# ECON_104_Project_3

Marc Luzuriaga, Takuya Sugahara, Daniel Day, Shabib Alam

2023-12-08

```
data(Cigar)
```

# 1 Panel Data: Cigar

## 1(a) Question

For the purposes of this project, we will ask "Which of the following variables over time: (1) population above the age of 16, (2) consumer price index, (3) state, (4) per capita disposable income (5) sales, (6) per capita disposable income, or (7) minimum price in adjoining states per pack of cigarettes have the largest impact on the price of cigarettes over time?"

In section 2, we will express our question more precisely using the following models: (1) Pooled Model, (2) Fixed Effects, (3) Random Effects, and we will determine which model is best fit to answer our question.

## 1(b) Data Set Summary

In this paper, we will be analyzing the Cigar Data Set from the PLM Package. The data set holds 46 time observations from 1963 to 1992 and 1380 cross-sectional units about Cigarette Consumption data in the United States. The dataset contains panel data with the following nine variables:

(1) state: state abbreviation. For the purposes of this project and simplification, we will only analyze 5 states (state 5, state 33, state 44, state 10, and state 39) in the panel data

(2) year: the year.

(3) price: price per pack of cigarettes.

(4) pop: population

(5) pop16: population above the age of 16.

(6) cpi: consumer price index (1983=100).

(7) ndi: per capita disposable income

(8) sales: cigarette sales in packs per capita

(9) pimin: minimum price in adjoining states per pack of cigaretttes

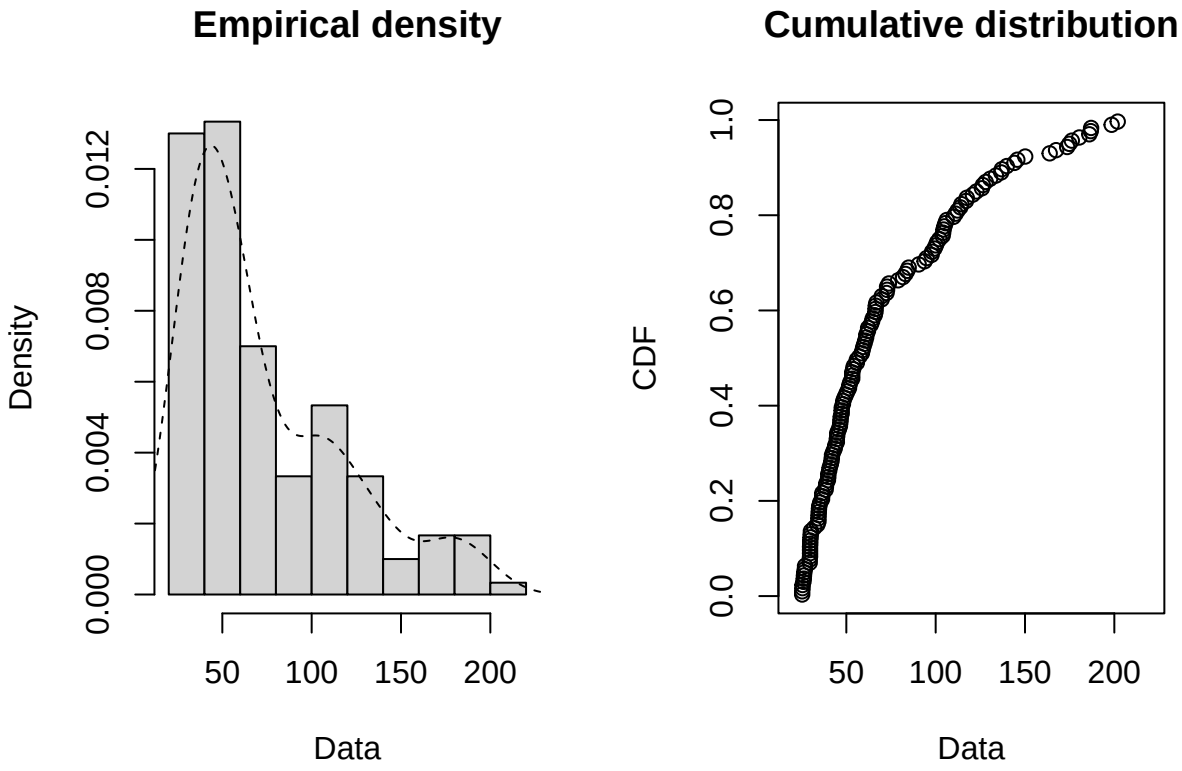
Since we have 46 time observations (T) and 1380 cross-sectional units (N), this panel dataset can be characterized as **short and wide**.

## 1(c) Variable Description

### 1(c).1 Graphs

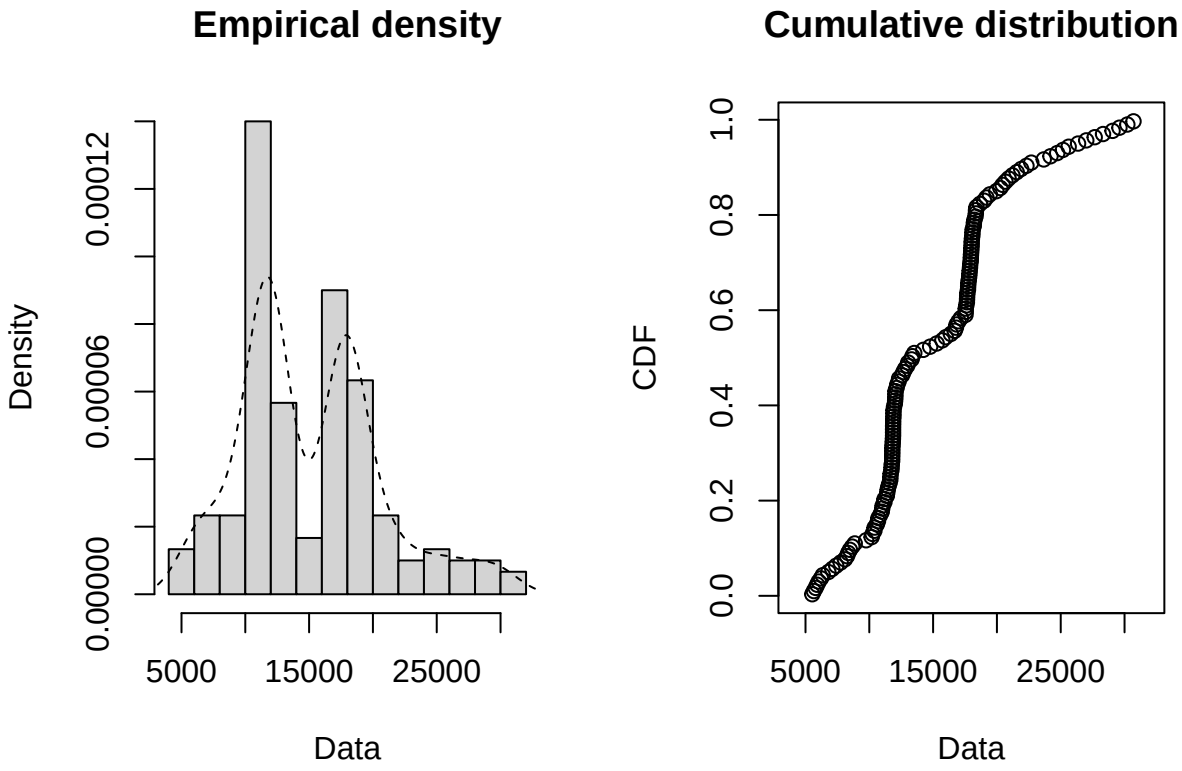
The graph below illustrates a histogram displaying the frequencies with respect to price. The graph portrays the fact that the majority of our observations for will fall between approximately between 0 to 50 gallons of consumption. The data also appears to right-skewed, indicating that we would need to perform a log transformation in order to normalize the data.

```
filtered_data <- filter(Cigar, state == 5 | state == 33 | state == 44 | state == 10
                        | state == 39)
plotdist(filtered_data$price, histo = TRUE, demp = TRUE)
```



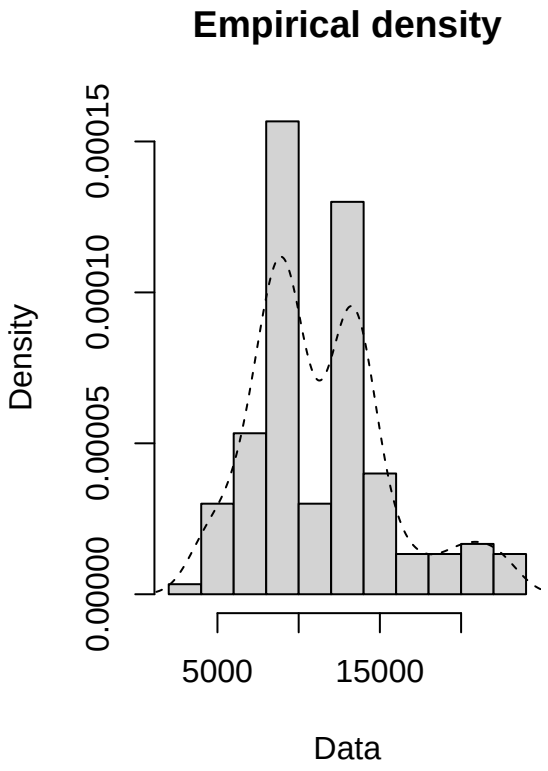

The graph below illustrates a histogram displaying the frequencies with respect to population. The data portrays that the relationship is bi modal (having two peaks).

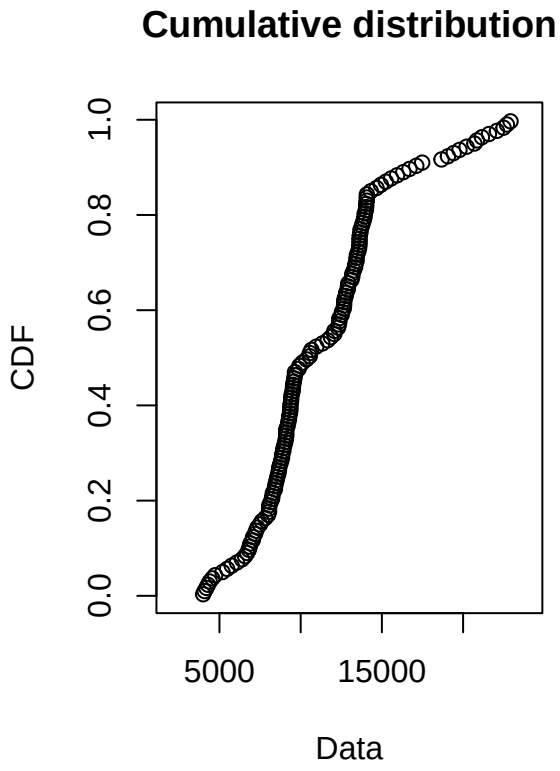
```
plotdist(filtered_data$pop, histo = TRUE, demp = TRUE)
```




The graph below illustrates a histogram displaying the frequencies with respect to population over 16 years old. The data portrays that the relationship is bi modal (having two peaks).

