Regression Modeling in Practice

**Assignment – Week 2**

**Test a Multiple Regression Model**

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

This week's assignment is to test a multiple regression model for the association between your primary explanatory variable and a response variable, if there is a relationship between variables primary variables when additional explanatory variables added to model, and to describe results. The other purpose of multiple regression test is to find confounding explanatory variables, and if there are such, how found confounder variable/s affect relationship between primary response and explanatory variables, including interpretation of diagnostic plots like q-q plot, standardized residuals for all observations and leverage plot.

**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, alcohol consumption, geographical region, and if there is a difference between female and male statistics.

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

**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 other 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)**

**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.

**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.

**urbanrate\_mc** (**centered** quantities explanatory Variable) – Urbanization rate in the country

**incomeperperson\_mc** (**centered** quantities explanatory Variable) – Income per person in the country

**hivrate\_mc** (**centered** quantities explanatory Variable) – HIV rate per 100,000 standard population in the country

**alcconsumption\_mc** (**centered** quantities explanatory Variable) – Alcohol consumption 100,000 standard population rate in the country

**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**

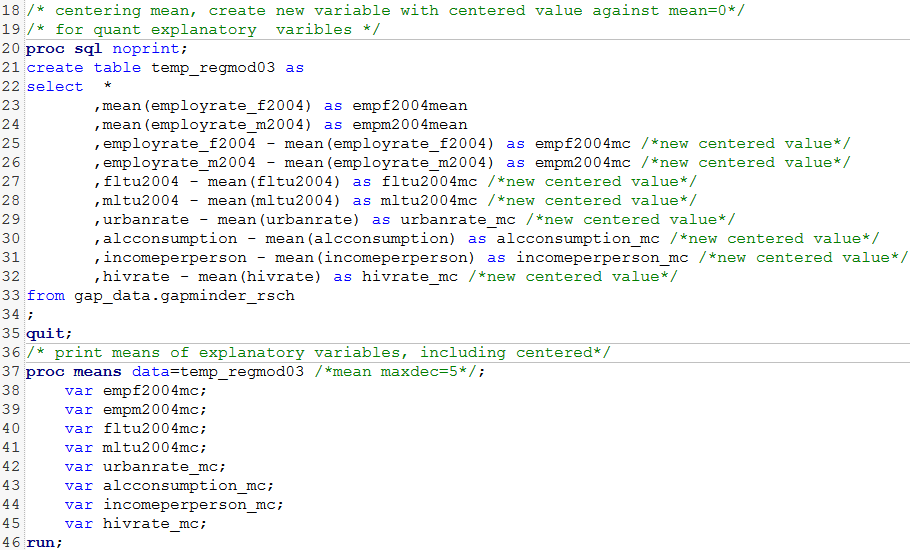
**urbanrate\_mc** is the mean centered variable for **urbanrate**

**incomeperperson\_mc** is the mean centered variable for **incomeperperson**

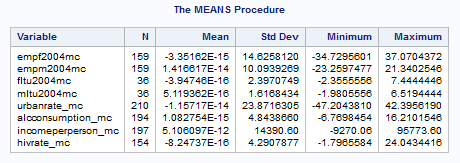
**hivrate\_mc** is the mean centered variable for **hivrate**

