HW2 STA521 Fall18

Zheng Yuan zy87 github loveyuanzheng

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*

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

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

```r
summary(UN3)
is.na(UN3)
sapply(UN3,class)
```

The result implies that there are 6 variables have missing data. All of the variables are quantitative.

2. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.
   Here, in order to be simple and clear, mean and standard deviation are rounded to two significant digits.

```r
library(knitr)
MeanSD<-sapply(UN3, function(x){signif(c(mean(x,na.rm=TRUE), sd(x,na.rm=TRUE)),2)})
rownames(MeanSD)<-c("mean","standard deviation")
kable(MeanSD,row.names = TRUE)
```

	ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
mean	39	1.4	6500	48	30000	3.2	56
standard deviation	23	1.1	9300	17	120000	1.7	24

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, we create a scatterolots of the predictors.

```
library(dplyr)
```

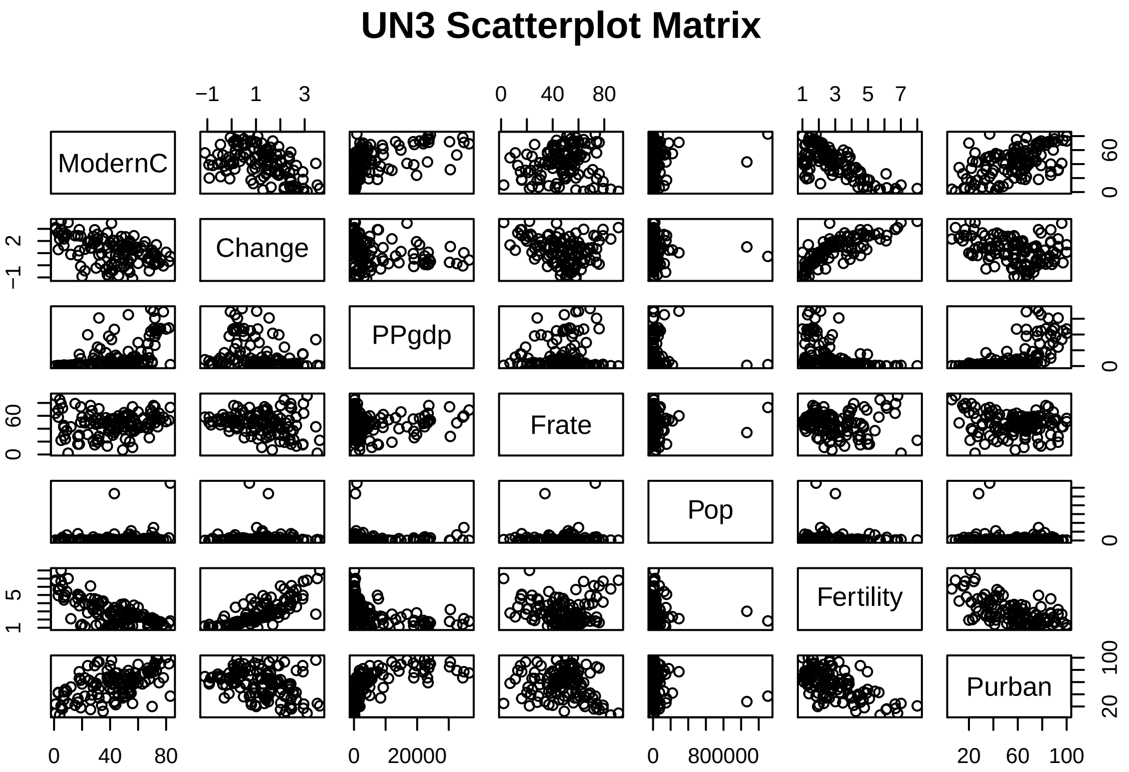
```
##
## Attaching package: 'dplyr'

## 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
```

```
pairs(~ModernC+Change+PPgdp+Frate+Pop+Fertility+Purban,data=UN3,
   main="UN3 Scatterplot Matrix",na.action = "na.omit")
```



The scatterplot matrix above implies that there exists an exponential relationship between PPgdp and ModernC as well as Fertility and ModernC because the plots of them appear to be exponential shape. Therefore, transfering PPgdp and Fertility into log(PPgdp) and log(Fertility) appear to be needed in this case. Besides, there seems to be no linear relationship between ModernC, Frate and Pop.

