Dana Hwu
9:30AM

3/26/14 STA371H Class Notes

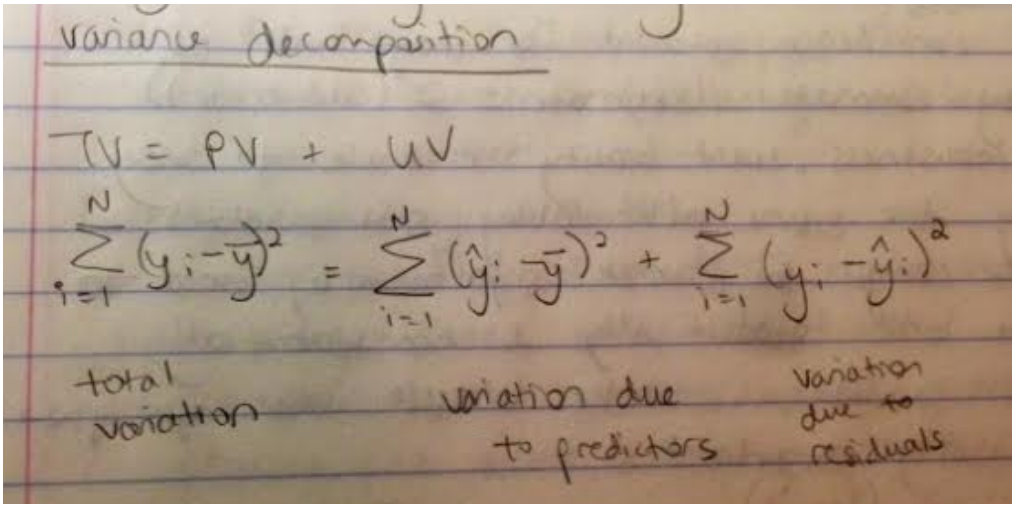
We began by reviewing homework assignment 7 that was due today. The R script is posted on the class website.

- First, you look at various plots to see a general trend of evaluation scores of the professors by each of the different variables. Just by eyeballing, you are able to see which seem to be making a difference in professor ratings.
- The plot for students and allstudents were squished at the bottom, so we can use log transformations for them to make them look more linear.
- Then we look at the summary of the linear model that includes the factors we care about:
    - >lm(eval ~ beauty + minority + gender + division + credits + log(allstudents) + tenure + native, data=profs)
- We can look at the estimate of the slope for beauty to see how beauty is affecting the professor rating
    - For each one point increase in beauty, the rating is increased by 0.16961
- Then we can look at the anova of the linear model
    - We want to focus on the "sum sq" and see which ones are smaller and may not be very contributing to the rating
    - Example is tenure which is only 0.126

```
Analysis of Variance Table

Response: eval
                  Df  Sum Sq Mean Sq F value    Pr(>F)
beauty             1   5.083  5.0830 19.5523 1.226e-05 ***
minority           1   0.979  0.9787  3.7646  0.052966 .
gender             1   3.948  3.9483 15.1874  0.000112 ***
division           1   1.679  1.6792  6.4591  0.011370 *
credits            1   8.934  8.9339 34.3652 8.797e-09 ***
log(allstudents)   1   2.011  2.0108  7.7349  0.005642 **
tenure             1   0.126  0.1261  0.4849  0.486570
native             1   1.452  1.4522  5.5859  0.018526 *
Residuals        454 118.026  0.2600
```

Next, we looked at variance decomposition and reviewed how total variation is composed of predictable and unpredictable variation.

Dana Hwu
9:30AM



While trying to create these models to explain a certain y variable by certain x variables, you have 2 goals to balance:

1. good predictive model
2. simplicity

Essentially, you want large predictability with as few predictors as possible.

Next, we learned how to use the drop1 function in R. This is like a mini version of the backwards step-wise selection. It drops each variable individually and shows the effect on the model.

- Dropping beauty has a large sum of sq of 7.9919 so we can tell that it is most likely a predictive factor for professor evaluations.
- You are trying to create the smallest (most negative) AIC, and it can be seen that by dropping tenure, you can achieve the smallest AIC of -616.60. Additionally, it goes with what we had seen before of its sum sq being small.

