Shivi Agarwal & Lisa Zhang
STA 371H (MW 11-12:30)
Professor Scott

Class 2/17/14 Notes

Interaction - The whole is greater than the sum of its parts. One variable modulates another.
Bicycle example: bike is harder to pedal if it is in high gear than low gear
bike is harder to pedal if biker is traveling uphill than downhill
If bike is traveling up a steep hill in high gear, it will be very difficult to pedal. But is it a greater overall total (more difficult to pedal than the individual constraints)?
If yes, then interaction term applies.

When one variable modulates another variable.

- Midterm before Spring Break (Wednesday, March 5th)
- 4 class days with new information + 1 review day before Midterm (on March 4th)
- New topics: Interaction term, Intro to Multiple Regression, & Hypothesis Testing
- Midterm will cover conceptual core understanding → you, pen, paper, thoughts
- Homework scripts for 04 and 05 are on the website

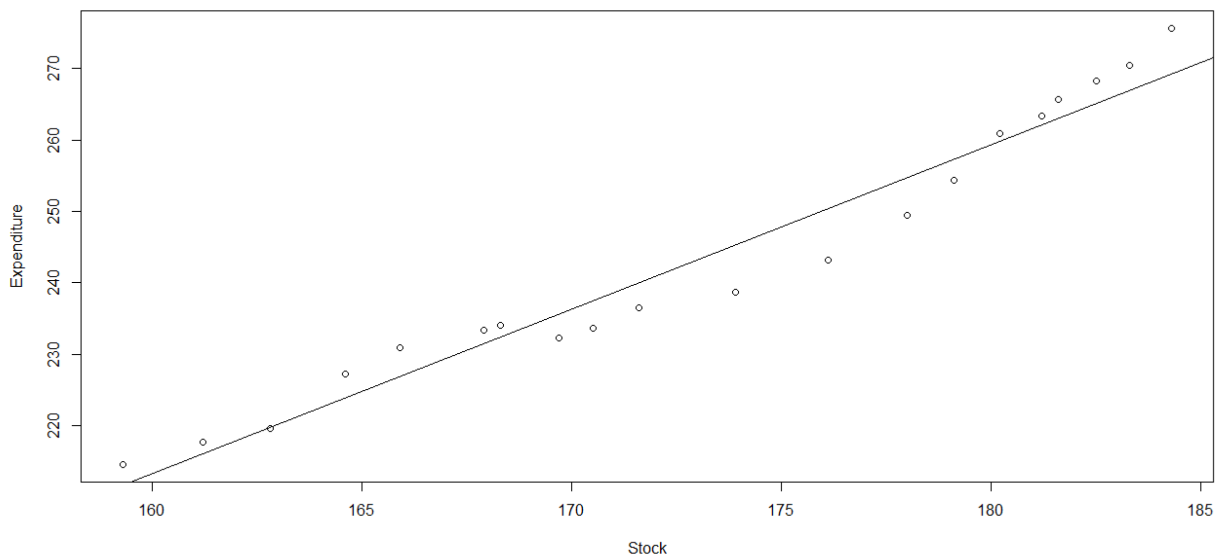
Homework Exercise 5 Review:
Problem #1: Quantify uncertainty
Bootstrapping & Regression nudges
Read course packet.

Problem #2: How much consumer spending is influenced by stock of money?
Federal Reserve meetings always discuss the money multiplier. Money that sits in the bank has a money multiplier of 0. Money paid to Professor Scott, who then pays the diner for a meal, who then pays the salary of worker, has a money multiplier of 2.
If stock of money increases, then consumer spending increases.

