**Logistic Regression and Bootstrap Method for Real Data Analysis**

R Script:

1. Load the Dataset and Data Processing:

A screenshot of a computer program

Description automatically generated

**Output:**

# Load the "Default" dataset  
data("Default", package = "ISLR")  
  
# Store the Default dataset in a dataframe  
default\_df <- Default  
  
# Display the first few rows of the dataframe  
head(default\_df)

## default student balance income  
## 1 No No 729.5265 44361.625  
## 2 No Yes 817.1804 12106.135  
## 3 No No 1073.5492 31767.139  
## 4 No No 529.2506 35704.494  
## 5 No No 785.6559 38463.496  
## 6 No Yes 919.5885 7491.559

# Data Processing:  
# Check the structure of the Default dataset  
str(Default)

## 'data.frame': 10000 obs. of 4 variables:  
## $ default: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ student: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 2 1 1 ...  
## $ balance: num 730 817 1074 529 786 ...  
## $ income : num 44362 12106 31767 35704 38463 ...

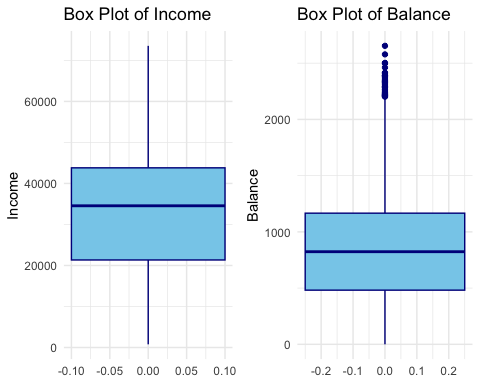
# Check summary statistics for the balance column  
summary(Default)

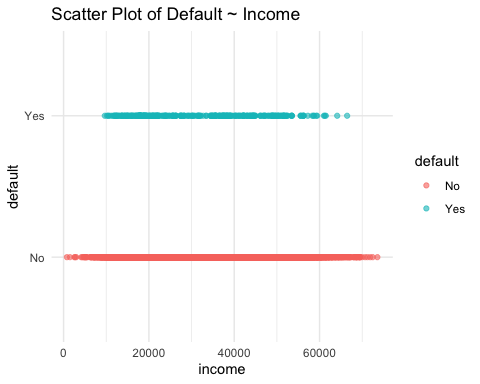
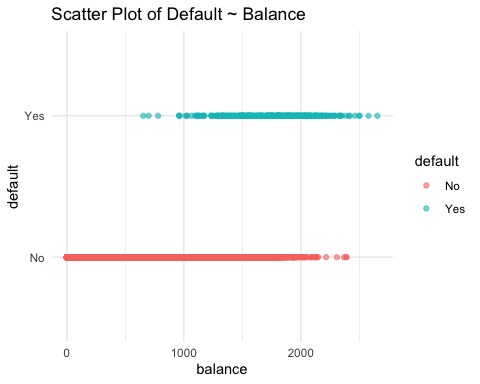
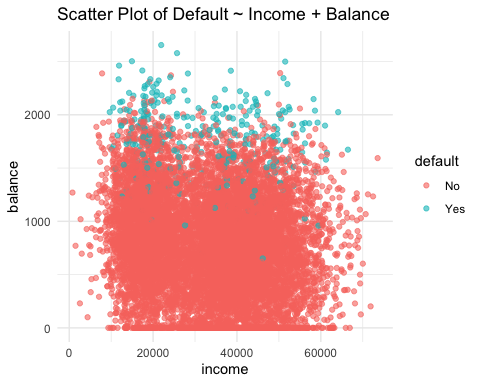
## default student balance income   
## No :9667 No :7056 Min. : 0.0 Min. : 772   
## Yes: 333 Yes:2944 1st Qu.: 481.7 1st Qu.:21340   
## Median : 823.6 Median :34553   
## Mean : 835.4 Mean :33517   
## 3rd Qu.:1166.3 3rd Qu.:43808   
## Max. :2654.3 Max. :73554

