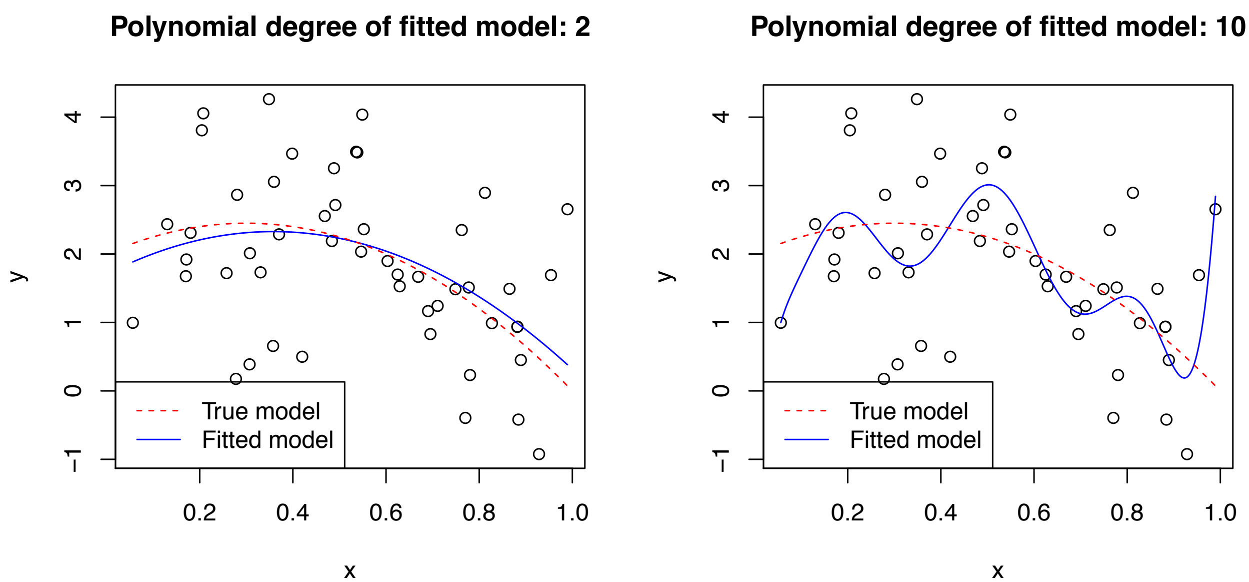
AAE 637 Lab 7: Model Selection[*]

3/18/2015

Model selection is the process of choosing the “best” type of statistical model and the most appropriate set of variables to include in that model. When we don’t have a strong theory to guide our empirical strategy, model selection becomes more art than science.

The problem of overfitting

Ideally, the results of a statistical model reflect the “signal” in the data rather than the “noise”.



[*]prepared by Travis McArthur, UW-Madison (http://www.aae.wisc.edu/tdmcarthur/teaching.asp)

Model selection is closely related to model fit

Theil's adjusted R^2 is:

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

where n is the number of observations and p is the number of parameters.

If one more regressor is added to the regression and the homoskedastic estimate of the standard error is used, $\bar{R}^2$ will increase if and only if the absolute value of t-stat of the coefficient of the regressor is greater than 1. This roughly corresponds to a p-value less than 0.16.

In an OLS framework, $\bar{R}^2$ can be used to help select the best model. Models with a higher $\bar{R}^2$ tend to be "preferred". Warning: systematically trying all combinations of regressors and dropping the regressors with a low t-stat is called "stepwise regression". This procedure has well-known flaws, such as a tendency to overfit the data.

Nested vs. nonnested models

- Model A is **nested** within model B if model A is a special case of model B. This usually means that some restrictions on the values of the parameters of model B can produce model A.
- Two models are **nonnested** if one cannot be represented as a special case of the other.

When we have nested models, we have a large array of tools (likelihood ratio, Wald, Lagrange multiplier tests) to test sharp hypotheses about which model is preferred.

AIC

When two models are nonnested, it is typical to resort to the Akaike information criterion (AIC) or its variants. It is a heuristic (rule of thumb) and does not itself yield hypothesis tests.

$$AIC = -2\ln\left(L\right) + 2k$$

where L is the likelihood value for a model and k is the number of parameters.

When comparing the AIC's of two different models, the model with the *lower* AIC value is the preferred one. Selection based on AIC is logically similar to selection based on $\bar{R}^2$.

The Vuong test

Let $L_{i,0}$ be the likelihood value for the i'th observation of model A.

Let $L_{i,1}$ be the likelihood value for the i'th observation of model B.

Then let $m_i = \ln L_{i,0} - \ln L_{i,1}$

The statistic generated for the Vuong test is:

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(m_i - \bar{m}\right)^2}}$$

In summary, V is $\sqrt{n}$ times the mean of m, divided by its standard deviation: $\sqrt{n}\,(\bar{m}/s_m)$.

Under the null hypothesis, V is distributed standard normal. The test is:

If $V < -1.96$, reject model A in favor of model B at the 5% significance level.

If $V > 1.96$, reject model B in favor of model A at the 5% significance level.

If $-1.96 < V < 1.96$, then we cannot conclude anything.

The Vuong test is discussed in more detail on pages 574-576 of the 7th edition of Greene.