Project 3

Juliesen Jaime, Rachel Asher, Caden Campbell

2023-11-29

Panel Data Model (a) Briefly discuss your data and the question you are trying to answer with your model.

We will be using the Cigar dataset from the PLM package. This is a panel dataset which includes 46 observations from 1963 to 1992. The data is relatively long and wide since there are 50 states and 46 observations. We will be observing states, year, price, population, CPI (consumer price index), and NDI (per capita disposable income), and sale variables in our data across 5 states from 1978 to 1992.

Question: How does the sale of cigars differ across states from 1978 to 1992?

```r
data("Cigar")
summary(Cigar)
```

```
##      state            year          price             pop
##  Min.   : 1.00   Min.   :63.0   Min.   : 23.40   Min.   :  319
##  1st Qu.:15.00   1st Qu.:70.0   1st Qu.: 34.77   1st Qu.: 1053
##  Median :26.50   Median :77.5   Median : 52.30   Median : 3174
##  Mean   :26.83   Mean   :77.5   Mean   : 68.70   Mean   : 4537
##  3rd Qu.:40.00   3rd Qu.:85.0   3rd Qu.: 98.10   3rd Qu.: 5280
##  Max.   :51.00   Max.   :92.0   Max.   :201.90   Max.   :30703
##      pop16             cpi             ndi            sales
##  Min.   :  215.2   Min.   : 30.6   Min.   : 1323   Min.   : 53.4
##  1st Qu.:  781.2   1st Qu.: 38.8   1st Qu.: 3328   1st Qu.:107.9
##  Median : 2315.3   Median : 62.9   Median : 6281   Median :121.2
##  Mean   : 3366.6   Mean   : 73.6   Mean   : 7525   Mean   :124.0
##  3rd Qu.: 3914.3   3rd Qu.:107.6   3rd Qu.:11024   3rd Qu.:133.2
##  Max.   :22920.0   Max.   :140.3   Max.   :23074   Max.   :297.9
##      pimin
##  Min.   : 23.40
##  1st Qu.: 31.98
##  Median : 46.40
##  Mean   : 62.90
##  3rd Qu.: 90.50
##  Max.   :178.50
```

```r
# Filter observations
filtered_Cigar <- Cigar %>%
  filter(state > 46)
filtered_Cigar <- filtered_Cigar %>%
  filter(year > 78)
head(filtered_Cigar)
```

```
##      state year price  pop  pop16   cpi       ndi sales pimin
## 1247    47   79  45.8 5197 3993.1  72.6  7396.398 151.8  43.4
## 1248    47   80  48.5 5346 4100.0  82.4  8230.751 148.9  46.3
## 1249    47   81  51.8 5430 4187.7  90.9  9028.556 149.9  49.4
## 1250    47   82  56.4 5491 4247.4  96.5  9754.307 147.4  56.3
## 1251    47   83  68.8 5550 4310.1  99.6 10561.510 144.7  66.4
## 1252    47   84  76.0 5636 4391.0 103.9 11529.526 136.8  75.4
```

(b) Provide a descriptive analysis of your variables. This should include relevant figures with comments including some graphical depiction of individual heterogeneity.

Our first variable is price. The histogram shows that price is rightly skewed and thus we need to log the variable since it is a lognormal distribution. The boxplot for price shows that price is skewed slightly to the right. The median of the price data across states is 102.40. The range is 153.4. The standard deviation 37.04398. Price is autoregressive. There is a clear upward trend in the plot means. Across states, however, price appears to be homoskedastic

Our second variable is sales. The histogram als depicts that it is rightly skewed so we used the log(sales). The boxplot for sales depicts the median of the data to be 110.6. The range is 119.9 and the standard deviation is 20.2686. Sales is autoregressive. There is a clear downward trend in the plot means. Across states, it is also heteroskedastic, as the data is wandering.

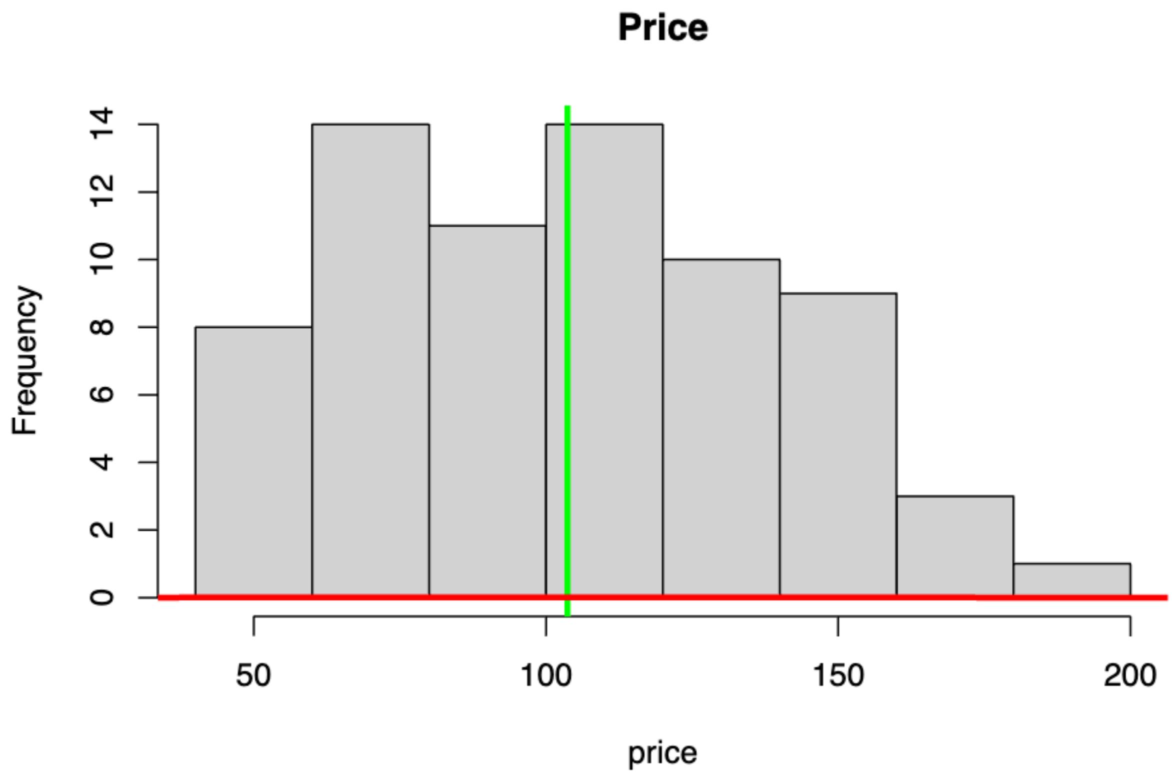
Our third variable is population. The histogram shows that population is bi-modal. The boxplot for population depicts that median of the data to be 4434. The range is 5904. The standard deviation is 2004.693.Population is stationary (though slightly increasing). This makes sense, logically. Meanwhile, the population across states is quite heteroskedastic.

Our fourth variable is CPI. The histogram for CPI depicts a fairly normal distribution.The boxplot for CPI depicts the median of the data to be 108.6. The range is 31.7. The standard deviation is 19.32959. CPI is strictly autoregressive across time, demonstrating a strong upward trend in the plot means, which again makes logical sense. Across states, it is exactly heteroskedastic, because CPI is representative.

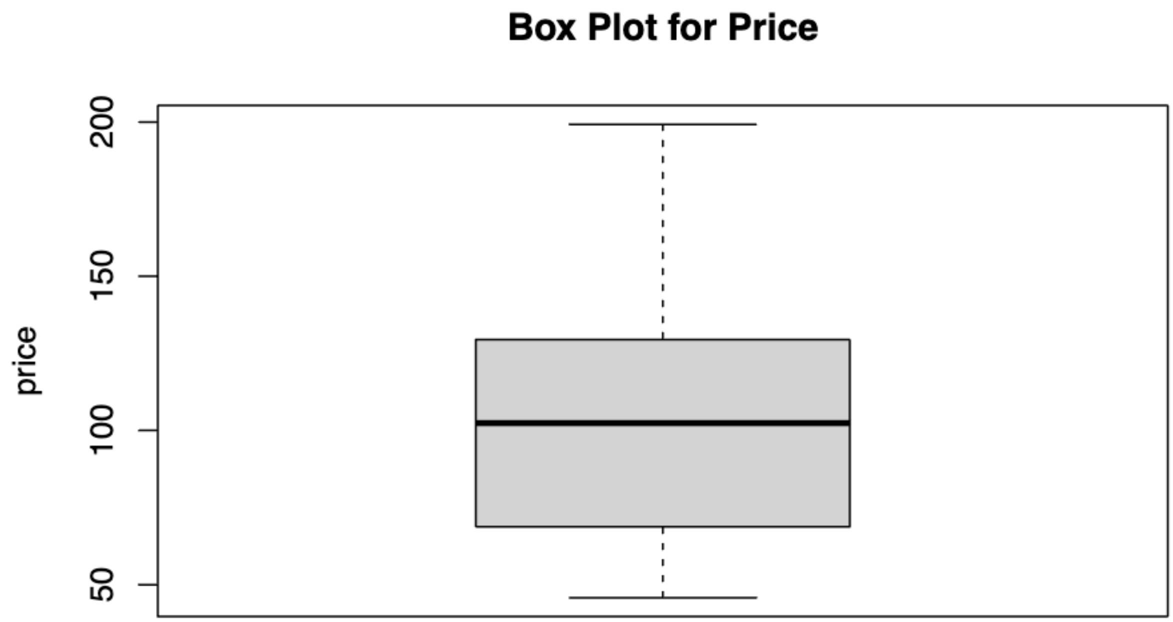
Our fifth variable is NDI. The histogram for NDI is fairly normal as well. However, it does appear to have a slightly skewed right tail. Since the boxplot displayed a relatively normal distribtion, we did not log our variable. The boxplot for NDI depicts that the median of the data is 11102. The range is 11958. The standard deviation is 2990.111. NDI is autoregressive as well, with another upward trend. Across states, it is heteroskedastic, wandering somewhat.

```r
#Price

hist(filtered_Cigar$price, xlab = 'price', ylab = 'Frequency', main = 'Price')
abline(v = mean(filtered_Cigar$pric), col='green', lwd = 3)
lines(density(filtered_Cigar$pric), col = 'red', lwd = 3)
```




```
boxplot(filtered_Cigar$price, main = "Box Plot for Price", ylab = "price")
```



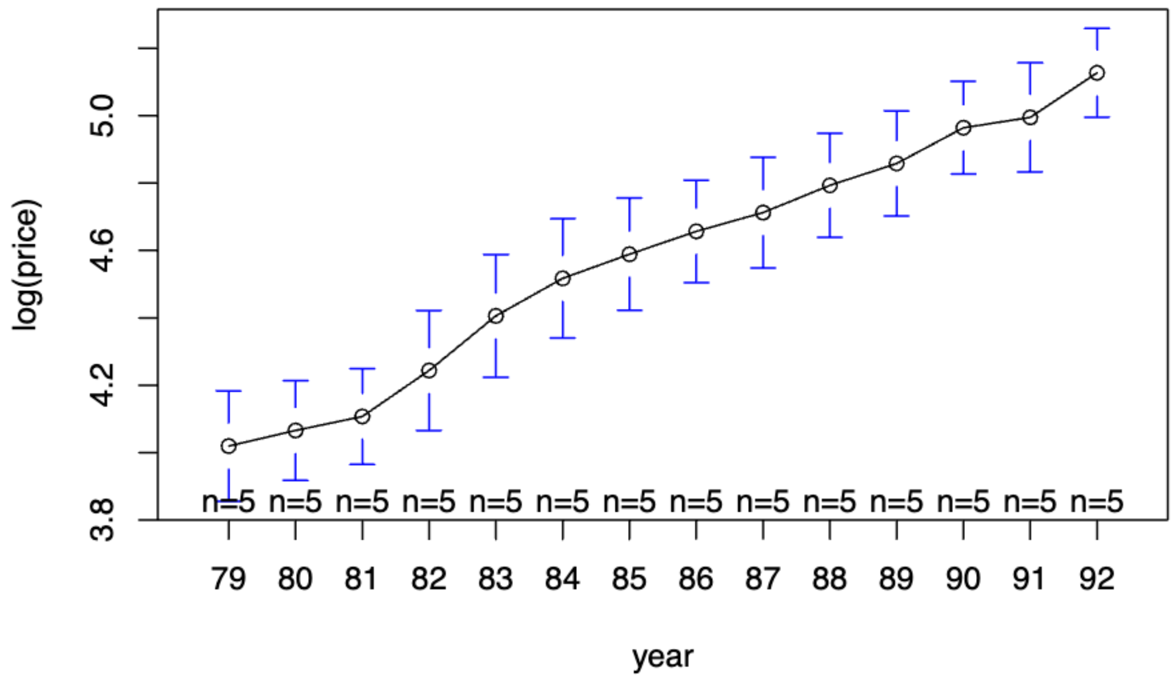

```r
summary(filtered_Cigar$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   45.80   69.35  102.40  103.65  129.28  199.20
```

