STATISTICS 151A FINAL PROJECT

TIMOTHY WANG
23073372

Table of Contents

CASE I

- HOW DO RACE AND INCOME AFFECT PRESIDENTIAL ELECTION RESULTS? -

1 Introduction

In several publications, statistician Andrew Gelman from Columbia University has displayed his finding on the relationship between income and voting patterns in the presidential elections. The results of his analyses revealed a consistently "positive association between income and probability of voting Republican," with the strength of this association depending on state-level income distributions. For poorer states, this association is stronger. [1]

Now I want to know whether race has anything to do with this association, "since African-Americans vote heavily Democratic and tend to be poorer than whites."[2] For this analysis, I will be using exit poll data from the National Annenberg Election Survey of the 2004 Presidential Election. This survey contains 69,663 observations total from the 48 contiguous states of the United States, i.e. all states besides Alaska and Hawaii.

2 Data Exploration

As mentioned above, my original data set has 69,663 observations. However, not all of the observations in this data set have complete information, and my concern lies in the observations without information for the variables in which I am interested. Since the Democratic and Republican Parties tend to be the ones that win the presidential elections, I will only be considering the observations that voted for either Bush (Republican) or Kerry (Democrat). Additionally, I am primarily interested in studying the effects of income and race on voting, specifically for those with race white or black. Thus, I can ignore the observations that either did not report income level, or are not indicated as white or black racially. After taking all of the irrelevant observations out of the data set, I am left with 38,562 observations, which is still a sufficient size to run my analysis.

One preliminary problem with the data set is the disparity between the amounts of data on whites (35,150) and on blacks (3,412). Thus, I will be using proportions of the total number of whites or blacks to help conduct my analysis in an attempt to normalize the data on the two separate populations.

As mentioned earlier, I have a slight inclination that African-Americans tend to vote more Democratic, and tend to be poorer than white people. Calculating

[1] From the assignment PDF file, page 2.
[2] Ibid.

these proportions (Figure 1.1) and creating some graphics comparing the distributions (Figures 1.2 and 1.3) seems to confirm this inclination.

FIGURE 1.1: DISTRIBUTIONS OF INCOME LEVEL AND PARTY VOTED BY RACE			**Party Voted**	**Voted Democrat**	**Voted Republican**
			Whites	45.474%	54.526%
			Blacks	90.211%	9.789%
Income Level	**Under $10,000**	**$10,000 - $15,000**	**$15,000 - $25,000**	**$25,000 - $35,000**	**$35,000 - $50,000**
Whites	18.100%	11.815%	37.650%	25.081%	7.354%
Blacks	30.803%	14.185%	35.111%	16.120%	3.781%

Figure 1.2: Proportion of Population in Income Groups

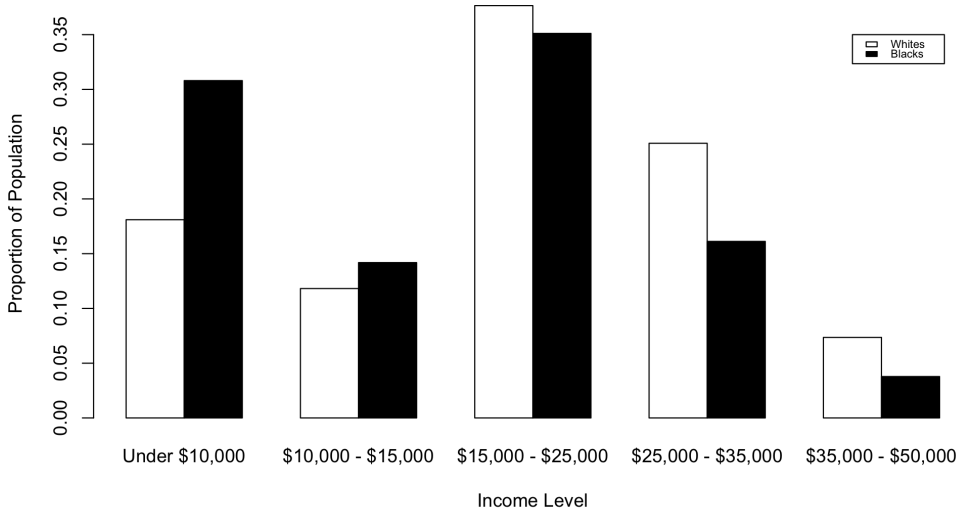
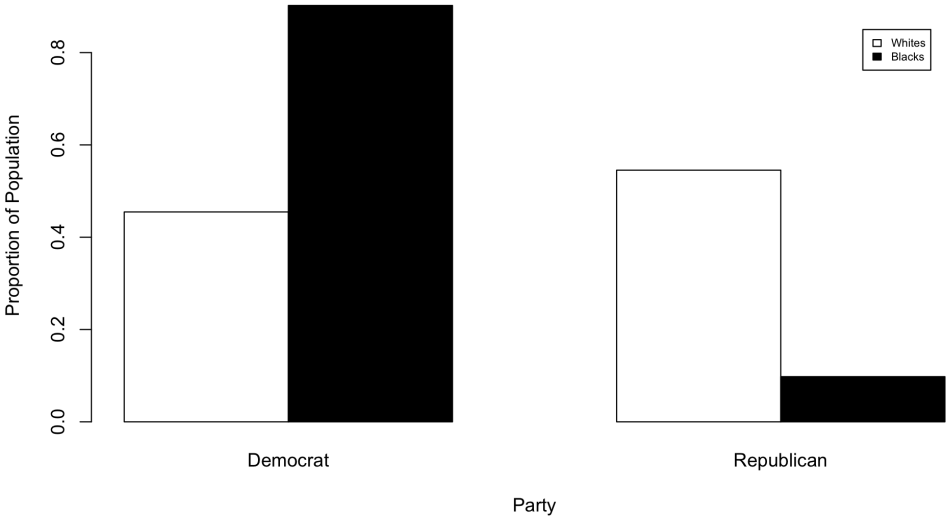


Figure 1.3: Proportion of Population by Party Voted



On another note, Gelman used three particular states as having weak, moderate, and strong income-voting relationships, which were Connecticut, Ohio, and Mississippi, respectively. Calculating the income distributions within these three states (Figure 1.4), I found that Connecticut was a relatively rich state, Ohio was relatively moderate, and Mississippi was a relatively poor state in 2004.

FIGURE 1.4: INCOME DISTRIBUTIONS IN MODEL STATES

Income Level	Under $10,000	$10,000 - $15,000	$15,000 - $25,000	$25,000 - $35,000	$35,000 - $50,000
Connecticut	14.621%	6.527%	30.026%	33.943%	14.883%
Ohio	19.896%	12.228%	40.570%	22.850%	4.456%
Mississippi	26.284%	14.199%	36.254%	18.127%	5.136%

Based on these three model states, I classified all of the other states based on a "chi-square test" approach. Taking the proportions of each model state's distribution as the expected value, and all of the other states' proportions as the observed value, I calculated three X^2 statistic for each state: one to compare that state to Connecticut, one to compare to Ohio, and one to compare to Mississippi.

$$X_{CT}{}^2 = \sum \frac{(O - E_{CT})^2}{E_{CT}}$$
$$X_{OH}{}^2 = \sum \frac{(O - E_{OH})^2}{E_{OH}}$$
$$X_{MS}{}^2 = \sum \frac{(O - E_{MS})^2}{E_{MS}}$$

Whichever X^2 statistic is smallest, the closer the state's distribution is to the model state. So for example, if the X^2 for Oregon is smallest for the comparison to Ohio, Oregon's income distribution is most similar to Ohio's and thus my algorithm would classify Oregon as a middle state.

