HW2 STA521 Fall18

Yuwei Xu, yx144, vivixu7

Due September 23, 2018 5pm

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*

```
library(alr3)
```

```
## Loading required package: car
```

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

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

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

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

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

According to the summary, 6 variables out of all 7, except Purban have missing data (nonzero number of Na's). All variables are quantitative variables.

```
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
```

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

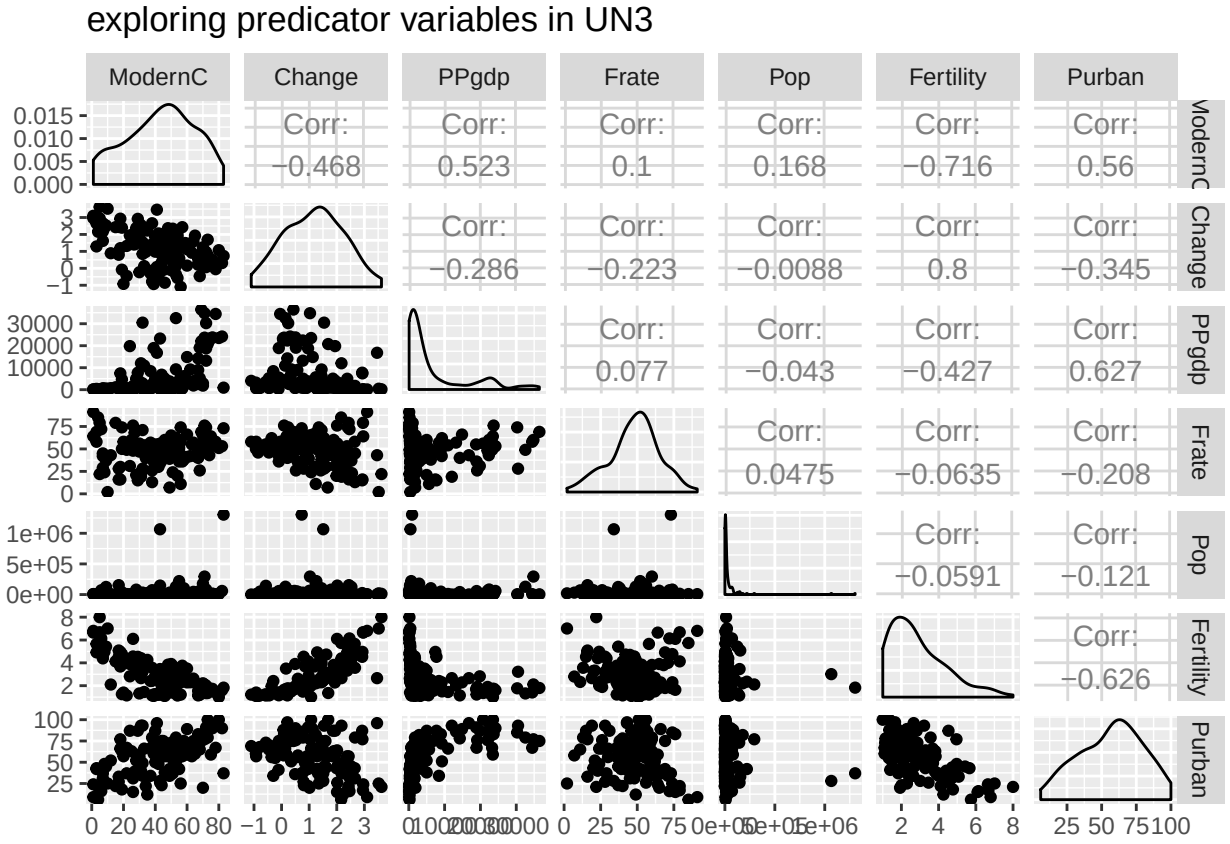
```
qp_mu <- sapply(UN3, mean, na.rm = TRUE)
qp_sd <- sapply(UN3, sd, na.rm = TRUE)
qpdata<-cbind(qp_mu,qp_sd)
kable(qpdata, digits=2,col.names=c("Mean", "Standard Deviation"))
```

	Mean	Standard Deviation
ModernC	38.72	22.64
Change	1.42	1.13
PPgdp	6527.39	9325.19
Frate	48.31	16.53
Pop	30281.87	120676.69
Fertility	3.21	1.71
Purban	56.20	24.11

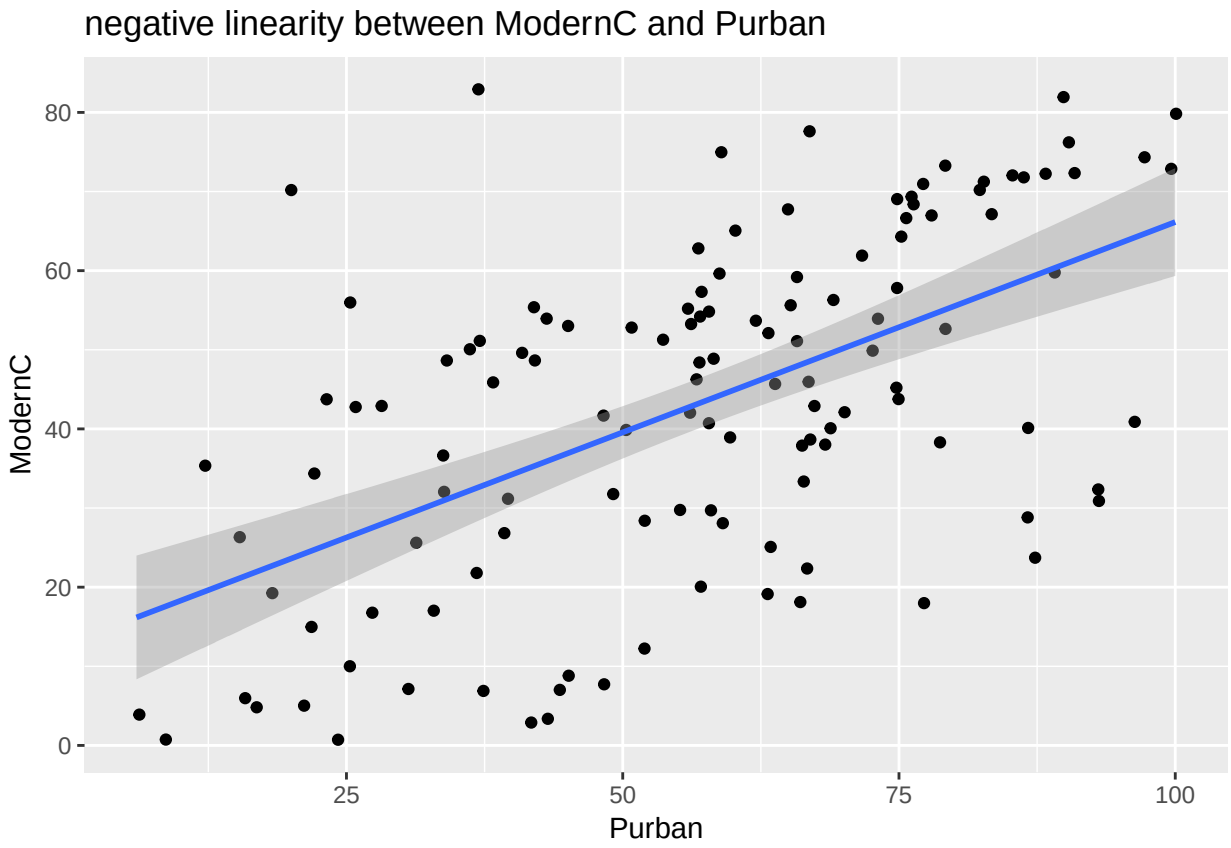
3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict `ModernC` from the other variables. Are there potential outliers, nonlinear relationships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

