HW2 STA521 Fall18

[Zeren Li, zl129 and zerenli1992]

Due September 19, 2018

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*

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

ModernC, Changem PPgdp, Frate, Pop,Fertility, Purban are quantitative.

Six out of seven variable, including ModernC, Change, PPgdp, Frate, Pop, Fertility have missing data, the amount of missing is shown in the table.

```
#summary
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
```

```
#missing data
map_df(UN3, function(x) sum(is.na(x)))
```

```
## # A tibble: 1 x 7
##   ModernC Change PPgdp Frate   Pop Fertility Purban
##     <int>  <int> <int> <int> <int>     <int>  <int>
```

```
## 1      58      1     9     43     2         10       0
```

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

```
#construct mean variable
mean <- map_df(UN3, function(x)
       {
        x %>%
          na.exclude() %>%
          mean()
       }
        )

#construct sd variable
sd <-  map_df(UN3, function(x)
       {
        x %>%
         na.exclude() %>%
        mean()
        }
        )

#summary
rbind( mean, sd)  %>%
  t.data.frame() %>%
  kable(col.names     = c("mean", "sd") )
```

	mean	sd
ModernC	38.717105	38.717105
Change	1.418373	1.418373
PPgdp	6527.388060	6527.388060
Frate	48.305389	48.305389
Pop	30281.871428	30281.871428
Fertility	3.214000	3.214000
Purban	56.200000	56.200000

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 `modern` from the other variables. Are there potential outliers, nonlinear relationships or transformations that appear to be needed based on your graphical EDA?

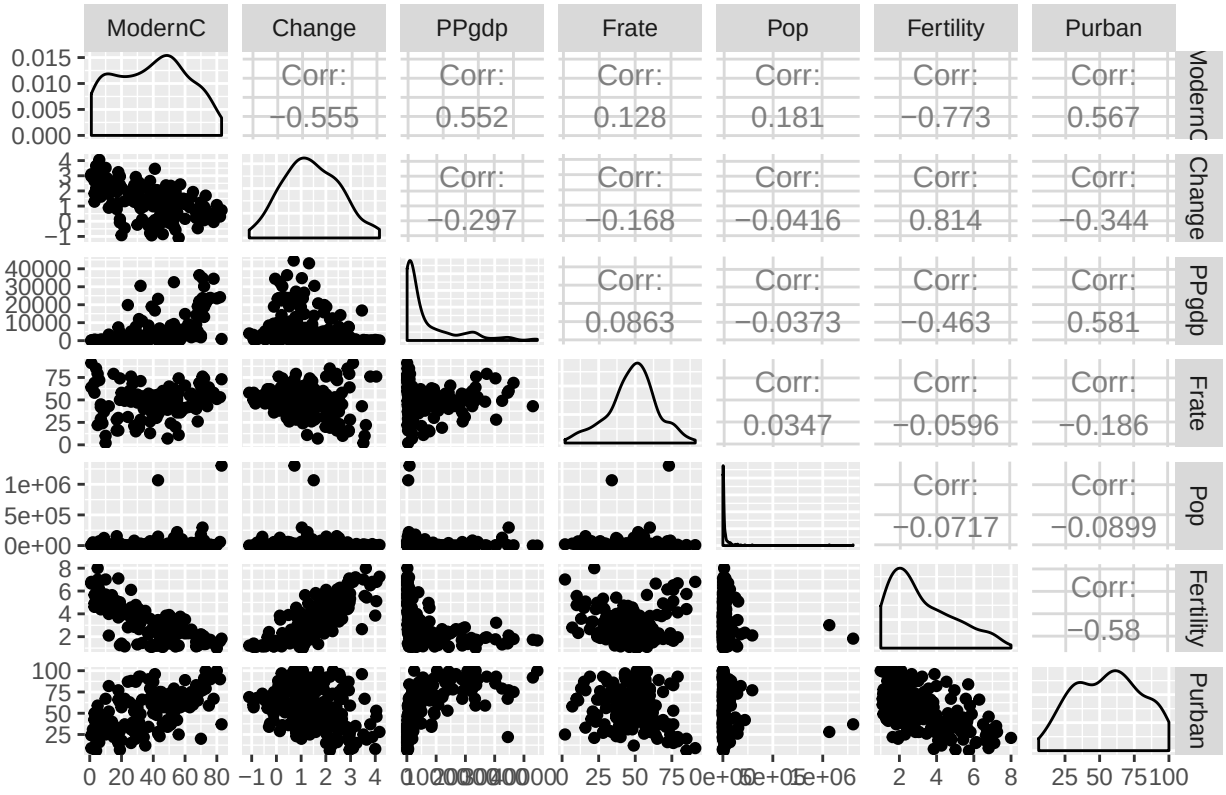
4. `ModernC` is correlated with `PPgdp`, `Change`, `Fertility`, and `Purban` over .5. The scatterplot also shows that `ModernC` has a negative correlation with Change and Fertility, a possible correlation with `PPgdp` and Purban.

5. There are two obvious outliers in Pop, which with value over 1e+06.

6. `Pop`, `Fertility` and `PPgdp` are very right-skewed, which requires further transformation. 4. Frate and `ModernC` has a U-shape relationship.

```
UN3 %>%
ggpairs(.,title = "correlation table" )
```

correlation table



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?

