HW2 STA521 Fall18

[Your Name Here, netid and github username here]

Due September 23, 2018 5pm

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

Exploratory Data Analysis

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 Purban - Percent of population that is urban, 2001

```
## Warning: package 'alr3' was built under R version 3.4.4

## Loading required package: car

## Warning: package 'car' was built under R version 3.4.4

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.4.4

## Warning: package 'GGally' was built under R version 3.4.4

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.4.4

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout

## Warning: package 'dplyr' was built under R version 3.4.4

##
## 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

## Warning: package 'stargazer' was built under R version 3.4.4

##
## 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

## Warning: package 'MASS' was built under R version 3.4.4

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## The following object is masked from 'package:plotly':
##
##     select

## The following object is masked from 'package:alr3':
##
##     forbes
```

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

The variables with missing values are described below in the table. Additionally, all of the variables are quantitative. However, some are integers which sometimes suggests a mis-classified factor variable. To check this, I also counted to the number of unique values for each variable. Upon inspection of results, I decided that for the sake of interpretability they should remain integer values rather than factors.

```r
Variable.Types = unlist(lapply(UN3, class))
Num.Unique = unlist(lapply(UN3, function(x) length(unique(x)) ))
Num.Missing = unlist(lapply(UN3, function(x) sum(is.na(x)) ))

kable(cbind(Variable.Types, Num.Unique, Num.Missing), caption = "Data Summary")
```

Table 1: Data Summary

	Variable.Types	Num.Unique	Num.Missing
ModernC	integer	74	58
Change	numeric	160	1
PPgdp	integer	196	9
Frate	integer	63	43
Pop	numeric	207	2
Fertility	numeric	154	10
Purban	integer	86	0

```
kable(summary(UN3))
```

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

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

```
Mean = unlist(lapply(UN3, mean, na.rm = TRUE))
SD = unlist(lapply(UN3, sd, na.rm = TRUE))

kable(cbind(Mean, SD), digits = 2, align = 'c')
```

	Mean	SD
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 plotshighlighting 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?

It looks like population has some outliers. Most of the remaining variables seem to have a linear relationship (Fertility being negatively related). Frate does not appear to have much of an impact. PPGDP has a nonlinear, albeit noticeable effect on Modern Contraception usage rate.

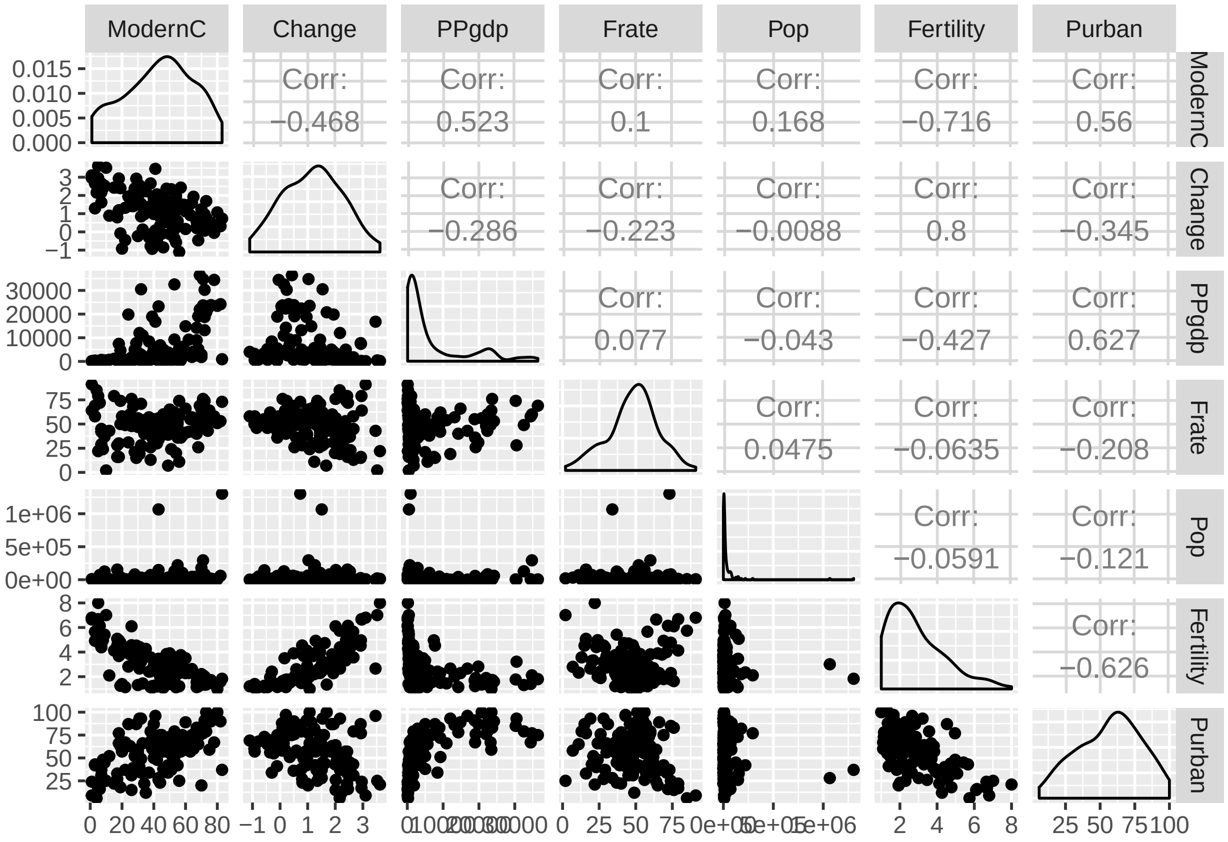
Population has a very large range. It's so large in fact that no relationship can be discerned between it and the other variables from the scatterplots though there does appear to be a positive correlation between ModernC and population. In most of the scatterplots that include population, it appears that there are two points that are outliers (India and China). These outliers may need to be removed later or may be addressed by a variable transformation depending on the results from upcoming diagnostic plots.

To a lesser degree, Fertility and PP GDP also appear to be positively skewed and seem to show nonlinear relationships in their scatterplots.

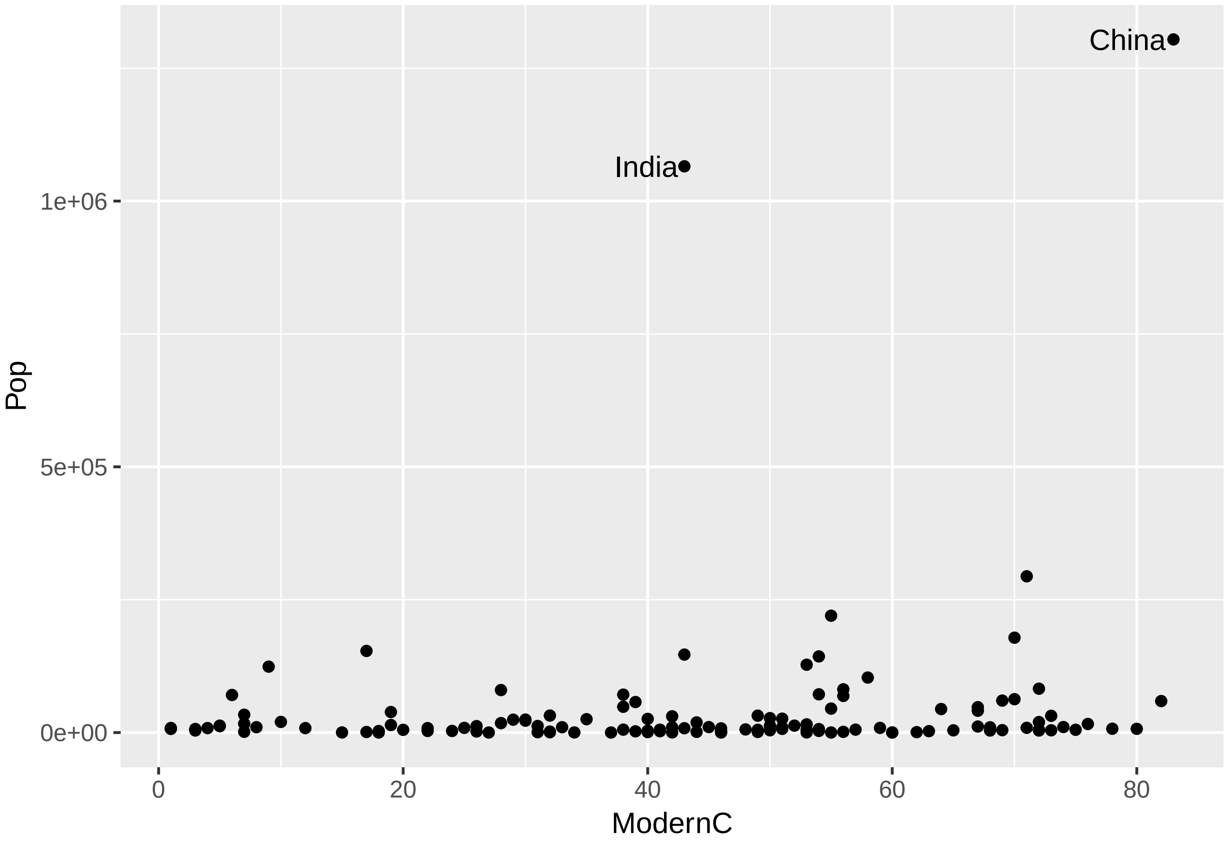
Many of the predictors are also related to each other. In some cases, the mean functions for the plots of predictor versus predictor appear to be linear; in other cases, they are not linear.

