Annotated Appendix

The name of my data is countries of the world, and I got this data from Kaggle. The data was created by Mr. Fernando Lasso and was last updated on 04/26/2018. The data is public and can be used by anyone.

Link for the dataset: <https://www.kaggle.com/fernandol/countries-of-the-world>

Description of the variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable name** | **Original definition** | **Units** | **Range or Levels** | **Rationale** |
| Net Migration | The net migration rate is the difference between the number of immigrants and the number of emigrants throughout the year, expressed in percentage | % | -100 to +100 (Negative if immigration is higher, positive is emigration is higher) | Response variable |
| GDP per Capita | GDP per capita measures the economic output of a nation per person | US $ | >0 | Main explanatory variable of interest |
| Literacy | the percentage of the population of a given age group that can read and write | % | 0-100 | Possible confounding variable |
| Population Density | the concentration of individuals within a species in a specific geographic locale | Per square mile | >0 | Possible confounding variable |
| Death Rate | death rate is a measure of the number of deaths in a particular population, scaled to the size of that population, per year | % | 0-100 | Possible confounding variable |
| Service | The percentage of composition of GDP by service sector businesses. | % | 0-100 | Possible confounding variable |
| Infant mortality | Infant mortality is the rate of a child born in a specific year or period dying before reaching the age of one | Per thousand births | 0 – 1000 | Possible confounding variable |
| Phones | The number of people who own and use a smartphone, expressed in per thousand population. | Per thousand | 0 – 1000 | Possible confounding variable |
| Region | an area, especially part of a country or the world having definable characteristics but not always fixed boundaries. |  | 0,1 | Possible confounding variable |

References:

Podra, Olha & Levkiv, Halyna & Koval, Ganna & Petryshyn, Nataliia & Bobko, Ulyana. (2020). The impact of migration processes on the economy of Ukraine: Trends, reasons, consequences. Journal of the Geographical Institute Jovan Civic, SASA. 70. 171-179. 10.2298/IJGI2002171P.

1st reference: This paper aims to research the impact of migration on the economic and social development of Ukraine’s population. The researchers conduct sociological research to understand the intentions and motives of potential migrants. Based on the conducted research and official domestic and international data, the paper concludes that migration growth in Ukraine is related to “structural changes in the economy and the labor market, economic growth slowdown, low wages, and quality of life compared to the nearest neighbor countries, visa liberalization with EU countries, and access to education abroad.” I have decided to refer this in my paper as it is directly related to the topic that I am investigating.

Khan, Khalid & Su, Chi-Wei & Tao, Ran & Yang, Lin. (2019). Does Remittance Outflow Stimulate or Retard Economic Growth? International Migration. 57. 10.1111/imig.12615.

2nd reference: This paper aims to evaluate the association between remittance outflow (RMO) and economic growth in the Gulf Cooperation Council (GCC) countries. The article uses the panel bootstrap Granger method as its methodology. Based on this method, the research paper finds that RMO has an important effect on GDP per capita for Bahrain, Oman, and Saudi Arabia. As the article explains the impact that remittance, a common incentive for migrants, has on economic growth, I believe that it is relevant to my research question.

### Data Cleaning

Getting rid of the column with missing data is probably not a good idea because by doing so, I lost the data for 48 countries.

Currently we have 110 missing values.

sum(is.na(data))

## [1] 110

The columns Industry and Agriculture had a lot of missing values and I was not going to use these variable for my model, so I decided to get ride of them.

cols.dont.want <- c("Agriculture", "Industry")  
data2 <- data[, ! names(data) %in% cols.dont.want, drop = F]  
  
sum(is.na(data2))

## [1] 79

Before, we had 110 missing values but after removing we have 79 which is a good progress.

I looked at the dataset and rows West Sahara and Walis and Futuna has a lot of missing values so I am going to get rid of them.

data3 = data2[-c(222), ]  
  
data4 = data3[-c(223), ]  
  
sum(is.na(data4))

## [1] 67

Now, we have 67 missing values in our dataset which are randomly scattered throughout the dataset so I am going to substitute missing cells with column mean.

#library(zoo)  
#na.aggregate(data4)  
  
#data6<- function(x) replace(x, is.na(x), mean(x, na.rm = TRUE))  
#data4[] <- lapply(data4, data6)  
  
data6 = na\_mean(data4)

We can see that all the missing values have been taken care of at this point.

sum(is.na(data6))

## [1] 0

The main variables that I want to use are Net Migration, GDP per capita, Literacy, Population density, Death Rate, Service, Infant Mortality, Phones, and Region.

### Plots and Summary Statistics

table(data6$Region)

##   
## ASIA (EX. NEAR EAST) BALTICS   
## 28 3   
## C.W. OF IND. STATES EASTERN EUROPE   
## 12 12   
## LATIN AMER. & CARIB NEAR EAST   
## 45 16   
## NORTHERN AFRICA NORTHERN AMERICA   
## 5 5   
## OCEANIA SUB-SAHARAN AFRICA   
## 20 51   
## WESTERN EUROPE   
## 28

favstats(data6$Net.migration)

## min Q1 median Q3 max mean sd n missing  
## -20.99 -0.91 0 0.99 23.06 0.038125 4.878343 225 0

favstats(data6$Pop..Density..per.sq..mi..)

## min Q1 median Q3 max mean sd n missing  
## 0 29.3 79.5 191.8 16271.5 382.152 1667.249 225 0

favstats(data6$GDP....per.capita.)

## min Q1 median Q3 max mean sd n missing  
## 500 1900 5600 15700 55100 9716.444 10063.55 225 0

favstats(data6$Literacy....)

