Midterm 1 - Akshaj Kammari

02/19/2024

```r
library(readr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

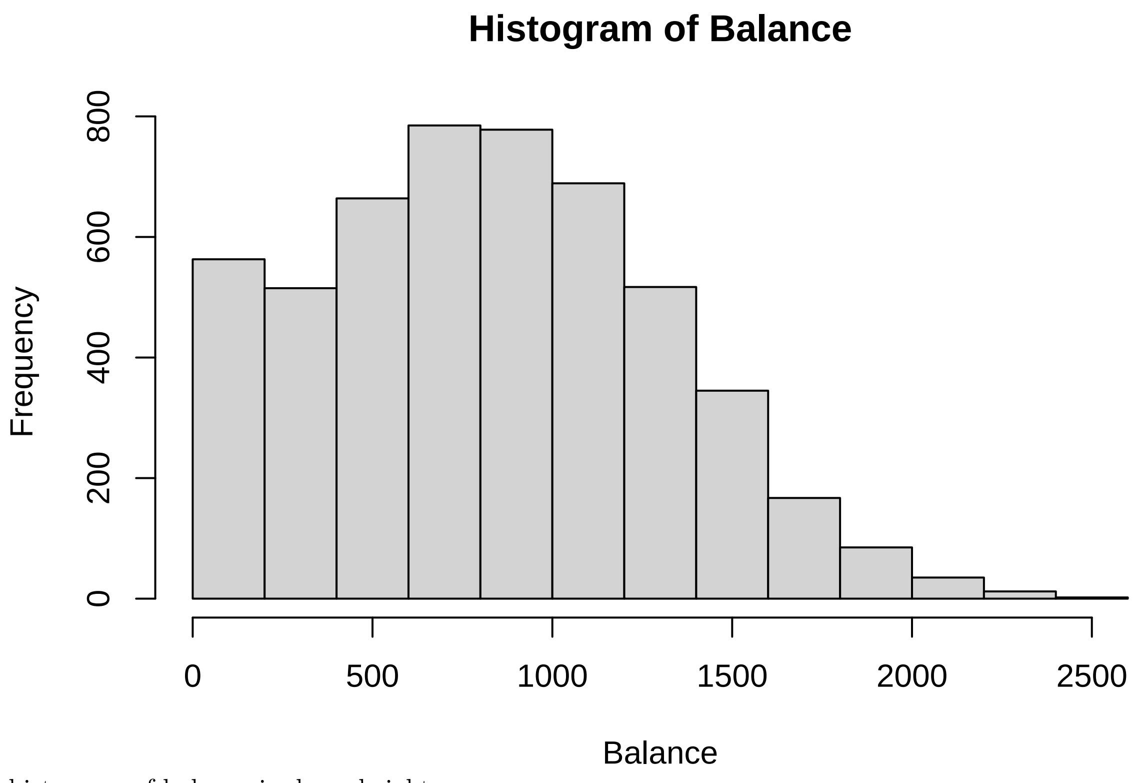
```r
library(moderndive)
library(epiDisplay)
```

