1/22/14 Class Notes

Admin:

- You can have homework groups across sections
- You do NOT have to type out math problems: you can handwrite!
- You do NOT have to memorize R-Script: it will NOT be on the exam!
- Calculus will be used in this course enough that you need to know it
    - You don't have to memorize the functions or rules (Just wiki it!)
    - Rather understand the intellectual uses of Calculus
- Laptop closed during lecture unless scribing/downloading R
- You do NOT have to print out R script- we care about methods/conclusions/pictures
- TA DAVID PUELZ Office Hours Help Session
    - W 5:30-7:00 CBA 4.304A

Homework 1 Discussion:

1.) Whenever you see a causal claim like the one found in *Contraceptive Used in Africa May Double Risk of H.I.V.*, in a newspaper, it is wise to distinguish the conclusions of the researcher from the conclusions of journalist. Do not fault the survey because of how journalist reports it. Do we see confounders being addressed appropriately? Yes, the general consensus was that there was a lot needing to still be addressed in the article.

2.) What could have tripped you up in dataset:
    i. Load mosaic library
    ii. Import oxford dataset
    iii. Sometimes have to tell R what dataset looks like: In this case, you have to tell R that the first line of the data set is a Header row, not observation, so check "Yes" for Heading before importing dataset.
    iv. Variables: Year/Years Since/Price
    v. BOXPLOT is only for quantitative thing versus qualitative variables. Thus if you used the quantitative years, then you got a goofy boxplot
    vi. Use factor command (found in mammalsasleep video) to turn a quantitative factor into a qualitative factor

3.) Substantively speaking, this problem is asking us:
    i. How do we choose the best number to minimize our target function- what is the one number we should use to summarize all the numbers in that group?
        i. This number is Θ
        ii. We take the derivative of the SSE single function to determine that the sample mean (Ybar) minimizes Θ. This is determined through 5 lines of algebra.
    ii. Why do we do these calculus problems?
        i. Deeper appreciation for why out of all the numbers we could chose for Θ, why the average (y) minimizes the function
        ii. The most important part of these problems is to gain insight and a conceptual grasp.

Class Lecture: Variability

Statistical modeling:

- Goal= Partition variability of dataset
    - Ex: In the Oxford data set, the component of variability was the features of the flats themselves of the year it is sold in.

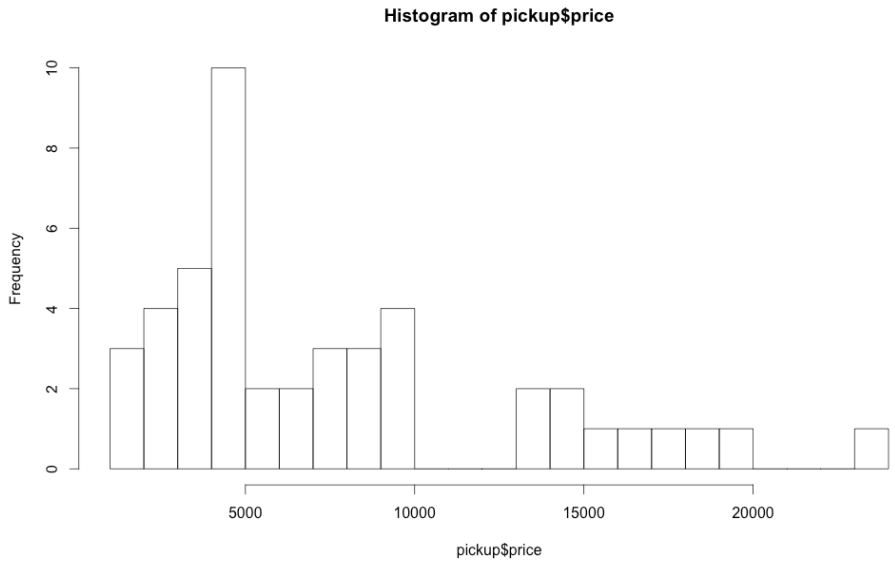
From this point forward, we used the data set pickup.R and pickup.csv

pickup.R Dataset

- Data of used cars in Austin from Craiglist
- Variables
    - Price
    - Mileage

Histogram

- This demonstrates variability in price
- Some trucks go for $2000 while others go for $20,000.



How do we measure variability?

1.) **COVERAGE INTERVAL**
    i. If we want 80% coverage interval, we will take 80% of data
    ii. We will pick the central data (aka from 10% to 90%)
    iii. The narrower the interval, the more concentrated around a point
    iv. Confidence intervals are subtler than coverage intervals. Coverage intervals are simpler; how do you cover __ percent of the points.