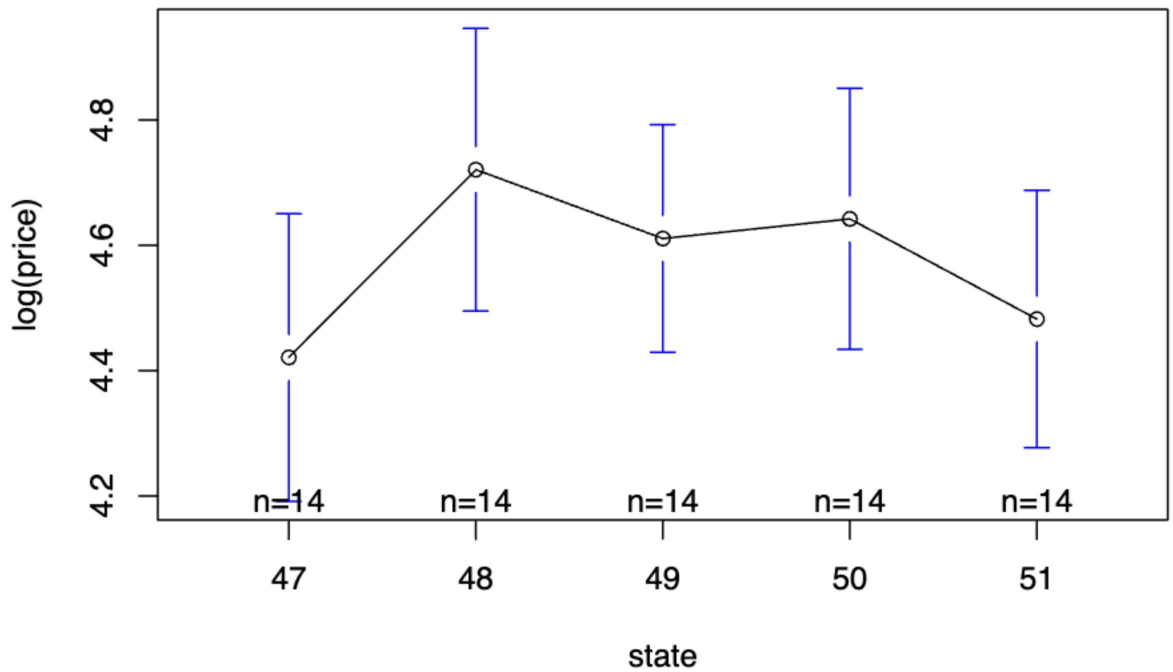
```r
sd(filtered_Cigar$price)
```

```
## [1] 37.04398
```

```r
#Heterogeneity across Time
plotmeans(log(price) ~ year, data=filtered_Cigar)
```

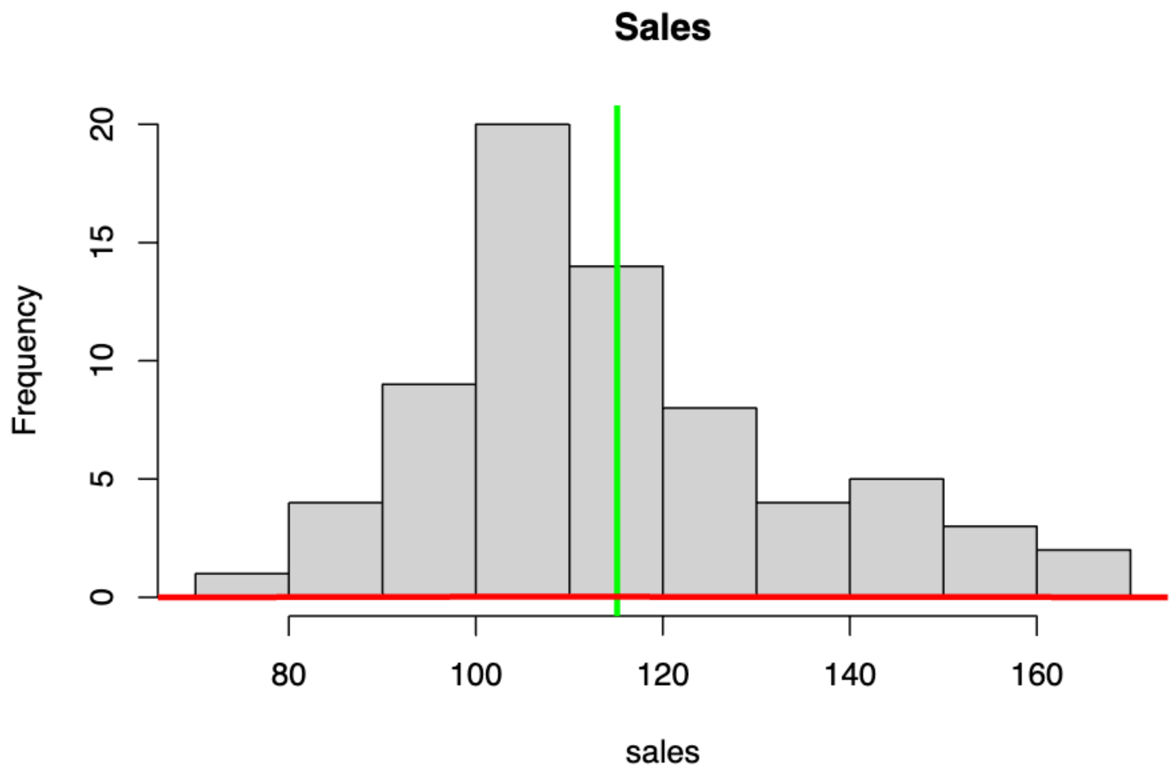



```
#Heterogeneity across States
plotmeans(log(price) ~ state, data=filtered_Cigar)
```

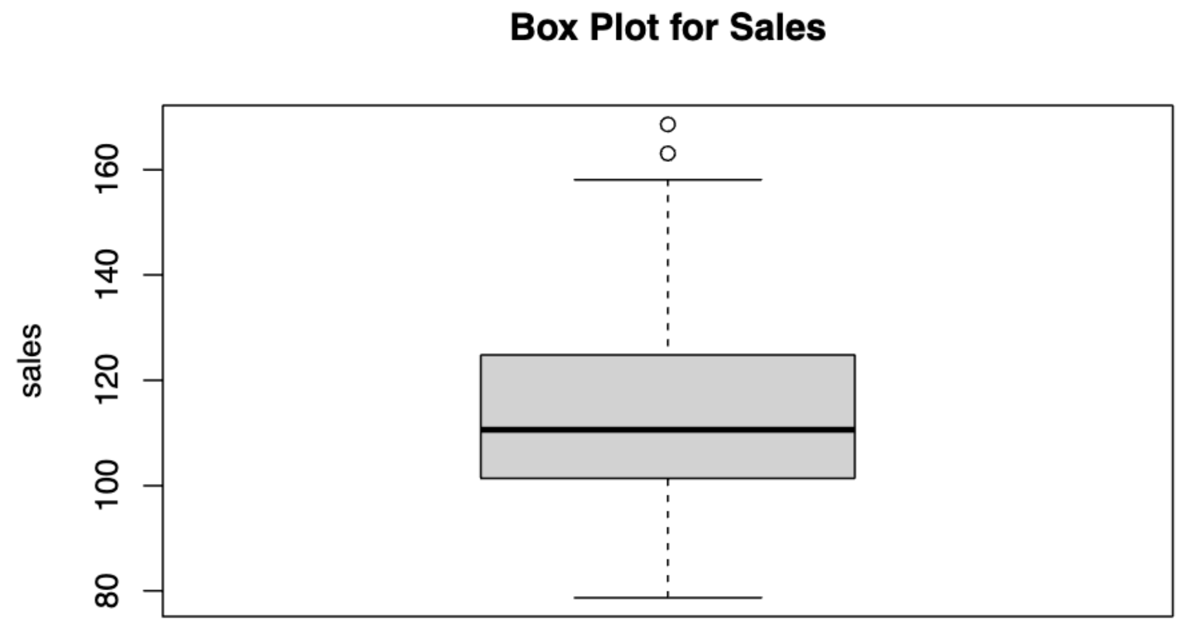



```r
#Sales

hist(filtered_Cigar$sales, xlab = 'sales', ylab = 'Frequency', main = 'Sales')
abline(v = mean(filtered_Cigar$sales), col='green', lwd = 3)
lines(density(filtered_Cigar$sales), col = 'red', lwd = 3)
```




```r
boxplot(filtered_Cigar$sales, main = "Box Plot for Sales", ylab = "sales")
```



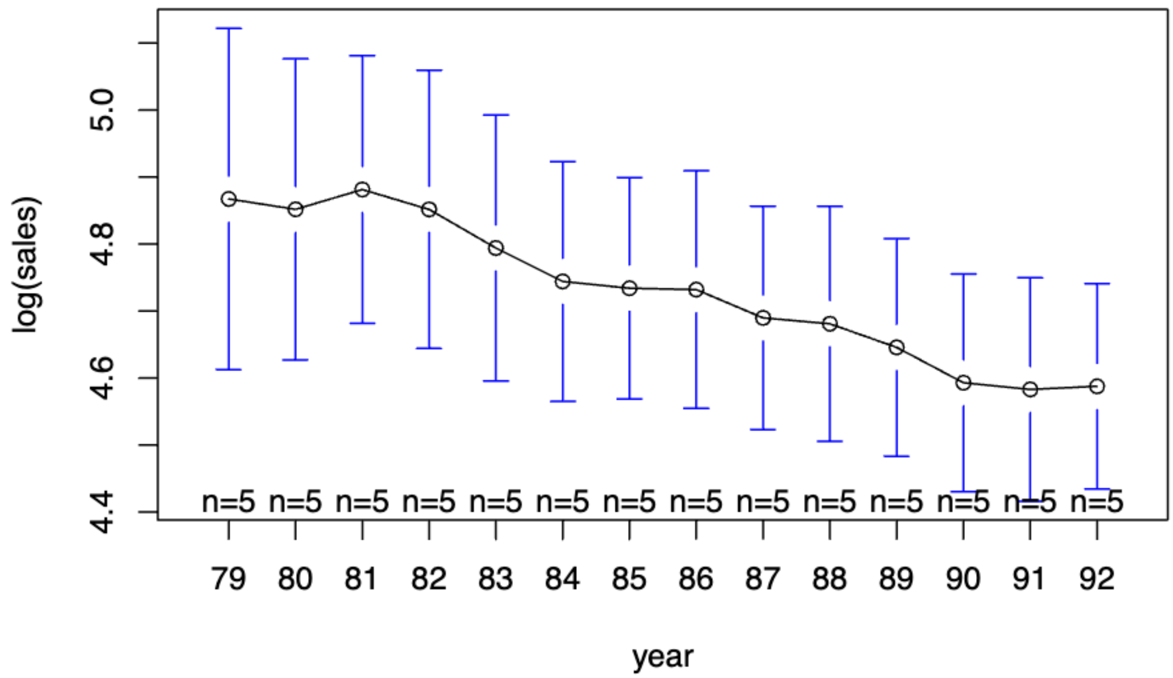

```r
summary(filtered_Cigar$sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    78.7   101.7   110.6   115.1   124.2   168.6
```

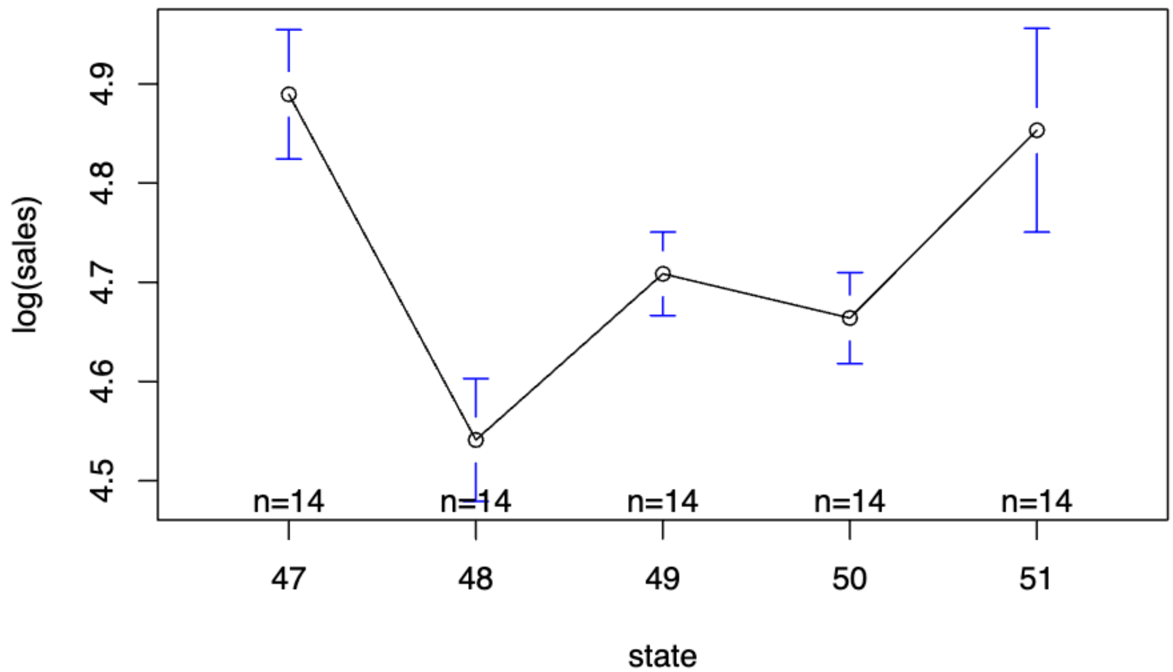
```r
sd(filtered_Cigar$sales)
```

```
## [1] 20.2686
```

```r
#Heterogeneity across Time
plotmeans(log(sales) ~ year, data=filtered_Cigar)
```