First clean the data by removing the "Na"s. Then using ggpairs and ggplots to explore correlations between any two variables. From the plots we notice four things: 1) ModernC seems postiviely, linearly related with Purban, negatively and linearly related with Fertility and Change. 2) India and China are two outliers in the scattorplot of Pop. 3) ModernC seems to have non-linear relation with Pop, Frate and PPgdp, thus needs transformation. 4) PPgdp seems to have non-linear relation with Pop, thus needs transformation.

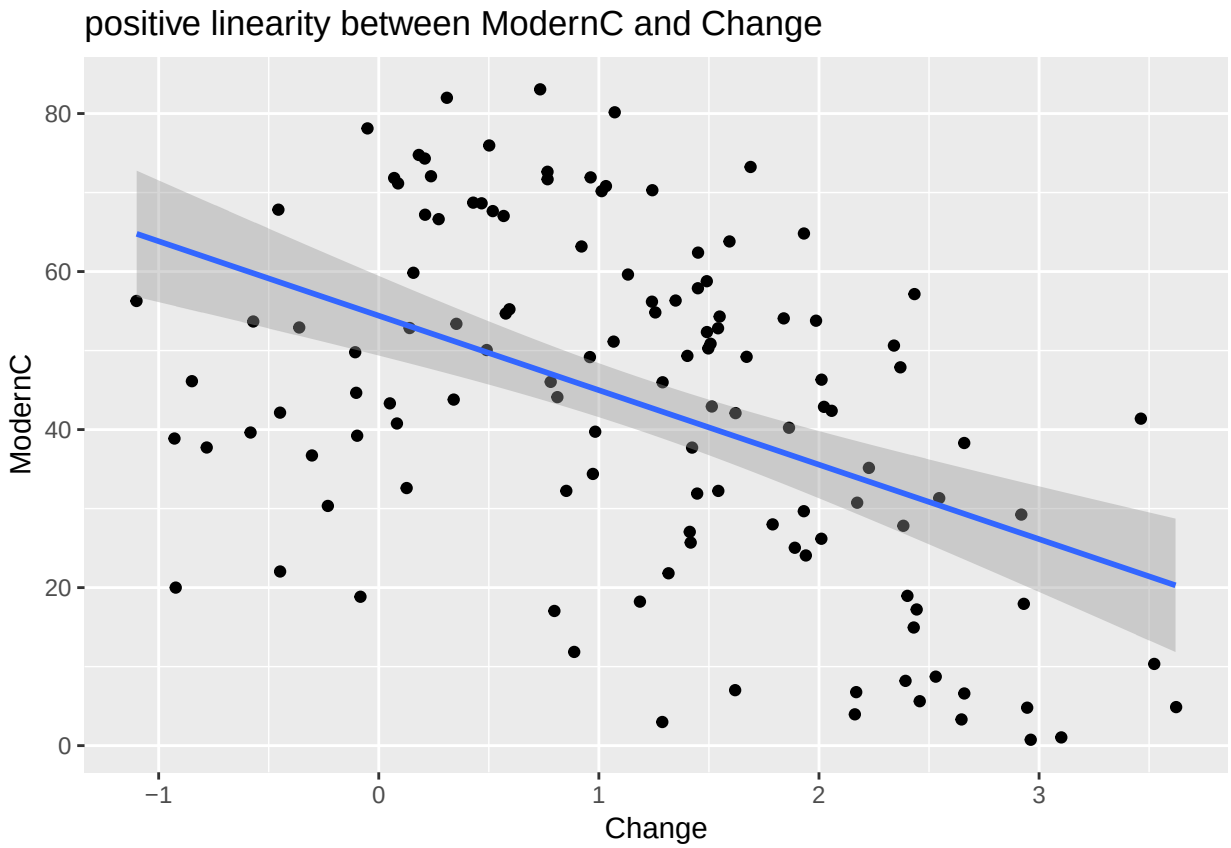
```
UN3clean<-UN3[complete.cases(UN3),]
ggpairs(UN3clean,progress=FALSE)+ggtitle("exploring predicator variables in UN3")
```



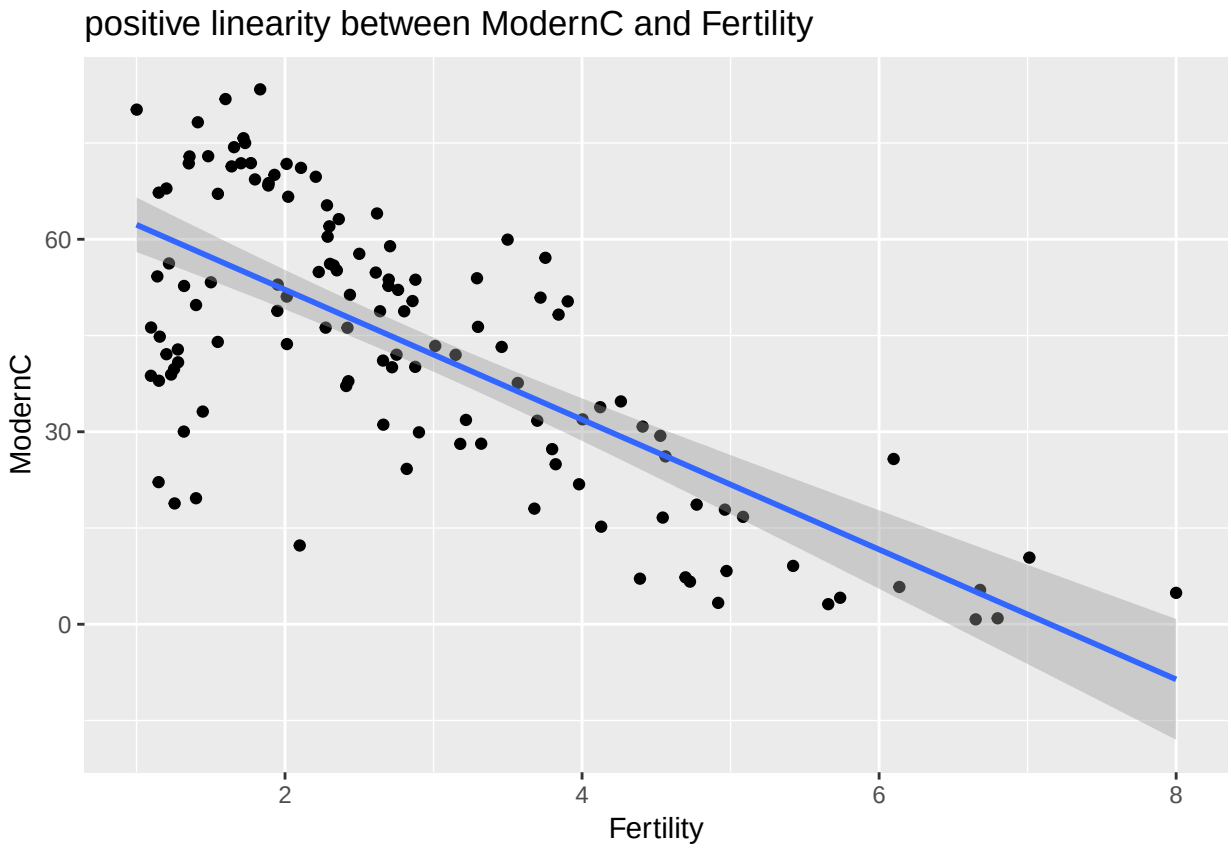

```
par(mfrow=c(2,2))
ggplot(UN3clean,aes(Purban,ModernC))+geom_jitter()+geom_smooth(method="lm")+ggtitle("negative linearity
```




```
ggplot(UN3clean,aes(Change,ModernC))+geom_jitter()+geom_smooth(method="lm")+ggtitle("positive linearity
```




```
ggplot(UN3clean,aes(Fertility,ModernC))+geom_jitter()+geom_smooth(method="lm")+ggtitle("positive linear
```

