HW2 STA521 Fall18

[Freyafu, zf43, freyafu326]

Due September 19, 2018

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

This exercise involves the UN data set from `alr3` package. Install `alr3` and the `car` packages and load the data to answer the following questions adding your code in the code chunks. Please add appropriate code to the chunks to suppress messages and warnings as needed once you are sure the code is working properly and remove instructions if no longer needed. Figures should have informative captions. Please switch the output to pdf for your final version to upload to Sakai. **Remove these instructions for final submission**

Exploratory Data Analysis

0. Preliminary read in the data. After testing, modify the code chunk so that output, messages and warnings are suppressed. *Exclude text from final*

```r
suppressWarnings(library(car))
```

```
## Loading required package: carData
```

```r
library(alr3)
data(UN3, package="alr3")
help(UN3)
library(car)
library(knitr)
library(GGally)
```

```
## Loading required package: ggplot2
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:GGally':
##
##     nasa
```

```
## The following object is masked from 'package:car':
##
##     recode
```

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

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

1. Create a summary of the data. How many variables have missing data? Which are quantitative and which are qualtitative?

```
summary(UN3)
```

```
##     ModernC          Change           PPgdp           Frate      
##  Min.   : 1.00   Min.   :-1.100   Min.   :   90   Min.   : 2.00  
##  1st Qu.:19.00   1st Qu.: 0.580   1st Qu.:  479   1st Qu.:39.50  
##  Median :40.50   Median : 1.400   Median : 2046   Median :49.00  
##  Mean   :38.72   Mean   : 1.418   Mean   : 6527   Mean   :48.31  
##  3rd Qu.:55.00   3rd Qu.: 2.270   3rd Qu.: 8461   3rd Qu.:58.00  
##  Max.   :83.00   Max.   : 4.170   Max.   :44579   Max.   :91.00  
##  NA's   :58      NA's   :1        NA's   :9       NA's   :43     
##       Pop               Fertility         Purban      
##  Min.   :      2.3   Min.   :1.000   Min.   :  6.00  
##  1st Qu.:    767.2   1st Qu.:1.897   1st Qu.: 36.25  
##  Median :   5469.5   Median :2.700   Median : 57.00  
##  Mean   :  30281.9   Mean   :3.214   Mean   : 56.20  
##  3rd Qu.:  18913.5   3rd Qu.:4.395   3rd Qu.: 75.00  
##  Max.   :1304196.0   Max.   :8.000   Max.   :100.00  
##  NA's   :2           NA's   :10
```

```
sapply(UN3, function(x) sum(is.na(x))) #Missing data for each variable
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
##        58         1         9        43         2        10         0
```

```
sapply(UN3, function(x) class(x))
```

```
##   ModernC    Change     PPgdp     Frate       Pop Fertility    Purban 
## "integer" "numeric" "integer" "integer" "numeric" "numeric" "integer"
```

Comments: All of the variables are quantitative.
ModernC: Percent of unmarried women using a modern method of contraception.
Change; Annual population growth rate, percent.
PPgdp: Per capita 2001 GDP, in US $.
Frate: Percent of females over age 15 economically active.
Pop: Population, thousands.
Fertility:Expected number of live births per female, 2000

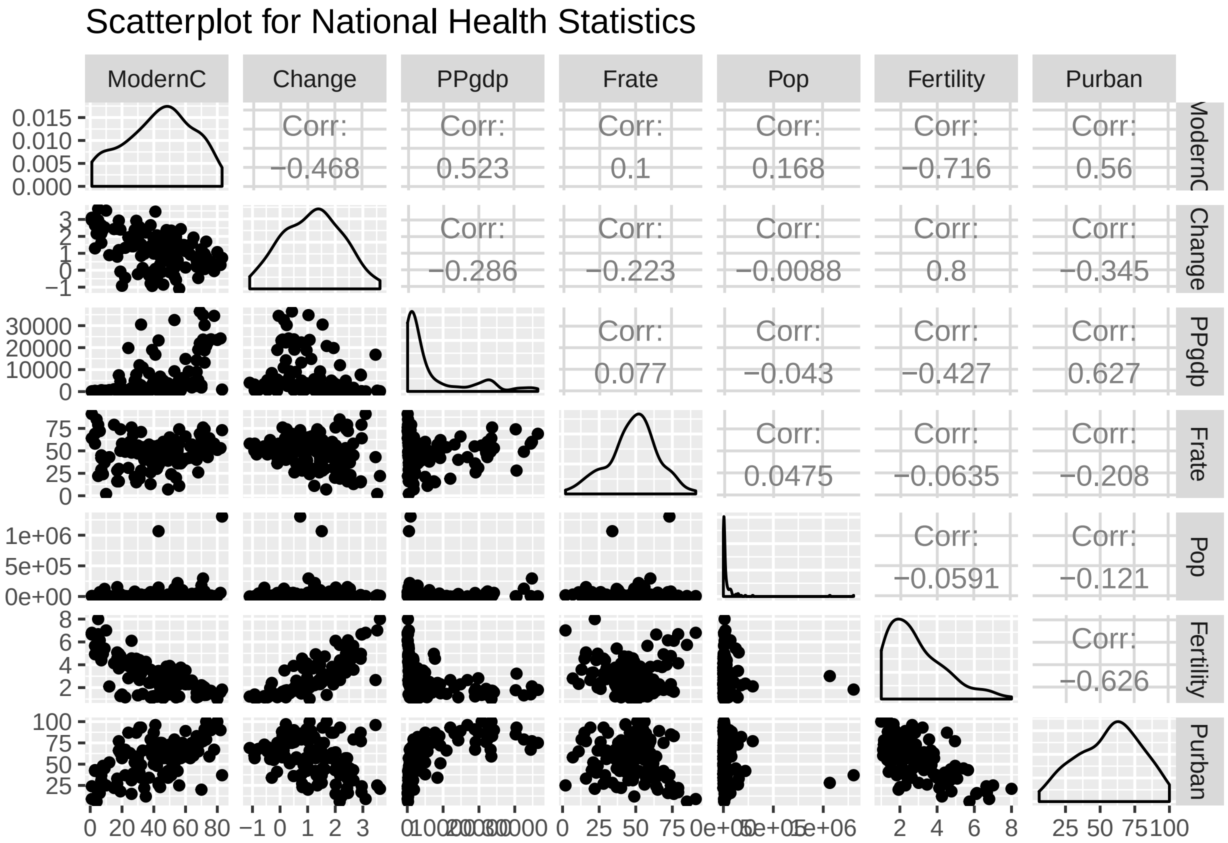
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```r
A<-matrix(NA,7,2) #create an empty matrix to store values
for (i in 1:7){
  A[i,1]=mean(na.omit(UN3[,i])) #assign mean to the first column vector
  A[i,2]=sd(na.omit(UN3[,i])) #assign sd to the second column vector
}
b<-colnames(UN3)
c<-cbind.data.frame(b,round(A,3))
colnames(c)<-c("Variables","Mean","Std")
kable(c,format = "markdown")
```

Variables	Mean	Std
ModernC	38.717	22.637
Change	1.418	1.133
PPgdp	6527.388	9325.189
Frate	48.305	16.532
Pop	30281.871	120676.694
Fertility	3.214	1.707
Purban	56.200	24.110

3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships or transformations that appear to be needed based on your graphical EDA?

```
UN3_0<-na.omit(UN3) #delate any rows that has missing values
ggpairs(UN3_0,columns = 1:7,title = "Scatterplot for National Health Statistics")
```



```
#delete data that has 3 or more missing values in a row
delete.na <- function(DF, n=2) {
  DF[rowSums(is.na(DF)) <= n,]
}
UN3_2<-delete.na(UN3)
#use the data set UN3_0 afterwards
```

Comments: In the scatterplot of ModernC compared with PPgdp and fertility, there seems to be a non-linear relationship. In the scatterplot of ModernC and Pop, there are 2 potential outliners as they are really deviated from the rest of the data points.

## Model Fitting

4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and comment on results regarding assumptions. How many observations are used in your model fitting?

