Classification Using Logistic Regression

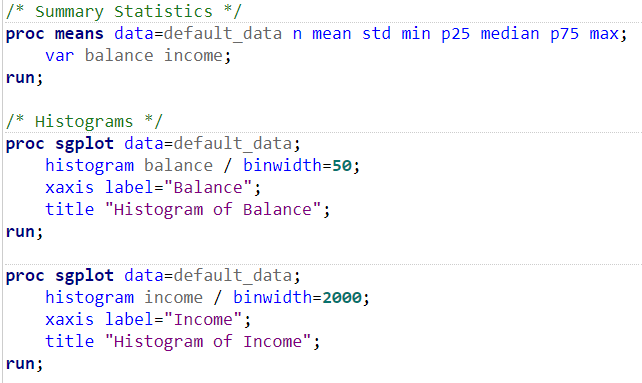
Hector J. Rodriguez-Mondragon

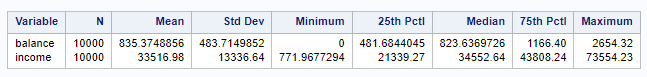
12/1/2024

# Default.csv

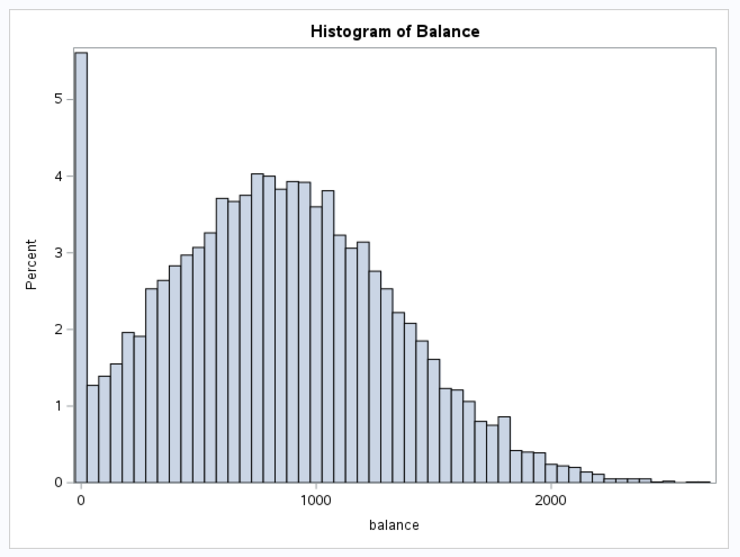
In this paper I will analyze summary statistics and histograms for balance and income based on combinations of default and student. Afterwards the logistic regression model will be done. The default.csv is set as the response variable while predictors will be set as student, balance, and income. Lastly, I will use the model to calculate the likelihood of default for a student with a balance of $621 and an income of $1,850.

## Summary Statistics

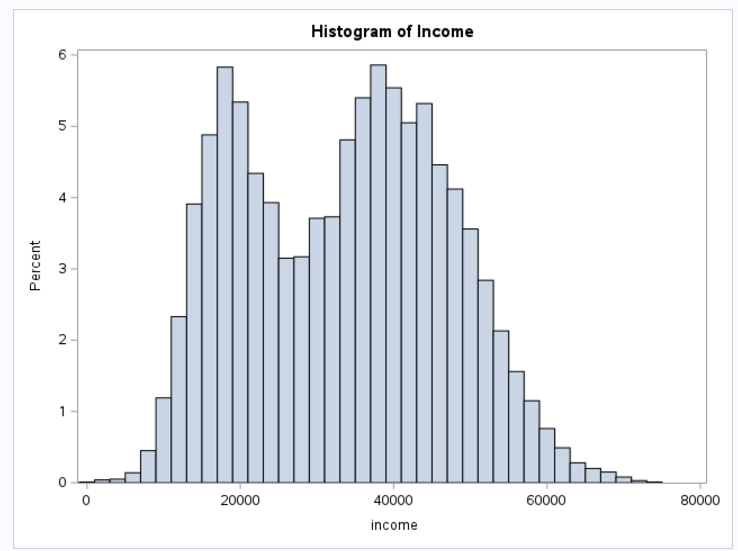




In the summary we can see the balance average is $835.37 with std deviation of $483.71. The ranges go from $0-2,654.32. As for income, the range is from $771.97-73,554.23. The average income is $33,516.98 with a std deviation of $13,336.64.



