Bayesian Cross-Validation Choice and Assessment of Statistical Models

by

Fatemah Ali Alqallaf

B.A., Kuwait University, 1994

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

September 1999

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Oct, 6th, 99

DE-6 (2/88)

Abstract

This thesis will be concerned with application of a *cross-validation* criterion to the choice and assessment of statistical models, in which observed data are partitioned, with one part of the data compared to predictions conditional on the model and the rest of the data.

We develop three methods, gold, silver, and bronze based on the idea of splitting data in the context of measuring prediction error; however, they can also be adapted for model checking. The gold method uses analytic calculations for the posterior predictive distribution; however, the silver method avoids this mathematical intensity, instead simulating many posterior samples, and the bronze method reduces the amount of sampling to speed up computation.

We also consider the Bayesian p-value in which the posterior distribution can be used to check model adequacy, in the context of cross-validation with repeated data splitting. Application to examples is detailed, using the discussed methodologies of estimation and prediction.

Contents

List of Figures

Acknowledgements

I would like to thank my supervisor Paul Gustafson for his guidance, help and support throughout the development of this thesis. Paul was always willing to give his advice and intuitive ideas which were greatly appreciated. I would also like to thank Bertrand Clarke for his careful reading of the manuscript and for his neverending encouragement and advice. Without these two people I may never have succeeded in finishing this degree. It is a pleasure to thank all fellow graduate students, especially Matias for his invaluable help with computing, and various other problems along the way. Many thanks also go to Nathan for his encouragement, and help. Finally, I would like to thank my husband Mohammed, for his support and continued patience.

FATEMAH ALI ALQALLAF

The University of British Columbia

September 1999

To my parents, and to Mohammed

Chapter 1

Introduction

Statistical procedures can often adequately predict the data that generated them, however they do not always perform so well when used to predict new data. Cross-validation is a data oriented method which seeks to improve reliability in this context. This thesis is concerned with application of a cross-validation criterion to the choice and assessment of statistical models. The concept of such assessment is an old one but nevertheless useful. It consists of the division of the data sample into two subsamples, the training sample and the validation sample. The estimation of a statistical model is based on the training sample, and then the assessment of its performance is made by measuring its predictions on the the validation sample. Herzberg (1969) made a detailed theoretical and numerical study of predictor construction methods, using cross-validatory assessment. Stone (1974) applied a generalized form of the cross-validation criterion to the choice and assessment of models using the data-analytic concept. Other examples include variable selection in linear

regression using cross-validated predictive squared error (Hjorth, 1994).

Model selection has attracted the attention of many researchers. Cross-validation is one of the most well-known methods. There are other methods for model selection, such as the Akaike information criterion (AIC) (Akaike 1974; Shibata 1981), the C_p (Mallows 1973), the jackknife, and the bootstrap (Efron 1983, 1986). All these methods are asymptotically equivalent to the cross-validation with validation sample size $n_v \equiv 1$ (Stone 1977a; Efron 1983), however, and thus they share the same deficiency; that is they are inconsistent. In the problem of selecting models, this deficiency of the cross-validation can be rectified by using a cross-validation with a large n_v depending on n. Shao (1993) showed that this inconsistency can be rectified by choosing n_v such that $n_v/n \to 1$ as $n \to \infty$. Herzberg and Tsukanov (1986) did some simulation comparisons between the cross-validation procedures with $n_v \equiv 1$ and $n_v \equiv 2$. They found that the leave-two-out cross-validation is sometimes better than the leave-one-out cross-validation, although the two procedures are asymptotically equivalent in theory. See also Geisser (1975), Burman (1989), and Zhang (1993).

There are a number of different cross-validation methodologies, and they largely differ in how the partitions are chosen. Much work has been done on the ordinary cross-validation, for example Stone (1974, 1977), Bowman (1984) and Härdle and Marron (1985). The approach taken here is to delete a single data point (x_1, y_1), fit the model to the reduced data set, and using the estimated model parameters, obtain a prediction $\hat{y}_1$ for the missing ob-

servation. This is repeated for each point (x_i, y_i). Then the cross-validatory score $C = \sum L(y_i, \hat{y}_i)/n$ is computed where L is an appropriate loss function; typically $L(y, \hat{y}) = (y - \hat{y})^2$. Clearly for large data sets the computational requirements may become excessive and alternative strategies are adopted, rather than omitting points singly. Burman (1989) considered two techniques in a study of the optimal transformations of variables, "v-fold" cross-validation and repeated learning-testing methods. The v-fold cross-validation uses v disjoint validation partitions by dividing the data randomly in v groups so that their sizes are as nearly equal as possible. This method with $v = n$ is the leave-one-out cross-validation. For model selection in linear regression, Burman (1989), Shao (1993), and Zhang (1993) have each investigated a particular cross-validation procedure where M partitions are generated at random independently with a fixed fraction β being used as validation samples, and $1 - \beta$ being used for parameter estimation in each case, Burman calls this repeated-learning-testing, and Shao calls it "Monte Carlo cross-validation". The main difference between the repeated-learning-testing method and the v-fold method is that with the former each data point may be used as a validation point more than once. These methods can be used in the place of ordinary cross-validation whenever the latter is computationally very expensive and asymptotically inconsistent. A comparative study of these methods has been carried out in detail by Burman (1989).

Consider the problem of selecting a model having the best predictive ability among a class of models. Cross-validation can be a method for model

selection according to the predictive ability of the models (Shao, 1993). Picard and Cook (1984) examined the cross-validation assessments of the predictive ability of a fitted multiple-regression model. Smyth (1998) applied the cross-validation approach to model selection in the sense that models are judged directly on their out-of-sample predictive performance.

The approach and methods we will introduce later are designed to fit with Bayesian statistics, which uses probability theory to describe uncertainty about the parameters and the observables. In the Bayesian approach, information and beliefs that are available before data are observed contribute to the specification of a *prior distribution.* After the data are observed the prior distribution is updated to the *posterior distribution* which is proportional to the product of the likelihood and the prior distribution. Inferences are made on the basis of the posterior distribution. The recent revolution in Bayesian statistics is due to Markov Chain Monte Carlo (MCMC) algorithms. These algorithms allow a user to draw inferences from a complex posterior distribution on a high-dimensional parameter space, by simulating a Markov chain with the posterior distribution as the stationary distribution. Then the posterior quantities are estimated from the simulation output, as the chain converges to its stationary distribution. Review material can be found in Neal (1993), Smith and Roberts (1994), Tierney (1994), and Besag et al. (1995).

Checking the model is critical in the statistical analysis. Therefore good Bayesian analysis should check how well the model fits the data and how good the posterior inferences are. In Bayesian statistics, a model can be checked

in at least three ways: (1) examining sensitivity of inferences to reasonable changes in the prior distribution and the likelihood; (2) checking that the posterior inferences are reasonable given the substantive context of the model; and (3) checking that the model fits the data. We address the third of these concerns using the posterior predictive distribution for a *discrepancy*, which extend classical test statistics to allow dependence on unknown parameters. Posterior predictive assessment was introduced by Guttman (1967), applied by Rubin (1981), and given a formal Bayesian definition by Rubin (1984). Gelman, Meng, and Stern (1996), GMS, contributed to further develop Rubin's (1984) work on posterior predictive assessment, as one possible way of measuring any discrepancies that may exist between the observables and their predictive distributions under a given model specification. This methodological contribution is the adoption of more general discrepancies, which allows more direct assessment of the discrepancy between data and the model.

That there is a pressing need to perform such calibrative work with any model has become clearer at a time when Markov Chain Monte Carlo (MCMC) methods permit the realization of modeling in much more complicated settings than ever before. In fact, Bayesians are dwelling so much these days on diagnostics for MCMC convergence-monitoring that a strong reminder of the value of model-checking is all the more welcome.

The GMS approach appears to be simple conceptually and computationally, and connected well to the classical goodness-of-fit methods with which most researchers are familiar. It is also very general, applicable for comparing

observations with model predictions in any form. Their applied work has benefited from the application of this method, as documented in their paper and also in many examples of application of this method in Gelman, Carlin, Stern and Rubin (1995) (also see Belin and Rubin (1995) and Upadhyay and Smith (1993)). Meng (1994) discusses a similar method for testing parameters values within a model rather than for the entire model. West (1986) and Gelfand, Dey, and Chang (1992) also present posterior predictive approaches to model evaluation, in the contexts of sequential data and cross-validation of the existing data set, respectively, rather than comparing to hypothetical replications of the entire data set.

We consider the prediction error by randomly splitting the data many times and averaging the squared prediction errors over the splits. Shao (1993) also averaged the squared prediction errors over the splits. We also consider the Bayesian p-value in which the posterior distribution can be used to check a statistical model, in the same context of cross-validation with repeated data splitting. In Smyth (1998), cross-validated likelihood is investigated as an appropriate score function for model selection in probabilistic clustering, in particular for choosing the number of component densities in finite mixture models. He splits the data $\mathbf{x}$ into $\mathbf{x}^{test}$ and $\mathbf{x}^{train}$, for each split compute the likelihood $f(\mathbf{x}^{test}|\hat{\theta}_{train})$, where $\hat{\theta}_{train}$ generated given the train sample. In our approach, we also split the data into samples, which we call validation and training samples. Then compute the prediction function, in the sense of Bayesian approach, $f(\mathbf{y}^{test}|\mathbf{x}^{test}, \hat{\theta}_{train})$.

The thesis is organized as follows. Chapter 2 consists of definitions, a discussion of computational issues, and an analytical illustration of the cross-validatory methodology for assessing prediction error. Chapter 3 presents a detailed illustration of measuring the prediction error in two regression models. Chapter 4 exhibits the posterior predictive assessment of model fit through the Bayesian p-value, including a brief comparison to other versions of p-values. Chapter 5 provides discussion of various issues related to the method of measuring the prediction error and the posterior predictive p-value.

Chapter 2

Methodology for Measuring Prediction Error

We propose methods for measuring prediction error in the context of cross-validation analysis. In this chapter we present our three methods, gold, silver and bronze, and the basic methodology of our approach in the context of measuring prediction error; however, they can also be adapted for other purposes, such as model checking, as will be seen in the later chapter on Bayesian p-values. We introduce the analytical theory of the gold method, basic methodology for the silver and the bronze methods, and the theory pertaining to the silver method.

2.1 Reference Distribution

In this section we introduce definitions and notations which will be used in the theoretical illustration of the posterior predictive approaches to prediction error. We use predictive distributions in a somewhat different way, by cross-validating the modeling process, in which we set aside some of the data, fit the current model to the rest, and then locate the observed values of the set-aside data in their respective predictive distributions given the chosen model.

2.1.1 Notations

Let y be the observed data, and θ be the vector of parameters. To distinguish between the observed data y and the replicated data we define y^{rep} as the replicated data that could have been observed. That is, y^{rep} is the data we would see if the experiment that produced y were replicated with the same model and the same value of θ that produced the observed data. Since the cross-validation methodology will be used, we introduce a split indicator s. This splits the data into training and validation sets. Specifically s is a vector of zeros and ones, the zeros indicating observations that go into training sample and the ones indicating observations that go into the validation sample. We consider the joint distribution of y^{rep}, θ, and the split s,

$$p(y^{rep}, \theta, s) = p(y^{rep}|\theta, s)p(\theta|s)p(s), \tag{2.1}$$

as a reference distribution in our methodological approach. The following explanation may be helpful.

The conditional distribution $p(y^{rep}|\theta, s)$ is the probability distribution of the replicated data y^{rep} given θ and the training sample indicated by s, which is equivalent to the distribution of the replicated data set y^{rep} given θ, expressed as $p(y^{rep}|\theta, s) = p(y^{rep}|\theta)$. In our approach we typically use only the validation part of y^{rep}, which we compare with the realized validation set. So we denote the distribution of the replicated data in the case of validation sample by $y_V^{rep}|s$ which is distributed as the predictive distribution for the validation sample given the training sample. And the distribution $p(\theta|s)$ is the posterior distribution of θ based on the training set y_T. The probability distribution over the split $s = (T, V)$ is $p(s)$. We chose to split the data in two with half being the training set and half being the validation set. All such 50-50 splits are equally likely under $p(s)$. Via cross-validation, we can estimate the predicted error. But instead of splitting the data once we split the data many times. Then by averaging over the splits can estimate the expected sum of squared predicted error with respect to the joint distribution $p(y^{rep}, \theta, s)$, which is given by

$$W = E\left\{\|y_V^{rep} - y_V\|^2\right\}. \tag{2.2}$$

That is, we view W as a summary of predictive performance. There are ways to estimate W by using the idea of splitting the data set. In the next section we introduce the three methods used to estimate W.

2.2 Description of the Methods

In principle we are interested in various expectations with respect to the joint distribution $p(y^{rep}, \theta, s)$. We develop our methodologies based on the idea of splitting data to estimating the expected prediction error, but they may also be extended to include other posterior predictive checks. In this section we briefly discuss the three methods of estimating W. The gold method uses analytic calculations for the posterior predictive distribution; this requires computational effort that most of the time can not be handled by hand. However, the silver method avoids this mathematical intensity, instead simulating many posterior samples. The bronze method reduces the amount of sampling to speed up computation. In the description of the techniques used for each method, we show how to conduct these methods, and then one can decide when it is appropriate to implement one rather than another.

2.2.1 Gold Method

To estimate the expected sum of squared predicted error,

$$W = E\left\{\sum (y_{V_i}^{rep} - y_{V_i})^2\right\},$$

we do some analytic calculations on W to get,

$$\begin{aligned} W &= E_s E_{y^{rep}|s}\left\{\sum_i \left(y_{V_i}^{rep} - E(y_{V_i}^{rep}|s) + E(y_{V_i}^{rep}|s) - y_{V_i}\right)^2\right\} \\ &= E_s\left\{\sum_i E_{y^{rep}|s}\left(y_{V_i}^{rep} - E(y_{V_i}^{rep}|s)^2\right)\right\} + E_s\left\{\sum_i E_{y^{rep}|s}\left(E(y_{V_i}^{rep}|s) - y_{V_i}\right)^2\right\} \end{aligned}$$

where the cross-term

$$E_{y^{rep}|s}(y_{V_i}^{rep} - E(y_{V_i}^{rep}|s))(E(y_{V_i}^{rep}|s) - y_{V_i}) = 0,$$

since $E_{y^{rep}|s}(y_{V_i}^{rep} - E(y_{V_i}^{rep}|s))$ can be written as

$$E_{y^{rep}|s}(y_{V_i}^{rep}) - E_{y^{rep}|s}(E(y_{V_i}^{rep}|s))$$

$$= E_{y^{rep}|s}(y_{V_i}^{rep}) - E_{y^{rep}|s}(y_{V_i}^{rep}).$$

Now,

$$W = E_s\left\{\sum_i Var(y_{V_i}^{rep}|s)\right\} + E_s\left\{\sum_i \left(E(y_{V_i}^{rep}|s) - y_{V_i}\right)^2\right\}.$$

To estimate W, we need to perform the following. First, sample independent splits $s_1, s_2, \ldots, s_r$ from $p(s)$. Then estimate W by

$$\widehat{W}_G = \frac{1}{r}\sum_{j=1}^{r}\left\{\sum_i Var(y_{V_i}^{rep}|s_j)\right\} + \frac{1}{r}\sum_{j=1}^{r}\left\{\|E(y_V^{rep}|s_j) - y_V\|^2\right\}. \qquad (2.3)$$

To compute $\widehat{W}_G$ (G stands for Gold) we need to compute the mean $E(y_i^{rep}|s_j)$ and the variance $Var(y_i^{rep}|s_j)$ of the predicted distribution of the i-th observation in the validation sample, given the training data set based on split s_j.

