

# Midterm 1 - Akshaj Kammari

02/19/2024

```
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
```

```
library(moderndiver)
library(epiDisplay)
```

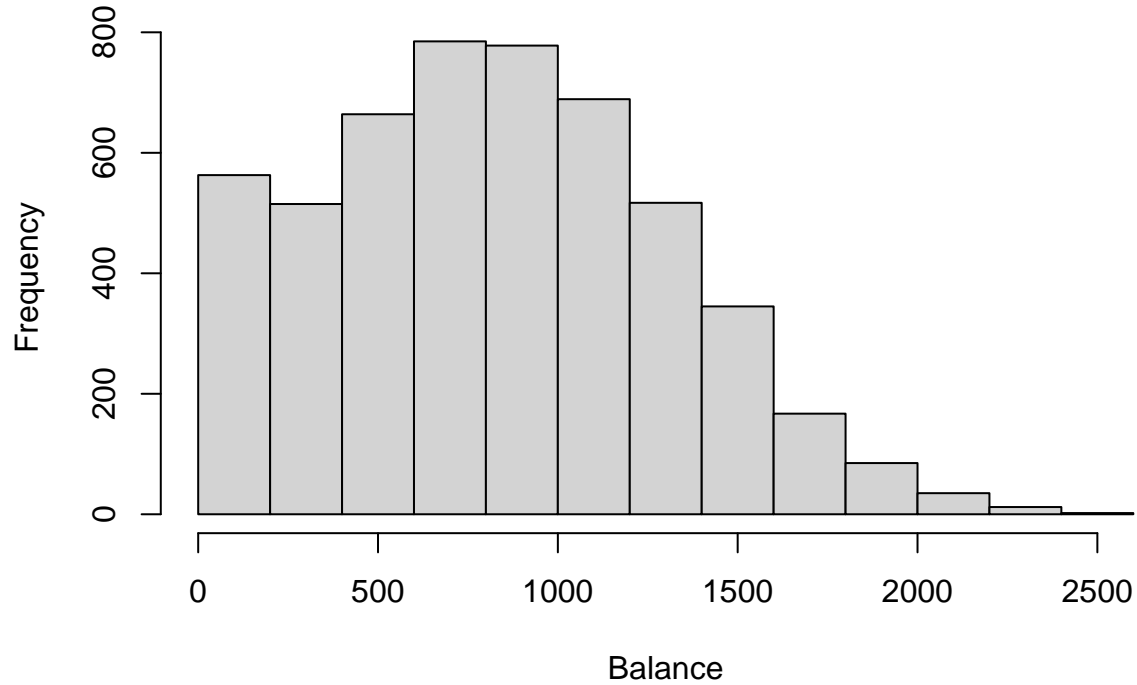
```
## Loading required package: foreign
## Loading required package: survival
## Loading required package: MASS
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##   select
## Loading required package: nnet
##
## Attaching package: 'epiDisplay'
## The following object is masked from 'package:ggplot2':
##
##   alpha
```

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

```
1. #histogram of balance
```