2.) **STANDARD (AVERAGE) DEVIATION**
    i. Standard deviation=Average Error (laments terms)

ii. How far are points from the mean on average (Yi-Ybar)
iii. We don't want positives and negatives to cancel each other out, so square them. Want to know the average distance, thus we sum them up then divide by (N-1)
iv. Why do we divide by N-1 instead of N? Mainly for convention and every piece of software implements standard deviation by dividing by N-1.
v. We then take the square root to get the units back into more understandable units.

Measuring Variability in R

i. Coverage Interval (If want 80%, chop off bottom 10% and top 10%)

```
> qdata(c(0.1, 0.9), price, data=pickup)
  10%   90%
2500 16240
```

qdata stands for quantile

ii. Standard Deviation

```
> sd(pickup$price)
[1] 5584.154
```

> mean(pickup$price)
[1] 7910.13

Thus, the average error of guessing 7910.13 for the mean is 5584.

Group Means:

Two conventions (each useful in its own context):

1. **Groupmeans Straight Up:** Present groupmeans straight up.

```
> mm(Price~Year, data=oxford)

Groupwise Model Call:
Price ~ Year

Coefficients:
  2000    2001    2002    2003    2004    2005    2006    2007    2008
295051  326919  375155  382481  407773  441218  448484  456027  381698
  2009    2010    2011
395887  424327  398897
```

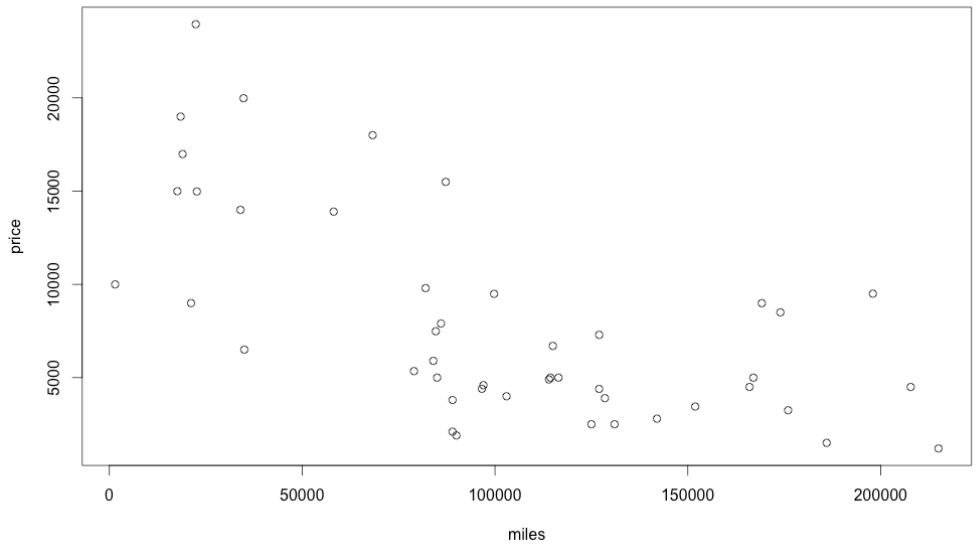
2. **Baseline/Offset Form:**
   i. The baseline is the first number. The offsets are the coefficients that you add to the baseline to get the group means.

ii. Store group means in Groupmeans.
iii. Use lm function for a linear model and store this model in an object titled lm1. The model is fit and ready for subsequent study.
iv. Then use coef(lm1) to view. Mean in 2003 says 87,430.49, which is quite a bit different from 400,000. The number being presented to you is difference between the group mean and the baseline (coefficient for 2000).
v. Finally take baseline and add the coefficient (offset) to determine the group mean.
vi. Why do we do this?
    1. Because statistical modeling is interested in differences (between years/people/drugs in clinical trials, etc.) and this baseline/offset form presents the differences directly
    2. Whenever you see ln, you will always get baseline/offset form
    3. This is interchangeable with dummy variable. We will discuss more later

Cbind function puts data in columns next to each other

Interpretation of fitting a line:

- Data from Pickup.R
    - Price-outcome variable
    - Mileage-predictor variable
- Create a scatter plot in R to plot data points and observe the trend
    - Do I need to show?
- Negative trend: As mileage goes up, the price goes down. Pay less for 50,000 mile truck than a 2,000 mile truck. (This is based on plot shown below)



How do we fit a linear regression line?

