HW2 STA521 Fall18

Evan Stump, eas90, eaxstump

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.

```
data(UN3, package="alr3")
help(UN3)
```

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
```

All the variables are quantitative. All the variables have missing data except for Purban.

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

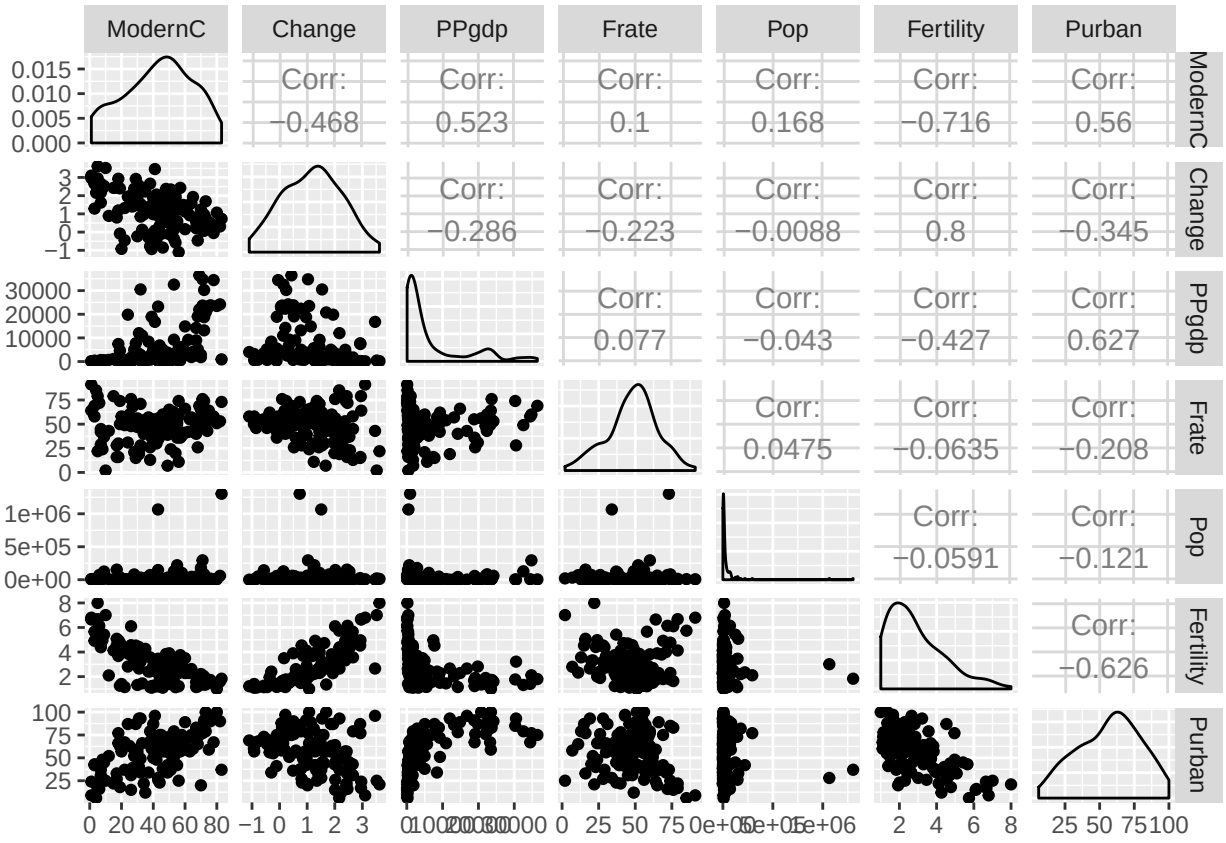
```
meanMat=sapply(UN3, mean, na.rm=TRUE)
sdMat=sapply(UN3, sd, na.rm=TRUE)
paramMat=cbind(meanMat, sdMat)

param.df=data.frame(matrix(nrow=7, ncol=3))
param.df=paramMat
paramcolNames=c("mean", "std")
colnames(param.df)=paramcolNames
kable(param.df, digits=c(3,3))
```

	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?