NOTE: I added a "na.omit" statement before anything was plotted. This is because the observations with at least one missing value will not be used in the model building below and I wanted the plots to reflect the data that was going to be modeled.

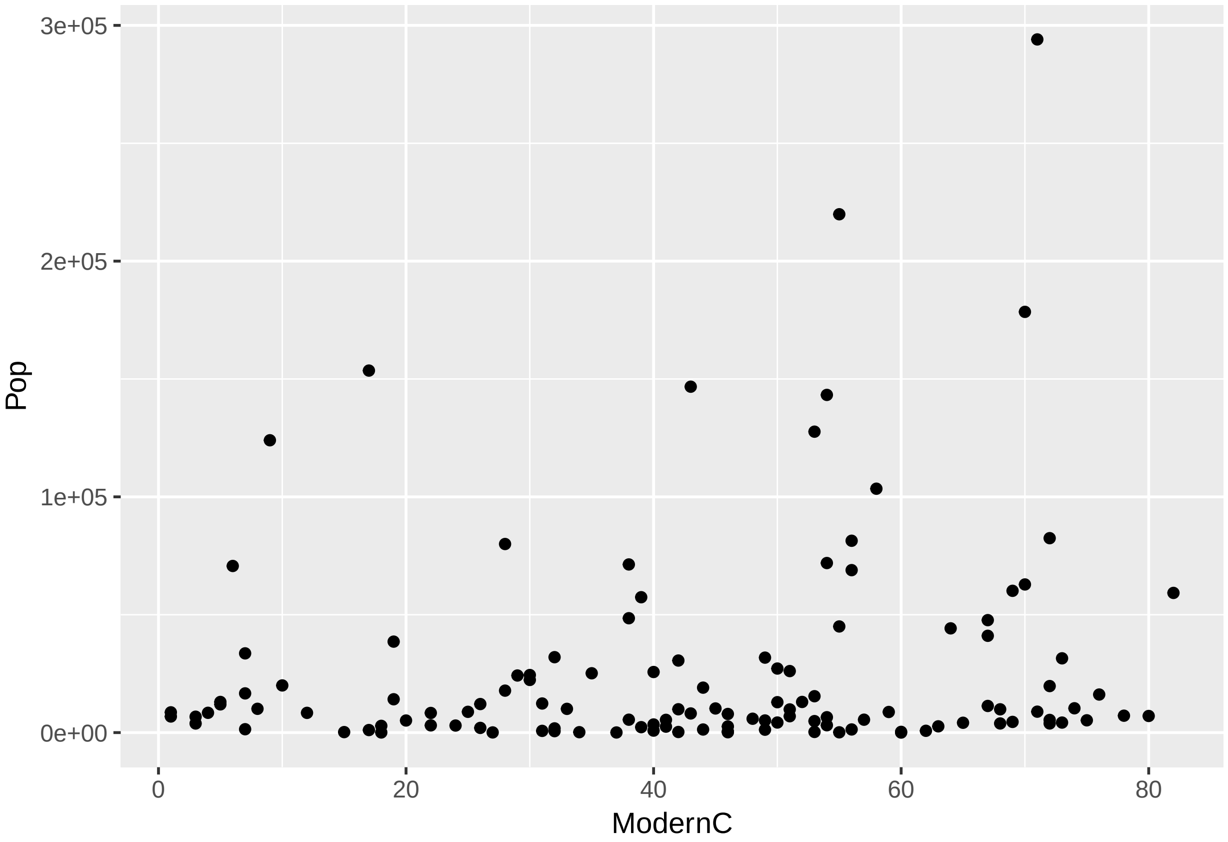
```
pm <- ggpairs(na.omit(UN3), progress = FALSE)
print(pm)
```



```
#Creates a great interactive 3D scatterplot, but doesn't appear in PDF
# plot_ly(x=UN3$Fertility, y=UN3$Change, z=UN3$ModernC, type="scatter3d", mode="markers",
# xlab = "PerPersonGDP", ylab = "Perc_Urban", zlab = "ModernContraception")

par(mfrow=c(1,2))
UN3.filter = UN3 %>% filter(!rownames(UN3) %in% c("India","China"))

qplot(data=na.omit(UN3),ModernC,Pop) + geom_text(aes(label=ifelse((Pop>800000),rownames(na.omit(UN3)),"
```



```r
qplot(data=na.omit(UN3.filter),ModernC,Pop)
```



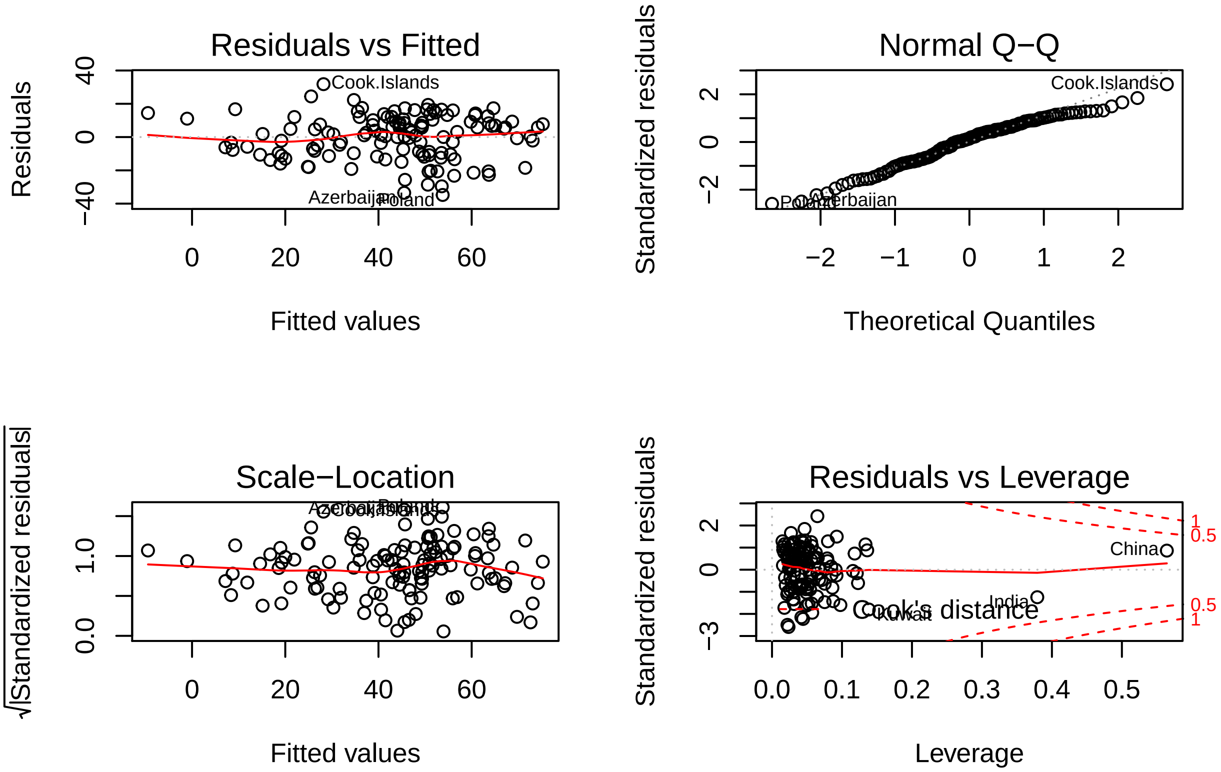
Model Fitting

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?