Mechanics:

i. If we want to fit a best trend line, use lm function:
   a. lm(price~miles, data=pickup)
ii. Computer draws the line with the formula: $Yi = b_0 + b_1x$
iii. Line is easy to fit with only two points because goes through both points
iv. When there are three points, you can't go through all points: you will miss
v. $Yi = b_0 + b_1x$ (trend model value fitted value)+ei (residual error)
   a. ei=random component which you can't predict systematically
   b. Residual=distance between fitted value and actual value
vi. The line uses Bo and B1 so as to make the sum of the squared residuals (ei^2) as small as possible.
   a. Why do we use sum of squared errors?
      i. You can use other measures like discrepancy, but sum of square errors is a standard for many reasons:
      ii. The calculus is a lot easier
      iii. Normal distribution is connected to least squares
      iv. Pythagorean Theorem (The most convincing)
      v. Every piece of software uses criteria of least squares
vii. HISTORY FUN FACT! Associated with French man Legendre- insert French accent- "take line, chose parameters of line, make sum of errors as small as possible and you will have best fitting line"

R Script Example:

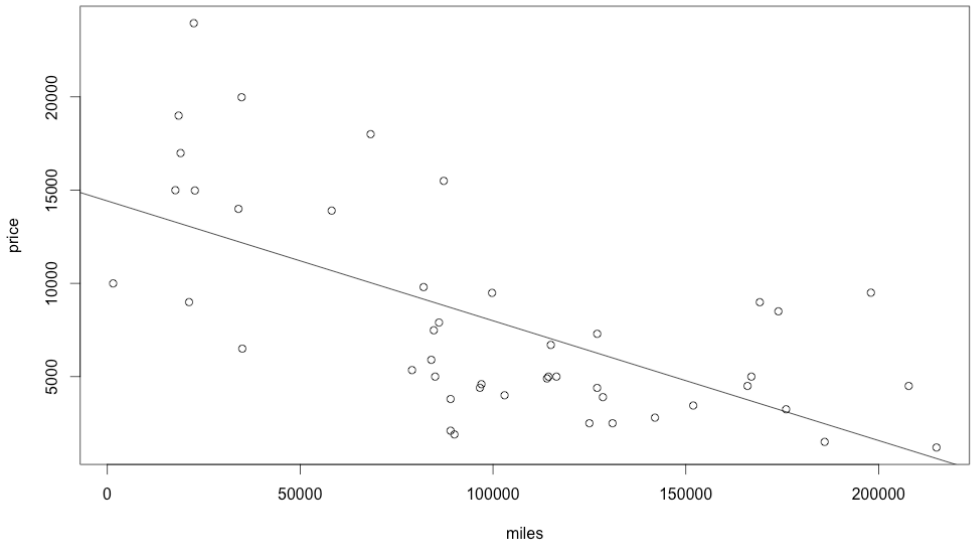
```
# Show a scatteplot of asking price versus mileage
plot(price~miles, data=pickup)

# Fit a trend line by least squares
lm(price~miles, data=pickup)

# Save this model in an object called model1
model1 = lm(price~miles, data=pickup)
```

Once you fit the model, put up the scatterplot, and then feed it the fitted model, it puts the least squares line on the plot.

```
# Show the scatterplot and add the trend-line
plot(price~miles, data=pickup)
abline(model1)
```



Four stories we can tell with regression model: *(Ranked from least to most subtle)*

1. **PLUG-IN PREDICTION**.
    i. Take new cases were you've seen predictor (x) value but not result (y) value, then make a forecast. In more detail on page 43 of course packet.
    ii. If you know the mileage (x-variable) of a new truck you have not seen before, but you do not know its mileage, what do you do? Plug in the x in the linear regression equation to find Y. Or reference the above graph. If it has 90,000 miles, the line predicts on average a fair price of around $9,0000.
    iii. R example. Compare 3 new trucks

```
# Story 1: plug in prediction
newx = c(25000,50000,100000)
yhat = 14419.3762 + (-0.0643)*newx
```

    Result will be: 12811.876

2. **SUMMARIZING THE TREND**.
    i. A lot of summary but useful summary. What you lose in information, you gain in clarity and ability to summarize.
        i. If there are 46 trucks (92 pieces of information. (mileage and price) Compress data to slope and intercept (go from 92 to 2 pieces of information)
    ii. Description and Interpretation of Variables:
        i. Intercept- where the line hits y-axis when x variable is zero
            i. Expected value of y-variable (price) when x-variable is identically zero.
                a. If we saw a used truck with zero miles on it, we would use the line to locate the intercept and pay that price.
                b. $14,000 in this model (reference graph above)
        ii. Slope- a few different ways to understand what this variable means

i. $Y(hat) = b_0 + b_1x + e$

ii. One way to interpret b_1:
   a. Take the derivative: Dyi/dx= 0+B1
   b. Rate of change (change in y over change in x) Change in dollar value per change in value (B1=0.065) For every change in mile, get a .065 change in price.

iii. Another way to interpret b_1:
   a. Work out units
   b. Price=Price+(B1)Miles (to make whole thing have unit of price, what does B1 unit have to be?
      i. (price/miles) is the only way units will work

iv. ex: You could sell a truck now and expect to get a certain price or drive for another year and sell it then. What do you think the difference in price would be? Slope times change (additional time plan to drive)

