Regression Modeling in Practice

**Assignment – Week 2**

**Test a Basic Linear Regression Model**

**By** [mapolarbear@gmail.com](mailto:mapolarbear@gmail.com)

This week's assignment is to test a basic linear regression model for the association between your primary explanatory variable and a response variable, how strong or weak relationship is, and to describe results.

**About My research**

My major interest is in demographics combined in **GAPMINDER** dataset [www.gapminder.org](http://www.gapminder.org) project.

My main research/study question is about **factors affecting female and male suicide rates in different countries** and possible affect, association and correlation of variables like: **employment**, breast cancer, urban rates, geographical region, and if there is a difference between female and male statistics.

**This assignment will test basic linear regression model for association of 2004 employment and suicide rates for female or male genders.**

**Sample**

The sample is taken from the GAPMINDER dataset with combined various collections of observational data provided by [www.gapminder.org](http://www.gapminder.org) for 213 countries (N=213). The combined GAPMINDER dataset created to study **female and male suicide and employment ratings, with sample data for 2002 and 2004**, for reported countries across the globe.

**Procedure**

Data were collected during 2002 and 2004 by various sources, including the US Census Bureau’s International Database, Institute for Health Metrics and Evaluation, United Nations Statistics Division, International Labor Organization and the World Bank. Each country presented with corresponding rating, including breakdown by female and male gender.

**Measures (current study)**

**empf2004mc** (**centered** quantitative explanatory Variable) - Female 2004 employment rates (% of total population) of employed female for age 15+ that has been employed during the given year collected by International Labor Organization.

**empf2004mc** (**centered** quantities explanatory Variable) - Male 2004 employment rates (% of total population) of employed male for age 15+ that has been employed during the given year collected by International Labor Organization.

**fsuicides2004** (quantitative response variable) - Female 2004 suicides ratings per 100,000 standard population, age adjusted, collected by WHO Global Bureau of diseases and Valance and Injury Prevention.

**msuicides2004** (quantitative response variable) - Male 2004 suicides ratings per 100,000 standard population, age adjusted, collected by WHO Global Bureau of diseases and Valance and Injury Prevention.

**Centering quantitative explanatory variables**

The main purpose to center quantitative variable is to calculate (center) variable’s value around variable mean value to create more precise linear recreation model. Centering a variable means that a constant has been subtracted from every value of a variable. In the code below, four new variables are created:

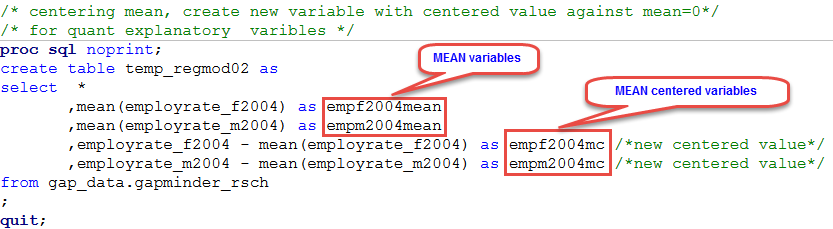
**empf2004mean** is the mean of **employrate\_f2004**

**Empm2004mean** is the mean of **employrate\_m2004**

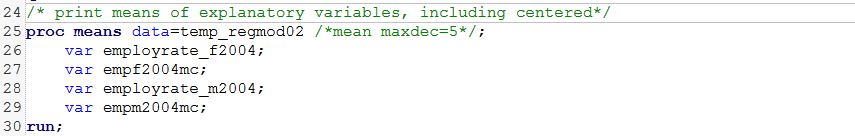
**Empf2004mc**  is the mean centered variable for **employrate\_f2004**

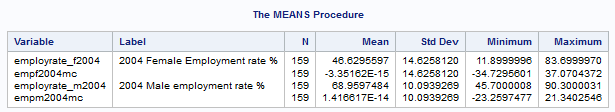
**Empm2004mc** is the mean centered variable for **employrate\_m2004**