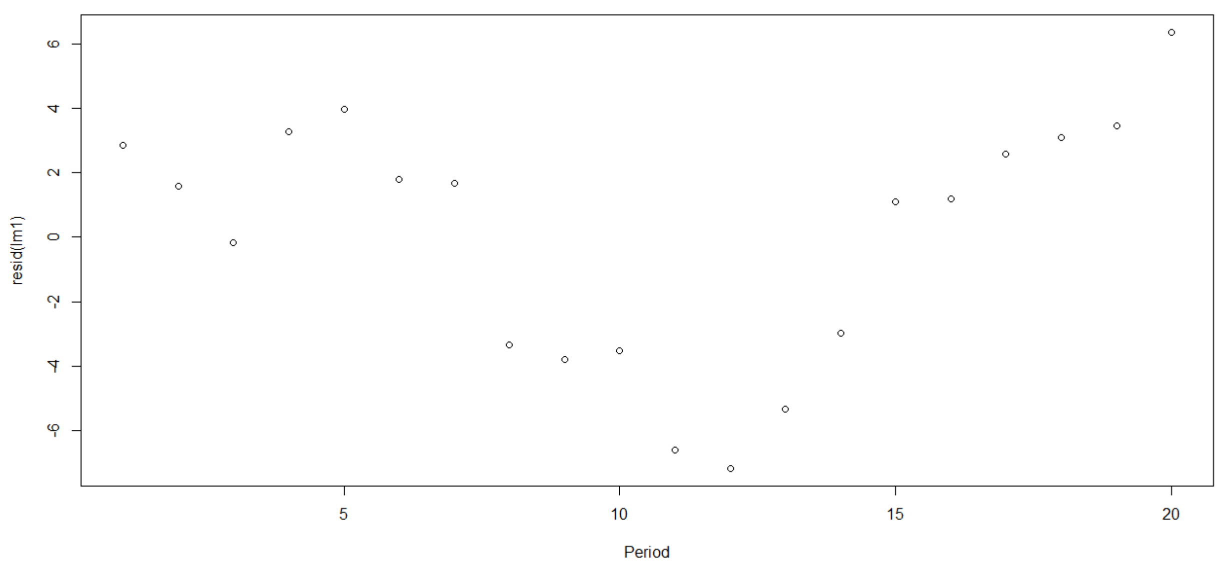
Load consumerexp.csv
line 14 → Add explicit time index to account for time
line 15→ new plot
Fit line adjusted for stock of money over time and look at residuals over time. Relatively consistent.

```
>plot(Expenditure ~ Stock, data=consumerexp)
```

```
>lm1 = lm(Expenditure ~ Stock, data=consumerexp)
>abline(lm1)
```




```
>plot(resid(lm1) ~ Period, data=consumerexp)
```



```
> summary(lm1)

Call:
lm(formula = Expenditure ~ Stock, data = consumerexp)

Residuals:
   Min     1Q Median     3Q    Max 
-7.176 -3.396  1.396  2.928  6.361 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -154.7192    19.8500  -7.794 3.54e-07 ***
Stock          2.3004     0.1146  20.080 8.99e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.983 on 18 degrees of freedom
Multiple R-squared:  0.9573, Adjusted R-squared:  0.9549 
F-statistic: 403.2 on 1 and 18 DF,  p-value: 8.988e-14

> 
> confint(lm1)
                 2.5 %      97.5 %
(Intercept) -196.422532 -113.015792
Stock          2.059693    2.541049
> 
```

Summary of model gives us the values of 2.3004 as the multiplier and 0.1196 as the Standard Error.
We retrieve 95% Confidence Interval as 2.06, 2.54.
Are we happy with the confidence interval (not with the precision but the certainty of it)?
It does not look correct. Confidence Interval for slope follows assumptions of a normal linear regression model. Thus, we must check to see if the assumptions are met.

<u>Normal Linear Model Assumptions</u>

1. Normal Distribution
2. Constant Variance
3. Independent of each other

Looking back at the residual plot, we find that the residuals appear to be correlated over time. Are the assumptions valid then? Residual plot should ideally be a random cloud of points to convey no correlated information. However adjacent residuals, like the residuals found in our plot , convey information. This should not occur and thus implies that the normal linear model is wrong. Because we discover that the model is incorrect and untrustworthy, it is unwise to trust the prior conclusion, and thus unwise to trust the confidence interval. (Metaphor**:** If a tree is poisoned, you cannot trust any fruit that the tree bears).

How to Check Assumptions
1. Normal Distribution → Histogram
2. Constant Variance → Fan(Bootstrapping)
3. Independent of each other → Residual plot

Problem # 3:
Price of Borden Cheese that week.
vol - # of units sold in grocery store
disp - dummy variable indicator of the presence of in-store display promotions
(inflation adjusted) price - cost at which unit were sold at
store - location of grocery units were sold at

