U.S. Crime Rates of the 1960s

Michael Albert
Montana Gwynn
Adam Rockett
gwynnm@wwu.edu

Section 1: Introduction

This report studied the connection between various economic and social factors in relation to crime rate. The data set is from Vandaele (1978), and includes the following variables: number of offenses reported to police per 100,000 population (R), number of individuals aged 14-24 per 1000 population (Age), whether a state is in the Southern U.S. (S), mean number of years of education (Ed), 1960 per capita expenditure on police by state and local government (Ex0), 1959 per capita expenditure on police by state and local government (Ex1), labor force participation rate per 1000 civilian urban males between ages 14 and 24 (LF), number of males per 1000 females (M), state population size (N), number of non-whites per 1000 population (NW), unemployment rate of urban males between ages 14 and 24 (U1), unemployment rate of urban males between ages 35 and 39 (U2), median value of family income/assets/transferable goods (W), and number of families per 1000 population who are earning less than half the median income (X).

The goal of this paper is to construct a linear model relating R to the above variables. Using the principle of Variance Inflation Factors (VIF), the following information was obtained about the variables.

Table 1.1 (VIF Values of All Variables)

Var	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
VIF	2.69	4.87	5.04	94.6	98.6	3.67	3.65	2.32	4.12	5.93	4.99	9.96	8.40

There is a strong multicollinearity effect with the variables Ex0 and Ex1, most likely with each other, since they are the amount of money spent by a state government in the years 1959 and 1960 respectively. For a given state, this amount wasn’t going to change drastically from year to year unless some kind of significant legislation was passed. It turns out that removing either Ex1 or Ex0 results in VIF values of smaller than 10 for all the variables.

Table 1.2 (VIF Values with Ex1 Removed)

Var	Age	S	Ed	Ex0	LF	M	N	NW	U1	U2	W	X
VIF	2.68	4.87	4.89	5.32	3.53	3.64	2.28	4.08	5.93	4.93	9.93	8.37

Section 2: The Regression Models

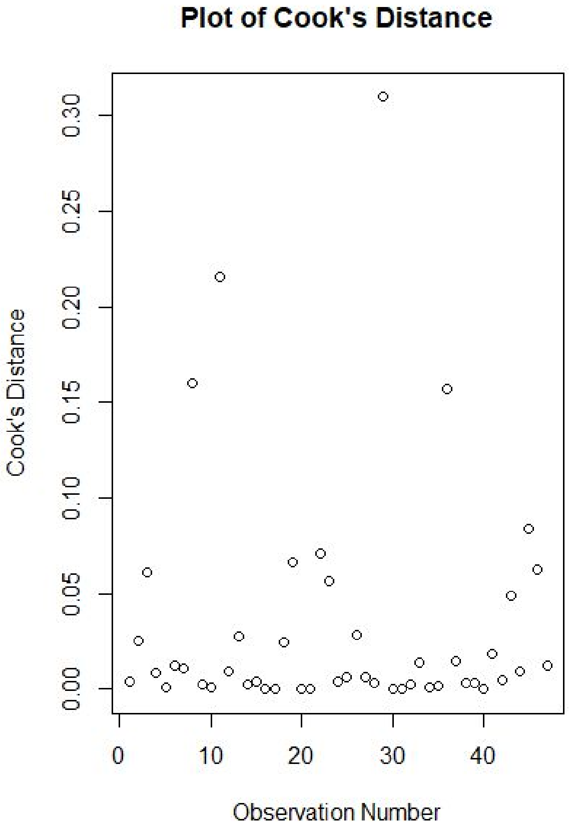
Using stepwise, forward, and backward regression with base model, R = β_0+ϵ and a full model of the variables in table 1.2, it was found each time that the model of most significance is

