack a sufficient amount of data-points to create pristine, predictive models.

Removing outliers and high leverage points is helpful in cleaning up the models, but their presence indicates not a blip, error or simple mutation. These data are nations, and thus should not be ignored. The difficulties associated with them are representative of a few things. There are likely outside variables that do not pertain to population that affect GDP per capita. These types of variables (e.g. war, corruption, etc.) can be difficult to account for, especially in 1992. In general, the accuracy of the data at hand is unknown. A general caution should be taken to them as some countries are likely to be harder to record data for than others.

What we can gather from our report is the following:

**Birth rates:** The birth rate variable was influential only for the combo group and the Middle East group, having a positive relation with GDP per capita

**Death rates:** Prominent in the Asia model, predictably having a negative relation with GDP per capita, death rate was also in the interaction model as its relation with GDP varied wildly between countries

**Infant mortality:** Only significant for the super-power group, not good for GDP p/c

**Life expectancies:** Highly related, one of these two was often in group models and the full model. Higher life expectancies were very indicative of better wealth

**Population:** An interesting case: was positive for GDP in two models, but negative in the full model and in Asia’s model. It shows how difficult it is to gauge these relationships in a full scale world model

**Age:** A positive indicator of GDP for the super group and South America, otherwise was fairly insignificant

**Migration:** Largely insignificant in most models, barring Asia where higher migration rates related to better GDP per capita.

**Dependency Ratio:** A strong variable, dependency showed up in most models as a negative indicator of GDP. Oddly, in South America it was a positive indicator of wealth.

**Appendix:**

Eastern Europe Model:

