STA3064

Regression Case Study

Background

In recent years, teen pregnancy has been a major issue in the United States. The objective of this study is to find out which economic indicators can be used to predict the teenage birth rate in large metropolitan areas. This study will focus on the city of Chicago. The hope is that by finding factors that contribute to teen pregnancy, a better understanding of prevention can be found to inform public policy makers.

Data Description and Variables

The data used comes from a larger data set containing 27 variables describing public health in various Chicago neighborhoods. The information was compiled by the Chicago Department of Public Health (CDPH). The accompanying file, *teen.csv*, contains eight variables and 77 observations. Designations of each variable follow. Variable names are given followed by a brief description in parentheses.

Predictor Variables:

BelowPovLev (Below poverty level -- percent of households)
Crowded (Crowded housing -- percent of occupied housing units)
Dependency (Percent of people aged less than 16 or more than 64 years old)
NoHSDiploma (No high school diploma -- percent of people aged 25 years or older)
Income (Per capita income -- 2011 inflation-adjusted dollars)
Unemployment (Percent of people not in labor force aged 16 years and older)

Response Variable:

BirthRate (Teen birth rate -- per 1,000 females aged 15-19)

Tasks to Complete:

1. Data Exploration:

a. Keeping the text file external to your SAS code (i.e., do not use data lines), read your data into SAS. Include the data step used to get your data into a SAS data set. Print the first 20 observations. Comment on any additional data manipulation that was necessary.
b. Using your SAS data set, produce a scatterplot matrix of all variables (PROC SGSCATTER) and a correlation matrix (PROC CORR) of all variables and all observations. Note any interesting characteristics in the relationships revealed by the above procedures.

2. Model Fitting and Analysis:

From your analysis in the Data Exploration section above, assume that you select the variable, Unemployment, as the predictor variable that looks most promising in predicting your response. Use the following items to guide this portion of your analysis:
a. Fit a simple linear regression model. Provide the model equation.
b. Interpret the R2 for your fit.
c. Perform a residual analysis to determine if all model assumptions are being met.

d. Explore potential transformations on your response variable (even if your residual analysis indicates a transformation is not required just to confirm). If a transformation is indicated, refit your model and assess.
e. For your original simple regression model (2a) produce a 95% confidence interval for the true slope of your regression line. Interpret.
f. Create a 95% bootstrap confidence interval for the slope based on quantiles from the bootstrap distribution of at least 1000 replications. Compare your results to your confidence interval based on normal theory above.
g. Conduct the ANOVA test for the slope for the original in 2(a). Discuss whether this test indicates that the simple linear model is effective.
h. Suppose a new value of Unemployment of interest is 1.2% (*x*= 1.2*). Produce both 95% confidence and prediction intervals around the predicted response for *x**. Interpret both intervals.

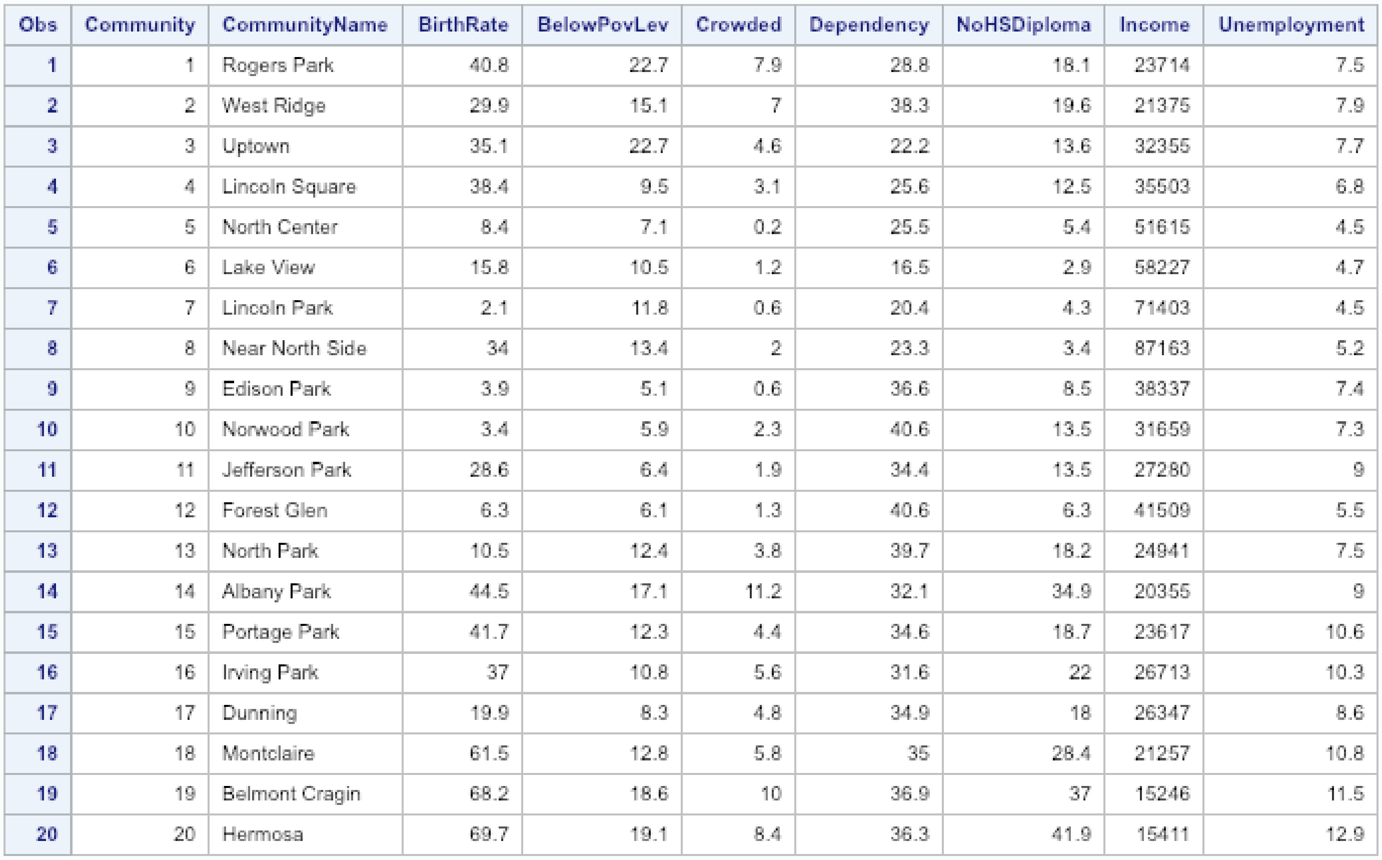
Now include all of your predictor variables in your analysis and use the following items to guide your study:

i. Explore the potential impact of multicollinearity on your full model using the original (non-transformed) data.
j. Use a variable selection method to assist in fitting your best multiple regression model (if you had to exclude any variables due to multicollinearity, do not include them in the variable selection procedure).
k. Interpret the R2 for your best model.
l. Perform a residual analysis to determine if all model assumptions are being met.
m. Perform a nested F-test comparing your full model (all predictors included) to a reduced model of interest.
n. Make note of any outliers, high-leverage points, and influential points. Describe how you think they may be impacting the fit of your model. Discuss whether removal of any points is justified.
o. At this point, state what you believe is your best model based on the above analysis. What would be your next step? (You do not need to perform the potential action, just discuss it.)

