Test a Basic Linear Regression Model

(Regression Modeling in Practice Week 2 Assignment)

Expected Activities

- Test a basic linear regression model for the association between the primary explanatory variable and a response variable, and to create a blog entry describing the results.
- Create a blog entry where 1) post the program and output, and 2) post a frequency table for (recoded) categorical explanatory variable or report the mean for centered explanatory variable. 3) Write a few sentences describing the results of the linear regression analysis.

Note:

1) If you have a categorical explanatory variable, make sure one of your categories is coded "0" and generate a frequency table for this variable to check your coding. If you have a quantitative explanatory variable, center it so that the mean = 0 (or really close to 0) by subtracting the mean, and then calculate the mean to check your centering.
2) Test a linear regression model and summarize the results in a couple of sentences. Make sure to include statistical results (regression coefficients and p-values) in your summary.

SAS Program

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

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

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

/* 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;
      LABEL urbanrate_c="Centered Urbanisation Rate";

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

```
/* Linear regression model for the association between primary
explanatory variable and a response variable */
PROC GLM;
MODEL lifeexpectancy=urbanrate_c/SOLUTION;

RUN;
```

Output

The MEANS Procedure

Analysis Variable : urbanrate Urbanisation Rate				
N	Mean	Std Dev	Minimum	Maximum
203	56.7693596	23.8449326	10.4000000	100.0000000

The MEANS Procedure

Analysis Variable : urbanrate_c Centered Urbanisation Rate				
N	Mean	Std Dev	Minimum	Maximum
203	5.9113279E-9	23.8449326	-46.3693596	43.2306404

The GLM Procedure

Number of Observations Read	213
Number of Observations Used	188

The GLM Procedure

Dependent Variable: lifeexpectancy Life Expectancy

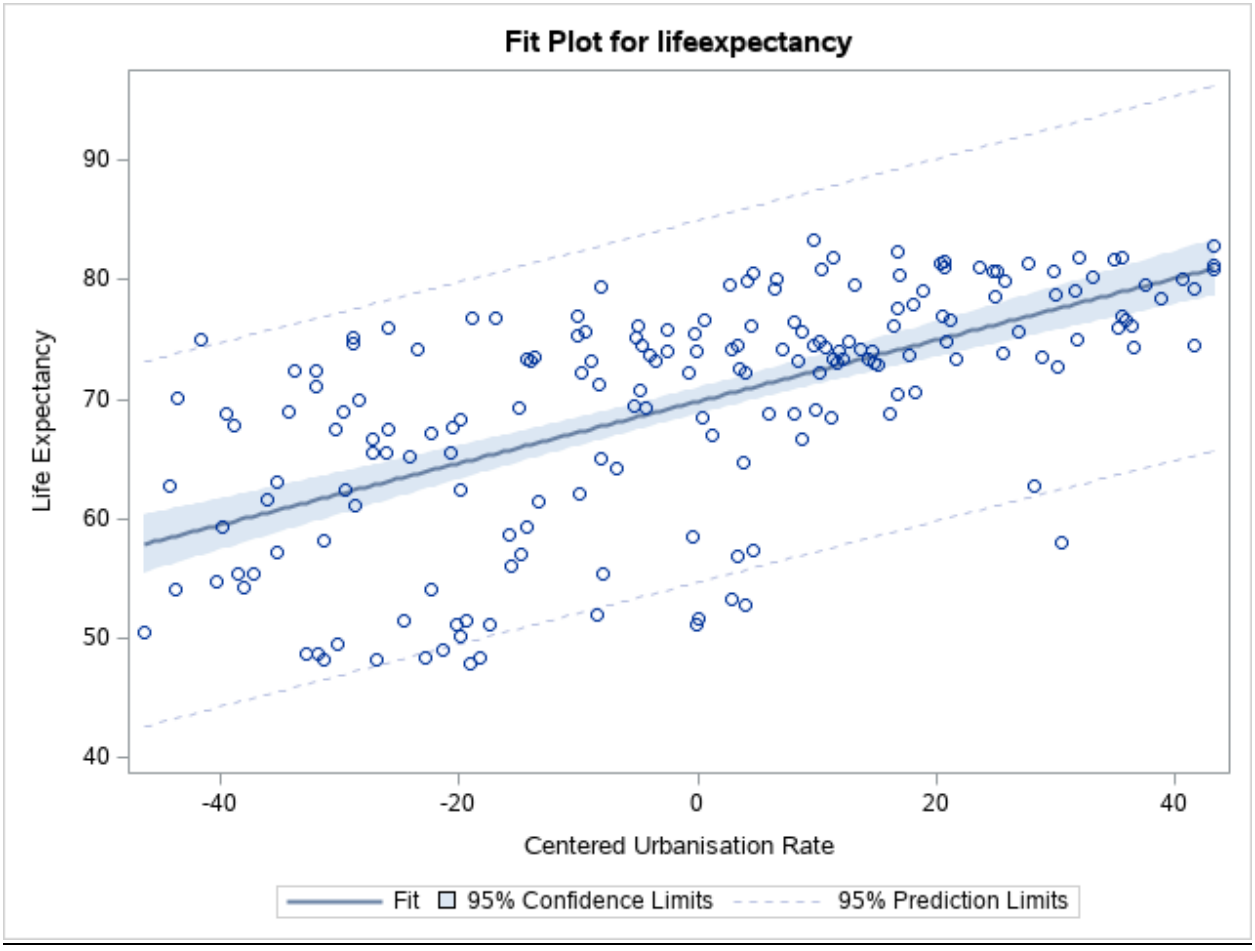
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6746.79068	6746.79068	115.36	<.0001
Error	186	10878.12428	58.48454		
Corrected Total	187	17624.91496			

R-Square	Coeff Var	Root MSE	lifeexpectancy Mean
0.382798	10.98770	7.647518	69.60070

Source	DF	Type I SS	Mean Square	F Value	Pr > F
urbanrate_c	1	6746.790681	6746.790681	115.36	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
urbanrate_c	1	6746.790681	6746.790681	115.36	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	69.81649586	0.55811436	125.09	<.0001
urbanrate_c	0.25794331	0.02401575	10.74	<.0001



After centering the explanatory quantitative variable (Urbanisation Rate) I obtained the new variable (Centered Urbanisation Rate), with the mean equal to zero (0), and I used it in the Linear regression model.

The results of the linear regression model indicate that Life Expectancy ($F = 115.36$, $p<.0001$) is significant and positively associated with the Centered Urbanisation Rate.

The parameter estimates show a coefficient value of 0.25794331 and an intercept value of 69.81649586. Therefore, the best fit line equation for the linear regression is:

Life Expectancy = 0.25794331 * Urbanisation Rate (Centered) + 69.81649586

The p-values for both the intercept and coefficient values are very small (both $p < 0.0001$). This indicates there is indeed a straight-line relationship between Life Expectancy and Urbanisation Rate.

The R-square value of 0.382798 indicates that the proportion of variance in the response variable that can be attributed to the explanatory variable is 38%.