```
goodUN3=na.omit(UN3)
ggpairs(goodUN3, columns=c(1,2,3,4,5,6,7))
```



ModernC is most correlated with PPgdp, Fertility, Change, and Purban. There's a correlation between Fertility and Change which makes intuitive sense. There seem to be two outliers in Pop. There might be some transformation needed to PPgdp and Pop.

Model Fitting

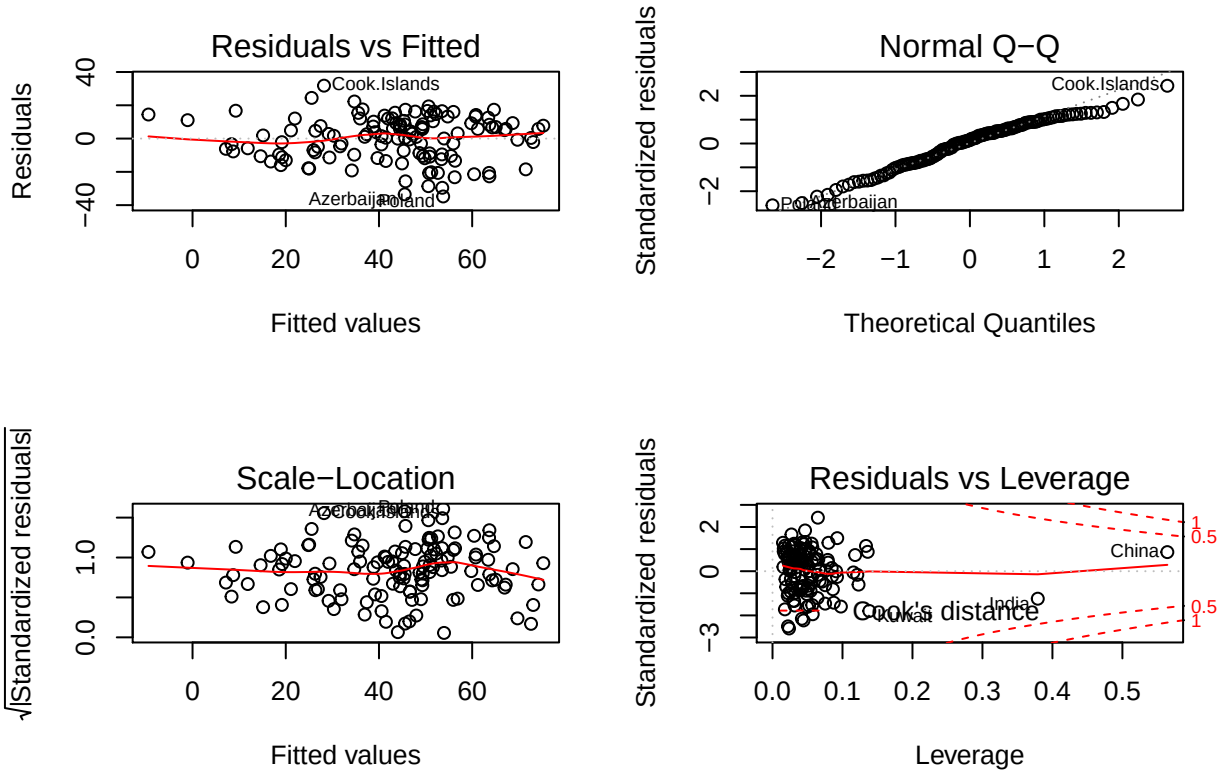
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining

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?

