Data Analysis Tools

**Assignment – Week 3**

**Generating a Correlation Coefficient**

**Pearson Correlation**

**By mapolarbear@gmail.com**

This assignment is about **Pearson Correlation** coefficient, how it is calculated and interpreted. **Pearson Correlation coefficient** assesses the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect, positive, linear relationship between the two variables. A correlation of -1 means there is a perfect, negative linear relationship between the two variables. In both cases, knowing the value of one variable, you can perfectly predict the value of the second.

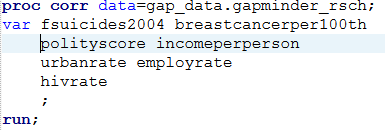
**Research question:** If 2004 suicide rates of females and males have linear correlation to indicators in GAPMINDER dataset and how strong correlation is. In addition, is decided to explore women breast cancer indicator to explore how strong linear correlation of this indicator with other indicators can be.

**Variables:**

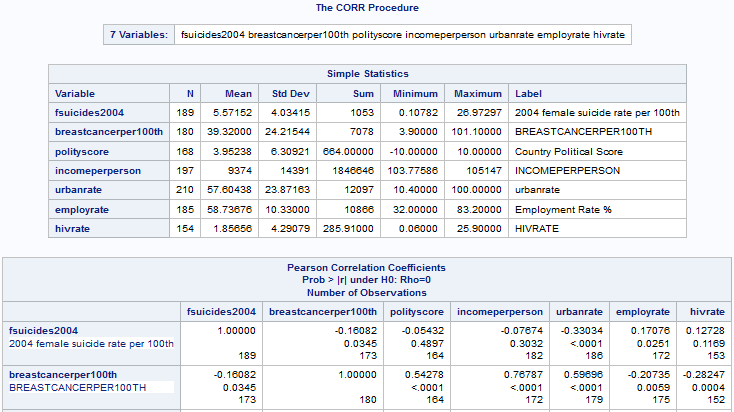
**fsuicides2004, polityscore, incomeperperson, urbanrate, employrate, breastcancerper100th, hivrate**

All variables in participating in my study are QUANTITATIVE.

**Program code:**

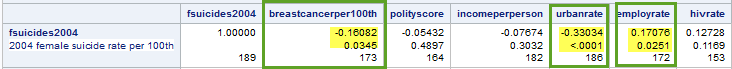


**Output of PROC CORR procedure:**



First, for the purpose of interpretation of PROC CORR output, only variables fsuicides2004 and breastcancerper100th with the linear correlation to other GAPMINDER indicators presented, the full output can be found in Appendix A.

Variable **fsuicides2004** linear correlations

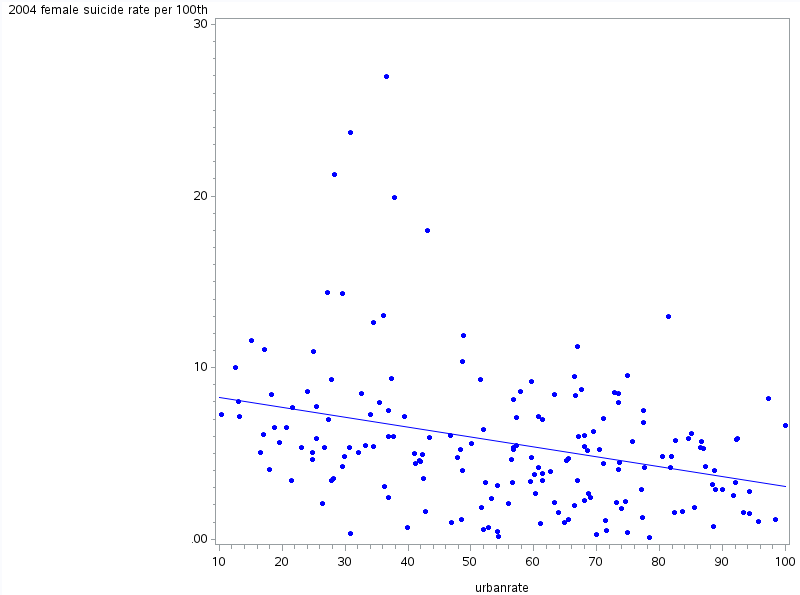


As we can see following outlined (green) variables (breastcancerper200th, urbanrate, employrate) have acceptable **p-values <= 0.05** and we can conclude that each from outlined variables more or less dependent to fsuicides2004, but based on Pearson Correlation coefficient, linear correlations of each variable not that strong, not that far from 0(zero).