FIGURE 1.5: STATE CLASSIFICATIONS BASED ON INCOME DISTRIBUTIONS OF CT, OH, AND MS

TYPE	STATES
Rich	NY, CA, VA, MA, MD, NJ, CT, NH
Middle	NC, WI, IL, MO, FL, WY, ME, PA, NV, OH, MI, GA, TX, MN, CO, IN, AZ, DE, LA, IA, KS, RI
Poor	ND, TN, OK, MS, SC, KY, AL, NM, WV, AR, ID, MT

After running this classifying algorithm, I get eight rich states, twenty-eight middle states, and twelve poor states (Figure 1.5). I will be using this classification as an indicator to test whether the individual income-voting relationships do depend on the income distribution of the state.

Before moving on to the analysis, I would like to point out that I am going to be treating state as a categorical variable, since the order in which the states are

numbered is arbitrary and carries no incremental value. Additionally, any pairs plots that I generated did not prove useful, as every pair of variables seemed relatively uncorrelated. Thus I will be proceeding without making any transformations to the data.

3 Analysis

Since I am trying to predict values of voting Democrat or Republican, I will be modeling the response variable as an indicator variable, which is 0 if the instance voted for Kerry (Democrat) and 1 if the instance voted for Bush (Republican). This way, I can predict the proportion of populations that vote Republican, which is helpful because the income-voting relationship Gelman derived is in terms of proportion voting Republican. Additionally, since I am predicting proportions, which fall into [0,1], I will be using logistic regression.

At first, I tried to make my generalized linear model using the Binomial family, but I soon realized that changing the family to Poisson to reflect the fact that I am working with counts would give me a better model. Sure enough, applying different families to my null model, which includes race, classification, state, and income, I get no change in degrees of freedom, yet my deviance in my Poisson model is less than half of that in my Binomial model.

Since I now know to use the Poisson as my family for logistic regression, I can start to build my model. In this experiment, I already know that the income-voting relationship exists, with its strength dependent upon how rich the state is. Thus, my null model in this case would predict the probability of voting Republican based on classification, income, and an interaction between these two explanatory variables. I find that adding state as another explanatory variable gives me a better model, so I decide to add it in to my null model.

FIGURE 1.6: ANALYSIS OF DEVIANCE TABLE

KEY: C = CLASSIFICATION, I = INCOME, S = STATE, R = RACE, * = INTERACTION

Model	Residual Degrees of Freedom	Residual Deviance	Change in Degrees of Freedom	Change in Deviance	P-Value
Null Model (C*I)	38,455	25,993			
C * I + S	38,410	25,739	45	254.24	0
R * C * I + S	38,395	23,994	15	1,744.41	0
R + C * I + S	38,409	24,020	-14	-25.98	0.026
R + S	38,421	24,204	-12	-183.51	0

Now, I have everything set up to determine whether the income-voting relationship is part of a race effect. After adding in race as an interaction to the classification-income interaction, adding in race as its own separate explanatory

variable, and taking out classification and income (Figure 1.6), I find that race is significantly better at predicting the proportion than classification and income. However, the best model is the model in which I add an interaction between race, classification, and income. Thus, my analysis tells me that while a large part of the income-voting relationship is a race effect, it is not entirely contained in the race effect.

Please note that since logistic regression does not have the same assumptions as ordinary least-squares regression, residual plots and qq-plots do not hold any particularly useful information.

4 Conclusion

Through logistic regression, I have seen that the income-voting relationship can be seen, for the most part, as a race effect. However, it is important to note that the race effect does not capture all of the information given by the income-voting relationship, and thus including this relationship would still benefit the logistic regression model. In other words, income and race both seem to affect the probability of voting Republican in a presidential election.

CASE II

- HOW DOES RESTAURANT MANAGEMENT COPE WITH MINIMUM WAGE INCREASES? -

1 Introduction

An increase in the minimum wage may seem like a good deed to society. After all, it's essentially making sure that the hard-working people in society earn more to support their families. But what are the underlying economic effects of raising this minimum wage bar? Some potential consequences include less employment or an increase in product prices to offset the extra money being paid per employee.

In this study, I will be analyzing these effects as commonly seen in the business world. More specifically, I will be monitoring changes in employment and changes in product prices. For my data set, I have a collection of 410 fast-food restaurants in two states. One of the states does not undergo an increase in minimum wage between the two interviews, while the other state does. Thus, for a sample of product prices, the data set records the price of the entrée, the price of fries, and the price of a soda.

2 Data Exploration

As mentioned above, I have a data set of 410 fast-food restaurants. Preliminarily, I decided to only consider the restaurants that gave a second response, since in both cases, my response variable is a measure of change between a period of time. Thus, I can leave out any observations in which a second interview response is not provided, which leaves me with 399 restaurants to consider. I can now re-organize my data by calculating the percent change of each continuous variable that had a before and after measurement, e.g. starting wage or hours open. I chose to use percent change in an attempt to normalize all of the data. I will also consider the following variables as categorical: special program indicator, cash bounty indicator, free/reduced price code (before and after), chain, co-owned indicator, and state. The only variable I will consider and do not need to manipulate is percent of staff affected by the wage change.

For the first part of my analysis, I will be exploring the percent change in employment. Again, I am using percent change because different restaurants have different standards and I want to normalize the data. If I have 10 fewer employees in a restaurant of originally 100 employees, it's different from the case of having 10 fewer employees in a restaurant of originally 15 employees. Additionally, I have decided to study change in total employment, which means that I have combined full-time, part-time, and managers into one statistic. This simplification compromises my ability to analyze more in depth the changes in the different types of employment, but given the unpredictability of people's

circumstances, that analysis might not have been as enlightening as I would like it to have been[3].

After converting all of the data to the appropriate formats (percent change, categorical, or keeping it the same), I decided to have a preliminary look at the variables in which I am interested, to decide if I should apply any transformations. A pairs plot (Figure 2.1A) reveals that my main explanatory variable of interest has a point of high leverage, which could potentially be influential. Thus, I decided to remove the point from my analysis to get a better feel for the general relationship between my variables. However, the pairs plot without the outlier (Figure 2.2), does not seem to reveal any meaningful relationship or transformation I should apply. Thus, I will move forward with my analysis without applying transformations to my variables.

