MThambeliyagodage\_Data605\_W12\_Assign12

Matheesha Thambeliyagodage

April 28, 2018

Table of Contents

## Warning: package 'ggplot2' was built under R version 3.4.3

## Warning: package 'dplyr' was built under R version 3.4.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# Data 605 Assignment 12

### Task

Using the cars dataset in R, build a linear model for stopping distance as a function of speed and replicate the analysis of your textbook chapter 3 (visualization, quality evaluation of the model, and residual analysis).

### Data Import

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | LifeExp | PropMD | PersExp | GovtExp | TotExp |
| Afghanistan | 42 | 0.0002288 | 20 | 92 | 112 |
| Albania | 71 | 0.0011431 | 169 | 3128 | 3297 |
| Algeria | 71 | 0.0010605 | 108 | 5184 | 5292 |
| Andorra | 82 | 0.0032973 | 2589 | 169725 | 172314 |
| Angola | 41 | 0.0000704 | 36 | 1620 | 1656 |
| Antigua and Barbuda | 73 | 0.0001429 | 503 | 12543 | 13046 |

Data is real-world World Health Organization data from 2008. It includes 190 observations for 10 variables. Data dictionary:

* Country: name of the country
* LifeExp: average life expectancy for the country in years
* InfantSurvival: proportion of those surviving to one year or more
* Under5Survival: proportion of those surviving to five years or more
* TBFree: proportion of the population without TB
* PropMD: proportion of the population who are MDs
* PropRN: proportion of the population who are RNs
* PersExp: mean personal expenditures on healthcare in US dollars at average exchange rate
* GovtExp: mean government expenditures per capita on healthcare, US dollars at average exchange rate
* TotExp: sum of personal and government expenditures

### Data Exploration

## Country LifeExp InfantSurvival   
## Afghanistan : 1 Min. :40.00 Min. :0.8350   
## Albania : 1 1st Qu.:61.25 1st Qu.:0.9433   
## Algeria : 1 Median :70.00 Median :0.9785   
## Andorra : 1 Mean :67.38 Mean :0.9624   
## Angola : 1 3rd Qu.:75.00 3rd Qu.:0.9910   
## Antigua and Barbuda: 1 Max. :83.00 Max. :0.9980   
## (Other) :184   
## Under5Survival TBFree PropMD PropRN   
## Min. :0.7310 Min. :0.9870 Min. :0.0000196 Min. :0.0000883   
## 1st Qu.:0.9253 1st Qu.:0.9969 1st Qu.:0.0002444 1st Qu.:0.0008455   
## Median :0.9745 Median :0.9992 Median :0.0010474 Median :0.0027584   
## Mean :0.9459 Mean :0.9980 Mean :0.0017954 Mean :0.0041336   
## 3rd Qu.:0.9900 3rd Qu.:0.9998 3rd Qu.:0.0024584 3rd Qu.:0.0057164   
## Max. :0.9970 Max. :1.0000 Max. :0.0351290 Max. :0.0708387   
##   
## PersExp GovtExp TotExp   
## Min. : 3.00 Min. : 10.0 Min. : 13   
## 1st Qu.: 36.25 1st Qu.: 559.5 1st Qu.: 584   
## Median : 199.50 Median : 5385.0 Median : 5541   
## Mean : 742.00 Mean : 40953.5 Mean : 41696   
## 3rd Qu.: 515.25 3rd Qu.: 25680.2 3rd Qu.: 26331   
## Max. :6350.00 Max. :476420.0 Max. :482750   
##

Looking at the range of personal and government expenditures (13 to 482,750), I thought it was interesting to see top and bottom countries.

Bottom 5 Countries by Total Expenditures

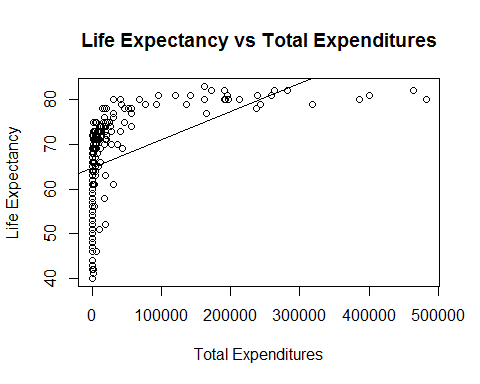
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | LifeExp | PropMD | PersExp | GovtExp | TotExp |
| Burundi | 49 | 0.0000245 | 3 | 10 | 13 |
| Ethiopia | 56 | 0.0000239 | 6 | 64 | 70 |
| Democratic Republic of the Congo | 47 | 0.0000961 | 5 | 66 | 71 |
| Nepal | 62 | 0.0001948 | 16 | 64 | 80 |
| Bangladesh | 63 | 0.0002749 | 12 | 75 | 87 |