The best observed linear correlation for fsuicides2004 variable:

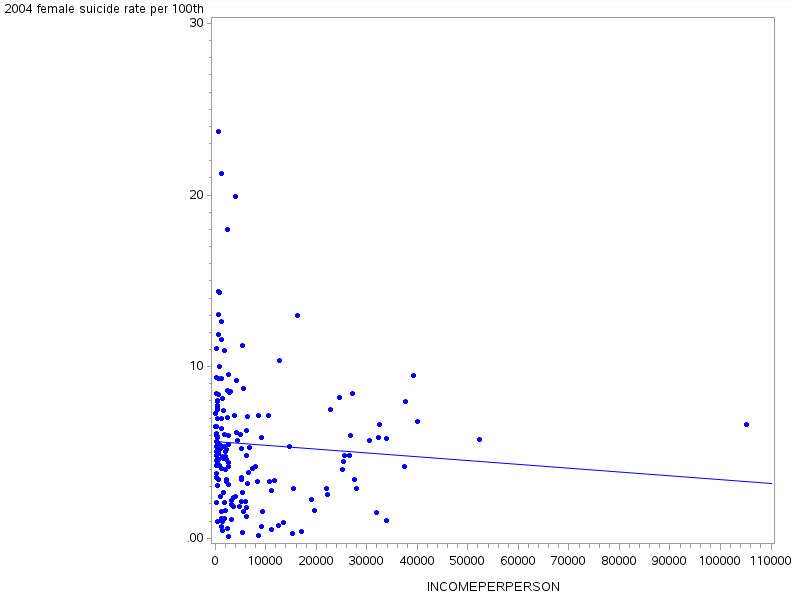
The linear correlation between **fcuicides2004 and urbanrate** is the strongest, having negative linear correlation with values of: **-0.33034 (Pearson Correlation coefficient)** and **p-value of <.0001**.

The -0.33034 is not that fare from 0 (Zero), data is not that concentrated along correlation line, but we can still **make conclusion that fsuicided2004 and urbanratre have strong linear relationship**, and corresponding PLOT output more or less supports my conclusion:



The weakest observed correlation for **fsuicides2004** variable:

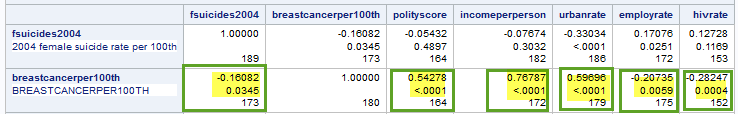
The weakest observed correlation is between **fsuicides2004** and **incomeperperson** with negative linear correlation, almost flat, with value of -0.07674 and p-value of 0.3032, which is way above 0.05 and represented by following plot:



Conclusion:

Variables **fsuicides2004** and **incomeperperson** have strong enough linear relationship.