```
## Loading required package: foreign

## Loading required package: survival

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: nnet

##
## Attaching package: 'epiDisplay'

## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```r
data <- read.csv("Default.csv")
```

1. #histogram of balance

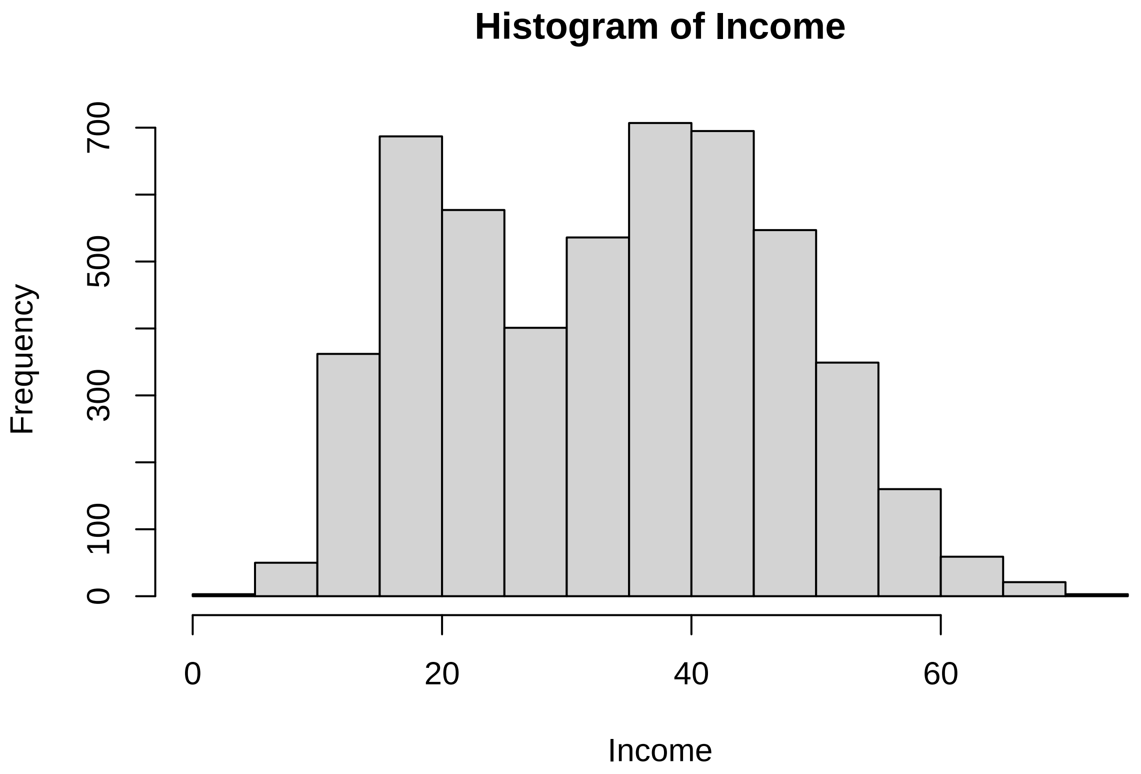
```r
hist(data$balance, main = "Histogram of Balance", xlab = "Balance")
```



##the histogram of balance is skewed right.

#histogram of income

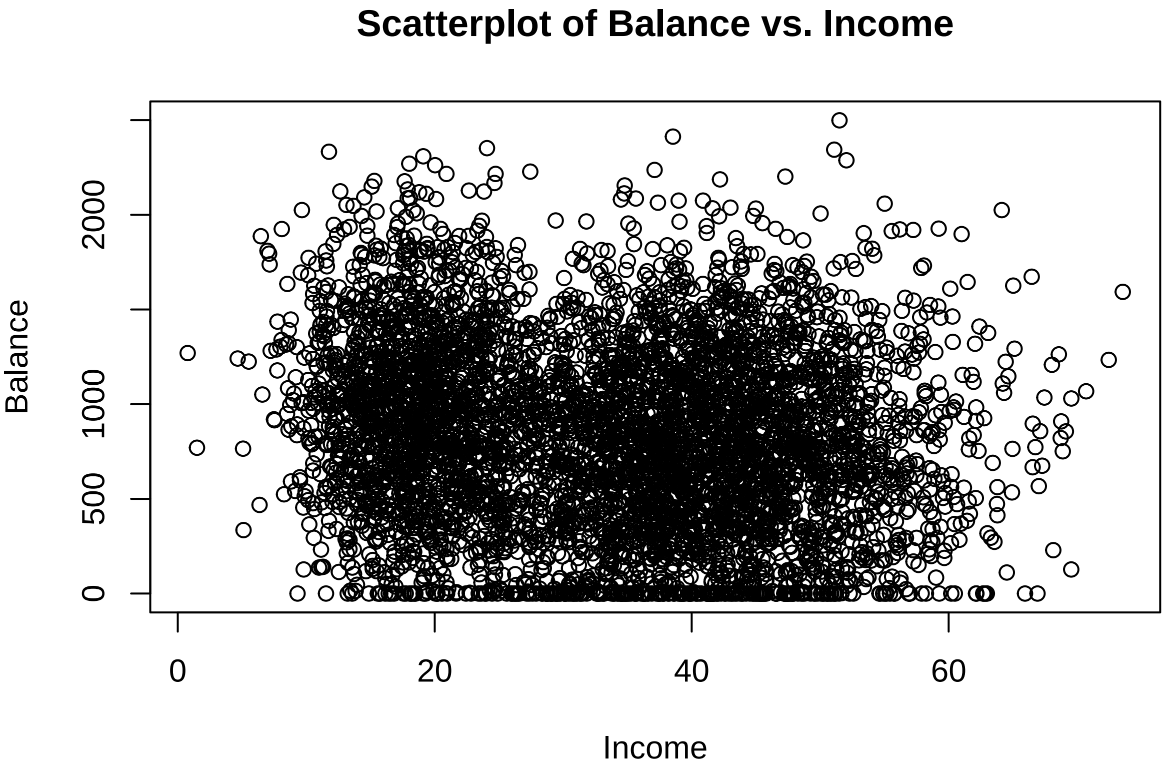
```r
hist(data$income, main = "Histogram of Income", xlab = "Income")
```



##the histogram of income seems to have a bell-curve

2.