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

## Histogram of Balance

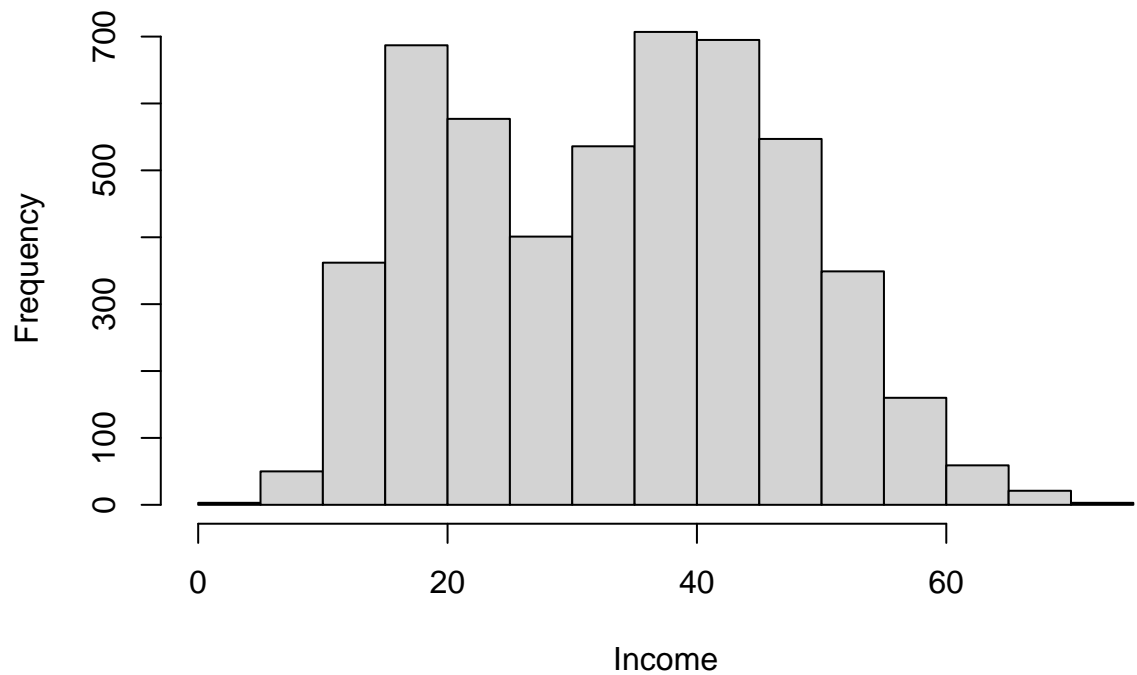


histogram of balance is skewed right.

#histogram of income

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

## Histogram of Income



histogram of income seems to have a bell-curve

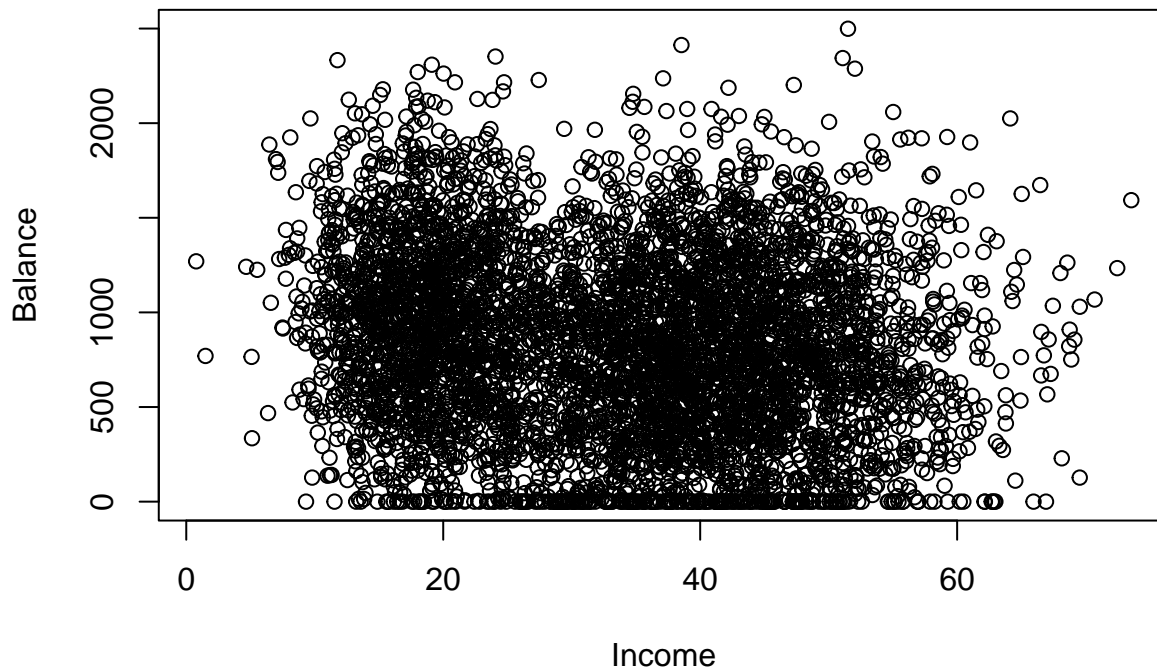
##the

##the

2.

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

### Scatterplot of Balance vs. Income



3.

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

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

4.

