Tommy Garber
11:00-12:30 AM

Scribe Notes: March 26

Agenda: Homework 7 review and introduction to time series analysis

Administrative Notes

- We are moving the unit on logistic regression to the end of the semester. For the last week of the semester, we will vote whether we want to do more decision analysis or logistic regression.
- For time series analysis, you do not have to know the entire reading that will be posted. Focus on the 2 main ideas covered in class (trends and seasonalities).

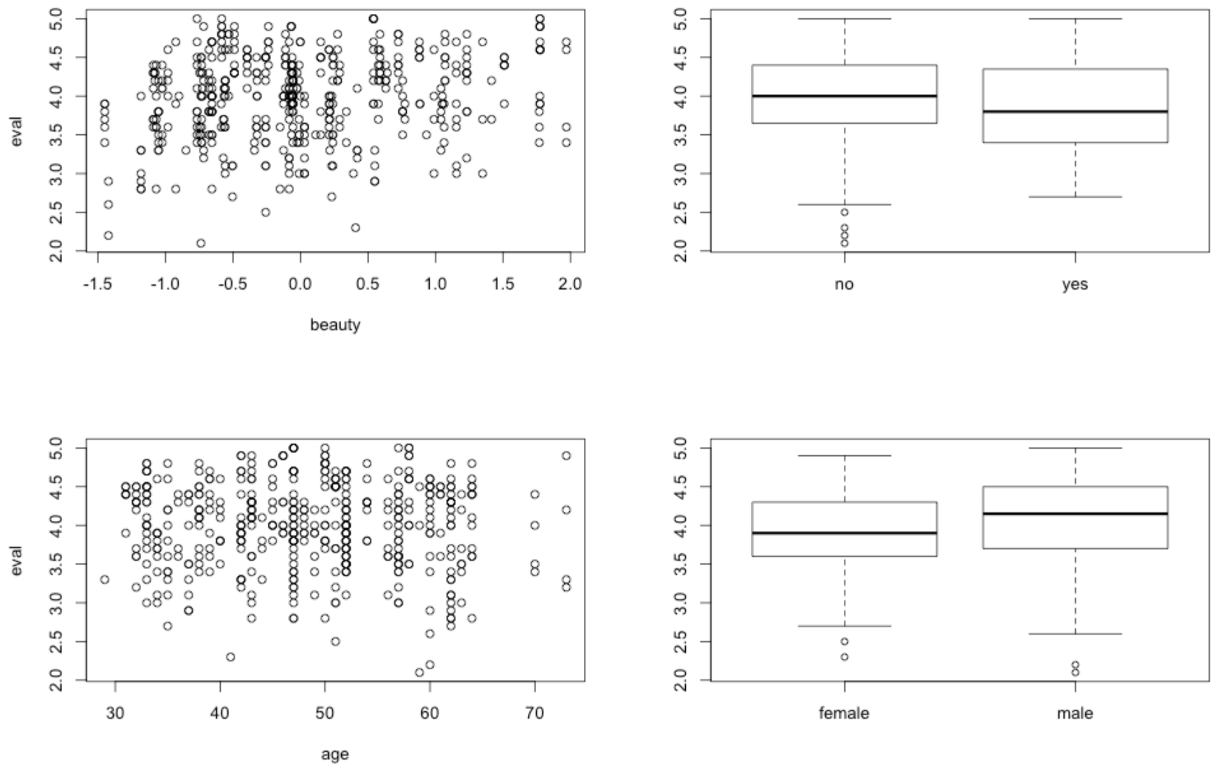
Homework 7 Review

The r script used is hw07.r from the website.

This homework was about whether a professor's attractiveness affects his/her course instructor evaluation score. Different approaches to this problem are OK as long as good statistical principles are implemented.

It is a good idea to start by plotting the data, especially when we have a focused question and are worried about the effects of confounding.

Looked at following 4 plots first: evaluation results vs. beauty, minority, age, gender



- At first glance, appears to be a positive trend between evaluation and beauty
- Appears that minority instructors have systematically lower evaluation results than white professors (top right plot above)
- Tough to see any trend in age plot
- Appears that female instructors systematically have lower evaluation results than males

We then plotted the rest of the variables (each vs. evaluation results) and analyzed them in the same way. We took the log of students and the log of allstudents because those two plots were squished.

