Regression Modeling in Practice

**Assignment – Week 4**

**Test a Logistics Regression Model**

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

This week's assignment is to test a **logistics regression model** for the association between your primary explanatory variable and response variable, **when both variables are categorical binary** (values 0 or 1), and if there is a relationship between primary variables. The goal is also to test logistics model additional explanatory variables added to model, and to describe results. The other purpose of logistics 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.

**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 logistics 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)**

**fsuicides2004bv** (**binary 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. Values are: 0 – rate below average, 1 – average and above average

**msuicides2004bv** **(binary 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. Values are: 0 – rate below average, 1 – average and above average

**employrate\_f2004bv** (**binary 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. Values are: 0 – rate below average, 1 – average and above average

**employrate\_m2004bv** (**binary 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. Values are: 0 – rate below average, 1 – average and above average

**urbanratebv** (**binary explanatory Variable**) – Urbanization rate in the country. Values are: 0 – rate below average, 1 – average and above average

**incomeperpersonbv** (**binary explanatory Variable**) – Income per person in the country. Values are: 0 – rate below average, 1 – average and above average

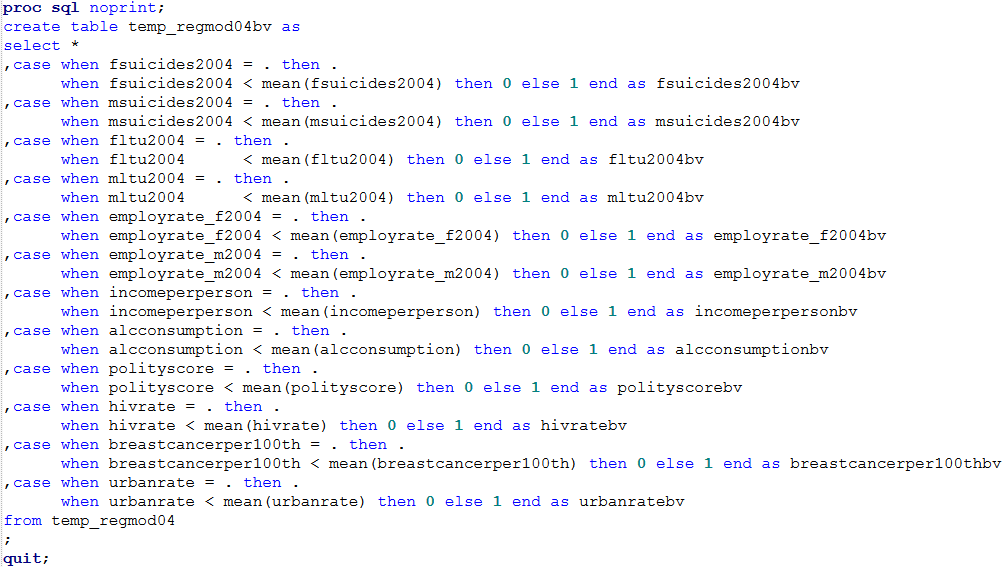
**hivratebv** (**bynary explanatory Variable**) – HIV rate per 100,000 standard population in the country. Values are: 0 – rate below average, 1 – average and above average. Values are: 0 – rate below average, 1 – average and above average

**alcconsumptionbv** (**binary explanatory Variable**) – Alcohol consumption 100,000 standard population rate in the country. Values are: 0 – rate below average, 1 – average and above average

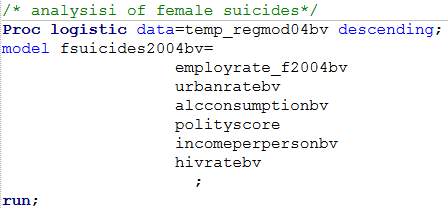
**Centering binary explanatory variables**

The **logistics regression requires to binary variables only**, meaning both response and explanatory variables must have values of 1 (as positive answer) or 0 (as negative answer). The GAPMINDER contains only quantities variables and because of that we have to bin each variable into 2 categories (1 and 0), which is very artificial, but we still need to run this study to complete assignment.

**Program code to create binary variables out of quantitative:**



**Code to run logistics regression** with ALL listed binary variables against female suicides ratings:



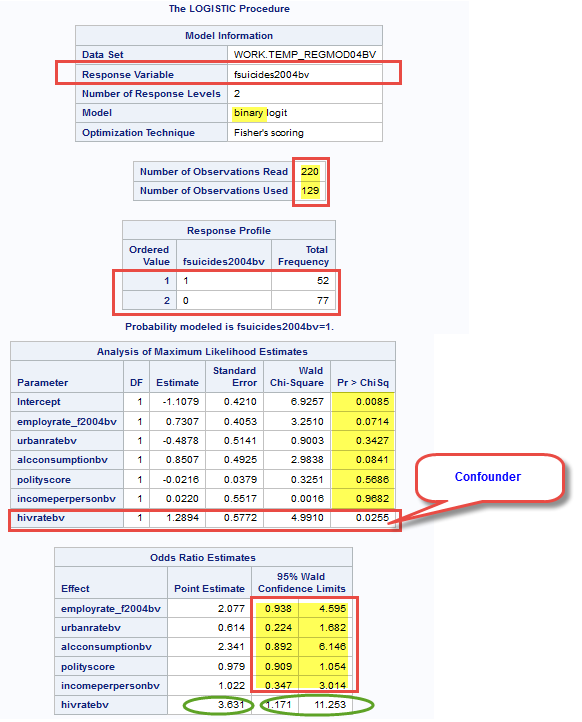
**Logistics Regression Model outputs and interpretations**