## min Q1 median Q3 max mean sd n missing  
## 17.6 76.6 90.7 97.8 100 82.99615 18.87707 225 0

favstats(data6$Deathrate)

## min Q1 median Q3 max mean sd n missing  
## 2.29 5.97 8.1 10.59 29.74 9.241345 4.967699 225 0

favstats(data6$Service)

## min Q1 median Q3 max mean sd n missing  
## 0.062 0.447 0.5660664 0.672 0.954 0.5660664 0.1605756 225 0

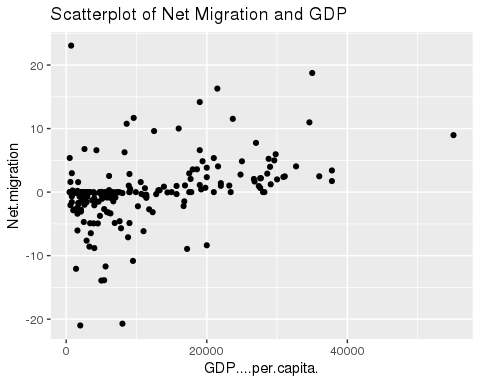
favstats(data6$Infant.mortality..per.1000.births.)

## min Q1 median Q3 max mean sd n missing  
## 2.29 8.19 21.03 55.51 191.19 35.50696 35.31082 225 0

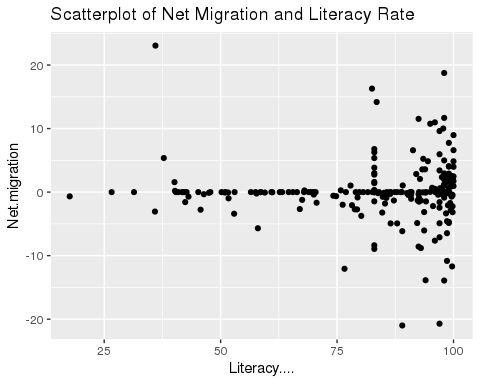
favstats(data6$Phones..per.1000.)

## min Q1 median Q3 max mean sd n missing  
## 0.2 38.4 181.6 384.9 1035.6 236.5905 226.8354 225 0

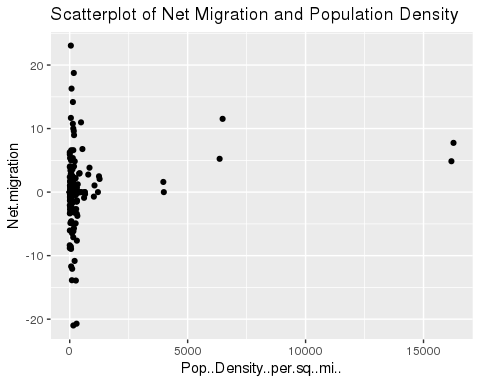
gf\_point(Net.migration~ GDP....per.capita., data=data6, title= "Scatterplot of Net Migration and GDP")



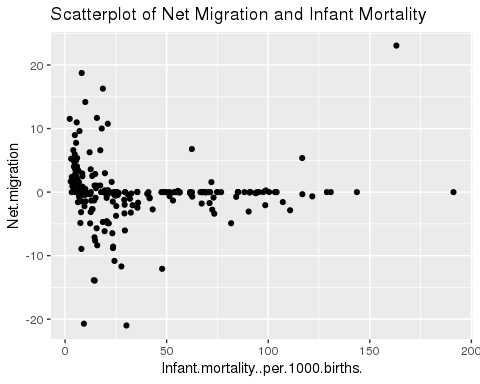
gf\_point(Net.migration~ Literacy...., data=data6, title = "Scatterplot of Net Migration and Literacy Rate")



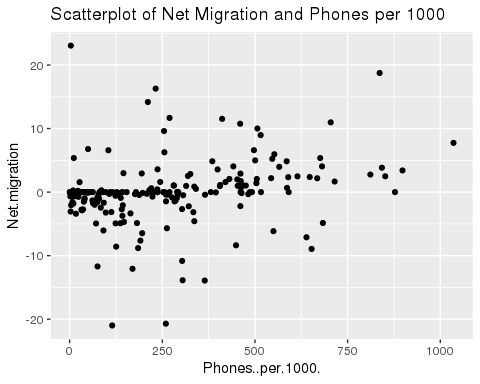
gf\_point(Net.migration ~ Pop..Density..per.sq..mi.., data=data6, title = "Scatterplot of Net Migration and Population Density")



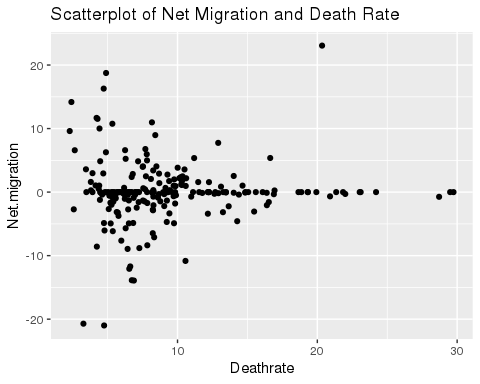
gf\_point(Net.migration ~ Infant.mortality..per.1000.births., data=data6, title = "Scatterplot of Net Migration and Infant Mortality")



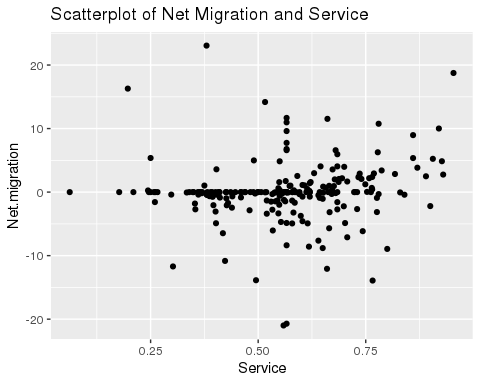
gf\_point(Net.migration ~ Phones..per.1000. , data=data6, title = "Scatterplot of Net Migration and Phones per 1000")