Top 5 Countries by Total Expenditures

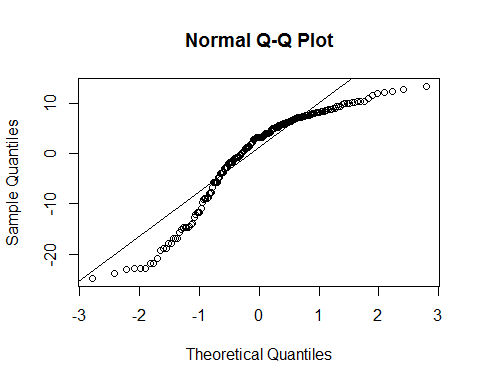
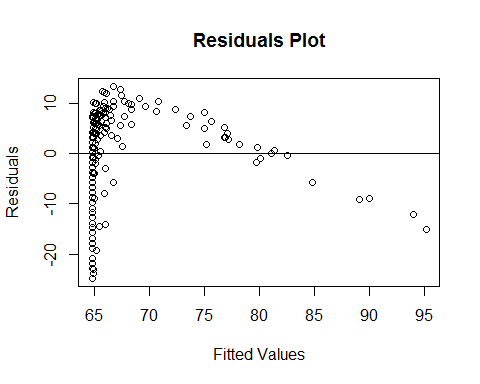
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | LifeExp | PropMD | PersExp | GovtExp | TotExp |
| Denmark | 79 | 0.0035519 | 4350 | 314588 | 318938 |
| Norway | 80 | 0.0037531 | 5910 | 380380 | 386290 |
| Iceland | 81 | 0.0037584 | 5154 | 395622 | 400776 |
| Monaco | 82 | 0.0056364 | 6128 | 458700 | 464828 |
| Luxembourg | 80 | 0.0027223 | 6330 | 476420 | 482750 |

### Question 1

Let us build a linear regression model for predicting life expectancy by total expenditures. Below scatterplot shows the relationship along with the linear regression line.



##   
## Call:  
## lm(formula = LifeExp ~ TotExp, data = who)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.764 -4.778 3.154 7.116 13.292   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.753374534 0.753536611 85.933 < 2e-16 \*\*\*  
## TotExp 0.000062970 0.000007795 8.079 0.0000000000000771 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.371 on 188 degrees of freedom  
## Multiple R-squared: 0.2577, Adjusted R-squared: 0.2537   
## F-statistic: 65.26 on 1 and 188 DF, p-value: 0.00000000000007714



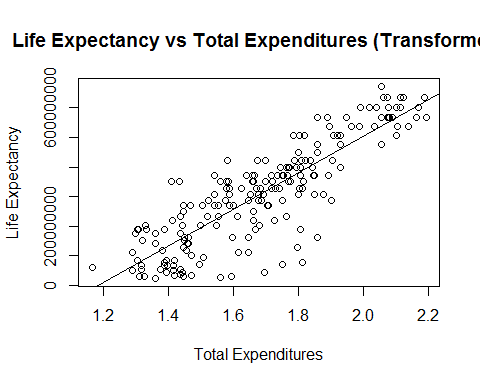
#### Results

**Residual standard error** is 9.371 and **F-statistic** is 65.26. Considering that average life expectancy is 67.38, the SE is not terrible and F-statistics is high. However, is only 0.2577 (so the model explains only 25.77% of variability). **P-value** is nearly 0, so the relationship is not due to random variation.

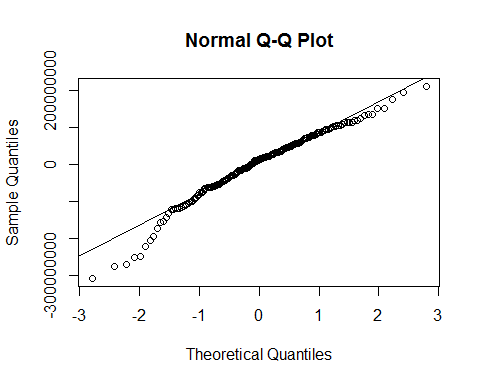
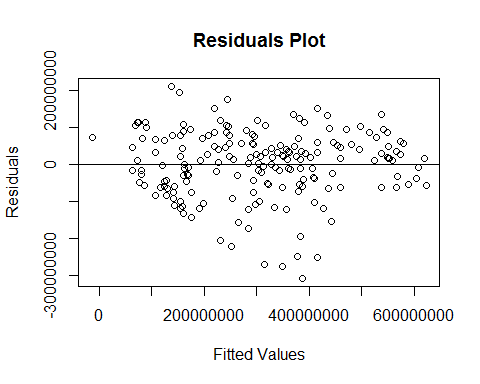
Looking at residuals plots it is clear that there is no constant variability and that residuals are not normally distributed. This is **not a good model** to describe the relationship. It is clear from the scatterplot that the relationship is not linear.