1) Data Exploration

1(A)

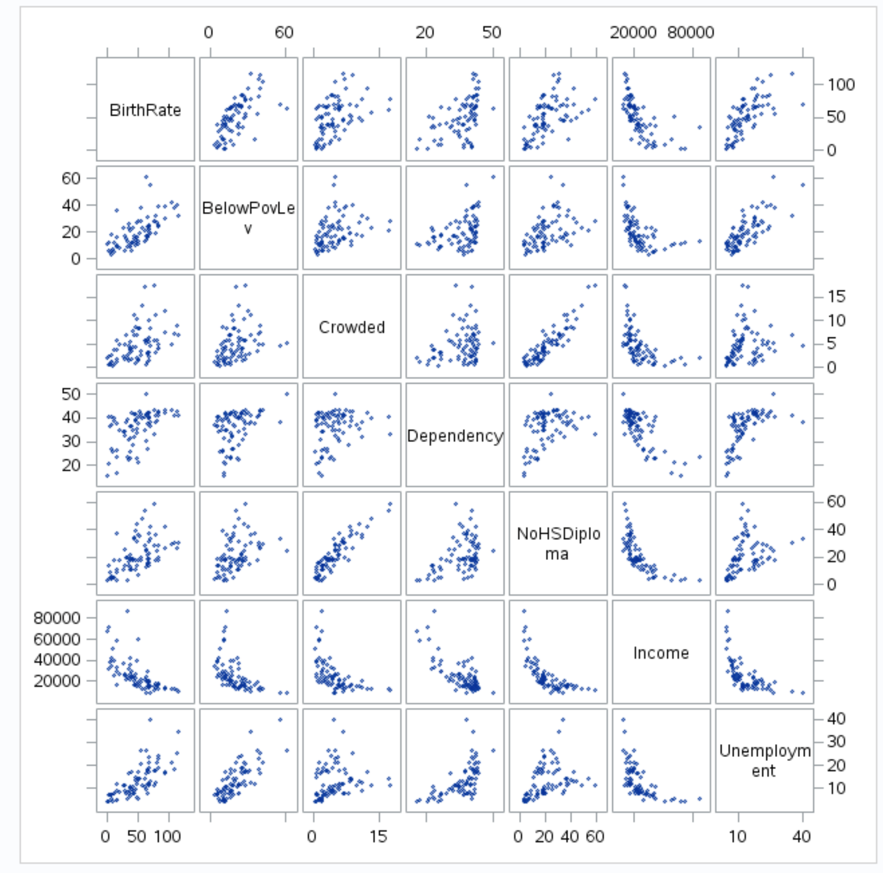
First 20 observations:



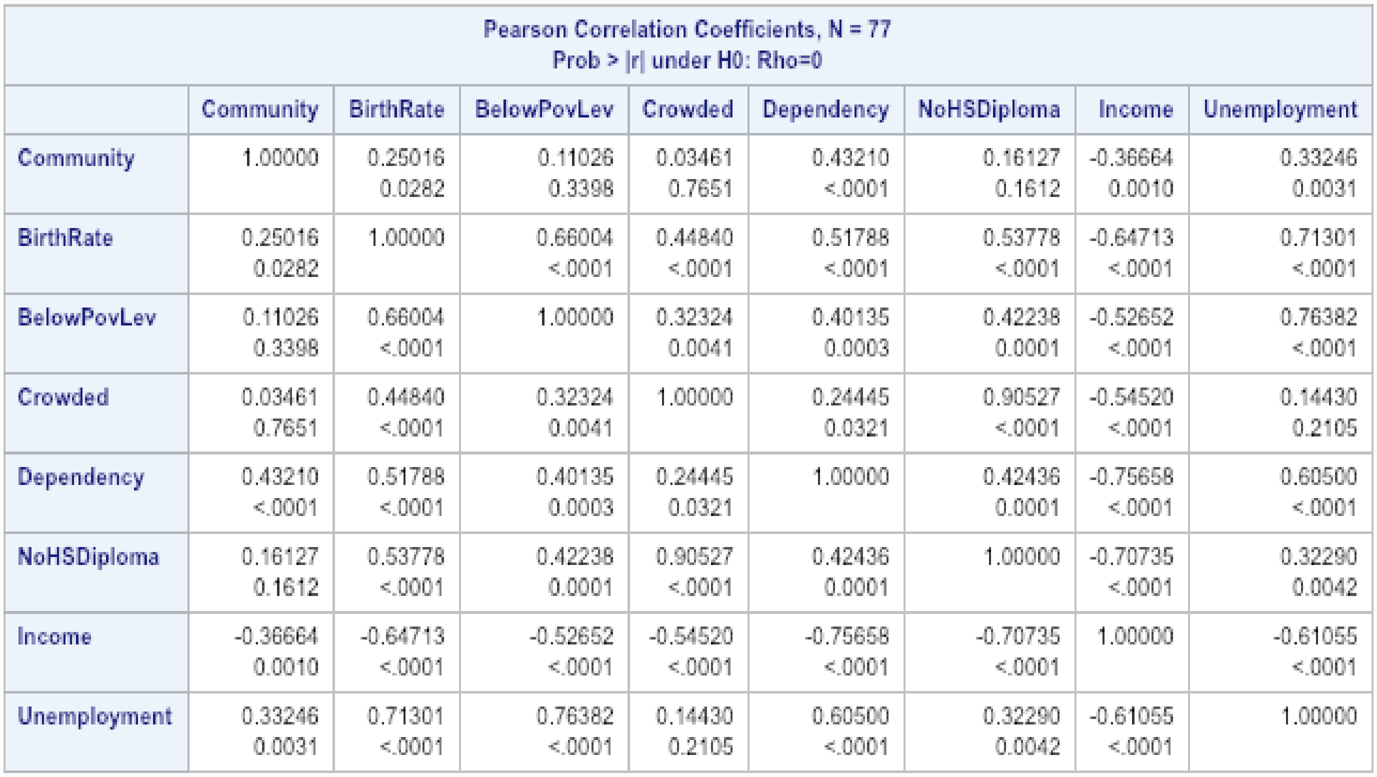
Obs	Community	CommunityName	BirthRate	BelowPovLev	Crowded	Dependency	NoHSDiploma	Income	Unemployment
1	1	Rogers Park	40.8	22.7	7.9	28.8	18.1	23714	7.5
2	2	West Ridge	29.9	15.1	7	38.3	19.6	21375	7.9
3	3	Uptown	35.1	22.7	4.6	22.2	13.6	32355	7.7
4	4	Lincoln Square	38.4	9.5	3.1	25.6	12.5	35503	6.8
5	5	North Center	8.4	7.1	0.2	25.5	5.4	51615	4.5
6	6	Lake View	15.8	10.5	1.2	16.5	2.9	58227	4.7
7	7	Lincoln Park	2.1	11.8	0.6	20.4	4.3	71403	4.5
8	8	Near North Side	34	13.4	2	23.3	3.4	87163	5.2
9	9	Edison Park	3.9	5.1	0.6	36.6	8.5	38337	7.4
10	10	Norwood Park	3.4	5.9	2.3	40.6	13.5	31659	7.3
11	11	Jefferson Park	28.6	6.4	1.9	34.4	13.5	27280	9
12	12	Forest Glen	6.3	6.1	1.3	40.6	6.3	41509	5.5
13	13	North Park	10.5	12.4	3.8	39.7	18.2	24941	7.5
14	14	Albany Park	44.5	17.1	11.2	32.1	34.9	20355	9
15	15	Portage Park	41.7	12.3	4.4	34.6	18.7	23617	10.6
16	16	Irving Park	37	10.8	5.6	31.6	22	26713	10.3
17	17	Dunning	19.9	8.3	4.8	34.9	18	26347	8.6
18	18	Montclaire	61.5	12.8	5.8	35	28.4	21257	10.8
19	19	Belmont Cragin	68.2	18.6	10	36.9	37	15246	11.5
20	20	Hermosa	69.7	19.1	8.4	36.3	41.9	15411	12.9

1(B)

Scatterplot for all variables plotted against each other:



The scatter plots for **Crowded** vs **NoHSDiploma** show a positive linear relationship between variables.



