Linear Regression: Residual Analysis, Transformation

Charlie Qu

September 18, 2017

This weekly summary involves the UN data set from ALR. Download the `alr3` library and load the data to answer the following questions adding your code in the code chunks. Please add appropriate code to the chuncks to suppress messages and warnings as needed once you are sure the code is working properly and remove instructions if no longer needed. Please switch the output to pdf for your final version to upload to Sakai.

Exploratory Data Analysis

1. Create a summary of the data. Check the missing, quantitative and qualtitative data.

All of the variables are quantitative, as described below (1)ModernC: Percent of unmarried women using a modern method of contraception. (2)Change: Annual population growth rate, percent. (3)PPgdp:Per capita 2001 GDP, in US $. (4)Frate:Percent of females over age 15 economically active. (5)Pop:Population, thousands. (6)Fertility:Expected number of live births per female, 2000. (7)Purban:Percent of population that is urban, 2001.

All of the variables except "Purban" have missing data by simply viewing the UN3 dataset summary. Specifically, here is a brief 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
```

Alternatively, let us look at the proportions of missing values for all the variables, respectively.

```
library(knitr)
missing.Mod=sum(is.na(UN3[,1]))/nrow(UN3)
missing.Ch=sum(is.na(UN3[,2]))/nrow(UN3)
missing.PP=sum(is.na(UN3[,3]))/nrow(UN3)
missing.Fr=sum(is.na(UN3[,4]))/nrow(UN3)
missing.Pop=sum(is.na(UN3[,5]))/nrow(UN3)
missing.Fe=sum(is.na(UN3[,6]))/nrow(UN3)
missing.Pur=sum(is.na(UN3[,7]))/nrow(UN3)
##Find the absolute missing values of all the variables ##
```

```
missing=rep(7)
for (i in 1:7){missing[i]=sum(is.na(UN3[,i]))}
##Find the missing proportions of all the variables ##
missing.per=rep(7)
for (i in 1:7){missing.per[i]=sum(is.na(UN3[,i]))/nrow(UN3)}
percent <- function(x, digits = 2, format = "f", ...) {
  paste0(formatC(100 * x, format = format, digits = digits, ...), "%")
}
## Gernerate a summary table for missing#
Variables=c("ModernC", "Change","PPgdp","Frate","Pop","Fertility","Purban")
missingsummary=data.frame(Variables,missing,percent(missing.per))
kable(missingsummary)
```

Variables	missing	percent.missing.per.
ModernC	58	27.62%
Change	1	0.48%
PPgdp	9	4.29%
Frate	43	20.48%
Pop	2	0.95%
Fertility	10	4.76%
Purban	0	0.00%

2. Find the mean and standard deviation of each quantitative predictor.

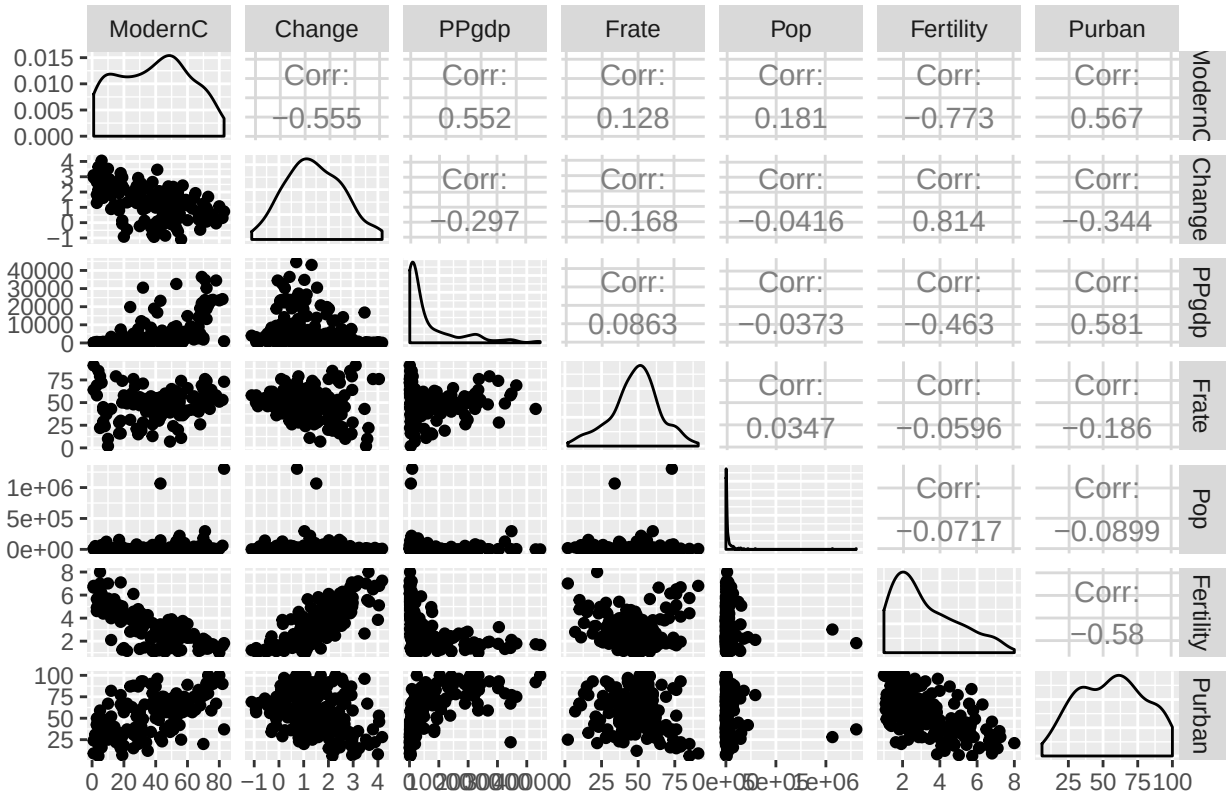
The summary of means and standard deviations are as below

