CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 43
Comparing Models

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Building a Model

- In general, we strive for parsimony
  - Find the simplest model consistent with the data and our knowledge of the problem
- If a simple model is not good enough, we can
  - Add more predictor variables
  - Add more complex functions of the predictor variables
  - Add interaction terms
- How do we know if the added terms are really helping, or just fitting the noise (overfitting)?
  - R^2 always improves when new model terms are added
  - We need something else to understand overfitting

Coefficient of Determination

- The Coefficient of Determination (R^2) is a measure of how much of the variation in Y is explained by the model

$$Regression\ Sum\ of\ Squares:\quad SSR=\sum(\hat{y}_i-\bar{y})^2$$

$$Error\ Sum\ of\ Squares:\quad SSE=\sum(y_i-\hat{y}_i)^2$$

$$Total\ Sum\ of\ Squares:\quad SSTO=\sum(y_i-\bar{y})^2$$

$$R^2=\frac{SSR}{SSTO}=1-\frac{SSE}{SSTO}\qquad SSTO=SSR+SSE$$

(for linear regression)

Adjusted Coefficient of Determination

- Adjust the SSE and SSTO by their degrees of freedom (p = # of adjustable model parameters)

$$R_a^2=1-\frac{SSE/(n-p)}{SSTO/(n-1)}=1-\frac{MSE}{MSTO}$$

$$R_a^2=R^2-\left(\frac{p-1}{n-p}\right)(1-R^2)$$

- If adding a new model term makes R_a^2 smaller, that term is probably not needed

Information Criteria

- Generic Information Criterion (xIC)

$$xIC=-2\ln(L)+complexity\ term$$

  - L = maximized likelihood, commonly returned by regression software
- We reward lower unexplained variance but penalize greater complexity
  - We try to lower the information criterion value

Log-Likelihood

- For iid normal errors,

$$Likelihood\ L=\left(\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\right)^n exp\left[-\frac{1}{2}\chi^2\right]\qquad \chi^2=\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\sigma_\varepsilon^2}$$

$$-2\ln L=\chi^2+n\ln(2\pi\sigma_\varepsilon^2)$$

- But, $E[\chi^2]=n-p$

$$E[-2\ln L]=n-p+n\ln(2\pi\sigma_\varepsilon^2)$$

Information Criteria

- Akaike Information Criterion (AIC)

$$AIC = -2\,ln(L) + 2p \xleftarrow{} \text{Complexity term}$$

- Log-likelihoods are computed up to an additive constant

$$Example\text{:}\ -2\,ln(L) = n + nln\left(\frac{SSE}{n}\right) + nln(2\pi) + \sum \ln(w_i)$$

- Schwarz's Bayesian Criterion (SBC or BIC)

$$BIC = -2\,ln(L) + p\,\ln(n)$$

Comparing Models

- When comparing models with different numbers of parameters, the "goodness of fit" measure must penalize models with too many parameters

$$R_a^2\ \ vs.AIC\ \ vs.BIC$$

Results in larger p

Most popular choice

Results in smaller p

Lecture 43: What have we learned?

- Why can't R^2 be used to compare models with different number of parameters?
- Explain the adjusted R^2 and how it is used
- What is an "information criterion" and how is it used?
- The use of which information criterion results in the most parsimonious model?