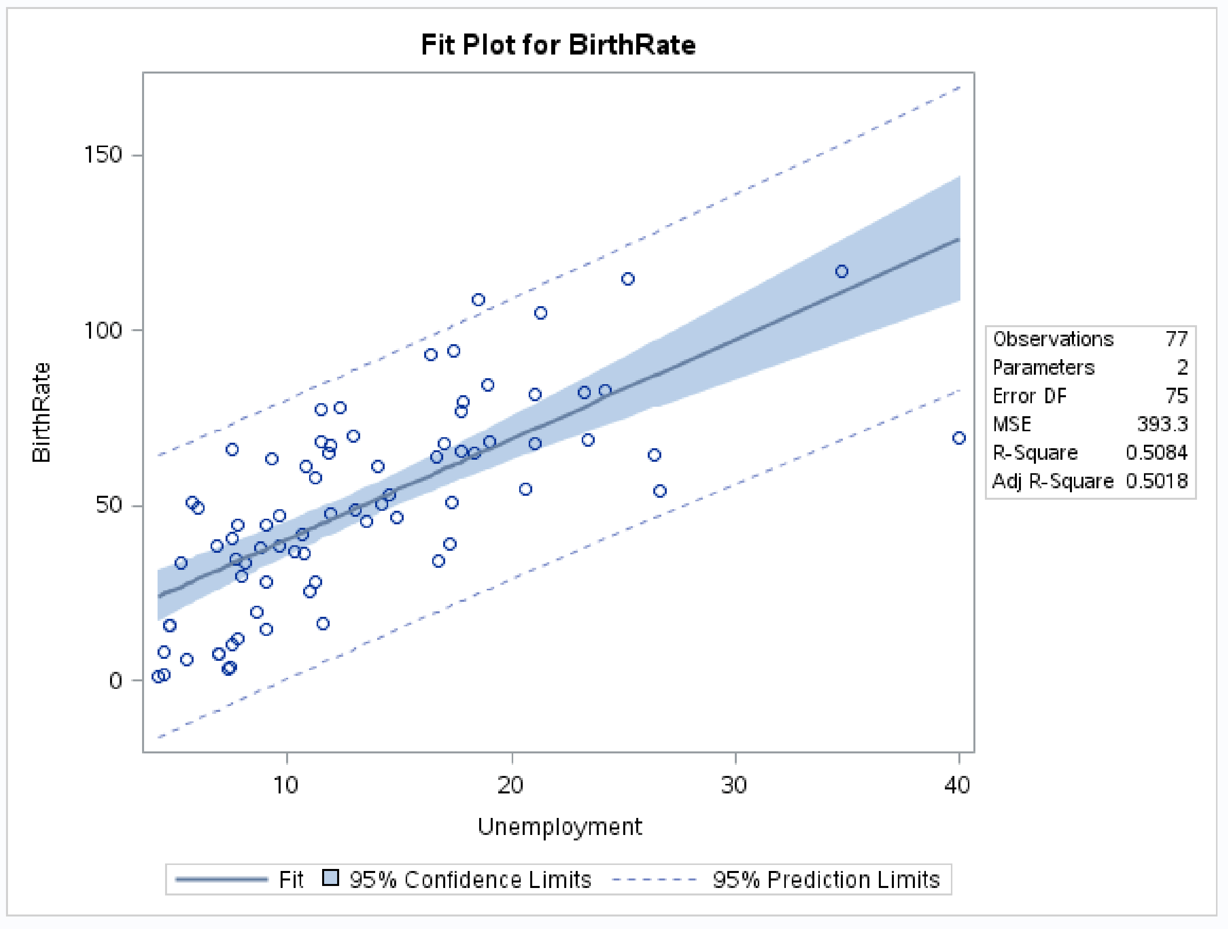
Pearson Correlation Coefficients, N = 77
Prob > |r| under H0: Rho=0

	Community	BirthRate	BelowPovLev	Crowded	Dependency	NoHSDiploma	Income	Unemployment
Community	1.00000	0.25016 0.0282	0.11026 0.3398	0.03461 0.7651	0.43210 <.0001	0.16127 0.1612	-0.36664 0.0010	0.33246 0.0031
BirthRate	0.25016 0.0282	1.00000	0.66004 <.0001	0.44840 <.0001	0.51788 <.0001	0.53778 <.0001	-0.64713 <.0001	0.71301 <.0001
BelowPovLev	0.11026 0.3398	0.66004 <.0001	1.00000	0.32324 0.0041	0.40135 0.0003	0.42238 0.0001	-0.52652 <.0001	0.76382 <.0001
Crowded	0.03461 0.7651	0.44840 <.0001	0.32324 0.0041	1.00000	0.24445 0.0321	0.90527 <.0001	-0.54520 <.0001	0.14430 0.2105
Dependency	0.43210 <.0001	0.51788 <.0001	0.40135 0.0003	0.24445 0.0321	1.00000	0.42436 0.0001	-0.75658 <.0001	0.60500 <.0001
NoHSDiploma	0.16127 0.1612	0.53778 <.0001	0.42238 0.0001	0.90527 <.0001	0.42436 0.0001	1.00000	-0.70735 <.0001	0.32290 0.0042
Income	-0.36664 0.0010	-0.64713 <.0001	-0.52652 <.0001	-0.54520 <.0001	-0.75658 <.0001	-0.70735 <.0001	1.00000	-0.61055 <.0001
Unemployment	0.33246 0.0031	0.71301 <.0001	0.76382 <.0001	0.14430 0.2105	0.60500 <.0001	0.32290 0.0042	-0.61055 <.0001	1.00000

The Pearson Correlation Coefficient chart above confirms the correlation between **Crowded** and **NoHSDiploma** by observing. It seem that the variables have a strong correlation (0.90527)**.**

2) Model Fitting and Analysis

2(A)



Observations	77
Parameters	2
Error DF	75
MSE	393.3
R-Square	0.5084
Adj R-Square	0.5018

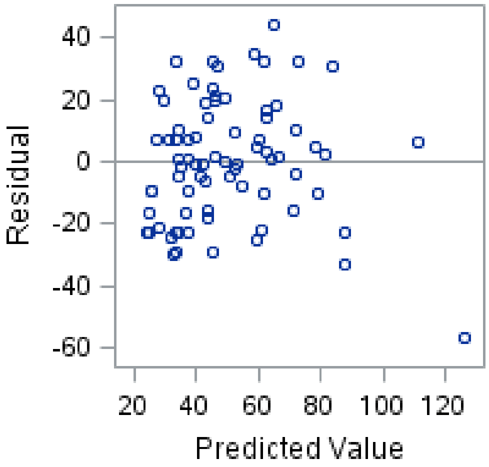
The equation for this model is

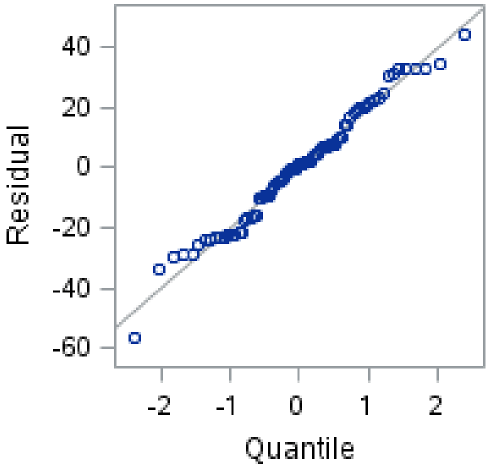
$$BirthRate = \beta_0 + \beta_1 Unemployment$$

2(B)

The R-Square is 0.5084 which means that **50.84%** of the variation of teenage female birth rates(**BirthRate** variable) can be explained by the percent of people not in the labor force aged 16 years and older(**Unemployment** variable).

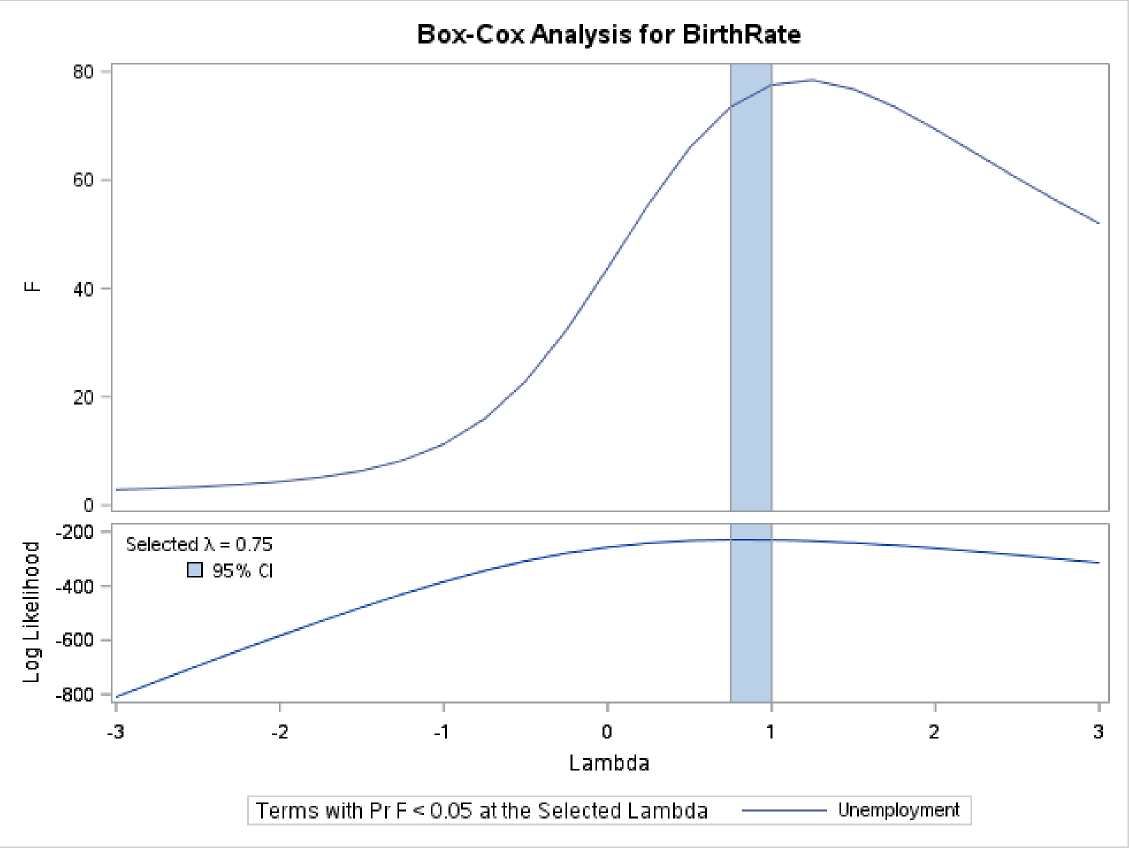
2(C)