**alcconsumption\_mc** is the mean centered variable for **alcconsumption**

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



**Display/print means:**



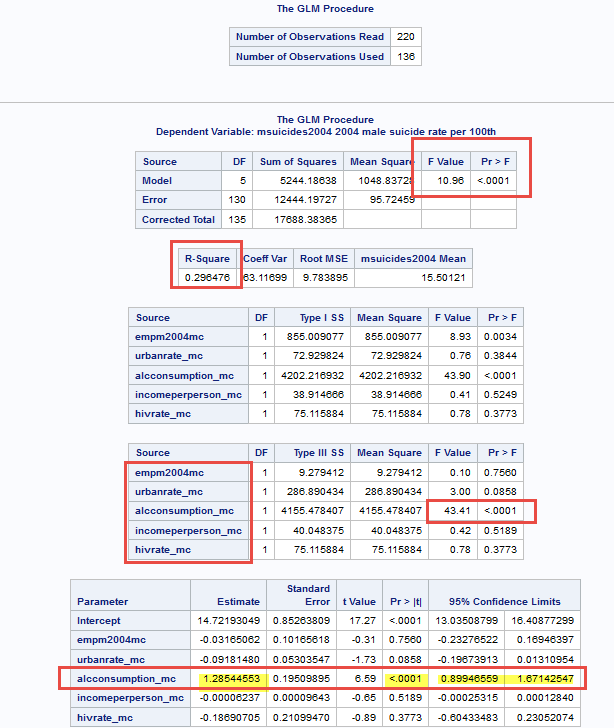
Based on **observed results of MEANS procedure**, we can conclude that mean values of all centered variables **are very close to corresponding center mean=0**.

**Multiple regression models – interpretation of results**

1. **Male model results.**

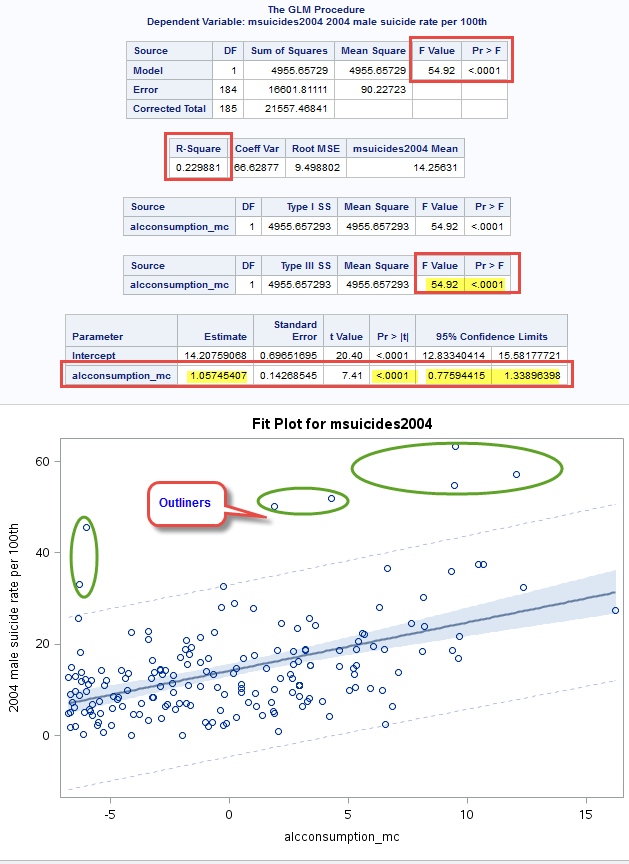
**Conclusion: Using Multiple Regression Test, Report does not support hypothesis that male suicide rating is related to corresponding employment rate. Instead, it shoes strong relationship of male suicide rating with alcohol consumption rate.**

The model shows variability of test sample **R-square= 0.30%** which is far enough from center of 0 and overall p-value of model in less than 0.05. Based on output results we can conclude that **alcconsumption\_mc** (centered explanatory variable) **is the confounder with estimated positive beta value = 1.30 and significant p-value > 0.0001**, showing strong effect and relationship with **msuicides2004** (male 2004 suicide rate). The **F-Value** of confounder **urbanrate\_ms = 12.04** which is very strong and provides additional support to the effect of confounding. Multiple regression model results **also clearly show that relationship between 2004 male suicide (fsuicides2004) and male employment rate (empm004mc) is weak having beta value = (-0.03 and p-value > 0.7560 and we cannot reject NULL Hypothesis.**