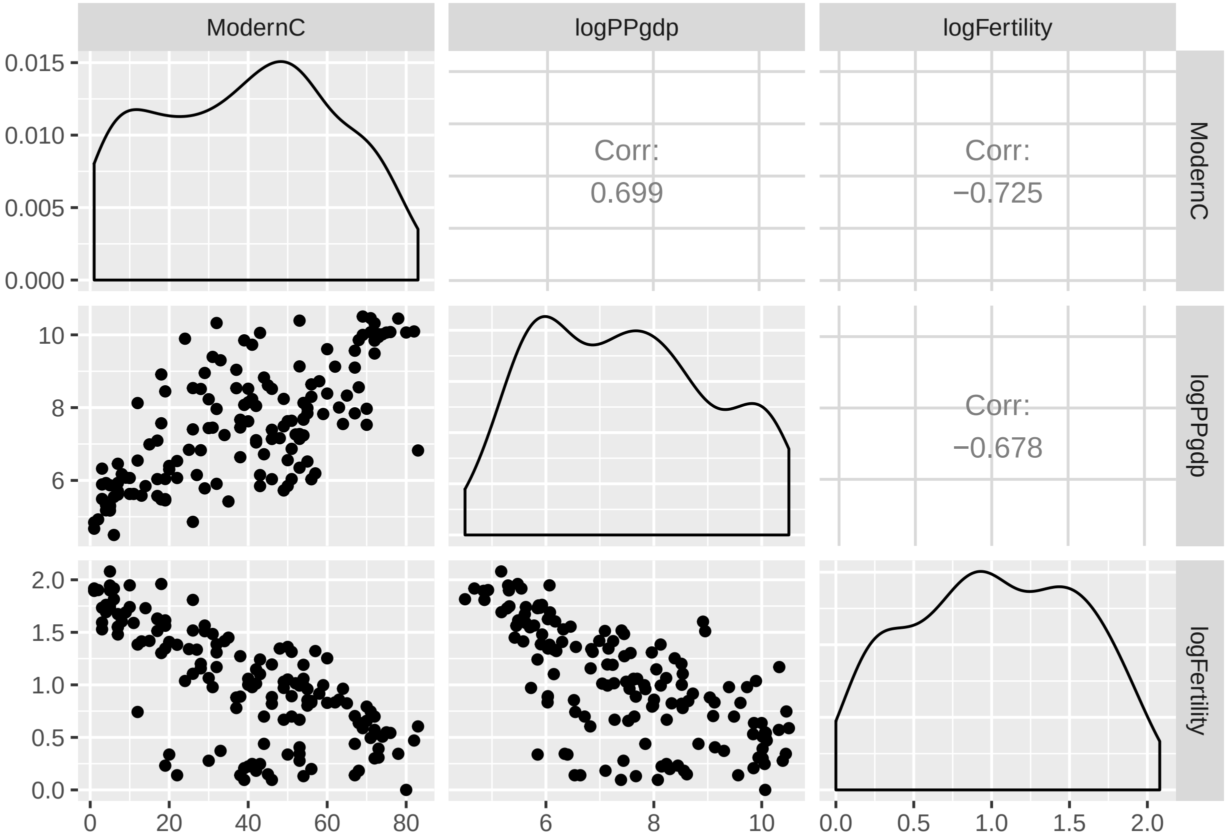
```
library(dplyr)
UN = dplyr::select(UN3, c(ModernC,Fertility, PPgdp)) %>%
  mutate(logPPgdp = log(PPgdp),
         logFertility = log(Fertility)) %>%
  na.omit()
library(GGally)
```

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

```
##
## Attaching package: 'GGally'

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

```
ggpairs(UN, c(1,4,5))
```



The plot above shows that after transfering PPgdp and Fertility into log(PPgdp) and log(Fertility) respectively, there appears to be a linear relationship bewtween ModernC and log(PPgdp) as well as log(Fertility).

Model Fitting

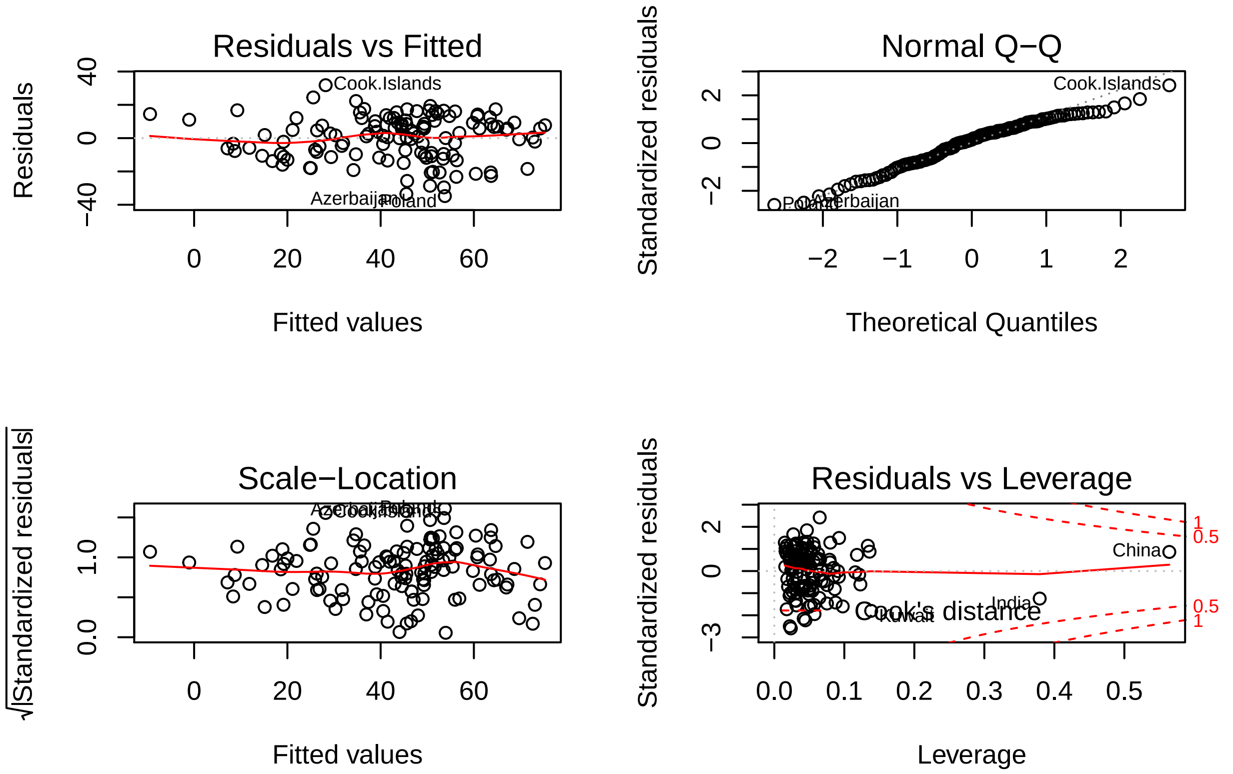
4. Use the `lm()` function to perform a multiple linear regression with `ModernC` as the response and all other variables as the predictors, using the formula `ModernC ~ .`, where the `.` includes all remaining variables in the dataframe. Create diagnostic residual plot from the linear model object and comment on results regarding assumptions. How many observations are used in your model fitting?