From all of these plots, which ones looked like they showed a trend?

Beauty, minority, gender, division, credits, log(allstudents), tenure, and native

Start by looking at basic model with all of these variables:
lm1 = lm(eval ~ beauty + minority + gender + division + credits + log(allstudents) + tenure + native, data=profs)

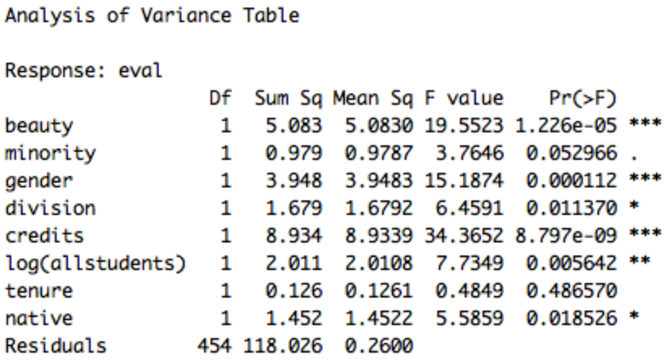
[Note: most students started just by running a stepwise pruning procedure or checked whether there was a significant p-value (and cut variables that did not have significant p-values). Both of these methods are fine, but they are not what Professor Scott did in class. He wanted to be very careful and conservative in deciding whether to include each variable, so he did not trust the "machete" of the stepwise selection method.]

When thinking about adding or deleting a variable from our starting model, we ask 3 questions (the first two are a review from previous classes):

1. Can we estimate that variable precisely (in terms of its effect)?
2. Does that variable improve the predictive ability of the model? (Can measure w/ R^2)
3. Does it make a difference? In this example, if I take out or add a new variable, does it change the effect of the beauty variable on evaluation score? This is called <u>robustness</u>.
    a. This may be the most useful question – the answer determines if you have to do other stuff. (If the variable does not make a difference, who cares?)

Useful new functions: Anova (analysis of variance) and Drop1 (drops 1 variable)

<u>Anova</u> - *anova(lm1)*



```
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                      
```

- Look at 'sums of squares' column in the output. Variables that are likely to be deleted have small sums of squares because of the following reasoning:

Decomposition of Variance

$$\text{TV} \quad = \quad \text{PV} \quad + \quad \text{UV}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Total variation = variation due to the predictors + variation due to residuals

*Anova function tells us which variables contribute most to the predictable variation piece. For example, from the output above, adding the 'minority' variable contributes 0.979 to the beauty model. The bigger this number, the more that variable contributes to the predictable variation. Therefore, natural candidates to delete are variables with small sums of squares.

Delete variables (tenure and division) from the model

- Here, tenure looks like a good candidate for deletion because it does not improve predictions much and (from earlier) had a high standard error. We checked this theory with a permutation test:

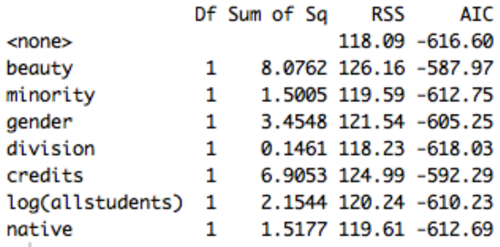
  *perm1 = do(1000)*lm(eval ~ beauty + minority + gender + division + credits + log(allstudents) + shuffle(tenure) + native, data=profs)*
- Following the appropriate steps of hypothesis testing, we failed to reject the null that there is no effect of tenure, so it is safe to exclude tenure from our model.
- Just to be sure, we actually fit the model where we dropped tenure. The coefficient and standard error for beauty did not change much at all, so it is definitely okay to exclude tenure from the model.

  lm2 = lm(eval ~ beauty + minority + gender + division + credits + log(allstudents) + native, data=profs)
- We can do the same procedure with the division variable and drop that from the model too.

Drop1 Function

*Another cool function is the "drop1" function. This function is like a baby version of stepwise selection.

