HW2 STA521 Fall18

Harshit Sahay, netID:hs239, github username: harshitsahay

Due September 24, 2018 9pm

Backgound Reading

Readings: Chapters 3-4 in Weisberg Applied Linear Regression

Exploratory Data Analysis

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

All variables are quantitative.

6 of the 7 variables (ModernC, Change, PPgdp, Frate, Pop, Fertility) have missing data. The only variable that doesn't have missing data is Purban.

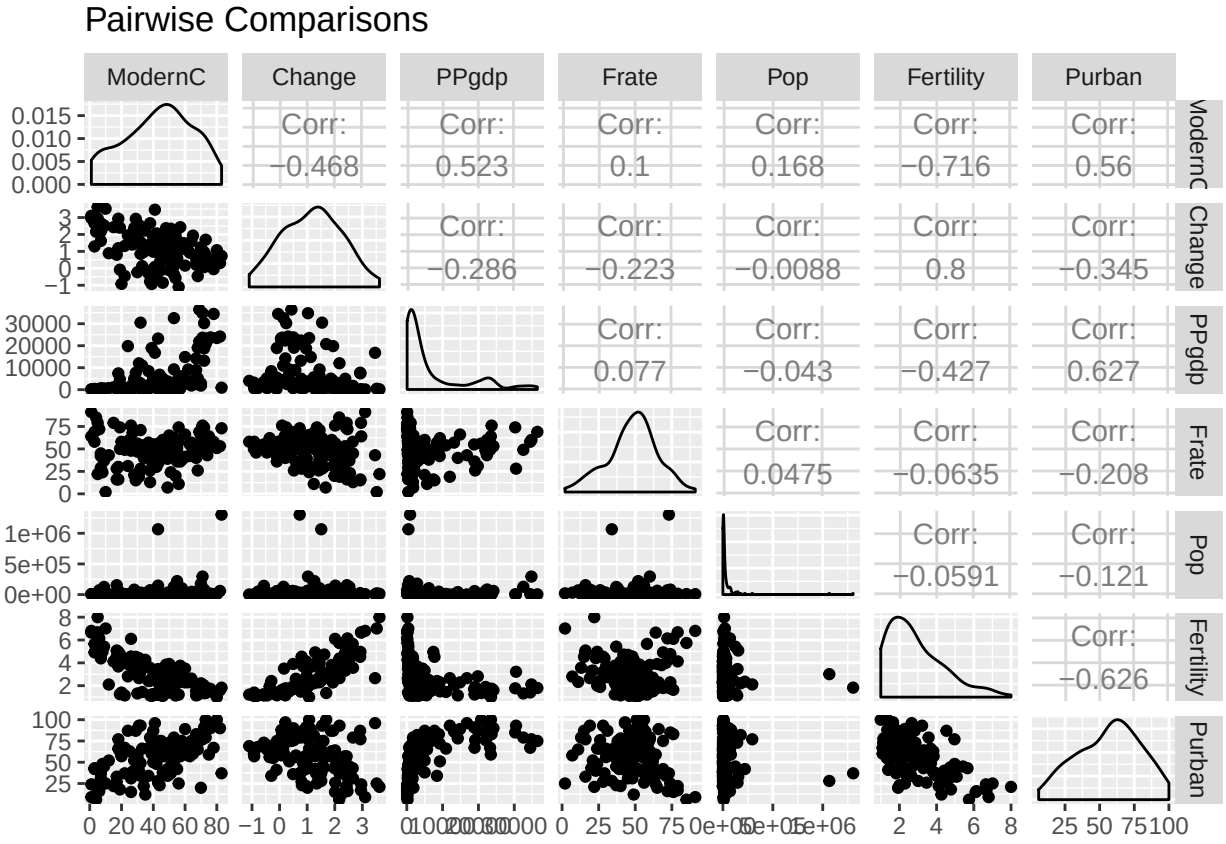
2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

```
data_mean= round(sapply(UN3,mean,na.rm=TRUE),3)
data_sd= round(sapply(UN3,sd,na.rm=TRUE),3)

kable(cbind(data_mean,data_sd), col.names =c('Mean','Std. Dev'))
```

	Mean	Std. Dev
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?



The Population plots suggests the presence of two outliers. As far as linearity is concerned, two variables seem useful- Fertility, with a negative relationship, and PUrban, with a positive relationship. There does seem to be some predictive power associated with PPgdp too, but this is not a linear relationship, and might need to be transformed. We can look at some more plots in closer detail.

```
ModC_fer=ggplot(na.omit(UN3),aes(x=Fertility,y=ModernC))+geom_point()+theme(plot.title = element_text(h
```

```
ModC_Change=ggplot(na.omit(UN3),aes(y=ModernC,x=Change)) + geom_point()+theme(plot.title = element_text
```

```
ModC_Purban=ggplot(na.omit(UN3),aes(y=ModernC,x=Purban)) + geom_point()+theme(plot.title = element_text
```

