SMIII lectures

Week 11

Model selection

Model selection

How do we decide on the best model for our data.

Two parts:

- Choice of algorithm.
- Choice of heuristic.

The forward selection algorithm with P-values

1. Begin with the null model.
2. For every term not currently included in the model, calculate a P-value for the inclusion of that term.
3. If the smallest P-value is less than the threshold p_{in} (usually chosen to be 0.05), add that term to the model.
4. Iterate (2), (3) until no further terms are significant.

The backwards elimination algorithm with P-values

(1) Begin with the most complicated model to be considered.

(2) For each term included in the model, calculate the P-value for the removal of that term.

(3) If the largest P-value is greater than the threshold p_{out} (usually chosen to be 0.05), remove that term from the model.

(4) Iterate (2), (3) until the model contains only significant terms.

The stepwise selection procedure with P-values

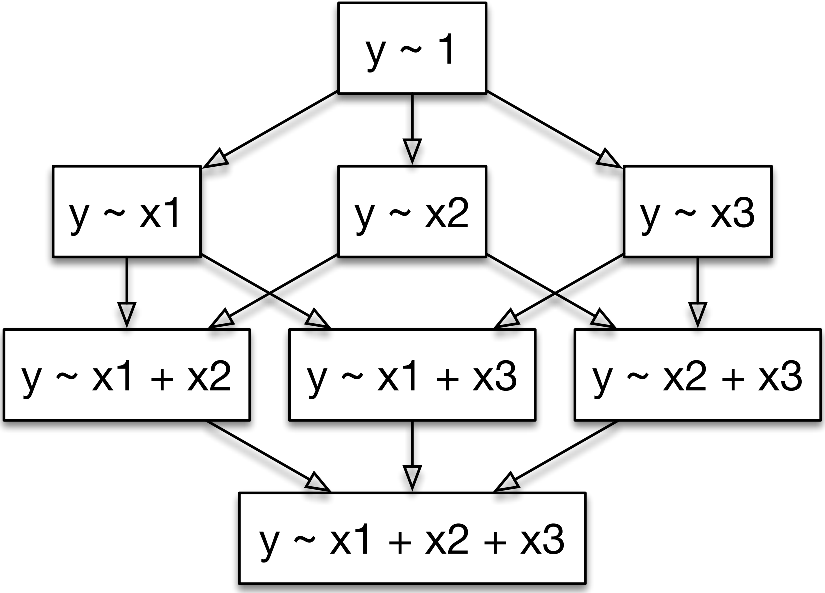
(1) Begin with the null model.

(2) Perform one step of forward selection using a liberal value of p_{in} such as 0.2 or 0.15.

(3) Perform one step of backward elimination with a value of P_{out} such as 0.05.

(4) Iterate (2), (3) until no further changes occur or the algorithm cycles.

Comparison of methods



Principle of marginality

Whenever an interaction term is included in the model, all implied lower order interactions and main effects must also be included.

For example if we find that we have an interaction term $x_{i1}x_{i2}$ in the model, the we must keep the main effects x_{i1} and x_{i2} in the model.

Measures

Root mean square error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}.$$

R-squared

$$R^2 = cor(y, \hat{f}(x))^2.$$

Mean absolute error (MAE)

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{f}(x_i)|.$$

Adjusted R^2

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

C_p

For a fitted least squares model with k predictors, the C_p estimate of test MSE is

$$C_p = \frac{1}{n}(RSS + 2k\hat{\sigma}^2),$$

where

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

and $\hat{\sigma}^2$ is an estimate of var(ϵ).

Mallow's C_p

Mallow's C_p is defined as

$$C_p^{'} = \frac{RSS}{\hat{\sigma}^2} + 2k - n$$

Akaike information criterion (AIC)

The Akaike information criterion (AIC) is defined as

$$AIC = 2k - 2\ln(\hat{L}),$$

where k is the number of parameters in the model, and $\ln(\hat{L})$ is the log likelihood evaluated at the maximum likelihood estimates. We choose the model with the lowest AIC.