Computation of the predicted distributions can be performed analytically for some simple models such as normal linear regression, as will be seen in the next chapter on examples of measuring prediction error. But in complicated models which may arise in practical applications, such as a hierarchical mixture model, we can not compute $E(y_{V_i}^{rep}|s_j)$ and $Var(y_{V_i}^{rep}|s_j)$ by hand. Therefore it is more easily accomplished via simulation. This is some extra

computational burden, but simulation is a standard tool for Bayesian analysis with complex models, and often it is the only option. The usual Bayesian simulation provides a set of draws of θ from the posterior distribution, $p(\theta|y)$. In the normal linear model the posterior distribution can be easily simulated using exact sampling; However, when the posterior distribution cannot be simulated directly, an indirect simulation method such as the Metropolis algorithm (MA) is often used. On the basis of this fact, we introduce the silver and the bronze methods, using simulation methods to sample from the posterior distribution so that the mathematical computation effort is lessen.

2.2.2 Silver Method

In this method we sample from posterior predictive distributions to estimate the predicted error. We consider the following algorithm for comparing the realized validation data vector $\mathbf{y}_V$ to $\mathbf{y}_V^{rep}$:

1. Generate random splits $s_1, \ldots, s_r$, from the $p(s)$ that we introduced before.

2. For each s, construct a sample of the parameter θ from the posterior distribution $p(\theta|s)$ based on the training sample.

   - Given θ, draw a replication data set y_V^{rep} from the sampling distribution $p(y^{rep}|\theta, s)$ given the training sample.

   - Having obtained y_V^{rep}, we can estimate the expected squared of predicted error $E(\|y_V^{rep} - y_V\|^2|s_i)$ by $\hat{w}(s_i) = \frac{1}{m}\sum_{j=1}^{m} \|y_{V_j}^{rep} - y_{V_j}\|^2$,

where m is the number of θ's generated using the i-th split.

3. Averaging over the multiple splits, the silver estimate of the predicted error is

$$\widehat{W}_S = \frac{1}{r}\sum_{i=1}^{r}\hat{w}(s_i). \tag{2.4}$$

In the silver method as we have seen, for every split there is a new simulation from the posterior distribution, since we simulate draws from the posterior distributions for as many as splits we have. We thought that we should reduce the number of simulations and that the number of simulations should not depend on the number of splits, because by simulating every time we split the data we slow down the estimating procedure, especially when using complicated models. In the next section we introduce the method that reduces the number of simulations.

2.2.3 Bronze Method

In this section, we propose the use of importance sampling to estimate the predicted error and improve upon the silver method in terms of computation time. The bronze method uses a similar concept as the silver method; however, we try to improve on the idea of drawing samples from the posterior distribution of θ based on one portion or split of the data set. Instead, we simulate a sequence of draws for θ based on the entire data set. Using importance sampling we make a correction so that the draws look as though they are drawn from the posterior distribution based on one portion of the data set, the training set. This approach is shown in mathematical terms below.

The basic computational strategy

Combining the densities $f(y_i|\theta)$ with a prior density, $p(\theta)$, yields the posterior density

$$p(\theta|s) \propto \left\{ \prod_{i \in T} f(y_i|\theta) \right\} p(\theta),$$

which is based on a specific training set. Here we roughly approximate $p(\theta|s)$ by a "squashed" version of the posterior based on all the data,

$$q(\theta|y) \propto \left\{ \prod_{i=1}^{n} f(y_i|\theta) \right\}^{\alpha} p(\theta),$$

where n is the size of the data set. It is intuitively sensible to choose $\alpha = \frac{1}{2}$ for an effective sample size of $\frac{n}{2}$, to match the training sample size, as we split the data into two halves.

A sample of $m > 1$ draws for a better approximation of posterior sampling can be simulated as follows.

1. Sample values $\theta_1, \ldots, \theta_m$ from $q(\theta|y)$, then sample $y_{i1}^*, \ldots, y_{in}^*$ from the$f(y|\theta_i)$.

2. Draw different splits $s_1, \ldots, s_r$ from $p(s)$. For each split s_j the probability of sampling each θ_i is proportional to the importance weight

$$w_{i,j} = \frac{w^*(\theta_i)}{\Sigma_{i=1}^{m} w^*(\theta_i)},$$

where

$$w^*(\theta_i) = \frac{p(\theta_i|s)}{q(\theta_i)} = \frac{\Pi_{j \in T} f(y_j|\theta_i)}{\Pi_{j=1}^{n} \{f(y_j|\theta_i)\}^{\alpha}}.$$

These weights $w_{i,j}$ which depend on θ_i and split j make the $\theta's$ look like a sample from $p(\theta|s_j)$ based on the training set.

3. For every different split, use the split indicator to split the replicated data into validation and training samples, $y^*_{i_V}$ and $y^*_{i_T}$. Also using the same split, the observed data are split into validation y_V, and training y_T samples.

Now for each split, we have $y^*_{i_V}$ which is equivalent to $y^{rep}_{V_i}$, the replicated observation of the validation sample with probability w. Therefore, to estimate the expected sum of squared predicted error

$$W = E\{\|y^{rep}_V - y_V\|^2\},$$

we should use these weights in our estimator. If n is the number of data points, then the bronze estimator is

$$\begin{aligned} \widehat{W}_B &= \frac{1}{r}\sum_{j=1}^{r}\sum_{i=1}^{m}\sum_{k=1}^{n} I_{s_j}(k)(y^{*i}_{V_k} - y^i_{V_k})^2 w_{i,j} && (2.5) \\ &= \sum_{i=1}^{m}\sum_{k=1}^{n}\left[\frac{1}{r}\sum_{j=1}^{r} I_{s_j}(k) w_{i,j}\right](y^{*i}_{V_k} - y^i_{V_k})^2, && (2.6) \end{aligned}$$

where $I_{s_j}(k)$ is an indicator, taking the value 1 is y_k is in the validation sample for split s_j, and 0 otherwise. Analogously, for each s we calculate the difference error

$$e_s = \sum_{i=1}^{m}\left[\sum_{k=1}^{\frac{n}{2}}(y^*_{v_k} - y_{V_k})^2\right] w_{i,j}.$$

Then the bronze estimate of the predicted error is

$$\widehat{W}_B = \frac{1}{r}\sum_{i=1}^{r} e_{s_i}. \qquad (2.7)$$

2.3 Theory Pertaining to the Silver Method

In this section we show the theoretical considerations for the silver method, regarding the number of splits versus the size of the posterior samples for each split. We discuss the question of whether a few big samples or many small samples is optimal. We aim to analyze this concept theoretically under conditions that make our choice of the size of the posterior samples and the number of splits optimal.

Let $x_{i1}, x_{i2}, \ldots, x_{iJ}$ be draws of a posterior quantity for i^{th} split, where $i = 1, \ldots, I$ correspond to different splits. To estimate the value $\theta = E(x)$ we may use the estimator

$$\hat{\theta} = \frac{1}{I} \sum_{i=1}^{I} \left\{ \frac{1}{J} \sum_{j=1}^{J} x_{ij} \right\} \tag{2.8}$$

Now how big should I and J be relative to one another so that $\hat{\theta}$ is a good estimator with small variance ?

Suppose that we use Markov Chain Monte Carlo (MCMC) to simulate draws from the posterior distributions. Then for each split there will be some burn-in time of say w observations, so we can think of the cost (c) of sampling as

$$c = I(J + w). \tag{2.9}$$

We may reformulate the problem as one of minimizing the variance of $\hat{\theta}$ subject to a fixed cost under the following reasonable assumptions.

As an ANOVA-style set-up, let

$$x_{ij} = \mu_i + \epsilon_{ij}$$

where $\mu_i \sim N(\mu, \tau^2)$, reflects split-to-split variation, and $\epsilon_{ij} \sim N(0, \sigma^2)$. All x_{ij}'s have the same distribution, with x_{ij}'s independent for different splits,

$$corr(\epsilon_{ij}, \epsilon_{kl}) = 0, \quad \text{for} \quad i \neq k,$$

but x_{ij}'s are dependent for the same split,

$$corr(\epsilon_{ij}, \epsilon_{ik}) = \rho^{|j-k|}, \quad \text{for} \quad i \neq k.$$

Now to minimize the variance of the estimator (2.8), an analysis proceeds as follows.

For simplicity let $\bar{x}_{i.} = \frac{1}{J}\sum_{j=1}^{J} x_{ij}$, so that

$$\hat{\theta} = \frac{1}{I}\sum_{i=1}^{I} \bar{x}_{i.},$$

and

$$var(\hat{\theta}) \quad = \quad \frac{1}{I^2}\sum_{i=1}^{I} var(\bar{x}_{i.}),$$

since the covariance between different splits is zero, i.e. $cov(\bar{\epsilon}_{i.}, \bar{\epsilon}_{j.}) = 0$ for all $i \neq j$. Now,

$$\begin{aligned} var(\bar{x}_{i.}) &= var\{\mu_i + \bar{\epsilon}_{i.}\} \\ &= \tau^2 + \frac{1}{J^2}\sum_{j=1}^{J} var(\epsilon_{ij}) + 2\sum_{j<k} cov(\epsilon_{ij}, \epsilon_{ik}) \\ &= \tau^2 + \frac{1}{J^2}\left\{J\sigma^2 + 2s\sigma^2\right\}, \end{aligned}$$

where $s = (J-1)\rho + (J-2)\rho^2 + \ldots + (1)\rho^{J-1}$.

Using the large J approximation

$$\begin{aligned} s &\approx J\sum_{j=1}^{J-1}\rho^j \\ &\approx J\frac{\rho}{1-\rho}, \end{aligned}$$

gives

$$\begin{aligned} Var(\bar{x}_{i.}) &\approx \tau^2 + \frac{1}{J^2}\left\{J\sigma^2 + 2\sigma^2 J\frac{\rho}{1-\rho}\right\} \\ &\approx \tau^2 + \frac{\sigma^2}{J}\left\{1 + 2\frac{\rho}{1-\rho}\right\}. \end{aligned}$$

Now

$$\begin{aligned} Var(\hat{\theta}) &\approx \frac{1}{I^2}\sum_{i=1}^{I}\left\{\tau^2 + \frac{\sigma^2}{J}\left(1 + \frac{2\rho}{1-\rho}\right)\right\} \\ &\approx \frac{J+w}{c}\left\{\tau^2 + \frac{\sigma^2}{J}\left(1 + \frac{2\rho}{1-\rho}\right)\right\} \end{aligned}$$

To find the minimum of the variance we differentiate with respect to J

$$\frac{\partial}{\partial J}var(\hat{\theta}) \approx \frac{\frac{\sigma^2(1+2\frac{\rho}{1-\rho})}{J} + \tau^2}{c} - \frac{(J+w)\sigma^2(1+2\frac{\rho}{1-\rho})}{cJ^2}$$

Thus solving $\frac{\partial}{\partial J}var(\hat{\theta}) = 0$, and clearly the second derivative yields

$$J = \sqrt{\frac{\sigma^2 w(1+\rho)}{\tau^2(1-\rho)}}$$

as the optimal posterior sample size, with I then determined by the cost constraint.

In light of these results, we can conclude that the optimal length of the chain of the parameters drawn from a posterior distribution is proportional to the square root burn-in time

$$\text{chain length} \propto \sqrt{\text{burn-in time}}.$$

For the silver method we need to draw posterior samples of the parameters every time we split the data set. Based on the mathematical results the optimal length of the chain depends on the square root of the burn-in time. Whether or not this chain length should be used when using the silver method is entirely dependent on the user's confidence, we just want to show how long the chain should be from a mathematical point of view. For the bronze method we simulate all the parameters once before splitting the data set, so maybe people would prefer one long chain at once as in the bronze method and tolerate the computational intensity in computing the importance weights for each split.

Chapter 3

Examples of Measuring Prediction Error

3.1 Introduction

Prediction errors for a new set of data can be used to assess the quality of model-based predictions. The approach taken in this chapter is to use standard output from regression analysis to get the prediction function. We illustrate the three methods of estimating expected prediction error with examples of constructing regression models. This chapter introduces Bayesian model building and inference for normal and Weibull regression models. Using a Bayesian approach we chose a noninformative prior distribution, with the understanding that this is no more than a convenient assumption for the purposes of exposition and can be extended to informative prior distributions. The Bayesian approach to prediction is potentially very rich, as one can focus attention on

the whole predictive distribution of unobserved future values, given predictors, prior data, and information. This can be more informative than just looking at point predictors. We compute the posterior distributions for these models, interpret the results, and compare the estimates of the prediction error for the different methods.

Throughout, we describe computations used in each of the methods for estimating expected prediction error. In particular, we show how simple simulation methods can be used to draw samples from posterior and predictive distributions, to incorporate uncertainty in the model parameters, and to draw samples for posterior predictive checks. We also check the sampling variability of the prediction error estimates of each method.

3.2 Normal Regression Model

In this section we investigate the three methods using an example of normal linear regression with a noninformative prior distribution. It is interesting both theoretically because of the elegance of the underlying theory, and from an applied point of view because of the wide variety of uses of linear regression. The normal model had been chosen basically to be a test case in which it is convenient to compare the three methods.

Data set

The data set used in this analysis was obtained from Weisberg (1995, pp. 144-5). The mammals data contains two variables, the average brain

weights and the average body weights for 62 species of mammals. These data were taken from a larger study and were collected for another purpose (Allison and Cicchetti, 1976). Once the parameters of a regression model have been estimated, they can be used to predict future observations of the brain weight from the explanatory variable, the body weight. So we thought this was a good example to test our procedures for estimating the prediction error by splitting the data into two halves. Given one half we predict the other half. Before we start, we check if any transformation is needed for the linear fit. We shall consider the problem of modeling brain weight y as a function of body weight x. An initial attempt to graph brain weight (in grams) versus body weight (in kilograms), as given in Figure (3.1), indicates that some transformation is required. The plot shows that the relationship between the body weight and the brain weight is not linear, with most of the points in the plot being jammed into the lower left corner and only a few stragglers elsewhere. Because of the wide variation of both variables, an exponential fit seems to be a good candidate. Assume that the correct functional relationship is of the form brain weight $= \alpha_0(\text{body weight})^{\alpha_1}$. The scatter plot in the log scale given in Figure (3.1), suggests that there is a strong linear relationship in the log scale. The transformation seems to achieve linearity and constant variance. Taking logarithms will reduce the model to a normal linear model with additive errors,

$$log(Y) = log(\alpha_0) + \alpha_1 log(X) + e,$$

where e have zero expectation and constant variance σ^2. To simplify the notations we consider $v = log(y)$, $u = log(x)$, $\beta_0 = log(\alpha_0)$, and $\beta_1 = \alpha_1$. So

we may rewrite the normal linear model as

$$V = \beta_0 + \beta_1 U + e. \tag{3.1}$$

The data points v follow a normal distribution with mean $\mu = \beta_0 + \beta_1 \mathbf{u}$ and variance σ^2. For a vector $\mathbf{v} = (v_1, \ldots, v_n)$ of iid observations, the likelihood is

$$p(v|\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(v_i - \beta_0 - \beta_1 u_i)^2\right). \tag{3.2}$$

Since we are conducting a Bayesian analysis, we need a prior distribution in addition to the likelihood based on (3.2). This will yield a posterior distribution on the unknown parameters $\theta = (\beta_0, \beta_1, \sigma^2)$ of the normal linear model and enable us to estimate them. The prior distribution on which our inference will be based is a standard noninformative prior,

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Now we illustrate the three methods through this example, and try to conduct a fair comparison between them. We thought if we could draw a histogram of repeated realizations $\widehat{W}$ (the estimate of the expected prediction error) we would gain some knowledge about the sampling distribution. The better the estimate the less variability there is in the sampling distribution. This can be done by repeating the process of getting $\widehat{W}$'s many times for different splits. Then by drawing histograms of the $\widehat{W}$'s we can visually compare the sample variances between them in each method.