1. A screenshot of a computer code

   Description automatically generatedData processing and Visualization:

**Output:**





1. Data Visualization and Logistic Regression Analysis:

A screenshot of a computer code

Description automatically generated

**Output:**

Model Summary Table:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Model Deviance | AIC | BIC |
| Income | 2916.687 | 2920.687 | 2935.108 |
| Balance | 1596.452 | 1600.452 | 1614.872 |
| Income + Balance | 1578.966 | 1584.966 | 1606.597 |

1. Logistic Regression Analysis Visualization:

A screenshot of a computer program

Description automatically generated

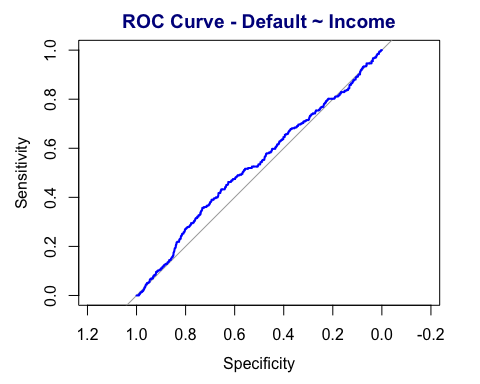
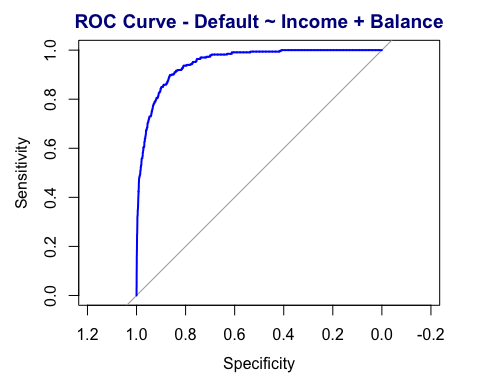
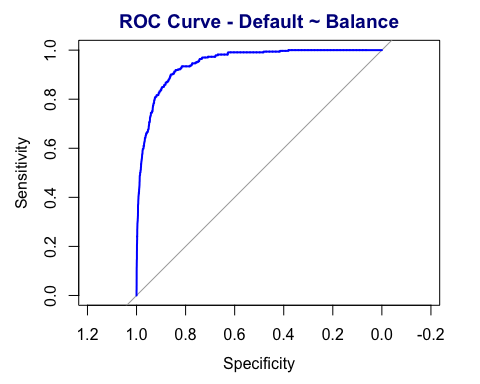
1. Interpretation of Logistic Regression:

A computer screen shot of text

Description automatically generated

**Output:**

**Figures:**



**Figure 1:** Shows ROC Curve for the Logistic Regression Model. (a) Shows ROC Curve for Default as a dependent vs Income as an independent variable. (b) Shows ROC Curve for Default as a dependent vs Balance as an independent variable. (c) Shows ROC Curve for Default as a dependent vs Income and Balance as two independent variables (Multiple Logistic regression). The Graph shows the ratio of sensitivity and specificity. Sensitivity represents the ability of the test to correctly identify positive cases. The higher the sensitivity, the lower the rate of false negatives. **1 - Specificity (False Positive Rate),** represents the probability of a false positive, indicating how often the test incorrectly identifies negative cases as positive. The farther the ROC curve is from the diagonal, the better the diagnostic performance.

(a)

(b)

(c)

**Discussion:**

The Default dataset is readily available in the ISLR package in R programming. The dataset comprises selected variables and data points related to credit card users, and the primary objective is to predict which customers are likely to default on their credit card payments. Variables in the dataset include factors such as balance, income, and student status. The target variable is binary, indicating whether a customer defaulted on their credit card payment (Yes/No). Logistic regression, a subset of generalized linear models (GLMs), is widely applied in various fields for binary classification tasks. It extends the general linear model by incorporating a logit link function, facilitating the modeling of the relationship between independent variables and the log-odds of the binary response variable.

