Regression Analysis on U.S. Life Expectancy

Wenxuan Zhang (Section W 10 AM)

Karen Zhao (Section T 3 PM)

S20 PSTAT 126 Final Project

Prof. Sudeep Bapat

June 7, 2020

1. Introduction

This project focuses on studying the prediction of life expectancy in the U.S. states based on the dataset ‘state.x77’ in R library, which is derived from the U.S. Department of Commerce, Bureau of the Census (1977) *Statistical Abstract of the United States*. We will examine the effects of the following 7 variables on life expectancy: population, income, illiteracy, murder rate (Murder), high school graduate rate (HS Grad), land area, and mean number of days with minimum temperature below freezing (Frost). We find that ‘Murder’, ‘HS Grad’, ‘Frost’, and “Population’ are the most related predictors.

2. Questions of Interest

Can we predict life expectancy of a region given its population, income, illiteracy, murder rate (Murder), high school graduate percent (HS Grad), land area, and mean number of days with minimum temperature below freezing (Frost) as predictors?

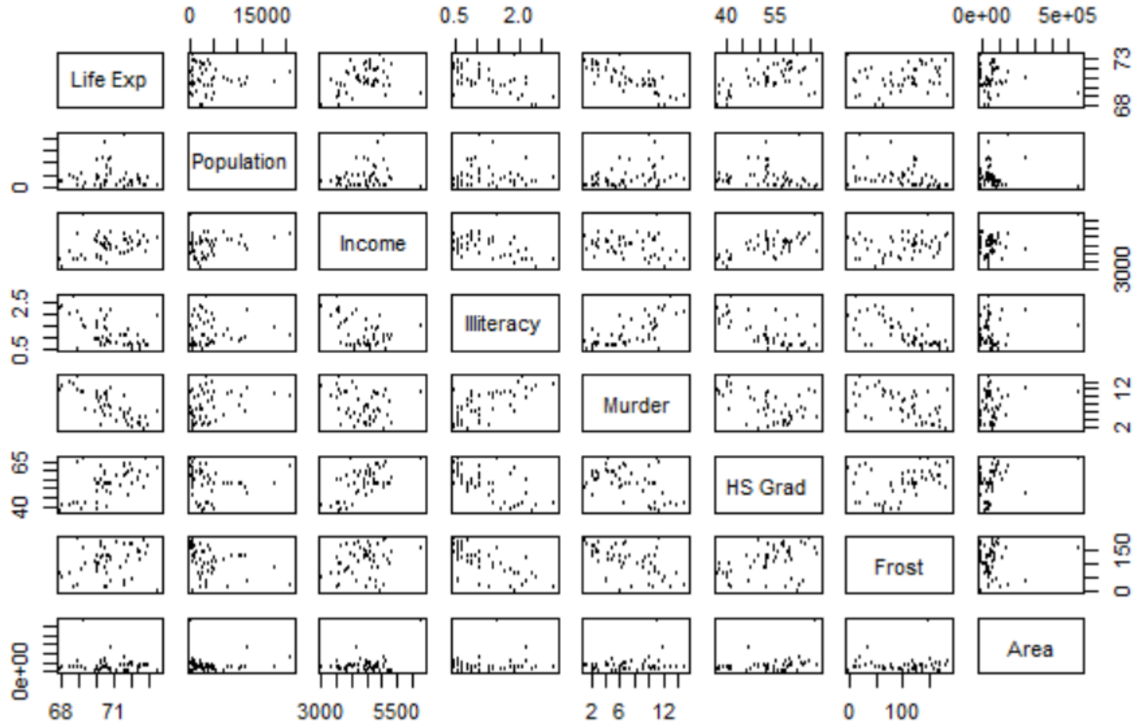
3. Regression Method

We will approach this question first by applying stepwise and best subsets regression on the 7 potential predictors to determine the best model. Then we will check LINE conditions on this model using residual analysis. If any of the assumptions are not met, we will transform the data and check LINE conditions for the new model. After fitting the model with transformed data, we will interpret our model and summarize our findings.

