ST 595
Winter 2020
2/9/20

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 $50k.

Objectives

Objectives for this analysis are listed below:

- Determine if sex expanatory variable is a significant predictor of the income response variable.

- Determine if native region explanatory variable is a significant predictor of the income response variable.

- Which other explanatory variables are not statistically significant to the binary income response variable?

Data Description

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

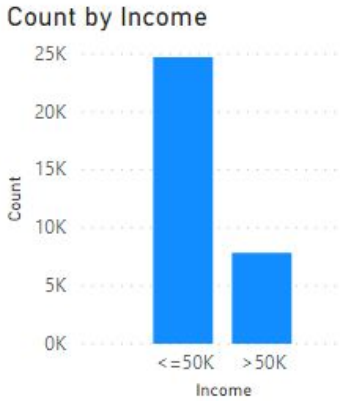

Figure 2 - Distribution of Income Response Variable

Transformation of Explanatory Variables
Since all the two different types of explanatory variables, all the explanatory variables need to be segmented into two different classifications.

Variables Types:

- Category Variables
- Continuous Variables

Based on the key callouts above, I will be creating a new adjusted dataset. Below are the two sets we will be evaluating are 1. Original Dataset - Consists of ~32560 with 15 variables 2. Adjusted Dataset - Reduced to ~30160 with 12 variables
Below are details on the adjusted dataset.

Variables Removed
There are a few variables that are either redundant or unique that need to be reviewed to see if the variables need to be removed from the dataset. I removed relationship, and education_num due to either the variables being represented in other variables. I also removed fnlwgt since every value is unique.

Grouping of Explanatory Variables

Grouped Categorical Variables
Grouped Native Regions
All the Native Regions have been group has been converted into a country based on geography. The new Native
Regions are Asia_East, Asia_Central, Central_America, South_America, Europe_West, Europe_East.

Grouped Continuous Variables
There are a few continuous variables that I grouped into buckets.

Hours worked per week - Grouped into buckets of 1_to_20, 21_to_40, 41_to_50, 51_to_60, 61_to_70 and more_than_71

Captial Gains - Grouped into buckets of none, 1_to_10000, 20001_to_30000, 30001_to_40000, and 40001_to_50000

Capital Losses - Grouped into buckets of none, 1_to_1000, 2001_to_3000, 3001_to_4000, 4001_to_5000, and more_than_5k

In Figure 3 below is a summary of the adjusted dataset.

```
## 'data.frame':    30162 obs. of  12 variables:
##  $ age           : int  39 50 38 53 28 ...
##  $ workclass     : Factor w/ 8 levels " Federal-gov",..: 7 6 4 4 4 ...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 ...
##  $ 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 ...
##  $ 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  : chr  "20001_to_30000" " none" ...
##  $ capital_loss  : chr  " none" " none" ...
##  $ hours_per_week: chr  " 21_to_40" " 1_to_20" ...
##  $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 ...
##  $ native_region : Factor w/ 8 levels " Central-America",..: 8 8 8 8 1 ...
##  - attr(*, "na.action")= 'omit' Named int  15 28 39 52 62 ...
##   ..- attr(*, "names")= chr  "15" "28" ...
```

Figure 3 - Adjustec Dataset Summary

Statistical Modeling

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 is a visual representation of the Logistic Regression model incontext of this dataset.

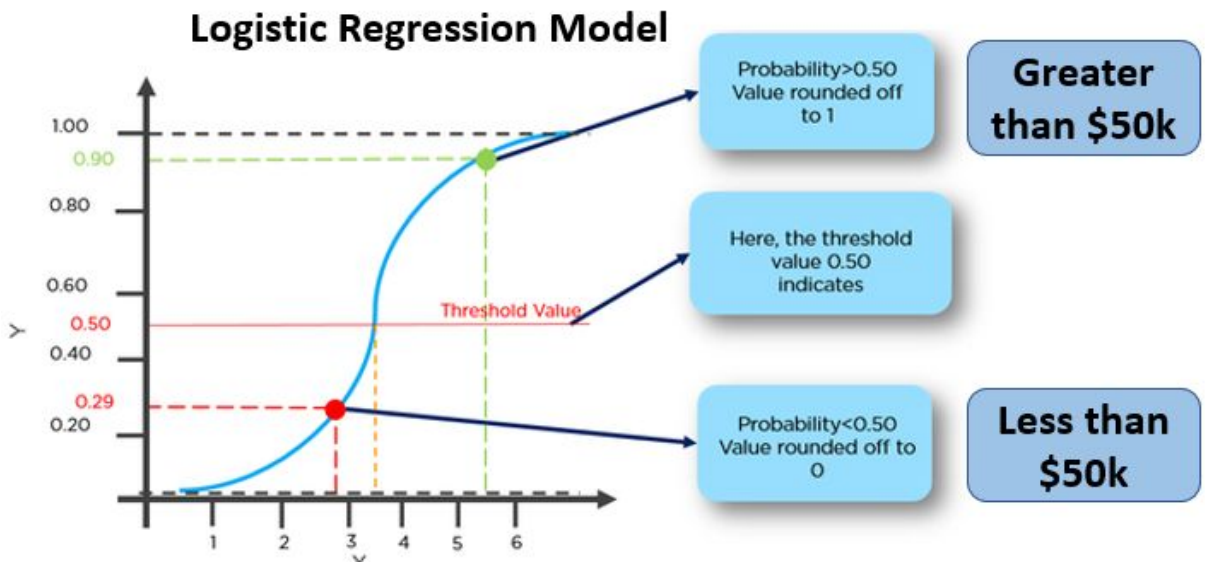

Figure 4 - Visual Representation of Logistic Regression Model

Analytical Questions:

- ***Is the Sex variable statistically significant being a male making over $50 a year?***

- ***Which Native Regions are statistically significant making over $50 a year?***

- **Which other explanatory variables are not statistically significant factors to the income response variable?**

Evaluation Metric

The P-value is the metric that will be evaluated to determine if a variable is statistically significant. The P-value is the probability that the null hypothesis gives for a specific experimental result to happen. In short, a low p-value means a higher chance of the hypothesis being true. In context of this analysis, a low p-value means that the variable is statistically significant. If the p-value is below 0.05, it is statistically significant.

Results

- ***Is the Sex variable statistically significant to the income response variable?***

After running a Logistic Regression model with income response variable to the sex explanatory response variable, being the sex variable is statistically significant. In other words, it is statistically significant to the income response variable. The p-value recorded 2e-16 which is significantly below 0.05. The summary of the results is displayed in figure 5.

```
Call:
glm(formula = income ~ sex + native_region, family = binomial(link = "logit"),
    data = income)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0794  -0.8795  -0.4972  -0.2456   2.6166

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.39019    0.10822 -31.326  < 2e-16 ***
sex Male                   1.27788    0.03535  36.147  < 2e-16 ***
```

Figure 5 - Logistic Regression Model focused on Income and Sex Variables

- ***Which Native Regions are statistically significant to the income response variable?***

After running a Logistic Regression model with income response variable to the native region explanatory response variable, most native regions are statistically significant. In other words, the Native Regions **Central Asia, East Asia, East Europe, West Europe, Outlying US**, and **United States** variables are statistically significant to the income response variable due to the p-value being extremely low (significantly less than 0.05).
Living in **South America** is not statistically significant due to a high p-value of 0.803 (significantly greater than 0.05). The summary of the results is displayed in figure 6.

```
Call:
glm(formula = income ~ sex + native_region, family = binomial(link = "logit"),
    data = income)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0794  -0.8795  -0.4972  -0.2456   2.6166

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.39019    0.10822 -31.326  < 2e-16 ***
native_region Central-Asia     1.87739    0.20214   9.288  < 2e-16 ***
native_region East-Asia        1.47270    0.16719   8.808  < 2e-16 ***
native_region Europe-East      1.23293    0.28117   4.385 1.16e-05 ***
native_region Europe-West      1.70820    0.15122  11.296  < 2e-16 ***
native_region Outlying-US      1.59447    0.15619  10.209  < 2e-16 ***
native_region South-America   -0.09600    0.38433  -0.250    0.803
native_region United-States    1.36188    0.10477  12.999  < 2e-16 ***
---
```

Figure 6 - Logistic Regression Model focused on Income and Native Region

Which other explanatory variables are not statistically significant factors to the income response variable?

After running a Logistic Regression model with income response variable to all of the explanatory response variable, below are a summary of the explanatory variables that are statistically significant and the explanatory variables that are statistically significant to the income response variable with their p-values.

Variables that are statistically significant to the income response variable are listed below in Figure 7

Variables that are not Statistically Significant	P-Value
workclass Without-pay	0.94782
marital_status Married-spouse-absent	6.97E-01
marital_status Separated	9.03E-01
marital_status Widowed	0.77827
occupation Armed-Forces	0.4021
occupation Craft-repair	6.73E-01
occupation Transport-moving	0.11271
race Black	0.05899
race Other	9.08E-01
capital_gain none	5.60E-01
capital_gain20001_to_30000	0.54623
capital_loss 4001_to_5000	9.83E-01
capital_loss none	0.42512
capital_loss1001_to_2000	5.89E-01

Variables that are not Statistically Significant	P-Value
capital_loss2001	0.94353
capital_loss2001_to_3000	8.25E-01
capital_lossmore_than_5k	5.62E-01
native_region Central-Asia	0.91895
native_region East-Asia	0.71249
native_region Europe-East	4.20E-01
native_region Outlying-US	0.14876
native_region South-America	5.05E-02
education 11th	9.44E-01
education 12th	1.34E-01
education 1st-4th	0.31729
education 5th-6th	0.08054
education 9th	2.30E-01
education Preschool	9.21E-01

Figure 7 - Variables that are not statistically significant to the income response variable

Conclusion

About 20% of the factored variables in the dataset are not statistically significant.
There are 28 factored variables in this dataset that are not statistically significant out of a total of 69 factored variables. A list of the factored variables that are not statistically significant is displayed above in Figure 7.

Future Steps in this Analysis

After reviewing my final project and putting together this technical report, there are a few areas that will be the focus to continuing to develop this analysis.

- Review and plot the ROC Curve to show overall predictive powers and decision. This will help determine the weights of each variable that are statistically significant

- Further explain the capital loss income due to the variable not being statistically significant

- Build a Logistic Regression prediction model