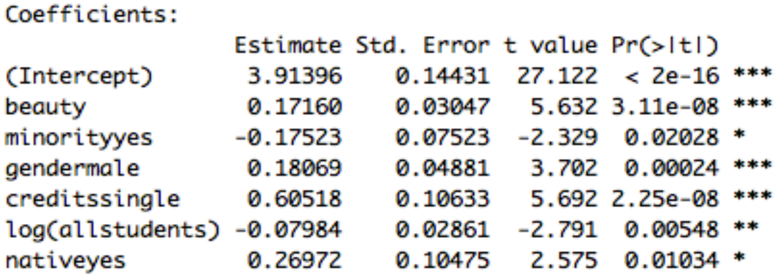
drop1(lm2)



```
                 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
```

- A smaller AIC is better, so we should consider dropping division (most negative AIC).
- Division also has the smallest sums of squares, so dropping division looks good.

We end up with this reasonable model: *lm3 = lm(eval ~ beauty + minority + gender + credits + log(allstudents) + native, data=profs)*



```
Coefficients:
                 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 *  
```

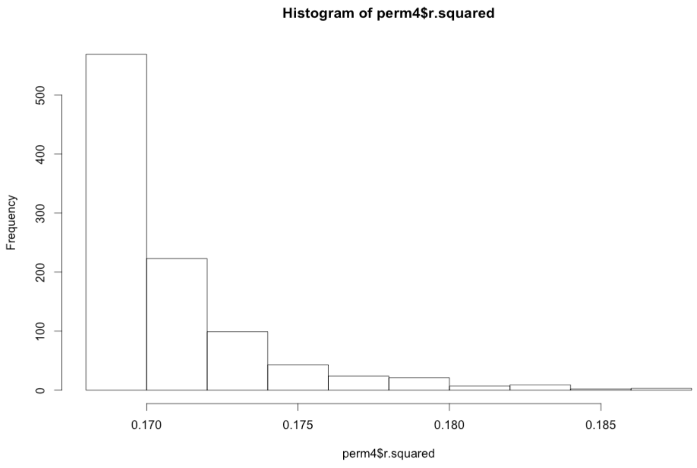
- This output tells us that, when all other factors are held constant, a 1-unit change in the beauty variable changes the y-variable by about 0.17.
  - What does this mean? Looking at the distribution, for example, a change from average beauty to about the 80th percentile in beauty corresponds to a change in survey results from 4.0 to about a 4.2 (not a small effect).

We can assess beauty with a permutation test (didn't do in class). This would show that we reject the null hypothesis that beauty has no effect on CIS results (so beauty does make a difference).

<u>Add interaction terms and do permutation test</u> (for male vs. female question)
We added interaction terms between gender and beauty. As a reminder, this interaction term says that we have different regression lines for men and for women; these lines can be higher/lower and steeper/shallower for the different genders.

We then did a permutation test to shuffle the interaction term.

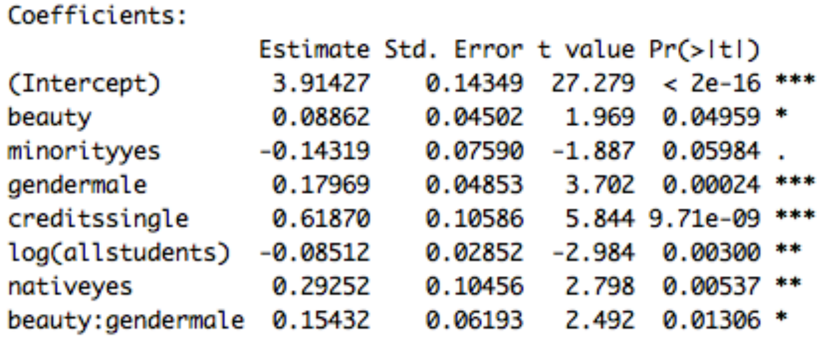
- *perm4 = do(1000)*lm(eval ~ beauty + minority + gender + credits + log(allstudents) + native + shuffle(gender):beauty, data=profs)*
- Our null hypothesis is that there is no interaction between beauty and gender (or, the slope is the same for men and women).
- We looked at the following sampling distribution for R^2 values:



- Set a rejection region at, say, 0.1775. Alpha is about 5%.

- Regressing the actual interaction (not our shuffled one), our R^2 is 0.18. This would technically cause us to reject the null, but it is very close to the rejection region boundary. Thus, we genuinely do not know whether there is a difference between men and women in terms of the slope of the beauty vs. evaluation rating line.