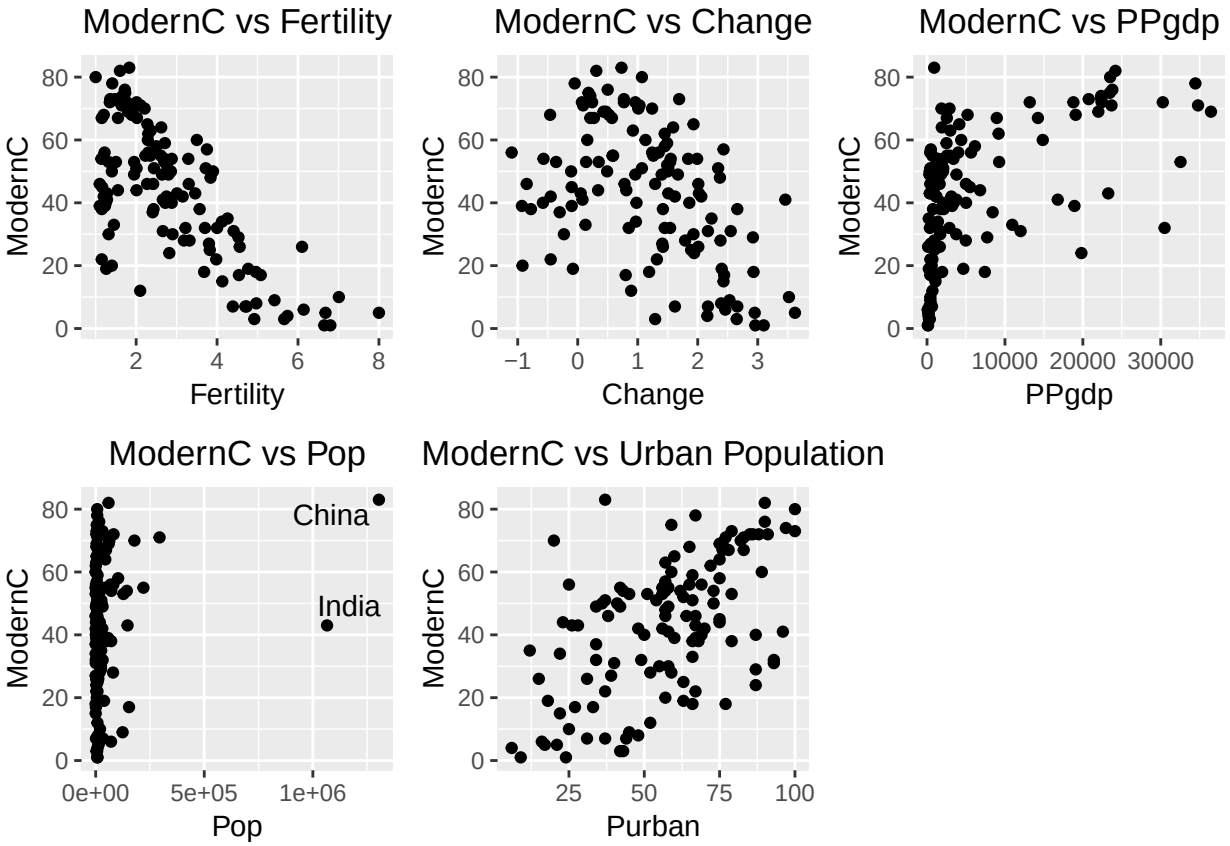
```
ModC_PPgdp=ggplot(na.omit(UN3),aes(y=ModernC,x=PPgdp)) + geom_point()+theme(plot.title = element_text(h
```

```
UN3_mod=mutate(UN3,log_PPgdp=log(PPgdp))
ModC_logPPgdp=ggplot(data=na.omit(UN3_mod),aes(x=log_PPgdp,y=ModernC))+geom_point()+theme(plot.title =
```

```
UN3_test=UN3
UN3_test$country=row.names(UN3)
ModC_Pop=ggplot(na.omit(UN3_test),aes(y=ModernC,x=Pop)) + geom_point()+theme(plot.title = element_text(
```

```
UN3_mod1=mutate(UN3,log_pop=log(Pop))
ModC_logPop=ggplot(data=na.omit(UN3_mod1),aes(x=log_pop,y=ModernC))+geom_point()+theme(plot.title = eler
```

```
grid.arrange(ModC_fer,ModC_Change,ModC_PPgdp,ModC_Pop,ModC_Purban,nrow=2)
```