1) Residuals vs Fitted plot shows a very weak non-linear pattern and observation Azerbaijan and Cook's Islands drive this pattern.

2) Normal QQ plot shows that the residuals are normally distributed, besides the distortion effect of two observations Azerbaijan and Cook.Islands.

3) ScaleLocation plot shows that residuals are spread equally along the ranges of predictors as the fitted line is flat. The results support the homoscedasticity assumption.

4) Residuals vs Leverage plot shows that China, India Cook's Island are the three most influential cases, which may drive the regression result.

5) 125 observation were used in the regression.

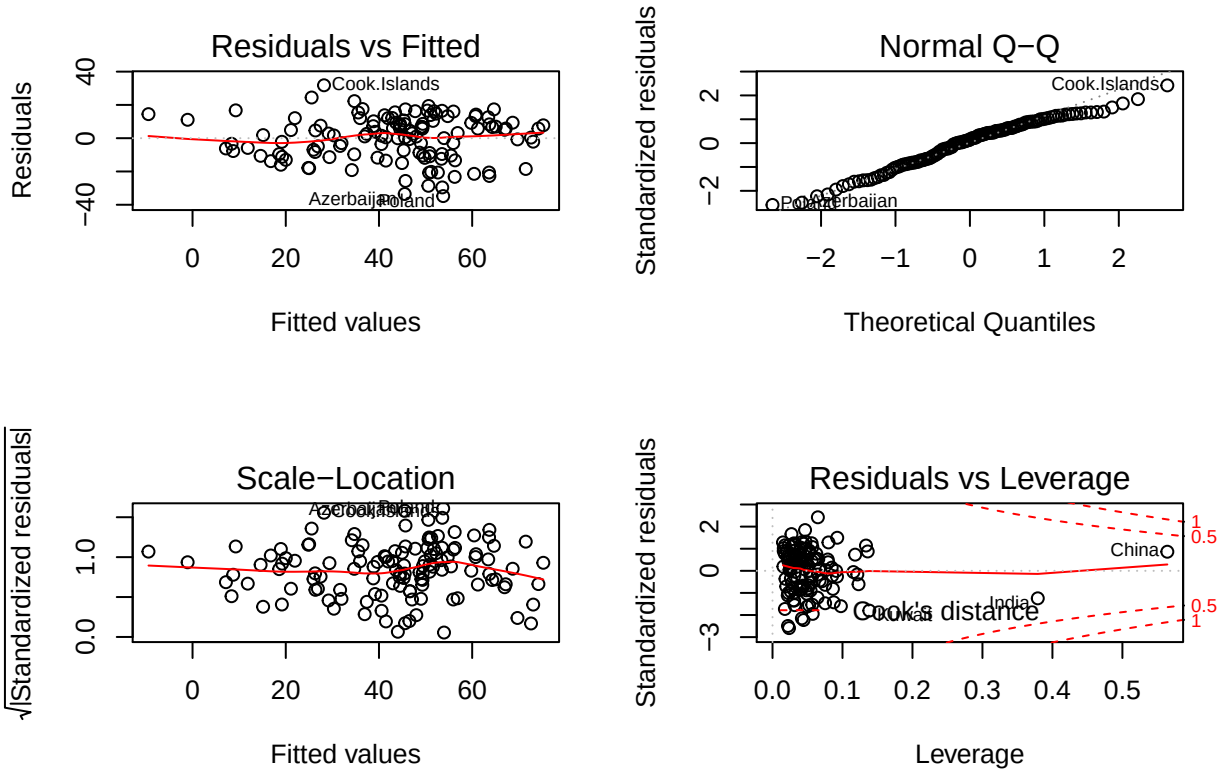
```
m1 <- lm(ModernC ~ ., data = UN3 )


# show number of observation in the regression 210 - 85 = 125
summary(m1)
```

```
##
## Call:
```

```
## lm(formula = ModernC ~ ., data = UN3)
##
## 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
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5989
## F-statistic: 31.85 on 6 and 118 DF,  p-value: < 2.2e-16
```

