Does education level has an effect on defaulting Credit Card payment?

Salman Quaiyum

1. Introduction

This is the final project for Udacity's Intro to Inferential Statistics course. It has been implemented using **R**.

2. Research Question & Hypothesis

For analysis purposes, Default of Credit Card Clients Dataset This dataset contains payment records of 30,000 clients from Taiwan - recording their age, gender, marital status, education level, amount of credit given, whether they would default their payment next month etc.

This analysis was done to explore whether their level of education is somehow related to the clients making payments towards their credit card dues. In other words, whether people of a certain education level would be more prone to defaulting their credit card payments.

3. 1st Experimental Design

The first step was to check whether the level of education is related to defaulting at all. For this purpose, a **Chi Square Test of Independence** was performed. The null and alternative hypotheses are given below.

Null Hypothesis: Defaulting of credit card payment is independent of the clients' education level. Clients' level of education has no effect on default payment.

Alternative Hypothesis: Defaulting of credit card payment is not independent of the clients' education level. Clients' level of education has an effect on default payment.

3.1 Coding

At first, the required libraries were imported.

```
library(data.table)
library(dplyr)
```

Before importing the data in R, I opened the excel file, took a look at the data, got rid of unnecessary rows and saved it as a csv file. Then I imported it into R.

```
setwd("D:/R/Data")
credit_data <- fread("default.csv", header = TRUE)
```

After importing, the `default_payment` column was converted into factors having two levels. This was done to represent the data in more legible format.

```
credit_data$default_payment <- factor(credit_data$default_payment, labels
= c("No", "Yes"), levels = c(0,1))
```

As the dataset description does not say what **0**, **5** and **6** stands for. I would consider **0** signifying no education and would merge **5** & **6** with **4**.

```
credit_data$EDUCATION[credit_data$EDUCATION > 4] <- 4
```

Then the dataset was summarized into a tabular format using `table` command.

```
credit_table <- table(credit_data$default_payment, credit_data$EDUCATION,
dnn = c("Default", "Education_level"))

##         Education_level
## Default     0     1     2     3     4
##     No     14  8549 10700  3680   421
##     Yes     0  2036  3330  1237    33
```

The column headings are different education levels:

1. **1** = Graduate School
2. **2** = University
3. **3** = High School
4. **4** = Others

Afterwards, the `credit_table` was converted into a `matrix` object, so that I could directly work with the table.

```
credit_table <- as.matrix(credit_table)
```

For performing the **Chi square test**, I created the following function. This funtion, called `chi2_test` takes a matrix as its input and outputs **Chi Square** value.

```
chi2_test <- function(mat){

  col_total <- apply(mat, 2, sum)
  row_total <- apply(mat, 1, sum)
  grand_total <- sum(row_total)

  exp_val <- matrix(0, nrow = nrow(mat), ncol = ncol(mat))
  chi2_val <- matrix(0, nrow = nrow(mat), ncol = ncol(mat))

  for (i in 1:nrow(mat)) {

    exp_val[i,] <- (col_total*row_total[i])/grand_total
    chi2_val[i,] <- ((mat[i,]-exp_val[i,])^2)/exp_val[i,]
  }

  chi2 <- sum(chi2_val)

  return(chi2)
```

```
}
```

Let us call this function with `credit_table` as its argument.

```
chi2 <- chi2_test(credit_table)
```

I later found out that the same thing could be done by utilizing the `chisq.test` function from the `stats` package.

```
chisq.test(credit_table)

##
##  Pearson's Chi-squared test
##
## data:  credit_table
## X-squared = 160.83, df = 4, p-value < 2.2e-16
```

I also calculated the **Cramer's V**. The dataset has **30,000** records and table created for Chi square test has **2** rows and **5** columns. So for calculating degrees of freedom, number of rows was considered.

```
sample_number <- nrow(credit_data)

df <- (nrow(credit_table) - 1)

cramers_v <- sqrt(chi2 / (sample_number * df))
```

The **Critcal Chi Square Value** for **95% Confidence Interval** (alpha = 0.05) for **4** degrees of freedom was found to be: **9.488**.

3.2 Results

Chi2(4) = 160.83, p < 0.05, Cramer's V = 0.07

Since our Chi square value is much greater than the critical value (9.488), it definitely falls in the critical region. Therefore, we **reject** the null hypothesis. As such, clients' defaulting of payment is dependent on their respective education level. **Cramer's V** value indicates that the effect size is small.