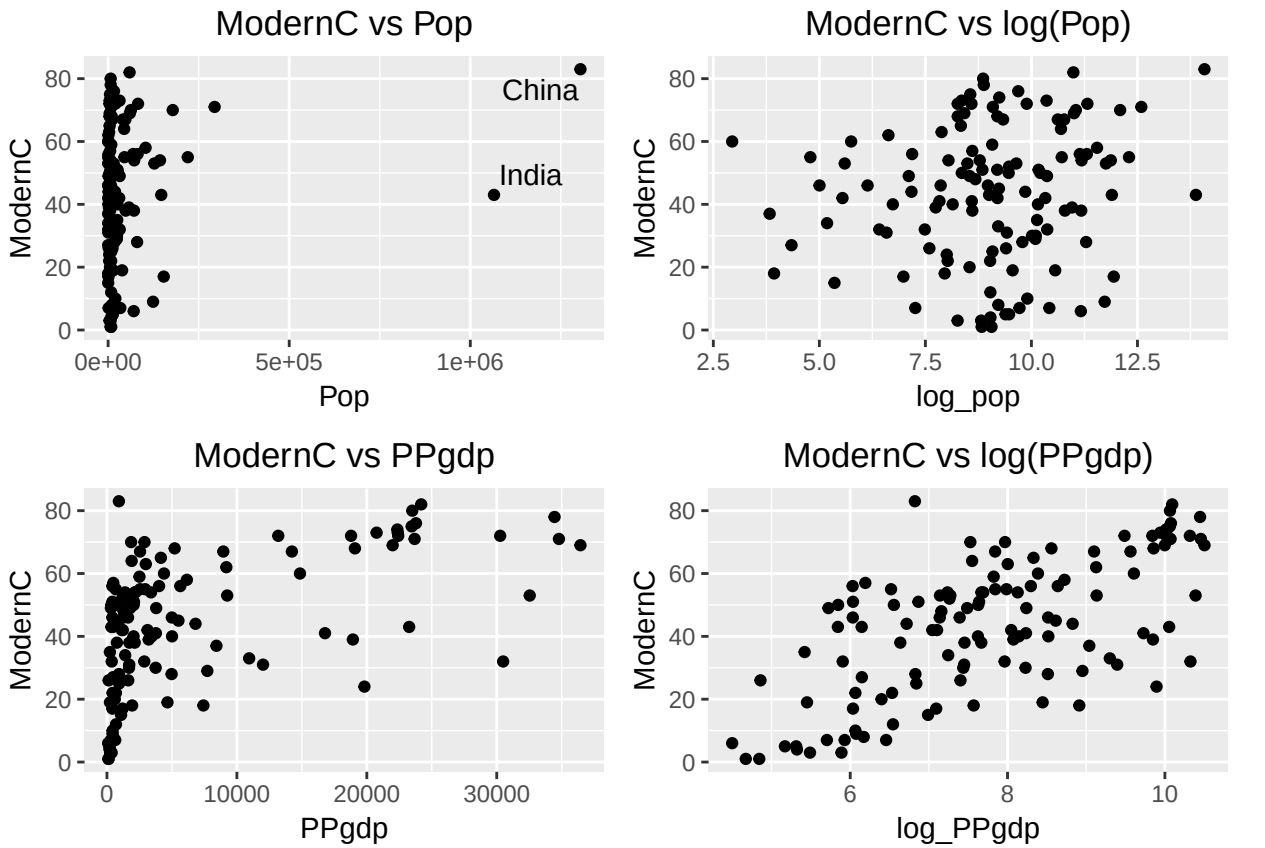


There is a negative, somewhat linear relationship we saw earlier for Fertility. A similar relationship can be seen for Change, with slightly less correlation (and hence predictive value). Similarly, for Purban, we see the positive linear relationship. All of these aren't exact, as data rarely is, but are good places to start. Out of the three of these, it it clear that Fertility is the best predictor, as could also be seen from the correlation value in the ggpairs plot.

Going back to the ggpairs plot suggests that Frate and Pop may not be useful predictors due to their low correlations with ModernC. However, PPgdp has correlations comparable to Change and Purban, but this is not clear from the plot, since the relationship isn't linear. A transformation would help.

The ModernC vs Pop graph suggests that India and China are outliers.

```
grid.arrange(ModC_Pop,ModC_logPop,ModC_PPgdp,ModC_logPPgdp,nrow=2)
```



We can also see that the log transforms for Pop and PPgdp represent the relationship better, as far as modelling with a linear model is concerned.

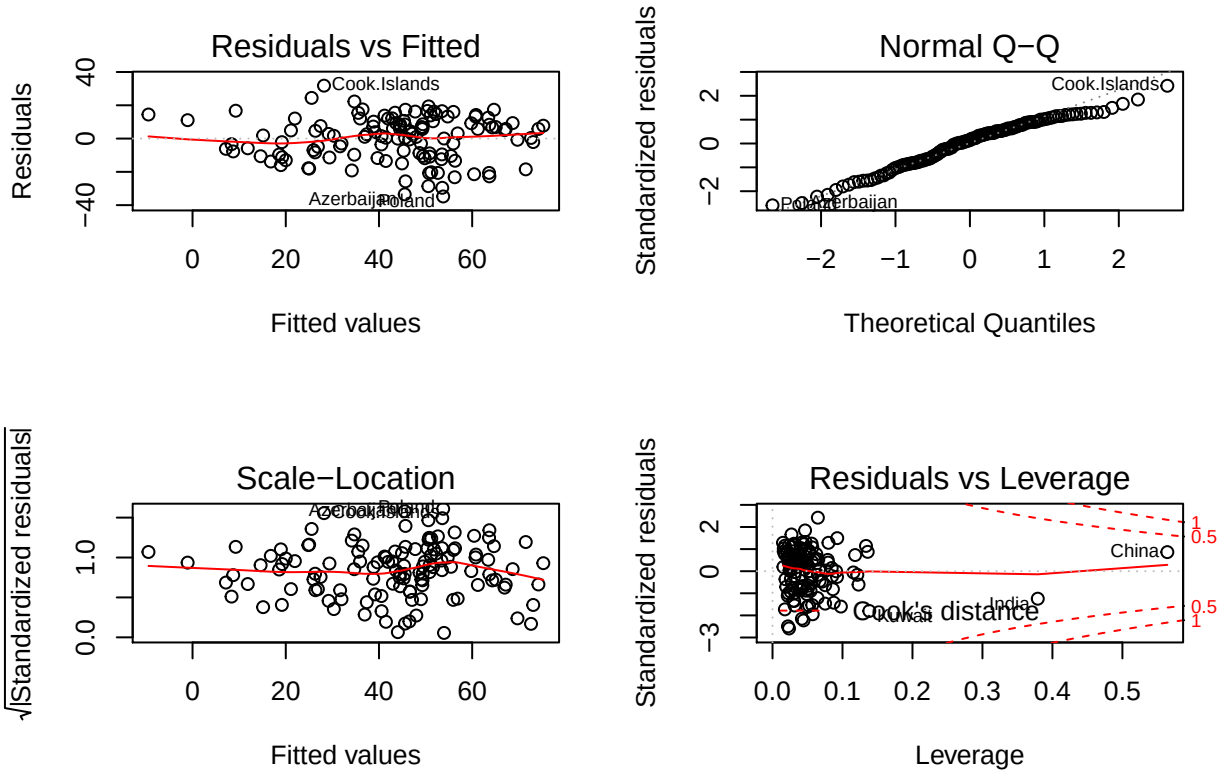
Model Fitting

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?

125 observations are used in the fitted model.

The Residuals-Fitted and Scale-Location plots suggest that the linearity assumptions isn't completely valid. The slight bulge in both plots suggests this. The Normal-QQ plot suggests that normality isn't valid here, the data appears to be left skewed. The final plot suggests that India and China have high leverage, but are still within the Cook's distance thresholds. At the threshold of 0.5, these points are not influential, but they still have a significantly higher leverage compared to other points.

```
model=lm(ModernC~.,data=na.omit(UN3))
par(mfrow=c(2,2))
plot(model)
```



