POLS/CS&SS 503:
Advanced Quantitative Political Methodology

MODEL SPECIFICATION AND FIT

May 12, 2015

Jeffrey B. Arnold

UNIVERSITY of WASHINGTON
DEPARTMENT OF POLITICAL SCIENCE

CENTER for STATISTICS and the SOCIAL SCIENCES

Overview

- Measures of Fit
  - R^2
  - Standard Error of the Regression
  - Information Criteria
  - Out-of-Sample and Cross-Validation Method
- General Advice on Model Selection

How To Choose Among Different Models?

- Depends on your purpose
- Some tools
  - Internal model validation: residuals, outliers
  - Overall model Fit statistics: out of sample is preferred

Measures of Fit

- R^2
- Standard Error of the Regression
- Information Criteria
- Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

Measures of Model Fit

Various measure of how the model fits the data, both *in-sample* and *out-of-sample*

Measures of Fit

R^2

Standard Error of the Regression

Information Criteria

Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

The Coefficient of Determination, R^2

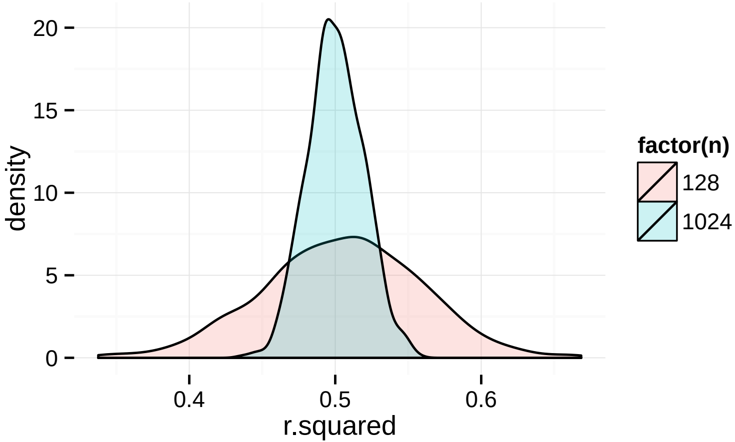
$$R^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}} = 1 - \frac{\text{Residual sum of squares}}{\text{Total sum of squares}}$$
$$= \frac{\sum(\hat{y} - \bar{y})^2}{\sum(\hat{y} - \bar{y})^2}$$
$$= 1 - \frac{\sum \hat{\epsilon}^2}{\sum(\hat{y} - \bar{y})^2}$$

- Commonly used
- Ranges between
- Why can it never be less than 0?
- What happens when you add a variable?
- What is the case when $R^2 = 1$
- Bivariate case: Cor$(y, x)^2$
- General case: Cor$(y, \hat{y})^2$

What R^2 does and doesn't say

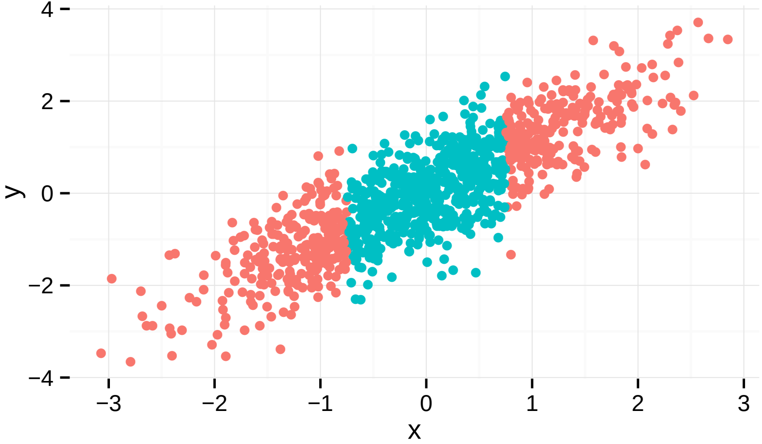
- Indirectly reports scatter around the regression line
- Only *in sample*
- Maximizing R^2 perverse:
  - Not usually interesting for explanation. Y regressed on itself, vote choice on vote intention.
  - Not usually best for prediction
- Not an estimate

Variation in sample R^2



Population $R^2 = 0.5$

R^2 is a function of variation in X



- Complete sample: $R^2 = 0.719$, $\hat{\sigma} = 0.652$
- Complete sample: $R^2 = 0.289$, $\hat{\sigma} = 0.66$

Adjusted R^2

What's adjusted?

$$
\begin{aligned}
\tilde{R}^2 &= 1 - \frac{S_E^2}{S_Y^2} \\
&= 1 - \frac{n-1}{n-k-1} \times \frac{RSS}{TSS}
\end{aligned}
$$

- Unlike R^2, treat squared error terms as estimates of populatio, not sample statistics.
- How does it change with respect to n? With respect to s_j?
- But it is an ad hoc adjustment

Standard Error of the Regression

$$\hat{\sigma} = S_E = \sqrt{\frac{\sum E_i^2}{n-k-1}}$$

- S_E is at least as useful to report as R^2
- S_E is the average error E_i
- On the same scale as y. Substantive significance can be clearer.
- Smaller S_E is better