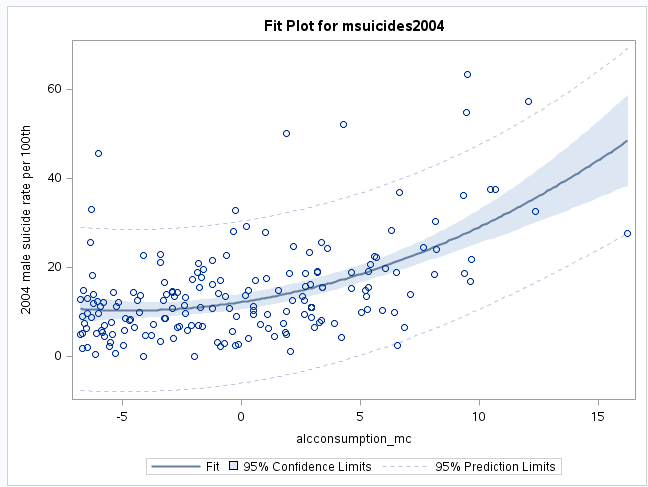
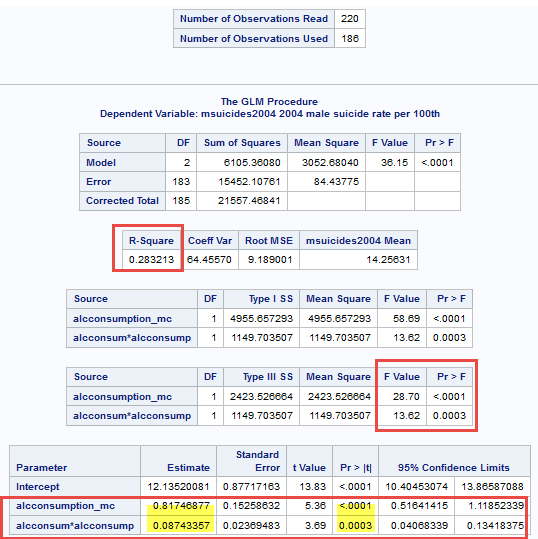
**Interpretations of plots and graphs (with confounding variable alcconsumption\_mc)**

Visual output of **GLM procedure** sows that majority of data is **within 95% limits with few outliners** and have increasing slope and **strong change per unit Beta value = 1.06 and very strong p-value > 0.0001** and we can safely **reject NULL hypothesis and accept alternate hypothesis that male suicide rating (msuicides2004) has strong relationship with confounder urban rate variable (alcconsumption\_mc).** Here output results with highlighted findings:



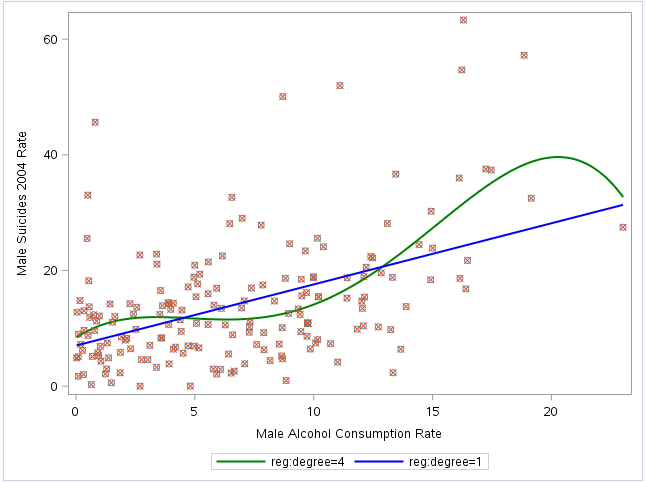
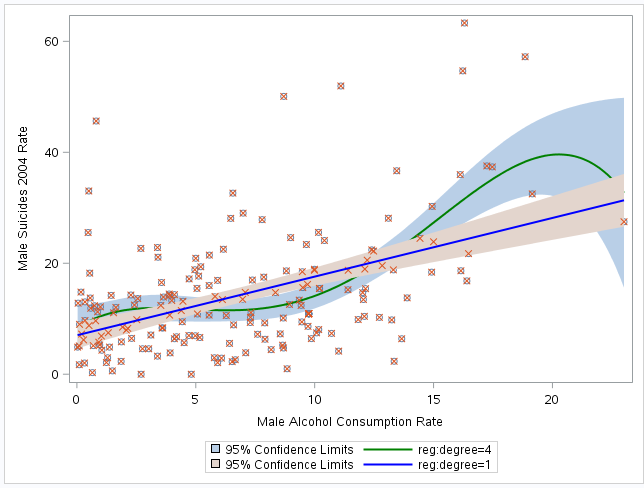
Visually examination of **GLM** procedure plots (linear regression) also show that data may have has some curvature, data distributed uneven on both sides of main linear regression line and we may **polynomial regression model**, including additional squared **alcconsumption\_mc** value into the model, including **SGPLOT** procedure.

