Econ 104 Project 3

Elaine Tran, Maritza Jimenez, Titania Le, Aanya Pramanik

2023-12-08

Panel Data Model

Question 1a

Christian Kleiber and Achim Zeileis (2008). Applied Econometrics with R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER

The data set "Municipalities" from the AER package is a panel data set for 265 Swedish municipalities that covers a time span of nine years (1979 to 1987). It has five variables (municipality, year, expenditures, revenues, and grants) with a total of 2,385 observations. The municipality variable is a factor with ID numbers for municipalities; the year variable is a factor identifying the year; the expenditures variable codes for total expenditures that include both capital and current expenditures; the revenues variable codes for total own-source revenues; and the grants variable includes the intergovernmental grants received by each municipality. For each municipality ID number, there seems to be the same number of observations each (9), with each observation ranging in year from 1979 to 1987. For the three other variables, they are expressed in million SEK which explains why their values are small (all under 0.035). The series data are in per capita form and are deflated using a municipality-specific price index. Expenditures, revenues, and grants have minimum values of 0.01424, 0.009854, and 0.003921, respectively, and have maximum values of 0.02887, 0.023431, and 0.005886, respectively.

The question we are trying to answer with our data and model is: Is there a model that depicts the relationship between the amount of expenditures and grants different municipalities receive to the revenues that they earn and do they vary across time?

Question 1b

```r
library(AER)
library(xtable)
library(plm)
library(gplots)
library(ggplot2)

data("Municipalities")

Municipal <- subset(Municipalities, municipality == "114" |
                      municipality == "115" | municipality == "120" |
                      municipality == "123" | municipality == "125")

Municipal_pd <- pdata.frame(Municipal, index = c("municipality", "year"))
head(Municipal_pd)
```

```
##          municipality year expenditures  revenues    grants
## 114-1979          114 1979    0.0229736 0.0181770 0.0054429
## 114-1980          114 1980    0.0266307 0.0209142 0.0057304
## 114-1981          114 1981    0.0273253 0.0210836 0.0056647
## 114-1982          114 1982    0.0288704 0.0234310 0.0058859
## 114-1983          114 1983    0.0226474 0.0179979 0.0055908
## 114-1984          114 1984    0.0215601 0.0179949 0.0047536
```

```r
summary(Municipal_pd)
```

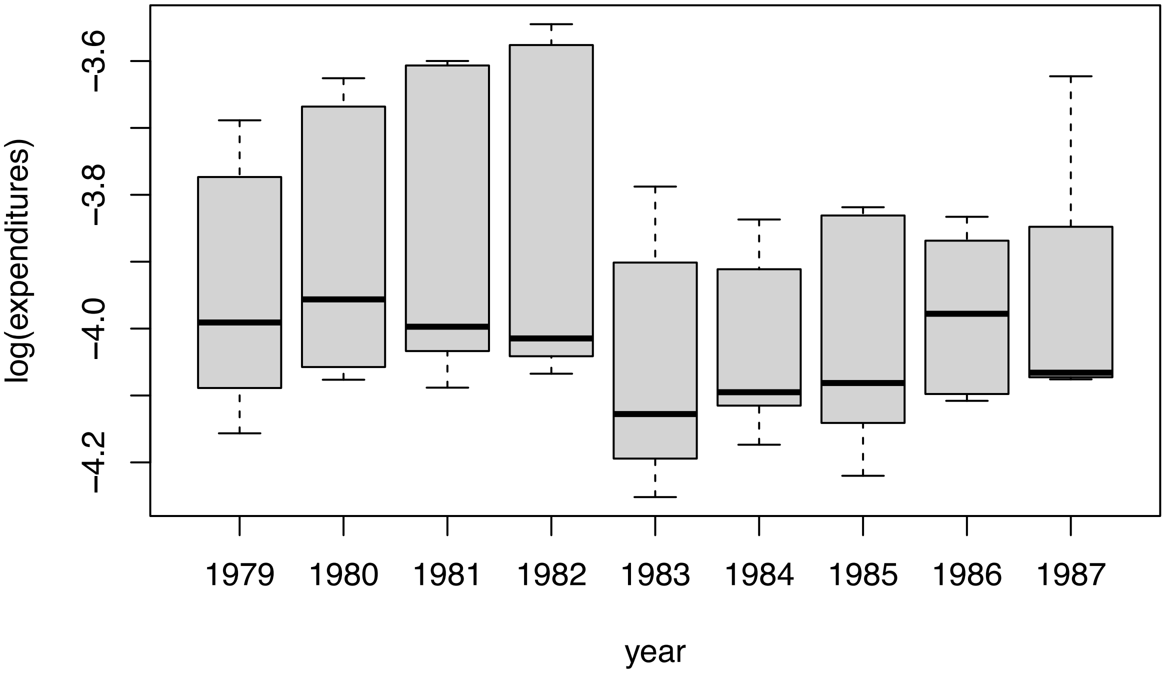
```
##  municipality      year     expenditures        revenues
##  114:9        1979   : 5   Min.   :0.01424   Min.   :0.009854
##  115:9        1980   : 5   1st Qu.:0.01676   1st Qu.:0.012287
##  120:9        1981   : 5   Median :0.01805   Median :0.013748
##  123:9        1982   : 5   Mean   :0.01965   Mean   :0.015045
##  125:9        1983   : 5   3rd Qu.:0.02169   3rd Qu.:0.017690
##               1984   : 5   Max.   :0.02887   Max.   :0.023431
##               (Other):15
##      grants
##  Min.   :0.003921
##  1st Qu.:0.004386
##  Median :0.004644
##  Mean   :0.004775
##  3rd Qu.:0.005272
##  Max.   :0.005886
##
```

```r
pdim(Municipal_pd)
```

