Model selection

Kerby Shedden

Department of Statistics, University of Michigan

November 8, 2021

Background

Suppose we observe data y and are considering a family of models f_θ that may approximately describe how y was generated.

If we are mainly interested in the individual model parameters, we will focus on how close $\hat{\theta}_j$ is to θ_j (e.g. in terms of its bias, variance, MSE).

Alternatively, our focus may be on the probability distribution f that is the data-generating model for y. In this case, we are more interested in whether $f_{\hat{\theta}}$ approximates f, rather than in whether $\hat{\theta}_j$ approximates θ_j.

Background

The term model model selection is used to describe statistical estimation in a context when the focus is more on the structure of the fitted model than on the individual parameters.

Model selection is a form of statistical inference, but in model selection, we are often contrasting two families of models $\{f_\theta\}$ and $\{g_\eta\}$. Often, the dimensions of η and θ will be different.

In contrast, when we do parameter estimation we are selecting a model from within one family $\{f_\theta\}$, where θ has a fixed dimension and usually lies in a simple domain like $\mathcal{R}^p$.

Model complexity and parsimony

The discrepancy between f_θ and $f_{\hat{\theta}}$ is strongly influenced by how complex of a model we decide to fit.

Suppose we have $p = 30$ covariates and $n = 50$ observations. We could consider the following two alternatives:

1. We could fit a model using all of the covariates. In this case, $\hat{\theta}$ is unbiased for θ (in a linear model fit using OLS). But $\hat{\theta}$ has very high variance.
2. We could fit a model using only the five strongest effects. In this case, $\hat{\theta}$ will be biased for θ, but it will have lower variance (compared to the estimate including all covariates).

If our goal is for $f_{\hat{\theta}}$ and f_θ to be close, either approach 1 or approach 2 could perform better, depending on the circumstances.

Assessing model fit

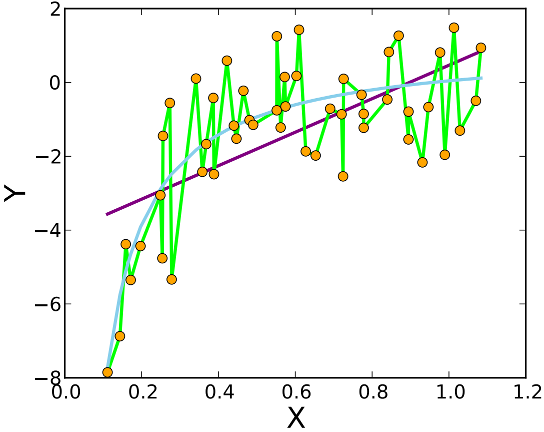
A more complex model will usually fit the data better than a more parsimonious (simpler) model. This is called overfitting.

Due to overfitting, we cannot simply compare models in terms of how well they fit the data (e.g. in terms of the residual sum of squares, or the height of the likelihood). The more complex model will always appear to be better if we do this.

To overcome this, most model selection procedures balance a measure of a model's fit with a measure of its complexity. To justify selecting the more complex model, it must fit the data substantially better than the simpler model.

Fit and parsimony

Example: The purple, green, and blue curves below are estimates of $E[y|x]$. The green line fits the data better but is more complex. Which estimate is closest to the truth?



F-tests

Suppose we are comparing two nested families of models $\mathcal{F}_1 \subset \mathcal{F}_2$, both of which are linear subspaces. An F-test can be used to select between $\mathcal{F}_1$ and $\mathcal{F}_2$.

Mallows' C_p

Suppose we postulate the model

$$y = X\beta + \epsilon$$

but in fact $E[y] \notin \mathrm{col}(X)$. We'll continue to assume that the homoscedastic variance structure $\mathrm{cov}(\epsilon|X) = \sigma^2 I$ holds. Denote this model as M.

Denote the error in estimating $E[y]$ under model M as

$$D_M = \hat{y}_M - E[y],$$

where $\hat{y}_M$ is the projection of y onto $\mathrm{col}(X)$.

Mallows' C_p

Write

$$E[y] = \theta_X + \theta_X^\perp,$$

where $\theta_X \in \mathrm{col}(X)$ and $\theta_X^\perp \in \mathrm{col}(X)^\perp$. Since $y = \theta_X + \theta_X^\perp + \epsilon$, it follows that $\hat{y} = \theta_X + \epsilon_X$, where ϵ_X is the projection of ϵ onto $\mathrm{col}(X)$.