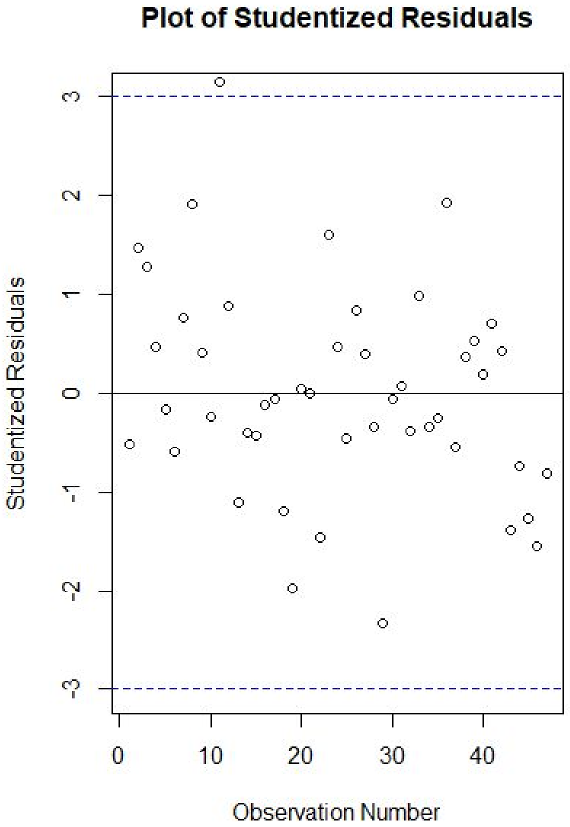
$R = \beta_0 + \beta_1 Ex0 + \beta_2 X + \beta_3 Ed + \beta_4 Age + \beta_5 U2 + \beta_6 W + \epsilon$. This model (Model A) will be analyzed first and terms with interaction will be considered later to coincide with the intuition about how crime rates respond to these variables.

Table 2.1 (Summary: Model A)

	Estimate	Std. Error	T value	Pr(>|t|)
Intercept	-618.5028	108.2456	-5.714	1.19e-06
Ex0	1.0507	0.1752	5.996	4.78e-07
X	0.8236	0.1815	4.538	5.10e-05
Ed	1.8179	0.4803	3.785	0.000505
Age	1.1252	0.3509	3.207	0.002640
U2	0.8282	0.4274	1.938	0.059743
W	0.1596	0.0939	1.699	0.097028

All of the variables are significant in the model at α=.05, except for U2 and W. Below the regression diagnostics for Model A are presented, first looking for outliers, and influential/ high leverage points.


Plot of Cook's Distance


Plot of Studentized Residuals

From the Cook's distance plot above was found that observation #11 and observation #29 had relatively high Cook's distances making them influential observations. Additionally it was found that observation #11 had a studentized residual greater than 3 and therefore was an outlier as well. It was concluded that observation #11 should be omitted from the data for these reasons. Upon removal of the observation, a new model was returned by stepwise regression which will be called Model B: R = β_0+ β_1Ex0 + β_2X + β_3Ed + β_4Age + β_5U2 + ϵ. This is the same as Model A, but without the variable W, which makes sense because it wasn't significant in Model A at the 0.05 level. As before, this same model was also returned by backwards and forwards regression. From here we consider the addition of interaction terms to Model B.

The model with interaction is R = β_0+ β_1Ex0 + β_2X + β_3S + β_4Age + β_5U2 + β_6S + β_7S*Ed + β_8Ed + ϵ (Model I). Here S is an indicator variable for whether or not a state is in the southern United States. It is well known that crime experienced a sudden surge in the year 1960 especially in southern states, which were less affluent than the northern states (Brooke). One way to potentially explain this is with the difference in years spent in education between the two categories of states. According to an article from Sharkey et. al (2016) violent crime is heavily associated with children’s peer networks and adult influences, so analyzing education is clearly justified.