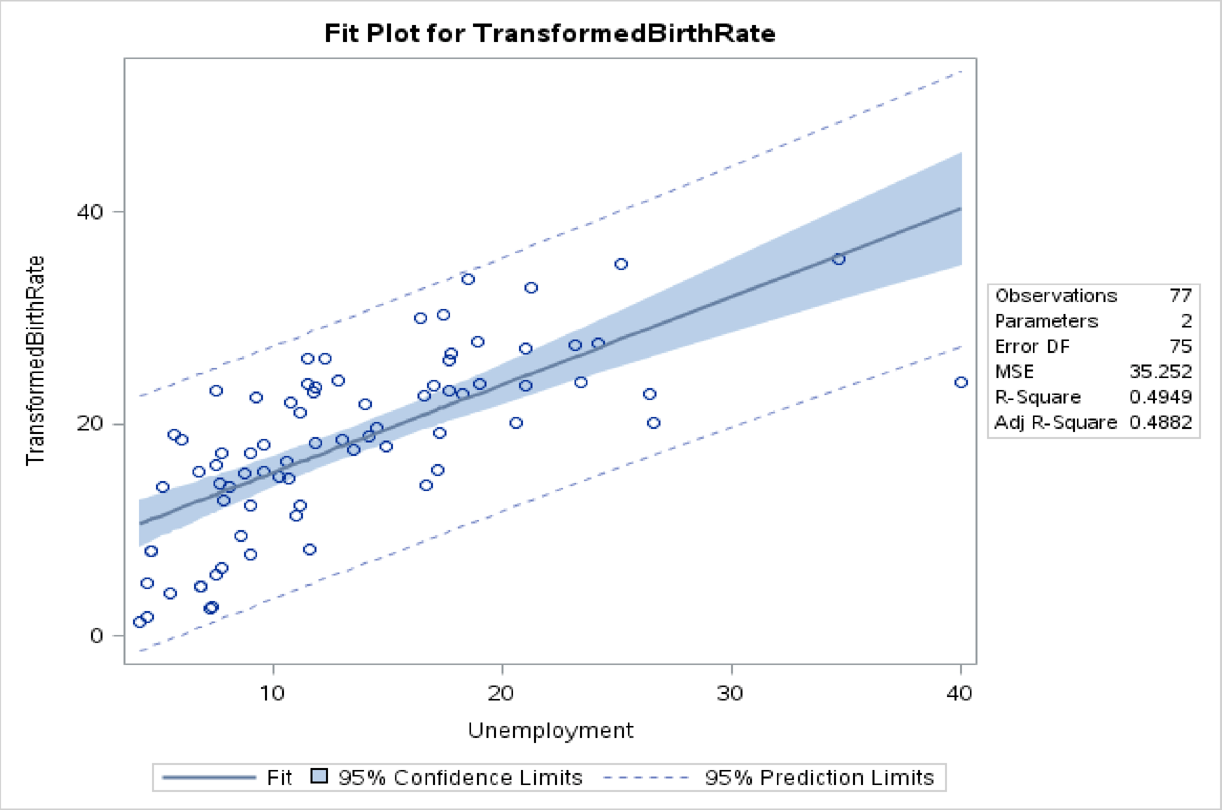


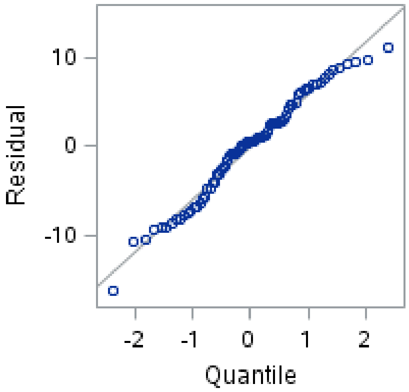
The graph of residual vs predicted values shows that the assumption of constant variance is met since the points are scattered randomly around 0 with what appears to be an outlier that has a predicted value above 120. The residual vs quantile plot graph show that the assumption of normality is also met.

2(D)



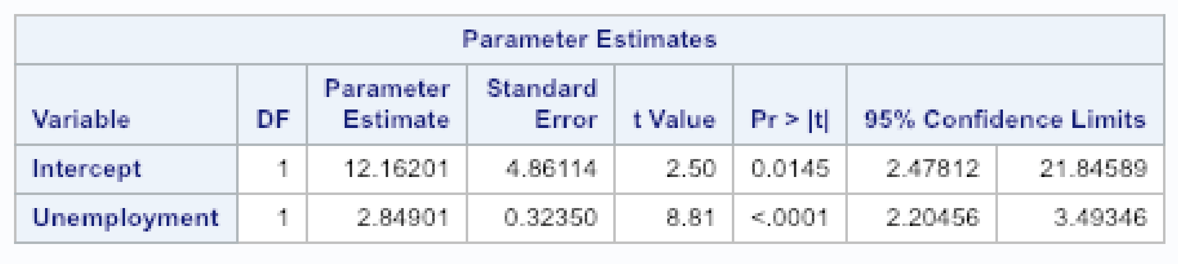
The Box-Cox method shows that the confidence Interval for the lambda value is between a number around 0.7 and 1. Having 1 in the interval indicates that this lambda value might not be useful for transformation since y^1=y. Nonetheless, I transformed the response using the value of lambda(0.75)





We can observe that R-Square has decreased and the residual vs quantile plot hasn't changed much indicating that this transformation is not be beneficial.

2(E)

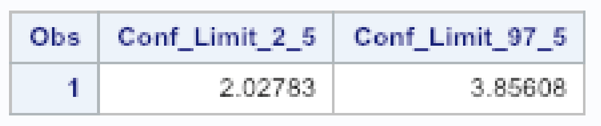


Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	12.16201	4.86114	2.50	0.0145	2.47812	21.84589
Unemployment	1	2.84901	0.32350	8.81	<.0001	2.20456	3.49346

The 95% confidence of the interval for the slope is (2.20456,3.49346). This means that we are 95% confident that the population slope falls within the interval (2.20456, 3.49346).

2(F)

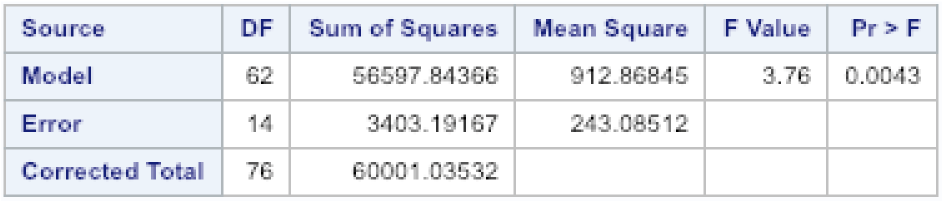
Bootstrap 95% Confidence Interval:



Obs	Conf_Limit_2_5	Conf_Limit_97_5
1	2.02783	3.85608

This shows that the bootstrap CI is narrower than the normal theory CI indicating that there is more variation on average for this model

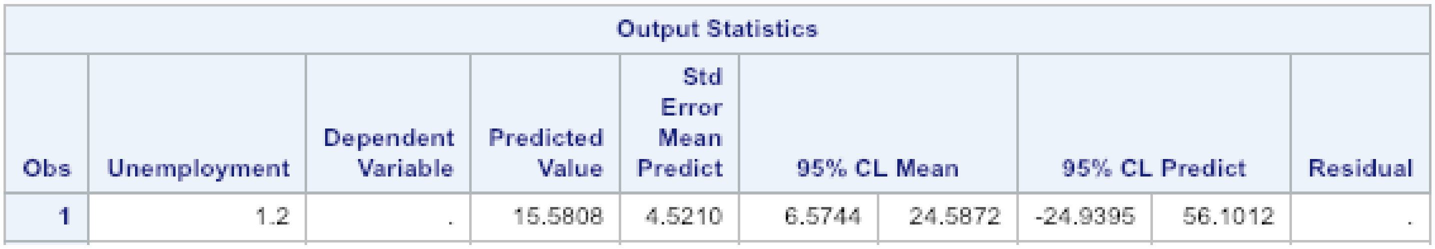
2(G)



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	62	56597.84366	912.86845	3.76	0.0043
Error	14	3403.19167	243.08512		
Corrected Total	76	60001.03532			

The ANOVA test for the model in 2(a) has a p-value of 0.0043. This indicates that at a 5% significance level, we can conclude that the simple linear model is effective.