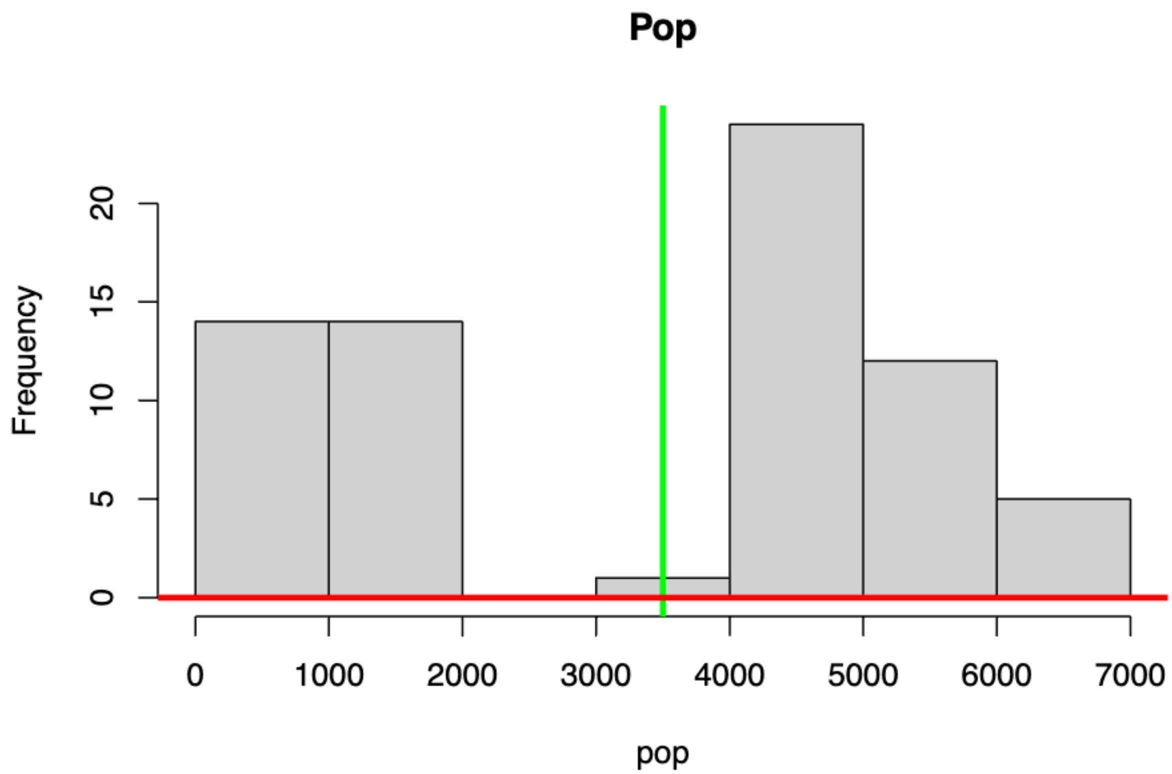



```
#Heterogeneity across States
plotmeans(log(sales) ~ state, data=filtered_Cigar)
```

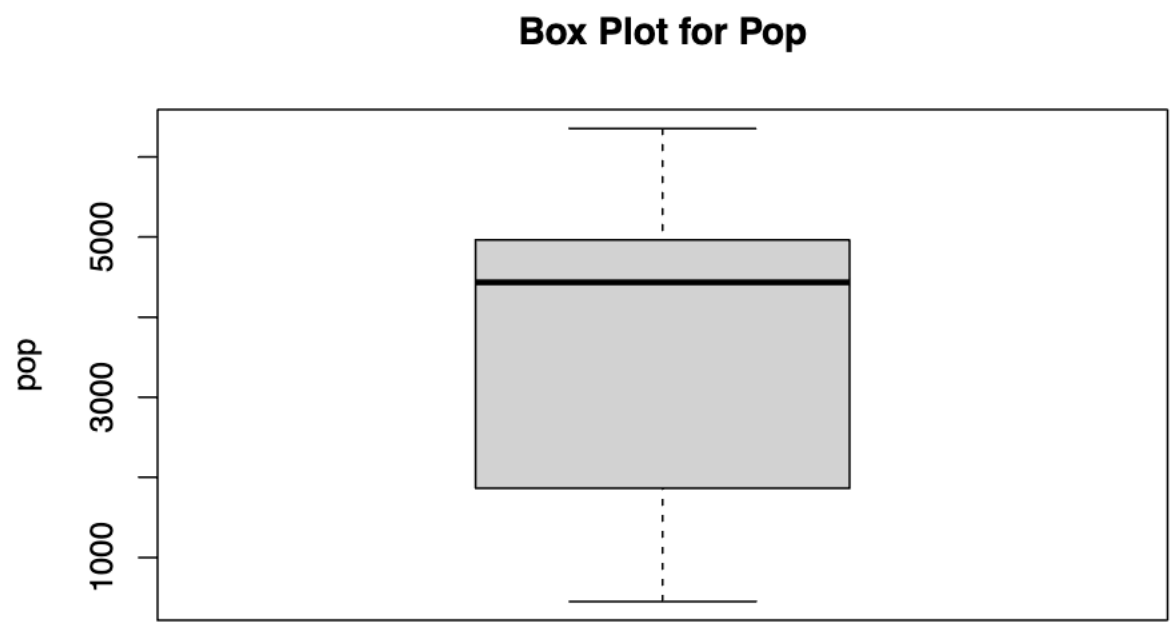



```
#Pop

hist(filtered_Cigar$pop, xlab = 'pop', ylab = 'Frequency', main = 'Pop')
abline(v = mean(filtered_Cigar$pop), col='green', lwd = 3)
lines(density(filtered_Cigar$pop), col = 'red', lwd = 3)
```




```r
boxplot(filtered_Cigar$pop, main = "Box Plot for Pop", ylab = "pop")
```



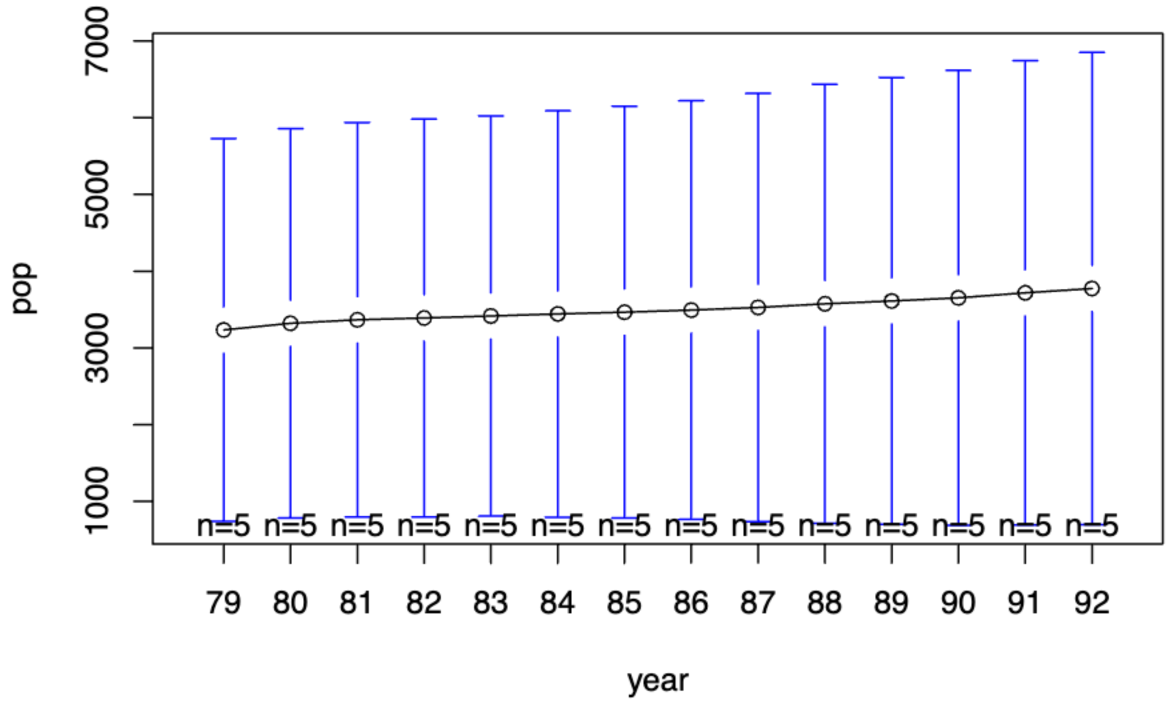

```r
summary(filtered_Cigar$pop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     450    1866    4434    3499    4949    6354
```

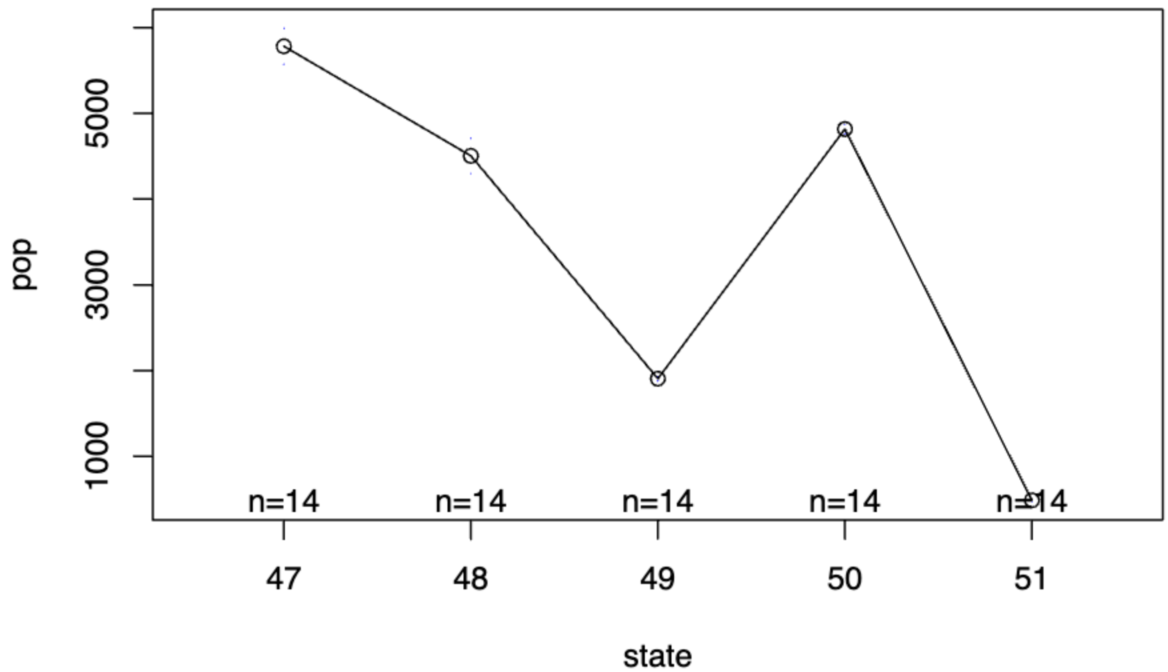
```r
sd(filtered_Cigar$pop)
```

```
## [1] 2004.693
```