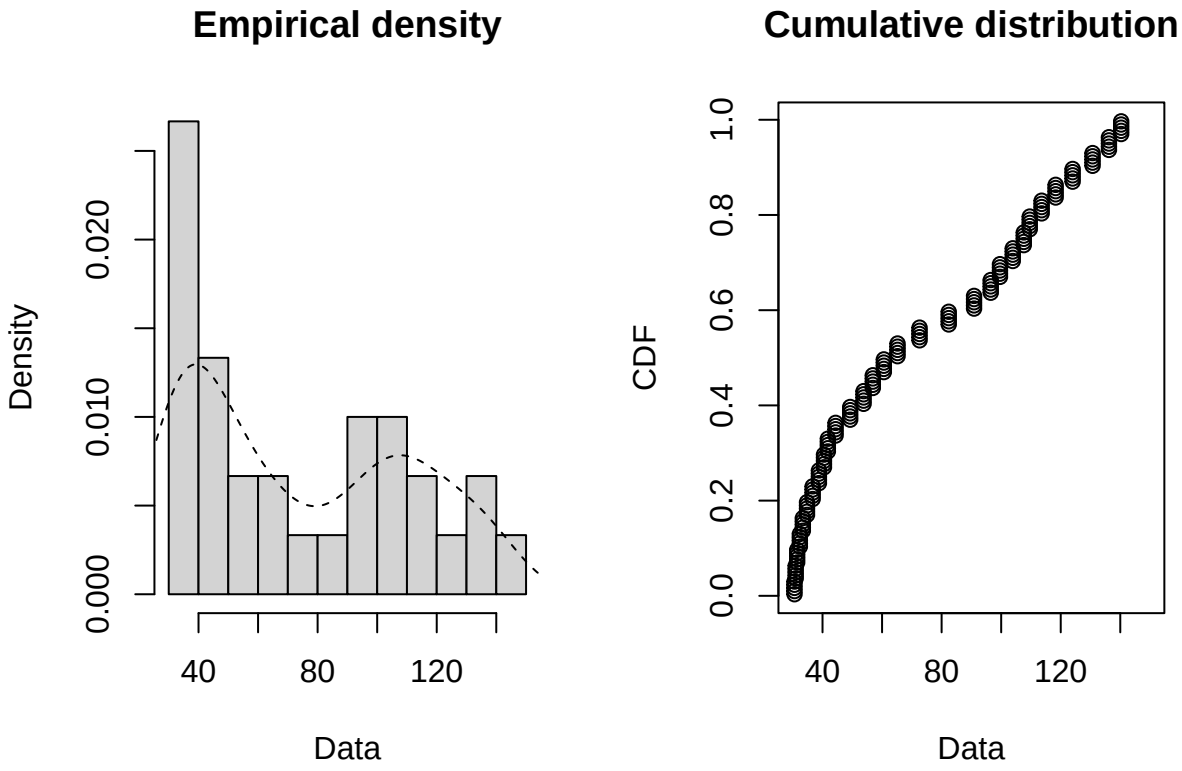
```
plotdist(filtered_data$pop16, histo = TRUE, demp = TRUE)
```



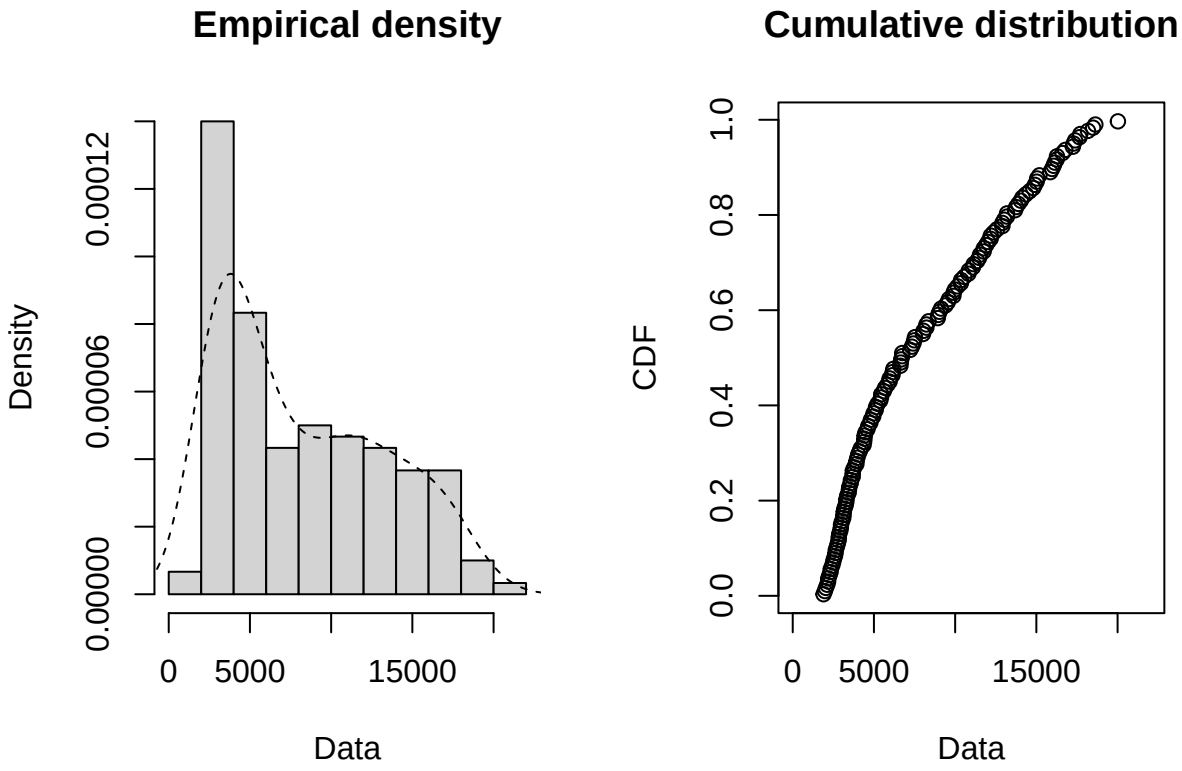

The graph below illustrates a histogram displaying the frequencies with respect to cpi. The data portrays that the relationship is right skewed. A log transformation is needed to normalize the distribution.

```
plotdist(filtered_data$cpi, histo = TRUE, demp = TRUE)
```



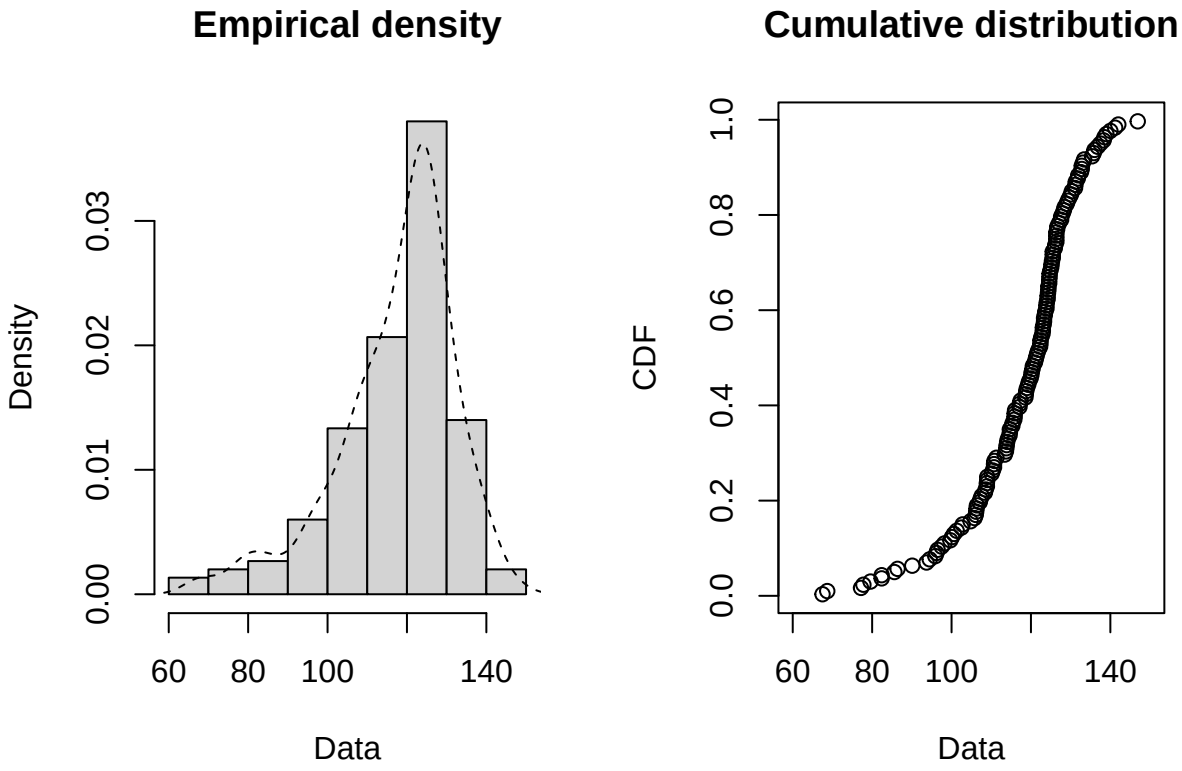

The graph below illustrates a histogram displaying the frequencies with respect to ndi. The data portrays that the relationship is right skewed. A log transformation is needed to normalize the distribution.

```
plotdist(filtered_data$ndi, histo = TRUE, demp = TRUE)
```



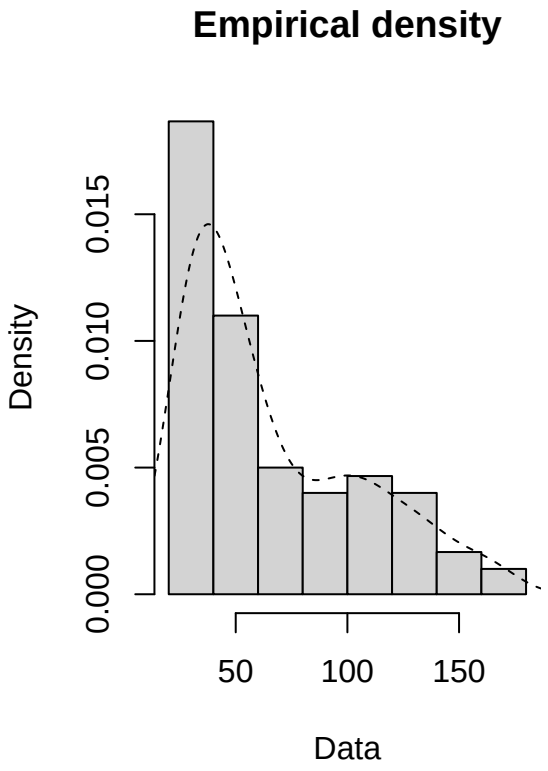

The graph below illustrates a histogram displaying the frequencies with respect to sales. The data portrays that the relationship is left skewed.

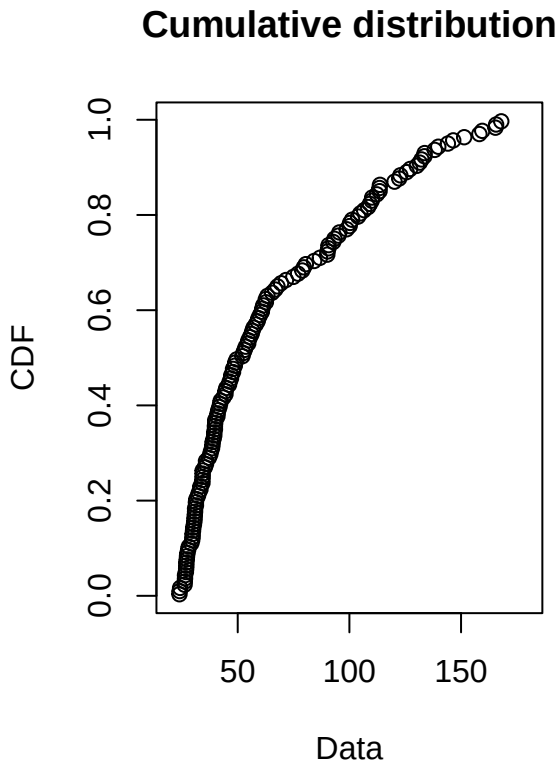
```
plotdist(filtered_data$sales, histo = TRUE, demp = TRUE)
```




The graph below illustrates a histogram displaying the frequencies with respect to pimin. The data portrays that the relationship is right skewed. A log transformation is needed to normalize the distribution.

```
plotdist(filtered_data$pimin, histo = TRUE, demp = TRUE)
```




Finally, we present the correlation matrix. The correlation between population and price is 0.3106. The correlation between the population over 16 and price is 0.3742. The correlation between cpi and price is 0.9477. The correlation between ndi and price is 0.9559. The correlation between the price and sales is -0.7448. The correlation between pimin and price is 0.99067.