2(H)

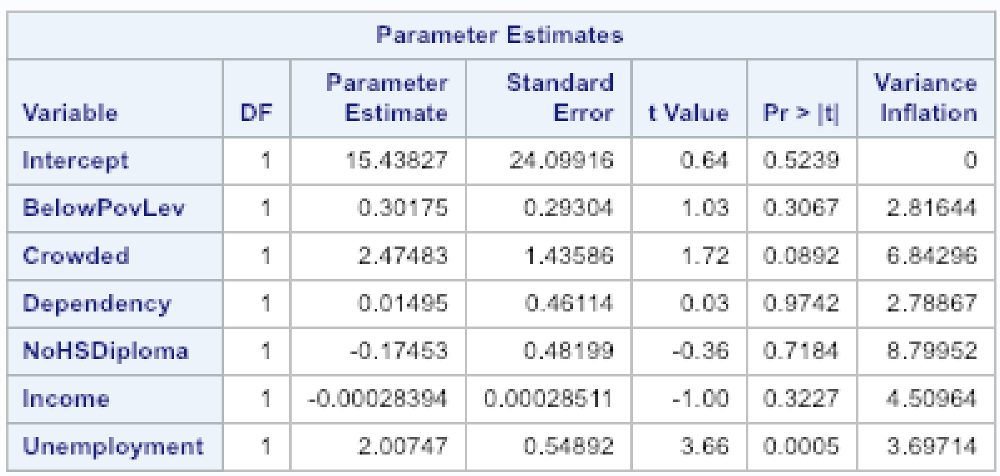


Output Statistics									
Obs	Unemployment	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	1.2	.	15.5808	4.5210	6.5744	24.5872	-24.9395	56.1012	.

The 95% confidence interval for the mean birthrate of female teenagers with 1.2 percent of unemployment is (6.5744,24.5872) and the 95% prediction interval is (-24.9395,56,1012). The width for the 95% confidence interval for the mean birthrate teenage females with 1.2 percent of unemployment is smaller than the width for the 95% confidence interval for a particular birthrate of a teenage female with 1.2 percent of unemployment because there is more variation in the a particular female teenager than if you take the mean of the female teenagers.

2(I)

Full Model VIF:

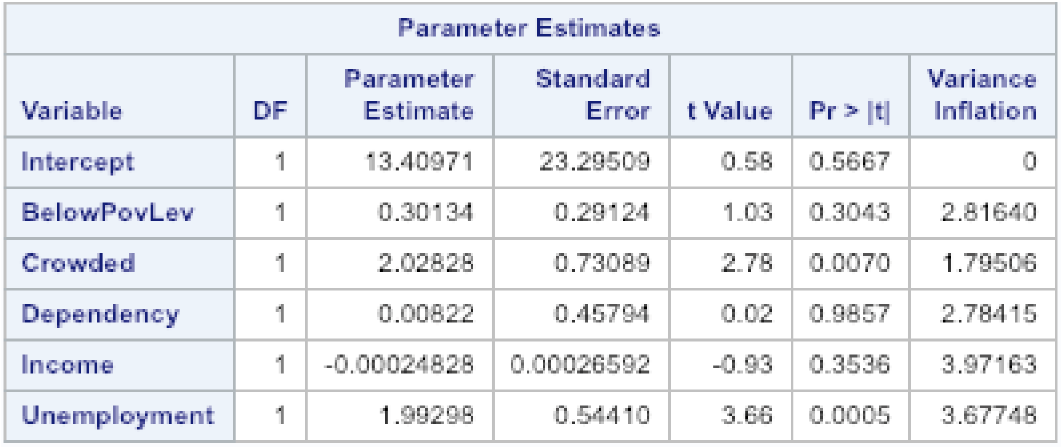


Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	15.43827	24.09916	0.64	0.5239	0
BelowPovLev	1	0.30175	0.29304	1.03	0.3067	2.81644
Crowded	1	2.47483	1.43586	1.72	0.0892	6.84296
Dependency	1	0.01495	0.46114	0.03	0.9742	2.78867
NoHSDiploma	1	-0.17453	0.48199	-0.36	0.7184	8.79952
Income	1	-0.00028394	0.00028511	-1.00	0.3227	4.50964
Unemployment	1	2.00747	0.54892	3.66	0.0005	3.69714

The estimates show that **Crowded** and **NoHSDiploma** have a variance inflation factor above 5. **NoHSDiploma** has the highest variance inflation factor so I decided to remove this variable from the model since it is a cause for multicollinearity and tried a regression model without **NoHSDiploma.**

Model Without NoHSDiploma VIF:

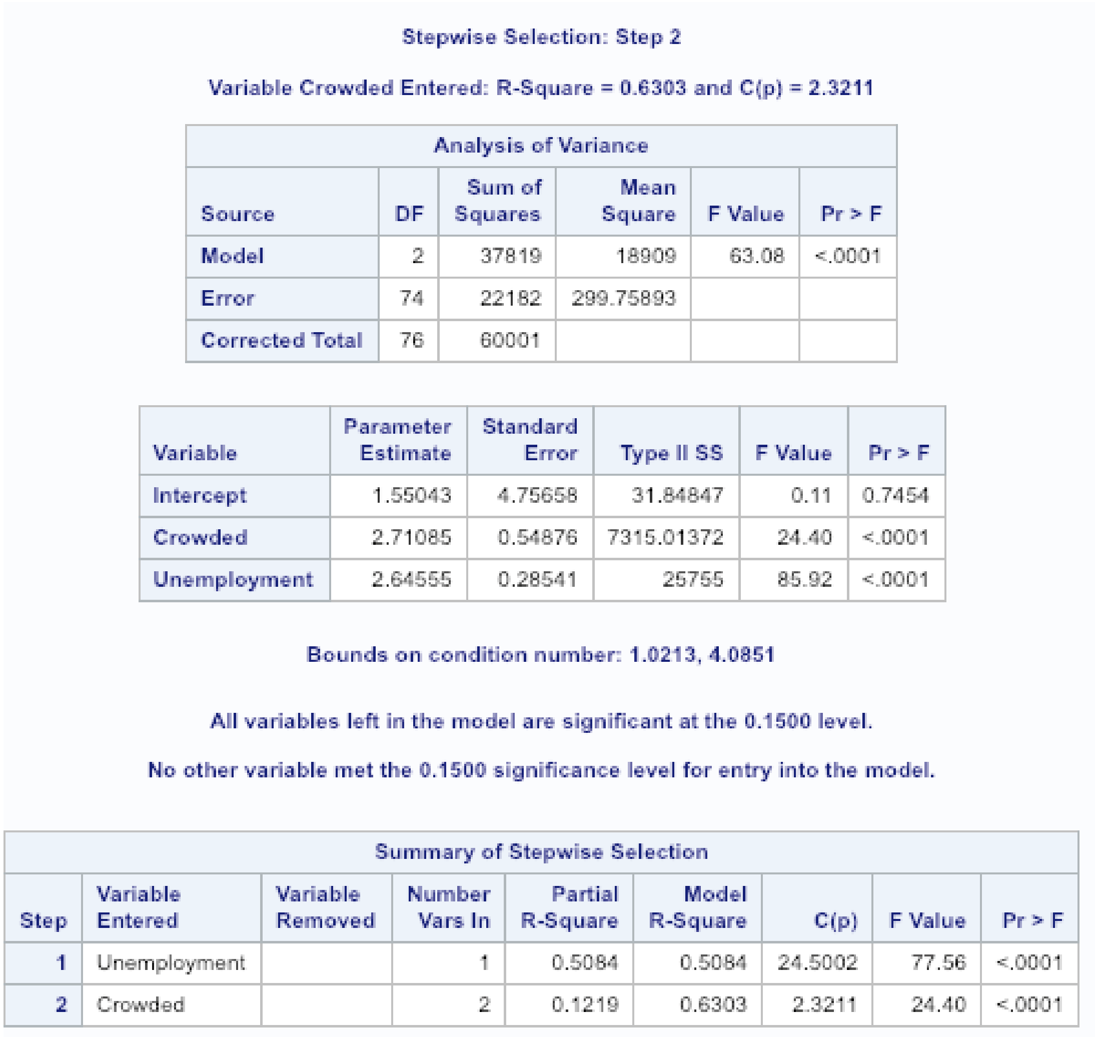


Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13.40971	23.29509	0.58	0.5667	0
BelowPovLev	1	0.30134	0.29124	1.03	0.3043	2.81640
Crowded	1	2.02828	0.73089	2.78	0.0070	1.79506
Dependency	1	0.00822	0.45794	0.02	0.9857	2.78415
Income	1	-0.00024828	0.00026592	-0.93	0.3536	3.97163
Unemployment	1	1.99298	0.54410	3.66	0.0005	3.67748

The variance inflation factor of all variables after removing **NoHSDiploma** are all below 5. No other variable will be removed from the model in this step.

2(J)



Stepwise Selection: Step 2

Variable Crowded Entered: R-Square = 0.6303 and C(p) = 2.3211

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	37819	18909	63.08	<.0001
Error	74	22182	299.75893		
Corrected Total	76	60001			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.55043	4.75658	31.84847	0.11	0.7454
Crowded	2.71085	0.54876	7315.01372	24.40	<.0001
Unemployment	2.64555	0.28541	25755	85.92	<.0001

Bounds on condition number: 1.0213, 4.0851

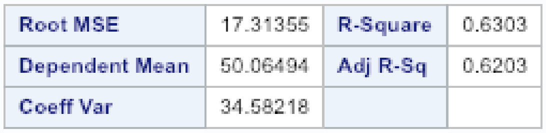
All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Unemployment		1	0.5084	0.5084	24.5002	77.56	<.0001
2	Crowded		2	0.1219	0.6303	2.3211	24.40	<.0001

Using the stepwise selection method, the charts show that the only significant variables are **Unemployment** and **Crowded**.

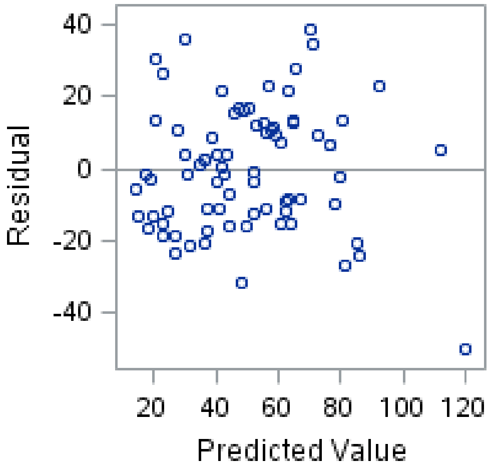
2(K)

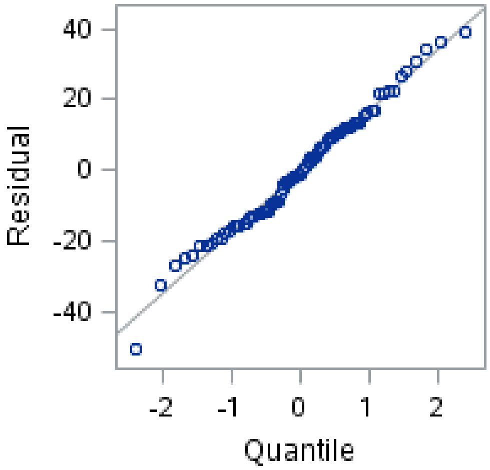


Root MSE	17.31355	R-Square	0.6303
Dependent Mean	50.06494	Adj R-Sq	0.6203
Coeff Var	34.58218		

The R-Square of the suggested best model in step 2(j) is 0.6303 which means that **63.03%** of the variation of teen birthrates(**BirthRate** variable) can be explained by the percent of people not in labor force aged 16 years and older(**Unemployment** variable), and the percent of occupied housing units(**Crowded** variable).

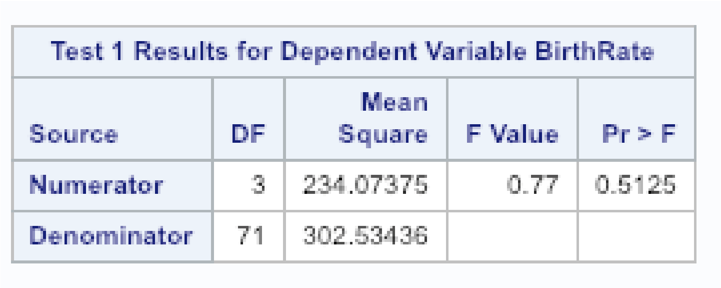
2(L)





The graph of residual vs predicted values shows that the assumption of constant variance is met since the points are scattered randomly around 0 with what appears to be an outlier that has predicted value above 120. The residual vs quantile plot graph display has improved from the model from 2(a), showing very light tails. This means the best model meets the normality assumption.

2(M)