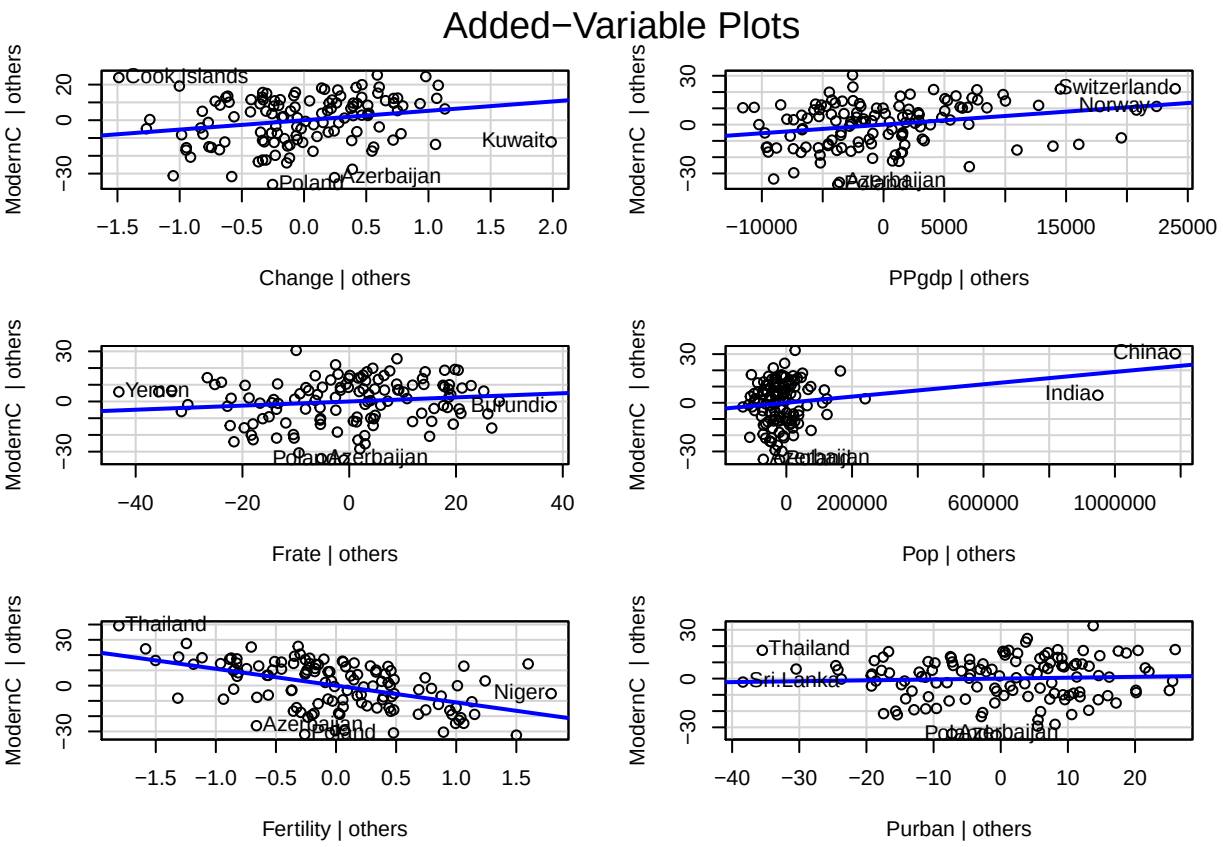
```
par(mfrow=c(2,2))
plot(m1)
```



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?

1) The direction of correlation in added variables plot is consistent with that in the previous regression.
2) Observation Kuwait drives the regression result between Change and `ModernC`.
3) Observations China and India drive the regression result between Pop and `ModernC`.
4) Norway and Switzerland drive the regression result between PPgdp and `ModernC`.
5) Change has a negative value, thus it requires transformation.
6) `PPgdp` and `Pop` have outliers, thus they require logged transformation.

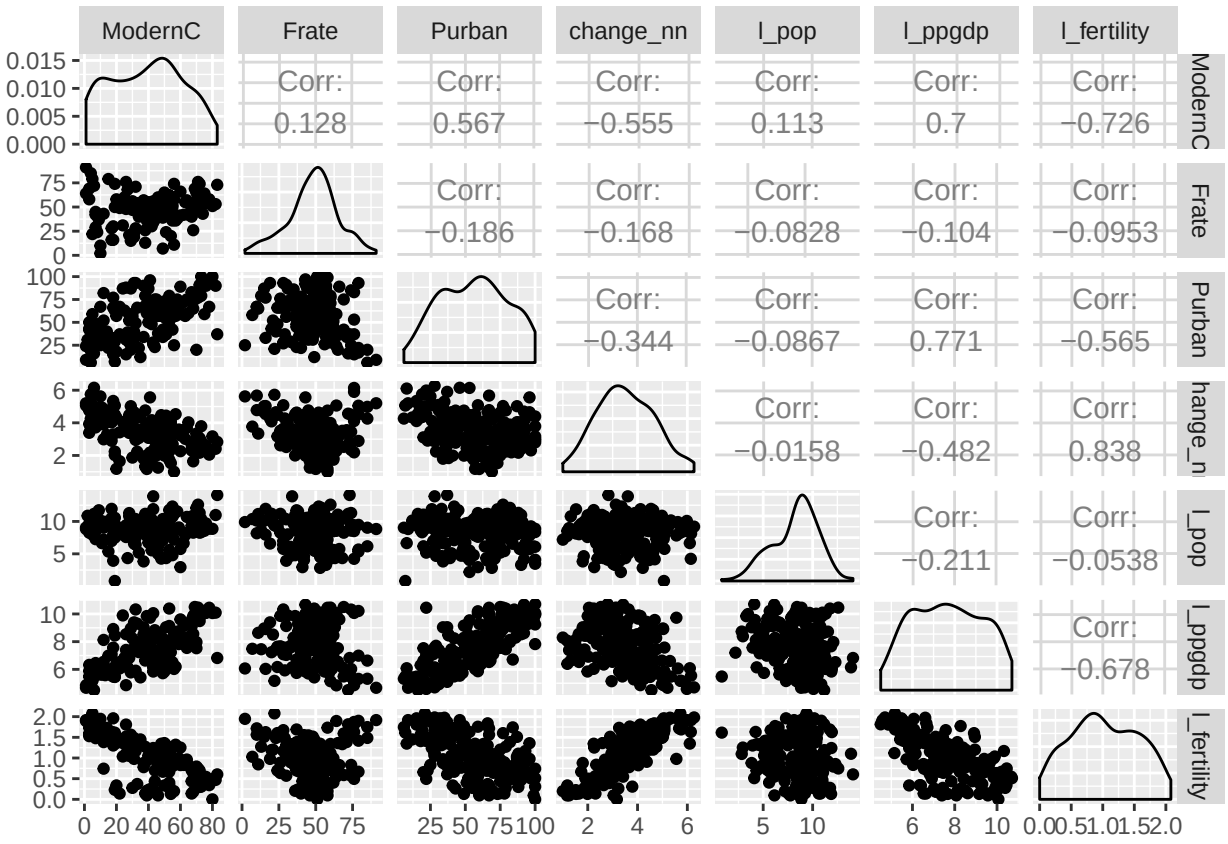
```
avPlots(m1)
```



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.