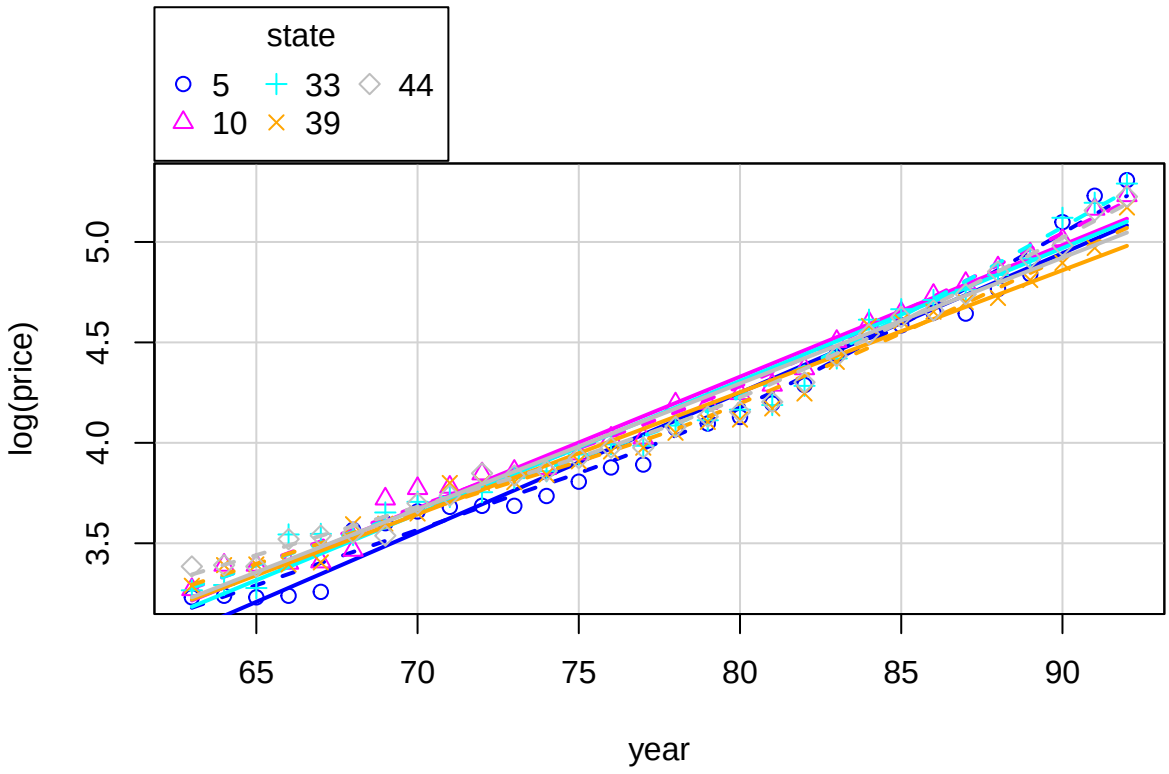
```
#Correlation Matrix
my_data <- filtered_data[, c(3,4,5,6,7,8,9)]
cor(my_data)
```

```
##            price        pop      pop16        cpi        ndi      sales
## price  1.0000000  0.3106000  0.3742244  0.9476577  0.9558833 -0.7447715
## pop    0.3106000  1.0000000  0.9905413  0.3270633  0.4211586 -0.3859187
## pop16  0.3742244  0.9905413  1.0000000  0.4077892  0.4986933 -0.3979319
## cpi    0.9476577  0.3270633  0.4077892  1.0000000  0.9849719 -0.6146341
## ndi    0.9558833  0.4211586  0.4986933  0.9849719  1.0000000 -0.6534560
## sales -0.7447715 -0.3859187 -0.3979319 -0.6146341 -0.6534560  1.0000000
## pimin  0.9906707  0.3407133  0.4050706  0.9643832  0.9718815 -0.7313841
##            pimin
## price  0.9906707
## pop    0.3407133
## pop16  0.4050706
## cpi    0.9643832
## ndi    0.9718815
## sales -0.7313841
## pimin  1.0000000
```

### 1(c).2 Data Visualization and Heterogeneity

In this section, we will explore heterogeneity across time (year) and states (id) as well as provide visualization of the variables against log of price in our dataset. Below, we have a scatterplot regressing the log of price against the year subject to the state. The graph below presents the differences or heterogeneity of price over time among the time-invariant variable that is states. We proceed to run the following commands below:

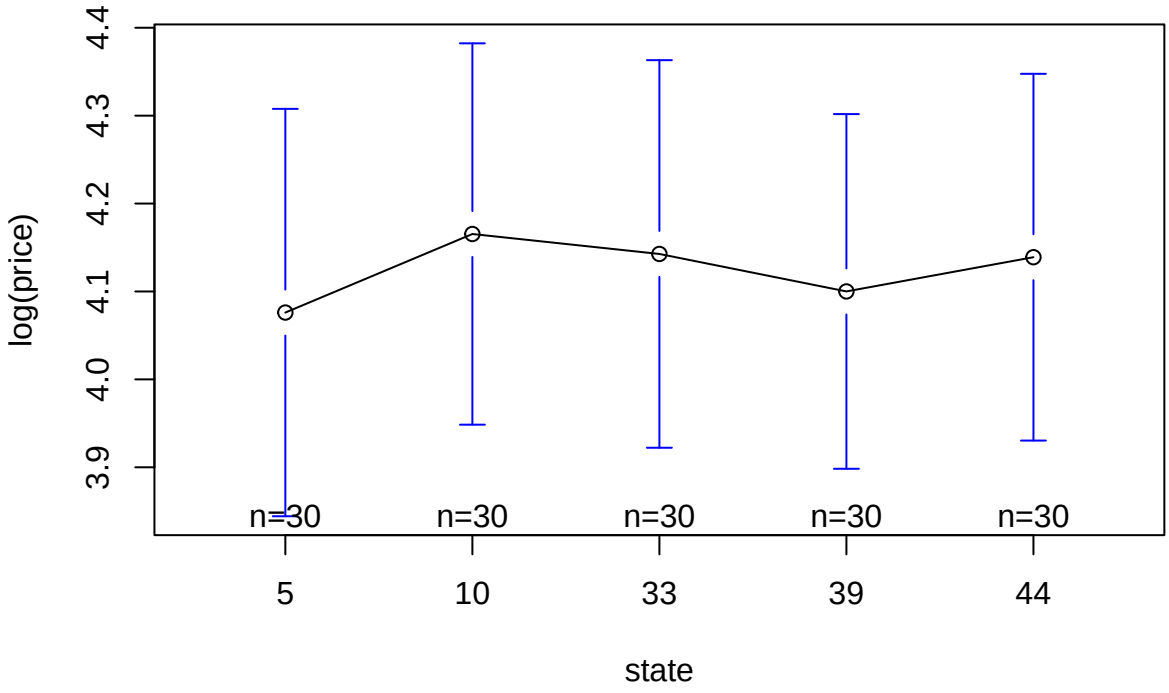
```
#Heterogeneity across Time
scatterplot(log(price) ~ year|state, data = filtered_data)
```




As depicted by the graph, there are differences among the intercepts of the states across time, implying that there is heterogeneity between the states with respect to time.

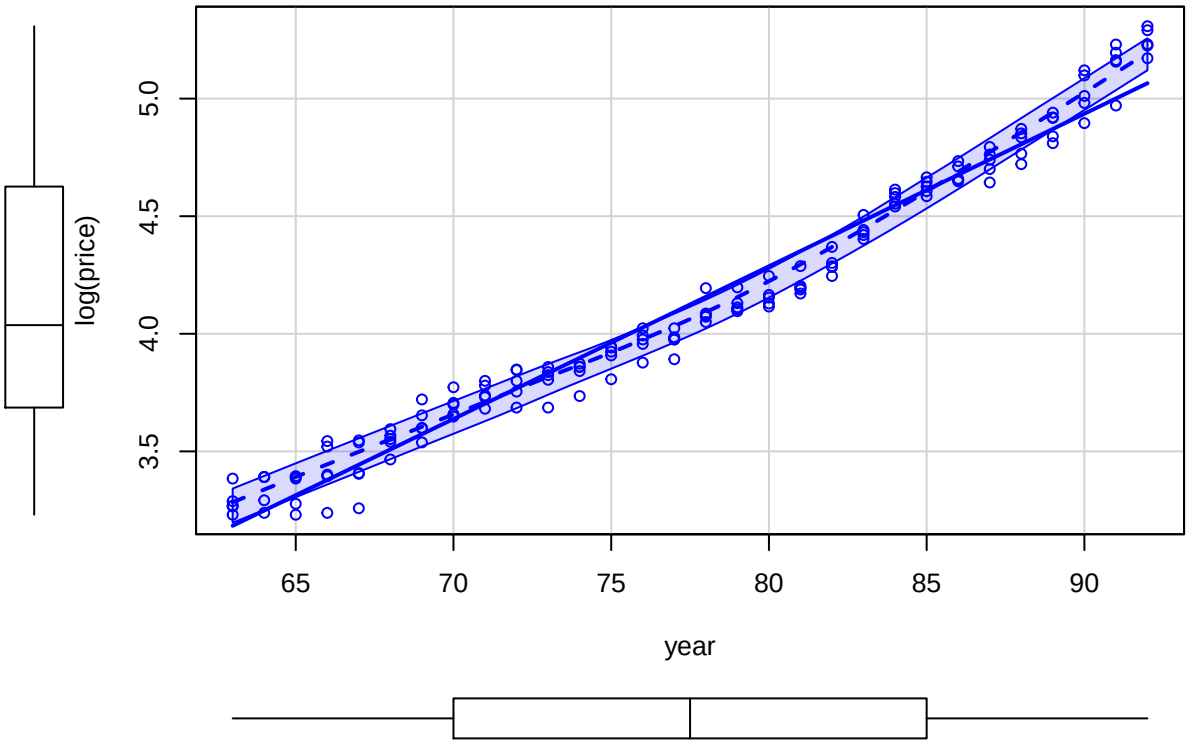
Next, we plot the means of log of price against the states. The graph below presents the differences or heterogeneity of average price against firms. We proceed to run the following commands below:

```
#Heterogeneity across Firms
plotmeans(log(price) ~ state, data = filtered_data)
```




As depicted by the plot above, there are differences and heterogeneity among the mean price across firms. Next, we graph a scatter plot of the log of price against the year.

```
scatterplot(log(price) ~ year, data = filtered_data)
```

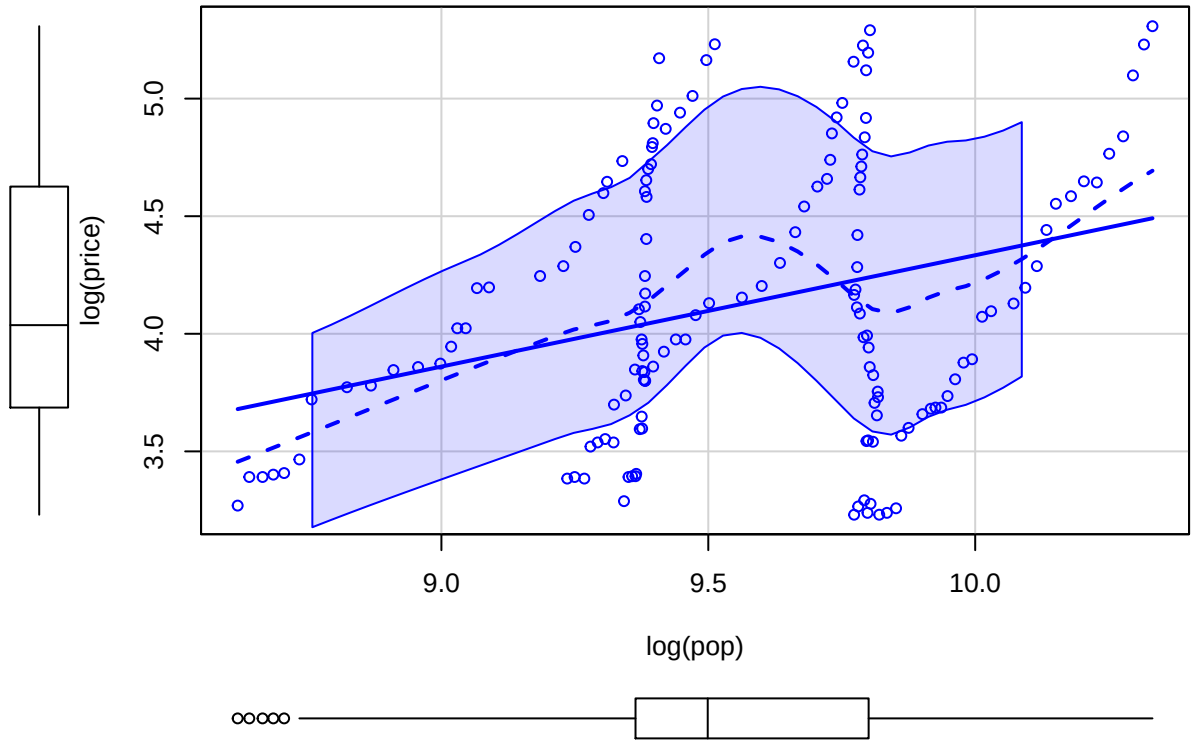



As depicted by the graph above, we see that the relationship between log of price and year has a positive linear relationship. We also see that the median log of price is approximately 4.0.

