Model Selection

Al Nosedal
University of Toronto

March 2, 2019

Background

A number of situations will be presented in which several possible models are proposed to fit a set of data. In these situations, both hypothesis testing and information criteria can be used for model selection. Both hypothesis testing and information criteria are rooted in the same basic idea: a more complex model will generally fit the data better than a simple model, so that the complexity of models and how well they fit the data must be balanced in making a decision about which model best fits the data.

Hypothesis tests in Simple Regression

Consider the model $y_j = \beta_0 + \beta_1 x_j + \epsilon_j$, with independent normal errors. The most common test in this situation is $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$. In other words, is there evidence of a linear trend in the plot x vs y?

Hypothesis tests in Simple Regression (cont.)

For example, consider a case where a line produced the data and both the usual t-test and the likelihood ratio test are used to verify this.

Example

Let us simulate 50 observations for a simple regression with normal errors and standard deviation 4.

```
set.seed(9999);

sigma_1<- 4;

n_1<- 50;

err_1<- rnorm(n_1, 0, sigma_1);

x_1<- c(1:n_1);

y_1<- 3.1 - 0.5*x_1 + err_1;
```

Example

```
# fitting data using lm( );

fit_1<-lm(y_1~ x_1);
```

Example

```
# fitting data using lm( );

summary(fit_1)$coef;

##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  6.1108561 1.03954580   5.878391 3.843339e-07
## x_1         -0.6105909 0.03547917 -17.209842 3.607873e-22
```

Example

There is overwhelming evidence (p-value $\approx 3.607873e-22$) a line fits the data compared to a simple mean (the null hypothesis).

Example

A second approach, one that generalizes to more complex situations, is the likelihood ratio test. Nested models are models in which one simpler model is a special case of a more general model. For example, in the case of regression the simpler model (reduced model) is a simple mean and the more complex model (full model) is a line. The reduced model can be viewed as a special case of the full model in that the reduced model is the full model when $\beta_1 = 0$.

Example

Let the full model have sum of squared residuals SSE_F and p_F free parameters, while the reduced model has sum of squared residuals SSE_R and free parameters p_R ($< p_F$). The p-value for a likelihood ratio test for such models, in the normal setting, is based on the claim that, asymptotically, when the null hypothesis is true: $\chi^2_{p_F - p_R} \sim nln(SSE_R) - nln(SSE_F)$. In other words, the p-value for the likelihood ratio test can be found by comparing the computed test statistic to a chi-square distribution whose degrees of freedom are the difference in the number of free parameters in each model.

Example

```
## SST;

SST<-(50-1)*var(y_1);

SST;

## [1] 4511.136

## SSE;

SSE<-sum(fit_1$residuals^2);

SSE;

## [1] 629.1341
```

Example

```
anova(fit_1);

## Analysis of Variance Table
##
## Response: y_1
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x_1        1 3882.0  3882.0  296.18 < 2.2e-16 ***
## Residuals 48  629.1    13.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Example (Table)

Model	SSE	Complexity (p)
One mean (R)	4511.1357567	1
Line (F)	629.1340703	2

Example (Table)

Our table is a tabular display of all candidate models with their fit (SSE) and complexity (p). The test statistic for the likelihood ratio test, based on information in our table and the sample size, is
50 ln (4511.1357567) - 50 ln (629.1340703) = 98.4979925.
In this case, with a chi-square of just one degree of freedom (2-1), the p-value is roughly 0. Obviously, there is overwhelming evidence that a line is a better fit to the data than a simple mean.

Information Criteria

There is an approach to model selection that does not require that models be nested. This approach considers only the complexity of the model and the size of the residual sums of squares. Although the theoretical derivations are quite sophisticated, these approaches can be understood in simple terms - find an optimal way to balance complexity and fit. In the special case of normal error models, the criteria are defined as

$$\text{Akaike's information} : AIC = nln(SSE) + 2(p+1) + C,$$

$$\text{Schwarz information} : BIC = nln(SSE) + pln(n) + C,$$

Information Criteria (cont.)

where C is an arbitrary constant - the numerical values are not intrinsically meaningful, but the differences are, so the fact that different authors have added various constants, which are ignored here, is not critical. In each case, the goal is to minimize the criterion.
(Note. A technical point is that all parameter estimates should be maximum likelihood estimates. When other estimates are used, the user should be aware that the AIC and BIC values are approximate).

Example (another table)

Model	SSE	Complexity (p)	AIC	BIC
(R)	4511.1357567	1	424.7152116	424.6272346
(F)	629.1340703	2	328.2172191	330.0412651

In this case, both methods pick the straight line by a wide margin over a single mean model.