```
library(knitr)
colSD<- function(data) {sapply(data, sd, na.rm = TRUE)}
means=colMeans(UN3[, 1:7],na.rm = TRUE)
means=format(round(means, 2), nsmall = 2)
sds=colSD(UN3[, 1:7])
sds=format(round(sds, 2), nsmall = 2)
UN3ms=data.frame(means,sds)
kable(UN3ms)
```

	means	sds
ModernC	38.72	22.64
Change	1.42	1.13
PPgdp	6527.39	9325.19
Frate	48.31	16.53
Pop	30281.87	120676.69
Fertility	3.21	1.71
Purban	56.20	24.11

3. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Some findings reported regarding trying to predict `ModernC` from the other variables. Check potential outliers, nonlinear relationships or transformations that appear to be needed.

```
library(GGally)
library(ggplot2)
ggp=ggpairs(UN3, columns= 1:7)
print(ggp + ggtitle("Scatterplot, correlation and histogram of UN3 quatitative predictors"))
```



(i) By running the ggpairs() we find that:

(1) For the relaltionships

(a) Strong correlated: Fertility and ModernC (-0.773), Fertility and Change (0.814)

(b) Moderdately strong correlated:change and ModernC(-0.555),PPgdp and ModernC(0.552),Purban and ModernC(0.567),Fertility and PPgdp(-0.463),Purban and Fertility(0.581),Pop and Purban(0.58);

(c) Weakly correlated: Pop and ModernC(0.181), PPgdp and Change(-0.297),Frate and Change(-0.168); Frate and Purban(-0.186);

(d) Ignored relation: Pop and Change (0.0416), Frate and PPgdp(0.0863), Pop and PPgdp(-0.0373), Pop and Frate(0.0347), Fertility and Frate (-0.0596), Fertility and Pop (-0.0717), Purban and Pop (-0.0899).

(2) For the distributions: Pop, PPgdp and Fertility are generally right skewed, Change and Frate are nearly prefect bell-shaped, Purban and ModernC are generally symmetric but of no particular pattern.

(ii) For ModernC, we find that Fertility is strongly correlated with it. There are potentially some influecial points in the left-lower region in their scatterplot. While Change, PPgdp and Purban are moderately strongly correlated to it, with much wider range of points in the scatterplot. So we can predict ModernC by regressing on Fertility and its transformation mainly, and check if adding the other three variables with their transformation would improve the model fitting or not. Moreover, since Change and Fertility are strongly correlated, and the others are also pairwisely moderate strongly correlated, their multicollineariy should be considered. In this situation, linear regression is good enough, no need to consider non-linear case.

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.

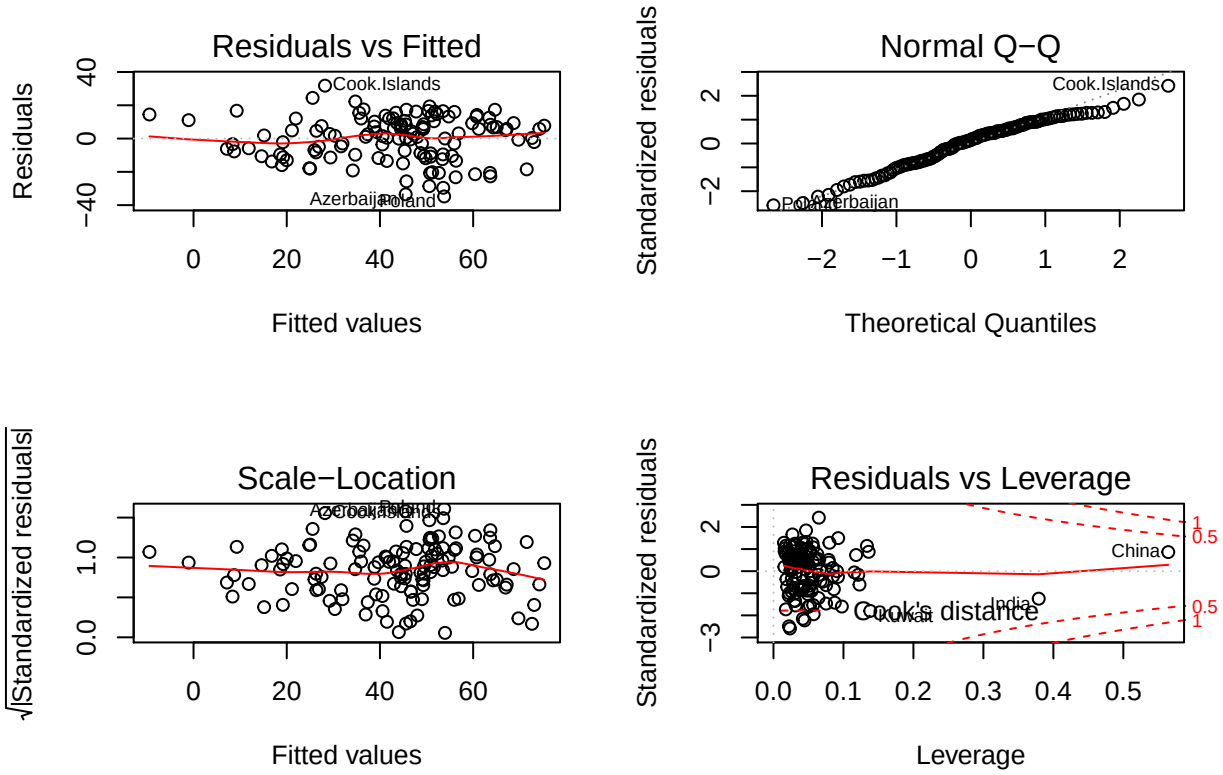
(i) We run the regression of ModernC on all the other predictors first, check the summary blow

