Bayesian Estimation of Prediction Error and Variable Selection in Linear Regression

Andrew A. Neath
Department of Mathematics and Statistics; Southern Illinois University Edwardsville;
Edwardsville, IL, 62025; e–mail: aneath@siue.edu

Joseph E. Cavanaugh
Department of Biostatistics; The University of Iowa;
200 Hawkins Drive, C22 GH; Iowa City, IA 52242; e–mail: joe-cavanaugh@uiowa.edu

Summary

An important statistical application is the problem of determining an appropriate set of input variables for modeling a response variable. A selection criterion may be formulated by constructing an estimator of a measure known as a discrepancy function. Such a measure quantifies the disparity between the true model and a fitted candidate model, where candidate models are defined by which input variables are included in the mean structure. A reasonable approach to gauging the propriety of a candidate model is to define a discrepancy function through the prediction error associated with the fitted model. An optimal set of input variables is then determined by searching for the candidate model that minimizes the discrepancy function. Although this type of variable selection problem has been extensively studied, attention is less often paid to the problem of accounting for the uncertainty inherent to the model selection problem. In this paper, we focus on a Bayesian approach to estimating a discrepancy function based on prediction error in linear regression. It is shown how this approach provides an informative method for quantifying model selection uncertainty.

Key words: Cp statistic; discrepancy function; model selection criterion

1 Introduction

An important topic in regression theory is the problem of determining which input variables are needed for estimating a response function. Typical methods for variable selection in linear regression are based on the goal of including inputs in the mean structure having nonzero true regression coefficients, and excluding inputs having true coefficients equal to zero. A regression model which excludes no input with a nonzero coefficient is said to be correctly specified. A gray area exists for the selection problem when an input variable has a small, yet nonzero, coefficient. Variable selection techniques based on finding correctly specified models do not directly address the problem of determining when an input coefficient is too small for inclusion.

Variable selection in linear regression may be facilitated by the use of a model selection criterion. A model selection criterion is often formulated by constructing an estimator of a measure known as a discrepancy function. Such a measure quantifies the disparity between the true model (i.e., the model which generated the observed data) and a candidate model. Candidate models in linear regression are defined by which input variables are included in the mean structure.

In selecting a discrepancy function, one must consider which aspect of a fitted model should be required to conform with the true model. A reasonable approach in linear regression is to define a discrepancy function through the prediction error associated with a candidate model. In this setting, the problem of discrepancy function estimation reduces to the problem of estimating prediction error.

Much of the discrepancy function based model selection literature is dedicated to the problem of discrepancy function estimation, and to the companion problem of developing model selection criteria. Less attention is paid to the problem of accounting for the uncertainty inherent to the model selection problem within the discrepancy function framework. The focus of the current paper is to provide a Bayesian approach to the problem of estimating a discrepancy function based on prediction error in linear regression. The goal of the companion model selection problem is to select those input variables corresponding to the candidate model with minimum prediction error. We discuss how the posterior distribution on the candidate model prediction errors can be used in quantifying model selection uncertainty. We present two applications which illustrate how a Bayesian approach leads to an improved understanding of the issues in a linear regression variable selection problem.

2 Linear Regression Variable Selection

We observe data following the linear regression model

$$y_i = x_i'\beta + e_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $e_1, \ldots, e_n$ are iid $N(0, \sigma^2)$ random variables and $x_1, \ldots, x_n$ are $(k+1) \times 1$ vectors of nonrandom input variables. The first entry of each x_i is 1, corresponding to an intercept parameter. Let $K^+ = \{0, 1, \ldots, k\}$ be a set of indices, where each index identifies k measured inputs along with an intercept. We will define candidate models by which inputs are included and which inputs are excluded. Let P denote a subset of K^+ with $|P| = p$ elements. We shall also use the notation P for a candidate model consisting of only the p inputs matching the indices in P. Let $Q = K^+ - P$ be the complement set with $|Q| = q$ elements. An all–subsets regression candidate class consists of models representing all possible subsets of K^+. In some situations, we may consider a candidate class smaller than all subsets.