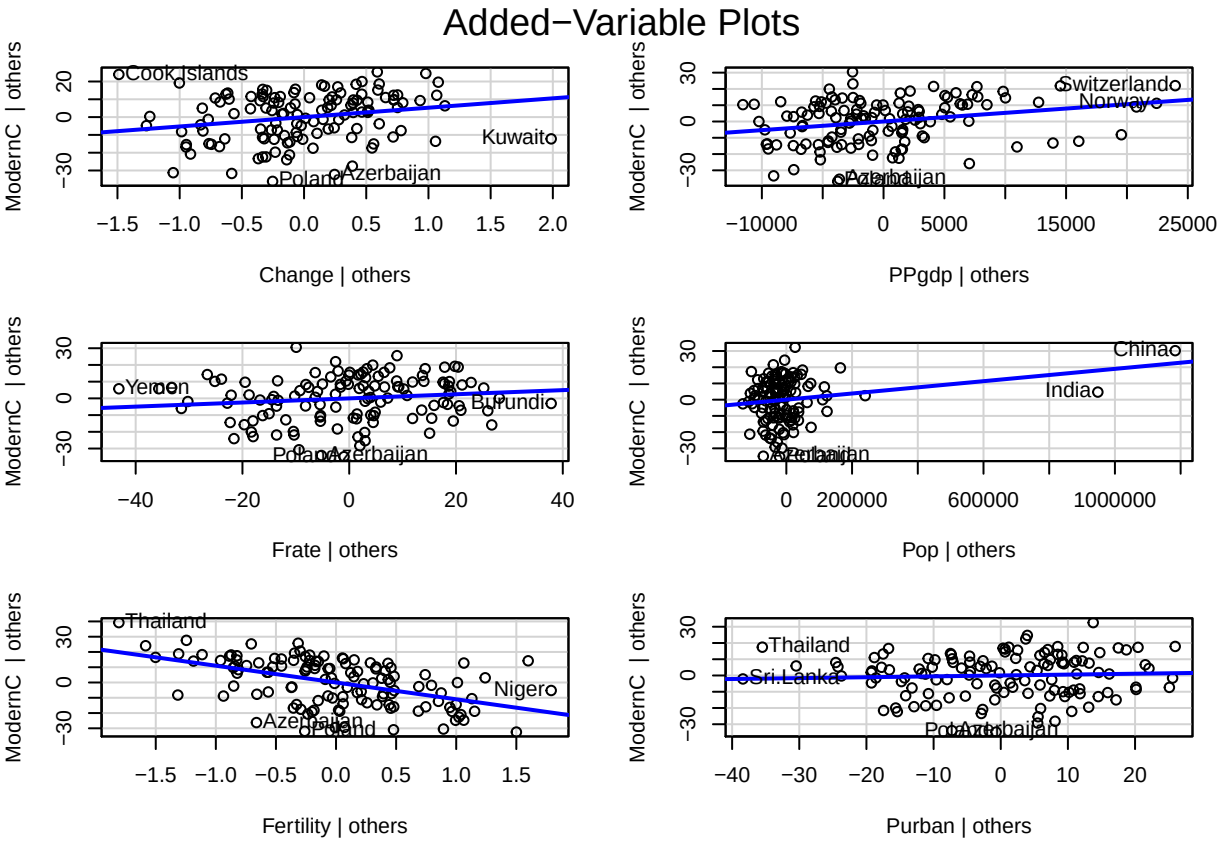
5. Examine added variable plots `car::avPlot` or `car::avPlots` for your model above. Are there any plots that suggest that transformations are needed for any of the terms in the model? Describe. Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?

We discussed earlier, and it's evident now that PPgdp does have predictive power. The scale of the axis on the residuals for it suggests that we might wanna try a log transformation. We can also see that Pop has predictive power, however, the scale of residuals, and also the outliers suggest that these might need transformations too. A log transformation might reduce the effect of these outliers (India and China).

We also see that Purban and Frate have little predictive value after the other variables have been accounted for.

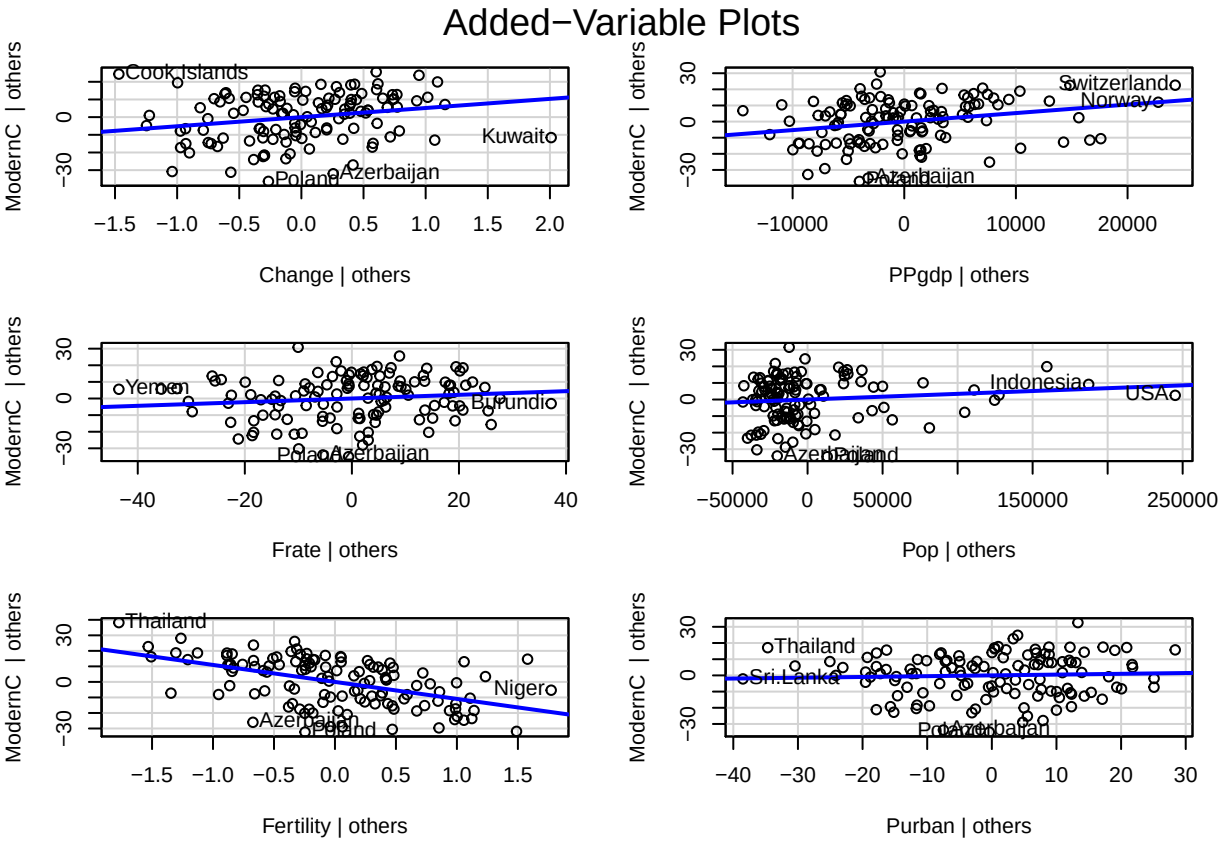
Localities that are influential are China and India for Pop. Cook Island and Kuwait for Change, Thailand for Fertility.

```
avPlots(model)
```