4. Regression Analysis, Results and Interpretation

Variable Selection

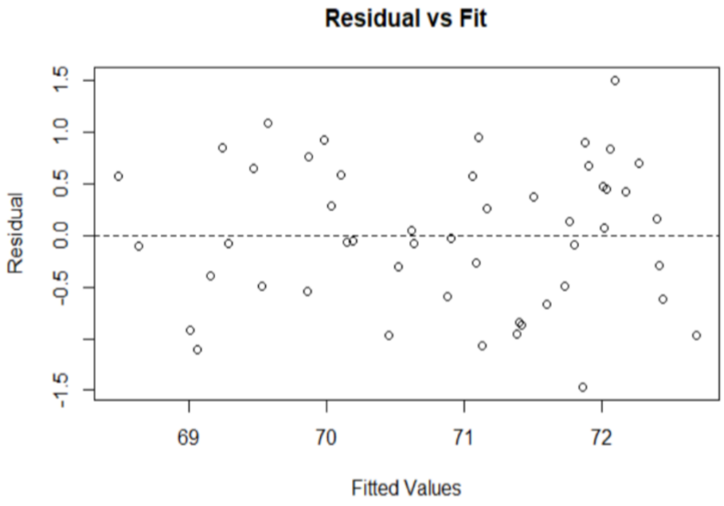
First, we look at the scatterplot matrix to gain some insight on the relationships between the variables in the data. From the scatterplot below, we can tell that there are some predictors like ‘Murder’ seems to be strongly related to ‘Life Expectancy’. Others like ‘Area’ and ‘Income’ seem to be moderately or weakly related.

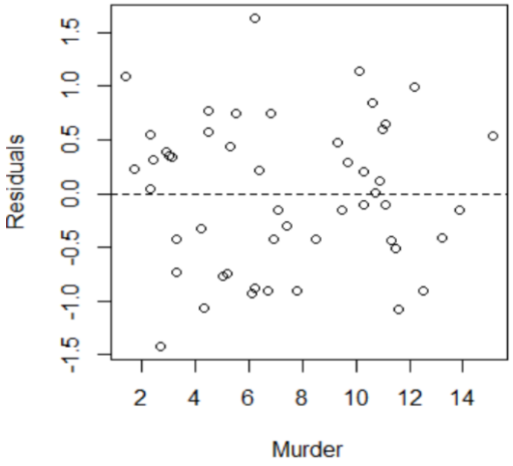


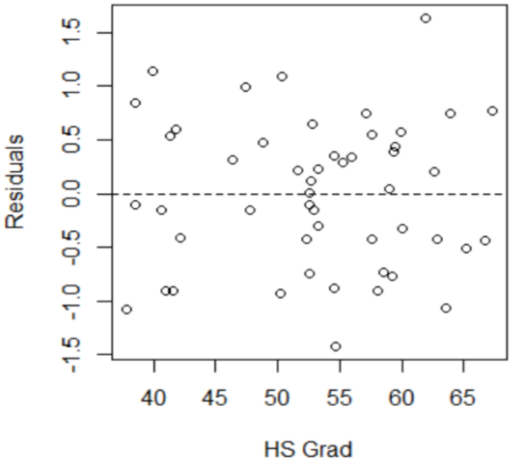
Secondly, we perform variable selection using stepwise regression, including AIC and partial F test, and the best subsets regression to determine the predictors. The results of our AIC test, partial F test, and adjusted R^2 criterion chooses four predictors: "*Murder*", "*HS Grad*", "*Frost*", and "*Population*". The Mallows' C_p criterion gives similar result except excluding the fourth predictor "*Population*". Therefore, we decide our model to be *`Life Exp` ~ Murder + `HS Grad` + Frost + Population.*