*Figure 1: Uniform balance histogram distribution with a clustering in the mean area.*



*Figure 2: Income Histogram showing a skewed distribution with a concentration around lower income levels*

## Assessment of four combinations of default and student

/\* Grouped Statistic Summaries and Histograms \*/

proc means data=default\_data n mean std min p25 median p75 max;

class default student;

var balance income;

run;

/\* Default = Yes, Student = Yes \*/

proc sgplot data=default\_data;

where default="Ye" and student="Yes";

histogram balance / binwidth=50;

xaxis label="Balance";

title "Histogram of Balance (Default=Yes, Student=Yes)";

run;

proc sgplot data=default\_data;

where default="Ye" and student="Yes";

histogram income / binwidth=2000;

xaxis label="Income";

title "Histogram of Income (Default=Yes, Student=Yes)";

run;

/\* Default = Yes, Student = No \*/

proc sgplot data=default\_data;

where default="Ye" and student="No";

histogram balance / binwidth=50;

xaxis label="Balance";

title "Histogram of Balance (Default=Yes, Student=No)";

run;

proc sgplot data=default\_data;

where default="Ye" and student="No";

histogram income / binwidth=2000;

xaxis label="Income";

title "Histogram of Income (Default=Yes, Student=No)";

run;

/\* Default = No, Student = Yes \*/

proc sgplot data=default\_data;

where default="No" and student="Yes";

histogram balance / binwidth=50;

xaxis label="Balance";

title "Histogram of Balance (Default=No, Student=Yes)";

run;

proc sgplot data=default\_data;

where default="No" and student="Yes";

histogram income / binwidth=2000;

xaxis label="Income";

title "Histogram of Income (Default=No, Student=Yes)";

run;

/\* Default = No, Student = No \*/

proc sgplot data=default\_data;

where default="No" and student="No";

histogram balance / binwidth=50;

xaxis label="Balance";

title "Histogram of Balance (Default=No, Student=No)";

run;

proc sgplot data=default\_data;

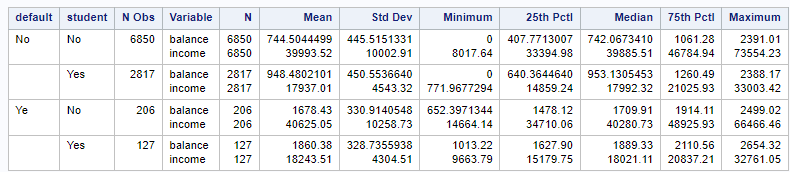
where default="No" and student="No";

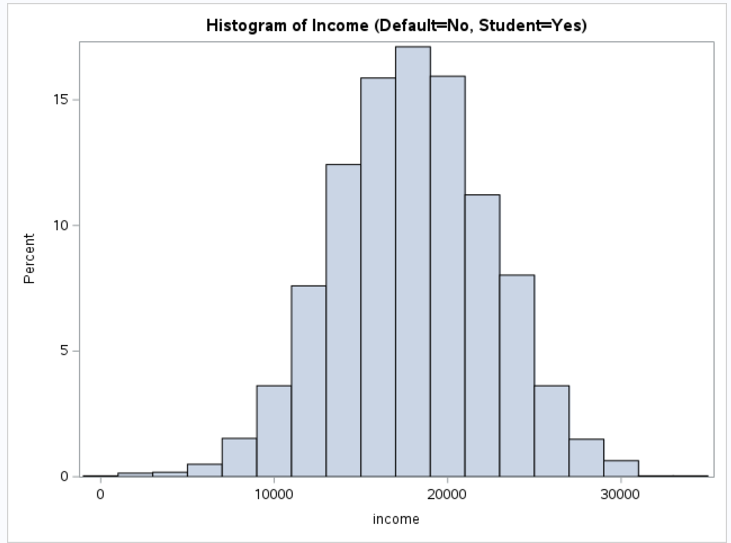
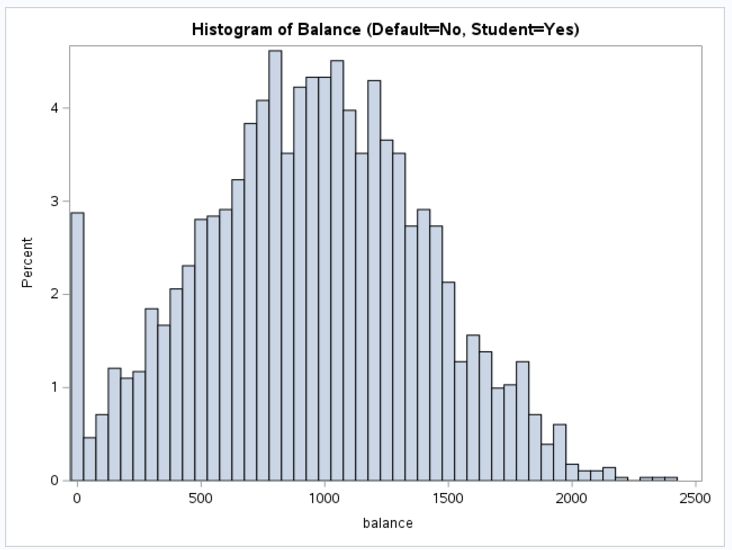
histogram income / binwidth=2000;

xaxis label="Income";

title "Histogram of Income (Default=No, Student=No)";