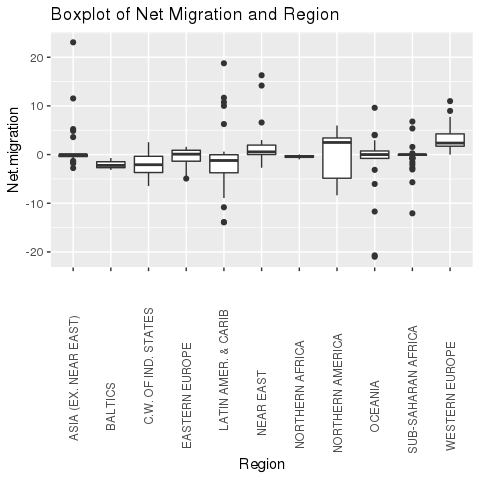
gf\_point(Net.migration~ Deathrate, data = data6, title = "Scatterplot of Net Migration and Death Rate")



gf\_point(Net.migration~ Service, data = data6, title = "Scatterplot of Net Migration and Service")



#gf\_boxplot(Net.migration~ Region, data = data6, las=2, par(mar = c(12, 5, 4, 2)+ 0.1))  
  
  
library(ggplot2)  
ggplot(data6, aes(Region, Net.migration)) + geom\_boxplot() + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 0.5))+ ggtitle("Boxplot of Net Migration and Region")

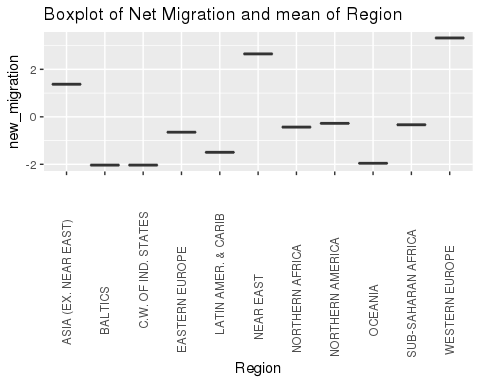


I decided to see the migration rate for each region and which region has the higest and lowest migration rate.

data7 = data6 %>% group\_by(Region) %>% summarise(new\_migration =mean(Net.migration))  
data7

## # A tibble: 11 × 2  
## Region new\_migration  
## <fct> <dbl>  
## 1 "ASIA (EX. NEAR EAST) " 1.37   
## 2 "BALTICS " -2.03   
## 3 "C.W. OF IND. STATES " -2.03   
## 4 "EASTERN EUROPE " -0.648  
## 5 "LATIN AMER. & CARIB " -1.49   
## 6 "NEAR EAST " 2.65   
## 7 "NORTHERN AFRICA " -0.432  
## 8 "NORTHERN AMERICA " -0.274  
## 9 "OCEANIA " -1.96   
## 10 "SUB-SAHARAN AFRICA " -0.335  
## 11 "WESTERN EUROPE " 3.32

ggplot(data7, aes(Region, new\_migration)) + geom\_boxplot() + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 0.5))+ ggtitle("Boxplot of Net Migration and mean of Region")



This piece of code shows which countries lie in which region.

data6 %>% group\_by(Region, Country)%>%summarize(n())

## `summarise()` has grouped output by 'Region'. You can override using the `.groups` argument.

## # A tibble: 225 × 3  
## # Groups: Region [11]  
## Region Country `n()`  
## <fct> <fct> <int>  
## 1 "ASIA (EX. NEAR EAST) " "Afghanistan " 1  
## 2 "ASIA (EX. NEAR EAST) " "Bangladesh " 1  
## 3 "ASIA (EX. NEAR EAST) " "Bhutan " 1  
## 4 "ASIA (EX. NEAR EAST) " "Brunei " 1  
## 5 "ASIA (EX. NEAR EAST) " "Burma " 1  
## 6 "ASIA (EX. NEAR EAST) " "Cambodia " 1  
## 7 "ASIA (EX. NEAR EAST) " "China " 1  
## 8 "ASIA (EX. NEAR EAST) " "East Timor " 1  
## 9 "ASIA (EX. NEAR EAST) " "Hong Kong " 1  
## 10 "ASIA (EX. NEAR EAST) " "India " 1  
## # … with 215 more rows

### 

### Correlation between the Variables

I am going to find the correlation between Net Migration and all the other variables to find out what variables should I add in my model.

cor(~cbind(Net.migration, Literacy...., Deathrate, Service),data=data6)

## Net.migration Literacy.... Deathrate Service  
## Net.migration 1.000000000 -0.008518261 0.03360515 0.1140510  
## Literacy.... -0.008518261 1.000000000 -0.38918788 0.4704585  
## Deathrate 0.033605151 -0.389187885 1.00000000 -0.3525842  
## Service 0.114051024 0.470458539 -0.35258422 1.0000000

cor(~cbind(Net.migration,GDP....per.capita., Phones..per.1000.),data = data6)

## Net.migration GDP....per.capita. Phones..per.1000.  
## Net.migration 1.0000000 0.3822726 0.2395682  
## GDP....per.capita. 0.3822726 1.0000000 0.8303195  
## Phones..per.1000. 0.2395682 0.8303195 1.0000000

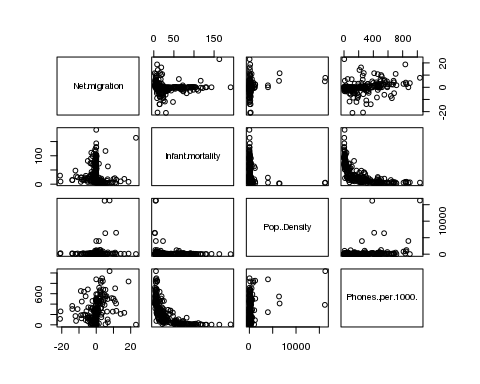
data7 = rename(data6, Infant.mortality= Infant.mortality..per.1000.births., Pop..Density=Pop..Density..per.sq..mi..)  
  