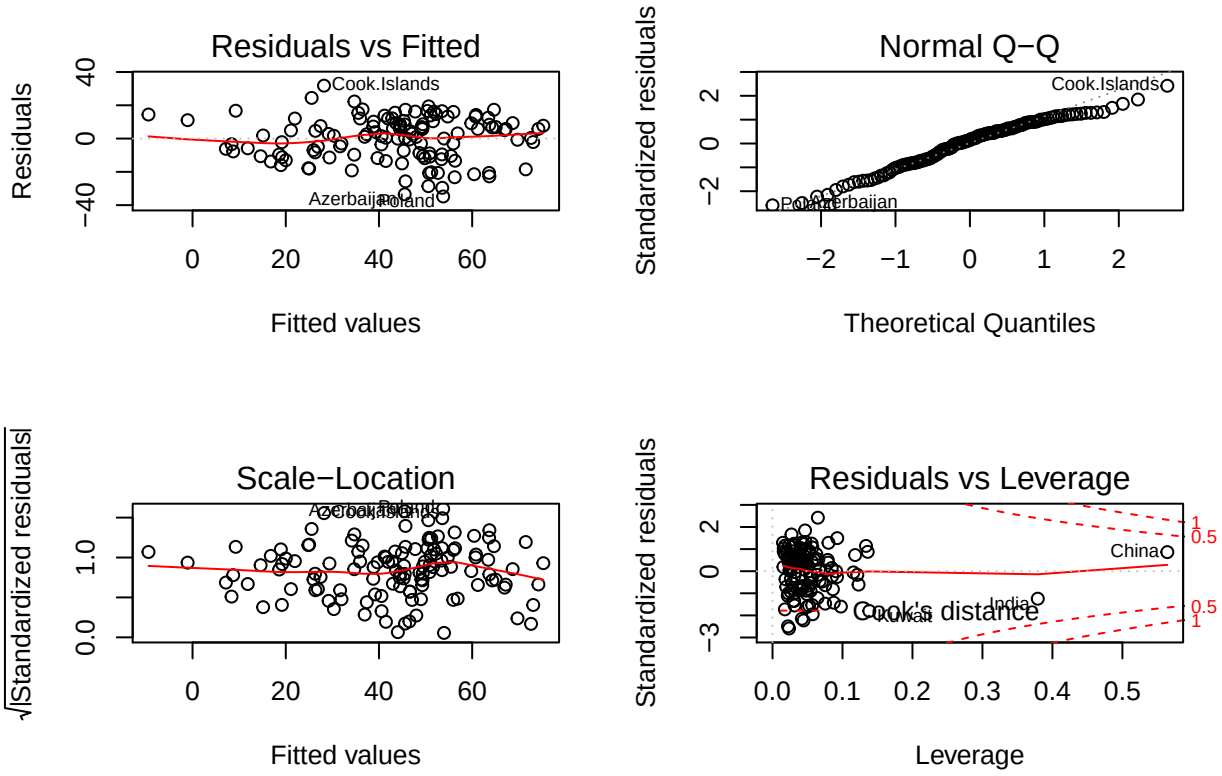


Model Fitting

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?

Since there were 85 missing datas, 125 observations are actually used in model fitting. The Residuals vs Leverage plot shows that China, India both have large leverages. The Residuals vs Fitted shows that Cook Island and Poland are the furthest away from fitted value.These observations, together with the Scale-Location plot show that he equal variance assumption isn't quite true. Furthermore, Normal Q-Q plot shows that normality assumption is not quite true.

```
nontr = lm(ModernC ~ ., data = UN3)
par(mfrow=c(2,2))
plot(nontr, ask=F)
```



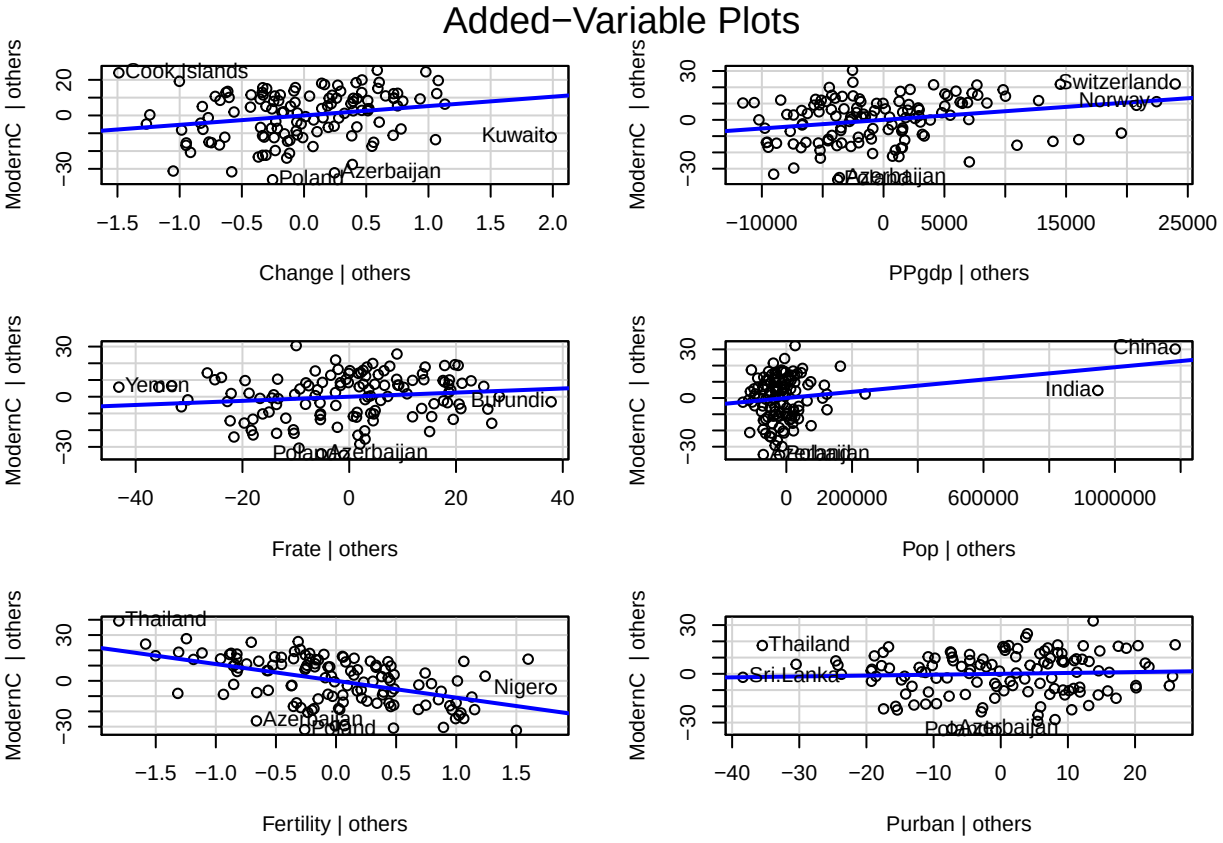

```
summary(nontr)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN3)
##
## 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
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5989
## F-statistic: 31.85 on 6 and 118 DF,  p-value: < 2.2e-16
```