```
model_1<-lm(ModernC~.,data = UN3_2)
model_2<-lm(ModernC~.,data=UN3_0)
model_1$df.residual
```

```
## [1] 118
```

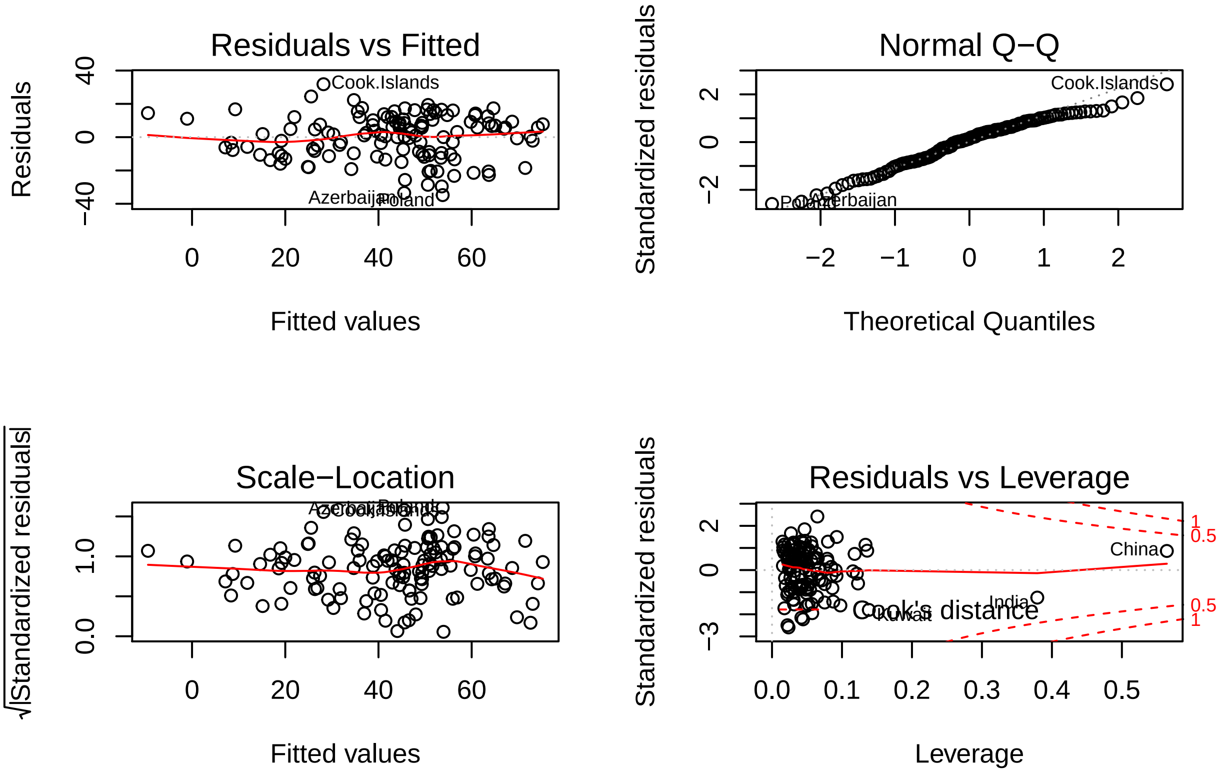
```
model_2$df.residual # the data used in the lm model is the same for both datasets. Only rows with no mi
```

```
## [1] 118
```

```
summary(model_2)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3_0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.781  -9.698   1.858   9.327  31.791
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.529e+01  9.467e+00   5.841 4.69e-08 ***
## Change       5.268e+00  2.088e+00   2.524  0.01294 *
## PPgdp        5.301e-04  1.770e-04   2.995  0.00334 **
## Frate        1.232e-01  8.060e-02   1.529  0.12901
## Pop          1.899e-05  8.213e-06   2.312  0.02250 *
## Fertility   -1.100e+01  1.752e+00  -6.276 5.96e-09 ***
## Purban       5.408e-02  9.285e-02   0.582  0.56134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.58 on 118 degrees of freedom
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5989
## F-statistic: 31.85 on 6 and 118 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model_2)
```

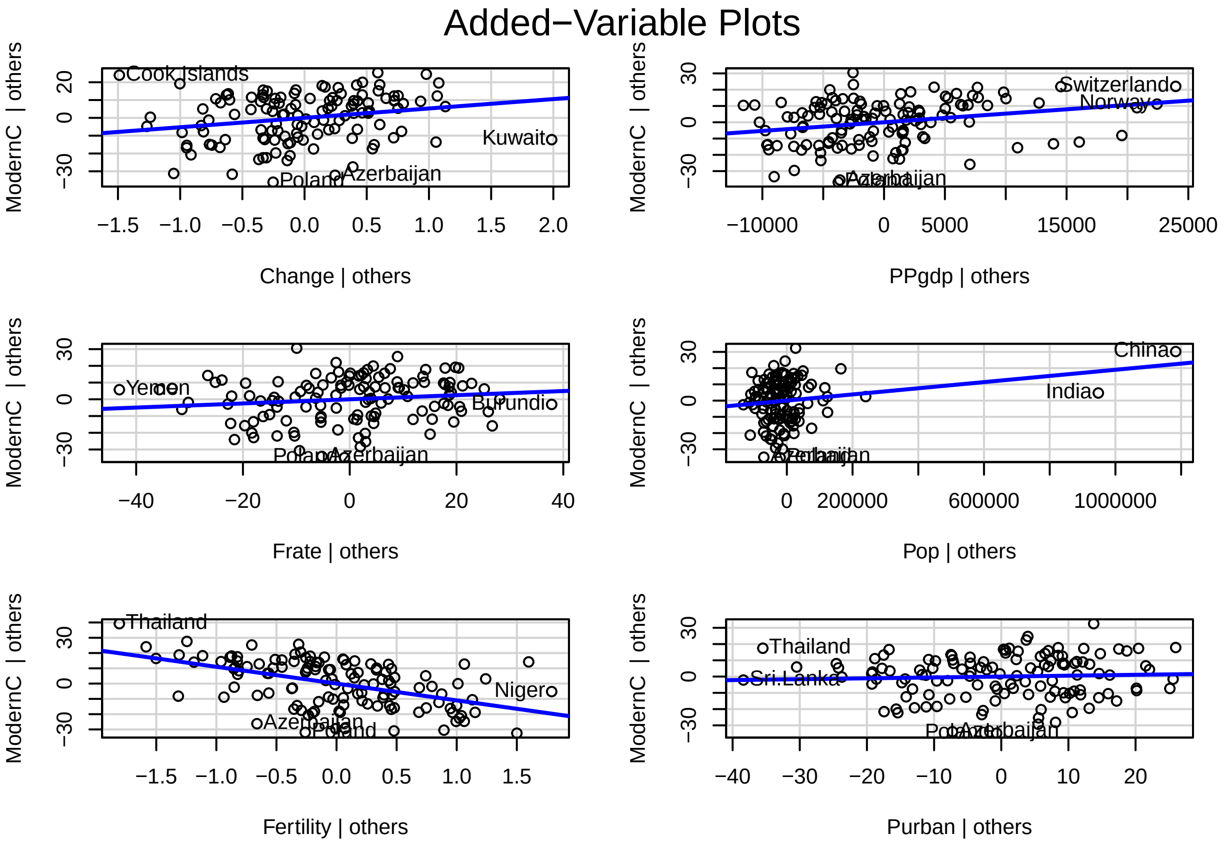


Comments: For the residual versus fitted value plot, we want to see that the residuals is randomly distributed among positive and negative values. The plot looks fine, so the assumption of $E(\epsilon) = 0$ is met
For the QQ plot, we want to check for nornality and see if the line is straight. The residuals are smaller than expected under normality. So it is lighter tailed, and the assumption of normality of residuals is not met. For the scale location plot, we want to see the spread of the residuals is constant over the range of fitted values. However, it looks like the variance is getting larger as the fitted values become larger. Thus, the assumption of constant variance of residuals are not met. For the residual versus leverage plot, we can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines. And we see that there even though points like China and india are high leverage, they have small residuals. So maybe there are no potential outliners and influential points.

There are 125 observations used in our model fitting

5. Examine added variable plots `car::avPlot` or `car::avPlots` for your model above. Are there any plots that suggest that transformations are needed for any of the terms in the model? Describe. Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?

