5.3 Selecting predictors

When there are many possible predictors, we need some strategy to select the best predictors to use in a regression model.

A common approach that is *not recommended* is to plot the forecast variable against a particular predictor and if it shows no noticeable relationship, drop it. This is invalid because it is not always possible to see the relationship from a scatterplot, especially when the effect of other predictors has not been accounted for.

Another common approach which is also invalid is to do a multiple linear regression on all the predictors and disregard all variables whose p-values are greater than 0.05. To start with, statistical significance does not always indicate predictive value. Even if forecasting is not the goal, this is not a good strategy because the p-values can be misleading when two or more predictors are correlated with each other (see Section 5/7).

Instead, we will use a measure of predictive accuracy. Five such measures are

Book information

About this book
Feedback on this book

Buy a printed copy

Rob J Hyndman
George Athanasopoulos

Forecasting: principles and practice

- Getting started

introduced in this section.

Adjusted R^2

Computer output for regression will always give the R^2 value, discussed in Section 5/1. However, it is not a good measure of the predictive ability of a model. Imagine a model which produces forecasts that are exactly 20% of the actual values. In that case, the R^2 value would be 1 (indicating perfect correlation), but the forecasts are not very close to the actual values.

In addition, R^2 does not allow for "degrees of freedom". Adding *any* variable tends to increase the value of R^2, even if that variable is irrelevant. For these reasons, forecasters should not use R^2 to determine whether a model will give good predictions.

An equivalent idea is to select the model which gives the minimum sum of squared errors (SSE), given by

$$\text{SSE} = \sum_{i=1}^{N} e_i^2.$$

Minimizing the SSE is equivalent to maximizing R^2 and will always choose the model with the most variables, and so is not a valid way of selecting predictors.

An alternative, designed to overcome these problems, is the adjusted R^2 (also called "R-bar-squared"):

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1},$$

where N is the number of observations and k is the number of predictors. This

is an improvement on R^2 as it will no longer increase with each added predictor. Using this measure, the best model will be the one with the largest value of $\bar{R}^2$.

Maximizing $\bar{R}^2$ is equivalent to minimizing the following estimate of the variance of the forecast errors:

$$\hat{\sigma}^2 = \frac{\text{SSE}}{N-k-1}.$$

Maximizing $\bar{R}^2$ works quite well as a method of selecting predictors, although it does tend to err on the side of selecting too many predictors.

Cross-validation

As discussed in Section 2/5, cross-validation is a very useful way of determining the predictive ability of a model. In general, leave-one-out cross-validation for regression can be carried out using the following steps.

1. Remove observation i from the data set, and fit the model using the remaining data. Then compute the error ($e_i^* = y_i - \hat{y}_i$) for the omitted observation. (This is not the same as the residual because the ith observation was not used in estimating the value of $\hat{y}_i$.)
2. Repeat step 1 for $i = 1, \ldots, N$.
3. Compute the MSE from $e_1^*, \ldots, e_N^*$. We shall call this the CV.

For many forecasting models, this is a time-consuming procedure, but for regression there are very fast methods of calculating CV so it takes no longer than fitting one model to the full data set. The equation for computing CV is given in Section 5/5.

Under this criterion, the best model is the one with the smallest value of CV.

Akaike's Information Criterion

A closely-related method is Akaike's Information Criterion, which we define as

$$\text{AIC} = N \log\left(\frac{\text{SSE}}{N}\right) + 2(k+2),$$

where N is the number of observations used for estimation and k is the number of predictors in the model. Different computer packages use slightly different definitions for the AIC, although they should all lead to the same model being selected. The $k+2$ part of the equation occurs because there are $k+2$ parameters in the model --- the k coefficients for the predictors, the intercept and the variance of the residuals.

The model with the minimum value of the AIC is often the best model for forecasting. For large values of N, minimizing the AIC is equivalent to minimizing the CV value.

Corrected Akaike's Information Criterion

For small values of N, the AIC tends to select too many predictors, and so a bias-corrected version of the AIC has been developed.

$$\text{AIC}_\text{c} = \text{AIC} + \frac{2(k+2)(k+3)}{N-k-3}.$$

As with the AIC, the AICc should be minimized.