We'll also look at the added variable plots after we remove India and China. This is to check if the added variable plot for Pop still looks similar after the outliers are removed. We see that there is still a relationship. The log transform will probably be a good idea to retain, and not something that's just an artefact due to these two data points.

```
UN3_outliers=subset(UN3,Pop<1000000)
model2=lm(ModernC~.,UN3_outliers)
avPlots(model2)
```

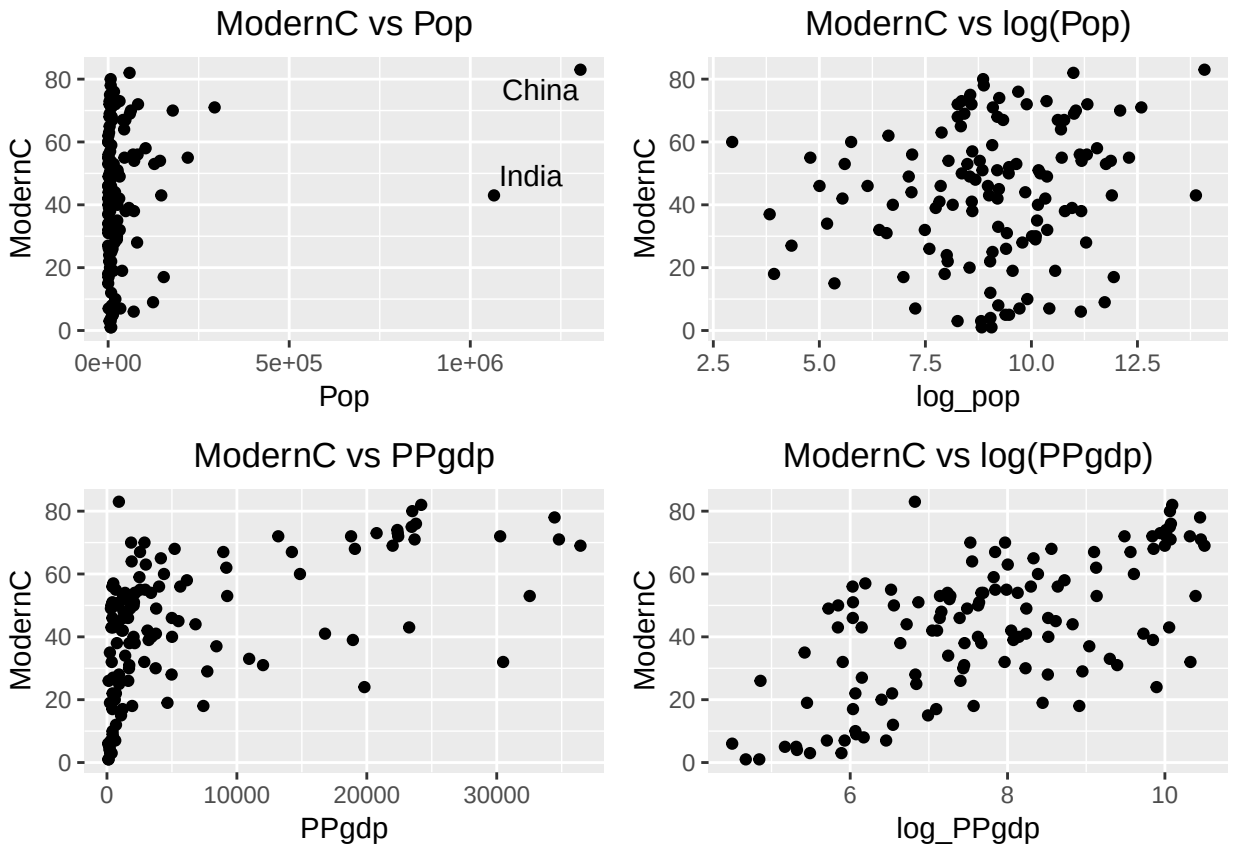


This plot with transformed variables again suggests that Purban doesn't have a lot of predictive power. Frate seems to have increased slightly in this respect, so we'll keep this in mind and keep it in the model for now.

6. Using the Box-Tidwell `car::boxTidwell`, `car::powerTransform` 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.

Note that graphical methods for appropriate transformations have already been provided in Q3. Reproducing those plots here, we basically see that log transforms for Pop and Ppgdp might be useful.

```
grid.arrange(ModC_Pop,ModC_logPop,ModC_PPgdp,ModC_logPPgdp,nrow=2)
```



We run boxTidwell tests for Pop and PPgdp, since these are the ones that we want to transform. We can also try dropping Purban since it doesn't have a lot of predictive value. To conform that we don't need this, we can use Anova later.

Our second test is to keep log transforms for Pop and PPgdp, and check if any of the variables that we have assumed to be linear need to be transformed. The following code chunk tests that. For the second test, we'll need to transform Change to make it positive. The best way to do this is probably to add a constant positive value. Since the lowest value is -1.10, adding say 2, is sufficient.

```
boxTidwell(ModernC~Pop+PPgdp,~Fertility+Change+Frate, data=na.omit(UN3))
```

```
##       MLE of lambda Score Statistic (z) Pr(>|z|)
## Pop         0.42227             -0.7941   0.4271
## PPgdp      -0.13620             -1.2872   0.1980
##
## iterations =  4
```

```
UN3_boxTid=mutate(UN3,Change=Change+2)
boxTidwell(ModernC~Fertility+Change+Frate,~log(Pop)+log(PPgdp),data=na.omit(UN3_boxTid))
```

```
## Warning in boxTidwell.default(y, X1, X2, max.iter = max.iter, tol = tol, :
## maximum iterations exceeded
```

```
##           MLE of lambda Score Statistic (z) Pr(>|z|)
## Fertility        2.3501             -0.2640  0.79178
## Change          71.4897             -2.3810  0.01727 *
## Frate            1.9542              1.0413  0.29775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## iterations =  26
```

All other transformations have no significance associated with them. So we'll keep the log transforms for Pop and PPgdp based on our earlier assessment. We'll also not transform Change, even though it's significant. This is because the exponent in question is too high to make sense, and might be due to the smaller dynamic range associated with this variable. Let's check anova like we decided to before. We can essentially see if any term we add is useful, starting with Fertility, which had the highest correlation.

