Abstract

This paper focuses on the superset model problem that arises in the context of regression. To address this problem, we take the Bayesian approach to measure its uncertainty. An illustrative example with the real dataset is provided.

1 Introduction

Regression analysis is a statistical method widely used in many research areas. It is often specified as the normal linear model, where coefficients are linear and the error term follows the normal distribution to simplify the analysis. This specification aims to approximate the state of nature and is often useful in prediction as well as discussing the causality.

Among its specifications, variable selection is a central issue in the regression analysis. It is important to select an appropriate set of explanatory variables partly because of the cost of collecting variables. Many methods are proposed for the variable selection problem.
In relation to the variable selection problem, this paper focuses on the superset model problem where the linear regression model selects a larger set of variables than the state of nature does (see the example provided in Section 2). The linear regression model tends to choose smaller set of variables when the state of nature is linear in variables due to its least squares loss. However, the nonlinear relationship is more likely and this case may lead to select a larger set of variables, which will be a deficiency of the linear regression model in terms of the data collection cost.

To evaluate the superset model problem, this paper utilizes the Bayesian approach, which provides a measure of uncertainty in the form of the posterior probability, and proposes an alternative model that will select the true set of variables when the sample size is large.

This paper is organized as follows. Section 2 describes the superset model problem by providing an example. Bayes’ Theorem is adopted to evaluate the problem in Section 3 and two regression models for specification is explained in Section 4. Section 6 illustrates the proposed method and discusses its robustness.

## 2 Superset model problem

Suppose the continuous response $Y$ is associated with the set of explanatory variables $x_T$. We are interested in its mean response conditional on $x_T$. We often assume it to be linear in practice, although it is more likely to be nonlinear in reality. To this end, a regression model is specified as

$$Y = \phi(x_T) + \epsilon,$$
where $\epsilon$ is an additive error term with mean zero. The functional form $\phi(\cdot)$, the distribution of the error term, and the true set of explanatory variables $x_T$ are all unknown. With this model, we make statistical inferences about the conditional mean response by estimating $\phi(\cdot)$ and the set of explanatory variables. Among problems about how this regression model should be specified, the variable selection problem focuses on the set of explanatory variables, based on the dataset $\{y_i, x_i\}_{i=1}^n$.

When the set of explanatory variables is known, the best fit is $E(Y \mid x_T)$ as an estimator of $\phi(x_T)$ when we use the squared loss. The linear regression model assumes $E(Y \mid x_T) = x_T'\beta$, where $\beta$ is referred to as the regression coefficient vector. However, in general, $E(Y \mid x_T) \neq x_T'\beta$, contrary to the linearity assumption.

For example, suppose

$$E[Y \mid x] = \alpha + \beta_1 x_T + \beta_2 x_T^2.$$  \hspace{1cm} (1)

The linear regression with $(x_T, x_U)$, where $x_U = x_T^2$, is better than the one with $x_T$, even though the latter selects the true explanatory variable. This is an example of the superset model problem. On the other hand, when $x_U'$ is independent of $x_T$, $x_U'$ should not be included in the regression to improve the fit.

Above example suggests that the knowledge about association among variables is helpful to examine the superset model problem, and hence the variable selection problem. One approach is to estimate the conditional expectation without linearity and compare it with the one implied by the normal linear model. If they are different and the latter contains more explanatory variables, there exists the superset model problem. Because the
dataset at hand is limited, it is difficult to determine whether the superset model problem exists or not. Rather, it is evaluated in a probability form, which is explained in the next section.

3 Superset model probability

Suppose $M^*$ is the set of explanatory variables in the state of nature, which is $x_T$ when vectorized, and is known for a moment. The current dataset $\{y_i, x_i\}_{i=1}^n$ is generated from this state of nature independently for each observation $i$ and is observed. Depending on a context, the normal linear model with $M^*$ explanatory variables would be a choice if it approximates the state of nature well. In this case, we do not have the superset model problem. On the other hand, a normal linear model with a set of explanatory variables indexed by $M(\neq M^*)$ is chosen independent of the state of nature in terms of, say, prediction, where the superset model problem arises when $\mathcal{M} \supset M^*$.

