HW2 STA521 Fall18

James Wang, cw261, jameswang1997

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

1. Preliminary read in the data. After testing, modify the code chunk so that output, messages and warnings are suppressed. *Exclude text from final*
2. Create a summary of the data. How many variables have missing data? Which are quantitative and which are qualtitative?

UN3 %>% summary() %>% kable

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | ModernC | Change | PPgdp | Frate | Pop | Fertility | Purban |
|  | Min. : 1.00 | Min. :-1.100 | Min. : 90 | Min. : 2.00 | Min. : 2.3 | Min. :1.000 | Min. : 6.00 |
|  | 1st Qu.:19.00 | 1st Qu.: 0.580 | 1st Qu.: 479 | 1st Qu.:39.50 | 1st Qu.: 767.2 | 1st Qu.:1.897 | 1st Qu.: 36.25 |
|  | Median :40.50 | Median : 1.400 | Median : 2046 | Median :49.00 | Median : 5469.5 | Median :2.700 | Median : 57.00 |
|  | Mean :38.72 | Mean : 1.418 | Mean : 6527 | Mean :48.31 | Mean : 30281.9 | Mean :3.214 | Mean : 56.20 |
|  | 3rd Qu.:55.00 | 3rd Qu.: 2.270 | 3rd Qu.: 8461 | 3rd Qu.:58.00 | 3rd Qu.: 18913.5 | 3rd Qu.:4.395 | 3rd Qu.: 75.00 |
|  | Max. :83.00 | Max. : 4.170 | Max. :44579 | Max. :91.00 | Max. :1304196.0 | Max. :8.000 | Max. :100.00 |
|  | NA’s :58 | NA’s :1 | NA’s :9 | NA’s :43 | NA’s :2 | NA’s :10 | NA |
| Purba | n is the only va | riable that doesn | ’t have missing | data |  |  |  |
| All o | f the variables | are quantative |  |  |  |  |  |

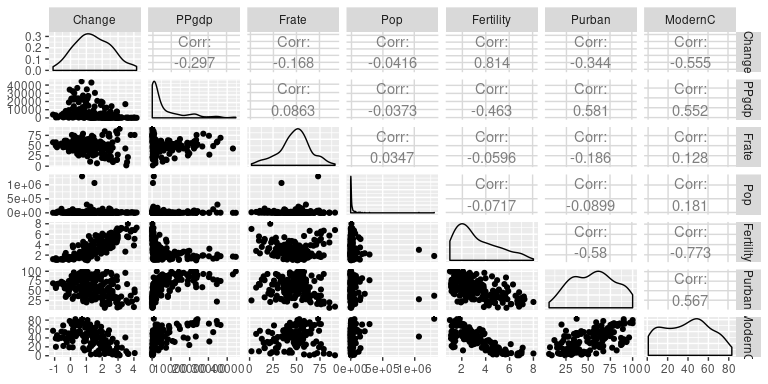
1. What is the mean and standard deviation of each quantitative predictor? Provide in a nicely formatted table.

t1 = sapply(UN3, function(col) c(mean(col, na.rm = T), sd(col, na.rm = T))) %>% t()  
colnames(t1) = c('mean', 'std')  
t1 %>% kable()

|  |  |  |
| --- | --- | --- |
|  | mean | std |
| ModernC | 38.717105 | 2.263661e+01 |
| Change | 1.418373 | 1.133133e+00 |
| PPgdp | 6527.388060 | 9.325189e+03 |
| Frate | 48.305389 | 1.653245e+01 |
| Pop | 30281.871428 | 1.206767e+05 |
| Fertility | 3.214000 | 1.706918e+00 |
| Purban | 56.200000 | 2.410976e+01 |