* **Female suicides ratings model – with ALL binary explanatory variables (finder confounder)**

As we can see total of **220 observations with 129 observations of complete data** were used in our model. It looks like 1/3rd of observations containing one or more observations with missing data were dropped, decreasing **variation of model sample size**. Parameter estimates and P-value table shows that there is only one variable: **HIVRATEBV with significant estimate value of 1.29 (beta= 1.29) and signioficant p-value of 0.0255** in the model, **all other variables have non sognigicant p-value > 0.05**. Based on results of logistics regression model we can conclude that **HIVRATEBV variable is the major confounder**.

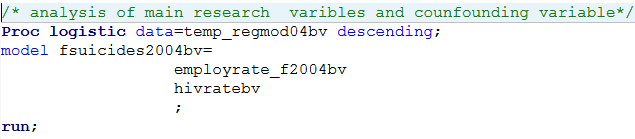
**Odds Ratio estimates**

In general terms, odds ratio is probability of event occuring in one group comparing to probability of event occuring in another group. Based on Odds Ratio Estimates table we can conclude that **females with (1) HIVRATEBV estimate of 3.63 are 3.6 times more likely to commit suicide than without (0).** The **95% Confidence limits of Odds Ration Estimates** also shows that in **95%, or samples will be in between values 1.17 and 11.25**. Another interpretation and **conclusion of odds ration limits** can be: **females with HIV in countries with HIV ratings from 1.17 to 11.25 are more likely commit suicide that without HIV.**



* **Female suicides ratings model – with MAIN binary explanatory variables and confounder**

**Program code:**



As we can see total of **220 observations with 139 observations of complete data** were used in our model. It looks like 1/3rd of observations containing one or more observations with missing data were dropped, decreasing **variation of model sample size**. Parameter estimates and P-value table shows that both variables **employrate\_f2004bv** and **hivratebv** are independently significant:

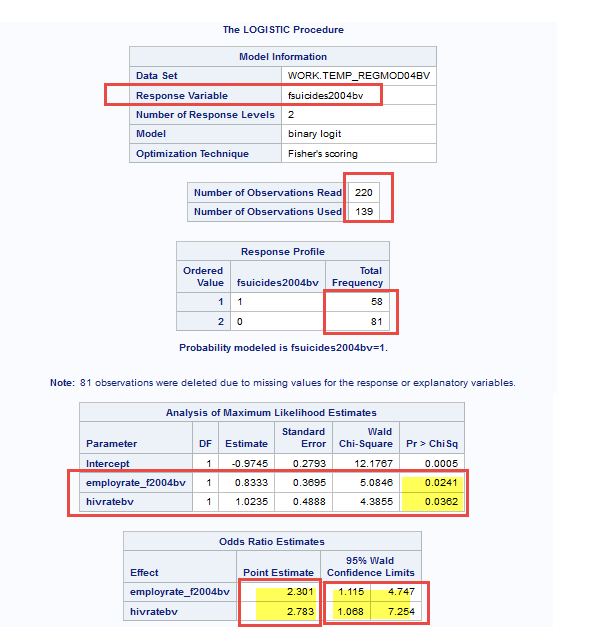
**employrate\_f2004bv with significant estimate value of 0.83 (beta= o.83) and significant p-value of 0.0241**  **hivratebv with significant estimate value of 1.02 (beta= 1.02) and significant p-value of 0.0362**

We can ignore NULL hypothesis and accept alternate hypothesis that

**Odds Ratio estimates**

Based on Odds Ratio Estimates table we can conclude that **females in countries with Employment rate of 2.3 are 2.3 times more likely commit suicides.** We can also create prediction that at the same time, females in countries **with (1) HIV rate estimate of 2.3 are 2.3 times more likely to commit suicide than without (0).** The **95% Confidence limits of Odds Ration Estimates** also shows that in **95%, or samples will be in between values 1.16 to 4.74 and 1.1 to 7.25**. But because of overlapping sample confidence limits, we cannot say that HIV rate is more related to female suicide rates that employment rate. Another interpretation and conclusion of odds ration limits can be: **Controlling female employment rate and HIV rate can affect female suicide rate in countries with female employment rates in between 1.1 to 4.5 and with HIV in countries with HIV ratings from 1.1 to 7.25.**

**Conclusion: Current results support previous finding and we can conclude that there is correlation between female suicides rate and employment rate. We can also conclude that the HIV rate also affects female suicide rating in corresponding country.**



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* end of week 4 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*