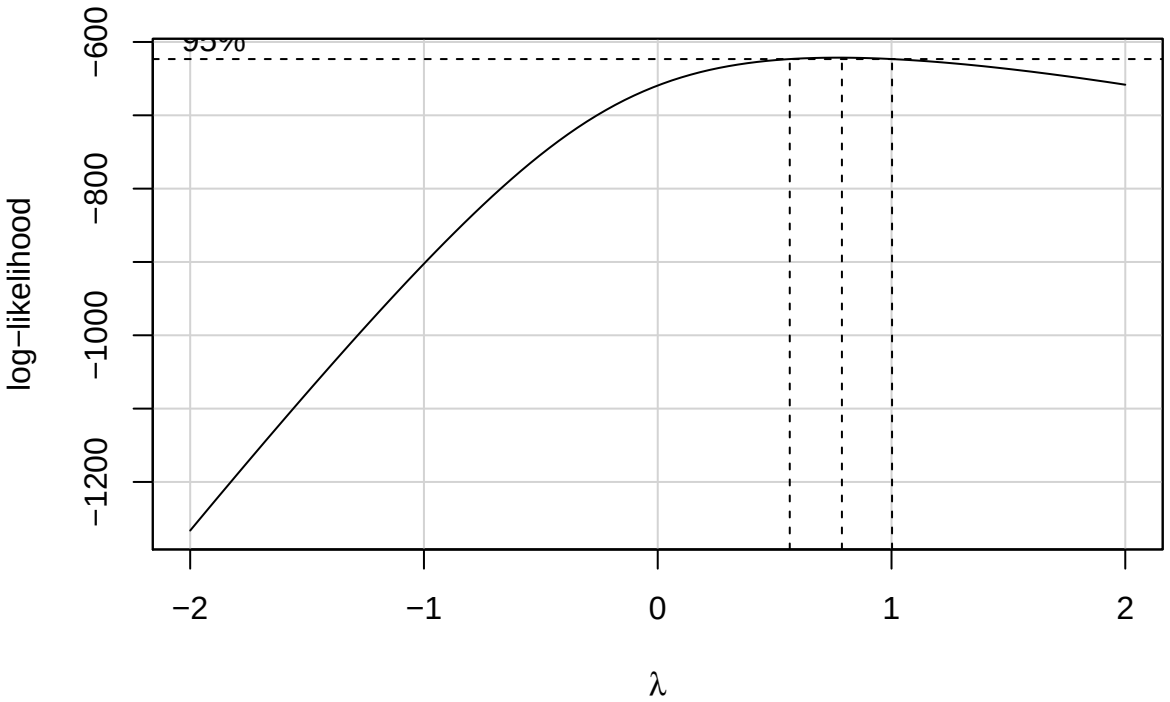
```
anova(lm(ModernC~Fertility+Change+log(PPgdp)+log(Pop)+Frate+Purban,data=na.omit(UN3)))
```

```
## Analysis of Variance Table
##
## Response: ModernC
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Fertility    1 29232.5 29232.5 161.7551 < 2.2e-16 ***
## Change       1  1727.7  1727.7   9.5599 0.0024821 **
## log(PPgdp)   1  2443.0  2443.0  13.5179 0.0003571 ***
## log(Pop)     1   938.1   938.1   5.1910 0.0245025 *
## Frate        1  1251.5  1251.5   6.9250 0.0096352 **
## Purban       1    95.0    95.0   0.5258 0.4698293
## Residuals  118 21325.0   180.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we said earlier, Purban doesn't add much to the model, and anova verifies that it doesn't have a significant reduction in RSS. So we'll drop it from the predictors, and use the log transforms for Pop and PPgdp.

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

```
bcx=boxCox(lm(ModernC~Fertility+Change+log(PPgdp)+log(Pop)+Frate,data=na.omit(UN3)),plotit=TRUE)
```



We can print the 95% confidence interval as:

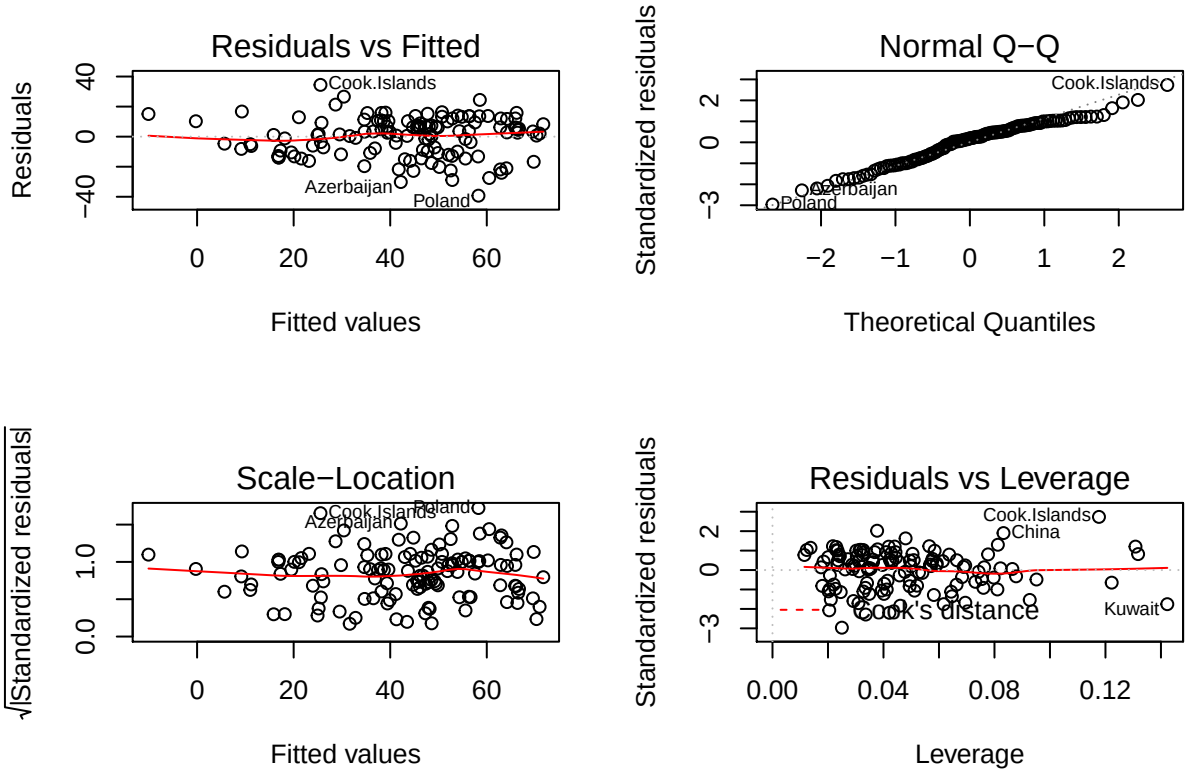
```
range(bcx$x[bcx$y > max(bcx$y)-qchisq(0.95,1)/2])
```