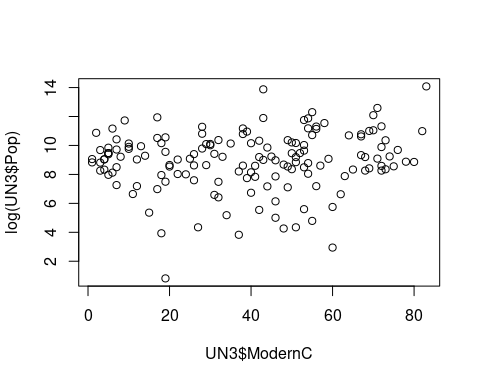
1. Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings regarding trying to predict ModernC from the other variables. Are there potential outliers, nonlinear relationships or transformations that appear to be needed based on your graphical EDA?

X\_col = names(UN3)[2:ncol(UN3)]  
UN3 %>% ggpairs( columns = c(X\_col, 'ModernC') )

 Strong Correlation with ModernC: Change (-), PPgdp, Fertility (-), Purban Note: fertity has a high correlation with Change, and PUrban

Outlier / log might needed: Pop After log(pop), the scatter plot looks better

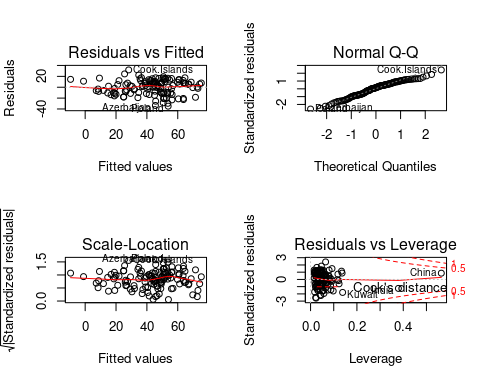
plot(UN3$ModernC, log(UN3$Pop))



## Model Fitting

1. 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?

fit1 = lm(  
 ModernC ~ . ,   
 data = UN3  
)  
par(mfrow = c(2,2))  
plot(fit1)

 Model Assumptions: Linearity: some outlier for big counties skew the result to right (leverage but not enough to be influential pts) Independence: countries are pretty randomly scattered Normailty: curving at both ends (might need log y) Constant variance: generally true

a = summary(fit1)   
a$df

## [1] 7 118 7

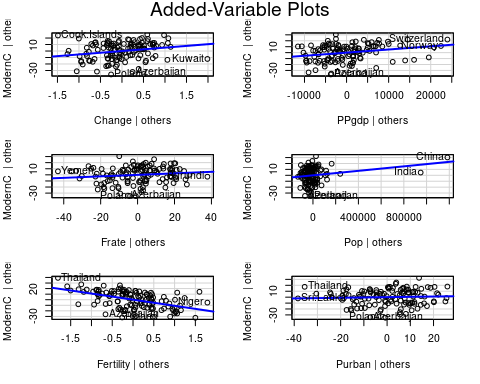
a$df[2] + a$df[3]

## [1] 125

(85 observations deleted due to missingness) So the N used in the model is 118 df + 6 predictor + 1 intercept = 125

1. 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(fit1)

 These added varible plots show that Change, Fertility, and PPgdp have linear relationship with ModernC. Pop does too but needs to be logged. Frate and Purhan show a less clear linear relationship

1. 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.

First, use Multivariate Imputation by Chained Equations to impute NAs

imp = mice(UN3)  
UN\_imp = complete(imp)

# center shift transform for Change to make all Change\_i non negative   
  
UN\_imp = UN\_imp %>% mutate(Change1 = Change + abs(min(Change)) + 1,  
 Pop1 = Pop,   
 Frate1 = Frate + 10,  
 Purban1 = Purban/100)  
  
# pop and frate doesn't work without log  
bt = car::boxTidwell(ModernC ~ Change1 + PPgdp + Fertility + Pop + Frate + Purban, data = UN\_imp)