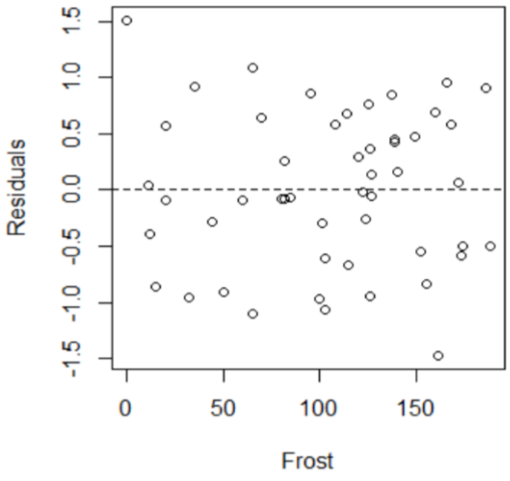
Diagnostic Checks and Transformation

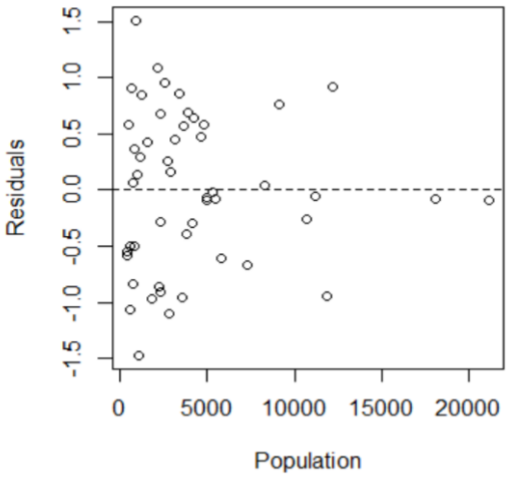
Thirdly, we check the LINE conditions for this model. We will not be checking the independence assumption, since we are not given data related to time order.


Residual vs Fit
Residual
1.5
1.0
0.5
0.0
-0.5
-1.5
69
70
71
72
Fitted Values

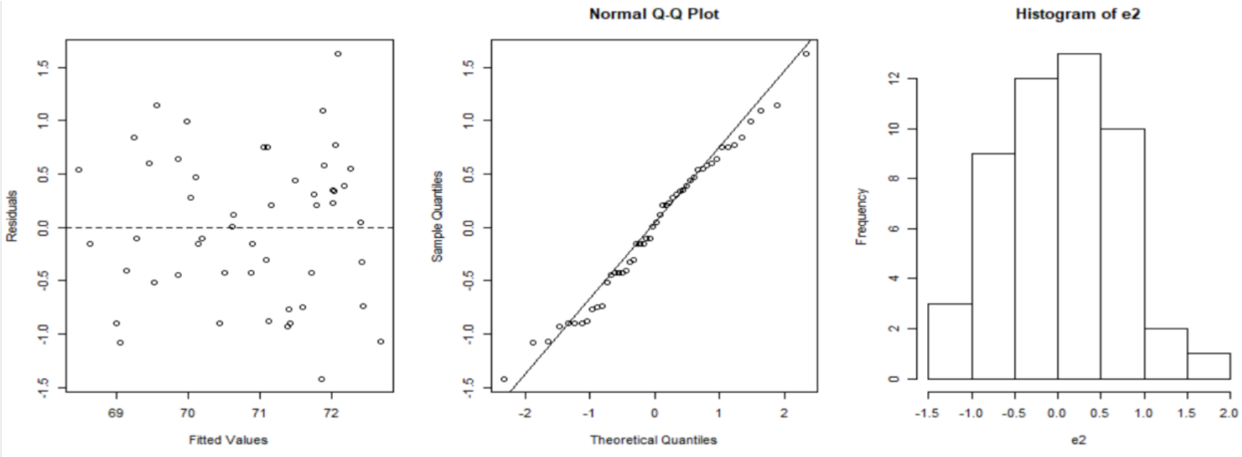

Residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
2
4
6
8
10
12
14
Murder