The summary table shows the results of a regression model with three different predictor variables: Income, Balance, and a combination of both (Income + Balance). The Deviance measures the goodness of fit of the model. Lower values indicate a better fit. The model with both Income and Balance together (Row 3) has the lowest deviance (i.e. 1578.966), suggesting the best fit among the models. AIC (Akaike Information Criterion)is a measure of the model's goodness while penalizing for the number of predictors. Lower AIC values indicate a better trade-off between fit and complexity. Again, the model with both Income and Balance (Row 3) has the lowest AIC (i.e. 1584.966), implying superior performance. Similarly, BIC (Bayesian Information Criterion) also penalizes model complexity. Lower BIC values indicate a better model. Once again, the model with both Income and Balance (Row 3) has the lowest BIC (i.e. 1606.597), signifying a more favorable balance between fit and complexity. Nevertheless, a comparison between income and balance models shows that the balance model is a better fit for predicting the outcome than the income model.

1. Continuation of #4:

A screenshot of a computer program

Description automatically generated

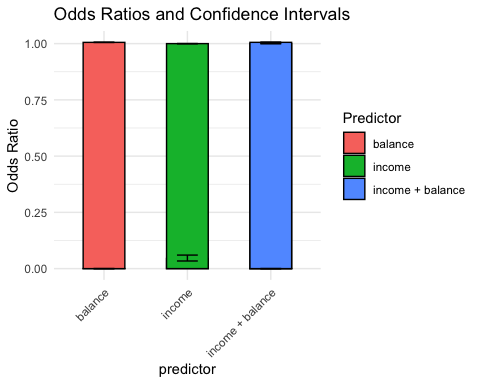
1. Continuation of #5:

A screenshot of a computer program

Description automatically generated

**Results:**

|  |  |  |
| --- | --- | --- |
| Variable | Coefficients | Odds\_ratios |
| Income | -8.352575e-06 | 0.9999916 |
| Balance | 5.498917e-03 | 1.0055141 |
| Income + Balance | 2.080898e-05 | 1.0000208 |



**Figure 2**: Graphical representation of odds\_ratios and Confidence Intervals. An odds ratio greater than 1 suggests an increase in the odds of the event. An odds ratio of less than 1 suggests a decrease in the odds of the event.

**Discussion:**

The coefficient for Income is negative, suggesting a decrease in the odds of the event (dependent variable) as Income increases. The odds ratio being slightly less than 1 reinforces this, indicating a small decrease in odds.The coefficient for Balance is positive, indicating an increase in the odds of the event as Balance increases. The odds ratio being slightly greater than 1 supports this, suggesting a small increase in odds.The combined effect of Income and Balance shows a positive coefficient, suggesting a joint increase in the odds. However, the odds ratio is very close to 1, indicating a minimal impact on the odds.This can also be observed in the visualized graph for odds\_ratios and confidence intervals. In summary, the logistic regression suggests that higher Balance is associated with increased odds, while higher Income is associated with slightly decreased odds. The combined effect of Income and Balance has a marginal impact on the odds.

1. Visualization of Logistic Regression and Bootstrapping Method:

A screenshot of a computer code

Description automatically generated

1. Boot Strap Method:

A computer screen shot of text

Description automatically generated

1. Statistical Results for BootStrap Method:

A screenshot of a computer code

Description automatically generated

1. Logistic Regression after Bootstrapping:

A screenshot of a computer code

Description automatically generated

**Results:**

|  |  |  |
| --- | --- | --- |
| Model | Lower\_CI | Upper\_CI |
| Income | -3.404395 | -2.805551 |
| Balance | -11.364361 | -9.967937 |
| Income + Balance | -12.479442 | -10.799186 |

