Bagging and the Bayesian Bootstrap

Merlise A. Clyde and **Herbert K. H. Lee**
Institute of Statistics & Decision Sciences
Duke University
Durham, NC 27708

Abstract

Bagging is a method of obtaining more robust predictions when the model class under consideration is unstable with respect to the data, i.e., small changes in the data can cause the predicted values to change significantly. In this paper, we introduce a Bayesian version of bagging based on the Bayesian bootstrap. The Bayesian bootstrap resolves a theoretical problem with ordinary bagging and often results in more efficient estimators. We show how model averaging can be combined within the Bayesian bootstrap and illustrate the procedure with several examples.

1 INTRODUCTION

In a typical prediction problem, there is a trade-off between bias and variance, in that after a certain amount of fitting, any increase in the precision of the fit will cause an increase in the prediction variance on future observations. Similarly, any reduction in the prediction variance causes an increase in the expected bias for future predictions. Breiman (1996a) introduced bagging as a method of reducing the prediction variance without affecting the prediction bias.

Bagging is short for "Bootstrap AGGregatING" which describes how it works. The idea is straightforward. Instead of making predictions from a single model fit to the observed data, bootstrap samples are taken of the data, the model is fit to each sample, and the predictions are averaged over all of the fitted models to get the bagged prediction. Breiman explains that bagging works well for unstable modeling procedures, i.e. those for which the conclusions are sensitive to small changes in the data, such as neural networks, classification and regression trees (CART), and variable selection for regression (Breiman, 1996b). He also gives a theoretical explanation of how bagging works, demonstrating the reduction in mean-squared prediction error for unstable procedures.

In this paper, we consider a Bayesian version of bagging based on Rubin's Bayesian bootstrap (1981). This overcomes a technical difficulty with the usual bootstrap in bagging, and it leads to a reduction in variance over the bootstrap for certain classes of estimators. Another Bayesian approach for dealing with unstable procedures is Bayesian model averaging (BMA) (Hoeting et al., 1999). In BMA, one fits several models to the data and makes predictions by taking the weighted average of the predictions from each of the fitted models, where the weights are posterior probabilities of models. We show that the Bayesian bootstrap and Bayesian model averaging can be combined. We illustrate Bayesian bagging in a regression problem with variable selection and a highly influential data point, a classification problem using logistic regression, and a CART model.

2 BOOTSTRAPPING

Suppose we have a sample of size n, with observed data $Z_1, \ldots, Z_n$ where Z_i is a vector in $\mathcal{R}^{p+1}$ and the Z_i are independent, identically distributed realizations from some distribution $F \in \mathcal{F}$. Here $\mathcal{F}$ is the class of all distribution functions on $\mathcal{R}^{p+1}$. The parameter of interest is a functional $T(F)$ where T is a mapping from $\mathcal{F}$ to $\mathcal{R}$ or $\mathcal{R}^k$ in the case of vectored valued functions, for example, the $p+1$ dimensional mean of the distribution, $\mu = \int z dF$.

In Efron's bootstrap and the Bayesian bootstrap, the class of distribution functions is restricted to a parametric model by restricting estimation to $F_n \in \mathcal{F}_n$, where F_n is represented as

$$F_n = \sum_{i=1}^{n} \omega_i \delta_{Z_i},$$

δ_{Z_i} is a degenerate probability measure at Z_i, and ω_i is the weight associated with Z_i, $\omega_i \geq 0$, $\sum \omega_i = 1$.

In the bootstrap, the distribution of $T(F)$ is obtained by repeatedly generating bootstrap replicates, where one bootstrap replicate is a sample drawn with replacement of size n from $Z_1, \ldots, Z_n$. The bootstrap distribution of $T(F)$ is based on considering all possible bootstrap replications $T(F_n^{(r)})$ where $\omega_i^{(r)}$ in $F_n^{(r)}$ corresponds to the proportion of times Z_i appears in the rth bootstrap replicate, with $\omega_i^{(r)}$ taking on values in $\{0, 1/n, \ldots, n/n\}$. For example, the bootstrap mean for the rth replicate is calculated as $\hat{Z}^{(r)} = \sum_{i=1}^n \omega_i^{(r)} Z_i$, and the bagged estimate of the mean is the average over all bootstrap replicates, $\sum_{r=1}^R \hat{Z}^{(r)}/R$, where R is the number of bootstrap replicates. Of course, one cannot usually consider all possible bootstrap samples, which is $\binom{2n-1}{n}$, and bagging is often based on a much smaller set of bootstrap replicates, say 25 to 50 (Breiman, 1996a).