However, the state of nature is usually unknown and is inferred from the dataset. The uncertainty from inference is evaluated by the posterior probability over possible subsets of explanatory variables. This paper approximates it by assuming a flexible model (see Model (5) in Section 4), which is denoted by $H_0$. Then, this posterior probability is calculated via Bayes' theorem, which is given by

$$
\Pr (M^* | \{y_i, x_i\}_{i=1}^n, H_0) = \frac{\Pr (\{Y_i\}_{i=1}^n | \{x_i\}_{i=1}^n, H_0, M^*) \Pr (M^*, H_0)}{\sum_{\bar{M}} \Pr (\{Y_i\}_{i=1}^n | \{x_i\}_{i=1}^n, H_0, \bar{M}) \Pr (\bar{M}, H_0)}.
$$

(2)

The numerator is the cross product of the marginal likelihood and the prior
belief about the set of explanatory variables. When the latter is uniform (which is assumed in the following empirical illustration), the posterior probability is proportional to the marginal likelihood under $H_0$.

When uncertainty from inference about the normal linear model is evaluated from its posterior probability as well, the overall superset model probability is calculated as

$$
\sum_{\mathcal{M}} \sum_{\mathcal{M}^*} I(\mathcal{M} \supseteq \mathcal{M}^*) \Pr(\mathcal{M}^* \mid \{y_i, x_i\}_{i=1}^n, H_0) \Pr(\mathcal{M} \mid \{y_i, x_i\}_{i=1}^n, H_1), \quad (3)
$$

where $H_1$ denotes the normal linear model (see Model 4 in Section 4).

We note that the above expression is general enough to include variables (the response and explanatory variables) that are continuous or discrete. The next section specifies two regression models $H_0$ and $H_1$, where the response is assumed to be continuous for simplicity.

### 4 Two regression models

First, the linear regression model $H_1$ is specified as

$$
Y_i = \alpha + \mathbf{x}_i' \mathbf{\beta} + \eta_i, \quad \eta_i \sim N(0, \lambda^2),
$$

where each of explanatory variables is standardized without loss of generality. To estimate model parameters ($\mathbf{\beta}, \lambda^2$), the hyper-$g$ prior is assumed. Then, the marginal likelihood is analytically tractable (see Miyawaki and MacEachern (na) for example).

Second, the model $H_1$ that is alternative to the normal linear regression model is specified as

$$
Y_i = \theta_x + \epsilon_x, \quad \epsilon_x \sim N(0, \sigma^2_x),
$$

5
given $\mathbf{x}_i = \mathbf{x}$. The normal error assumption is made because we have no other knowledge on it. Further, it makes the conditional mean estimation simpler, in terms of the number of parameters as well as the computational burden. The main purpose of this semiparametric model is to estimate conditional means of $Y$ in a flexible manner, and to captures the association between $Y$ and $\mathbf{x}$ in the state of nature as the sample size increases, which can be viewed as as an extreme of the local constant estimation (see, e.g., Fan and Gijbels (2003) for the local estimation).

When the dataset is fixed, the covariate space becomes sparse as its dimension gets larger. Then, the marginal likelihood (hence, the superset model probability) under this alternative model is strongly dependent on the prior specification. To mitigate this influence, this paper takes the $m$-fold cross-validation approach, which is described in details below.

The dataset is randomly divided into $m$ groups. One of them is used as the test set, while the remainings belong to the training set. Explanatory variables in the training set are standardized, and those in the test set are standardized by the mean and standard deviation of those in the training set. Let $\mathcal{D}_0$ and $\mathcal{D}_1$ be the sets of identification numbers of observations which belongs to the training and test sets. More precisely, $\mathcal{D}_0 = \{i \mid \text{the } i\text{-th observation is in the training set, } i = 1, \ldots, n\}$ and $\mathcal{D}_1 = \{i \mid \text{the } i\text{-th observation is in the test set, } i = 1, \ldots, n\}$. Given a choice of $\mathcal{D}_0$ and $\mathcal{D}_1$, we construct the prior and conditional marginal likelihood in the following manner.

The prior for $\theta_x$ in the model (5) given $\mathbf{x}_i = \mathbf{x}$ and $i \in \mathcal{D}_1$ is assumed as

$$\theta_x \sim N(\hat{y}_x, t^2_x),$$

(6)
where $\hat{y}_x = \bar{y}_0 + x' \hat{\beta}$, $\bar{y}_0$ is the sample average of the response in $D_0$,

$$\hat{\beta} = \left(\sum_{i \in D_0} x_i x_i'\right)^{-1} \sum_{i \in D_0} x_i y_i, \quad (7)$$

$$t_x^2 = s^2 \left\{ \frac{1}{|D_0|} + x' \left(\sum_{i \in D_0} x_i x_i'\right)^{-1} x \right\}, \quad (8)$$

$$s^2 = \frac{1}{|D_0| - k - 1} \sum_{i \in D_0} \left(y_i - \bar{y}_0 - x'_i \hat{\beta}\right)^2. \quad (9)$$

When $A$ is a set, $|A|$ is the number of elements in the set. This prior is constructed from classical OLS estimates of mean and standard deviation of $Y_i$ at $x_i = x$. By using the prior that is obtained from the linear model and using the model that focuses on the local observation, we are able to combine local and global information. It is possible to use other estimates such as corresponding normal linear regression estimates under the hyper-g prior. However, to keep the methodology as simple as possible, we take the above prior specification.