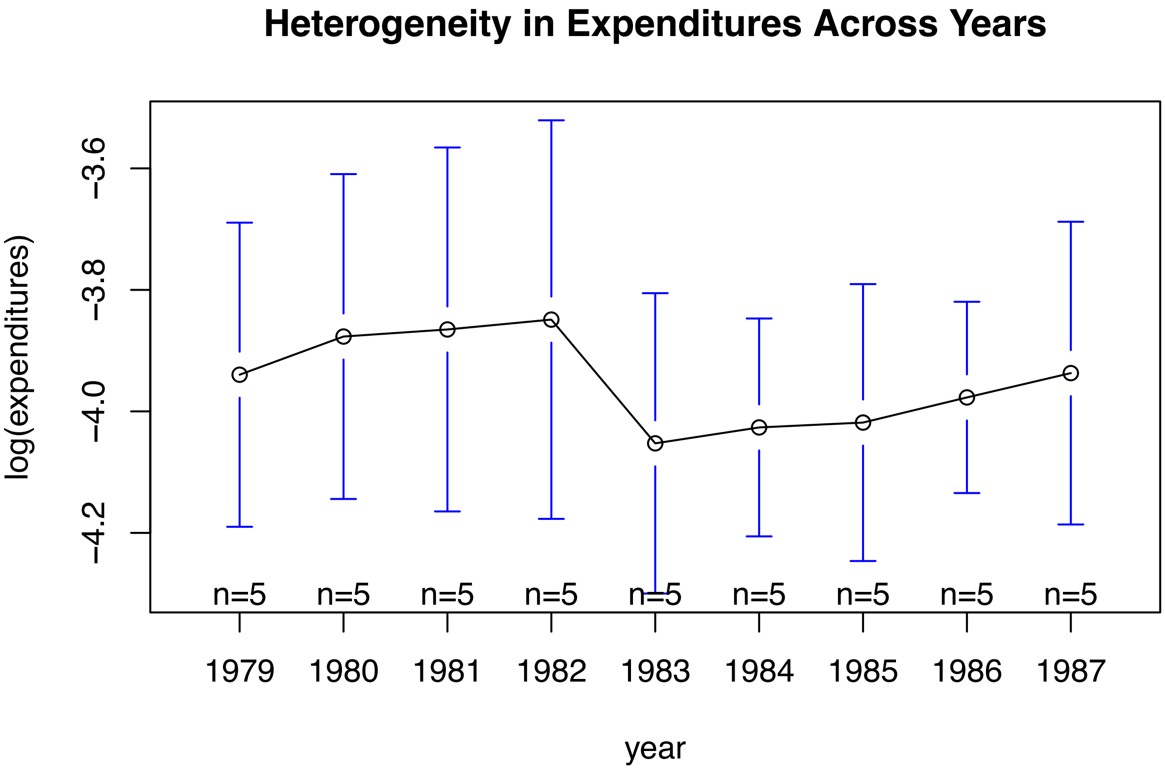
```
## Balanced Panel: n = 5, T = 9, N = 45
```

Because the dataset is so large, a subset of Municipalities was taken to make subsequent analyses easier to evaluate and understand. 5 municipalities were chosen, each with 9 observations related to it. The data was then converted to be a panel data and the function pdim() was used to extract the dimensions of the panel data to determine whether it is balanced or not. The results indicate that it is a balanced panel (n = 5, T = 9, N = 45) in which all units are observed in all periods. If the panel were unbalanced, that suggests some units are missing in some periods. In addition, since the cross-sectional dimension (N) is much larger than the longitudinal dimension (T), this panel can be characterized as "short and wide" rather than a "long and narrow" panel where the opposite (N<T) is true.

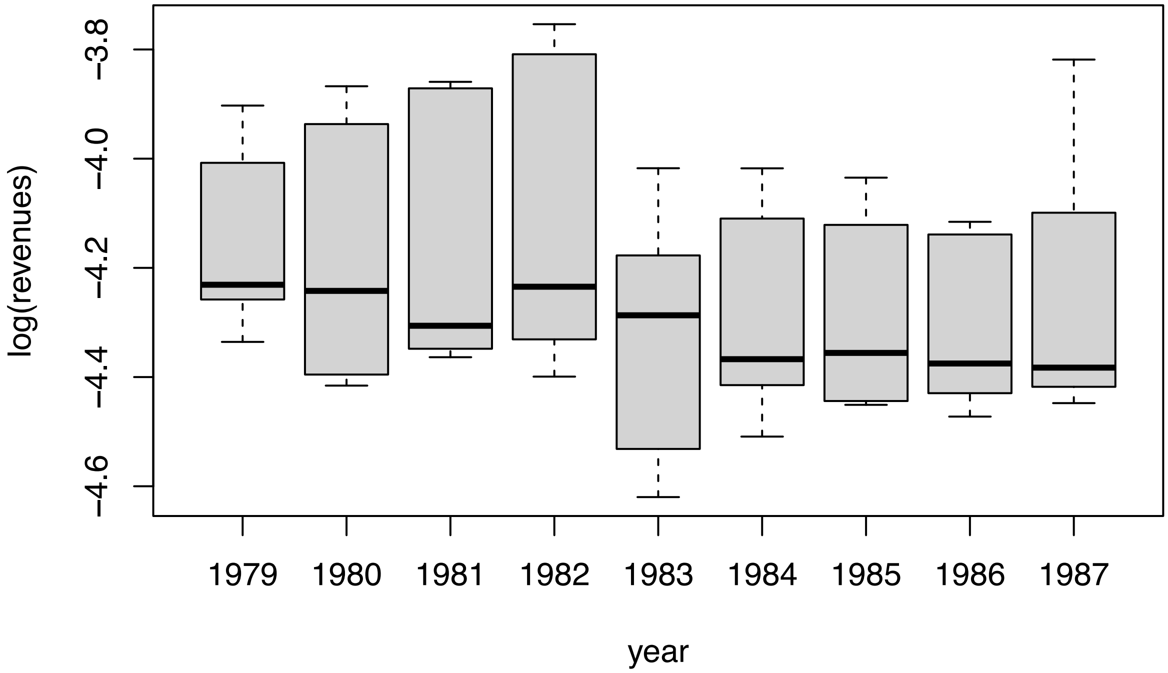
```
scatterplot(log(expenditures) ~year|municipality, data = Municipal_pd)
```



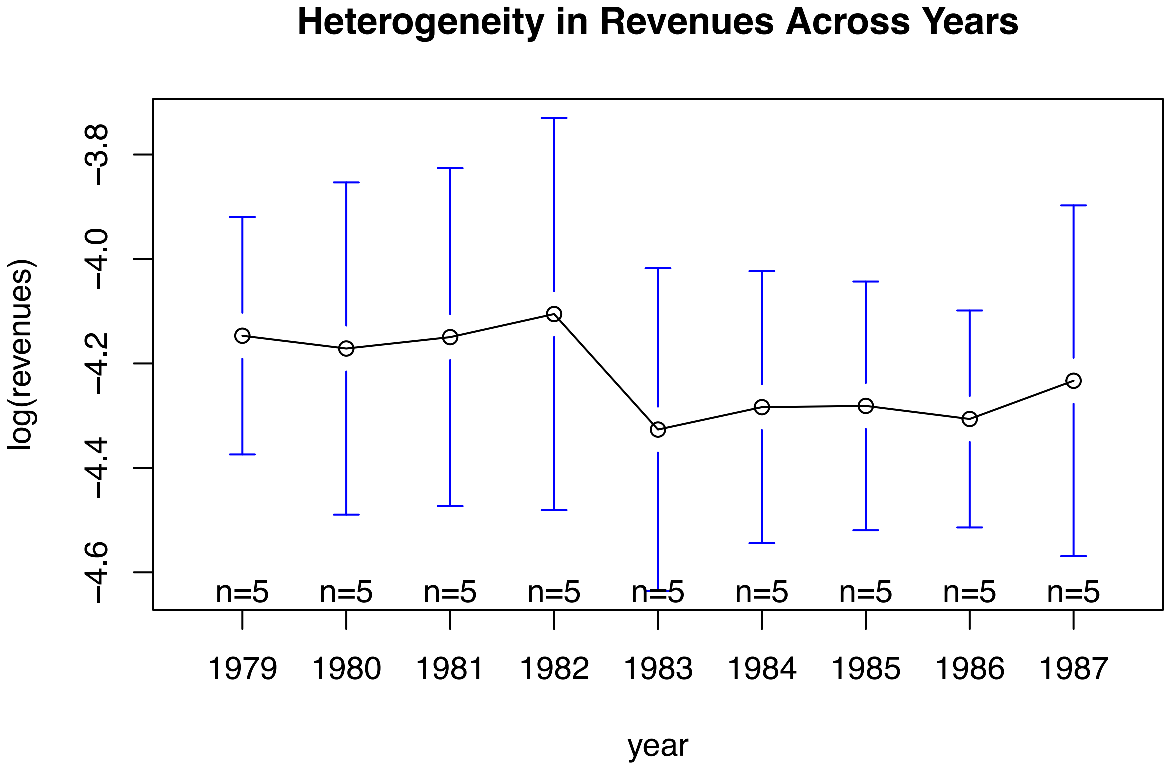
```
plotmeans(log(expenditures) ~year,
          main = 'Heterogeneity in Expenditures Across Years', data = Municipal_pd)
```



```
scatterplot(log(revenues) ~year|municipality, data = Municipal_pd)
```