The model is fitting only 125 observations out of the original 210. 85 observations were excluded from this analysis because at least one of the variables was missing.

Q-Q Plot -Does not follow normal distribution exactly most noticeably because it has heavy tails. The scale-location plot shows that the residuals appear to be homoscedastic, though there are some observations that are far from the fitted values Residuals vs leverage - There are a few points that show a very large leverage. That is, the predicted values for the remaining observations change significantly with the inclusion of these observations

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



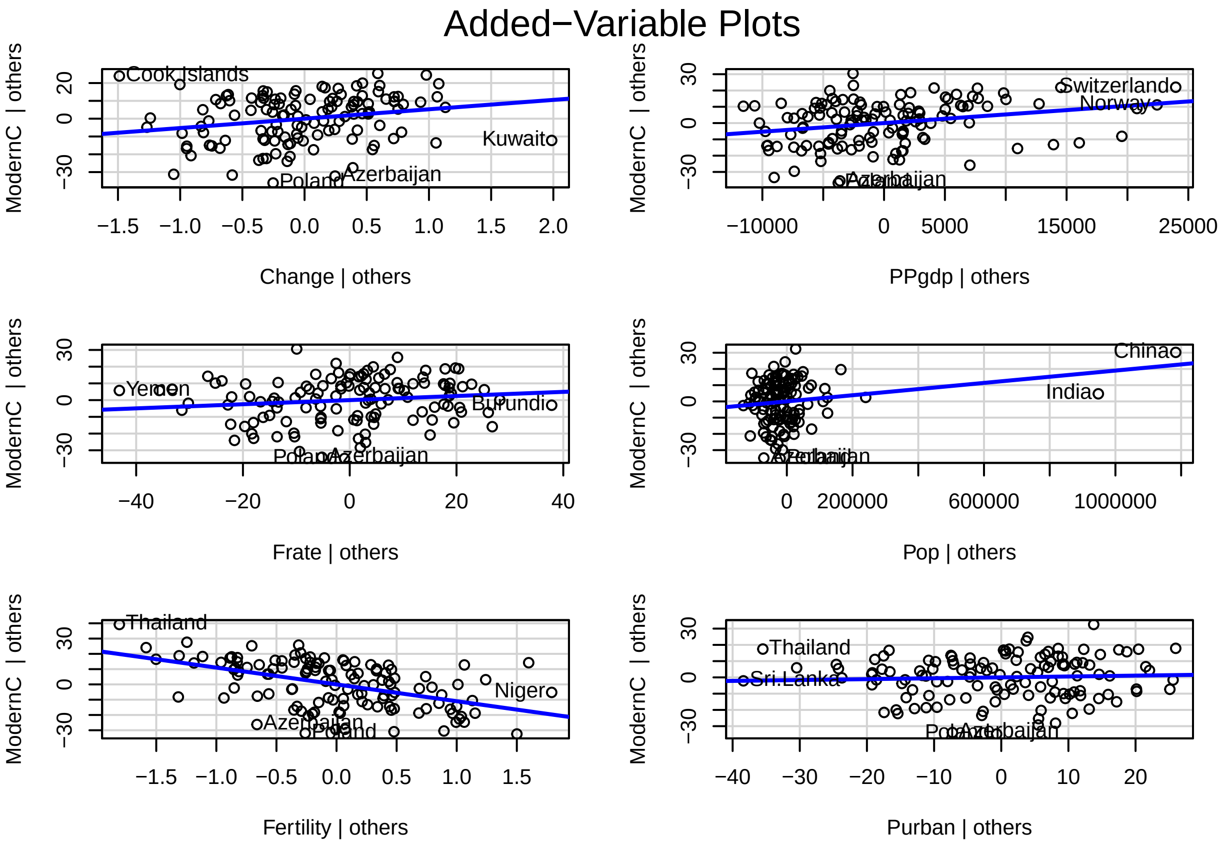
```
# summary(Initial.Pred)
# nobs(Initial.Pred)
```

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?

These added variable plots show the relationship between Y: variability that cannot be explained by the predictors (excluding one variable) and Xi: the variability in that one variable X_i that cannot be explained by the other predictors. Therefore, a relationship between these residuals shows that there is some additional variability that can be accounted for by that X_i. The code below performs this procedure for each of the variables in our original model.

It appears that there are several predictors that can account for additional variation beyond the other predictors in the model. Namely, the av plot for Fertility shows that it is negatively related to ModernC. Some of the other plots suggest that there may be some transformation needed before their impact on ModernC can be completely assessed. Most noticeably, Population has two very large outliers that make it difficult to discern a relationship between it and ModernC. Similarly, PPgdp and, to a lesser degree, Fertility show higher magnitude positive residuals than negative residuals. This would suggest that a transformation may be required.

```
avPlots(lm(ModernC ~ ., data = UN3))
```



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 the results from 3,4, and 5, I decided to include 3 variables in the Box Tidwell method to find the appropriate transformation of the predictors: Fertility, Pop, and PPgdp. When this was conducted there were no variables that were statistically significant, however Population and PPgdp still clearly needed transformation.

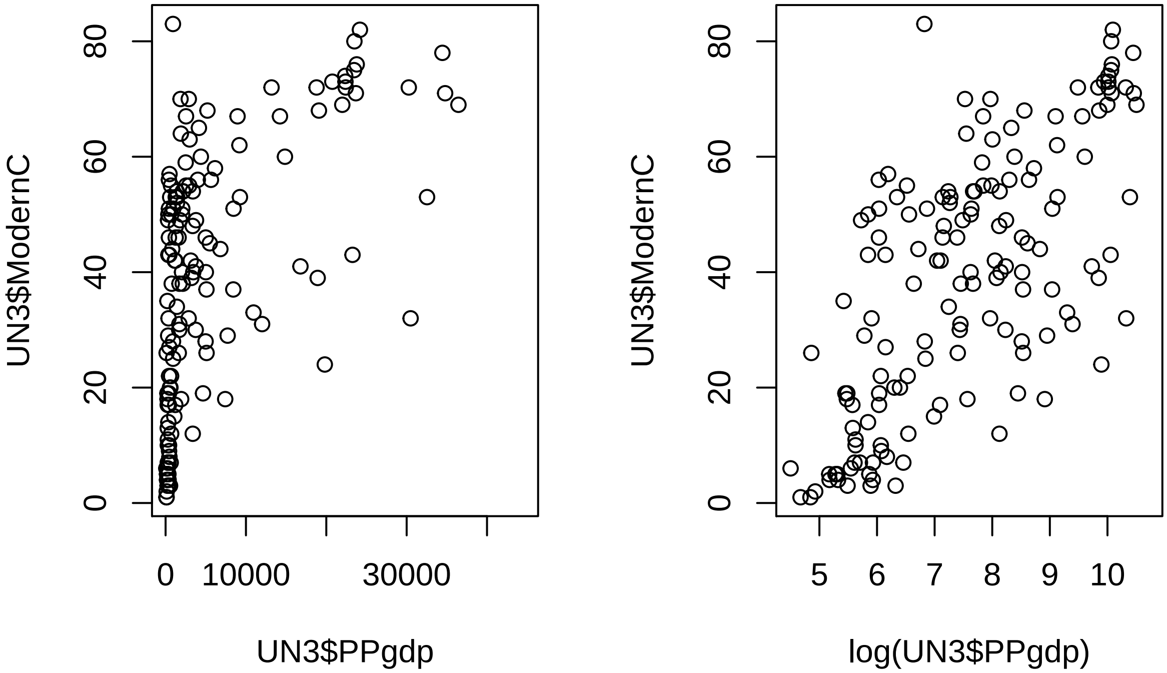
The Box-Tidwell estimate for population was not statistically significant, though, as mentioned in the Applied Linear Regression text, variables with such a large range are often transformed using log. Therefore, based on the previous visualizations and intuition about population, it follows that the log of population would be an appropriate transformation for this model.

Additionally, for PPgdp, we notice that its p-value for the estimate in the Box-Tidwell results is not statistically significant by traditional standards (i.e. p>.05). However, this model did not allow us to contribute our prior beliefs to this estimate which, especially considering other economic studies, tend to favor a log transform of this variable. Additionally, the estimate itself is quite close to 0 which would suggest a log transformation. Therefore, a log transformation was also done to this variable.

