ST 595
Fall 2020
10/25/20

Technical Report:

Logistic Regression Statistical Model Analysis in R

Prepared By:

Josh Olsen

Table of Contents

INTRODUCTION

Being able to describe which explanatory variables are statistically significant to the income response variable is extremely valuable.
Government agencies both local and federal can use this formation for planning purposes.
Credit Card companies can determine if someone is a high credit risk.

This project explores logistic regression using the UCI Adult Income data set. The dataset contains anonymous information (explanatory variables) such as age, occupation, education, working class, etc. with a response variable which has two possible values “greater than $50k” or “less than equal to $50k.

Question of Interests

1. Are there explanatory variables that are highly correlated with each other? If so, which ones?

2. Which explanatory variables are not statistically significant to the income predictor?

3. How accurate is the Generalized Linear Model predict “greater than 50k” or “less than equal to 50k” binary response variable?

REVIEW OF THE DATASET

As stated in the introduction, the dataset is from UCI which contains Adult Income information.. The dataset contains anonymous information (explanatory variables) such as age, occupation, education, working class, etc. with a response variable which has two possible values “greater than $50k” or “less than $50k.

Description of dataset

There are about 32,560 observations with 15 variables with various of factors. A few examples of the variables range from age, education (11th, 12th, etc.) race (various amount of factors), sex (male and felame), native country (various amount of factors), hour per week worked, etc.
The response variable we will be analyzing is the income variable.
This consists of categorical binary data. For example, <=50k and >50k.