**Program code to create centered mean centered variables:**



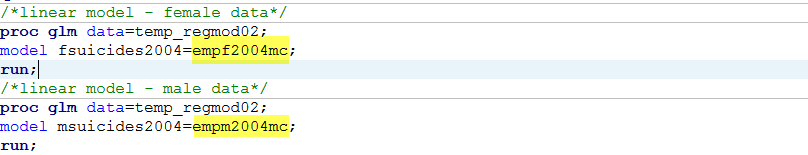
**Display/print means:**



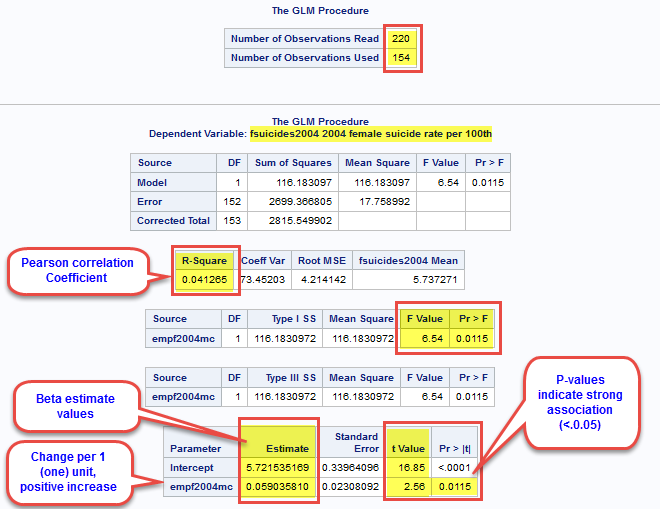


Based on **observed results of MEANS procedure**, we can conclude that mean values of centered variables **empf2004mc** and **empm2004mc** **are very close to center mean=0** and minimum and maximum values **are in range of -34.7 to 37.1 and -23.26 to 21.34 respectively**.

**Basic Linear Rogation model – Program code**



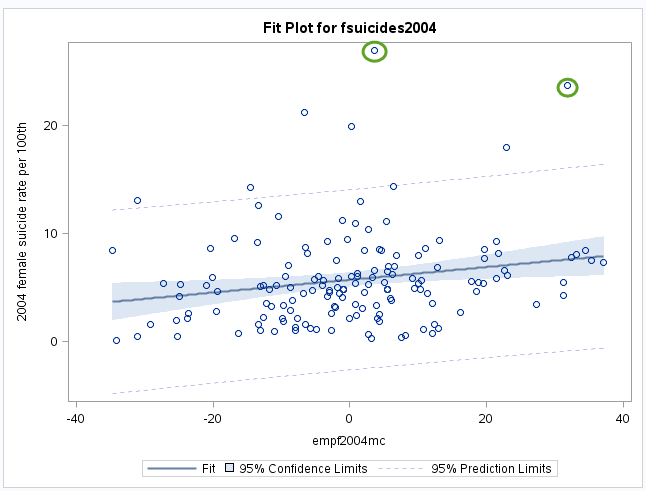
**Output statistics - Female:**



Based on created statistics, **154 observations participated in test**, **R-Square value** (Person Correlation Coefficient) of **0.041 is not far from 0 and does not indicate strong linear relationship**, at the same time **F Value of 6.54 is moderate, P-value of 0.0115 indicates strong relationship** (below 0.05 alpha value) and we can **reject NULL Hypothesis.** **Beta 0** estimate value of intercept is **5.72. Beta 1** estimate value of **empf2004mc indicate positive change or 0.06 per one unit of change**, meaning slight positive incline/slope of regression model line.

Bata values **Beta0=5.72, Beta1=0.06** (fsuicides2004=5.72 + 0.06\*empf2004mc)

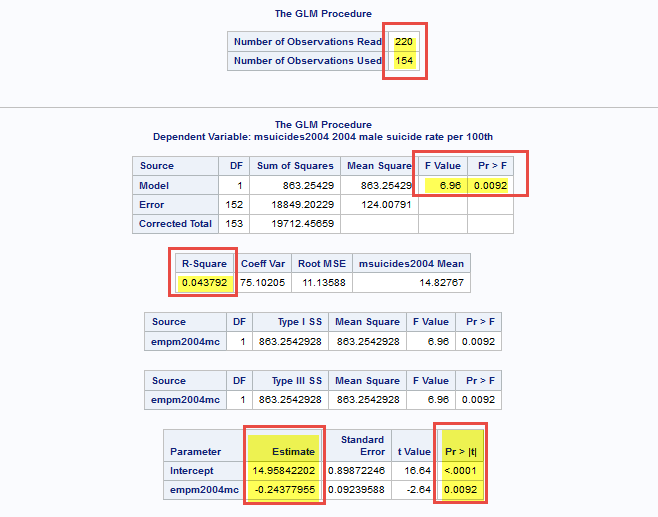
**Generated plot:**



Based on visual analysis of created plot there are couple outliners, with large values and cannot be excluded, because they represent observed values collected from specific countries.