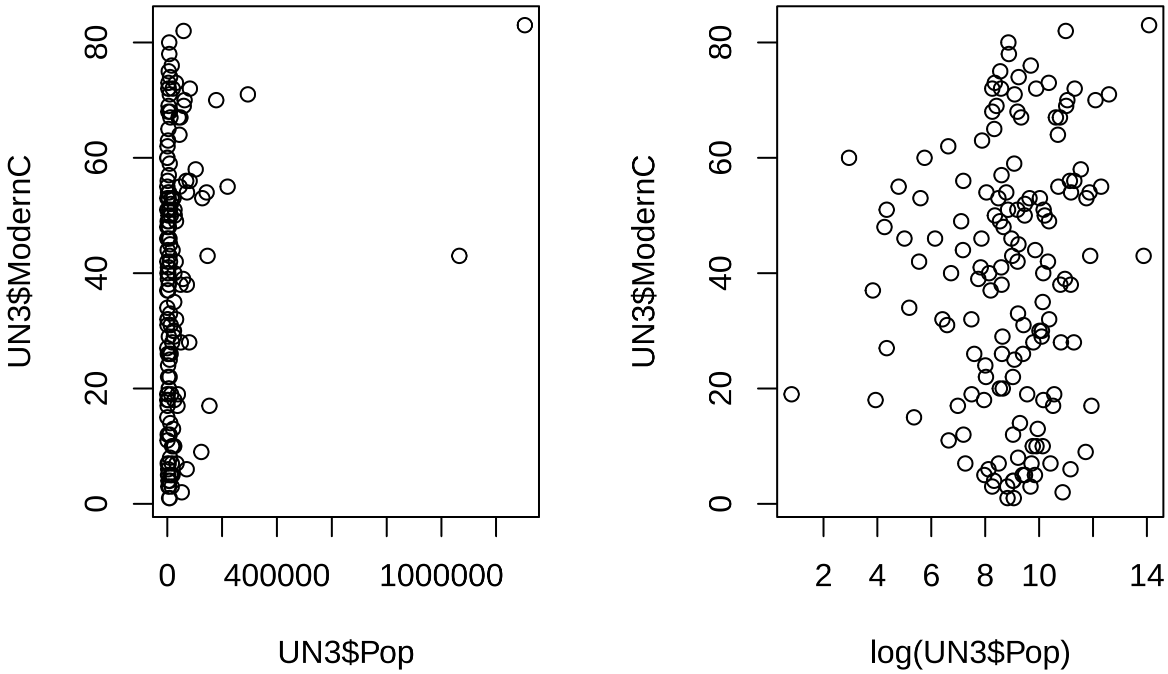
```
#Transform Variables to be nonnegative
BoxTid.UN3 = na.omit(UN3)
BoxTid.UN3$Change = BoxTid.UN3$Change + 1.1 + 1

boxTidwell(ModernC~Pop+PPgdp+Fertility, other.x = ~Frate+Change+Purban, data=na.omit(BoxTid.UN3), max.i
```

```
##           MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop            0.374984             -0.9042   0.3659
## PPgdp          -0.035767             -1.2324   0.2178
## Fertility       1.346874             -1.7985   0.0721 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  22
```

```
## I begin by looking at the plots to find
#Check PPgdp
par(mfrow=c(1,2))
plot(UN3$PPgdp, UN3$ModernC)
plot(log(UN3$PPgdp), UN3$ModernC)
```



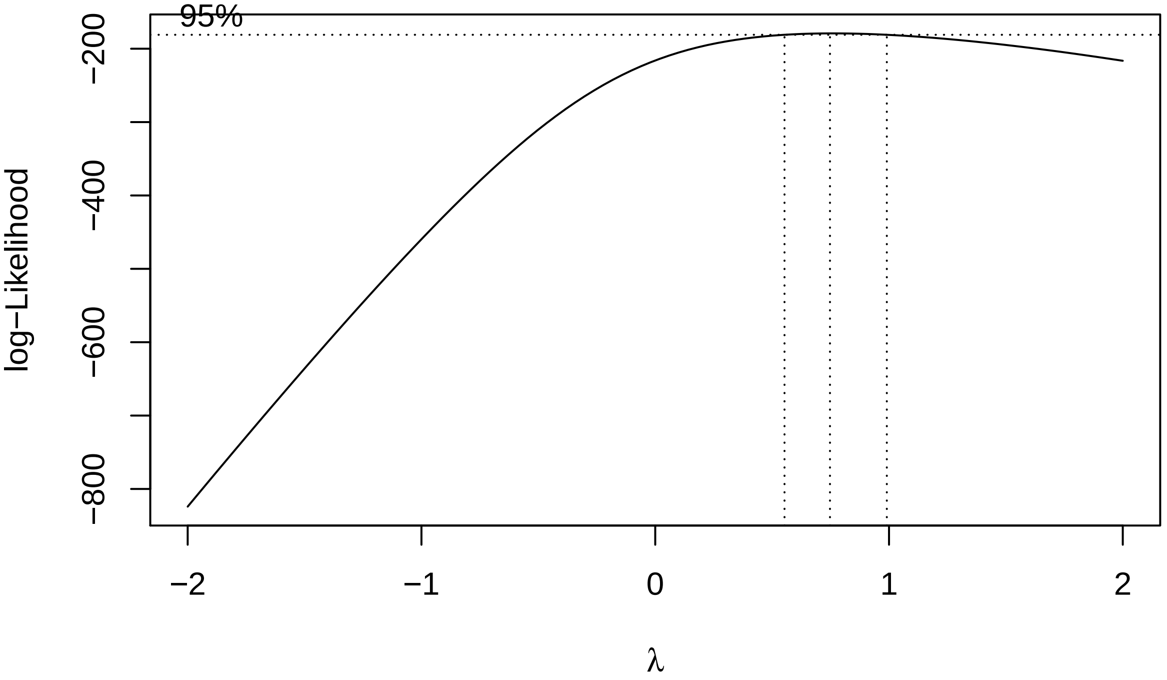
```
#Check Population - remove outliers
plot(UN3$Pop, UN3$ModernC)
plot(log(UN3$Pop), UN3$ModernC)
```



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

Based on the results of the Box-Cox test, it appears that though the confidence interval is not centered at 1, it is close to the interval. While in most circumstances it may be preferable to assume that the interval is close enough to one so that no transformations would be required. For the sake of comparison later on, I will continue with the transformation.

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



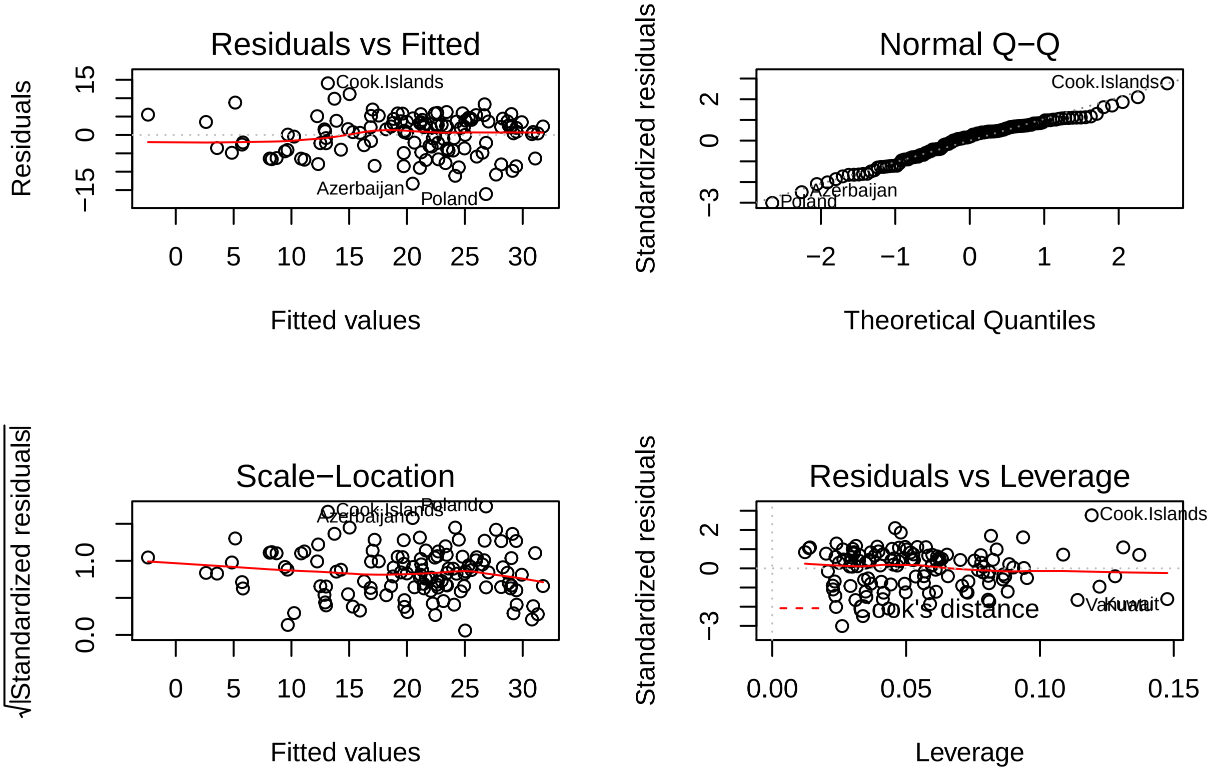
```
lambda = BC$x[which.max(BC$y)]
```

```r
UN3_BC.Transform <- UN3 %>% mutate(ModernC_t = (ModernC^lambda-1)/lambda)# %>% select(ModernC_t, Change
rownames(UN3_BC.Transform) = rownames(UN3)

Iteration2_transform = lm(ModernC_t~log(Pop)+log(PPgdp)+Frate+Change+Purban+Fertility, data=UN3_BC.Trans
# summary(Iteration2_transform)

# Untransformed outcome
Iteration2= lm(ModernC~log(Pop)+log(PPgdp)+Frate+Change+Purban+Fertility, data=UN3_BC.Transform)
# summary(Iteration2)
```