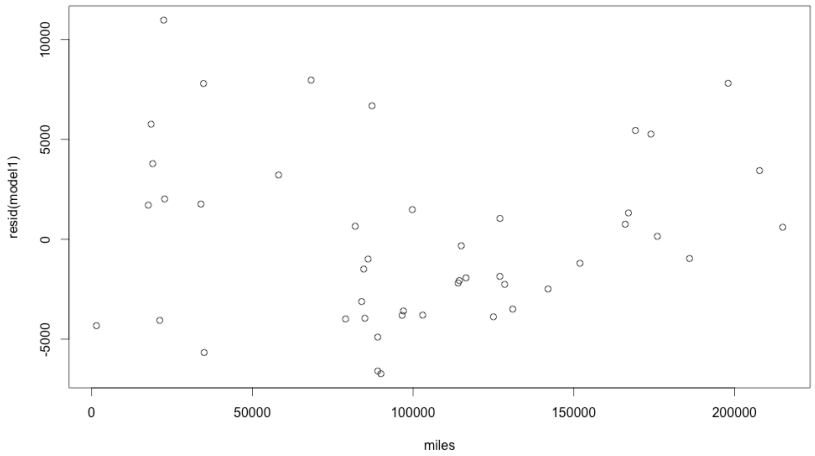
v. This allows you to reason about marginal return. Is it worth losing $642.99 to drive for another year?

```
# Story 2: summarizing the trend
coef(model1)
-6.429944e-02 * 10000
```

3. **STATISTICAL ADJUSTMENT.** *"Taking the X-ness out of Y"*
   i. Adjusting for systematic features you know, then determining the best deal or value
   ii. Ex: Moneyball, Sports, Unemployment reports
      i. The unemployment rate decreased to 7.55 *on a seasonally adjusted basis*. In other words, these are not the raw numbers.
   iii. Model has both fitted values and residuals. Can use R to print these
      coef(model1)
      fitted(model1)
      resid(model1)
   iv. Can you see x-ness in Y?
      i. YES! This is the trend demonstrated in the linear regression graph above. There is a negative relationship between X and Y
   v. How to remove the "x-ness"?
      i. We fit the plot, and then plot the residuals against Y, and there is no systematic trend as illustrated below.

```
# Story 3: taking the X-ness out of Y
# Which truck looks like the best deal, adjusting for mileage?
# Compare with http://usedcars.truecar.com
plot(resid(model1) ~ miles, data=pickup)
```

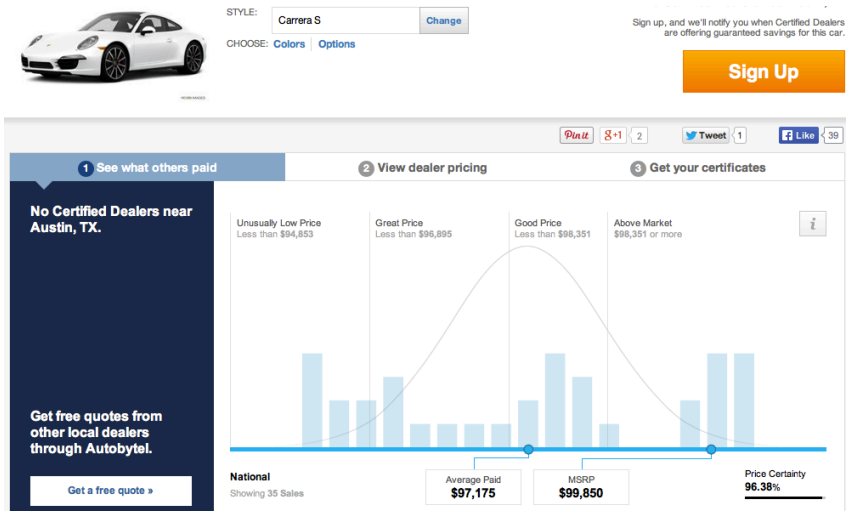
ii. The y-axis is a mileage-adjusted truck. Looking at the scatterplot, you should look for the points that have smallest residual (largest negative number) to find the best value.
    i. In this case, the best deal for a mileage-adjusted truck is around a 100,000 mileage truck



    ii. You can also determine this in a nonvisual manner by calling the minimum residual

```
# What's the minimum residual, and which truck is it?
min(resid(model1))
which.min(resid(model1))
```

vi. Ex: True Car Website



4.) **REDUCING UNCERTAINTY:**
    i. Quantification of relevancy.
    ii. This was only briefly mentioned in class and is covered in more detail in the textbook on pages 47-49 of course packet