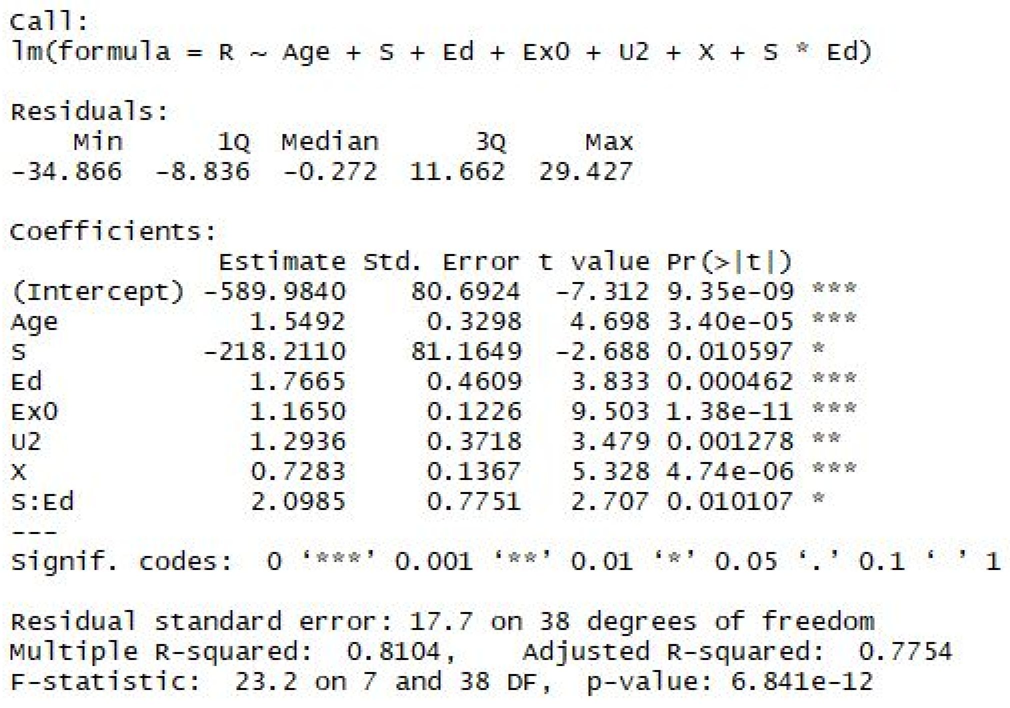
In order to determine which of Model B and Model I was a better fit, a partial F-test was run between them with a null hypothesis of $\beta_6 = \beta_7$=0.

Analysis of Variance Table
Model 1: R ~ Ex0 + X + Ed + Age + U2
Model 2: R ~ Ex0 + X + Ed + Age + U2 + S + S * Ed

Model	Res. Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	39	14172				
2	37	11822	2	2350.2	3.6778	0.03493

A p-value of 0.03493 was returned from the partial F test, which lead to a rejection of the null hypothesis. Therefore, we conclude that β_6 and/or β_7 are nonzero and Model I better represents the data than Model B does.

When best subset selection was run on the full model with the interaction term, the following R output was presented:



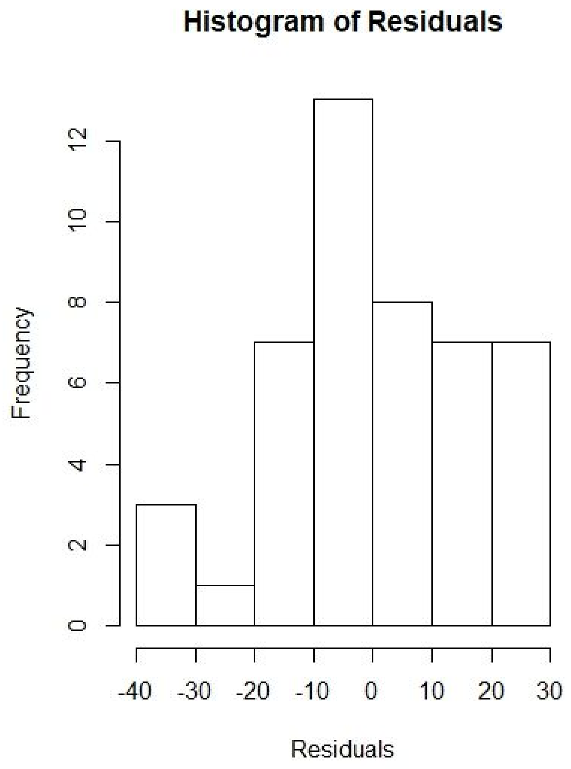
```
Call:
lm(formula = R ~ Age + S + Ed + Ex0 + U2 + X + S * Ed)

Residuals:
    Min      1Q  Median      3Q     Max 
-34.866  -8.836  -0.272  11.662  29.427 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -589.9840    80.6924  -7.312 9.35e-09 ***
Age            1.5492     0.3298   4.698 3.40e-05 ***
S           -218.2110    81.1649  -2.688 0.010597 *  
Ed             1.7665     0.4609   3.833 0.000462 ***
Ex0            1.1650     0.1226   9.503 1.38e-11 ***
U2             1.2936     0.3718   3.479 0.001278 ** 
X              0.7283     0.1367   5.328 4.74e-06 ***
S:Ed           2.0985     0.7751   2.707 0.010107 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.7 on 38 degrees of freedom
Multiple R-squared:  0.8104,	Adjusted R-squared:  0.7754 
F-statistic:  23.2 on 7 and 38 DF,  p-value: 6.841e-12
```