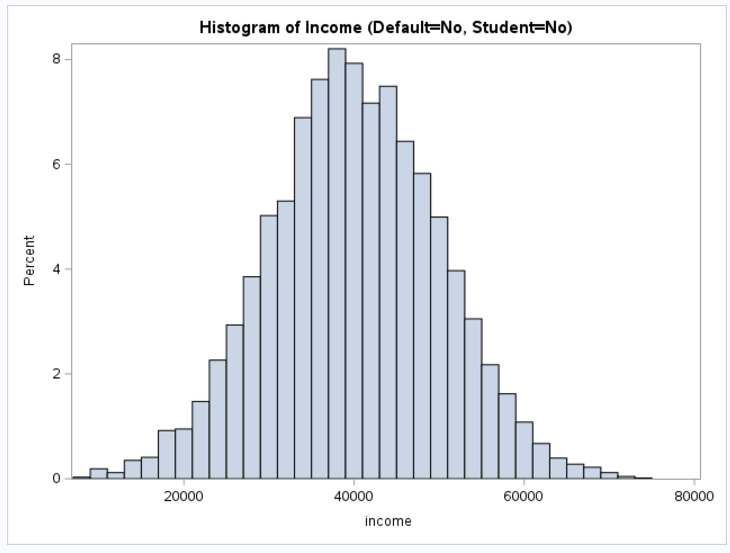
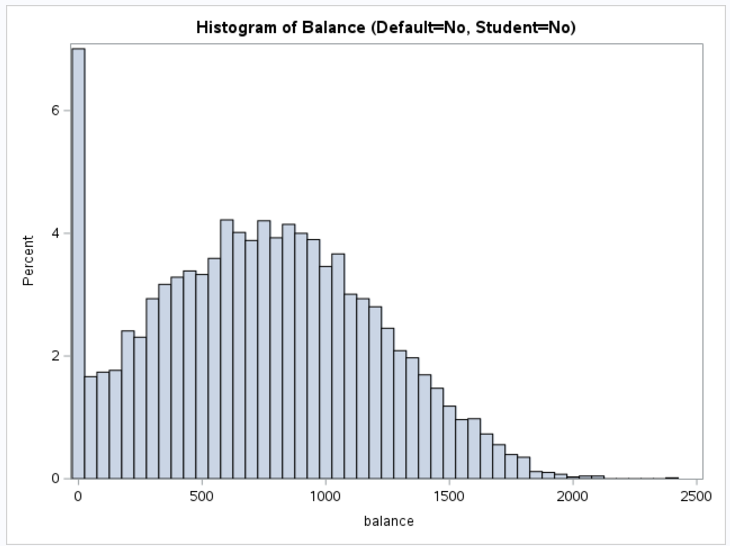
run;