The output results are showing that **data fit well polynomial regression model** expectation, **data have curve**. The **R-square value** of variation of sample was **improved from 0.22 to 0.28**, but not significantly. The squared **alcconsumption\_mc** **has beta value of 0.81 and p-value > 0.0001 and the square of alcconsumption\_mc has beta value of 0.087 and p-value > 0.0003 and we can safely reject NULL Hypothesis and accept alternate hypothesis: male suicides are related to alcohol consumption.**

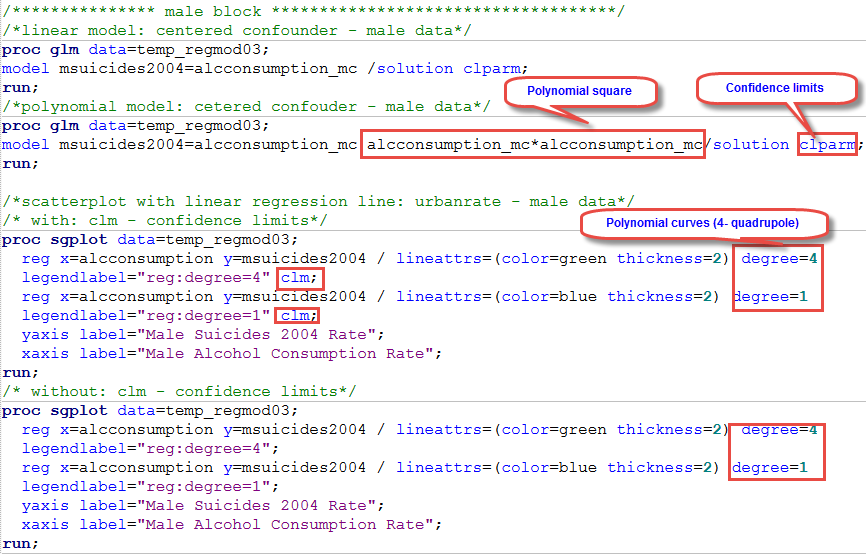


**Output of SQPLOT procedure with CLM (confidence limits parameter) and without:**

Result of **multivariate polynomial regression** with degree=4 shows the best fit without much bias or overfitting.