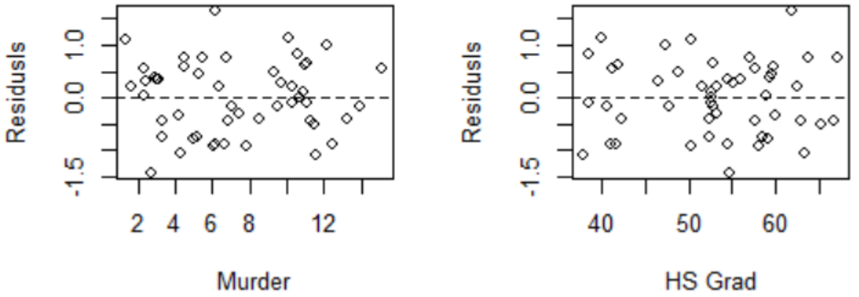

Residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
40
45
50
55
60
65
HS Grad

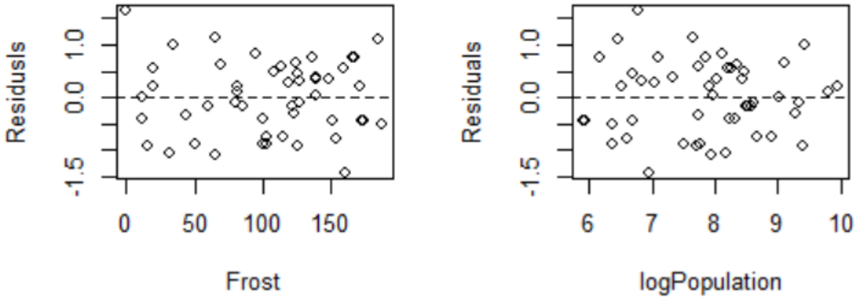

Residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
50
100
150
Frost


Residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
5000
10000
15000
20000
Population

The Residual v.s. Fitted plot shows that residuals "bounce randomly" and roughly form a "horizontal band" around the *y=0* line. However, when looking at the "Residuals vs Predictor" plot, and see a strong funneling effect for the "Residuals v.s. Population" plot. Since a log function has the ability to "spread out" smaller values and bring in larger ones, we will perform log transformation on "Population". Our model is now *`Life Exp` ~ Murder + `HS Grad` + Frost + log(Population)*. Then we check our LINE conditions again.





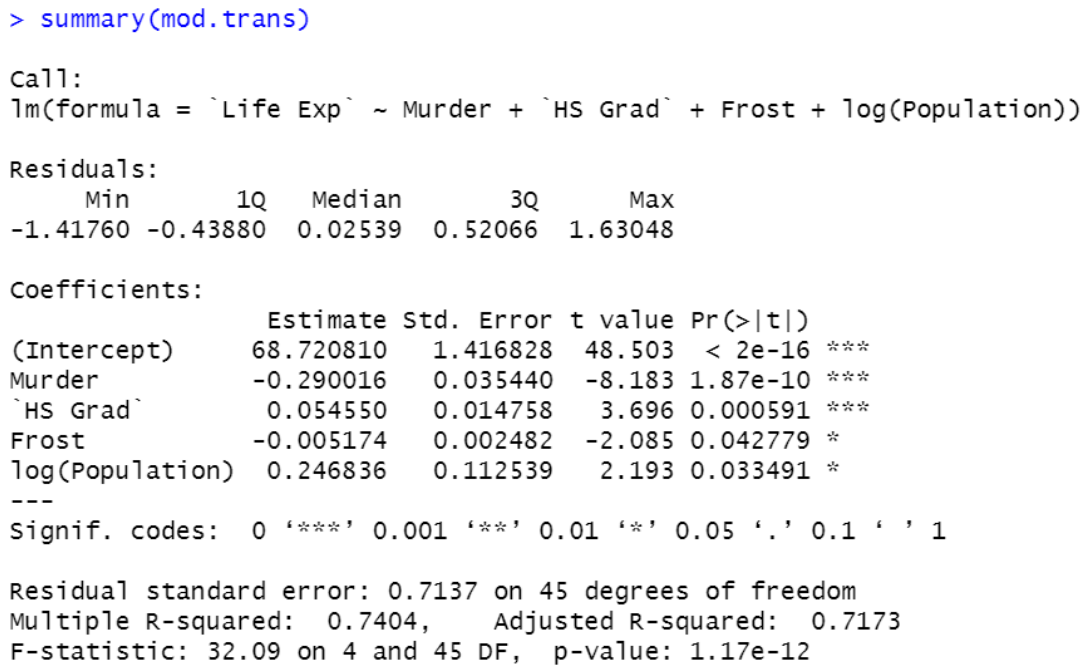


The “Residuals vs Predictor” plot for *log(Population)* is well-behaved now. The Residual v.s. Fit plot and Normal Q-Q plot are both well-behaved. There are no unequal variance or nonlinearity problems.