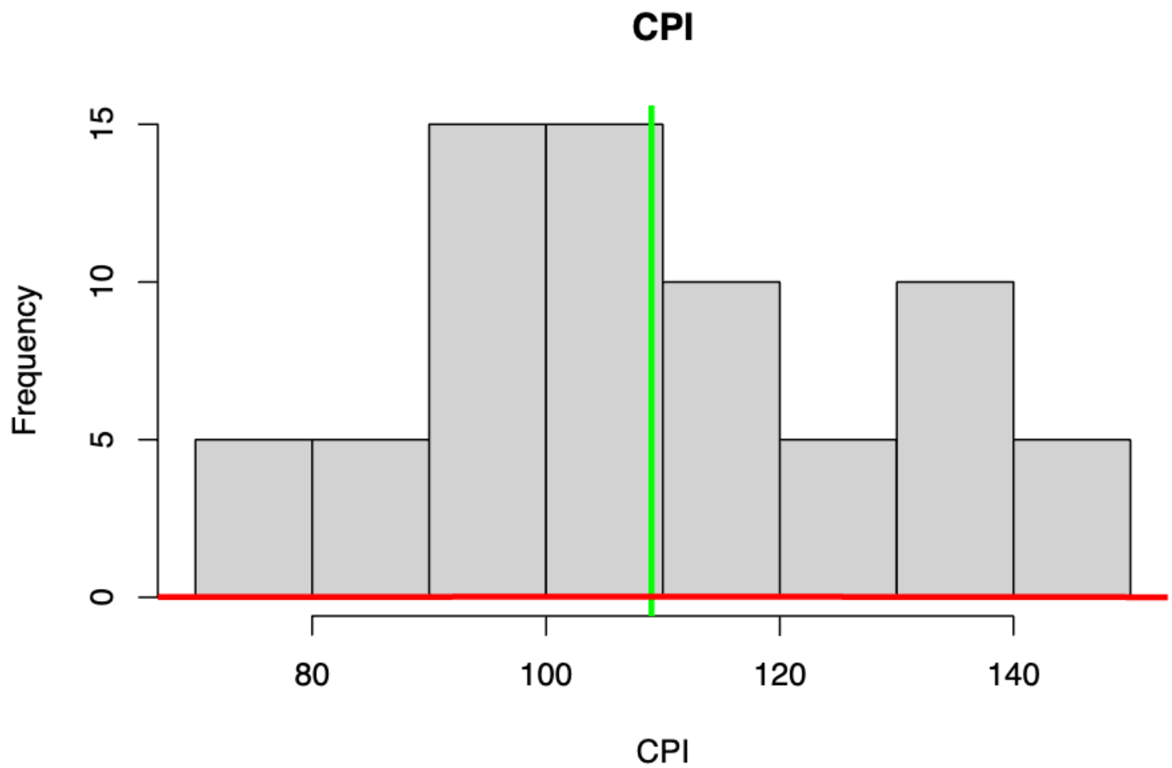
```r
#Heterogeneity across Time
plotmeans(pop ~ year, data=filtered_Cigar)
```



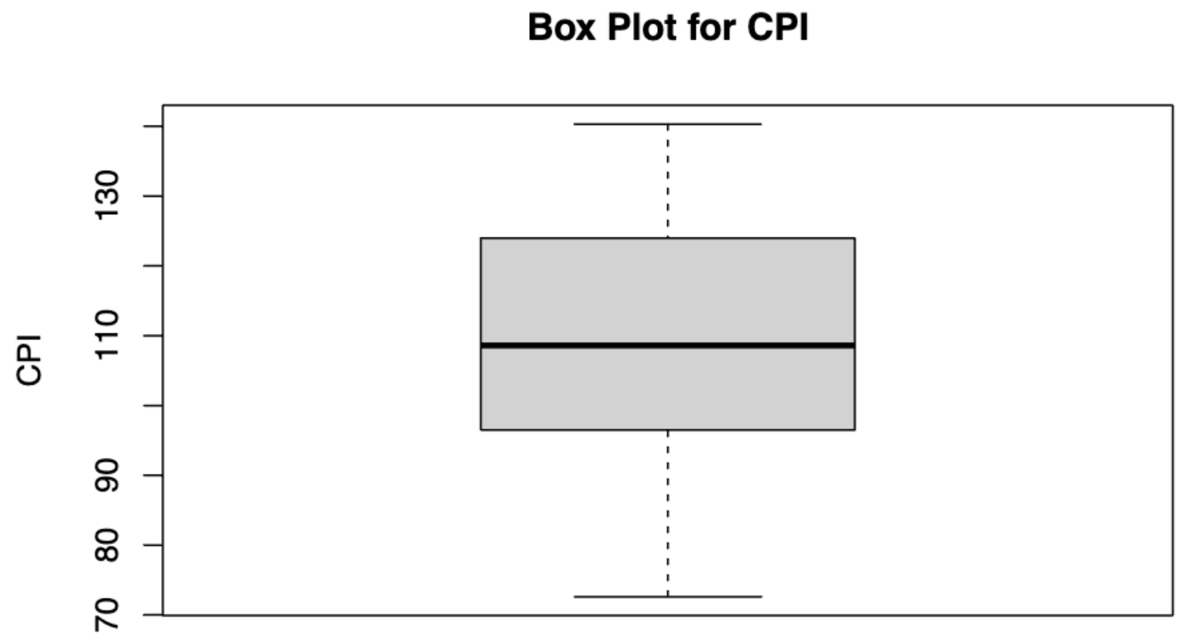

```r
#Heterogeneity across States
plotmeans(pop ~ state, data=filtered_Cigar)
```




```r
#CPI
hist(filtered_Cigar$cpi, xlab = 'CPI', ylab = 'Frequency', main = 'CPI')
abline(v = mean(filtered_Cigar$cpi), col='green', lwd = 3)
lines(density(filtered_Cigar$cpi), col = 'red', lwd = 3)
```




```r
boxplot(filtered_Cigar$cpi, main = "Box Plot for CPI", ylab = "CPI")
```



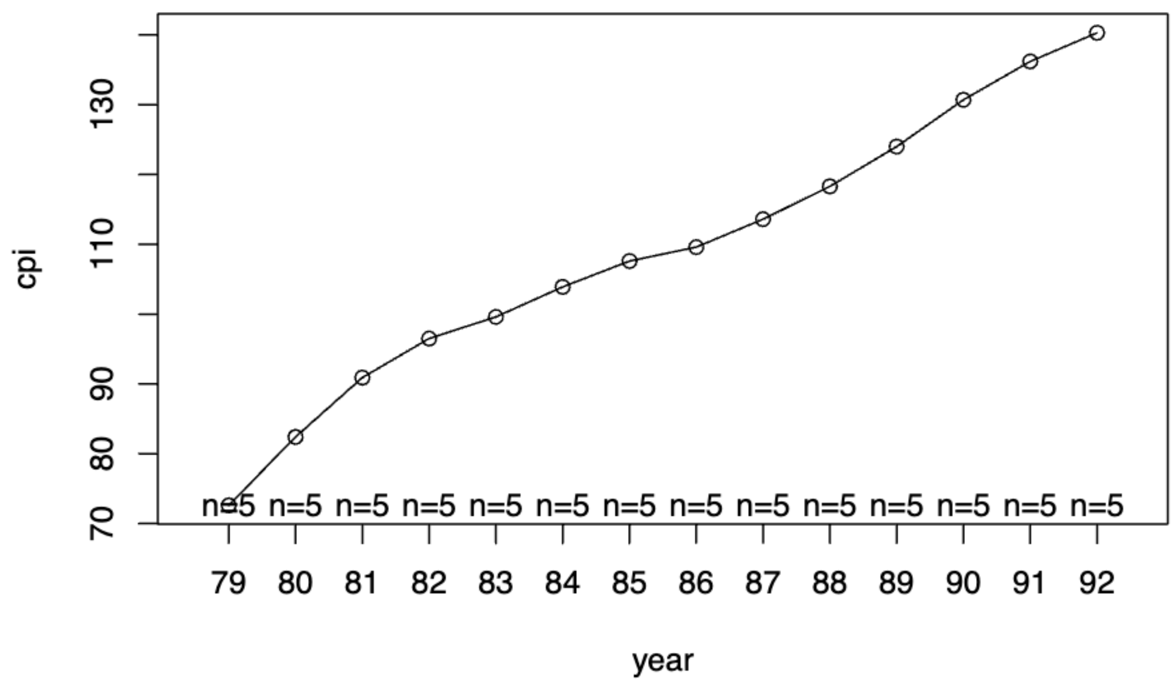

```r
summary(filtered_Cigar$cpi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    72.6    96.5   108.6   109.0   124.0   140.3
```

```r
sd(filtered_Cigar$cpi)
```

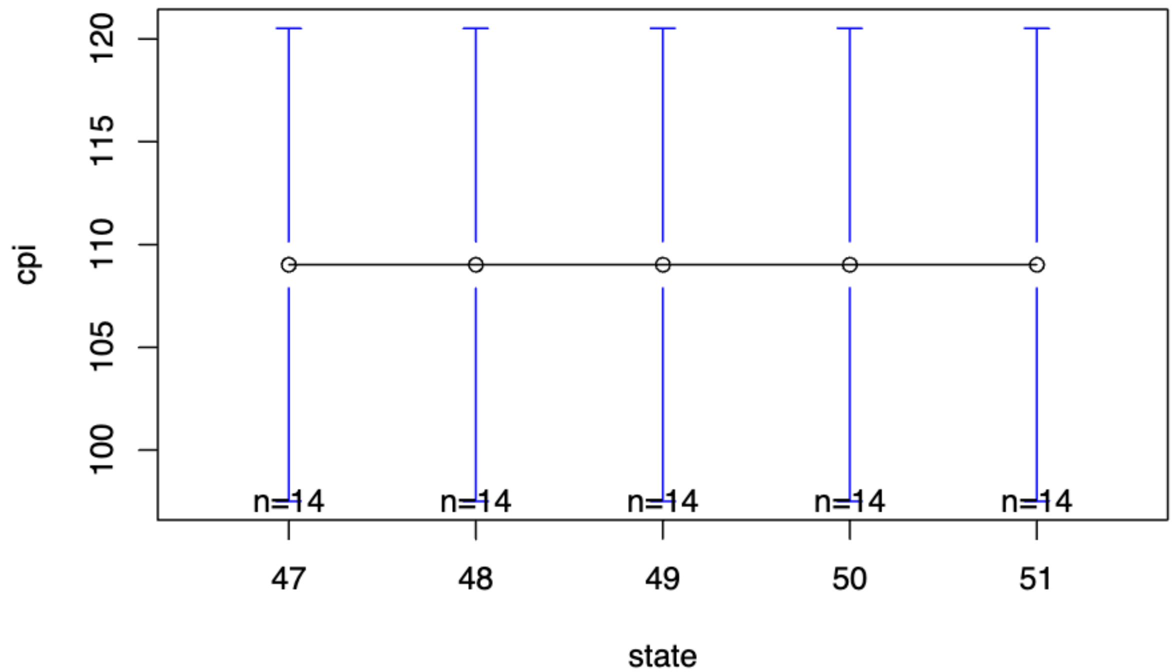
```
## [1] 19.32959
```

```r
#Heterogeneity across Time
plotmeans(cpi ~ year, data=filtered_Cigar)
```



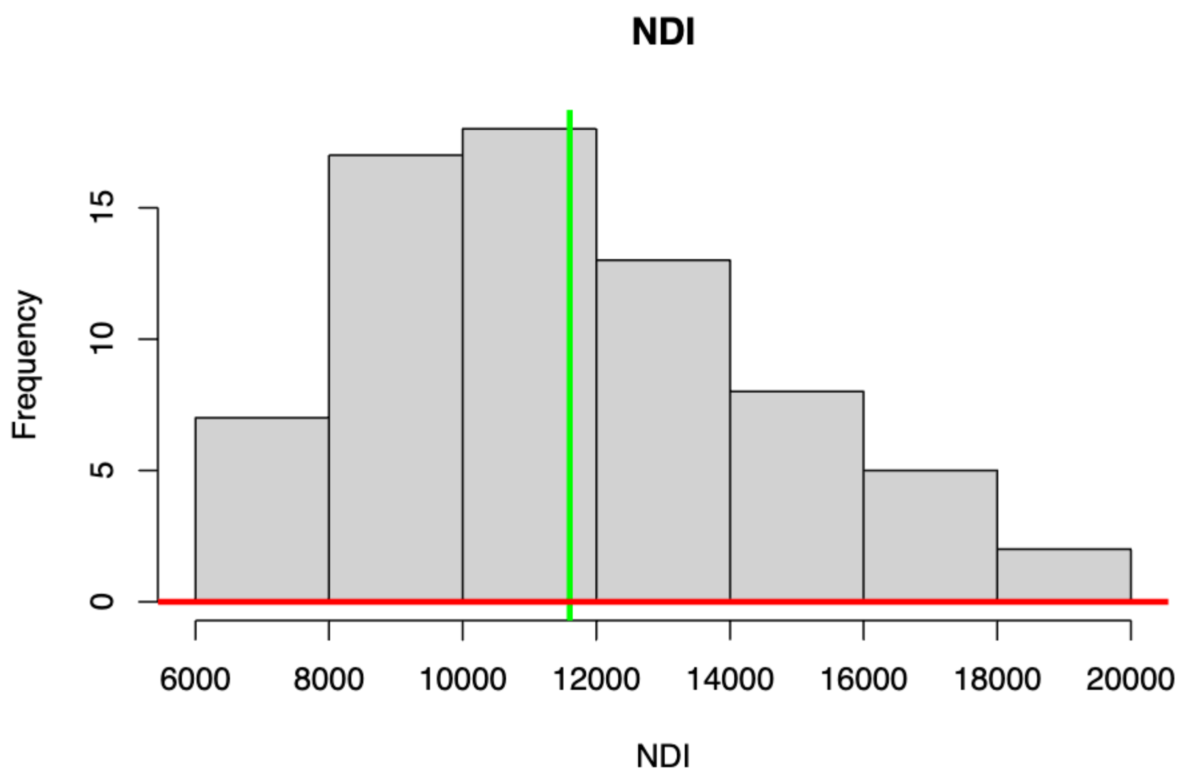

```
#Heterogeneity across States

plotmeans(cpi ~ state, data=filtered_Cigar)
```