```
summary(lm(ModernC ~ Change+PPgdp+Frate+Pop+Fertility+Purban, data=UN3))
```

```
##
## Call:
## lm(formula = ModernC ~ Change + PPgdp + Frate + Pop + Fertility +
##     Purban, 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
```

(ii) Then we create a set of diagnostic residual plots as below

```
par(mfrow=c(2,2))
plot(lm(ModernC ~ Change+PPgdp+Frate+Pop+Fertility+Purban, data=UN3,ask=F))
```



(1) There is no obvious curved pattern for the residual plot, indicating that the variance is contant.

(2) Normal Q-Q plots is slightly negative skewed since the tail of the group of points is obviously above off the straight line.

(3) Similarly as (1), the curved plot indicates the variance is nearly ideally contant.

(4) There are severl obvious outliers for this model, such as China, India and Kerwait. Besides,there are bunch of potentially infuencial points off the center of the points, which indicates the model may be fitted better with transformations. For example, if we take log on the predictors and regressor, the influecial points issue may be improved significantly.

5. Using the Box-Tidwell `boxTidwell` from library `car` 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.

Since by the summary we find that "Change" ranges from -1.10, we add 1.2 to it to make it non-negative. (i) Firstly, check the hypotheses that if it is necessary to conduct power transformations on X's.

(a) On "Change"

```
library(car)
UN3$Change=UN3$Change+1.20
boxTidwell(ModernC~Change,data=UN3)
```

```
##  Score Statistic p-value MLE of lambda
##        -4.590535 4.4e-06      2.296942
##
## iterations =  20
```

(b) On "PPgdp"

```
boxTidwell(ModernC~PPgdp,data=UN3)
```

```
##  Score Statistic p-value MLE of lambda
##        -4.969335   7e-07    -0.2519939
##
## iterations =  4
```

(c) On "Frate"

```
boxTidwell(ModernC~Frate,data=UN3)
```

```
##  Score Statistic   p-value MLE of lambda
##        -3.383727 0.0007151     -34.45257
##
## iterations =  26
```

(d) On "Pop"

```
boxTidwell(ModernC~Pop,data=UN3)
```

```
##  Score Statistic   p-value MLE of lambda
##        -1.198873 0.2305774     0.5909587
##
## iterations =  11
```

(e) On "Fertility"

```
boxTidwell(ModernC~Fertility,data=UN3)
```

```
##  Score Statistic   p-value MLE of lambda
##       -0.7605769 0.4469098      1.141199
##
## iterations =  8
```

(f) On "Purban"