**Conclusion:** Basic Linear Regression test model for **fsuicides2004** response variable and centered **empf2004mc** explanatory variable corresponds to our previous observations and shows strong association between tested variables

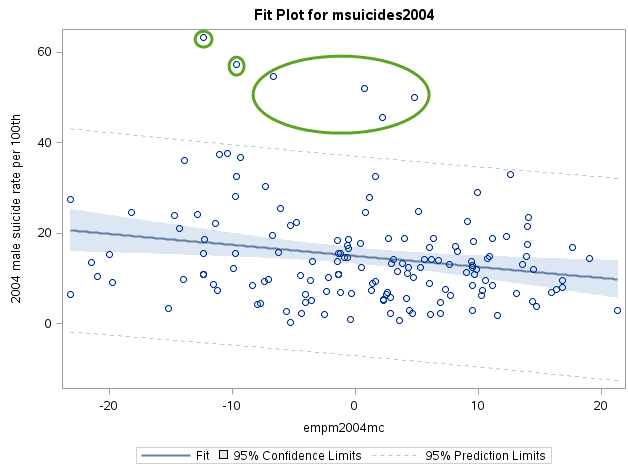
**Output Statistics – Male:**



Based on created statistics, **154 observations participated in test**, **R-Square value** (Person Correlation Coefficient) of **0.043 is not far from 0 and does not indicate strong linear relationship**, at the same time **F Value of 6.96 is moderate, P-value of 0.0092 indicates strong relationship** (below 0.05 alpha value) and we can **reject NULL Hypothesis.** **Beta 0** estimate value of intercept is **14.96. Beta 1** estimate value of **empm2004mc indicate negative change of -0.24 per one unit of change**, meaning negative decline/slope of regression model line.

Bata values **Beta0=14.96, Beta1=-0.24** (msuicides2004=14.96 + (-0.24)\*empf2004mc)

**Generated plot:**

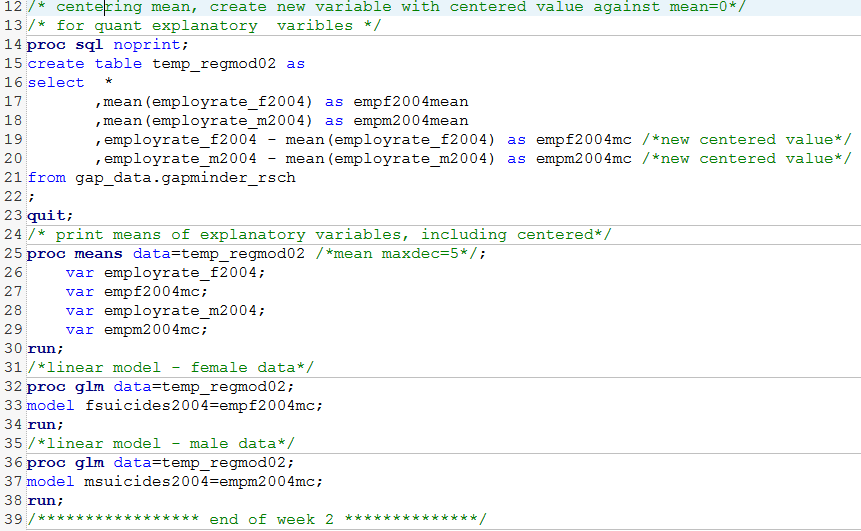


Based on visual analysis of created plot there are couple outliners, with large values and cannot be excluded, because they represent observed values collected from specific countries.

**Conclusion:** Basic Linear Regression test model for **msuicides2004** response variable and centered **empm2004mc** explanatory variable corresponds to our previous observations and shows strong association between tested variables

**Appendix 1.**

Full program code



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* End of week 2 assignment \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

http://mapolarbear-da.blogspot.com/2016/11/coursera-w-regmod-assignment-week-2.html