cor(~cbind(Net.migration,Infant.mortality, Pop..Density),data=data7)

## Net.migration Infant.mortality Pop..Density  
## Net.migration 1.00000000 -0.02501535 0.1780254  
## Infant.mortality -0.02501535 1.00000000 -0.1442283  
## Pop..Density 0.17802539 -0.14422830 1.0000000

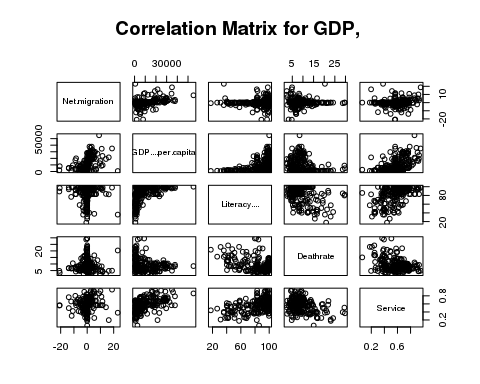
cor(~cbind(Net.migration,Infant.mortality,GDP....per.capita., Deathrate, Region, Literacy....), data=data7)

## Net.migration Infant.mortality GDP....per.capita.  
## Net.migration 1.000000000 -0.02501535 0.3822726  
## Infant.mortality -0.025015350 1.00000000 -0.6004780  
## GDP....per.capita. 0.382272643 -0.60047803 1.0000000  
## Deathrate 0.033605151 0.65573399 -0.2013079  
## Region 0.069515394 0.14915959 0.1905230  
## Literacy.... -0.008518261 -0.75091320 0.4971458  
## Deathrate Region Literacy....  
## Net.migration 0.03360515 0.06951539 -0.008518261  
## Infant.mortality 0.65573399 0.14915959 -0.750913203  
## GDP....per.capita. -0.20130785 0.19052297 0.497145823  
## Deathrate 1.00000000 0.33015175 -0.389187885  
## Region 0.33015175 1.00000000 -0.193784256  
## Literacy.... -0.38918788 -0.19378426 1.000000000

pairs(~Net.migration + Infant.mortality + Pop..Density + Phones..per.1000. , data=data7)



pairs(~Net.migration + GDP....per.capita. + Literacy....+ Deathrate+ Service , data=data7, main= "Correlation Matrix for GDP, ")