```
boxTidwell(ModernC~Purban,data=UN3)
```

```
##  Score Statistic  p-value MLE of lambda
##       -0.0476013 0.962034      0.979598
##
## iterations =  2
```

Since the P-values for Change, PPgdp and Frate are less than 0.05, no BT transformations are needed. But those tests for Pop, Fertility and Purban indicate that we need to make BT transformation on them.

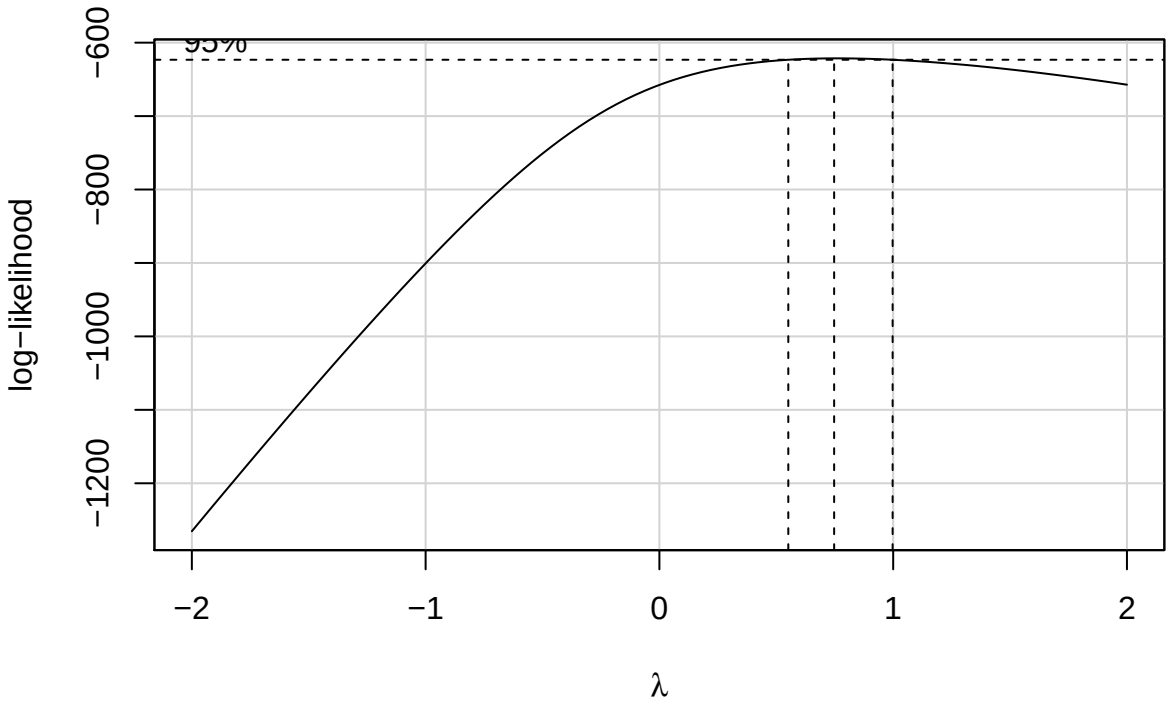
(ii) The final model with transformed X's is as below

$$ModernC = \beta_0+\beta_c\times Change+\beta_{PP}\times PPgdg+\beta_{Fr}\times Frate+\beta_{Pop}\times Pop^{0.5910}+\beta_{Fer}\times Fertility^{1.141}+\beta_{Pur}\times Purban^{0.9796}$$

6. Given the selected transformations of the predictors, select a transformation of the response and justify.

By running boxCox(),

```
Poptrans=UN3$Pop^0.5909587
Fertilitytrans=UN3$Fertility^1.141199
Purbantrans=UN3$Purban^0.979598
translm=lm(ModernC ~ Change+PPgdp+Frate+Poptrans+Fertilitytrans+Purbantrans, data=UN3,ask=F)
bc=boxCox(translm)
```



we find that the optimal λ for BT transformation on y is about 0.8, where log-likelihood is maximized

Let us check the regression summary

```
Modernt=(UN3$ModernC^0.8-1)/0.8
summary(lm(Modernt ~ Change+PPgdp+Frate+Poptrans+Fertilitytrans+Purbantrans, data=UN3))
```