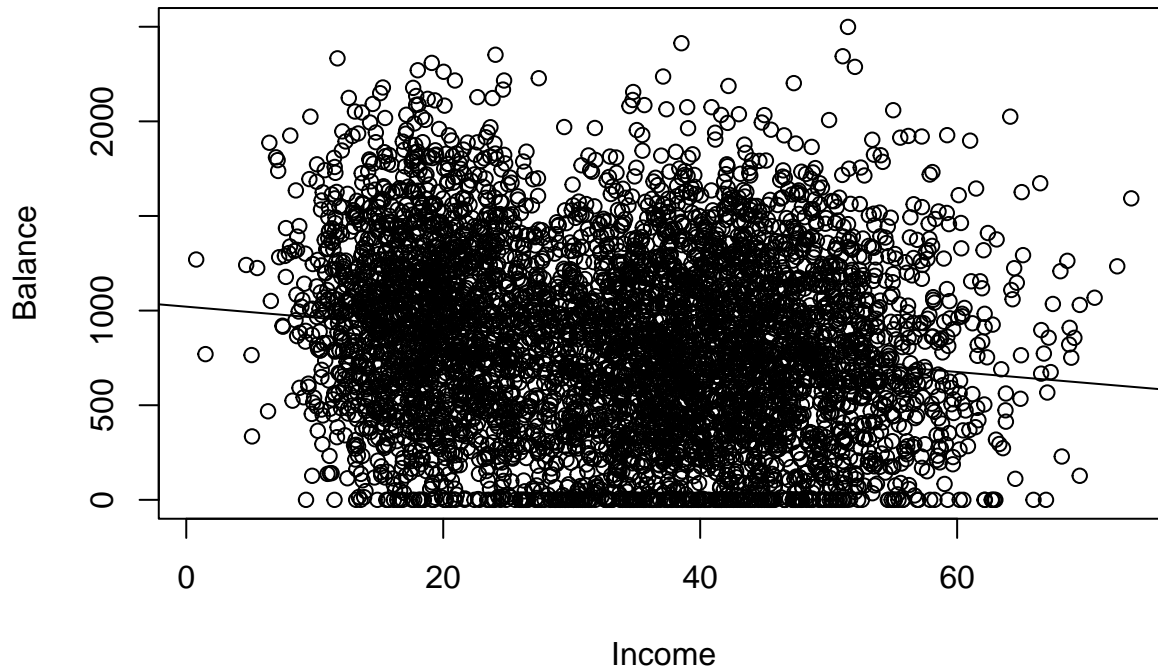
```
model <- lm(balance ~ income, data = data)
summary(model)
```

```
##
## 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. Income")  
abline(model)
```

## Scatterplot of Balance vs. Income



```
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)
```

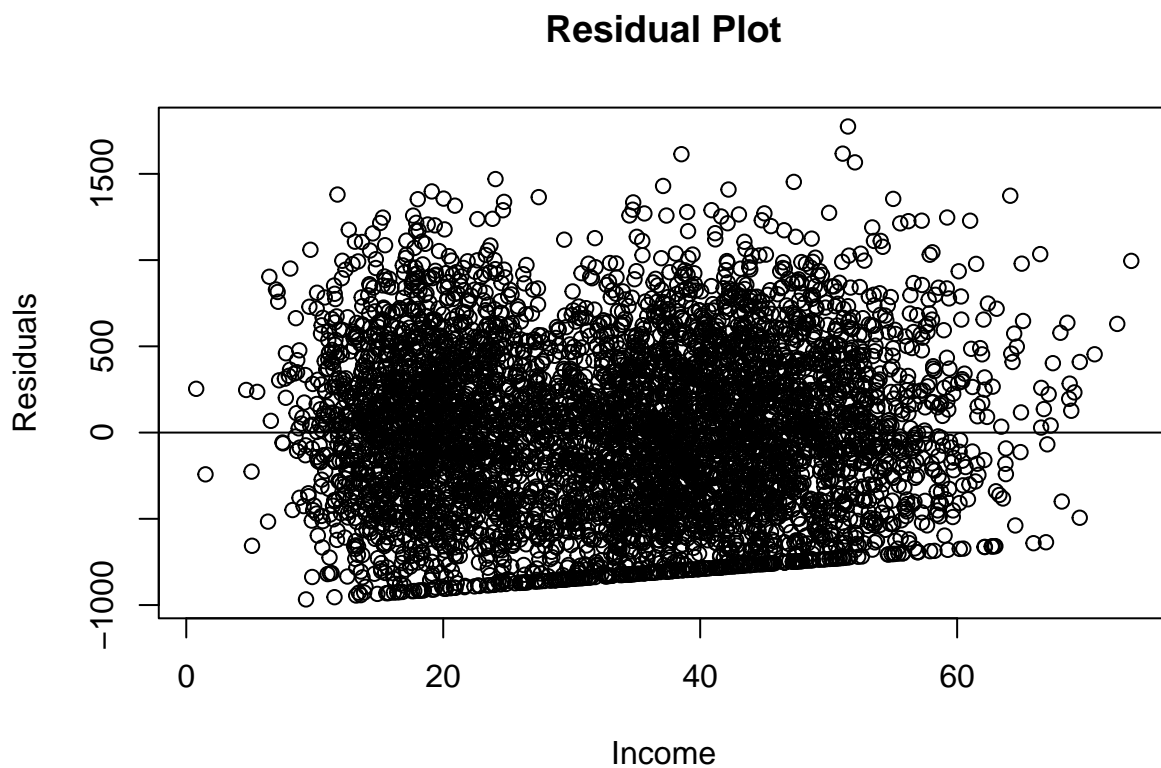
```
##          1  
## -459179.5
```

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

7.