Therefore

$$\begin{aligned} ED_M D_M' &= E(\hat{y}_M - Ey)(\hat{y}_M - Ey)' \\ &= E(\epsilon_X - \theta_X^\perp)(\epsilon_X - \theta_X^\perp)' \\ &= \theta_X^\perp \theta_X^{\perp\prime} + \sigma^2 P_X \end{aligned}$$

where P_X is the projection matrix onto $\mathrm{col}(X)$.

Mallows' C_p

Taking the trace of both sides, yields

$$E\|D_M\|^2 = \|\theta_X^\perp\|^2 + (p+1)\sigma^2,$$

where $p+1$ is the rank of P_X.

Mallows' C_p aims to estimate

$$C_p^* = E\|D_M\|^2/\sigma^2 = \|\theta_X^\perp\|^2/\sigma^2 + p + 1$$

The model that minimizes C_p^* is the closest to the true model in this particular sense.

Mallows' C_p

We need an estimate of C_p^*.

To begin, we can derive the expected value of

$$\hat{\sigma}^2 = \|y - \hat{y}_M\|^2/(n-p-1)$$

in the case where $E[y]$ is not necessarily in $\mathrm{col}(X)$:

$$\begin{aligned} E\hat{\sigma}^2 &= E[y'(I-P)y/(n-p-1)] \\ &= E[(\theta_X + \theta_X^\perp + \epsilon)'(I-P)(\theta_X + \theta_X^\perp + \epsilon)/(n-p-1)] \\ &= E[\mathrm{tr}(I-P)(\theta_X^\perp + \epsilon)(\theta_X^\perp + \epsilon)'/(n-p-1)] \\ &= \|\theta_X^\perp\|^2/(n-p-1) + \sigma^2. \end{aligned}$$

Mallows' C_p

Now suppose we have an unbiased estimate of σ^2. This could come from a regression against a much larger design matrix that is thought to contain $E[y]$. Call this estimate σ^{*2}. Then

$$(n-p-1)E(\hat{\sigma}^2 - \sigma^{*2}) = \|\theta_X^\perp\|^2.$$

Therefore we can estimate C_p^* using

$$C_p = (n-p-1)(\hat{\sigma}^2 - \sigma^{*2})/\sigma^{*2} + p + 1.$$

The model M with the smallest value of C_p is selected.

AIC

Suppose we are selecting from a family of linear models with design matrices X_M, for $M \in \mathcal{M}$.

For each X_M, the model parameters (slopes and error variance) can be estimated using least squares (and method of moments for the error variance) as a vector $\hat{\theta}_M$. This allows us to construct a predictive density:

$$p(y; X_M, \hat{\theta}_M).$$

AIC

The Kullback-Leibler divergence ("KL-divergence") between the predictive density and the actual density $p(y)$ is

$$E_y \log\left(\frac{p(y)}{p(y; X_M, \hat{\theta}_M)}\right) = \int \log\left(\frac{p(y)}{p(y; X_m, \hat{\theta}_M)}\right) p(y) dy \geq 0.$$

Here we are considering $\hat{\theta}_M$ to be fixed.

Small values of the KL divergence indicate that the predictive density is close to the actual density.

AIC

Akaike's Information Criterion (AIC) aims to estimate the KL divergence between a candidate model and the data-generating model $p(y)$ unbiasedly. We can then select the candidate model that has the smallest estimated KL-divergence relative to $p(y)$.

The KL-divergence can be written

$$E_y \log(p(y)) - E_y \log(p(y; \hat{\theta}_M, X_M).$$

We can ignore the first term since it doesn't depend on M. Thus it will be equivalent to select the model that maximizes the predictive log likelihood:

$$E_y \log(p(y; \hat{\theta}_M, X_M)) = \int \log(p(y; X_M, \hat{\theta}_M))p(y)dy.$$

AIC

The predictive log likelihood is the expected value of

$$\log p(y^*; X_M, \hat{\theta}_M(y))$$

taken over the joint distribution of y and y^*, which are independent copies of the data.

The parameter estimates $\hat{\theta}_M = \hat{\theta}_M(y)$ are determined from y, which you can think of as a "training set", and the log-likelihood is evaluated at y^*, using $\hat{\theta}_M(y)$ to set the parameters.

AIC

Since we don't have both y and y^*, it is natural to use the plug-in estimator of the predictive log likelihood:

$$\log p(y; X_M, \hat{\theta}_M(y))$$

But this is biased upward, due to overfitting.

Surprisingly, this upward bias can be shown to be approximately equal to the dimension of M, which is $p+1$ for regression ($p+2$ if you count σ^2).