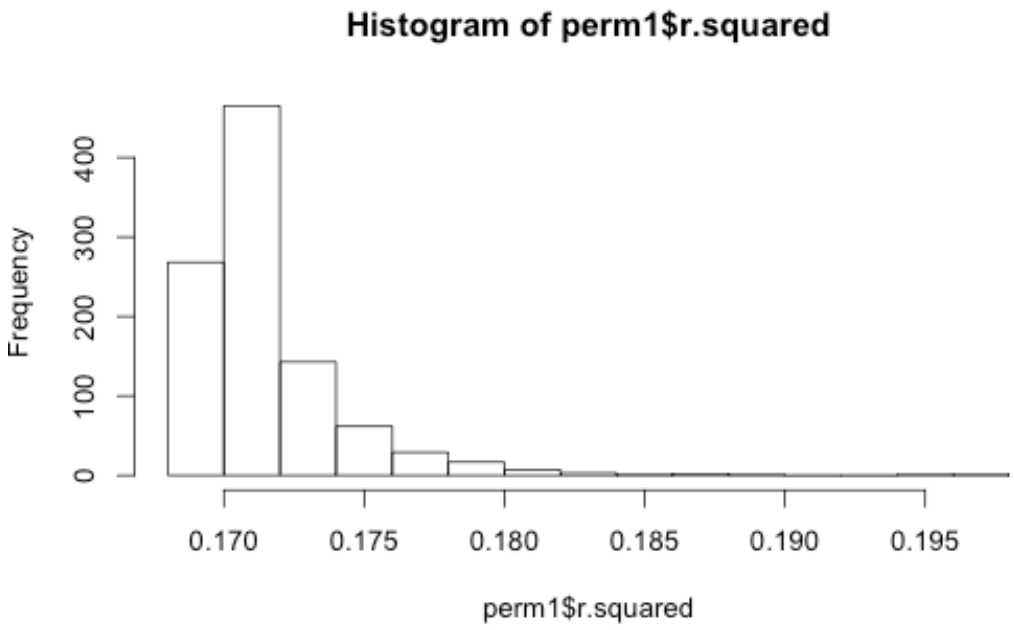
```
Model:
eval ~ beauty + minority + gender + division + credits + log(allstudents) +
    tenure + native
                Df Sum of Sq    RSS     AIC
<none>                      118.03 -614.84
beauty           1    7.9919 126.02 -586.50
minority         1    1.3872 119.41 -611.43
gender           1    3.5124 121.54 -603.26
division         1    0.1320 118.16 -616.32
credits          1    6.0731 124.10 -593.61
log(allstudents) 1    1.9956 120.02 -609.07
tenure           1    0.0606 118.09 -616.60
native           1    1.4522 119.48 -611.17
```

Dana Hwu
9:30AM

Next we want to do a permtest to confirm our decision. We shuffle tenure to see what the r-squared would be like through random chance.

```
>perm1 = do(1000)*lm(eval ~ beauty + minority + gender + division + credits +
log(allstudents) + shuffle(tenure) + native, data=profs)
hist(perm1$r.squared)
```

With our model, the r-squared is 0.1702, which is well within the bulk of the histogram and is definitely not in the rejection region. Therefore, we fail to reject the null hypothesis that tenure is a significant factor, and we can proceed to drop it.



Next, we might consider dropping division.
We can use drop1 again to see the following output. Dropping division makes the AIC the smallest, and the sum of sq is also small.

```
Model:
eval ~ beauty + minority + gender + division + credits + log(allstudents) +
    native
                Df Sum of Sq    RSS     AIC
<none>                      118.09 -616.60
beauty           1    8.0762 126.16 -587.97
minority         1    1.5005 119.59 -612.75
gender           1    3.4548 121.54 -605.25
division         1    0.1461 118.23 -618.03
credits          1    6.9053 124.99 -592.29
log(allstudents) 1    2.1544 120.24 -610.23
native           1    1.5177 119.61 -612.69
```

You can form a confidence interval for the estimate of beauty's partial slope by taking the estimate (0.17160) and adding and subtracting 2 times the std. error (0.03047).

