Lecture 6 Model selection criteria

Model selection criteria

When we start adding predictors to the model, a natural question arises of whether these extra predictors add value. We run into an issue of how to compare different models (where models are defined by the predictors included in them). Including more predictors typically improves a model's fit (reduces *variance*), however, the fit could improve even if predictors are truly irrelevant, hence more predictors could increase the *bias* of the model. When the bias is high, taking our model to a new data set could lead to bad performance.

One way to balance this trade-off is to use *model selection criteria*. The two most popular model selection criteria are the Akaike Information Criterion (AIC) and the Bayesian (Schwartz) Information Criterion (BIC). The AIC formula is given by

$$\text{AIC} = 2k + n \ln \frac{SSR}{n} + n \ln(2\pi) + n,$$

where k is the number of estimated coefficients, n is the number of observations, and SSR is the residual sum of squares.

> ⚠ **Warning**
>
> Some formulas use the estimated number of coefficients + 1 for k when computing the AIC for the OLS. The function `AIC` in `R` does that.

When a new predictor is added to the model (k increases)

- the first term $2k$ increases
- the second term $n \ln \frac{SSR}{n}$ (weakly) decreases
- the third term $n \ln(2\pi) + n$ remains constant

This leads to a trade-off. One should include an additional predictor if the corresponding decrease in the first term is larger than the increase in the second term. When we compare different models, we pick the model with the *smallest* value of the AIC.

> ⚠ **Warning**
>
> When computing the AIC for the OLS models, some prefer to omit the constant term $n \ln(2\pi) + n$. This omission does not affect the comparison between the models but does change the absolute value of the AIC. The function `extractAIC` in `R` does that. It also uses the number of coefficients for k, not the number of coefficients + 1.

The formula for the BIC slightly differs from the AIC in the first term:

$$\text{BIC} = k \ln n + n \ln \frac{SSR}{n} + n \ln(2\pi) + n.$$

The BIC penalizes extra predictors more than the AIC, hence the values of the BIC are typically higher than the AIC for a given model.

Info

In case you are wondering where the $n \ln(2\pi) + n$ term comes from, the formulas for the AIC and the BIC are actually derived for models estimated using the Maximum Likelihood Estimation method. A model estimated using this method has a *likelihood*, L. The original AIC formula is

$$\text{AIC} = 2k - 2\ln(L).$$

However, for the linear models, one can show that the logarithm of the likelihood can be written in terms of SSR as follows:

$$\ln L = -\frac{n}{2} \ln \frac{SSR}{n} - \frac{n}{2} \ln(2\pi) - \frac{n}{2},$$

which, after substitution, yields the formula in the beginning.

When using the information criteria, there are a few of points worth keeping in mind. First, the criteria are used to compare different models, they do not tell you about a model's fit. You need a measure like R-squared for this purpose. Second, you should not compare two models that use different transformations of the outcome variable. For example, you should not compare a model where the outcome variable is untransformed with the model where the outcome variable is logged. At least, not directly, there are ways to correct for this. E.g., see `R` function `performance_aic` from the `performance` package. Third, it is a good idea to check both criteria, however, they do not always give the same results.

Example

Let's go back to the trade data and compare four different models. The first model will be our very first simple linear regression

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + u_i.$$

The second model will be the so-called gravity model that adds distance as a predictor

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + u_i.$$

The third model will add another predictor to the gravity model: the degree of a country's liberalization:

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + \beta_3 liberal + u_i.$$

Finally, the fourth model will add a country's area to the previous model

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + \beta_3 liberal + \beta_4 \ln(area) + u_i.$$

The table below shows the estimation results from each model and shows the R-squared, adjusted R-squared, the AIC and the BIC.

	simple	gravity	add liberal	add area
(Intercept)	-5.785	4.670	2.774	2.451
	(2.199)	(2.181)	(2.183)	(2.132)
log(gdp)	1.078	0.976	0.941	1.030
	(0.088)	(0.064)	(0.062)	(0.077)
log(distance)		-1.075	-0.973	-0.888
		(0.157)	(0.153)	(0.156)
liberal			0.497	0.333
			(0.193)	(0.207)
log(area)				-0.159
				(0.086)
Num.Obs.	48	48	48	48
AIC	166.333	134.103	129.337	127.634
BIC	171.947	141.588	138.693	138.861
R2	0.767	0.886	0.901	0.908
R2 Adj.	0.762	0.881	0.894	0.900

Looking at R-squared, we see that it always increases with the inclusion of additional predictors. The adjusted R-squared also increases, although it's values are lower than the unadjusted R-squared. The AIC decreases with the inclusion of additional predictors. The BIC, however, is minimized for model 3 ("add liberal"). Notice that the BIC for a given model is higher than the AIC.

Gauss-Markov theorem