```r
avPlots(model_2, terms=~.)
```



Comments: It looks like the variable Pop need transformation as the plot for it is least linear compared with others. The ppdgp may be another variable that needs transformation because it looks like the data points are clustered between -10000 to 5000.

Fot influential terms, it looks like China and India in the Pop plot maybe influential as their localities are around 1,000,000. It also looks like Cooks Island and Kuwaito maybe influential as their localities in the change plot are -1.5, 2.0 each.

6. Using the Box-Tidwell `car::boxTidwell` or graphical methods find appropriate transformations of the predictor variables to be used as predictors in the linear model. If any predictors are negative, you may need to transform so that they are non-negative. Describe your method and the resulting transformations.

```
#car::boxTidwell(ModernC~PPgdp,data=UN3_0)
#car::boxTidwell(ModernC~PPgdp,other.x=~Change+Frate+Fertility+Purban+Pop,data=UN3_0)
car::boxTidwell(ModernC~Pop+PPgdp,other.x=~Change+Frate+Fertility+Purban,data=UN3_0)
```

```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop         0.40749             -0.7874   0.4310
## PPgdp      -0.12921             -1.1410   0.2539
##
## iterations =  4
```

```
range(UN3_0$Change)# the variable change has negative values
```

```
## [1] -1.10  3.62
```

```
UN3_2<-UN3_0
UN3_2$Change<-UN3_2$Change+2 # add a constant to change to make it unnegative
powerTransform(as.matrix(UN3_2)~.,family = "bcnPower",data = UN3_0)
```

```
## Estimated transformation power, lambda
## [1] 0.9999779 0.9992009 0.9999977 0.9999845 0.3251017 0.9993778 0.9999833
##
## Estimated location, gamma
## [1] 1.000000e-01 1.000000e-01 3.013179e+00 1.000000e-01 1.304196e+06
## [6] 1.000000e-01 1.000000e-01
```

Comments: From the avplots above, we think Pop and PPgdp are potential variables that may need transformation.

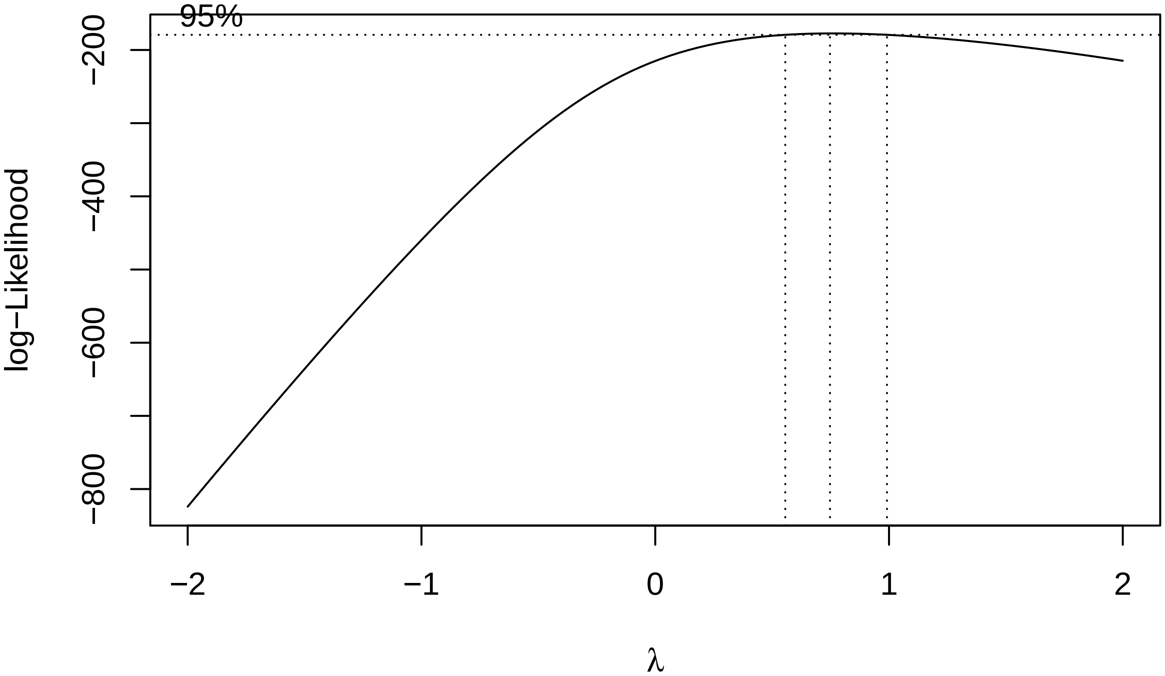
The method we tried first is the boxTidewell power transform. We put variables like change,frate, Fertility and Purba in the other.x as the variables that we do not want to transform, and variables like Pop and PPgdp as the variables that we want to transform. For the variables that has MLE close to 1, we do not need transformation. The λ value variable Pop for is 0.407, so we can round it up to 0.5 and take the square root of Pop as transformation. For PPgdp, the λ value for variable PPgdp is close to -0.129, so we can round it up to 0 and take the log of PPgdp as transformation. However, the p-value for the boxTidwell are not significant.

We want to further justify what variables we want to transform. The second method we used is function powerTransform. For the variables that has MLE close to 1, we do not need transformation. The λ value for variable Pop is 0.325, so we can round it up to 0.5 and take the square root of Pop as transformation.

We decide only to transform two variables, PPgdp and Pop.

7. Given the selected transformations of the predictors, select a transformation of the response using `MASS::boxcox` or `car::boxCox` and justify.

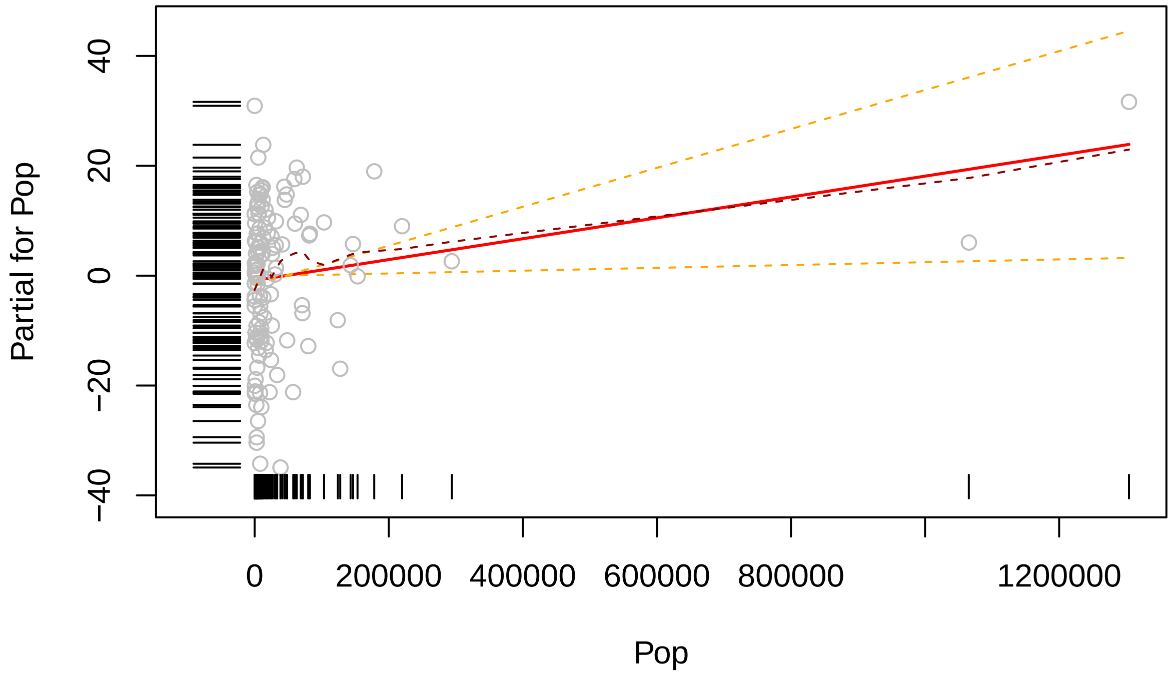
```
pop_sqrt<-sqrt(UN3_0$Pop)
PPgdp_log<-log(UN3_0$PPgdp)
MASS::boxcox(ModernC~pop_sqrt+PPgdp_log+Change+Frate+Fertility+Purban, data = UN3_0,lambda = seq(-2, 2,
```



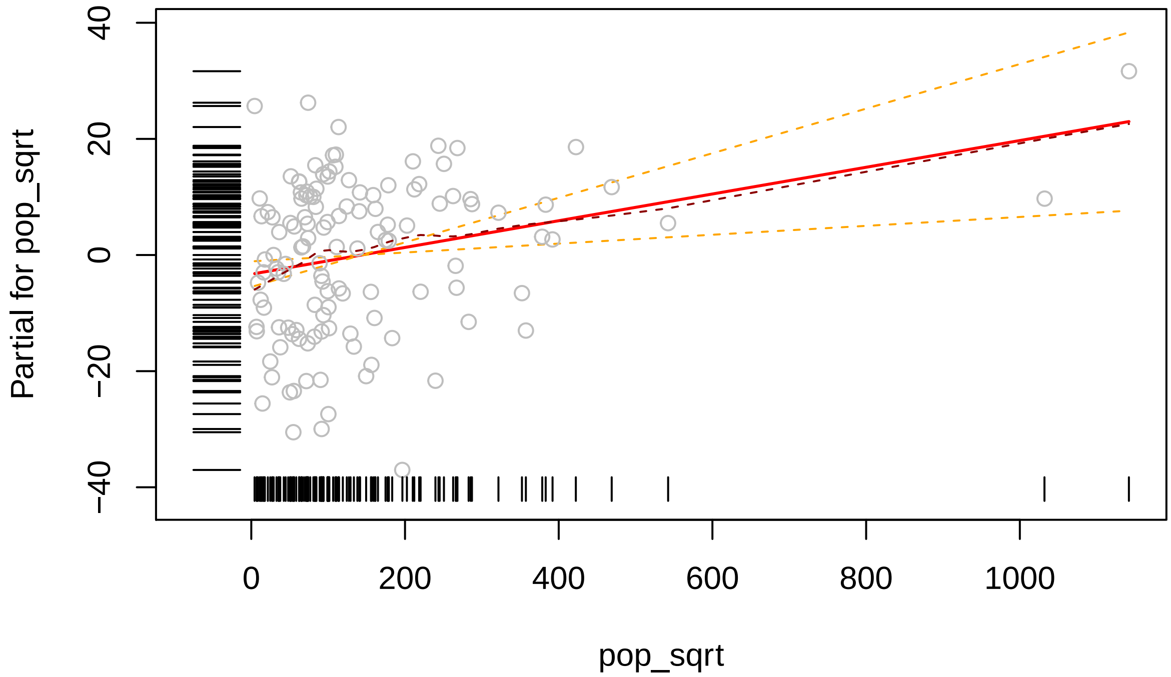
Comments: It looks like the MLE for lamda for the predictor is around 0.8, which can round up to 1. So we do not need to transform the predictor.

8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