Then, we are able to derive the marginal likelihood conditional on the nuisance parameter $\sigma_x^2$ for each $x_i = x$ and $i \in D_1$. Let $M_x = \{i \mid x_i = x, i \in D_1\}$ and $n_x = |M_x|$. Then, this conditional marginal likelihood is given by

$$m^*(\{Y_i\}_{i \in M_x} \mid \{x_i\}_{i \in M_x}, \sigma_x^2, \{y_i, x_i\}_{i \in D_0})$$

$$= \frac{\tau_x}{(\sqrt{2\pi} \sigma_x)^{n_x} t_x} \exp \left\{ -\frac{1}{2} \left( -\frac{\mu_x^2}{\tau_x^2} + \frac{\sum_{i \in M_x} y_i^2}{\sigma_x^2} + \frac{\hat{y}_x^2}{t_x^2} \right) \right\},$$

$$\mu_x = \tau_x \left( \frac{\sum_{i \in M_x} y_i}{\sigma_x^2} + \frac{\hat{y}_x}{t_x} \right), \quad \tau_x = \left( \frac{n_x}{\sigma_x^2} + \frac{1}{t_x^2} \right)^{-1}.$$

The full Bayes analysis specifies a prior on the nuisance parameter $\sigma_x^2$ as well. However, because the data are sparse at $x$, how we specify it af-
fects the (unconditional) marginal likelihood much. To mitigate this problem, this paper takes the empirical Bayes approach. The marginal likelihood for each \( x_i = x \) is the conditional marginal likelihood \( m^*\{Y_i\}_{i \in M_x} \mid \{x_i\}_{i \in M_x}, \sigma^2_x, \{y_i, x_i\}_{i \in D_0} \) maximized over \( \sigma^2_x \). More precisely,

\[
m(x_i \mid \{x_i\}_{i \in M_x}, \{y_i, x_i\}_{i \in D_0}) \equiv \max_{\sigma^2_x} m^*\{Y_i\}_{i \in M_x} \mid \{x_i\}_{i \in M_x}, \sigma^2_x, \{y_i, x_i\}_{i \in D_0}.
\]

See the next section for this maximization in details. By multiplying it over all distinct \( x \) in \( D_1 \) and taking the geometric mean, we have the marginal likelihood for \( D_1 \) per one observation, which is given by

\[
\left\{ \prod_x m(x_i \mid \{x_i\}_{i \in M_x}, \{y_i, x_i\}_{i \in D_0}) \right\}^{1/|D_1|}.
\]

The geometric mean is to take care of different sample sizes in different test sets.

Finally, we repeat above process until all \( m \) groups are used as the test set and calculate above marginal likelihood for each test group selection. After averaging \( m \) marginal likelihoods, we raise it to the power of \( n \) to obtain the final marginal likelihood estimate for the model (5).

The robustness of this approach is of interest. The approach will be more useful if we know the upper and lower bounds of the superset model probability under \( H_0 \) when its specification changes. Two points are discussed regarding robustness.

First, we consider the robustness to the number of folds in the cross-validation. In the following empirical analysis in Section 6, we use the 10-fold cross-validation. When the number of folds changes from 2 to 15, we see the resulting probability does not change much with the diabetes dataset (see Figure 1).
Figure 1: Superset model probability change as the number of folds increases. (The dotted line indicates the 10-fold cross-validation, which will be used in Section 6)

Second, it is important to check the robustness to the prior. One approach would be to use the ϵ-contamination class prior (see Section 4.7.4 of [Berger 1985]), and to show the sensitivity of the marginal likelihood, which will be our future analysis.