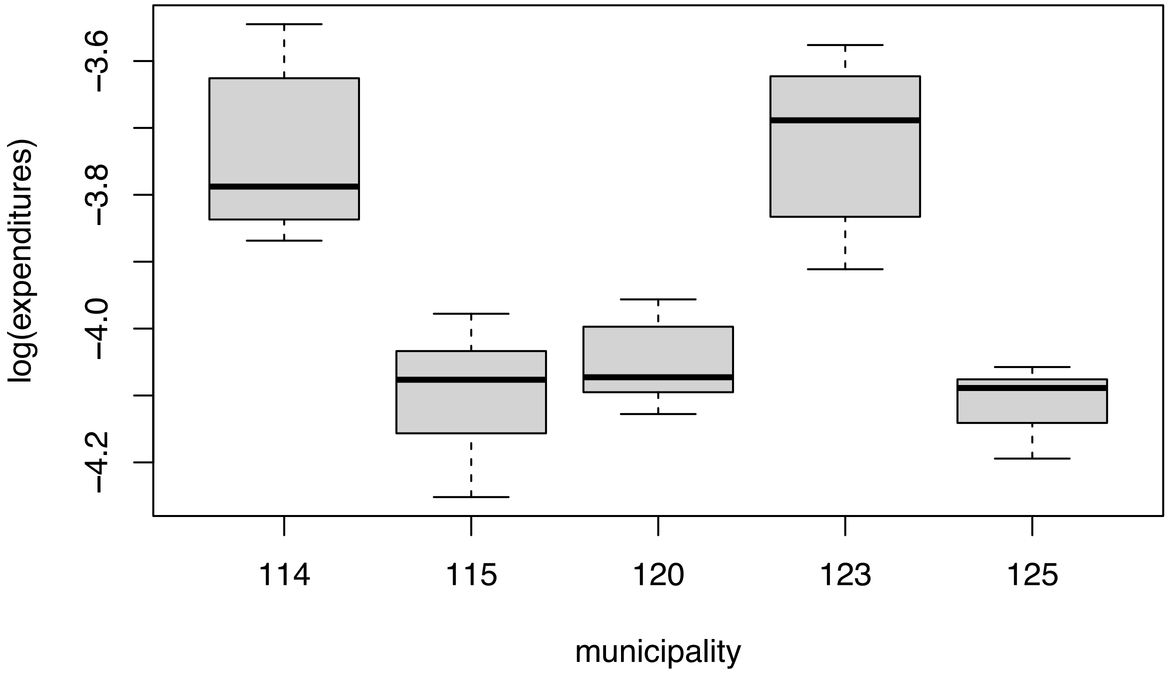


```
plotmeans(log(revenues) ~year,
          main = 'Heterogeneity in Revenues Across Years', data = Municipal_pd)
```

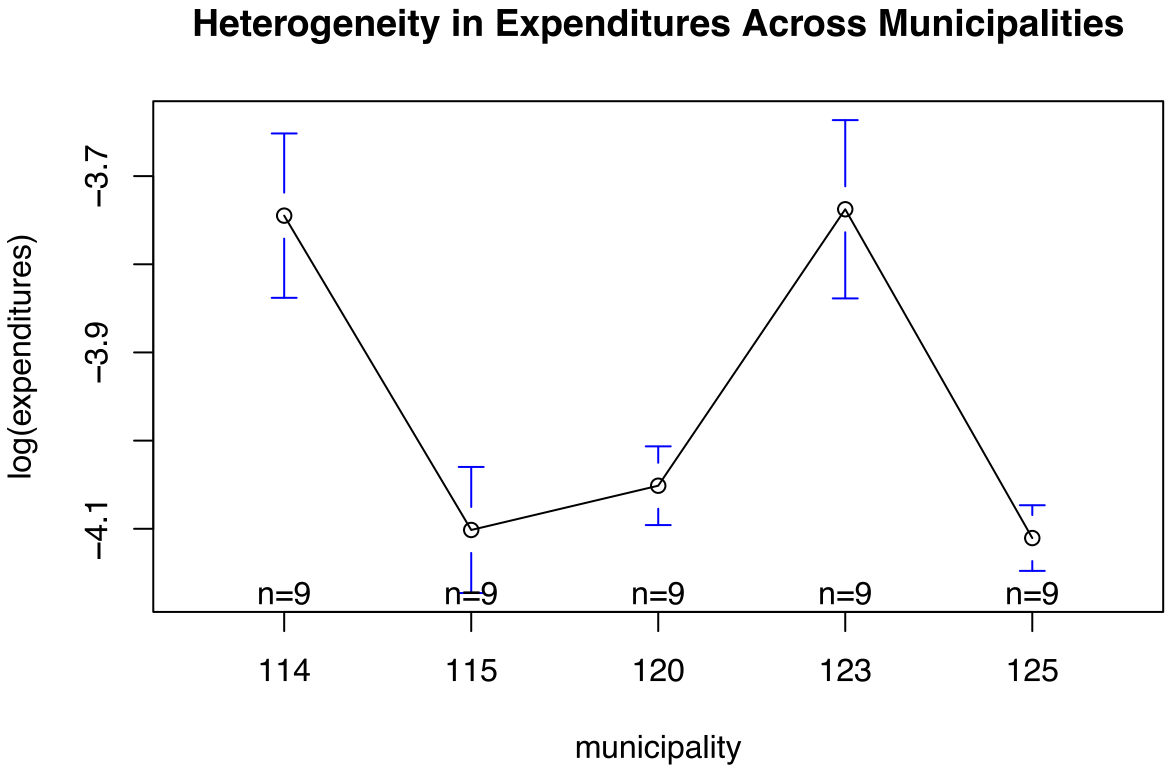


The four graphs above are a depiction of heterogeneity across time (years) for expenditures and revenues separately. We are able to visualize cell means with their interquartile ranges for each year, and it appears they do not stay around a constant average value but instead vary widely across individual units.

