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.

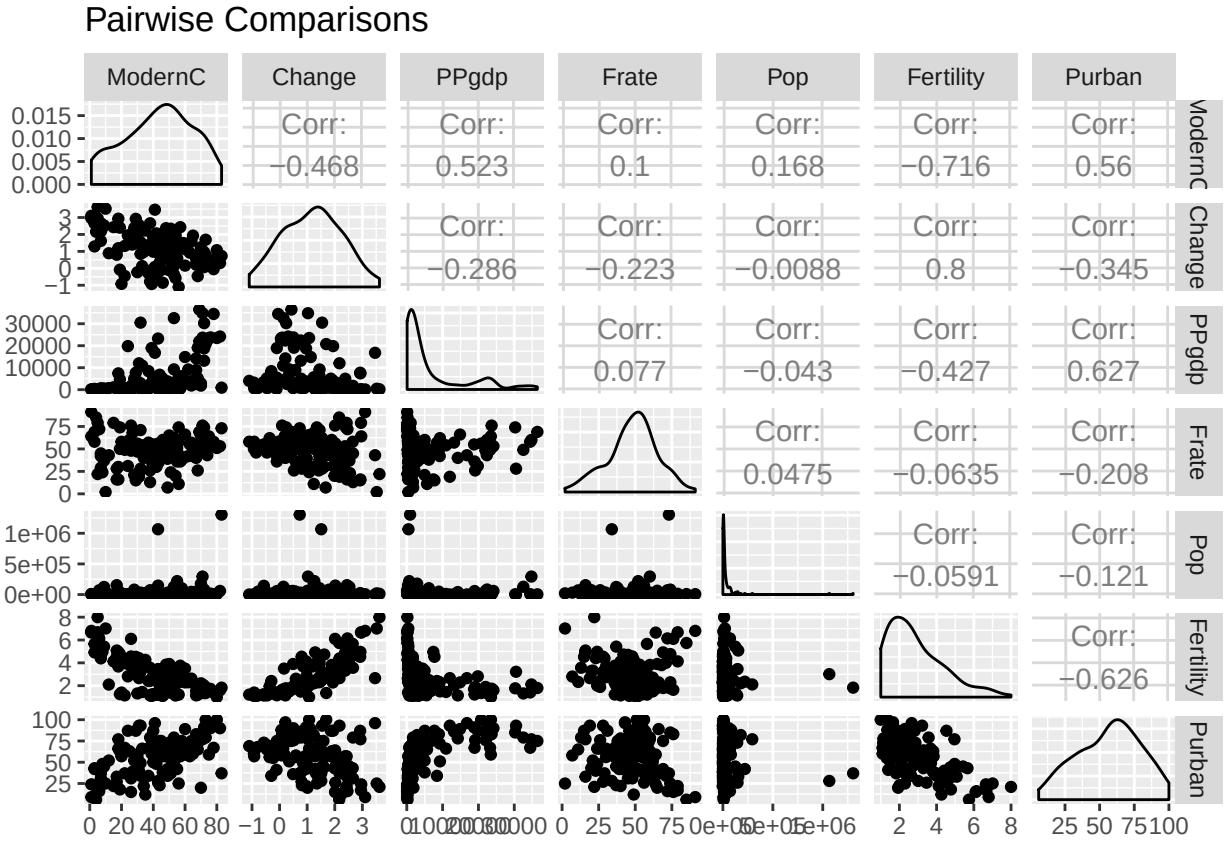
```
data_mean= round(sapply(UN3,mean,na.rm=TRUE),3)
data_sd= round(sapply(UN3,sd,na.rm=TRUE),3)

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?

```
## Warning in ggmatrix_gtable(x, ...): Please use the 'progress' parameter
## in your ggmatrix-like function call. See ?ggmatrix_progress for a few
## examples. ggmatrix_gtable 'progress' and 'progress_format' will soon be
## deprecated.TRUE
```



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
```

This is the negative, somewhat linear relationship we saw earlier. 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.

```
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
```

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.

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

We can see that a log transform better showcases the linear trend that can be modelled.

```
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 = 
```