Figure 2.1A: Pairs Plot of Variables of Interest - Before

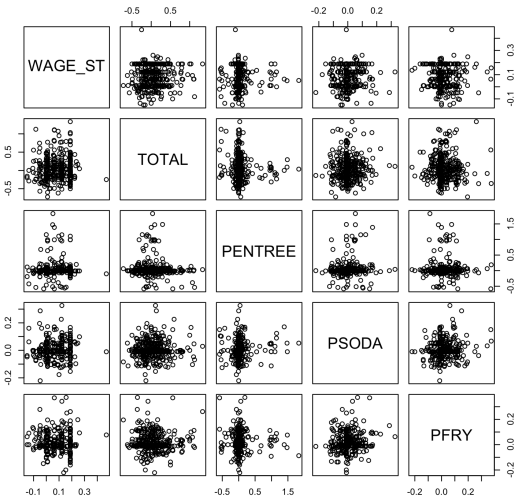
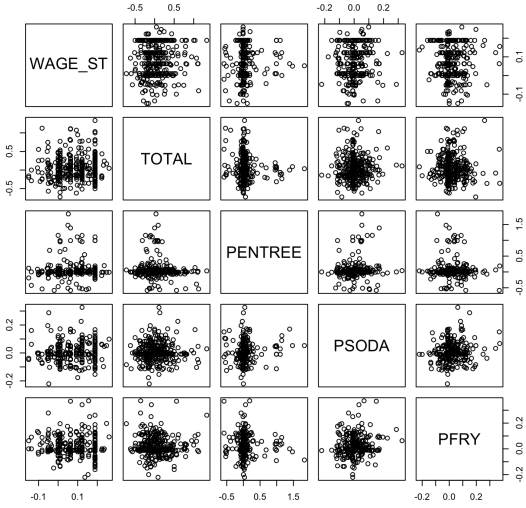


Figure 2.1B: Pairs Plot of Variables of Interest - After



3 Analysis – Employment

I start out running my full model with all of the variables, which turns out to be a horrible model.[4] Thus, I decide to run all-subsets regression[5] to find a smaller model that will keep as much as the predictive value possible. The model that

[3] I say this because there might be confounding variables. Perhaps some workers can no longer work full-time due to family circumstances. Perhaps others were not laid off, but quit due to personal reasons. It's just easier to look at employment trends as a whole.

[4] My full model achieves an adjusted-$R^2 < 0$. Adjusted-R^2 is a measure of how well a model fits; it penalizes the model based on the number of explanatory variables are used.

[5] To decide which subset of variables will make the best model, I use Mallow's C_p, which is another measure of goodness of fit for a model that penalizes a model based on how many variables the model uses.

came out of this algorithm only used “change in number of hours open”, and “change in usual amount of first raise.” Upon creating this new model, the added-variable (Figure 2.2A) and component-plus-residual plots (Figure 2.2B) seem to hint at a transformation. However, I first notice that one of my data points has a leverage point in “change in number of hours open,” so I decide to take out the leverage point to get a better view of my plots.

Figure 2.2A: Added Variable Plots

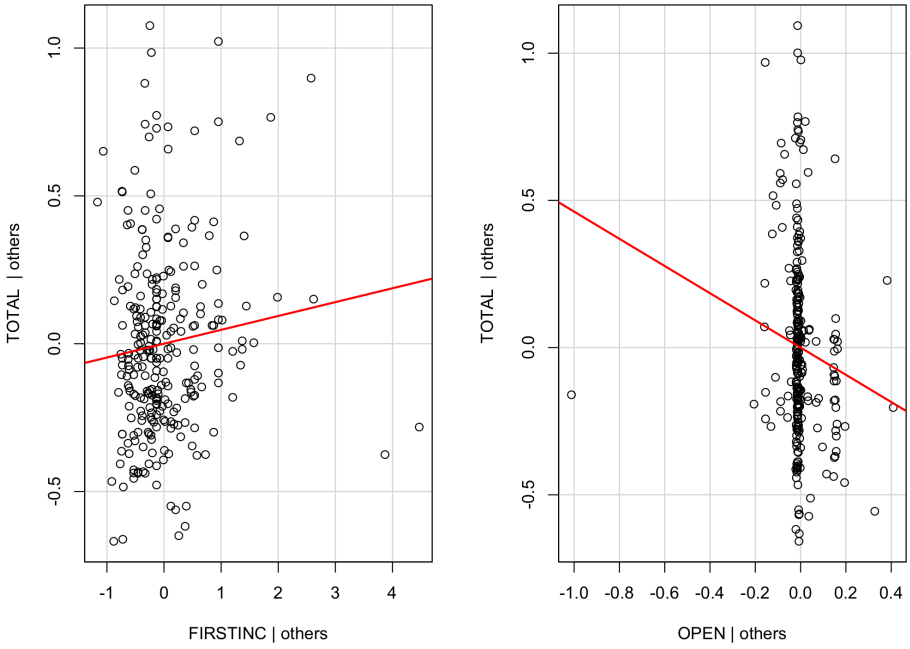
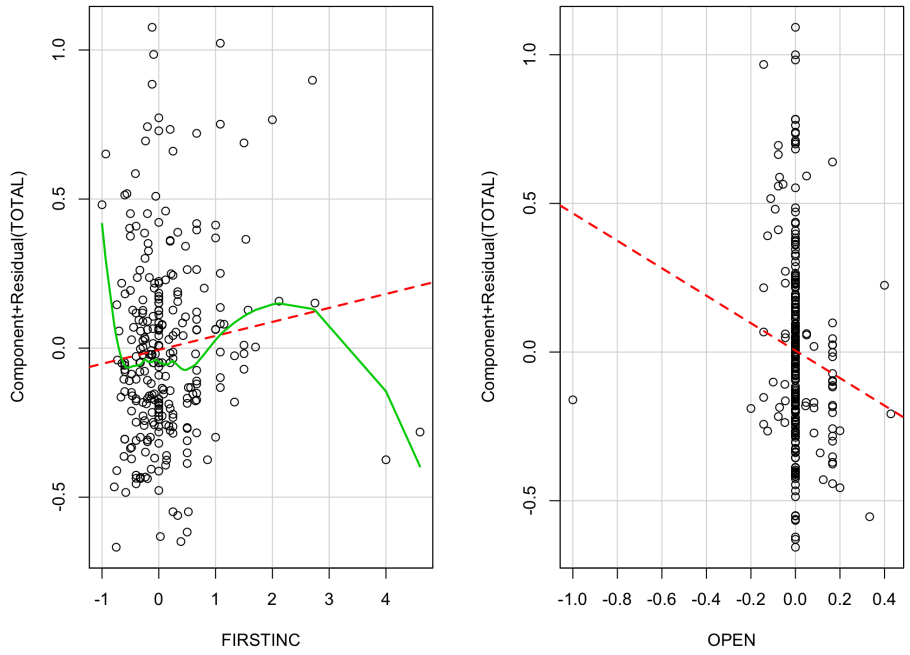


Figure 2.2B: Component + Residual Plots



After I remove my leverage point (Figures 2.3A and 2.3B), I still do not see a transformation I should use on "change in number of hours open," but I do see that I should try using a cubic function of "change in usual amount of first raise."

