Test a Multiple Regression Model

(Regression Modeling in Practice Week 3 Assignment)

Expected Activities

- Test a multiple regression model.
- Write a blog entry that summarize in a few sentences
    1) what you found in your multiple regression analysis. Discuss the results for the associations between all of your explanatory variables and your response variable. Make sure to include statistical results (Beta coefficients and p-values) in your summary.
    2) Report whether your results supported your hypothesis for the association between your primary explanatory variable and the response variable.
    3) Discuss whether there was evidence of confounding for the association between your primary explanatory and response variable (Hint: adding additional explanatory variables to your model one at a time will make it easier to identify which of the variables are confounding variables); and
    4) Generate the following regression diagnostic plots:
        - q-q plot
        - standardized residuals for all observations
        - leverage plot
        - Write a few sentences describing what these plots tell you about your regression model in terms of the distribution of the residuals, model fit, influential observations, and outliers.

Note:

1) If your response variable is categorical, you will need to identify a quantitative variable in the data set that you can use as a response variable for this assignment. Variables with response scales with 4-5 values that represent a change in magnitude (eg, "strongly disagree to strongly agree", "never to often") can be considered quantitative for the assignment.

SAS Program

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " ACCESS=readonly;

DATA new;
      SET mydata.gapminder;
      KEEP country urbanrate incomeperperson lifeexpectancy;
      LABEL lifeexpectancy="Life Expectancy";
      LABEL urbanrate="Urbanisation Rate";
      LABEL incomeperperson="Income per Person";

      /* Find the mean of explanatory variable */
PROC MEANS;
      VAR urbanrate incomeperperson;
```

```
      /* For quantitative explanatory variable, center it so that
      the mean = 0 (or really close to 0) by subtracting the mean */
DATA new2;
      SET new;
      urbanrate_c=urbanrate - 56.7693596;
      incomeperperson_c=incomeperperson - 8740.97;
      LABEL urbanrate_c="Centered Urbanisation Rate";
      LABEL incomeperperson_c="Centered Income per Person";

      /* Calculate the mean to check centering */
PROC MEANS;
      VAR urbanrate_c incomeperperson_c;

      /* Multiple regression model for the association between two
      explanatory variables and a response variable */
PROC GLM;
      MODEL lifeexpectancy=urbanrate_c incomeperperson_c/SOLUTION;

      /* Generate regression diagnostic plots */
PROC GLM PLOTS(UNPACK)=ALL;
      MODEL lifeexpectancy=urbanrate_c incomeperperson_c/SOLUTION CLPARM;
      OUTPUT RESIDUAL=RES STUDENT=stdres OUT=results;

      /* Standardized residuals for observations */
PROC GPLOT;
      LABEL stdres="Standardized Residual";
      LABEL country="Country";
      PLOT stdres*country/VREF=0;