Load cheese.csv
summary(cheese) → discover wide spread

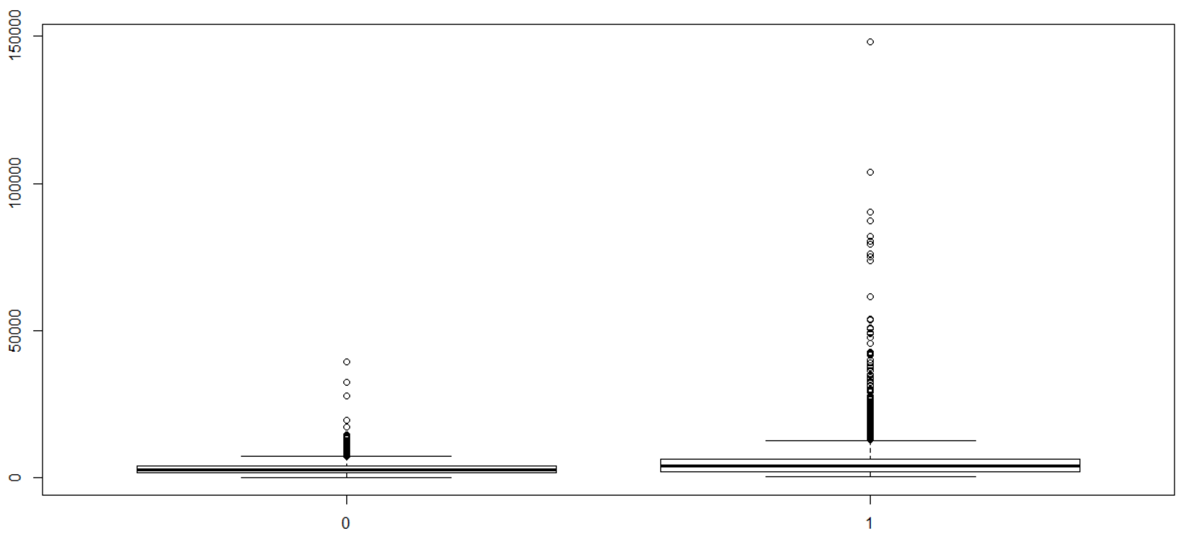
```
> summary(cheese)
                            store          price             vol              disp       
 BALTI/WASH - SAFEWAY          :  68   Min.   :1.320   Min.   :   231   Min.   :0.0000  
 BALTI/WASH - SUPER FRESH      :  68   1st Qu.:2.457   1st Qu.:  1990   1st Qu.:0.0000  
 BIRMINGHAM/MONTGOM - KROGER   :  68   Median :2.703   Median :  3408   Median :1.0000  
 BOSTON - STAR MARKET          :  68   Mean   :2.869   Mean   :  4771   Mean   :0.6457  
 BUFFALO/ROCHESTER - TOPS MARKETS:  68   3rd Qu.:3.203   3rd Qu.:  5520   3rd Qu.:1.0000  
 BUFFALO/ROCHESTER - WEGMANS   :  68   Max.   :4.642   Max.   :148109   Max.   :1.0000  
 (Other)                       :5147                                                     
> 
```

xtabs(~store, data=cheese) → we find that there are around 50-70 observations for each store

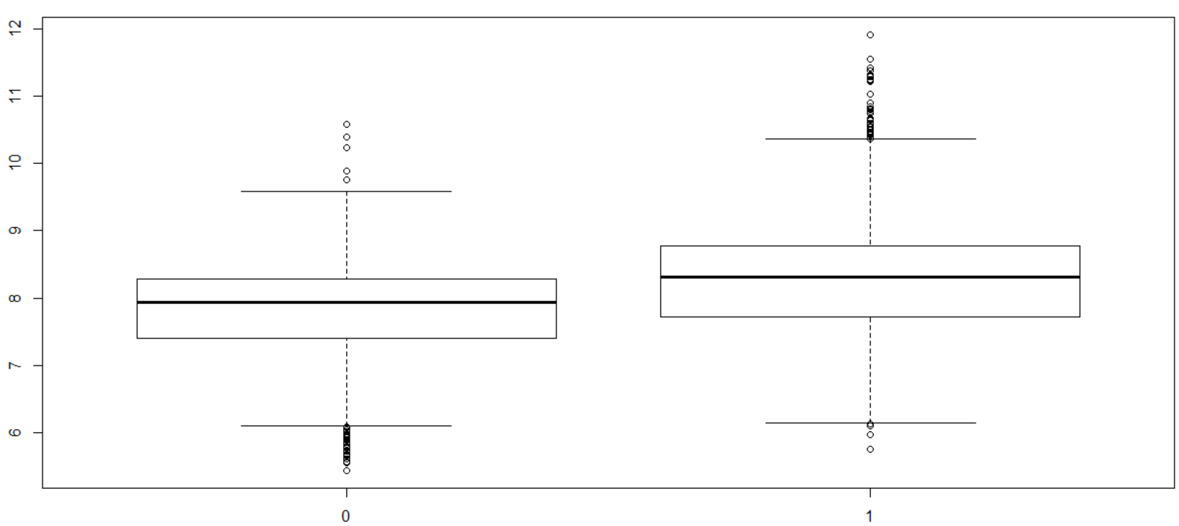
```
> xtabs(~store, data=cheese)
store
         ALBANY,NY - PRICE CHOPPER              ATLANTA - KROGER CO            ATLANTA - WINN DIXIE 
                                61                               61                              61 
       BALTI/WASH - GIANT FOOD INC             BALTI/WASH - SAFEWAY        BALTI/WASH - SUPER FRESH 
                                61                               68                              68 
       BIRMINGHAM/MONTGOM - BRUNOS      BIRMINGHAM/MONTGOM - KROGER BIRMINGHAM/MONTGOM - WINN DIXIE 
                                61                               68                              61 
                   BOSTON - SHAWS             BOSTON - STAR MARKET           BOSTON - STOP & SHOP 
                                61                               68                              61 
    BUFFALO/ROCHESTER - TOPS MARKETS      BUFFALO/ROCHESTER - WEGMANS             CHARLOTTE - BI LO 
                                68                               68                              61 
             CHARLOTTE - FOOD LION       CHARLOTTE - HARRIS TEETER         CHARLOTTE - WINN DIXIE 
                                61                               61                              61 
               CHICAGO - DOMINICK                 CHICAGO - JEWEL                  CHICAGO - OMNI 
```