Finally, the rest of this section plots the relationship between price and the rest of the variables, which include "pop", "cpi", "ndi", "sales", and "pimin":

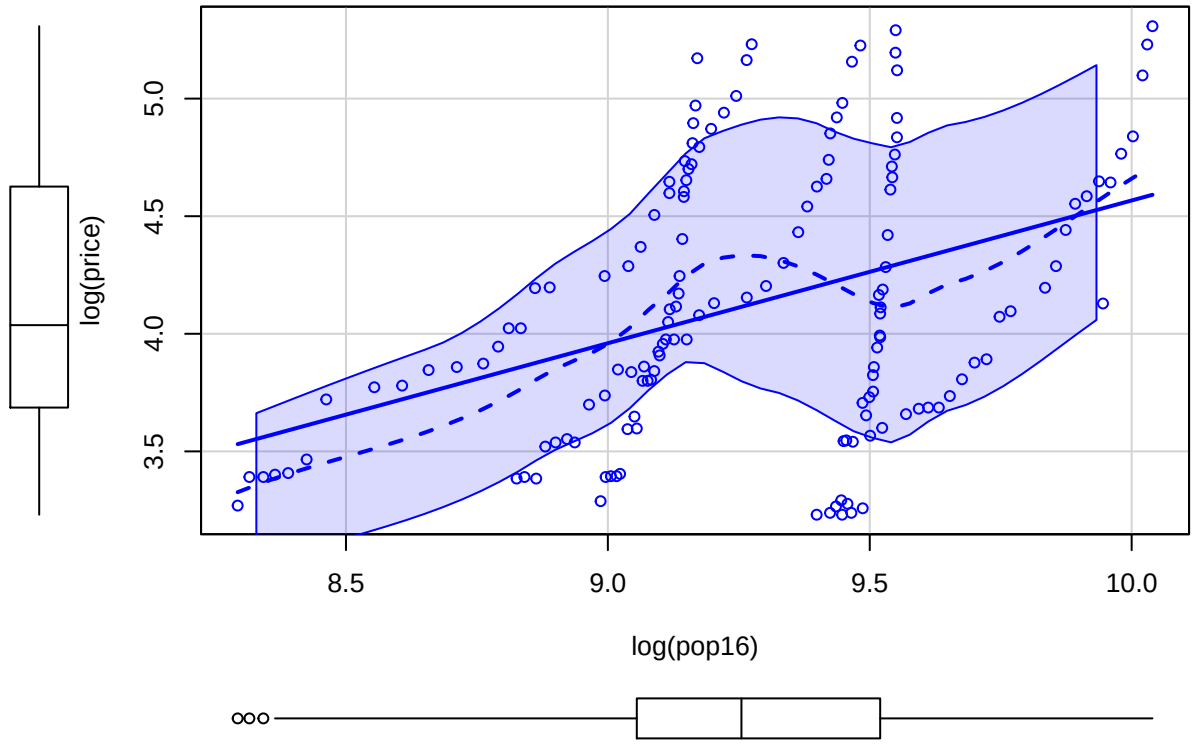
The following is a scatterplot of log(price) vs log(pop). The scatterplot indicates a positive but highly variable relationship between the two variables.

```
scatterplot(log(price) ~ log(pop), data = filtered_data)
```



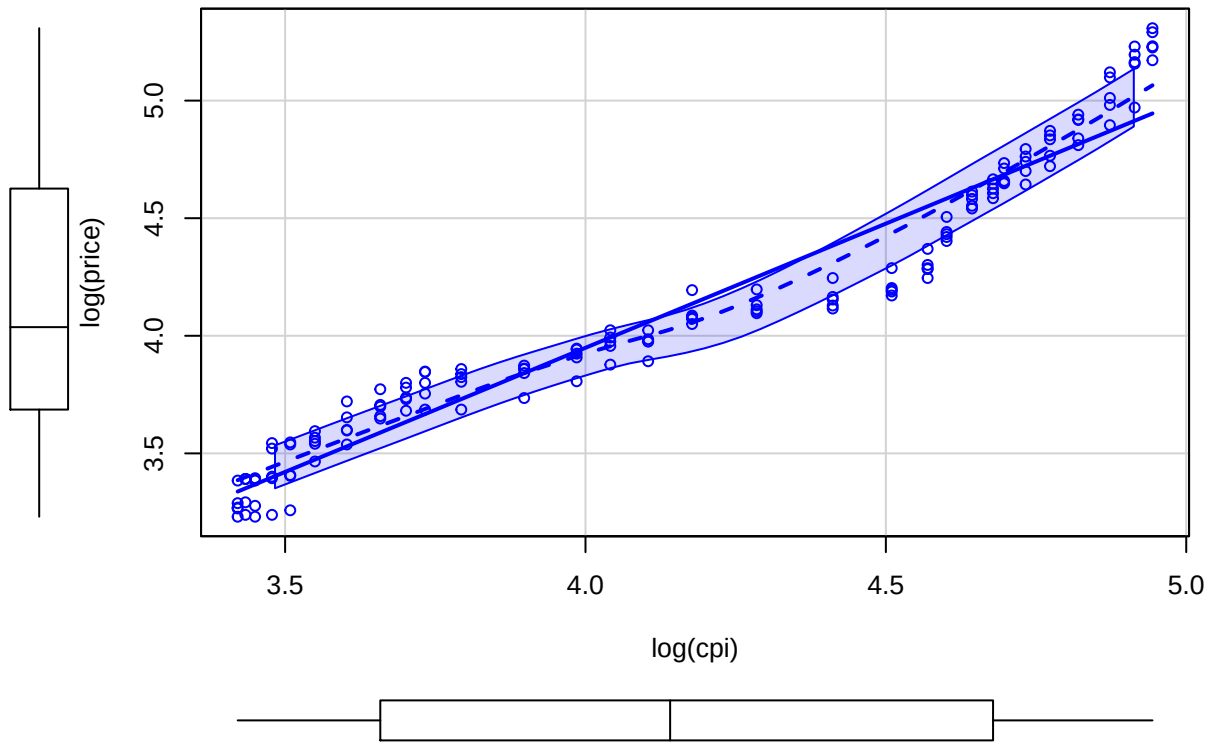

The following is a scatterplot of log(price) vs log(pop16). The scatterplot indicates a positive but highly variable relationship between the two variables.

```
scatterplot(log(price) ~ log(pop16), data = filtered_data)
```



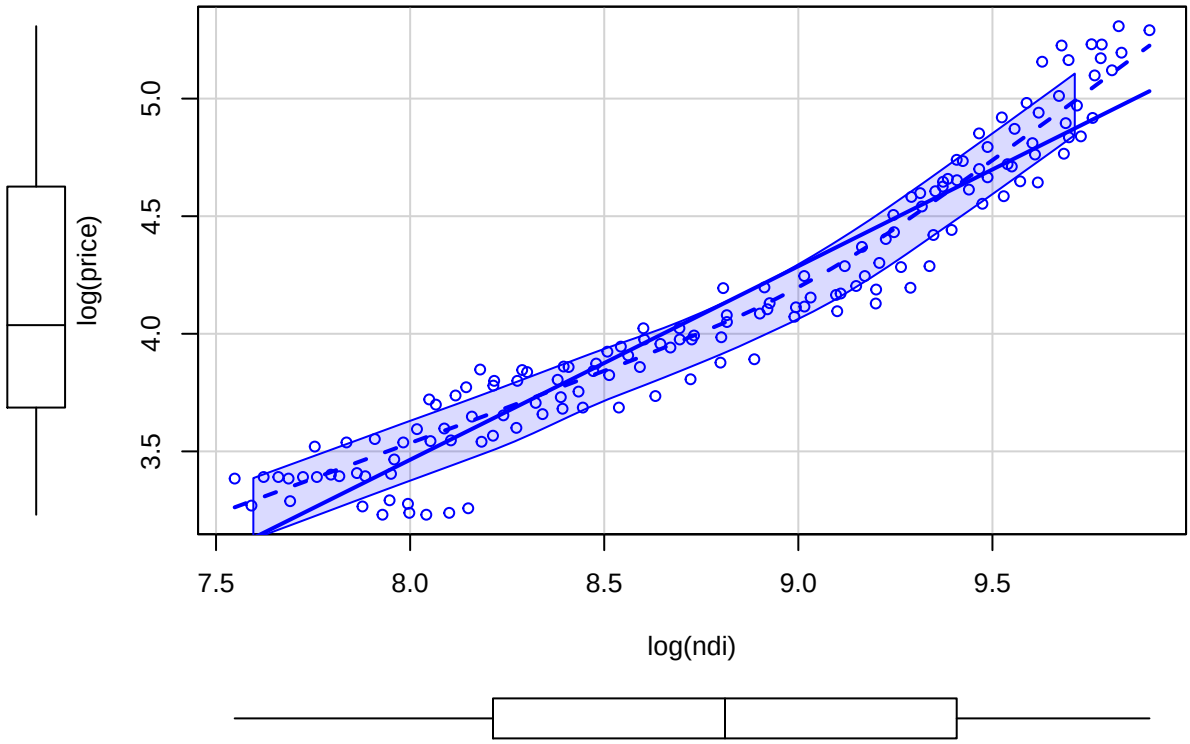

The following is a scatterplot of log(price) vs log(cpi). The scatterplot indicates a positive but tight relationship between the two variables.

```
scatterplot(log(price) ~ log(cpi), data = filtered_data)
```