RUN;
```

Output

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
urbanrate	Urbanisation Rate	203	56.7693596	23.8449326	10.4000000	100.0000000
incomeperperson	Income per Person	190	8740.97	14262.81	103.7758572	105147.44

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
urbanrate_c	Centered Urbanisation Rate	203	5.9113279E-9	23.8449326	-46.3693596	43.2306404
incomeperperson_c	Centered Income per Person	190	-0.0039237	14262.81	-8637.19	96406.47

The GLM Procedure

Number of Observations Read	213
Number of Observations Used	176

The GLM Procedure

Dependent Variable: lifeexpectancy Life Expectancy

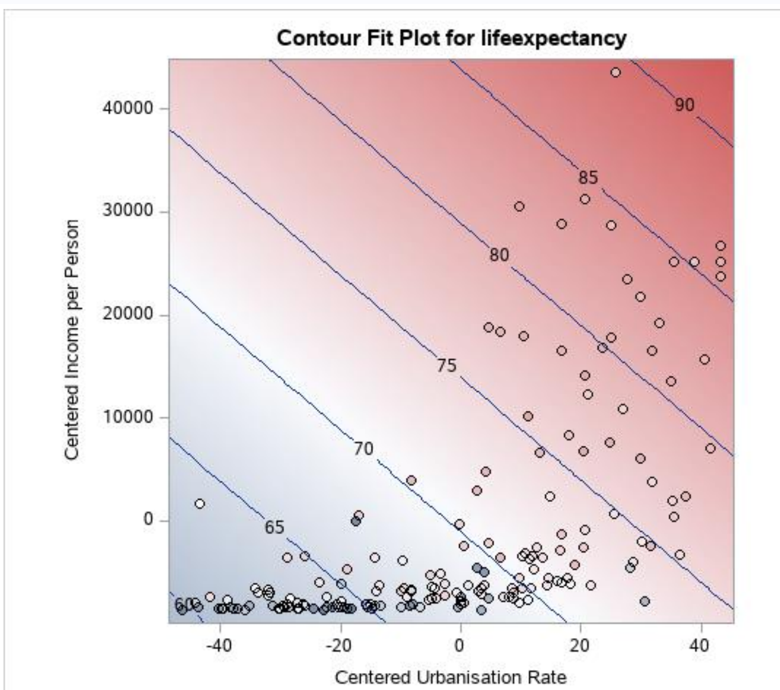
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7624.56072	3812.28036	73.76	<.0001
Error	173	8941.56557	51.68535		
Corrected Total	175	16566.12629			

R-Square	Coeff Var	Root MSE	lifeexpectancy Mean
0.460250	10.32127	7.189252	69.65473

Source	DF	Type I SS	Mean Square	F Value	Pr > F