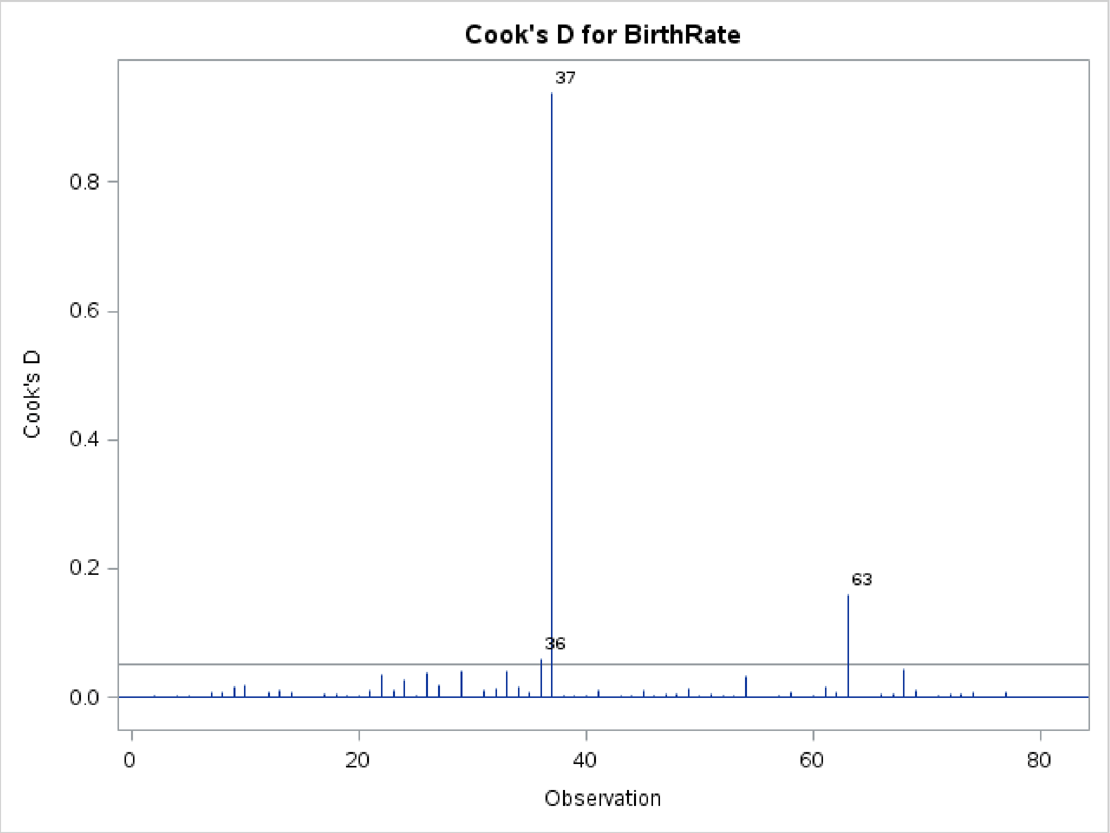


Test 1 Results for Dependent Variable BirthRate

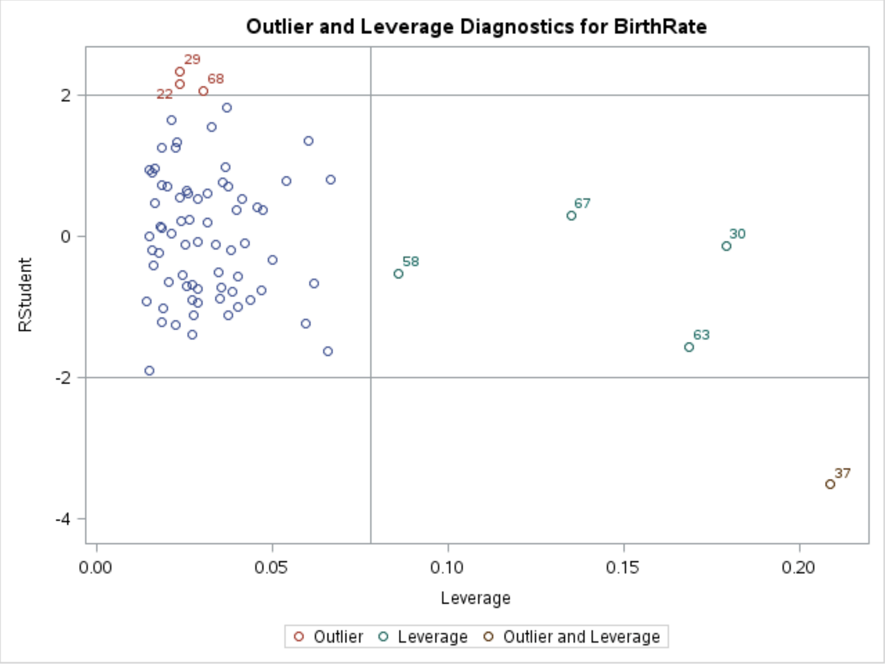
Source	DF	Mean Square	F Value	Pr > F
Numerator	3	234.07375	0.77	0.5125
Denominator	71	302.53436		

The nested f-test has a p-value of 0.5125 meaning the variables **BelowPovLevel**, **Dependency** and **Income** of the reduced model are not significant at a 5% significance level when predicting **BirthRate**.

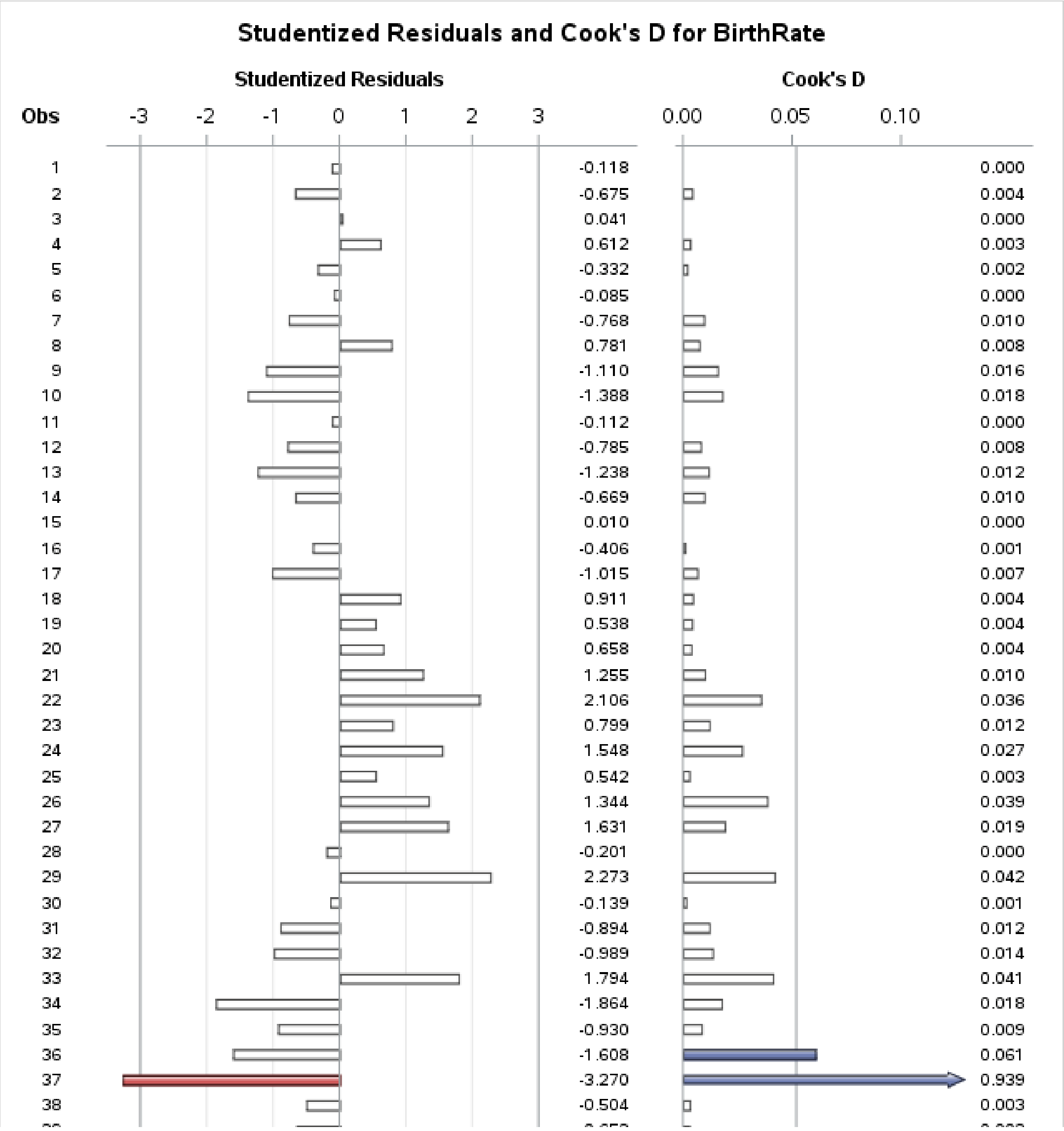
2N)



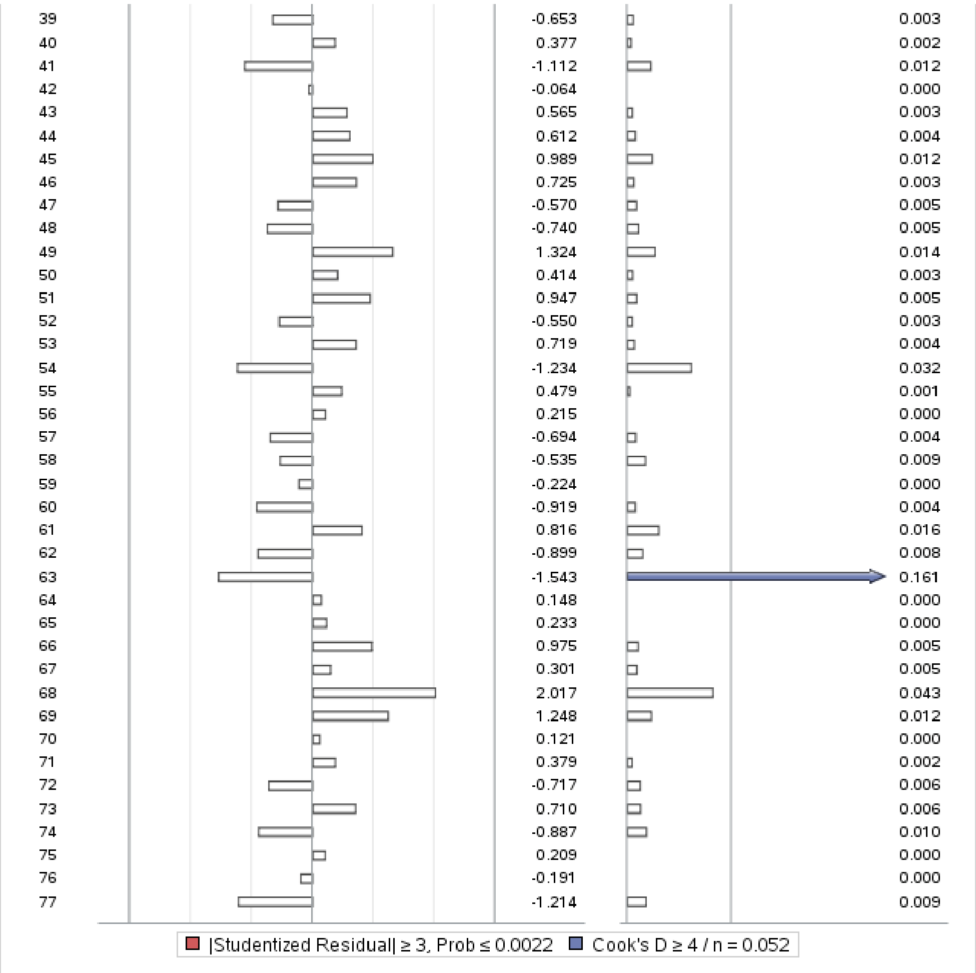
Cook's D for BirthRate
Cook's D
0.0
0.2
0.4
0.6
0.8
Observation
0
20
40
60
80
37
63
36