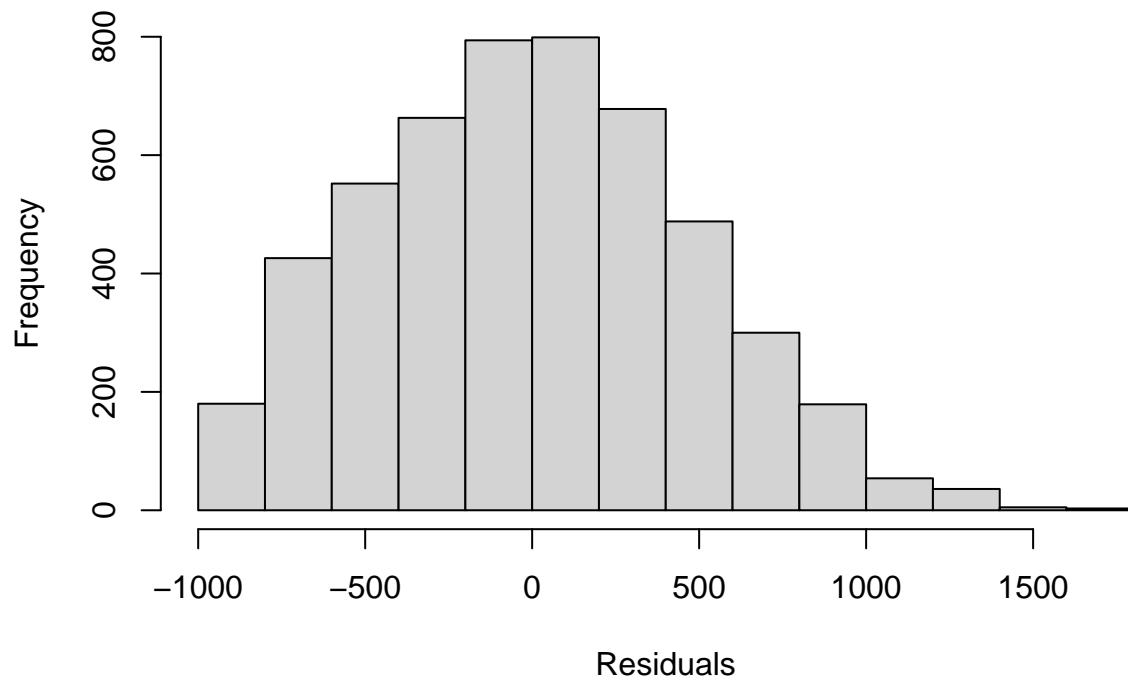
```
residuals <- residuals(model)
```

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



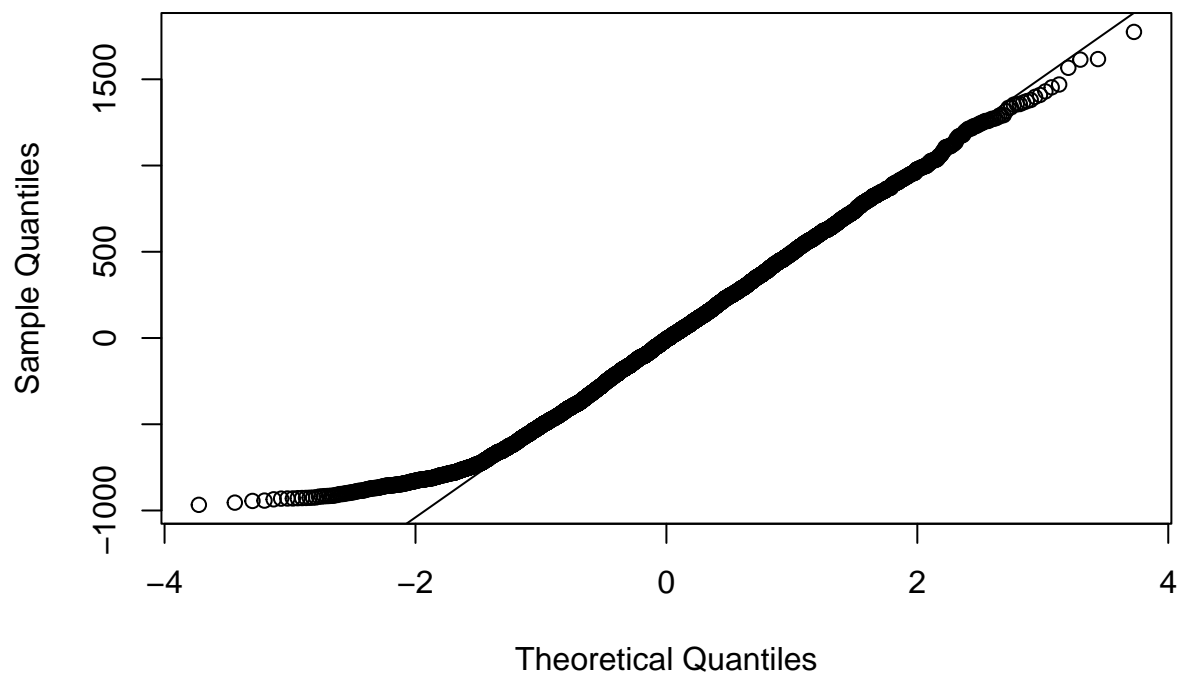
```
hist(residuals, main = "Histogram