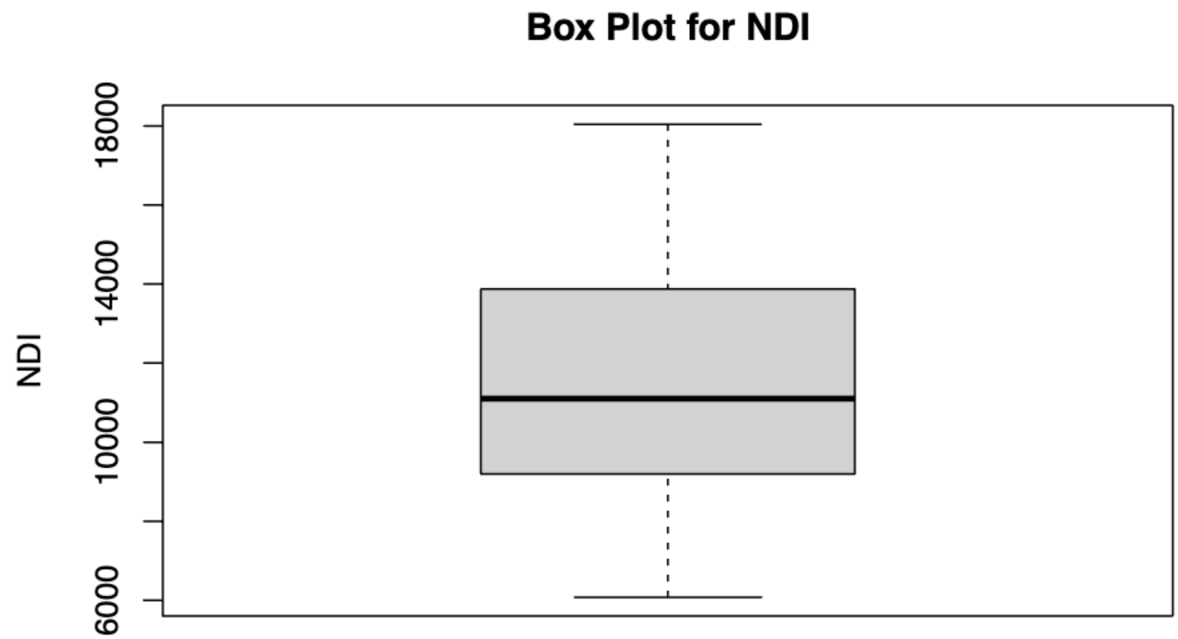



```r
#NDI

hist(filtered_Cigar$ndi, xlab = 'NDI', ylab = 'Frequency', main = 'NDI')
abline(v = mean(filtered_Cigar$ndi), col='green', lwd = 3)
lines(density(filtered_Cigar$ndi), col = 'red', lwd = 3)
```




```r
boxplot(filtered_Cigar$ndi, main = "Box Plot for NDI", ylab = "NDI")
```



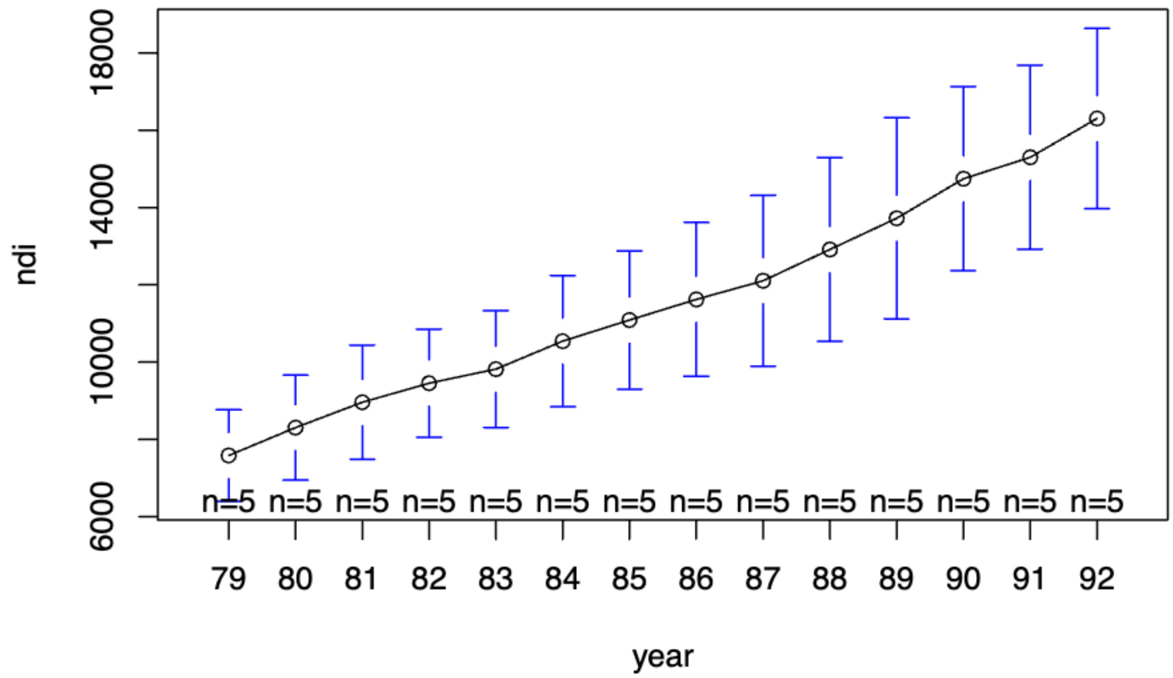

```r
summary(filtered_Cigar$ndi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6080    9248   11102   11603   13785   18038
```

```r
sd(filtered_Cigar$ndi)
```

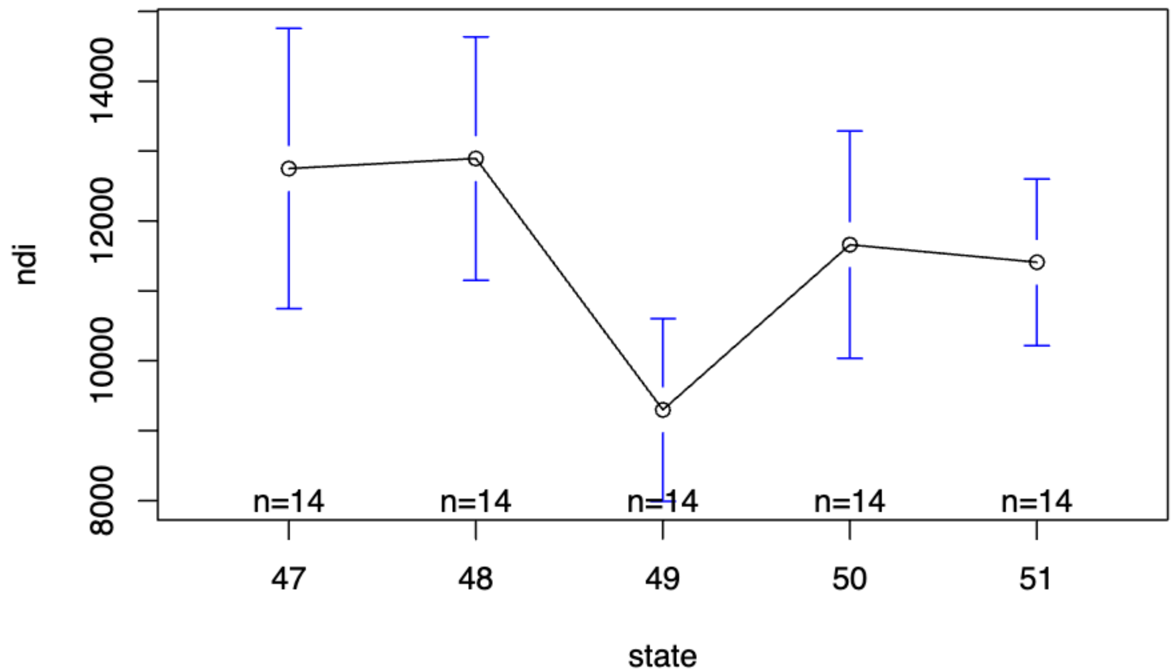
```
## [1] 2990.111
```

```r
#Heterogeneity across Time
plotmeans(ndi ~ year, data=filtered_Cigar)
```




```
#Heterogeneity across States

#NDI
plotmeans(ndi ~ state, data=filtered_Cigar)
```



(c) Fit the three models below, and identify which model is your preferred one and why. Make sure to include your statistical diagnostics to support your conclusion, and to comment on your findings. • Pooled Model • Fixed Effects • Random Effects

For our models, we regressed log(sales) on ndi, cpi, log(price), and population to determine what effects these variables have on sales across states and time. An interesting find in our regression was that in the fixed effects model that incorporated time CPI was omitted from the regression (twoway model and time model). However, all variables were included in the fixed effect model with effect = individual, pooled model, and random effect model. This could have something to do with the fact that CPI varies across time but not across states. Thus, we picked to analyze the regression with respect to states (fm_state).

```
Cig <- pdata.frame(filtered_Cigar, c("state", "year"))


#Pooled Model

fm_state <- plm(log(sales) ~ ndi + cpi + log(price) +pop,
data = Cig, model = "pooling")

print(fm_state)
```

```
##
## Model Formula: log(sales) ~ ndi + cpi + log(price) + pop
##
## Coefficients:
```

```
## (Intercept)         ndi         cpi  log(price)         pop
##  7.6607e+00  1.1993e-05  9.0506e-03 -8.7686e-01 -1.2343e-05
```

```r
#Fixed Effects Model

fm_full <- plm(log(sales) ~ ndi + cpi + log(price) + pop,
data = Cig, model = "within", effect = "twoways")

print(fm_full)
```

```
##
## Model Formula: log(sales) ~ ndi + cpi + log(price) + pop
##
## Coefficients:
##         ndi  log(price)         pop
##  7.9192e-05 -4.7986e-01 -1.4131e-04
```

```r
fm_time <- plm(log(sales) ~ ndi + cpi + log(price) + pop,
data = Cig, model = "within", effect = "time")

print(fm_time)
```

```
##
## Model Formula: log(sales) ~ ndi + cpi + log(price) + pop
##
## Coefficients:
##         ndi  log(price)         pop
##  1.0666e-05 -1.0545e+00 -1.0571e-05
```

```r
fm_state <- plm(log(sales) ~ ndi + cpi + log(price) + pop,
data = Cig, model = "within", effect = "individual")

print(fm_state)
```

```
##
## Model Formula: log(sales) ~ ndi + cpi + log(price) + pop
##
## Coefficients:
##         ndi         cpi  log(price)         pop
##  2.9612e-05 -1.7906e-03 -3.9851e-01 -2.5666e-05
```

```r
#Random Effects Model
fm_rstate <- plm(log(sales) ~ ndi + cpi + log(price) + pop,
data = Cig, model = "random")

print(fm_rstate)
```

```
##
## Model Formula: log(sales) ~ ndi + cpi + log(price) + pop
##
## Coefficients:
## (Intercept)         ndi         cpi  log(price)         pop
##  7.3487e+00  1.9741e-05  5.6733e-03 -7.4501e-01 -1.6065e-05
```

In order to test which model was best for our regression, we first used the PLM test which concluded that the p-value was 1.894e-05.Thus, we reject the null confirming that we should not use pooled model and instead use a fixed effect model. Next, we used the Hausman test in which we got a p-value of 2.2e-16, suggesting that we should reject the null and use a fixed effect model. This correlates with the PLM test. The best model is the fixed effect model with effect = individual.

```
#Test for Best Model

wageReTest <- plmtest(fm_state, effect= "individual")
wageReTest
```

```
##
##  Lagrange Multiplier Test - (Honda)
##
## data:  log(sales) ~ ndi + cpi + log(price) + pop
## normal = 4.1201, p-value = 1.894e-05
## alternative hypothesis: significant effects
```

```
#Hausman Test

phtest(fm_state, fm_rstate)
```

```
##
##  Hausman Test
##
## data:  log(sales) ~ ndi + cpi + log(price) + pop
## chisq = 148.3, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Binary Dependent Variables

(a) Briefly discuss your data and the question you are trying to answer with your model.

We are using the Swisslabor dataset in the AER package. The data is cross-sectional pulled from health survey in Switzerland in 1981. The variables in this data include income, age, education, number of children under 7 years (youngkids), number of children over 7 years (old kids), if the person is foreign, and participation. Our data has 2 binary variables which are participation and foreign. We will be using participation as our dependent variable.