```
##
## Call:
## lm(formula = Modernt ~ Change + PPgdp + Frate + Poptrans + Fertilitytrans +
##     Purbantrans, data = UN3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.365  -4.406   1.243   4.328  15.835
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.599e+01  4.419e+00   5.880 3.90e-08 ***
## Change          2.539e+00  9.829e-01   2.583  0.01102 *
## PPgdp           2.361e-04  8.535e-05   2.766  0.00659 **
## Frate           4.817e-02  3.891e-02   1.238  0.21822
## Poptrans        2.445e-03  1.104e-03   2.215  0.02867 *
## Fertilitytrans -4.211e+00  6.068e-01  -6.939 2.27e-10 ***
## Purbantrans     2.042e-02  4.943e-02   0.413  0.68026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 6.548 on 118 degrees of freedom
##   (85 observations deleted due to missingness)
## Multiple R-squared:  0.6375, Adjusted R-squared:  0.6191
## F-statistic: 34.59 on 6 and 118 DF,  p-value: < 2.2e-16
```

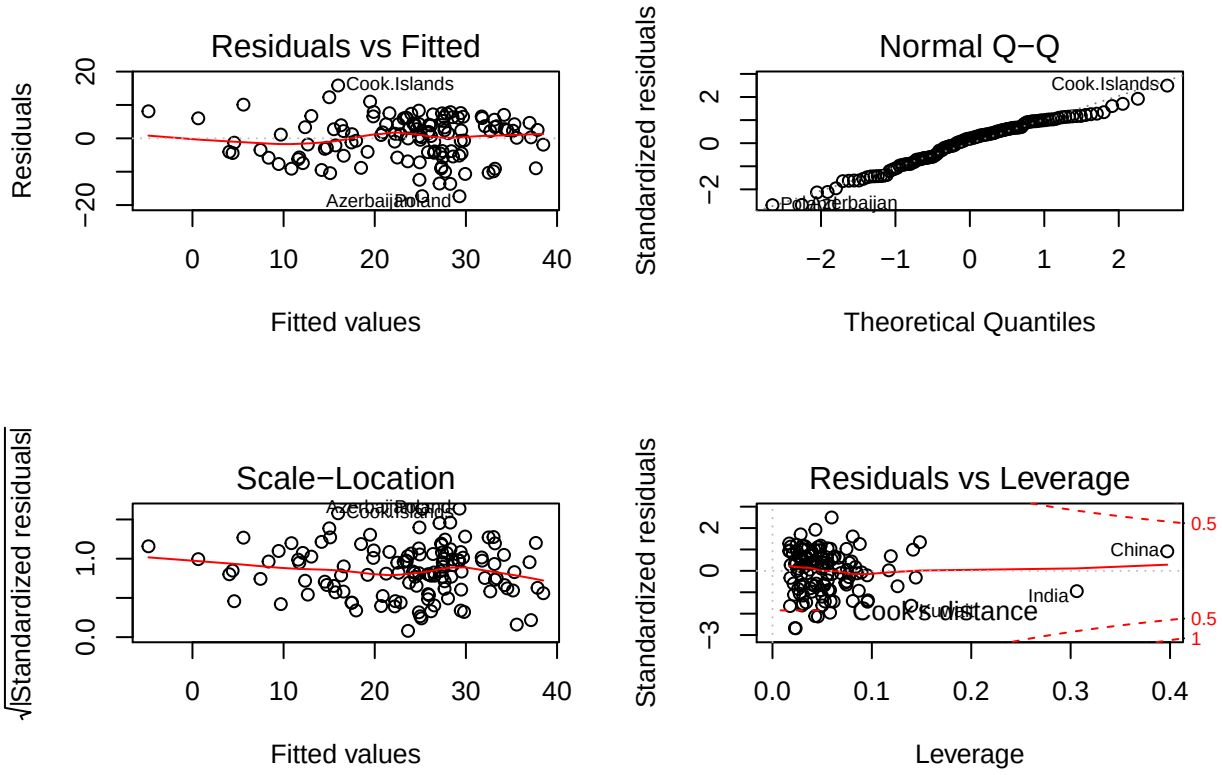
7. Fit the regression using the transformed variables. Provide residual plots and comment. Provide summaries of coefficients with 95% confidence intervals in a nice table with interpretations.

(i) The model is finally fitted as

$$\frac{ModernC^{0.8}-1}{0.8} = \beta_0+\beta_c\times Change+\beta_{PP}\times PPgdg+\beta_{Fr}\times Frate+\beta_{Pop}\times Pop^{0.5910}+\beta_{Fer}\times Fertility^{1.141}+\beta_{Pur}\times Purb$$

(ii) The residual plots are as below

```
par(mfrow=c(2,2))
lmt=lm(Modernt ~ Change+PPgdp+Frate+Poptrans+Fertilitytrans+Purbantrans, data=UN3)
plot(lmt)
```



(iii) The confidence intervals are

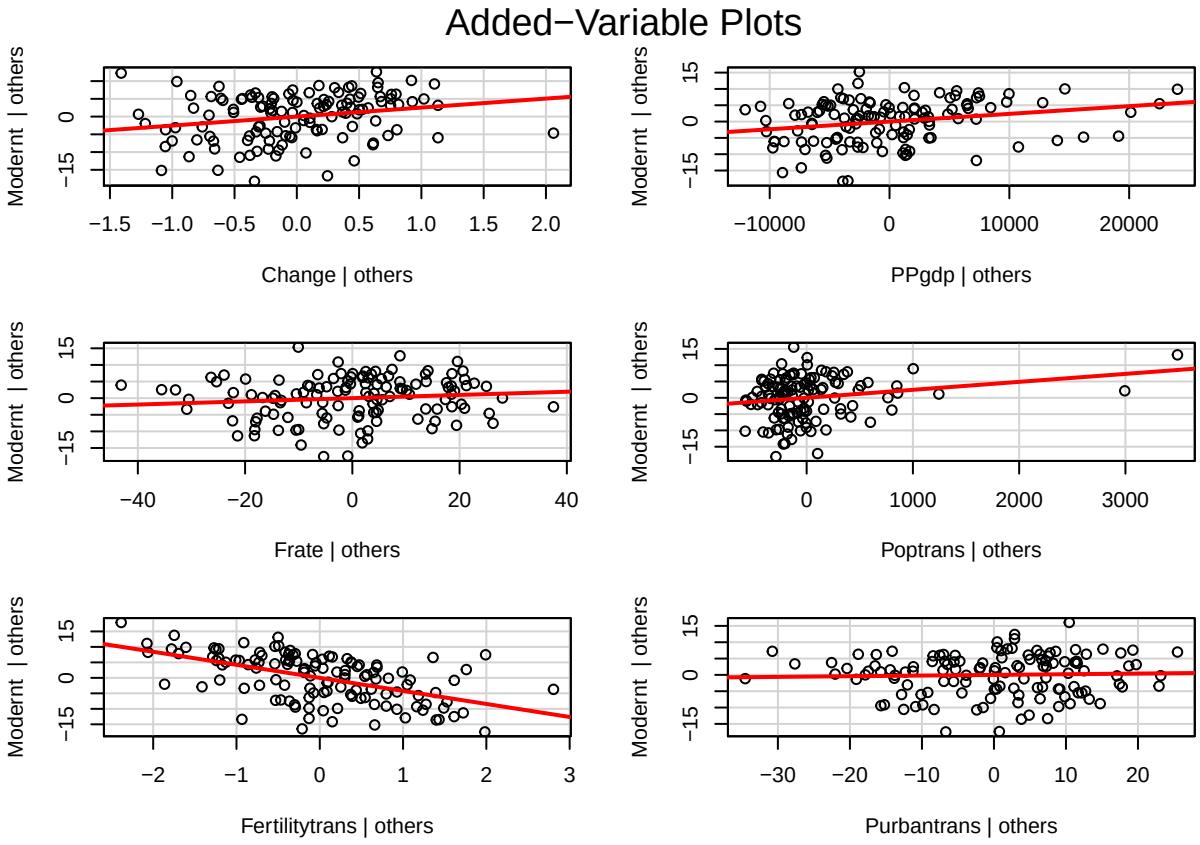
```
Modernt=(UN3$ModernC^0.8-1)/0.8
tval <- -qt((1-0.95)/2, df=nrow(UN3)-2)
LB=rep(6)
for (i in 1:6){LB[i]=summary(lmt)$coefficients[i,1]-summary(lmt)$coefficients[i,2]*tval}
UB=rep(6)
for (i in 1:6){UB[i]=summary(lmt)$coefficients[i,1]+summary(lmt)$coefficients[i,2]*tval}
```