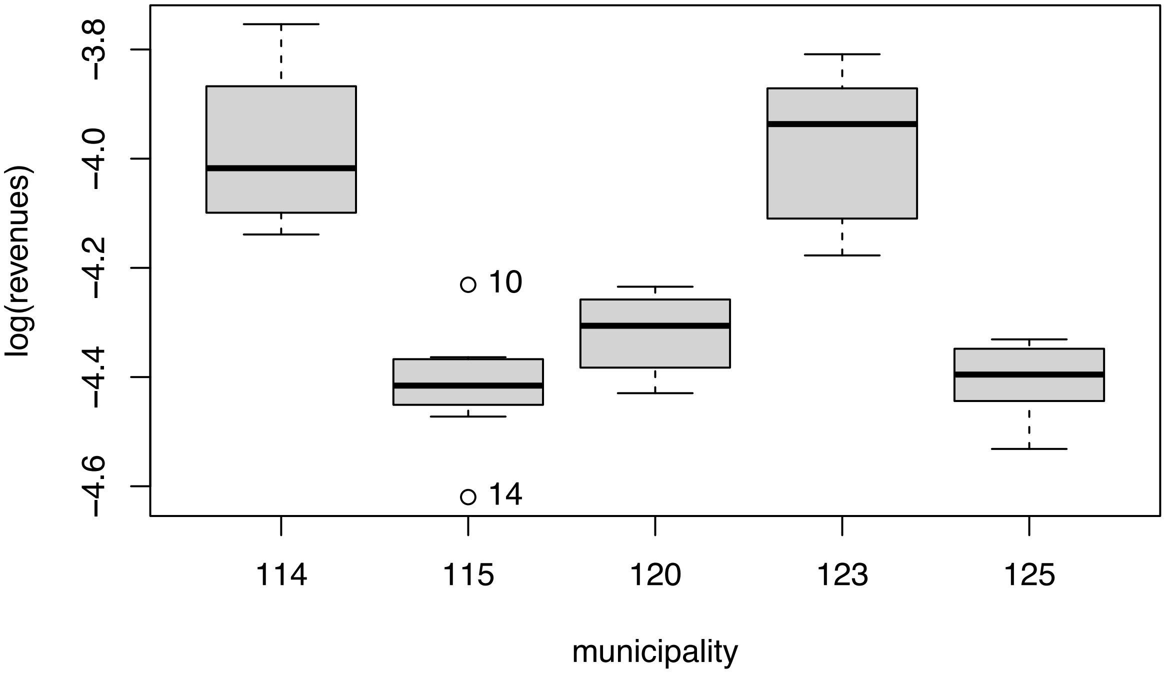
```r
scatterplot(log(expenditures) ~municipality|year, data = Municipal_pd)
```



```r
plotmeans(log(expenditures) ~municipality,
          main = 'Heterogeneity in Expenditures Across Municipalities',
          data = Municipal_pd)
```

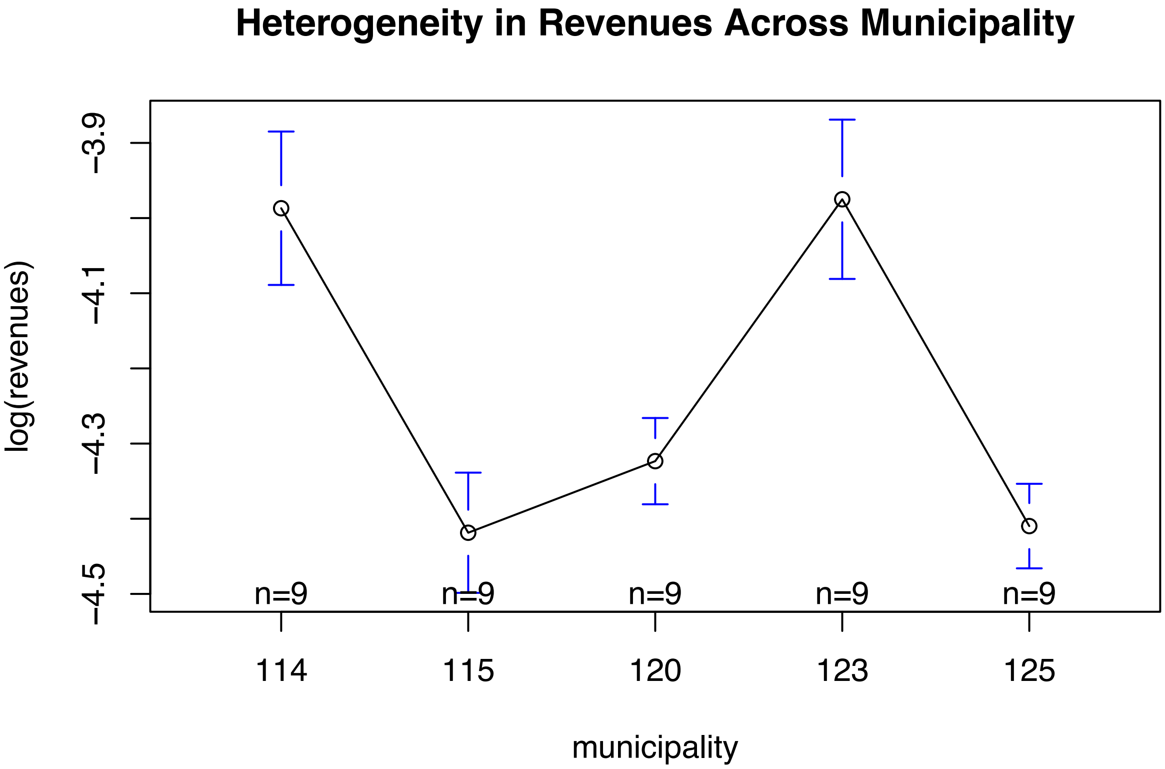


```
scatterplot(log(revenues) ~municipality|year, data = Municipal_pd)
```



