Project Proposal

Isaac Cheong, Sam Tan, Max Zhang

Introduction

rz

Data Description

EDA

In our Exploratory Data Analysis, we will examine each response and explanatory variable to check if these variables fit the basic assumptions of linear regression to provide an accurate and concise context for our future analysis, including model selection and model diagnostic.

If we see the histogram of our response variable`Income` (See Appendix), we can observe that this histogram is heavily skewed to the right, violating our normality assumption of the response variable. To fix this issue, we will apply a log transformation to the variable income. As you see in the Appendix, our log-transformed `income` histogram looks approximately normal distribution, and this is much more aligned with the linear regression assumption of response variable normality. We are not excluding the maximum point of the histogram because it is reasonable that someone earns $17000 per month, and it is hard to consider this point as an error from the data collection process.

Our data analysis on explanatory variables will be more complicated since we have many variables. We will mainly use visual tools in the ggplot2 package to examine and analyze the distribution of each explanatory variable. Possible visual tools will be a correlation matrix, multi-level boxplot, and bivariate scatter plot. Through this analysis of this explanatory variable, we should be able to answer the following questions.

(1) What will be the correlation between `income` and our explanatory variables? What will be the correlation between explanatory variables? Can we identify any collinearity between variables?

(2) Using our graphical demonstration of explanatory variables, can we identify any unusual data point in our explanatory variables? Can we determine if the data point is an outlier?

(3) Should we re-group our categorical variable to dichotomous variables like “Yes” or “No”? Can we change our categorical variable to a numerical variable by checking the linearity of the boxplot between income and categorical variables?

(4) Can we identify any possible interaction variables using relevant co-plot or boxplot?

Our research paper hopes to answer all of these questions from this exploratory data analysis to make a better conclusion and more accurate causal inference.

Model Selection

Model Diagnostics

Prediction

Conclusion

Appendix

Loading the dataset and assigning the data to the variable "cils"

```
load('ICPSR_20520/DS0001/20520-0001-Data.rda')  # creates df called da20520.0001
cils <- da20520.0001
```

Sample of dataet

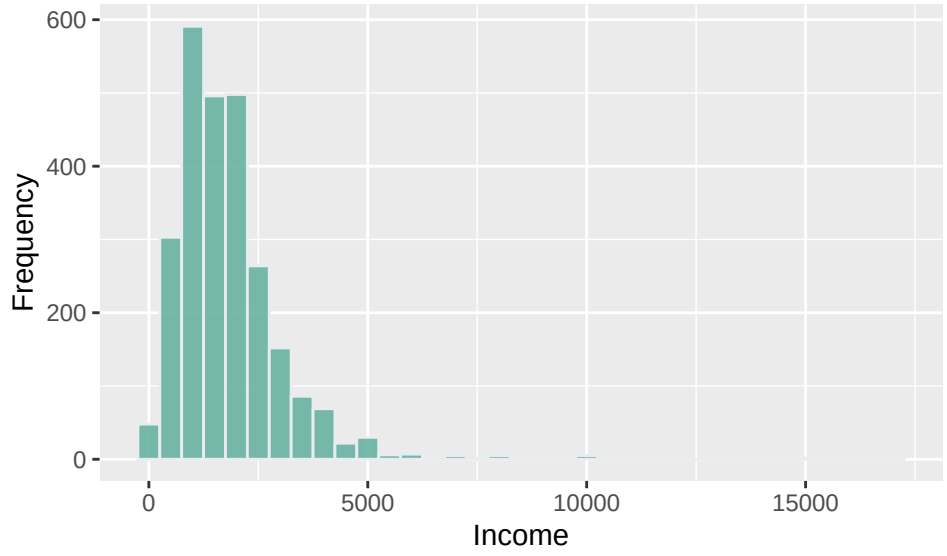
As you see it here, the dataset cils can be loaded into R and looks beautiful.

```
head(cils[ ,1:5], 5)
```

```
##   CASEID   V1                 V2             V4                 V5
## 1      1  257          (1) Miami  (07) School 7 (08) Eighth grade
## 2      2 2347          (1) Miami (13) School 13  (09) Ninth grade
## 3      3  860          (1) Miami (12) School 12  (09) Ninth grade
## 4      4 5178 (3) Ft. Lauderdale (20) School 20  (09) Ninth grade
## 5      5 1984          (1) Miami (12) School 12  (09) Ninth grade
```

Distribution of `income`

```
ggplot(cils, aes(x = V421)) +
  geom_histogram(na.rm = TRUE, binwidth = 500,
                 fill="#69b3a2", color="#e9ecef", alpha=0.9) +
  xlab("Income") +
  ylab("Frequency")
```



Maximum income

```
max(cils$V421, na.rm = TRUE)
```

```
## [1] 17000
```

Histogram of Log Transformed `income`

```
ggplot(cils, aes(x = log(V421))) +
  geom_histogram(na.rm = TRUE,binwidth = 0.4,
                 fill="#69b3a2", color="#e9ecef", alpha=0.9) +
  xlab("Income") +
  ylab("Frequency")
```