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 ...
##  $ workclass     : Factor w/ 8 levels " Federal-gov",..: 7 6 4 4 4 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 ...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 ...
##  $ education_num : int  13 13 9 7 13 ...
##  $ marital_status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 ...
##  $ occupation    : Factor w/ 14 levels " Adm-clerical",..: 1 4 6 6 10 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 ...
##  $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 ...
##  $ capital_gain  : int  2174 0 0 0 0 ...
##  $ capital_loss  : int  0 0 0 0 0 ...
##  $ hours_per_week: int  40 13 40 40 40 ...
##  $ native_country: Factor w/ 41 levels " Cambodia"," Canada",..: 39 39 39 39 5 ...
##  $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 ...
```

Figure 1 - Original Dataset Summary

Figure 1 displays all the explanatory variables and the response variable in the dataset.

The distribution of income is displayed in Figure 2 below.
About 75% are “<=50” and 25% are “>50k”.

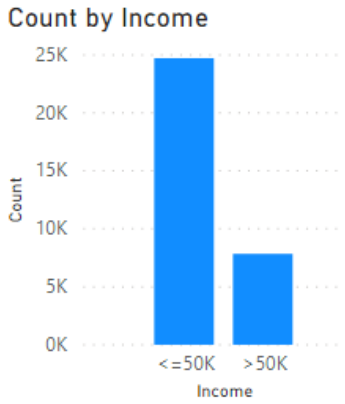

Figure 2 - Distribution of Income Response Variable

OBSERVATIONS

Missing Data

Missing data can be problematic for any statistical model. Below in **Figure 3** is a visual of missing data within the dataset. There is a very low amount of missing data. There appears to be only 3 explanatory variables that contain missing values. Those explanatory variables are workclass, occupation, and native county.

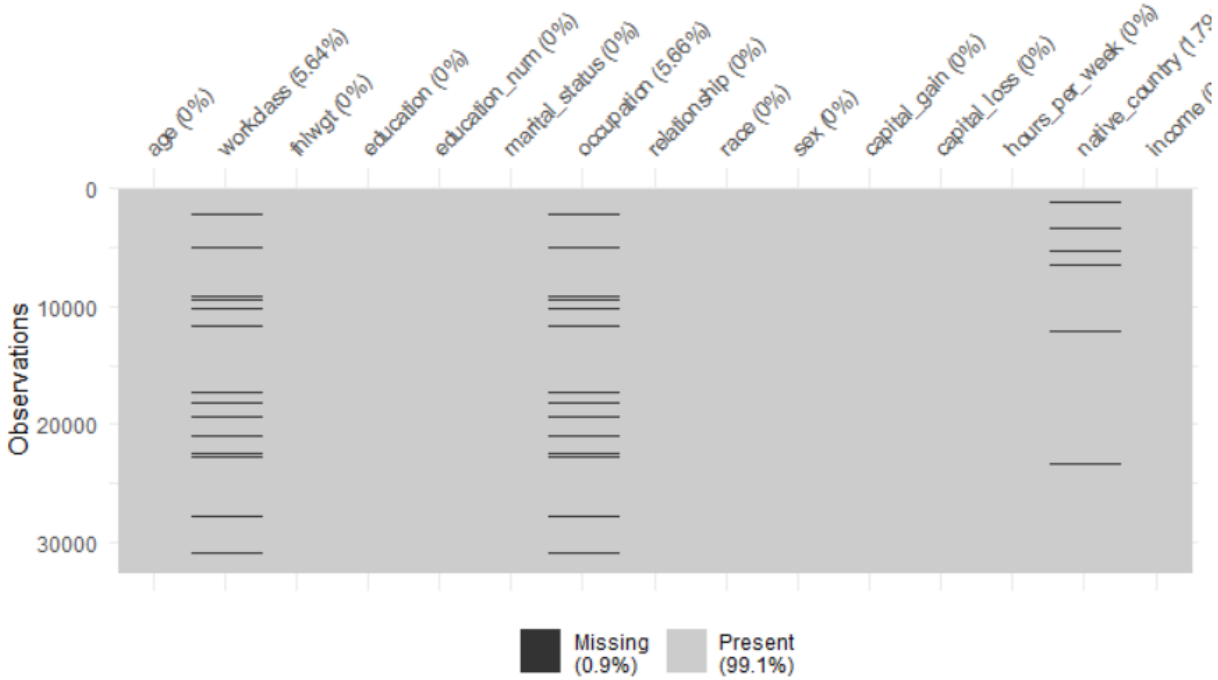

Figure 3 – Visual of Missing Data

Addressing Missing Data

The explanatory variables which are missing values are workclass, occupation, and native county. It appears that the data was not collected or unknown. All missing values are replaced with “Not-Known”. Below in **Figure 4** displays a visual of missing data after replacing missing values with “Not-Known”.

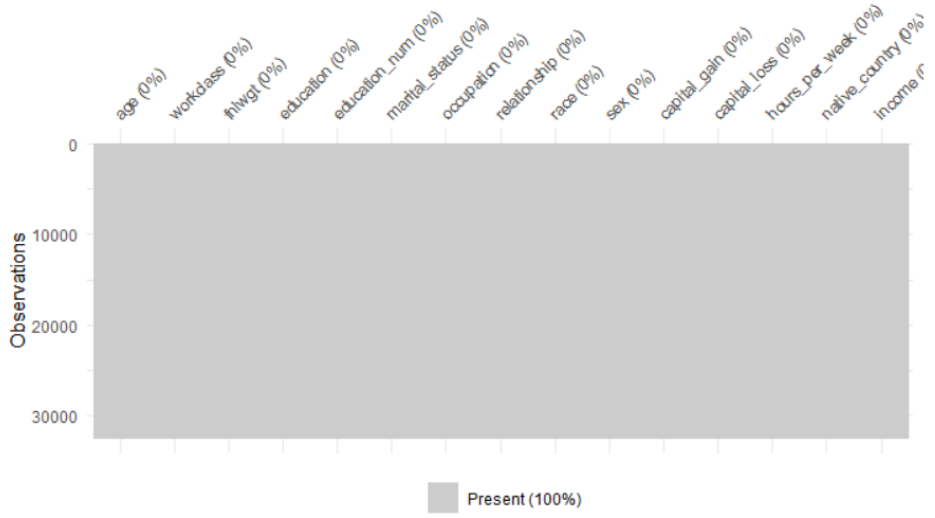

Figure 4 – Updated Visual of Missing Data

After addressing for missing values, **Figure 4** displays that there are not any missing values.

Cleaning Up Variables

Re-Grouping Native County to Native Region

All the Native Counties have been re-organized into a region based on geography.
The new Native Regions are Asia_East, Asia_Central, Central_America, South_America, Europe_West, Europe_East.

Skewness in Continuous Variables

Continuous variables need to analyzed to determine of there distribution is skewed. The two common types of skewness is left-skewed and right skew-ed. The continuous variables in this dataset are age, fnlwgt, education_num, captial_gain, captial_loss, and hours_per_week.

Fnlwgt, capital_gain, and capital_loss display significant amount of skewness.
Each of these variables will need to be address in their own unique way.

Address Skewness in Fnlwgt

Fnlwgt continuous variable is heavily skewed. Figure 6 displays that this variable is left skewed. The log of fnlwgt is performed to address the skewness.

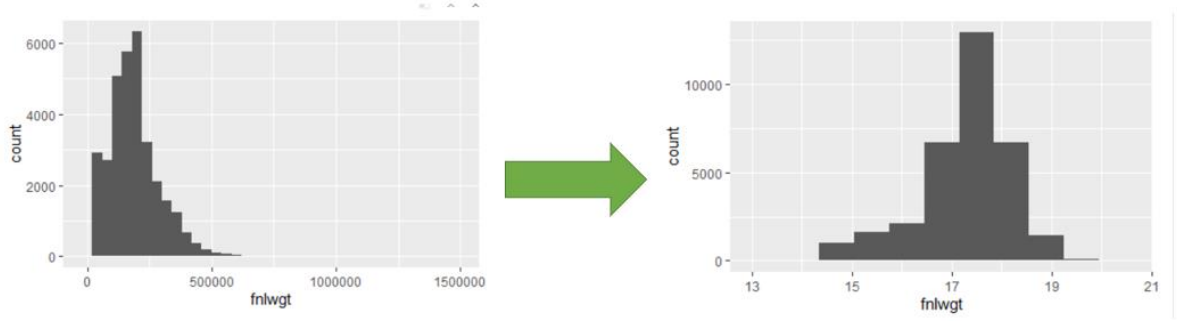

Figure 6 – Transformation of fnlwgt to log of fnlwgt

After taking the log of fnlwgt, the distribution is not skewed anymore. The plot on the right in **Figure 6** displays. This means that we have a normal distribution.

Address Skewness in Capital Gains and Capital Loss

Capital Gains and Captial Loss are two continuous variables that behave very similarly. There are a significant amount of 0's mean that are no capital gains or losses. I will address the skewness by transforming the Capital Gains and Capital Loss continuous variables to categorical variables.

Captial Gains - Grouped into buckets of None, 1_to_10000, 10001_to_20000, 20001_to_30000, and more_than_30k.

Capital Losses - Grouped into buckets of None, 1_to_1000, 1001_to_2000, more_than_2k

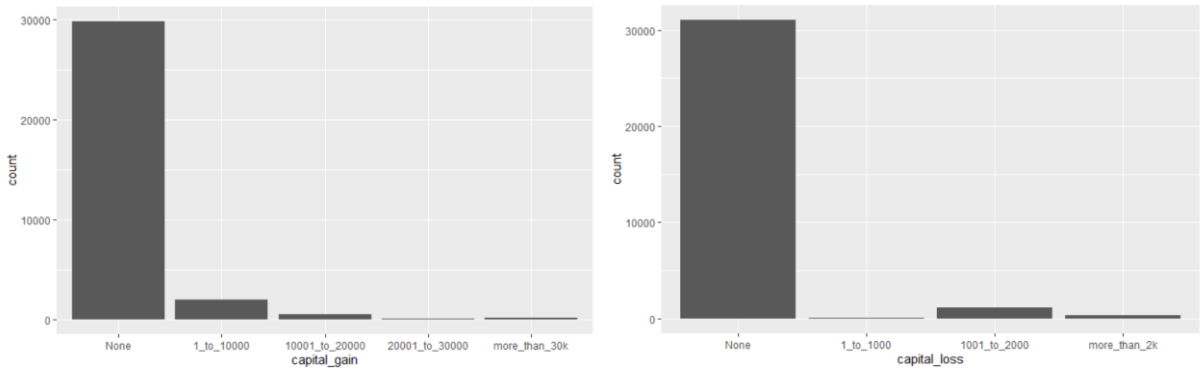

Figure 7 and 8 – Capital Gains and Capital Loss transformed to a Categorical Variable

Figure 7 and **8** display the transformation of Capital Gains and Capital Loss variable into a Categorical variable.

Different Types of Categorical Explanatory Variables

There are two different types of categorical variables, unordered and ordered. There are 4 categorical explanatory variables that need to be ordered since they have a higher sequence of order.

Ordered Categorical Variables:

- Capital Gains
- Capital Loss
- Education_num
- Education

Capital Gains, Capital Loss, Education_num, and Education were transformed into ordered categorical variables based on their sequence of order.

Updated Dataset