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?

From the Added-Variable Plots, clearly China and India are quite influential for Pop. Switzerland and Norway have some influences on PPgdp. Cook Islands and Kuwait are influential to Change but their influences are in opposite directions. Overall, due to these influential localities, Pop and PPgdp need transformation to better estimate their influence on ModernC.

```
avPlots(nontr)
```



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.

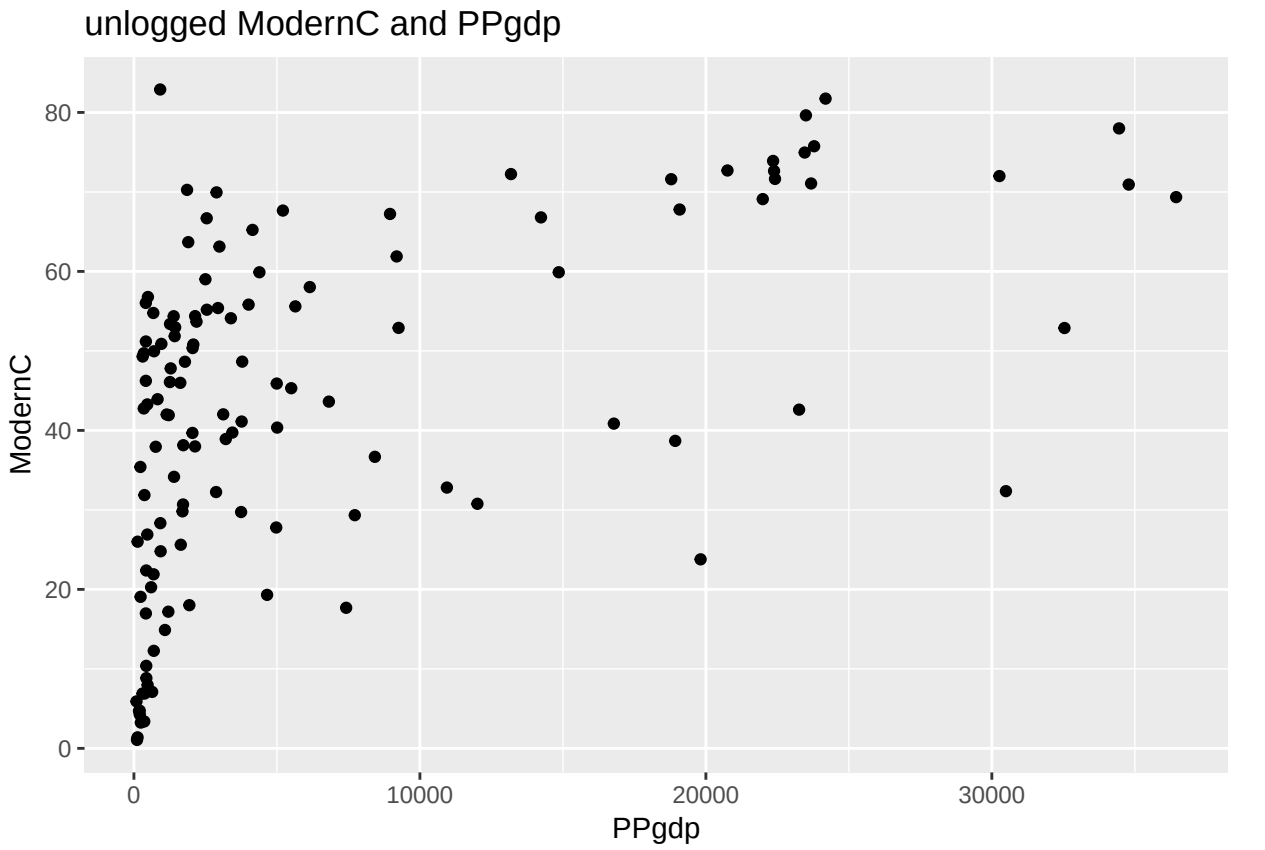
Based on my conclusions from 5, Pop and PPgdp needs transformation, so we use Box-Tidwell on these two predictors to find corresponding appropriate transformation. After transforming negative predictors Change, the result show that both P-values of PPgdp and Pop are not statistically significant if our cut-off is 0.05. However, we must also account for the large values in Pop and PPgdp. Typically it makes sense to take log when the variable values are large. Thus, I decide that the appropriate transformation for both Pop and PPgdp is taking log. Attached plots show comparison between un-transformed and transformed relation between Pop, PPgdp vs. ModernC. Clearly taking log does clear things out.The log-transformed PPgdp vs. ModernC looks linear.The log-transformed Pop vs. ModernC have a better scale that enables us to examine relationships within the Pop.

```
#Change has nonnegative values, so transform to non-negative first
UN3clean$Change = UN3clean$Change + 1.1 + 1
boxTidwell(ModernC ~ PPgdp+Pop, ~ Change+Frate+Purban+Fertility, data = UN3clean)
```