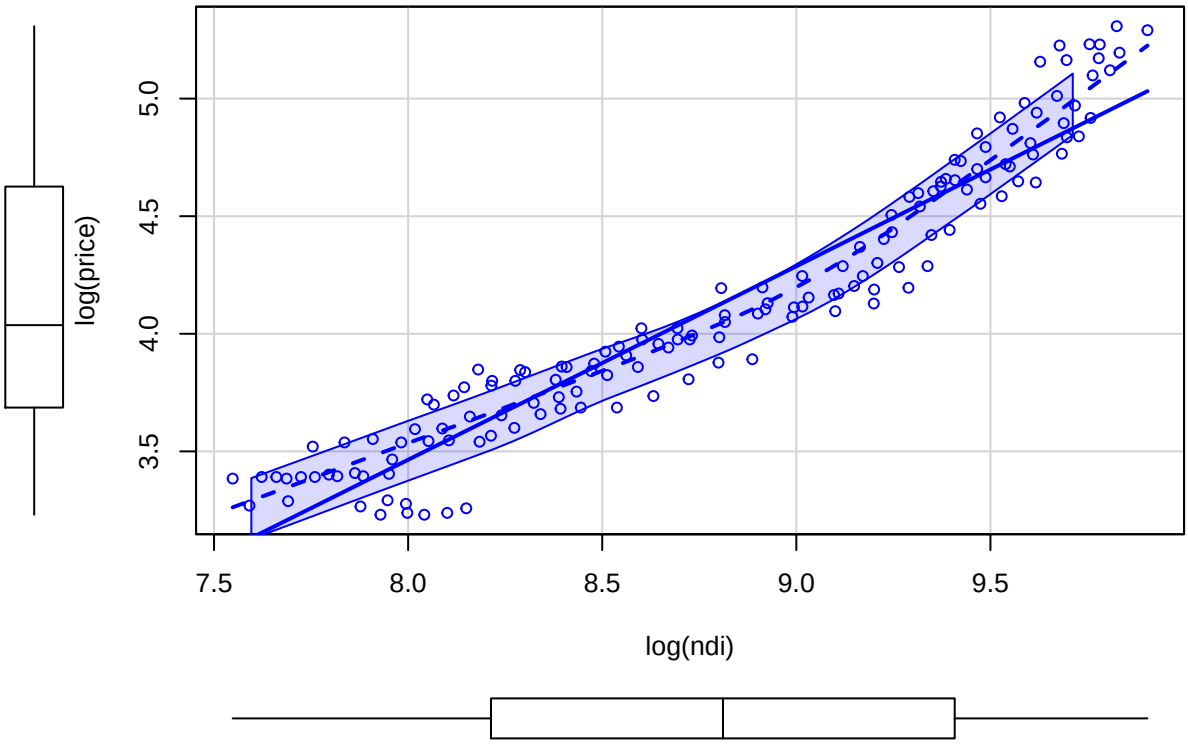

The following is a scatterplot of log(price) vs log(ndi). The scatterplot indicates a positive but tight relationship between the two variables.

```
scatterplot(log(price) ~ log(ndi), data = filtered_data)
```



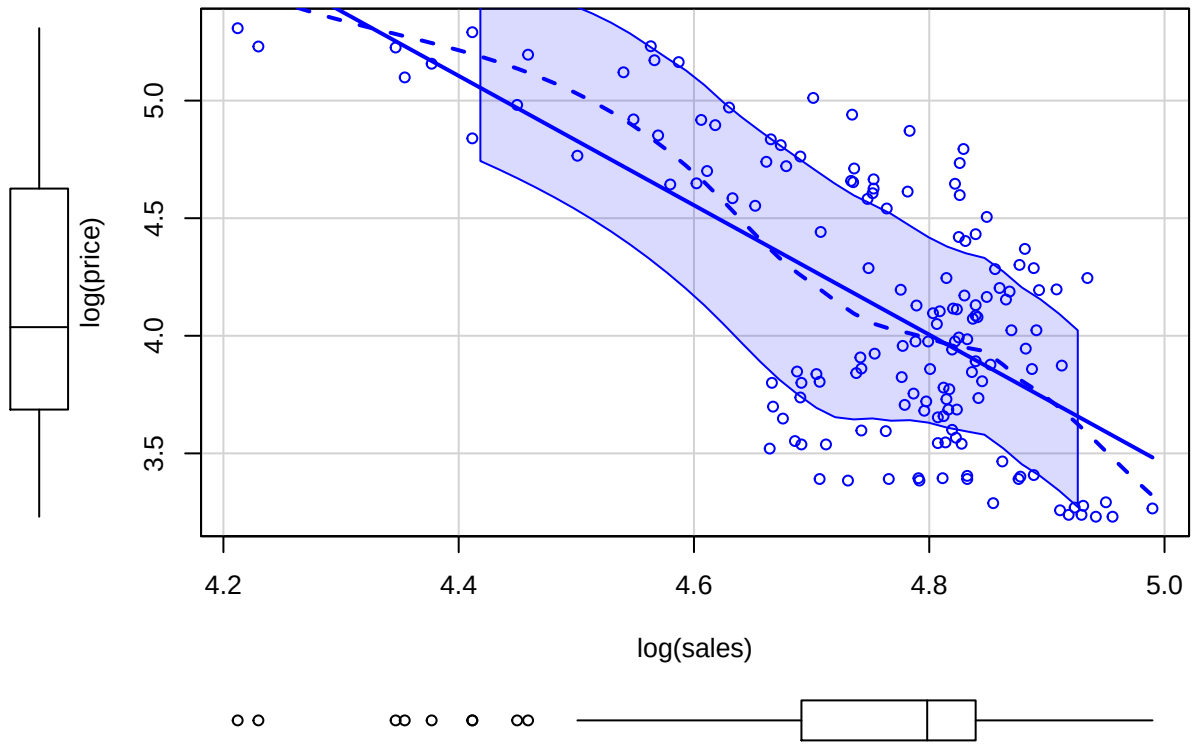

The following is a scatterplot of log(price) vs log(ndi). The scatterplot indicates a positive relationship between the two variables.

```
scatterplot(log(price) ~ log(ndi), data = filtered_data)
```



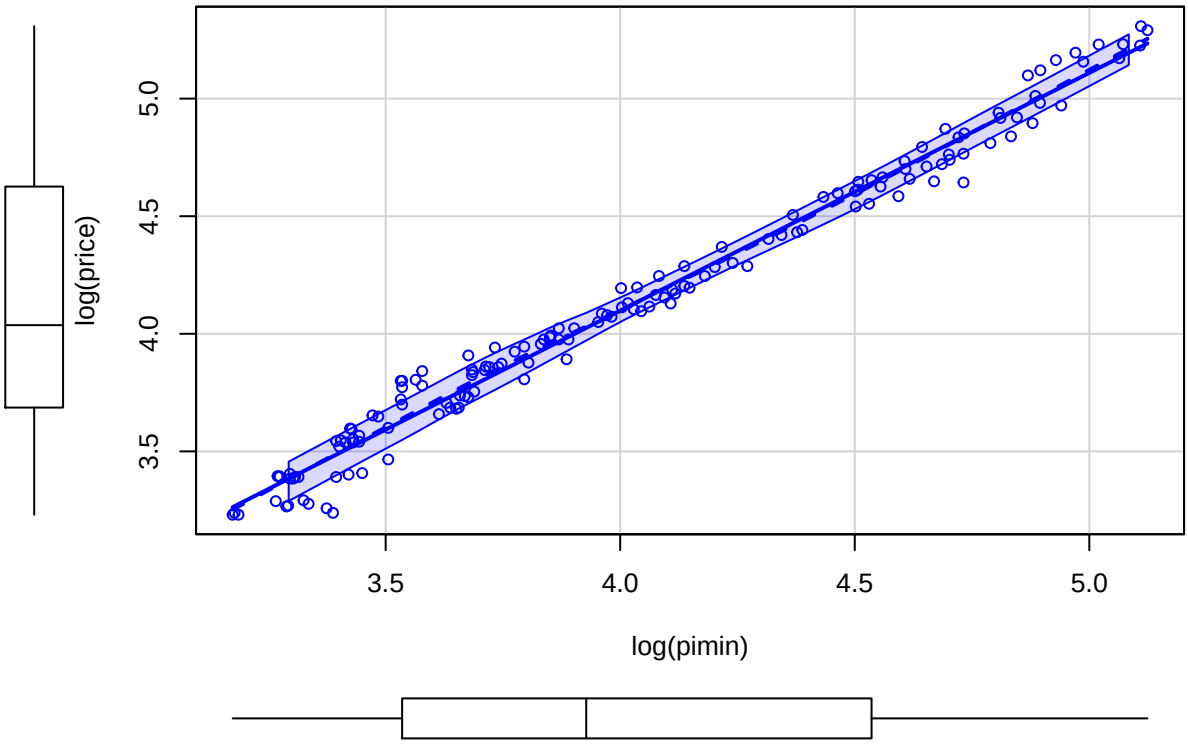

The following is a scatterplot of log(price) vs log(sales). The scatterplot indicates a positive but highly variable relationship between the two variables.

```
scatterplot(log(price) ~ log(sales), data = filtered_data)
```




The following is a scatterplot of log(price) vs log(pimin). The scatterplot indicates a positive but tight relationship between the two variables.

```
scatterplot(log(price) ~ log(pimin), data = filtered_data)
```




# 2 Panel Data Models: Cigar

## 2(a) Pooled Model

In this section, we will be creating a pooled model to model the question that we provided in section 1(a). We proceed to plot the pooled regression with the following commands below:

```
pooledreg1 <- plm(price~pop+pop16+cpi+ndi+sales+pimin,model="pooling",data=filtered_data)
crse<- coeftest(pooledreg1, vcov=vcovHC(pooledreg1,
type="HC0",cluster="group"))
stargazer(pooledreg1, crse, column.labels = c("\\textit{Pooled}", "\\textit{Pooled(prse)}"),
model.names = FALSE,type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##              ---------------------------------
##                         price
##                         Pooled          Pooled(prse)
##                          (1)                (2)
## ------------------------------------------------
## pop                     -0.001             -0.001
##                        (0.001)            (0.001)
```

```
##
## pop16                   0.001               0.001
##                        (0.001)             (0.001)
##
## cpi                    -0.156*              -0.156
##                        (0.092)             (0.125)
##
## ndi                     0.001               0.001
##                        (0.001)             (0.001)
##
## sales                  -0.144**            -0.144***
##                        (0.056)             (0.050)
##
## pimin                  1.162***            1.162***
##                        (0.082)             (0.072)
##
## Constant              24.775***           24.775***
##                        (8.201)             (7.720)
##
## ------------------------------------------------
## Observations             150
## R2                      0.984
## Adjusted R2             0.983
## F Statistic  1,464.068*** (df = 6; 143)
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

From the pooled regression model, we could note that cpi, sales, pimin, and the constant are statistically significant. However, pop, pop16, and ndi are not statistically significant. We also note that the R-squared is reasonably high with the value of 0.984 being between 0.5 and 0.99.

However, the estimated pooled model assumes that the intercept value between the states are the same. Recall our observation from 1(c).2 that the data exhibits heterogeneity across firms and across time. This indicates to us that we need to relax the first assumption that the slope coefficients are constant but the intercepts varies across the cross sectional units. Thus, in the subsequent models, we will relax this constraint by using the fixed effects and random effects model in order to better model our question that we presented in the beginning of the paper.

## 2(b) Fixed Effects Model

```
fixed_effects_model <- plm(price ~ pop + pop16 + cpi + ndi + sales + pimin,model =
                           "within", data = filtered_data)
summary(fixed_effects_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = price ~ pop + pop16 + cpi + ndi + sales + pimin,
##     data = filtered_data, model = "within")
##
## Balanced Panel: n = 5, T = 30, N = 150
##
```

```
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -21.18629  -1.99985   0.39021   2.40882  14.50521
##
## Coefficients:
##          Estimate  Std. Error t-value  Pr(>|t|)
## pop    0.00135628  0.00118281  1.1467    0.2535
## pop16 -0.00195075  0.00154140 -1.2656    0.2078
## cpi    0.10226444  0.13328282  0.7673    0.4442
## ndi   -0.00014308  0.00113119 -0.1265    0.8995
## sales -0.31236323  0.07063072 -4.4225 1.953e-05 ***
## pimin  0.99512082  0.08830076 11.2697 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    296220
## Residual Sum of Squares: 4034.2
## R-Squared:      0.98638
## Adj. R-Squared: 0.9854
## F-statistic: 1677.9 on 6 and 139 DF, p-value: < 2.22e-16
```

The dataset has a balanced panel with 5 entities (cross-sectional units), 30 time periods, and a total of 150 observations.The residuals (differences between observed and predicted values) have a minimum of -21.19 and a maximum of 14.51. Interpretations for significant variables: sales: For a one-unit increase in sales, the price decreases by approximately 0.3124 units. pimin: For a one-unit increase in pimin, the price increases by approximately 0.9951 units. sales and pimin are highly statistically significant (p-values $< 0.01$), indicating their strong impact on the dependent variable. The R-squared value is 0.9864, suggesting that the Fixed Effects model explains a substantial portion of the variation in the dependent variable. The F-statistic is 1677.9 with a very low p-value, indicating that the overall model is statistically significant. ## 2(c) Random Effects Model

```
cor(filtered_data[, c("price", "pop", "pop16", "cpi", "ndi", "sales", "pimin")])
```

```
##            price        pop      pop16        cpi        ndi      sales
## price  1.0000000  0.3106000  0.3742244  0.9476577  0.9558833 -0.7447715
## pop    0.3106000  1.0000000  0.9905413  0.3270633  0.4211586 -0.3859187
## pop16  0.3742244  0.9905413  1.0000000  0.4077892  0.4986933 -0.3979319
## cpi    0.9476577  0.3270633  0.4077892  1.0000000  0.9849719 -0.6146341
## ndi    0.9558833  0.4211586  0.4986933  0.9849719  1.0000000 -0.6534560
## sales -0.7447715 -0.3859187 -0.3979319 -0.6146341 -0.6534560  1.0000000
## pimin  0.9906707  0.3407133  0.4050706  0.9643832  0.9718815 -0.7313841
##            pimin
## price  0.9906707
## pop    0.3407133
## pop16  0.4050706
## cpi    0.9643832
## ndi    0.9718815
## sales -0.7313841
## pimin  1.0000000
```

Here we can see that there has high multicoleanirity between price~cpi, price~ndi, and price~pimin. Need to remove those three variables that are causing the problem.

