

POLS/CS&SS 503:
Advanced Quantitative Political Methodology
MODEL SPECIFICATION AND FIT

May 12, 2015

Jeffrey B. Arnold



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

References

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

References

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

References

The Coefficient of Determination, R^2

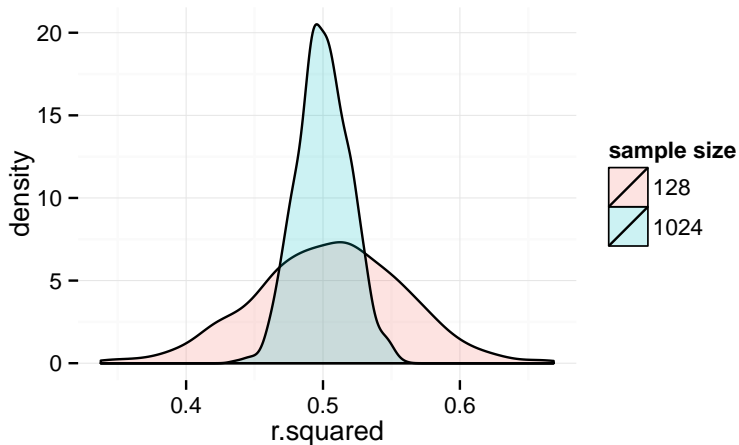
$$\begin{aligned} R^2 &= \frac{\text{Explained sum of squares}}{\text{Total sum of squares}} = 1 - \frac{\text{Residual sum of squares}}{\text{Total sum of squares}} \\ &= \frac{\sum(\hat{y} - \bar{y})^2}{\sum(\hat{y} - \bar{y})^2} \\ &= 1 - \frac{\sum \hat{\epsilon}^2}{\sum(\hat{y} - \bar{y})^2} \end{aligned}$$

- 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: $\text{Cor}(y, x)^2$
- General case: $\text{Cor}(y, \hat{y})^2$

What R^2 does and doesn't say

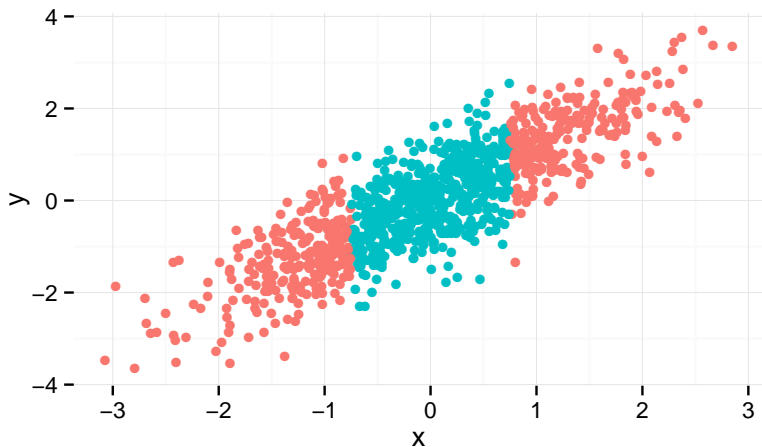
- Indirectly reports scatter around the regression line
- Only *in sample*
- Maximizing R^2 perverse:
 - Not usually interesting for explanation. Y regressed on itself, vote choice on vote intention.
 - Not usually best for prediction
- Not an estimate

R^2 varies between samples



R^2 of samples drawn from a linear model with a population $R^2 = 0.5$.

R^2 is a function of variation in X



- Complete sample (red + blue): $R^2 = 0.72$, $\hat{\sigma} = 0.65$
- Restricted sample (blue only): $R^2 = 0.29$, $\hat{\sigma} = 0.66$

Adjusted R^2

What's adjusted?