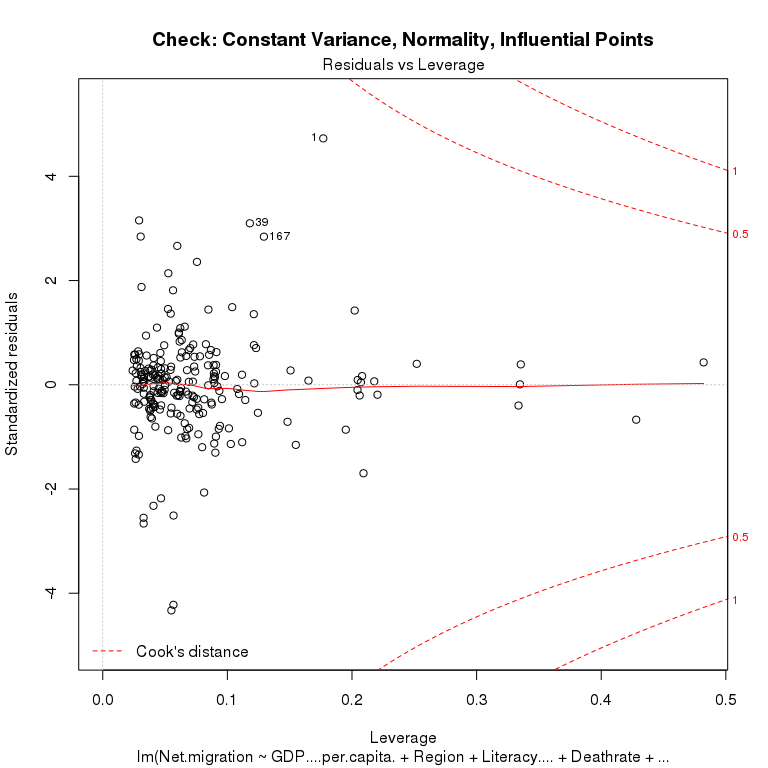
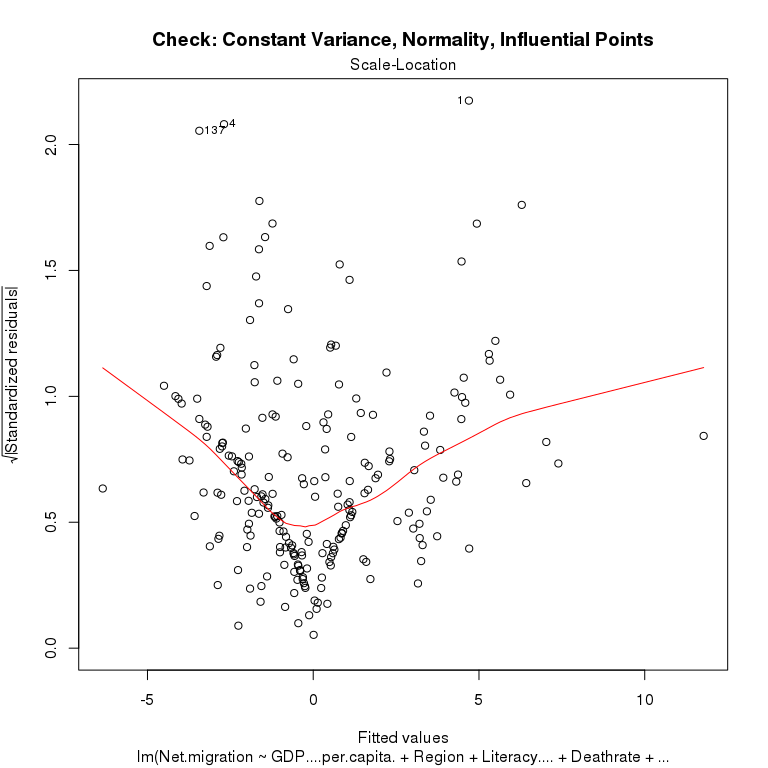
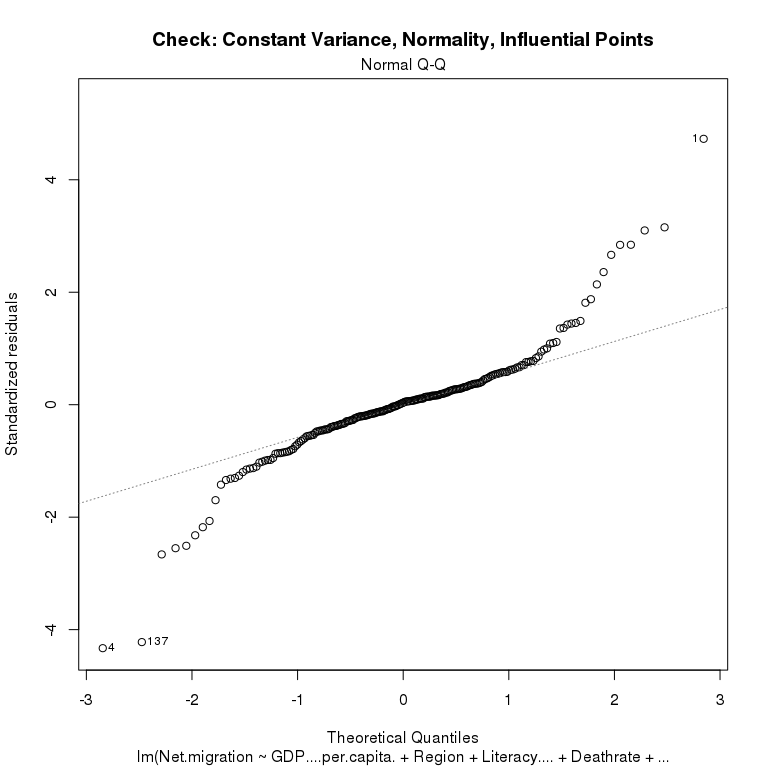
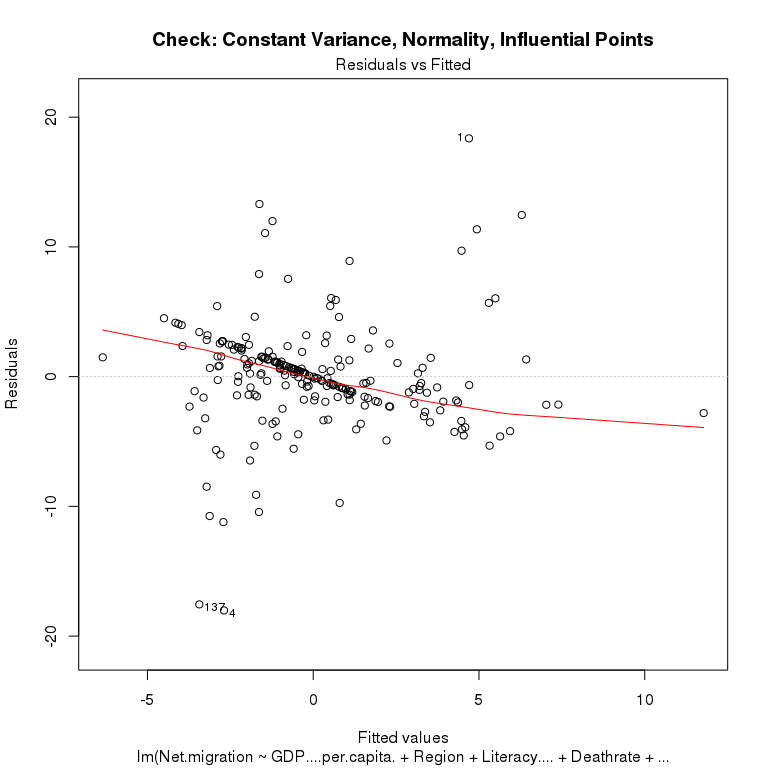
### 

### Modelling

FirstMod = lm(Net.migration~ GDP....per.capita.+ Region+ Literacy.... + Deathrate + Service + Infant.mortality + Pop..Density , data=data7)  
summary(FirstMod)