```
Vs=c("Change","PPgdp","Frate","Pop","Fertility","Purban")
kable(data.frame(Vs,LB,UB))
```

Vs	LB	UB
Change	17.2745463	34.7000266
PPgdp	0.6009895	4.4762457
Frate	0.0000678	0.0004043
Pop	-0.0285448	0.1248832
Fertility	0.0002691	0.0046217
Purban	-5.4068307	-3.0144331

8. Examine added variable plots and term plots for you model above. Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?

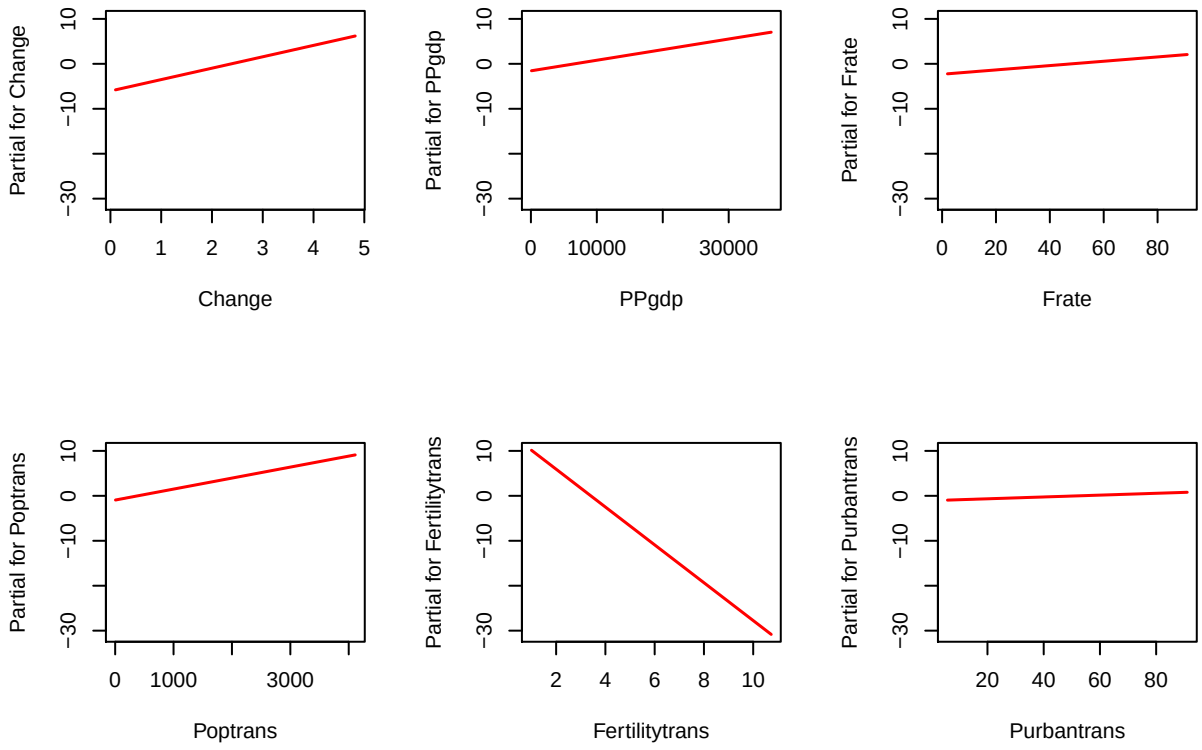
```
av.plots(lmt)
```



For Change, the points beyond 2.0 on x-axis seems to be influencial. For PPgdp, the series of points beyong 10000 on x-axis seem to be influencial. For Frate, the series of points beyong 30 and -40 on x-axis seem to be influencial. For Poptrans, the series of points beyong 1000 on x-axis seem to be influencial. For Fertility, the series of points beyong 2 and -2 on x-axis seem to be influencial. For Purban, the series of points beyong -20 ahd 20 on x-axis seem to be influencial.

The term plots are as below