Note: look at the baseline-offset form of our actual model.



```
Coefficients:
                  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 *  
```

- According to this model, the bump (drop) for an attractive (ugly) female professor is not nearly as big as the bump (drop) for an attractive (ugly) male professor.
    - Slope is about 3 times as big for males: About .08 for women (without offset) vs. about 0.23 for men (.08 + offset of .15 =.23)

<u>Confounders</u>

- Think about professor rating websites that judge attractiveness. What if students self-selected classes based on hot professors? This results in systematically different populations of students within each course – actual students that are in these courses are not random. Hot professors have students that are in their class because the professor is hot, so the students are more inclined to like the professor.
- Note: you might want to account for the difficulty of the class – maybe by adjusting for average GPA. However, the beauty ratings were assigned by an impartial jury – not by the people in the class.

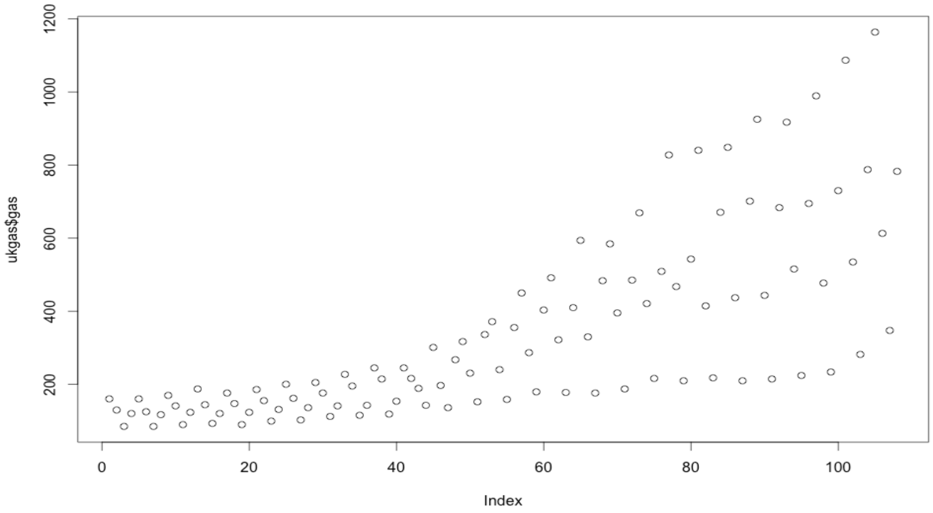
Time-Series Analysis

The r script used is ukgas.r and the data is from ukgas.csv.

This data set is about gas consumption in the United Kingdom. It includes data by year and quarter with the following quantitative variables: gas consumption (in billions of therms), population, and GDP (in tens of billions of dollars).

First, we scrubbed some missing values. (4 quarters at the end of the data set had NA for gas consumption that was not included for those periods.)

We plotted gas consumption over time:



- Clearly gas consumption grows nonlinearly over time.
- The plot fans out.
- The plot is a little squished along the bottom.
- *All 3 of these factors suggest a log transformation.

<u>Time Series Analysis</u>
2 things get you about 75-80% of the way there for time series analysis: Trends and Seasonalities

1. Trends: simple – does stuff go up over time, go down over time, or stay flat over time?
2. Seasonalities
    - Example: which month is air conditioning usage highest/lowest? It is always highest in June-July-August no matter which year. This is a seasonal pattern.
    - Example: which month does Apple sell the most iPhones? December
      Fewest iPhones sold? January This is another seasonal pattern that will hold each year.

Strategies

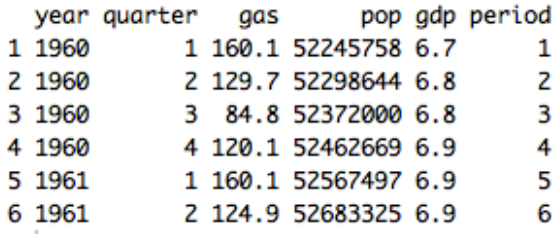
- To deal with trends, regress on a "period index."
- For seasonalities, use seasonal dummy variables.

<u>Regress on a period index</u> (with *ukgas.r*)
Gas consumption is growing over time – how much is it growing? This is all we are doing by regressing on a time index. (How much does something change on average over the period?)