```
plot(data$income, data$balance, xlab = "Income", ylab = "Balance", main = "Scatterplot of Balance vs. I
```



3.

```
correlation <- cor(data$balance, data$income)
correlation
```

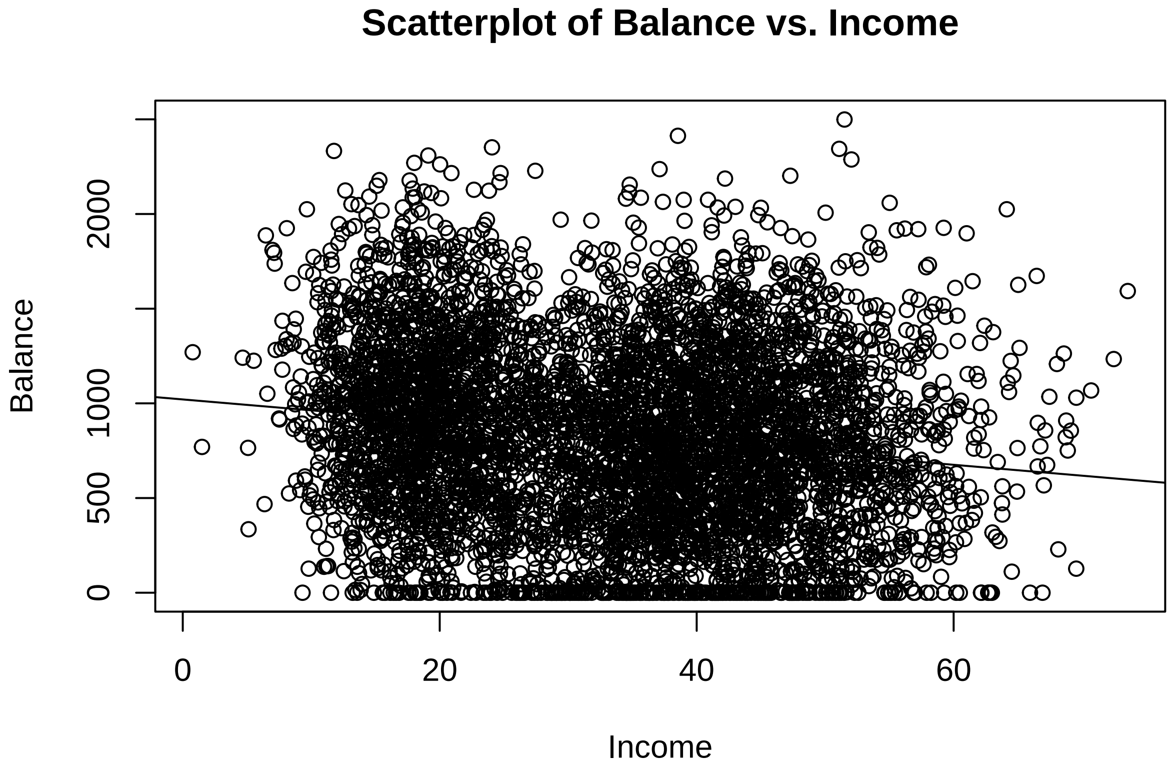
```
## [1] -0.1592327
```

4.