## Call:

## lm(formula = gdp ~ lem + pop + dep, data = est[-4, ])

##

## Residuals:

## Min 1Q Median 3Q Max

## -658.27 -366.28 88.14 220.01 860.15

##

## Coefficients:

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

## (Intercept) -5.291e+03 7.133e+03 -0.742 0.48622

## lem 2.353e+02 1.060e+02 2.219 0.06828 .

## pop -1.378e-02 5.922e-03 -2.327 0.05886 .

## dep -1.140e+02 1.976e+01 -5.768 0.00118 \*\*

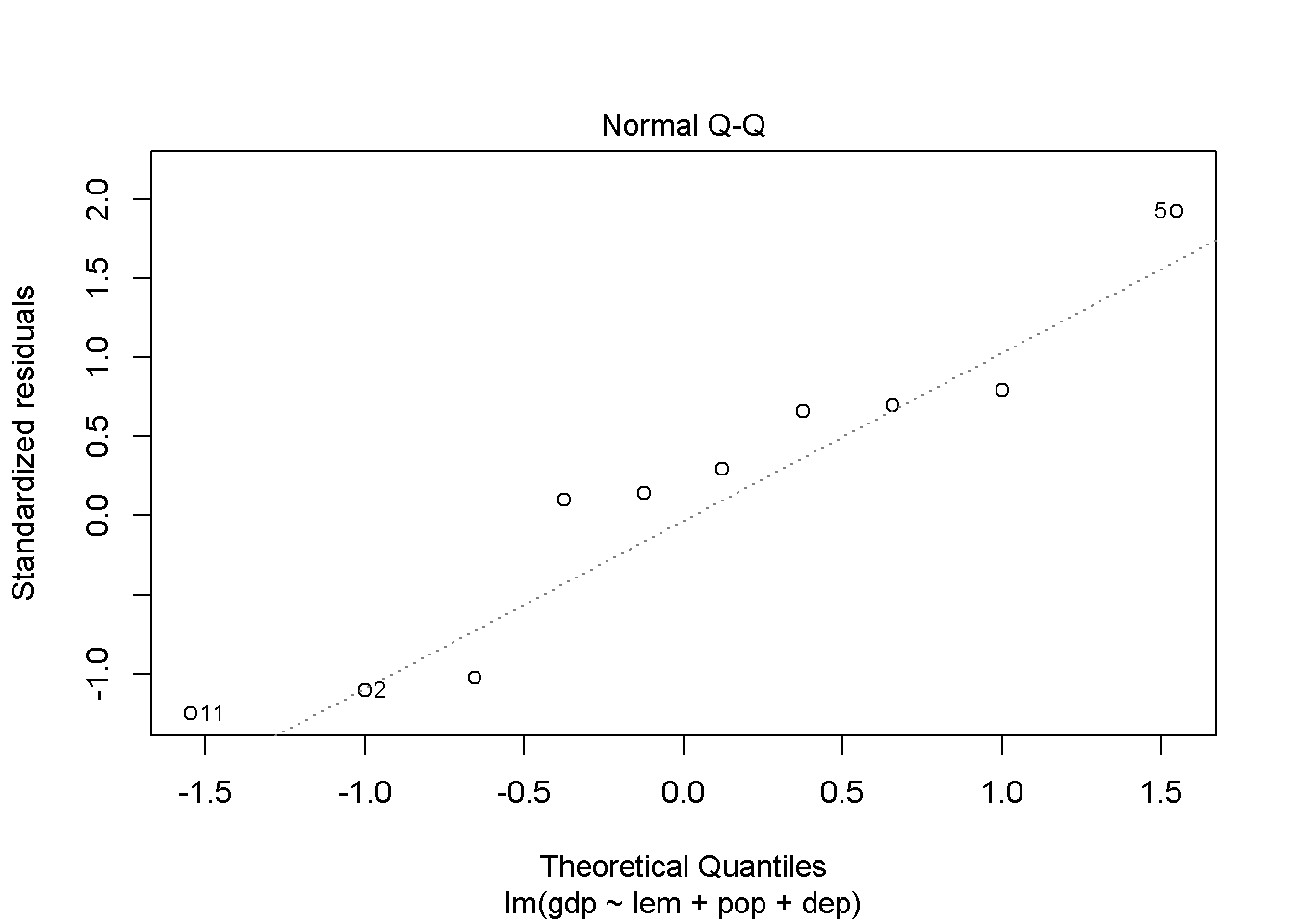
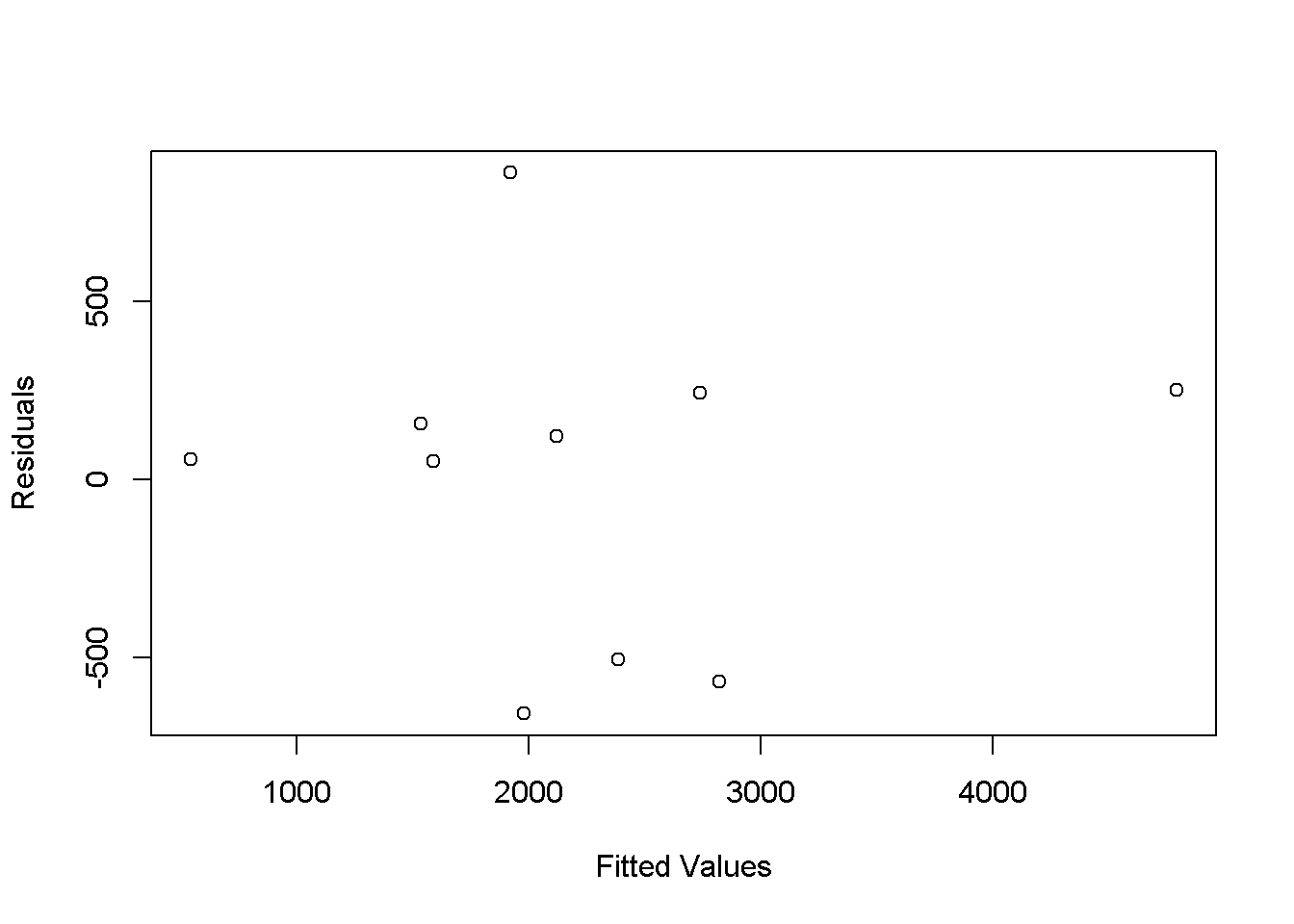
## ---