```
ModernC.lm=lm(ModernC ~., data=goodUN3 )
summary(ModernC.lm)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = goodUN3)
##
## 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
```

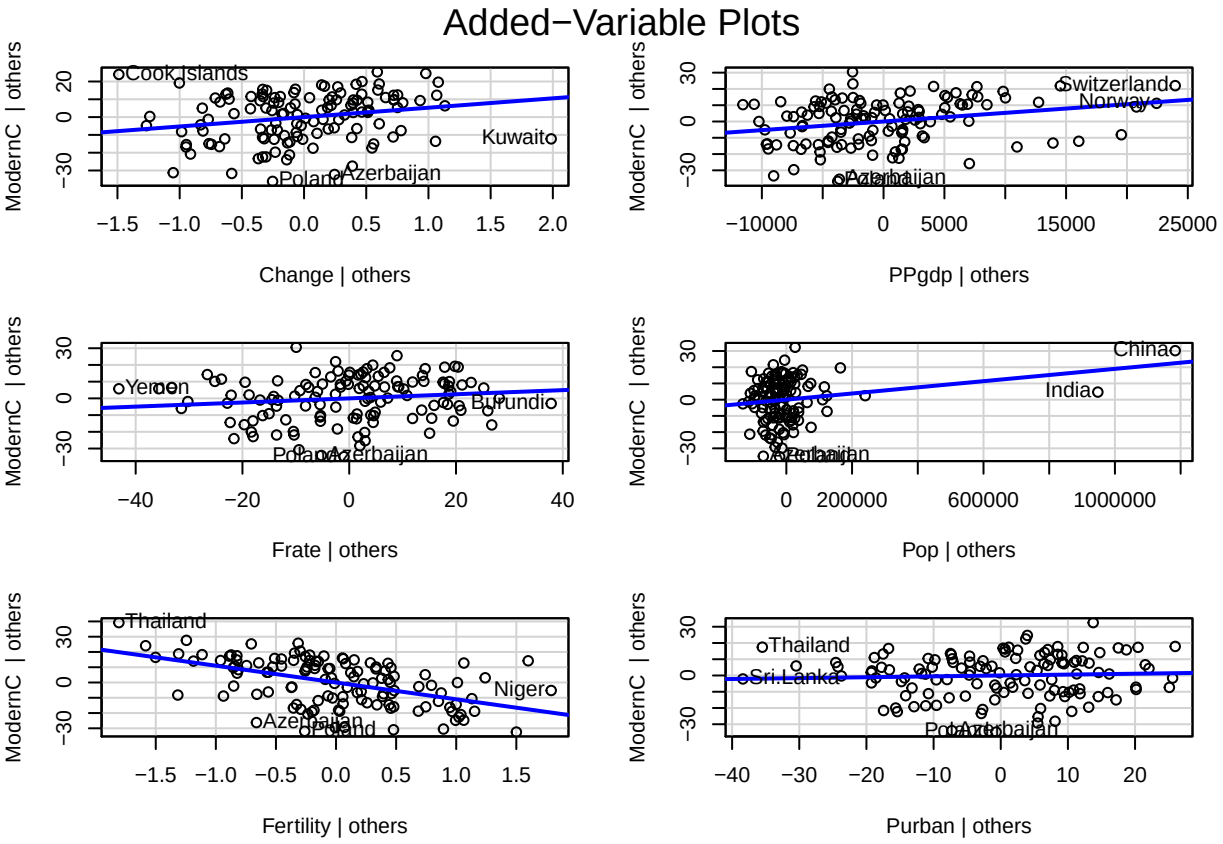
```
par(mfrow=c(2,2))
plot(ModernC.lm)
```



125 observations are used to fit the model. Looking at the residual plots we see they have a standard mean and a constant variance. On the Q-Q plot should approximate the x=y line, there seems to be a longer left tail suggested that the distribution isn't perfectly normal. From the ggpairs plots we expected 2 outliers, China and India but they are not influential points since they don't cross any Cooke's distance contours.

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?

```
car::avPlots(lm(ModernC.lm))
```



Ideally we should see a straight line in the added variable plots. This would suggest that adding a variable gives no new information to the model and the existing variables are uncorrelated with the new one. These plots suggest that we need