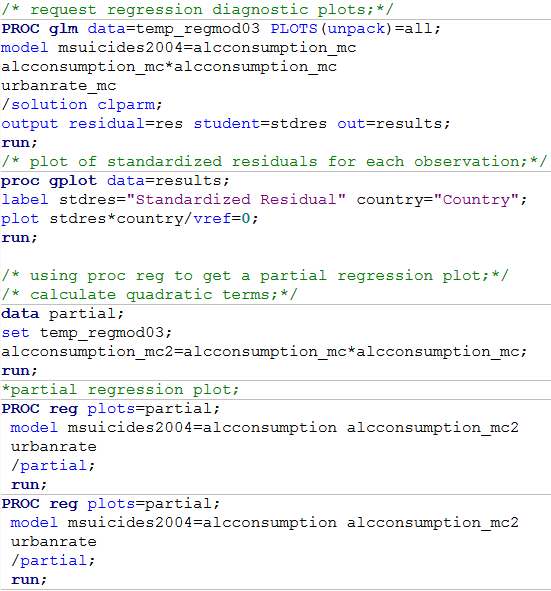


**Program code related to above tests:**

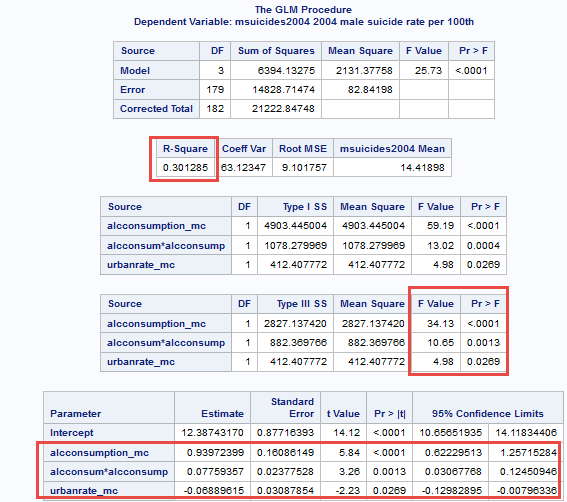


**Interpretation of regression diagnostic plots**

**Program code:**

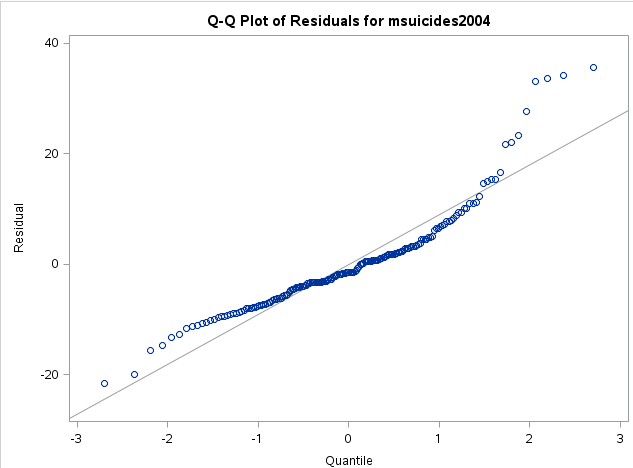


Variable **urbanrate\_mc** added to multiple regression model test and corresponding statistics of GLM procedure show **improvements in R-Square test variability increase to 0.30%** as well as **in beta values for all variables to 0.93, 0.07 and negative -0.07 for urbanrate\_mc with significant p-value > 0.0269**, which definitely improves model fit.

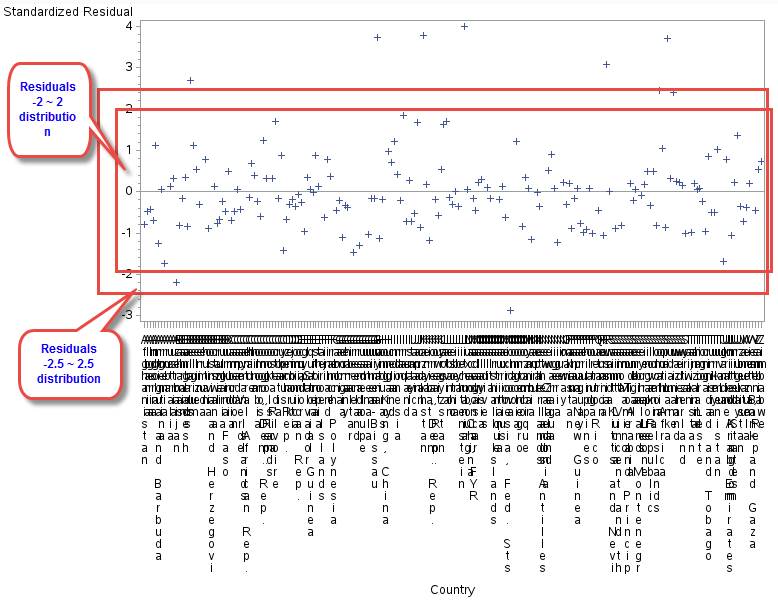


**Q-Q Plot interpretation**

Taking in consideration outliners, affecting distribution, general distribution of data around linear regression line is acceptable, mostly around line.