```
UN1<-UN3%>%
  na.omit()
fit1<-lm(ModernC~.,data=UN1)
summary(fit1)
```

```
##
## Call:
## lm(formula = ModernC ~ ., data = UN1)
##
```

```
## 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
```

```r
par(mfrow=c(2,2))
plot(fit1, ask=F)
```



First, there exists a linear relationship residuals and fitted values, which means that our linear model fits the data.

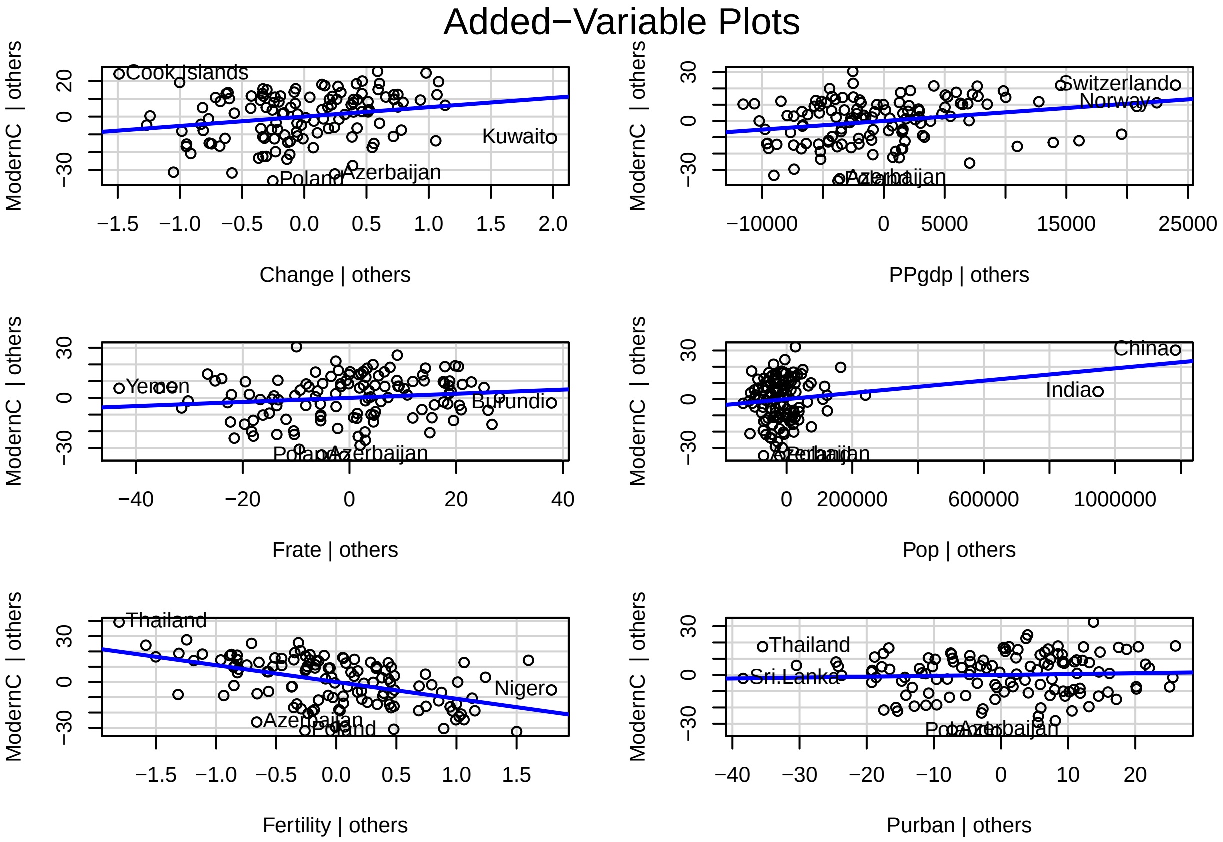
Next, in the second plot, standardized residuals are lined well on the straight dashed line despite some outliers, this implies that the residuals are normally distributed.

Then, the scale-location plot shows that residuals are not spread equally along the ranges of predictors, the residuals begin to spread wider along the x-axis before it passes 50.

Lastly, since there are not any points outside the dashed line, Cook Distance, there are no influential points

in this case.
By the way, 125 observations are used in my model fitting.
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(fit1)
```



