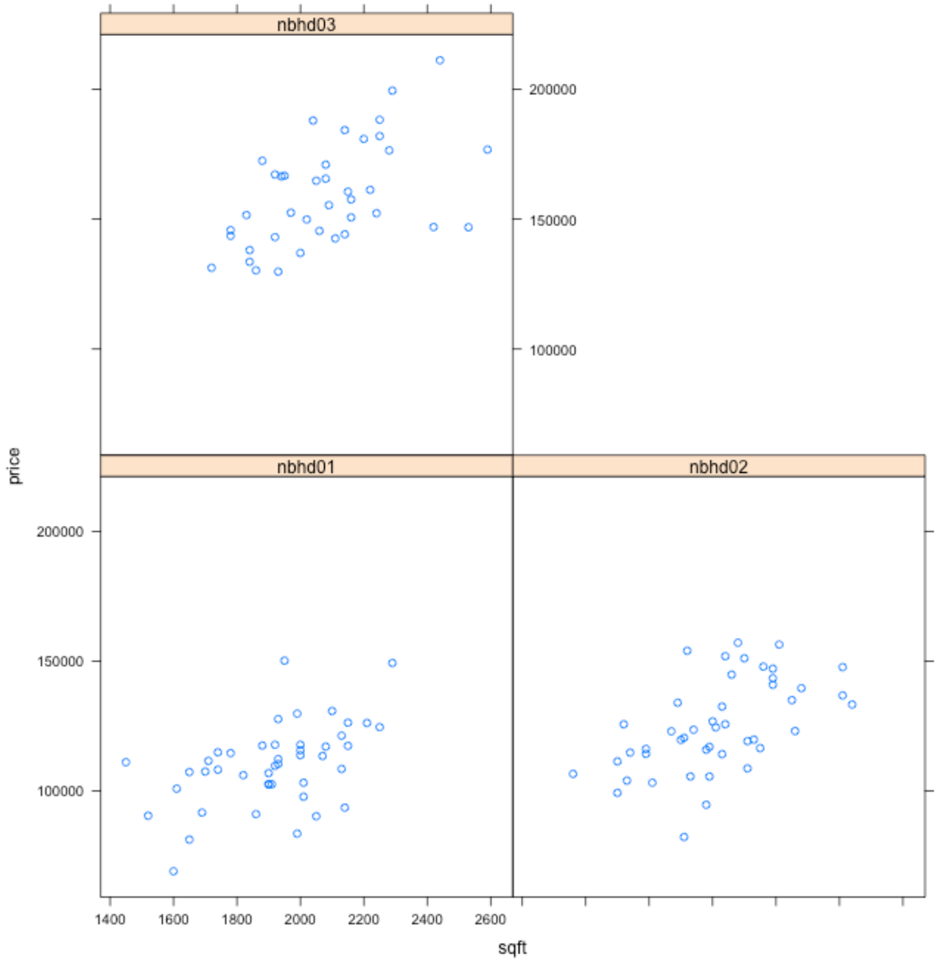
Hayden McMurrey and Christin Urso
STA371H—11 A.M.
February 24, 2014

HW6 recap—R Script posted on the website

#1) House Prices

Step 1: `plot(price ~ sqft, data=house)`

Step 2: `xyplot(price ~ sqft | nbhd, data=house)`



Step 3: Evaluate summary of model for price determined by square foot and determine the presence of an aggregation paradox and need for dummy variables. Aggregation paradox: when you get the wrong estimate of a coefficient by inaccurately aggregating across groups. Seen in the baseball data.

Step 4:

```
lm2 = lm(price ~ sqft + nbhd, data=house)
```

```
summary(lm2)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21241.174  13133.642   1.617  0.10835
sqft           46.386      6.746   6.876 2.67e-10 ***
nbhdnbhd02  10568.698   3301.096   3.202  0.00174 **
nbhdnbhd03  41535.306   3533.668  11.754  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 15260 on 124 degrees of freedom
Multiple R-squared:  0.6851,     Adjusted R-squared:  0.6774
F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

Based on the summary statistics for the linear model creating a dummy variable for nbhd, we see that the value of the house goes up about $46,000 per additional 1000 sqft. R is splitting up dummy variables and finding coefficients for each, so we see that the intercept for nbhd 2 is $10,000 more than nbhd 1 and the intercept for nbhd 3 is $41,000 more than nbhd 1. All neighborhoods are assumed to have the same slope, increasing by $46 in value per sqft. We also see that R squared determines the predictive power to be 68.5% with an error average forecasting error of about $15,000 (residual standard error).

Step 4: Now we add an interaction term between nbhd and sqft

```
lm3 = lm(price ~ sqft + nbhd + nbhd:sqft, data=house)
summary(lm3)
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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 15360 on 122 degrees of freedom
Multiple R-squared:  0.6861,     Adjusted R-squared:  0.6732
F-statistic: 53.32 on 5 and 122 DF,  p-value: < 2.2e-16
```

After reviewing the summary statistics, we see that R squared is 68.6%, which is only .1% higher than the previous model without the interaction term. The modest improvement in predictive power brings up the tradeoff between precision vs. simplicity. This is an important factor to consider in the real world when sharing results with a tax collector, manager, etc. We also have to consider the error bars on the coefficients themselves. The model estimates nbhd 2 to be 9.1 and nbhd 3 to be 9.0, but the accuracy interval is very wide. This results in a double whammy for this model, because not only does it not have a significant difference in prediction power, but the estimates have an error range that is ±2*standard error of about 13 from the estimate. This could be the result of confounding variables that have an effect on the prediction.

Algorithm for Forecasting Y Variable