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

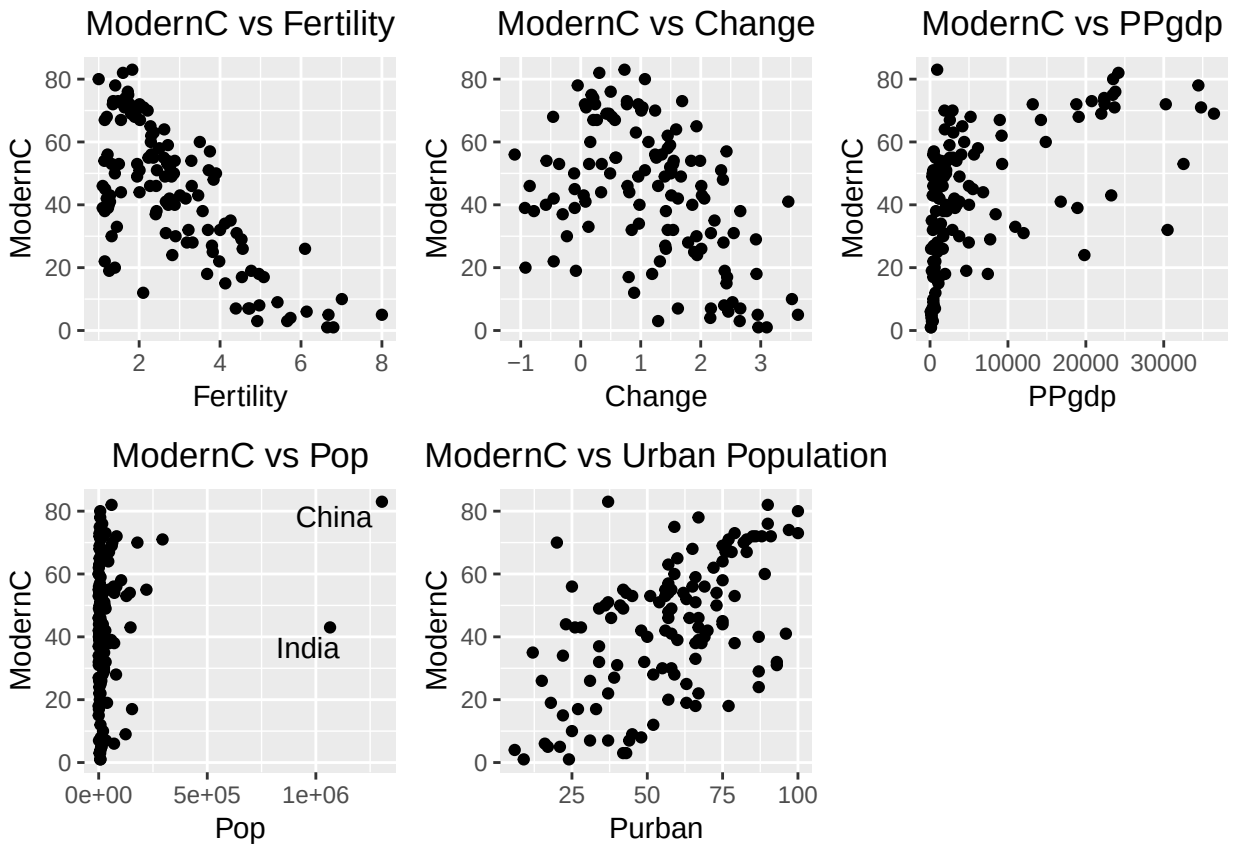
```
UN3_test=UN3
UN3_test$country=row.names(UN3)
ModC_Pop=ggplot(na.omit(UN3_test),aes(y=ModernC,x=Pop)) + geom_point()+theme(plot.title = element_text(
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

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



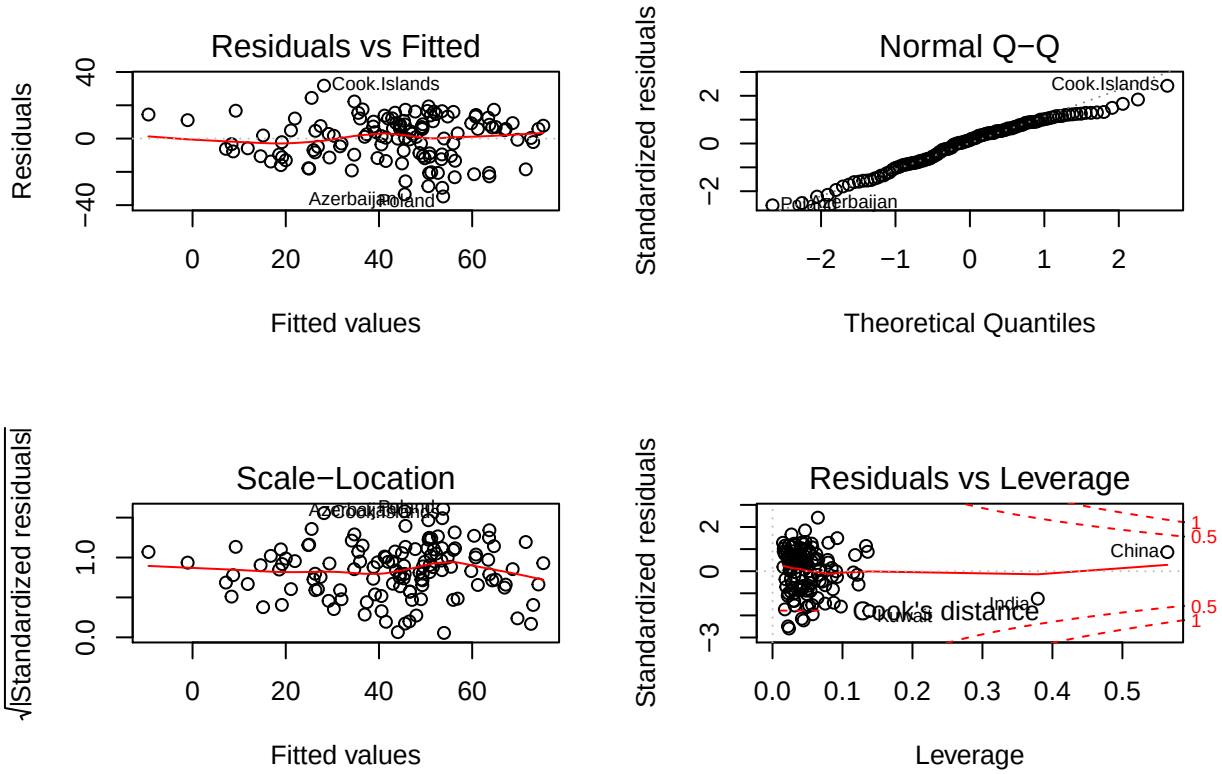
Model Fitting

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.

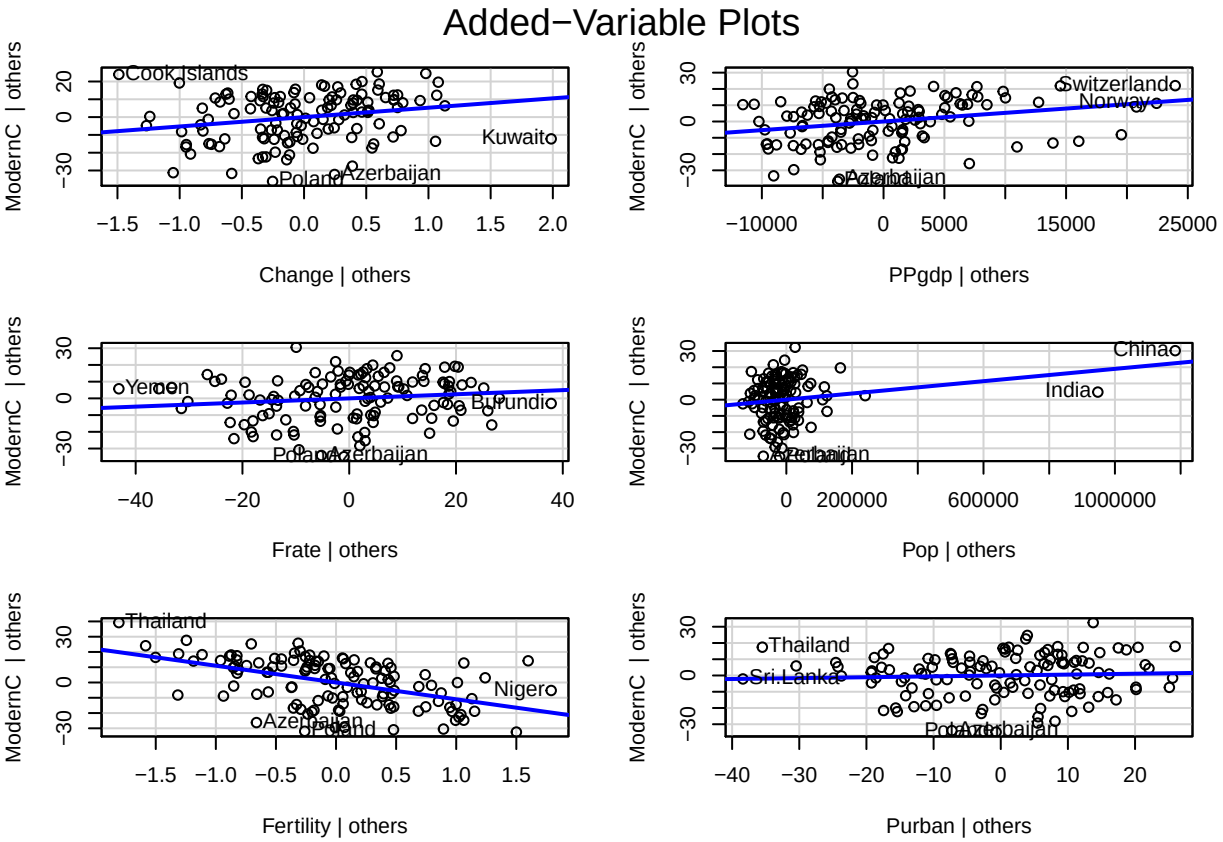
The Residuals-Fitted and Scale-Location plots suggest that the linearity assumptions isn't completely valid. The slight bulge in both plots suggests this. The Normal-QQ plot suggests that normality isn't valid here, the data appears to be left skewed. The final plot suggests that India and China have 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?

```
avPlots(model)
```



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. We might not need these in the model.

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

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.

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

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.

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?

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.

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!

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

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.*

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.