We decided to have 100 $\widehat{W}$'s for each method. However, for comparison purposes we fix the split indicators that we use to split the data to be the

same for the three methods. Before we discuss each method individually we may roughly explain the procedure of fixing the splits to draw histograms of 100 $\widehat{W}$'s.

We agreed on using 50 splits to compute each $\widehat{W}$, and we have 62 data points. To generate all split indicators for the 100 $\widehat{W}$'s, we create an array with dimensions $100 \times 50 \times 62$. To compute $\widehat{W}$ we need to split the data 50 times and get the same samples of the validation and the training for each split in each method. Using that array properly we will make sure that each $\widehat{W}$ in different methods is using the same splits of the data set.

For the gold method, we need to compute the mean and the variance of the predictive distribution of each observation in the validation sample, given the training data based on the split. In the ordinary linear model, working through this mathematical computation is quite easy. The silver method uses many posterior samples, we should simulate θ per split, then based on θ generate the replicated data set. We first try to sample one θ per split. But this may lead to some criticism so we decide to sample 100 θ's per split for a more confident results as sampling one θ only is less reliable. Finally, for the bronze method we generate 100 θ's at once and simulate replicated data for each of them. Then based on the same splits we estimate the 100 predicted errors $\widehat{W}$ by splitting the replicated data set and estimating the expected sum of squared error. Also we simulate 50 θ's for the bronze method, so as to have a fair comparison to the silver method in the case of sampling one θ per split.

3.2.1 The Gold Method

The gold method is easy to apply for our ordinary linear model because the full conditional posterior distributions $p(y^{rep}|\theta, s)$ and $p(\theta|s)$ have standard forms with the noninformative prior distribution, and the mean and variance of the predictive distribution can be easily computed. (For further details see Appendix A.)

We computed a set of 100 estimates which estimate the expected sum of predicted error W by (GOLD W). The simulation is set up as follows.

- For each iteration $l = 1, \ldots, 100$, we split the data set 50 times by sampling the splits indicator s_i, $i = 1, \ldots, 50$. For each split we compute a predictive mean and variance for the validation sample given the training sample, $E(y_{v_i}^{rep}|y_{T_i})$, $var(y_{v_i}^{rep}|y_{T_i})$ using (A.7) and (A.8), so the expected predicted error is

$$\hat{w}_i = \|y_{V_i} - E(y_{V_i}^{rep}|y_{T_i})\|^2 + \sum_i var(y_{V_i}^{rep}|y_{T_i}).$$

- Average over splits $\widehat{W} = \frac{1}{50}\sum_{i=1}^{50} \hat{w}_i$.

- Repeat this procedure 100 times to obtain 100 $\widehat{W}$'s, and draw the histogram as shown in Figure (3.2).

The histogram shows little variability amongst the 100 estimates.

3.2.2 The Silver Method

We illustrate the silver method in estimating the predictive error with the mammals example, and we consider the same ordinary linear regression model.

As with the normal distribution with unknown mean and variance, we simply use the result derived by Gelman et al. (1995) to determine the posterior distribution for β, conditioning on σ^2, and then the marginal posterior distribution for σ^2. That is we factor the joint posterior distribution for β and σ^2 as

$$p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)p(\sigma^2|y),$$

following the factorization of the posterior distribution given by (A.2) and (A.5). Using exact sampling to estimate the posterior parameters of the linear regression on the mammals data set, we first draw a random value of $\sigma^2 \sim Inv{-}\chi^2(29, s^2)$, as $29s^2$ divided by a random draw from the $\chi^2(29)$ distribution (see Appendix B). Then given this value of σ^2, we draw β from its conditional posterior distribution, $N(\hat{\beta}, V_\beta\sigma^2)$, where $\hat{\beta}$ is from (A.3).

Based on 50 simulated values of (β, σ^2), we estimate the expected sum of predicted error W. We carried out the idea of splitting the data as follows.

For each iteration $l = 1, \ldots, 100$:

1. Sample multiple splits s_i, $i = 1, \ldots, 50$.

2. For a single split, compute $\hat{w}(s_i) = \sum_{i=1}^{31}(y_{V_i}^{rep} - y_{V_i})^2$.

3. Having performed the step 50 times, we obtain $\widehat{W}$ by averaging over the 50 $\hat{w}(s)'s$,

$$\widehat{W} = \frac{1}{50}\sum_{i=1}^{50}\hat{w}(s_i).$$

We combine the results from the 100 iterations, and draw a histogram of the 100 $\widehat{W}$'s as shown in Figure (3.2). The silver standard estimates are more

spread out than the gold standard estimates, as expected. However, we could get less sampling variability among the estimates by increasing the number of simulated values (β, σ^2) to 100 for each split as shown in Figure (3.2). Even though using larger posterior samples takes more computing time, it is worth trying for better results.

3.2.3 The Bronze Method

In this section we demonstrate the use of importance sampling through the example of the body and brain weights of the mammals to check and improve upon the procedure of splitting the data set and generating the posterior parameters for each split. We present the same ordinary linear model on the mammals with parameters $\theta = (\beta_0, \beta_1, \sigma^2)$, with the use of importance weights in drawing the posterior parameters to look as though they were generated from the posterior distribution given the training set. For the bronze method we start first by generating the marginal and the conditional posterior distributions for θ and the replicated data set based on the whole data set. This suggests the following simulation algorithm:

- Simulate a sample of size $m = 100$ from the flattened posterior distribution given by

$$q(\theta|y) \propto \left\{ \prod_{i=1}^{n} \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} exp \left\{ -\frac{1}{2} \left(\frac{y_i - \beta_0 - \beta_1}{\sigma} \right)^2 \right\} \right\}^{\frac{1}{2}} \frac{1}{\sigma^2}. \qquad (3.3)$$

  Sample $\sigma^2 \sim Inv - \chi^2(\alpha(62) - 2, s^2)$, and conditioning on σ^2 sample $\beta \sim N(\hat{\beta}, V_\beta \frac{\sigma^2}{\alpha})$.

- Given θ_i, $i = 1, \ldots, 100$, generate a sample of $m = 100$ from $p(y|\theta_i)$.
- Sample splits s_i, $i = 1, \ldots, 50$.
- For each split, compute the importance weights w_i^s, $i = 1, \ldots, 100$.

  The estimate of the predictive error is

$$\hat{w}(s_i) = \sum_{i=1}^{m} \left[\|y_V^{rep} - y_V\|^2 \right] w_i^s.$$

- Average over splits to obtain $\widehat{W} = \frac{1}{50} \sum_{i=1}^{50} \hat{w}(s_i)$.

Repeating the whole procedure 100 times using the same split indicators that the gold and silver methods used, we have 100 $\widehat{W}$'s shown in the histogram in Figure (3.2). Also for a fair comparison with the silver method of 50 simulated values of (β, σ^2), one value per split, we generated a sample of size $m = 50$ from the flattened posterior distribution before splitting the data and carrying out the same procedure to draw the histogram of the 100 $\widehat{W}$ as in Figure (3.2). This is more spread out than the histograms of the $\widehat{W}$'s from the 100 θ's generated for the 50 splits.

By looking at the histograms in Figure (3.2), we can conclude that the gold method has the smallest variance, as expected from the gold estimator which has the bias correction term, the variances of the replicates, added to the prediction errors to provide an adequate estimate of W. Therefore the gold $\widehat{W}$'s have less sampling variability. By sampling 50 θ's at once for the bronze method and one θ per split for the 50 splits in the silver method, we wanted to conduct a fair comparison between the two methods. Based on the results from this example they seem to have the same sample variability. Also, from a

practical point of view sampling one θ it is less reliable so we needed to sample more than one θ every time for each split. The 100 θ's histogram of the silver sample estimates shows less spread than the bronze sample estimates, which was expected. For the silver method for each different split there is a new sample of the parameter θ, so we expect to have less sample variability than the bronze. However, we should mention that the silver method takes more computing time than the bronze for sampling from the posterior distribution in each split. On the other hand, for the bronze method we need to calculate the likelihood for each θ and in each split we need to calculate those weights. In this example, the gold method turned to be the best but for complicated models it will not be feasible. For us, we see the bronze method works better than the silver, and takes less computing time. We can not judge these methods based on one example only, so we decided to try another model, which we discuss in the next section.

3.3 Weibull and Extreme Value Regression Model

Until this point we have dealt exclusively with the ordinary linear model. In practice many situations involve heterogeneous populations, and it is important to consider the relationship between lifetime and other factors. We thought we would try the methods on regression models, where the dependence of lifetime on regressor variables is explicitly recognized. The methods for estimating expected prediction error are treated in detail in this section via a regression method using data on the time to breakdown of a type of electrical insulting fluid subject to a constant voltage stress. The silver and the bronze methods are implemented for estimating the predictive error, whereas the gold method is not involved in this example as it is more mathematically intense to figure out the predictive distribution, and this is not the purpose of this research.

Data set

The data set used in this analysis were obtained from Lawless (1982). Nelson (1970a) presents the data, which are breakdown times for seven groups of specimens, each group involving a different voltage level. The data are uncensored, and times to breakdown are given in minutes. The results of an accelerated life test experiment on a type of electrical insulation were presented. In the experiment, specimens of insulation were subjected to seven voltage levels, 26, 28, 30, 32, 34, 36, and 38 kilovolts (KV). Engineers thought the appropriate model was that for a fixed voltage level x_i, the lifetime distribution for the items is a Weibull distribution with shape δ and scale parameter

$\alpha_i = exp\{\beta_0 + \beta_1 x_i\}$. Note that distributions corresponding to different voltage levels are considered to differ only with respect to their scale parameters α_i, the shape parameter δ being the same for different levels. It is shown that there is no evidence against the hypothesis of equality of shape parameters δ (for more details see Lawless (1982)).

The Weibull distribution can be extended to include regressor variables in different ways. However, the most commonly used is that for which the p.d.f. of lifetime, given the vector $\mathbf{x}$ of regressor variables, is

$$f(t) = \frac{\delta}{\alpha(\mathbf{x})}\left(\frac{t}{\alpha(\mathbf{x})}\right)^{\delta-1} exp\left\{-\left(\frac{t}{\alpha(\mathbf{x})}\right)^{\delta}\right\}, t \geq 0.$$

We may often work with log lifetime, $Y = log(T)$, with the p.d.f. given $\mathbf{x}$ being

$$f(y|\mathbf{x}) = \frac{1}{\sigma} exp\left\{\frac{y-\mu(\mathbf{x})}{\sigma} - exp\left(\frac{y-\mu(\mathbf{x})}{\sigma}\right)\right\}, -\infty < y < \infty, \qquad (3.4)$$

where $\mu(\mathbf{x}) = log\alpha(\mathbf{x})$ and $\sigma = \delta^{-1}$. So the linear model can be written as

$$y = \beta_0 + \beta_1 x + \sigma z \qquad (3.5)$$

with $\mu(x) = \beta_0 + \beta_1 x$, where z has a standard extreme value distribution, $-\infty < z < \infty$.

Prediction

To try out the prediction error in our Weibull regression model, we carried out the procedures from the silver and the bronze methods. In this case it was not possible to sample directly from the posterior distribution. Therefore,

random numbers were generated from a Markov chain with the posterior as the stationary distribution. This method is well known as the Markov Chain Monte Carlo (MCMC) method. We used the Metropolis-Hastings algorithm with independent increments. In the next section we give a brief review of the Metropolis algorithm.

3.3.1 The Metropolis Algorithm

The Metropolis algorithm was originally introduced by Metropolis, Roseenbuth, Teller and Teller (1953) for computing properties of substances composed of interacting individual molecules. This algorithm has been used extensively in statistical physics. A variety of these kinds of algorithms were proposed by many researchers in the past.

The Hastings version of the algorithm constructs a Markov chain with Π as its stationary distribution as follows. In the Bayesian approach Π is the posterior distribution which would be the target distribution. If the chain is currently at a point $X_n = x$, then it generates a candidate value Y for the next location X_{n+1} from a transition density $q(x, \cdot)$. With probability $\alpha(x, y)$ this candidate is accepted and the chain moves to $X_{n+1} = y$. Otherwise, the step is rejected and the chain remains at $X_{n+1} = x$ with probability 1- $\alpha(x, y)$, where

$$\alpha(x, y) = min \left\{ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1 \right\}$$

is the acceptance probability. Note that this algorithm depends on π only

through the ratios of the form $\pi(y)/\pi(x)$; thus π only needs to be known up to a normalizing constant (Hastings, 1970).

If Π is a continuous univariate distribution on $\mathbb{R}$, the random walk Metropolis algorithm (Tierney, 1994) proceeds as follows. To update from $X_n = x_n$ to $X_{n+1} = x_{n+1}$, we add noise (usually taken to be normal with mean zero and variance σ^2) to the current state. As such, $y = x_n + z$, where $z \sim N(0, \sigma^2)$ and

$$\alpha = min\left(\frac{\pi(y)}{\pi(x)}, 1\right).$$

Then,

$$x_{n+1} = \begin{cases} y & \text{with probability } \alpha, \\ x_n & \text{with probability } 1 - \alpha. \end{cases}$$

Implementing MCMC

We thought we would try out the MCMC algorithms before conducting them in the methods, so we could make sure the chains converge and have an appropriate acceptance rate. As for the Bayesian approach, we must specify a prior distribution in addition to the likelihood based (3.4), so that samples of the unknown parameters $\theta = (\beta_0, \beta_1, \sigma)$ can be simulated from the posterior distribution. Since we do not have enough information to construct a prior distribution, we chose a standard noninformative prior, $\pi(\beta, \sigma) \propto \frac{1}{\sigma}$, so that inferences are unaffected by information external to the current data. We used the random walk Metropolis algorithm with independent increments for exploring the posterior distribution. The candidate state is obtained by adding

noise to the current state. the noise is generated from the normal distribution $Z \sim N(0, \tau_j)$, where τ_j is the tuning parameter set up by trial and error to adjust how high the acceptance rate of the chain should be.