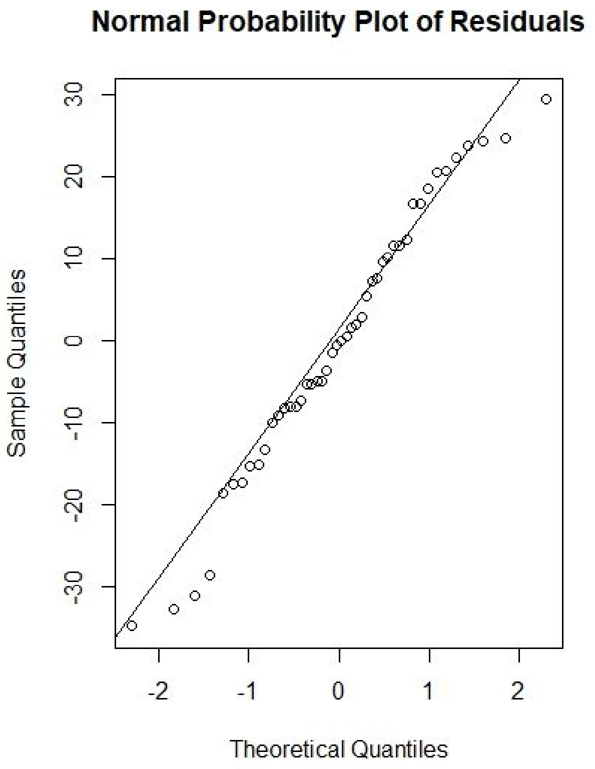
From the R output above, Age, S, Ed, Ex0, U2, X, and the interaction term, S*Ed, are all significant at a 0.05 confidence level. This is the same model that was given to us with the stepwise regression, so therefore it is a good candidate for the final model based on this set of data.

Section 3: Model Diagnostics

Our model was developed under the assumption that observation #11 was an outlier and should be removed from the data. Without this assumption, our model would have included an extra variable, W, which was not included upon removal of the outlier. However, the high Cook's distance and studentized residual of the point indicated that it was not representative of the general trend of the data so its removal was appropriate.

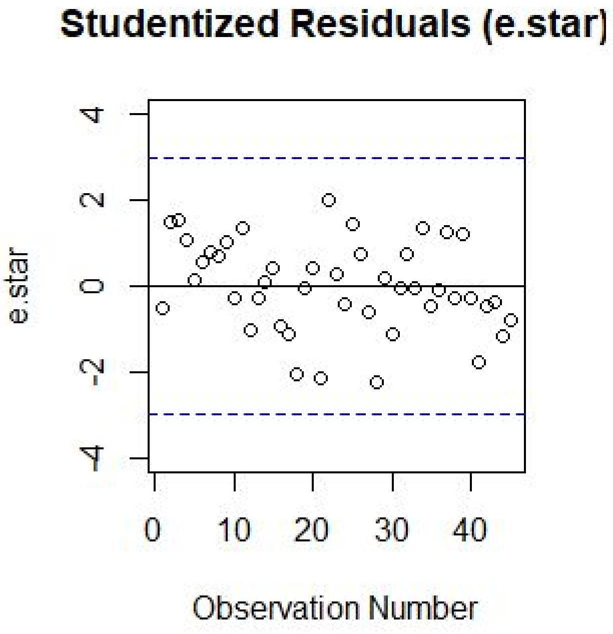
In analyzing our model's adequacy, we tested for the normality of its residuals by generating a normal probability plot, the histogram of residuals, and by running a Shapiro-Wilk normality test on the residuals.





The points representing the sample quantities and theoretical quantities tend to fall on the 45 degree line, indicating that the residuals do indeed follow a normal distribution.

When the Shapiro-Wilk normality test was run on the data, it returned a p-value of 0.3592. From the high p-value returned by the Shapiro-Wilk normality test (0.3592>0.05), we conclude that there is no evidence against the normality of the residuals of our model. This backs up the conclusion drawn from the normal probability plot and supports the adequacy of the model.