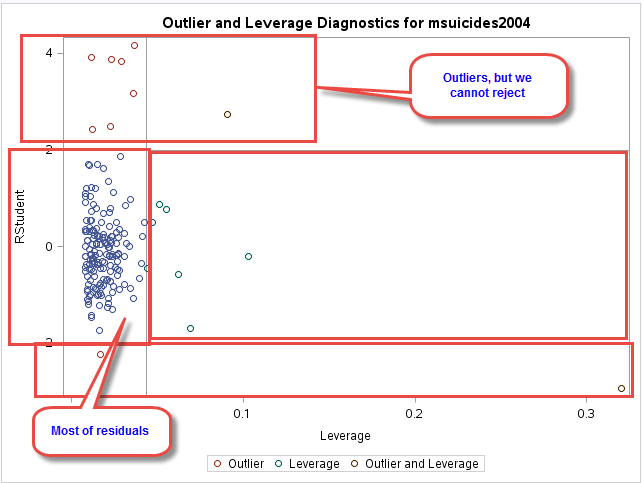
**Standardized residuals for all observations**



Residuals very well distributed within **-2 to 2** **64%** range with few outliners **in -3 to 3** range and few in above **95%** coefficient range and we may conclude that proposed model fit well data.

**Leverage Plot interpretation**

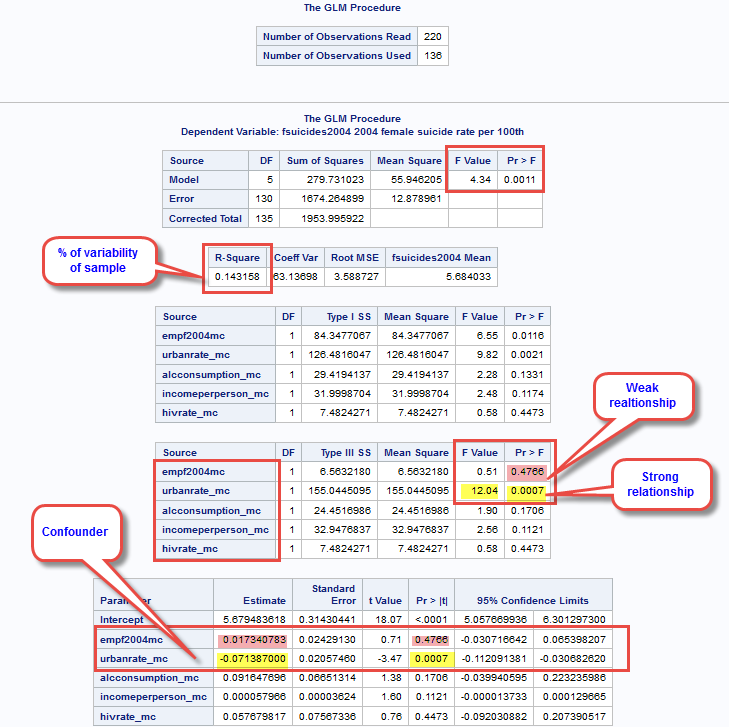
With some outliers data/residuals distributed well, fit model assumptions.



1. **Female model results.**

**Conclusion: Report does not support hypothesis that female suicide rating is related to corresponding employment rate. Instead, it shoes strong relationship of female suicide rating with urban rate.**

