# Machine Learning for Data Analysis

**Assignment – Week 2**

**Running a Lasso Regression Analysis**

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

This week's assignment deals with **LASSO Regression** and interpretation of results.

Lasso regression is what is called the Penalized regression method, often used in machine learning to select the subset of variables. It is a supervised machine learning method. Specifically, LASSO is a Shrinkage and Variable Selection method for linear regression models. LASSO, is an acronym for **Least Absolute Selection and Shrinkage Operator**.

The LASSO imposes a constraint on the sum of the absolute values of the model parameters, where the sum has a specified constant as an upper bound. This constraint causes regression coefficients for some variables to shrink towards zero.

The LASSO regression it can provide greater prediction accuracy. If the true relationship between the response variable and the predictors is approximately linear and you have a large number of observations, then OLS regression parameter estimates will have low bias and low variance. LASSO Regression can increase model interpretability.

With Lasso Regression, the regression coefficients for unimportant variables are reduced to zero which effectively removes them from the model and produces a simpler model that selects only the most important predictors.

**About My research**

For research purposes of Machine Learning course we are advised to use ADDHEALTH modified dataset and I’m interested in how substances like **alcohol, marijuana, smoking** etc. Affecting adolescent life and what we can predict and up to what level. I will be using ADDHEALTH for the first time, for my previous courses I was using GAPMINDER, so, this is bit new for me.

Predicting of ALCEVER1 (if adolescent ever drank alcohol) and effect and importance of other variables like

**Sample**

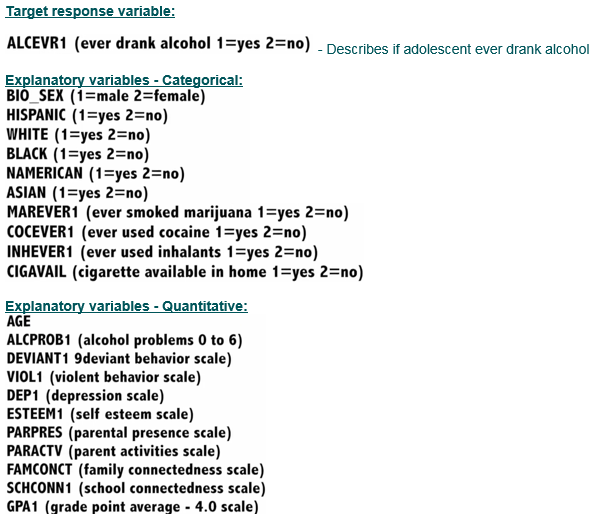
ADDHEALTH - **The sample to** be used represent adolescent various data collected in 2004.

**Procedure**

Data were during 2004.

**Measures (current study)**

Following **target**, **explanatory binary categorical** and e**xplanatory quantitative** variables used in **LASSO Regression:**



**Program code**

The response variable ALCEVR1 value of 2 was assigned as ‘NEVER DRUNC ALCOHOL’ replacing value of 0 (zero).



**Interpretation of results:**

The output of SURVEYSELECT procedure shows that training set created with as 70% (N=3203) from total of 4575 observations input dataset NEW.

The output of GLMSELECT procedure basic as well as detailed information about LASSO Regression method used,



The model was validated against test set (**N=1372**) using **K-Fold (least angle regression algorithm) with K=10**.

**Out of 24 predictors only 13 were chosen/retained in current model.**

Following best set of predictor variables created as an output of each step (adding predictors) to validate change in cross validation average (mean) squared error. Based on observed results, variables **MAREVER1 (Ever tried marihuana), ALCPROBS (alcohol problems), DEVIANT1 (Deviant behavior), TREG1 (regular, non-regular smoker) and AGE are more positively associated with ALCEVR1 (ever drank alcohol)**.