The problem of variable selection requires a decision as to which input variables are needed for modeling the response variable. With the preceding notation, variable selection is defined by the selection of a model P from among the candidate class. For a survey of the variable selection problem, see George (2000), and Clyde & George (2004). A standard Bayesian approach to model selection proceeds as follows. Denote the candidate class of models as $M_1, M_2, \ldots, M_L$. Assume candidate model M_l is uniquely parameterized by a vector θ_l. A prior $\pi\left(M_l\right)$ is placed on the candidate class of models, and a prior $\pi\left(\theta_l|M_l\right)$ is placed on the parameters for each model. Data is observed according to $\pi\left(y|\theta_l, M_l\right)$. Model selection is based on the posterior distribution over the candidate class of models, where $\pi\left(M_l|y\right)$ represents the probability that candidate model M_l is the true model. In the regression variable selection problem, the structure of M_l coincides with the structure of the true model if the input variables included in M_l have nonzero true coefficients, and the excluded input variables in M_l have true coefficients equal to zero. The philosophy behind this approach leads one to infer that excluded input variables from models with high probability are likely to have true coefficients equal to zero.

In this paper, we will consider a discrepancy function approach to variable selection. A discrepancy function is a measure which quantifies the disparity between the model which generated the data and a candidate model. Write the full model in (1) as

$$y = X\beta + e, \tag{2}$$

where y is an $n \times 1$ response vector, X is the $n \times (k+1)$ input matrix with full column rank, β is the $(k+1) \times 1$ coefficient vector, and e is an $n \times 1$ error vector. Let Z_P denote the $n \times p$ input matrix for candidate model P obtained from X by deleting the columns matching the indices in Q. Let $M_P = Z_P \left(Z_P' Z_P\right)^{-1} Z_P'$ project onto the column space of Z_P. Define $\widehat{\beta}_P$ as an estimate of β obtained by setting the coefficients with subscripts in Q equal to zero, and estimating the remaining coefficients using least squares as $\left(Z_P' Z_P\right)^{-1} Z_P' y$. A discrepancy for candidate model P, based on the difference between the fitted model and the true model, is given as

$$J_P = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(x_i' \widehat{\beta}_P - x_i' \beta \right)^2 .$$

Since J_P is a random variable, it is better to think of $\Delta_P = E\left(J_P\right)$ as the discrepancy function. We will refer to Δ_P as the prediction error for candidate model P.

The prediction error can be decomposed into a variance contribution and a bias contribution as

$$\Delta_P = V_P + \frac{1}{\sigma^2} B_P, \tag{3}$$

where

$$V_P = |P| = p \tag{4}$$

and

$$B_P = \left\| X\beta - M_P X\beta \right\|^2 . \tag{5}$$

The variance contribution arises from the need to estimate unknown regression coefficients in a model, whereas the bias contribution is due to possible model misspecification. One can easily see that the variance contribution increases as the number of nonzero regression coefficients to be estimated increases. The bias contribution is the squared distance between the true mean response vector $X\beta$ and the approximating mean space determined through the candidate model P by the column space of Z_P. As the number of nonzero regression coefficients in a candidate model increases, the approximating mean space grows larger, and the bias contribution decreases.

We define the *best* model in the candidate class as that model which minimizes the prediction error Δ_P. A *correctly specified* model is one whose bias contribution is zero. The best model in the candidate class does not need to be a correctly specified model. Input variables not included in a correctly specified model necessarily have true regression coefficients equal to zero. Assuming that the best model satisfies this property may be unrealistic in some applications. Consider a case where the excluded input variables have true coefficients which are near, but not equal to, zero. Least squares estimates of these coefficients will be unbiased, but will introduce an increase to the variance contribution. It may be better under the prediction error framework to set these estimated coefficients to zero, introduce a small increase to the bias contribution, but without introducing any additional variance contribution.

The best model according to prediction error is one which balances the variance contribution and the bias contribution. Within this framework, we are not looking to exclude just those input variables with true coefficients equal to zero, but rather we are looking to exclude those input variables with true coefficients near enough to zero that an estimate set to zero is more accurate than an estimate computed from the data.