To run the random walk (RW) algorithm we should start with initial values for the parameters that we want to sample from their posterior distributions. The simple way to do that is to fit a simple linear model and let the least squares estimates of the parameters be the initial values for β_0 and β_1. And we let σ^2 equal the variance of the log(lifetime). With these settings RW chains of length $n = 500$ are simulated in Splus. We obtained unsatisfactory results of chains moving slowly through out of the support of the target distribution and others having high jumps with high range of acceptance rates between the parameters as shown in Figure (3.3), which indicates that the chains will never converge. So we thought we should use some methods which are useful for improving lack of convergence or slow convergence due to bad starting values, or high posterior correlations. The posterior parameters are correlated which may slow down the mixing in the chains. A simple remedy is to work with centered covariates $x_i' = x_i - \bar{x}$ (Gilks and Roberts, 1996). Doing that with the same first settings, and adjusting the tuning parameters, we achieve adequate mixing of a Markov chain and a range of acceptance rates for updating the parameters between 50% and 75% which are reasonable rates based on Gelman et al. (1995). Plots of these are given in Figure (3.4).

In the next section we discuss the silver and the bronze methods through this same example. And for a fair comparison we generate the whole split

indicator by creating arrays so we use the same splits for both methods at the same time. We are using MCMC to sample from the posterior distributions. Based on some tests done to check how fast the chains converge, we decided to run the chains for a total of 500 iterations, with 300 burn-in iteration so we have a sample of size $m = 200$ of θ's to be our posterior sample.

3.3.2 The Silver Method

We tried out the silver method for estimating the predictive error in the voltage data, and we consider the extreme value model. Using the random walk Metropolis-Hastings algorithm to sample from the posterior distribution of $\theta = (\beta_0, \beta_1, \sigma)$, we carried out the procedure for estimating the expected error as follows.

- Sample splits s_i, $i = 1, \ldots, I$.
- For each split:
    1. Set up the initial value for the MCMC method which are $\beta_0 = E(y_T)$, the mean of the training sample of the independent variables; $\beta_1 = 0$. And $\sigma^2 = var(y_T)$, the variance of the sample of the independent variables.
    2. Run MCMC to estimate a sample of size $m = 200$ θ's.
    3. Based on the 200 θ's, draw a hypothetical replication y_V^{rep}.
    4. Compute the squared predictive error for the 200 sets of the repli-

cation,

$$s = ||y_V^{rep} - y_V||^2.$$

5. The estimate of the predictive error would be

$$\hat{w}_i = mean(s).$$

- After multiple splits, average over the $\hat{w}_i$'s to obtain the estimate of the expected squared error

$$\widehat{W} = \frac{1}{I}\sum_{i=1}^{I} \hat{w}_i.$$

We first tried the silver method with number of splits $i = 50$, and sample size $m = 100$ θ's using MCMC over 1000 iterations. The estimate of the expected squared error was $\widehat{W} = 212.28$, which is a sensible estimate based on the results of the mean of the squared predictive error which have small variance for each split. The problem in for this setting was the computing time that took very long to get one $\widehat{W}$, which is not practical as we need many of them so we could compare with the bronze method estimates. Therefore, we cut down the number of splits to $i = 25$ and the number of iterations to 500. This time the silver method procedure took less time so we could use it for many estimates of $\widehat{W}$'s. In the next section we present the bronze method, then we plot graphs of the estimates for both methods for comparison purposes and checking the bias of the estimates.

3.3.3 The Bronze Method

As we discussed previously for the bronze method, we first sample all θ's needed for the replication values, then split the data many times for each to get the estimate of the predicted error. We tried the following algorithm for the voltage data.

1. Simulate a sample of $m = 200$ θ's. By using the random walk Metropolis MCMC algorithm, with $\beta_0 = E(y)$, $\beta_1 = 0$, and $\sigma = var(y)$ as the initial settings for the MCMC chain. We use the likelihood function

$$L(\theta) = \left\{ \prod_{i=1}^{n} \frac{1}{\sigma} exp\left(\frac{y_i - (\beta_0 + \beta 1 x_i)}{\sigma} - exp\left(\frac{y_i - (\beta_0 + \beta 1 x_i)}{\sigma} \right) \right) \right\}^{\alpha},$$

   where $\alpha = \frac{1}{2}$ as explained in the previous chapter.

2. Given θ's, compute the log densities for the regressor variables (voltages) that are needed to figure out the importance weights for each θ, given by

$$log(L(\theta)) = -nlog(\sigma) + \sum_{i=1}^{n} \left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma} \right) - \sum_{i=1}^{n} exp\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma} \right).$$

3. For each θ, sample replicated data y^{rep} from the extreme value distribution.

4. Sample splits $i = 1, \ldots, I$. For each split s

   - For each θ_i, $i = 1, \ldots, m$, compute the importance weight

$$w_j = \sum_{j=1}^{m} ((s - \alpha) log(L(\theta_j))).$$

- Compute the squared difference between the replicated validation sample y_V^{rep} and the realized validation sample y_V $\|y_V^{rep} - y_V\|^2$.
- Compute the split-specific estimate of the expected squared predicted error

$$\hat{w}_i = \sum_{j=1}^{m} (\|y_V^{rep} - y_V\|_j^2 \times w_j).$$

5. Average over splits, $\widehat{W} = \frac{1}{I} \sum_{i=1}^{I} \hat{w}_i$.

Based on the same first settings of the silver method with $i = 50$ splits and sample size $m = 100$ θ's, using MCMC over 1000 iterations, for the first set of splits we obtained $\widehat{W} = 214.53$ which is approximately the same as the silver estimate, with much less computing time. Then based on the ultimate setting of the silver method, we wanted to check if there is any bias in the estimates of the silver and the bronze. We drew a plot shown in Figure (3.5), which obtained by repeating the process of computing $\widehat{W}$ 20 times for the silver and the bronze methods. The plot shows no sign of bias in either of the methods, which makes us more confident with our results. Also we compute 50 $\widehat{W}$'s for each of the silver and the bronze methods to draw histograms and test the sampling variability for each of them as shown in Figure (3.6). From this histogram, we see the silver method estimates have less sample variability than the bronze method estimates. These results would give us an idea which method to choose, Each of them has some advantage over the other. For example, the bronze method has less computing time and we do not have to fit the model for each split to sample θ as is the case in the silver method. However, for the bronze method there are a lot of calculations such us the

log-likelihood for every θ and the importance weights. As we will see in the next chapter, prediction methods, while useful in their own right, can also be used as a tool in model checking.

Figure 3.1: *Scatterplot of the mammals data for the average brain and body weights.*