3a) Plot boxplot.

>boxplot(vol~disp, data=cheese)



There is a long upper tail and squished box for the boxplot with the in-store display variable compared to the boxplot with the no in-store display variable. (1 - with in-store display, 0 - without in-store display). Thus, we must take the log of the volume variable and plot boxplot again using log volume.
>boxplot(log(vol)~disp, data=cheese)



We then discover that the mean volume with the in-store display variable is higher than without the in-store display variable. (1 - with in-store display, 0 - without in-store display)

We then fit linear model with groupwise mean to retrieve the baseline-offset form. *Important thing to check:* Did we inappropriately aggregate data by store? We discover that we did because different stores have different volumes of sales. To account for the differences, we must put in a dummy variable to estimate the store "nudges" that move the mean volume up or down. Volume will change based on if there is a display or not.

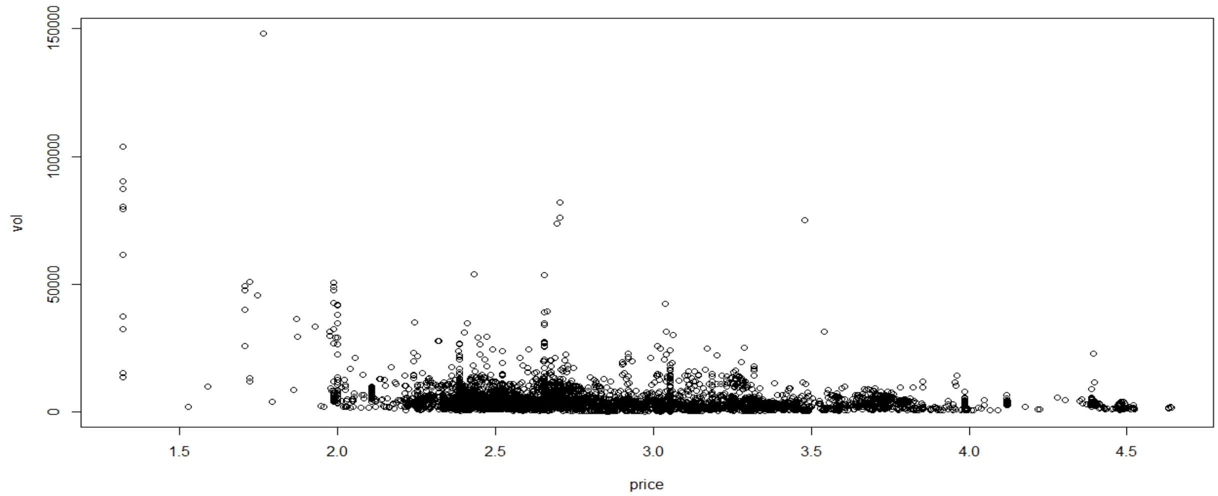
```
>lm2 = lm(vol~disp + store, data=cheese)
>summary(lm2)
```

```
Call:
lm(formula = vol ~ disp + store, data = cheese)

Residuals:
   Min     1Q Median     3Q    Max
-15722  -1143   -261    559 121273

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         588.11     597.67   0.984 0.325151
disp                               1970.13     163.91  12.019  < 2e-16 ***
storeDALLAS/FT. WORTH - ALBERTSONS 2939.56     840.57   3.497 0.000474 ***
storeDALLAS/FT. WORTH - KROGER CO  2541.10     841.14   3.021 0.002531 **
storeDALLAS/FT. WORTH - TOM THUMB  2821.11     840.72   3.356 0.000797 ***
storeDALLAS/FT. WORTH - WINN DIXIE -370.06     881.17  -0.420 0.674526
```

Disp= 1970
For Kroger DFW when disp = 0, our baseline + offset = 588.11 + 2541.10
For Kroger DFW when disp = 1, our baseline + offset = 588.11 + 2541.10 + 1970.13
From this data, we can see that the volume is higher for stores with in-store display promotion.