A variable selection criterion can be created by deriving an estimator of Δ_P for each candidate model. Some of the earliest selection criteria were based on minimizing the prediction error. For example, Mallows (1973) introduced the now famous statistic

$$\mathrm{C}_P = \left(\frac{RSS_P}{\widehat{\sigma}^2_{K^+}} - n \right) + 2p,$$

where $RSS_P = \sum_{i=1}^{n} \left(y_i - x_i' \widehat{\beta}_P \right)^2$ is the residual sum of squares for candidate model P, and $\widehat{\sigma}^2_{K^+} = RSS_{K^+} / \left(n - k - 1 \right)$ is an unbiased estimator of the variance σ^2. It can be shown that $E\left(\mathrm{C}_P\right) \approx \Delta_P$, so C_P is an approximately unbiased estimator of the prediction error for candidate model P. Fujikoshi & Satoh (1997) introduced the modified C_P statistic by creating an exactly unbiased estimator of Δ_P. Furthermore, Davies, Neath & Cavanaugh (2006) showed that the modified C_P statistic is an optimal estimator of the prediction error in the sense of being minimum variance unbiased. A variable selection criterion would be defined by the selection of the input variables in P for which C_P, or any other estimator of Δ_P, is minimum.

As we have discussed, the best model in the candidate class is the one for which Δ_P is minimum. The use of a variable selection criterion would select that model for which the criterion statistic is minimum. For a given set of regression data, a reasonable question to ask is how certain are we that the selection criterion has truly selected the best model? A more general problem is to assess how likely it is for each of the candidate models to truly be the best. More general still is the problem of quantifying the uncertainty in the measure that serves as the basis for the model selection: namely, the prediction error for each of the models in the candidate class.

3 Bayesian Approach to Model Selection Uncertainty

We will describe the uncertainty inherent to variable selection based on prediction error using a Bayesian approach. The uncertainty inherent to the specification of the full model (2) is characterized through the uncertainty associated with the parameters β and σ^2. In an effort to stay objective, we take a noninformative prior on these parameters although it is not necessary to follow this convention if good prior information is available. After observing response vector y from this model, the posterior distribution updates easily (see Gelman, Carlin, Stern & Rubin, 2003, for example) to become

$$\beta \,|\, \sigma^2, y \sim N_{k+1}\left(\widehat{\beta}_{K^+}, \sigma^2 \left(X'X\right)^{-1}\right), \tag{6}$$

$$\sigma^2 \,|\, y \sim \frac{RSS_{K^+}}{\chi^2\left(n-k-1\right)}. \tag{7}$$

Our inferential goals are focused on the prediction errors Δ_P, as defined by (3) for each of the models in the candidate class. From (4) and (5), we see how Δ_P is a function of the parameters β and σ^2. Thus, the joint posterior distribution on the set of prediction errors $\{\Delta_P\}$ over the candidate class is induced from the posterior distributions in (6) and (7). Since it is easy to generate outcomes from the multivariate normal and chi–square distributions, the posterior distribution on $\{\Delta_P\}$ can be obtained via simulation.

We can use the joint posterior distribution on $\{\Delta_P\}$ to quantify the uncertainty inherent to the problem of selecting the model with minimum prediction error. In the next section, we present some examples of various ways one can use the information from this joint distribution. For now, we offer a Bayesian model selection criterion to serve as a solution to a question posed in the last section: how likely is it that model P is the candidate model which minimizes prediction error? We look to calculate the posterior probability

$$\Pi_P = \Pr\left[\Delta_P \text{ is min} \mid y\right].$$

Consider the following algorithm for calculating these posterior probabilities.

1. Generate $\left(\beta, \sigma^2\right)_{(1)}, \ldots, \left(\beta, \sigma^2\right)_{(N)}$ from the posterior distributions in (6) and (7).
2. Calculate $\Delta_{P(1)}, \ldots, \Delta_{P(N)}$ for each candidate model.
3. Determine $P^*_{(j)} = \text{ArgMin}_P\left[\Delta_{P(j)}\right]$ for each simulation outcome j $(j = 1, \ldots, N)$.
4. Calculate $\Pi_P = \frac{1}{N}\sum_{j=1}^{N} I\left\{P = P^*_{(j)}\right\}$ for each candidate model.

The algorithm begins with repeated simulated outcomes generated from the posterior distribution on the parameters β and σ^2. For each simulation, we calculate prediction errors for the models in the candidate class. For the j^{th} simulation, call this set $\left\{\Delta_{P(j)}\right\}$. We then determine the candidate model which yields a minimum on the set $\left\{\Delta_{P(j)}\right\}$. Define this model as $P^*_{(j)}$, the best model from the candidate class for the j^{th} simulation. We calculate the posterior probability of candidate model P being best as the proportion of simulations for which candidate model P yields a minimum on the set $\left\{\Delta_{P(j)}\right\}$. Instead of merely a selection of a candidate model, we can use the statistic Π_P to quantify the degree of plausibility for a model truly being best in the sense of minimum prediction error.