Measures of Fit

R^2

Standard Error of the Regression

Information Criteria

Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

Likelihood

- Likelihood is the probability of observing the data given a statistical model.
- For a normal linear model, the likelihood is

$$p(y) = \prod_i N(y_i | X_i \beta, \sigma_\epsilon^2) = \prod_i \frac{1}{\sigma_\epsilon} \text{exp} \left(- \frac{(y_i - x_i' \beta)^2}{2\sigma_\epsilon^2} \right) = \prod_i \frac{1}{\sigma_\epsilon \sqrt{2}}$$

- For computational stability (the product of probabilities is a small number), the log likelihood is usually used

$$\text{log}\, p(y) \propto \sum_i \epsilon_i^2$$

- The

Information Criteria

- Information criteria are the Log Likelihod + a penalty for complexity
- The two Most common are AIC and BIC:

$$AIC_j = -2 \log L(\hat{\theta}) + 2k$$
$$BIC_j = -2 \log L(\hat{\theta}) + k \log n$$

- Lower is better
- Smaller values = better fit

Measures of Fit

- R^2
- Standard Error of the Regression
- Information Criteria
- Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

Out of Sample Methods

- Compare models on how well they do on data that was not used to estimate their parameters.
- In practice, serves as a good check against spurious findings
- Even if our goal is explanation, not prediction, scientific models strive for generality
- Usual caveat: best fitting may not be the only criteria for the model

Out of Sample Goodness of Fit

- Method
    1. Split data into training $(X_{\text{training}}, y_{\text{training}})$, test data, $(X_{\text{test}}, y_{\text{test}})$.
    2. Fit model to training data, $(X_{\text{training}}, y_{\text{training}})$, obtain $\hat{\beta}_{\texttt{training}}$
    3. Calcuate fitted $\hat{y}_{\text{test}}$ for the test sample $(X_{\text{test}}, y_{\text{test}})$.
    4. Calculate predicted mean squared error of the **test** data

$$\hat{\sigma}_{\text{test}} = \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} y_i - X_i \hat{\beta}_{\text{training}}$$

- Usually MSE of test data lower than MSE of training data. In-sample fit statistics are overly optimistic.

Cross-Validation

Multipe in-sample

- Method
  1. Split data into training $(X_{\text{training}}, y_{\text{training}})$, test data, $(X_{\text{test}}, y_{\text{test}})$.
  2. Fit model to training data, $(X_{\text{training}}, y_{\text{training}})$, obtain $\hat{\beta}_{\texttt{training}}$
  3. Calcuate fitted $\hat{y}_{\text{test}}$ for the test sample $(X_{\text{test}}, y_{\text{test}})$.
  4. Calculate predicted mean squared error of the **test** data

$$\hat{\sigma}_{\text{test}} = \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} y_i - X_i \hat{\beta}_{\text{training}}$$

- Best model minimizes MSE
- Usually MSE of test data lower than MSE of training data. In-sample fit statistics are overly optimistic.
- Test data should be representative (you can also "overfit" the test data).

Cross Validation

Reuse data for multiple in-sample and out-of-sample tests.

- Method
  1. Select all but $1/k$th of the data: $(y_{\text{training}}, X_{\text{training}})$
  2. Repeat out of sample tests k times
- Usual methods:
  - Leave-one-out (LOO-CV).
  - 5– or 10–fold cross-validation
- Best model minimizes MSE

Measures of Fit

R^2

Standard Error of the Regression

Information Criteria

Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

Fox on Model Selection

Problems

- Simultaneous inference
- Fallacy of affirming the consequent
- Impact of large samples on hypothesis tests
- Exaggerated precision

Fox on Model Selection

Strategies

- Alternative model-selection criteria (not stat sig)
- Compensating for simulaneous inference
- Avoiding model selection: maximally complex and flexible model.
- Model averaging: select many models.

Fox on Model Selection

General Advice

- It is problematic to use stat. hypoth. tests for model selection. Simultaneous inference, biased results. Complicated models in large n, exaggerated prediction. (p. 6008)
- Most methods maximize *predication* not interpretation
- When purpose is interpretation, simplify based on substantive considerations, even if that includes removing small, but stat sig coefficients. (p. 622)
- **validation**: using separate model choice and inference

Gelman and Hill's Rules for Building a Regression Model for Prediction

- Include all input variables expected to be important in predicting outcome (substantively)
- Not always necessary to include these separately, e.g. indices
- For inputs with large effects, consider including interactions
- Whether to exclude a varaible from prediction based on significance
  - Not stat sig, expected sign: keep. Will not help much, but will not hurt predictions.
  - Not stat sig, not expected sign: consider removing
  - Stat sig, not expected sign: **Think hard** Are there lurking variables?
  - Stat sig, expected sign: keep
- Think hard before the model; but adjust to new information
- Gelman and Hill use *predictaion* differently than Fox.

Gelman and Hill, p. 69