```
model <- lm(balance ~ income, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = balance ~ income, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -967.15 -362.86   -7.45  326.04 1774.51
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1020.7665    17.9691   56.81   <2e-16 ***
## income        -5.7525     0.4967  -11.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 476.1 on 5155 degrees of freedom
## Multiple R-squared:  0.02536,    Adjusted R-squared:  0.02517
## F-statistic: 134.1 on 1 and 5155 DF,  p-value: < 2.2e-16
```

```
plot(data$income, data$balance, xlab = "Income", ylab = "Balance", main = "Scatterplot of Balance vs. I
abline(model)
```



```
p_value_slope <- summary(model)$coefficients["income", "Pr(>|t|)"]
p_value_slope
```

```
## [1] 1.236754e-30
```

##the model is statistically significant

5.

```
coefficient_income <- coef(model)["income"]
expected_change <- coefficient_income * 1000
expected_change
```

```
##    income 
## -5752.503
```

6.

```
income_estimation <- 40000
income_prediction <- 80000

estimated_average_balance <- predict(model, newdata = data.frame(income = income_estimation))

predicted_balance <- predict(model, newdata = data.frame(income = income_prediction))

#estimated average balance fro income = 40k:
round(estimated_average_balance, 2)
```

```
##         1
## -229079.3
```

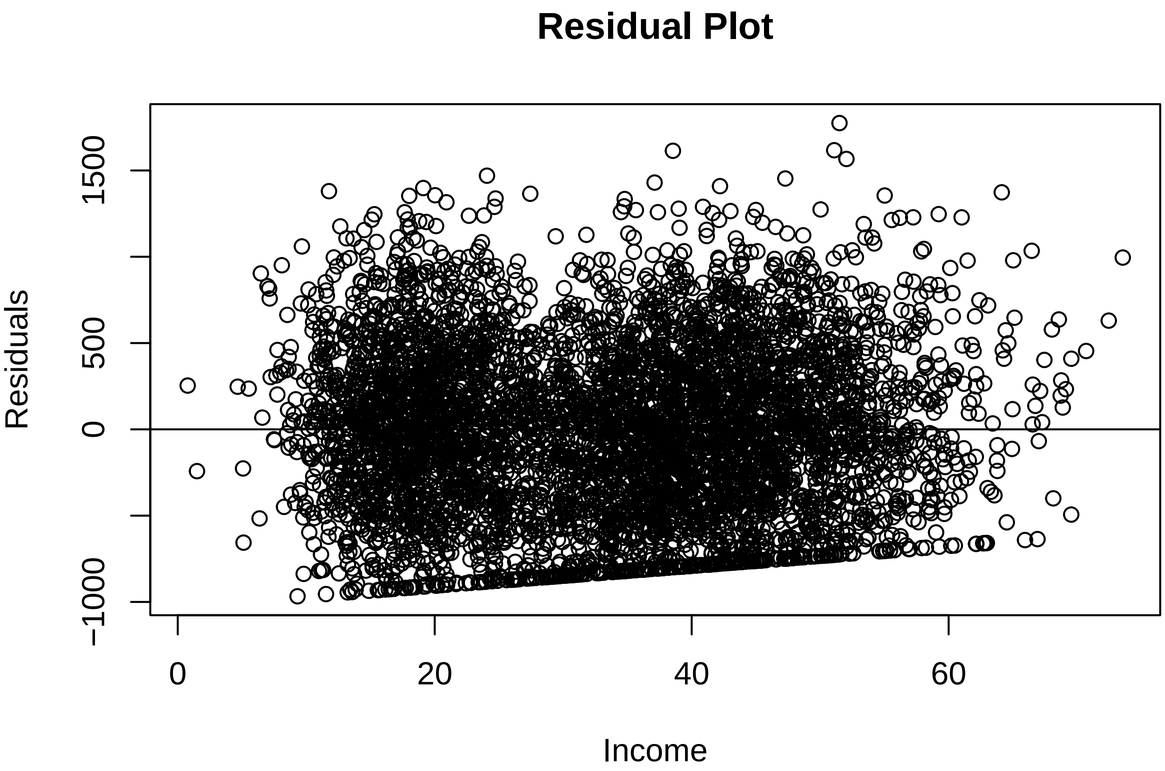
```
#predicted balance for income = 80k
round(predicted_balance, 2)
```