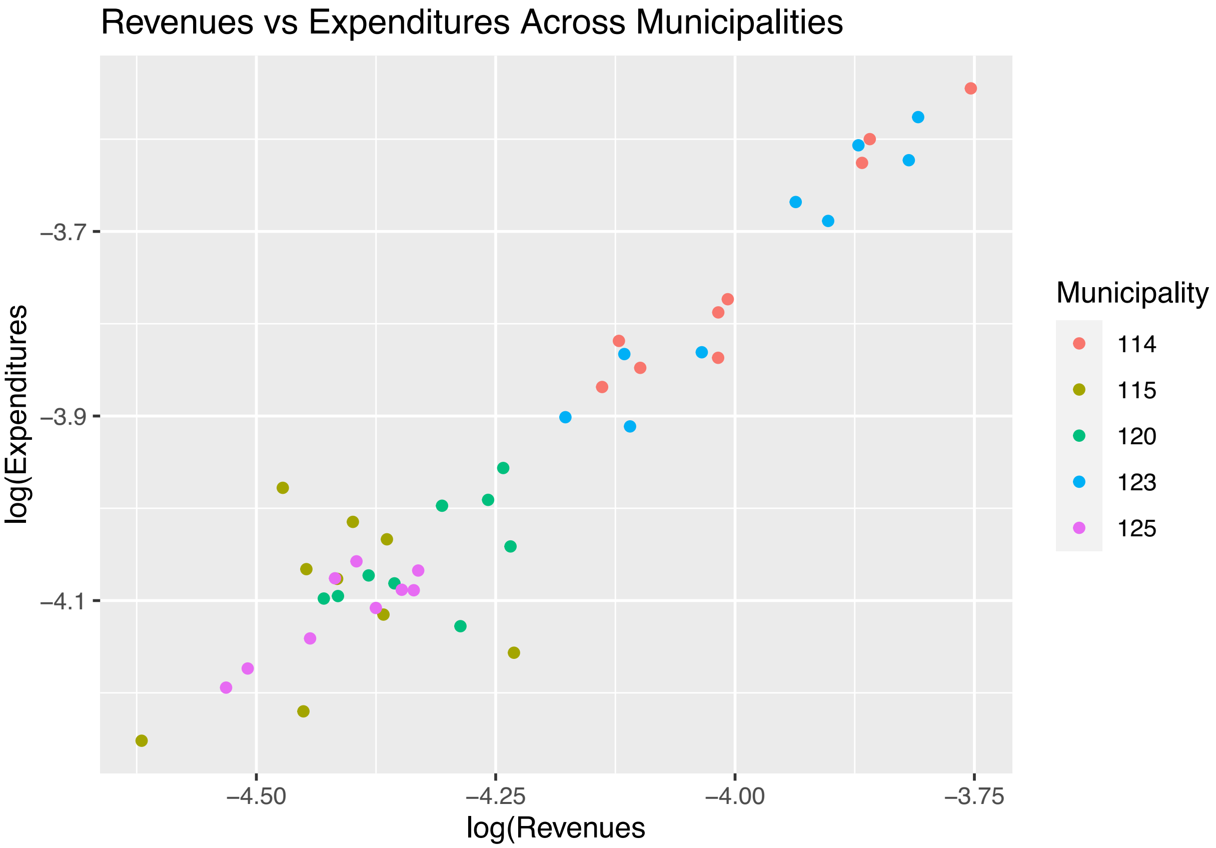
```
## [1] "14" "10"
```

```
plotmeans(log(revenues) ~municipality,
          main = 'Heterogeneity in Revenues Across Municipality', data = Municipal_pd)
```



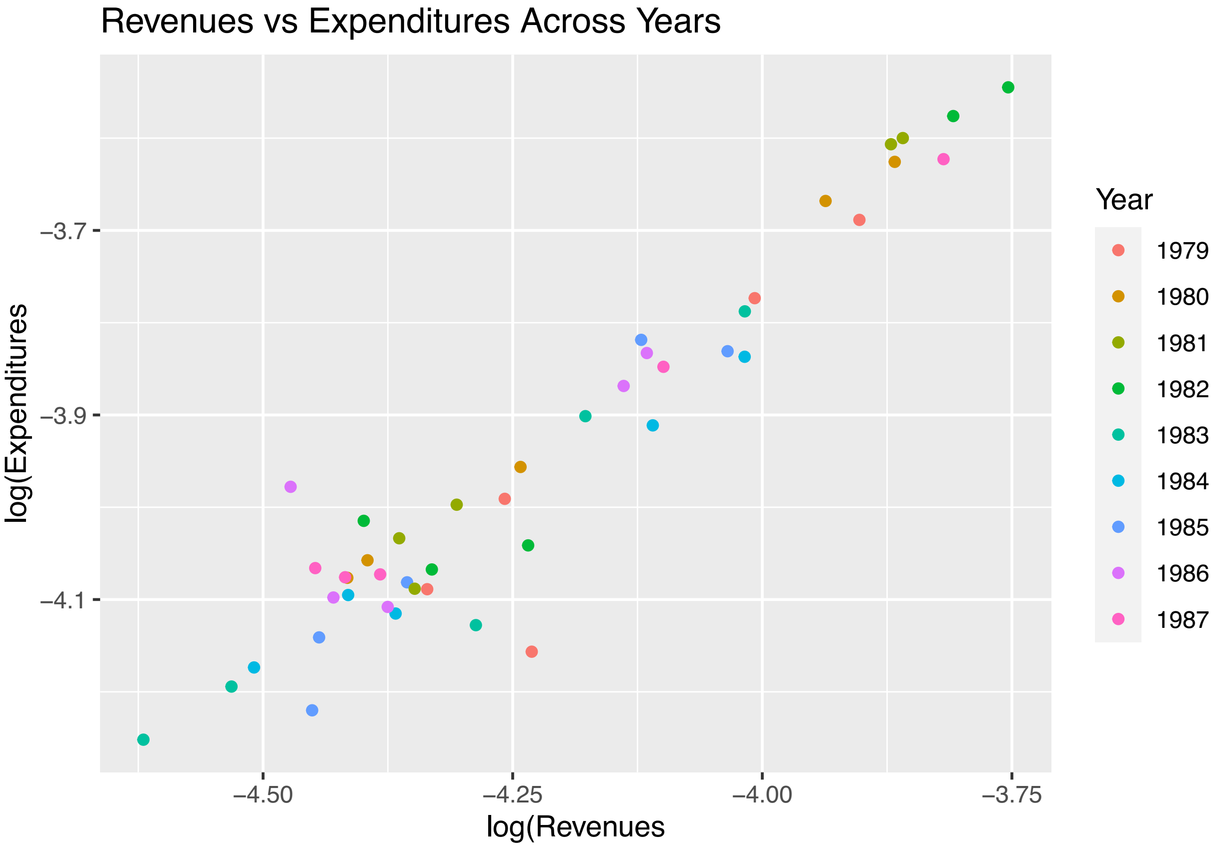
The four graphs above are another depiction of heterogeneity across municipalities for expenditures and revenues separately. There is significant variation seen regarding the means of both variables across municipalities, indicating that there is significant individual heterogeneity in the data.

```r
ggplot(Municipal_pd, aes(x=log(revenues),y=log(expenditures),colour=factor(municipality)))+
  geom_point()+ggtitle('Revenues vs Expenditures Across Municipalities') +
  xlab("log(Revenues") +ylab("log(Expenditures")+scale_colour_discrete(name="Municipality")
```



With the plotted data, we can visualize Revenues vs Expenditures by Municipality. A positive correlation can be seen between the two variables where it seems revenues and expenditures of municipalities 114 and 123 move together at the top right quadrant whereas there's paralleled movement with the other three municipalities in the lower left quadrant. This could indicate possible heterogeneity that could affect subsequent data analyses.

```
ggplot(Municipal_pd, aes(x=log(revenues), y=log(expenditures), colour=factor(year))) +
  geom_point() + ggtitle('Revenues vs Expenditures Across Years') + xlab("log(Revenues") +
  ylab("log(Expenditures") + scale_colour_discrete(name="Year")
```



The above graph is a visualization of Revenues vs Expenditures by Year. Like the previous graph, a similar positive relationship can be seen between the two variables across time (years). What is different is there is not any groups of years that move with one another. This could indicate possible heterogeneity in the data that should be considered when creating a best fitting model.