3 BAYESIAN BOOTSTRAP

The Bayesian bootstrap was introduced by Rubin (1981) as a Bayesian analog of the original bootstrap. Instead of drawing weights ω_i from the discrete set $\{0, \frac{1}{n}, \ldots, \frac{n}{n}\}$, the Bayesian approach treats the vector of weights ω in F_n as unknown parameters and derives a posterior distribution for ω, and hence $T(F)$. Rubin (1981) used a non-informative prior, $\prod_{i=1}^n \omega_i^{-1}$, which when combined with the multinomial likelihood for Z, leads to a Dirichlet$(1, \ldots, 1)$ distribution for the posterior distribution of ω. The posterior distribution of $T(F)$ is estimated by Monte Carlo methods: generate $\omega^{(b)}$ from a Dirichlet$(1, \ldots, 1)$ distribution and then calculate $T(F_n^{(b)})$ for each sample $\omega^{(b)}$. The average of $T(F_n^{(b)})$ over the samples corresponds to the Monte Carlo estimate of the posterior mean of $T(F)$ and can be viewed as a Bayesian analog of bagging.

Although there are differences in interpretation, operationally the ordinary bootstrap and Bayesian bootstrap differ primarily in how the values of ω are drawn. As Rubin (1981) shows, the expected values of the weights ω are equal to $1/n$ under both bootstrap methods. As the expectations of the weights are the same, both ordinary bagging and Bayesian bagging will have the same expectation for functions $T(F_n)$ that are linear in the weights ω, such as means. There are situations, which will be discussed later, where the ordinary bootstrap distribution is not well defined, and the two approaches may yield different answers. Both approaches also lead to the same correlation between weights. However, the variability of the weights ω under the ordinary bootstrap is $(n+1)/n$ times the variance of ω under the Bayesian bootstrap. For linear functionals, the variance of the estimate under the Bayesian bootstrap is therefore strictly less than the variance under the ordinary bootstrap. This applies directly for CART models; for other estimators that are not necessarily linear in the weights, in our experience the Bayesian bootstrap has also empirically exhibited less variability than the ordinary bootstrap. We illustrate this reduction in the rat liver example.

4 BAGGING VIA THE BAYESIAN BOOTSTRAP

Bagging is used primarily in prediction problems, and with that in mind we partition each Z_i into a response Y_i (which could be continuous or categorical) and a p dimensional vector of input variables x_i for predicting Y. In matrix form, the data are $Y = (y_1, \ldots, y_n)'$ with a $n \times p$ matrix of covariates X with rows x_i.

Under the nonparametric model for the data, the only unknown quantity is the distribution F, or the parameters ω under the restricted class of distribution functions. The posterior distribution on ω induces a posterior distribution on a functional $T(F)$ for predicting Y given X. We first consider the case where interest is in regression-type estimates, then extend the procedure to allow for variable selection, nonlinear functions, categorical responses and model uncertainty.

4.1 LINEAR REGRESSION

For making predictions based on linear combinations of Y, we consider functionals of the form

$$\begin{aligned} \hat{\beta} = T(F) &= \arg\min_{\beta} \int ||Y - X\beta||^2 dF \qquad (1) \\ &= \arg\min_{\beta} \sum_{i=1}^n \omega_i (y_i - x_i\beta)^2 \\ &= (X'WX)^{-1} X'WY \end{aligned}$$

where W is a diagonal matrix of weights ω. The values of $\hat{\beta}$, that minimize (1) with the restriction to $\mathcal{F}_n$, are equivalent to weighted least squares estimates using weights ω.