Figure 2.3A: Added Variable Plots Without Leverage Point

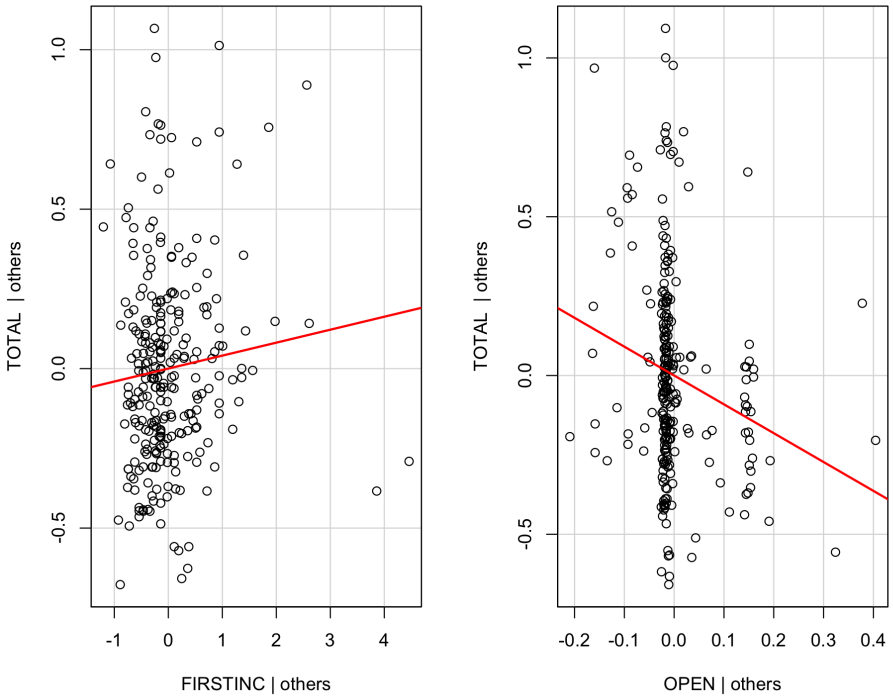
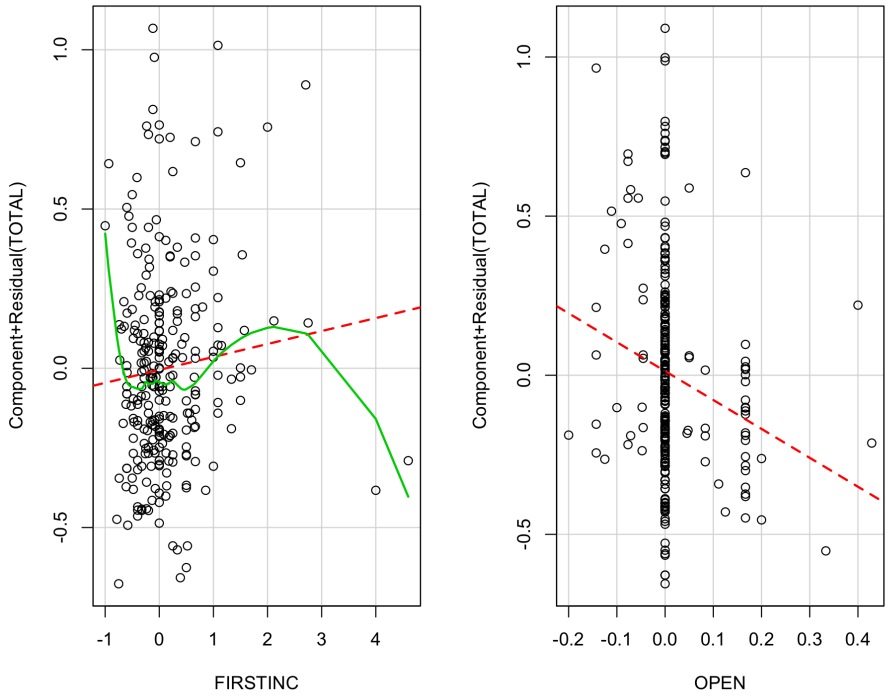


Figure 2.3B: Component + Residual Plots Without Leverage Point



Indeed, making the model a cubic function of "change in usual amount of first raise" improves my model; the added-variable and component-plus-residual plots all indicate (roughly) linearity, the fitted-versus-residual plot (Figure 2.4A) looks roughly homoscedastic, and the qq-plot (Figure 2.4B) suggests rough normality.

Figure 2.4A: Residuals vs Fitted

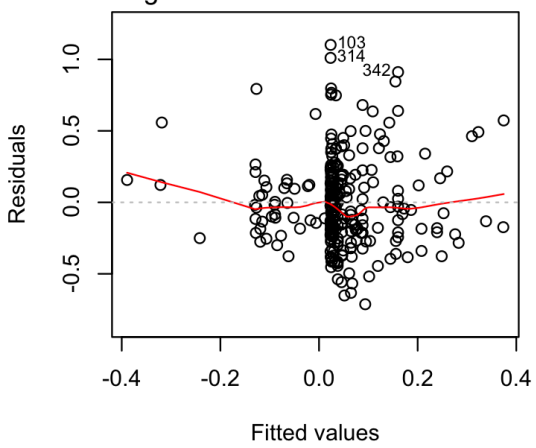
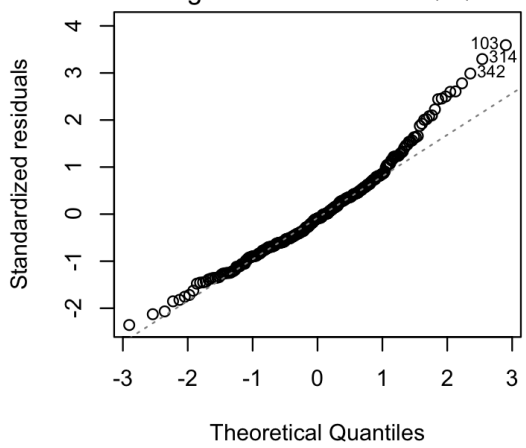


Figure 2.4B: Normal Q-Q



Now that I have a good model, I can add in "state" as an interaction variable to test if the raise in minimum wage truly made a difference. From analysis of variance (Figure 2.5), the model adding "state" is not significantly different from the model without "state." Thus, it does not seem that in the case of fast-food restaurants, the change in minimum wage affects overall employment.

FIGURE 2.5: ANALYSIS OF VARIANCE TABLE WITH AND WITHOUT "STATE"

Model	Residual Degrees of Freedom	Residual Sum of Squares (RSS)	Change in Degrees of Freedom	Change in RSS	P-value
Without "state"	265	25.053			
With "state"	260	24.625	5	0.42823	0.477015

4 Analysis – Prices

Since every observation has control over the prices of its entrée, soda, and fries, unlike employment situations, I was able to run analyses on each change in price separately, and together as an overall change.

As an overall view, none of the models for each change in price analysis required transformations, upon looking at the added-variable plots and component-plus-

residual plots. However, interestingly enough, each change in price analysis used different explanatory variables in the all-subsets regression best model.

FIGURE 2.6: OVERALL VIEW OF MODELS IN ANALYSIS OF PRICE DIFFERENCE				
Key: * = interaction				
Analysis (Change in Prices of ____)	**Explanatory Variables Used[6]**	**Adjusted-R^2**	**P-value**	**Does adding "state" make a significant model?**
Entrée	STATE * CHAIN * (OPEN + PCTAFF)	24.12%	$< 10^{-11}$	Yes
Fries	CHAIN + FIRSTINC + PCTAFF	4.207%	0.01049	No
Soda	STATE * CHAIN * (MEALS + NREGS + WAGE_ST + OPEN)	17.95%	$< 10^{-5}$	Yes
Overall	STATE * CHAIN * (NREGS + FIRSTINC + TOTAL + PCTAFF)	25.39%	$< 10^{-7}$	Yes

From my analysis, it looks as though the change in minimum wage did affect prices overall, but seemed to have little effect on the price of fries.