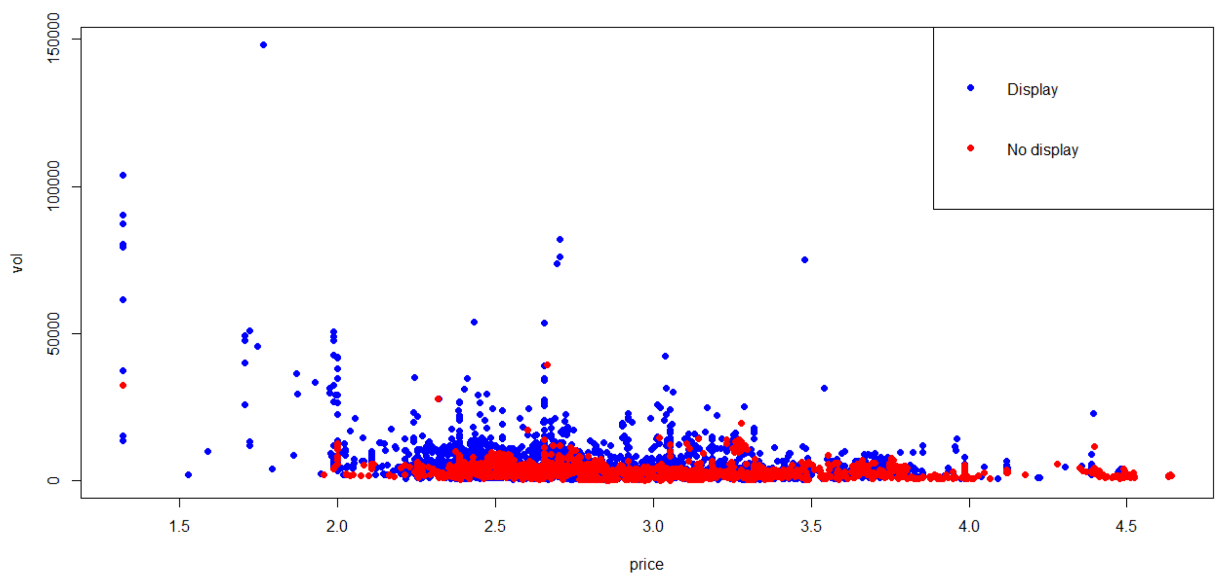
3b) Is the presence of in-store display promotion correlated with the price/pricing strategy of cheese? (Cheese on sale or cheap cheese with display)
Plot Volume vs Price.

```
>plot(vol~price, data=cheese)
```



We need a model that accounts for both in-store display and no in-store display.

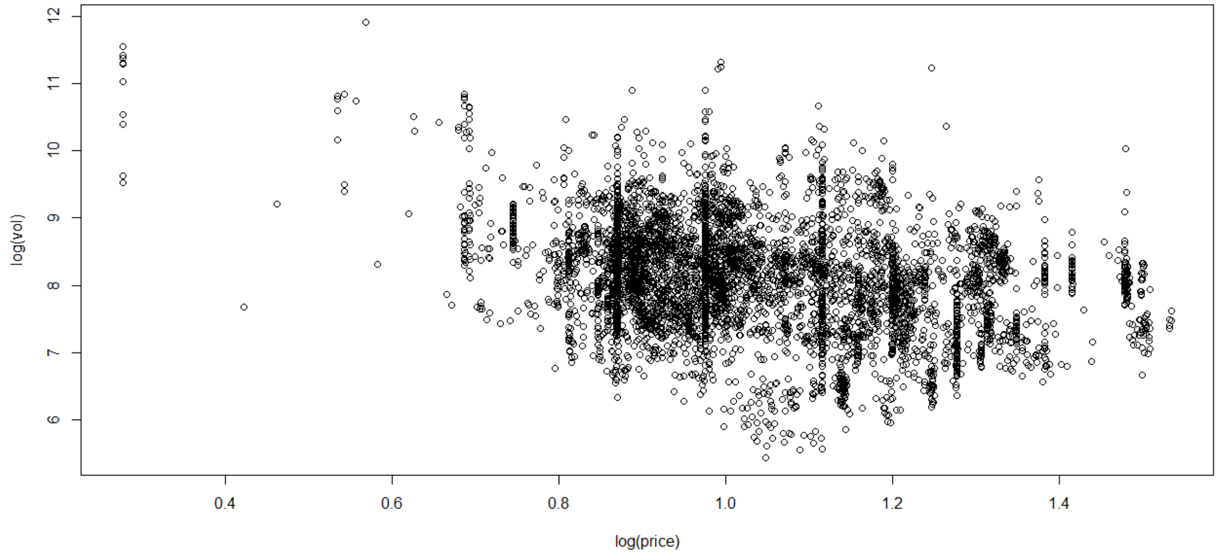
```
>points(vol~price, data=subset(cheese, disp==1), col='blue', pch=19)
>points(vol~price, data=subset(cheese, disp==0), col='red', pch=19)
>legend("topright", legend=c("Display", "No display"), pch=19, col=c('blue', 'red'))
```