4 Applications

We begin with the well known Hald data as an illustration of our Bayesian approach to quantifying model selection uncertainty. See, for instance, Draper & Smith (1998), who comment: "This particular problem illustrates some typical difficulties that occur in regression analysis." A summary of an all–subsets regression using C_P is presented in Table 1. The model with inputs X_1 and X_2 is selected based on the C_P model selection criterion. However, several other models in the candidate class might plausibly be best in terms of minimizing prediction error, if one were to account for the uncertainty of C_P as an estimate of Δ_P.

Table 1. C_P for some Hald data candidate models.

Model	C_P
4	138.7
1,2	2.7
1,4	5.5
3,4	22.4
1,2,4	3.0
1,2,3	3.0
1,3,4	3.5
2,3,4	7.3
1,2,3,4	5.0

The posterior probability of minimizing prediction error is calculated for each model in the candidate class to quantify the degree of plausibility for each model truly being best in the sense of minimum Δ_P. The results are displayed in Table 2. Enthusiasm for the minimum C_P model should be appropriately tempered. Although Model 1,2 has the greatest probability of actually minimizing Δ_P, the Π_P statistic places this probability at only .26. It is interesting to note that a Bayesian approach can do more than provide an accompanying probability for each model. We see in this example that the order and degree of preference for some candidate models is different for Π_P than for C_P. In particular, Model 1,2,3 has a higher probability of minimizing Δ_P than Model 1,2,4 even though their C_P values are essentially the same.

Table 2. Π_P for some Hald data candidate models.

Model	C_P	Π_P
1,2	2.7	.26
1,2,4	3.0	.10
1,2,3	3.0	.21
1,3,4	3.5	.15
1,2,3,4	5.0	.09
1,4	5.5	.10
2,3,4	7.3	.09

We can use the joint posterior on $\{\Delta_P\}$ to further investigate properties of the candidate models. A pairwise comparison between Model 1,2,3 and Model 1,2,4 is based on calculating $\Pr[\Delta_{124} < \Delta_{123} \mid y] = .5144$. The two models have nearly equal probability in a pairwise comparison. So, what accounts for the difference in preference when all candidate models are considered? An explanation may lie in noting that input variables X_2 and X_4 are highly correlated ($r_{24} = -.973$). Model 1,2,4 with both X_2 and X_4 can provide a good fit, and hence a small C_P value, but good fitting models without this redundancy of input information may be preferred. We can investigate this premise. Since X_2 and X_4 carry similar information, we suspect that in cases where Model 1,2,4 is preferred to Model 1,2,3, there is a good chance that a simpler model is better. This notion can be checked. Table 3 contains posterior probabilities conditional on whether or not Model 1,2,4 is preferred to Model 1,2,3. If $\Delta_{124} < \Delta_{123}$, a model without X_4 (Model 1,2) and a model without X_2 (Model 1,3,4) are both more likely to minimize prediction error than Model 1,2,4 with both X_2 and X_4. If $\Delta_{123} < \Delta_{124}$, then the most likely scenario is that Model 1,2,3 is best overall.

Table 3. Probabilities of minimizing Δ_P.

Model	unconditional	given $\Delta_{124} < \Delta_{123}$	given $\Delta_{123} < \Delta_{124}$
1,2	.26	.25	.28
1,2,4	.10	.18	0
1,2,3	.21	0	.43
1,3,4	.15	.27	.02
1,2,3,4	.09	.02	.18
1,4	.10	.14	.06
2,3,4	.09	.14	.03

As a second illustration, we consider an application based on data from a cardiac rehabilitation program at the University of Iowa Hospitals and Clinics. The data consist of measurements based on 35 patients who have had a myocardial infarction and have completed the program. The response variable is the final score on a test that reflects the capability of the patient to physically exert himself / herself. The score is in units of metabolic equivalents (METs). One MET corresponds to the rate of oxygen consumption for an average person at rest. The input variables are the patient's initial score on the exertion test (I), the patient's age (A), the patient's gender (G), the patient's baseline body mass index, dichotomized as greater than 30 or not (B), and interactions for initial score / gender (IG), initial score / body mass index (IB), age / gender (AG), age / BMI (AB). The candidate class consists of those subset models that satisfy the following criteria: the initial score is included as an input, and if an interaction is included, then both inputs represented in the interaction are also included.