Akaike information criterion corrected (AICc)

To adjust for small sample sizes, the AICc is used. It is defined as

$$AICc = AIC + \frac{2k(k+1)}{n-k-1},$$

where n is the sample size.

Bayesian information criterion (BIC)

A more stingent criterion with respect to the number of parameters is the Bayesian information criterion (BIC). It is defined as

$$BIC = \ln(n)k - 2\ln(\hat{L}).$$

Cross-validation

Cross-validation

A useful method to access how good a model is for prediction is cross-validation.

For k-fold cross-validation, you split the data into k parts.

In each step, train the model on $k - 1$ parts and test for the kth part.

Cross-validation

Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test

Notation

Label each part

$$C_1, C_2, \ldots, C_K$$

Let the number of observations in C_k be n_k. So that

$$n = \sum_{k=1}^{K} n_k,$$

i.e. n is the total number of observations.

Prediction error

The cross validation estimate of the prediction error is

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k,$$

where

$$MSE_k = \sum_{i \in C_k} \frac{(y_i - \hat{y}_i)^2}{n_k},$$

where $\hat{y}_i$ is the fitted value for observation i for the model with part k removed.

Ridge regression

Ridge regression

An alternative approach is to use a penalized regression method to control model complexity.

Consider the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i$$

where

$$e_i \sim N(0, \sigma^2)$$

independently for $i = 1, 2 \ldots, n$.

Ridge regression

The ridge regression estimator, for $\lambda \geq 0$ is defined by

$$\hat{\boldsymbol{\beta}}_{(\lambda)} = \mathrm{argmin}_{\beta} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}_1\|^2$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}.$$

The complexity parameter λ

The parameter λ controls the complexity of the fitted model.

- When $\lambda = 0$, there is no penalty for model complexity and the resulting estimator is the ordinary least squares estimator.
- As $\lambda \to +\infty$, the components of $\boldsymbol{\beta}_1$ are forced to 0 and, in the limit, the predicted value becomes $\hat{y}_i = \bar{y}$ for all i. The ridge regression estimator, $\hat{\boldsymbol{\beta}}_{(\lambda)}$, is called a **shrinkage** estimator as the coefficients are shrunk toward 0.

Dealing with the intercept term

The intercept term is excluded from the penalty term.

A simple way to deal with intercept parameter is to **centre** the data before prior to analysis.

That is, the sample mean is subtracted from each variable so that y_i is replace by $y_i - \bar{y}$ and similarly for each of the predictor variables x_i.

It can be shown that be shown the ridge regression estimate of β_0 is always 0 for centred data.

For this reason, it is convenient to assume

- The data have been centred prior to analysis;
- No intercept is included in the model.

Ridge regression for centred data

For centred data, the ridge regression estimator becomes

$$\hat{\beta}_{(\lambda)} = \mathrm{argmin}_{\beta} \|\boldsymbol{y} - X\beta\|^2 + \lambda\|\beta\|^2.$$

For given λ, the solution to the ridge regression problem is

$$\hat{\beta}_{(\lambda)} = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}.$$

Scaling

Ordinary linear regression is invariant under scaling of the predictor variables.

- For example, in the regression of `FEV` on `Height` measure in cm, $\hat{\beta} = 0.052$ with standard error 0.00116. If we convert `Height` to metres, by dividing by 100, regression coefficient becomes $\hat{\beta} = 5.2$ with standard error 0.116.

This invariance does not apply with ridge regression.

- Therefore the ridge regression estimate depends on the units of measurements chosen.
- To resolve this ambiguity, the predictor variables are usually scaled to have unit variance prior to fitting a ridge regression model.

Model complexity

The complexity parameter, λ, can be chosen to minimise the cross-validated error.

The complexity of a ridge regression can also be described by the effective degrees of freedom

$$df(\lambda) = tr(X(X^TX + \lambda I)^{-1}X^T).$$

The effective degrees of freedom are defined in analogy to the ordinary degrees of freedom, p, in a linear regression model, noting that

$$p = tr(X(X^TX)^{-1}X^T)$$