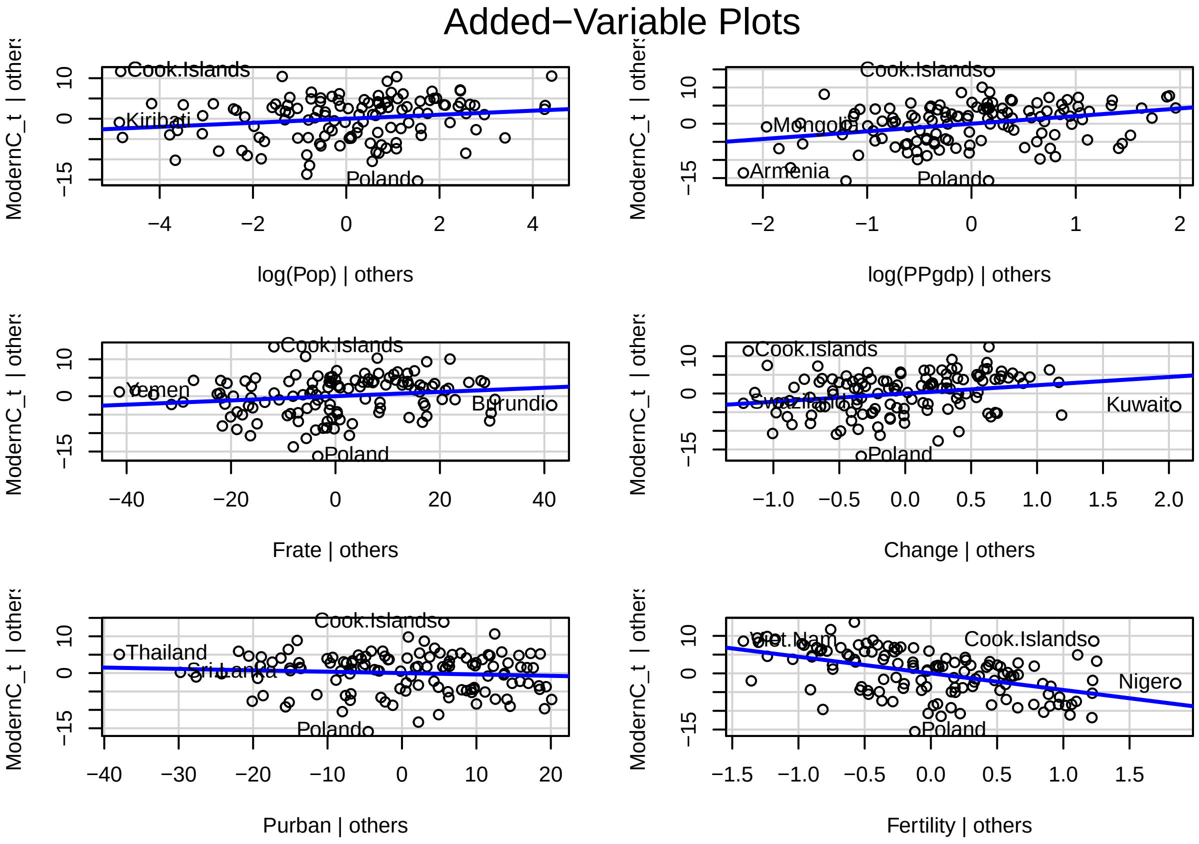
8. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

Overall, the diagnostic plots seem to show results that are more consistent with our assumptions about OLS regression. Notably, the log transform of population reduced the leverage India and China and there do not appear to be any other worrisome high leverage points. The QQ plot shows that there is slightly less spread on this distribution than a typical normal distribution, though this may be because of the structure of the response variable. That is, the response variable has a strict range from 0 to 100 so a more dispersed normal distribution would not be possible. Additionally, the Scale-Location and Residuals vs Fitted plots (with the fitted lines) appear to show residuals that are more consistent with our assumptions about OLS regression.

```r
# Iteration2 = lm(ModernC~Fertility+log(Pop)+log(PPgdp)+Frate+Change+Purban, data=UN3)
# plot(Iteration2, ask=F)
# summary(Iteration2)

Iteration2_transform = lm(ModernC_t~log(Pop)+log(PPgdp)+Frate+Change+Purban+Fertility, data=UN3_BC.Trans
par(mfrow=c(2,2))
plot(Iteration2_transform, ask=F)
```



```
avPlots(Iteration2_transform)
```



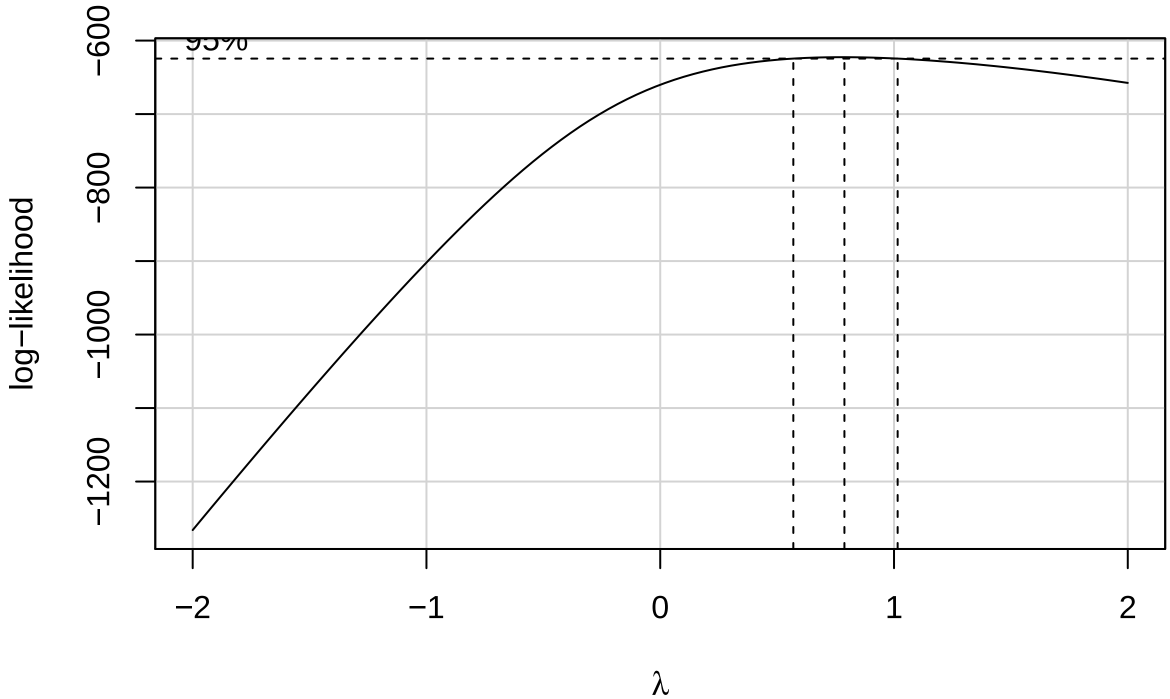
```
# summary(Iteration2_transform)


### Add stargazer table to show comparison to original model's p-values
```

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?