### Question 2

Let us transform variables and re-run the simple linear regression model - and .



##   
## Call:  
## lm(formula = LifeExp4.6 ~ TotExp0.06)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -308616089 -53978977 13697187 59139231 211951764   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -736527910 46817945 -15.73 <2e-16 \*\*\*  
## TotExp0.06 620060216 27518940 22.53 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 90490000 on 188 degrees of freedom  
## Multiple R-squared: 0.7298, Adjusted R-squared: 0.7283   
## F-statistic: 507.7 on 1 and 188 DF, p-value: < 2.2e-16



#### Results

**Residual standard error** is 90,490,000 and **F-statistic** is 507.7. The F-statistic is good, but the SE is a bit high considering that it corresponds to 53.67 years if we reverse the transformation). is 0.7298, which is considerably better than in the first model (the model explains 72.98% of variability). **P-value** is again nearly 0, so the relationship is not due to random variation.

Looking at residuals plots, variability is fairly constant with a few outliers and distribution of residuals is nearly normal with some deviation at the tails. This is **a fairly good model** to describe the relationship and it is significantly better than the first model. The linear relationship between transformed variables is clear from the scatterplot.

### Question 3

## fit lwr upr  
## 1 63.31153 35.93545 73.00793  
## 2 86.50645 81.80643 90.43414

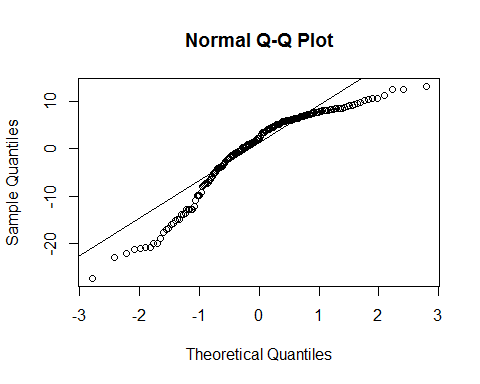
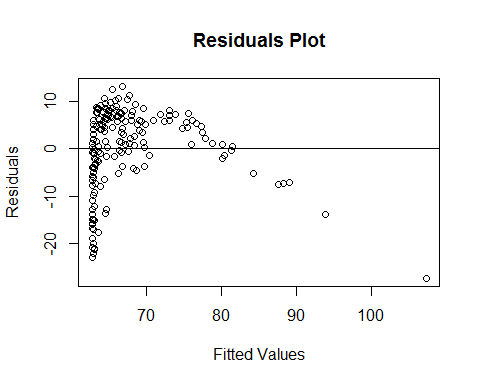
Based on the second model, prediction for total expeditures of $860.705 () is 63.31 years with a 95% confidence interval between 35.94 and 73.01.

Prediction for total expeditures of $4,288,777 () is 86.51 years with a 95% confidence interval between 81.81 and 90.43.

### Question 4

Let us build the following model: .

##   
## Call:  
## lm(formula = LifeExp ~ PropMD + TotExp + TotExp:PropMD, data = who)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.320 -4.132 2.098 6.540 13.074   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 62.772703255 0.795605238 78.899 < 2e-16 \*\*\*  
## PropMD 1497.493952519 278.816879652 5.371 0.0000002320602774 \*\*\*  
## TotExp 0.000072333 0.000008982 8.053 0.0000000000000939 \*\*\*  
## PropMD:TotExp -0.006025686 0.001472357 -4.093 0.0000635273294941 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.765 on 186 degrees of freedom  
## Multiple R-squared: 0.3574, Adjusted R-squared: 0.3471   
## F-statistic: 34.49 on 3 and 186 DF, p-value: < 2.2e-16



#### Results

**Residual standard error** is 8.765 and **F-statistic** is 34.49. Considering that average life expectancy is 67.38, the SE is not terrible and F-statistics is fairly high (but lower than in the first model). is only 0.3574, so the model explains only 35.74% of variability, which is not high. **P-value** is nearly 0, so the relationship is not due to random variation.

Looking at residuals plots it is clear that there is no constant variability and that residuals are not normally distributed. This is **not a good model** to describe the relationship. Kind of similar to the first model.

### Question 5

Consider forecast based on the last model with and .

## fit lwr upr  
## 1 107.696 84.24791 131.1441

The prediction is 107.70 years with 95% confidence interval between 84.25 and 131.14. The prediction is completely unrealistic. We do have individuals livings into their 100s; however, consider that the total expenditures of $14 is just a tad higher than the minimum value of $13 for Burundi and the life expectancy there is 49 years. The highest life expectancy in the data is 83 years. There is nothing in our data to support this prediction and it goes against common sense. As stated under question 4, this is not a good model.