Question 1c

Pooled Model

```
pooledreg1 <- plm(grants ~ expenditures + revenues, model = "pooling",
                  data = Municipal_pd)

pooledreg1
```

```
##
## Model Formula: grants ~ expenditures + revenues
##
## Coefficients:
##  (Intercept) expenditures     revenues
##    0.0026204    0.1118317   -0.0028283
```

Fixed Effects Model

```
fixedeffect <- plm(grants ~ expenditures + revenues, model = "within",
                   data = Municipal_pd)

fixef(fixedeffect)
```

```
##       114       115       120       123       125
## 0.0040357 0.0036102 0.0032419 0.0037597 0.0033418
```

```
fixedeffect
```

```
##
## Model Formula: grants ~ expenditures + revenues
##
## Coefficients:
## expenditures     revenues
##     0.071050    -0.014535
```

Random Effects Model

```
randomeffects <- plm(grants ~ expenditures + revenues, model = "random",
                     data = Municipal_pd)

randomeffects
```

```
##
## Model Formula: grants ~ expenditures + revenues
##
## Coefficients:
##  (Intercept) expenditures     revenues
##    0.0028366    0.0961231    0.0033148
```

Test to Determine Best Model

```r
REtest <- plmtest(pooledreg1, effect = "individual")
REtest
```

```
##
##  Lagrange Multiplier Test - (Honda)
##
## data:  grants ~ expenditures + revenues
## normal = 3.4154, p-value = 0.0003184
## alternative hypothesis: significant effects
```

The random effects test produced a p-value (0.0003184) much smaller than our significance level of 0.05, so we reject the null hypothesis and conclude heterogeneity among individuals may be significant. This provides evidence that the pooled model should not be used.

```r
hausmen_test <- phtest(fixedeffect, randomeffects)
hausmen_test
```

```
##
##  Hausman Test
##
## data:  grants ~ expenditures + revenues
## chisq = 8.9164, df = 2, p-value = 0.01158
## alternative hypothesis: one model is inconsistent
```

The Hausmen test produces a p-value (0.01158) that is less than 0.05, so we reject the null hypothesis and conclude that the fixed effect model is a better fit than the random effects model. In this case, the Hausman Taylor estimator may be used.

Binary Dependent Variables

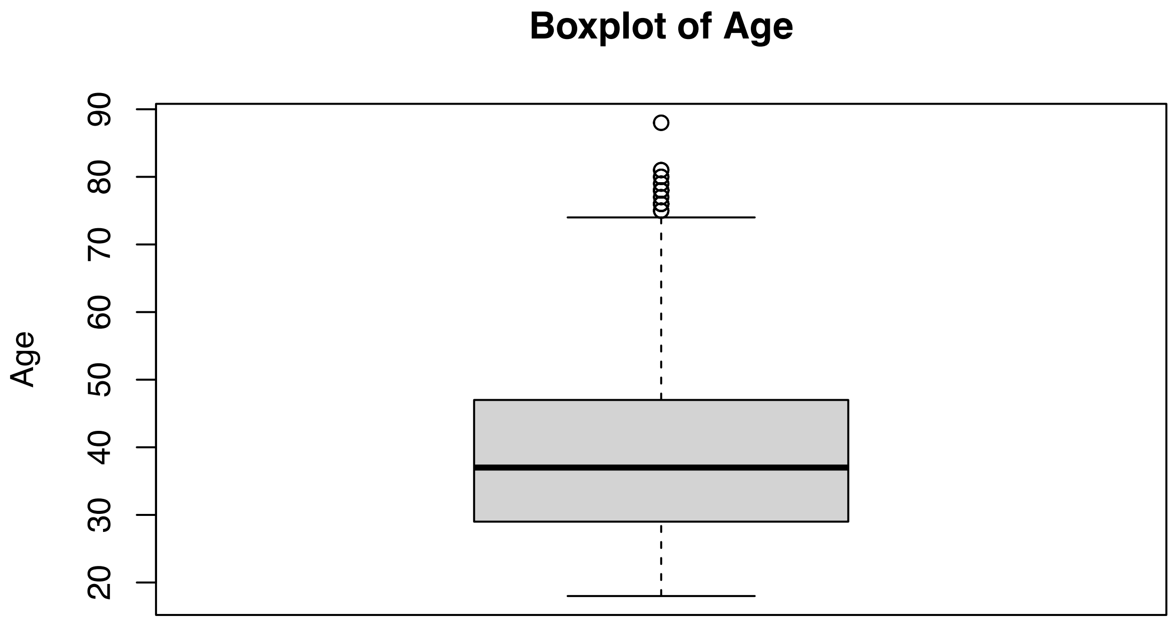
Question 2a

SmokeBan is a cross-sectional data set found in the AER package that estimates the impacts of workplace smoking bans on smoking of indoor workers. It contains 10,000 observations and 7 variables: smoker, ban, education, hispanic, gender, and age being the only numeric variable.

Question: Do factors such as age, gender, and race affect smoking behavior more than smoking bans do?

Question 2b

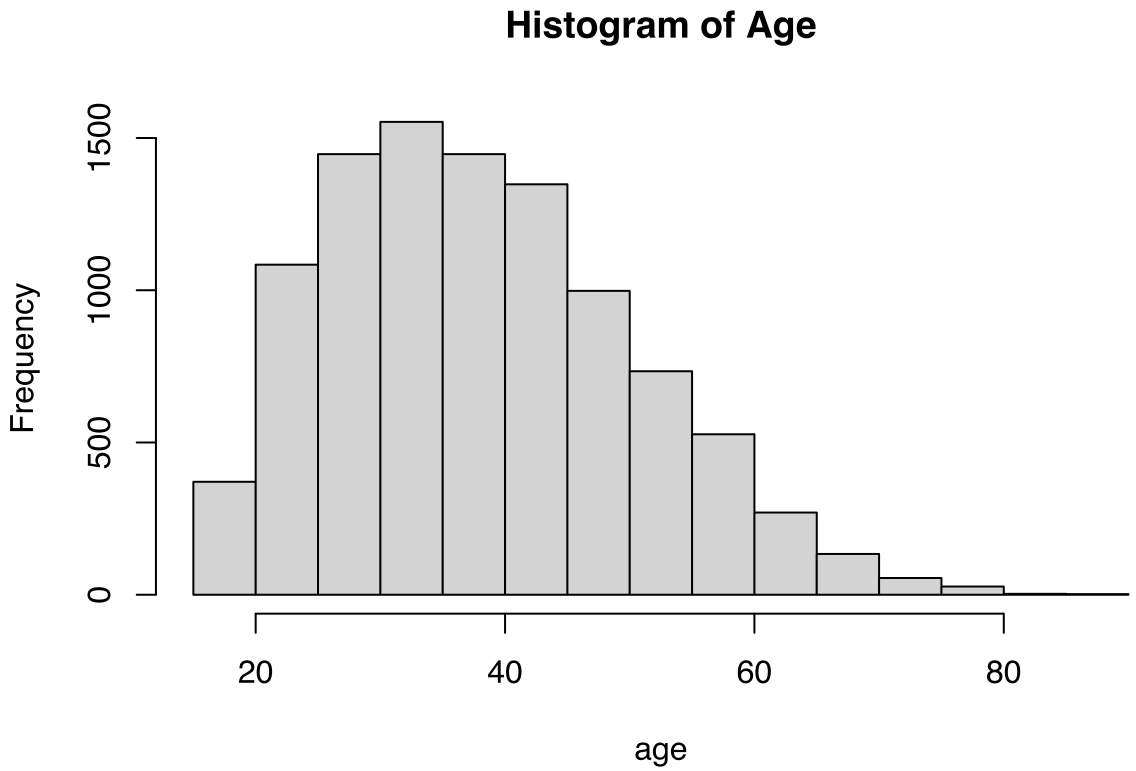
```
boxplot(SmokeBan$age, main= "Boxplot of Age", ylab= "Age")
```



```
summary(SmokeBan$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   29.00   37.00   38.69   47.00   88.00
```

```
hist(SmokeBan$age, main="Histogram of Age", xlab="age")
```