```r
random_effects_model <-plm(price ~ pop + pop16 + sales,model = "random", data = filtered_data)
summary(random_effects_model)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = price ~ pop + pop16 + sales, data = filtered_data,
##     model = "random")
##
## Balanced Panel: n = 5, T = 30, N = 150
##
## Effects:
##                   var std.dev share
## idiosyncratic 459.13   21.43     1
## individual      0.00    0.00     0
## theta: 0
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -104.1053  -17.6907   -2.5142   14.2518   54.7440
##
## Coefficients:
##                Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept) 309.0180265  20.0136756  15.440 < 2.2e-16 ***
## pop          -0.0234984   0.0026029  -9.028 < 2.2e-16 ***
## pop16         0.0316058   0.0034298   9.215 < 2.2e-16 ***
## sales        -2.0442292   0.1441926 -14.177 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    296990
## Residual Sum of Squares: 83502
## R-Squared:      0.71884
## Adj. R-Squared: 0.71306
## Chisq: 373.281 on 3 DF, p-value: < 2.22e-16
```

The Random Effects model provides a reasonable fit to the data, considering the significant variables and the explained variation in 'price.' The inclusion of individual effects helps capture entity-specific variability, addressing the heterogeneity observed in the dataset.

## 2(d) Which Model Should We Use?

1. Pooled Model: R-Squared: 0.984 Adjusted R-Squared: 0.983 F-Statistic: 1,464.068*** (p-value $< 0.01$) Significant Coefficients: cpi, sales, pimin (at various significance levels)
2. Fixed Effects Model: R-Squared: 0.98638 Adjusted R-Squared: 0.9854 F-Statistic: 1,677.9*** (p-value $<$ 2.22e-16) Significant Coefficients: pop, pop16, sales, pimin (at various significance levels)
3. Random Effects Model: R-Squared: 0.71884 Adjusted R-Squared: 0.71306 Chisq: 373.281 (p-value $<$ 2.22e-16) Significant Coefficients: Intercept, pop, pop16, sales (at various significance levels)

Based on the results and considerations, the Fixed Effects Model seems to be the preferred choice in this scenario. It has a higher Adjusted R-squared, a significant F-statistic, and all coefficients are significant.

# 3 Conclusion: Cigar

The Fixed Effects Model seems to provide the best fit (highest Adjusted R-squared). The Pooled Model performs well but may overlook individual-specific effects. The Random Effects Model, while capturing individual-specific effects, has a lower overall fit compared to the Fixed Effects Model.

Preferred Model: Fixed Effects Model High Adjusted R-squared. Captures individual-specific effects. Addresses potential heterogeneity across entities.

---

# 4 Binary Dependent Variables: Credit Card

## 4(a) Question

```
data("CreditCard")
```

For the purposes of this project, we will ask "Which of the following variables : (1) report, (2)age, (3) income, (4)expenditure, (5)dependents, (6) months, (7) active, have the largest impact on the whether someone will obtain a new Credit Card?"

In section 2, we will express our question more precisely using the following models: (1) Linear probability model, (2) Probit Model, (3) Logit model, and we will determine which model is best fit to answer our question.

## 4(b) Data Set Summary

In this paper, we will be analyzing the Credit Card Data Set from the PLM Package. Credit Card data in the United States. The dataset contains panel data with the following twelve variables:

(1) card: Factor. Was the application for a credit card accepted?

(2) reports: Number of major derogatory reports

(3) age: Age in years plus twelfths of a year

(4) income: Yearly income (in USD 10,000)

(5) share: Ratio of monthly credit card expenditure to yearly income.

(6) expenditure: Average monthly credit card expenditure

(7) owner: Factor. Does the individual own their home?

(8) selfemp: Factor. Is the individual self-employed?

(9) dependents: Number of dependents

(10) months: Months living at current address.

(11) majorcards: Number of major credit cards held

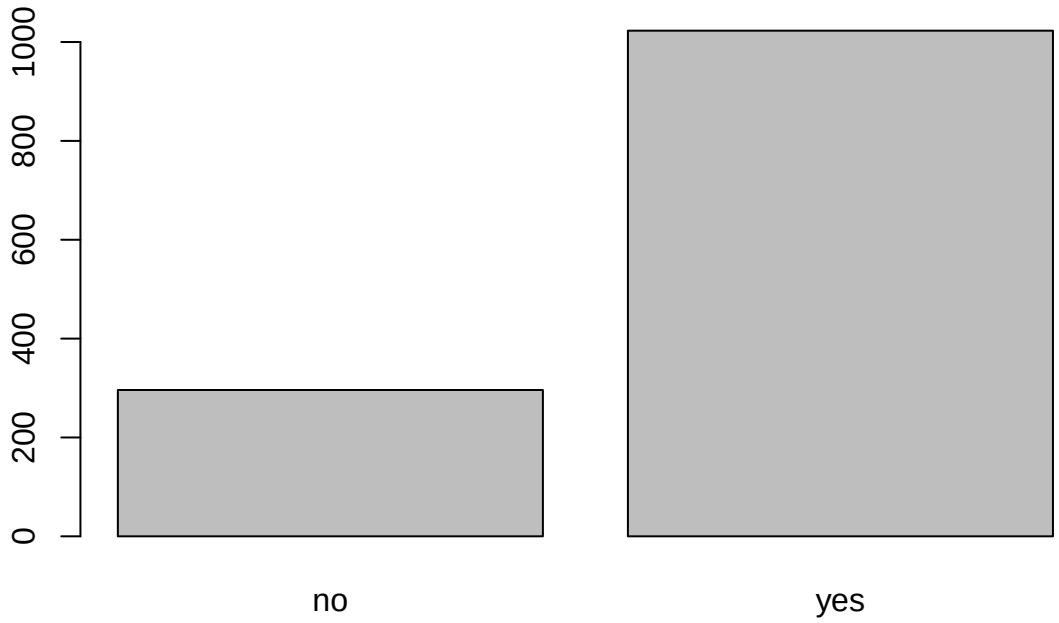
(12) active: Number of active credit accounts

## 4(c) Variable Description

### 4(c).1 Graphs

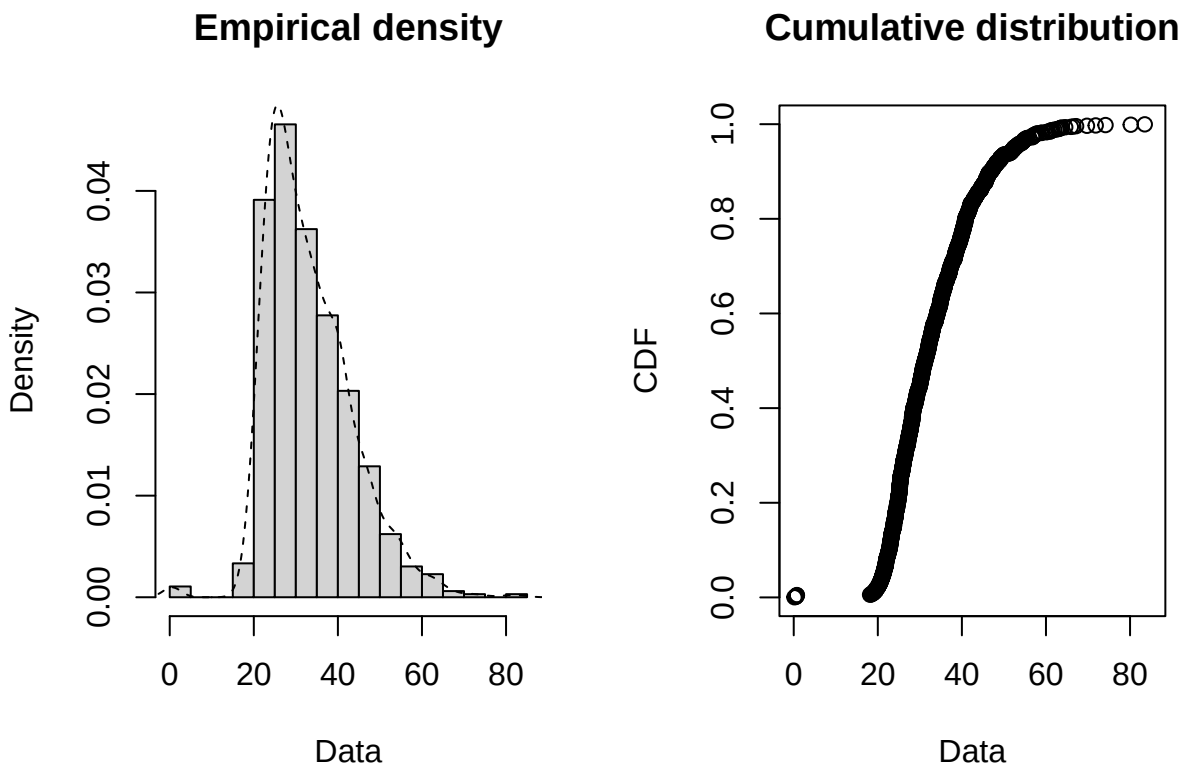
The graph below illustrates a histogram displaying the frequencies with respect to price. The graph portrays the fact that the majority of our observations for will fall between approximately between 0 to 50 gallons of consumption. The data also appears to right-skewed, indicating that we would need to perform a log transformation in order to normalize the data.

