# Report – CA2

|  |  |
| --- | --- |
| Module Title: | Data Mining |
| Code: | B9DA103 |
| Module Leader: | Bahman Honari |
| Student code: | 10543928 |
| Student name: | Tim Browning |
| Instruction for CA 2: | In the dataset provided, the first column is labelled as “Income” and is the column to be imputed.  Other variables are the potential predictors in the models explained below.  The first step is to determine the missing values of “Income” that are actually equal to 0. To do this,  you are required to define a new variable, say “Binary\_Income” with 0, 1 and missing values, for the  cases which have 0, non-zero and missing values in the “Income” variable, respectively. Then use the  cases with “Binary\_Income” equal to either 0 or 1, to develop a logistic regression model. The  outcome of this model would allow you to predict the missing values in the “Binary\_Income”  variable, as 0 or 1. For any case which “Binary\_Income” is predicted as 0, put the actual “Income”  equal to 0 as well.  Those cases who their “Binary\_Income” is predicted as 1, are used along with other cases with nonzero “Income” to develop a linear regression model. The outcome of such model would allow you to  predict the missing values for the “Income” variable, for the cases which their “Binary\_Income”  variable is predicted as 1 and therefore are expected to have non-zero “Income”.  Report:  - The results for both logistic and linear regression models.  - The number of cases with predicted “Binary\_Income” as either 0 or 1.  - The Mean and SD for the non-zero predicted “Income” |

Contents

[Assessment Brief 1](#_Toc38718893)

[Part 1: The first step is to determine the missing values of “Income” that are actually equal to 0. To do this, you are required to define a new variable, say “Binary\_Income” with 0, 1 and missing values, for the cases which have 0, non-zero and missing values in the “Income” variable, respectively. 3](#_Toc38718894)

[\*Part 2: Then use the cases with “Binary\_Income” equal to either 0 or 1, to develop a logistic regression model. (Report P1) 4](#_Toc38718895)

[Part 3: The outcome of this model would allow you to predict the missing values in the “Binary\_Income” variable, as 0 or 1. 4](#_Toc38718896)

[Part 4: For any case which “Binary\_Income” is predicted as 0, put the actual “Income” equal to 0 as well. 5](#_Toc38718897)

[\*Part 5: Those cases who their “Binary\_Income” is predicted as 1, are used along with other cases with nonzero “Income” to develop a linear regression model. (Report P1) 5](#_Toc38718898)

[Part6: In R file 6](#_Toc38718899)

[\*Report 2: The number of cases with predicted “Binary\_Income” as either 0 or 1. 6](#_Toc38718900)

[\*Report 3: The Mean and SD for the non-zero predicted “Income”. 6](#_Toc38718901)

## Part 1: The first step is to determine the missing values of “Income” that are actually equal to 0. To do this, you are required to define a new variable, say “Binary\_Income” with 0, 1 and missing values, for the cases which have 0, non-zero and missing values in the “Income” variable, respectively.