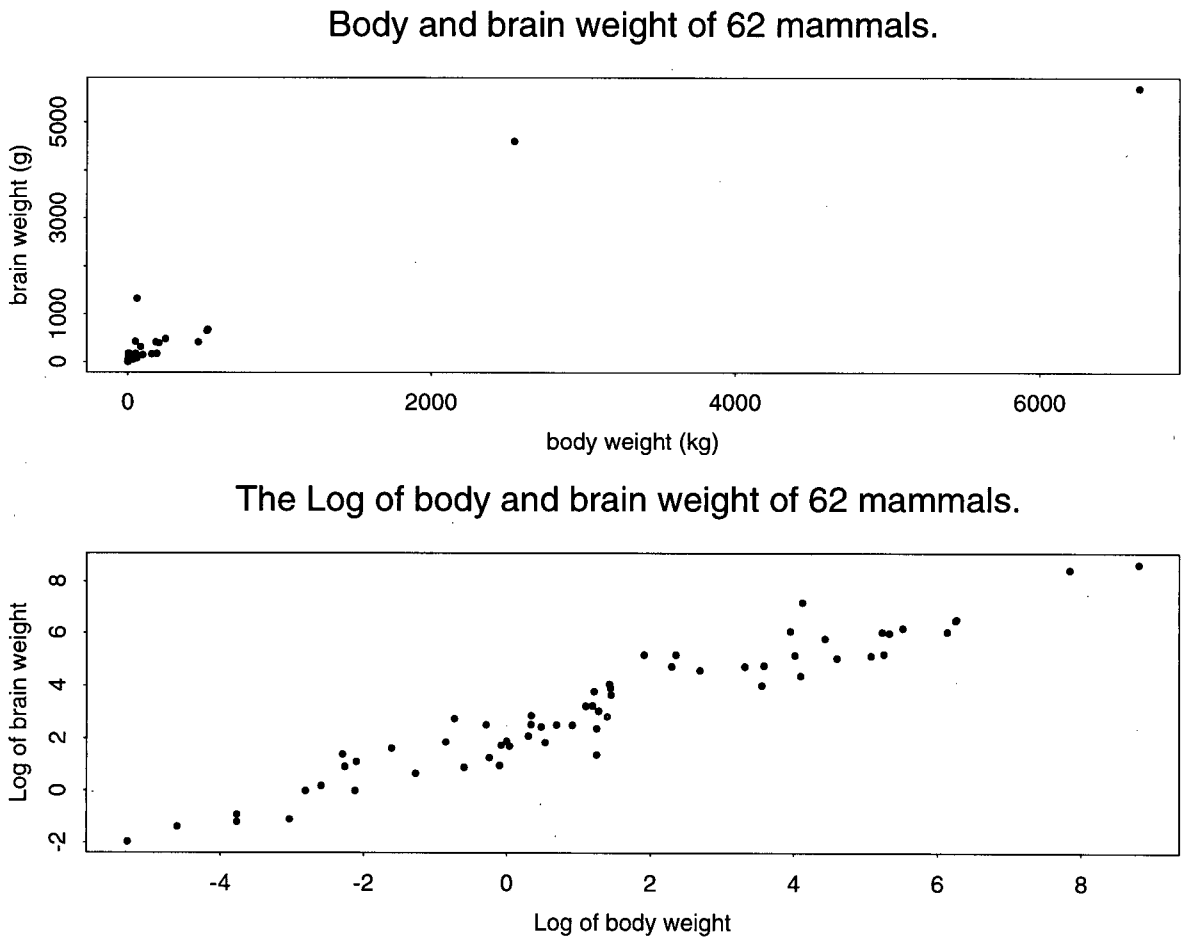


Figure 3.2: *Histograms of 100 $\widehat{W}$'s.*

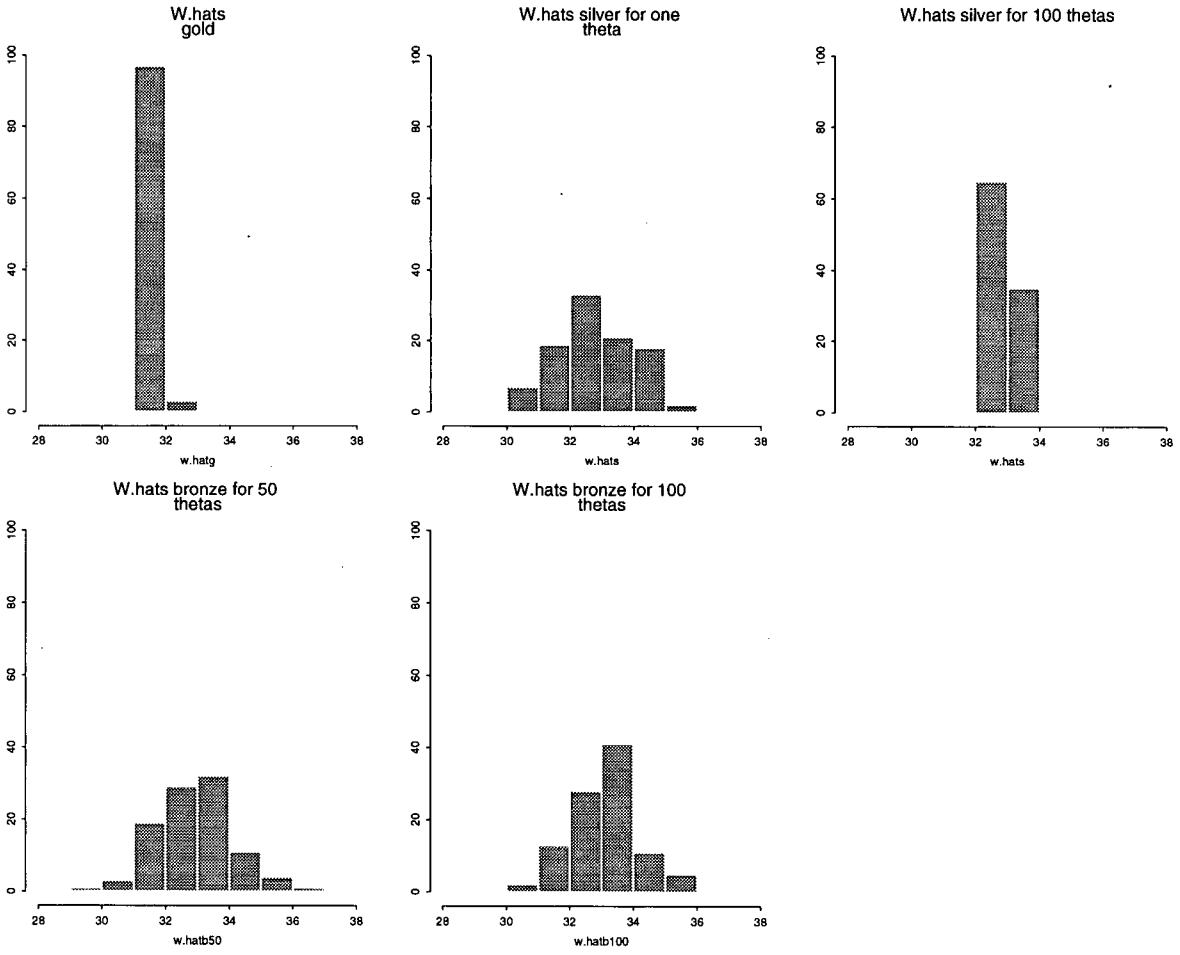


Figure 3.3: *MCMC of uncentered voltage for insulating fluid failure data.*

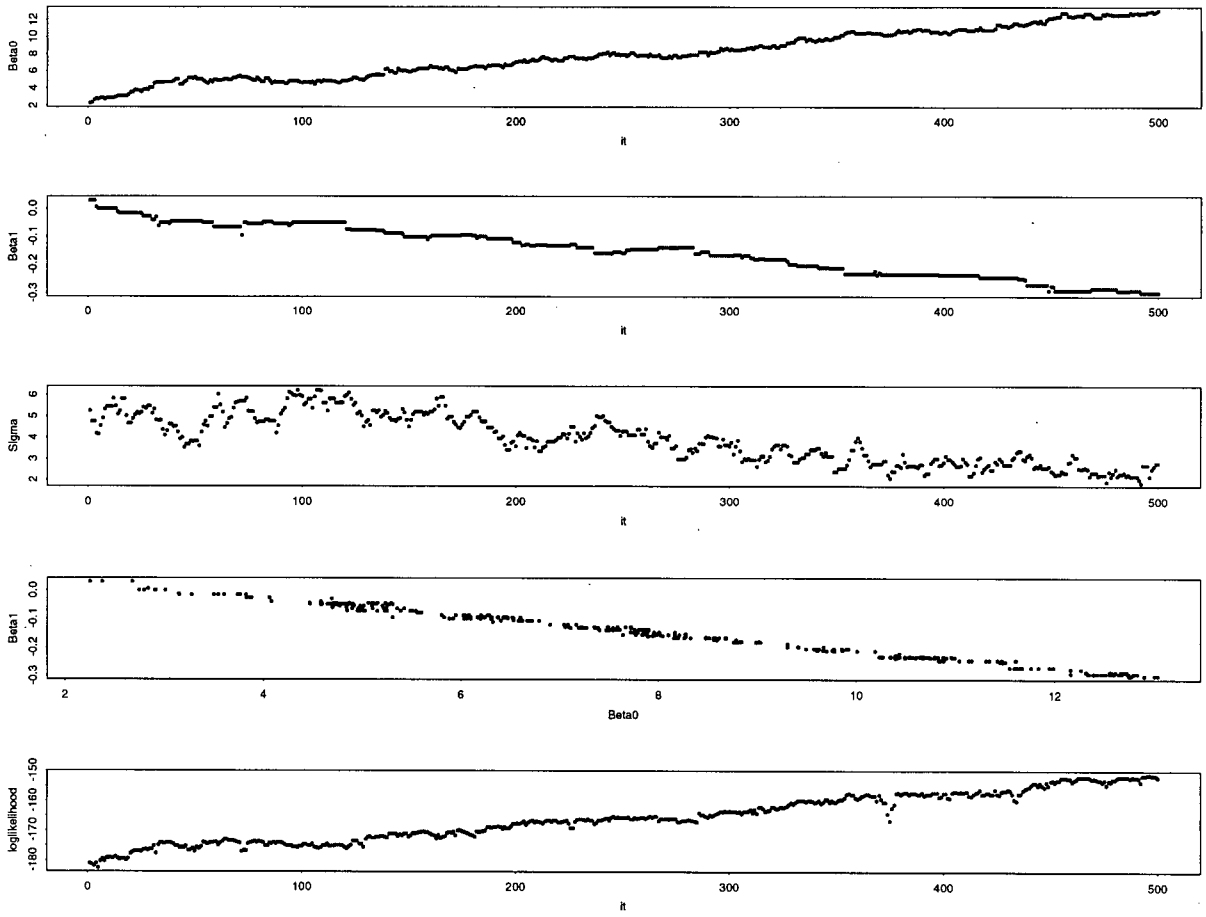


Figure 3.4: *MCMC of centered voltage for insulating fluid failure data.*

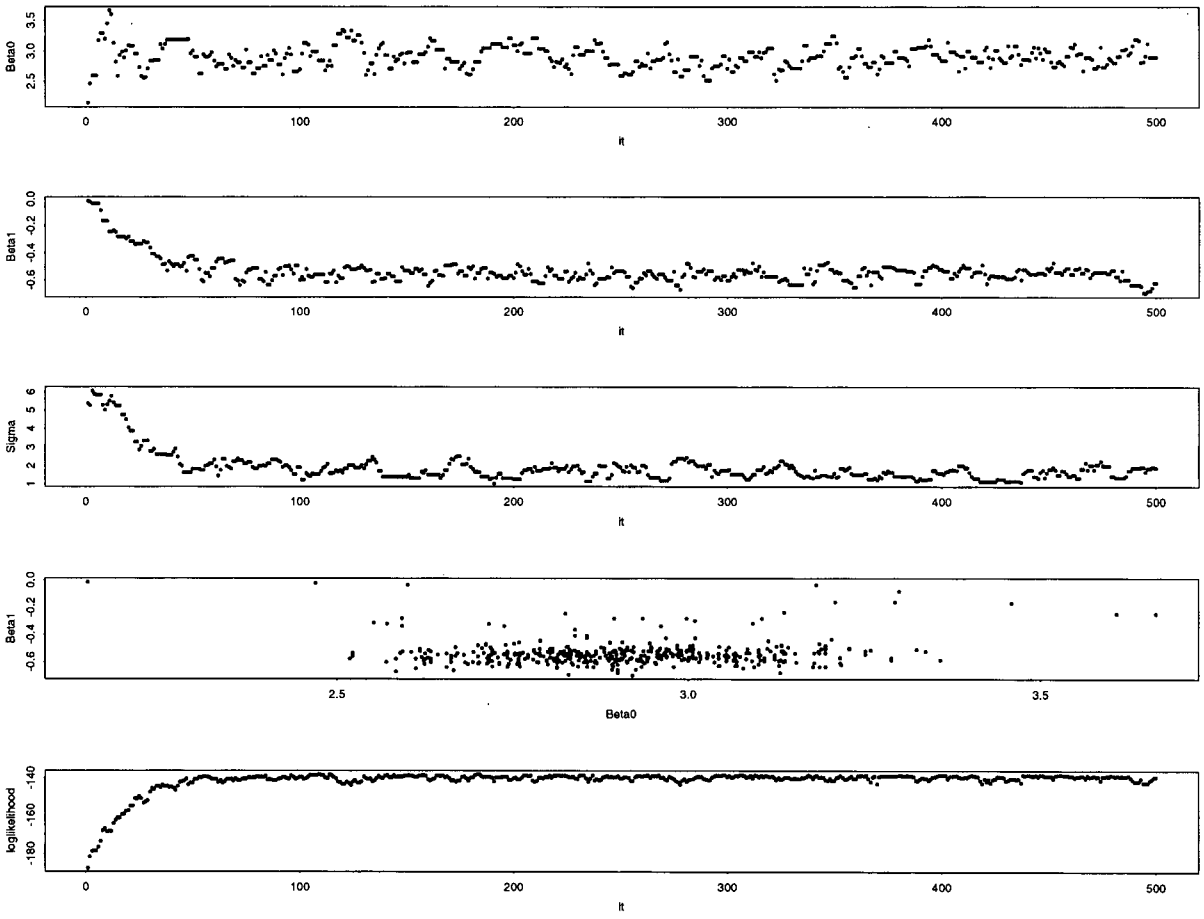


Figure 3.5: *Silver and bronze estimates prediction error.*

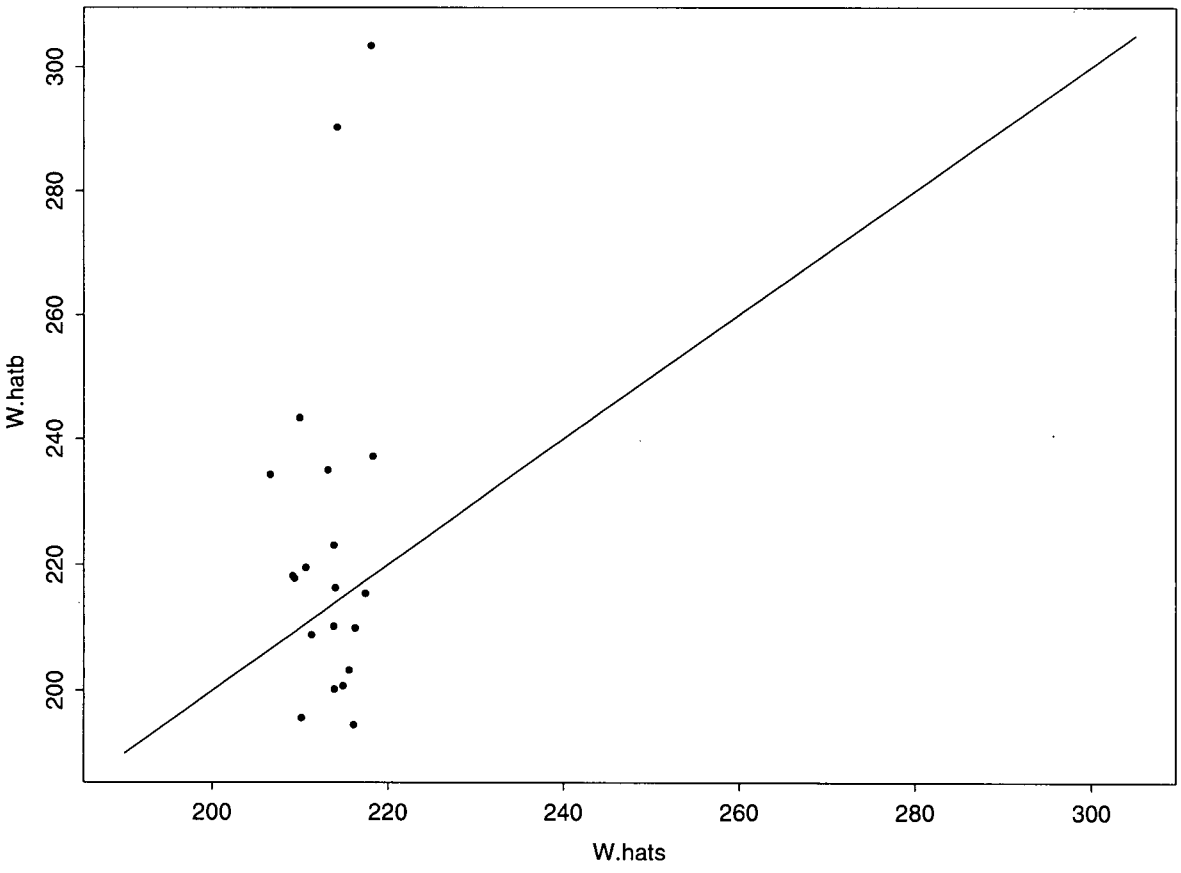
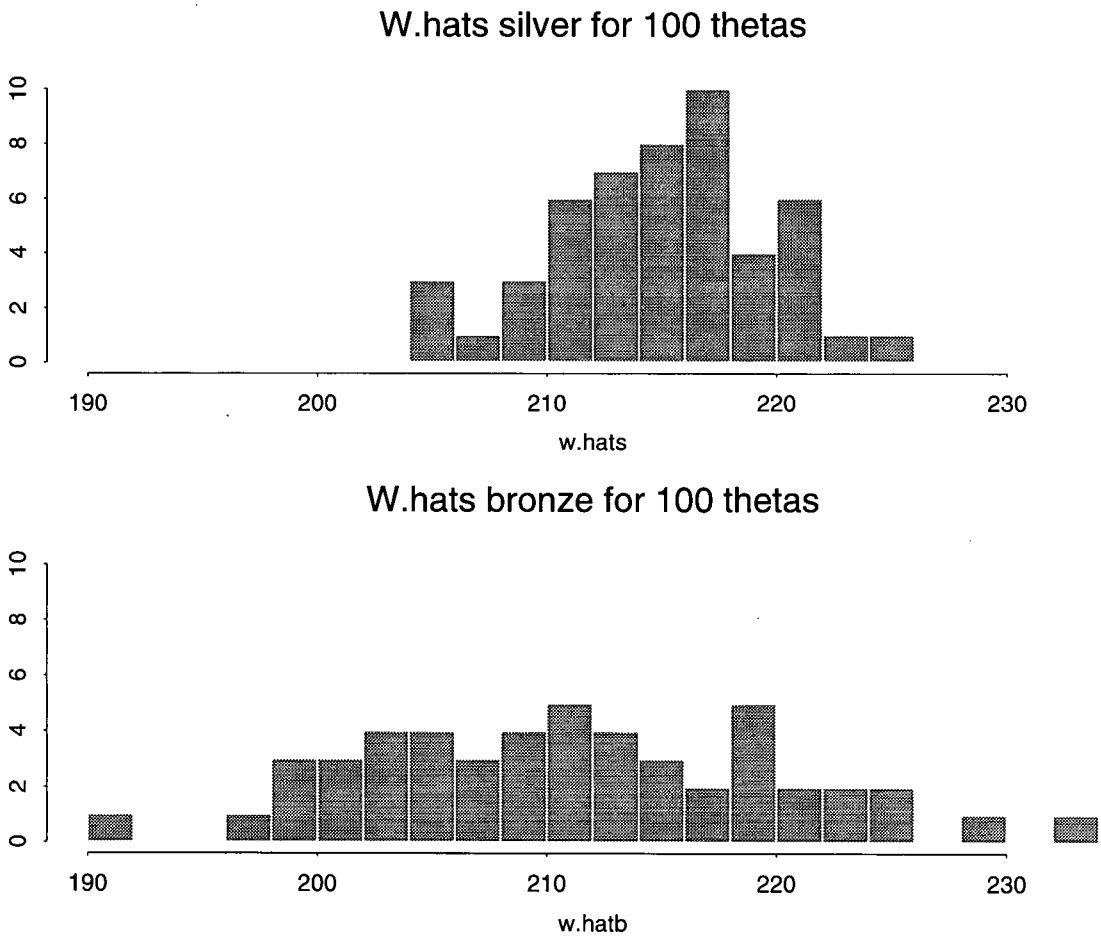


Figure 3.6: *Histograms of 50 $\widehat{W}$'s.*



Chapter 4

Bayesian p-Value

4.1 Introduction

The p-value is a statistical notion that has been widely used and yet seriously criticized for a long time. The main controversy is whether a p-value provides adequate evidence against a null hypothesis. Most criticism leveled at the p-value has come from the Bayesian school, as mentioned in Hwang, Casella, Robert, Well and Farrel (1992), because the calculation of a p-value involves averaging over sample values that have not occurred, which is a clear violation of the likelihood principle. This has led to several formulations of a "Bayesian p-value." In particular, Rubin (1984) uses the posterior predictive distribution of a statistic to calculate the tail-area probability corresponding to the observed value of the statistic. As Rubin pointed out, such a frequency calculation is relevant because it helps the process of model diagnosis, a fundamental part of any Bayesian analysis. Meng (1994) intended to illustrate

the utility of the posterior predictive p-value. The Bayesian formulation, using posterior predictive replications of the data, allows a test statistic to depend on both data and unknown parameters and thus permits a direct measure of the discrepancy between sample and population quantities. The extension to include generalized discrepancies $D(y;\theta)$ rather than just statistics $T(y)$, as in Rubin (1984), may be practically quite important because of the potentially tremendous computational advantages when averaging rather than maximizing over θ, especially with multi-modal likelihoods. This posterior predictive p-value can also be viewed as a generalization of a classical p-value, averaging over the posterior distribution of parameters under the null hypothesis, and thus it provides a general method for dealing with the unknown parameters.

In Bayesian statistics, a model can be checked in at least three ways. One way is checking that the model fits the data using the posterior predictive distribution, in the presence of unknown parameters. Gelman, Meng and Stern (1996), and Gelman, Carlin, Stern and Rubin (1995) discussed the approach of comparing the posterior predictive distribution of future observations to the data that have actually occurred. It is simple, both conceptually and computationally, and is applicable for comparing observations with model predictions in any form, without requiring any more substantive information than is in the existing data and model. In this chapter our purpose is to introduce a different notion of Bayesian p-value which includes averaging over different training/validating splits of the data. We illustrate this notion of p-value with an example used in Gelman, Carlin, Stern and Rubin (1995).

Draper (1996) also proposed cross-validating in the modeling process as an alternative to posterior predictive assessment of χ^2 goodness-of-fit measures described by Gelman, Meng and Stern (1996). Our approach is different in that we split the data many times rather than once. We compare our approach to Gelman et al. (1995) and investigate how the idea of splitting the data may effect the interpretation of the posterior predictive p-value. Also we provide a formal definition of a posterior predictive p-value.