1) I transferform the variable `Change` by adding 1 and then minusing it by the minimum of change.
2) As the distribution of `Pop`, `PPgdp`,`Fertility` are right-skewed, I use logged transformation on these variables.
3) Using Box-Tidwell and the correlation scatterplot, I find a non-linear relationship between `ModernC` and `l_pop`. I use a quadratic form of `l_pop`.

```
# transform the variable
UN3_new <- UN3 %>%
          mutate(change_nn =  Change + 1 - min(Change, na.rm = T) ,
                 l_pop = log(Pop),
                 l_ppgdp = log(PPgdp),
                 l_fertility = log(Fertility)) %>%
          select(-c("Change", "Pop", "PPgdp", "Fertility"))
# EDA
ggpairs(UN3_new)
```




```
# Tidwell box
boxTidwell(ModernC ~  l_pop  ,  ~ + change_nn + l_ppgdp + l_fertility + Purban + Frate , data=UN3_new, r
```

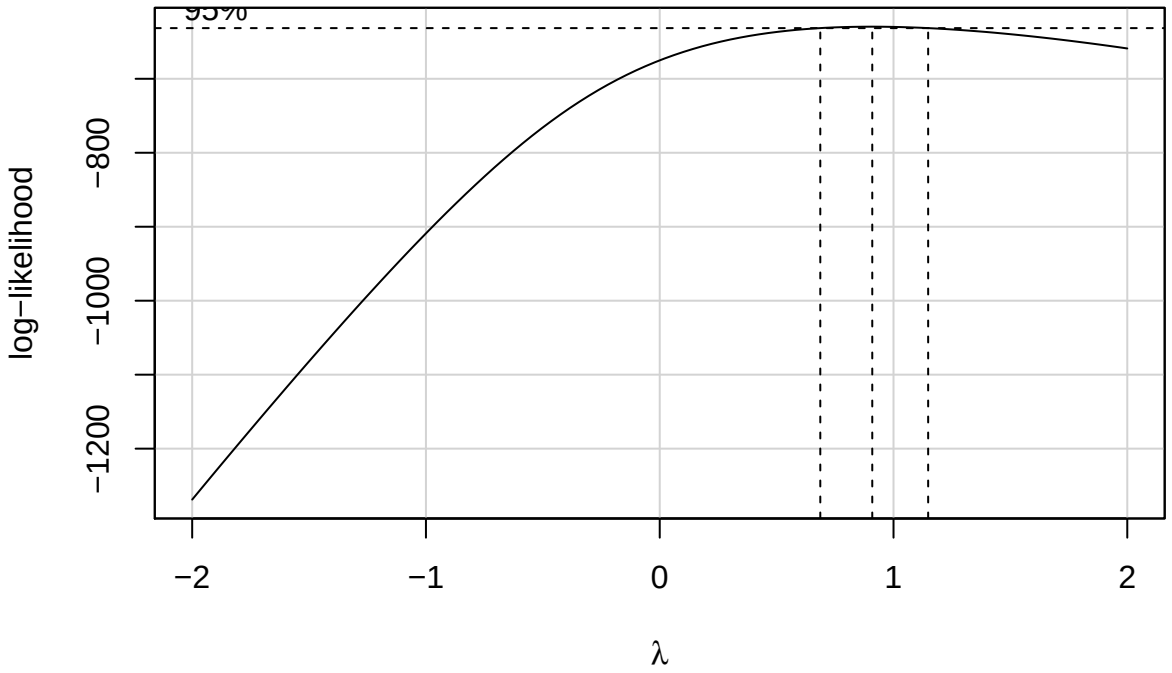
```
##  MLE of lambda Score Statistic (z) Pr(>|z|)
##         6.8611               2.426  0.01526 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  3
```

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

1)I use a quadratic form transformation of l_pop.

2) BoxCox plot shows that we don't need any transformation for our response variable and the point estimate of λ is around 1.

```
m2 <- lm(ModernC ~ poly( l_pop,2)  + change_nn + l_ppgdp + l_fertility + Purban + Frate, data= UN3_new )
boxCox(m2)
```



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.