to apply some transformation to the Population and PPgdp variables. We see China and India seem to be outliers in the Population case, but looking at the Cooke's Distnace plot on the previous problem they are not influential to the overall model, however they may be influential for determining the fit coefficient for the population variable.

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.

The variables that have the most offenseive added variable plots are Ppgdp and Pop and they need a transform more than other variables.

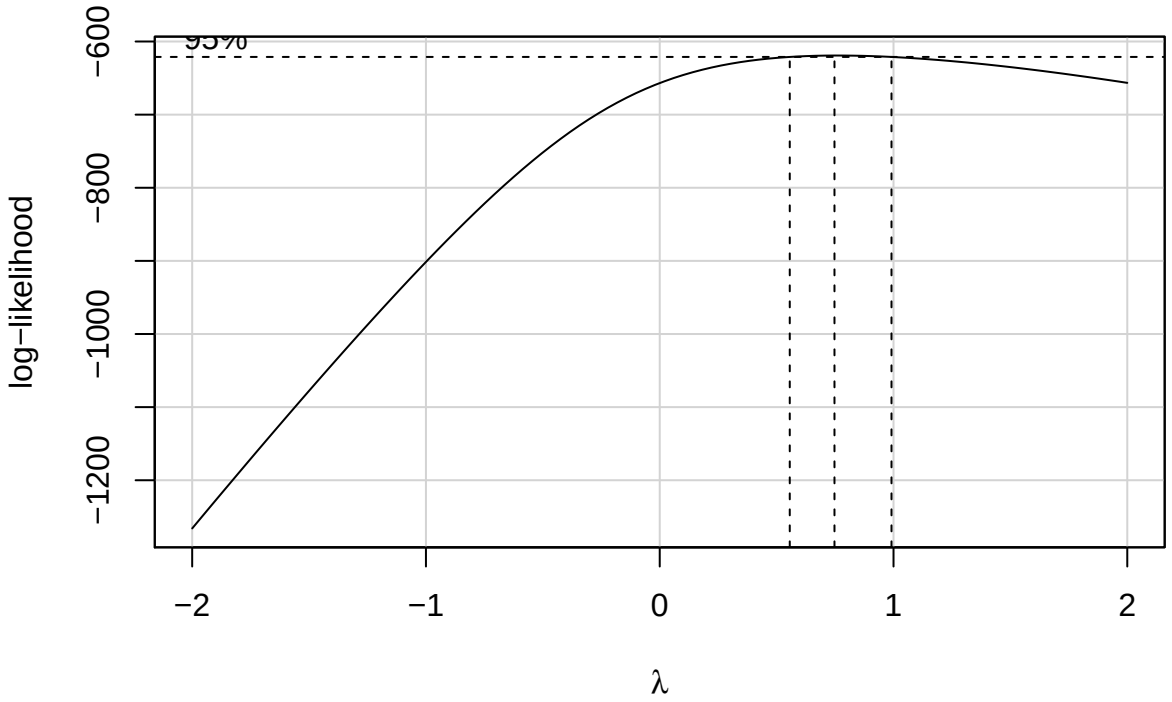
```
boxTidwell(ModernC ~ Pop + PPgdp, ~ Change+Frate+Fertility+Purban, data=goodUN3)
```

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

The ideal transformations are to raise pop to the .4 power and PPgdp to the -.12 power. For interpretability of the model we'll take the square root of population, since it's close to the power .5, and the log of the population because it's close to the 0 power.

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