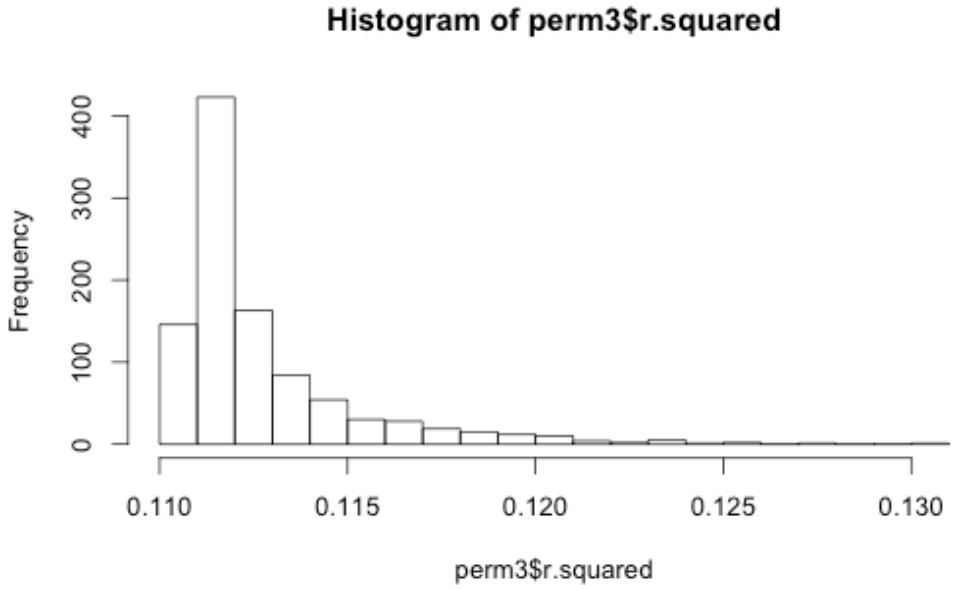
```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.91396    0.14431  27.122  < 2e-16 ***
beauty            0.17160    0.03047   5.632 3.11e-08 ***
minorityyes      -0.17523    0.07523  -2.329  0.02028 *
gendermale        0.18069    0.04881   3.702  0.00024 ***
creditssingle     0.60518    0.10633   5.692 2.25e-08 ***
log(allstudents) -0.07984    0.02861  -2.791  0.00548 **
nativeyes         0.26972    0.10475   2.575  0.01034 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5092 on 456 degrees of freedom
Multiple R-squared:  0.1688,    Adjusted R-squared:  0.1578
F-statistic: 15.43 on 6 and 456 DF,  p-value: 3.894e-16
```

We want robustness, and we want the beauty estimate to stay nearly the same when we take out various variables.

- If the estimate does change significantly, we need to decide both statistically and non-statistically which model is better because what we just did was probably important.

In the linear model summary above, the beauty estimate is 5 times the std error. This is unlikely and therefore, we are probably very sure in this estimate. We will reject the null hypothesis.

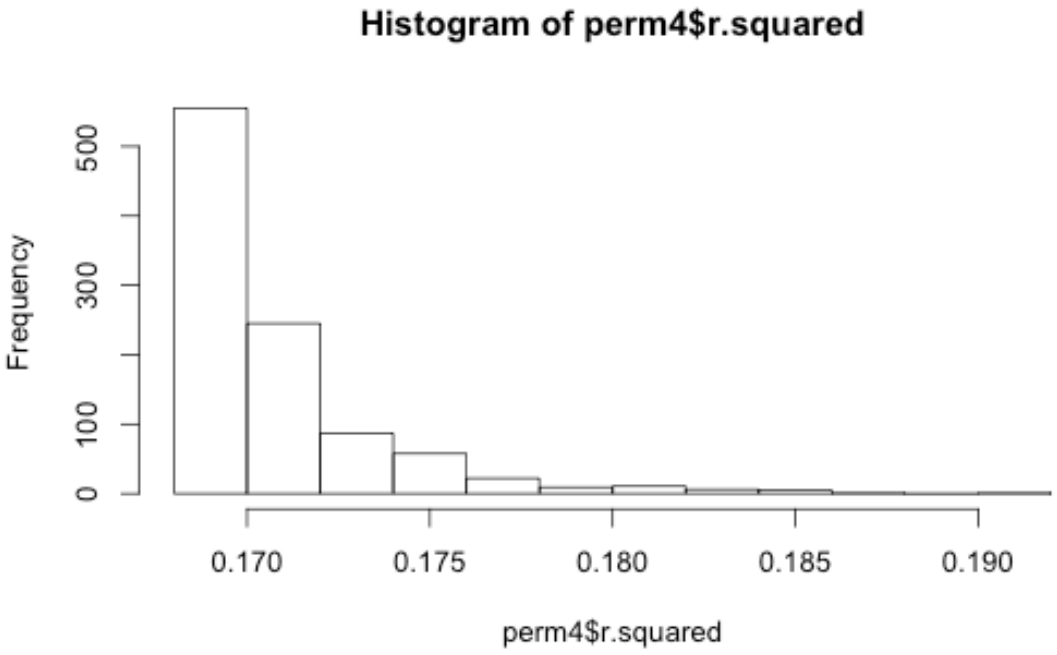
- We do another permtest and our r-squared is 0.1688. This is extremely unlikely on the histogram and is definitely in the rejection region.