The plots above imply that Kuwaito, Poland and Azerbajian are influential for “Change”; Poland and Azerbajian are influential for “PPgdp”,“Frate”,“Fertility” and “Purban”; “Azerbajian” is also influential for Pop.
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.

```
boxTidwell(ModernC~PPgdp+Pop+Fertility,~Change^2+Purban+Frate,data=UN1)
```

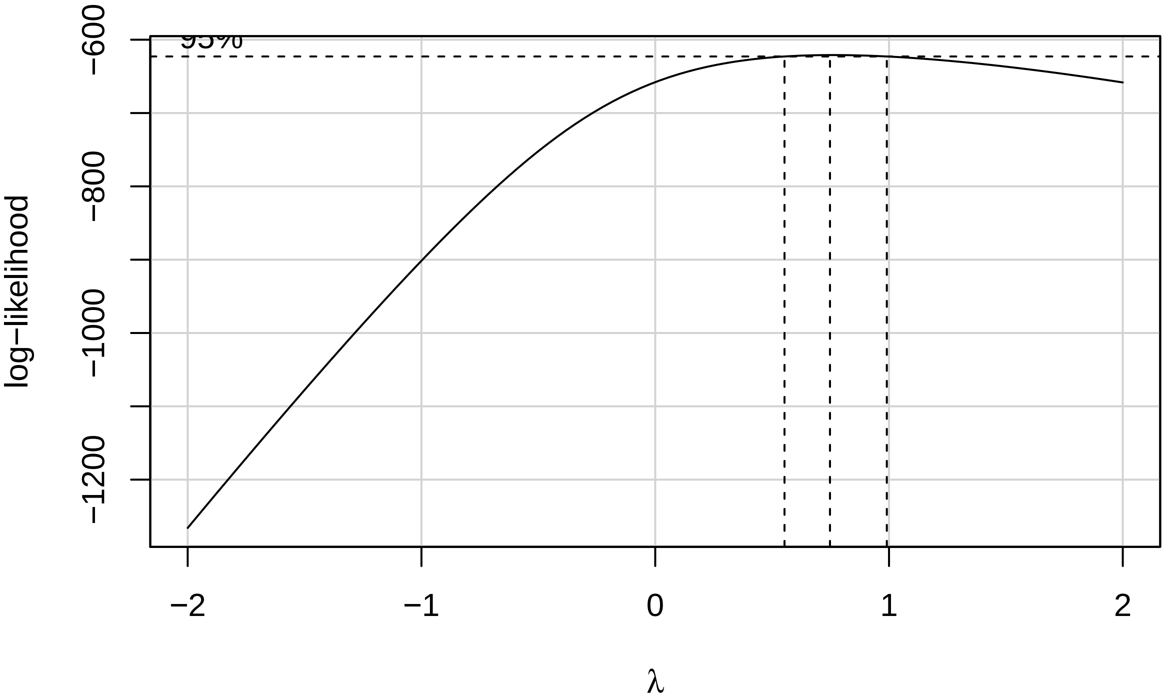
```
##            MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp          -0.035767             -1.2324   0.2178
## Pop             0.374984             -0.9042   0.3659
## Fertility       1.346874             -1.7985   0.0721 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1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

First, from the scatterplot matrix, I realise that the predictors which potentially need transformation are PPgdp, Pop and Fertility. Besides,since predictor Change has some negative values, I’d better transform it into its squared. Therefore, I use the boxTidewell function as above. According to the result, the MLE of lambda for PPgdp and Pop is nearly 0, which means we should transfer PPgdp into log(PPgdp) and tranfer

Pop into log(Pop).

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

```r
fit2<-lm(ModernC~log(PPgdp)+log(Pop)+Fertility+Change^2+Purban+Frate,data=UN1)
boxCox(fit2, plotit=TRUE)
```