```
Y=33000 + 40*sqft - 7224*1{Nbhd 2}
                    + 22752*1{Nbhd 3}
                    + 9.1*{Nbhd 2}*sqft       -- Interaction
                    + 9.0*{Nbhd 3}*sqft       -- Interaction
                    + error
```

- - 1{Nbhd2} refers to the dummy variable. If the dummy variable is set as a 1, it means that you keep the dummy variable. If it is a 0, you would be multiplying by 0 which takes it away.

Step 5: To check if there are confounding variables, we add a dummy variable for brick to the model.

```
lm4 = lm(price ~ sqft + nbhd + nbhd:sqft + brick, data=house)
summary(lm4)
coef(lm4)
round(coef(lm4), 1)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.078e+04  1.876e+04   1.108    0.270
sqft            4.504e+01  9.723e+00   4.632 9.21e-06 ***
nbhdnbhd02      3.817e+03  2.677e+04   0.143    0.887
nbhdnbhd03      3.208e+04  2.780e+04   1.154    0.251
brickYes        1.913e+04  2.466e+03   7.756 3.08e-12 ***
sqft:nbhdnbhd02 9.183e-01  1.358e+01   0.068    0.946
sqft:nbhdnbhd03 2.341e+00  1.384e+01   0.169    0.866
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 12600 on 121 degrees of freedom
Multiple R-squared:  0.7903,     Adjusted R-squared:  0.7799
F-statistic: 76.01 on 6 and 121 DF,  p-value: < 2.2e-16
```

After reviewing the summary statistics, we see that brick has some predictive power because R-squared is now higher at 79%. The Dummy variable interactions went from 9 to .9 and 2.3, which shows how much the estimates are affected by the addition of brick to the model.

However, this model is still imprecise because the standard error is still very large at about 13 for nbhd 2 and 3. **You can force R to round to one decimal point rather than using scientific notation for the coefficients with `>round(coef(lm4), 1)`.

Step 6: We now add dummy variables for bedrooms and bathrooms to examine their effect on the model.

```
lm5 = lm(price ~ sqft + nbhd + nbhd:sqft + brick + bedrooms +
bathrooms, data=house)
summary(lm5)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     18210.081  18306.267   0.995 0.321878
sqft               35.748      9.954   3.591 0.000479 ***
nbhdnbhd02       7406.659  26075.159   0.284 0.776865
nbhdnbhd03      30280.004  27034.309   1.120 0.264944
brickYes        18509.650   2424.603   7.634 6.28e-12 ***
bedrooms         1905.369   1923.190   0.991 0.323826
bathrooms        6849.041   2585.457   2.649 0.009167 **
sqft:nbhdnbhd02    -1.256     13.240  -0.095 0.924602
sqft:nbhdnbhd03     1.836     13.472   0.136 0.891809
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 12250 on 119 degrees of freedom
Multiple R-squared:  0.8051,     Adjusted R-squared:  0.792
F-statistic: 61.45 on 8 and 119 DF,  p-value: < 2.2e-16
```

The summary for this model (lm5) shows that the interaction terms between nbhd and sqft are small and very imprecisely estimated. Therefore, we will exclude them from the model.

Step 7: Drop the interaction variables from the model.

