Homework 1
Due Date: May 22, 2018

Homework 1: Part 1 (Linear Regression)

Objective:
We will use Linear Regression to study the impact of a variety of factors such as Education, Experience, Occupation and Gender on Wages.

Data:
The data we will be using is the Current Population Survey (CPS) 85 dataset, available here: http://lib.stat.cmu.edu/datasets/CPS_85_Wages

A tab separated file will also be uploaded here:
https://drive.google.com/file/d/1Xv9NiX-KK4TyOozmQCdnneZhcg47s1wY/view?usp=sharing

Note:
Some notes about getting started with Python or R are available here:
https://docs.google.com/document/d/1ZFKsL3aQnC7Rq3zar68tRAe7m7a_C7kdAN4xIlhfvDs/edit?usp=sharing

There is no strict requirement to use either Python or R, but if you do not have a strong preference for some other ML package, either of the two is good.

Tasks

1. Data Exploration 5%
    a. Read the dataset description, and load the data from the file.
    b. Plot a histogram of the WAGE variable. What is the mean, median, minimum and maximum WAGE present in the dataset? Is the WAGE data skewed?
    c. Plot the WAGE vs. EDUCATION. Comment briefly on any observed trends
    d. Create box plots of the WAGE vs. OCCUPATION and WAGE vs. SEX. What are your observations?
    e. Excluding WAGE, which two variables have the highest correlation?

2. Model Fitting 25%
    a. Before fitting any regression models, convert any categorical variables such as OCCUPATION or SECTOR to a categorical type.
    b. Fit a linear regression model, with the WAGE variable as the dependant variable and the other variables as predictors. Report the R-squared and adjusted R-squared metrics. Explain these metrics.
    c. What is the regression coefficient of the EDUCATION variable? How can this coefficient be interpreted?
    d. What are the regression coefficients of all SECTOR related categorical variables? How can these be interpreted?

e. Observe the p-values of the regression coefficients. Does a higher p-value indicate more or less certainty of a regression coefficient?
f. Report the AIC (Akaike Information Criteria) metric of the above model. This is another metric to evaluate the model fit, adjusting for model complexity. Lower values indicate a better fit.
g. The residuals provide diagnostics about the model fit. For a well behaved model, the residuals should be Normally distributed with 0 mean and constant variance, and have no correlation with the predicted value. Plot the Residuals vs. the predicted value, as well as a Quantile-Quantile plot (QQplot) of the Residuals with respect to the standard Normal distribution.

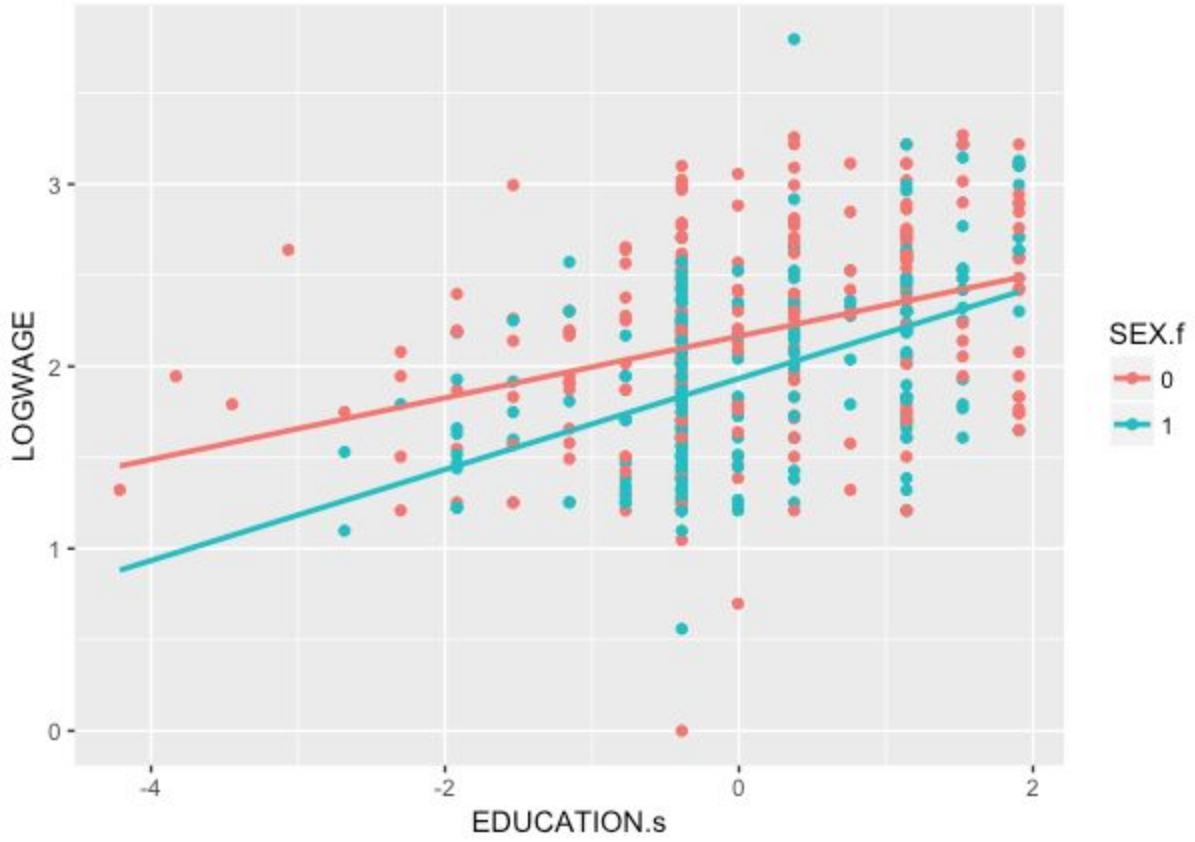
3. Feature Transformations 10%
   a. The Residual diagnostic plots should indicate that the residuals do not follow a Normal distribution. The behavior gets worse as the predicted values increase. Recall that the WAGE distribution in part 1 was skewed, which may cause this problem. Calculate the log of the WAGE, and plot it's histogram. Is this less skewed?
   b. Fit a new regression model, with the log(WAGE) as the dependant variable. Report the R-squared, adjusted R-squared and AIC metrics. Is the quality of the model fit better or worse?
   c. Plot the Residual vs. predicted values for this new model, as well as the QQPlot. Is the Residual distribution closer to a Normal distributions.
   d. Report the p-values for the EDUCATION and AGE variables.