Operationally, Bayesian bagging (BB) proceeds by taking a sample $\omega^{(b)}$, from a Dirichlet$(1, \ldots, 1)$ distribution, and then using weighted least squares to obtain

$$\hat{\beta}^{(b)} = (X'W^{(b)}X)^{-1} X'W^{(b)} Y$$

where $W^{(b)}$ is a diagonal matrix of weights $\omega^{(b)}$. This is repeated for $b = 1, \ldots, B$, where B is the total number of Monte Carlo samples, to obtain the posterior distribution of $\hat{\beta}$, and the posterior distribution of $\hat{Y} = X\hat{\beta}$ or other functions of ω. Let $\hat{Y}^{(b)} = X\hat{\beta}^{(b)}$. The BB estimate of $\hat{Y}$ given X is the Monte Carlo average

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^{B} \hat{Y}^{(b)} . \qquad (2)$$

4.2 VARIABLE SELECTION

While linear regression is a stable procedure, where bagging does not lead to substantial improvements, variable selection is viewed as being unstable. The BB procedure is modified to combine model selection with parameter estimation, where for each sample of $\omega^{(b)}$, one selects a model $M^{(b)}$ using an appropriate model selection criterion and then estimates $\hat{\beta}_M^{(b)}$ under model $M^{(b)}$. The posterior distribution for $\hat{Y}$ and the posterior mean are now based on multiple models where the BB estimate of Y given X is

$$\hat{Y} = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}_M^{(r)}$$

where $\hat{Y}_M^{(r)} = X_{M^{(r)}}\hat{\beta}_M^{(r)}$ is the prediction using design matrix $X_{M^{(r)}}$ based on model M. Although not equivalent to Bayesian model averaging as described in Hoeting et al. (1999), the above estimator is a variant of model averaging as the bootstrap aggregation results in averaging over different models.

4.3 MODEL AVERAGING

Bayesian model averaging can be introduced into Bayesian bootstrap estimates by replacing (1) by the Bayes risk for squared error loss with model uncertainty. For sample b, the BMA estimate of $\hat{Y}^{(b)}$ is the weighted average

$$\hat{Y}_{\text{BMA}}^{(b)} = \sum_M \pi(M|X, Y, \omega^{(b)})\hat{Y}_M^{(b)} \tag{3}$$

where $\pi(M|X, Y, \omega^{(b)})$ is the "posterior probability" of model M and the predicted values $\hat{Y}_M^{(b)} = X\hat{\beta}_M^{(b)}$ and coefficients $\hat{\beta}^{(b)}$ are calculated given model M, using weights $\omega^{(b)}$. These are combined to form the BB BMA predictions,

$$\hat{Y} = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}_{\text{BMA}}^{(b)}$$

which incorporate any instability in the model weights due to changes in the data. Posterior model probabilities may be based on the BIC (Schwarz, 1978),

$$\begin{aligned} \text{BIC}_M &= \text{RSS}_M + p_M \log(n) & (4) \\ \pi(M|X, Y, \omega^{(b)}) &= \frac{\exp(-.5\text{BIC}_M)}{\sum_M \exp(-.5\text{BIC}_M)} & (5) \end{aligned}$$

where RSS_M is the residual sum of squares under model M using weighted least squares and p_M is the number of parameters under model M. Of course, other prior specifications may lead to other posterior model probabilities, however, for many cases BIC does lead to consistent model selection and is a useful default (Hoeting et al., 1999).

4.4 NEURAL NETWORKS AND OTHER NONLINEAR ESTIMATORS

For continuous responses, the linear predictor $X\beta$ in (1) can be replaced by nonlinear functions as in neural nets or generalized additive models. While we no longer have an explicit solution for $\hat{\beta}^{(b)}$, any code for fitting neural networks (or other nonlinear models) that allows weights can be used to construct the BB predictions, where one substitutes $\hat{Y}^{(b)}$ using predictions from the neural network in (2). For model averaging with neural networks with continuous responses, model probabilities based on (4-5) are still appropriate.

4.5 EXPONENTIAL FAMILY MODELS, GLMS AND CART

For continuous responses, linear regression predictions were based on minimizing a residual sum of squares (1) which is equivalent to maximizing a normal likelihood. While the nonparametric bootstrap model implies a multinomial likelihood for the data Z, the use of likelihood score functions based on alternative distributional assumptions to provide estimates and predictions is in the same spirit as generalized estimating equations (Liang and Zeger, 1986). In this vein, we can extend the BB approach to other model classes, such as exponential families, CART models, and neural networks for categorical responses, to allow for categorical and discrete response variables. The connection between iteratively reweighted least squares and maximum likelihood estimation provides the basis for computations using the Bayesian bootstrap weights ω.