Our final step is checking for outliers and leverage. After computing for both internally studentized residuals and studentized deleted (or externally studentized) residuals, none of them are larger than 3 in absolute value. Thus, there are no unusual Y observations. After computing the hat values, we find that none of the points has higher hat value than $3\frac{p}{n} = 0.3$. Therefore, there are no outliers or leverage points. And we will not need to investigate for any potentially influential points. Our model has met the LINE conditions.

Interpretation

We are now able to observe our model with 4 predictors: Murder, HS Grad, Frost, log(Population). $Life\ Expectancy = -0.29Murder + 0.0546\ HSGrad - 0.051\ Frost + 0.24\ log(Population)$



```
> summary(mod.trans)

Call:
lm(formula = `Life Exp` ~ Murder + `HS Grad` + Frost + log(Population))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.41760 -0.43880  0.02539  0.52066  1.63048 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     68.720810   1.416828  48.503  < 2e-16 ***
Murder          -0.290016   0.035440  -8.183 1.87e-10 ***
`HS Grad`        0.054550   0.014758   3.696 0.000591 ***
Frost           -0.005174   0.002482  -2.085 0.042779 *  
log(Population)  0.246836   0.112539   2.193 0.033491 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7137 on 45 degrees of freedom
Multiple R-squared:  0.7404,	Adjusted R-squared:  0.7173 
F-statistic: 32.09 on 4 and 45 DF,  p-value: 1.17e-12
```

From the above summary table of our model, the adjusted $R^2$is 0.7173, telling us that about 71.73% percent variation in life expectancy is explained by our model. Also, the associated p-value $1.17e^{-12}$of the whole model is very small, indicating our model is significant.

“Murder” has negative coefficients -0.29, meaning that we predict a 1 percent increase in murder rate would result in -0.29 years decrease in the mean life expectancy. Similarly, “Frost” has a coefficient -0.00517, indicating that we expect a 1 unit increase in the mean number of days under freezing would bring 0.00517 years decrease in the mean life expectancy. On the other hand, the positive coefficient of “HS Grad” indicates that 1 percentage increase in high school graduation increase mean life expectancy by 0.0546 years. And we expect mean life expectancy to increase 0.5684 years for each ten-fold increase in population. (0.56836 = $\mathbf{0.246836} \times \boldsymbol{ln}(\mathbf{10})$)

5. Conclusion

In conclusion, we are able to predict the mean life expectancy of people in a U.S. state given its population, local murder rate, high school graduation percentage, and the mean number of days with minimum temperature below freezing. In general, states with higher population and high school graduation percentage would have longer life expectancy, while higher murder rates and more days in freezing temperatures would result in shorter life expectancy.

Given that the size of the dataset is limited (including only statistics from each state), the accuracy could be improved if we are able to draw more data by smaller region, for example, census by county. It would also be helpful if we could draw more possible related predictors into the dataset, for example, the elevation of the region, unemployment rate, healthcare coverage, air quality, etc. We should also note that the data we draw is from the US census in 1977, which means that necessary adjustment is needed with updated data for contemporary prediction.