Therefore, we can confirm once again that we will reject the null hypothesis, and beauty is a significant predictor for rating professors.

Now we want to check to see if the slope is different for men and women. We did this with interaction terms.

- We make a linear model, do another permtest, and shuffle the interaction term
- The null hypothesis is that we don't need the interaction term
- lm4 = lm(eval ~ beauty + minority + gender + credits + log(allstudents) + native + gender:beauty, data=profs)
- perm4 = do(1000)*lm(eval ~ beauty + minority + gender + credits + log(allstudents) + native + shuffle(gender):beauty, data=profs)
- Then we look at the histogram for the permtest.
    - There is a borderline conclusion because our r-squared is 0.18, and that could or could not be in the rejection region.



When we look at the summary the slope for women is 0.08 and the slope for men is 0.08 + 0.154 because it is written in baseline-offset form. We can see that the advantage of beauty is 3 times better for men than it is for women.

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.91427    0.14349  27.279  < 2e-16 ***
beauty             0.08862    0.04502   1.969  0.04959 *
minorityyes       -0.14319    0.07590  -1.887  0.05984 .
gendermale         0.17969    0.04853   3.702  0.00024 ***
creditssingle      0.61870    0.10586   5.844 9.71e-09 ***
log(allstudents)  -0.08512    0.02852  -2.984  0.00300 **
nativeyes          0.29252    0.10456   2.798  0.00537 **
beauty:gendermale  0.15432    0.06193   2.492  0.01306 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5063 on 455 degrees of freedom
Multiple R-squared:   0.18,     Adjusted R-squared:  0.1673
```

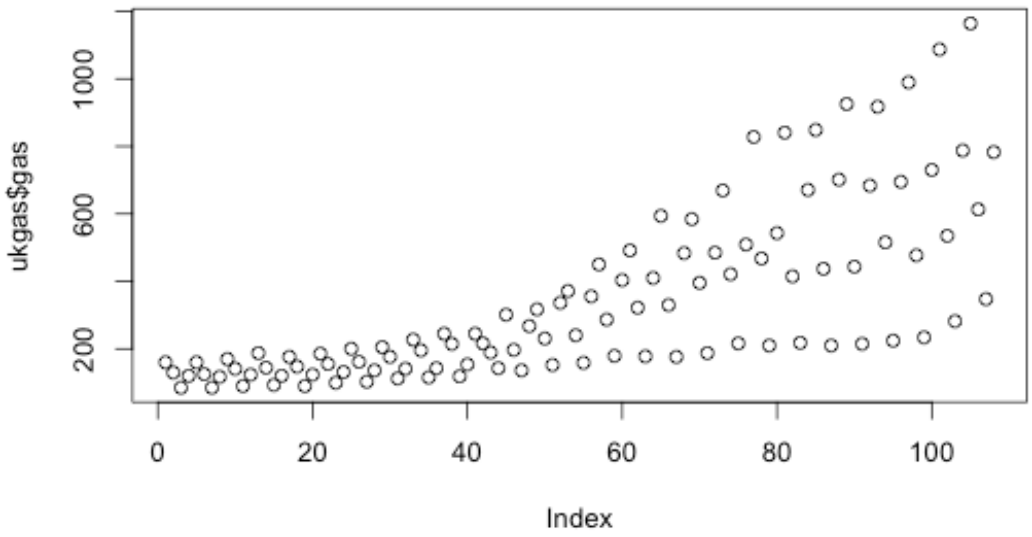
For the last homework question of playing devil's advocate, the main reason is that there is probably not a random selection of students evaluating the professors. There are too many possible biases, and there could be self-selection. Maybe some students like later classes, some like easier classes, or some like classes with more attractive professors.

Next we discussed the ukgas.csv file. The R script is also on the website.
This file relates to "Time Series", which has two parts.

1. trends
    - Strategy: regress on a time index
2. seasonality
    - Strategy: use seasonal dummy variables
    - Example of seasonality is iPhone sales
        - Sales are always greatest in December because of the holiday season and lowest in January because people just bought iPhones
        - Another example is air conditioning

For the ukgas data set, there are a few missing values in the dataset. We scrub them away using the na.omit function.

Dana Hwu
9:30AM



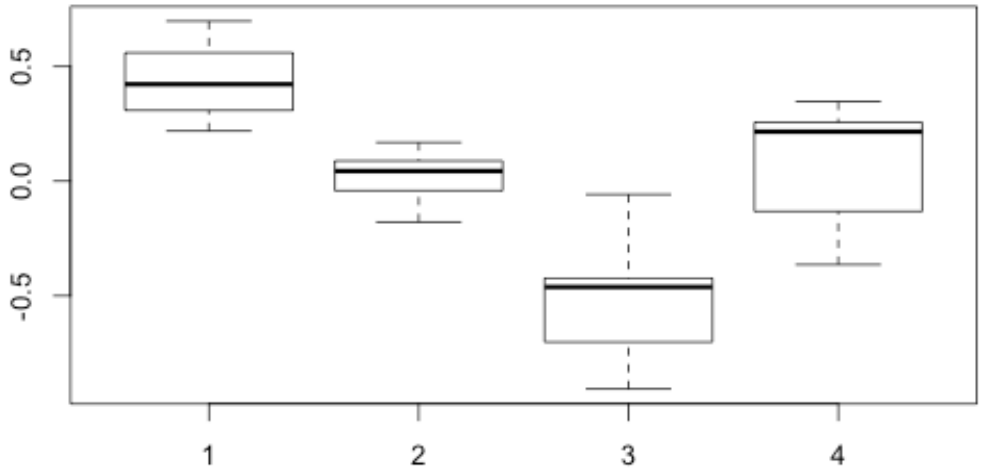
- We also need to use a log transformation to make the data less squished and more linear.
- We regress on a time index by adding a variable that counts each point
    - We add period
    - This makes it so it will not reset the quarters each year but instead continue counting throughout the years

	year	quarter	gas	pop	gdp	period
1	1960	1	160.1	52245758	6.7	1
2	1960	2	129.7	52298644	6.8	2
3	1960	3	84.8	52372000	6.8	3
4	1960	4	120.1	52462669	6.9	4
5	1961	1	160.1	52567497	6.9	5
6	1961	2	124.9	52683325	6.9	6

We find that the good predictors are log(gdp), period, and population. We then drop the trend term (period) because looking at anova, it doesn't change the sum sq very much.