```
card <- table(CreditCard$card)
barplot(card)
```



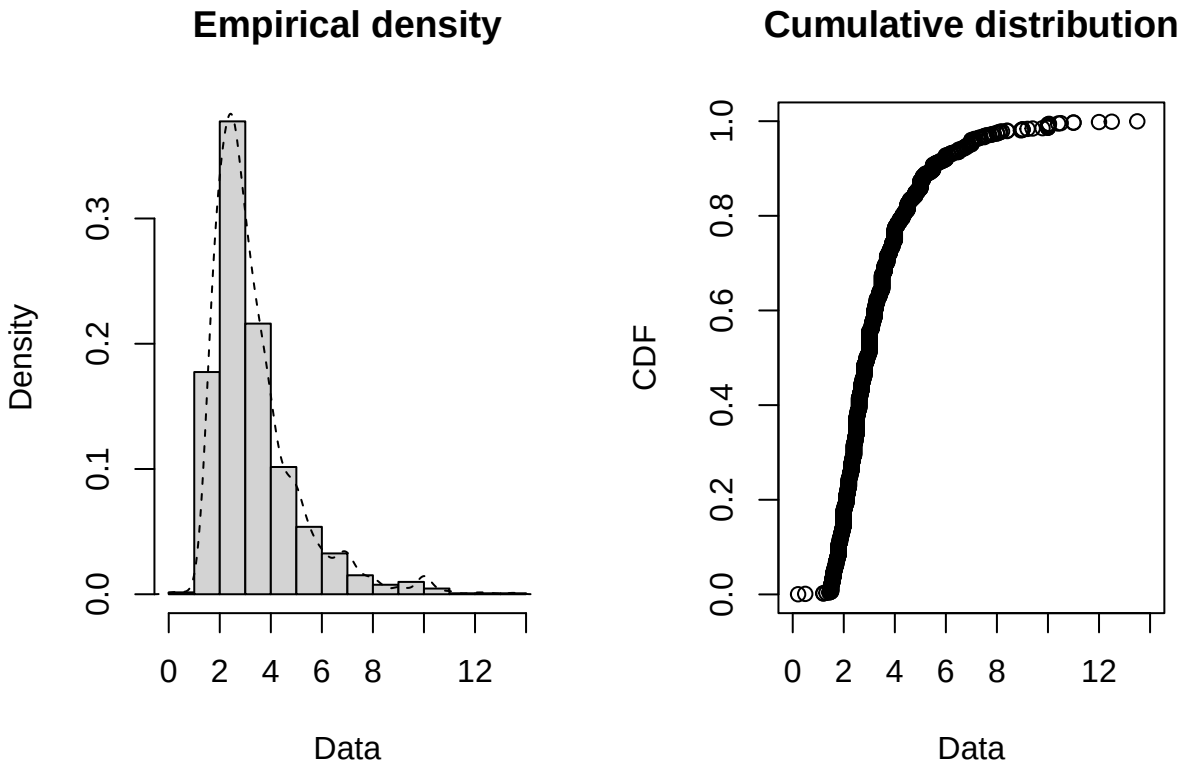

```
CreditCard$card<-(CreditCard$card=="yes")*1
```

```
#Change the variable of "Age"
plotdist(CreditCard$age, histo = TRUE, demp = TRUE)
```



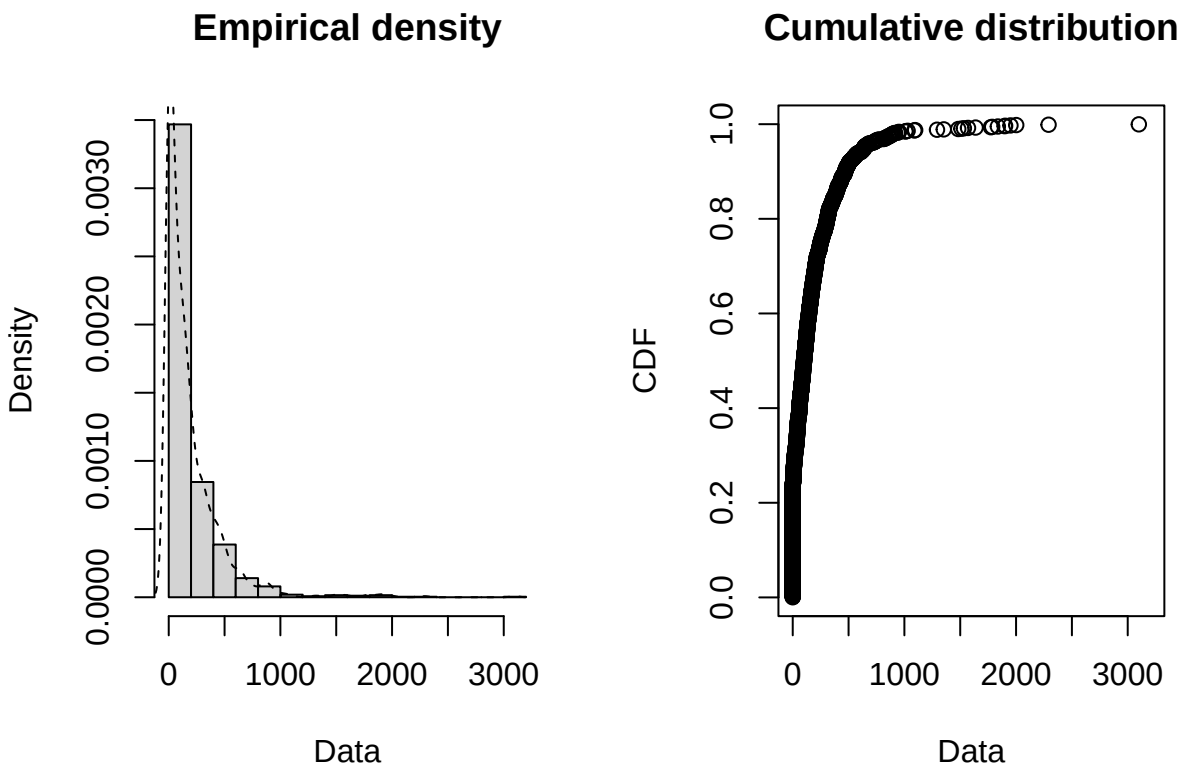

The graph above illustrates a histogram displaying the frequencies with respect to age. The data portrays that the relationship is left skewed.

```
plotdist(CreditCard$income, histo = TRUE, demp = TRUE)
```



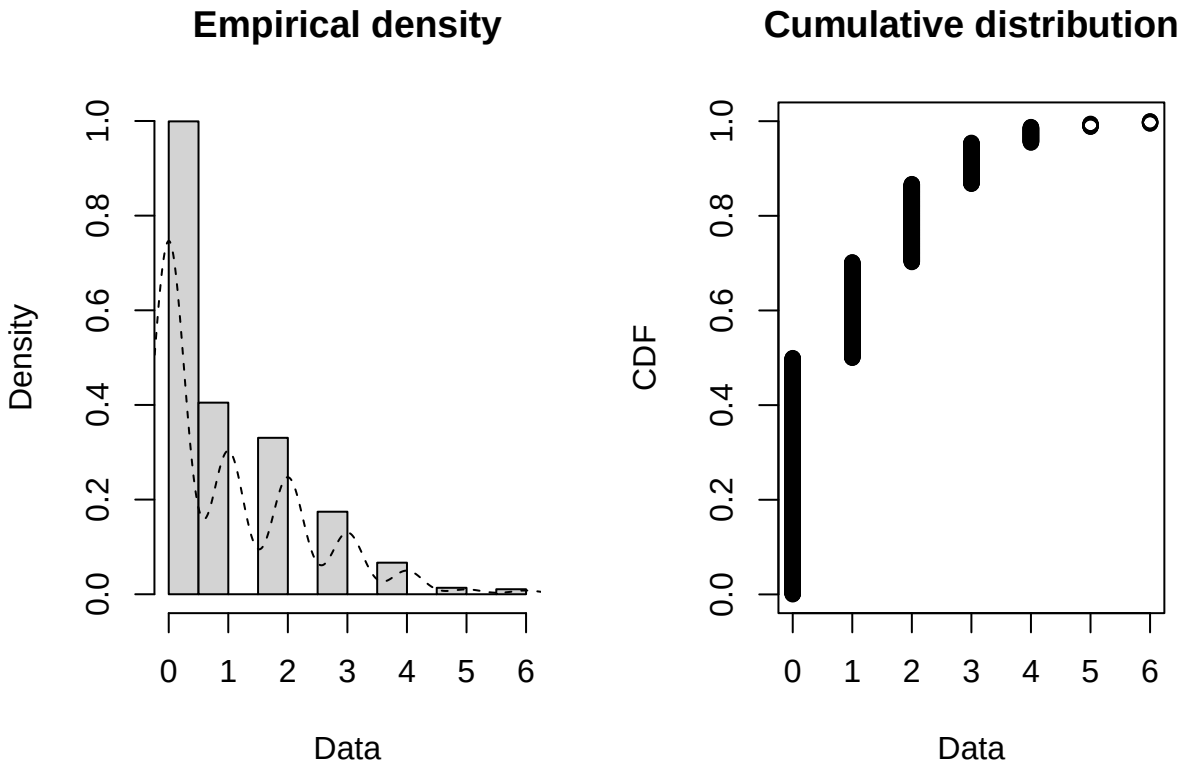

The graph above illustrates a histogram displaying the frequencies with respect to income (USD 10,000). The data portrays that the relationship is right skewed

```
plotdist(CreditCard$expenditure, histo = TRUE, demp = TRUE)
```



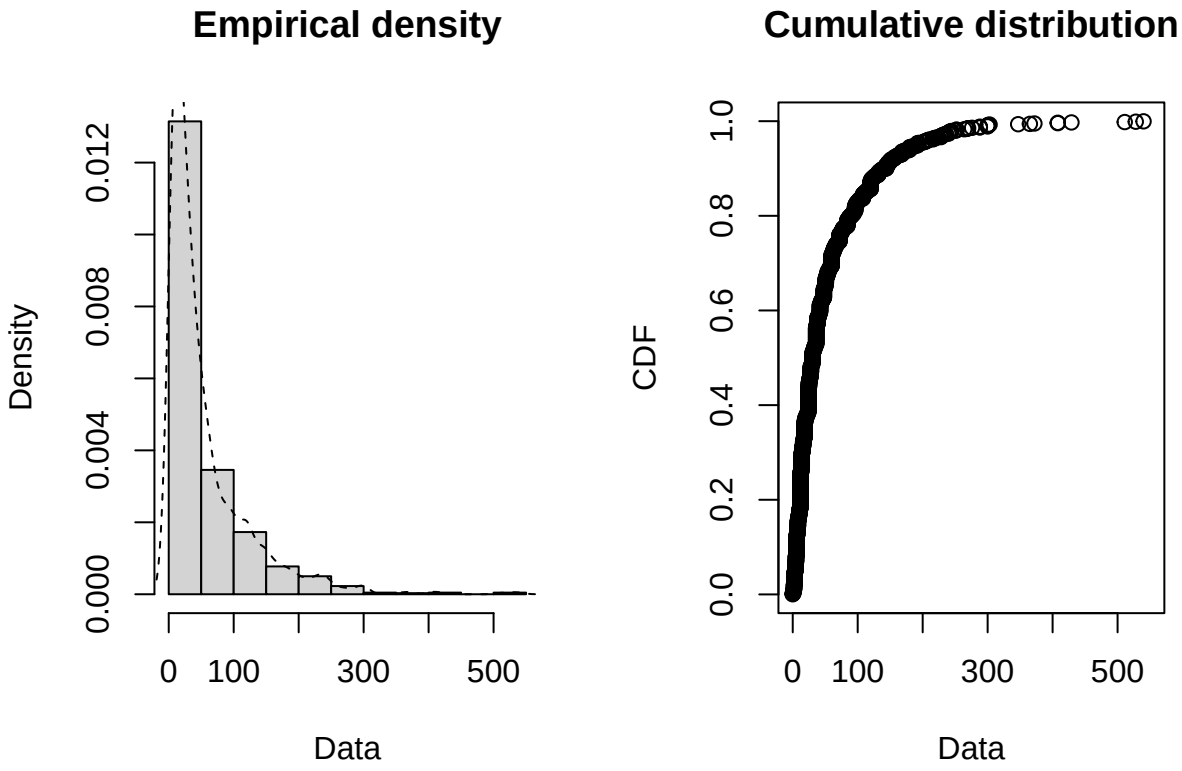

The graph above illustrates a histogram displaying the frequencies with respect to expenditure. The data portrays that the relationship is right skewed.