5 Maximize the marginal likelihood

Letting

\[ \bar{y}_x = \frac{1}{n_x} \sum_{i \in G_x} y_i, \]  
(10)

\[ s_y^2 = \frac{1}{n_x} \sum_{i \in G_x} (y_i - \bar{y}_x)^2 = \frac{1}{n_x} \sum_{i \in G_x} y_i^2 - \bar{y}_x^2, \]  
(11)
Table 1: Local marginal likelihood function

| Condition                          | Local Marginal Likelihood |
|------------------------------------|---------------------------|
| $\sigma^2_x \to 0$                 | $0$                       |
| $0 < \sigma^2_x < \infty$          | $\sigma^2_x \to \infty$ |
| $s_y^2 > 0$                         | $\frac{1}{\sqrt{2\pi t_x}} \exp \left\{ -\frac{1}{2t_x} (\bar{y}_x - \hat{y}_x)^2 \right\}$ See Lemma 2 | $0$ |
| $s_y^2 = 0$                         | $0$                       |

the local marginal likelihood function is characterized by Theorem 1.

**Theorem 1.** The local marginal likelihood function has extremal values at positive solutions to the cubic equation (20) given in the proof when $s_y^2 > 0$; at 0 when $s_y^2 = 0$, $t^2_x - (\bar{y}_x - \hat{y}_x)^2 \geq 0$; and at 0 and $-t^2_x + (\bar{y}_x - \hat{y}_x)^2 < 0$ when $s_y^2 = 0$, $t^2_x - (\bar{y}_x - \hat{y}_x)^2 < 0$.

See Appendix A for its proof. Table 1 summarizes the result. By this theorem, we are able to set a value of $\sigma^2_x$ to maximize the marginal likelihood, instead of placing a prior on it.

### 6 Illustrative example

The diabetes data (see [Efron et al. (2004)](https://www-stat.stanford.edu/~tibs/ElemStatLearn)) are used to illustrate our method. This dataset contains 442 observations. For the analysis below, we use the logarithm of the diabetes progression measure as the response and use remaining 10 variables are used as explanatory variables. The proposed method is applied, and the superset model probability for this dataset with 10-fold cross-validation is estimated to be 22.37%.

As discussed by [MacEachern and Miyawaki (2022)](https://www-personal.umich.edu/~nemahech/), the diabetes dataset seems to be collected from least two different sources. In particular, the preci-
sion of two explanatory variables (the blood pressure and fourth blood serum measurement) consists of a mix of finer and coarser observations. When the data are divided into two groups by this precision, MacEachern and Miyawaki (2022) suggests these two datasets have different characteristics.

This conclusion is also confirmed in terms of the superset model probability. When the dataset for observations with finer variables is used, it is 22.19%. When, on the other hand, that for observations with coarser variables is used, it is 10.72%. The superset model problem is more likely to occur with the former dataset than the latter one, which would be due to the difference in characteristics of these two datasets.

A Proof of Theorem 1

The local marginal likelihood function is characterized in the following two lemmas. The proof of the theorem directly follows from them.

**Lemma 1.** The local marginal likelihood function converges to a finite value as \( \sigma^2_x \) approaches zero or infinity.

**Proof.** Observe that

\[
g (\sigma^2_x) = - \frac{\mu^2_x}{\tau^2_x} + \sum_{i \in G_x} \frac{y_i^2}{\sigma^2_x}
\]

\[= \frac{1}{\sigma^2_x} \left\{ \sum_{i \in G_x} y_i^2 - \frac{\tau^2_x}{\sigma^2_x} \left( \sum_{i \in G_x} y_i \right)^2 \right\} - 2 \frac{\tau^2_x}{\sigma^2_x} \frac{n_x \bar{y}_x \hat{y}_x}{t^2_x} - \tau^2_x \hat{y}_x^2
\]

\[= \frac{1}{\sigma^2_x} \left\{ n_x s_y^2 + \left( \frac{1}{n_x} - \frac{\tau^2_x}{\sigma^2_x} \right) n_x^2 \bar{y}_x^2 \right\} - 2 \frac{\tau^2_x}{\sigma^2_x} \frac{n_x \bar{y}_x \hat{y}_x}{t^2_x} - \tau^2_x \hat{y}_x^2.
\] (12)
First, we consider the convergence when $\sigma^2_x \to \infty$. Because $\tau^2_x \to t^2_x$ and $\frac{\tau^2_x}{\sigma^2_x} \to 0$, $g(\sigma^2_x) \to \frac{\bar{y}_x}{t^2_x}$, as $\sigma^2_x \to \infty$. Hence, the local marginal likelihood function converges to zero in this case.