## Residual standard error: 565.3 on 6 degrees of freedom

## Multiple R-squared: 0.8526, Adjusted R-squared: 0.7788

## F-statistic: 11.57 on 3 and 6 DF, p-value: 0.006611

Final model plots below:



**South America Extras:**

## Call:

## lm(formula = gdp ~ lef + pop + age + dep, data = sam[-2, ])

##

## Residuals:

## Min 1Q Median 3Q Max

## -409.3 -104.9 -53.4 158.3 365.5

##

## Coefficients:

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

## (Intercept) -2.227e+04 3.957e+03 -5.628 0.001346 \*\*

## lef 2.329e+02 4.522e+01 5.150 0.002115 \*\*

## pop 1.547e-02 2.096e-03 7.379 0.000318 \*\*\*

## age 1.109e+02 5.507e+01 2.015 0.090584 .

## dep 4.532e+01 1.347e+01 3.366 0.015122 \*

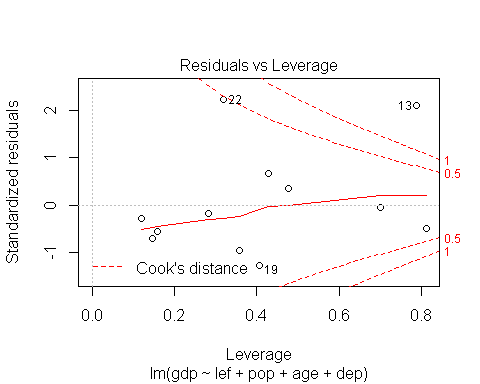
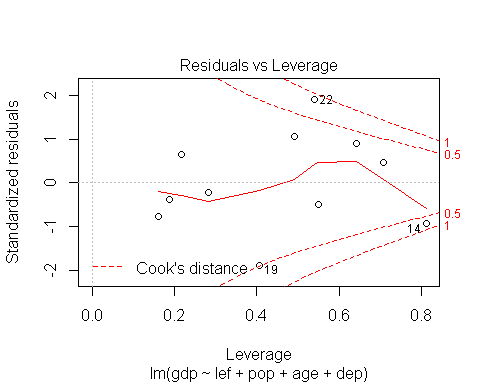
## ---