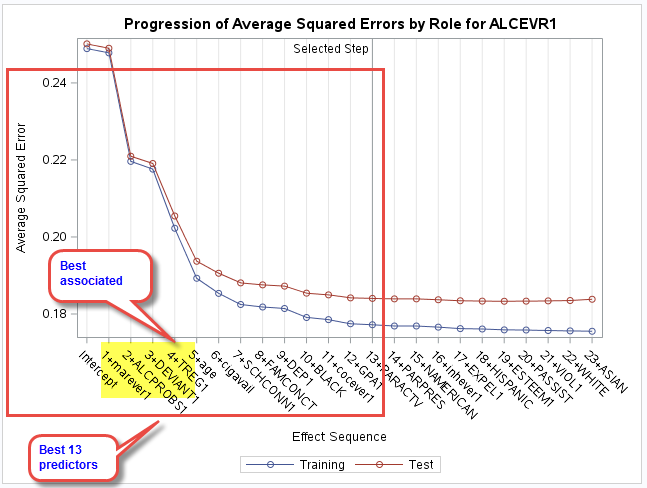
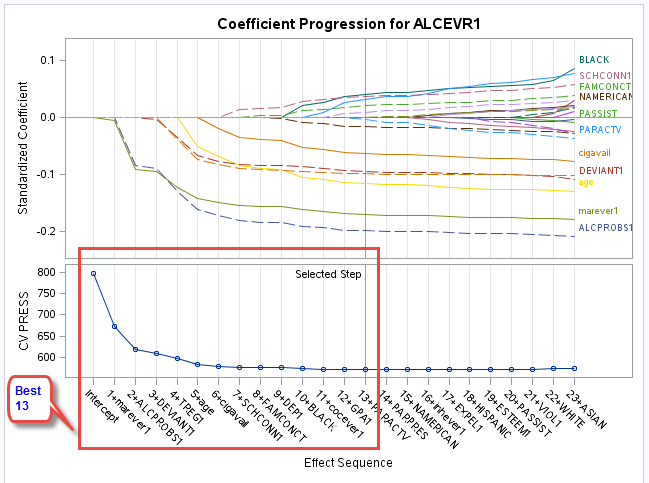
We can also conclude that cigarette availability (CIGAVIL), school consecutiveness (CSHCONN1), family connectivity (FAMCONCT) and other predictors within group of 13 predictors are positively associated with alcohol problems.

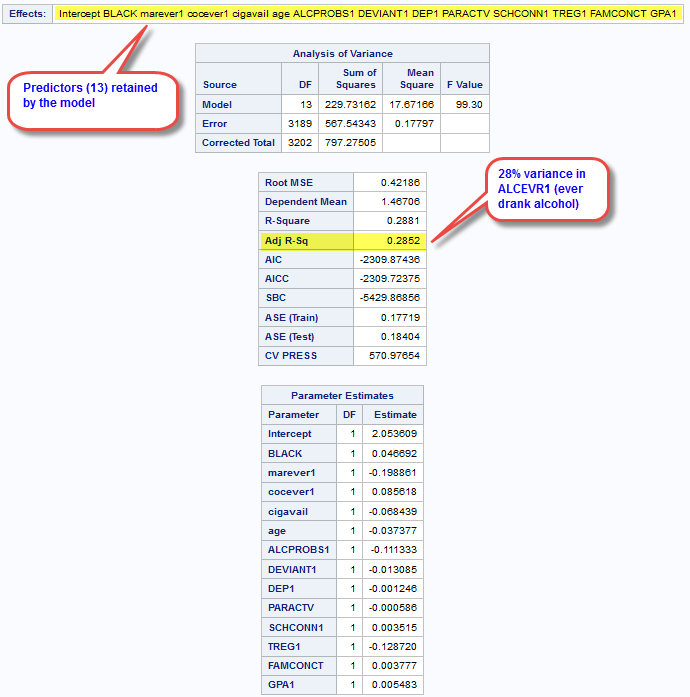
**These 13 predictors represent 28% of variance in adolescent ever drank alcohol (ALCEVR1).**

Attached are various tabular and graphical output supporting above conclusions.

In addition, current model correspond to Classification Tree previously created.





http://mapolarbear-da.blogspot.com/2017/01/ml-assignment-week3-lasso.html

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*