Part 6: Model Selection and Intro to ML

Chris Conlon

March 6, 2021

Applied Econometrics II

AIC/BIC and KLIC

Overview

How many components should we include in our model?

- Too few: under-fitting and large residuals.
- Too many: over-fitting and poor out of sample prediction.

How do we choose?

- X variables.
- Instrumental Variables.

When do we have too much data?

- On the internet!
- Hedonics: What really determines the price of your house?
- Prediction: What really determines loan defaults?
- Consideration Sets: How many products do consumers really choose among on the shelf?
- Which elements of financial filings really matter?

What do people mostly do in practice?

- Regress Y on X with all variables included.
- Drop some variables if they aren't significant?
- Re-run with some things dropped
- Add in some other things that may or may not be significant.

Nested and Non-nested Models

What makes a model nested or non-nested?

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 z_i + \varepsilon_i$$

A nested model can be written as a restricted version of the larger model

- ie: all of the following are nested within the model above

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_i + \beta_3 z_i + \varepsilon_i$$
$$y_i = \beta_0 + \beta_2 w_i + \beta_3 z_i + \varepsilon_i$$

- ie: this model is non-nested (because of s_i)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 z_i + \gamma s_i + \varepsilon_i$$

What we teach undergrads?

Start with sum of squared errors (If you want $\frac{1}{n}$'s imagine them):

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{total sum of squares}} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i(\theta))^2}_{\text{residual sum of squares}} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i(\theta) - \bar{y})^2}_{\text{explained sum of squares}}$$

Let $\dim(\theta) = p$ (the number of parameters).

What we teach undergrads

Three traditional ways to select the number of components in a model:

$$
\begin{aligned}
\overline{R}^2 &= 1 - SSR(p)/TSS - SSR(p)/TSS \cdot \frac{p}{N-p-1} \\
AIC(p) &= \ln\left(\frac{SSR(p)}{N}\right) + (p+1)\frac{2}{N} \\
BIC(p) &= \ln\left(\frac{SSR(p)}{N}\right) + (p+1)\frac{\ln N}{N}
\end{aligned}
$$

These are designed for strictly nested models.

Review AIC/BIC

- AIC tends to select larger models than BIC since it penalizes the number of parameters less heavily.
- These usually depend on ordering potential models by p the number of components and then sequentially fitting them.
- AIC is not consistent: as $N \to \infty$ it may still select too many parameters.
- BIC is consistent: as $N \to \infty$ it will select the correct number of parameters.
- Of course for finite-sample $N < \infty$ anything can happen.

What is KLIC?

Kullback-Leibler information criterion:

$$
\begin{aligned}
KLIC(f,g) &= \int \mathbf{f}(\mathbf{y}) \log\left(\frac{\mathbf{f}(\mathbf{y})}{\mathbf{g}(\mathbf{y})}\right) d\mathbf{y} \\
&= \int f(y)\log(f(y))\partial y - \int f(y)\log(g(y))\partial y \\
&= C_f - \mathbb{E}_f \log(g(y))
\end{aligned}
$$

Observe $KLIC(f,g) \geq 0$ and $KLIC(f,g) = 0$ IFF f, g are the same distribution!
C_f we ignore (doesn't depend on g).

Where does it come from?

How do we come up with these penalized regressions?

- AIC/BIC arise from considering the likelihood ratio test (LRT) of a maximum likelihood estimator and making a lot of assumptions.
- AIC arises from minimizing the Expected KLIC.
- Picking a model with best AIC means picking a model based on (estimated) expected KLIC (if g includes the correct model).
- Low values of KLIC mean the models are similar.

Where does it come from?

How do we come up with these penalized regressions?

- Recall that OLS is a ML estimator in the case where ε is normally distributed.

$$D = -2\ln\left(\frac{\text{Likelihood } H_0}{\text{Likelihood } H_a}\right) = -2\ln\underbrace{\left(\frac{(\sup L(\theta|x) : \theta \in \Theta_0)}{(\sup L(\theta|x) : \theta \in \Theta)}\right)}_{\Lambda(x)}$$

- If the models are nested then $\Theta_0 \subset \Theta$ and $\dim(\Theta) - \dim(\Theta_0) = q$ then as $N \to \infty$ we have that $D \to^d \chi^2(q)$.

Non-nested cases

Many cases we are interested in are not strictly nested

- Should I include x_2 OR x_3 in my regression? (partially overlapping)
- Is the correct distribution $f(y|x, \theta)$ normal or log-normal? (non-overlapping)

Non-nested cases

- Cox (1961) suggested the following (often infeasible solution) by assuming that F_θ is the true model.

$$LR(\hat{\theta}, \hat{\gamma}) = L_f(\hat{\theta}) - L_g(\hat{\gamma}) = \sum_{i=1}^{N} \ln \frac{f(y_i|x_i, \hat{\theta})}{g(y_i|x_i, \hat{\gamma})}$$

- Depending on which the true model is you could reject F_θ for G_γ and vice versa!
- Deriving the test statistic is hard (and specific to F_θ) because we must obtain $E_f[\ln \frac{f(y_i|x_i,\hat{\theta})}{g(y_i|x_i,\hat{\gamma})}]$.
- Similar to AIC in that we are minimizing KLIC over F_θ.

Vuong Test

$$H_0 : E_{h(y|x)}\left[\frac{f(y|x,\theta)}{g(y|x,\gamma)}\right] = 0$$
$$\rightarrow E_h[\ln(h/g)] - E_h[\ln(h/f)] = 0$$

- Instead of taking expectation with respect to one of two distributions, we take it with respect to $h(y|x)$ the unknown but true distribution.
- Same as testing whether two densities (f, g) have same KLIC.
- The main result is that (details in 8.5 of CT):

$$\frac{1}{\sqrt{N}} LR(\hat{\theta}, \hat{\gamma}) \rightarrow^d N[0, \omega_*^2], \quad \omega_*^2 = V_0\left[\ln \frac{f(y|x,\hat{\theta})}{g(y|x,\hat{\gamma})}\right]$$

Model Comparison

- Model selection is not the same thing as significance of β.
- AIC/BIC (even $\overline{R}^2$) compare models based on goodness of fit.
- BIC selects model on highest posterior probability of being the true model.
- AIC selects model that minimizes expected KLIC to the data.
- In practice both assume something like a likelihood and construct a penalty term.