```
model_3<-lm(ModernC~pop_sqrt+PPgdp_log+Change+Frate+Fertility+Purban,data = UN3_0)
termplot(model_2, terms = "Pop",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
```

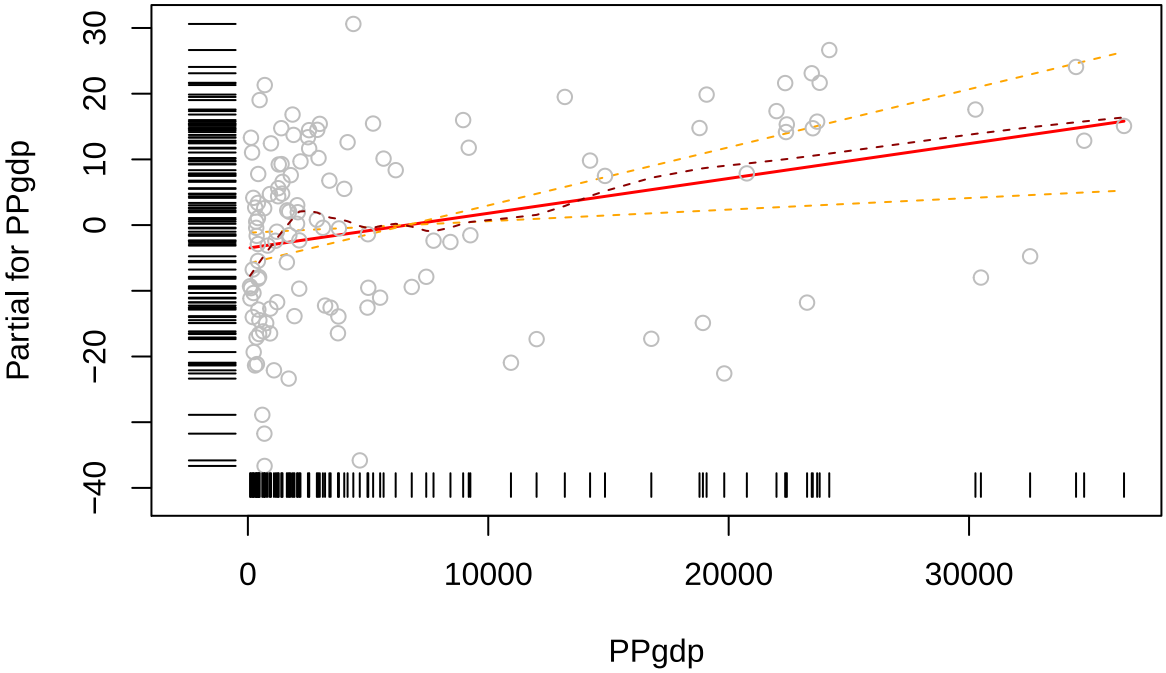


```r
termplot(model_3, terms = "pop_sqrt",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
```

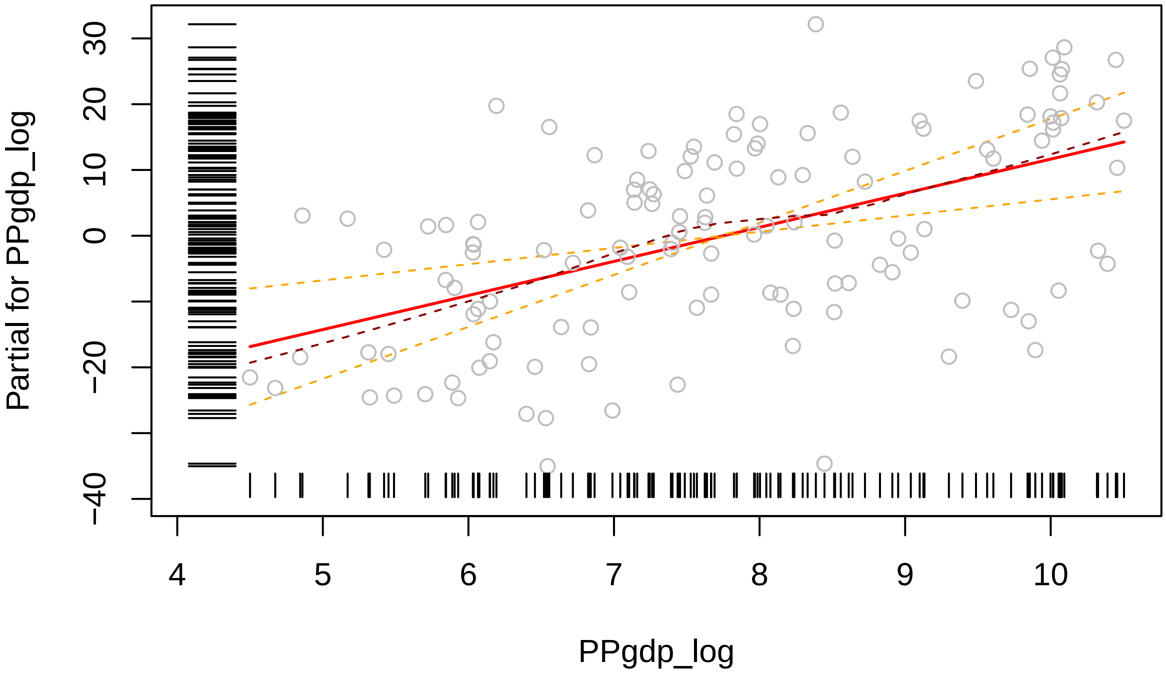


Comments: We can see that the termplot looks much better for the variable population, as it follows a more linear trend for the data points