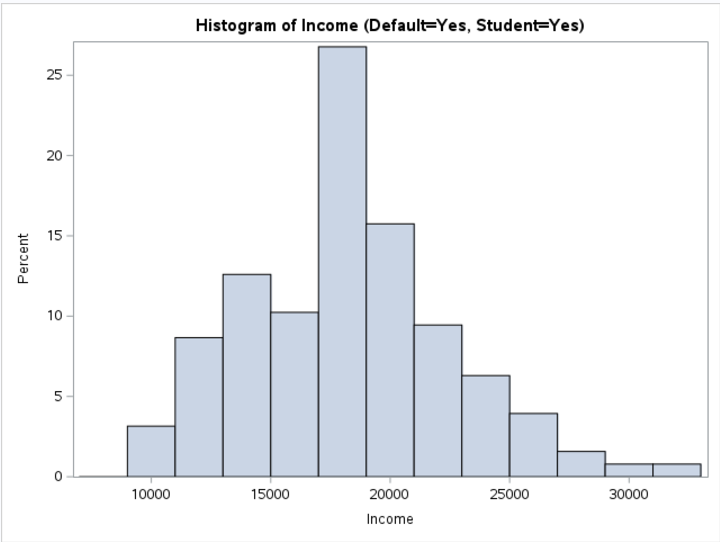
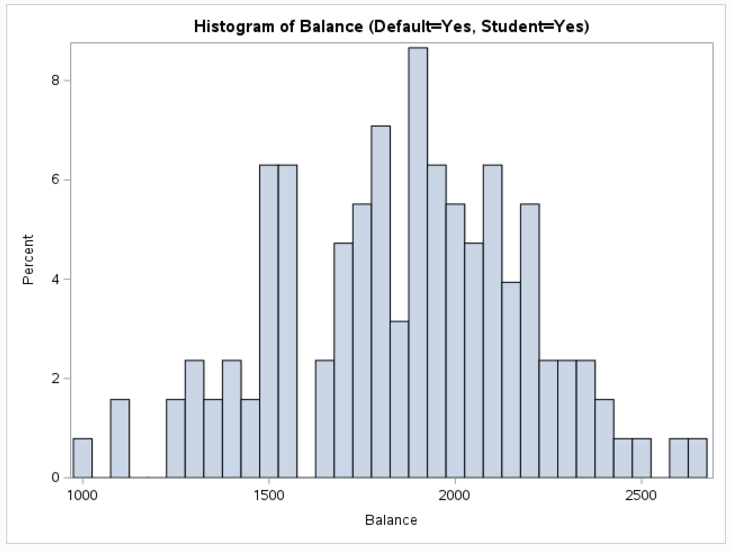
Students in this category have a moderate average balance of $948.48 and a lower average income of $17,937.01 compared to non-students. The histograms reveal that most balances are below $1,200, while incomes are clustered primarily below $30,000, reflecting financial constraints among students who do not default.

## WHERE Default=No

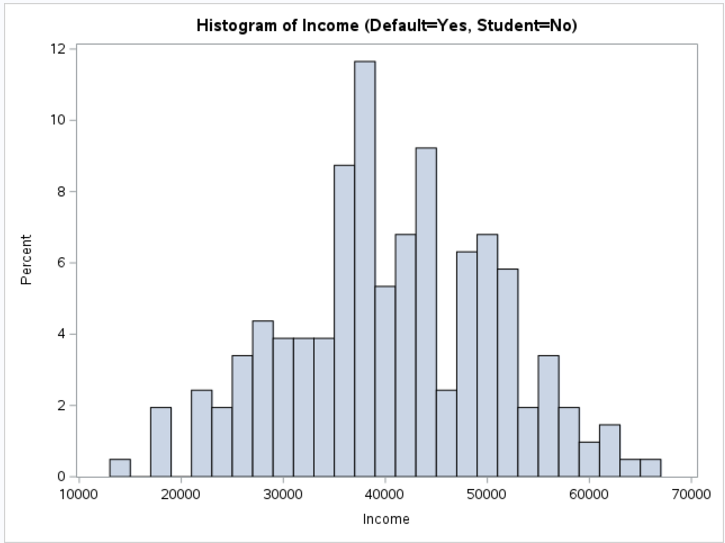
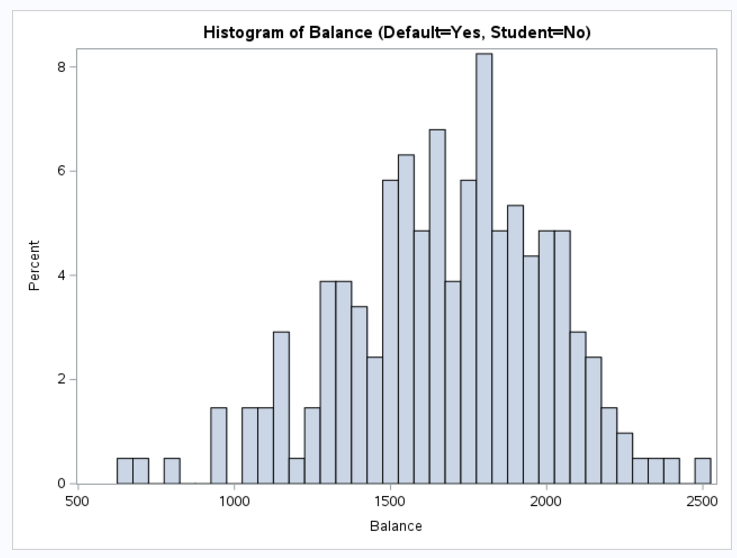


This group has the lowest average balance from all others at $744.50, with the highest average income of $39,993.52. The incomes are distributed in bell formation, showing better financial stability and management among non-students who do not default.