Further, it can be see from the studentized residuals plot that the data does not have any outliers in relation to our model.

Section 4: Value Prediction

Using values for the variables in Model (I) corresponding to data for Washington State in the year 2015, a 95% prediction interval for R is constructed and this is compared with the known crime rate. This tests the validity of Model (I) as a predictive tool.

Not all of the data gathered below was formulated in the same way as in Vandalae (1978), so some approximations were made. For example, from the U.S Census in 2017, it was found that the number of people under age 18 in Washington State per 1000 is 221, but the number of people aged 14-24 per 1000 was not available, so an estimate of 300 was made for this number.

Further, education in terms of years of schooling completed is not a statistic that is used as of 1990 (Census Bureau), so the average of the data set from this report is taken as the value of Ed for the prediction interval.

In considering X, one must recognize that in general income inequality has gone up since 1978, and is generally considered to be at historic and dangerous levels, according to a recent UN report (Pilkington). Therefore, the value of X in this computation will be multiple standard deviations from the mean of X in this data set, which will affect the validity of this model. The median income in Washington is around $60,000 and taking the national percentage of households earning below that amount (43.1%) , the value of X = 431 is used in this computation. Data for U2 from Washington State was not available so the mean of U2 from this data set was used.

Washington State Data Estimations:

Table 4.1

S	Ed (mean)	Ex0	Age	X	U2 (mean)	R (known)
0	105	450	300	431	33	3,984.7

The 95% prediction interval is (760.7845, 1121.426) with a fitted value of R=941.1052. Clearly the actual rate of reported crime for Washington State is not accurately predicted by the model.

Section 5: Conclusions

Based on model I, some of the coefficients in the least squares equation are intuitive in relation to their definition. For example, the coefficient on X, which measures poverty and income inequality, is .702. It is expected that as poverty increases, the overall crime rate should increase in turn. Further, the coefficient on U2, a measure of unemployment, is 1.29. Again, it is expected that as unemployment increases, crime rate goes up. Also Age, a measure of the proportion of young people in the population, has a coefficient (1.5) that makes intuitive sense. Young people are more often associated with crime, especially property crime, which in the case of Washington State, made up about 90% of total reported crimes. However, it doesn't make sense that the coefficient for Ed and Ex0 are positive. One would expect that as education and government expenditure on police increase, crime rate goes down. The notion that more education implies a more devious and exploitative population could be entertained, but this is unlikely. The model is limited in predicting current crime rates.

Works Cited

“Current Population Demographics and Statistics for Washington by Age, Gender and Race.” *SuburbanStats.org*, 2017, suburbanstats.org/population/how-many-people-live-in-washington.

Dietrich-Williams, Ayn. “The FBI Releases 2015 Crime Statistics for Washington State.” *FBI*, FBI, 26 Sept. 2016.

Pilkington, Ed. “Trump's 'Cruel' Measures Pushing US Inequality to Dangerous Level, UN Warns.” *The Guardian*, Guardian News and Media, 1 June 2018.

“Police and Corrections Expenditures.” *Urban Institute*, 20 Apr. 2018, www.urban.org/policy-centers/cross-center-initiatives/state-local-finance-initiative/state-and-local-backgrounders/police-and-corrections-expenditures.

Sharkey, Patrick, et al. “Poverty and Crime.” *Oxford Handbooks*, Oxford University Press, 16 June 2017.

“U.S. Household Income Distribution.” *Statista*, Sept. 2017.

Vandaele, Walter. Participation in Illegitimate Activities: Ehrlich Revisited, 1960. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1992-02-16. https://doi.org/10.3886/ICPSR08677.v1

