2/24 Scribed Notes
Amy Yu, Montana Blair — 9:30 – 11 am

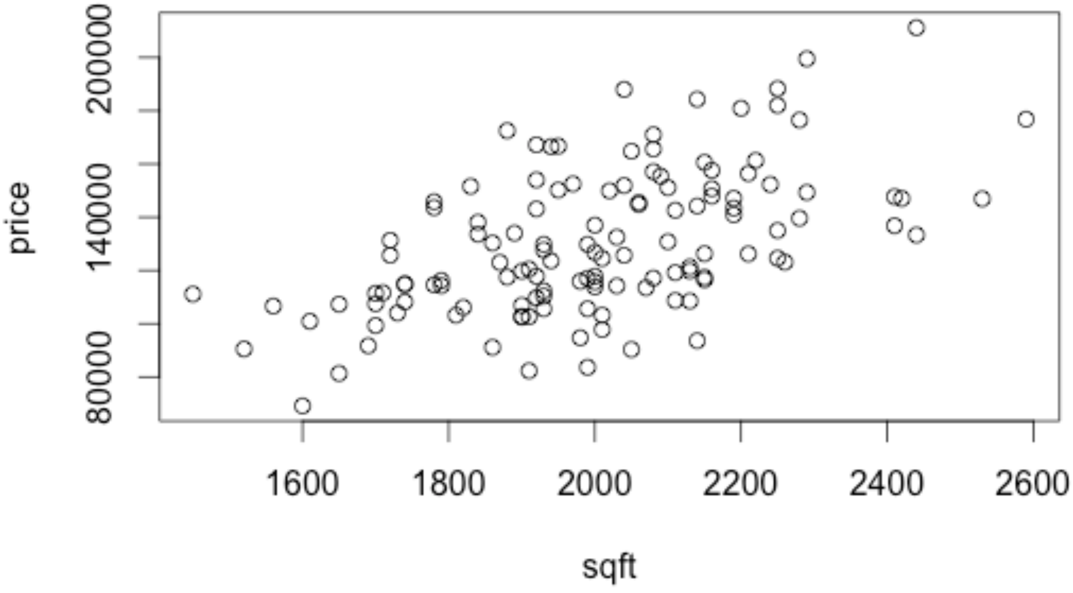
The Midterm:
Old pen and paper — simple, straightforward prompts about the core concepts. We will have a review session and go over everything point by point. Aim for concise, direct answers — no expectation to write a giant essay.

Short answers, long answers, everything in between — true/false, an essay are all possibilities. There is no R. In-class review session next Monday and review session Tuesday evening.

Today: analysis of variance (ANOVA) and permutation tests (to be continued Wednesday, 2/26)

Exercise 6 Recap:

1. We see 8 variables, one of which we do not need—"offers". We plot price v. sqft and there is a positive trend.



- Aggregation paradox: when we don't include something important, like a dummy variable; is a poor estimate of an effect size

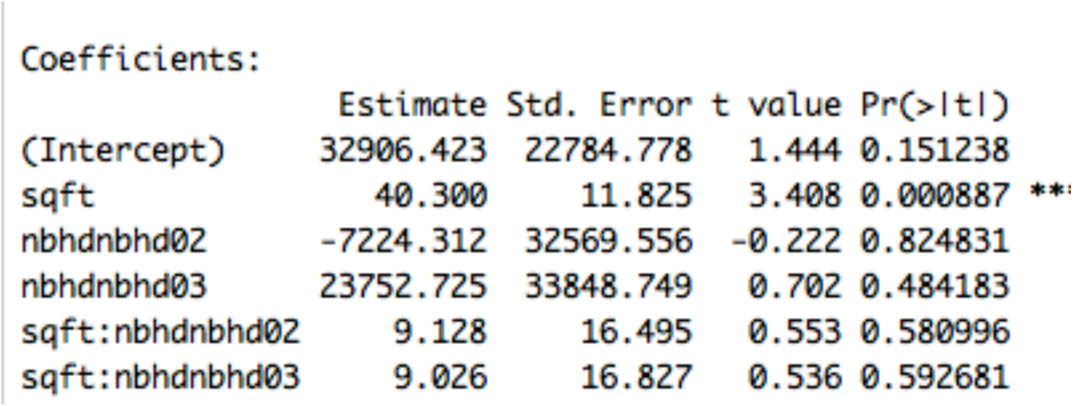
    - We need dummy variables. It’s obvious that neighborhood 3 is the most expensive.
    - If we don’t look at neighborhood, it improves the property value by about $70. If we include neighborhood, it improves property value by $46.
    -

  - Will there always be that change in slope with the aggregation paradox? No; we have inappropriately aggregated houses across neighborhoods here, so that changes the slope.
  - This is like the baseball data set as we fit 3 regression models.
  - Neighborhood is a dummy variable in R because it’s represented as a categorical variable.