```
## [1] 0.5858586 0.9898990
```

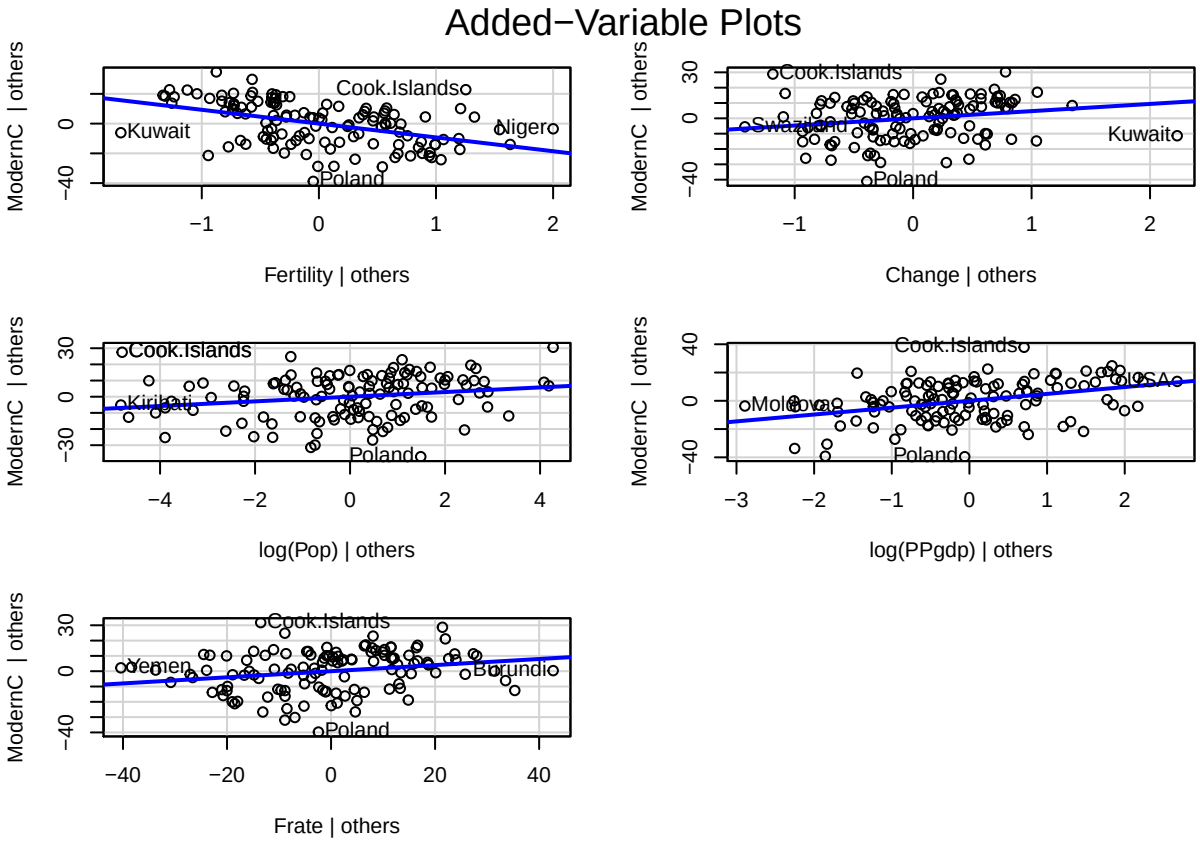
No integer liesi in the confidence interval but 1 is pretty close. Examination of the graph also confirms that there is no dip at 1, and that choosing it should be okay. Rather than go with a value in the confidence interval, for the purposes of interpretability, we'll take lambda to be 1. That means, we are using ModernC without any transformations as the response.

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.

```
final_model=lm(ModernC~Fertility+Change+log(Pop)+log(PPgdp)+Frate,data=na.omit(UN3))
par(mfrow=c(2,2))
plot(final_model)
```




