Model Selection Criteria: AIC & BIC

This document discusses some of the mathematical details of Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC). We assume the reader knowledge of the matrix form for multiple linear regression.Please see Matrix Form of Linear Regression for a review.

Maximum Likelihood Estimation of $\boldsymbol{\beta}$ and σ

To understand the formulas for AIC and BIC, we will first briefly explain the likelihood function and maximum likelihood estimates for regression.

Let $\mathbf{Y}$ be $n \times 1$ matrix of responses, $\mathbf{X}$, the $n \times (p+1)$ matrix of predictors, and $\boldsymbol{\beta}$, $(p+1) \times 1$ matrix of coefficients. If the multiple linear regression model is correct then,

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2) \tag{1}$$

When we do linear regression, our goal is to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2 from (1). In Matrix Form of Linear Regression, we showed a way to estimate these parameters using matrix alegbra. Another approach for estimating $\boldsymbol{\beta}$ and σ^2 is using *maximum likelihood estimation.*

A **likelihood function** is used to summarise the evidence from the data in support of each possible value of a model parameter. Using (1), we will write the likelihood function for linear regression as

$$L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^T (Y_i - \mathbf{X}_i\boldsymbol{\beta}) \right\} \tag{2}$$

where Y_i is the i^{th} response and $\mathbf{X}_i$ is the vector of predictors for the i^{th} observation. One approach estimating $\boldsymbol{\beta}$ and σ^2 is to find the values of those parameters that maximize the likelihood in (2), i.e. **maximum likelhood estimation**. To make the calculations more manageable, instead of maximizing the likelihood function, we will instead maximize its logarithm, i.e. the log-likelihood function. The values of the parameters that maximize the log-likelihood function are those that maximize the likelihood function. The log-likelihood function we will maximize is

$$\begin{aligned} \log L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^T (Y_i - \mathbf{X}_i\boldsymbol{\beta}) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{3}$$

[–insert details MLES–]

The maximum likelihood estimate of $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \qquad \hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{n}RSS \tag{4}$$

where RSS is the residual sum of squares. Note that the maximum likelihood estimate is not exactly equal to the estimate of σ^2 we typically use $\frac{RSS}{n-p-1}$. This is because the maximum likelihood estimate of σ^2 in (4) is a *biased* estimator of σ^2. When n is much larger than the number of predictors p, then the differences in these two estimates are trivial.

AIC

Akaike's Information Criterion (AIC) is

$$AIC = -2\log L + 2(p+1) \tag{5}$$

where $\log L$ is the log-likelihood. This is the general form of AIC that can be applied to a variety of models, but for now, let's focus on AIC for mutliple linear regression.

$$\begin{aligned} AIC &= -2\log L + 2(p+1) \\ &= -2\Big[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\Big] + 2(p+1) \\ &= n\log\big(2\pi\frac{RSS}{n}\big) + \frac{1}{RSS/n}RSS \\ &= n\log(2\pi) + n\log(RSS) - n\log(n) + 2(p+1) \end{aligned} \tag{6}$$

BIC

[—]