of Residuals", xlab = "Residuals", ylab = "Frequency")
```

### Histogram of Residuals



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

### Normal Q-Q Plot



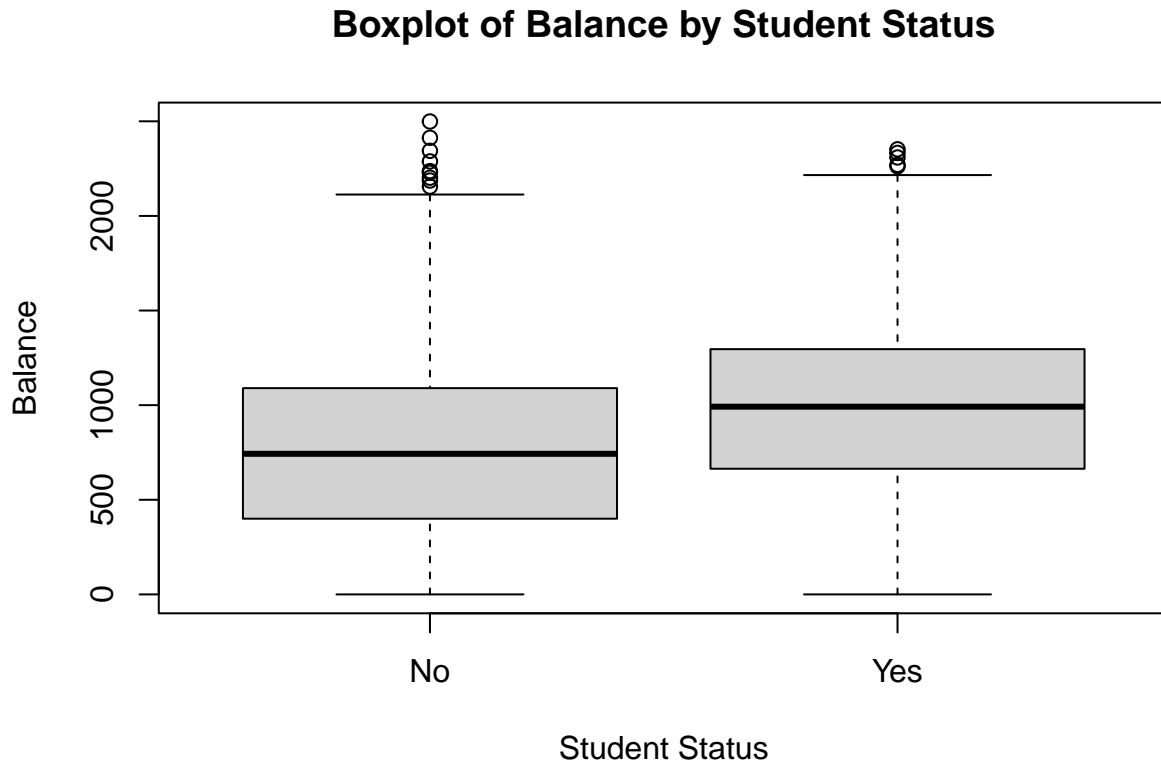
8.