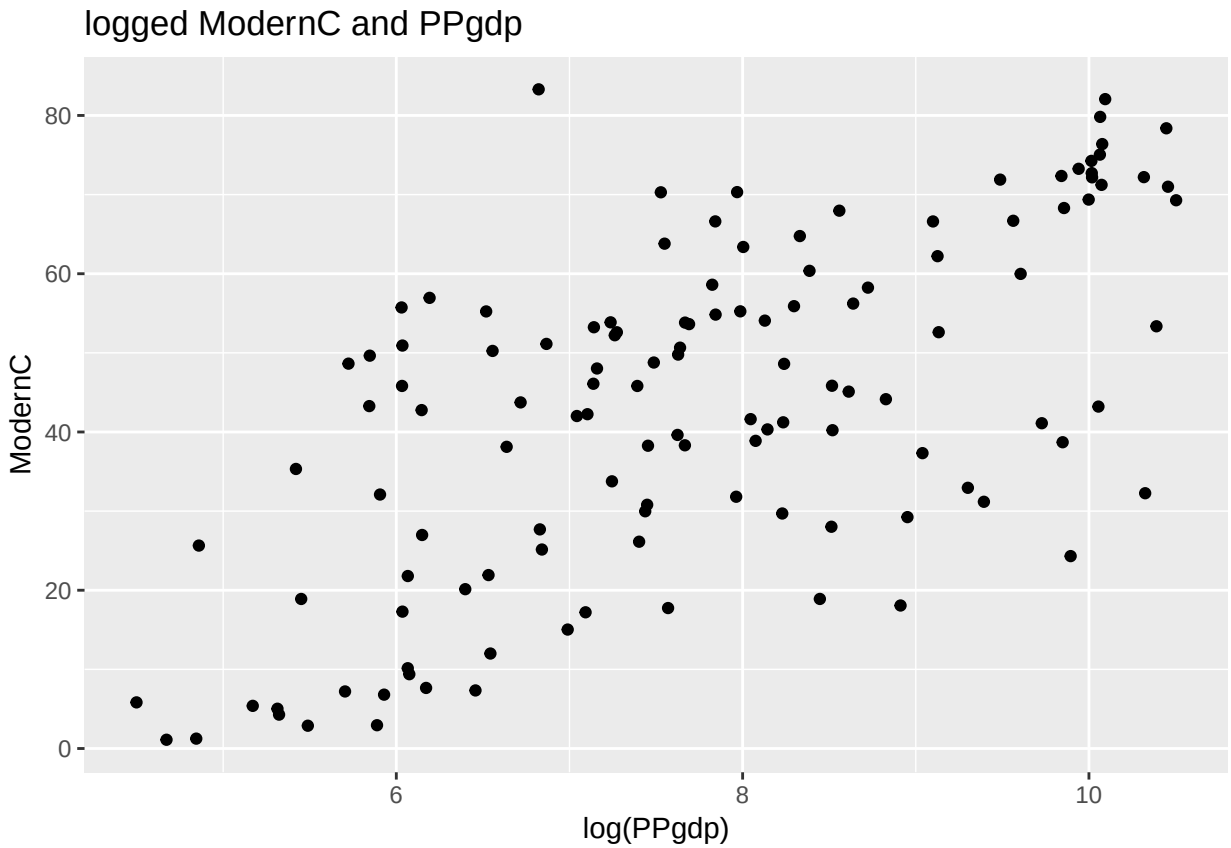
```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp      -0.12921             -1.1410   0.2539
## Pop         0.40749             -0.7874   0.4310
```

```
##
## iterations =  4
```

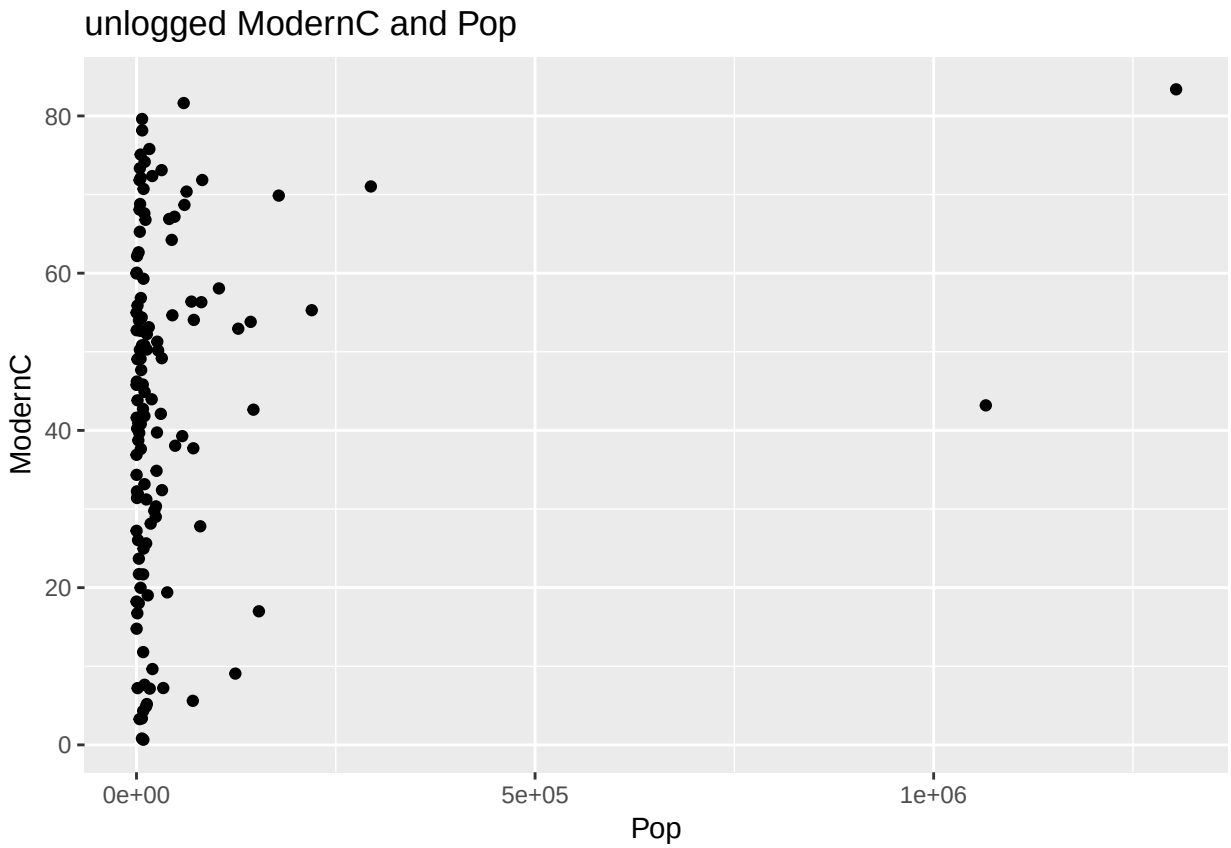
```
par(mfrow=c(2,2))
ggplot(UN3clean,aes(PPgdp,ModernC))+geom_jitter()+ggtitle("unlogged ModernC and PPgdp")
```