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

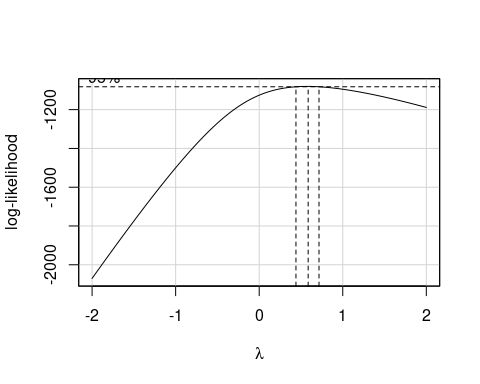
bt # we shall log change

## MLE of lambda Score Statistic (z) Pr(>|z|)   
## Change1 -0.99633 -2.1396 0.03239 \*  
## PPgdp 0.77108 -0.2274 0.82009   
## Fertility 0.94647 0.6350 0.52545   
## Pop 0.51101 -0.9266 0.35413   
## Frate 1.41549 0.5790 0.56257   
## Purban 3.74706 0.8109 0.41744   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 26

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

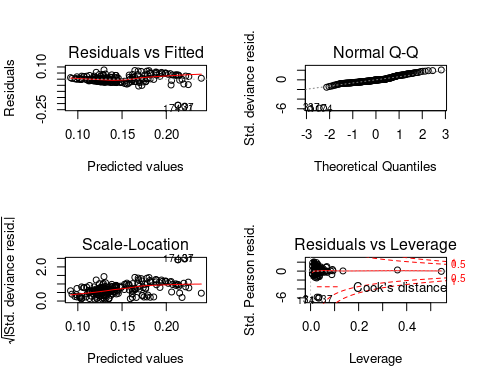
mean lambda is more toward 1 than 0, around .6. The transformation on y did not significantly improve the model. The diagnostic plots show a handful of datapoints are not explained by the transformed model. I decide not to do any transform on y

with(UN\_imp, car::boxCox(ModernC ~ log(Change1) + PPgdp + Fertility + Purban + Pop))



lambda1 = powerTransform( ModernC ~ log(Change1) + PPgdp + Fertility + Purban + Pop, data = UN\_imp)[['roundlam']]

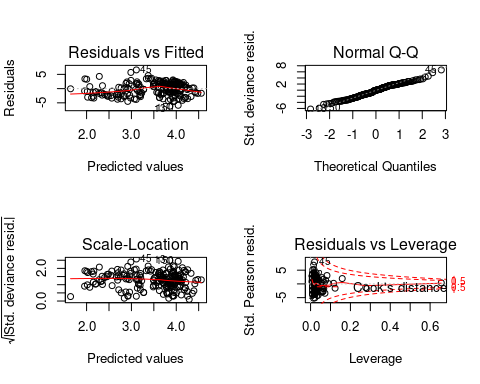
fit\_tran = glm( (ModernC^lambda1 -1) / ModernC ~ log(Change1) + PPgdp + Fertility + Purban + Pop, data = UN\_imp)  
par(mfrow = c(2,2))  
fit\_tran %>% plot



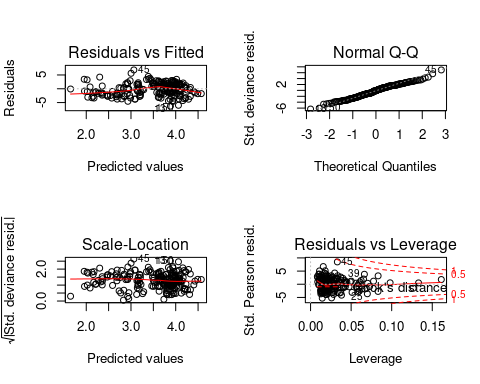
1. Fit the regression using the transformed variables. Provide residual plots and added variables plots and comment. If you feel that you need additional transformations of either the response or predictors, repeat any steps until you feel satisfied.