The histogram of age depicts data that is skewed right with the majority of workers between the ages of 25-45. Both the boxplot and 5 number summary of age show that the youngest workers are 18, the youngest 25% are roughly 29, the average age is 37, the oldest 25% are 47 and the oldest workers are 88.

```
smoker_prop<-table(SmokeBan$smoker)
smoker_perc<-prop.table(smoker_prop)*100
smoker_perc
```

```
##
##    no   yes
## 75.77 24.23
```

75.77% of workers in the data set are currently non-smokers while 24.23% are currently smokers.

```
ban_prop<-table(SmokeBan$ban)
ban_perc<-prop.table(ban_prop)*100
ban_perc
```

```
##
##    no   yes
## 39.02 60.98
```

39.02% of the observations in the data set do not have a work area smoking ban while 60.98% do.

```
educ_prop<-table(SmokeBan$education)
educ_perc<-prop.table(educ_prop)*100
educ_perc
```

```
##
##  hs drop out           hs some college      college       master
##         9.12        32.66        28.02        19.72        10.48
```

9.12% of the workers in the data set are high school drop outs, 32.66% are high school graduates, 28.02% have some college education, 19.72% are college graduates, and 10.48% of the workers have a masters degree.

```
afam_prop<-table(SmokeBan$afam)
afam_perc<-prop.table(afam_prop)*100
afam_perc
```

```
##
##    no   yes
## 92.31  7.69
```

92.31% of workers in the data set are not African American while 7.69% are.

```
hispanic_prop<-table(SmokeBan$hispanic)
hispanic_perc<-prop.table(hispanic_prop)*100
hispanic_perc
```

```
##
##    no   yes
## 88.66 11.34
```

88.66% of workers in the data set are not Hispanic while 11.34% are.

```
gender_prop<-table(SmokeBan$gender)
gender_perc<-prop.table(gender_prop)*100
gender_perc
```

```
##
##   male female
##  43.63  56.37
```

43.63% of workers in the data set are male while 56.37% are female.

Question 2c

Linear Probability Model

```
SmokeBan$smokernum <- as.numeric(SmokeBan$smoker) - 1
smoker.lpm <- lm(smokernum ~ ban + age + education + afam + hispanic + gender,
                 data=SmokeBan)
kable(tidy(smoker.lpm), digits=4, align='c', caption="Linear Probability Model")
```

Table 1: Linear Probability Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.5112	0.0216	23.7080	0.0000
banyes	-0.0453	0.0087	-5.1970	0.0000
age	-0.0014	0.0004	-3.8666	0.0001
educationhs	-0.0858	0.0163	-5.2742	0.0000
educationsome college	-0.1537	0.0166	-9.2722	0.0000
educationcollege	-0.2684	0.0176	-15.2420	0.0000
educationmaster	-0.3099	0.0197	-15.6944	0.0000
afamyes	-0.0265	0.0158	-1.6826	0.0925
hispanicyes	-0.1037	0.0139	-7.4389	0.0000
genderfemale	-0.0329	0.0085	-3.8454	0.0001

```r
summary(smoker.lpm)
```

```
##
## Call:
## lm(formula = smokernum ~ ban + age + education + afam + hispanic +
##     gender, data = SmokeBan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48682 -0.28725 -0.17239 -0.03619  0.99792
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.5111949  0.0215621  23.708  < 2e-16 ***
## banyes                -0.0453435  0.0087250  -5.197 2.07e-07 ***
## age                   -0.0013543  0.0003503  -3.867 0.000111 ***
## educationhs           -0.0858065  0.0162692  -5.274 1.36e-07 ***
## educationsome college -0.1537486  0.0165818  -9.272  < 2e-16 ***
## educationcollege      -0.2683776  0.0176077 -15.242  < 2e-16 ***
## educationmaster       -0.3099189  0.0197471 -15.694  < 2e-16 ***
## afamyes               -0.0265034  0.0157518  -1.683 0.092491 .
## hispanicyes           -0.1037449  0.0139463  -7.439 1.10e-13 ***
## genderfemale          -0.0328743  0.0085489  -3.845 0.000121 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.417 on 9990 degrees of freedom
## Multiple R-squared:  0.05397,    Adjusted R-squared:  0.05312
## F-statistic: 63.33 on 9 and 9990 DF,  p-value: < 2.2e-16
```

```r
models <- list(smoker.lpm)
model.names <- c('LPM')
aictab(cand.set = models, modnames = model.names)
```

```
##
## Model selection based on AICc:
##
##      K     AICc Delta_AICc AICcWt Cum.Wt       LL
## LPM 11 10895.47          0      1      1 -5436.72
```

```
hcErrors <- coeftest(smoker.lpm,vcov.=hccm(smoker.lpm,type="hc1"))

### Probit Model
smoker.probit <- glm(smoker ~ ban + age + education + afam + hispanic +
      gender, family=binomial(link="probit"),
      data=SmokeBan)
kable(tidy(smoker.probit), digits=4, align='c', caption="Probit Model")
```