```
termplot(model_2, terms = "PPgdp",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
```

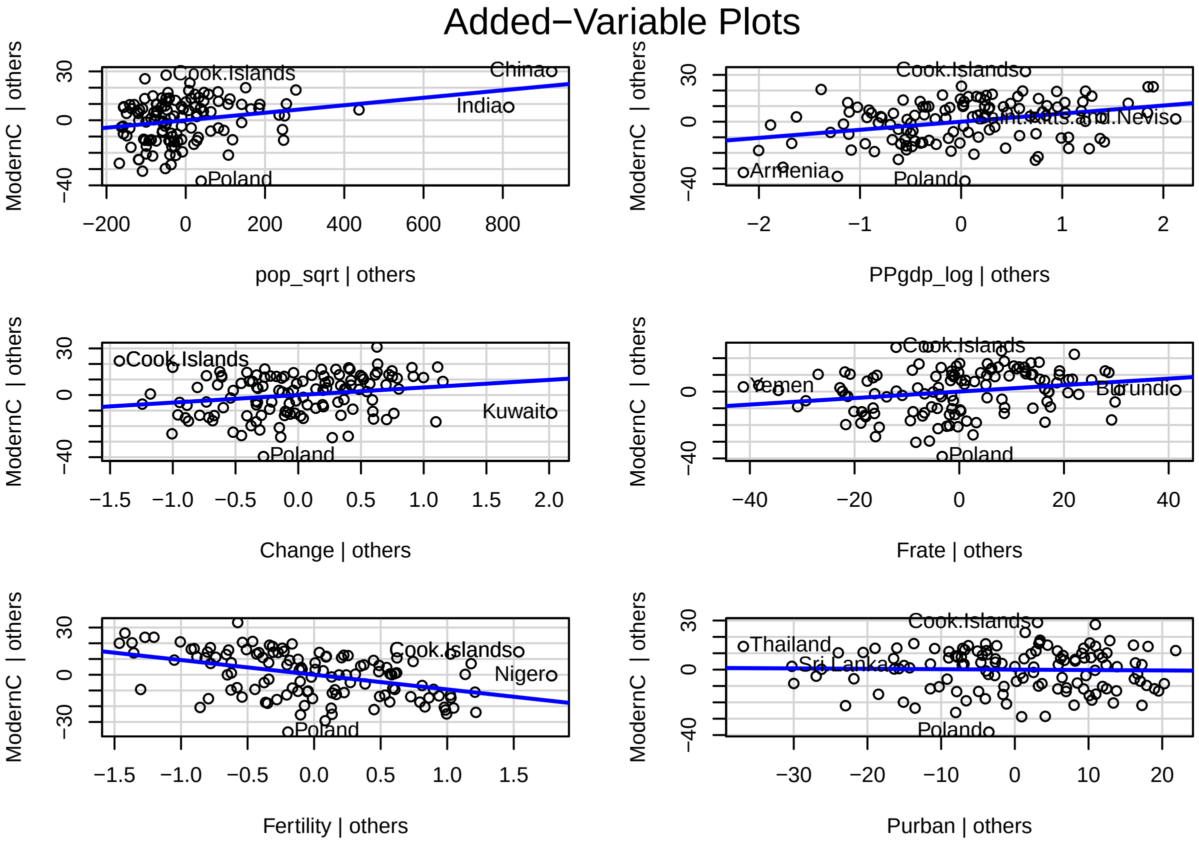


```
termplot(model_3, terms = "PPgdp_log",
         partial.resid = T, se=T, rug=T,
         smooth = panel.smooth)
```

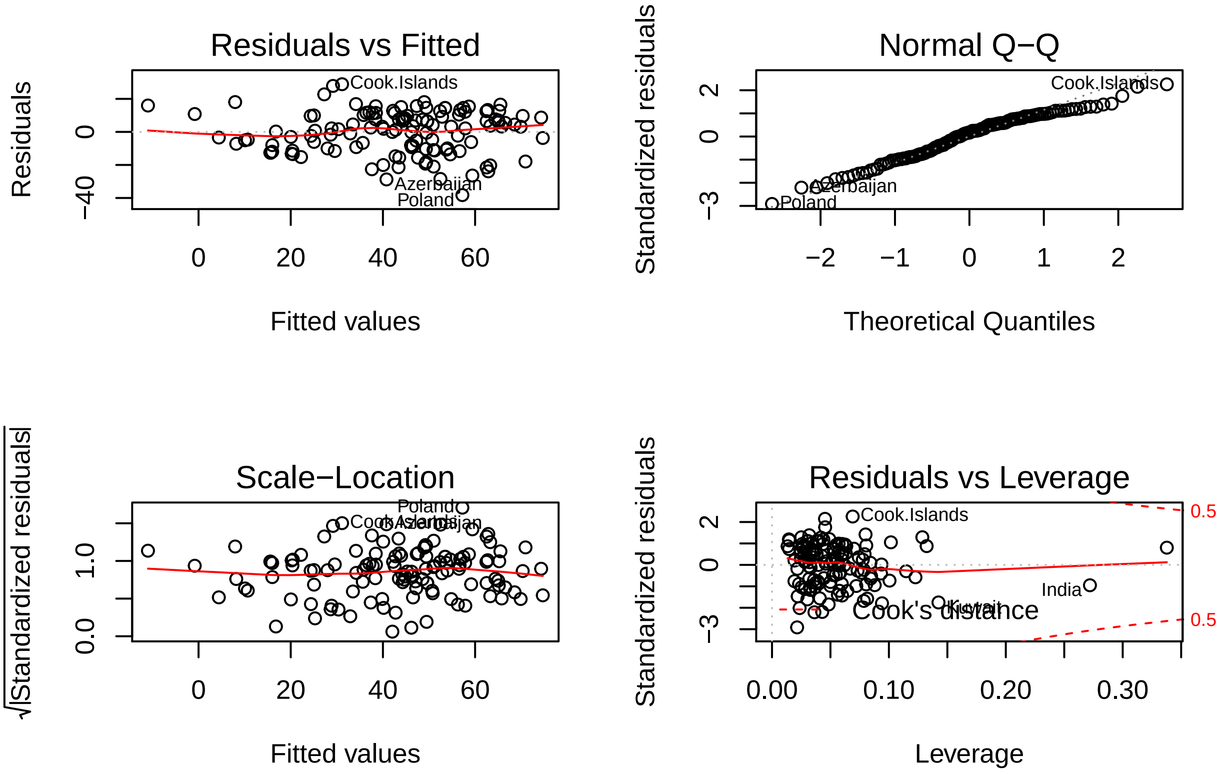


Comments: We can see that the termplot looks much better for the variable GDP as well