```
boxCox(lm(ModernC~log(PPgdp)+sqrt(Pop)+Change+Fertility+Purban+Frate, data=goodUN3))
```



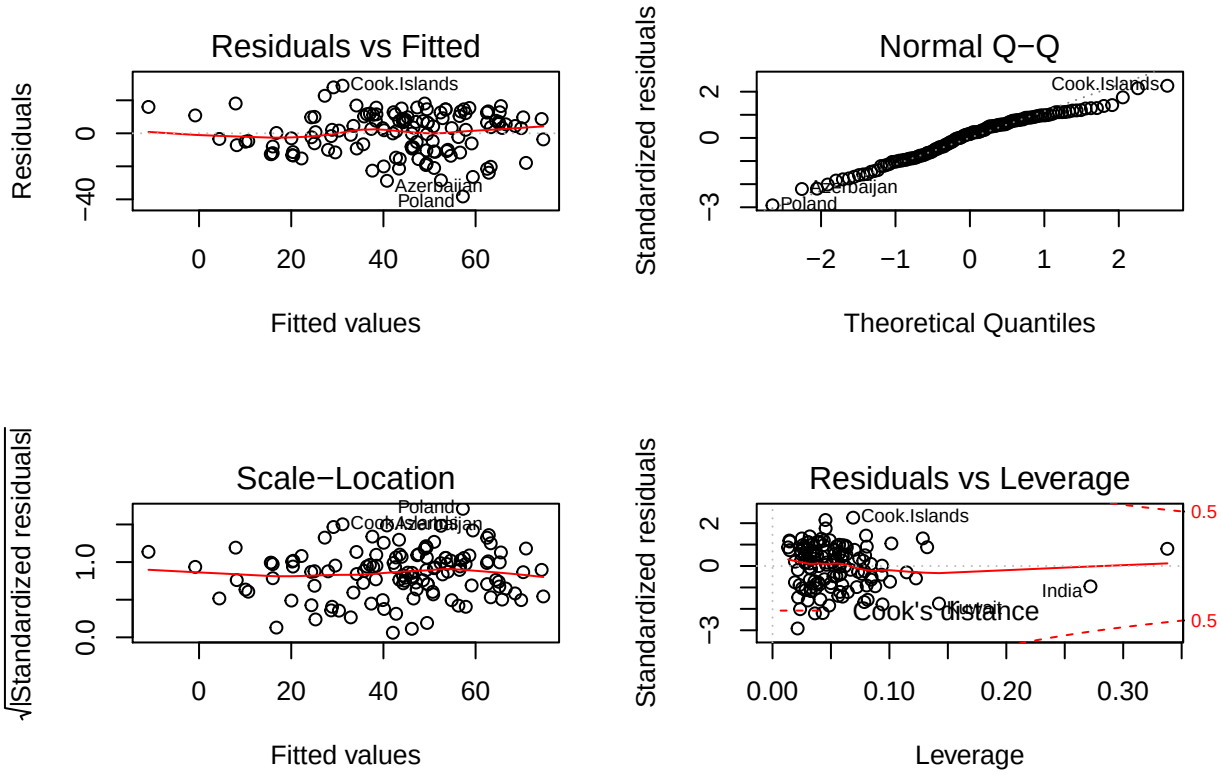
These intervals are close to 1, which means that the fit is close to linear with respect to the logarithmic transformations of the variables.

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.

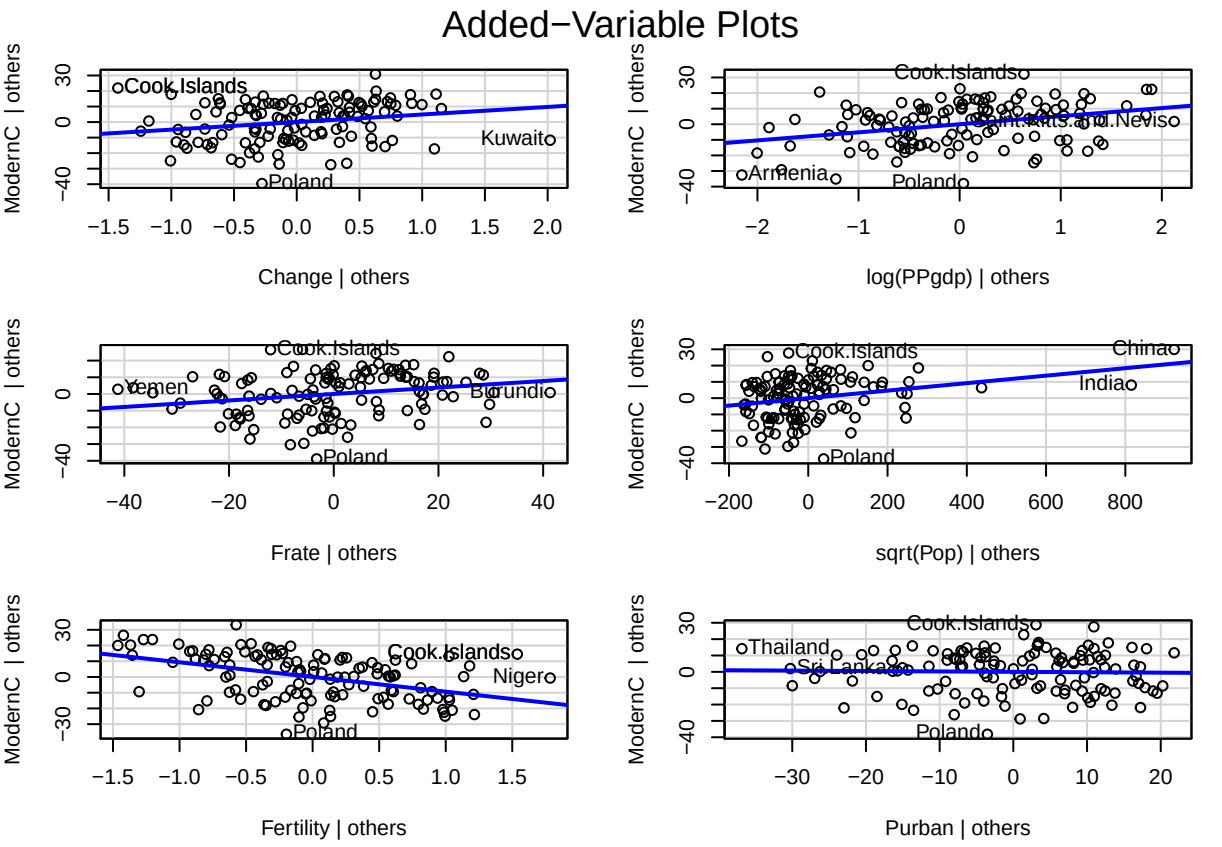
```
test.lm= lm(ModernC~ Change+log(PPgdp)+Frate+sqrt(Pop)+Fertility+Purban, data=goodUN3)
kable(summary(test.lm)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.9680651	12.2479918	1.0587911	0.2918575
Change	4.8691469	2.0427656	2.3836053	0.0187395
log(PPgdp)	5.1824148	1.3590763	3.8131890	0.0002197
Frate	0.1940100	0.0760527	2.5509960	0.0120207
sqrt(Pop)	0.0230166	0.0076852	2.9949129	0.0033458
Fertility	-9.3275721	1.7497734	-5.3307314	0.0000005
Purban	-0.0250716	0.0965569	-0.2596558	0.7955817

```
par(mfrow=c(2,2))
plot(test.lm)
```




```
car::avPlots((test.lm))
```



