STA375H NOTES 2/25/2015

Intro:

Discussed impending test: be sure you can answer essay questions (like the ones on the homework), especially about concepts, essential ideas (i.e. Bootstrapping vs. Linear; what IS bootstrapping?; assumptions of bootstrapping, describe what you would do with bootstrapping, and what is the point of bootstrapping? What is an interaction term, what does the slope represent? What is a regression line?)

Practical thing you might do – try to make very explicit connections between lectures, walkthroughs, and the way those skills are playing through on the exercises, in a substantive sense, not mathematical sense > you'll be in great shape for the midterm; all you need is a blue book and a pen; generally between 3-7 question in varying length

Going over Homework 5

There is a clear relationship between log volume and display, at least as shown through boxplot:

R Code:

lm1=lm(log(vol)~factor(disp), data=cheese)

coef(lm1)

summary(lm1)

*vary narrow confidence interval

- Remember to convert back from log when you have a log function, as is the case here

log (sales) = $b_0 + b_1$ (1 if disp=1)

sale = e^{B0} $e^{B1(disp=1)}$

*when there is no display, we get a linear model of only sales=e^{B0}

- Estimated sales percent change variable is = exp*0.43 = 1.54, so 54% higher/ or you could use the function exp(confint(lm1))

Part B

-need to check if grouping variables are affecting response

- Look at store by store, you can see the drastic difference
- Incorporate store, but adding it to the regression model =

  R code:

lm(log(vol ~disp +store, data=cheese)

coef(lm2)

- Look at display variable confidence interval …somewhere between 35 and 43 percent change
  R code: (exp(confin(lm2))
- Probably using display ads to give attention to a sale or discount – is there a correlation between price and sales volume, because of the display ad advertising the discounted price?
- Look at boxplot(price ~disp, data=cheese) = when there is a display, there is a systematic increase in sales

  Lm3= lm(log(vol)~log(price)+store, data=cheese)

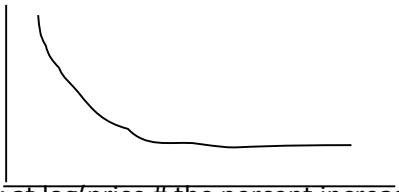
  Lm4 = ml(log(vol)~ log(price) + store + disp > exp(confint(lm4)) > changes the disp interval= 17 to 22% change

Looking specifically at different stores

- Log sales = b_0 + $b_1$log*price + b_2 **1(disp)** + B_3(**1(kroger) =** 9.4 +0.18*1(if disp)-2.5 log(price) + 1.4

Part C

- Sales = e^B0 price ^B1
- Demand curve >>>



- Exp(confin(lm4([‘disp,] > look at log(price # the percent increase associated with display, adjusting for price
- the more negative beta one is , the more sensitive consumers are to price
- Price elastic of demand = how sensitive you are about the price adjustment of cheese…two scenarios

  -scenario 1: you go in, no ads, just want to get cheese, not really aware of its price

  -scenario 2: go in see sale on cheese, $7 a package. If there’s a display ad, it looks like the cheese is more expensive than they would expect, consumers would be more price sensitive when they see something that’s supposed to be on sale.

Normal Linear Regression Model

- $Y_i = B_0 + B_1X_i + E_i$
- $E_i \sim N(0, \text{sigma}^2)$ assumptions

  Math and prob. theory

  ⇩

  Standard errors

- Showed mathematical equations within notes, figures 5.6-5.9
- Remind you what standard errors mean = when you take samples from population, a finite sample won't get us the true line, a variety of lines (a fan built up over time) of what the fitted line of a given sample is…these numbers, the sigma squared 0 means the variance of the sampling distribution of the intercepts, or how spread out are those intercepts from all your samples
- Two different ways of quantifying standard error > normality and bootstrapping (they both require assumptions) > if they disagree, it's because of their assumptions, some failure of one set of assumptions or the other
- How to interpret standard error within the store variables (example), under the assumptions of the normal linear regression: if we had taken a different sample, that standard deviation would have been different by 0.05 (std error column within R program)

Example of Galapagos Data

plot(Species~Area, data=gala)

plot(log(Species)~log(Area), data=gala)

lm1=lm(log(Species)~log(Area), data=gala)

abline(lm1)

coef(lm1)

plot(log(Species)~log(Elevation), data=gala)

-plot shows a trend; question arises: are these two variables a confounder for each other?

-if you are going to investigate a hypothesis of biological diversity how can you isolate the effect of elevation, which strips away the idea of area, considering that there is also a relationship

-adjust for the effect of area as well

- checked the residuals for a systematic correlation with elevation

lm1b = lm(resid(lm1)~log (elevation), data=gala)

- The regression shows that once I've taken species into account with area, that we have a negative correlation

- Once you take out those residuals, they have a negative correlation with elevation

- If elevation and area are correlated, if you are implicitly stripping out the area, you are stripping out elevation as well, thus is it difficult to tell which is driving the change in species count

coef(lm1b)

log(Area)~log(Elevation)

Tennis example:

Imagine that you play tennis, and your coach wants you to practice hitting the ball 100 times with a forehand swing into a little box he's drawn on the court. You count how many times you achieve this, which is 80/100. Your coach then tells you to improve by transferring more weight into your swing and changing your grip. You then get 90/100. But what contributed to the improvement, the extra weight or the change in grip?

It is impossible to explain the variation, just like in the problem, where elevation and area are competing.