```
avPlots(model_3, terms=~.)
```



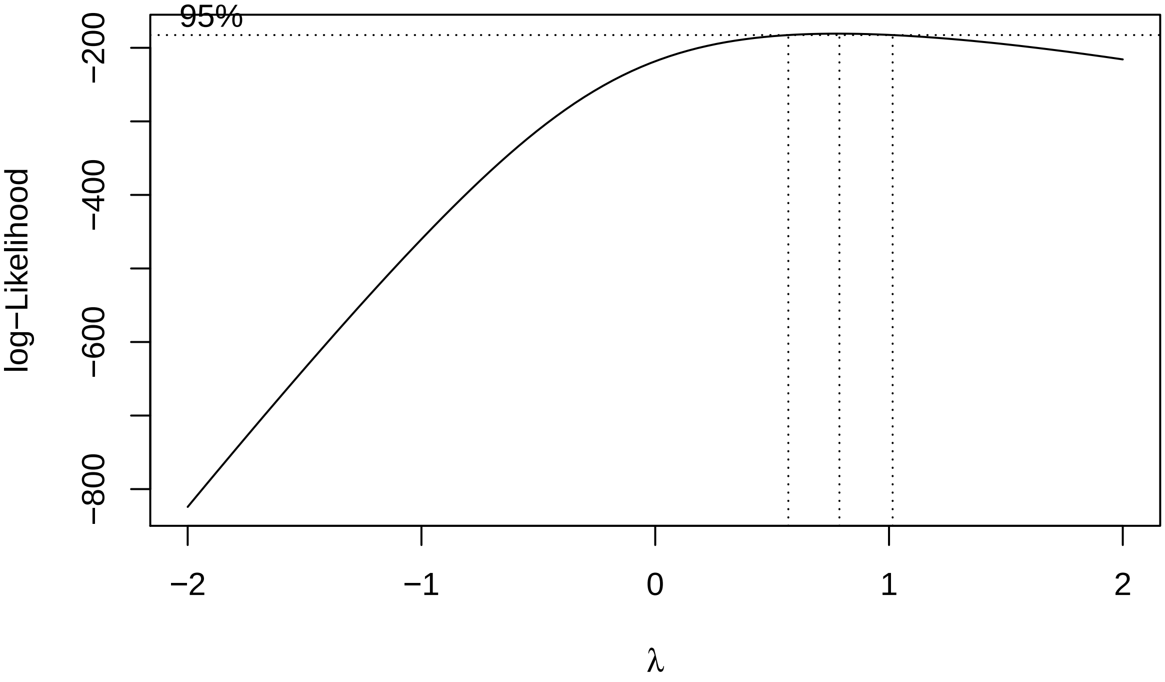
```
par(mfrow=c(2,2))
plot(model_3)
```



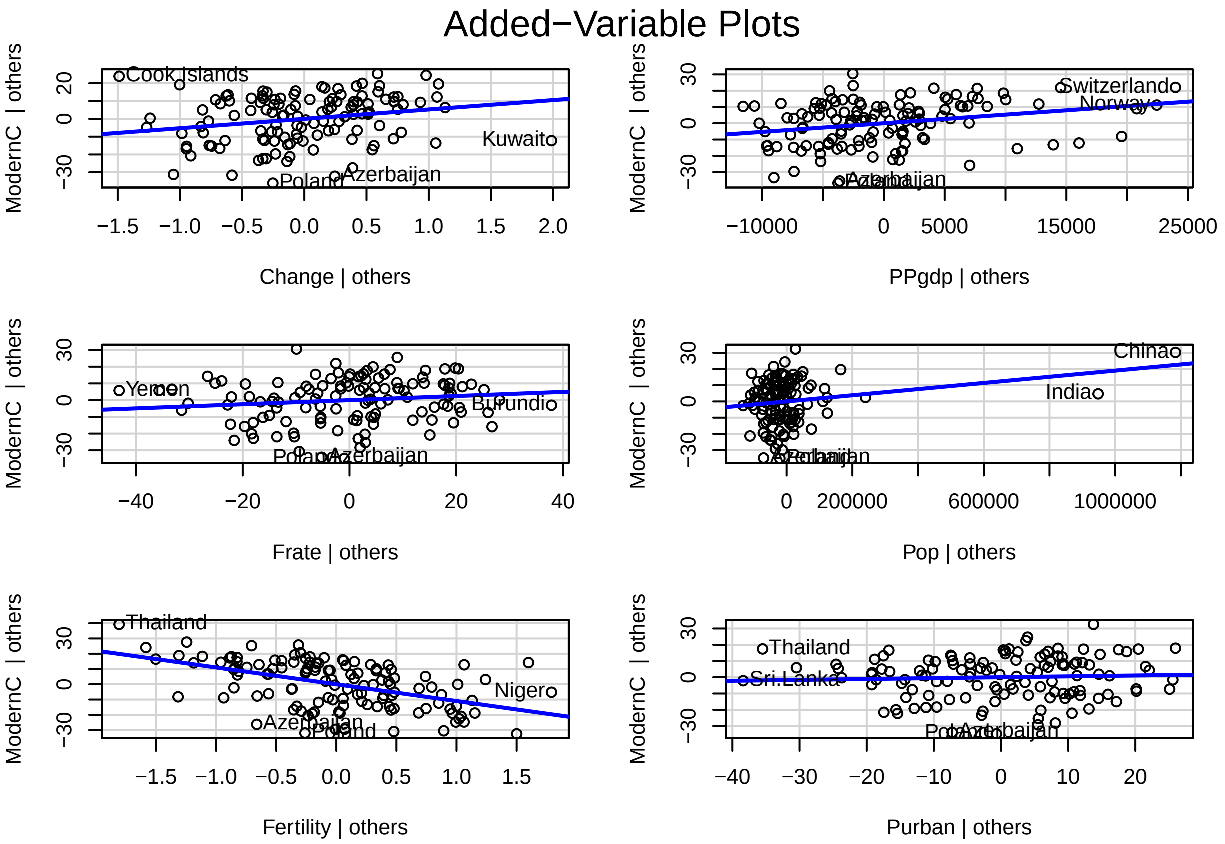
Comment: We see that the AVplot for PPgdp and Pop has improved a lot compared with the untransformed model. The residual plot is too. Though QQ-plot still looks heavy-tailed, the scale location plot suggests constant variance.

9. Start by finding the best transformation of the response and then find transformations of the predictors. Do you end up with a different model than in 8?

```
MASS::boxcox(ModernC~., data = UN3_0,lambda = seq(-2, 2, length = 20))
```



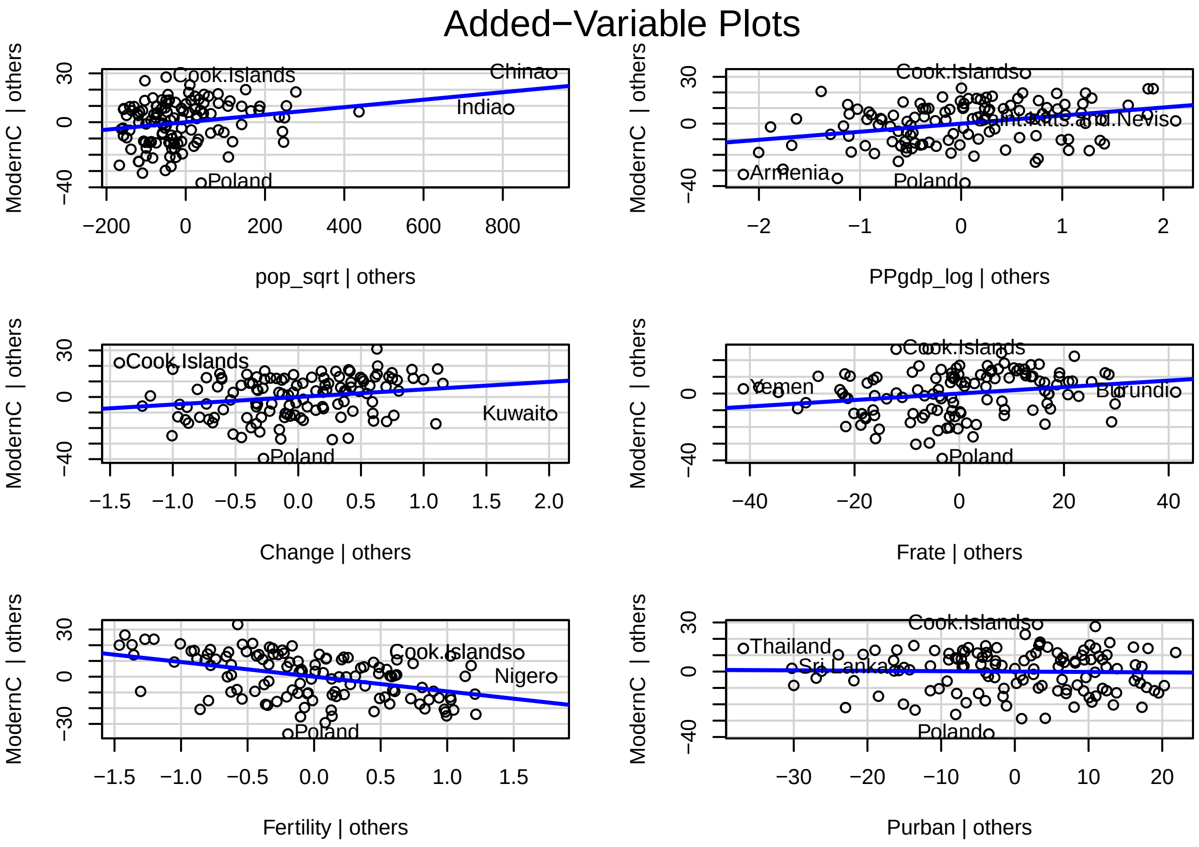
```
#Comments: It looks like the MLE for lamda for the predictor is around 0.8, which can round up to 1
# So we do not need to transform the predictor.
model_4<-lm(ModernC~.,data = UN3_0)
avPlots(model_4, terms=~.) # the avplot suggest Pop and PPgdp may need transformation
```