Add a time index – add a variable called "Period" as the last column. It just counts up to show how many periods it has been since the beginning of the data set.

N = nrow(ukgas)
ukgas$period = 1:N

This results in output that looks like this:



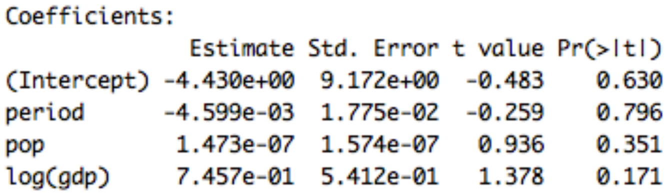
```
  year quarter   gas      pop gdp period
1 1960       1 160.1 52245758 6.7      1
2 1960       2 129.7 52298644 6.8      2
3 1960       3  84.8 52372000 6.8      3
4 1960       4 120.1 52462669 6.9      4
5 1961       1 160.1 52567497 6.9      5
6 1961       2 124.9 52683325 6.9      6
```

Fit a model

- A simple model fitting gas vs. period gives us an intercept of 13.52 and a slope of 5.95. Gas consumption starts at about 13 and grows by about 6 each period.
  - o *lm1 = lm(gas ~ period, data=ukgas)*
- A log transformation makes the residuals look healthier (more constant variance).
  - o *lm2 = lm(log(gas) ~ period, data=ukgas)*

Gas consumption grows over time, but so does the population and economy. Is gas going up once we adjust for population and economy?

- Add population and log(GDP) to the model.
  - o *lm3 = lm(log(gas) ~ period + pop + log(gdp), data=ukgas)*
- *Now we see that there is not much effect from period. On a per population or per factory basis, it appears that gas consumption is not changing much.

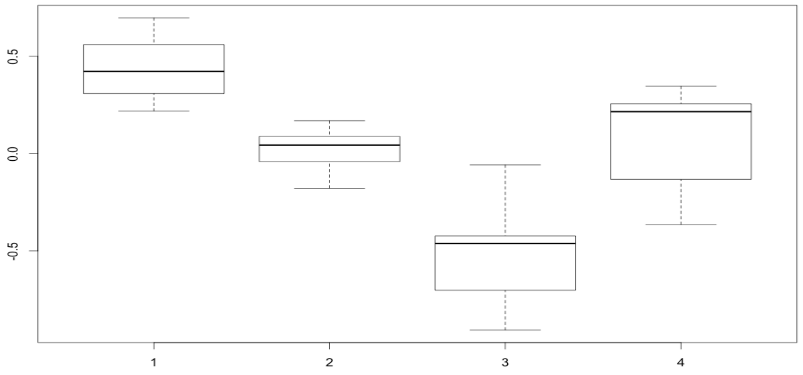


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.430e+00  9.172e+00  -0.483    0.630
period      -4.599e-03  1.775e-02  -0.259    0.796
pop          1.473e-07  1.574e-07   0.936    0.351
log(gdp)     7.457e-01  5.412e-01   1.378    0.171
```

- So we can safely drop the Period term. Now we are left with the following model:
  - o *lm4 = lm(log(gas) ~ pop + log(gdp), data=ukgas)*

Seasonal effects

- There are big seasonal effects in this data set.
- Look at the boxplot for residuals (from the above model) by quarter.



- The first boxplot (January-February-March) is clearly higher than the third boxplot (July-August-September) because people use more gas for heat in the winter months.

How do we deal with seasonal effects? Add seasonal dummy variables

- Think of season as our grouping variable, and include it in the model.
- The following model includes a dummy variable for quarter:
    - *lm(log(gas) ~ pop + log(gdp) + factor(quarter), data=ukgas)*
    - (Use "*factor(quarter)*" to have R treat it as a group instead of as a number.)
- This model has a much higher R^2 (>90% vs. 66% before) than the previous models. We have taken a huge jump in forecasting ability by accounting for seasonal trends.

Wrap-Up

- 2 new lessons: trends (things can change over time) and seasonality (any periodic effect, doesn't necessarily have to be spring/summer/fall/winter - could be a week or a day)

For next class: we will look at peak electricity demand for Raleigh, NC. We will spend 30 minutes next class exploring this data set to close our unit on time series analysis.