urbanrate_c	1	6259.858023	6259.858023	121.11	<.0001
incomeperperson_c	1	1364.702692	1364.702692	26.40	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
urbanrate_c	1	1630.573301	1630.573301	31.55	<.0001
incomeperperson_c	1	1364.702692	1364.702692	26.40	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	70.32419950	0.54713692	128.53	<.0001
urbanrate_c	0.16549598	0.02946463	5.62	<.0001
incomeperperson_c	0.00033277	0.00006476	5.14	<.0001



The GLM Procedure

Number of Observations Read	213
Number of Observations Used	176

The GLM Procedure

Dependent Variable: lifeexpectancy Life Expectancy

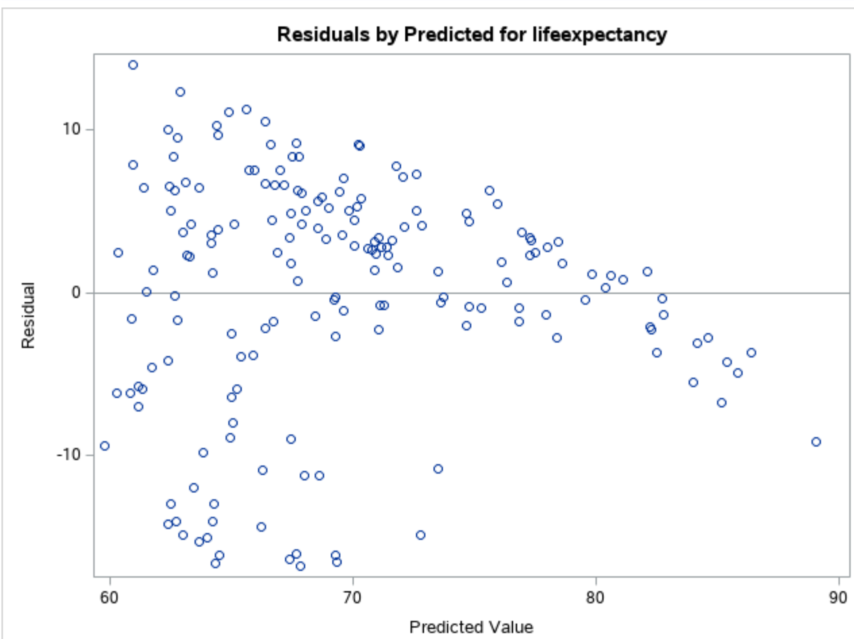
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7624.56072	3812.28036	73.76	<.0001
Error	173	8941.56557	51.68535		
Corrected Total	175	16566.12629			

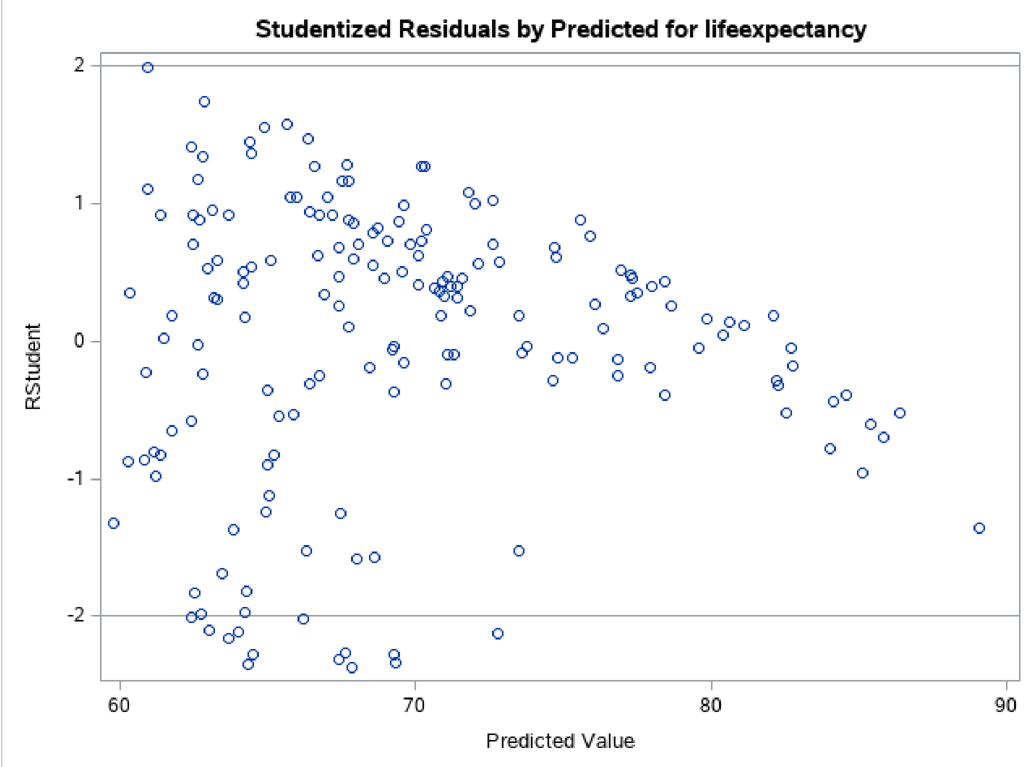
R-Square	Coeff Var	Root MSE	lifeexpectancy Mean
0.460250	10.32127	7.189252	69.65473