First I fit m2 with a quadratic form of l_pop in the previous chunk. The quadratic form of l_pop is significant and its residual diagnostics shows that the fitted model is less good than the m1, the original one. I use l_lop without polynomial transformation instead as m3.

```
# fit the model
m3 <- lm(ModernC ~ l_pop + change_nn + l_ppgdp + l_fertility + Purban + Frate, data= UN3_new )

# compare with m2
anova(m2, m3)
```

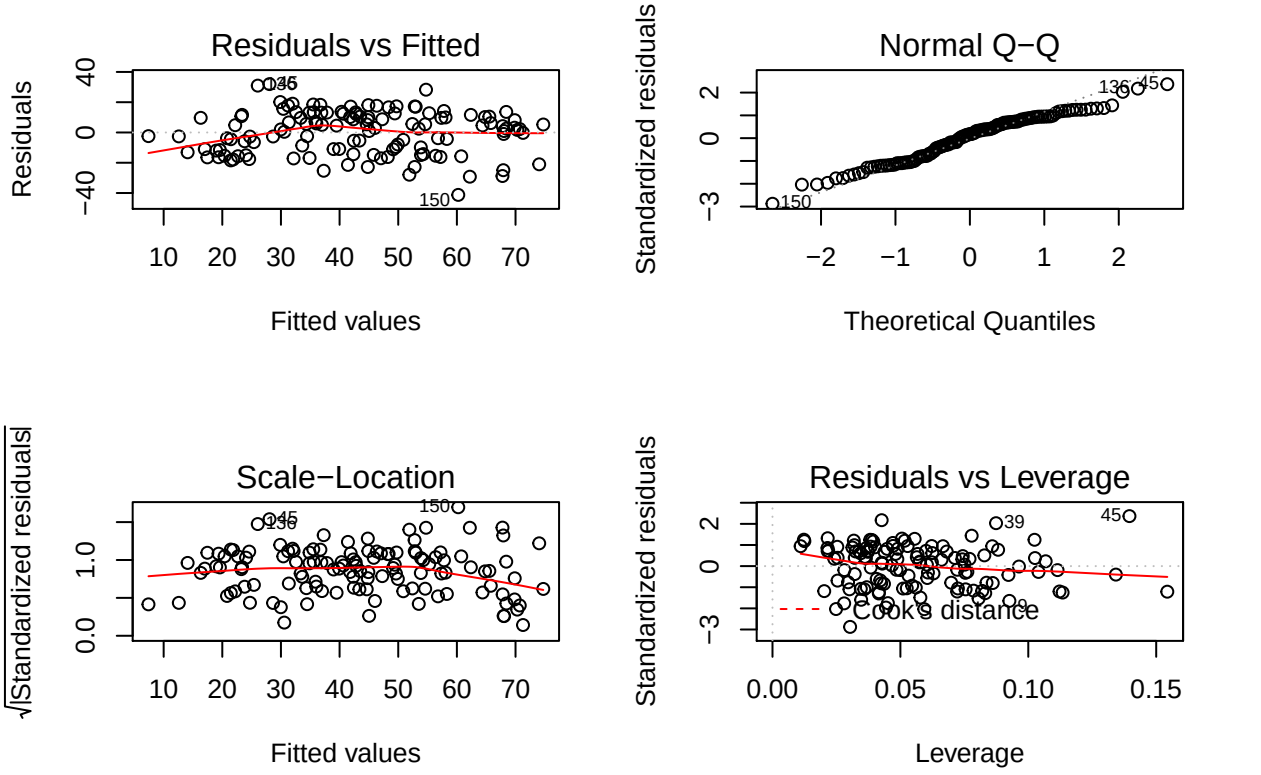
```
## Analysis of Variance Table
##
## Model 1: ModernC ~ poly(l_pop, 2) + change_nn + l_ppgdp + l_fertility +
##     Purban + Frate
## Model 2: ModernC ~ l_pop + change_nn + l_ppgdp + l_fertility + Purban +
##     Frate
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    117 23816
## 2    118 24998 -1   -1181.5 5.8043 0.01755 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m3)
```