Points with no in-store display promotion have generally higher prices, implying that there are sales or cheaper prices with the presence of in-store displays

Demand Curve: Q≈ K * price^β

Plot log Volume vs log Price, aggregating by all stores



Fit linear model between logs of x & logs of y. Y must be able to change based on store & presence of in-store display.

>lm3 = lm(log(vol)~log(price) + store, data=cheese)

>summary(lm3)

```
Call:
lm(formula = log(vol) ~ log(price) + store, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8553 -0.1559 -0.0180  0.1346  3.3308 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       9.56181    0.05201 183.854  < 2e-16 ***
log(price)                       -2.64380    0.03457 -76.479  < 2e-16 ***

storeDALLAS/FT. WORTH - ALBERTSONS  1.59834    0.05469  29.224  < 2e-16 ***
storeDALLAS/FT. WORTH - KROGER CO   1.48359    0.05465  27.149  < 2e-16 ***
storeDALLAS/FT. WORTH - TOM THUMB   1.44191    0.05467  26.377  < 2e-16 ***
storeDALLAS/FT. WORTH - WINN DIXIE  0.42424    0.05706   7.435 1.21e-13 ***
```

Is demand curve shifted up or down by presence of display? Disp = .1850 → causes shift up

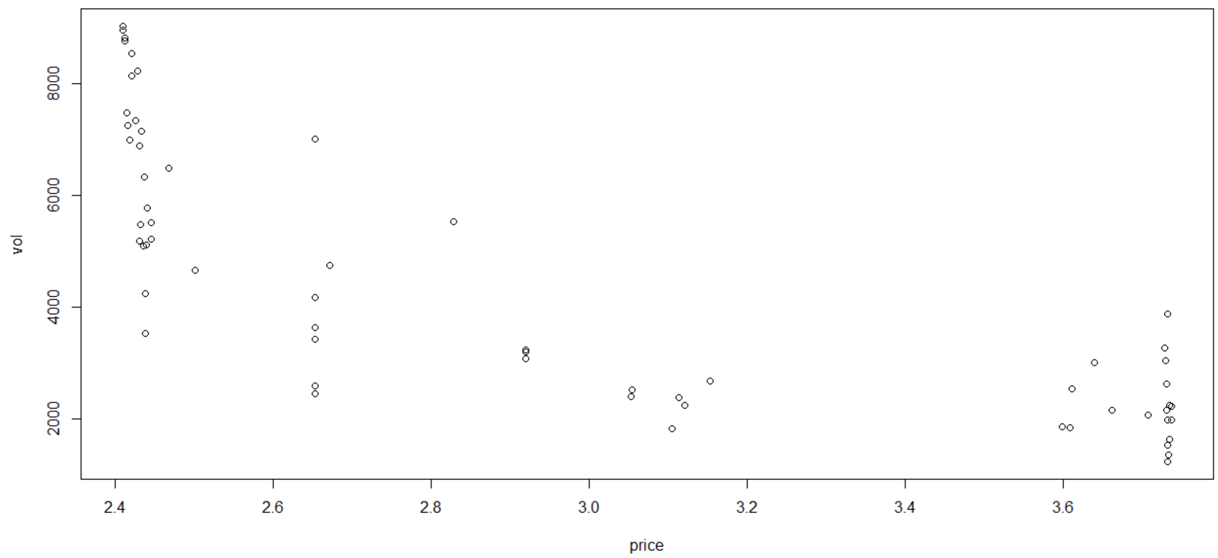
Create two subset for adv (in-store display) and no adv (no in-store display).