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

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

Measures of Fit

$$R^2$$

Standard Error of the Regression

Information Criteria

Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

References

Standard Error of the Regression

The standard error of the regression is the estimate of the population σ :

$$\hat{\sigma}_{\epsilon} = S_E = \sqrt{\frac{\sum E_i^2}{n - k - 1}} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n - k - 1}}$$

- S_E is at least as useful to report as R^2
- S_E : on average, how much does the fitted value miss the actual value.
- On the same scale as y . Easier for interpretation and substantive importance.

Measures of Fit

$$R^2$$

Standard Error of the Regression

Information Criteria

Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

References

Likelihood Function

- Likelihood is the probability of observing the data given a statistical model.
- The **likelihood** of a linear model with normal errors:

$$\begin{aligned}L(\hat{\beta}, \hat{\sigma}_\epsilon) &= p(y|\hat{\beta}, \hat{\sigma}) = \prod_i N(y_i|X_i\hat{\beta}, \hat{\sigma}_\epsilon^2) \\&= \left(\frac{1}{\hat{\sigma}_\epsilon\sqrt{2\pi}}\right)^n \prod_i \exp\left(-\frac{(y_i - x'_i\hat{\beta})^2}{2\hat{\sigma}_\epsilon^2}\right) \\&= \left(\frac{1}{\hat{\sigma}_\epsilon\sqrt{2\pi}}\right)^n \prod_i \exp\left(-\frac{\hat{\epsilon}_i^2}{2\hat{\sigma}_\epsilon^2}\right)\end{aligned}$$

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

$$\log L(\hat{\beta}, \hat{\sigma}_\epsilon) = -n \log \hat{\sigma}_\epsilon - \frac{1}{2} \log 2\pi - \frac{1}{2\hat{\sigma}_\epsilon^2} \sum_i \hat{\epsilon}_i^2$$

Information Criteria

- Information criteria include log Likelihood + a penalty for complexity
- The two Most common are AIC and BIC:

$$AIC = -2 \log L(\hat{\beta}, \hat{\sigma}_\epsilon) + 2k$$

$$BIC = -2 \log L(\hat{\beta}, \hat{\sigma}_\epsilon) + k \log n$$

- Lower is better
- Smaller values = better fit
- See Fox for justifications
- AIC = approx leave one out cross-validation; BIC = a specific k-fold cross-validation

Measures of Fit

R^2

Standard Error of the Regression

Information Criteria

Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

References

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 $\hat{\beta}_{\text{training}}$
3. Calculate fitted \hat{y}_{test} for the test sample $(X_{\text{test}}, y_{\text{test}})$.
4. Calculate predicted mean squared error of the **test** data

$$RMSE_{\text{prediction}} = \hat{\sigma}_{\text{test}} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} \hat{\epsilon}_i^2}$$

- Usually MSE of test data lower than MSE of training data. In-sample fit statistics are overly optimistic.
- Good rule of thumb: 70–75% training, 30–25% test
- Can use other prediction statistics to evaluate models

Cross Validation

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

- k -fold cross validation
 1. Select all but $1/k$ th of the data: $(y_{\text{training}}, X_{\text{training}})$
 2. Repeat out of sample tests k times
- Leave-one-out (LOO-CV): $k = n$.
- 5- or 10-fold cross-validation; generally the best in terms of bias / variance tradeoff.
- The best model minimizes prediction RMSE
- **Important:** the test and training data should be from same “population”. Randomly sampled in cross-section. Need to be careful in panel, blocked, or time-series.

Measures of Fit

$$R^2$$

Standard Error of the Regression

Information Criteria

Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

References

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 simultaneous 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 variable 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 *prediction* differently than Fox.

Gelman and Hill, p. 69

Measures of Fit

$$R^2$$

Standard Error of the Regression

Information Criteria

Out-of-Sample and Cross-Validation Method

General Advice on Model Selection

References

References

- John Fox, *Applied Regression Analysis and Generalized Linear Models*, Ch. 22, “Model Selection, Averaging, and Validation”.
- Christopher Adolph (Spring 2014) “Linear Regression: Specification and Fitting” [Lecture slides].
| <http://faculty.washington.edu/cadolph/503/topic5.pw.pdf> |.