

# Model Selection

Suppose we have two families of models,  $\mathcal{F}_1$  and  $\mathcal{F}_2$ .

We can use maximum likelihood or another procedure to identify the best fitting models  $f_1 \in \mathcal{F}_1$  and  $f_2 \in \mathcal{F}_2$  to our data.

But how do we decide between  $f_1$  and  $f_2$ ?

This is the problem of model selection.

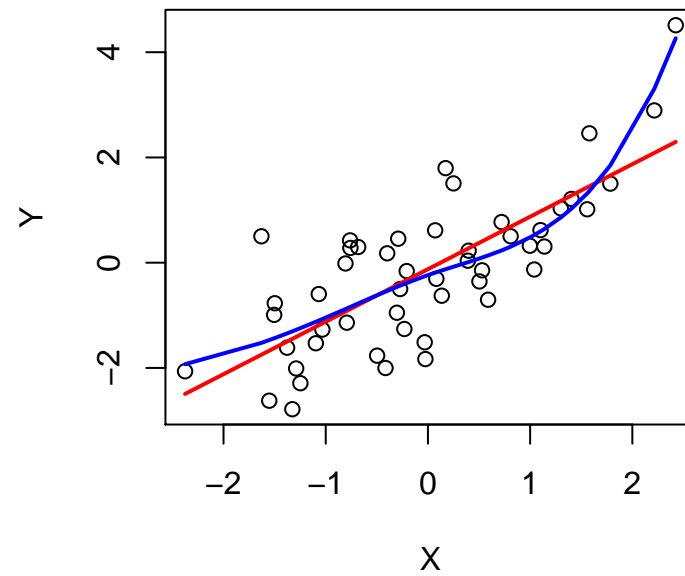
## **Model complexity and parsimony**

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

Due to “overfitting,” this will happen even if the simpler model is closer to the true model.

Therefore model selection must balance a measure of a model’s fit with a measure of its complexity.

The red and blue curves in this figure are estimates of  $E(Y|X)$ . The blue line fits better but is more complex. Which is closer to the truth?



## **F-tests**

If  $\mathcal{F}_1 \subset \mathcal{F}_2$  are nested and are both linear subspaces, then 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  $EY \notin \text{col}(X)$ . We'll continue to assume that the variance structure  $\text{cov}(\epsilon|X) = \sigma^2 I$  holds.

Denote the error in estimating  $EY$  by

$$D = \hat{Y} - EY,$$

where  $\hat{Y}$  is the usual projection of  $Y$  onto  $\text{col}(X)$ .

## Mallows' $C_p$ (continued)

Write

$$EY = \theta_X + \theta_X^\perp,$$

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

Therefore

$$\begin{aligned} EDD' &= E(\hat{Y} - EY)(\hat{Y} - EY)' \\ &= E(\epsilon_X - \theta_X^\perp)(\epsilon_X - \theta_X^\perp)' \\ &= \theta_X^\perp \theta_X^{\perp'} + \sigma^2 P_X \end{aligned}$$

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

## Mallows' $C_p$ (continued)

Taking the trace of both sides, yields

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

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

Mallows'  $C_p$  aims to estimate

$$C_p^* = E\|D\|^2/\sigma^2 = \underbrace{\|\theta_X^\perp\|^2/\sigma^2}_{\text{fit}} + \underbrace{p + 1}_{\text{complexity}}$$

The model that minimizes  $C_p$  has the strongest evidence for being close to the data generating model.

## Mallows' $C_p$ (continued)

We need an estimate of  $C_p^*$ . We can derive the expected value of  $\hat{\sigma}^2 = \|Y - \hat{Y}\|^2 / (n - p - 1)$  when  $EY \notin \text{col}(X)$ :

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

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

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

Therefore we can estimate  $C_p^*$  using

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

## AIC

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

For each model  $X_M$ , the model parameters (slopes and variances) can be estimated using least squares as  $\hat{\theta}_M$ . This allows us to construct a “predictive density”

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

## AIC (continued)

The Kullback-Leibler divergence (“KL-divergence”) between the predictive density and the actual density  $P(Y)$  is

$$E_Y \log \left( \frac{p(Y)}{p(Y; X_M, \hat{\theta}_M)} \right) = \int \log \left( \frac{p(Y)}{p(Y; X_m, \hat{\theta}_M)} \right) p(Y) dY \geq 0.$$

Note that here we are considering  $\hat{\theta}_M$  to be based on a data set  $Y_{\text{train}}$  that is equal in distribution to, but independent of  $Y$ .

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

## AIC (continued)

Our aim is to select the model with the least KL-divergence between the predictive density and the density of the data generating model  $p(Y)$ . 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

$$E_Y \log(p(Y; \hat{\theta}_M, X_M)) = \int \log(p(Y; X_M, \hat{\theta}_M))p(Y)dY.$$

## AIC (continued)

A naive estimate of the KL-divergence is

$$\log(p(Y_{\text{train}}; X_M, \hat{\theta}_M)).$$

But since  $Y_{\text{train}}$  and  $\hat{\theta}_M$  are dependent, this will be 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  $\sigma^2$ ).

Thus we may take

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

as a model selection statistic (commonly this is multiplied by -2). This is called the Akaike Information Criterion (AIC).

## AIC (continued)

To apply the AIC to linear models, we assume the error values  $\epsilon$  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  $\beta$ , we get  $-n \log(\hat{\sigma}^2)/2$  (plus a constant). Therefore maximizing the AIC is equivalent to maximizing

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

## Bayesian Information Criterion (BIC)

A different criterion that we will not derive here is the “Bayesian information criterion” (BIC)

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

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

$$\hat{Y}_{M,\text{test}} = X_{M,\text{test}}\hat{\theta}_M$$

and the quality of prediction can be assessed, for example with the “prediction mean squared error”

$$\|Y_{\text{test}} - \hat{Y}_{M,\text{test}}\|^2/n.$$

## 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  $S_1 \cup \dots \cup S_k = \{1, \dots, n\}$ .

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

$$\text{CV}_k = n^{-1} \sum_j \sum_{i \in S_j} (Y_i - X_i' \hat{\beta}_j)^2$$

This is an approximately unbiased (but potentially very imprecise) estimate of the expected squared prediction error on a true test set.

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

## Cross-validation (continued)

For OLS regression,  $CV_1$  can be computed rapidly as it is the sum of squared PRESS residuals:

$$CV_1 = n^{-1} \sum_i R_i^2 / (1 - P_{ii})^2.$$

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

$$GCV_1 = n^{-1} \sum_i R_i^2 / (1 - \text{tr}(P)/n)^2 = n^{-1} \frac{\|Y - \hat{Y}\|^2}{(1 - \text{tr}(P)/n)^2}.$$