>dfwkroger = subset(cheese, store=='DALLAS/FT. WORTH - KROGER CO')

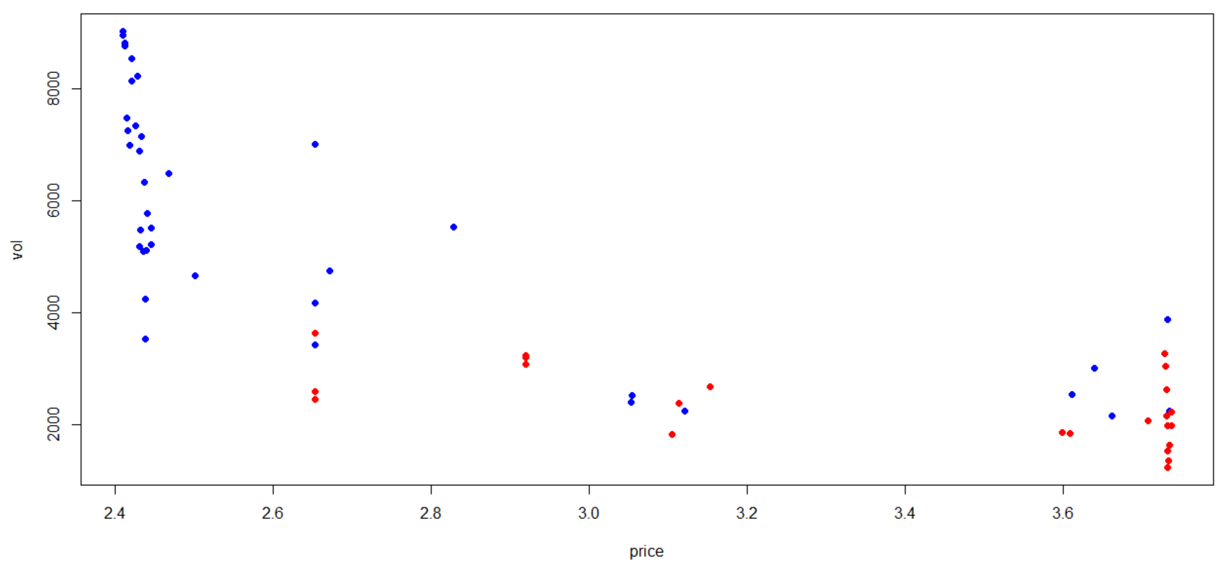
>sub1 = subset(dfwkroger, disp==1)

>sub0 = subset(dfwkroger, disp==0)

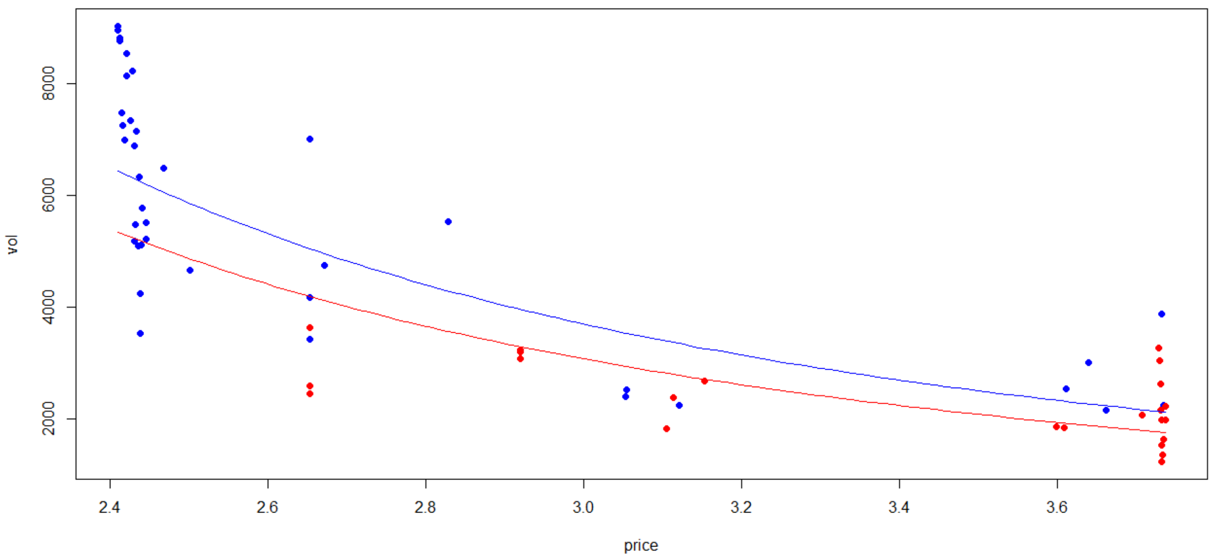
>plot(vol~price, data=dfwkroger)



>points(vol~price, data=sub1, col='blue', pch=19)
>points(vol~price, data=sub0, col='red', pch=19)



Plot Demand Curve for Kroger DFW: baseline + offset + dummy, slope
>curve(exp(9.37579 + 1.43461)*x^(-2.53159), add=TRUE, col='red')
>curve(exp(9.37579 + 1.43461 + 0.18540)*x^(-2.53159), add=TRUE, col='blue')



no adv: 9.38 + 1.43 + 0, -2.53
adv: 9.38 + 1.43 + .1850, -2.53

Must disaggregate by stores.