Variable **breastcancerper100th** linear correlations

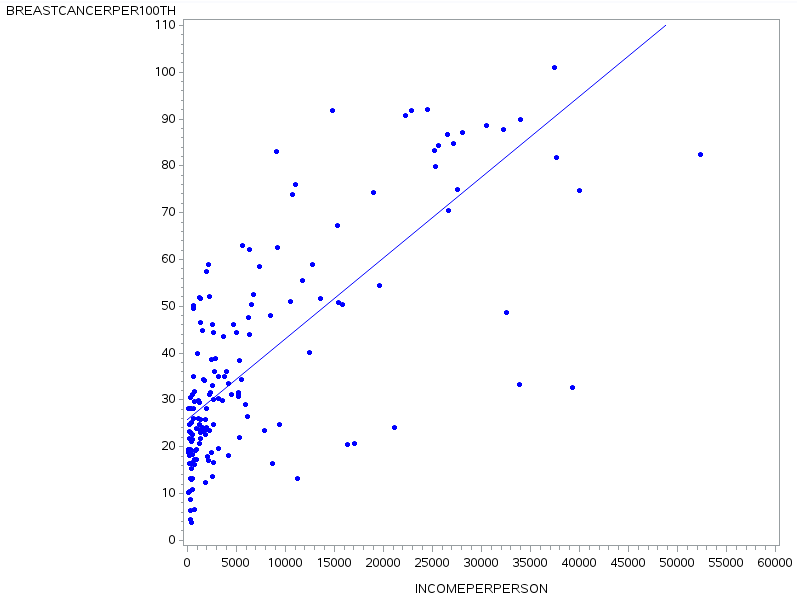


As we can see all outlined (green) variables (**breastcancerper200th, urbanrate, employrate, incomeperperson, employrate, hivrate**) have acceptable **p-values <= 0.05** and we can conclude that each from outlined variables more or less dependent to **breastcancerper100th**, but based on Pearson Correlation coefficient, linear correlations of each variable varies from strong (**0.76787 incomeperperson**) to weak **(fsuicides2004 -0.16082**), but still, we can conclude that all sited variables have linear relationship (positive or negative) to **breastcancerper100th** variable.

The best observed linear correlation for **breastcancerper100th** variable:

The linear correlation between **breastcancerper100th and incomeperperson** is the strongest, having positive linear correlation with values of: **0.76787 (Pearson Correlation coefficient)** and **p-value of <.0001**.

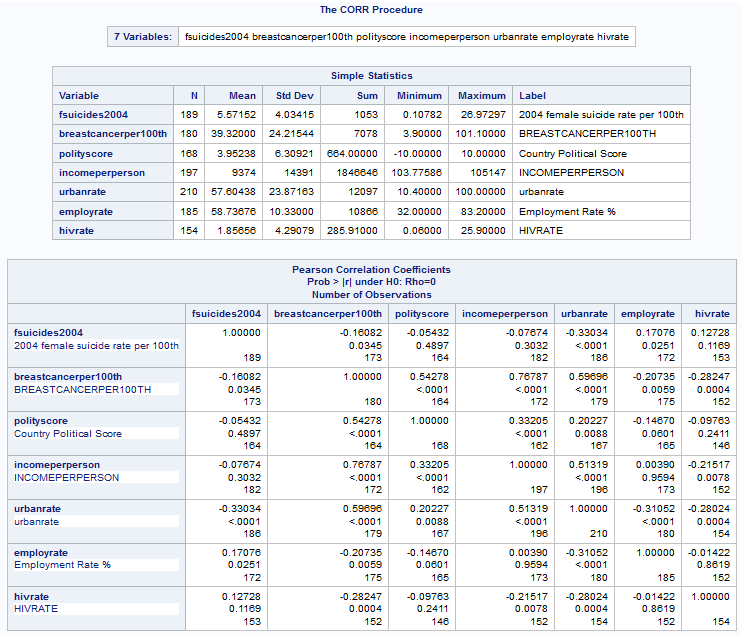
The value of 0.76787 is very close to positive 1 and we can make **conclusion that breastcancerper100th and incomeperperson have strong linear relationship**. Performing analysis of generated plot I also noticed that data is bit shifter to upper side of main linear correlation line and I’m curious if we can add additional interpretation. In addition to that, it was noticed that majority of breast cancer indicator concentrated in low income per person, and maybe we can go further and study low income per person range somewhere between 0 and 4000? Because most of the data is in that range and I think gradual increase of income in that range gives us general vector destination of linear correlation.



Conclusion:

Variables **breastcancerper100th and incomeperperson** have strong linear relationship.

**Appendix A.**



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* End of week 3 assignment \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

<http://coursera-sas-tools-a-week-1.tumblr.com/post/152079337232/data-analysis-tools-assignment-week-1-running>

<http://coursera-sas-dm-week4.tumblr.com/post/151798525721/data-management-and-visualization-assignment>