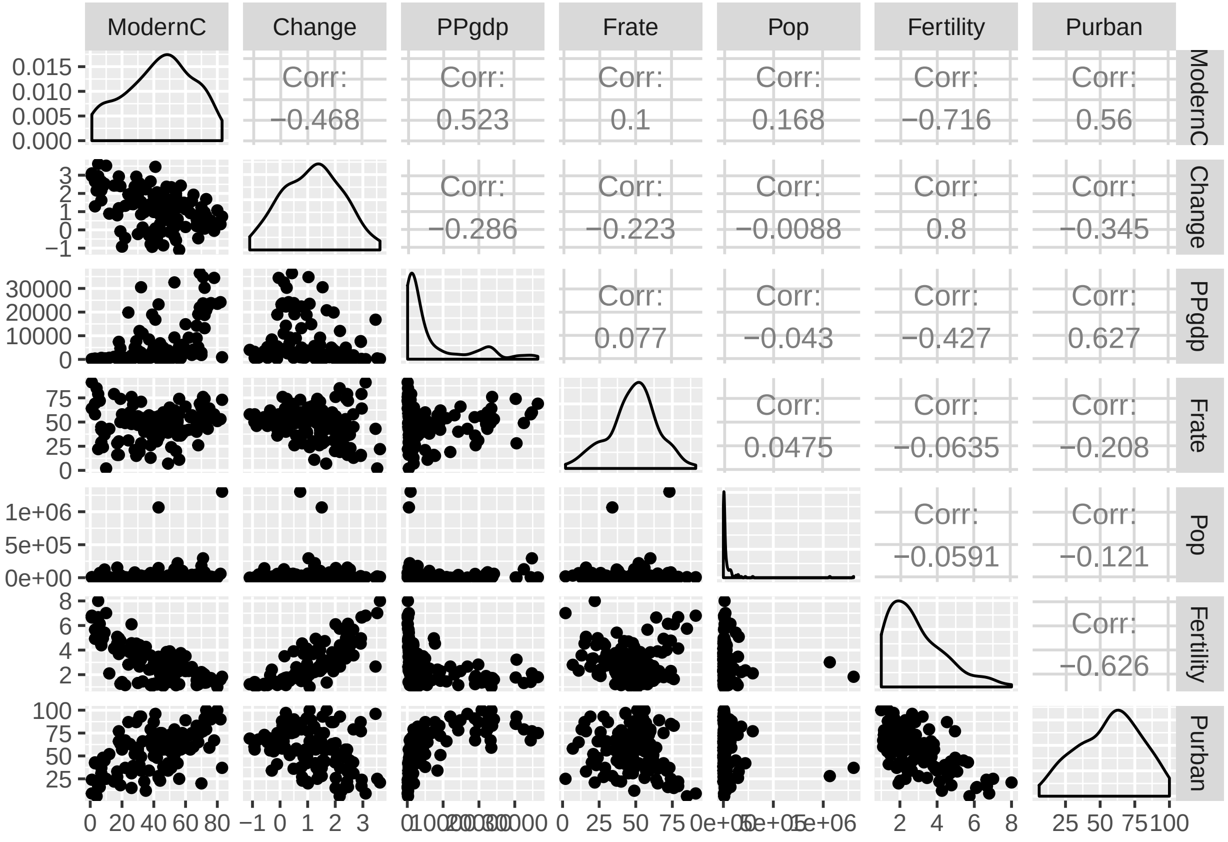
When running the Box-Cox transformation of the response variable first, there is a subtle change in the confidence interval for the lambda. However, this minor change in the confidence interval was significant because the interval contained 1. Therefore there was no change that needed to be made to the response variable and the predictors could be transformed based on the same rationale stated in question 6. Diagnostic plots for this model are shown below.

Now it is worth mentioning that when choosing a final model, the decision comes down to two models with identical predictors but slightly different response variables. The two tables are compared in the table below.

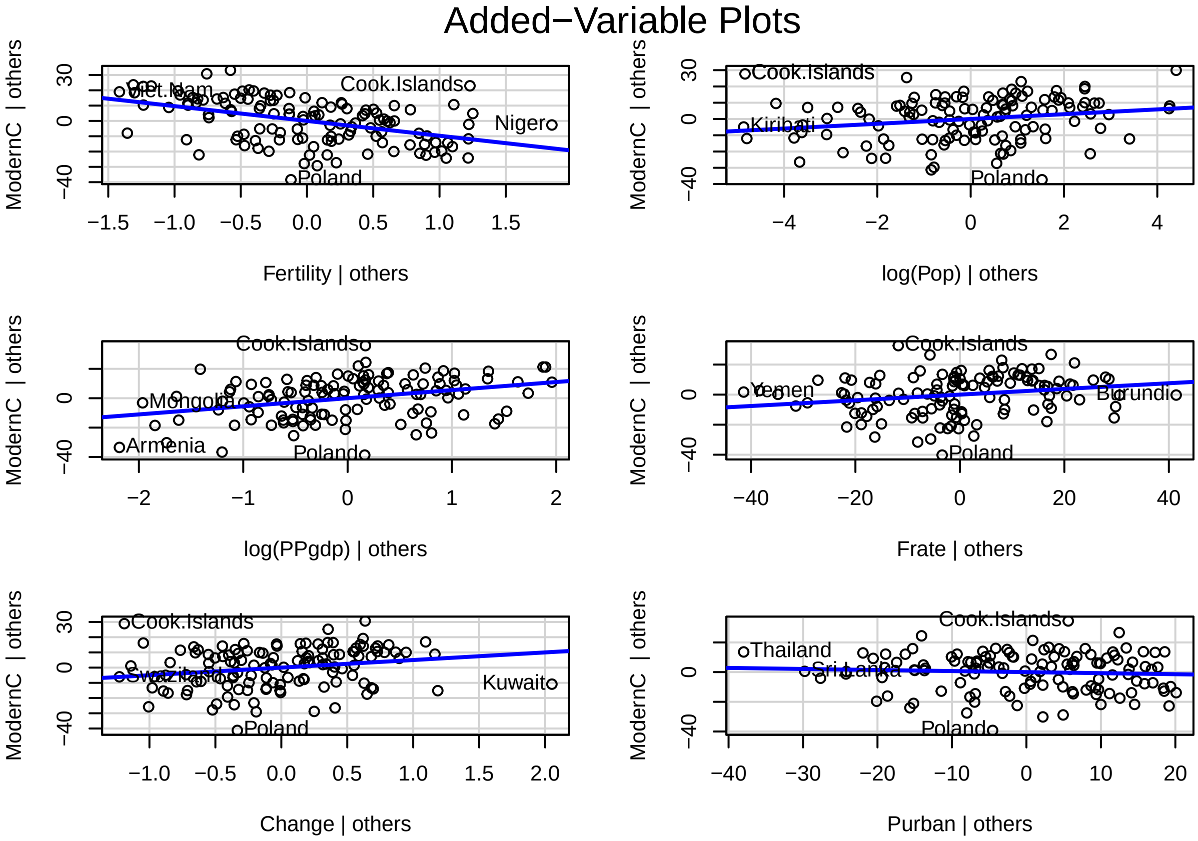
```
# Using boxcox to find the appropriate variable transformation:
boxCox(lm(ModernC~Fertility+Pop+PPgdp+Frate+Change+Purban, data=UN3))
```