```r
summary(fit2)
```

```
##
## Call:
## lm(formula = ModernC ~ log(PPgdp) + log(Pop) + Fertility + Change^2 +
##     Purban + Frate, data = UN1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.597  -9.540   2.238  10.024  34.840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.11547   14.50854   0.284 0.777169
## log(PPgdp)   5.50728    1.40505   3.920 0.000149 ***
## log(Pop)     1.47207    0.62875   2.341 0.020897 *
## Fertility   -9.67594    1.76561  -5.480 2.44e-07 ***
## Change       4.99296    2.07709   2.404 0.017781 *
## Purban      -0.07077    0.09760  -0.725 0.469829
## Frate        0.18939    0.07711   2.456 0.015500 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.44 on 118 degrees of freedom
## Multiple R-squared:  0.626,  Adjusted R-squared:  0.6069
## F-statistic: 32.91 on 6 and 118 DF,  p-value: < 2.2e-16
```

The plot show that the optimal λ here is about 0.8, so the according transformation will be $\frac{ModernC^{0.8}-1}{0.8}$
Then we will justify it.

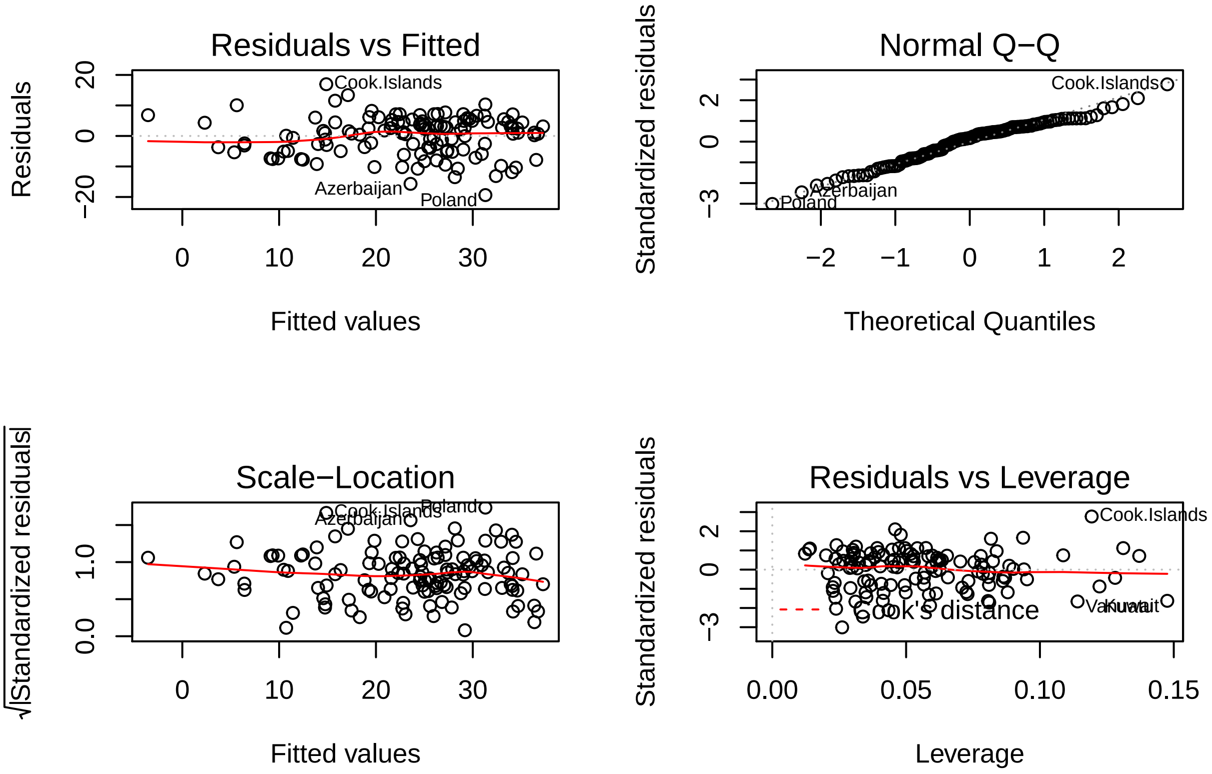
```
fit3<-lm(((ModernC^(0.8)-1)/0.8)~log(PPgdp)+log(Pop)+Fertility+Change+Purban+Frate,data=UN1)
summary(fit3)
```

```
##
## Call:
## lm(formula = ((ModernC^(0.8) - 1)/0.8) ~ log(PPgdp) + log(Pop) +
##     Fertility + Change + Purban + Frate, data = UN1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.375  -4.725   1.088   4.496  16.950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.74187    7.04119   1.242 0.216871
## log(PPgdp)   2.57082    0.68189   3.770 0.000256 ***
## log(Pop)     0.62411    0.30514   2.045 0.043047 *
## Fertility   -5.20407    0.85688  -6.073 1.57e-08 ***
## Change       2.61292    1.00804   2.592 0.010745 *
## Purban      -0.04282    0.04737  -0.904 0.367861
## Frate        0.07443    0.03742   1.989 0.049021 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.524 on 118 degrees of freedom
## Multiple R-squared:  0.6402, Adjusted R-squared:  0.6219
## F-statistic: 34.99 on 6 and 118 DF,  p-value: < 2.2e-16
```

Compared to the original model, the adjusted R-squared value does not change too much.

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.