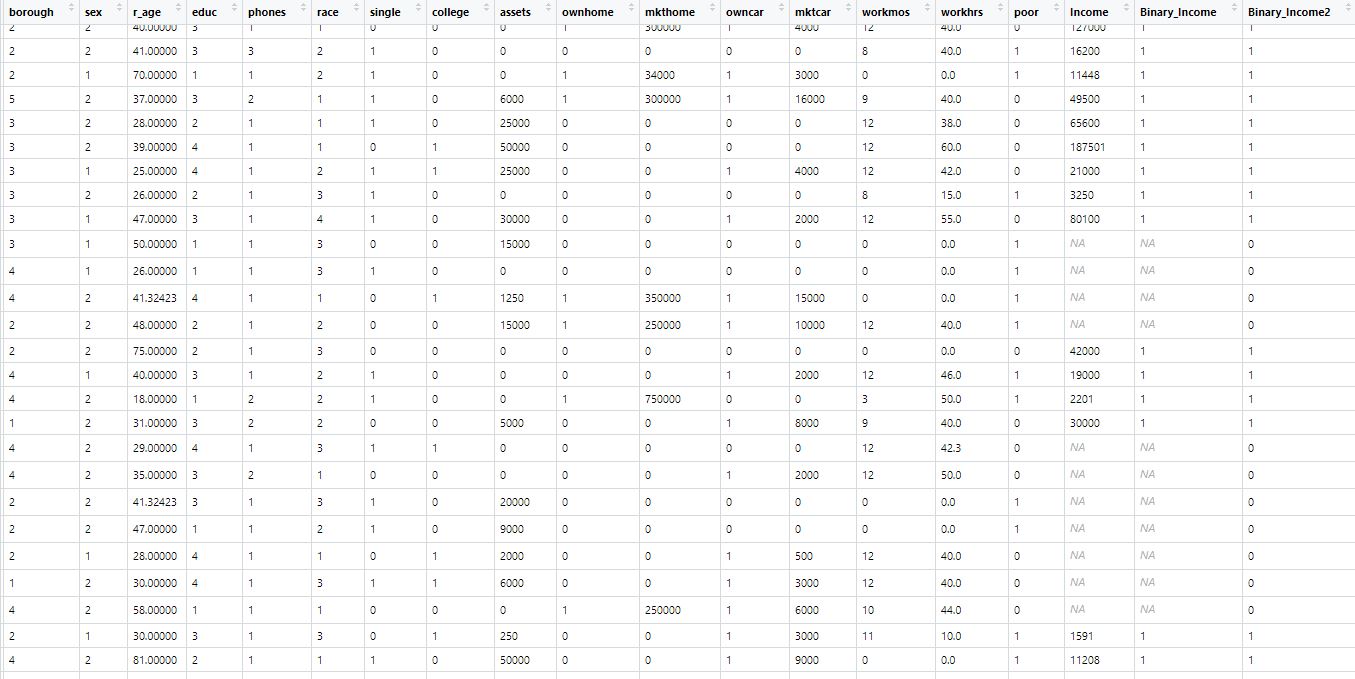


Figure 1 - New Dataset with Columns Binary\_Income & Binary\_income2

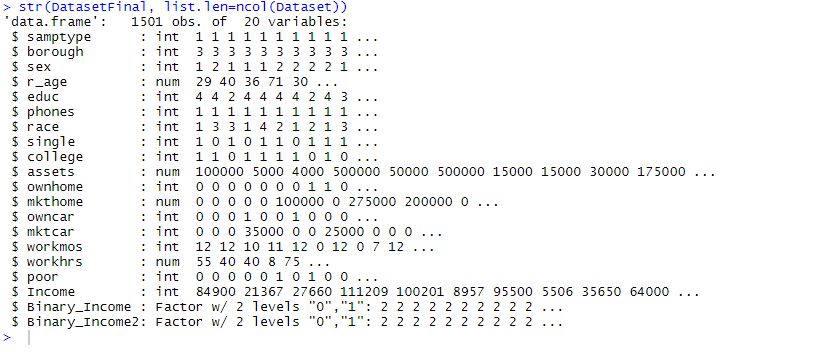


Figure 2 - STR on Dataset

## \*Part 2: Then use the cases with “Binary\_Income” equal to either 0 or 1, to develop a logistic regression model. (Report P1)

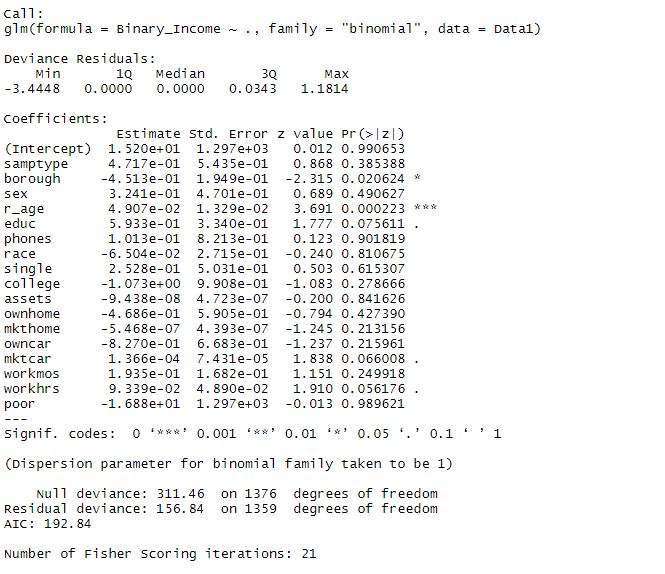


Figure 3 - Report 1 – GLM

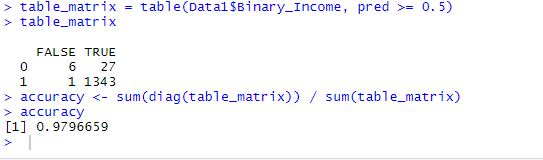
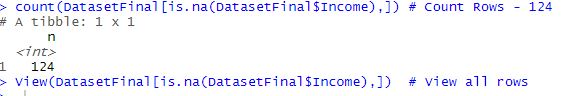
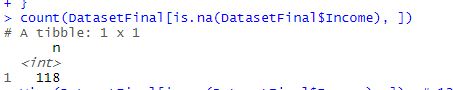


Figure 4 - Accuracy of GLM

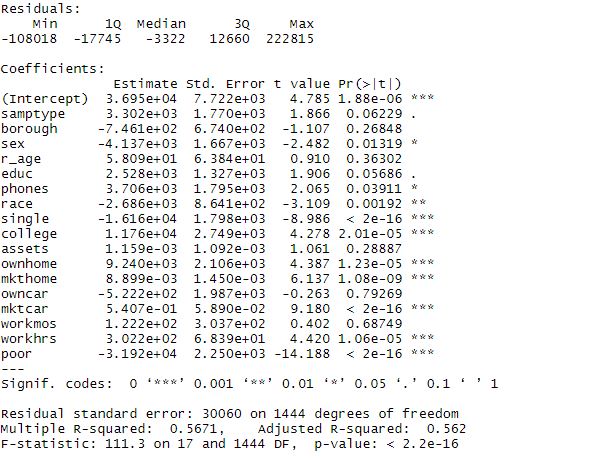
## Part 3: The outcome of this model would allow you to predict the missing values in the “Binary\_Income” variable, as 0 or 1.



## Part 4: For any case which “Binary\_Income” is predicted as 0, put the actual “Income” equal to 0 as well.

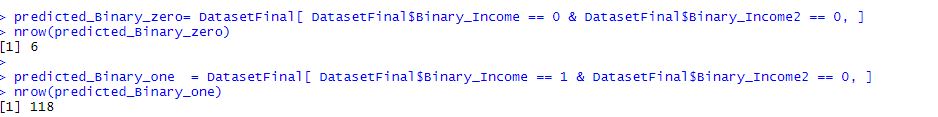


## \*Part 5: Those cases who their “Binary\_Income” is predicted as 1, are used along with other cases with nonzero “Income” to develop a linear regression model. (Report P1)



## Part6: In R file

## \*Report 2: The number of cases with predicted “Binary\_Income” as either 0 or 1.



## \*Report 3: The Mean and SD for the non-zero predicted “Income”.