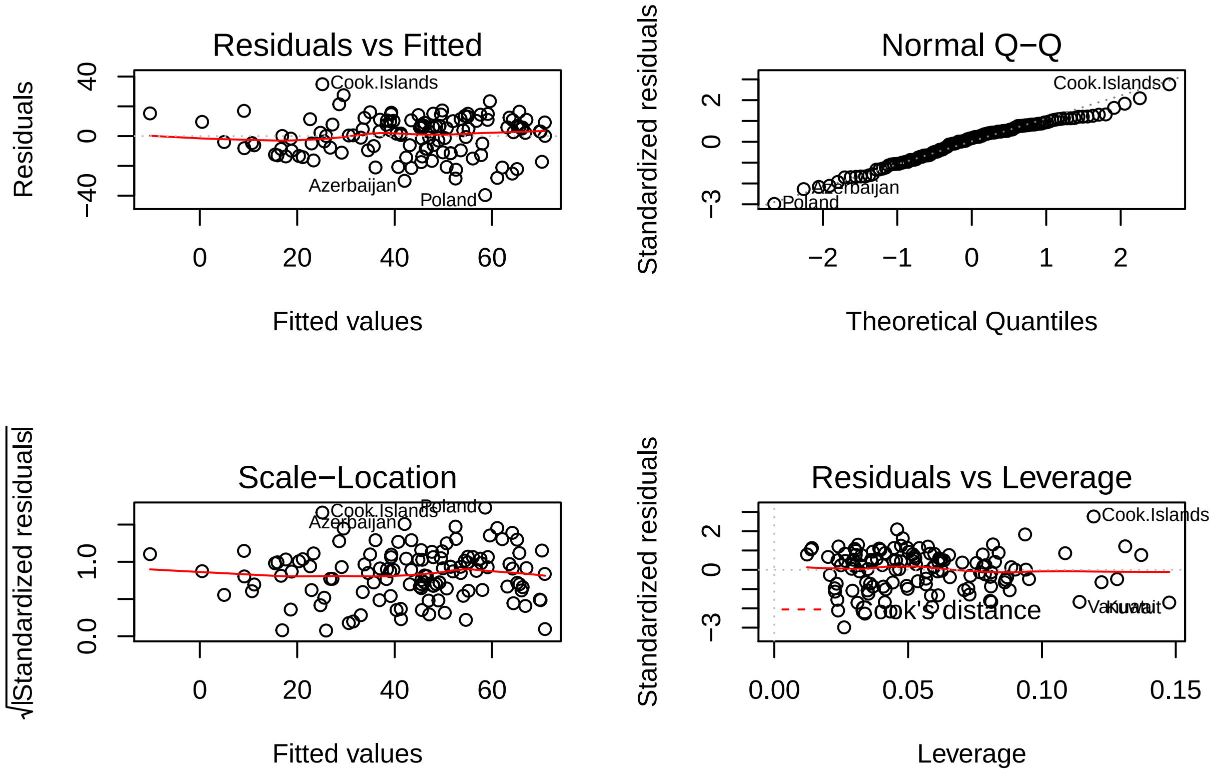
```r
## Make the variable transformation:
pm <- ggpairs(na.omit(UN3), progress = FALSE)
print(pm)
```



```
#added variable plots
avPlots(lm(ModernC ~Fertility+log(Pop)+log(PPgdp)+Frate+Change+Purban, data = UN3))
```



```
## Box Tidwell of Predictors
# BoxTid.UN3_t = na.omit(UN3)
# BoxTid.UN3_t$Change = BoxTid.UN3_t$Change + 1.1 + 1
#
# boxTidwell(ModernC~log(Pop)+log(PPgdp)+Fertility, other.x = ~Frate+Change+Purban, data=UN3_t, max.ite
# boxTidwell(ModernC~Pop+PPgdp+Fertility, other.x = ~Frate+Change+Purban, data=UN3_t, max.iter = 100)

# Transforming Fertility
# lambda = 1.421
# UN3_t = UN3_t %>% mutate(Fertility_t = (Fertility^lambda-1)/lambda) %>% select(ModernC_t, Change, PPg
# rownames(UN3_t) <- rownames(UN3)


Iteration3 = lm(ModernC~Fertility+log(Pop)+log(PPgdp)+Frate+Change+Purban, data=UN3)
par(mfrow=c(2,2))
plot(Iteration3, ask=F)
```



```
##
# 1. Show that the response variable does not need to be transformed
# 2. Show that the transformed predictor variables do not require further transformation
# 3. Compare the two models
```

```
stargazer(Iteration2_transform,Iteration3, header = FALSE, type = "latex", single.row = TRUE, table.pla
```

Table 4:

	Dependent variable:	
	ModernC_t	ModernC
	(1)	(2)
log(Pop)	0.497* (0.253)	1.472** (0.629)
log(PPgdp)	2.109*** (0.566)	5.507*** (1.405)
Frate	0.057* (0.031)	0.189** (0.077)
Change	2.209*** (0.837)	4.993** (2.077)
Purban	−0.038 (0.039)	−0.071 (0.098)
Fertility	−4.433*** (0.711)	−9.676*** (1.766)
Constant	9.026 (5.845)	4.115 (14.509)
Observations	125	125
R^2	0.644	0.626
Adjusted R^2	0.626	0.607
Residual Std. Error (df = 118)	5.416	13.443
F Statistic (df = 6; 118)	35.545***	32.912***
Note:		*p<0.1; **p<0.05; ***p<0.01

Comparing the results of the two models, it seems that the model with the transformed response variable is marginally better than the model where we chose to keep ModelC as is. However, it is also worth taking into account that the coefficients are now in terms of the transformed response variable and therefore lose a lot of interpretability. Therefore, it seems to make the most sense to use this latest iteration with the untransformed ModernC as the final model.

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.

We noticed from prior added plots of population that India and China were both influential points because of their extraordinarily large population. These two countries had a noticeable impact on the predictions of the other observations as seen through their high leverage values, though they did not have a very large Cook's distance which also takes into account their residuals. However, this observation occurred before the variable transformation to population and with that update they did not appear to be influential points anymore. To confirm there are no new influential points, I used the Bonferroni test which does not show any statistically significant outliers.

```
abs.ti = abs(rstudent(Iteration3))
pval= 2*(1- pt(abs.ti, Iteration3$df - 1))
min(pval) < .05/nrow(Iteration3)
```

```
## logical(0)
```

```
sum(pval < .05/nrow(Iteration3))
```

```
## [1] 0
```

```
# UN3.rm_IC = UN3 %>% filter(!rownames(UN3)%in% c("India","China"))
# Iteration3.rm_IC = lm(ModernC~log(PPgdp)+log(Pop)+Frate+Change+Fertility, data=na.omit(UN3.rm_IC))

# plot(Iteration3, ask=F)
```

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!

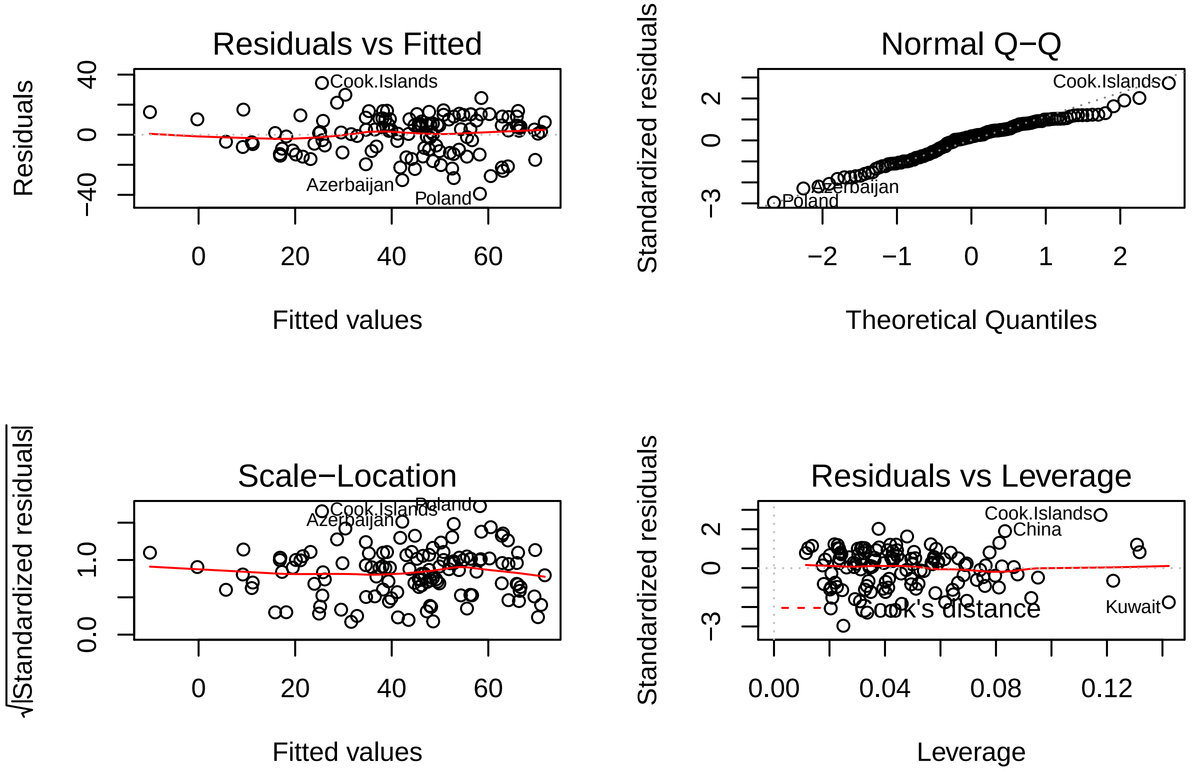
Before the confidence intervals are constructed, I created a comparison of models to arrive at a final model based on the information above. I developed 4 models with the same outcome variable (untransformed Modern C) and different predictors as summarized below:
1) log(PPgdp), log(Pop), Frate
2) log(PPgdp), log(Pop), Frate, Change
3) log(PPgdp), log(Pop), Frate, Change, Fertility
4) log(PPgdp), log(Pop), Frate, Change, Fertility, Purban

```
Iteration2 = lm(ModernC~log(Fertility)+Pop+log(PPgdp)+Frate+Change+Purban, data=UN3)

ModernC1.lm = lm(ModernC~log(PPgdp)+log(Pop)+Frate, data=na.omit(UN3))
ModernC2.lm = lm(ModernC~log(PPgdp)+log(Pop)+Frate+Change, data=na.omit(UN3))
ModernC3.lm = lm(ModernC~log(PPgdp)+log(Pop)+Frate+Change+Fertility, data=na.omit(UN3))
ModernC4.lm = lm(ModernC~log(PPgdp)+log(Pop)+Frate+Change+Fertility+Purban, data=na.omit(UN3))

# Summarize Models
Model.Comp = anova(ModernC1.lm, ModernC2.lm, ModernC3.lm, ModernC4.lm)
kable(Model.Comp, digits = 2, align = 'c', format='markdown' )
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
121	28549.15	NA	NA	NA	NA
120	26943.29	1	1605.86	8.89	0.00
119	21420.03	1	5523.26	30.56	0.00
118	21325.02	1	95.02	0.53	0.47