4. Multicollinearity 5%
   a. The high correlation between EDUCATION and AGE for part 1, along with their high p-values indicates that indicates that these variables are not independent, and compete to predict the WAGE. This is known as Multicollinearity. Another diagnostic for Multicollinearity is the Variance Inflation Factor (VIF). Which variables have the highest VIF?
   b. Drop the AGE variable, and refit the model. What is the p-value for EDUCATION?
   c. Did the R-squared, adjusted R-squared and AIC metrics change significantly?

5. Conclusion 5%
   a. One remaining issue is comparing the effect of continuous variables such as EDUCATION that cover a large range vs. Categorical variables such as SEX. To aid in interpretation, standardize the EDUCATION and EXPERIENCE variables and refit the model. Report the R-squared, adjusted R-squared and AIC metrics.
   b. What are the final coefficients for the EDUCATION, SEX and OCCUPATION variables on the WAGE? What is your interpretation of these coefficients?

6. Extra Credit

a. Let's visually look at the effect of EDUCATION and SEX on the wages:



It appears that at higher levels of EDUCATION, the wage gap based on SEX decreases. Quantify this impact, by fitting a new regression model with an extra predictor, the Interaction of EDUCATION and SEX. What are the regression coefficients on SEX, EDUCATION and new Interaction predictor?

Homework 1: Part 2 (Logistic Regression)

Objective:
For this assignment, we examine the Census Income dataset available at the UC Irvine Machine Learning Repository. We aim to predict whether an individual's income will be greater than $50,000 per year based on several attributes from the census data.

Introduction:
The US Adult Census dataset is a repository of 48,842 entries extracted from the 1994 US Census database.

Tasks

1. Getting the Data 5%
    a. Downloadable from http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
    b. Fields and values:
        i. age: continuous.
        ii. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
        iii. fnlwgt: continuous.
        iv. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
        v. education-num: continuous.
        vi. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
        vii. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
        viii. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
        ix. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
        x. sex: Female, Male.
        xi. capital-gain: continuous.
        xii. capital-loss: continuous.
        xiii. hours-per-week: continuous.
        xiv. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
        xv. Label whether the income of that person is > 50k or <= 50k

2. Preprocessing the data 5%

    Some of these columns have a large number of factors. We can clean these columns by combining similar factors, thus reducing the total number of factors. For Example:
    a. We can combine similar work classes and reduce them
    b. Marital Status combining can group people with almost similar marital status
    c. Number of different countries can be mapped to their continents
    d. For missing data for any field, instead of showing ? we can count them as value "NA"
3. Model training 15%

    We are going to build a logistic regression model is to classify people into two groups, below 50k or above 50k in income.

Split the data into a Training and Validation set, with an 80/20 split. Train the model using the Training set.
Save the coefficient/weights for each feature/field in a file after training the model

4. Prediction 10%
   After training, we then predict the salary class using the test group given by:
   http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test
   For that, we are supposed to read the coefficients from the file saved in the previous step, and fit that trained and saved model on the test data

5. Reporting the results: 15%
   - This task consists of following subtasks:
     - Report the precision, recall and accuracy on the Training and Validation sets
     - Generate the confusion matrix for Train and Validation sets
     - Show the ROC curve

Extra Credit:

- Instead of a splitting the data into a single Train and Validation set, perform 5-fold cross validation and report the cross validated precision, recall and accuracy scores