```
lm6 = lm(price ~ sqft + nbhd + brick + bedrooms + bathrooms,
data=house)
summary(lm6)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17919.446  10474.046   1.711  0.08967 .
sqft           35.930      6.404   5.610 1.30e-07 ***
nbhdnbhd02   4865.694   2721.805   1.788  0.07633 .
nbhdnbhd03  34083.719   3168.987  10.755  < 2e-16 ***
brickYes    18507.779   2396.302   7.723 3.65e-12 ***
bedrooms     1902.169   1902.270   1.000  0.31933
bathrooms    6826.925   2562.812   2.664  0.00878 **
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 12150 on 121 degrees of freedom
Multiple R-squared:  0.805,Adjusted R-squared:  0.7954 
F-statistic: 83.27 on 6 and 121 DF,  p-value: < 2.2e-16
```

When we drop the interaction, the resulting model (lm6) is more precise than the original model because we are adjusting for confounding variables. It is also a better predicting model than the models that include interaction terms (which determine change in slope for each nbhd) because there is no evidence to suggest that there is a marginal effect on price from nbhd. Therefore, we keep slope the same across neighborhoods by dropping the interaction term.

The coefficient on square foot in lm6 is smaller than the coefficient from lm2 (36 vs 46). Why? Some variation in y that was originally predicted by square footage is now predicted by bedrooms and bathrooms. When you add bed/bathrooms, you sap some explanatory mojo of square footage. As a result, you see a higher standard error. If you start with a model and add new predictors, and some of the new predictors are correlated to the old predictors, the estimates from the old predictors are going to change. If the predictors are positively correlated, your old correlation will probably decrease.

Analogy: tennis lesson. You hit 100 forehands into a box in the back corner. Eighty out of 100 in the box. Your coach tells you to focus on weight transfer and hit with more topspin. Now 90 out of 100 land in box. Did your improvement result from the weight transfer or the topspin? You don’t know. Perfect collinearity: there is no way to contribute the improvement to one variable, since they weren’t changed in isolation. They aren’t perfectly correlated but are nonetheless positively correlated. Bedrooms, bathrooms, and square footage are positively correlated. Positive correlation can also be caused by confounding.

The most important things to look at when evaluating a new predictor are:
1) The predictor’s effect on the R-squared value (whether it improves R-squared or not)
2) The predictive ability of the new predictor (look at the residual standard error)
3) The precision with which you can estimate the coefficients of the new predictor (look at the error bars associated with the predictor)
4) The practical effect size of the predictor variable on the y-variable (depending on your data, the practical effect size may or may not be meaningful, so it’s important to consider this in context)

#2) Price elasticity of demand for cheese
Last week, we fit a power law for log(volume) versus log(price), but with dummy variables for store and a different dummy variable for display. We have a single slope for the log-price variable.

Do we need different slopes depending on whether there is an in-store display? Assess whether we need different slopes by using an interaction term between disp and log(price). The interaction term gives a reasonably precise estimate. R-squared, however, is pretty similar, indicating that we are not improving the predictive ability of the model. Context determines what constitutes an important bump in R-squared. The more data points and the more predictors, the harder it is to improve R-squared.

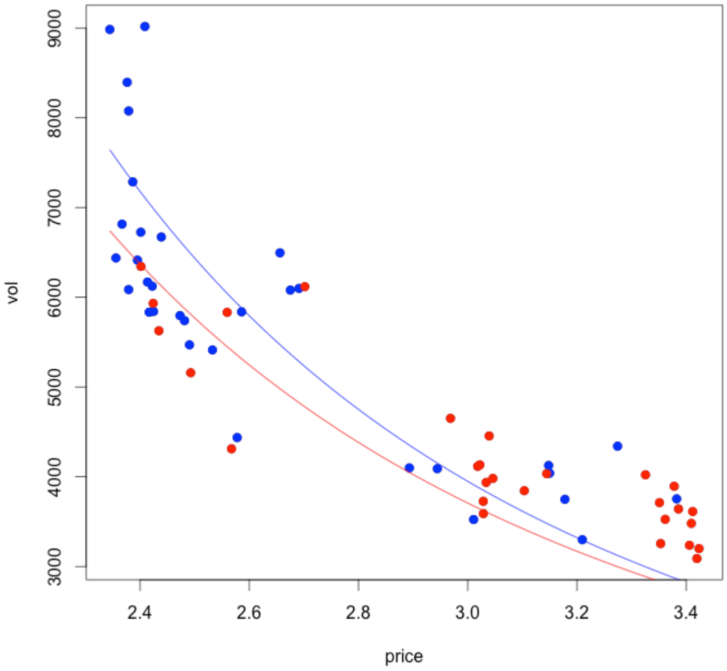
The curve for the in-store display includes the offset for the intercept and the offset for the slope. The curves are farther apart to the left of the plot (lower price), which mirrors real-life consumer trends. For higher prices of cheese on sale, there's not that much of a difference in sales volume. Consumers are more likely to respond to "good" sales—those with low prices. The correct answer to this question is murky regarding whether or not you need the interaction term. The importance is understanding the process.

```
lm4 = lm(log(vol)~log(price) + store + disp, data=cheese)
summary(lm4)
#how much did display improve sales?
exp(.18540)
=1.2037

lm5 = lm(log(vol)~log(price) + store + disp + disp:log(price),
data=cheese)
summary(lm5)
```

The curve equations generated from the model are as follows:

```
curve(exp(9.26095 + 1.62211)*x^(-2.42583), add=TRUE, col='red')
     curve(exp(9.26095 + 1.62211 + 0.34153)*x^(-2.53159 - 0.14761),
                              add=TRUE, col='blue')
```



New material

Cards were distributed in a biased sample. It appeared that there was a systematic correlation between card color and sex. How do we prove beyond a reasonable doubt that the sample is biased? Take a test statistic, such as the sampling distribution of what fraction of red cards went to men. By doing the resampling, you're basically doing a chi-squared test.

Looking at the Georgia vote data (HW6 #3):
Very squished at bottom end, so we do a log transformation. The summary of lm0 gives you a pretty small R-squared. How small? If you think of equipment as like the cards, you see small association between undercount rate and the equipment. But if you assert that there is actually a difference, can you prove it?

What if we can take every county and "deal out the cards" again (randomly assign types of voting)? Shuffling the equipment takes the original equipment possessed by a county and draws type at random.

Bottom line: It is totally implausible that undercount rate is correlated to a shuffled deck of cards. No way that the random process can be correlated with the outcome. To be continued next class.