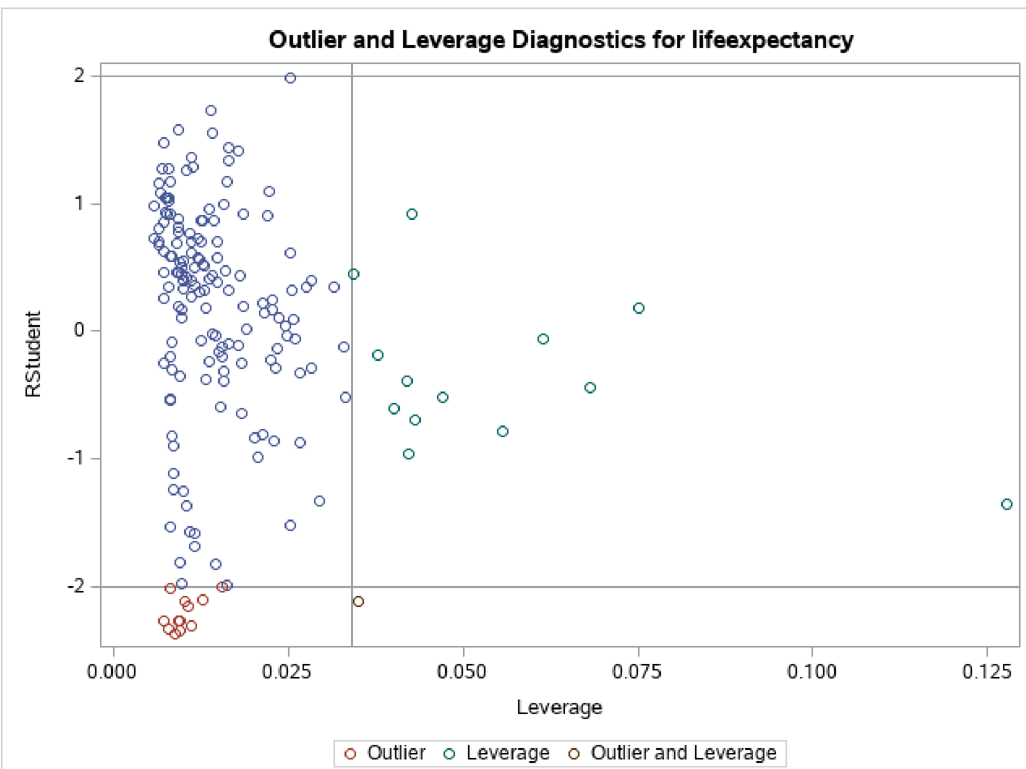
Source	DF	Type I SS	Mean Square	F Value	Pr > F
urbanrate_c	1	6259.858023	6259.858023	121.11	<.0001
incomeperperson_c	1	1364.702692	1364.702692	26.40	<.0001

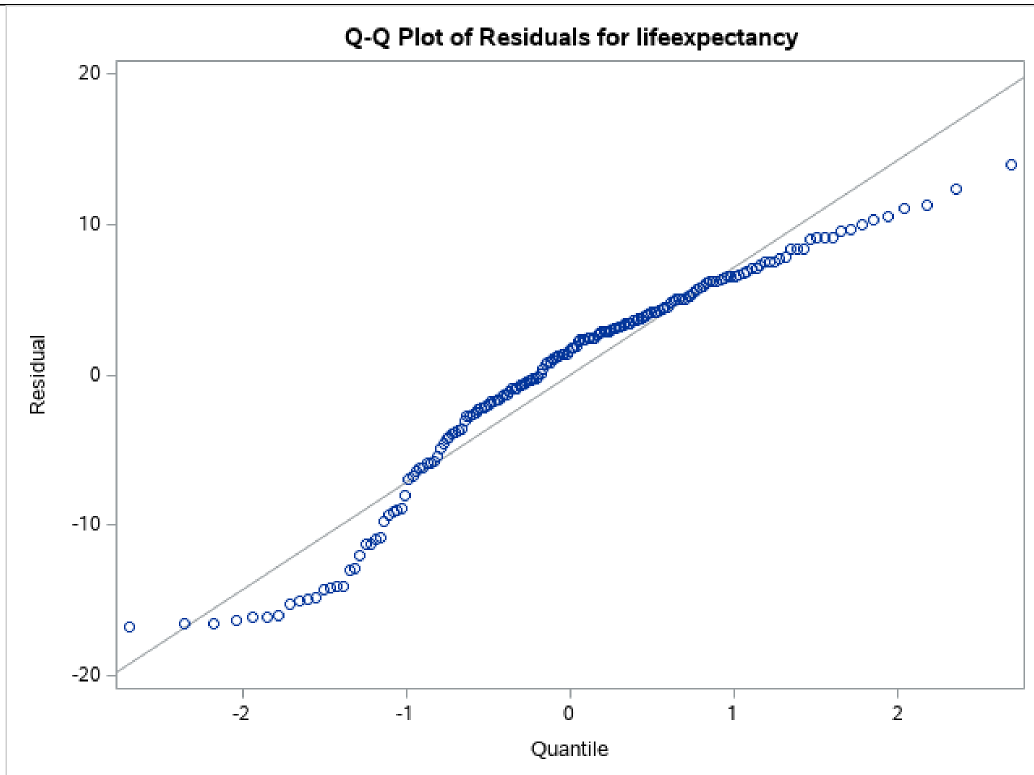
Source	DF	Type III SS	Mean Square	F Value	Pr > F
urbanrate_c	1	1630.573301	1630.573301	31.55	<.0001
incomeperperson_c	1	1364.702692	1364.702692	26.40	<.0001

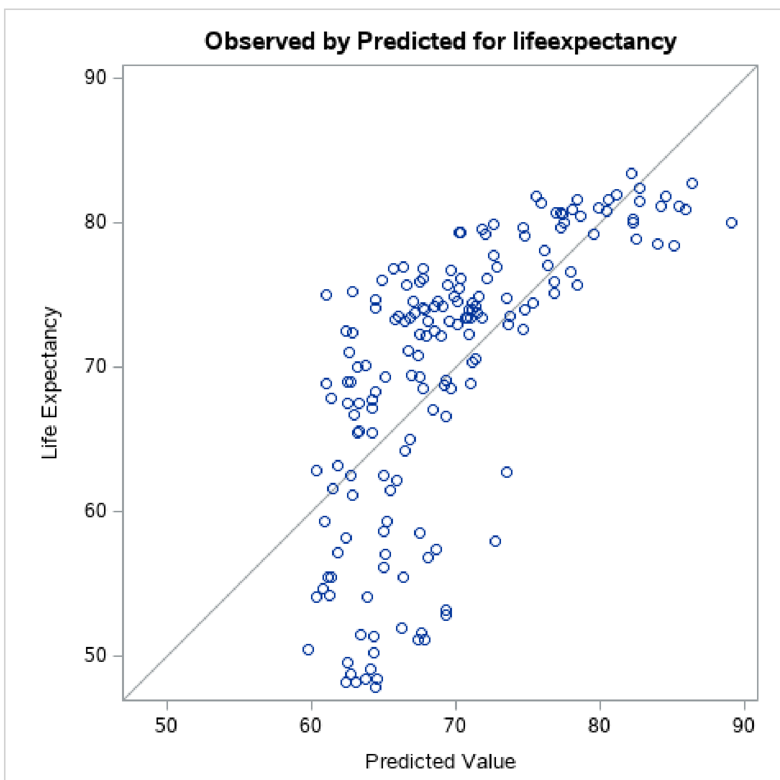
Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	70.32419950	0.54713692	128.53	<.0001	69.24427632	71.40412267
urbanrate_c	0.16549598	0.02946463	5.62	<.0001	0.10733953	0.22365242
incomeperperson_c	0.00033277	0.00006476	5.14	<.0001	0.00020495	0.00046059

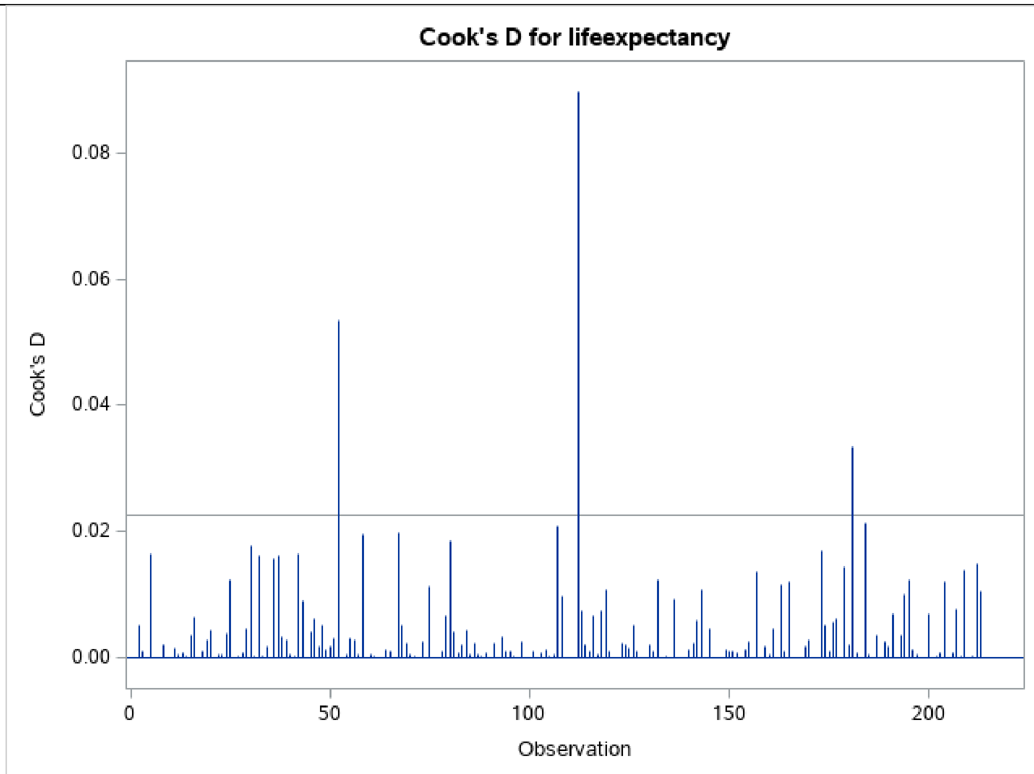


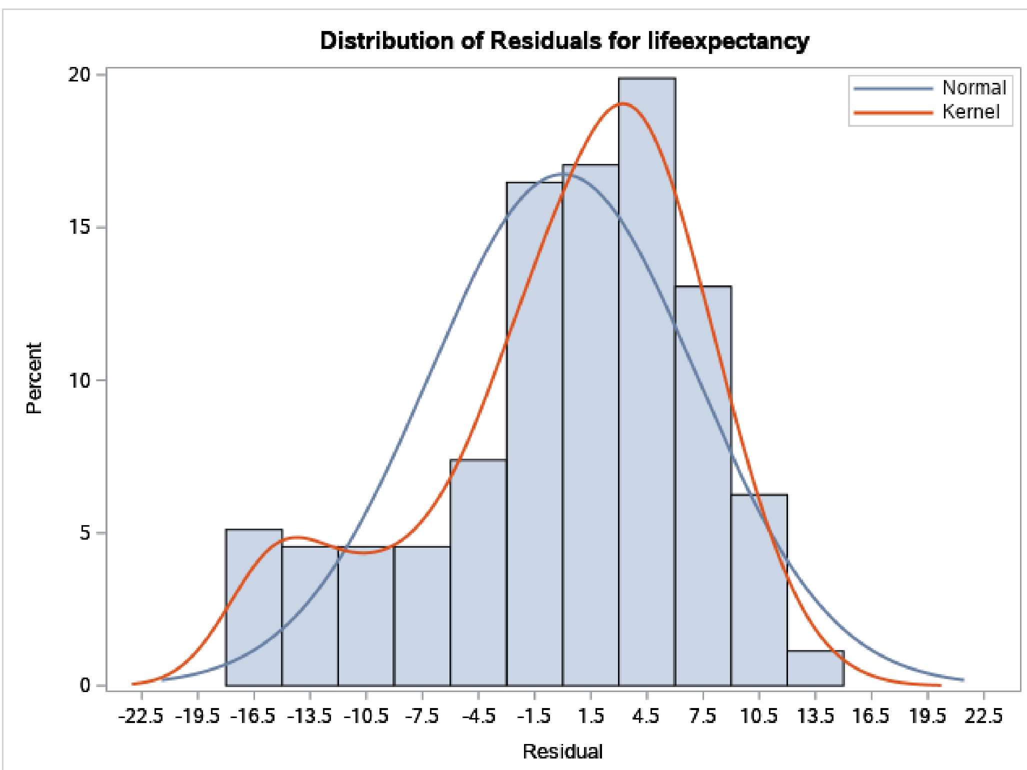

Studentized Residuals by Predicted for lifeexpectancy
RStudent
Predicted Value
2
1
0
-1
-2
60
70
80
90

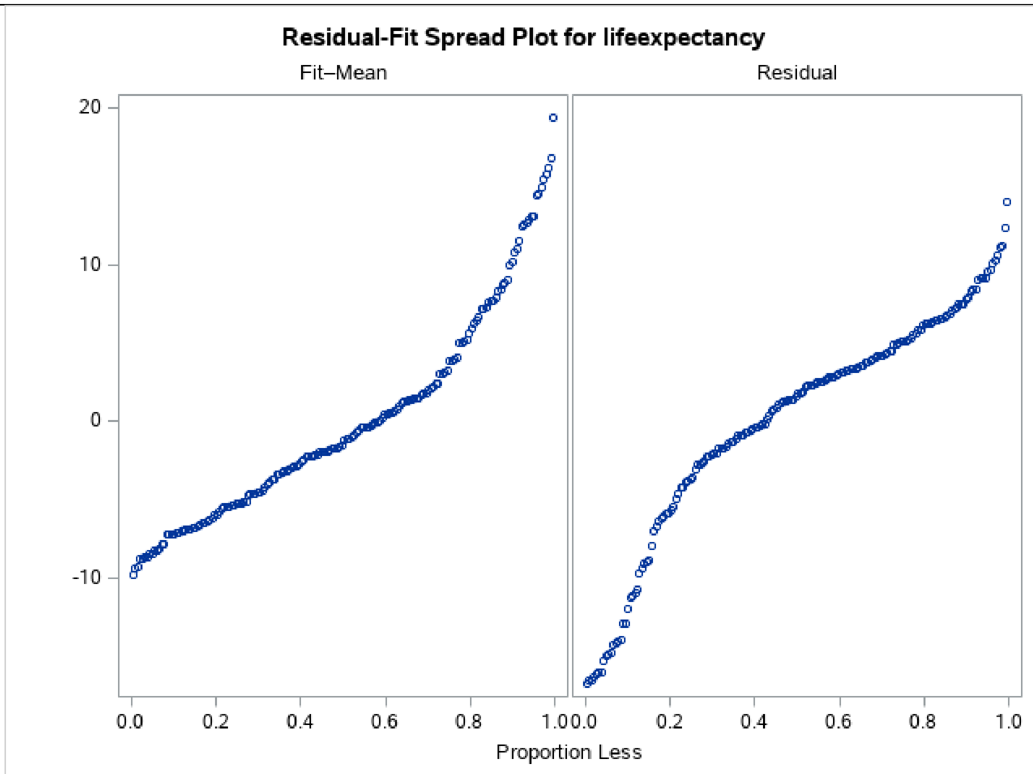