##   
## Call:  
## lm(formula = Net.migration ~ GDP....per.capita. + Region + Literacy.... +   
## Deathrate + Service + Infant.mortality + Pop..Density, data = data7)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.0191 -1.6122 0.1388 1.4457 18.3635   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -3.134e+00 3.085e+00 -1.016  
## GDP....per.capita. 2.966e-04 5.115e-05 5.798  
## RegionBALTICS -2.247e+00 2.767e+00 -0.812  
## RegionC.W. OF IND. STATES -1.643e+00 1.614e+00 -1.018  
## RegionEASTERN EUROPE -7.088e-01 1.624e+00 -0.437  
## RegionLATIN AMER. & CARIB -1.880e+00 1.123e+00 -1.674  
## RegionNEAR EAST 1.546e+00 1.392e+00 1.111  
## RegionNORTHERN AFRICA -3.172e-01 2.155e+00 -0.147  
## RegionNORTHERN AMERICA -5.280e+00 2.255e+00 -2.341  
## RegionOCEANIA -2.259e+00 1.312e+00 -1.722  
## RegionSUB-SAHARAN AFRICA -1.418e+00 1.237e+00 -1.146  
## RegionWESTERN EUROPE -1.844e+00 1.473e+00 -1.252  
## Literacy.... -1.202e-02 2.676e-02 -0.449  
## Deathrate -3.955e-02 1.015e-01 -0.390  
## Service 1.938e+00 2.499e+00 0.775  
## Infant.mortality 4.975e-02 2.002e-02 2.485  
## Pop..Density 2.292e-04 1.823e-04 1.258  
## Pr(>|t|)   
## (Intercept) 0.3108   
## GDP....per.capita. 2.46e-08 \*\*\*  
## RegionBALTICS 0.4177   
## RegionC.W. OF IND. STATES 0.3099   
## RegionEASTERN EUROPE 0.6629   
## RegionLATIN AMER. & CARIB 0.0956 .   
## RegionNEAR EAST 0.2680   
## RegionNORTHERN AFRICA 0.8831   
## RegionNORTHERN AMERICA 0.0202 \*   
## RegionOCEANIA 0.0865 .   
## RegionSUB-SAHARAN AFRICA 0.2530   
## RegionWESTERN EUROPE 0.2120   
## Literacy.... 0.6538   
## Deathrate 0.6971   
## Service 0.4390   
## Infant.mortality 0.0137 \*   
## Pop..Density 0.2100   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.281 on 208 degrees of freedom  
## Multiple R-squared: 0.2848, Adjusted R-squared: 0.2298   
## F-statistic: 5.177 on 16 and 208 DF, p-value: 4.467e-09

plot(FirstMod,main="Check: Constant Variance, Normality, Influential Points")



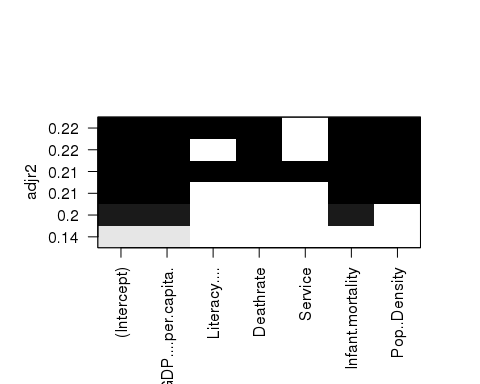
library(leaps)  
  
new= regsubsets(Net.migration ~ GDP....per.capita.+ Literacy.... +   
 Deathrate + Service + Infant.mortality + Pop..Density ,   
 data = data7)  
summary(new)

## Subset selection object  
## Call: regsubsets.formula(Net.migration ~ GDP....per.capita. + Literacy.... +   
## Deathrate + Service + Infant.mortality + Pop..Density, data = data7)  
## 6 Variables (and intercept)  
## Forced in Forced out  
## GDP....per.capita. FALSE FALSE  
## Literacy.... FALSE FALSE  
## Deathrate FALSE FALSE  
## Service FALSE FALSE  
## Infant.mortality FALSE FALSE  
## Pop..Density FALSE FALSE  
## 1 subsets of each size up to 6  
## Selection Algorithm: exhaustive  
## GDP....per.capita. Literacy.... Deathrate Service Infant.mortality  
## 1 ( 1 ) "\*" " " " " " " " "   
## 2 ( 1 ) "\*" " " " " " " "\*"   
## 3 ( 1 ) "\*" " " " " " " "\*"   
## 4 ( 1 ) "\*" " " "\*" " " "\*"   
## 5 ( 1 ) "\*" "\*" "\*" " " "\*"   
## 6 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## Pop..Density  
## 1 ( 1 ) " "   
## 2 ( 1 ) " "   
## 3 ( 1 ) "\*"   
## 4 ( 1 ) "\*"   
## 5 ( 1 ) "\*"   
## 6 ( 1 ) "\*"

summary(new)$adjr2

## [1] 0.1423034 0.2044519 0.2140668 0.2175498 0.2180923 0.2145076

plot(new,scale="adjr2")



Before using a bottom up approach using interactions and transformations a stepwise approach was taken to determine the best model using the coefficients to predict net migration. The best model is shown as containing GDP, literacy, deathrate, infant mortality, and population density.

For the second model, I got rid of the variables Service as seen in the setpwise regression.

SecondMod = lm(Net.migration ~ GDP....per.capita.+ Region+ (Literacy....) + (Deathrate) + Infant.mortality+ (Pop..Density) , data=data7)  
  
summary(SecondMod)

##   
## Call:  
## lm(formula = Net.migration ~ GDP....per.capita. + Region + (Literacy....) +   
## (Deathrate) + Infant.mortality + (Pop..Density), data = data7)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.1328 -1.6679 0.1192 1.4515 18.6657   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -1.943e+00 2.673e+00 -0.727  
## GDP....per.capita. 3.053e-04 4.984e-05 6.127  
## RegionBALTICS -2.077e+00 2.755e+00 -0.754  
## RegionC.W. OF IND. STATES -1.623e+00 1.612e+00 -1.007  
## RegionEASTERN EUROPE -6.453e-01 1.620e+00 -0.398  
## RegionLATIN AMER. & CARIB -1.699e+00 1.098e+00 -1.548  
## RegionNEAR EAST 1.477e+00 1.388e+00 1.064  
## RegionNORTHERN AFRICA -4.964e-01 2.141e+00 -0.232  
## RegionNORTHERN AMERICA -5.192e+00 2.250e+00 -2.307  
## RegionOCEANIA -2.165e+00 1.305e+00 -1.659  
## RegionSUB-SAHARAN AFRICA -1.312e+00 1.228e+00 -1.068  
## RegionWESTERN EUROPE -1.796e+00 1.470e+00 -1.222  
## Literacy.... -1.306e-02 2.670e-02 -0.489  
## Deathrate -4.280e-02 1.013e-01 -0.422  
## Infant.mortality 4.570e-02 1.931e-02 2.367  
## Pop..Density 2.476e-04 1.805e-04 1.372  
## Pr(>|t|)   
## (Intercept) 0.4680   
## GDP....per.capita. 4.4e-09 \*\*\*  
## RegionBALTICS 0.4517   
## RegionC.W. OF IND. STATES 0.3152   
## RegionEASTERN EUROPE 0.6908   
## RegionLATIN AMER. & CARIB 0.1230   
## RegionNEAR EAST 0.2887   
## RegionNORTHERN AFRICA 0.8169   
## RegionNORTHERN AMERICA 0.0220 \*   
## RegionOCEANIA 0.0985 .   
## RegionSUB-SAHARAN AFRICA 0.2867   
## RegionWESTERN EUROPE 0.2231   
## Literacy.... 0.6253   
## Deathrate 0.6731   
## Infant.mortality 0.0188 \*   
## Pop..Density 0.1717   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.277 on 209 degrees of freedom  
## Multiple R-squared: 0.2827, Adjusted R-squared: 0.2313   
## F-statistic: 5.492 on 15 and 209 DF, p-value: 2.372e-09

