In-Sample vs Out-of-Sample evaluation

In my view, the main difference between an "explanation" and "prediction" approach is the way in which models are evaluated. Think about it, if you want to explain the data then you are implicitly under weighting how important it is for the model to explain new data. But prediction is all about performance on new data, so explaining the sample is less important. It is a question of how important is it for the model to "generalize" to unseen samples.

Traditionally, explanatory models

have been evaluated "in-sample". That is, the metric of interest is how well the model explains the sample. In practical terms, this means that the model is evaluated on the same data it was trained on.

This is often the approach chosen by traditional econometricians and, i think, explains why econometric models usually have very low predictive power. Yet, there are good reasons for this preference. The main one being lack of data.

Out-of-Sample evaluation consists of portioning the dataset into training-testing sets (70-30 is a common split). The idea is to learn

enough from the training set but not so much we over fit. Overfitting happens when the model explains the training data very well (so great explanatory model perhaps! You'll need theory to justify it though) but predicts very poorly on the test set.

Modern data science models usually employs architectures (i.e., models) that have hyperparameters. These are parametes the modeler can manipulate, in contrast with normal parameters which the modeler is trying to estimate. When you hear people "tuning" a model, they probably refer to the process of finding the best values for these

hyperparameters. To test different combinations, a third data set is constructed. Called the Validation set. So a modern out-of-sample evaluation technique (eg. Cross-Validation) usually partitions your data into training, validation, and testing.

You can clearly see that this would require at least a few thousand observations in the full dataset to work well. In Econometrics it is not rare to work with less than 100 observations. Hence we are limited to explanatory models. This has been changing in Microeconomics with high

frequency consumption data, but remains a challenge with macro data.

Because of this feature, the evaluation metrics we use try to approximate out-of-sample performance. Let's see three metrics commonly used:

Adjusted R^2

Recall the normal R^2 metric we saw a while back.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\varepsilon^T \varepsilon}{y^T M_0 y}$$

$\varepsilon = y - X\beta$ and $M_0 = I - \frac{1}{n} 11^T$

↳ Centering matrix

The problem with R^2 is that it will always increase with more variables. It gives a measure of fit but does not account for model complexity.
For this reason, it is best to use its "adjusted" version which penalizes extra variables.

$$\bar{R}^2 = 1 - \frac{n-1}{n-K} \cdot \frac{\varepsilon^T \varepsilon}{y^T M_0 y}$$

Because of the $n-K$ denominator, this metric may decrease if the extra variable does not explain enough more (on the margin) to justify the loss in the degrees of freedom.

It's an intuitive metric (amount of total variance explained) but it falls short. Use it as a starting point but don't rely on it much.

Akaike Information Criteria (AIC)

One big drawback of $\bar{R}^2$ is that the penalty is pretty arbitrary. We are implicitly assuming that variance is a good measure of information. But perhaps this is not the case. There is a whole field that studies information, called information Theory. The main difference is that information is studied as probability distributions, rather than simply variance

explained. Specifically,

(Kullback-Leibler (KL) divergence)

$$D_{KL}(f \mid g) = \int f(y) \log\left(\frac{f(y)}{g(y|\theta)}\right) dy$$

where

$f(y)$ is the true but unknown DGP

$g(y|\theta)$ is the model's proposed DGP

D_{KL} serves as a "distance measure" between the true DGP and a candidate model.

Obviously, we don't know $f(y)$ so we cannot compute D_{KL} directly.

Instead, we can estimate the relative KL divergence between

models.

AIC is one way to do so.

$$AIC = -2 \log L(\hat{\theta}) + 2K$$

- $L(\hat{\theta})$ is the maximized likelihood function
- K is the # of parameters

For linear regression with normal errors:

$$AIC = n \cdot \log\left(\frac{RSS}{n}\right) + 2K + \text{constant}$$

$$= n \cdot \log\left(\frac{\varepsilon^T \varepsilon}{n}\right) + 2K + C$$

The lower the AIC score, the better the model. Also, because it uses the likelihood function, it can

be applied to non-linear models as well.

Bayesian Information Criteria (BIC)

Also called the Schwartz criterion, this is another metric from information theory. Unlike AIC, the BIC imposes a stronger penalty on model complexity. So it favors more persimonious models.

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

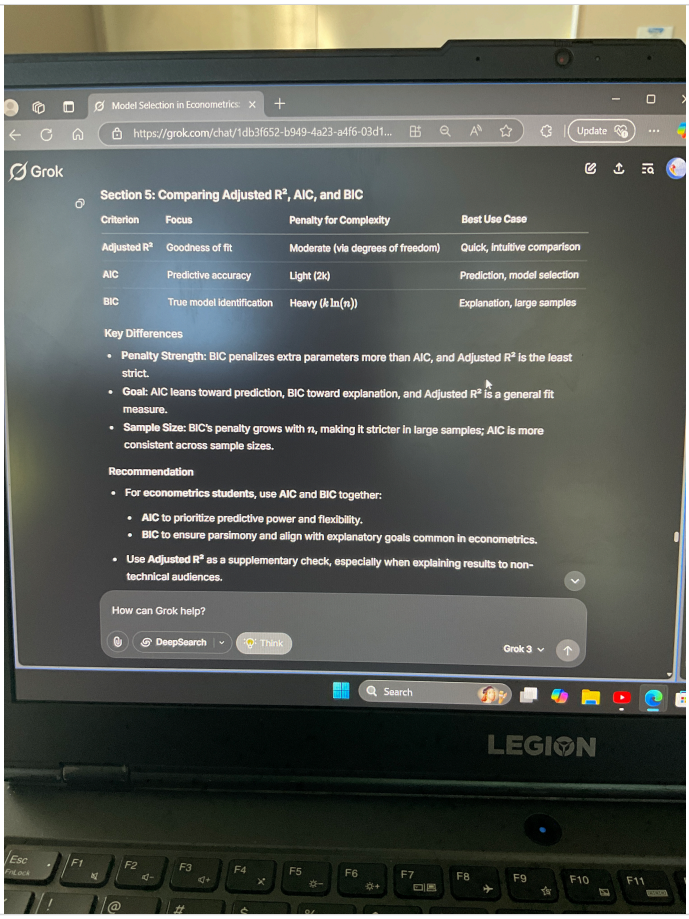
for linear reg. w/normal errors:

$$BIC = n \cdot \log\left(\frac{\varepsilon^T \varepsilon}{n}\right) + k \log(n) + c$$

In practice, use a combination of these to help you choose amongst candidate models.

There are obviously many more out there. But often they are different versions of AIC or BIC.

Here is a nice summary by Grok:



Congratulations! You made it all through ECON 167. We might have an additional lecture on a topic of your interest, but your focus should shift towards the

final project now. We've covered a lot of material. Mastering it cannot be done in a semester, you need practice now. I'm even still learning the details of basic econometrics!

Go craft a project you'll be proud to talk about with colleagues and interviewers A good portfolio of projects can make all the difference when competing for jobs or graduate school positions.

I hope you enjoyed being in class as much as I've enjoyed teaching it and learning from you guys.


STAY STRONG
THE SCHOOL YEAR IS ALMOST OVER!
mememaker.ne