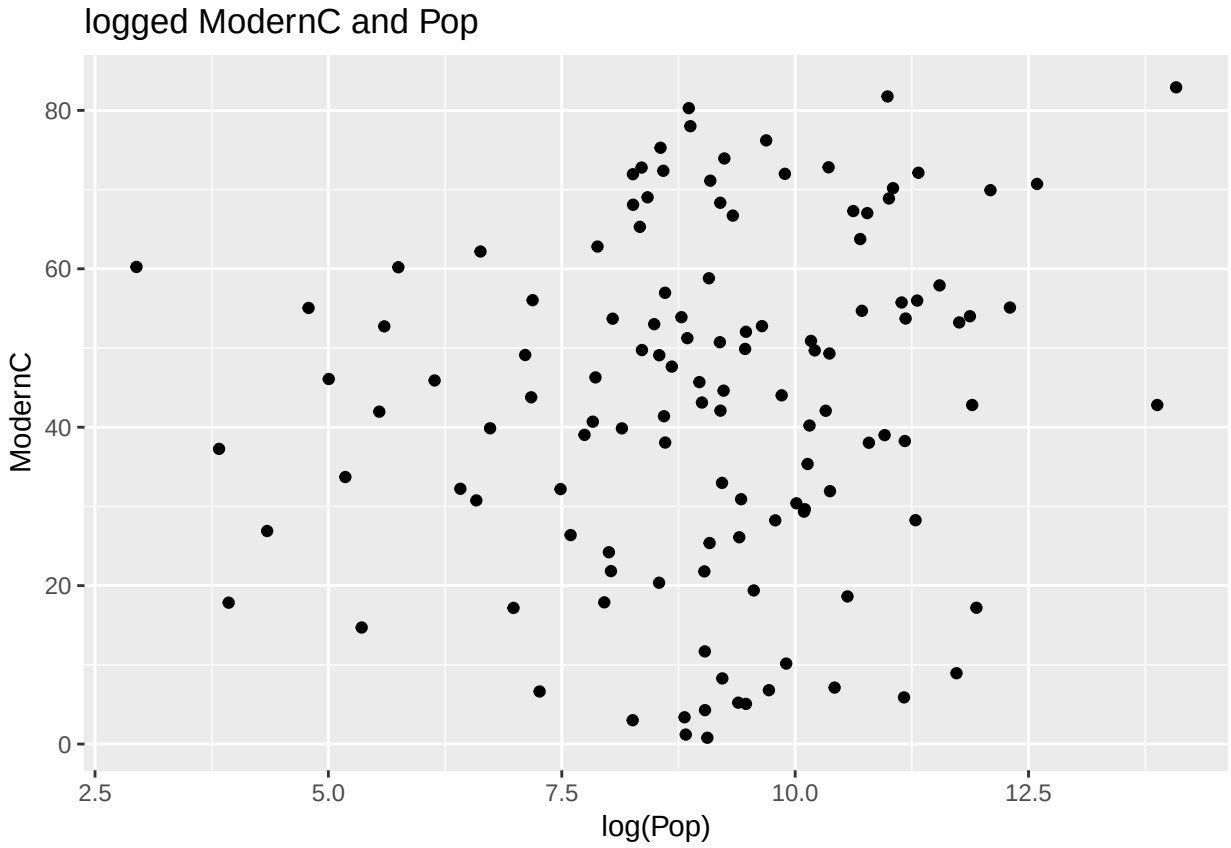

```
ggplot(UN3clean,aes(log(PPgdp),ModernC))+geom_jitter()+ggtitle("logged ModernC and PPgdp")
```




```
ggplot(UN3clean,aes(Pop,ModernC))+geom_jitter()+ggtitle("unlogged ModernC and Pop")
```



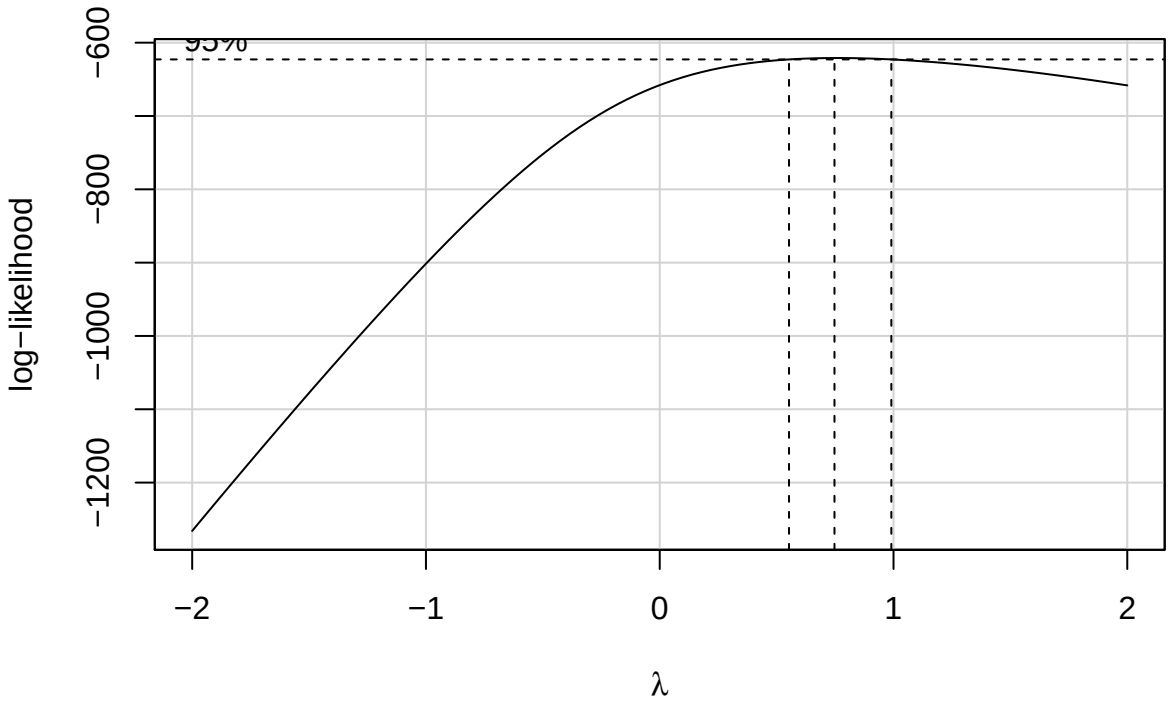

```
ggplot(UN3clean,aes(log(Pop),ModernC))+geom_jitter()+ggtitle("logged ModernC and Pop")
```



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

According to Boxcox, lambda is close to 1, therefore the transformation of the response is not necessary.