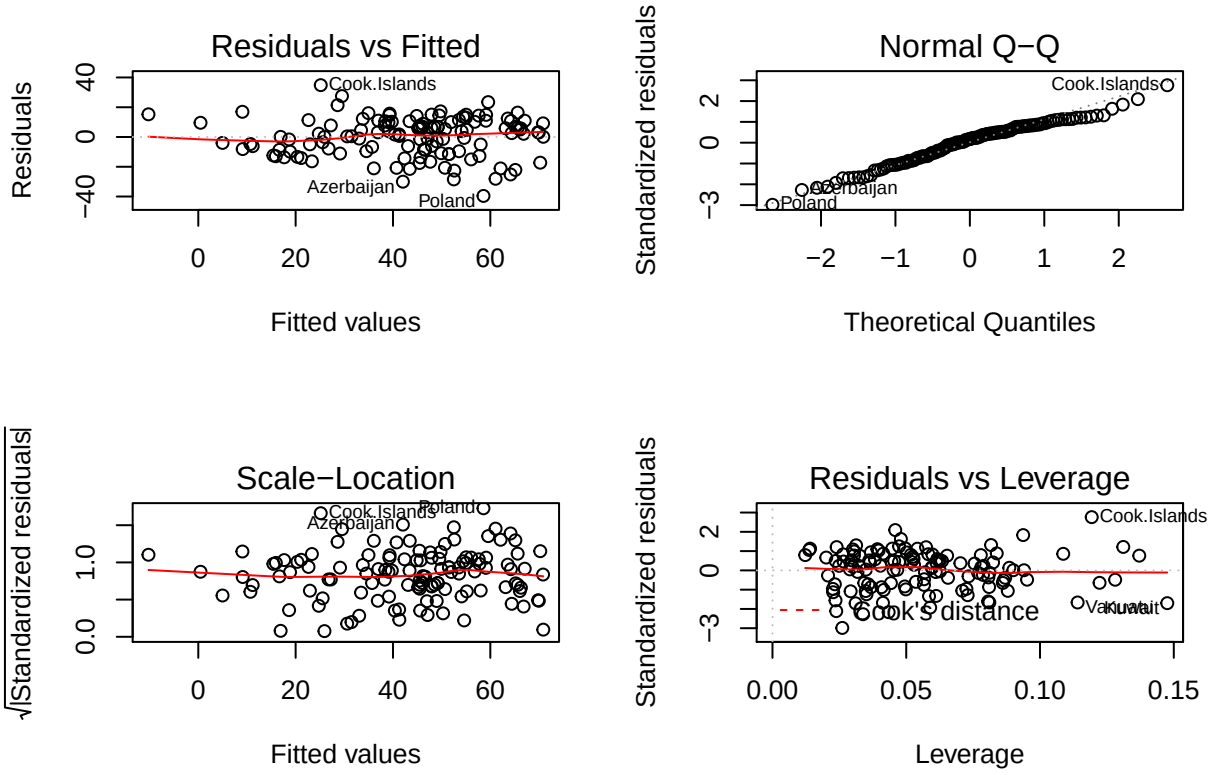
Square root might not have been the best transformation to apply to the population predictor. As we see