```
'data.frame':   32561 obs. of  15 variables:
 $ fnlwgt        : num  16.2 16.3 17.7 17.8 18.4 ...
 $ age           : num  39 50 38 53 28 37 49 52 31 42 ...
 $ hours_per_week: num  40 13 40 40 40 40 16 45 50 40 ...
 $ capital_gain  : Ord.factor w/ 5 levels "None"<"1_to_10000"<..: 2 1 1 1 1 1 1 1 3 2 ...
 $ capital_loss  : Ord.factor w/ 4 levels "None"<"1_to_1000"<..: 1 1 1 1 1 1 1 1 1 1 ...
 $ education_num : Ord.factor w/ 16 levels "1"<"2"<"3"<"4"<..: 13 13 9 7 13 14 5 9 14 13 ...
 $ education     : Ord.factor w/ 16 levels " Preschool"<" 1st-4th"<..: 13 13 9 7 13 14 5 9 14 13 ...
 $ workclass     : Factor w/ 9 levels " Federal-gov",..: 7 6 4 4 4 4 4 6 4 4 ...
 $ marital_status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
 $ occupation    : Factor w/ 15 levels " Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...
 $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
 $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
 $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
 $ native_region : Factor w/ 8 levels " Central-America",..: 8 8 8 8 1 8 1 8 8 8 ...
 $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