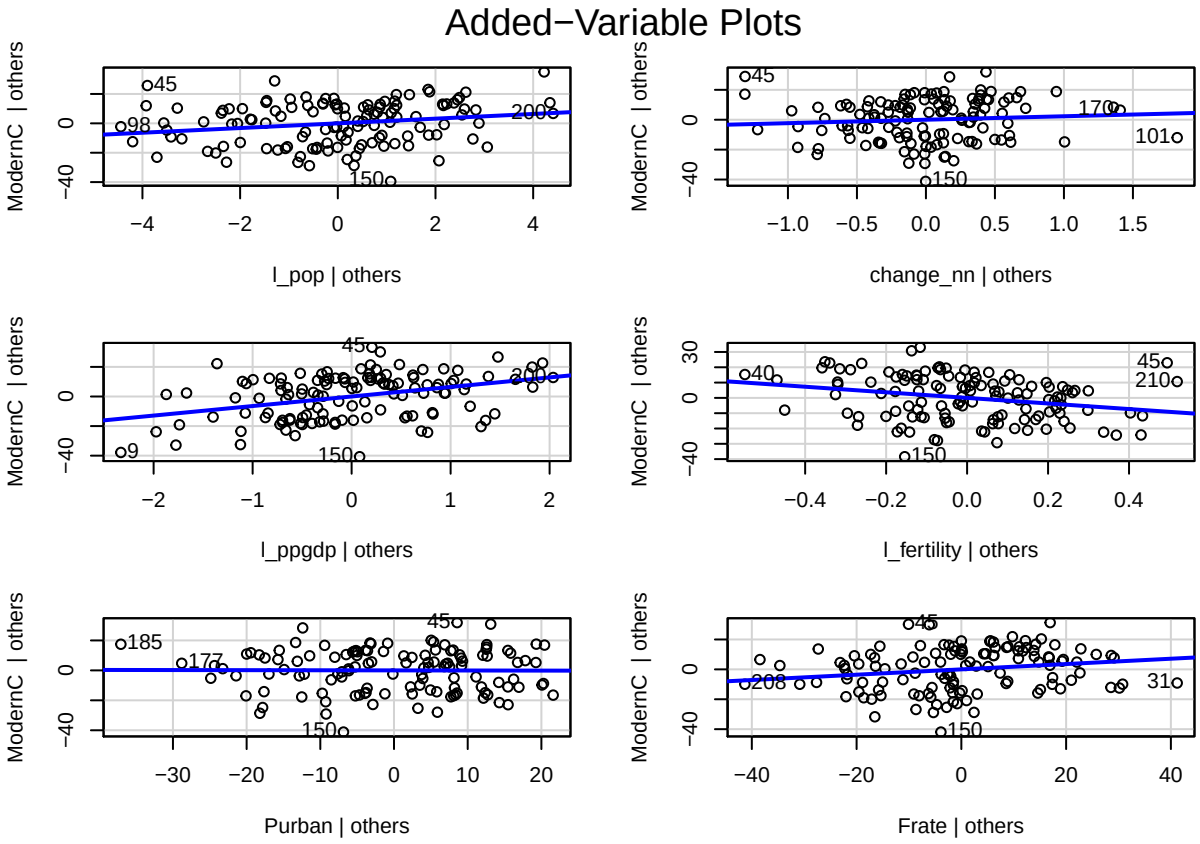
```
##
## Call:
## lm(formula = ModernC ~ l_pop + change_nn + l_ppgdp + l_fertility +
##     Purban + Frate, data = UN3_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -41.235 -11.589   2.498  10.748  31.954
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.958066  15.253221  -1.308  0.19326
## l_pop         1.595611   0.699289   2.282  0.02430 *
## change_nn     2.310274   2.560728   0.902  0.36879
## l_ppgdp       6.445713   1.508057   4.274 3.91e-05 ***
## l_fertility -18.237639   6.336680  -2.878  0.00475 **
## Purban       -0.007352   0.106591  -0.069  0.94513
## Frate         0.178242   0.083567   2.133  0.03500 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.55 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.5615, Adjusted R-squared:  0.5392
## F-statistic: 25.19 on 6 and 118 DF,  p-value: < 2.2e-16
```

```r
# residual plot
par(mfrow=c(2,2))
plot(m3)
```




```r
# Added Variable Plot
avPlots(m3)
```



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?

As results of BoxCox plot suggest, we don't have to transformation of the response variable. The model is the same as that in 8.

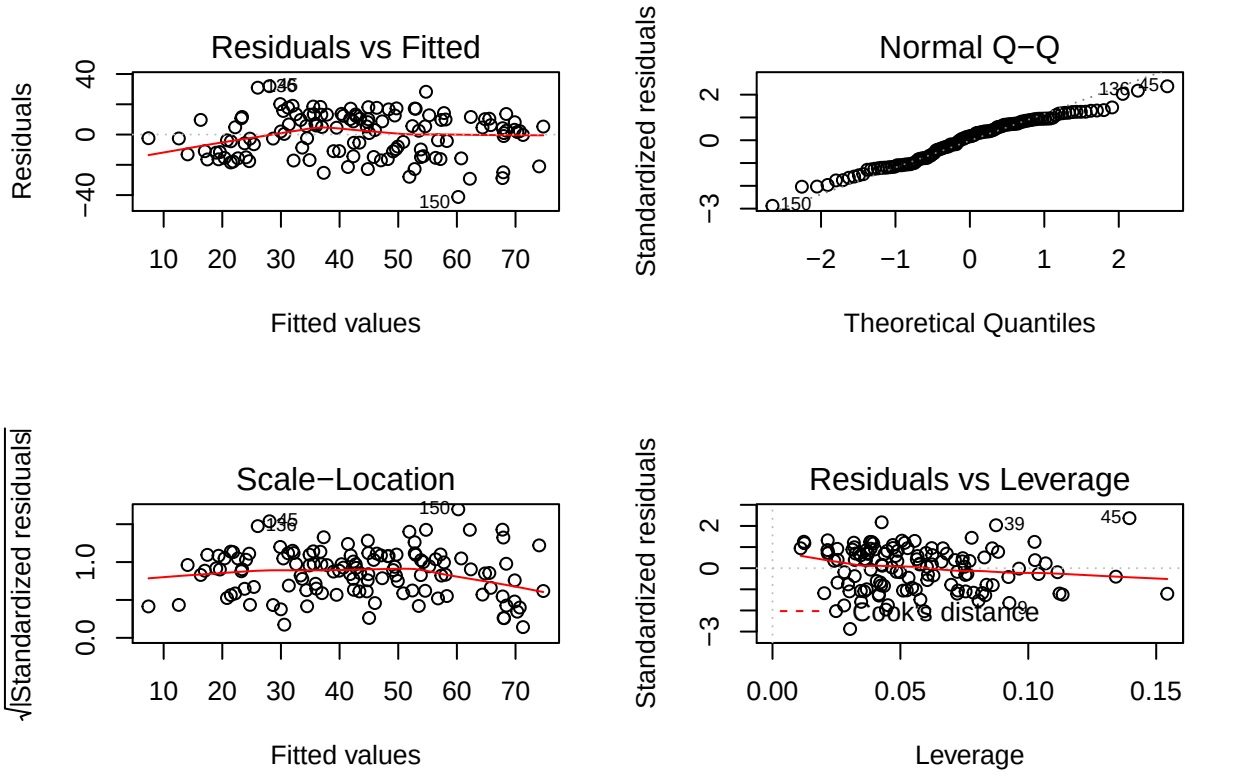
```
# "best" model
m4 <- lm(ModernC ~ l_pop + change_nn + l_ppgdp + l_fertility  + Purban + Frate, data= UN3_new )

# summary model 4
summary(m4)
```

```
##
## Call:
## lm(formula = ModernC ~ l_pop + change_nn + l_ppgdp + l_fertility +
##     Purban + Frate, data = UN3_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.235 -11.589   2.498  10.748  31.954
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.958066  15.253221  -1.308  0.19326
## l_pop         1.595611   0.699289   2.282  0.02430 *
## change_nn     2.310274   2.560728   0.902  0.36879
## l_ppgdp       6.445713   1.508057   4.274 3.91e-05 ***
## l_fertility -18.237639   6.336680  -2.878  0.00475 **
```

```
## Purban       -0.007352   0.106591  -0.069  0.94513
## Frate         0.178242   0.083567   2.133  0.03500 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.55 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.5615, Adjusted R-squared:  0.5392
## F-statistic: 25.19 on 6 and 118 DF,  p-value: < 2.2e-16
```

```r
# diagnostics
par(mfrow=c(2,2))
plot(m4)
```