New Information

Interaction Term

Learn different slopes in 1 model

Y_1= log10 salary

x_{i1}= classes (MLB, AAA, AA) [Categorical]

X_{i2}= Batting average [Numerical]

Last Class:

MLB: $Y_i = \beta_o^{(MLB)} + \beta_2 X_{i2} + e_i$

AAA: $Y_i = \beta_o^{(AAA)} + \beta_2 X_{i2} + e_i$

AA: $Y_i = \beta_o^{(AA)} + \beta_2 X_{i2} + e_i$

β_o changes based on class (x_{i1}). β_2 remains the same regardless of change in class
Generates 3 regression equations with 1 slope and 3 different intercepts

Dummy Variable:

$$Y_i = \beta_o + \beta_2 X_{i2} + \beta_1^{(MLB)} \mathbf{1}_{2\{Xi= MLB\}} + e_i$$
$$+ \beta_1^{(AAA)} \mathbf{1}_{2\{Xi= AAA\}} + e_i$$

Interaction Term:

Assumes Slopes are different for each league

AA: $Y_i = \beta_o^{(AA)} + \beta_1^{(AA)*} X_{i2} + e_i$

AAA: $Y_i = \beta_o^{(AAA)} + \beta_1^{(AAA)*} X_{i2} + e_i$

MLB: $Y_i = \beta_o^{(MLB)} + \beta_1^{(MLB)*} X_{i2} + e_i$

Is there interaction between Batting average and league?

Baseline Offset Form:

$$Y_i = \beta_o + \beta_2 X_{i2'} + \beta_1^{(AAA)}\mathbf{1}_{2\{X1i= AAA\}} + \beta_1^{(MLB)}\mathbf{1}_{2\{X1i= MLB\}} + \gamma_2^{(AAA)}\mathbf{1}_{2\{X1i= AAA\}}{}^*X_{i2} + \gamma_2^{(MLB)}\mathbf{1}_{2\{X1i= MLB\}}{}^*X_{i2}$$

***Note: It does not matter which category is the baseline

- 6 coefficients, 3 different intercepts, 3 different slopes

What is regression equation when X_{i1} = AA? When X_{i2} = AAA? When X_{i2} =s MLB?

AA: $Y_i = \beta_o + \beta_2 X_{i2} + 0 + 0$

AAA: $Y_i = \beta_o + \beta_2 X_{i2} + \beta_1^{(AAA)}\mathbf{1} + 0 + \gamma_2^{(AAA)}\mathbf{1}{}^*X_{i2} + 0 + e_i$

$= [\beta_o + \beta_1^{(AAA)}] + [\beta_2 + \gamma_2^{(AAA)}]{}^*X_{i2} + e_i$

Baseline Offset

Different intercept Different Slope

MLB: $Y_i = \beta_o + \beta_2 X_{i2} + 0 + \beta_1^{(MLB)}\mathbf{1} + 0 + \gamma_2^{(MLB)}\mathbf{1}{}^*X_{i2} + e_i$

$= [\beta_o + \beta_1^{(MLB)}] + [\beta_2 + \gamma_2^{(MLB)}]{}^*X_{i2} + e_i$

R-Script format:

```
>lm3 = lm(Log10Salary ~ BattingAverage + Class + Class:BattingAverage,
data=baseballsalary)
>summary(lm3)
```

```
Coefficients:
                          Estimate Std.
(Intercept)                 2.8488
BattingAverage              5.3985
ClassAAA                    1.7936
ClassMLB                    0.3148
BattingAverage:ClassAAA    -2.6468
BattingAverage:ClassMLB     6.0100
---
```

β_o = (Intercept)

β_2 = Batting Average

$\beta_1^{(AAA)}$ = ClassAAA

$\beta_1^{(MLB)}$ = ClassMLB

$\gamma_2^{(AAA)}$ = BattingAverage: ClassAAA

$\gamma_2^{(MLB)}$ = BattingAverage: ClassMLB