```
fit3<-lm(((ModernC^(0.8)-1)/0.8)~log(PPgdp)+log(Pop)+Fertility+Change^2+Purban+Frate,data=UN1)
par(mfrow=c(2,2))
plot(fit3, ask=F)
```



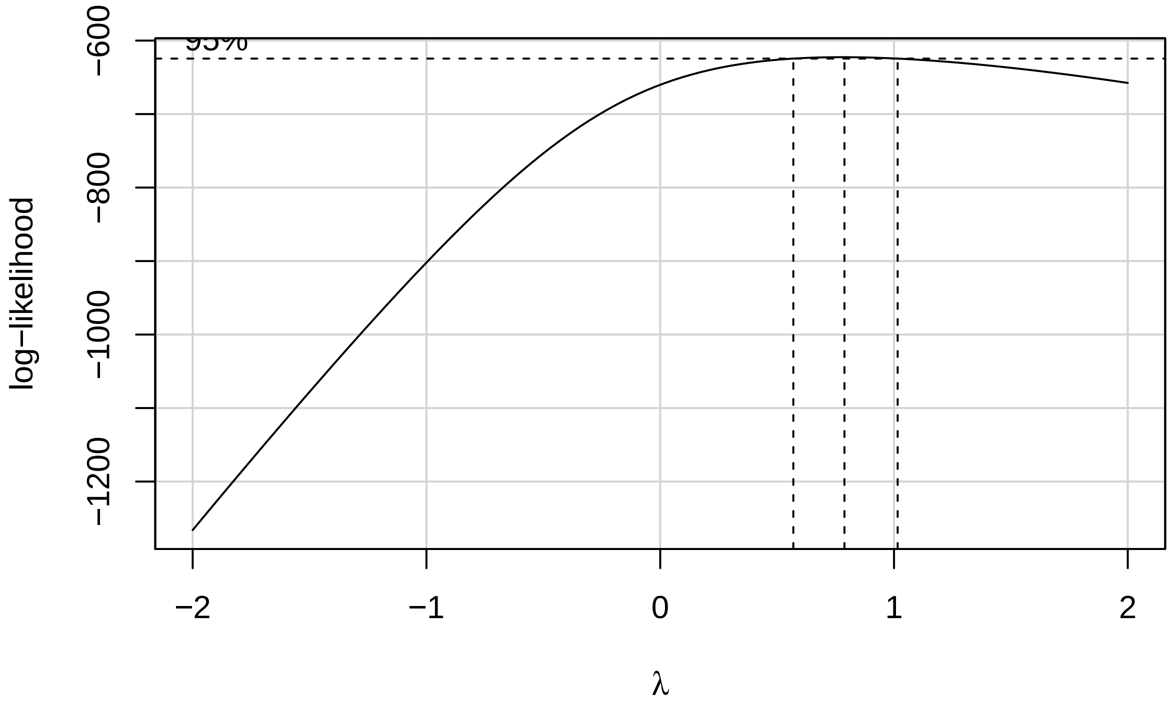
First, there exists a linear relationship residuals and fitted values, which means that our linear model fits the data.

Next, in the second plot, standardized residuals are lined well on the straight dashed line despite some outliers, this implies that the residuals are normally distributed.

Then, in the third plot, the points are now randomly distributed on both sides of a nearly horizonal line, better than before, which implies that residuals are spread equally along the ranges of predictors.

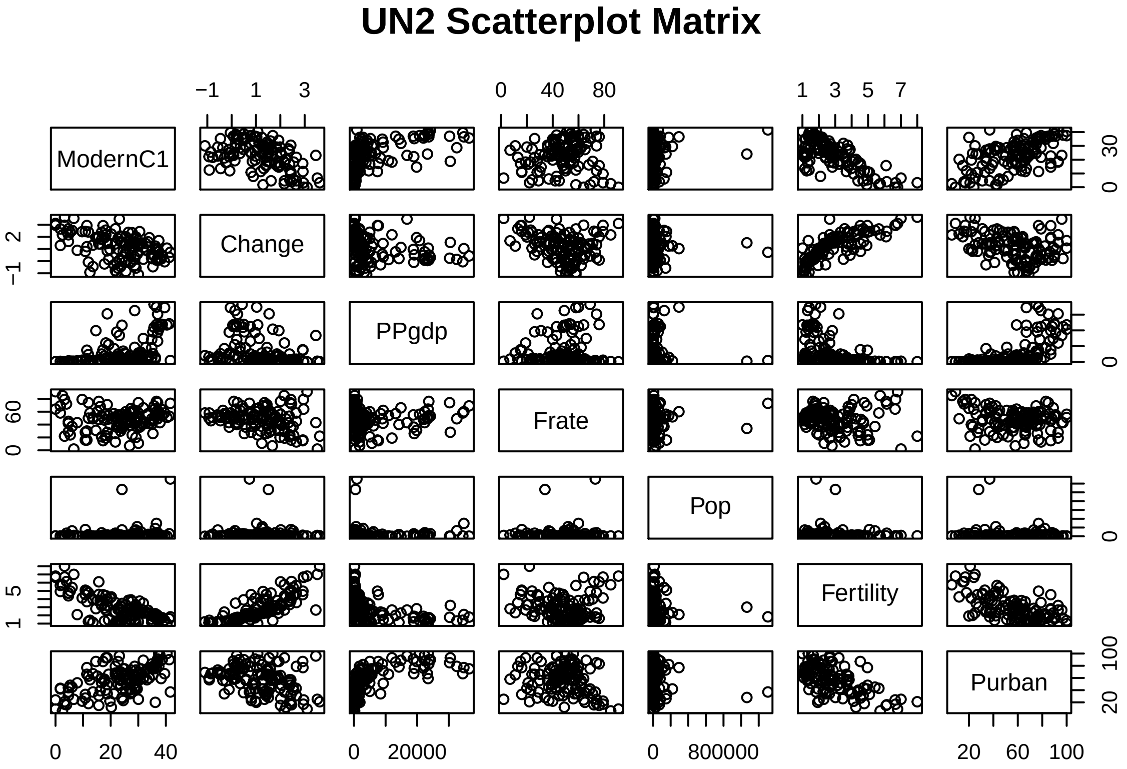
9. Start by finding the best transformation of the response and then find transformations of the predictors. Do you end up with a different model than in 8?

```
fit4<-lm(ModernC~PPgdp+Pop+Fertility+Change+Purban+Frate,data=UN1)
boxCox(fit4, plotit=TRUE)##First, we look for the bset transformation of the responseby boxcox
```



The plot shows that the optimal λ for response is still about 0.8.
Then we start to look for the transformations of the predictors.