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

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

9.

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



```
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.

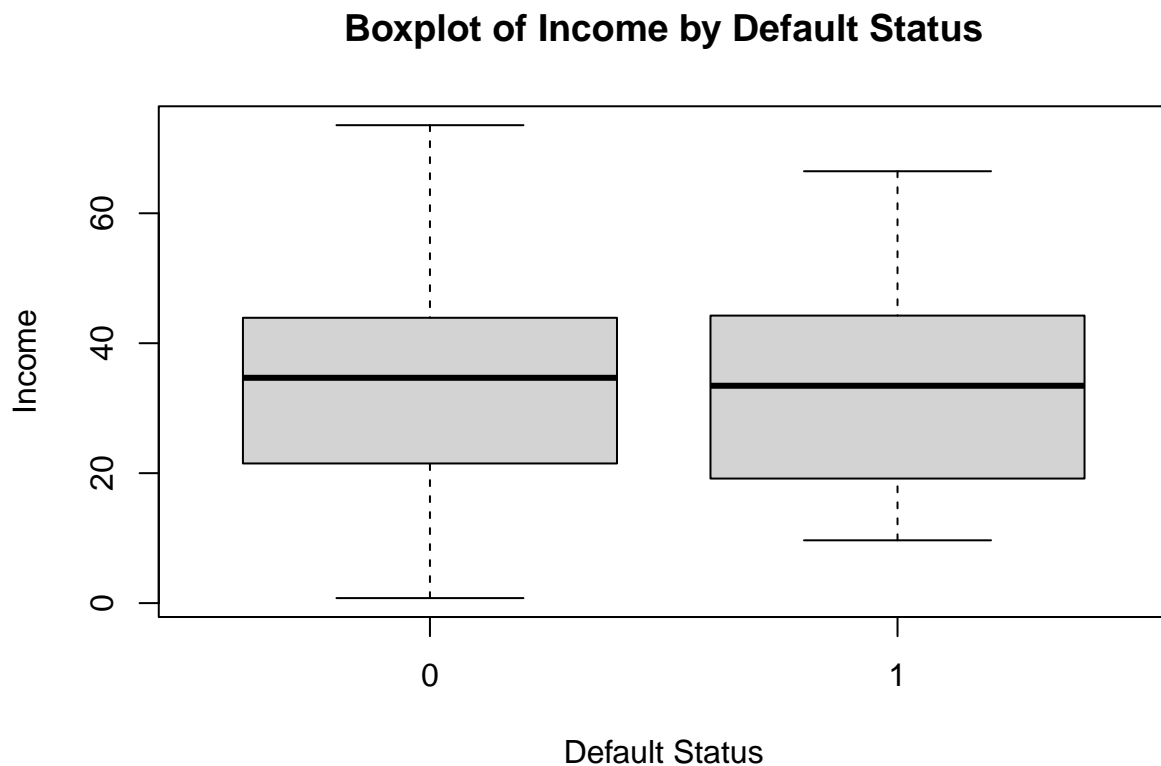
11.

```
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 Income by Default Status")
```



13.

14.



```
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",
```

A scatter plot showing the relationship between Income (X-axis) and Fitted Probability of Default (Y-axis). The X-axis ranges from 0 to 70, and the Y-axis ranges from 600 to 1000. The data points are represented by open circles, and a solid blue line represents the linear regression fit. The fitted probability of default decreases as income increases.

Income	Fitted Probability of Default
2	1020
3	1010
6	990
7	985
8	980
9	975
10	970
11	965
12	960
13	955
14	950
15	945
16	940
17	935
18	930
19	925
20	920
21	915
22	910
23	905
24	900
25	895
26	890
27	885
28	880
29	875
30	870
31	865
32	860
33	855
34	850
35	845
36	840
37	835
38	830
39	825
40	820
41	815
42	810
43	805
44	800
45	795
46	790
47	785
48	780
49	775
50	770
51	765
52	760
53	755
54	750
55	745
56	740
57	735
58	730
59	725
60	720
61	715
62	710
63	705
64	700
65	695
66	690
67	685
68	680
69	675
70	670
71	665
72	660
73	655
74	650
75	645
76	640
77	635
78	630
79	625
80	620
81	615
82	610
83	605
84	600
85	595
86	590
87	585
88	580
89	575
90	570
91	565
92	560
93	555
94	550
95	545
96	540
97	535
98	530
99	525
100	520

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

9