```r
#Build Final Model
FinalModel = lm(ModernC~log(PPgdp)+log(Pop)+Frate+Change+Fertility, data=na.omit(UN3))
par(mfrow=c(2,2))
plot(FinalModel, ask = F)
```



The ANOVA shows that Purban does not provide enough additional information about the outcome variable (which evident through the change in SSE) to justify its inclusion in the model. Based on these results, it seems reasonable that we exclude Purban from the list of predictors in the final model.

Now, with this final model we can construct a 95% confidence interval for each predictor where the numbers represent change in the same units as Modern C.

```r
CI = confint(FinalModel)
CI["log(Pop)",] = CI["log(Pop)",] * log(1.1)
CI["log(PPgdp)",] = CI["log(PPgdp)",] * log(1.1)

Estimate = coef(FinalModel)
Estimate["log(PPgdp)"] = Estimate["log(PPgdp)"] * log(1.1)
Estimate["log(PPgdp)"] = Estimate["log(PPgdp)"] * log(1.1)

table = cbind(Estimate, CI)
```

```r
rownames(table)[2:3] = c("PPgdp (10% increase)", "Pop (10% increase)")
# Add mean estimates for coefficients
kable(table, digits = 2, align = 'c', format='markdown' )
```

	Estimate	2.5 %	97.5 %
(Intercept)	4.10	-24.57	32.77
PPgdp (10% increase)	0.04	0.26	0.67
Pop (10% increase)	1.44	0.02	0.26
Frate	0.20	0.05	0.35
Change	4.70	0.67	8.72
Fertility	-9.28	-12.60	-5.96

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

```r
stargazer(FinalModel, header = F, type = "latex", single.row = T, table.placement = "h")
```

Table 7:

	Dependent variable:
	ModernC
log(PPgdp)	4.859*** (1.082)
log(Pop)	1.441** (0.626)
Frate	0.200*** (0.076)
Change	4.698** (2.033)
Fertility	−9.278*** (1.675)
Constant	4.102 (14.480)
Observations	125
R^2	0.624
Adjusted R^2	0.609
Residual Std. Error	13.416 (df = 119)
F Statistic	39.547*** (df = 5; 119)
Note:	*p<0.1; **p<0.05; ***p<0.01

The modeling process showed that Population, PPgdp, Frate, Change and Fertility have a statistically significant impact on the prediction of national usage rate of Modern Contraception (ModernC).

However, during our analysis, we found that PUrban does not have a statistically significant impact on ModernC. This means that, when taking all other variables into account, PUrban did not provide any insights beyond that which was already provided by the other variables. We notice that of these other variables, Fertility had a negative relationship with ModernC. Our estimates show that for every one unit increase in Fertility, we would expect on average a 9.3 unit decrease in Modern C.

On the other hand, Frate, Change, Population and PPgdp showed a positive relationship with ModernC. Following a similar pattern to Fertility, one unit increases in Change and Frate showed 4.7 and .2 unit increases in Modern C. Population and PPgdp showed a "log-linear relationship" with Modern C which allows us to explain their impact on Modern C as they change proportionately (e.g. 10% increase in population, doubling of PPgdp). Intuitively, this makes sense as we would not expect a 200,000 unit change in population to have the same impact in China as it would in Great Britain. Therefore, we expect a 10% increase in population to increase Modern C by approximately 0.14 units. Similarly, we would expect a 10% increase in PPgdp to increase Modern C by 0.46 units.

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.*

We begin by defining the equation for the added variable scatterplot:_

$$\hat{e}_{(Y)} = \mathbf{1}_n^T\hat{\beta}_0 + \hat{\beta}_1\hat{e}_{X_j}$$

Where X_j represents the jth column of the design matrix that we are using for our added variable plot.

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

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

We can expand $\hat{\beta_1}$ by using its definition in this case as $(X^TX)^{-1}XY$ where X is substituted for $(I-H)X_j$.

$$X_j^T(I-H)Y = X_j^T\mathbf{1}_n^T\hat{\beta}_0 + X_j^T[X_j^T(I-H)X_j]^{-1}X_j^T(I-H)Y(I-H)X_j$$

Multiplying both sides by X_j^T.

$$X_j^T(I-H)Y = X_j^T\mathbf{1}_n^T\hat{\beta}_0 + X_j^T\underbrace{[X_j^T(I-H)X_j]^{-1}}_{scalar}\underbrace{X_j^T(I-H)Y}_{scalar}(I-H)X_j$$

Rearranging the scalars, we can cancel out the inverse of $[X_j^T(I-H)X_j]$.\

$$X_j^T(I-H)Y = \Sigma X_j\hat{\beta}_0 + X_j^T(I-H)Y$$

$$0 = \Sigma X_j\hat{\beta}_0$$

This last statement must then be true when $\beta_0 = 0$ or, in other words, when the intercept of the added variable plot is 0.

We can also proceed by using the definition of $\hat{e}_y$ which is…

$$\hat{e}_Y = (I-H)Y$$

Multiplying both sides by 1_N^T we get

$$\mathbf{1}_N^T\hat{e}_Y = 1_N^T(I-H)Y$$

From what we were given above, we see that $1_n^T(I-H)=0$ so

$$\mathbf{1}_N^T\hat{e}_Y = \Sigma_{i=1}^N\hat{e}_{i,Y} = 0$$

Multiplying both sides by $\frac{1}{n}$ we see that the mean predicted residuals for y is 0.

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.

We first can specify the design matrix and response for this problem. Let the `Xj` be Fertility. Then the remaining `X`'s will be Pop, PPgdp Frate, Change, Purban. With this, we will construct two linear models and find their residuals.

```
Y = lm(ModernC ~ PPgdp + Pop + Frate + Change + Purban, data = na.omit(UN3))
X = lm(Fertility ~ PPgdp + Pop  + Frate + Change + Purban, data =  na.omit(UN3))

av.plot = lm(residuals(Y) ~ residuals(X))
summary(av.plot)$coefficients
```

```
##                    Estimate Std. Error       t value     Pr(>|t|)
## (Intercept)   5.263128e-16   1.189756  4.423702e-16 1.000000e+00
## residuals(X) -1.099843e+01   1.716484 -6.407537e+00 2.845226e-09
```

```
row1 = coef(summary(av.plot))["residuals(X)",c("Estimate","t value")]
row2 = coef(summary(Initial.Pred))["Fertility",c("Estimate","t value")]

kable(data.frame(round(rbind(row1,row2), digits=2), row.names = c("AV Estimate", "Initial Estimate")), 
```

	Estimate	t.value
AV Estimate	-11	-6.41
Initial Estimate	-11	-6.28

We notice that the two estimates are equal from the table. It is also worth noting that though the estimates are the same, the t statistics differ slightly. This is likely because of the different degrees of freedom in the two tests.