```
library(dplyr)
UN2=mutate(UN1,ModernC1=(ModernC^(0.8)-1)/0.8)#Transform the response at first
pairs(~ModernC1+Change+PPgdp+Frate+Pop+Fertility+Purban,data=UN2,
   main="UN2 Scatterplot Matrix")##Plot the scatterplot matrix
```



In fact, the shape of scatterplot has nothing different from that of original dataset.Therefore, I use the same form of boxTidewell function.

```
boxTidwell(ModernC1~PPgdp+Pop+Fertility,~Change^2+Purban+Frate,data=UN2)
```

```
##           MLE of lambda Score Statistic (z) Pr(>|z|)
## PPgdp          -0.10022             -1.2716  0.20353
## Pop             0.36613             -0.9210  0.35703
## Fertility       1.41657             -2.2705  0.02318 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  18
```

The result shows that we end up with a same model compared with Problem 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.

First thing to do is to detect outliers and influential points using "outlierTest" function in car package.

```
library(car)
outlierTest(fit3)
```

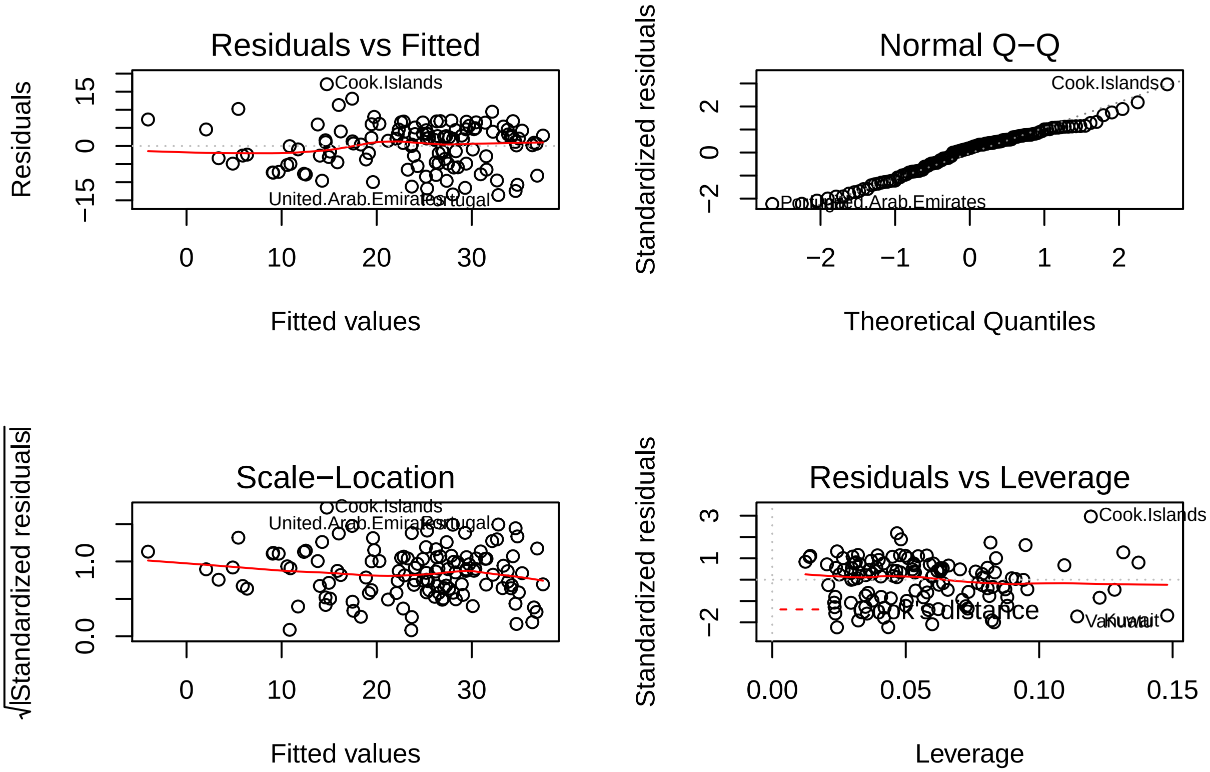
```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##         rstudent unadjusted p-value Bonferonni p
## Poland -3.118425          0.0022891      0.28614
```

The result shows that we'd better remove "Poland" from our dataset because it's the most extreme outliers. Besides, as we detect above, Azerbaijan is also an outlier, we'd better remove it too.

```
UN4 <- subset(UN1,!UN1$Pop%in%c(38588,8370))
```

Then we refit our model with new dataset.

```
fit5<-lm(((ModernC^(0.8)-1)/0.8)~log(PPgdp)+log(Pop)+Fertility+Change+Purban+Frate,data=UN4)
par(mfrow=c(2,2))
plot(fit5,ask=F)
```



As we can see in the Fitted values vs Residuals plots, points in it seems to be closer to zero line than before, which means our new model fits better after we delete the outliers. ## 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!