Figure 9 -Updated Dataset Summary

Figure 9 displays all the explanatory variables and the response variable in the dataset. Capital_gain, Capital_loss, education_num, and education are ordered categorical variables.

Detecting Collinearity

It is important to detect if there is any collinearity between the explanatory variables before any statistical model.

A test of Goodman and Kruskal's tau measure for all pairs of explanatory variables was performed. This will display if there is any association between the explanatory variables.

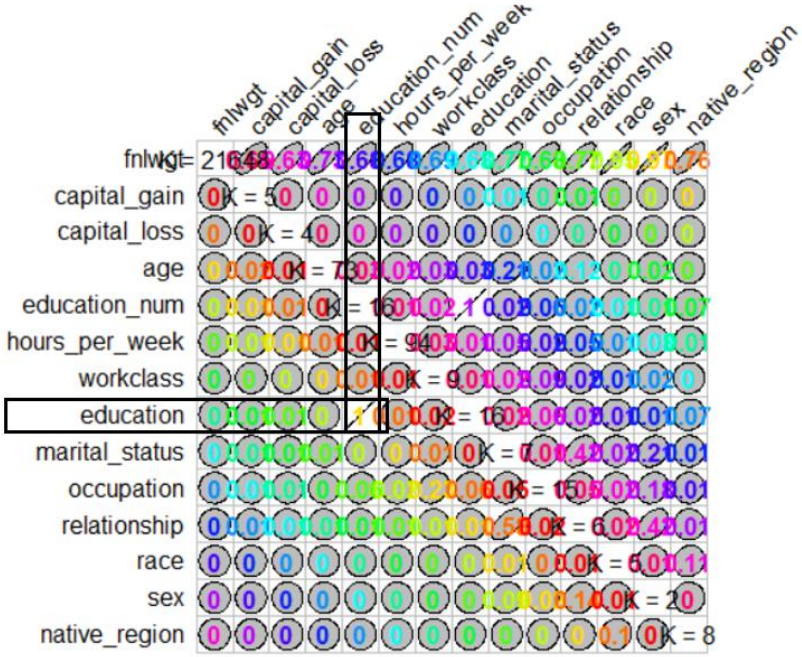

Figure 10 - Goodman and Kruskal's Matrix

Addressing Collinearity

There are two explanatory variables that need to be reviewed due to their high association with each other. **Figure 10** displays that two variables are education and education_num.

No surprise that they are associated with each other because they are almost the same explanatory variable.

In order to address this issue, we will drop the education_num from the dataset.

STATISTICAL MODELING

Logistic Regression Model

Logistic is used to model the probability of a certain class or event such as a pass/fail, win/lose, alive/dead or yes/no. Logistic Regression is a statistical model that uses a logistic function to model a binary dependent variable.

Below in **Figure 11** is a visual representation of the Logistic Regression model in context of this dataset.