Table 4 contains a list of the leading candidate models according to the C_P selection criterion, which selects Model I,G,A,B,AG. Also included in Table 4 are the posterior probabilities of minimizing prediction error. As in the first example, the order and degree of preference for the candidate models is different for Π_P than for C_P. In particular, the larger candidate models are favored according to the posterior probabilities at the expense of models with fewer input variables. So, what accounts for the difference in preference in this application?

Table 4. C_P and Π_P for rehabilitation data candidate models.

Model	C_P	Π_P
I,G,A,B,AG	5.11	.11
I,G,A,B,IG	5.33	.12
I,G,A,AG	5.36	.07
I,G,A,B,AG,IG	5.60	.10
I,G,A,AG,IG	6.15	.05
I,G,A,B	6.86	.03
I,G,A,IG	6.90	.03
I,G,A,B,AG,IG,IB	7.31	.17
I,G,A,B,AG,IG,AB	7.56	.12
I,G,A	8.38	.01
I,G,A,B,AG,IG,AB,IB	9.0	.19

An explanation may lie in the sampling variability of C_P as an estimate of Δ_P for models with fewer inputs relative to the full model. This phenomenon was studied by Mallows

(1995). In settings where no subset of inputs is strongly preferred, the C_P statistic for a smaller model may greatly underestimate the prediction error for that model. The behavior of Π_P is consistent with Mallows' result on the sampling variability of C_P for smaller models. The Π_P statistic accounts for this sampling variability and shifts its probability to the larger models in the candidate class.

Mallows (1995) derived an asymptotic approximation of the true prediction error for the "minimum C_P" model as $\min \mathrm{C}_P + 2q$. For our application, the minimum C_P model (I,G,A,B,AG) has $q = 3$ inputs excluded from the full model. Mallows' approximation of the prediction error for this model is $5.11{+}2{*}3 = 11.11$. We have the capability to investigate this idea further. Since we have the joint posterior distribution on $\{\Delta_P\}$, we have the posterior distribution on the prediction error for Model I,G,A,B,AG . The mean of this distribution is computed to be 11.02, right in line with the asymptotic approximation. Yet we can do more than provide a point estimate. For example, from the posterior distribution, we can compute a 50% Bayesian confidence interval for this prediction error to be $(8.275, 13.096)$, and a 90% Bayesian confidence interval to be $(6.701, 18.343)$. It is indeed quite possible that the prediction error for Model I,G,A,B,AG is greatly underestimated by its C_P statistic.

As the two applications illustrate, the Bayesian approach can present information on the variable selection process well beyond the capabilities of a model selection criterion.

5 Conclusion

Linear regression variable selection based on prediction error seeks to determine the model from among a candidate class for which Δ_P is minimum. An estimate of Δ_P can be used as a model selection criterion in this framework. Without accounting for the uncertainty inherent to a model selection problem, model selection criteria alone are not presenting a complete solution. In this paper, we have taken a Bayesian approach to the problem of estimating prediction error. This approach provides a method for quantifying model selection uncertainty. Two applications have been presented to illustrate the efficacy of a Bayesian approach to variable selection in linear regression.

References

Clyde, M. & George, E.I. (2004). Model uncertainty. *Statistical Science*, **19**, 81–94.

Davies, S.L., Neath, A.A. & Cavanaugh, J.E. (2006). Estimation optimality of corrected AIC and modified Cp in linear regression. *International Statistical Review,* **74**, 161–168.

Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis.* New York: Wiley.

Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and Cp in multivariate linear regression. *Biometrika,* **84**, 707–716.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2003). *Bayesian Data Analysis,* 2nd ed. London: Chapman and Hall.

George, E.I. (2000). The variable selection problem. *Journal of the American Statistical Association*, **95**, 1304–1308.

Mallows, C.L. (1973). Some comments on Cp. *Technometrics,* **15**, 661–675.

Mallows, C.L. (1995). More comments on Cp. *Technometrics,* **37**, 362–372.