```
par(mfrow = c(2,3))
termplot(lmt)
```



9. Are there any outliers in the data? Explain. If so refit the model after removing any outliers.

By Bonferonni Correction & Multiple Testing, we have

```
pval = 2*(1 - pt(abs(rstudent(lmt)), lmt$df -1))
View
```

```
## function (x, title)
## {
##     check <- Sys.getenv("_R_CHECK_SCREEN_DEVICE_", "")
##     msg <- "View() should not be used in examples etc"
##     if (identical(check, "stop"))
##         stop(msg, domain = NA)
##     else if (identical(check, "warn"))
##         warning(msg, immediate. = TRUE, noBreaks. = TRUE, domain = NA)
##     if (missing(title))
##         title <- paste("Data:", deparse(substitute(x))[1])
##     as.num.or.char <- function(x) {
##         if (is.character(x))
##             x
##         else if (is.numeric(x)) {
##             storage.mode(x) <- "double"
##             x
##         }
##         else as.character(x)
##     }
##     x0 <- as.data.frame(x)
##     x <- as.list(format.data.frame(x0))
```

```
##     rn <- row.names(x0)
##     if (any(rn != seq_along(rn)))
##         x <- c(list(row.names = rn), x)
##     if (!is.list(x) || !length(x) || !all(sapply(x, is.atomic)) ||
##         !max(lengths(x)))
##         stop("invalid 'x' argument")
##     if (grepl("darwin", R.version$os))
##         check_for_XQuartz()
##     invisible(.External2(C_dataviewer, x, title))
## }
## <bytecode: 0x0000000024b31b60>
## <environment: namespace:utils>
```

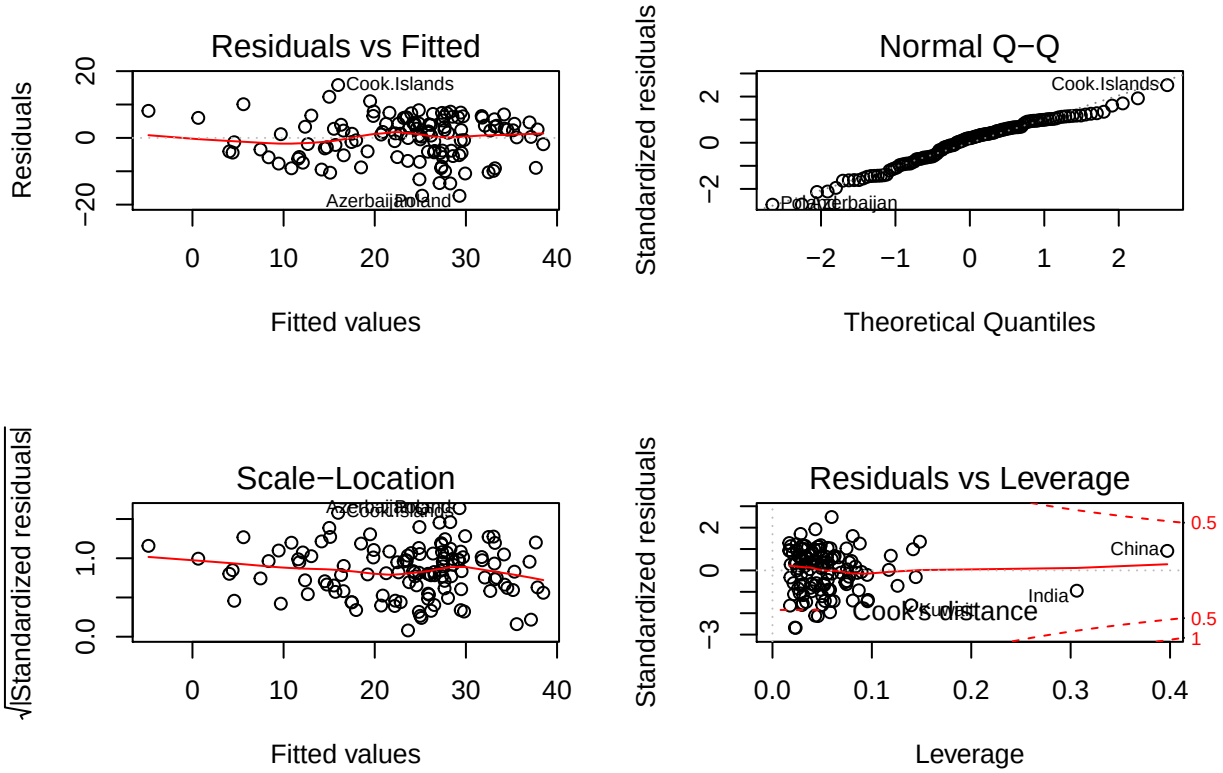
```
rownames(UN3)[pval < .05/nrow(UN3)]
```

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

Thus, there are no outliers in this case.

By checking the Cook's distances, we refit the residual plots, as below