AIC

Thus we may take

$$\log(p(y_{\text{train}}; X_M, \hat{\theta}_M)) - p - 1$$

as a model selection statistic to be maximized (commonly this is multiplied by -2, in which case it is to be minimized). This quantity is the AIC.

AIC in linear models

To apply the AIC to linear models, we assume the error values ϵ are multivariate normal, so the log-likelihood becomes

$$-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|y - X\beta\|^2.$$

If we work with the profile likelihood over β, we get $-n\log(\hat{\sigma}^2)/2$ (plus a constant). Therefore maximizing the AIC is equivalent to maximizing

$$\underbrace{-n\log(\hat{\sigma}^2)}_{\text{fit}} \quad - \quad \underbrace{2(p+1)}_{\text{complexity}}$$

The conventional form of the AIC is a scaled version of the expression above, which is to be minimized:

$$\mathrm{AIC} = n\log(\hat{\sigma}^2) + 2(p+1).$$

AIC and likelihood ratios

The AIC does not require the models being compared to be nested, but let's consider this special case.

Let L_1 and L_0 be the maximized log likelihoods for two nested models (so $L_1 \geq L_0$).

We know that $2(L_1 - L_0)$ approximately follows a χ^2_q distribution, where q is the difference between the number of parameters of the two models.

If $q = 1$, we select the larger model if $L_1 - L_0 \geq 1.92$ (the usual likelihood ratio test at the 0.05 type I error rate).

If the additional parameters are not needed, then $E[L_1 - L_0] = 0.5$ (so 0.5 is the lowest possible threshold for $L_1 - L_0$ that could ever be considered).

Under AIC, we select the larger model if $L_1 - L_0 > 1$ (less strict than the likelihod ratio test).

Bayesian Information Criterion (BIC)

A different criterion that we will not derive here is the "Bayesian information criterion" (BIC). The BIC defines complexity differently from the AIC:

$$\underbrace{-n\log(\hat{\sigma}^2)}_{\text{fit}} \quad - \quad \underbrace{(p+1)\log(n)}_{\text{complexity}}$$

The conventional definition of the BIC is $n\log(\hat{\sigma}^2) + (p+1)\log(n)$. The best-fitting model under the BIC minimizes this quantity.

The complexity penalty in BIC, $\log(n)(p+1)$, will always be larger than the corresponding AIC penalty, which is $2(p+1)$. Thus the BIC will always favor simpler models than the AIC.

Model selection based on prediction

Many approaches to model selection attempt to identify the model that predicts best on independent data.

If independent "training" and "test" sets are available, for each model M the parameters of M can be fit using the training data, yielding $\hat{\theta}_M$. Predictions can then be made on the test set

$$\hat{y}_{\mathrm{M,test}} = X_{\mathrm{M,test}}\hat{\theta}_M$$

and the quality of prediction can be assessed, for example, using the Mean Squared Prediction Error (MSPE):

$$\|y_{\mathrm{test}} - \hat{y}_{\mathrm{M,test}}\|^2/n.$$

Cross-validation

Separate training and test sets are usually not available. Cross validation is a direct method for obtaining unbiased estimates of the prediction mean squared error when only training data are available.

In k-fold cross validation, the data are partitioned into k disjoint subsets ("folds"), denoted $S_1 \cup \cdots \cup S_k = \{1, \ldots, n\}$.

Let $\hat{\beta}_j$ be the fitted coefficients omitting the j^{th} of these subsets, and let

$$\mathrm{CV}_k = n^{-1} \sum_{j=1}^{k} \sum_{i \in S_j} (y_i - X_i' \hat{\beta}_j)^2$$

This is an approximately unbiased (but potentially very imprecise) estimate of the MSPE on a test set.

The special case of leave one out cross validation (LOOCV) is when $k = n$.

Cross-validation

For OLS regression, CV_n (also known as "prediction residual error sum of squares", or PRESS), can be computed rapidly:

$$\mathrm{CV}_n = n^{-1} \sum_i R_i^2/(1 - P_{ii})^2.$$

The generalized cross-validatation (GCV) criterion replaces P_{ii} with the average diagonal element of P, which is $\mathrm{trace}(P)/n$:

$$\mathrm{GCV}_n = n^{-1} \sum_i R_i^2/(1 - \mathrm{tr}(P)/n)^2 = n^{-1} \frac{\|y - \hat{y}\|^2}{(1 - \mathrm{tr}(P)/n)^2}.$$