Table 2: Probit Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.1100	0.0696	1.5804	0.1140
banyes	-0.1518	0.0289	-5.2426	0.0000
age	-0.0042	0.0012	-3.5963	0.0003
educationhs	-0.2424	0.0507	-4.7774	0.0000
educationsome college	-0.4450	0.0524	-8.4980	0.0000
educationcollege	-0.8718	0.0585	-14.8961	0.0000
educationmaster	-1.0942	0.0716	-15.2928	0.0000
afamyes	-0.0797	0.0527	-1.5115	0.1306
hispanicyes	-0.3327	0.0480	-6.9312	0.0000
genderfemale	-0.1106	0.0288	-3.8439	0.0001

```
summary(smoker.probit)
```

```
##
## Call:
## glm(formula = smoker ~ ban + age + education + afam + hispanic +
##     gender, family = binomial(link = "probit"), data = SmokeBan)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.109990   0.069597   1.580 0.114018
## banyes                -0.151762   0.028948  -5.243 1.58e-07 ***
## age                   -0.004203   0.001169  -3.596 0.000323 ***
## educationhs           -0.242373   0.050733  -4.777 1.78e-06 ***
## educationsome college -0.444975   0.052362  -8.498  < 2e-16 ***
## educationcollege      -0.871756   0.058523 -14.896  < 2e-16 ***
## educationmaster       -1.094230   0.071552 -15.293  < 2e-16 ***
## afamyes               -0.079690   0.052721  -1.512 0.130650
## hispanicyes           -0.332704   0.048001  -6.931 4.17e-12 ***
## genderfemale          -0.110625   0.028779  -3.844 0.000121 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11074  on 9999  degrees of freedom
## Residual deviance: 10505  on 9990  degrees of freedom
## AIC: 10525
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
### Logit Model
smoker.logit <- glm(smoker ~ ban + age + education
       + afam + hispanic + gender,
       family=binomial(link="logit"), data=SmokeBan)
kable(tidy(smoker.logit), digits=4, align='c', caption="Logit Model")
```

Table 3: Logit Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.2343	0.1160	2.0196	0.0434
banyes	-0.2507	0.0492	-5.0999	0.0000
age	-0.0075	0.0020	-3.7509	0.0002
educationhs	-0.4078	0.0831	-4.9089	0.0000
educationsome college	-0.7510	0.0865	-8.6807	0.0000
educationcollege	-1.5063	0.1007	-14.9658	0.0000
educationmaster	-1.9311	0.1313	-14.7117	0.0000
afamyes	-0.1495	0.0900	-1.6609	0.0967
hispanicyes	-0.5848	0.0831	-7.0391	0.0000
genderfemale	-0.1887	0.0491	-3.8432	0.0001

```
summary(smoker.logit)
```

```
##
## Call:
## glm(formula = smoker ~ ban + age + education + afam + hispanic +
##     gender, family = binomial(link = "logit"), data = SmokeBan)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.234310   0.116019   2.020 0.043427 *
## banyes                -0.250735   0.049164  -5.100 3.40e-07 ***
## age                   -0.007452   0.001987  -3.751 0.000176 ***
## educationhs           -0.407770   0.083067  -4.909 9.16e-07 ***
## educationsome college -0.750995   0.086513  -8.681  < 2e-16 ***
## educationcollege      -1.506322   0.100651 -14.966  < 2e-16 ***
## educationmaster       -1.931075   0.131261 -14.712  < 2e-16 ***
## afamyes               -0.149472   0.089994  -1.661 0.096732 .
## hispanicyes           -0.584845   0.083085  -7.039 1.93e-12 ***
## genderfemale          -0.188720   0.049105  -3.843 0.000121 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11074  on 9999  degrees of freedom
## Residual deviance: 10502  on 9990  degrees of freedom
## AIC: 10522
##
## Number of Fisher Scoring iterations: 4
```

```r
### 3 Binary Dependent Variable Models
stargazer(hcErrors, smoker.probit, smoker.logit,
  header=FALSE,
  title="Three Binary Choice Models for the $SmokeBan$ Problem",
  type="text",
  keep.stat="n",digits=4, single.row=FALSE,
  intercept.bottom=FALSE,
  model.names=FALSE,
  column.labels=c("LPM","probit","logit"),
  omit.table.layout="n")
```

```
##
## Three Binary Choice Models for the SmokeBan Problem
## ======================================================
##                             Dependent variable:
##                        --------------------------------
##                                        smoker
##                           LPM       probit     logit
##                           (1)        (2)        (3)
## ------------------------------------------------------
## Constant              0.5112***    0.1100    0.2343**
##                        (0.0232)   (0.0696)   (0.1160)
##
## banyes                -0.0453*** -0.1518*** -0.2507***
##                        (0.0090)   (0.0289)   (0.0492)
##
## age                   -0.0014*** -0.0042*** -0.0075***
##                        (0.0003)   (0.0012)   (0.0020)
##
## educationhs           -0.0858*** -0.2424*** -0.4078***
##                        (0.0184)   (0.0507)   (0.0831)
##
## educationsome college -0.1537*** -0.4450*** -0.7510***
##                        (0.0186)   (0.0524)   (0.0865)
##
## educationcollege      -0.2684*** -0.8718*** -1.5063***
##                        (0.0188)   (0.0585)   (0.1007)
##
## educationmaster       -0.3099*** -1.0942*** -1.9311***
##                        (0.0194)   (0.0716)   (0.1313)
##
## afamyes                -0.0265    -0.0797    -0.1495*
##                        (0.0161)   (0.0527)   (0.0900)
##
## hispanicyes           -0.1037*** -0.3327*** -0.5848***
##                        (0.0140)   (0.0480)   (0.0831)
##
## genderfemale          -0.0329*** -0.1106*** -0.1887***
##                        (0.0086)   (0.0288)   (0.0491)
##
## ------------------------------------------------------
## Observations                       10,000     10,000
## ======================================================
### Average Marginal Effects
margins(smoker.lpm)
```

```
## Average marginal effects

## lm(formula = smokernum ~ ban + age + education + afam + hispanic +     gender, data = SmokeBan)

##        age   banyes educationhs educationsome college educationcollege
##  -0.001354 -0.04534    -0.08581               -0.1537          -0.2684
##  educationmaster afamyes hispanicyes genderfemale
##          -0.3099 -0.0265     -0.1037     -0.03287
```

```
margins(smoker.probit)
```

```
## Average marginal effects

## glm(formula = smoker ~ ban + age + education + afam + hispanic +     gender, family = binomial(link

##        age   banyes educationhs educationsome college educationcollege
##  -0.001245 -0.04555    -0.08922               -0.1564          -0.2687
##  educationmaster  afamyes hispanicyes genderfemale
##          -0.3104 -0.02306    -0.08969     -0.03293
```

```
margins(smoker.logit)
```

```
## Average marginal effects

## glm(formula = smoker ~ ban + age + education + afam + hispanic +     gender, family = binomial(link

##        age   banyes educationhs educationsome college educationcollege
##  -0.001293 -0.04414    -0.09215               -0.1601           -0.272
##  educationmaster  afamyes hispanicyes genderfemale
##          -0.3137 -0.02518    -0.09074     -0.03297
```

The preferred model for the SmokeBan data is the logit model based on the AIC values, which are 10895.47, 10525, and 10522 for the linear probability model, probit model, and logit model, respectively. Since the logit model has the lowest AIC, it is the best fit for the data.

The average marginal effects for all the models, including the preferred logit model, demonstrate that all factors (ban, age, education, race, gender) reduce the probability that one smokes. The education factor appears to have the greatest negative effect on smoking, with higher levels of education resulting in a lower smoking probability (higher negative magnitude of marginal effects). This could indicate that workplace smoking bans do not actually have a significant impact on smoking habits, whereas factors like education do.