```
plotdist(CreditCard$dependents, histo = TRUE, demp = TRUE)
```



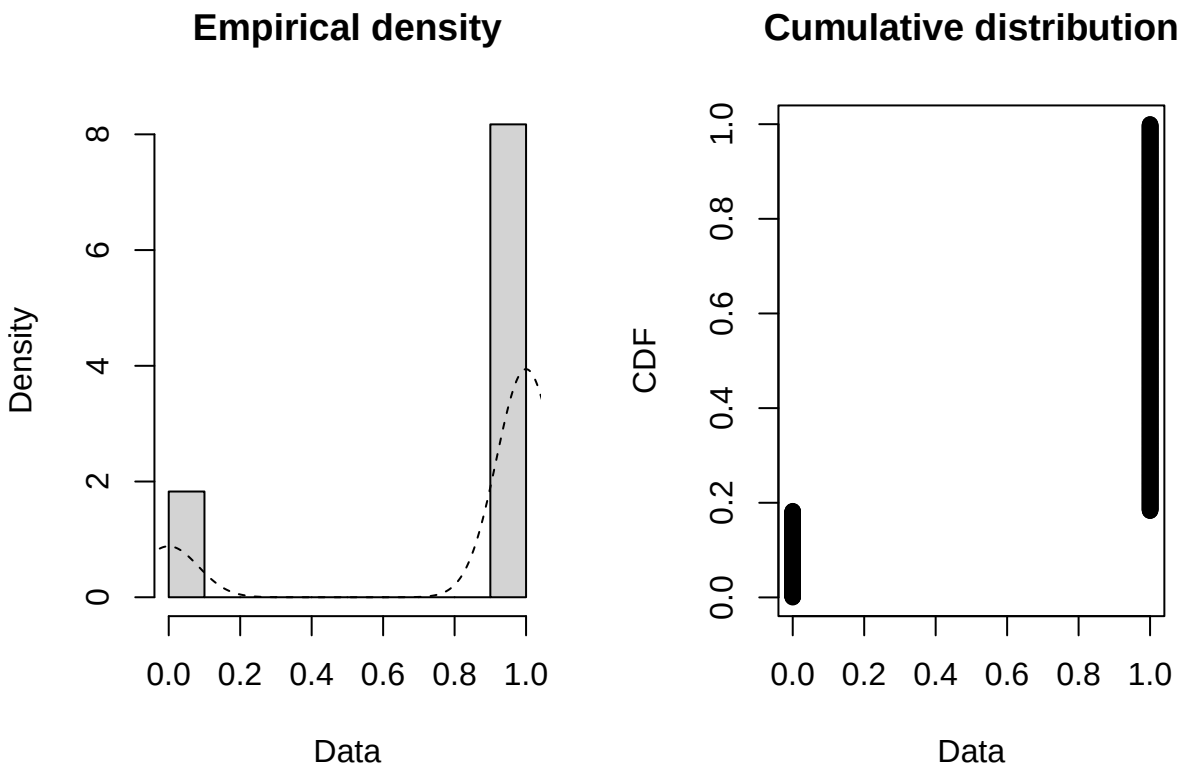

The graph above illustrates a histogram displaying the frequencies with respect to dependents. The data portrays that the relationship is right skewed.

```
plotdist(CreditCard$months, histo = TRUE, demp = TRUE)
```



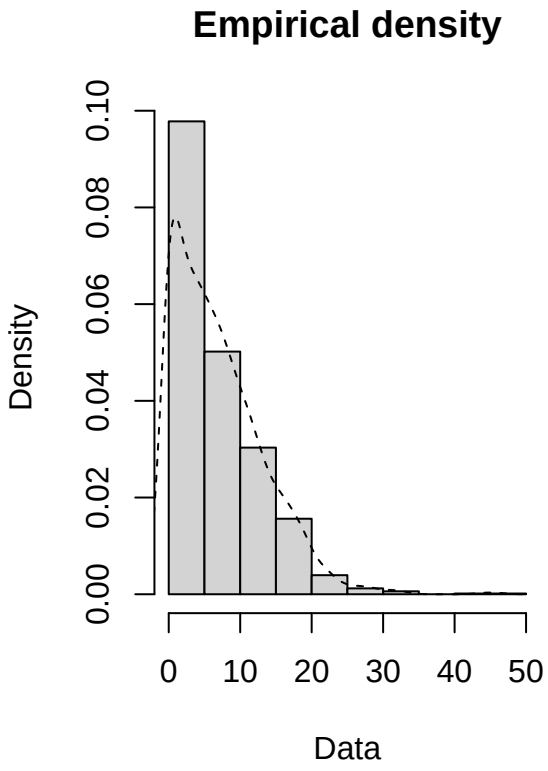

The graph above illustrates a histogram displaying the frequencies with respect to months. The data portrays that the relationship is right skewed. A log transformation is needed to normalize the distribution.

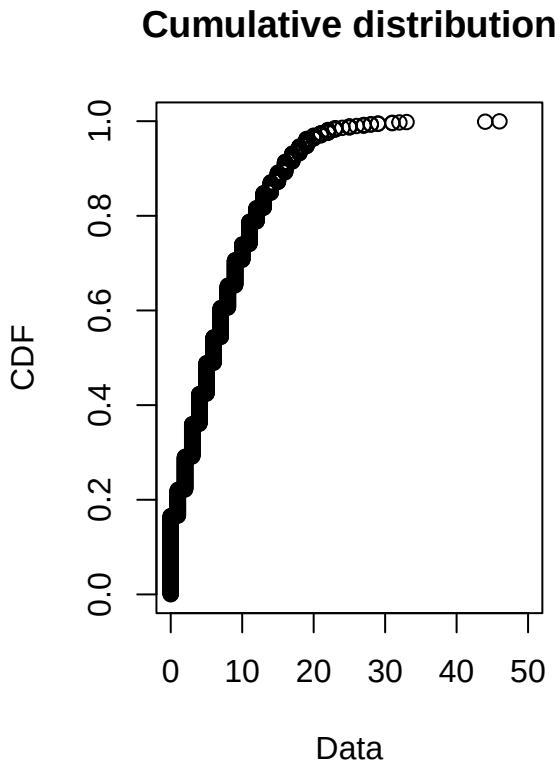
```
plotdist(CreditCard$majorcards, histo = TRUE, demp = TRUE)
```




The graph above illustrates a histogram displaying the frequencies with respect to months. The data portrays that the relationship has two peak.

```
plotdist(CreditCard$active, histo = TRUE, demp = TRUE)
```




The graph above illustrates a histogram displaying the frequencies with respect to months. The data portrays that the relationship is right skewed.

Finally, we present the correlation matrix. The relationships between each variable in our dataset is presented in the matrix below:

```
#Correlation Matrix
data1 <- CreditCard[, c(3,4,5,6,9,10,12)]
cor(data1)
```

```
##                    age      income       share expenditure  dependents
## age          1.0000000  0.32465320 -0.11569704  0.01494770  0.21214643
## income       0.3246532  1.00000000 -0.05442926  0.28110402  0.31760130
## share       -0.1156970 -0.05442926  1.00000000  0.83877932 -0.08261776
## expenditure  0.0149477  0.28110402  0.83877932  1.00000000  0.05266406
## dependents   0.2121464  0.31760130 -0.08261776  0.05266406  1.00000000
## months       0.4364255  0.13034627 -0.05534756 -0.02900660  0.04651197
## active       0.1810697  0.18054026 -0.02347440  0.05472424  0.10713276
##                  months      active
## age          0.43642554  0.18106971
## income       0.13034627  0.18054026
## share       -0.05534756 -0.02347440
## expenditure -0.02900660  0.05472424
## dependents   0.04651197  0.10713276
## months       1.00000000  0.10002764
## active       0.10002764  1.00000000
```

# 5 Binary Dependent Variables: Credit Card

## 5(a) Linear Probability Model

The Linear Dependent Variable Model is presented below:

```
linear_model <- lm(card ~ reports + age + income + expenditure + dependents + months
                   + active, data = CreditCard)

summary(linear_model)
```

```
##
## Call:
## lm(formula = card ~ reports + age + income + expenditure + dependents +
##     months + active, data = CreditCard)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33609 -0.08292  0.13885  0.22647  1.09368
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7028715  0.0347059  20.252  < 2e-16 ***
## reports     -0.1380140  0.0072762 -18.968  < 2e-16 ***
## age         -0.0004990  0.0011041  -0.452   0.6514
## income       0.0016318  0.0064430   0.253   0.8001
## expenditure  0.0004574  0.0000368  12.431  < 2e-16 ***
## dependents  -0.0203807  0.0080771  -2.523   0.0117 *
## months       0.0001337  0.0001591   0.840   0.4009
## active       0.0107197  0.0015756   6.804 1.55e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3435 on 1311 degrees of freedom
## Multiple R-squared:  0.3263, Adjusted R-squared:  0.3227
## F-statistic: 90.72 on 7 and 1311 DF,  p-value: < 2.2e-16
```

```
AIC(linear_model)
```

```
## [1] 934.0028
```

```
BIC(linear_model)
```

```
## [1] 980.6645
```

## 5(b) Probit model

The Probit Model is presented below:

```
probit_model <- glm(card ~ reports + age + income + expenditure + dependents +
                      months + active, family = binomial(link = "probit"),
                    data = CreditCard)

summary(probit_model)
```

```
##
## Call:
## glm(formula = card ~ reports + age + income + expenditure + dependents +
##     months + active, family = binomial(link = "probit"), data = CreditCard)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.542490   0.428702  -3.598 0.000321 ***
## reports     -1.107742   0.418944  -2.644 0.008190 **
## age          0.018571   0.012631   1.470 0.141504
## income      -0.046363   0.094617  -0.490 0.624131
## expenditure 13.243780  23.951130   0.553 0.580298
## dependents  -0.352375   0.148957  -2.366 0.018000 *
## months      -0.001534   0.002133  -0.719 0.472016
## active       0.050354   0.019247   2.616 0.008890 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1404.57  on 1318  degrees of freedom
## Residual deviance:  115.88  on 1311  degrees of freedom
## AIC: 131.88
##
## Number of Fisher Scoring iterations: 25
```

```
BIC(probit_model)
```

```
## [1] 173.3591
```

## 5(c) Logit Model

The Logit Model is presented below:

```
logit_model <- glm(card ~ reports + age + income + expenditure + dependents + months
                   + active, family = binomial(link="logit"), data = CreditCard)

summary(logit_model)
```

```
##
## Call:
## glm(formula = card ~ reports + age + income + expenditure + dependents +
##     months + active, family = binomial(link = "logit"), data = CreditCard)
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.719551   0.794143  -3.425 0.000616 ***
## reports      -2.395237   1.007250  -2.378 0.017407 *
## age           0.033564   0.022315   1.504 0.132550
## income       -0.076170   0.179790  -0.424 0.671813
## expenditure  33.030879 159.800610   0.207 0.836244
## dependents   -0.666992   0.300960  -2.216 0.026677 *
## months       -0.002594   0.003883  -0.668 0.504153
## active        0.094839   0.035220   2.693 0.007085 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1404.57  on 1318  degrees of freedom
## Residual deviance:  116.12  on 1311  degrees of freedom
## AIC: 132.12
##
## Number of Fisher Scoring iterations: 25
```

```
BIC(logit_model)
```

```
## [1] 173.5977
```

## 5(d) Which Model Should We Use?

The probit is the better model with the lowest AIC and BIC of 131.88 and 173.3591 respectively, compared to the linear model (AIC = 934.0028, BIC = 980.6645) and the logit Model (AIC = 132.12, BIC = 173.5977).

## 6 Which Model Should We Use?

All three models had reports, dependents, and actives as statically significant. Only the linear model had expenditure as significant. Reports had the largest effect that is statically significant. This makes sense since a history of delinquency would serious hinder someone getting a credit card. Our group thought it was interesting that income and expenditure did not have a significant role in determining whether someone got approved or not, since a person with a high income better able pay of their credit card statement and since a person who spends a lot who probably have a great risk of spending too much on their credit card. However, the other variables are better suited at predicting the truth.

Thus, our recommended model would be $Pr[Card = 1] = \Phi[-1.54 - 1.11 * reports + 0.02 * age - 0.05 * income + 13.24 * expenditure - 0.35 * dependents - 0.002 * months + 0.05 * active]$