STAT 215
Multicollinearity and Model Selection

Colin Reimer Dawson

Oberlin College

November 3, 2017

Outline

So many models...

- How to decide among all these models?
  1. Understand the subject area! Build sensible models.
  2. Nested F-tests
  3. Model quality measures

ASSESS: Coefficient of Determination

Can we use R^2 to identify the best model?

As before, $R^2 = \frac{SS_{Model}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$

What Happens if We Add Useless Predictors?

Handout

What Makes a Good Model?

Fit

High R^2
Small SSE
Large F

Validity

Strong evidence for predictors
Simple (Parsimonious)
Generalizes outside sample

Why Does Parsimony Matter?

Don't we just care about good predictions?

Not exclusively...

- We also use models to *understand* the world (harder with more complexity)

And even so...

- We really care about making predictions for data we *haven't seen yet*.

Balancing Fit and Parsimony

- R^2 can only go up as we add predictors, because at worst, we can choose $\beta_{k+1} = \beta_{k'} = 0$ and get the same SSE. Usually we can pick coefficients to do somewhat better.
- Would like to "penalize" unnecessary predictors.

Adjusted R^2

$$
\begin{aligned}
R^2_{adj} &= 1 - \frac{SS_{Error}/(n-k-1)}{SS_{Total}/(n-1)} \\
&= 1 - \frac{\hat{\sigma}^2_\varepsilon}{s^2_Y} \\
1 - R^2_{adj} &= \frac{1 - R^2}{df_{Error}/df_{Total}}
\end{aligned}
$$

Criteria to "score" models

1. high R^2/low SSE/low $\hat{\sigma}_\varepsilon^2$: always prefers more complex models
2. Adj. R^2: balances fit and complexity
3. Mallow's C_p / Akaike Information Criterion (AIC): estimates mean squared prediction error based on $\hat{\sigma}_\varepsilon^2$ from a "full" model
4. Out-of-sample predictive accuracy

Mallow's C_p / AIC

Two measures that reduce to the same thing in the case of MLR with independent, equal variance, Normal residuals. For a "reduced" model with $p_{reduced}$ total parameters (including the intercept) which is nested in a "full" model with p_{full} parameters, both fit using n observations:

$$C_p = \frac{SSE_{\text{reduced}}}{MSE_{\text{full}}} + 2p_{\text{reduced}} - n \tag{1}$$

$$= p_{\text{reduced}} + \Delta_p \cdot F_{\text{nested}} \tag{2}$$

where $\Delta_p = p_{\text{full}} - p_{\text{reduced}}$ is the number of parameters in the full model that have been left out of the reduced model. Should we prefer larger or smaller values?

Validating on Held Out Data

What data should we use to

(a) Fit the models?

(b) Evaluate the models?

Two answers

1. Use all the data for both (what we've done so far)
2. Separate the data set into distinct "training" and "validation" sets.

In-Sample vs. Out of Sample Prediction

- Idea: A good model should make accurate predictions on data it hasn't seen
- Evaluating in-sample is subject to **overfitting**: Since we try to minimize SSE (and maximize SSM), we are liable to extract too much "signal". Some of the SSM will really be "noise".
- This is particularly likely if we have lots of model df.
- Approaches such as adjusted R^2 and Mallow's C_p try to account for overfitting, but why not actually try to predict on different data than used for fitting?

Cross-Validation

Cross-validation is a technique whereby the full dataset is divided into training and validation (held-out) sets. The first is used for fitting parameters; the second for evaluating predictive power.

Versions:

1. Two-fold: Divide data (randomly) in half. Fit two models, exchanging roles of training and validation.
2. k-fold: Divide data into k equal sized sets, fit k models letting each set as the validation set.
3. Leave-one-out (n-fold): Let each observation be its own validation set. Requires fitting n models.

Practicing Cross-Validation

Lab