Outlier and Leverage Diagnostics for lifeexpectancy
RStudent
Leverage
2
1
0
-1
-2
0.000
0.025
0.050
0.075
0.100
0.125
Outlier
Leverage
Outlier and Leverage

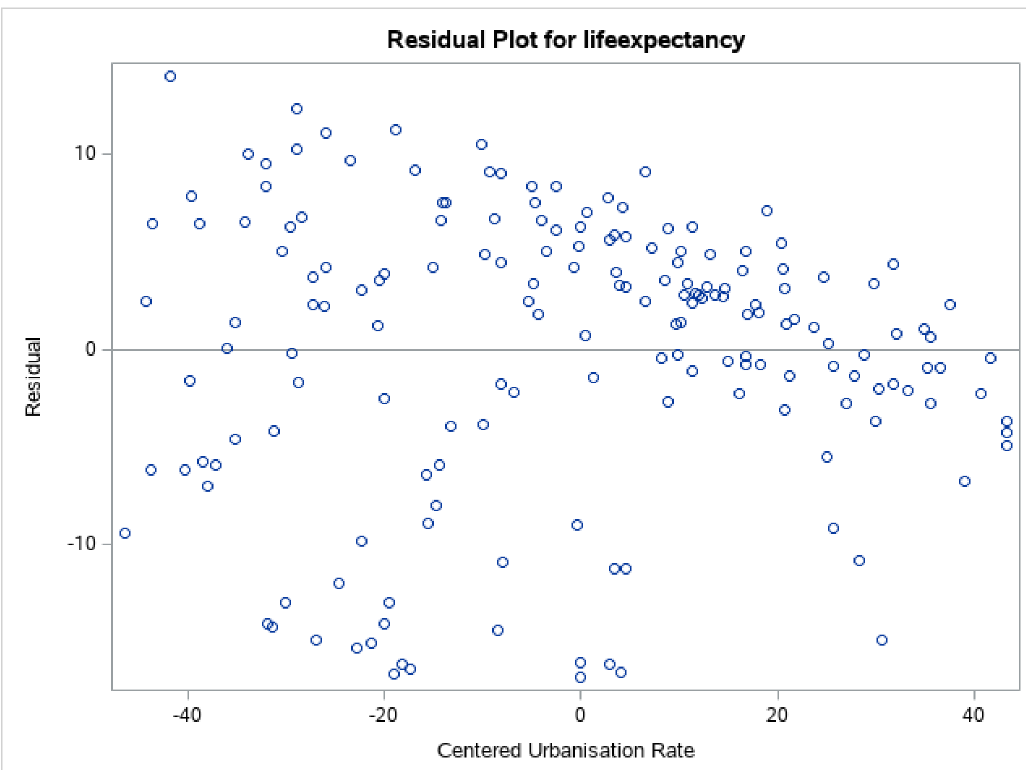

Q-Q Plot of Residuals for lifeexpectancy
Residual
Quantile
20
10
0
-10
-20
-2
-1
0
1
2

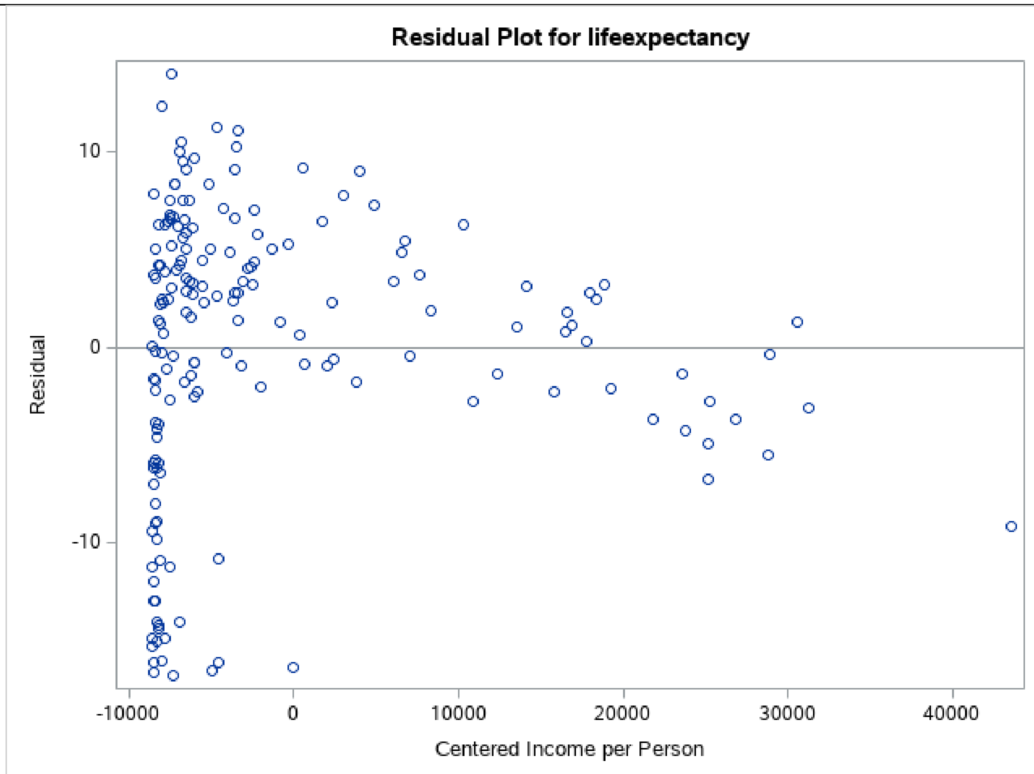

Observed by Predicted for lifeexpectancy
Life Expectancy
Predicted Value
90
80
70
60
50
50
60
70
80
90

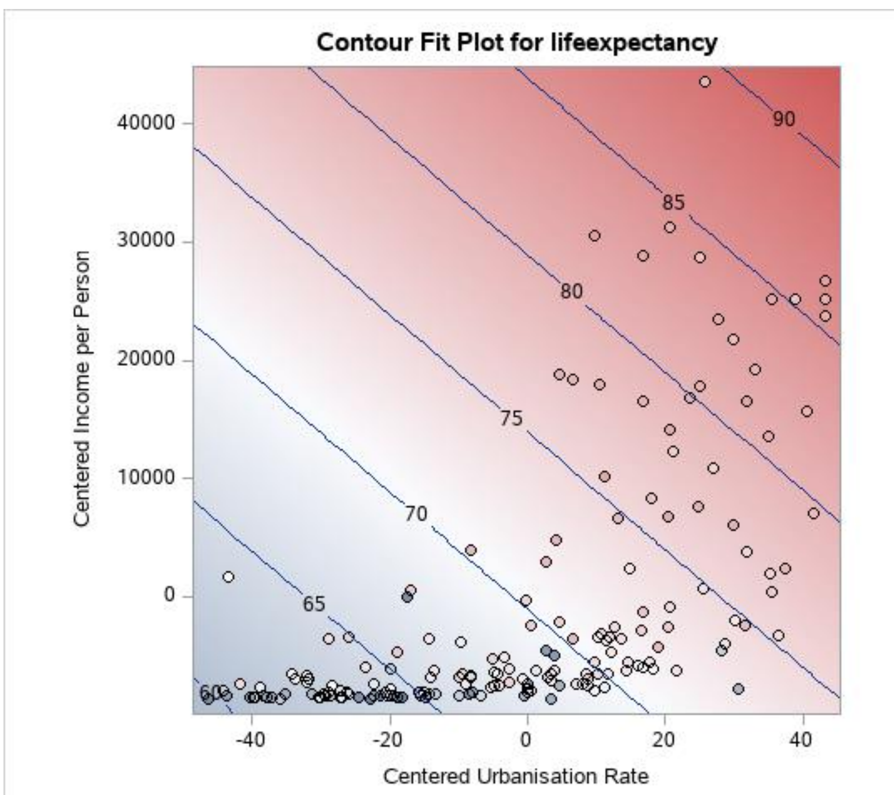

Cook's D for lifeexpectancy
Cook's D
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
Observation

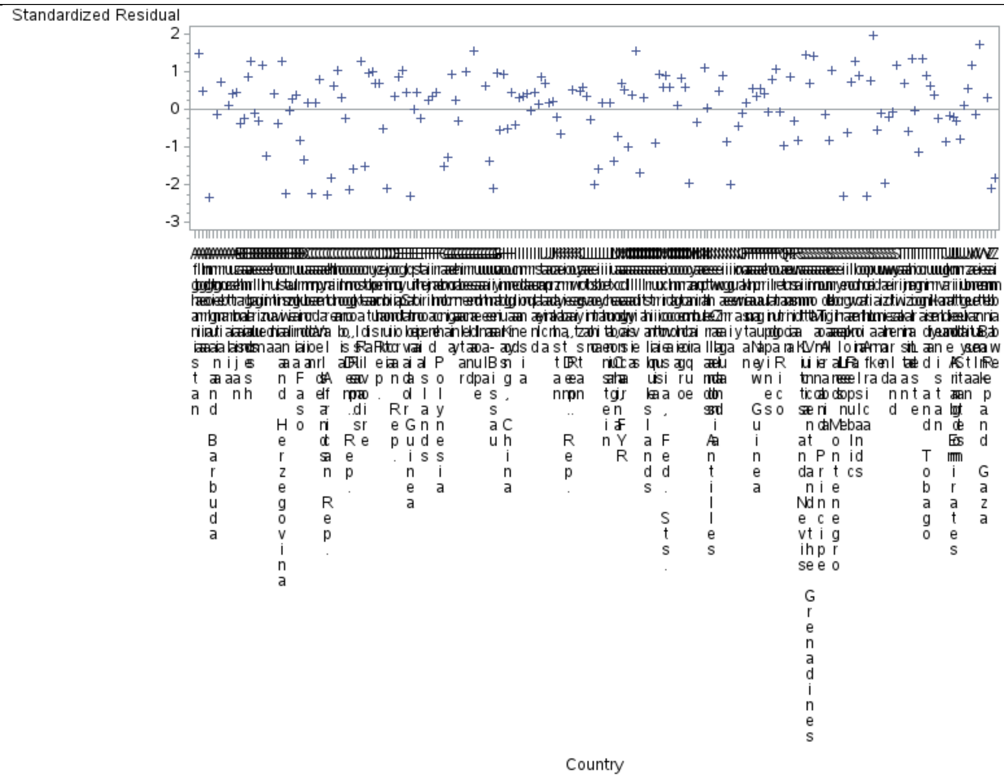

Distribution of Residuals for lifeexpectancy
Normal
Kernel
Percent
20
15
10
5
0
-22.5 -19.5 -16.5 -13.5 -10.5 -7.5 -4.5 -1.5 1.5 4.5 7.5 10.5 13.5 16.5 19.5 22.5
Residual


Residual-Fit Spread Plot for lifeexpectancy
Fit–Mean
Residual
20
10
0
-10
0.0
0.2
0.4
0.6
0.8
1.0
Proportion Less


Residual Plot for lifeexpectancy
10
0