Got rid of population density and literacy as they are not significant predictor for net migration according to the 2nd model.

ThirdMod = lm(Net.migration ~ GDP....per.capita.+ Region+ Deathrate+ Infant.mortality , data=data7)  
summary(ThirdMod)

##   
## Call:  
## lm(formula = Net.migration ~ GDP....per.capita. + Region + Deathrate +   
## Infant.mortality, data = data7)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.168 -1.496 0.076 1.609 18.651   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -2.783e+00 1.185e+00 -2.349  
## GDP....per.capita. 3.086e-04 4.925e-05 6.267  
## RegionBALTICS -2.524e+00 2.739e+00 -0.921  
## RegionC.W. OF IND. STATES -2.153e+00 1.515e+00 -1.421  
## RegionEASTERN EUROPE -1.018e+00 1.598e+00 -0.637  
## RegionLATIN AMER. & CARIB -2.063e+00 1.069e+00 -1.930  
## RegionNEAR EAST 1.298e+00 1.368e+00 0.949  
## RegionNORTHERN AFRICA -6.105e-01 2.096e+00 -0.291  
## RegionNORTHERN AMERICA -5.595e+00 2.228e+00 -2.512  
## RegionOCEANIA -2.522e+00 1.282e+00 -1.967  
## RegionSUB-SAHARAN AFRICA -1.443e+00 1.212e+00 -1.191  
## RegionWESTERN EUROPE -2.022e+00 1.452e+00 -1.393  
## Deathrate -4.823e-02 9.974e-02 -0.484  
## Infant.mortality 4.880e-02 1.608e-02 3.034  
## Pr(>|t|)   
## (Intercept) 0.01974 \*   
## GDP....per.capita. 2.04e-09 \*\*\*  
## RegionBALTICS 0.35794   
## RegionC.W. OF IND. STATES 0.15692   
## RegionEASTERN EUROPE 0.52470   
## RegionLATIN AMER. & CARIB 0.05500 .   
## RegionNEAR EAST 0.34394   
## RegionNORTHERN AFRICA 0.77115   
## RegionNORTHERN AMERICA 0.01277 \*   
## RegionOCEANIA 0.05049 .   
## RegionSUB-SAHARAN AFRICA 0.23485   
## RegionWESTERN EUROPE 0.16520   
## Deathrate 0.62920   
## Infant.mortality 0.00271 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.279 on 211 degrees of freedom  
## Multiple R-squared: 0.2754, Adjusted R-squared: 0.2307   
## F-statistic: 6.168 on 13 and 211 DF, p-value: 9.212e-10

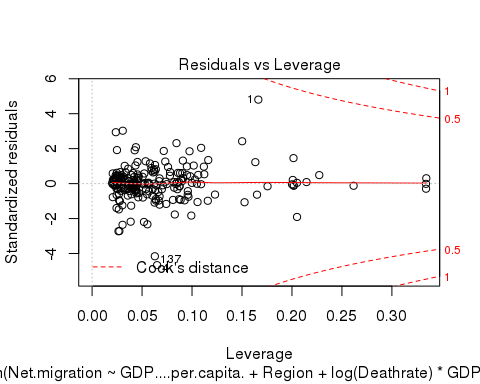
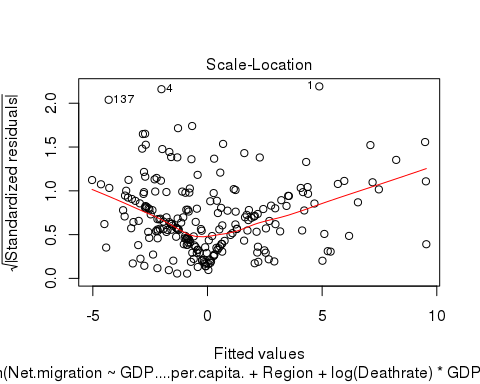
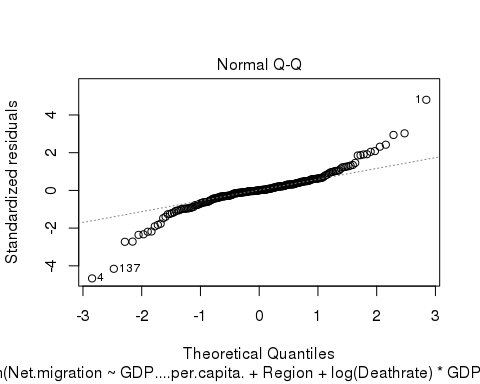
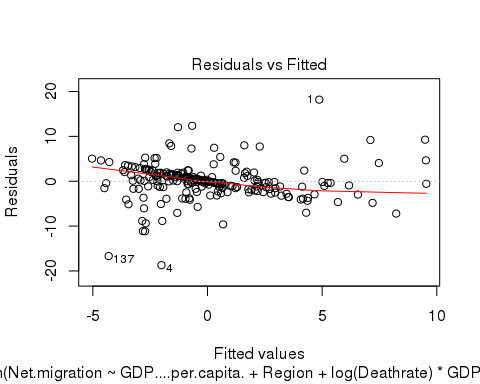
As we can see in the previous model, I decided to interact GDP with death rate and added log to death rate and transformed it to better the model.