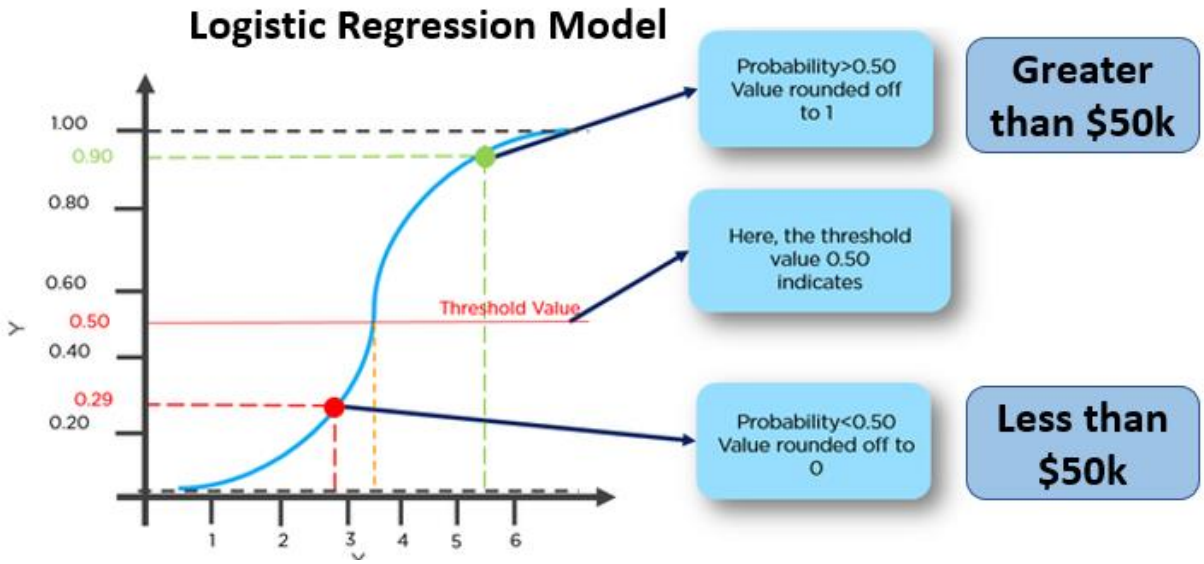

Figure 11 - Visual Representation of Logistic Regression Model

Generalized linear models (G.L.M.) with family Functional binomial (link = “logit”) will be modeled in R.

Below is the first model in the entire dataset.

```
mod1_1 <- glm(
    income ~ fnlwgt + capital_gain + capital_loss + age + education + hours_per_week +
    workclass + marital_status + occupation + relationship + race + sex + native_region,
    data=income, family=binomial(link="logit"))
```

Model Diagnostics

Goodness of Fit Test

A series of tests will be performed to study the appropriateness of the model.
To test the goodness of fit, we will review the null deviance and the deviance.

null.deviance = 27062.38
deviance = 16152.81

The null deviance shows how well the response variable is predicted by a model that includes only the grand mean where the deviance how well given all the observed response is within the given set of predictors. Residual deviance is significantly below null deviance. The means we pass our goodness of fit test.

Variable Significance

Checking to see if all the variables are important to the GLM model, a ANOVA test was performed on the model.

```
NULL
fnlwgt          0.804302
capital_gain   < 2.2e-16 ***
capital_loss   < 2.2e-16 ***
age            < 2.2e-16 ***
education      < 2.2e-16 ***
hours_per_week < 2.2e-16 ***
workclass      < 2.2e-16 ***
marital_status < 2.2e-16 ***
occupation     < 2.2e-16 ***
relationship   < 2.2e-16 ***
race            0.000126 ***
sex            < 2.2e-16 ***
native_region  2.335e-06 ***
```

Figure 12 – Analysis of Deviance Test

The ANOVA test in **Figure 12** displays that the fnlwgt variable is not statistically significant due to the high p-value. The fnlwgt variable will be dropped from the updated model.

Training and Test Datasets

In order to test the performance of the updated model, training and testing sets were cut from the dataset. The Training Set is 75% of dataset while the Testing Set is 25% of dataset.

Updated GLM Model with the Training Set

The training set was applied to the updated model.

```
mod1_2 <- glm(income ~ capital_gain + capital_loss + age + education + hours_per_week +
occupation + workclass + marital_status + relationship + race + sex + native_region,
data=training_set, family=binomial(link="logit"))
```

Model Performance

We can view the performance of the updated G.L.M. by displaying a confusion matrix against the testing set.

Figure 13 displays the confusion matrix from the update G.L.M. of the testing set.

```
        predicted
actual   FALSE TRUE
   <=50K  5503  675
   >50K    636 1289
```

Figure 13 – Confusion Matrix

We can see from the table in **Figure 13** that the updated G.L.M. has 5503 true positives and 1280 true negatives. The model has 675 of false positive and 636 false negatives.

The accuracy of the model is 83.8%.

CONCLUSION

To conclude this technical report, we will review and answer the questions of interest.

1. Are there explanatory variables that are highly correlated with each other? If so, which ones?

   The variables **education** and **education_num** are highly correlated with each other. This is no surprise since both variables are almost identical.

2. Which explanatory variables are not statistically significant to the income predictor?

   The **fnlwgt** variable was not statistically significant to the income predictor.

3. How accurate is the Generalized Linear Model predict “greater than 50k” or “less than equal to 50k” binary response variable?

   The Generalized Linear Model is **83.8% accurate** against the testing dataset.

FUTURE WORK

Next step would be to see how well the Generalized Linear Model performs against other classification models with the training and test datasets. I plan on exploring other classification models such as a decision tree, support vector machine, and random forest.