```
##         1
## -459179.5
```

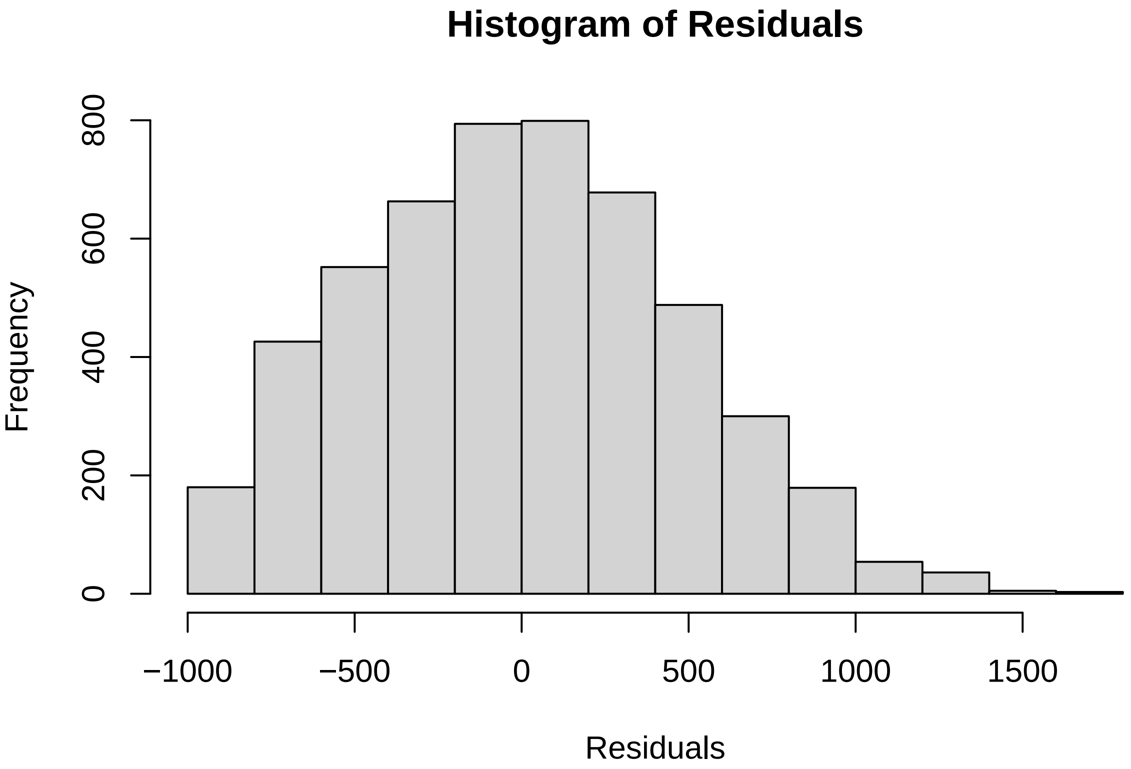
##it may be important to make sure that the assumptions of LINE are met to ensure that the model's predictions are reliable.

7.

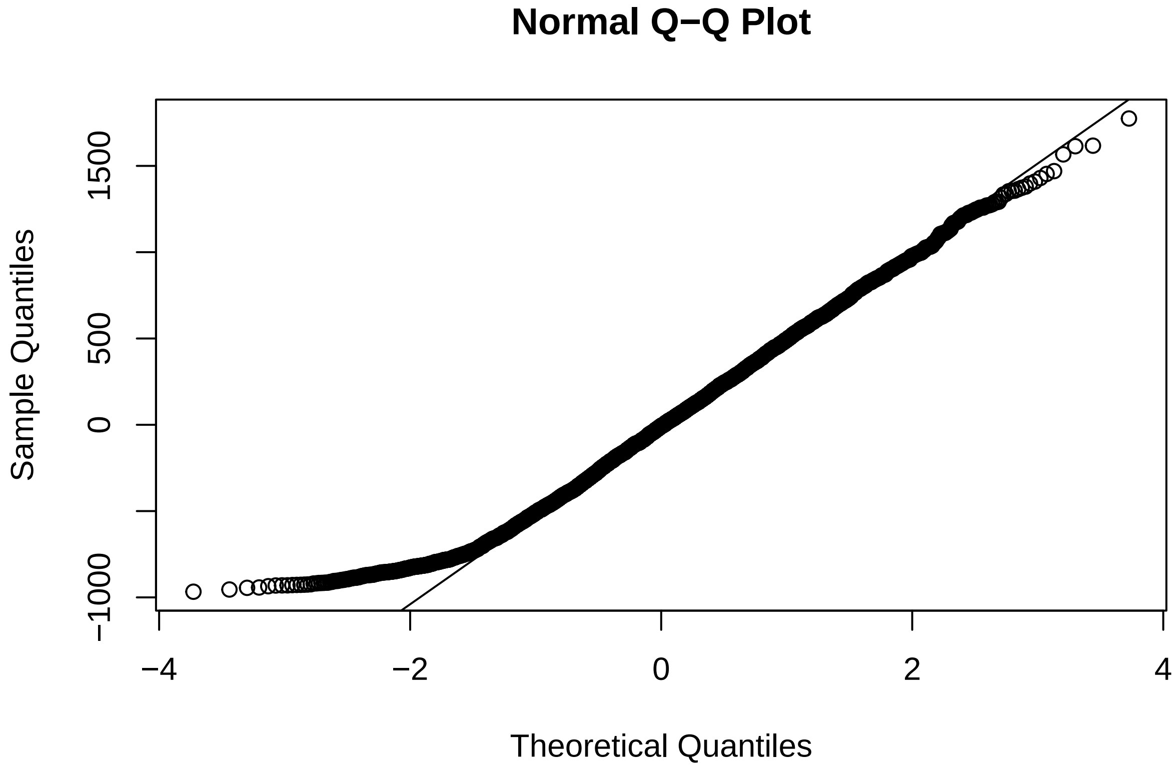
```
residuals <- residuals(model)