FourthMod= lm(Net.migration ~ GDP....per.capita.+ Region+ log(Deathrate)\*GDP....per.capita. + Infant.mortality , data=data7)  
summary(FourthMod)

##   
## Call:  
## lm(formula = Net.migration ~ GDP....per.capita. + Region + log(Deathrate) \*   
## GDP....per.capita. + Infant.mortality, data = data7)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.7023 -1.4861 0.0125 1.6542 18.1844   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -7.2119057 2.7332656 -2.639  
## GDP....per.capita. 0.0010825 0.0002188 4.947  
## RegionBALTICS -1.2048517 2.6906818 -0.448  
## RegionC.W. OF IND. STATES -1.8925636 1.4870212 -1.273  
## RegionEASTERN EUROPE -0.1553955 1.5828601 -0.098  
## RegionLATIN AMER. & CARIB -1.9029096 1.0357040 -1.837  
## RegionNEAR EAST 0.8025584 1.3624491 0.589  
## RegionNORTHERN AFRICA -0.2129674 2.0615880 -0.103  
## RegionNORTHERN AMERICA -3.8067944 2.2075662 -1.724  
## RegionOCEANIA -2.5250862 1.2445566 -2.029  
## RegionSUB-SAHARAN AFRICA -1.7681694 1.1770326 -1.502  
## RegionWESTERN EUROPE 1.0549540 1.6123519 0.654  
## log(Deathrate) 2.3028252 1.5014248 1.534  
## Infant.mortality 0.0323361 0.0169673 1.906  
## GDP....per.capita.:log(Deathrate) -0.0004177 0.0001162 -3.593  
## Pr(>|t|)   
## (Intercept) 0.008949 \*\*   
## GDP....per.capita. 1.55e-06 \*\*\*  
## RegionBALTICS 0.654769   
## RegionC.W. OF IND. STATES 0.204524   
## RegionEASTERN EUROPE 0.921888   
## RegionLATIN AMER. & CARIB 0.067577 .   
## RegionNEAR EAST 0.556457   
## RegionNORTHERN AFRICA 0.917821   
## RegionNORTHERN AMERICA 0.086102 .   
## RegionOCEANIA 0.043730 \*   
## RegionSUB-SAHARAN AFRICA 0.134541   
## RegionWESTERN EUROPE 0.513638   
## log(Deathrate) 0.126594   
## Infant.mortality 0.058044 .   
## GDP....per.capita.:log(Deathrate) 0.000408 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.145 on 210 degrees of freedom  
## Multiple R-squared: 0.3232, Adjusted R-squared: 0.2781   
## F-statistic: 7.162 on 14 and 210 DF, p-value: 4.623e-12

Normality, constant variance and no influential points are a good sign.

plot(FourthMod)



I calculated the confidence intervals from the final model.

confint(FourthMod)

## 2.5 % 97.5 %  
## (Intercept) -1.260006e+01 -1.8237513724  
## GDP....per.capita. 6.511019e-04 0.0015138862  
## RegionBALTICS -6.509059e+00 4.0993559032  
## RegionC.W. OF IND. STATES -4.823965e+00 1.0388380737  
## RegionEASTERN EUROPE -3.275727e+00 2.9649359534  
## RegionLATIN AMER. & CARIB -3.944619e+00 0.1387993693  
## RegionNEAR EAST -1.883271e+00 3.4883880922  
## RegionNORTHERN AFRICA -4.277027e+00 3.8510920497  
## RegionNORTHERN AMERICA -8.158624e+00 0.5450355569  
## RegionOCEANIA -4.978511e+00 -0.0716608972  
## RegionSUB-SAHARAN AFRICA -4.088483e+00 0.5521441166  
## RegionWESTERN EUROPE -2.123515e+00 4.2334232530  
## log(Deathrate) -6.569708e-01 5.2626212275  
## Infant.mortality -1.111996e-03 0.0657841309  
## GDP....per.capita.:log(Deathrate) -6.468289e-04 -0.0001884964

Then I conducted the anova F test to see if a significant portion of variability in net migration is explained by the model. For most of the values F>1 and the p-value is very small indicating that the model is useful.

anova(FourthMod)

## Analysis of Variance Table  
##   
## Response: Net.migration  
## Df Sum Sq Mean Sq F value Pr(>F)   
## GDP....per.capita. 1 779.0 779.00 45.3409 1.556e-10 \*\*\*  
## Region 10 485.9 48.59 2.8284 0.0025737 \*\*   
## log(Deathrate) 1 5.7 5.72 0.3328 0.5646439   
## Infant.mortality 1 230.3 230.35 13.4071 0.0003172 \*\*\*  
## GDP....per.capita.:log(Deathrate) 1 221.8 221.78 12.9082 0.0004075 \*\*\*  
## Residuals 210 3608.0 17.18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

This is the five variable model given by the stepwise regression function above. The anova test clearly shows the added variable in the first model does not better the model.

anova(ThirdMod,FirstMod, test='F')

## Analysis of Variance Table  
##   
## Model 1: Net.migration ~ GDP....per.capita. + Region + Deathrate + Infant.mortality  
## Model 2: Net.migration ~ GDP....per.capita. + Region + Literacy.... +   
## Deathrate + Service + Infant.mortality + Pop..Density  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 211 3862.9   
## 2 208 3812.6 3 50.312 0.915 0.4346