4. 2nd Experimental Design

I decided to perform a **One Way ANOVA (Analysis of Variance) Test** to further investigate the claim. For this purpose, I designed a new feature called `CREDIT_SCORE`. This sums up the scores from past paymenet records (`PAY_0` to `PAY_6`).

Null Hypothesis: The credit scores of clients having different levels of education is same. There is no relation between clients defaulting and their corresponding level of education.

Alternative Hypothesis: The credit scores of clients having different levels of education is different. There is a relation between clients defaulting and their corresponding level of education.

4.1 Coding

At first, I added a new calculated column called `CREDIT_SCORE` to our data and took only the sample of clients who have defaulted their payments.

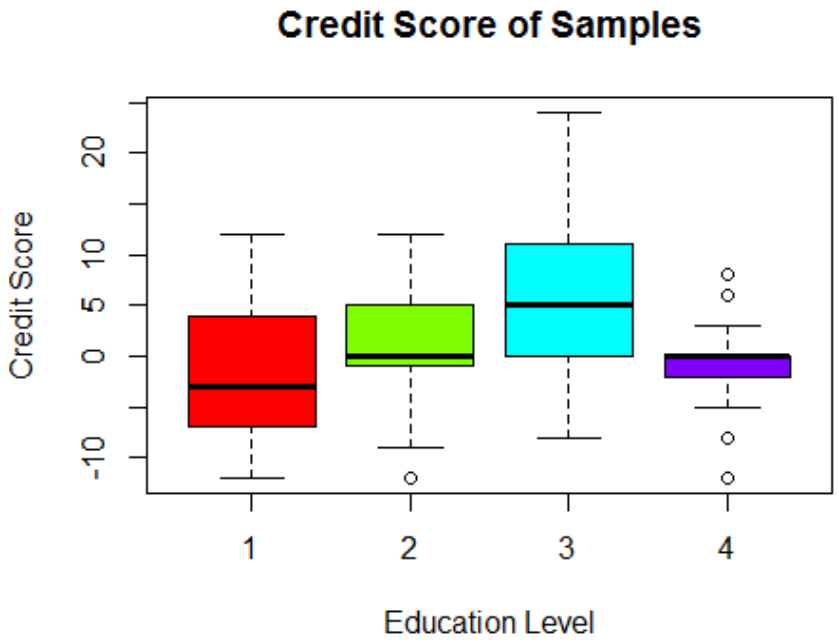
```
credit_edu <- credit_data %>%
              mutate(CREDIT_SCORE = PAY_0 + PAY_2 + PAY_3 + PAY_4 + PAY_5
+ PAY_6) %>%
              filter(default_payment == "Yes") %>%
              select(ID, EDUCATION, CREDIT_SCORE)
```

Then I took **33** samples from each group and performed One Way Anova test on them.

```
edu_level <- sort(unique(credit_edu$EDUCATION))

sample_size <- 33

edu_samp <- matrix(0, nrow = sample_size, ncol = length(edu_level))

for (i in edu_level) {
  data <- credit_edu %>%
          filter(EDUCATION == i) %>%
          select(CREDIT_SCORE)
  edu_samp[,i] <- data[1:sample_size,]
}
```

Let's take a look at our samples using the `summary` command:

1	2	3	4
Min. :-12.000	Min. :-12.0000	Min. :-8.000	Min. :-12.000
1st Qu.: -7.000	1st Qu.: -1.0000	1st Qu.: 0.000	1st Qu.: -2.000
Median : -3.000	Median : 0.0000	Median : 5.000	Median : 0.000
Mean : -1.242	Mean : 0.4545	Mean : 4.879	Mean : -1.485
3rd Qu.: 4.000	3rd Qu.: 5.0000	3rd Qu.:11.000	3rd Qu.: 0.000
Max. : 12.000	Max. : 12.0000	Max. :24.000	Max. : 8.000



Then I would calculate Sum of Squares of between group variability (SSb).

```
grand_mean <- mean(edu_samp)

mean_edu <- colMeans(edu_samp)

SSb <- sample_size * sum((mean_edu - grand_mean)^2)
```

Then I moved on to calcuating Sum of Squares of within group variablility (SSw).

```
ssw_mat <- matrix(0, nrow = sample_size, ncol = length(edu_level))

for (i in edu_level) {
  ssw_mat[,i] <- (edu_samp[,i] - mean_edu[i])^2
}

SSw <- sum(ssw_mat)
```