```
data("SwissLabor")
summary(SwissLabor)
```

```
##  participation     income            age          education
##  no :471       Min.   : 7.187   Min.   :2.000   Min.   : 1.000
##  yes:401       1st Qu.:10.472   1st Qu.:3.200   1st Qu.: 8.000
##                Median :10.643   Median :3.900   Median : 9.000
##                Mean   :10.686   Mean   :3.996   Mean   : 9.307
##                3rd Qu.:10.887   3rd Qu.:4.800   3rd Qu.:12.000
##                Max.   :12.376   Max.   :6.200   Max.   :21.000
##    youngkids         oldkids         foreign
##  Min.   :0.0000   Min.   :0.0000   no :656
```

```
## 1st Qu.:0.0000   1st Qu.:0.0000   yes:216
## Median :0.0000   Median :1.0000
## Mean   :0.3119   Mean   :0.9828
## 3rd Qu.:0.0000   3rd Qu.:2.0000
## Max.   :3.0000   Max.   :6.0000
```

(b) Provide a descriptive analysis of your variables. This should include RELEVANT histograms and fitted distributions, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments. For binary variables, you can simply include the proportions of each factor.

Our first variable is participation. Participation exhibits a nearly even split—54/46 (N/Y) to be precise. Because this variable is binary, we converted it to numeric with N being 0 and Y being 1.

Our second variable is income. The histogram for income has a normal distribution. The boxplot for income indicates the median of the data to be 10.643. The range is 5.189. The standard deviation is 0.4124888.

Our third variable is age. The histogram for age has a relatively normal distribution.The boxplot for age indicated the median to he 3.9. The range is 4.2. The standard deviation is 1.055167.

Our fourth variable is education. The histogram for education is normally distributed. The boxplot for education indicates the mean to be 9.0. The range of the data is 20.0. The standard deviation is 3.036259.

Our fifth variable is youngkids. The histogram for young kids is skewed right with a high population of people having 0 babies. The boxplot for young kids indicated the median to be 0.0. The range is 3.0. The standard deviation is 0.61287.

Our sixth variable is old kids. The histogram for old kids is skewed right with about half having no old kids. The boxplot for old kids indicates the median of the data to be 1.0. The range is 6.0. The standard deviation is 1.086786.

Cross correlation: The only variables which exhibit significant cross correlations are income with education and youngkids with age. Needless to say, economic intuition is congruent with these findings. Interestingly though, some boxes are blank.

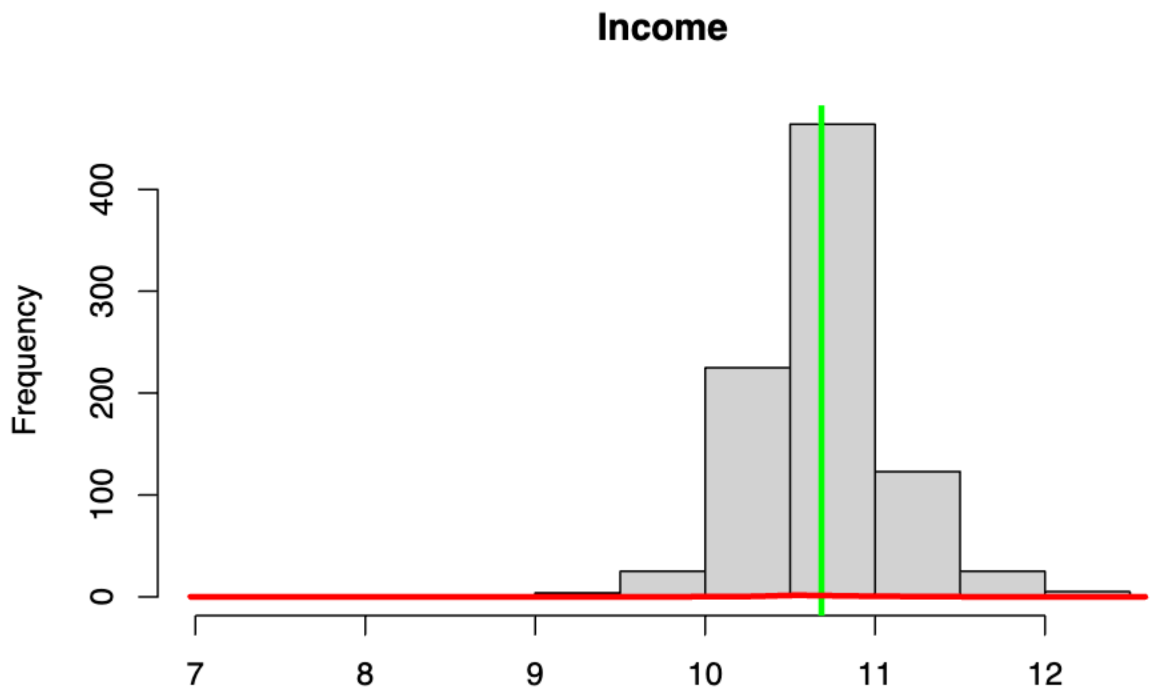
```r
#Participation
summary(SwissLabor$participation)
```

```
##  no yes
## 471 401
```

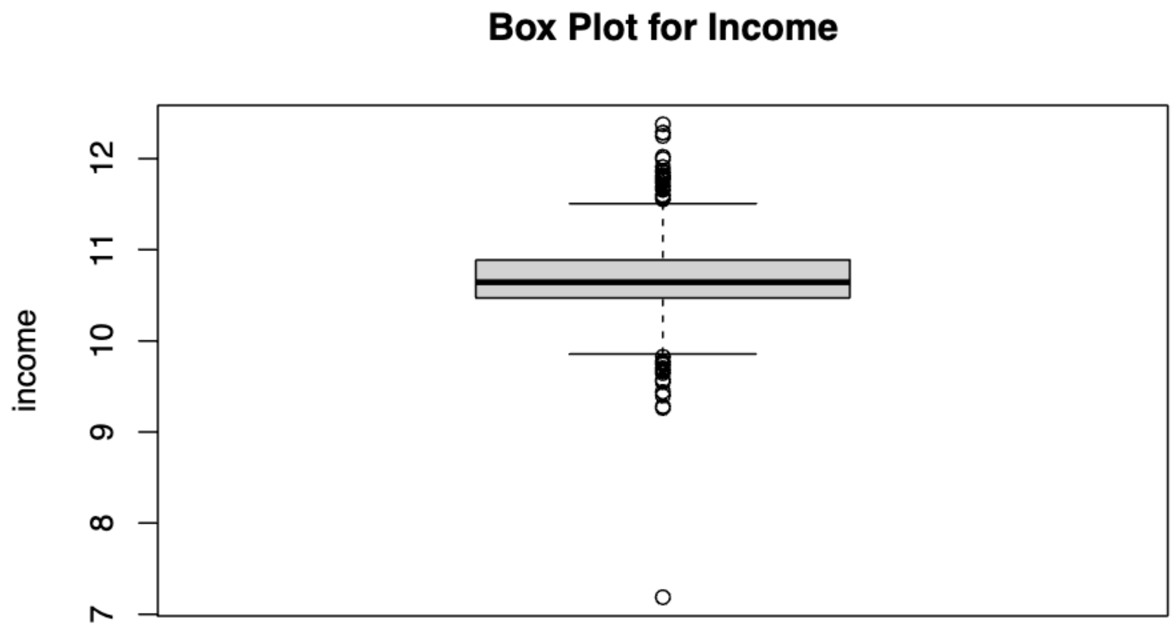
```r
participation_table <- table(SwissLabor$participation)
proportion_table <- prop.table(participation_table)
print(proportion_table)
```

```
##
##        no       yes
## 0.5401376 0.4598624
```

```r
#Income
hist(SwissLabor$income, xlab = '', ylab = 'Frequency', main = 'Income')
abline(v = mean(SwissLabor$income), col='green', lwd = 3)
lines(density(SwissLabor$income), col = 'red', lwd = 3)
```




```r
boxplot(SwissLabor$income, main = "Box Plot for Income", ylab = "income")
```



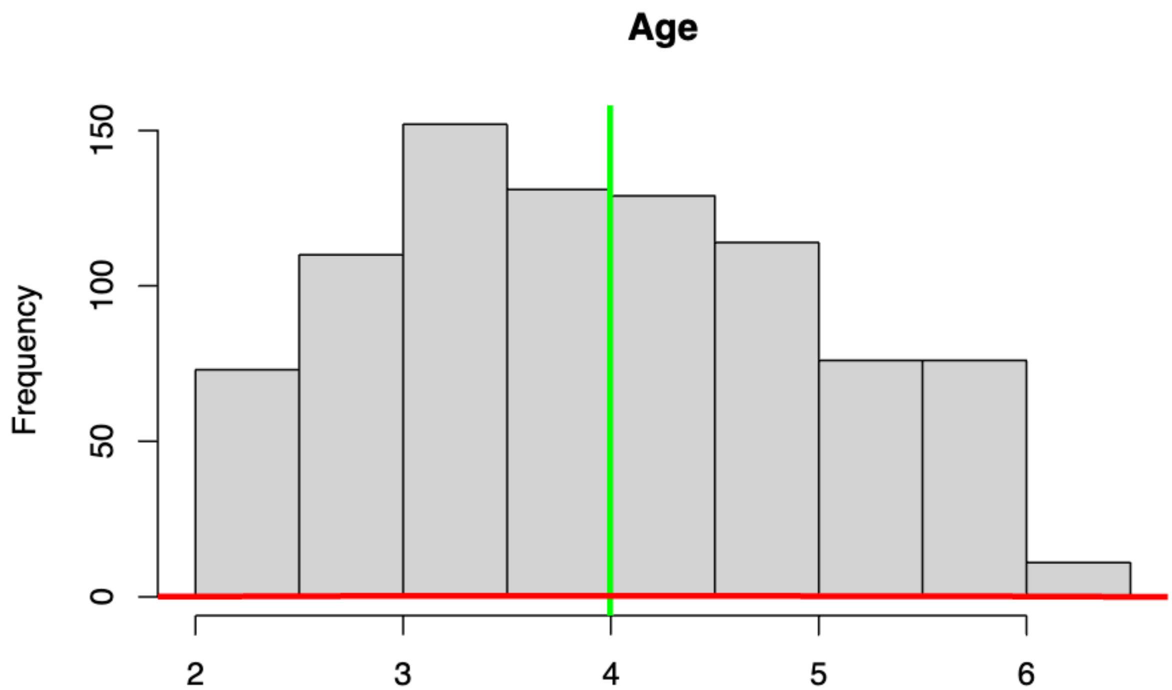

```r
summary(SwissLabor$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.187  10.472  10.643  10.686  10.887  12.376
```

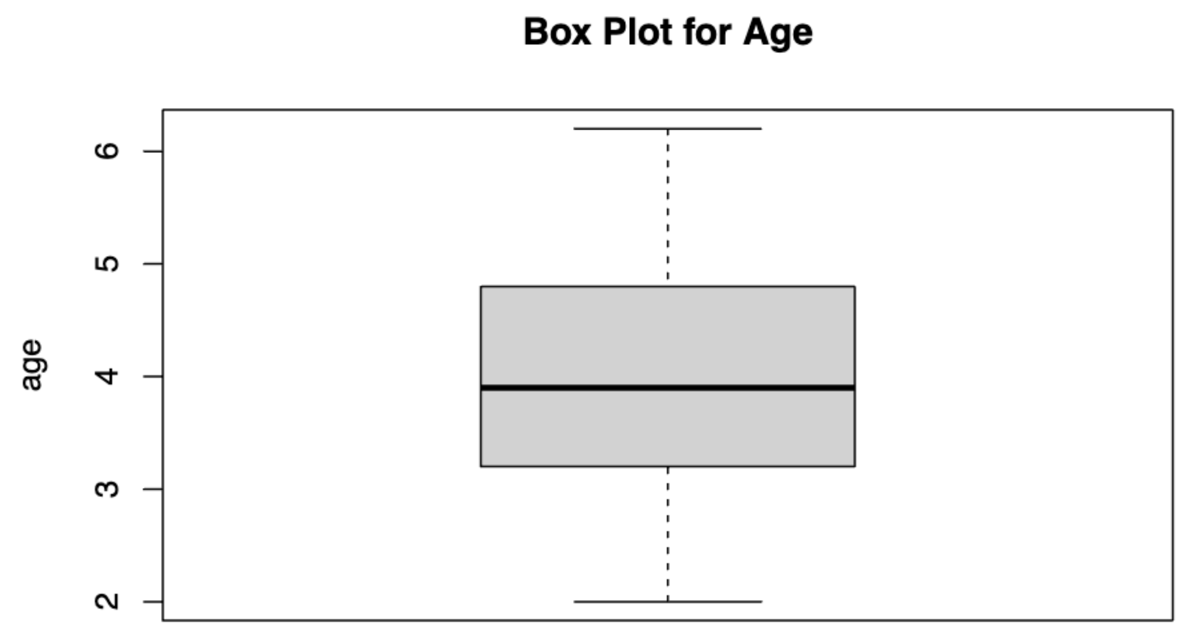
```r
sd(SwissLabor$income)
```