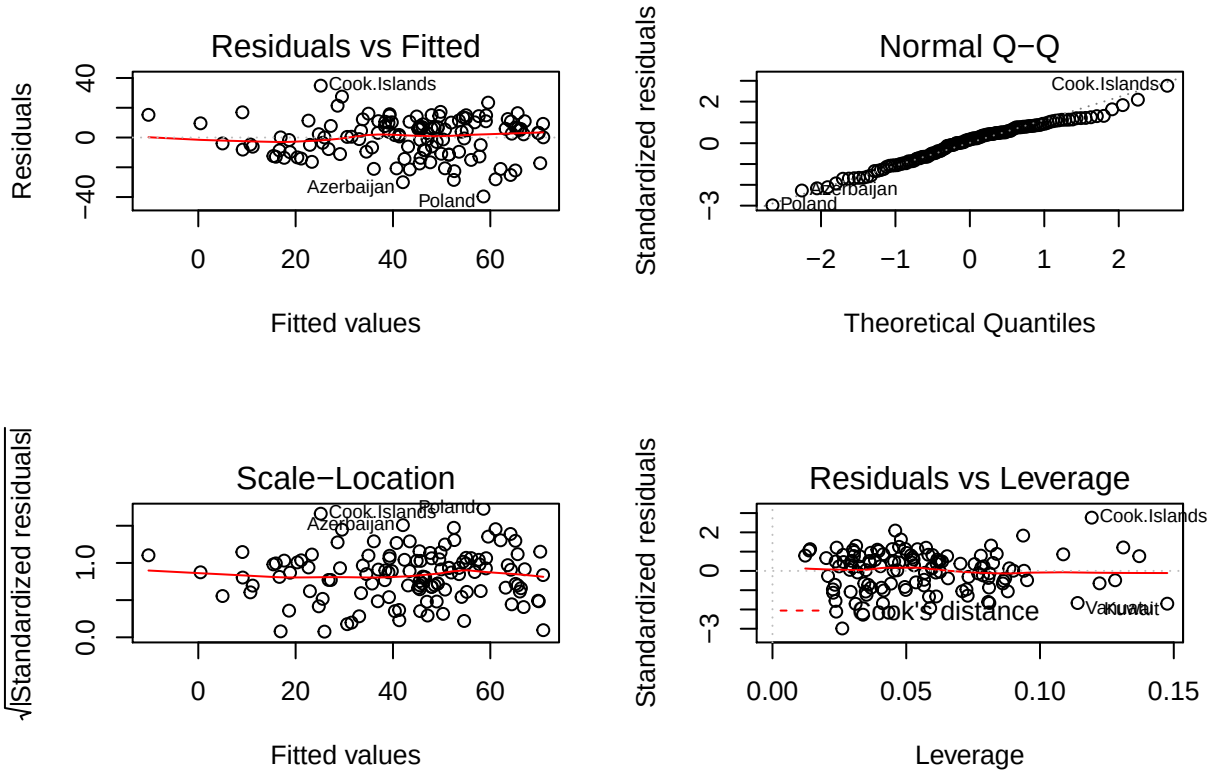
```
trns = lm(ModernC ~ log(PPgdp)+log(Pop)+Frate+Change+Purban+Fertility, data = UN3clean)
lambda = boxCox(trns)
```



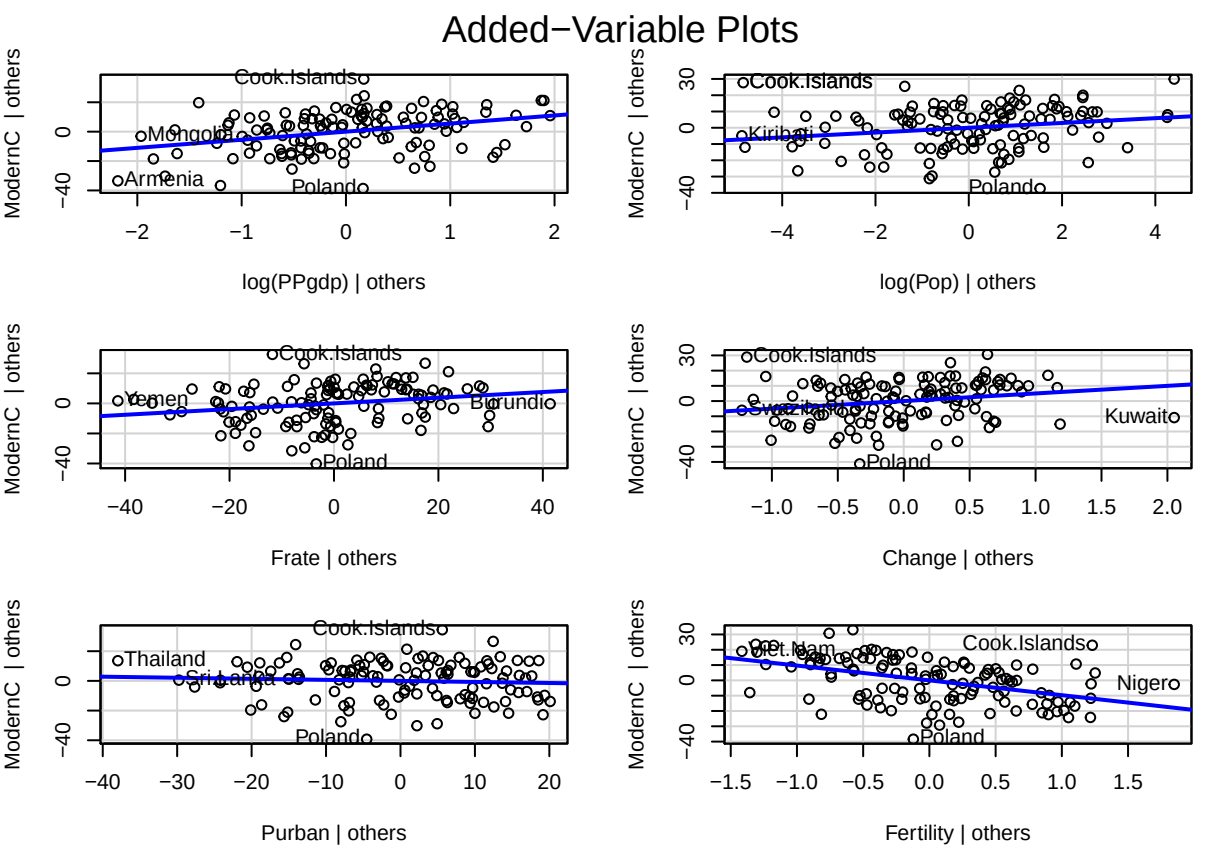
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

The plots show some improvements. Particularly, China and India no longer have large leverages in Residuals vs Leverage plot. Yet Cook Island and Kuwait now becomem relative outliers, yet still pretty close. The Residual vs Fitted also centered more around the fitted, however a few outliers. Normality doesn't show significant improve in Normal Q-Q plot.

```
par(mfrow = c(2, 2))
plot(trns, ask = F)
```




```
avPlots(trns)
```