Schwarz Bayesian Information Criterion

A related measure is Schwarz's Bayesian Information Criterion (known as SBIC, BIC or SC):

$$\text{BIC} = N\log\left(\frac{\text{SSE}}{N}\right) + (k+2)\log(N).$$

As with the AIC, minimizing the BIC is intended to give the best model. The model chosen by BIC is either the same as that chosen by AIC, or one with fewer terms. This is because the BIC penalizes the number of parameters more heavily than the AIC. For large values of N, minimizing BIC is similar to leave-v-out cross-validation when $v = N[1 - 1/(\log(N) - 1)]$.

Many statisticians like to use BIC because it has the feature that if there is a true underlying model, then with enough data the BIC will select that model. However, in reality there is rarely if ever a true underlying model, and even if there was a true underlying model, selecting that model will not necessarily give the best forecasts (because the parameter estimates may not be accurate).

R code

```
# To obtain all these measures in R, use
CV(fit)
```

Example: credit scores (continued)

In the credit scores regression model we used four predictors. Now we can check if all four predictors are actually useful, or whether we can drop one or more of them. With four predictors, there are $2^4 = 16$ possible models. All 16 models were fitted, and the results are summarized in the table below. An X indicates that the variable was included in the model. The best models are given at the top of the table, and the worst models are at the bottom of the table.

Savings	Income	Address	Employ.	CV	AIC	AICc	BIC	Adj R^2
X	X	X	X	104.7	2325.8	2325.9	2351.1	0.4658
X	X	X		106.5	2334.1	2334.2	2355.1	0.4558
X		X	X	107.7	2339.8	2339.9	2360.9	0.4495
X		X		109.7	2349.3	2349.3	2366.1	0.4379
X	X		X	112.2	2360.4	2360.6	2381.5	0.4263
X			X	115.1	2373.4	2373.5	2390.3	0.4101
X	X			116.1	2377.7	2377.8	2394.6	0.4050
X				119.5	2392.1	2392.2	2404.8	0.3864
	X	X	X	164.2	2551.6	2551.7	2572.7	0.1592
	X	X		164.9	2553.8	2553.9	2570.7	0.1538
	X		X	176.1	2586.7	2586.8	2603.6	0.0963
		X	X	177.5	2591.4	2591.5	2608.3	0.0877
		X		178.6	2594.6	2594.6	2607.2	0.0801
	X			179.1	2595.3	2595.3	2607.9	0.0788
			X	190.0	2625.3	2625.4	2638.0	0.0217
				193.8	2635.3	2635.3	2643.7	0.0000

The model with all four predictors has the lowest CV, AIC, AICc and BIC values and the highest $\bar{R}^2$ value. So in this case, all the measures of predictive accuracy indicate the same "best" model which is the one that includes all four predictors. The different measures do not always lead to the same model being selected.

Best subset regression

Where possible, all potential regression models can be fitted (as was done in the above example) and the best one selected based on one of the measures discussed here. This is known as "best subsets" regression or "all possible

- Using R
- Resources
- Reviews

subsets" regression.

It is recommended that one of CV, AIC or AICc be used for this purpose. If the value of N is large enough, they will all lead to the same model. Most software packages will at least produce AIC, although CV and AICc will be more accurate for smaller values of N.

While $\bar{R}^2$ is very widely used, and has been around longer than the other measures, its tendency to select too many variables makes it less suitable for forecasting than either CV, AIC or AICc. Also, the tendency of BIC to select too few variables makes it less suitable for forecasting than either CV, AIC or AICc.

Stepwise regression

If there are a large number of predictors, it is not possible to fit all possible models. For example, 40 predictors leads to $2^{40} > 1$ trillion possible models! Consequently, a strategy is required to limit the number of models to be explored.

An approach that works quite well is **backwards stepwise regression**:

- Start with the model containing all potential predictors.
- Try subtracting one predictor at a time. Keep the model if it improves the measure of predictive accuracy.
- Iterate until no further improvement.

It is important to realise that a stepwise approach is not guaranteed to lead to the best possible model. But it almost always leads to a good model.

If the number of potential predictors is too large, then this backwards stepwise regression will not work and the starting model will need to use only a subset of

potential predictors. In this case, an extra step needs to be inserted in which predictors are also added one at a time, with the model being retained if it improves the measure of predictive accuracy.

‹ 5.2 Some useful predictors up 5.4 Residual diagnostics ›