```
## [1] 0.4124888
```

```r
#Age
hist(SwissLabor$age, xlab = '', ylab = 'Frequency', main = 'Age')
abline(v = mean(SwissLabor$age), col='green', lwd = 3)
lines(density(SwissLabor$age), col = 'red', lwd = 3)
```




```r
boxplot(SwissLabor$age, main = "Box Plot for Age", ylab = "age")
```



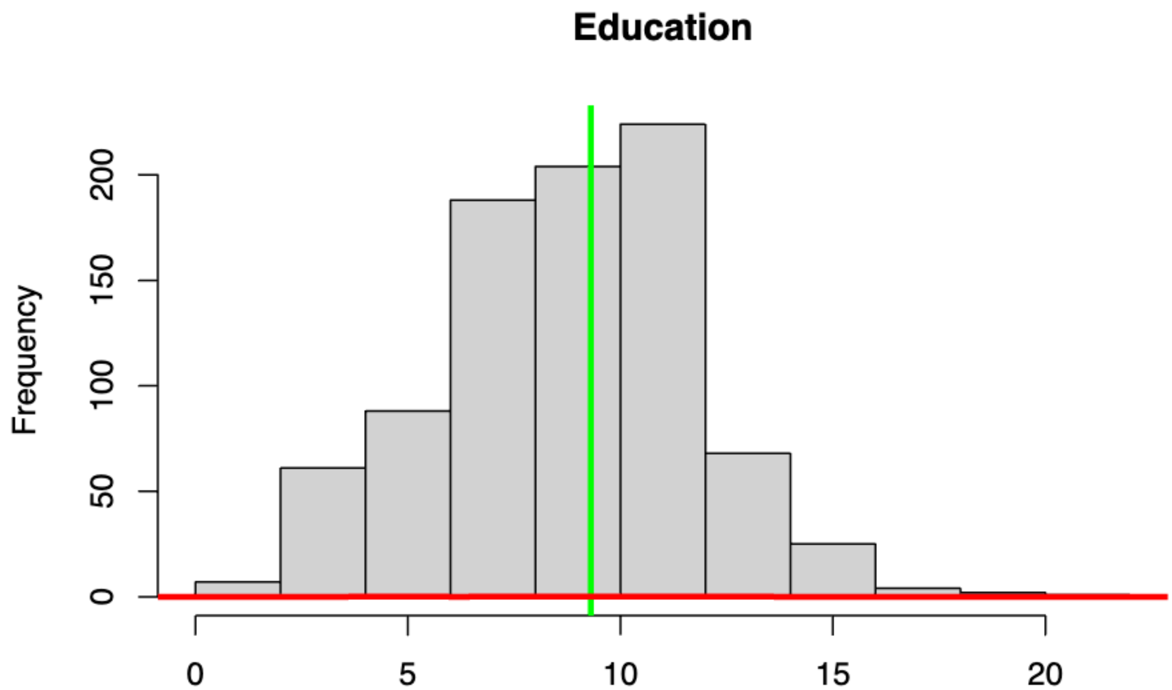

```
summary(SwissLabor$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.200   3.900   3.996   4.800   6.200
```

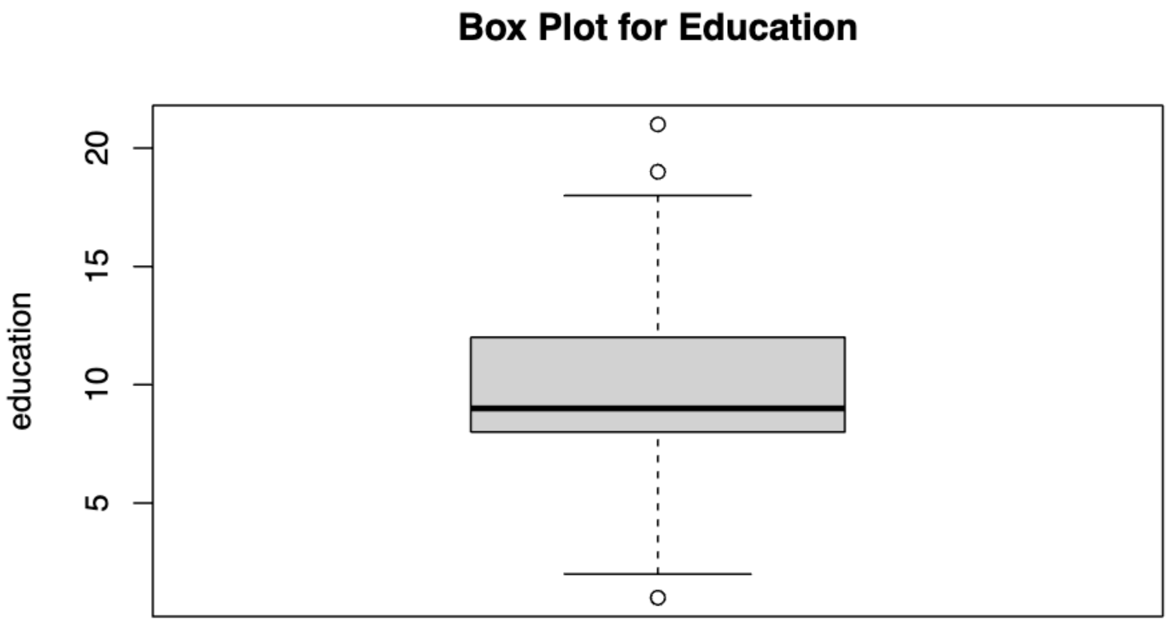
```
sd(SwissLabor$age)
```

```
## [1] 1.055167
```

```
#Education
hist(SwissLabor$education, xlab = '', ylab = 'Frequency', main = 'Education')
abline(v = mean(SwissLabor$education), col='green', lwd = 3)
lines(density(SwissLabor$education), col = 'red', lwd = 3)
```




```r
boxplot(SwissLabor$education, main = "Box Plot for Education", ylab = "education")
```



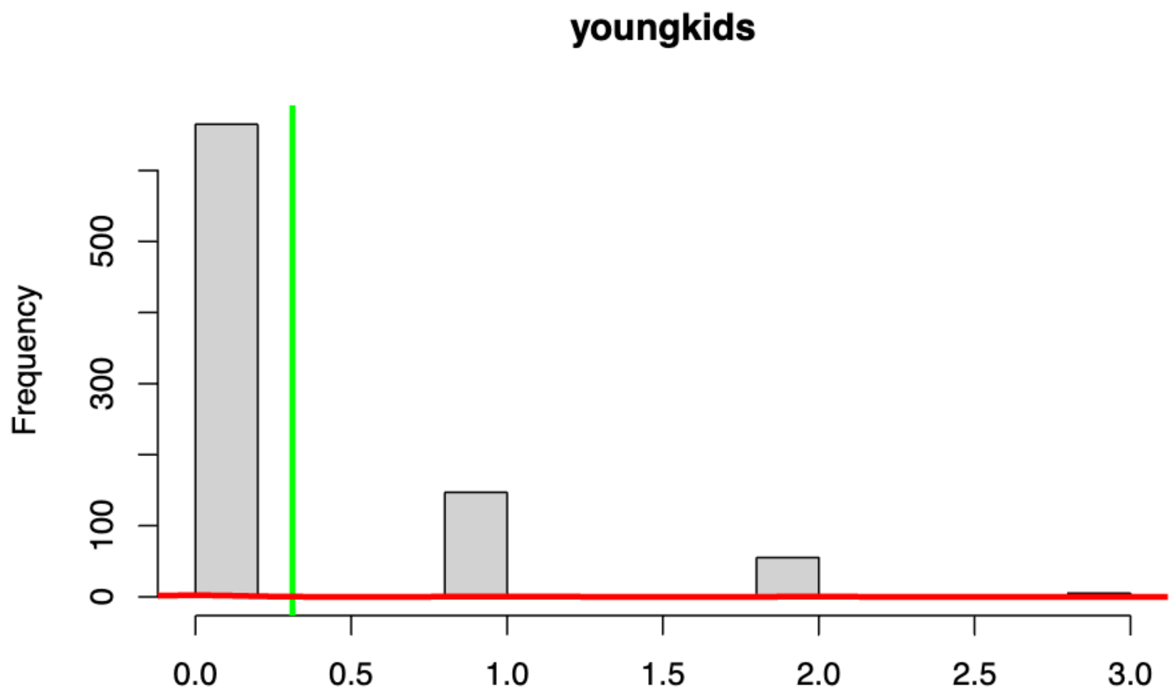

```r
summary(SwissLabor$education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   8.000   9.000   9.307  12.000  21.000
```

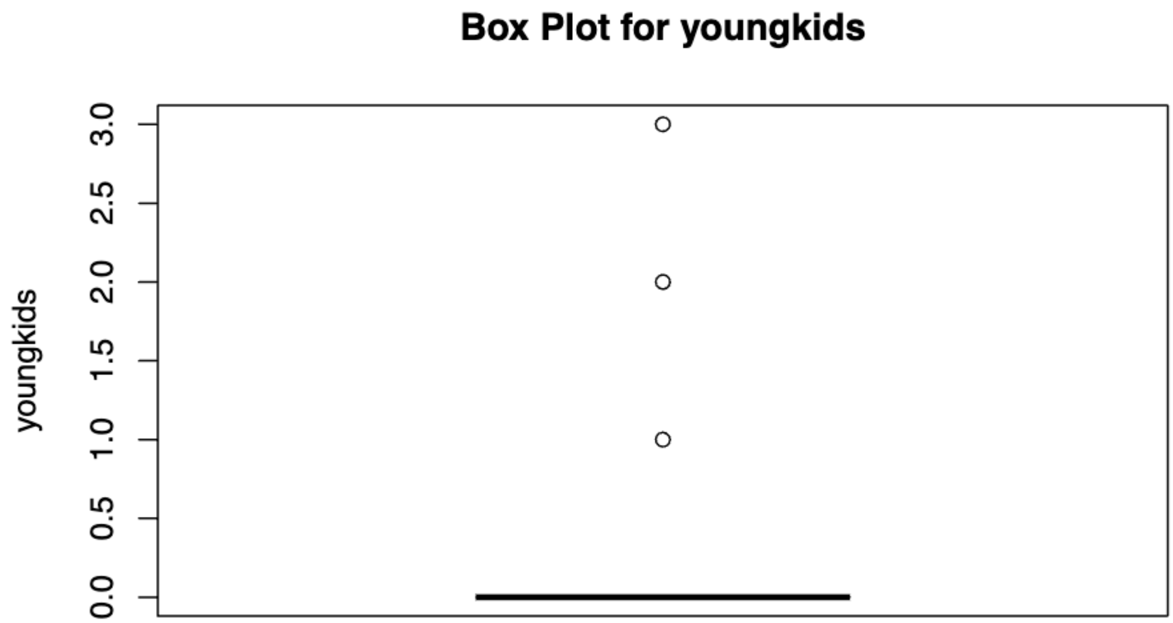
```r
sd(SwissLabor$education)
```

```
## [1] 3.036259
```

```r
#Youngkids
hist(SwissLabor$youngkids, xlab = '', ylab = 'Frequency', main = 'youngkids')
abline(v = mean(SwissLabor$youngkids), col='green', lwd = 3)
lines(density(SwissLabor$youngkids), col = 'red', lwd = 3)
```