## Residual standard error: 281.7 on 6 degrees of freedom

## Multiple R-squared: 0.9295, Adjusted R-squared: 0.8824

## F-statistic: 19.76 on 4 and 6 DF, p-value: 0.00133

Cook’s distance for South America & Mexica with Bolivia and without Bolivia:

**Combo Group Extras**

## Call:

## lm(formula = gdp ~ bir + inf + pop + age, data = com[-14, ])

##

## Residuals:

## Min 1Q Median 3Q Max

## -3366.6 -1998.4 -127.5 2264.2 4392.9

##

## Coefficients:

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

## (Intercept) -2.636e+04 1.649e+04 -1.599 0.133861

## bir 1.053e+03 4.041e+02 2.607 0.021719 \*

## inf -1.866e+03 3.834e+02 -4.868 0.000307 \*\*\*

## pop 2.427e-02 1.119e-02 2.168 0.049297 \*

## age 1.256e+03 3.414e+02 3.681 0.002770 \*\*

## Residual standard error: 2858 on 13 degrees of freedom

## Multiple R-squared: 0.829, Adjusted R-squared: 0.7763

## F-statistic: 15.75 on 4 and 13 DF, p-value: 6.613e-05

**Middle East Extras:**

## Call:

## lm(formula = log(gdp) ~ bir + lem + dep, data = swa)

## Residuals:

## Min 1Q Median 3Q Max

## -0.7418 -0.2890 0.1468 0.2776 0.4913

##

## Coefficients:

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

## (Intercept) 1.354967 2.107258 0.643 0.53821

## bir 0.077913 0.031899 2.443 0.04041 \*

## lem 0.117888 0.023761 4.961 0.00110 \*\*

## dep -0.027211 0.007996 -3.403 0.00932 \*\*

## ---

## Residual standard error: 0.5022 on 8 degrees of freedom

## Multiple R-squared: 0.8942, Adjusted R-squared: 0.8546

## F-statistic: 22.55 on 3 and 8 DF, p-value: 0.0002947

**Asia Extras:**

## Call:

## lm(formula = log(gdp) ~ dea + pop + mig + dep, data = asi)

## Residuals:

## Min 1Q Median 3Q Max

## -1.05011 -0.40032 0.06253 0.47913 0.77045

##

## Coefficients:

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

## (Intercept) 1.109e+01 6.575e-01 16.860 1.01e-09 \*\*\*

## dea -1.156e-01 4.298e-02 -2.690 0.019660 \*

## pop -1.024e-06 4.939e-07 -2.073 0.060412 .

## mig 5.773e-02 1.927e-02 2.996 0.011156 \*

## dep -3.337e-02 6.792e-03 -4.913 0.000358 \*\*\*

## ---

## Residual standard error: 0.6556 on 12 degrees of freedom

## Multiple R-squared: 0.8391, Adjusted R-squared: 0.7855

## F-statistic: 15.65 on 4 and 12 DF, p-value: 0.0001047

Summary table for Africa:

## Call:

## lm(formula = log(gdp) ~ lef, data = afr)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1.5575 -0.4801 -0.1066 0.4896 1.4855

##

## Coefficients:

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

## (Intercept) 0.45429 1.13297 0.401 0.692

## lef 0.10610 0.02076 5.111 2.8e-05 \*\*\*

## ---

## Residual standard error: 0.7446 on 25 degrees of freedom

## Multiple R-squared: 0.511, Adjusted R-squared: 0.4914

## F-statistic: 26.12 on 1 and 25 DF, p-value: 2.797e-05

Data obtained from:

<http://esa.un.org/unpd/wpp/DataQuery/>

<http://www.amstat.org/publications/jse/datasets/poverty.dat.txt>