4.2 Posterior Predictive p-value

The p-value of the test quantity as defined in the Bayesian context can be used to measure lack of fit of the data with respect to the posterior predictive distribution. Before we introduce our notion of the p-value, we briefly present the Gelman et al. notion of the p-value and define the classical p-value. Also we define Draper's notion of the p-value, as it is somehow related to our approach by splitting the data once rather than many times.

Using the same notation as in Chapter 2, the Gelman et al. work with the joint posterior distribution of y^{rep} and θ given the existing data y,

$$p(y^{rep}, \theta|y) = p(y^{rep}|\theta)p(\theta|y).$$

Then the posterior predictive distribution is

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta.$$

First we define the classical setting of the p-value for the test statistic $T(y)$ to

be

$$\text{C}p - \text{value} = Pr\{T(y^{rep}) \geq T(y)|\theta\},$$

where the probability is over the distribution of y^{rep} *with fixed* θ.

In the Bayesian approach, the test quantity $T(y, \theta)$ is a function of the unknown parameters and data as it is evaluated over draws from the posterior distribution of the unknown parameters. So the Gelman et al. definition of the posterior predictive p-value is the probability that the replicated data could be more extreme than the observed data,

$$\text{B}p - \text{value} = Pr\{T(y^{rep}, \theta) \geq T(y, \theta)|y\},$$

where the probability is over the posterior predictive distribution of y^{rep} *and the posterior distribution of* θ. The Gelman et al. approach to measuring the p-value by using posterior simulations of (θ, y^{rep}) is as follows.

- Draw $\theta_1, \ldots, \theta_L$ independently from the posterior distribution of θ.
- For each simulated θ^l, draw a replicated observation, $y^{rep,l}$ from the distribution, $y|\theta = \theta^l$.
- The estimated p-value is the proportion of the L test quantities for which the predictive test quantities $T(y^{rep,l}, \theta^l)$ equal or exceed the realized test quantity $T(y, \theta^l)$.

Draper suggested in his comment on The Gelman, Meng and Stern (1996) paper proposed using predictive distributions in a somewhat different

way-by cross-validation. Draper's p-value is "split-specific" which means specific to one split s. The split-specific p-value is defined as follows.

$$p_s = Pr\{T(y_V^{rep}, \theta) \geq T(y_V, \theta)|S = s\} \tag{4.1}$$

Draper thinks that this discrepancy measure can better highlight which part of the data set fits the model badly, and thus it prevents using the same data twice problem.

We choose to use a different approach for measuring the p-value, following the methodology of splitting the data. So we define the "split-averaged" p-value as

$$p = Pr\{T(y_V^{rep}, \theta) \geq T(y_V, \theta)\}, \tag{4.2}$$

with respect to the "reference distribution" (2.1). Note that $p = E\{p_s\}$ with respect to the reference distribution (2.1). To compute the p-value, the silver method can be used in the following way.

- Sample θ_i^l from the posterior distribution $p(\theta_i^l|y_{T_i})$, $l = 1, \ldots, L$.
- Sample $y_{V_i}^{rep,l}$ from the probability distribution $p(y_{V_i}|\theta_i^l)$.
- Compare the realized validation test quantities $T(y_{V_i}, \theta_i^l)$ and the predictive validation test quantities $T(y_{V_i}^{rep}, \theta_i^l)$, that is, test whether $T(y_V^{rep}, \theta) \geq T(y_V, \theta)$.

To estimate the p-value (4.2), note that

$$\begin{aligned} Pr(T(y_V^{rep}, \theta) \geq T(y_V, \theta)|y_T) &= E\left\{I_{\{T(y_V^{rep}, \theta) \geq T(y_V, \theta)\}}|y_T\right\} \\ &= EE\left\{I_{\{T(y_V^{rep}, \theta) \geq T(y_V, \theta)\}}|\theta, y_T\right\}, \end{aligned}$$

and we sample from the joint distribution on (s, θ) given by

$$p(s, \theta) = p(s)p(\theta|y_T).$$

This suggests the following estimator, for a given split

$$\hat{p}_s = \frac{1}{L}\sum_{i=1}^{L} I_{\{T(y_V^{rep},\theta_i) \geq T(y_V,\theta_i)\}} \tag{4.3}$$

where $\theta_1, \ldots, \theta_L \sim \theta|s$. And thus

$$\hat{p} = \frac{1}{m}\sum_{i=1}^{m} \hat{p}_{s_i}$$

where $s_1, \ldots, s_m \sim p(s)$. Thus we are averaging over independent splits.

4.3 Example: Comparing speed of light measurements to the posterior predictive distribution.

We illustrate the split-averaged p-value through the example used by Gelman, Carlin, Stern and Rubin (1995). We calculate the p-value using the Gelman et al. approach and ours with two different testing functions. We compare the adequacy of the two methods for model checking.

Data set

The data set used to illustrate the concept of posterior predictive p-value is from an experiment set up by Simon Newcomb in 1882 to measure the speed of light. Newcomb measured the amount of time required for light

to travel a distance of 7442 meters. The 66 measurements are given in a histogram as shown in Figure (4.1). The histogram shows that there are two unusually low measurements and then a cluster of measurements that are approximately symmetrically distributed. It is inappropriate to assume that all 66 measurements are independent draws from a normal distribution with mean μ and variance σ^2, with a noninformative uniform prior distribution $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$. It is obvious that the outlying measurements do not fit the normal model. We discuss the Bayesian p-value for measuring the lack of fit for these data by comparing the observed data to what we expect to be observed under the posterior distribution. Any systematic differences between the simulations and the data indicate potential failings of the model.

4.3.1 The Gelman et al. Approach

The Gelman et al. approach uses graphical comparisons of summaries of the data to summaries from posterior predictive simulations, and the notion of posterior predictive p-value to measure the statistical significance of the lack of fit. They use the sample variance as the test quantity. The histogram of the observed sample variance and the 200 simulated variances from the posterior predictive distribution, and the estimated p-value close to $\frac{1}{2}$, indicate that the model fit the data well, which is not true. They justify this result by noting that the sample variance is not a good test statistic because it is a sufficient statistic of the model, which will be well fit in the posterior distribution.

4.3.2 The Split-Averaged p-value

We tried our approach of splitting the data with the speed of light measurements using the sample variance as a test quantity. We adapt the silver method to estimate the posterior predictive p-value as follows.

- Sample s_i, $i = 1, \ldots, 50$, to be used for splitting the data into equal halves.
- Given the training sample y_T, we first draw a random value of $\sigma^2 \sim Inv-\chi^2(32, s^2)$ as $32s^2$ divided by a random draw from the χ^2_{32} distribution (see Appendix B). Having obtained σ^2, we draw μ from the conditional posterior distribution, $N(\bar{y}_T, \sigma^2/33)$.
- Simulate a sample of posterior predictive observations y^{rep} from the posterior predictive distribution $N(\mu, \sigma^2)$.
- Compute the sample variance of the validation predictive observations y_V^{rep}.

For each split we have the sample variances of the actual validation sample and the replicated validation sample. The scatterplot in Figure (4.2) for the two test quantities shows two clusters, but the estimated p-value is close to $\frac{1}{2}$, similar to the Gelman et al. result. However, the scatterplot implies that the model is inadequate for the data, so we should try to justify this. To investigate further, for each split we sampled 200 replicated validation samples, leading to 200 replicated validation variances. These are compared to the sample variance of the actual validation sample. In particular, for each split

we can locate the actual validation variance in the histogram of the replicated variances. The histograms show the actual validation variance is either to the far right or the far left of the distribution of the simulated variances, which was expected in the way that the outliers may be split between the training and the validation samples. And this implies that the model is inadequate for the data. Rather than looking over all those histograms, we could try to summarize them in one plot and then in a single number corresponding to the posterior predictive p-value. For each split s, we estimate the split-specific p-value p_s (4.1) by $\hat{p}_s$(4.3). In this example, $\hat{p}_s$ is either zero or one.

The histogram of $\hat{p}_s$'s shown in Figure (4.3) should look uniformly distributed between zero and one if the model is adequate. For comparison purposes, we generate 66 data points from the standard normal and use the silver method of measuring the split-specific p-value. Based on 50 splits each with 200 simulated validation sample variances, the split-specific p-values look uniformly distributed between zero and one as the histogram shows in Figure (4.3).

For the sake of a single p-value which people always look for, we used the χ^2 test to check how uniform the $\hat{p}_s$'s are. To conduct the χ^2 test, we set up the null hypothesis to be

$$H_0 : \hat{p}_s \sim U(0, 1).$$

Each histogram is divided into bins (for our purposes, 5 bins to match the histograms), then we compute the frequency in each cell and the expected frequency. For the speed light measurements the p-value was less than 0.01

which means reject the null hypothesis, as expected. For the generated standard normal, the p-value was 0.9384 which implies that the p-values follow a uniform distribution. So we could get a single p-value to present the fit of the model. Based on results from this example we showed that the model is inadequate for the data; however, the Gelman et al. could not prove that without data splitting from the same test quantity that we used.

The Gelman et al. approach for calculating the p-value has been criticized for using the data twice, once to determine the posterior distribution of θ, and then again for replication. For the split-specific p-value, we overcome this problem by splitting the data into two, so that one part is used in determining the posterior distribution and the other for prediction. The point of model checking is to see if the current model is good enough. How do you know if a particular predictive p-value is precise and correct? The satisfying answer to this question must certainly include an attempt to quantify the utility of the p-value. The speed of light is a simple example so the model inadequacy was obvious, but in more complex models things may not be so clear. In particular, the Gelman et al. method may be very sensitive to how close the test quantity $T(\cdot)$ is to a sufficient statistic. Our method to model checking does not have that problem. It might be best applying this interesting notion of p-value with more complex models where its message fits into the overall picture of Bayesian model checking, by comparing the observed data to what we expect to be observed under the specified posterior distribution.

In the next section we discuss the same example with different test function that depends on the parameter θ. Gelman et al. could demonstrate the poor fit of the normal model to the speed of light data using $\min(y_i)$ as the test quantity. They suggested that the model can be inadequate for some purposes but adequate for others. They assessed the adequacy of the model except for the extreme tails, as the normal model clearly does not capture the variation that Newcomb observed. Using the same test quantity, we want to see how our method works and compare with the Gelman et al. result.

4.3.3 A Model Check Based on a Test Quantity Sensitive to Asymmetry in the Centre of the Distribution

Since the model looks adequate except for the extreme tails, Gelman et al. chose a test quantity sensitive to asymmetry in the centre of the distribution,

$$T(y, \theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|,$$

where $y_{(61)}$ and $y_{(6)}$ are the 61st and 6th order statistics representing approximately the 90% and 10% points of the distribution. We tried the Gelman et al. method, and the resulting scatterplot in Figure (4.4) shows the test quantity for the observed data and the test quantity evaluated for the simulated data for 200 simulations from the posterior distribution of (θ, σ^2). The estimated p-value is 0.24 which is computed as the proportion of points in the upper-left half of the plot. The test quantity should be scattered about zero for a

symmetric distribution. As shown in Figure (4.4), the test quantities are more skewed to the positive side. Gelman et al. explain any observed asymmetry in the middle of the distribution by sampling variability.

The Split-Averaged p-Value

Next, we tried the split-averaged p-value. We chose to split the data 200 times, for each split sampling a replicated validation sample from the posterior distribution of (θ, σ^2), and calculating the test quantity

$$T(y_V, \theta) = |y_{V_{(31)}} - \theta| - |y_{V_{(3)}} - \theta|,$$

for the actual observed validation sample and the replicated validation sample. Figure (4.5) shows the scatterplot and the estimated p-value is 0.38, implying the symmetry in the middle of the distribution. We expected to get clusters in the scatterplot to show the inadequacy of the model, as in the case of the sample variance test quantity. We interpreted the results as follows. There are clusters but they may not be visible. While splitting the data the outliers either go into the validation sample or into the training sample. When the outliers are in the training samples, then the test quantities are more positive as in the Gelman et al. case. This is because the posterior distribution is more effected by the outliers to the left. However, when the outliers are in the validation sample the test quantities are more about zero, since the posterior distribution is more in the middle. We tried for each split sampling 200 replicated validation samples, leading to 200 replicated validation and actual validation test quantities. We did not see any clusters in the scatterplots of 10 splits. We thought also that test quantity is quite sensitive to asymmetry

in the middle of the distribution which was obvious from the histogram of the speed of light data. The split-averaged p-value shows the symmetry in the centre of the data.

Figure 4.1: *Histogram of Simon Newcomb's measurements for estimating the speed of light.*

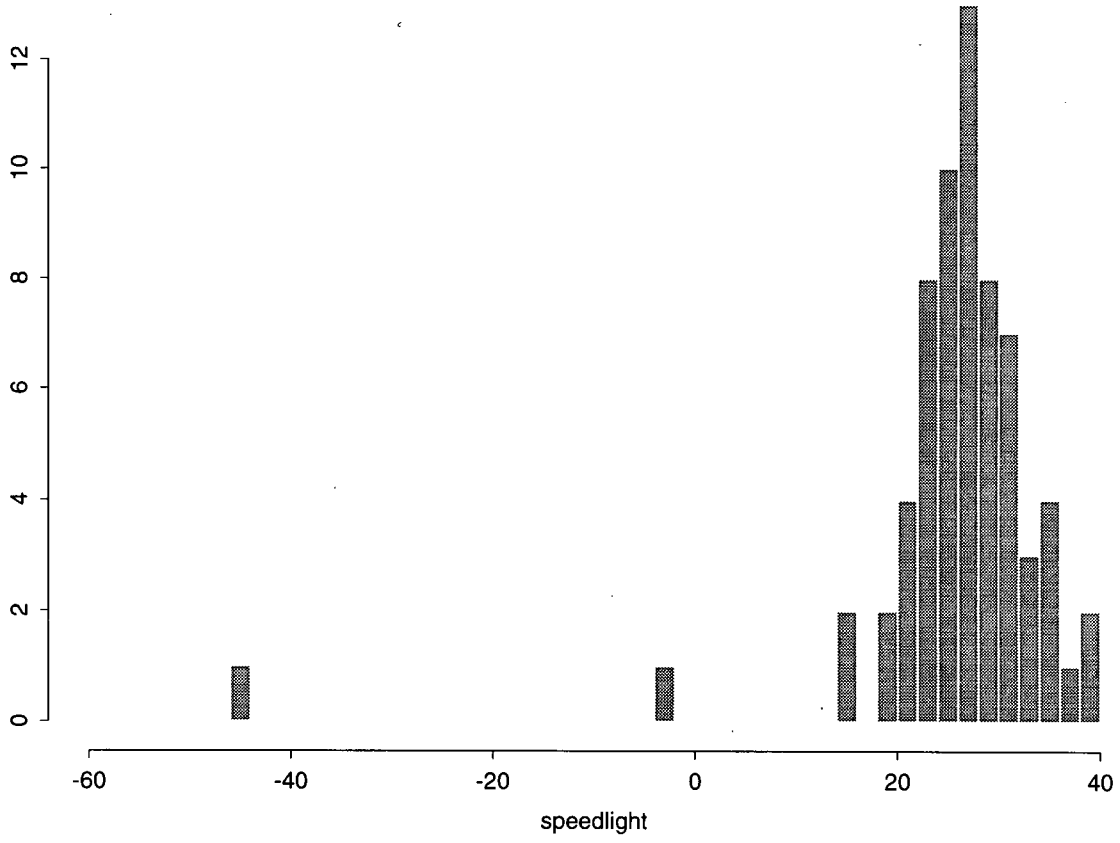


Figure 4.2: *Scatterplot showing the sample variances of the actual validation sample and the replicated validation sample.*