```
Model:
log(gas) ~ period + pop + log(gdp)
          Df Sum of Sq    RSS     AIC
<none>                  16.754 -193.26
period     1  0.010813 16.765 -195.19
pop        1  0.141153 16.895 -194.35
log(gdp)   1  0.305886 17.060 -193.30
```

Next we look at the box plots by quarter and can see obvious seasonal trends.



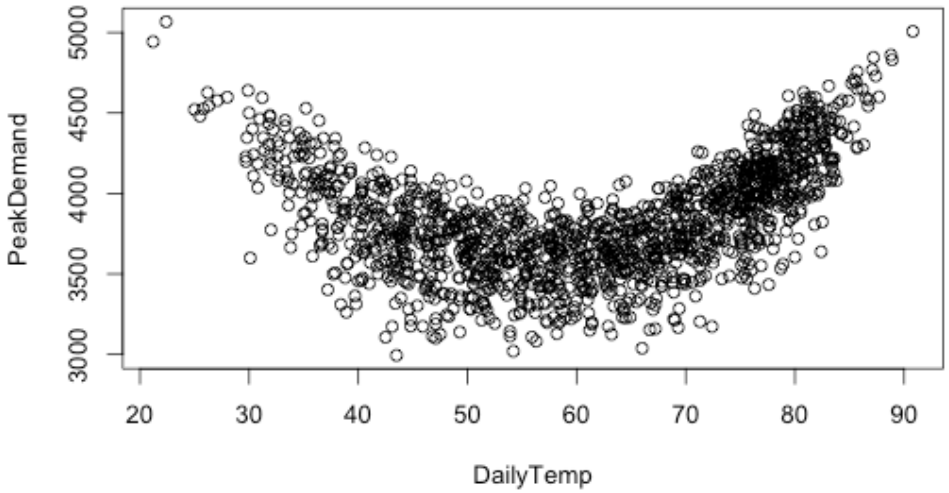
So then we add dummy variables for quarter using the factor function. For the linear model summary, the baseline is quarter one and the offsets are the other quarters.

- >lm5 = lm(log(gas) ~ pop + log(gdp) + factor(quarter), data=ukgas)

The r-squared is 0.9346 and compared to the earlier linear model with r-squared=0.6696, we have done much better and need the time predictor.

Next we started on peakdemand.csv. We were instructed to build a good model for peak demand of gas usage while accounting for time series effects.

- Most people plotted >plot(PeakDemand~DailyTemp, data=peakdemand)
- According to below, you can see that there is more gas usage in the cold and hot times (winter and summer) and less in the spring and fall.



Looking at this plot, we are assuming there is a quadratic relationship because it looks like a parabola.