```
library(knitr)
coe<-confint(fit5)
kable(coe,row.names =TRUE,digits=2,caption="Coefficients with 95% confidence intervals")
```

Table 2: Coefficients with 95% confidence intervals

	2.5 %	97.5 %
(Intercept)	-1.84	24.73
log(PPgdp)	1.11	3.67
log(Pop)	0.09	1.23
Fertility	-7.09	-3.87
Change	0.67	4.43
Purban	-0.13	0.04
Frate	0.00	0.14

Interpretation:
To interpret this, first let ModernC1=$\frac{ModernC^{0.8}-1}{0.8}$ and ModernC'=$exp(ModernC1)$, so when PPgdp=Pop=1 and other predictors are all 0, ModernC1 would be in $e^{-1.84}$ to $e^{24.73}$ with a probablity of 0.95; 10% increase in PPgdp would cause $1.1^{1.11}-1$ to $1.1^{3.67}-1$ increase in ModernC';10% increase in Pop would cause $1.1^{0.09}-1$ to $1.1^{1.23}-1$ increase in ModernC'; 1 unit increase in Fertility would cause $1-e^{-7.09}$ to$1-e^{-3.87}$ decrease in ModernC'; 1 unit increase in Change would cause $e^{0.67}-1$ to $e^{4.43}-1$ increase in ModernC'; 1

unit increase in Purban would cause $1 - e^{-0.13}$ decrease to $e^{0.04} - 1$ increase in ModernC';1 unit increase in Frate would cause up to $e^{0.04} - 1$ increase in ModernC'.

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

First, in the original dataset, almost every variable has missing values, therefore, we would omit those observations with missing values. Then, after plotting and analysis, we find that "Poland" and "Azerbeijan" are extreme outliers and may influence the model very much so they are also removed from the dataset. After some anaysis, we make some transformations to some predictors as well as response, and then obtain the final model, that is,

$$\frac{ModernC^{0.8} - 1}{0.8} = 11.45 + 2.39 * log(PPgdp) + 0.66 * log(Pop) - 5.48 * Fertility + 2.55 * Change - 0.05 * Purban - 0.07 * Frate$$

The interpretation of each coefficient is in the Problem 11. The summary of the model shows that ModernC has strong relationship with log(PPgdp), log(Pop), Fertility and Change, so these variables are the important factors that decide the percent of unmarried women while Purban and Frate are not.

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.
    Proof:
    First, let the residual from regressing Y on all the predictors except for X_j be $e_{(y)}$ and the residual from regressing X_j on all the predictors be $e_{(X_j)}$ Then our regression model would be

$$\begin{aligned}
e_{(y)} &= \hat{\beta}_0 + \hat{\beta}_1 e_{(X_j)} \\
(I - H)Y &= \hat{\beta}_0 + \hat{\beta}_1 (I - H)X_j \\
(I - H)Y &= \hat{\beta}_0 1_n + [X_j^T (I - H)(I - H)X_j]^{-1}((I - H)X_j)^T (I - H)Y(I - H)X_j \\
(I - H)Y &= \hat{\beta}_0 1_n + (X_j^T (I - H)X_j)^{-1} X_j^T (I - H)Y(I - H)X_j \\
X_j^T (I - H)Y &= X_j^T \hat{\beta}_0 1_n + X_j^T (X_j^T (I - H)X_j)^{-1} X_j^T (I - H)Y(I - H)X_j \\
X_j^T (I - H)Y &= X_j^T 1_n \hat{\beta}_0 + X_j^T (I - H)X_j (X_j^T (I - H)X_j)^{-1} X_j^T (I - H)Y \\
X_j^T (I - H)Y &= X_j^T 1_n \hat{\beta}_0 + X_j^T (I - H)Y \\
X_j^T 1_n \hat{\beta}_0 &= 0 \\
\sum_{i=1}^{n} X_j^{(i)} \hat{\beta}_0 &= 0 \\
\hat{\beta}_0 &= 0
\end{aligned}$$

    Next we prove that the sample mean of residuals will always be zero if there is an intercept, note that $\sum e$ can be written as $1_n^T e$ which equals $1_n^T(I - H)Y$. Since intercept is included here by assumption, $(I - H)$ is the project matrix which contains a column of ones, then $1_n^T(I - H) = 0$. It follows that $1_n^T e = 1_n^T(I - H)Y = 0$ Therefore, the sample will always be 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.

First, let us recall the model we fit in Ex.10.

```
summary(fit5)
```

The final model is

$$\frac{ModernC^{0.8}-1}{0.8}=11.45+2.39*log(PPgdp)+0.66*log(Pop)-5.48*Fertility+2.55*Change-0.05*Purban-0.07*Frate$$

Here let us take the predictor Fertility for an example, the estimate for the slope of it in our full model is -5.48. Then let us construct the added variable model in two steps.

First, let us make regression of $\frac{ModernC^{0.8}-1}{0.8}$ on all of the predictors except for Fertility and then we extract the residual.

```
fit6<-lm(((ModernC^(0.8)-1)/0.8)~log(PPgdp)+log(Pop)+Change+Purban+Frate,data=UN4)
res1<-residuals(fit6)
```

Then we make regression of Fertility on the other predictors and also extract the residual.

```
fit7<-lm(Fertility~log(PPgdp)+log(Pop)+Change+Purban+Frate,data=UN4)
res2<-residuals(fit7)
```

Finally we make regression of res1 on res2 and see its coefficients

```
fit8<-lm(res1~res2)
summary(fit8)
```

According to the rusult, the slope of our manually constructed added variable plot for predictor Fertility is -5.48, which is the same as the estimate from our model.