Big Data and Security

Jeffrey Borowitz, PhD
Lecturer
Sam Nunn School of International Affairs

Generalizations of Parametric Models, Part 2

How Do We Choose Between Models?

- Economists or other social scientists have a role for their theories here
  - If you're trying to understand people's behavior, you have a testable hypothesis
    - E.g. People with more education make more money
- But sometimes theory doesn't really apply: e.g. how many polynomial terms in X do we use?
- In big data applications, we sometimes don't use theory, so then what?
- We use model selection criteria!

Georgia Tech

How Do We Know If We Did A Good Job? R^2

- We did a good job in linear regression if we find low average residuals
  - The residuals (ε) are how much of the outcome variable we can't explain
  - If we have less total residuals, that's better
  - Except:
    - If you imagine explaining wages in yen instead of dollars, yen residuals would be better
    - So a term R^2 is a measure of the size of the residuals, controlling for the fact that units can be different
  - R^2 can go from 0 (no explanatory power) to 1 (perfect explanatory power)
  - There are other criteria too, but we won't discuss this too much

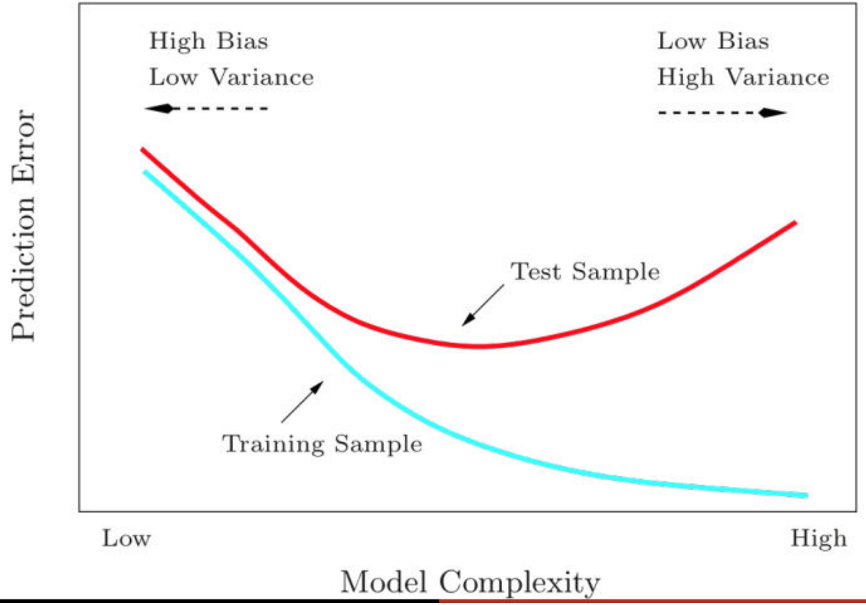
Problem with R^2: Increasing with # of Xs

- If you add more X variables, R^2 can only increase
  - Think about it:
    - You could explain Y just as well if you added another variable but then didn't use it at all, so you can't do worse!
    - If the other variable helps at all, then R^2 will go up.
- So this makes it seem like the more stuff you have in your model, the better
- This isn't right!
- So we find various ways to penalize more complicated models
  - A nice one is called AIC

Stepwise Regression

- Start with some set of Xs
- Fit your model and compute AIC, or another criteria
- Try adding another variable, fitting, and computing AIC
    - If it’s higher, you’re doing better, keep that variable.
    - If not, try adding another
- You can also do this with subtraction
- At the end, you have the “best” model according to this criteria!
- Example: In Google Flu Trends, they added and subtracted combinations of search terms to optimize correlation with flu search variation

Overfitting



Another Solution: Testing/Training

- AIC is a stab at not overfitting your model, which it does by using a theoretical calculation with assumptions about what your parameters are like
- Another way to do this, with less assumption, is to compare models based on a withheld **testing** dataset.
  - Split the data set into **test** and **training**
  - Fit the model on the training data
  - Look at the R^2 or other criteria on the training data
  - Since you didn't use this test data to fit the actual data, you don't have to worry about overfitting
  - Just take the model that does better on the test data

Regularized Regression

- Instead of fitting

$$F(\alpha, \beta) = \sum_i \varepsilon_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$

- We fit

$$F(\alpha, \beta) = \sum_i \varepsilon_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2 - \lambda(|\alpha| + |\beta|)$$

- So the objective function gets smaller if α or β get larger in magnitude
- This is another approach to not overfitting, by explicitly penalizing big coefficients at rate λ

Georgia Tech

Cross Validation

- Cross validation is like using testing and training data, but is more general:
    - Split the data into chunks, and withhold one chunk at a time.
    - Average all the model coefficients
    - The best model is the one that does the best on the withheld data
- Because you are not using all the data when you fit the model each time, you are avoiding overfitting
- If you use k different chunks, it's called k-fold cross validation

So Why Don’t We Always Use Cross-Validation?

- It doesn’t use whatever piece of the data you are withholding, so that’s a waste of data
  - Or, minimizing this problem, there is leave one out cross validation (LOOCV), which requires estimating your model N times
- Compared to thinking theoretically about what should be in a model
  - When have things changed so we need a new model?
  - When can we use our old model?
  - If we have a theory, we can answer those questions
  - Example:
    - Google Flu trends didn’t initially work for H1N1
- It doesn’t actually get around the extrapolation problem
  - Subsets of current data are part of current (not outside) data

Georgia Tech

A Little Intellectual History

- Statistics (the academic discipline)
  - Concerned with the study of different estimators (functions for estimating $\hat{y}$)
  - Are β, α estimated consistently? Under what assumptions?
- Econometrics (This ones is me!)
  - It's a science (or wants to be) (it is!)
  - Science has theory about α, β, which we want to test
  - **Goal is to estimate α, β as well as possible**
- Machine Learning (the empirical discipline of folks in Silicon Valley, quintessential big data types to me)
  - How do we have a non-intelligent machine make a decision, based on example inputs (X) and decisions (Y)
  - **The goal is to predict Y as well as possible**

Georgia Tech

Lesson Summary

- R^2 is an indicator of how well any model fits the data
  - R^2 is a value ranging from 0 to 1
  - The more X variables, R^2 can only increase
  - Stepwise Regression is a method to change the set of X variables to find the best fitting model
- Cross Validation helps you get around the overfitting problem but is not a panacea