**[Regression and Exploratory Analysis](https://blackboard.iit.edu/webapps/assignment/uploadAssignment?content_id=_499415_1&course_id=_58224_1&group_id=&mode=view)**

**Analyzing the Labor and Income Dynamic Set**

**Introduction:**

The main purpose of the report is to study and analyze the labor and Income Dynamics of Canada. Linear regression is performed on the dataset using response variable wages. Analysis is performed on other variables as well in the dataset and studied the significance to model the linear regression.

**Objective/Problem Statement:**

We study the data of labor and income data using 5 different variables in the labor department produced by various factors. The wages vary with respect to education, gender and other factors. The dataset consists of 3988 observations and 4 features aiming to predict one real valued response.

Here – wages are taken as response (Y1) variables. The analysis is performed on the below variables as mentioned below in order to check the trend of response variables.

**Information about 4 variables:**

* **x1 = Education**
* **x2 = Age**
* **x3 = sex (Male = 1, Female =2)**
* **x4 = language**

**Data Set Characteristics**:  Multivariate

**Attribute Characteristics**: Categorical, Integer, Real

**Associated Tasks**: Classification, Regression

**Number of Instances**: 3988

**Number of Attributes**: 5

The dataset contains eleven attributes (or features, denoted by X1...X4) and one response variable (or outcomes, denoted by y1). The aim is to use the eight features to predict each of the two responses.

For the above variables, wage is the response variables. First of all, we calculated the correlation of the response variable and other variables individually.

Details on Dataset:

**File Name:** SLID.csv

**Source (link):** <https://vincentarelbundock.github.io/Rdatasets/doc/car/SLID.html>

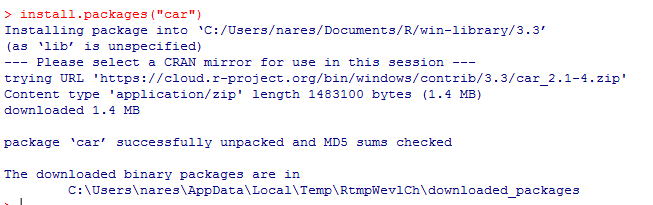
<https://vincentarelbundock.github.io/Rdatasets/datasets.html>

**Total number of rows in the raw/final dataset:** 3988

**Time period of the dataset:** 1994

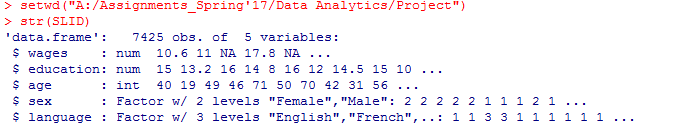
**Linear Regression Analysis on SLID:**

**Installing car Package**



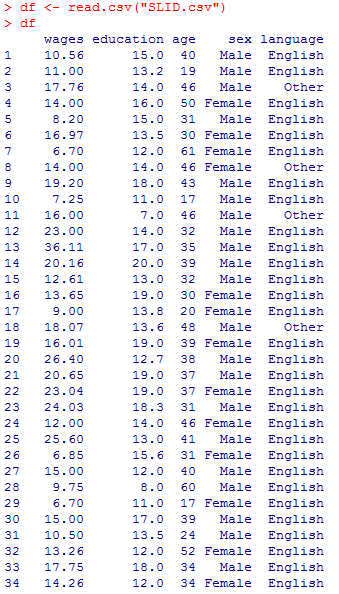
**Structure of the source file**

**Str(SLID)**



**Data Frame holding SLID dataset**

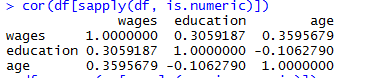




First, we are loading data into dataframe (df) and we are displaying the structure using str function. From the structure, we identified 4 independent variables. That will have significance impact on wages.

Also, we are determining the correlation between the variables and a scatter plot between those variables.

