Kristi Tamberelli

Class notes 3/19/14

STA371H

Part 1:

We started class out by reading an article on SAP that can be found in the zip file named sap.zip under data on Professor Scott's website. The article claims that SAP users on average are 20% less profitable according to ROE analysis. We were asked to first read the article, and then try to replicate the test mentioned in the article in order to get the same results.

- When we tried to replicate the study, we found that the researchers most likely found the 20% in a way that is too simple for the data, and essentially the wrong way
- The article disregarded the different solutions (CRP, ERP, etc.) and just took the mean for the firms and the mean for the industry and divided the first mean by the second to arrive at the 20% better statistic

```
> mean(sap$roefirm)
[1] 0.1263704
> mean(sap$roeindustry)
[1] 0.1569877
> 0.1263704/0.1569877
[1] 0.8049701
```

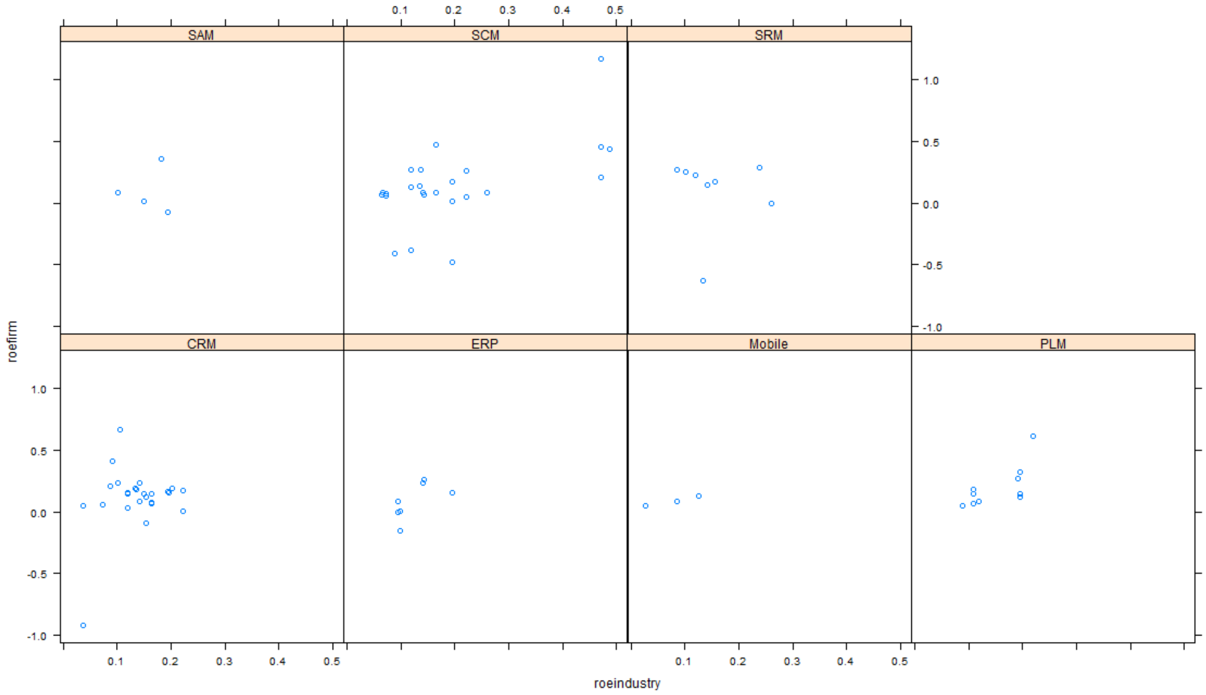
- We then looked for a better way to test the hypothesis that there's no significant difference between the firm and industry ROEs.
- In RStudio:

```
xyplot(roefirm~roeindustry | solution, data=sap)

lm1 = lm(roefirm ~ roeindustry, data=sap)

summary(lm1)
```

  - The summary data showed that our r-squared value is .1795
  - lm1 seems sensible to use in order to see if firm roe and industry roe line up on average (slope of 1)



- We then added in solutions as a predictor, which caused r-squared to increase to .2042

  lm2 = lm(roefirm ~ roeindustry + solution, data=sap)

  summary(lm2)