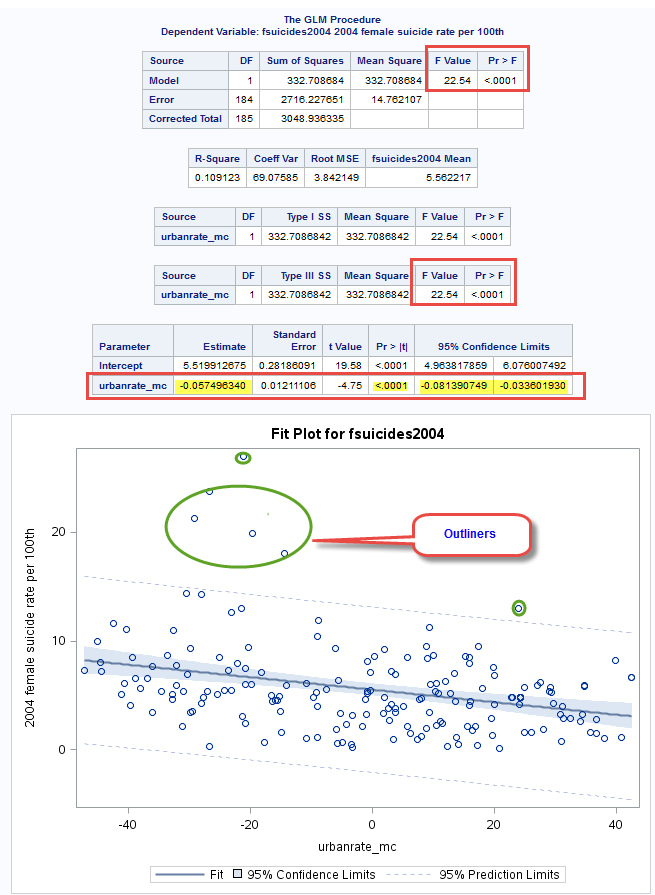
The model shows variability of test sample **R-square= 0.14.31%** which is far enough from center of 0 and overall p-value of model in less than 0.05. Based on output results we can conclude that **urbanrate\_mc** (centered explanatory variable) **is the confounder with estimated negative beta value = (-0.07) and significant p-value > 0.0007**, showing strong effect and relationship with **fsuicides2004** (female 2004 suicide rate). The **F-Value** of confounder **urbanrate\_ms = 12.04** which is very strong and provides additional support to the effect of confounding. Multiple regression model results **also clearly show that relationship between 2004 female suicide (fsuicides2004) and employment rate (empf2004mc) is weak having beta value = 0.01 and p-value > 0.4766 and we cannot reject NULL Hypothesis.**



**Corresponding program** code can be found in **Appendix A**.

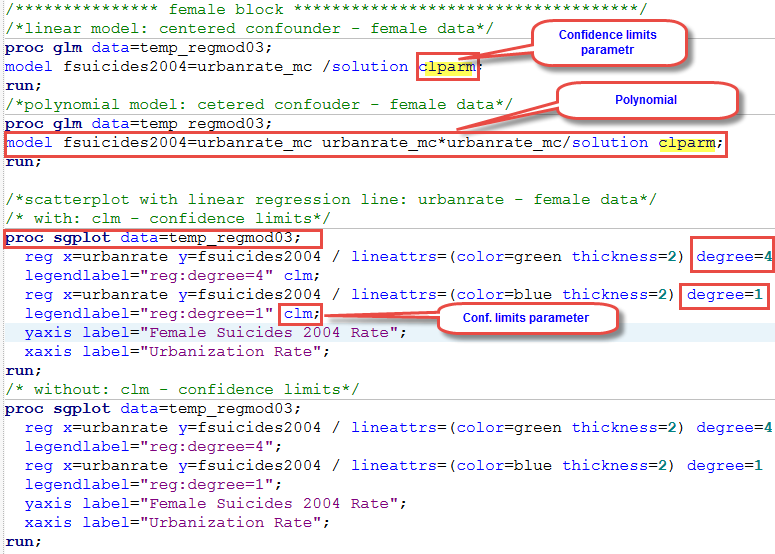
**Interpretations of plots and graphs (with confounding variable urbanrate\_mc)**