```
car::boxTidwell(ModernC~Pop+PPgdp,other.x=~Change+Frate+Fertility+Purban,data=UN3_0)
```

```
##        MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop          0.40749             -0.7874   0.4310
## PPgdp       -0.12921             -1.1410   0.2539
##
## iterations =  4
```

```
#repeat the BoxTidewell power transformation again to see what values should Pop and PPgdp to be transf
model_4<-lm(ModernC~pop_sqrt+PPgdp_log+Change+Frate+Fertility+Purban,data = UN3_0)
avPlots(model_4, terms=~.) # avplot looks much better
```



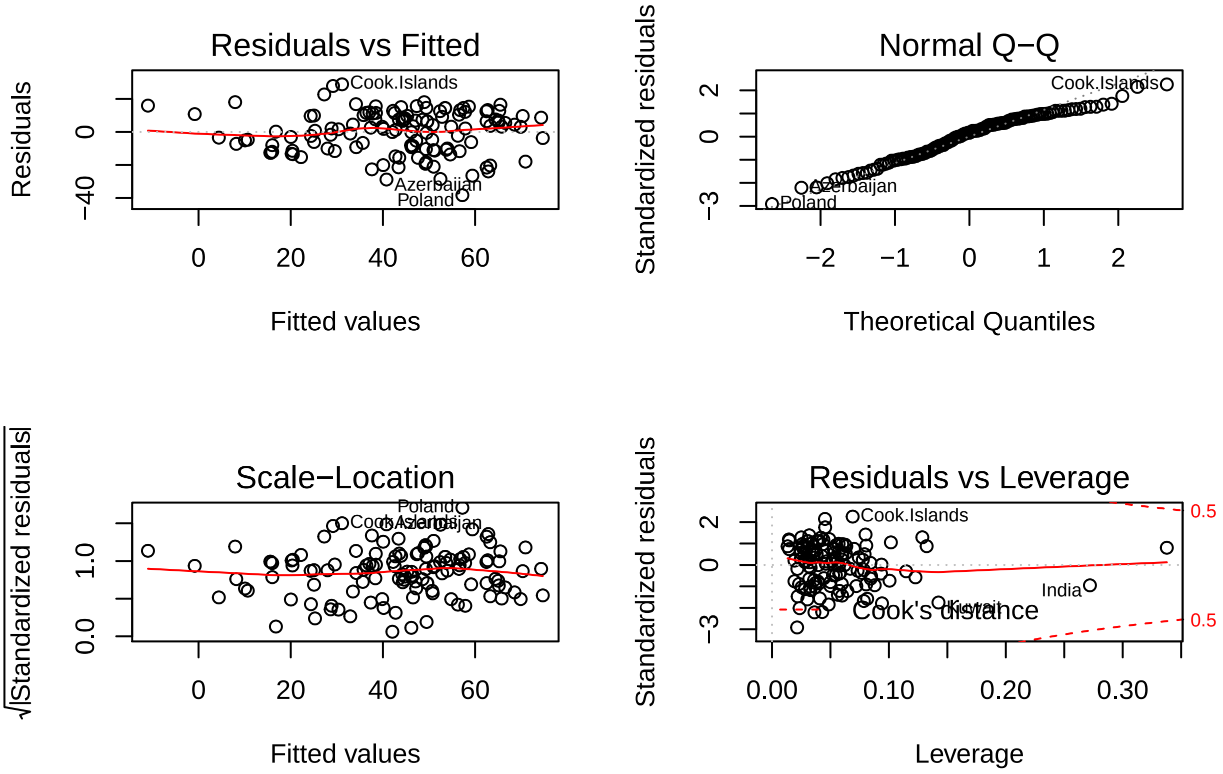
Comments: We find the transformation from response to predictor to be exactly the same

10. Are there any outliers or influential points in the data? Explain. If so, refit the model after removing any outliers and comment on residual plots.

```
outlierTest(lm(ModernC~., data=UN3_0)) # test for outliners
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##         rstudent unadjusted p-value Bonferonni p
## Poland -2.657549          0.0089714           NA
```

```
#UN3_1<-UN3_0[-c(25,50),]
par(mfrow=c(2,2))
plot(model_3)
```



Comments: The outlinerTest suggest there is no studentized residuals with Bonferonni p<0.05. Thus, we think there is no outliners.

We will use model 3. For the residual versus leverage plot, we can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines. And we see that there even though points like China and india are high leverage, they have small residuals. So maybe there are no potential outliners and influential points.

Summary of Results

11. For your final model, provide summaries of coefficients with 95% confidence intervals in a nice table with interpretations of each coefficient. These should be in terms of the original units!

```
C<-summary(model_3)$coefficient
A<-data.frame(confint(model_3))
B<-data.frame(C[,1],A)
colnames(B)<-c("Estimate","Lower Bound","Upper Bound")
B<-round(B,5)
kable(B,format = "markdown")
```

	Estimate	Lower Bound	Upper Bound
(Intercept)	12.96807	-11.28629	37.22242
pop_sqrt	0.02302	0.00780	0.03824
PPgdp_log	5.18241	2.49107	7.87376
Change	4.86915	0.82391	8.91438
Frate	0.19401	0.04341	0.34461
Fertility	-9.32757	-12.79260	-5.86254
Purban	-0.02507	-0.21628	0.16614

Comments: A one unit increase in the square root of Population would result a 0.02302 unit increase in Modern C (Percent of unmarried women using a modern method of contraception) holding other variables constant. And the 95% CI for the square root of population is [0.008, 0.038]

A one unit increase in the log of gdp would result in a 5.18 unit increase in Modern C (Percent of unmarried women using a modern method of contraception) holding other variables constant. And the 95% CI for the log of gdp is [2.49 7.87]
A one unit increase in change would result in a 4.869 unit increase in Modern C (Percent of unmarried women using a modern method of contraception) holding other variables constant. And the 95% CI for the change is [0.824 8.914]
A one unit increase in the Frate would result in a 0.194 unit increase in Modern C (Percent of unmarried women using a modern method of contraception) holding other variables constant. And the 95% CI for Frate is [0.043 0.345]
A one unit increase in the Fertility would result in a -9.327 unit decrease in Modern C (Percent of unmarried women using a modern method of contraception) holding other variables constant. And the 95% CI for Fertility is [-12.793 -5.863]
A one unit increase in the Purban would result in a -0.025 unit decrease in Modern C (Percent of unmarried women using a modern method of contraception) holding other variables constant. And the 95% CI for Purban is [-0.216 0.166]

12. Provide a paragraph summarizing your final model and findings suitable for the US envoy to the UN after adjusting for outliers or influential points. You should provide a justification for any case deletions in your final model