[6] Key: STATE = indicator of whether the restaurant is in the state where minimum wage was increased,
CHAIN = categorical variable indicating which chain the restaurant is a part of
OPEN = change in number of hours open
PCTAFF = percentage of staff affected by minimum wage change
FIRSTINC = change in usual amount of first raise
MEALS = categorical variable indicating free/reduced price
NREGS = number of registers in the store
WAGE_ST = change in starting wage
TOTAL = change in total employment

5 Conclusion

Through my analysis of the effects of changing the minimum wage on employment and prices in fast-food restaurants, I was not able to find a significant relationship between change in minimum wage and change in employment. However, I did find a significant relationship between increase in minimum wage and increase in food prices. Though the price of fries did not seem to change, entrees and soda prices went up, causing overall prices in fast-food restaurants to increase. Thus, it seems that in the fast-food industry, a change in minimum wage is offset by an increase in prices that customers must pay, and not by decreasing employment in the restaurants.

APPENDIX

```
############################
#######  Question 1  #######
############################
{
 ############################
 #Set-up and Reading in Data#
 ############################

 require(foreign)
 require(ggplot2)
 setwd("/Users/timwang/Desktop/Stat
 151A/Final Exam")
 annen2004 <-
 read.dta("annen2004_processed.dta"
 )
 summary(annen2004)
 dim(annen2004)

 ########################
 ##### Data-cleaning #####
 ########################

 cand <-
 !is.na(annen2004$votechoice) #
 only Bush and Kerry votes
 annen2004 <- annen2004[cand,]
 #summary(annen2004)
 cand <- !is.na(annen2004$income) #
 no NA income
 annen2004 <- annen2004[cand,]
 #summary(annen2004)
 cand <- annen2004$race == "white"
 | annen2004$race == "black" # only
 black and white race
 annen2004 <- annen2004[cand,]
 #summary(annen2004)
 names(annen2004)
 dim(annen2004)

 ############################
 #   EXPLORATORY ANALYSIS   #
 ############################
 # Distribution of income across
 race types #
 #  FIGURES 1.1 - 1.3 IN TEXT #
 ############################

 ########################
```

```
 ### FIGURE 1.1 IN TEXT ###
 ########################

 dataWhite <-
 annen2004[annen2004$race ==
 "white",]
 dim(dataWhite)
 whiteCount <- sapply(1:5,
 function(x){sum(as.numeric(dataWhi
 te$income) == x)})
 whiteCount2 <- sapply(0:1,
 function(x){sum(as.numeric(dataWhi
 te$votechoice) == x)})
 whiteProp <- whiteCount /
 sum(whiteCount)
 whiteProp2 <- whiteCount2 /
 sum(whiteCount)
 whiteProp # income distribution
 for whites
 # [1] 0.18099573 0.11815078
 0.37650071 0.25081081 0.07354196
 whiteProp2 # voting distribution
 for whites
 # [1] 0.4547368 0.5452632

 dataBlack <-
 annen2004[annen2004$race ==
 "black",]
 dim(dataBlack)
 blackCount <- sapply(1:5,
 function(x){sum(as.numeric(dataBla
 ck$income) == x)})
 blackCount2 <- sapply(0:1,
 function(x){sum(as.numeric(dataBla
 ck$votechoice) == x)})
 blackProp <- blackCount /
 sum(blackCount)
 blackProp2 <- blackCount2 /
 sum(blackCount)
 blackProp # income distribution
 for blacks
 # [1] 0.30803048 0.14185229
 0.35111372 0.16119578 0.03780774
 blackProp2 # voting distribution
 for whites
 # [1] 0.9021102 0.0978898

 par(mfcol = c(1,2))
```

```
#########################
### FIGURE 1.2 IN TEXT ###
#########################

tograph <-
data.frame(t(as.matrix(data.frame(
whiteProp, blackProp))))
names(tograph) <- c("Under
$10,000", "$10,000 - $15,000",
"$15,000 - $25,000", "$25,000 -
$35,000", "$35,000 - $50,000")

barplot(as.matrix(tograph),
main="Figure 1.2: Proportion of
Population in Income Groups",
ylab="Proportion of Population",
beside=TRUE, col=c("white",
"black"), xlab = "Income Level")
legend(13, 0.35, c("Whites",
"Blacks"), cex=0.6,
fill=c("white", "black"))

#########################
### FIGURE 1.3 IN TEXT ###
#########################

tograph2 <-
data.frame(t(as.matrix(data.frame(
whiteProp2, blackProp2))))
names(tograph2) <- c("Democrat",
"Republican")

barplot(as.matrix(tograph2),
main="Figure 1.3: Proportion of
Population by Party Voted",
ylab="Proportion of Population",
beside=TRUE, col=c("white",
"black"), xlab = "Party")
legend(5, 0.85, c("Whites",
"Blacks"), cex=0.6,
fill=c("white", "black"))

## Pairs Plot
{
    inds <- match(c("race",
"income", "state", "votechoice"),
names(annen2004))
    # pairs(annen2004[,inds]) #
Not useful, try to de-factorize
income
}

## De-factorize income
{
    base <- c(0, 10000, 15000,
25000, 35000)
    ind <-
as.numeric(annen2004$income)
    baseInc <- base[ind]
    inc <- c(10000, 5000, 10000,
10000, 15000)
    incInd <- inc[ind]
    randInc <-
runif(length(incInd), 0, incInd)
    randAmt <- baseInc + randInc
    annen2004$income2 <- randAmt
    summary(annen2004)

    ## Median incomes
    med <- c(5000, 12500, 20000,
30000, 42500)
    annen2004$medIncome <-
med[ind]
}

## Pairs Plot, Round 2!
{
    inds <- match(c("race",
"income2", "state", "votechoice"),
names(annen2004))
    #pairs(annen2004[,inds]) #
Not useful, try log income
    annen2004$logIncome <-
log(annen2004$income2)
    inds <- match(c("race",
"logIncome", "state",
"votechoice"), names(annen2004))
    #pairs(annen2004[,inds]) #
Not useful, try to de-factorize
income
}
## Pairs Plot is not useful...

## Explore the 3 states: CT - 7,
OH - 35, MS - 24
{