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.

1) Yes, observation 45 is an outlier.

2) Residuals vs Fitted plot shows a very weak nonlinear pattern.

3) Normal QQ plot shows that the residuals are normally distributed.

4) ScaleLocation plot shows that residuals are spread equally along the ranges of predictors as the fitted line is flat. The results support the homoscedasticity assumption.

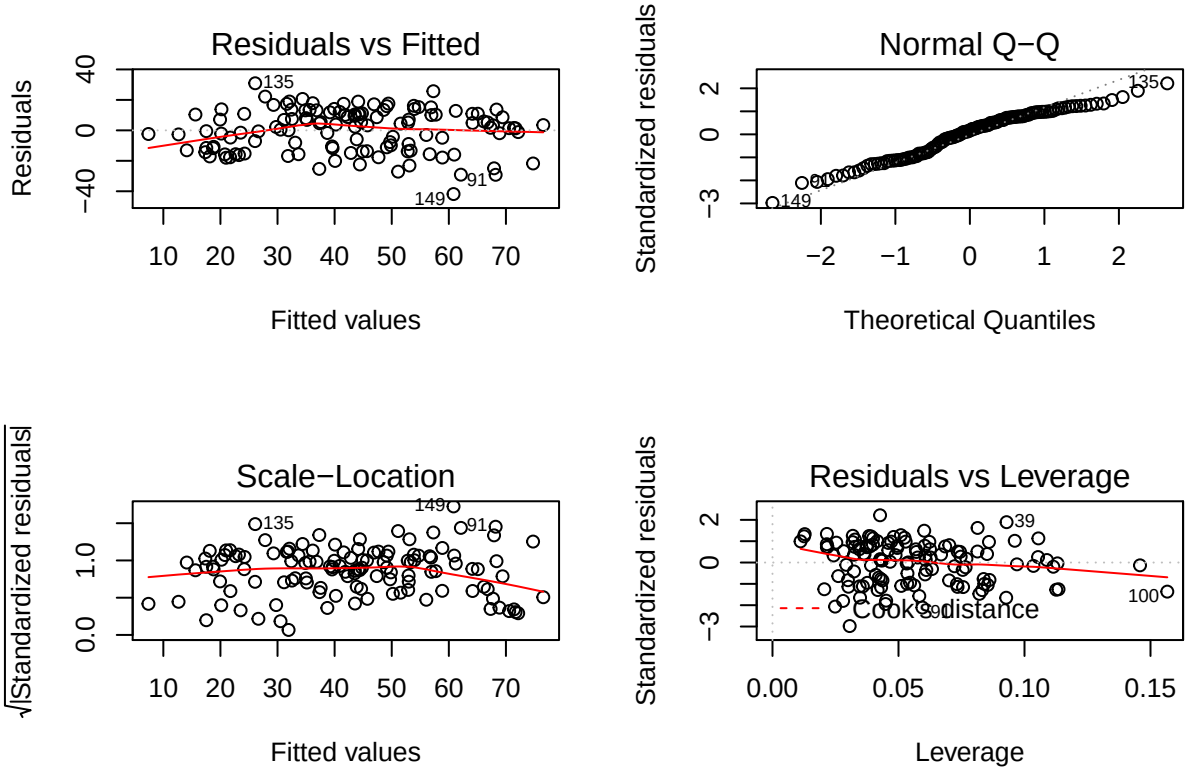
5. Residuals vs Leverage plot shows that no observation has a cook's distance over .5.

```r
m5 <- lm( ModernC ~ l_pop  + change_nn + l_ppgdp + l_fertility + Purban + Frate, data= UN3_new %>% slic

# summary
summary(m5)
```

```
##
## Call:
## lm(formula = ModernC ~ l_pop + change_nn + l_ppgdp + l_fertility +
##     Purban + Frate, data = UN3_new %>% slice(-45))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.828 -11.661   2.337  10.850  30.912
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.98384   15.04294  -1.594 0.113554
## l_pop         1.92979    0.69923   2.760 0.006714 **
## change_nn     3.81671    2.58624   1.476 0.142690
## l_ppgdp       6.36387    1.47849   4.304  3.5e-05 ***
## l_fertility -21.71644    6.37572  -3.406 0.000904 ***
## Purban       -0.02441    0.10471  -0.233 0.816083
## Frate         0.19059    0.08207   2.322 0.021943 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.27 on 117 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.5588
## F-statistic: 26.96 on 6 and 117 DF,  p-value: < 2.2e-16
```

```
# diagnostics
par(mfrow=c(2,2))
plot(m5)
```