```
boxplot(SwissLabor$youngkids, main = "Box Plot for youngkids", ylab = "youngkids")
```



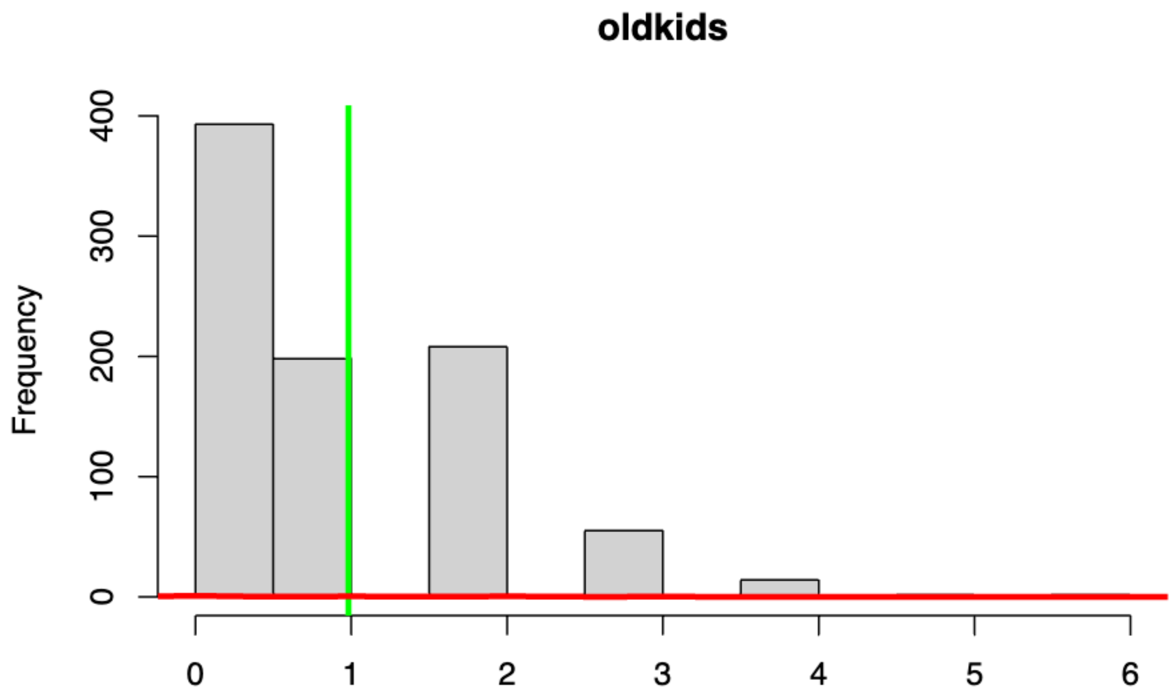

```r
summary(SwissLabor$youngkids)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3119  0.0000  3.0000
```

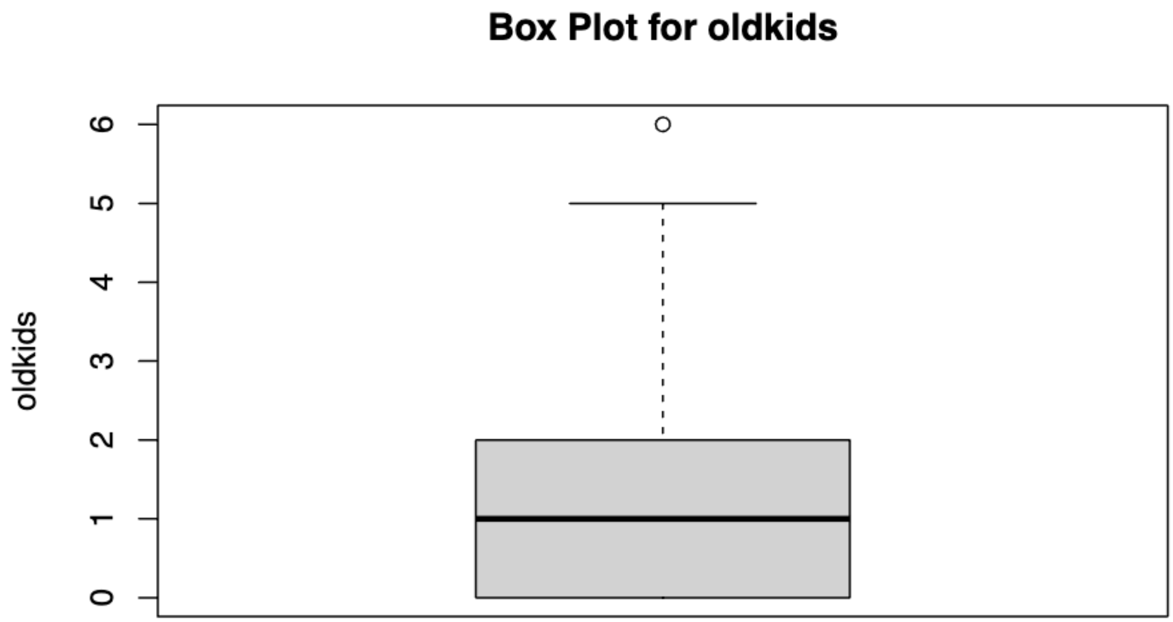
```r
sd(SwissLabor$youngkids)
```

```
## [1] 0.61287
```

```r
#Oldkids
hist(SwissLabor$oldkids, xlab = '', ylab = 'Frequency', main = 'oldkids')
abline(v = mean(SwissLabor$oldkids), col='green', lwd = 3)
lines(density(SwissLabor$oldkids), col = 'red', lwd = 3)
```




```r
boxplot(SwissLabor$oldkids, main = "Box Plot for oldkids", ylab = "oldkids")
```




```r
summary(SwissLabor$oldkids)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.9828  2.0000  6.0000
```

```r
sd(SwissLabor$oldkids)
```

```
## [1] 1.086786
```

```r
#cross correlation

filtered_SwissLabor <- SwissLabor[, !names(SwissLabor) %in% c("participation", "foreign")]
str(filtered_SwissLabor)
```

```
## 'data.frame':    872 obs. of  5 variables:
##  $ income   : num  10.8 10.5 11 11.1 11.1 ...
##  $ age      : num  3 4.5 4.6 3.1 4.4 4.2 5.1 3.2 3.9 4.3 ...
##  $ education: num  8 8 9 11 12 12 8 8 12 11 ...
##  $ youngkids: num  1 0 0 2 0 0 0 0 0 0 ...
##  $ oldkids  : num  1 1 0 0 2 1 0 2 0 2 ...
```

```r
S <- cor(filtered_SwissLabor)
corrplot(S, method = 'number')
```

(c) Fit the three models below, and identify which model is your preferred one and why. Make sure to include statistical diagnostics to support your conclusion, and to comment on your findings. • Linear Probability Model • Probit Model • Logit Model

For our models, we are choosing to regress participation on all the variables except for foreign to see what effects each variable has on particiaption.

```
#Linear Probability Model

data(SwissLabor)

levels(SwissLabor$participation)
```

```
## [1] "no"  "yes"
```

```
# Convert Factor to Numeric
# No = 0 and Yes = 1
SwissLabor$participation_numeric <- as.numeric(SwissLabor$participation) - 1

head(SwissLabor)
```

```
##   participation   income age education youngkids oldkids foreign
## 1            no 10.78750 3.0         8         1       1      no
## 2           yes 10.52425 4.5         8         0       1      no
```

```
## 3            no 10.96858 4.6         9         0        0       no
## 4            no 11.10500 3.1        11         2        0       no
## 5            no 11.10847 4.4        12         0        2       no
## 6           yes 11.02825 4.2        12         0        1       no
##   participation_numeric
## 1                     0
## 2                     1
## 3                     0
## 4                     0
## 5                     0
## 6                     1
```

```r
participation.lpm<-lm(participation_numeric~age+youngkids+oldkids+education+income,data=SwissLabor)

kable(tidy(participation.lpm), digits=4,align='c', caption=
"Linear Probability Model for the $participation$ Problem")
```

Table 1: Linear Probability Model for the *participation* Problem

term	estimate	std.error	statistic	p.value
(Intercept)	3.1412	0.4314	7.2823	0.0000
age	-0.1224	0.0187	-6.5340	0.0000
youngkids	-0.2482	0.0326	-7.6075	0.0000
oldkids	-0.0014	0.0161	-0.0848	0.9325
education	-0.0099	0.0057	-1.7379	0.0826
income	-0.1892	0.0417	-4.5349	0.0000

```r
summary(participation.lpm)
```

```
##
## Call:
## lm(formula = participation_numeric ~ age + youngkids + oldkids +
##     education + income, data = SwissLabor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8660 -0.4334 -0.1653  0.4678  1.1030
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.141225   0.431352   7.282 7.38e-13 ***
## age         -0.122417   0.018735  -6.534 1.09e-10 ***
## youngkids   -0.248238   0.032631  -7.608 7.28e-14 ***
## oldkids     -0.001363   0.016079  -0.085   0.9325
## education   -0.009871   0.005680  -1.738   0.0826 .
## income      -0.189190   0.041718  -4.535 6.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4719 on 866 degrees of freedom
## Multiple R-squared:  0.1097, Adjusted R-squared:  0.1046
## F-statistic: 21.34 on 5 and 866 DF,  p-value: < 2.2e-16
```

```r
#Probit Model

participation.probit <- glm(participation~age+youngkids+oldkids+education+income,data=SwissLabor, famil

summary(participation.probit)
```

```
##
## Call:
## glm(formula = participation ~ age + youngkids + oldkids + education +
##     income, family = binomial(link = "probit"), data = SwissLabor)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.656197   1.276056   6.000 1.97e-09 ***
## age         -0.340421   0.053317  -6.385 1.72e-10 ***
## youngkids   -0.708289   0.101074  -7.008 2.42e-12 ***
## oldkids     -0.007107   0.044251  -0.161   0.8724
## education   -0.026645   0.015811  -1.685   0.0919 .
## income      -0.555513   0.122178  -4.547 5.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1203.2  on 871  degrees of freedom
## Residual deviance: 1100.2  on 866  degrees of freedom
## AIC: 1112.2
##
## Number of Fisher Scoring iterations: 4
```

```r
#Logit Model

participation.logit <- glm(participation~age+youngkids+oldkids+education+income,data=SwissLabor, family

summary(participation.logit)
```

```
##
## Call:
## glm(formula = participation ~ age + youngkids + oldkids + education +
##     income, family = binomial(link = "logit"), data = SwissLabor)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 12.43322    2.14404   5.799 6.67e-09 ***
## age         -0.56395    0.08891  -6.343 2.26e-10 ***
## youngkids   -1.22101    0.17579  -6.946 3.76e-12 ***
## oldkids     -0.01635    0.07229  -0.226   0.8211
## education   -0.04585    0.02597  -1.765   0.0775 .
## income      -0.89414    0.20406  -4.382 1.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1203.2  on 871  degrees of freedom
## Residual deviance: 1098.9  on 866  degrees of freedom
## AIC: 1110.9
##
## Number of Fisher Scoring iterations: 4
```

After using the AIC to test for which model was the best for our regression, we determined that the Logit model was the most efficient since it has the lowest AIC of 1110.914 compared to the LPM with an AIC of 1172.812 and probit model with an AIC of 1112.161. The marginal effects for our logit regression is as follows: age with a marginal effect of -0.56395, young kids with a marginal effect of -1.22101, old kids with a marginal effect of -0.01635, education with a marginal effect of -0.04585 , and income with a with a marginal effect of -0.89414. Thus, we can conclude that having an additional kid lowers participation. The same can be said about an increase in education and income. All the variables have a negative marginal effect.

```r
# Calculate AIC for the model
aic_lpm <- AIC(participation.lpm)
print("AIC:")
```

```
## [1] "AIC:"
```

```r
print(aic_lpm)
```

```
## [1] 1172.812
```

```r
aic_probit <- AIC(participation.probit)
print("AIC:")
```

```
## [1] "AIC:"
```

```r
print(aic_probit)
```

```
## [1] 1112.161
```

```r
aic_logit <- AIC(participation.logit)
print("AIC:")
```

```
## [1] "AIC:"
```

```r
print(aic_logit)
```

```
## [1] 1110.914
```

```r
head(marginal_effects(participation.logit))
```

```
##      dydx_age dydx_youngkids dydx_oldkids dydx_education dydx_income
## 1 -0.13225745     -0.2863493 -0.003833718   -0.010752526 -0.20969391
## 2 -0.14062104     -0.3044572 -0.004076152   -0.011432485 -0.22295432
## 3 -0.13599034     -0.2944314 -0.003941923   -0.011056009 -0.21561238
## 4 -0.05101484     -0.1104517 -0.001478756   -0.004147504 -0.08088405
## 5 -0.13035804     -0.2822369 -0.003778660   -0.010598103 -0.20668240
## 6 -0.13644931     -0.2954251 -0.003955227   -0.011093323 -0.21634007
```

```r
print(participation.logit)
```

```
##
## Call:  glm(formula = participation ~ age + youngkids + oldkids + education +
##     income, family = binomial(link = "logit"), data = SwissLabor)
##
## Coefficients:
## (Intercept)          age    youngkids      oldkids    education       income
##    12.43322     -0.56395     -1.22101     -0.01635     -0.04585     -0.89414
##
## Degrees of Freedom: 871 Total (i.e. Null);  866 Residual
## Null Deviance:       1203
## Residual Deviance: 1099  AIC: 1111
```