Outlier and Leverage Diagnostics for BirthRate
RStudent
2
0
-2
-4
Leverage
0.00
0.05
0.10
0.15
0.20
29
68
22
67
30
58
63
37
○ Outlier
○ Leverage
○ Outlier and Leverage



Studentized Residuals and Cook's D for BirthRate

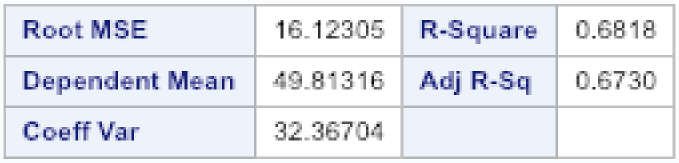


39	-0.653	0.003
40	0.377	0.002
41	-1.112	0.012
42	-0.064	0.000
43	0.565	0.003
44	0.612	0.004
45	0.989	0.012
46	0.725	0.003
47	-0.570	0.005
48	-0.740	0.005
49	1.324	0.014
50	0.414	0.003
51	0.947	0.005
52	-0.550	0.003
53	0.719	0.004
54	-1.234	0.032
55	0.479	0.001
56	0.215	0.000
57	-0.694	0.004
58	-0.535	0.009
59	-0.224	0.000
60	-0.919	0.004
61	0.816	0.016
62	-0.899	0.008
63	-1.543	0.161
64	0.148	0.000
65	0.233	0.000
66	0.975	0.005
67	0.301	0.005
68	2.017	0.043
69	1.248	0.012
70	0.121	0.000
71	0.379	0.002
72	-0.717	0.006
73	0.710	0.006
74	-0.887	0.010
75	0.209	0.000
76	-0.191	0.000
77	-1.214	0.009

🟥 |Studentized Residual| ≥ 3, Prob ≤ 0.0022
🟦 Cook's D ≥ 4 / n = 0.052

By looking at these charts, we can observe that there are 3 observations with higher than normal Cook's distance values(36,37,38) with one been almost 1(37) , 5 with high leverage(58,67,30, 63,37)and one with a very low R-Student value(37). Observations with a large R-Student values(in magnitude) indicate unusual response values, observations with high leverage indicate covariate values that are extreme(far from the center of the distribution) and observations with a high Cook's D indicate values that a high influence on the estimated parameters and the predicted values. I decided to remove observation 37 since it is highly influential, has a high leverage and a very low R-Student Value.

By removing the influential observation, we can see that the R-Square improved by 5%.



Root MSE	16.12305	R-Square	0.6818
Dependent Mean	49.81316	Adj R-Sq	0.6730
Coeff Var	32.36704		

This indicates that now the predictors explain more in the variation of the response.

2(O)

I would suggest using the best model without the influential/outlier observation 37 in order to improve the prediction.

SAS CODE

```
/* Eric Fernandez Case Study */
/* I certify that the SAS code given is my original and exclusive work */
/* Part 1 DATA EXPLORATION */
/* 1(a) */
/* Datastep */
/*
 * To read the file:
 * -Locate the file 'Teen.csv'
 * -Right click on 'Teen.csv' and select 'Properties'
 * -Copy and paste the path name to the FILENAME statement
 */
FILENAME CSV "/home/eff100/datasets/Teen.csv" TERMSTR=CRLF;
PROC IMPORT DATAFILE=CSV
                    OUT=teen
                    DBMS=CSV
                    REPLACE;
RUN;
/* Print first 20 observations */
PROC PRINT data=teen(obs=20);
RUN;

/* 1(b) */
/* Scatter plot for all the variables plotted against each other */
PROC SGSCATTER data=teen;
        matrix BirthRate BelowPovLev Crowded Dependency NoHSDiploma Income Unemployment;
RUN;

/* Output correlation coefficient table */
PROC CORR data=teen;
RUN;

/* Part 2 MODEL FITTING AND ANALYSIS */
/* 2(a), 2(b), 2(c) and 2(d) */
/* Look at ANOVA test and R-square*/
PROC REG data=teen;
        model BirthRate = Unemployment;
RUN;



/* Perform Box-Cox test to obtain lambda for power transformation */
PROC TRANSREG data=teen;
        model Boxcox(BirthRate)=Identity(Unemployment);
RUN;

/* Transform response using the lambda obtained from Box-Cox */
DATA TeenTransformed;
        set teen;
        TransformedBirthRate=BirthRate**0.75;
RUN;
/*
 * Look at R2 and residual vs quantile plot to asses if transformation
 * is beneficial
 */
PROC REG data=TeenTransformed;
```

```
        model TransformedBirthRate = Unemployment;
RUN;

/* 2(e) */
/* Confidence intervals -- 95% default */
PROC REG data=teen;
        model BirthRate = Unemployment/clb; /* Confidence interval for the slope */
RUN;

/* 2(f) */
/* Bootstrap for slope Confidence Interval */
/* Generate 1000 replications with equal probability and with replacement(URS). */
PROC SURVEYSELECT Data=teen out=boot
        seed=4321 samprate=1
        method=urs outhits rep=1000;
RUN;
/* Regression on every sample to find the slopes */
PROC REG data=boot outest=betas noprint;
        model BirthRate= Unemployment;
        by replicate;
RUN;
PROC UNIVARIATE data=betas noprint;
        var Unemployment;
        output out=BootCI pctlpts= 2.5 97.5  pctlpre=Conf_Limit_;
RUN;
/* Print 95% Confidence interval of the bootstrap set */
PROC PRINT data=BootCI;
RUN;

/* 2(g) */
/* Perform ANOVA on the model of 2(a) to verify model's effectiveness */
PROC ANOVA data=teen;
        class Unemployment;
        model BirthRate = Unemployment;
RUN;

/* 2(h) */
/* Create dataset with Unemployment=1.2 */
DATA NewTeen;
Input BirthRate Unemployment;
datalines;
. 1.2
;
RUN;

/* Create new Data set with old values plus the value created. */
DATA Teen2;
        set NewTeen teen; /* Concatenate Data Sets */
RUN;

/* Produce both 95% confidence and prediction intervals around the predicted response for x*.  */
PROC REG data=Teen2;
        model BirthRate = Unemployment/cli clm;
        id unemployment;
RUN;

/* 2(i) */
PROC REG data=Teen;
```

```
        model BirthRate = BelowPovLev Crowded Dependency NoHSDiploma Income Unemployment /VIF; /* VIF checks for
multicolinearity */
RUN;
/* Model with no NoHSDiploma due to high VIF*/
PROC REG data=Teen;
        model BirthRate = BelowPovLev Crowded Dependency Income Unemployment /VIF; /* VIF checks for multicolinearity
*/
RUN;

/* 2(j) */
/* Stepwise method for variable selection. This model does not include NoHSDiploma*/
PROC REG data=Teen;
        model BirthRate = BelowPovLev Crowded Dependency Income Unemployment/selection=stepwise;
RUN;

/* 2(k), 2(l) and 2(n) */
/* Do a regression using only the variables obtained from the stepwise selection method */
PROC REG data=Teen plots(label)=(CooksD RStudentByLeverage);
        model BirthRate = Crowded Unemployment/r influence;
RUN;

data TeenNew ;
 set teen;
 if community = 37 then delete;/* remove obs 37*/
RUN;

/* fit model with new dataset */
PROC REG data=TeenNew;
        model BirthRate = Crowded Unemployment/r influence;
RUN;

/* 2(m) */
/*Nested F test*/
PROC REG data=Teen;
        model BirthRate = BelowPovLev Crowded Dependency Income Unemployment;
        test BelowPovLev,Dependency,Income ;
RUN;
```