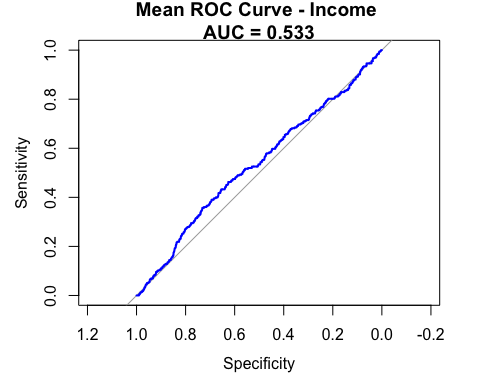
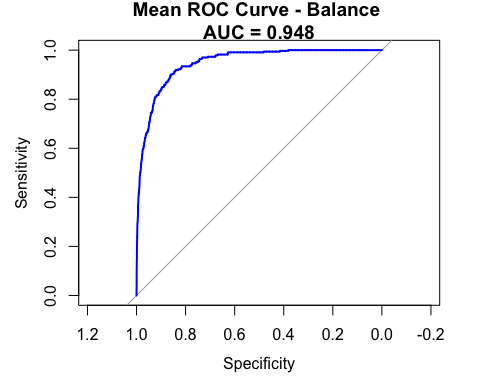
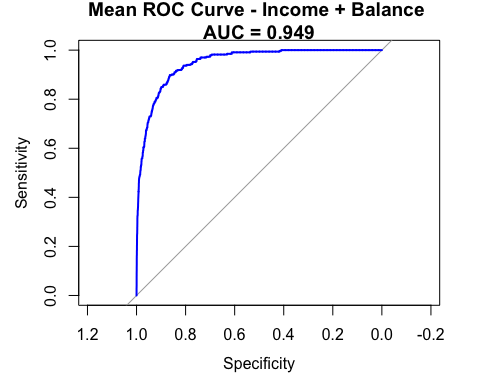
1. Visualization of Logistic Regression after Bootstrapping:

A screenshot of a computer code

Description automatically generated

**Output:**

**Figure 3:** Shows ROC Curve for the Logistic Regression Model after Bootstrapping. (a) Shows ROC Curve for Default as a dependent vs Income as an independent variable. (b) Shows ROC Curve for Default as a dependent vs Balance as an independent variable. (c) Shows ROC Curve for Default as a dependent vs Income and Balance as two independent variables (Multiple Logistic regression). The Graph shows the ratio of sensitivity and specificity. Sensitivity represents the ability of the test to correctly identify positive cases. The higher the sensitivity, the lower the rate of false negatives. **1 - Specificity (False Positive Rate),** represents the probability of a false positive, indicating how often the test incorrectly identifies negative cases as positive. The farther the ROC curve is from the diagonal, the better the diagnostic performance.



(a)

(b)

(c)

1. Comparison of Regression Analysis Before and After Bootstrapping:

A close up of text

Description automatically generated

1. Visualization:

A screen shot of a computer program

Description automatically generated

**Output:**

A graph with a bar and a bar chart

Description automatically generated with medium confidenceA graph of a bar

Description automatically generated

**Discussion:**

Bootstrapping is employed to estimate the variability of model parameters, quantify uncertainties in coefficient estimates, and enhance the accuracy of predictive distributions. By generating numerous resampled datasets from the original dataset through random sampling with replacement, bootstrapping allows for the construction of empirical confidence intervals for logistic regression coefficients and facilitates a more comprehensive understanding of the model's stability [3]. The result shows that the bootstrapped confidence interval (CI) for the coefficient of Income ranges from approximately -3.40 to -2.81. This suggests the true coefficient of Income lies within a 95% confidence interval. The bootstrapped CI for the coefficient of Balance ranges from around -11.36 to -9.97. Like Income, the negative values imply a negative relationship between Balance and the outcome. Also, the actual coefficient of Balance falls within a 95% confidence interval. The bootstrapped CI for the coefficient of the combined effect of Income and Balance is between approximately -12.48 and -10.80. Just like the individual variables, the negative values suggest a negative association with the outcome. The interval represents our 95% confidence in the true coefficient of the combined effect. These intervals provide a range of plausible values for the coefficients based on the bootstrapping method. The negative associations indicate that higher values of Income, Balance, or their combination are associated with a decrease in the log odds of the event happening, holding other variables constant.