residual plot still not independent (constant var) I think need to log population

fit2 = glm(ModernC ~ log(Change1) + PPgdp + Fertility + Purban + Pop, data = UN\_imp, family = 'poisson')  
par(mfrow = c(2,2))  
fit2 %>% plot

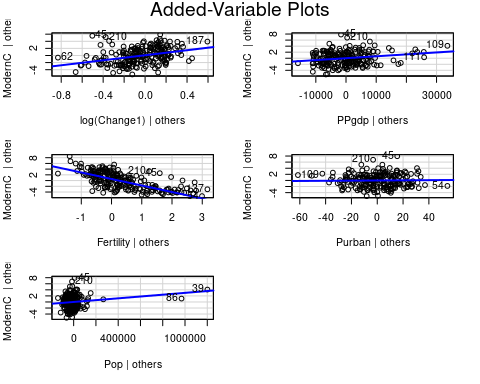


fit3 = glm(ModernC ~ log(Change1) + PPgdp + Fertility + Purban + log(Pop),  
 data = UN\_imp,  
 family = 'poisson')  
par(mfrow = c(2,2))  
fit3 %>% plot

 avplot

linear correlation slightly stronger than before, still could be better

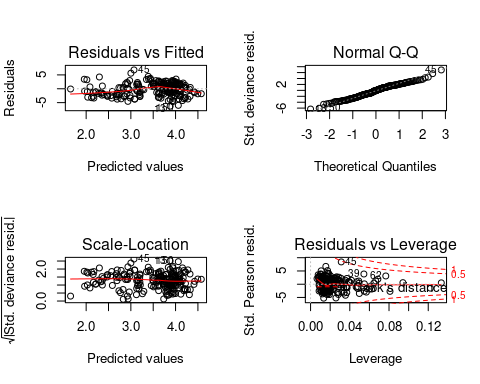
avPlots(fit2)



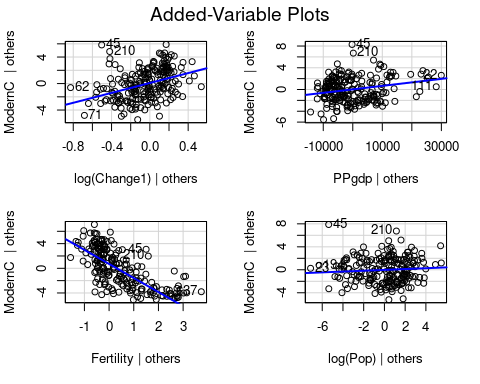
fit4 = step(fit3)

## Start: AIC=2169.31  
## ModernC ~ log(Change1) + PPgdp + Fertility + Purban + log(Pop)  
##   
## Df Deviance AIC  
## - Purban 1 1064.4 2167.7  
## <none> 1064.0 2169.3  
## - log(Pop) 1 1067.1 2170.4  
## - PPgdp 1 1104.3 2207.6  
## - log(Change1) 1 1209.6 2312.8  
## - Fertility 1 1975.2 3078.4  
##   
## Step: AIC=2167.71  
## ModernC ~ log(Change1) + PPgdp + Fertility + log(Pop)  
##   
## Df Deviance AIC  
## <none> 1064.4 2167.7  
## - log(Pop) 1 1067.6 2168.9  
## - PPgdp 1 1110.0 2211.3  
## - log(Change1) 1 1216.7 2318.0  
## - Fertility 1 2163.3 3264.6

# ModernC ~ log(Change1) + PPgdp + Fertility  
# stepwise variable selection shows we should take out Pop and Purban, which do not have strong correlation with Y  
par(mfrow = c(2,2))  
plot(fit4)



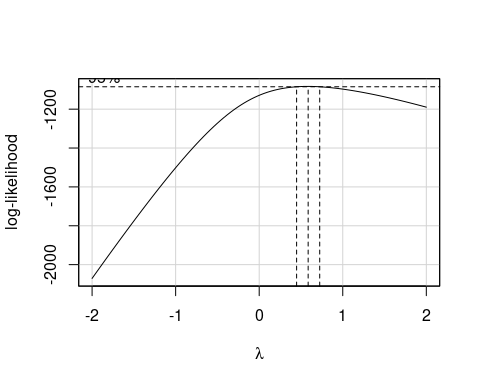
avPlots(fit4)

 final model: ModernC = B0 + B1 \* log(Change1) + B2 \* PPgdp + B3 \* Fertility

1. 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?

