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: $SSR = \sum (\hat{y}_i - \bar{y})^2$
  - Error Sum of Squares: $SSE = \sum (y_i - \hat{y}_i)^2$
  - Total Sum of Squares: $SSTO = \sum (y_i - \bar{y})^2$
  - $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$  (for linear regression)

**Adjusted Coefficient of Determination**

- Adjust the SSE and SSTO by their degrees of freedom ($p = \#$ of adjustable model parameters)
  - $R^2_a = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \frac{MSE}{MSTO}$
  - $R_a^2 = R^2 - \frac{(p-1)}{(n-p)}(1 - R^2)$
- If adding a new model term makes $R_a^2$ smaller, that term is probably not needed

**Information Criteria**

- Generic Information Criterion ($xIC$)
  - $xIC = -2 \ln(L) + \text{complexity term}$
  - $L$ = 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,
  - $L = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[ -\frac{1}{2} \chi^2 \right]$
  - $\chi^2 = \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\sigma^2}$
  - $-2 \ln L = \chi^2 + n\ln(2\pi\sigma^2)$
- But, $E[\chi^2] = n - p$
  - $E[-2 \ln L] = n - p + n\ln(2\pi\sigma^2)$
Information Criteria

• Akaike Information Criterion (AIC)
  \[ AIC = -2 \ln(L) + 2p \]
  Complexity term

• Log-likelihoods are computed up to an additive constant
  Example: \[ -2 \ln(L) = n + n \ln \left( \frac{SSE}{n} \right) + n \ln(2\pi) + \sum \ln(w_i) \]

• 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

  \[ R^2_a \text{ vs. AIC vs. BIC} \]

Results in larger p  Results in smaller p
Most popular choice

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?