Next, we consider the case when $\sigma^2_x \to 0$. As $\sigma^2_x \to 0$, $\tau^2_x \to 0 \text{ and } \tau^2_x \sigma^2_x \to 1$. Because $\frac{1}{\sigma^2_x} \left( \frac{1}{n_x} - \frac{\tau^2_x}{\sigma^2_x} \right) = \frac{1}{n_x (n_x t^2_x + \sigma^2_x)}$, we have

$$g(\sigma^2_x) \to \begin{cases} \infty, & \text{when } s^2_y > 0, \\ \frac{\bar{y}_x^2}{t^2_x} - 2 \frac{\bar{y}_x \bar{y}_x}{t^2_x}, & \text{when } s^2_y = 0, \end{cases}$$

as $\sigma^2_x \to 0$. Thus, the local marginal likelihood function converges to

$$\begin{cases} 0, & \text{when } s^2_y > 0, \\ \frac{1}{\sqrt{2\pi t^2_x}} \exp \left\{ -\frac{1}{2 t^2_x} (\bar{y}_x - \hat{y}_x)^2 \right\}, & \text{when } s^2_y = 0, \end{cases}$$

when $\sigma^2_x \to 0$. \hfill \Box

**Lemma 2.** For $0 < \sigma^2_x < \infty$, at least one positive extremal value of the local marginal likelihood function exists when $s^2_y > 0$, while at most one such a value exists otherwise.

**Proof.** The first order condition for maximizing the log local marginal likelihood function is

$$m' \left( \{Y_i\}_{i \in G_x} \mid \{x_i\}_{i \in G_x}, \sigma^2_x \right) m \left( \{Y_i\}_{i \in G_x} \mid \{x_i\}_{i \in G_x}, \sigma^2_x \right) = 0. \quad (16)$$

The left hand side is calculated as

$$\frac{1}{2} \frac{(\tau^2_x)^2}{(\sigma^2_x)^3} n_x \left\{ a_1 \left( \sigma^2_x \right)^3 + a_2 \left( \sigma^2_x \right)^2 + a_3 \sigma^2_x + a_4 \right\}, \quad (17)$$
where

\[ a_1 = -\frac{1}{t_x^4}, \quad a_2 = \frac{1}{t_x^2} \left(1 - 2n_x\right) + \frac{1}{t_x^4} \left\{s_y^2 + (\bar{y}_x - \hat{y}_x)^2\right\}, \quad (18) \]

\[ a_3 = -n_x^2 + n_x + \frac{2n_x s_y^2}{t_x^2}, \quad a_4 = n_x^2 s_y, \quad (19) \]

Then, the extremal values are solutions to the following cubic equation:

\[ a_1 \left(\sigma_x^2\right)^3 + a_2 \left(\sigma_x^2\right)^2 + a_3 \sigma_x^2 + a_4 = 0. \quad (20) \]

When \( s_y^2 > 0 \), \( a_4 > 0 \). Thus, at least one solution to this cubic equation is positive when \( s_y^2 > 0 \).

Remark 1. Let \((\alpha, \beta, \gamma)\) be three solutions to this cubic equation. Then, it is factorized as \((\sigma_x^2 - \alpha)(\sigma_x^2 - \beta)(\sigma_x^2 - \gamma) = 0\). By comparing coefficients, \(-\alpha \beta \gamma = \frac{a_4}{a_1} < 0\). This implies that solutions are nonzero and either one of three possibilities: (i) three distinct real solutions, (ii) one real and two complex conjugate solutions, or (iii) three real, but at least two are the same, solutions. Then, we conclude that at least one solution is positive.

When \( s_y^2 = 0 \), the cubic equation reduces to

\[ \left(\sigma_x^2\right)^3 + \left\{t_x^2 - (\bar{y}_x - \hat{y}_x)^2\right\} \left(\sigma_x^2\right)^2 = 0. \quad (21) \]

If \( t_x^2 - (\bar{y}_x - \hat{y}_x)^2 \geq 0 \), no extremal value exists for the local marginal likelihood function over \( 0 < \sigma_x^2 < \infty \). If, on the other hand, \( t_x^2 - (\bar{y}_x - \hat{y}_x)^2 < 0 \),

\[ \sigma_x^2 = -t_x^2 + (\bar{y}_x - \hat{y}_x)^2 \quad (22) \]

is the extremal value within the same range.
References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer series in statistics. New York: Springer-Verlag.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics* 32(2), 407–499.

Fan, J. and I. Gijbels (2003). *Local Polynomial Modelling and Its Applications*. Boca Raton: Chapman & Hall/CRC.

MacEachern, S. N. and K. Miyawaki (2022). A regression approach to the two-dataset problem.

Miyawaki, K. and S. N. MacEachern (n/a). Economic variable selection. *Canadian Journal of Statistics* n/a(n/a), n/a. to appear.