# since we get the same y decison (not transform), we end up with the same model

a= with(UN\_imp, car::boxCox(ModernC ~ Change1 + PPgdp + Fertility + Purban + Pop))

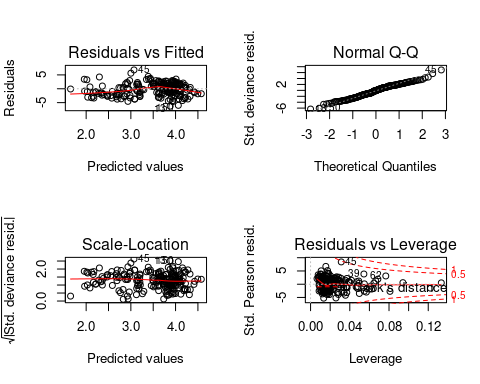


powerTransform(ModernC ~ Change1 + PPgdp + Fertility + Purban + Pop, data = UN\_imp)

## Estimated transformation parameter   
## Y1   
## 0.5803382

1. 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.

# no influential points  
par(mfrow = c(2,2))  
plot(fit4)



## Summary of Results

1. 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!

cbind(OddRatio = fit4$coefficients, confint.default(fit4)) %>% kable

|  |  |  |  |
| --- | --- | --- | --- |
|  | OddRatio | 2.5 % | 97.5 % |
| (Intercept) | 4.0763203 | 3.9570337 | 4.1956069 |
| log(Change1) | 0.5928944 | 0.4974447 | 0.6883440 |
| PPgdp | 0.0000078 | 0.0000056 | 0.0000100 |
| Fertility | -0.4425300 | -0.4705017 | -0.4145582 |
| log(Pop) | 0.0085151 | -0.0008758 | 0.0179059 |

cbind(OddRatio = fit4$coefficients, confint.default(fit4)) %>% exp() %>% kable

|  |  |  |  |
| --- | --- | --- | --- |
|  | OddRatio | 2.5 % | 97.5 % |
| (Intercept) | 58.9282316 | 52.3019527 | 66.3940122 |
| log(Change1) | 1.8092174 | 1.6445137 | 1.9904167 |
| PPgdp | 1.0000078 | 1.0000056 | 1.0000100 |
| Fertility | 0.6424091 | 0.6246888 | 0.6606321 |
| log(Pop) | 1.0085514 | 0.9991246 | 1.0180672 |
| $1 in 2001 GDP | per capita wo | uld result in | a non practically signifant ratio increase in percent of unmarried woman, given all else constant |
| A one point inc | rease in expe | cted number o | f live birth per female would result in a decrease in percent of unmarried woman to a factor .641 of before, given all else constant |
| If we increase | Annual popula | tion growth r | ate by 1%, we would expect to see the percent of unmarried women increase by .666% |

1. 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

The Percent of unmarried women using a modern method of contraception in each country has significant negative correlation with Fertility rate and positive correlation with Population Growth Rate, all else held constant. GDP/capital has a negligible positive correlations, while Urbanization rate, female percentage over 15, and population have no significant correlation. The missing data have been imputed with multivariate imputation. No case was deleted.

## Methodology

1. Prove that the intercept in the added variable scatter plot will always be zero. *Hint: use the fact that if is the project matrix which contains a column of ones, then . Use this to show that the sample mean of residuals will always be zero if there is an intercept.*
2. 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.