- Next we performed a permutation test
  - If a difference is due to chance, our p-value would fall somewhere on the sampling distribution
  - We picked an r-squared of somewhere around .325 as our cut off for an alpha of 5%, meaning that anything to the left of .325 on the histogram would not be significant, and anything to the right would be significant
  - We eye-balled the .325 as approximately an alpha of 5%, but you can really just pick whatever you want as your cutoff
  - We got 20% for the actual r-squared, so we couldn't reject that the differences were due to chance, or our data was not statistically significant (maybe there is a difference, but it's small and there's no systematic difference)

Conclusion for this part of class: The research in the article didn't seem very reliable, so we chose this method as a way to address the article's claim and see if there was a difference between SAP firms and industry firms. We found that there was no statistical difference. The method we chose is not the only possible method of proving this.

Part 2:

Model choice: “Which predictors should I include?”

Two goals of regression analysis –

1) Prediction or forecasting
2) Estimating effects (focused questions)
    a. Looking at the effect of a particular x on the y, holding other variables constant

Prediction or forecasting

We looked at googleflu.R as an example of a pure forecasting model

- Googleflu can build up a pretty good idea of the flu activity for a week

In R:

Line 10 - omits missing data and values

Line 13 - the plot’s y axis is activity of flu and the x axis is the week

Line 18 – the “.—week” means include everything not already explicitly named in the formula, except week

We always want a balance of fit and simplicity. The more simple the model, the better you can generalize.

Occam’s Razor

Says we should make models as complicated as they need to be, but no more

- We prefer simple models
- We cut off what is not helpful (include the variables that are useful, and exclude the others)
- Chapter 7 is the related course packet reading material

Example #1:

R^2 (subscript A) = adjusted R^2 = $1 - \frac{UV}{TV}\left(\frac{n-1}{n-1-p}\right)$

Adjusted R^2 can actually go **down** when you add a predictor

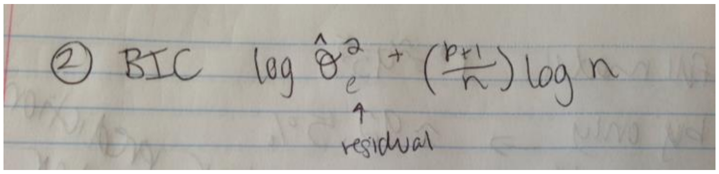
n = number of observations in the dataset

p = number of parameters to estimate

Regular R^2 is NOT good for Occam’s Razor because it doesn’t balance simplicity and fit

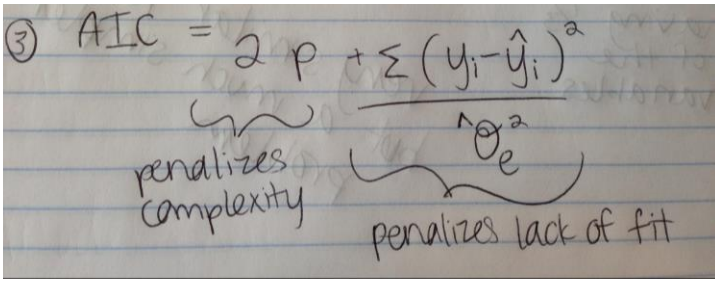
$$R^2 = 1 - \frac{UV}{TV}$$

Example #2:



$$② \ BIC \quad \log \underset{\substack{\uparrow \\ \text{residual}}}{\hat{\theta}_e^2} + \left(\frac{p+1}{n}\right) \log n$$

Example #3:



$$③ \ AIC = \underbrace{2p}_{\text{penalizes complexity}} + \underbrace{\frac{\sum (y_i - \hat{y}_i)^2}{\hat{\theta}_e^2}}_{\text{penalizes lack of fit}}$$

Strategy – Greedy backwards selection

- Start with something large and start pruning from there
  - Prune until you can’t get a better AIC
- The lower the AIC, the better
- So you start with the full equation and do the best possible additions or deletions you can and then you end with a final model that balances simplicity and fit
  - Full model: R^2=93.5%
  - With 1/3 of the variables: R^2=92.5%
    - Very similar predictive power, but a much simpler problem