**Correlation values for all pairs of variables**



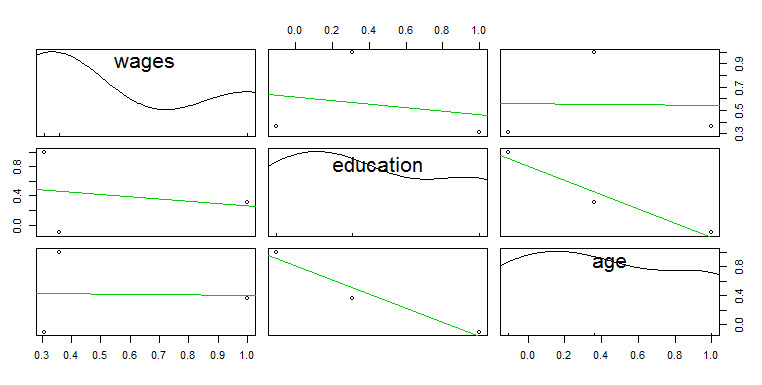
**lm**

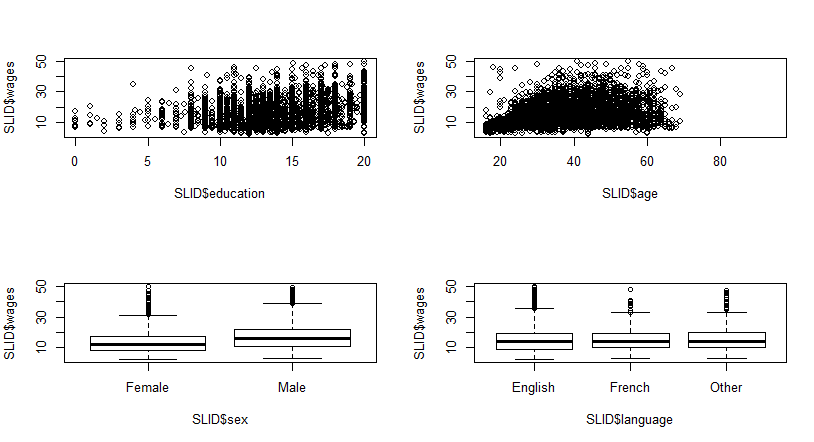
According to the correlation matrix, the correlation between two response variables is very high (i.e.0.35). Therefore, by analyzing one of the response variables, we can predict the value of other response variable. So, we have considered wage as the only response variable.

**Study of Variables (Scatter Plot):**

The pairwise scatter plot is created with the variables in the dataset to study the variables behavior against the target and then we conduct the linear regression for the same data set. These scatter plot provide an initial idea and a high-level diagnosis report before conducting the statistical analysis.







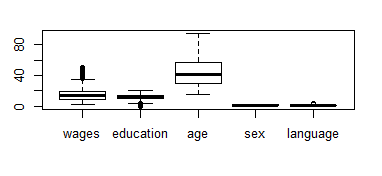
We find the below inference from the scatter plot visualization.

*Education (Upper left) – There is a positive relationship between education & wage*

*Age (Upper right) – There is a positive relationship between age and wage*

*Sex (Lower left) – box plot represents that wages of males are higher than females*

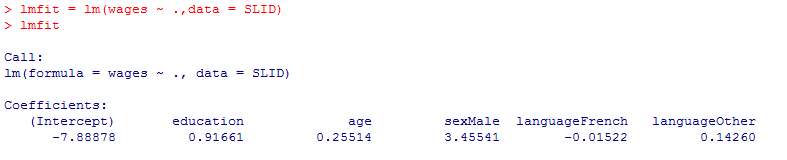
*Language (lower right) – The box plot between the three different languages and wages represents that the correlation is not obvious.*

**Boxplot of Target and Predictors:**

**1) Use R-squared value as % influence. Include F-value and p-value validation to check association of you X value.**

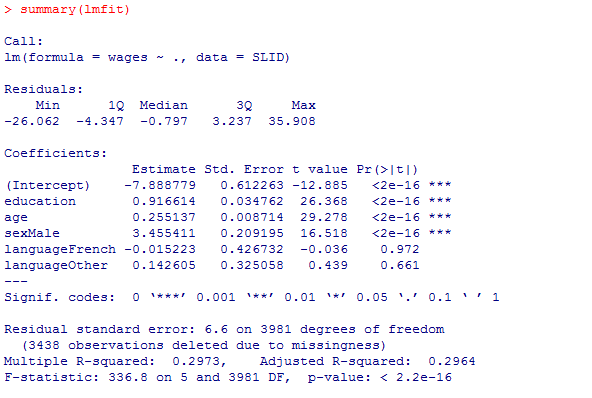
**Fitting a model**

Now, we will declare education, sex, age and language as predictor variables to analyze and summarize the regression model.



**Summary**

Summary of the fit model using summary function. The output includes multiple R-squared correlation among the determined parameter estimates, residual standard errors and standard errors



*Y=β0+ β1X1+ β2X2+ β3X3+ β4X4*

*Y = Wages*

*X1 = Education*

*X2 = Age*

*X3 = Sex*

*X4 = Language*

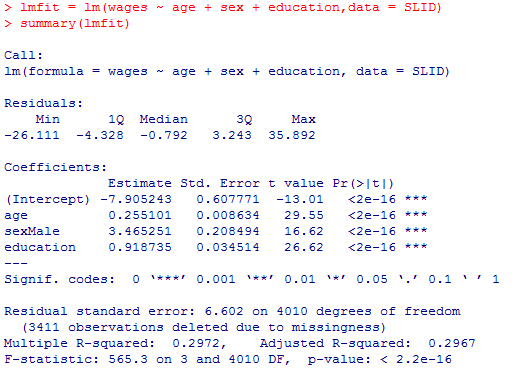
*From the above values*

*Y= -7.888779 + 0.916614 education + 0.255137 sex + 3.45411 language*

*From the above regression analysis, we identified that education, sex and age are creating significant impact but not language as their pr > 0.05. Hence, we are ignoring the language variable as it is not creating any significance on wage variable.*

*F-statistic value = 336.8*

Dropping one of the attribute (language), and checking the analysis with rest of the attributes and create the model