India and China are still outliers in the transformed population, a transformation that reigns in extreme values, and was also close in the original boxTidwell calculation is the log.

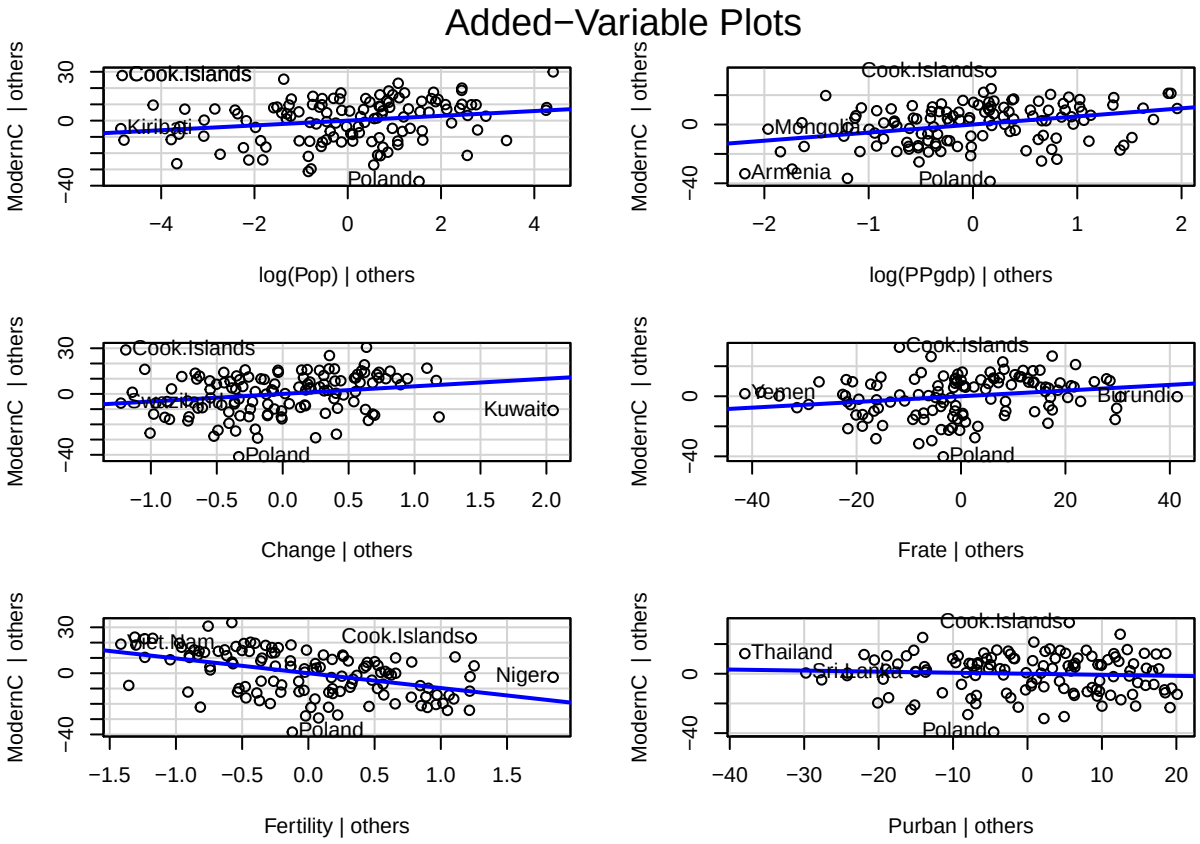
9. Start by finding the best transformation of the response and then find transformations of the predictors. Do you end up with a different model than in 8?

```
test2.lm= lm(ModernC~ log(Pop)+log(PPgdp)+Change+Frate+Fertility+Purban, data=goodUN3)

par(mfrow=c(2,2))
plot(test2.lm)
```




```
car::avPlots(test2.lm)
```




```
kable(summary(test2.lm)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.1154711	14.5085388	0.2836586	0.7771692
log(Pop)	1.4720744	0.6287542	2.3412557	0.0208965
log(PPgdp)	5.5072784	1.4050465	3.9196415	0.0001492
Change	4.9929573	2.0770920	2.4038209	0.0177813
Frate	0.1893936	0.0771102	2.4561402	0.0155002
Fertility	-9.6759414	1.7656122	-5.4802189	0.0000002
Purban	-0.0707680	0.0975983	-0.7250948	0.4698293

We do end up with a different model than the one from the previous question.

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.

There are no influential outliers in the updated linear model.

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!

```
kable(summary(test2.lm)$coef, digits=c(3,3,2,4))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.115	14.509	0.28	0.7772
log(Pop)	1.472	0.629	2.34	0.0209
log(PPgdp)	5.507	1.405	3.92	0.0001

	Estimate	Std. Error	t value	Pr(>|t|)
Change	4.993	2.077	2.40	0.0178
Frate	0.189	0.077	2.46	0.0155
Fertility	-9.676	1.766	-5.48	0.0000
Purban	-0.071	0.098	-0.73	0.4698

```
kable(confint(test2.lm), digits = c(3,3))
```

	2.5 %	
(Intercept)	-24.615	
log(Pop)	0.227	
log(PPgdp)	2.725	
Change	0.880	
Frate	0.037	
Fertility	-13.172	
Purban	-0.264	
The population	and PPgdp	variables are best fit with a logarithmic transformation. This means that if we fix the popul