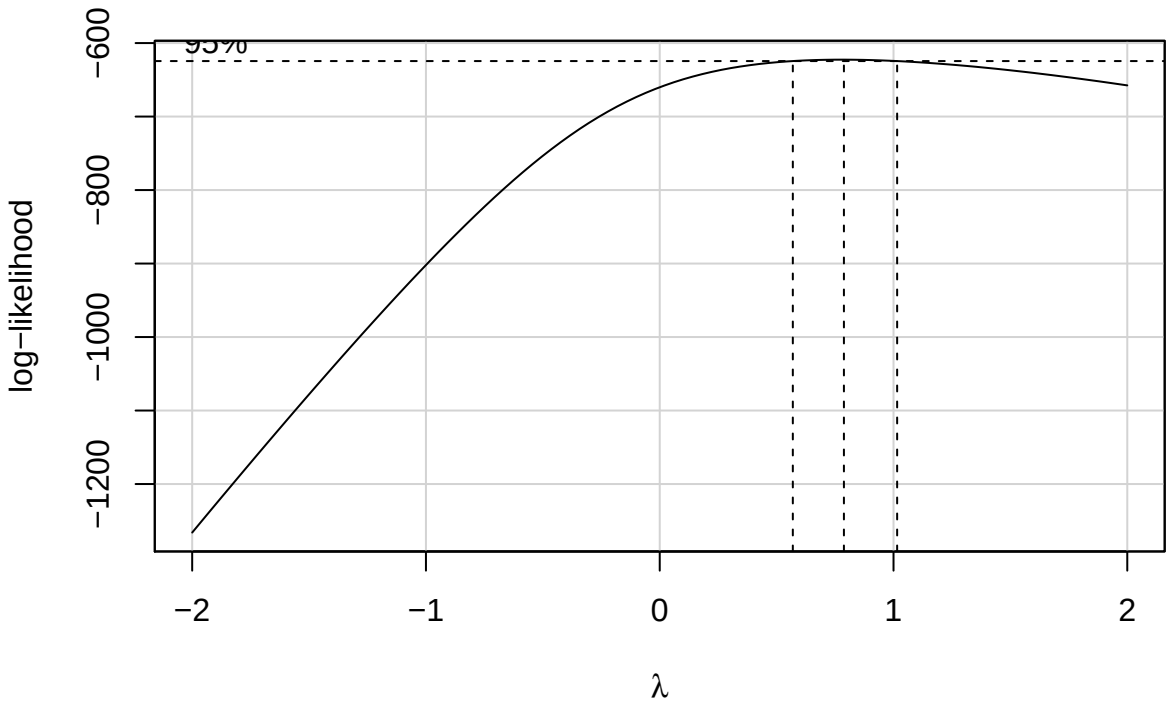


9. Start by finding the best transformation of the response and then find transformations of the predictors.

Do you end up with a different model than in 8?

The boxCox method shows that lambda value is still close to one when running transformation of the response.Thus no transformation needed for the response.We will perform log-transformation on two predictors. In the end, we end up with the same model in 8.

```
boxCox(lm(ModernC~Pop+PPgdp+Frate+Change+Purban+Fertility, data=UN3))
```



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. Since China and India, the most influential points previously have lost their influence after log-transformation. There are no longer influential points in the data.

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!

```
confco <- data.frame(LowB = confint.lm(trns)[, 1],HighB = confint.lm(trns)[, 2], Betas = trns$coefficie

kable(confco, col.names = c("Coef. Est.", "2.5%", "97.5%"),digits=2,caption ="coefficients with 95% con
```

Table 2: coefficients with 95% confidence interval

	Coef. Est.	2.5%	97.5%
(Intercept)	-34.54	21.80	-6.37
log(PPgdp)	2.72	8.29	5.51
log(Pop)	0.23	2.72	1.47
Frate	0.04	0.34	0.19
Change	0.88	9.11	4.99
Purban	-0.26	0.12	-0.07
Fertility	-13.17	-6.18	-9.68

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

My final model uses five predictors: PPgdp, Pop, Frate, Change and Fertility. log-transformations are performed on two predictors: PPgdp and Pop. In my final deletion choice, I deleted Purban as a predictors because after performing stargazer as model comparison, the p-value suggests that Purban is not statistically significant as a predictor. After the deletion, the summary table shows that every predictors are strongly statistically significant with p-values less than 0.05. R-squared value also improves from 0.607 to 0.609, which suggests that the new final model fits the data slightly better. In the final model, only Fertility has negative relation with the response, every other predictors have positive relation with the response. Intuitively this makes sense, becauase as GDP and population grows, people are more interested in personal development and better educated about the usage and benefits of contraception.

```
#summary(trns)
pre<-lm(ModernC~log(PPgdp)+log(Pop)+Frate+Change+Fertility+Purban,data=UN3clean)
Final<-lm(ModernC~log(PPgdp)+log(Pop)+Frate+Change+Fertility,data=UN3clean)
#stargazer(pre,Final,type="latex",single.row = TRUE)
```

Table 3:

	Dependent variable:	
	ModernC	
	(1)	(2)
log(PPgdp)	5.507*** (1.405)	4.859*** (1.082)
log(Pop)	1.472** (0.629)	1.441** (0.626)
Frate	0.189** (0.077)	0.200*** (0.076)
Change	4.993** (2.077)	4.698** (2.033)
Fertility	−9.676*** (1.766)	−9.278*** (1.675)
Purban	−0.071 (0.098)	
Constant	−6.370 (14.224)	−5.763 (14.172)
Observations	125	125
R^2	0.626	0.624
Adjusted R^2	0.607	0.609
Residual Std. Error	13.443 (df = 118)	13.416 (df = 119)
F Statistic	32.912*** (df = 6; 118)	39.547*** (df = 5; 119)
Note:		*p<0.1; **p<0.05; ***p<0.01

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.*

Notice that by definition, we have

$$\hat{e}_Y = \hat{\beta}_1 \hat{e}_{X_i} + \hat{\beta}_0$$

Mutiply both sides by $\mathbf{1}_n^T$, we get

$$\mathbf{1}_n^T \hat{e} = \mathbf{1}_n^T (I-H)Y = \mathbf{1}_n^T \hat{\beta}_0 + (I-H)\hat{\beta}_1 X_i$$

.We can substitude $\hat{\beta}_1$ using its definition $\hat{\beta}_1 = ((I-H)x_i)^T((I-H)x_i)]^{-1}((I-H)x_i)^T(I-H)Y$ and then

multiply both sides by x_i^T, then the equality becomes

$$(I-H)x_i^TY = \sum x_i\hat{\beta}_0 + x_i^T(I-H)x_i[x_i^T(I-H)x_i]^{-1}x_i^T(I-H)Y = \sum x_i\hat{\beta}_0 + x_i^T(I-H)Y$$

.Therefore

$$\sum x_i\hat{\beta}_0 = 0$$

, i.e., the intercept is zero.

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 a manually constructed added variable plot for one of the predictors in Ex. 10 is the same as the estimate from your model.

According to the problem, let X_j be the predictor Change. We will construct e_X using other predictors. The table shows that the slope in this manually constructed added variable plot for Change in my final model in Ex.10 is the same as the estimate e_X.

```
e_Y <- residuals(lm(ModernC~log(Pop)+log(PPgdp)+Fertility+Frate, UN3clean))
e_X <- residuals(lm(Change~log(Pop)+log(PPgdp)+Fertility+Frate, UN3clean))
eyex <- lm(e_Y~e_X)


coef_ex = summary(trns)$coefficients['Change',c('Estimate','t value')]
coef_final = summary(eyex)$coefficients['e_X', c('Estimate', 't value')]
co<-rbind(coef_ex,coef_final)
kable(round(co,2))
```

	Estimate	t value
coef_ex	4.99	2.40
coef_final	4.70	2.35