```
lmt2 = lm(Modernt ~ Change+PPgdp+Frate+Poptrans+Fertilitytrans+Purbantrans,
          data=UN3,subset=!cooks.distance(lmt)>1)
par(mfrow=c(2,2)); plot(lmt2)
```



Summary of Results

10. Provide a brief paragraph summarizing your final model and findings suitable for the US envoy to the UN after adjusting for outlierd or influential points.

(i) The final model fitted is

$$\frac{ModernC^{0.8}-1}{0.8}$$
$$=\beta_0+\beta_c\times Change+\beta_{PP}\times PPgdg+\beta_{Fr}\times Frate+\beta_{Pop}\times Pop^{0.5910}+\beta_{Fer}\times Fertility^{1.141}+\beta_{Pur}\times Purban^{0.9796}$$

(ii) We select not to remove the outliers and stick to the original model.

(iii) There is no obvious curved pattern for the residual plot, indicating that the variance is contant.

(iv) Normal Q-Q plots is slightly negative skewed since the tail of the group of points is obviously above off the straight line.

(v) The curved plot of standardized residual vs fitted values indicates the variance is nearly ideally contant.

(vi)63.75% of the variation in ModernC is explained by this model.

Theory

11. Using $X^TX = X_{(i)}^TX_{(i)} + x_ix_i^T$ where the subscript (i) means without the ith case, show that

$$(X_{(i)}^TX_{(i)})^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-h_{ii}}$$

where h_{ii} is the ith diagonal element of $H = X(X^TX)^{-1}X^T$.

Proof

$$X^TX = X_{(i)}^TX_{(i)} + x_ix_i^T \Rightarrow$$
$$(X_{(i)}^TX_{(i)})^{-1} = (X^TX - x_ix_i^T)^{-1} \Rightarrow$$
$$(X_{(i)}^TX_{(i)})^{-1} = (X^TX - \frac{(X^TX)x_ix_i^T(X^TX)}{(X^TX)^2})^{-1} \Rightarrow$$
$$(X_{(i)}^TX_{(i)})^{-1} = (X^TX - \frac{(X^TX)x_ix_i^T(X^TX)}{(X^TX)^2})^{-1} \Rightarrow$$

Thus,

$$(X_{(i)}^TX_{(i)})^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-h_{ii}}$$

12. Use 11 to show that

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X^TX)^{-1}x_ie_i}{1-h_{ii}}$$

where $\hat{\beta}_{(i)} = (X_{(i)}^TX_{(i)})^{-1}X_{(i)}^TY_{(i)}$ and $e_i = y_i - x_i^T\hat{\beta}$. *Hint write* $X_{(i)}^TY_{(i)} = X^TY - x_iy_i$.

proof Since $X_{(i)}^T Y_{(i)} = X^TY - x_iy_i $ by 11.

$$\hat{\beta}_{(i)} = (X_{(i)}^TX_{(i)})^{-1}X_{(i)}^TY_{(i)}$$
$$= (X_{(i)}^TX_{(i)})^{-1}(X^TY - x_iy_i)$$
$$= ((X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-h_{ii}})(X^TY - x_iy_i)$$
$$= (X^TX)^{-1}X^TY - (X^TX)^{-1}x_iy_i + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-h_{ii}}X^TY - \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-h_{ii}}x_iy_i$$

$$= \hat{\beta} - \frac{(1-h_{ii})(X^TX)^{-1}x_iy_i}{1-h_{ii}} + \frac{(X^TX)^{-1}x_ix_i^T\hat{\beta}}{1-h_{ii}} - \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-h_{ii}}x_iy_i$$

$$= \hat{\beta} - \frac{(X^TX)^{-1}[x_iy_i(1-h_{ii}) - x_ix_i^T\hat{} + x_ix_i^T(\hat{X^TX})^{-1}x_i\hat{\beta}y_i}{1-h_{ii}}$$

$$= \hat{\beta} - \frac{(X^TX)^{-1}[x_iy_i - x_ix_i^T\hat{} + x_ix_i^T(\hat{X^TX})^{-1}x_i\hat{\beta}y_i - h_{ii}x_iy_i]}{1-h_{ii}}$$

Since $h_{ii} = x_ix_i^T(X^TX)^{-1}$ by definition, the above reduces to

$$= \hat{\beta} - \frac{(X^TX)^{-1}[x_iy_i - x_ix_i^T\hat{\beta}]}{1-h_{ii}}$$

$$e_i = y_i - x_i^T\hat{\beta} \Rightarrow$$

$$x_ie_i = x_iy_i - x_ix_i^T\hat{\beta} \Rightarrow$$

Thus,

$$= \hat{\beta} - \frac{(X^TX)^{-1}x_ie_i}{1-h_{ii}}$$