6. Appendix

```
# data preparation
dat=as.data.frame(state.x77)
attach(dat)
names(dat)
pairs(dat[c(4,1,2,3,5,6,7,8)], cex=0.4) #scatterplot matrix
cor(dat)

# Stepwise regression using AIC
mod0=lm(`Life Exp`~1)
mod.all = lm(`Life Exp`~., data=dat) # including all predictors in lm()
step(mod0, scope = list(lower = mod0, upper = mod.all))
mod.AIC = lm(`Life Exp` ~ Murder + `HS Grad` + Frost + Population, data=dat)

# Stepwise regression using F-test
mod0=lm(`Life Exp`~1)
add1(mod0, ~.+Population+Income+Illiteracy+Murder+`HS Grad`+Frost+Area, test = 'F')

# 1.choose 'Murder'
mod1 = update(mod0, ~.+Murder)
add1(mod1, ~.+Population+Income+Illiteracy+`HS Grad`+Frost+Area, test = 'F')

# 2.choose 'HS Grad'
mod2 = update(mod1, ~.+`HS Grad`)
summary(mod2)       #check significance after adding 'HS Grad'
add1(mod2, ~.+Population+Income+Illiteracy+Frost+Area, test = 'F')

# 3.choose 'Frost'
mod3 = update(mod2, ~.+Frost)
summary(mod3)       #check significance after adding 'Frost'
```

```
add1(mod3, ~.+Population+Income+Illiteracy+Area, test = 'F')

# 4.choose 'Pop'
mod4 = update(mod3, ~.+Population)
summary(mod4)       #check significance after adding 'Pop'
add1(mod4, ~.+Income+Illiteracy+Area, test = 'F')

# no more significant predictors, p-values > 0.15
# same model as what we found in AIC

###############################################################

# best subset regression
library(leaps)
mod = regsubsets(cbind(Population, Income, Illiteracy, Murder, `HS Grad`, Frost, Area), `Life Exp`)
summary.mod = summary(mod)
summary.mod$which
names(summary.mod)
summary.mod$adjr2
# from 3rd to 4th, increased almost 2%
# from 4th to 5th, dropping
# so we choose 4 predictors, look back at matrix, find that same as what we found in stepwise regression
summary.mod$cp
# only C_p close to p is the third one, 3.74 close to p=4, three predictors

# residuals analysis
yhat=mod.AIC$fitted.values
e=mod.AIC$residuals
plot(yhat, e, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fit')
```

```
abline(h = 0, lty = 2)

## residual v.s. predictors
par(mfrow=c(2,2))
plot(Murder, e, xlab = 'Murder', ylab = 'Residuals' )
abline(h = 0, lty = 2)
plot(`HS Grad`, e, xlab = 'HS Grad', ylab = 'Residuals' )
abline(h = 0, lty = 2)
plot(Frost, e, xlab = 'Frost', ylab = 'Residuals' )
abline(h = 0, lty = 2)
plot(Population, e, xlab = 'Population', ylab = 'Residuals' )
abline(h = 0, lty = 2)

# residuals analysis for new model
 mod.trans <- lm(`Life Exp` ~ Murder + `HS Grad` + Frost + log(Population))
e2 = resid(mod.trans)
yhat2 = fitted(mod.trans)

## residual v.s. Predictors transformed
par(mfrow=c(2,2))
plot(Murder, e2, xlab = 'Murder', ylab = 'Residuals' )
abline(h = 0, lty = 2)
plot(`HS Grad`, e2, xlab = 'HS Grad', ylab = 'Residuals' )
abline(h = 0, lty = 2)
plot(Frost, e2, xlab = 'Frost', ylab = 'Residuals' )
abline(h = 0, lty = 2)
plot(log(Population), e2, xlab = 'logPopulation', ylab = 'Residuals' )
abline(h = 0, lty = 2)

## normal q-q
par(mfrow=c(1,3))
```

```
plot(yhat, e2, xlab = 'Fitted Values', ylab = 'Residuals' )
abline(h = 0, lty = 2)
qqnorm(e2)
qqline(e2)
hist(e2)

rs=rstandard(mod.trans) # internally studentized residuals
sort(rs) # 2.67 < 3
rsd=rstudent(mod.trans) # studentized deleted residuals
sort(rsd) # 2.88 < 3

n=length(e2)
p=4+1     # four predictors + 1
3*p/n     # rules of thumb, 3 times the mean leverage value
hv=hatvalues(mod.trans)
sort(hv)  # 0.282 .3

summary(mod.trans)
coef(mod.trans)[5]*log(10)
```