Afterwards, I calculated F value for this sample.

```
k <- length(edu_level)

df_b <- k - 1

df_w <- (sample_size*k) - k

MSb <- SSb / df_b

MSw <- SSw / df_w

F_val <- MSb / MSw
```

```
eta2 <- SSb / (SSb + SSw)
```

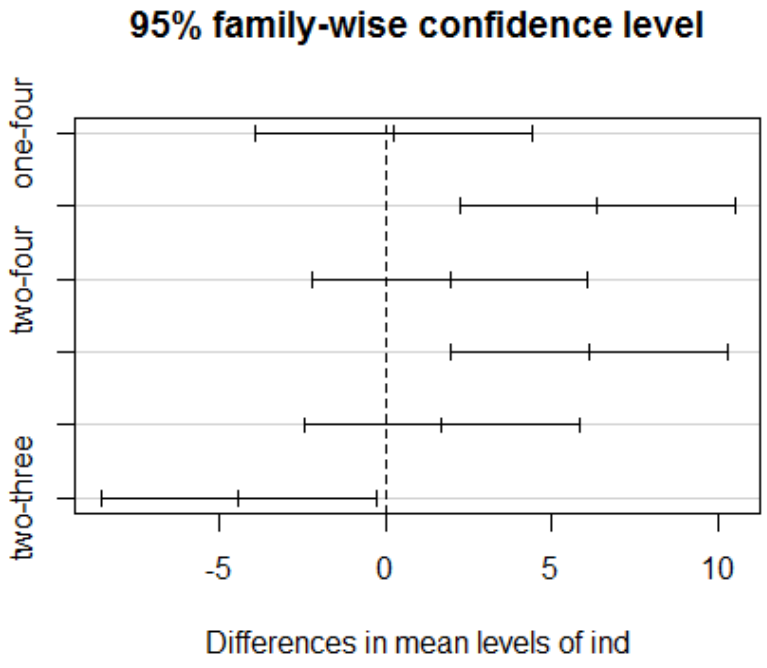
The same analysis could be done using the built in aov function.

```
result <- aov(values ~ ind, data = edu_samp2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## ind           3    860  286.66   6.856 0.000255 ***
## Residuals   128   5352   41.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is also possible to calculate the **Tukey's Honsestly Significant Difference (HSD)** using `TukeyHSD` command and plot them.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = values ~ ind, data = edu_samp2)
##
## $ind
##                  diff       lwr       upr     p adj
## one-four    0.2424242 -3.901410  4.386259 0.9987354
## three-four  6.3636364  2.219802 10.507471 0.0006158
## two-four    1.9393939 -2.204440  6.083228 0.6164124
## three-one   6.1212121  1.977378 10.265047 0.0010726
## two-one     1.6969697 -2.446865  5.840804 0.7107791
## two-three  -4.4242424 -8.568077 -0.280408 0.0314600
```



4.2 Results

SSbetween = 859.97, dfbetween = 3, MSbetween = 286.66

SSwithin = 5352, dfwithin = 128, MSwithin = 41.81

F(3, 128) = 6.86, $p < 0.05$, eta2 = 0.14

Since the F value is greater than critical values of F (2.70 at alpha = 0.05) we **reject** the null hypoethesis.

Therefore, the credit levels of clients having different levels of education are different.

14% of total variation in credit scores was due to different levels of education.

Looking at the results of **Tukey's HSD**,

1. There is no significant difference between groups One & Four (p = 0.99), Two & Four (p = 0.62), Two & One (p = 0.71)

2. There are significant differences between groups Three & Four (p = 0.0006), Three & One (p = 0.001) and group Two & Three (p = 0.03). These groups also do not cross the 0 in Tukey's plot.

5. Conclusion

Unfortunately, it is not possible to find which group of clients are more prone to defaulting their credit card payment. It can only be stated that there is a difference between clients of differing levels of education. However, since the credit scores of group Three differs more with other groups, it might be assumed that clients belonging to group **Three**, people who have recived education up to High School, are more likely to default their credit card payments.

I have learned a lot of new things about statistics while doing this project. Though I have a long way to go, before I am really able to really extract meaningful insights from data. I would appreciate any sort of feedback / comment / suggestion.

https://about.me/salmanq