#########################
##   FIGURE 1.4 IN TEXT ##
#########################

    ## Get distribution of income
```

```
levels in 3 model states
    nonNA <-
annen2004[!is.na(annen2004$state),
]
    levs <- 1:5
    ctData <- nonNA[nonNA$state
== 7,]
    ctCount <- sapply(levs,
function(x){sum(as.numeric(ctData$
income) == x)})
    ctProp <- ctCount /
sum(ctCount)
    ohData <- nonNA[nonNA$state
== 35,]
    ohCount <- sapply(levs,
function(x){sum(as.numeric(ohData$
income) == x)})
    ohProp <- ohCount /
sum(ohCount)
    msData <- nonNA[nonNA$state
== 24,]
    msCount <- sapply(levs,
function(x){sum(as.numeric(msData$
income) == x)})
    msProp <- msCount /
sum(msCount)
    ctProp
    # [1] 0.14621410 0.06527415
0.30026110 0.33942559 0.14882507
    ohProp
    # [1] 0.19896373 0.12227979
0.40569948 0.22849741 0.04455959
    msProp
    # [1] 0.26283988 0.14199396
0.36253776 0.18126888 0.05135952

    ## Classify all states

    summary(factor(annen2004$stat
e)) ## State numbers 2 and 11 are
missing.
    stateInd <- 1:50
    stateInd <- stateInd[c(-2, -
11)]
    classification <- rep(0,
times = 50) # 1 = rich, 2 =
middle, 3 = poor
    for(i in stateInd) ## Use a
chi-square-type approach to
classify.
    {
        stateData <-
nonNA[nonNA$state == i,]
        stateCount <-
sapply(1:5,
function(x){sum(as.numeric(stateDa
ta$income) == x)})
        stateProp <- stateCount
/ sum(stateCount)
        x1 <- sum((stateProp -
ctProp)^2 / ctProp) # (O-E)^2 / E
        x2 <- sum((stateProp -
ohProp)^2 / ohProp)
        x3 <- sum((stateProp -
msProp)^2 / msProp)
        finzzz <- c(x1, x2, x3)
# Take the smallest X^2 statistic.
        classification[i] =
match(min(finzzz), finzzz)
    }
    annen2004$classification <-
factor(classification[annen2004$st
ate])
    inds <- match(c("race",
"classification", "state",
"votechoice"), names(annen2004))
    pairs(annen2004[,inds]) #
Still not very useful.

##########################
##   FIGURE 1.5 IN TEXT ##
##########################


    unique(annen2004[!is.na(annen
2004$classification) &
annen2004$classification ==
1,]$cst) # Rich states
    # [1] "NY" "CA" "VA" "MA"
"MD" "NJ" "CT" "NH"


    unique(annen2004[!is.na(annen
2004$classification) &
annen2004$classification ==
2,]$cst) # Middle states
    #  [1] "NC" "WI" "IL" "MO"
"FL" "WY" "ME" "PA" "NV" "OH" "MI"
"GA" "TX" "MN" "CO" "IN" "AZ" "DE"
"LA" "IA" "KS" "RI"
    # [23] "OR" "UT" "WA" "NE"
"VT" "SD"
```

```
    unique(annen2004[!is.na(annen
2004$classification) &
annen2004$classification ==
3,]$cst) # Poor states
    #  [1] "ND" "TN" "OK" "MS"
"SC" "KY" "AL" "NM" "WV" "AR" "ID"
"MT"


    length(unique(annen2004[!is.n
a(annen2004$classification) &
annen2004$classification ==
1,]$cst)) # Rich states

    length(unique(annen2004[!is.n
a(annen2004$classification) &
annen2004$classification ==
2,]$cst)) # Middle states

    length(unique(annen2004[!is.n
a(annen2004$classification) &
annen2004$classification ==
3,]$cst)) # Poor states
}

####################
##### ANALYSIS #####
####################

## GLM's and ANOVA
{
    require(car)
    annen2004$state <-
factor(annen2004$state)
    voteNull <-
glm(votechoice~race+classification
+state+income, family =
"binomial", data = annen2004)
    voteAlt2 <-
glm(votechoice~race +
classification + state + income,
family = "poisson", data =
annen2004)
    anova(voteNull, voteAlt2,
test = "LRT") # Binomial VS
Poisson


    voteAltNoRaceState <-
glm(votechoice~classification*inco
me, family = "poisson", data =
annen2004)
    voteAltNoRace <-
glm(votechoice~classification*inco
me + state, family = "poisson",
data = annen2004)
    voteAltRace <-
glm(votechoice~race +
classification*income + state,
family = "poisson", data =
annen2004)
    curious <-
glm(votechoice~race + state,
family = "poisson", data =
annen2004)
    voteAltRaceInt <-
glm(votechoice~race*classification
*income + state, family =
"poisson", data = annen2004)

#########################
##   FIGURE 1.6 IN TEXT ##
#########################

    anova(voteAltNoRaceState,
voteAltNoRace, voteAltRaceInt,
voteAltRace, curious, test =
"LRT") # Comparisons: Adding
state, adding race, adding race
interaction, removing income-
classification interaction


}

}

################################
#########  Question 2  #########
################################

##################
###   Set-up   ###
##################

require(foreign)
require(ggplot2)
setwd("/Users/timwang/Desktop/Stat
 151A/Final Exam")
fastfood <-
```

```
 read.csv("fastfood.csv",
 na.strings = c('.', 'NA'), header
 = T)
fastfood$TotalEMP1 <-
 fastfood$EMPFT + fastfood$EMPPT +
 fastfood$NMGRS
fastfood$TotalEMP2 <-
 fastfood$EMPFT2 + fastfood$EMPPT2
 + fastfood$NMGRS2
fastfood <-
 fastfood[fastfood$STATUS2 == 1,] #
 We only want restaurants that
 replied for the second time.

head(fastfood)
summary(fastfood)
names(fastfood)

################################
#  Factorize categorical variables
 #
################################

dfastfood = data.frame(1:399)
dfastfood$SPECIAL2 <-
 factor(fastfood$SPECIAL2)
dfastfood$BONUS <-
 factor(fastfood$BONUS)
dfastfood$MEALS <-
 factor(fastfood$MEALS)
dfastfood$MEALS2 <-
 factor(fastfood$MEALS2)
dfastfood$CHAIN <-
 factor(fastfood$CHAIN)
dfastfood$CO_OWNED <-
 factor(fastfood$CO_OWNED)
dfastfood$STATE <-
 factor(fastfood$STATE)

############################
#  Compute all % changes  #
############################

dfastfood$PFRY <- fastfood$PFRY2 /
 fastfood$PFRY - 1