Appendix (R code)

```
# Reads data into a table to manipulate
data=read.table("project_data.txt", header=TRUE)
data
attach(data)

# Correlation Matrix to check for collinearity
pairs(data)
cor(data)
# Remove Ex1 from the model since it shows strong collinearity with Ex0

# The full model
modelFull=lm(R ~ Age + S + Ed + Ex0 + LF + M + N + NW + U1 + U2 + W + X, data=data)
summary(modelFull)

# Checks the VIF values of the variables in the data set
install.packages("car")
library(car)
vif(modelFull)

### Stepwise Regression
model0=lm(R~1, data=data) ### constant mean model
MSEk = anova(modelFull)["Residuals", "Mean Sq"]
step(model0, scope=list(lower=~1, upper=modelFull), direction="both", scale=MSEk)
# Returns model R ~ Ex0 + X + Ed + Age + U2 + W

### Forward Selection
model0=lm(R~1, data=data) ### constant mean model
MSEk = anova(modelFull)["Residuals", "Mean Sq"]
step(model0, scope=list(lower=~1, upper=modelFull), direction="forward", scale=MSEk)
# Returns model R ~ Ex0 + X + Ed + Age + U2 + W

### Backwards Elimination
model0=lm(R~1, data=data) ### constant mean model
MSEk = anova(modelFull)["Residuals", "Mean Sq"]
step(modelFull, scope=list(lower=~1, upper=modelFull), direction="backward", scale=MSEk)
# Returns model R ~ Age + Ed + Ex0 + U2 + W + X
# The model given from Different Variable Selection methods
modelA = lm(R ~ Age + Ed + Ex0 + U2 + W + X)
summary(modelA)

# Residuals for modelA
resid=modelA$residuals
resid
y.hat=predict(modelA)

install.packages("MASS")
library(MASS)
e.star = studres(modelA)

#Plot of Cook's Distancs
influence.measures(modelA)
cook.d.model=influence.measures(modelA)$infmat[,"cook.d"]
par(mfrow=c(1,2))
plot(cook.d.model, main="Plot of Cook's Distance", ylab="Cook's Distance", xlab="Observation
Number")

#Plot of Studentized Residuals
plot(e.star, ylim=c(-3,3), ylab="Studentized Residuals",
    xlab="Observation Number", main="Plot of Studentized Residuals")
abline(h=3, col="blue", lty=2)
abline(h=-3, col="blue", lty=2)
abline(h=0)

# Remove observation 11 from the data set
data=data[-11,]
data
attach(data)

### Stepwise Regression
model0=lm(R~1, data=data) ### constant mean model
MSEk = anova(modelFull)["Residuals", "Mean Sq"]
step(model0, scope=list(lower=~1, upper=modelFull), direction="both", scale=MSEk)
# Returns model R ~ Ex0 + X + Ed + Age + U2

# Model given after outlier was removed
```

```
modelB=lm(R ~ Ex0 + X + Ed + Age + U2)
summary(modelB)

# Model with interaction term
modelI=lm(R ~ Ex0 + X + Ed + Age + U2 + S + S*Ed)
summary(modelI)

# ANOVA Table comparing models with and without interaction term
anova(modelB, modelI)
# Since p-value is small, we decide that the larger model is the better fit for the data

### Best subset selection
### use install.packages("leaps") to download and install the leaps package.
install.packages("leaps")
library(leaps)
best.subset.cp=leaps(x=cbind(Age,S,Ed,Ex0,LF,M,N,NW,U1,U2,W,X,S*Ed), y=R,
method="Cp")
best.subset.cp

min.cp.value=min(best.subset.cp$Cp)
min.cp.value

min.cp.location=which.min(best.subset.cp$Cp)
min.cp.location

best.subset.cp$which[min.cp.location,]
# Model returned is the same as modelI

# Residuals for modelI
resid=modelI$residuals
resid
y.hat=predict(modelI)

e.star = studres(modelI)

#Plot of Cook's Distancs for modelI
influence.measures(modelI)
cook.d.model=influence.measures(modelI)$infmat[,"cook.d"]
par(mfrow=c(1,2))
```

```
plot(cook.d.model, main="Plot of Cook's Distance", ylab="Cook's Distance", xlab="Observation Number")

#Plot of Studentized Residuals for modelI
plot(e.star, ylim=c(-3,3), ylab="Studentized Residuals",
    xlab="Observation Number", main="Plot of Studentized Residuals")
abline(h=3, col="blue", lty=2)
abline(h=-3, col="blue", lty=2)
abline(h=0)

# Histogram of Residuals, and Normality probability plot of residuals
resid=modelB$residuals
par(mfrow=c(1,2))
hist(resid, main="Histogram of Residuals", xlab="Residuals")
qqnorm(resid, main="Normal Probability Plot of Residuals")
qqline

# Prediction Interval
newdata=data.frame(S=0, Age=300, Ed=105, Ex0=450, U2=33, X=431)
predict(modelI, newdata, interval="predict")
```