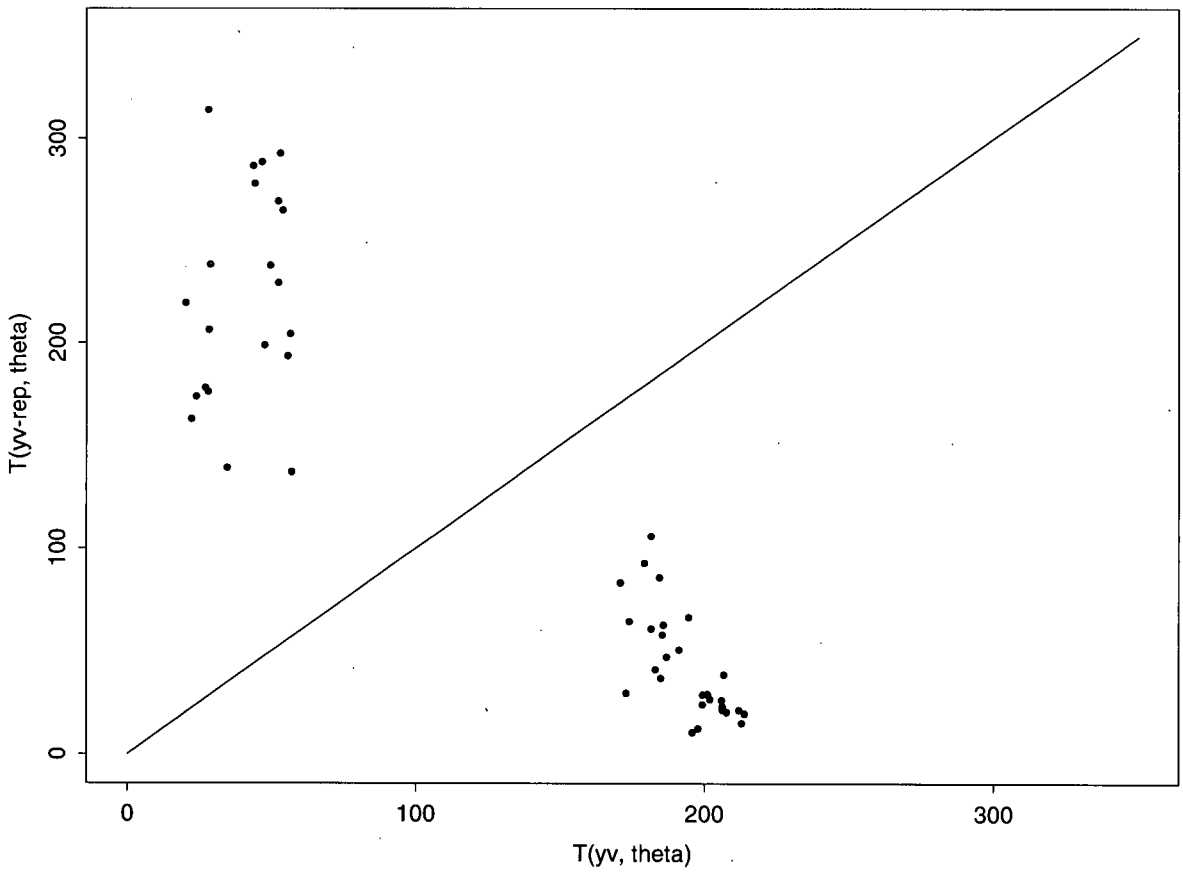


Figure 4.3: *Histograms of the estimate split-specific p-values for 50 splits.*

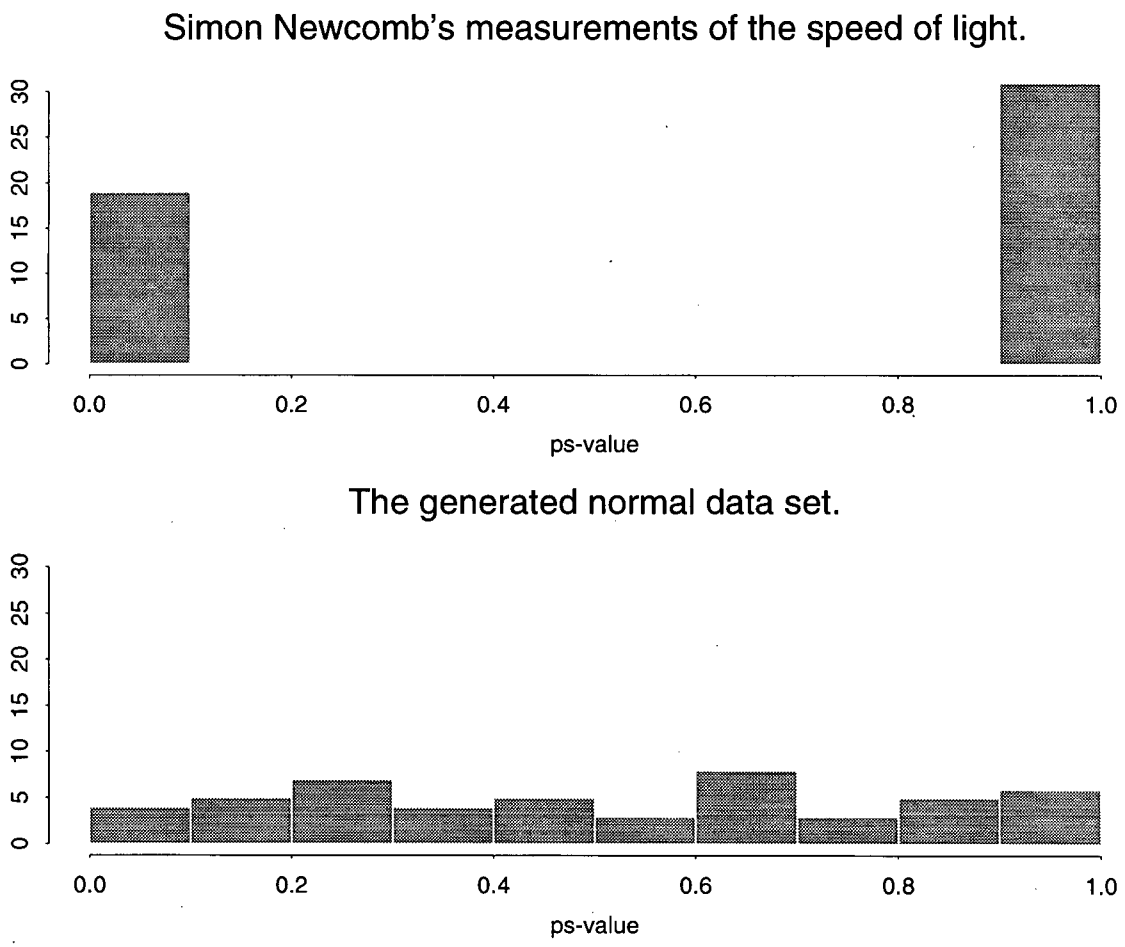


Figure 4.4: *Scatterplot showing prior and posterior simulations of a test quantity* $T(y, \theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|$, *based on 200 simulations from the posterior distribution of* (θ, y^{rep}).

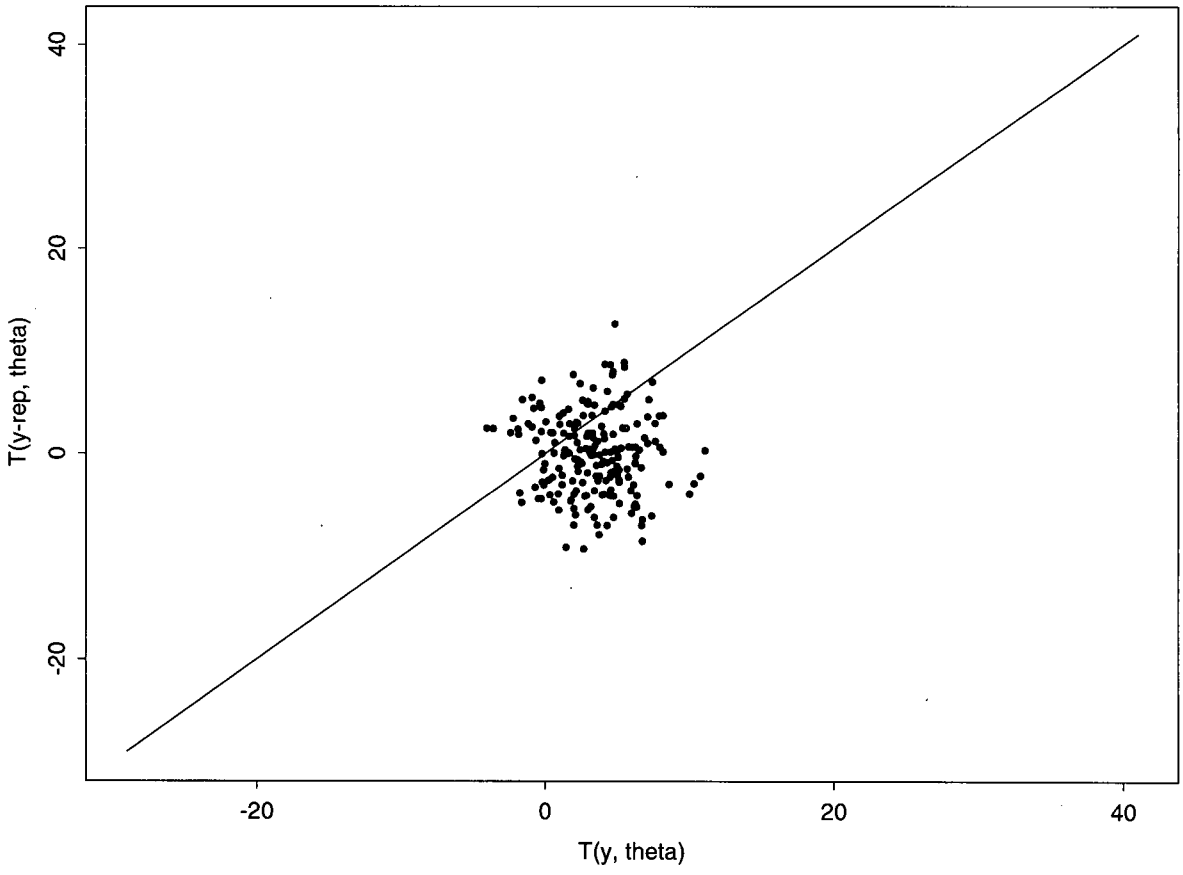
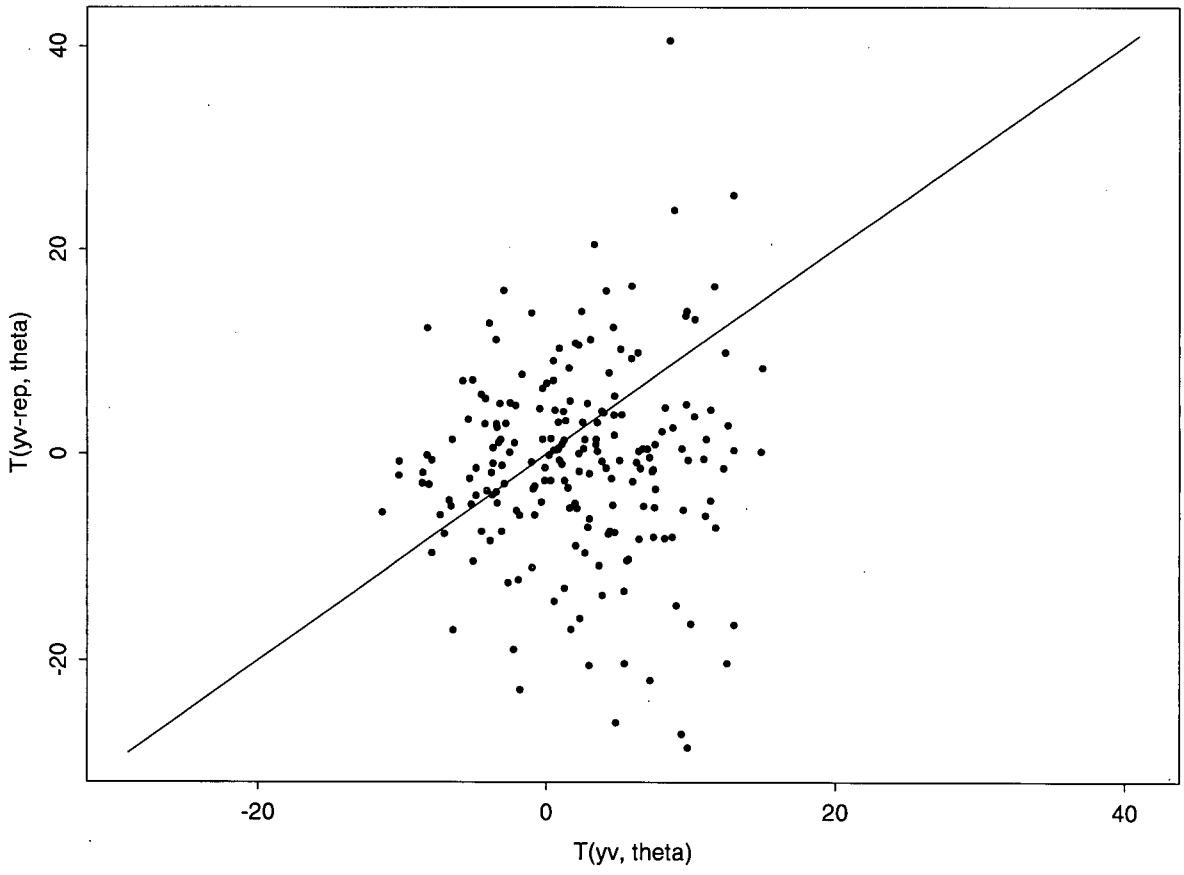


Figure 4.5: *Scatterplot showing a test quantity* $T(y_V, \theta) = |y_{V_{(31)}} - \theta| - |y_{V_{(3)}} - \theta|$ *for the actual observed validation sample and the replicated validation sample, based on 200 splits.*



Chapter 5

Discussion and Concluding Remarks

In principle cross-validation can be used in a great variety of situations, because of the lack of assumptions behind the cross-validation method. However, it seems to work best for unstructured data, or for insufficiently specified models. Cross-validation appears to have disadvantages in some classical situations. Data from designed experiments are usually highly structured. Deletion of a single observation destroys that structure, and therefore increases the computing effort. The use of cross-validation in regression is perhaps more dangerous. For simple models, the values of y corresponding to extreme values of x have disproportionate influence on the fitted parameters (Atkinson, 1985). When the data point (x_i, y_i) is omitted, the difference between the predicted value y_i^{rep} and the actual value y_i is at least as big as the i-th residual from the model fitted to the full data set. For extreme values x_i the

predictions y_i^{rep} are extrapolations, and so the difference can be expected to be larger than those for interior points x_i. Thus the disproportionate influence of the extreme data is increased by cross-validation.