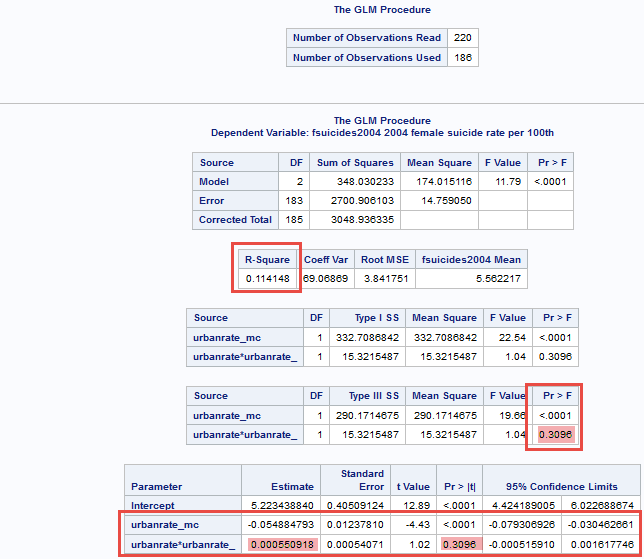
Visual output of GLM procedure sows that majority of data is **within 95% limits with very few outliners** and have decreasing slope and strong change per unit Beta value = (-0.058) and very strong p-value > 0.0001 and we can safely reject NULL hypothesis and accept alternate hypothesis that female suicide rating (fsuicides2004) has strong relationship with confounder urban rate variable (urbanrate\_mc). Here output results with highlighted findings:



Visually examination of plots also shows that data may have has some curvature, data distributed uneven on both sides of main linear regression line and we may **polynomial regression model**, including additional squared **urbanrate\_ms** value into the model, including SGPLOT procedure.

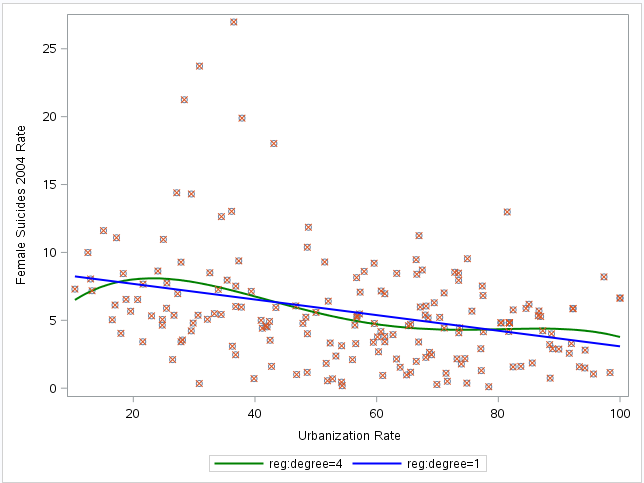
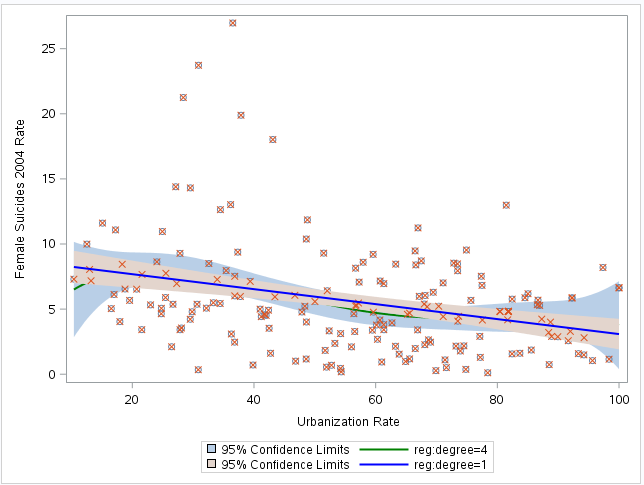
The output results are showing that data does not fit polynomial regression model expectation, **data does not have curve**. The **R-square value** of variation of sample was **improved from 0.10 to 0.11**, but not significantly. The squared **urbanrate\_mc** **has beta value of 0.000 and p-value > 0.3096 and we can safely say that data does not polynomial regression model.**





**Output of SQPLOT procedure with CLM (confidence limits parameter) and without:**

Result of multivariate polynomial regression with degree=4 shows evidence of data bias and overfitting.



Plots are not presented because of multiple regression model failure. **Plots described in male section only.**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* end of week 3 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*