```
avPlots(final_model)
```

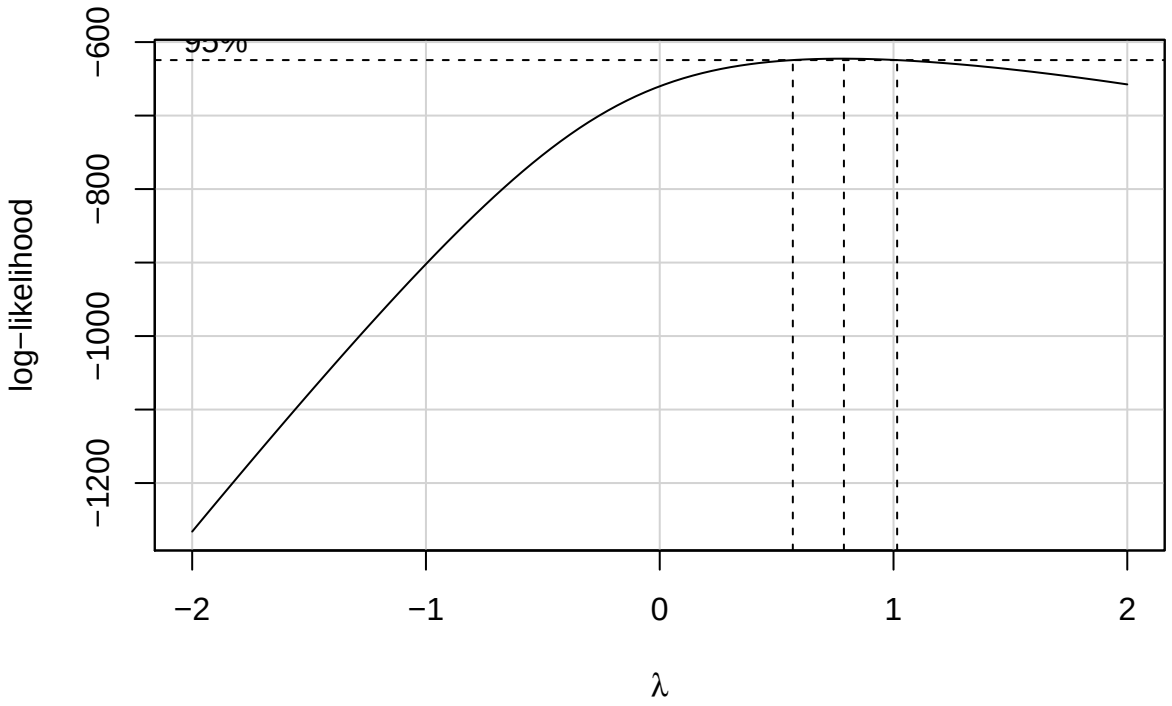


The response isn't completely linear. This is expected based on what we've seen, but for the most parts the Residual vs Fitted and Scale vs Location plots are relatively okay. The QQ plot suggests a slightly lighter tail, but is mostly okay.

While Kuwait and Cook Islands (an outlier in some other plots) have high leverage values, they don't have a significant Cook's distance, and are not influential.

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?

```
boxCox(lm(ModernC~.,data=na.omit(UN3)),plotit = TRUE)
```



Not transforming the predictors first also has similar effects, 1 is in the confidence interval this time. So we will end up with the same model, because after using lambda=1 for ModernC we'll perform similar analysis as before, dropping Purban and using log transforms for Pop and PPgdp.

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 aren't any influential points since all points are within Cook's Distance. We can assume that they aren't affecting the regression slop much.

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!

```
kable(cbind(round(final_model$coefficients,3),round(confint(final_model),3)),col.names=c("Coefficient",
```

	Coefficient	2.5%	97.5%
(Intercept)	4.102	-24.569	32.773
Fertility	-9.278	-12.595	-5.962
Change	4.698	0.673	8.723
log(Pop)	1.441	0.202	2.681
log(PPgdp)	4.859	2.717	7.002
Frate	0.200	0.050	0.349

Interpretation for these is as follows. Holding all other variables constant:

1) Fertility: A nation where the expected fertility is higher by 1 child per female has a 9.2% decrease in the number of married women using a modern method of contraception.

2) Change: For every 1% increase in the population growth rate, the use of modern contraception increases by 4.6%.

3) Frate: For every 1% increase in the number of economically active females over 15, the access to contraception increases by 0.2%.

4) PPgdp: We used a log transform, so a change of 1 unit for this variable is equal to 2.71 (the value of e) times the original value. This tells us that for every 10% increase in a nation's per capita GDP, the access to contraception rises by 0.46%.

5) Pop: For every 10% increase in a nation's population, we see a 0.13% increase in access to contraception.

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

These findings suggest that access to modern contraception has a strong relationship with several predictors. Countries with higher access to modern contraception usually have lower fertility rates, provided all other factors are constant. Increase in GDP rates also affects access to these methods, albeit much less significantly. The model doesn't have any case deletions. The final model looks like this, and can be used accordingly for parameter estimates to shape policies in question: It explains 62% of the variance.

$$ModernC = 4.102 - 9.278 * Fertility + 4.698 * Change + 1.441 * log(Pop) + 4.859 * log(PPgdp) + 0.200 * Frate$$

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 projection matrix for X which contains a column of ones, then $1_n^T(I-H) = 0$ or $(I-H)1_n = 0$. Use this to show that the sample mean of residuals will always be zero if there is an intercept.*

We can start with the following

$$e_{(Y)} = \hat{\beta}_0 + \hat{\beta}_1 e_{(X_i)}$$

We also have

$$e_{(Y)} = (I - H)Y$$

and

$$e_{(X_i)} = (I - H)X_i$$

This gives us :

$$(I - H)Y = \hat{\beta}_0 + \hat{\beta}_1 (I - H)X_i$$

Note that for $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, we have $\hat{\beta}_1 = [X^T X]^{-1} X^T Y$. Here $Y = (I - H)Y and X = (I - H)X_i$. Substituting this, \

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

Pre-multiplying by X_i^T:

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

\ Note that $X_i^T(I - H)Y$ is a scalar, so we can rearrange to give:

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

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

$$\sum_i X_i^{(1)^n} \hat{\beta}_0 = 0$$

. Hence, slope of the intercept has to be zero, assuming X_i is not 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.

We can take X_j to be Fertility

```
e_Y=residuals(lm(ModernC~Change+log(Pop)+log(PPgdp)+Frate,data=na.omit(UN3)))
e_X=residuals(lm(Fertility~ Change+log(Pop)+log(PPgdp)+Frate,data=na.omit(UN3)))
av_model=lm(e_Y~e_X)
kable(cbind(coef(av_model)["e_X"],coef(final_model)["Fertility"]),col.names = c("e_X","Fertility"),row.
```

e_X	Fertility
-9.278421	-9.278421