## WHERE Default=Yes



“Yes,yes” group has the highest average balance at $1,860.38, significantly higher than other groups. However, it also has the lowest average income of $18,243.51, telling us that students with high balances are more likely to default.



The average balance for this group is $1,678.43, which is lower than the "Yes, Yes" seen above. Despite this, their average income is higher at $40,625.05, suggesting that non-students who default have better financial resources but still struggle with repayment.

**Logistic Regression Model**

/\* Logistic Regression \*/

data default\_data\_binary;

set default\_data;

default\_binary = (default = "Ye");

student\_binary = (student = "Yes");

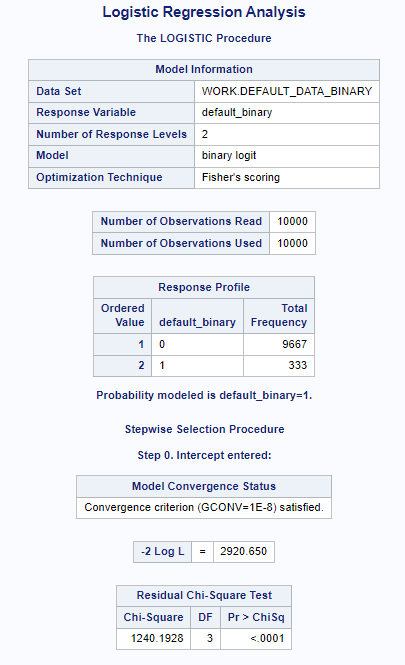
run;

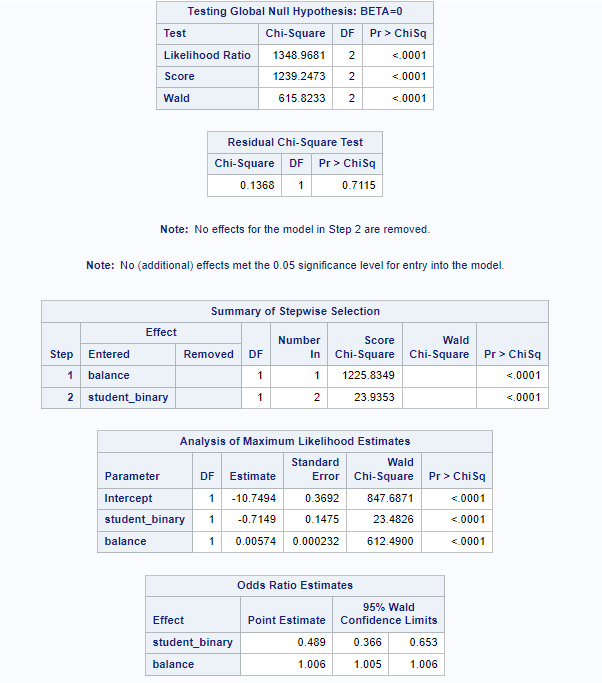
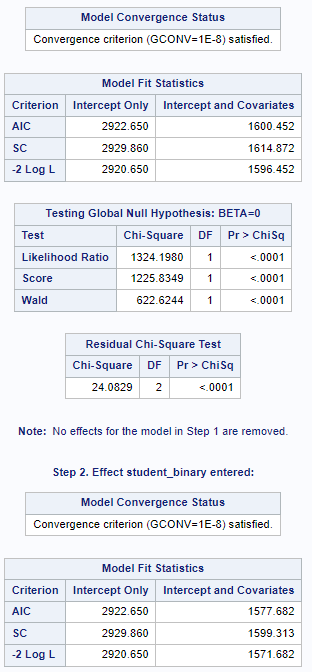
proc logistic data=default\_data\_binary outmodel=model\_store;

model default\_binary(event='1') = student\_binary balance income / selection=stepwise;

title "Logistic Regression Analysis";

run;





From 1000, 333 had defaulted on their loans, while 9,667 had not. First, the variable balance (the amount owed) was included. This significantly improved the model, showing that higher balances increase the chance of default. Student\_binary helped with showing students are less likely to default compared to non-students.

# Results

/\* Example Prediction \*/

data predict;

input student\_binary balance income;

datalines;

1 621 1850

;

run;

proc logistic inmodel=model\_store;

score data=predict out=predicted;

run;

proc print data=predicted;

title "Predicted Default Probability";

run;

Balance: For each additional unit of balance, the odds of defaulting slightly increase.

Student Status: Students are less likely to default than non-students.

As an example, a student with a balance of 621 and an income of 1,850 was predicted to have a very low probability of defaulting (about 0.037%). This matched the actual result, as the person did not default.