```
summary(model_3)
```

```
##
## Call:
## lm(formula = ModernC ~ pop_sqrt + PPgdp_log + Change + Frate +
##     Fertility + Purban, data = UN3_0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.239  -9.995   2.133   9.961  28.861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.968065  12.247992   1.059  0.29186
## pop_sqrt     0.023017   0.007685   2.995  0.00335 **
## PPgdp_log    5.182415   1.359076   3.813  0.00022 ***
## Change       4.869147   2.042766   2.384  0.01874 *
## Frate        0.194010   0.076053   2.551  0.01202 *
## Fertility   -9.327572   1.749773  -5.331 4.77e-07 ***
## Purban      -0.025072   0.096557  -0.260  0.79558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.26 on 118 degrees of freedom
## Multiple R-squared:  0.6362, Adjusted R-squared:  0.6177
## F-statistic:  34.4 on 6 and 118 DF,  p-value: < 2.2e-16
```

Comments: The final model is:

ModernC ~ sqrt(Pop)+Log(PPgdp)+Change+Frate+Fertility+Purban

Findings: We think the larger the population growth rate and the larger the population is, the more likely for unmariied women in a country to use modern method of contraception

The larger the per capital GDP is, the more likely for unmariied women in a country to use modern method of contraception

The larger the percent of females over age 15 economically active is, the more likely for unmariied women in a country to use modern method of contraception

The larger the expected number of live births per female is, the more likely for unmariied women in a country to use modern method of contraception

We did not delete and outliners or influential points. But in our model we only considered the data with no missing values because when we use the lm model fit, lm would only consider data that does not contain NA values.

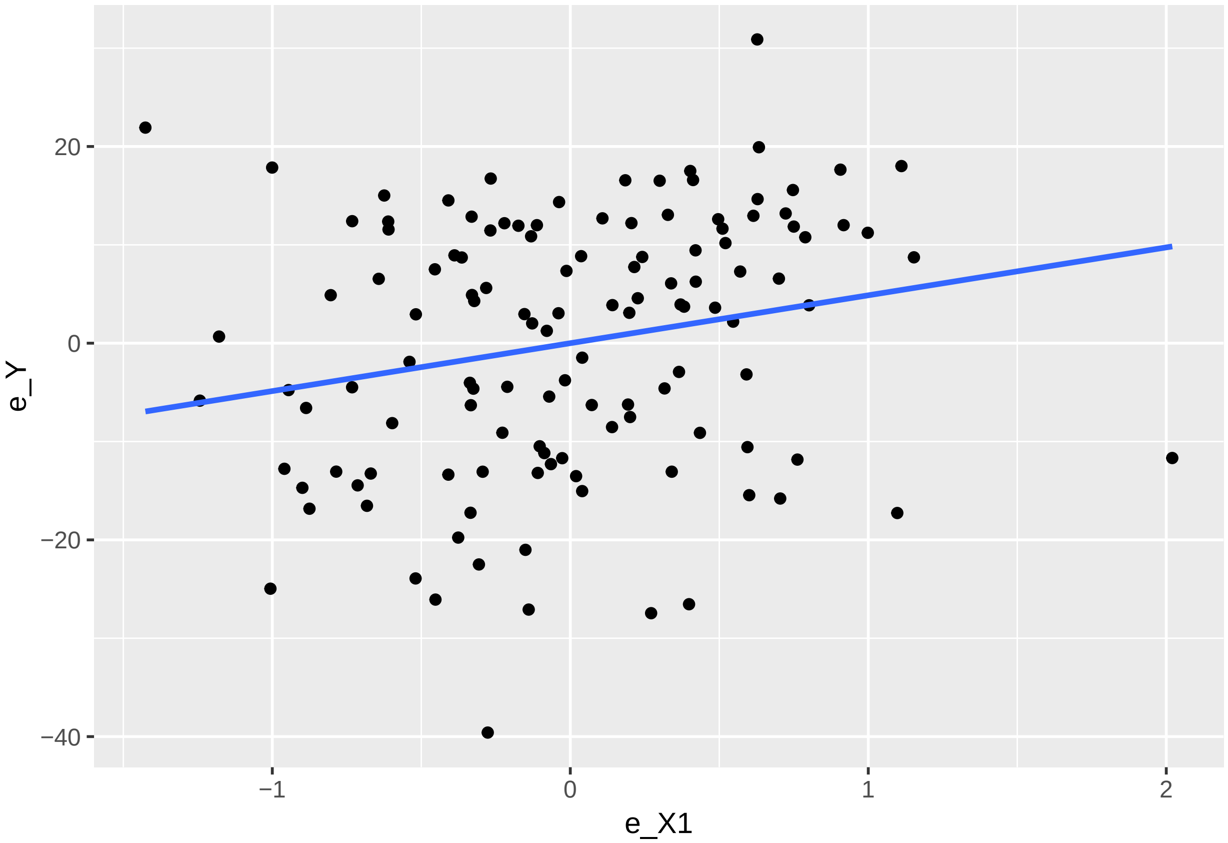
Methodology

13. Prove that the intercept in the added variable scatter plot will always be zero. *Hint: use the fact that if H is the project matrix which contains a column of ones, then $1_n^T(I-H)=0$. Use this to show that the sample mean of residuals will always be zero if there is an intercept.*

$$
\begin{aligned}
e_{(y)} &= \hat{\beta}_0 + \hat{\beta}_1 e_x \ , \ H = X(X^TX)^{-1}X^T \\
\underbrace{(I-H)y}_{y} &= \hat{\beta}_0 I + \hat{\beta}_1 \underbrace{(I-H)x_3}_{x} \\
(I-H)y &= \hat{\beta}_0 I + [x_3^T(I-H)^T(I-H)x_3]^{-1}[(I-H)x_3]^T(I-H)y(I-H)x_3 \\
(I-H)y &= \hat{\beta}_0 I + [x_3^T(I-H)x_3]^{-1}x_3^T(I-H)y(I-H)x_3 \\
(I-H)y &= \hat{\beta}_0 I + [x_3^T(I-H)x_3]^{-1}x_3^T(I-H)y(I-H)x_3 \ , \ I \ - \ H \ is \ idempotent \\
x_3^T(I-H)y &= x_3^T\hat{\beta}_0 I + x_3^T[x_3^T(I-H)x_3]^{-1}x_3^T(I-H)y(I-H)x_3 \\
x_3^T(I-H)y &= x_3^T\hat{\beta}_0 I + [x_3^T(I-H)x_3][x_3^T(I-H)x_3]^{-1}x_3^T(I-H)y \ , \ constant \\
x_3^T(I-H)y &= x_3^T\hat{\beta}_0 I + x_3^T(I-H)y \\
0 &= x_3^T\hat{\beta}_0 I \\
\sum_{i=1}^{n} x_3^{(i)}\hat{\beta}_0 &= 0 \\
Thus, \ \hat{\beta}_0 \ is \ 0
\end{aligned}
$$

14. For multiple regression with more than 2 predictors, say a full model given by `Y ~ X1 + X2 + ... Xp` we create the added variable plot for variable `j` by regressing `Y` on all of the `X`'s except `Xj` to form `e_Y` and then regressing `Xj` on all of the other X's to form `e_X`. Confirm that the slope in the manually constructed added variable plot for one of the predictors in Ex. 10 is the same as the estimate from your model.

```r
e_Y = residuals(lm(ModernC~pop_sqrt+PPgdp_log+Frate+Fertility+Purban,data=UN3_0))
e_X1 = residuals(lm(Change ~ pop_sqrt+PPgdp_log+Frate+Fertility+Purban,data=UN3_0))
df = data.frame(e_Y = e_Y, e_X1 = e_X1)
ggplot(data=df, aes(x = e_X1, y = e_Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```
summary(lm(ModernC~pop_sqrt+PPgdp_log+Frate+Fertility+Purban+Change,data=UN3_0))$coef
```

```
##                Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 12.96806511 12.247991809  1.0587911 2.918575e-01
## pop_sqrt     0.02301658  0.007685224  2.9949129 3.345754e-03
## PPgdp_log    5.18241483  1.359076309  3.8131890 2.197102e-04
## Frate        0.19401002  0.076052657  2.5509960 1.202073e-02
## Fertility   -9.32757211  1.749773407 -5.3307314 4.768203e-07
## Purban      -0.02507157  0.096556923 -0.2596558 7.955817e-01
## Change       4.86914689  2.042765633  2.3836053 1.873953e-02
```

```
summary(lm(e_Y ~ e_X1, data=df))$coef
```

```
##                  Estimate Std. Error       t value   Pr(>|t|)
## (Intercept) -5.733491e-16   1.161417 -4.936633e-16 1.00000000
## e_X1         4.869147e+00   2.000815  2.433581e+00 0.01638773
```

Comments: Coefficient for Change is the same, which is 4.869.