dfastfood$PSODA <- fastfood$PSODA2
 / fastfood$PSODA - 1
dfastfood$PENTREE <-
 fastfood$PENTREE2 /
 fastfood$PENTREE - 1
dfastfood$HRSOPEN <-
 fastfood$HRSOPEN2 /
 fastfood$HRSOPEN - 1
dfastfood$NREGS <- fastfood$NREGS2
 / fastfood$NREGS - 1
dfastfood$NREGS11 <-
 fastfood$NREGS112 /
 fastfood$NREGS11 - 1
dfastfood$WAGE_ST <-
 fastfood$WAGE_ST2 /
 fastfood$WAGE_ST - 1
dfastfood$FIRSTINC <-
 fastfood$FIRSTIN2 /
 fastfood$FIRSTINC - 1
# dfastfood$EMPFT <-
 fastfood$EMPFT2 / fastfood$EMPFT -
 1
# dfastfood$EMPPT <-
 fastfood$EMPPT2 / fastfood$EMPPT -
 1
# dfastfood$NMGRS <-
 fastfood$NMGRS2 / fastfood$NMGRS -
 1
# dfastfood$WRKRS <-
 (fastfood$EMPFT2 +
 fastfood$EMPPT2) / (fastfood$EMPFT
 + fastfood$EMPPT) - 1
dfastfood$TOTAL <-
 fastfood$TotalEMP2 /
 fastfood$TotalEMP1 - 1
dfastfood$INCTIME <-
 fastfood$INCTIME2 /
 fastfood$INCTIME - 1
fastfood$OPEN[fastfood$OPEN == 0]
 <- NA
dfastfood$OPEN <- fastfood$OPEN2R
 / fastfood$OPEN - 1

################################
# Keep "% of staff affected" #
################################

dfastfood$PCTAFF <-
 fastfood$PCTAFF
# dfastfood$INCTIME <-
 fastfood$INCTIME
# dfastfood$INCTIME2 <-
 fastfood$INCTIME2
# dfastfood$OPEN <- fastfood$OPEN
# dfastfood$OPEN2R <-
 fastfood$OPEN2R
dfastfood <- dfastfood[, -1]
```

```
############################
# Remove NA's from workers #
############################

remNA = match(c("TOTAL"),
 names(dfastfood))
# remNA = match(c("WRKRS"),
 names(dfastfood))
# remNA = match(c("TOTAL",
 "WAGE_ST"), names(dfastfood))
for(i in remNA)
{
 dfastfood <-
 dfastfood[!is.na(dfastfood[, i]),
 ]
}

# for(j in
 1:length(names(dfastfood)))
# {
 # dfastfood <-
 dfastfood[is.finite(dfastfood[,
 j]),]
# }

dim(dfastfood)
# [1] 378  19

################################
# Explore variables of interest#
#      FIGURE 2.1 IN TEXT      #
################################

indi <- match(c("WAGE_ST",
 "TOTAL", "PENTREE", "PSODA",
 "PFRY"), names(dfastfood))
par(mfcol = c(1, 2))
pairs(dfastfood[, indi], main =
 "Figure 2.1A: Pairs Plot of
 Variables of Interest - Before")
dim(dfastfood[dfastfood$WAGE_ST <
 0.4,])
dfastfood <-
 dfastfood[dfastfood$WAGE_ST <
 0.4,]
pairs(dfastfood[, indi], main =
 "Figure 2.1B: Pairs Plot of
 Variables of Interest - After")

#########################
```

```
###      ANALYSIS      ###
#########################

#############
# Full Model #
#############

require(car)
pairs(dfastfood)
mod <- lm(TOTAL ~ ., data =
 dfastfood)
# mod <- lm(WRKRS ~ ., data =
 dfastfood)
summary(mod)
avPlots(mod)
crPlots(mod)

##################################
 ########################
# All-subsets to find best model
 without transformations #
#      FIGURE 2.2 IN TEXT      #
##################################
 ########################

require(leaps)
regs <- regsubsets(TOTAL ~ ., data
 = dfastfood, nvmax =
 dim(dfastfood)[2])
summary(regs)
bestModel <-
 match(min(summary(regs)$cp),
 summary(regs)$cp)
coefs <-
 which(summary(regs)$which[bestMode
 l,])[-1]
coefs
# FIRSTINC     OPEN
#       22       24
mod2 <- lm(TOTAL~FIRSTINC + OPEN,
 data = dfastfood)
summary(mod2)
par(mfrow = c(2, 2))
avPlots(mod2, main = "Figure 2.2A:
 Added Variable Plots")
crPlots(mod2, main = "Figure 2.2B:
 Component + Residual Plots") #
 There is an influential point for
 OPEN; remove to get better
 picture.
```

```
############################
# Remove outliers in OPEN #
#    FIGURE 2.3 IN TEXT    #
############################

dim(dfastfood) # 378
dfastfood <-
 dfastfood[dfastfood$OPEN > -0.8, ]
dim(dfastfood) # 376

mod2woout <- lm(TOTAL~FIRSTINC +
 OPEN, data = dfastfood)
summary(mod2woout)
avPlots(mod2woout, main = "Figure
 2.3A: Added Variable Plots Without
 Leverage Point")
crPlots(mod2woout, main = "Figure
 2.3B: Component + Residual Plots
 Without Leverage Point") # No
 outlier does not particularly
 reveal a shape anyways.

################################
# Use results from AV and CR
 Plots. #
#      FIGURE 2.4 IN TEXT      #
################################

mod3 <- lm(TOTAL~I(FIRSTINC^3) +
 I(FIRSTINC^2) + FIRSTINC + OPEN,
 data = dfastfood)
summary(mod3)
avPlots(mod3)
crPlots(mod3)
anova(mod2woout, mod3, test =
 "LRT")
par(mfrow=c(2,2))
plot(mod3, caption = list("Figure
 2.4A: Residuals vs Fitted",
 "Figure 2.4B: Normal Q-Q", "Scale-
 Location", "Cook's distance",
 "Residuals vs Leverage",
 expression("Cook's dist vs
 Leverage  " * h[ii] / (1 -
 h[ii]))))

################################
# Add STATE as interaction to
 indicate change in MW. #
#        See if this addition is
 significant.        #
################################

mod4 <-
 lm(TOTAL~STATE*(I(FIRSTINC^3) +
 I(FIRSTINC^2) + FIRSTINC + OPEN),
 data = dfastfood)
summary(mod4)
avPlots(mod4)
crPlots(mod4)
anova(mod2woout, mod3, mod4, test
 = "LRT")

################################
# Analyze change in prices - SETUP
 #
#      FIGURE 2.6 IN TEXT      #
################################

# par(mfcol = c(2, 2))
# hist(dfastfood$PENTREE)
# hist(dfastfood$PFRY)
# hist(dfastfood$PSODA)
# dfastfood$PRICE <-
 dfastfood$PENTREE + dfastfood$PFRY
 + dfastfood$PSODA
# hist(dfastfood$PRICE)
# inds <- match(c("PENTREE",
 "PFRY", "PSODA"),
 names(dfastfood))
dataPENTREE <- dfastfood[, -
 match(c("PFRY", "PSODA"),
 names(dfastfood))]
dataPFRY <- dfastfood[, -
 match(c("PENTREE", "PSODA"),
 names(dfastfood))]
dataPSODA <- dfastfood[, -
 match(c("PENTREE", "PFRY"),
 names(dfastfood))]
# for(i in inds)
# {
#    dataP <-
 dataP[!is.na(dfastfood[, i]), ]