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!

The model shows that: 1) If population increases by 1 percent, we would expect `ModernC` would increase by 0.01920205 (1.92979*log(1.01)), holdings variables are constant. 2) If GDP per capita increases by 1 precent, we would expect* ***`ModernC`*** *would increase by 0.06332261 (6.36387*log(1.01)). 3) If live births per female increases by 1 percent, we would expect `ModernC` would decrease by -0.2160858 -21.71644*log(1.01). 3) If the percentage of females over age 15 economically active increases by 1 percent, we would expect* ***`ModernC`*** *would increase by 0.001896434(0.19059*log(1.01)).

```
library(xtable)

# point estimate
point_est  <- xtable(m5) %>% select(Estimate)

# 95 ci
ci <- xtable(confint(m5))

# table
cbind(point_est, ci) %>%
  arrange()
```

```
##                 Estimate        2.5 %     97.5 %
## (Intercept) -23.98383781 -53.77558314  5.8079075
## l_pop         1.92978920   0.54499759  3.3145808
## change_nn     3.81671444  -1.30519630  8.9386252
## l_ppgdp       6.36386834   3.43579107  9.2919456
## l_fertility -21.71643771 -34.34322175 -9.0896537
## Purban       -0.02440963  -0.23178774  0.1829685
## Frate         0.19058615   0.02805826  0.3531140
```

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

No influential point or outlier is founded using Bonferroni Correction and Cook's Distance method. Overall, we find that there are more unmarried women in countries where has better economic growth, lower fertility rate, and larger female labor. Based on the regression result, we find that the fertility rate has the largest impact on unmarried women, however, the correlation makes sense that most women will have a baby after marriage. However, it's important to focus on the labor market and economic development if one wants to adopt some policy to adjust the proportion of unmarried women in the entire population.

```
#Find influential point using Bonferroni Correction
abs.ti = abs(rstudent(m5))
pval= 2*(1- pt(max(abs.ti), m5$df - 1))
min(pval) < .05/nrow(UN3_new)
```

```
## [1] FALSE
```

```
sum(pval < .05/nrow(UN3_new))
```

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

```
#Find outliers using Cook's Distance
rownames(UN3_new)[cooks.distance(m5) > .5]
```

```
## character(0)
```

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.*

1)

$$
\begin{aligned}
(I-H)Y &= \beta_0 + \beta_1(I-H)X \\
(I-H)Y &= \beta_0 + (X^TX)^{-1}X^TY(I-H)X \\
(I-H)Y &= \beta_0 + (X_\beta^T(I-H)(I-H)X_\beta^T)^{-1}X_\beta^T(I-H)Y(I-H)X_\beta \\
X_\beta^T(I-H)Y &= X_\beta^T 1\beta_0 + X_\beta^T(X_\beta^T(I-H)X_\beta)^{-1}X_\beta^T(I-H)Y(I-H)X_\beta \\
X_\beta^T(I-H)Y &= \sum_1^m X_\beta^T 1\beta_0 + X_\beta^T(I-H)Y \\
\sum_1^m X_\beta^T 1\beta_0 &= 0 \\
\beta_0 &= 0
\end{aligned}
$$

2) $e = \sum_1^m X_\beta^T 1\beta_0 = 0$

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 the manually constructed added variable plot for one of the predictors in Ex. 10 is the same as the estimate from your model.

Suppose l_pop is `Xj`. In m5, the point estimate of it is 1.930, which is the same as the result of added variable regression.

```
# dependent variable of avplot
e_Y =  lm( ModernC ~  change_nn + l_ppgdp + l_fertility + Purban + Frate, data= UN3_new %>% slice(-45) %


# indepdent variable of avplot
e_X = lm( l_pop ~  change_nn + l_ppgdp + l_fertility + Purban + Frate, data= UN3_new %>% slice(-45) %>%

avplot <- data.frame( e_Y , e_X)
lm(e_Y ~  e_X, avplot ) %>%  summary()
```

```
##
## Call:
## lm(formula = e_Y ~ e_X, data = avplot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.828 -11.661   2.337  10.850  30.912
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.439e-16  1.255e+00   0.000  1.00000
## e_X          1.930e+00  6.848e-01   2.818  0.00564 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.97 on 122 degrees of freedom
## Multiple R-squared:  0.06112,    Adjusted R-squared:  0.05343
## F-statistic: 7.942 on 1 and 122 DF,  p-value: 0.005636
```