*Y=β0+ β1X1+ β2X2+ β3X3*

*Y = Wages*

*X1 = Age*

*X2 = Sex*

*X3 = Education*

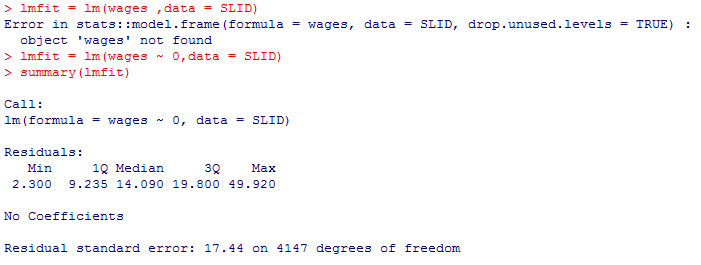
*From the above values*

*Y= -7.905243 + 0.255101 sex + 3.465251 language + 0.918735 education*

*As all the dependent variable pr values are below 0.05, hence this is the best good fit model.*

*Here the F-Statistic value is appreciated from 336.8 to 565.3*

Y value when X = 0

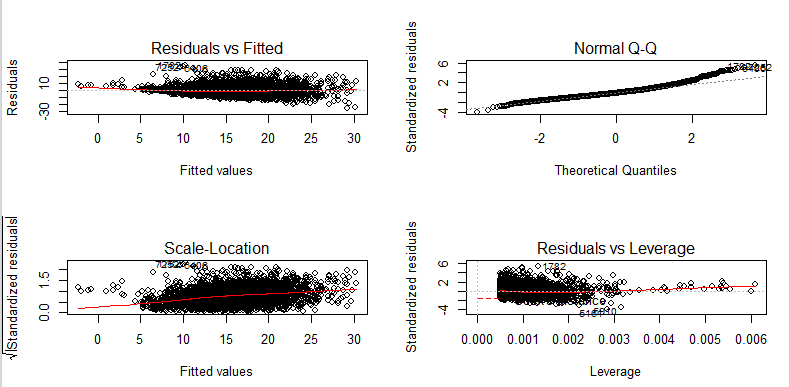


*When all dependent variables are equal to 0 then residual error = 17.44*

As wages are ranging over several orders of magnitude, to induce the

**Creating a diagnostic plot of lmfit**

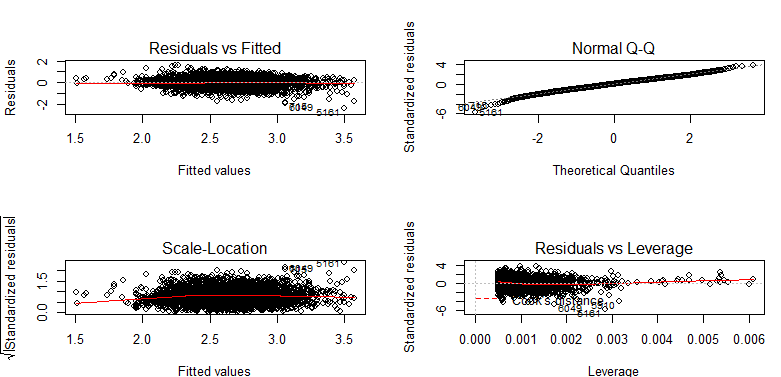




*Here, all four plots (residuals vs fitted, scale – location, and residuals vs leverage) are biased towards regression model. Since wages*

Replot the lmfit with log wage values and with same independent variable and creating a diagnostic limit





**Multi-Collinearity**

Determining the multi-collinearity of the regression model

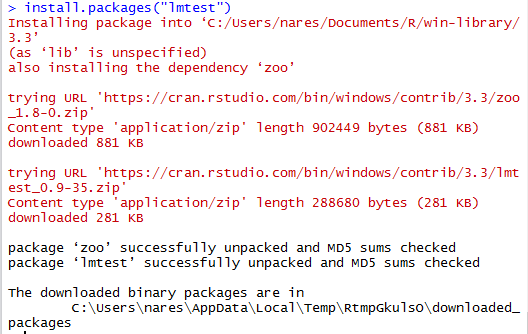
We are conducting this test to check whether multi-collinearity exists or not in the model.



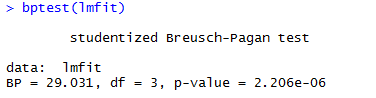


**Heteroscedasticity**

Diagnose the heteroscedasticity of the regression model with the bptest function



**Breusch-Pagen Test for heteroscedasticity:**



We are performing this test, to check whether heteroscedasticity exists in the model. The assumption is that the variance of the error is homogenous or constant across observations. But in heteroscedasticity, the variance is unequal across the observations which mislead our analysis or test conducted on the dataset.

In order to identify the heteroscedasticity, we can perform the Breusch-Pagen test within the lmtest package. In this case, p value is < 0.5 implies there are no null hypothesis of homoscedasticity.

**Robust Standard Errors**

We conduct this test to increase the significance of truly significant parameters correct the standard errors and correct the standard error. We have to conduct the least squares model as rms accepts only fitted models.