Try the Interaction Model:

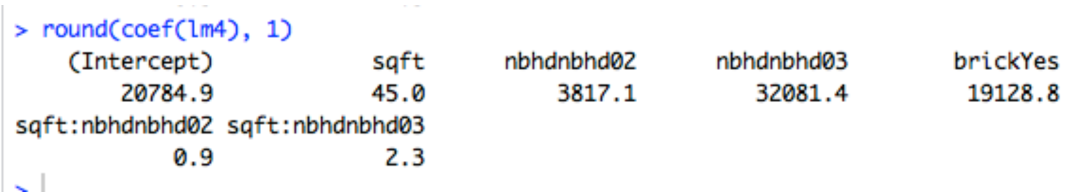
- lm3 = lm(price ~ sqft + nbhd + nbhd:sqft, data=house)
    - results: constant slope across neighborhoods so maybe the price premium is constant, so we fit 3 different lines to the 3 models

- If we compare lm2 (constant slope w/ no interaction terms) to lm3, we have an r-squared of .6861.
    - The difference in predictor terms here is small; so the effect of including interaction terms is not that large.
    - The confidence interval is another check —due to the wide interval we know that the data is unable to estimate these terms very precisely.
- We may not want to put all interaction terms into the model if 1) confidence interval is too wide to be precise and 2) if the r-sq is not significantly larger



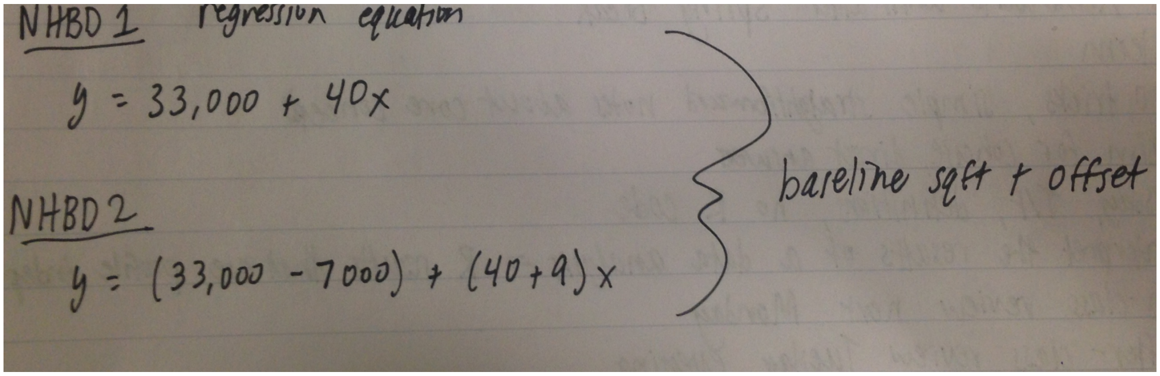
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     32906.423  22784.778   1.444 0.151238
sqft               40.300     11.825   3.408 0.000887 **
nbhdnbhd02      -7224.312  32569.556  -0.222 0.824831
nbhdnbhd03      23752.725  33848.749   0.702 0.484183
sqft:nbhdnbhd02     9.128     16.495   0.553 0.580996
sqft:nbhdnbhd03     9.026     16.827   0.536 0.592681
```

What happens if we include brick in the model? (lm4)