# }
# dim(dataP)
# # [1] 376  19
modPENTREE <- lm(PENTREE ~ ., data
 = dataPENTREE)
modPFRY <- lm(PFRY ~ ., data =
 dataPFRY)
modPSODA <- lm(PSODA ~ ., data =
```

```
 dataPSODA)
summary(modPENTREE)
avPlots(modPENTREE)
crPlots(modPENTREE)
summary(modPFRY)
avPlots(modPFRY)
crPlots(modPFRY)
summary(modPSODA)
avPlots(modPSODA)
crPlots(modPSODA)

################################
# Analyze change in entree prices.
 #
################################

regsPENTREE <- regsubsets(PENTREE
 ~ ., data = dataPENTREE, nvmax =
 dim(dataPENTREE)[2])
summary(regsPENTREE)
bestModelPENTREE <-
 match(min(summary(regsPENTREE)$cp)
 , summary(regsPENTREE)$cp)
coefsPENTREE <-
 which(summary(regsPENTREE)$which[b
 estModelPENTREE,])[-1]
coefsPENTREE
#   CHAIN4 STATE1   OPEN PCTAFF
#       12     14     22     23
modP2E <- lm(PENTREE ~ CHAIN +
 STATE + OPEN + PCTAFF, data =
 dataPENTREE)
summary(modP2E)
avPlots(modP2E)
crPlots(modP2E)
plot(modP2E)

modP3E <- lm(PENTREE ~ STATE *
 (CHAIN + OPEN + PCTAFF), data =
 dataPENTREE)
summary(modP3E)
avPlots(modP3E)
crPlots(modP3E)
plot(modP3E)
anova(modP2E, modP3E, test =
 "LRT")

modP4E <- lm(PENTREE ~ STATE *
 CHAIN * (OPEN + PCTAFF), data =
 dataPENTREE)
summary(modP4E)
avPlots(modP4E)
crPlots(modP4E)
anova(modP2E, modP3E, modP4E, test
 = "LRT")
par(mfcol = c(2,2))
plot(modP4, caption = list("Figure
 2: Residuals vs Fitted", "Normal
 Q-Q", "Scale-Location", "Cook's
 distance", "Residuals vs
 Leverage", expression("Cook's dist
 vs Leverage  " * h[ii] / (1 -
 h[ii]))))

################################
# Analyze change in fries prices.
 #
################################

regsPFRY <- regsubsets(PFRY ~ .,
 data = dataPFRY, nvmax =
 dim(dataPFRY)[2])
summary(regsPFRY)
bestModelPFRY <-
 match(min(summary(regsPFRY)$cp),
 summary(regsPFRY)$cp)
coefsPFRY <-
 which(summary(regsPFRY)$which[best
 ModelPFRY,])[-1]
coefsPFRY
#     CHAIN3 FIRSTINC   PCTAFF
#         11       19       23
modP2F <- lm(PFRY ~ CHAIN +
 FIRSTINC + PCTAFF, data =
 dataPFRY)
summary(modP2F)
avPlots(modP2F)
crPlots(modP2F)
plot(modP2F)

modP3F <- lm(PFRY ~ CHAIN *
 (FIRSTINC + PCTAFF), data =
 dataPFRY)
summary(modP3F)
avPlots(modP3F)
crPlots(modP3F)
plot(modP3F)
anova(modP2F, modP3F, test =
 "LRT")

modP4F <- lm(PFRY ~ STATE * CHAIN
 * (FIRSTINC + PCTAFF), data =
```

```
 dataPFRY)
summary(modP4F)
avPlots(modP4F)
crPlots(modP4F)
plot(modP4F)
anova(modP2F, modP3F, modP4F, test
 = "LRT")
par(mfcol = c(2,2))
plot(modP4, caption = list("Figure
 2: Residuals vs Fitted", "Normal
 Q-Q", "Scale-Location", "Cook's
 distance", "Residuals vs
 Leverage", expression("Cook's dist
 vs Leverage  " * h[ii] / (1 -
 h[ii]))))

################################
# Analyze change in soda prices. #
################################

regsPSODA <- regsubsets(PSODA ~ .,
 data = dataPSODA, nvmax =
 dim(dataPSODA)[2])
summary(regsPSODA)
bestModelPSODA <-
 match(min(summary(regsPSODA)$cp),
 summary(regsPSODA)$cp)
coefsPSODA <-
 which(summary(regsPSODA)$which[bes
 tModelPSODA,])[-1]
coefsPSODA
#    MEALS1  CHAIN4  STATE1
 NREGS WAGE_ST    OPEN
#         4      12      14
 16      18      22
modP2S <- lm(PSODA ~ MEALS + CHAIN
 + STATE + NREGS + WAGE_ST + OPEN,
 data = dataPSODA)
summary(modP2S)
avPlots(modP2S)
crPlots(modP2S)
plot(modP2S)

modP3S <- lm(PSODA ~ STATE *
 (MEALS + CHAIN + NREGS + WAGE_ST +
 OPEN), data = dataPSODA)
summary(modP3S)
avPlots(modP3S)
crPlots(modP3S)
plot(modP3S)
anova(modP2S, modP3S, test =
 "LRT")

modP4S <- lm(PSODA ~ STATE * CHAIN
 * (MEALS + NREGS + WAGE_ST +
 OPEN), data = dataPSODA)
summary(modP4S)
avPlots(modP4S)
crPlots(modP4S)
plot(modP4S)
anova(modP2S, modP3S, modP4S, test
 = "LRT")
par(mfcol = c(2,2))
plot(modP4, caption = list("Figure
 2: Residuals vs Fitted", "Normal
 Q-Q", "Scale-Location", "Cook's
 distance", "Residuals vs
 Leverage", expression("Cook's dist
 vs Leverage  " * h[ii] / (1 -
 h[ii]))))

################################
 ###
# Analyze change in overall
 prices. #
################################
 ###

par(mfcol = c(2, 2))
hist(dfastfood$PENTREE)
hist(dfastfood$PFRY)
hist(dfastfood$PSODA)
dfastfood$PRICE <-
 dfastfood$PENTREE + dfastfood$PFRY
 + dfastfood$PSODA
hist(dfastfood$PRICE)
inds <- match(c("PENTREE", "PFRY",
 "PSODA"), names(dfastfood))
dataP <- dfastfood[, -inds]
dataP <-
 dataP[!is.na(dataP$PRICE), ]
dim(dataP)
# [1] 313  17
modP <- lm(PRICE ~ ., data =
 dataP)
summary(modP)
avPlots(modP)
crPlots(modP)

regsP <- regsubsets(PRICE ~ .,
 data = dataP, nvmax =
 dim(dataP)[2])
```

```
summary(regsP)
bestModelP <-
 match(min(summary(regsP)$cp),
 summary(regsP)$cp)
coefsP <-
 which(summary(regsP)$which[bestMod
 elP,])[-1]
coefsP
#  CHAIN4   STATE1    NREGS
 FIRSTINC    TOTAL   PCTAFF
#      12       14       16
 19       20       23
modP2 <- lm(PRICE ~ CHAIN + STATE
 + NREGS + FIRSTINC + TOTAL +
 PCTAFF, data = dataP)
summary(modP2)
avPlots(modP2)
crPlots(modP2)
plot(modP2)

modP3 <- lm(PRICE ~ CHAIN * (STATE
 + NREGS + FIRSTINC + TOTAL +
 PCTAFF), data = dataP)
summary(modP3)
avPlots(modP3)
crPlots(modP3)
plot(modP3)
anova(modP2, modP3, test = "LRT")

modP4 <- lm(PRICE ~ STATE * CHAIN
 * (NREGS + FIRSTINC + TOTAL +
 PCTAFF), data = dataP)
summary(modP4)
avPlots(modP4)
crPlots(modP4)
anova(modP2, modP3, modP4, test =
 "LRT")
par(mfcol = c(2,2))
plot(modP4, caption = list("Figure
 2: Residuals vs Fitted", "Normal
 Q-Q", "Scale-Location", "Cook's
 distance", "Residuals vs
 Leverage", expression("Cook's dist
 vs Leverage  " * h[ii] / (1 -
 h[ii]))))
```