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

A study was conducted to assess the sexual well being of women in a given country. Data was collected measuring the annual change in the population, the per capita gdp, population, and percent of urban population of a given country; as well as the percentage of women using a modern contraception method, the percent of females over 15 that are economically active, and te expected number of live births per female. There are many countries where data collection is difficult, and much of the dataset has missing entries. We studied the data to see if we could predict the percent of women using a modern cotnraception method with the other predictors and use this model to predict the parameters for countries with missing data. We found that modern contraception usage is most correlated with population, fertility rate, and gdp and these are the most influential predictors. This makes intuitive sense, a country that is more urbanized and has a higher gdp is more modernized and it is more likely modern contraception methods are available and women have money to spend on them. This results in a relativcely small population change and a lower fertility rate. We removed entries with missing data from the dataset since they're not used in the computation of the model anyway.

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.

Let there be a linear model

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

Since the residual distribution is normal, it follows the same relationship as the linear model

$$e_y = \hat{\beta}_0 + \hat{\beta}_1 e_x$$

We know that the residual can be defined in terms of a projection matrix since we're projecting the observation into a vector space defined by the predictor variables. We know:

$$\begin{aligned} e(y) =& Y - \hat{Y} = (I - H)Y \\ H =& x(X'X)^{-1}x' \\ \hat{\beta} =& (X'X)^{-1}X'Y \\ (I - H)Y =& \hat{\beta}_0 + \hat{\beta}_1(I - H)X \end{aligned}$$

let X_j be a subset of X without the jth term.

$$X = (I - H)X_j$$

Exploiting the fact that $(I - H)$ is an idempotent matrix and symmetric $(I - H)' = (I - H)$

$$\begin{aligned} (I - H)Y =& \hat{\beta}_0 + (X'X)^{-1}X'Y(I - H)X \\ (I - H)Y =& \hat{\beta}_0 + (((I - H)X_j)'(I - H)X_j)^{-1}((I - H)X_j)'Y(I - H)(I - H)X_j \\ (I - H)Y =& \hat{\beta}_0 + (X_j'(I - H)'(I - H)X_j)^{-1}X_j'(I - H)'Y(I - H)(I - H)X_j \\ (I - H)Y =& \hat{\beta}_0 + (X_j'(I - H)X_j)^{-1}X_j'(I - H)'Y(I - H)X_j \end{aligned}$$

We then left multiply both sides by X_j

$$X_j'(I - H)Y = X_j'\hat{\beta}_0 + X_j'(X_j'(I - H)X_j)^{-1}X_j'(I - H)'Y(I - H)X_j$$

The quantities $X_j'(I - H)'Y$ $(X_j'(I - H)X_j)$ are scalar and can be moved freely.

$$\begin{aligned} X_j'(I - H)Y =& X_j'\hat{\beta}_0 + X_j'(X_j'(I - H)X_j)^{-1}X_j'(I - H)'Y(I - H)X_j \\ X_j'(I - H)Y =& X_j'\hat{\beta}_0 + X_j'(I - H)'X_j'(X_j'(I - H)X_j)^{-1}Y(I - H)X_j \\ X_j'(I - H)Y =& X_j'\hat{\beta}_0 + X_j'Y(I - H)X_j \\ (I - H)Y =& \hat{\beta}_0 + Y(I - H)X_j \end{aligned}$$

which can only be true if $\hat{\beta}_0 = 0$, i.e. it doesnt change

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.

```
e_Y= residuals(lm(ModernC~ log(PPgdp)+Change+Frate+Fertility+Purban, data=goodUN3))
e_X= residuals(lm(log(Pop)~ +log(PPgdp)+Change+Frate+Fertility+Purban, data=goodUN3))

Ris.lm=lm(e_Y~e_X, data=goodUN3)

kable(summary(test2.lm)$coef, digits=c(3,3,3,2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.115	14.509	0.284	0.78
log(Pop)	1.472	0.629	2.341	0.02
log(PPgdp)	5.507	1.405	3.920	0.00
Change	4.993	2.077	2.404	0.02
Frate	0.189	0.077	2.456	0.02
Fertility	-9.676	1.766	-5.480	0.00
Purban	-0.071	0.098	-0.725	0.47

```
kable(summary(Ris.lm)$coef, digits=c(3,3,3,2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.000	1.178	0.00	1.00
e_X	1.472	0.616	2.39	0.02

Looking at the estimate of the parameters, we see the parameter of predicting the log of the population are the same in both models, 1.472.