A REVIEW OF CLASSIC PROCEDURES

Zheng Yuan

July 5, 2019

1 Introduction

Variable selection is essential for improving inference and interpretation in multivariate linear regression. Although a number of alternative variable selection criteria have been suggested, the most prominent and widely used are the Akaike information criterion (AIC), Bayesian information criterion (BIC), Wilks test and their modifications. Suppose we have a set of models, usually not all of the same dimension, and want to decide which of them fits a data set best. For the Wilks test, suppose that we had an m-dimensional model H_0 included in a d-dimensional model H_1, where m < d. The maximum of the likelihood over H_1 would always be at least as large, and usually larger, than over H_0 because of the inclusion. But, if the maximum likelihood over H_0 was not too much smaller than over H_1, then in the test, H_0 is not rejected.

2 Wilks Test

Assume that observations $X_1, ..., X_n$ are i.i.d. with likelihood function $f(\theta, x)$ for some $\theta \in H_1$. We want to test the hypothesis that $\theta \in H_0$. Let $L(\theta) = \prod_{j=1}^{n} f(\theta, X_j)$ be the likelihood function. Let ML_d be the maximum of the likelihood for θ in H_1. Let ML_m be, likewise, the maximum of the likelihood for θ in H_0. Then $ML_m \leq ML_d$ because $H_0 \subset H_1$. Let Λ be the likelihood ratio, $\Lambda = ML_m/ML_d$, so that $0 < \Lambda \leq 1$. Simply, we would want to reject H_0 if Λ is small, or sufficiently less than 1, depending on n, but not reject it if Λ is close to 1.
S. S. Wilks (1938) proposed the following test: let

$$W = -2log\Lambda,$$

so that $0 \leq W < \infty$. Wilks found that if the hypothesis H_0 is true, then the distribution of W converges a χ^2 distribution with $(d - m)$ degrees of freedom as $n \to \infty$, not depending on the true $\theta = \theta_0 \in H_0$. Thus, H_0 would be rejected if W is too large in terms of the tabulated $\chi^2(d - m)$ distribution.

3 AIC And BIC

In "model selection," there are m models $M_1, ..., M_m$, where usually $m > 2$. The models may be "nested," with inclusions $M_1 \subset M_2 \subset ... \subset M_m$, or they may not be. Rather than testing multiple hypotheses on the models two at a time, to see if we reject one or the other, it's convenient to have a criterion for selecting one of the models. Arbitrary levels such as 0.05 may not be appropriate. But, as in the Wilks test, we want to avoid simply choosing the model with maximum likelihood, which in the nested case would mean always selecting M_m. That could well be "overfitting." It is natural to consider maximum log likelihoods rather than likelihoods themselves. Let ML_i be the maximum likelihood over the ith model and $MLL_i = ln(ML_i)$ the maximum log likelihood over the ith model. Let d_i be the dimension of the ith model M_i. Different "penalties" have been proposed to be subtracted from MLL_i to avoid overfitting. Perhaps the first was the AIC or "Akaike information criterion"

$$AIC_i = MLL_i - d_i$$

(Akaike, 1974). Later, G. Schwarz (1978) proposed a different penalty giving the "Bayes information criterion,"

$$BIC_i = MLL_i - \frac{1}{2} d_i log n.$$

For either AIC or BIC, we would select the model with the largest value of the criterion.
Schwarz (1978) proved that under some conditions, the BIC is consistent, meaning that if one of the models $M_1, ..., M_m$ is correct, so that there is a true θ_0 in that model, then as n becomes large, with probability approaching 1, BIC will select the best model, namely the smallest model (model of lowest dimension) containing θ_0. Poskitt (1987) and Haughton (1988) extended and improved Schwarz's work, showing that consistency held also under less restrictive conditions. The AIC is not necessarily consistent in this sense, as will be shown. Although that may make the BIC seem preferable, the problem is that if none of the models $M_1, ..., M_m$ is actually correct, and in such a case it is not so clear which criterion, if either, is best to use.

4 Asymptotic Essence Of Wilks Test, AIC And BIC

Suppose we have just two models M_1 and M_2 with $M_1 \subset M_2$, and M_i has dimension d_i with $d_1 < d_2$. To fit with the assumptions of the Wilks test, suppose that there is a true $\theta = \theta_0 \in M_2$. Then M_1 is the best model if $\theta_0 \in M_1$, otherwise M_2 is. For any of three methods, the Wilks test, AIC, and BIC, given a data set, we'd evaluate the maximum log likelihoods MLL_i for i = 1,2. For the Wilks test, with test statistic $W = -2log\Lambda$ as above, where

$\Lambda = ML_1/ML_2$, so $W = 2(MLL_2 - MLL_1)$, for n large enough, and some $\alpha > 0$, we would reject M1 (and so select M2) if $W \geq \chi^2_{1-\alpha}(d2 - d1)$, otherwise select M1. If $\theta_0 \notin M_1$, so M_2 is the best model, then ML_1/ML_2 will approach 0 exponentially as $n \to \infty$, and $W \sim cn$ for some $c > 0$, so we will make the correct choice with probability $\to 1$ as $n \to \infty$. If $\theta \in M_1$, the Wilks test will correctly select M_1 with a probability converging to $1 - \alpha$.
The AIC will select M_2 if $W > 2(d2 - d1)$, which if $\theta_0 \notin M_1$ will occur and give the correct choice with probability converging to 1 as $n \to \infty$. On the other hand, if $\theta_0 \in M_1$, W will converge in distribution to $\chi^2(d2 - d1)$ as $n \to \infty$, so the probability of incorrectly rejecting it will again not go to 0 as n becomes large (as in the Wilks test for fixed $\alpha > 0$) because

$$Pr(W > 2k) \to Pr(\chi^2(k) > 2k) > 0$$

for k = d2-d1.
To this sense, we can also see that leave one out cross validation is asymptotically equivalent to AIC ,at least in linear models, in terms of their error rates.
The BIC will select M_2 if $W > (d2 - d1)logn$. If $\theta_0 \in M_1$, the probability of selecting M_2 will go to 0 as $n \to \infty$, as $(d2 - d1)logn$ eventually becomes larger than $\chi^2_{1-\alpha}(d2 - d1)$ for any $\alpha > 0$. This illustrates the consistency of BIC, that it will select a lower dimensional model when it is best. If M_2 is the best model, then BIC will select it with probability $\to 1$ as $n \to \infty$, as n becomes larger than log n. So of the three criterias, BIC is the only consistent one.