plot(data$income, residuals, xlab = "Income", ylab = "Residuals", main = "Residual Plot", ylim = range(
abline(h = 0)
```



```
hist(residuals, main = "Histogram of Residuals", xlab = "Residuals", ylab = "Frequency")
```



```
qqnorm(residuals)
qqline(residuals)
```



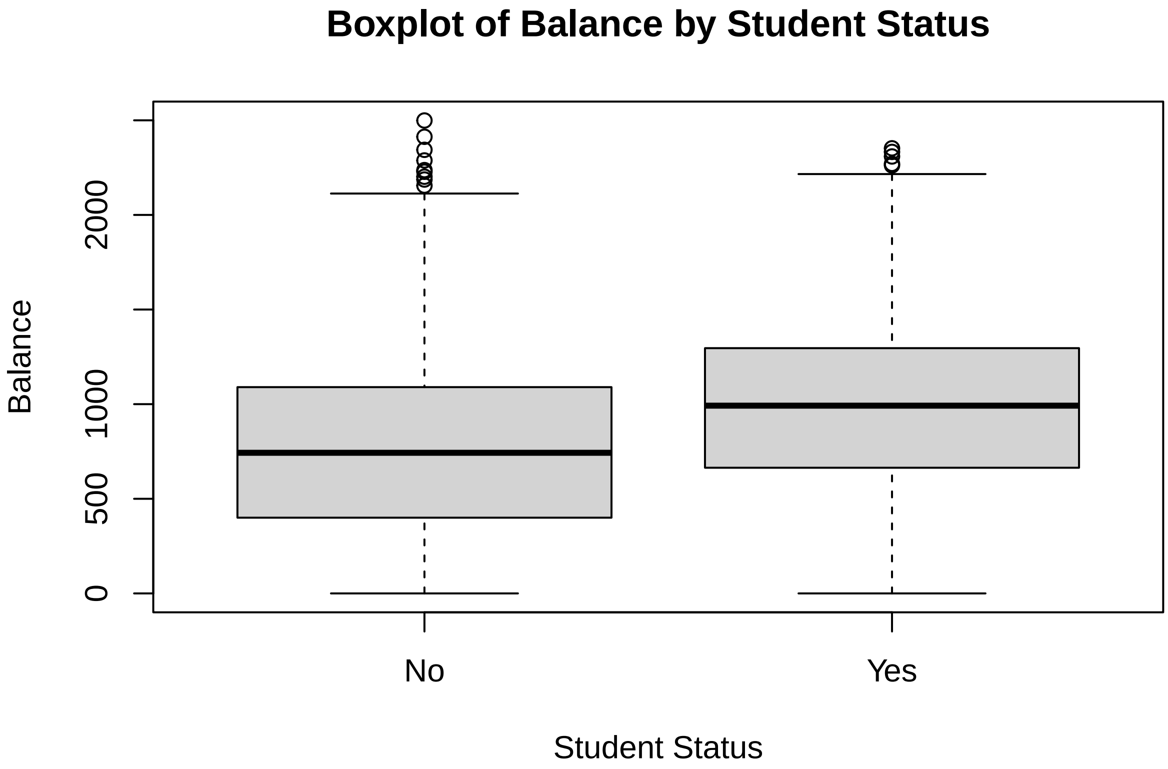
8.