-10
Residual
-40
-20
0
20
40
Centered Urbanisation Rate


Residual Plot for lifeexpectancy
Residual
10
0
-10
-10000
0
10000
20000
30000
40000
Centered Income per Person


Contour Fit Plot for lifeexpectancy
Centered Income per Person
40000
30000
20000
10000
0
90
85
80
75
70
65
-40
-20
0
20
40
Centered Urbanisation Rate



Variables Used:

- Urbanisation Rate
- Life Expectancy
- Income per Person

All variables used in the analysis are quantitative.

As the first step, I centered the explanatory variables and checked the coding by using the means procedure. The data analysis done for previous assignment revealed a strong positive correlation between Life Expectancy (response variable) and Urbanisation Rate (explanatory variable). In this assignment, one more explanatory variable (Income per Person) is included to run a multiple linear regression.

Hypothesis:

There is a significant association between two explanatory variables and one response variable.

Summary:

After adding the second explanatory variable, the correlation between Life Expectancy (response variable) and Urbanisation Rate (initial explanatory variable) remained significantly and positively associated (b= 0.16549598, $p < 0.0001$). Also, it appeared that that there is a slight association between Income per Person of the country (second explanatory variable) and the Life Expectation of its citizens (b= 0.00033277, $p < 0.0001$). This would suggest that this explanatory variable is not confounding the results.

Results obtained support my hypothesis. The assumption that both explanatory variables are significantly correlated with the response variable proved to be correct.

Using a second explanatory variable slightly increases the R-squared value of the model. The R-square value of 0.460250 indicates that the proportion of variance in the response variable that can be attributed to the explanatory variable is 46%.

Q-Q Plot

The Q-Q Plot shows that the residuals generally follow a straight line, but deviate somewhat at lower and highest quantiles, i.e. the residuals do not follow perfect normal distribution.

Standard Residuals

This procedure shows that almost the same number of countries have standard residuals grater and lower than 0. Only one of them are greater than or equal to 2 and only a few are lower than -2, making this model acceptable.

Outliers and Leverage

The Outlier and Leverage Diagnostics plot shows that the majority of the points have close to zero leverage and are within a residual standardized value of 2. That is, the majority of the observations have no leverage on the model. However, there are a few observations that are outliers (red) and another small set of observations have high leverage (green). There is only one point which is both an outlier and have high leverage.