To overcome this problem of cross-validation we implemented the repeated learning-testing methods shown by Burman (1989) to work very well for model selection purposes. We repeatedly split the data randomly into two parts. For each split, estimates are developed based on the data in the training set and then these are tested on the data in the validation set. So the extreme values, if any, are split between the two sets.

It seems to be more natural to use predictive distributions by cross-validating the modeling process (for an example see Draper (1995b)). This prescription uses a non-omnibus discrepancy measure that can better highlight which parts of the data set are badly fit by the current model. And it avoids the inherent diagnostic overfitting problem.

Chapter 2 presented the methodology for measuring the prediction error and called for the Bayesian approach in the context of cross-validation, which was revealed to be a fairly natural way to investigate the prediction error.

The examples and comparisons between the proposed methods carried out in Chapter 3 found the bronze method to perform reasonably, in terms of computing and the posterior sampling comparing to the silver method. However, the gold method gives the best estimates with the least sampling variability. To use the gold standard, the user must be able to compute the posterior expectation $E(y^{rep}|s)$ and variance $var(y^{rep}|s)$ with respect to the reference

distribution analytically, and that is not always possible depending on the complexity of data set. As for more realistic application, the silver method becomes more practical with the rapid computing developments. Therefore the reliability of the simulated result is important and requires much attention. When the posterior distribution cannot be simulated directly by exact sampling, an indirect simulation method (Markov Chain Monte Carlo) such as the Metropolis algorithm is often used. In the silver methodology, we sample from the posterior distribution and then sample the replicated data set from the posterior replicated distribution given the training set to compute the squared prediction error $||y^{rep} - y||$. We could try the silver method to estimate the expectation $\widehat{E}(y^{rep}|s)$ and the $\widehat{var}(y^{rep}|s)$ by sampling many times from the posterior distribution. We thought we may try this procedure in the future. However, this methodology apparently will not differ much from the current silver method in terms of posterior sampling for each split which we try to reduce by implementing the bronze method.

The recent rapid development of Bayesian computation allows us to fit more realistic and sophisticated models than previously possible, and thus there is a corresponding need for general methods to assess the fit of these model when classical tests are not applicable. An example is demonstrated by Gelman, Meng, and Stern (1996). Model checking is always a vital component of model building. The frequentist approach relies on the clever choice of discrepancy measures that are pivotal and whose distribution under the hypothesized model is known, at least approximately. The Bayesian approach

described by Gelman, Meng, and Stern (1996) is more general. The discrepancy measures used do not have to be pivotal or to have a known distribution, and so can be chosen to check aspects of the model that cause concern.

Gelman, Carlin, Stern, and Rubin (1995) introduced the posterior p-value method to check goodness-of-fit for parameteric Bayesian models. In a sense their method can be criticized for using the data twice, as their method checks how well the data predict themselves under the model. In Chapter 4 we introduced the cross-validation methodology with the split-averaged p-value using the silver method. This involves seeing how one portion of the data predicts another portion, which avoids using the data set twice. In Chapter 4 we illustrated the split-averaged p-value with the example used by Gelman et al., showing our procedure is simple yet efficient. Gelman et al. used the sample variance to be the discrepancy function, and thus they could not show the poor fit of the model. They accounted for this by noting that this discrepancy function is a sufficient statistic of the model and thus is automatically fitted in the posterior distribution. However, using the same discrepancy function, the split-averaged p-value could succeed in proving the inadequacy of the model, as was obvious from the histogram of the data showing the two clusters in Figure (4.2), resulting from the two outliers. In light of these encouraging results, we would like to try the split-averaged p-value on a more complicated data structures, such as hierarchical models.

Finally, although we could not try it in the scope of this thesis, the idea of implementing the Markov Chain Monte Carlo methods (MCMC) in the

procedure of splitting the data for each posterior sampling seems promising in the silver method. Referring to the same reference distribution in Chapter 2, we have $s = (T, V)$, y^{rep}, and θ, then the joint distribution

$$p(y^{rep}, \theta, T, V) = p(y^{rep}|\theta, T, V)p(\theta|T, V)p(T, V),$$

where $p(y^{rep}|\theta, T, V) = p(y^{rep}|\theta)$ in principle. And

$$p(\theta|T, V) \sim c(T, V)L(\theta|T, V)p(\theta),$$

where $c(T, V)$ is a normalizing constant.

To update (T, V), the target distribution is proportional to $c(T, V)p(T, V)$. Then we face the problem of calculating the constant that normalize the target distribution, making this approach difficult to apply.

Bibliography

[1] Akaike, H. (1974). A New Look at Statistical Model Identification. *IEEE Transactions on Automatic Control,* **19**, 716-732.

[2] Allison, T. and D. V. Cicchetti (1976). Sleep in Mammals: Ecological and Constiutional Correlates. *Science,* **194**, 732-734.

[3] Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford University Press, Oxford.

[4] Belin, T. R. and Rubin, D. B. (1995). The Analysis of Repeated-Measures Data on Schizophrenic Reaction Times Using Mixture Models. *Statistics in Medicine*, to Appear.

[5] Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian Computation and Stochastic System (With Discussion). *Statistical Science,* **10**, 10-36.

[6] Bowman, A. W. (1984). Jakknife Approximation to Bootstrap Estimate. *Annals of Statistics,* **12**, 101-18.

[7] Burman, P. (1989). A comparative Study of Ordinary Cross-Validation v-Fold Cross-Validation, and the Repeated Learning-Testing Methods. *Biometrika,* **76**, 503-514.

[8] Burman, P. (1989). Estimation of Optimal Transformations Using v-Fold Cross Validation and Repeated Learning-Testing Methods. *Sankhya,* **A 51**.

[9] Draper, D. (1995b). Discussion of "Model Uncertainty, Data Mining, and Statistical Inference," by C. Chatfield. *Journal of Royal Statistical Society, Ser. A,* **158**, 419-466.

[10] Draper, D. (1996). Comment: Utility, Sensitivity Analysis, and Cross-Validation in Bayesian Model-Checking. *Statistica Sinica,* **6**, 29-35.

[11] Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvment on Cross-Validation. *Journal of the American Statistical Association,* **78**, 316-331.

[12] Efron, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule?. *Journal of the American Statistical Association,* **81**, 461-470.

[13] Geisser, S. (1975). The Predictive Sample Reuse Method With Applications. *Journal of the American Statistical Association,* **70**, 320-328.

[14] Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model Determination Using Predictive Distributions, With Implementing Via Sampling-Based Methods. *In Bayesian Statistics,* **4**, 147-167.

[15] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis.* New York: Chapman and Hall.

[16] Gelman, A., Meng, X. L. and Stern, H. S. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies (With Discussion). *Statistica Sinica,* **6**.

[17] Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996). *Strategies for Improving MCMC. In Markov Chain Monte Carlo in Practice.* New York: Chapman and Hall.

[18] Guttman, I. (1967). The Use of the Concept of a Future Observation in Goodness-of-Fit Problems. *Journal of Royal Statistical Society, Ser. B,* **29**, 83-100.

[19] Härdle, W., Marron, J. S. (1985). Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *Annals of Statistics,* **12**, 76-86.

[20] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika,* **57**, 97-109.

[21] Herzberg, P. A. (1969). The Parameters of Cross-Validation. Monograph Supplement to *Psychometrika,* **34**.

[22] Herzberg, G., and Tsukanov, S. (1986). A Note on Modifications of the Jackknife Criterion on Model Selection. *Utilitas Mathematics,* **29**, 209-216.

[23] Hjorth, J. S. U. (1994). *Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap.* UK: Chapman and Hall.

[24] Hwang, J. T., Casell, G., Robert, C., Wells, M. T. and Farrell, R. H. (1992). Estimation of Accuracy in Testing. *Annals of Statistics,* **20**, 490-509.

[25] Lawless, J. F. (1982). *Satatistical Models and Methods for Lifetime Data.* New York: John Wiley & Sons.

[26] Mallows, C. L. (1973). Some Comments on C_p. *Technometrics,* **15**, 661-675.

[27] Meng, X. L. (1994). Posterior Predictive p-values. Annals of Statistics, **22**, 1142-1160.

[28] Metroplis, Roseenbuth, Teller and Teller (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics,* **21**, 1087-1092.

[29] Neal, M. R. (1993). *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* Technical Report **CRG-TR-93-1**, Department of Computer Science, University of Toronto.

[30] Nelson, W. B. (1970a). Statistical Methods for Accelerated Lifetest Data-the Inverse Power Law Model. General Electric Co. Technical Report **71-C-011**. New York: Schenectady.

[31] Picard, R. R., and Cook, R. D. (1984). Cross-Validation of Regression Models. *Journal of American Statistical Association,* **79**, 575-583.

[32] Rubin, D. B. (1981). Estimation in Parallel Randomized Experiments. *Journal of Educational Statistics,* **6**, 377-400.

[33] Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statisticain. *Annals of Statistics,* **12**, 1151-1172.

[34] Shibata, R. (1981). An Optimal Selection of Regression Variables. *Biometrika,* **68**, 45-54.

[35] Smith, A. F. M. and Roberts, G. O. (1994). Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods (With Discussion). *Journal of the Royal Statistical Society, Ser. B,* **55**, 3-23.

[36] Smyth, P. (1998). *Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood.* Technical Report **UCI-ICS 98-09**, Information and Computer Science, University of California, Irvine.

[37] Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Prediction. *Journal of Royal Statistical Society, Ser. B,* **36**, 111-133.

[38] Stone, M.(1977). Cross Validation: A Review. *Math. Oper. Statist., Ser. Statist.,* **9**, 127-39.

[39] Stone, M. (1977a). An Asymptotic Equivalence of Choice and Models by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society, Ser. B,* **39**, 44-47.

[40] Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of American Association,* **88**, 486-494.

[41] Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions (with Discussion). *Annals of Statistics,* **22**, 1701-1762.

[42] Upadhyay, S. K. and Smith, A. F. M. (1993). *A Bayesian Approach to Model Comparison in Reliability Via Predictive Simulation.* Technical Report, Department of Mathematics, Imperial College, London.

[43] Weisberg, S. (1985). *Applied Linear Regression. Second Edition.* New York: John Wiley & Sons.

[44] West, M. (1986). Bayesain Model Monitoring. *Journal of Royal Statistical Society, Ser. B,* **48**, 70-78.

[45] Zhang, P. (1993). Model Selection Via Multifold Cross-Validation. *Annals of Statistics,* **21(1)**, 299-313.

Appendix A

Bayesian Analysis of the Classical Regression Model

The normal linear model is simple enough that we can determine the posterior predictive distribution analytically. In the ordinary linear regression, the observation errors are independent and have equal variance;

$$y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I),$$

where X is the $n \times k$ matrix of explanatory variables and I is the $n \times n$ identity matrix.

Prior distribution

It is convenient in the normal regression model, to choose the noninformative prior distribution to be uniform on $(\beta, log\sigma)$ or, equivalently,

$$p(\beta, \sigma^2) \propto \sigma^{-2}.$$

It gives acceptable results and requires less effort than specifying prior information in probabilistic terms. However, this is only true when we have a large sample size and a few parameters, but if we have a small sample size or a large number of parameters, then it is important to specify prior knowledge or perhaps use hierarchical models.

The posterior distribution

To characterize the joint posterior distribution for β and σ^2 as

$$p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)p(\sigma^2|y), \tag{A.1}$$

it is convenient to draw simulations of σ and then $\beta|\sigma^2$. First we introduce the conditional distribution of the (vector) parameter β given σ^2

$$\beta|\sigma^2, y \sim N(\hat{\beta}, V_{\beta}). \tag{A.2}$$

By completing the square we have

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{A.3}$$

$$V_{\beta} = (X'X)^{-1}. \tag{A.4}$$

The posterior distribution of σ^2 from (A.1) can be written as

$$p(\sigma^2|y) = \frac{p(\beta, \sigma^2|y)}{p(\beta|\sigma^2, y)},$$

which has a scaled inverse-χ^2

$$\sigma^2|y \sim Inv - \chi^2(n-k, s^2), \tag{A.5}$$

where

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})'(y - X\hat{\beta}). \tag{A.6}$$

The density function of σ^2 is $p(\sigma^2) = Inv - \chi^2(\sigma^2|n-k, s^2)$, which is useful for variance parameters in normal models. Note that the scaled Inverse-χ^2 is a special case of the inverse-gamma distribution with $\alpha = \frac{n-k}{2}$ and $\beta = \frac{n-k}{2}s^2$.

The posterior predictive distribution

Based on the standard results by Gelman et al. (1995), we determine the posterior predictive distribution for the linear regression model. Their approach is to consider first the conditional posterior predictive distribution $p(\tilde{y}|\sigma^2, y)$ where $\tilde{y}$ is the future observable, then to average over the posterior $\sigma^2|y$. The predicted outcome $\tilde{y}$ has a normal distribution given σ^2 and by averaging over β its mean is given by

$$\begin{aligned} E(\tilde{y}|y) &= E(E(\tilde{y}|\beta, \sigma^2, y)|\sigma^2, y) \\ &= E(\tilde{X}\beta|\sigma^2, y) \\ &= \tilde{X}\hat{\beta} \end{aligned}$$

where $\tilde{X}$ is the matrix of explanatory variables.

Similarly, the variance of the future observation $\tilde{y}$ is

$$\begin{aligned} Var(\tilde{y}|\sigma^2, y) &= E[var(\tilde{y}|\beta, \sigma^2, y)|\sigma^2, y] + var[E(|\beta, \sigma^2, y)|\sigma^2, y] \\ &= E[\sigma^2 I|\sigma^2, y] + var[\tilde{X}\beta|\sigma^2, y] \\ &= (I + \tilde{X}V_\beta\tilde{X}')\sigma^2. \end{aligned}$$

We determine the posterior predictive distribution by averaging over the posterior predictive distribution of σ^2. Then the posterior predictive distribution $p(\tilde{y}|y)$ is multivariate t with center β, $n-k$ degrees of freedom, mean

$E(\tilde{y}|y)$, and variance $var(\tilde{y}|\sigma^2, y)$. We may extend these results to fit into the case of splitting the data into training and validation samples. Based on the training sample we first compute $\hat{\beta}$ and V_β, which in practice is easy using any standard linear regression software. Then we derive the mean and the variance of the predictive distribution of the validation sample given the training sample by

$$E(y^{rep}|y_T) = \tilde{X}_V\hat{\beta}, \tag{A.7}$$

$$Var(y^{rep}|y_T) = E(\sigma^2|y_T)\{I + \tilde{X}_V V_\beta \tilde{X}_V'\}, \tag{A.8}$$

where $E(\sigma^2|y_T)$ based on the training set is equal to s^2 (A.6).

Appendix B

Inverse Chi-square

The inverse-χ^2 is a special case of the gamma distribution, with $\alpha = v/2$ and $\beta = \frac{1}{2}$, where v is the degrees of freedom. In addition to the standard parameterization, we also define the scaled inverse chi-square distribution. To obtain a simulation draw θ from the Inv-$\chi^2(v, s^2)$ distribution, first draw X from the χ^2_v distribution and then let $\theta = vs^2/X$.