```
rsquared <- summary(model)$r.squared
percentage_explained <- rsquared * 100
round(percentage_explained, 2)
```

```
## [1] 2.54
```

9.

```
boxplot(balance ~ student, data = data, xlab = "Student Status", ylab = "Balance", main = "Boxplot of Ba
```



```
model_student <- lm(balance ~ student, data = data)
summary(model_student)
```

```
##
## Call:
## lm(formula = balance ~ student, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -989.54 -351.54  -12.02  320.46 1737.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  761.023      7.784   97.77   <2e-16 ***
## studentYes   228.513     14.448   15.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 470.9 on 5155 degrees of freedom
## Multiple R-squared:  0.04628,    Adjusted R-squared:  0.0461
## F-statistic: 250.2 on 1 and 5155 DF,  p-value: < 2.2e-16
```

10.

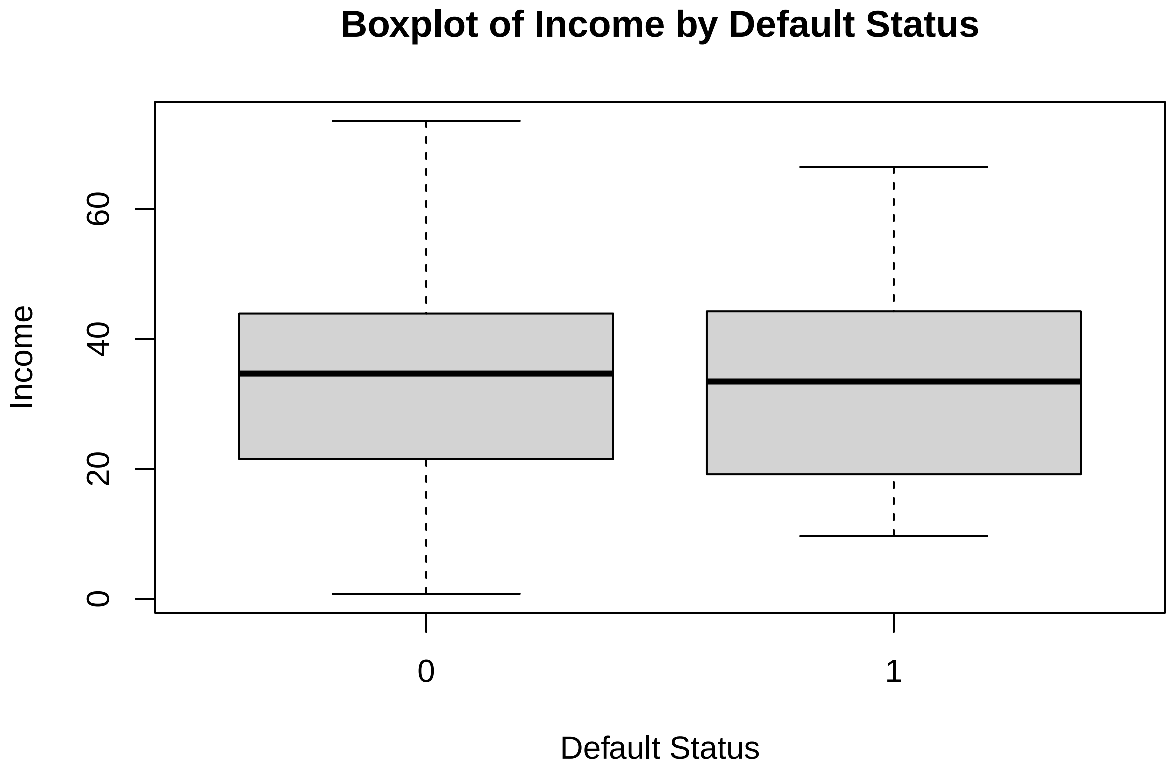
11.