```
> round(coef(lm4), 1)
   (Intercept)            sqft      nbhdnbhd02      nbhdnbhd03        brickYes
       20784.9            45.0          3817.1         32081.4         19128.8
sqft:nbhdnbhd02 sqft:nbhdnbhd03
           0.9             2.3
```

- The interaction terms are much smaller now
- We also get a big bump in r-squared up to .79. We have a smaller, precise confidence interval, so including the interaction terms has practical implications. We keep bricks.
- Baseline offset form: = (45+.9)*1000 = $45900 bump in this model



NHBD 1 regression equation

$$y = 33{,}000 + 40x$$

NHBD 2

$$y = (33{,}000 - 7000) + (40+9)x$$

baseline sqft + offset

- Lm5: add bedrooms and bathrooms (footballs and clouds! We're in 4 dimensions)
    - The estimate for the square foot slope goes down because we allocate more variance to bedrooms and bathrooms. Makes sense because additional square feet have to go towards more rooms
    - So due to collinearity, the slopes are all smaller together than if we estimated them individually.

Collinearity: Let's pretend you're practicing forehands and you get 80/100 in the box.

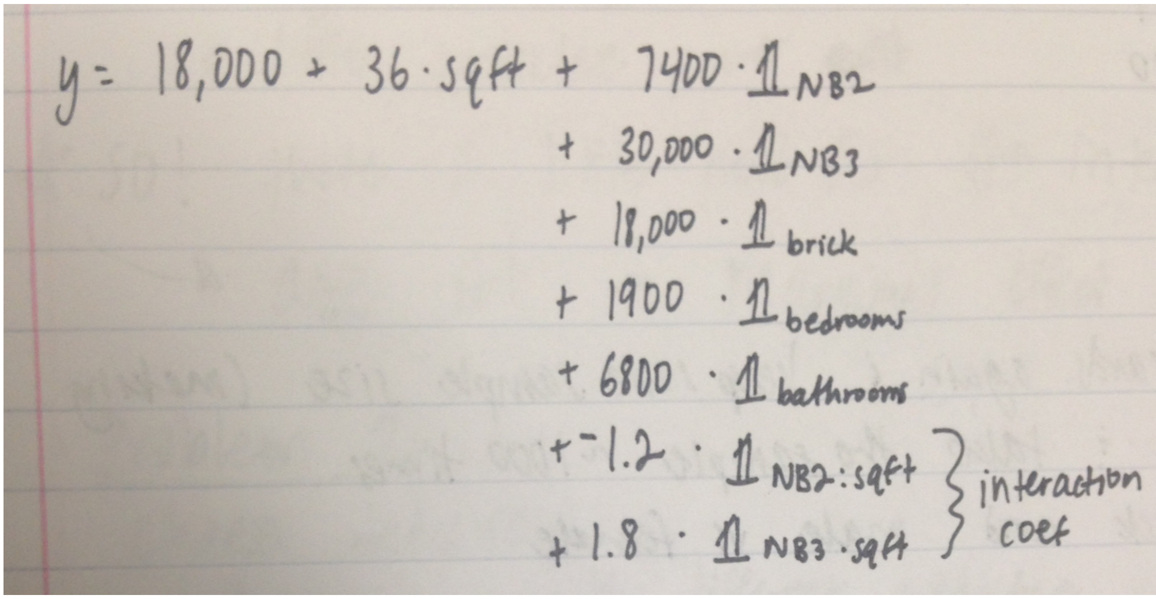
- Coach tells you to change weight and topspin
- Change 1: drive weight forward, increase topspin = 90/100 balls
- Change 2: do the opposite – drive weight back, decrease topspin = 70/100 balls
- You don't know which movement made your accuracy change AT ALL.
- You always want small, isolated changes so you can decouple marginal effects of different variables. So run multiple regression when this isn't possible

How do we know what to include in the model?

- Reason it out—is there practical significance? Is this range of effect likely?
- Often based on situation and type of data

Coefficient summary output:

- Each coefficient shows the relationship of each interaction term with price
- Allocation increases as you increase the amount of allocation terms
- Categorical and dummy variables change intercepts
- Ex: bedrooms and bathrooms change slope
- How to put the output into an regression equation:



$$
\begin{aligned}
y = 18{,}000 + 36 \cdot \text{sqft} &+ 7400 \cdot \mathbb{1}_{NB2} \\
&+ 30{,}000 \cdot \mathbb{1}_{NB3} \\
&+ 18{,}000 \cdot \mathbb{1}_{\text{brick}} \\
&+ 1900 \cdot \mathbb{1}_{\text{bedrooms}} \\
&+ 6800 \cdot \mathbb{1}_{\text{bathrooms}} \\
&\left.\begin{aligned}
&+ {}^{-}1.2 \cdot \mathbb{1}_{NB2:\text{sqft}} \\
&+ 1.8 \cdot \mathbb{1}_{NB3 \cdot \text{sqft}}
\end{aligned}\right\} \begin{array}{c}\text{interaction} \\ \text{coef}\end{array}
\end{aligned}
$$

Individual predictors are like wickers of a basket: they are stronger collectively ☺

How do you know how to go into the models?

- If we drop the interaction terms, very wide error bars and r-squared goes down. Since the difference is tiny, axe the terms
- There's little need for different price-sqft relationships

2. There's a practical difference between add and no-add.

- Does the price elasticity of demand change in the presence of the display? Yes? INTERACTION TERM!
  - With the interaction term, the R-sq is small
  - If it doesn't improve r-sq and you can't estimate coefficients precisely, axe the variable. But in this case we have a narrower confidence interval only and no bump in R-sq
  - What counts as a reasonable bump in R? The answer changes depending on how big your data set is.

Permutation Testing

If you notice a difference between the two groups of people with black/red cards and the dealer says he did it randomly—how can you prove it?

	Male	Female
Black	1	10
Red	4	1

- If you want to know if I shuffled the cards randomly the first time: make me do it again and ensure it's random
    - Keep sample size, from same population
- Then compare the discrepancy with the first trial

3. Define undercount and undercount rate (as fraction of count/ballots).
    - If we look at residuals, the punch cards seem to have higher rates of undercounts and lever cards have lower rates.
    - Now we put in a dummy variable for equipment. We can tell if the cards were systematically dealt to different counties.