```
model_other_level <- lm(balance ~ student, data = data)
summary(model_other_level)
```

```
##
## Call:
## lm(formula = balance ~ student, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -989.54 -351.54  -12.02  320.46 1737.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  761.023      7.784   97.77   <2e-16 ***
## studentYes   228.513     14.448   15.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 470.9 on 5155 degrees of freedom
## Multiple R-squared:  0.04628,    Adjusted R-squared:  0.0461
## F-statistic: 250.2 on 1 and 5155 DF,  p-value: < 2.2e-16
```

12.

```
boxplot(income ~ default, data = data,xlab = "Default Status", ylab = "Income", main = "Boxplot of Inco
```

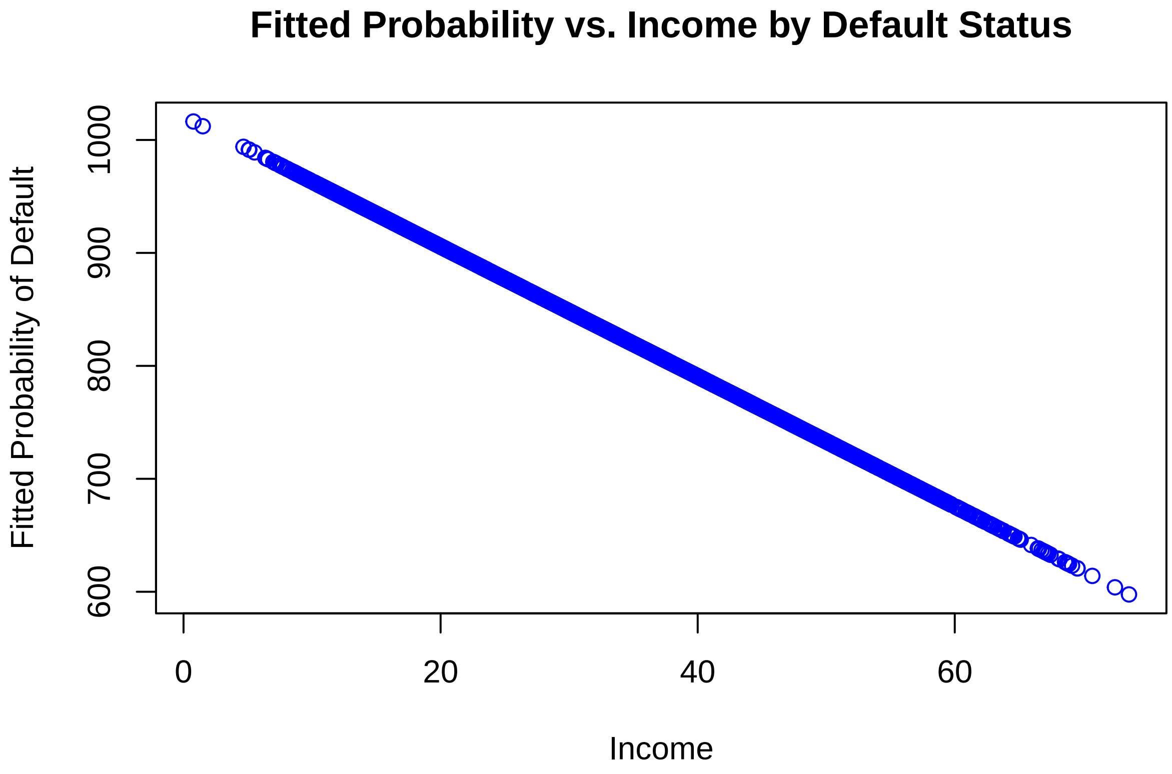


13.

14.

15.

```r
fitted_probs <- predict(model, newdata = data, type = "response")
scatter_data <- data.frame(income = data$income, default = data$default, fitted_prob = fitted_probs)
plot(scatter_data$income, scatter_data$fitted_prob, col = ifelse(scatter_data$default == "Yes", "red",
```



```r
new_data <- data.frame(income = 60000)
predicted_prob <- predict(model, newdata = new_data, type = "response")
round(predicted_prob, 4)
```

```
##         1
## -344129.4
```

16.