For exponential family models, the log likelihood can be written as

$$l(\theta) = \sum_{i=1}^{n} \frac{v_i}{\phi}(y_i\theta_i - b(\theta)) + c(y_i, \phi)$$

where θ is the canonical parameter, v_i is a known prior weight, and ϕ is a dispersion parameter; the mean parameter is $\mu_i = b'(\theta_i)$ (McCullagh and Nelder, 1989). As in GLMs, we express μ_i as a function of x_i and β, $\mu = f(X, \beta)$ (although not necessarily through a linear predictor $X\beta$). Incorporating the bootstrap weights ω_i, into the exponential family weights v_i, so that $w_i^{(b)} = v_i\omega_i^{(b)}$, we find the bootstrap estimate $\hat{\beta}^{(b)}$ that maximizes $l(\theta(\beta))$ using weights $v_i^{(b)}$. This is repeated to provide $b = 1, \ldots, B$ samples and a Monte Carlo estimate of the posterior distribution of $f(X, \hat{\beta})$. The BB estimate $\hat{Y}$ is

$$\hat{Y} = \frac{1}{B}\sum_{b=1}^{B} f(X, \hat{\beta}^{(b)}).$$

Any GLM or CART software that allows weights can be used to construct the estimates $f(X, \hat{\beta}^{(b)})$ for BB. BB with BMA can be carried out by replacing the residual sum of squares in the expression for the BIC (4) with the residual deviance for the model.

5 EXAMPLES

5.1 RAT LIVERS

Weisberg (1985, p. 121) describes a dataset on drug uptake in rat livers. The experimental hypothesis is that because dose was matched to weight, there would be no relationship between the percent dose in the liver and the three input variables (body weight, liver weight, and relative dose). One rat (case 3), however, received a large dose relative to its weight and is an influential point, leading to a rejection of the experimental hypothesis (the null model which is believed to be true). Regression is normally thought of as a stable procedure, where methods such as bagging will not help. However, in the presence of outliers and influential points (e.g., case 3), regression is no longer stable. Variable selection further contributes to instability.

The left plot of Figure 1 shows body weight versus dose, where one can see both the high correlation as well as the highly influential point (case 3) at weight 190, which causes instability in the linear models. The right plot of Figure 1 shows predicted percent dose in the liver by body weight. The experimenters expected no relationship between percent dose in the liver and body weight; deviations from a horizontal line are because of nonzero regression coefficients for input variables. Predictions from the full model provide the greatest deviations, demonstrating the trouble caused by case 3. Bagging and BB of the full model (Bag/BB in the plot) have nearly identical predictions, both providing additional shrinkage towards the overall mean, reducing the effect of case 3. BMA produces the best fit, with virtually no change under bagging BMA or BB BMA, as the null model receives the highest posterior probability under the complete data.

The series of boxplots in Figure 5.1 highlights the variation in predictions for case 3. Of particular interest is that BB estimates show a large reduction in variation over estimates using bagging (without or with BMA). BMA is more stable, with greater shrinkage towards the overall mean of the data (without case 3), which is in the direction expected by the experimenters. Even with BMA, bagging occasionally produces large deviations. While estimates under bagging and BB are comparable, the sampling distribution under the bootstrap exhibits more variability than the posterior distribution under the Bayesian bootstrap.

5.2 OZONE

Ground level ozone data were analyzed in the original bagging paper (Breiman, 1996a). The dataset consists of daily readings of the maximum ozone concentration at ground level in Los Angeles over the course of a year, with 9 meteorological predictors. Eliminating cases with missing data leaves 330 complete records. Following Breiman, we: (i) randomly selected a test set of 15 cases; (ii) fit a single regression tree using ten-fold cross-validation on the remaining cases (the training data), and then used this fitted tree to predict on the test data; (iii) for $b = 1, \ldots, 25$ generated $\omega^{(b)}$ which were used as weights to fit a regression tree using ten-fold cross validation on the training data, and then used this fitted tree to predict on the test data; the average of the 25 predictions is the BB prediction. This process was repeated 500 times. The average mean squared error (MSE) of prediction was calculated for the single tree model and the BB method. The MSE for the single tree model was 23.9% (standard error 0.47), and the MSE for the BB predictions was 18.6% (standard error 0.35) resulting in a 22.0% reduction in error due to the BB, comparable to Breiman's results.

5.3 DIABETES

Smith et al. (1988) introduced a dataset on the prevalence of diabetes in Pima Indian women. The goal is to predict the presence of diabetes using seven health-related covariates. There are 532 complete records, of which 200 are used as a training set and the other 332 are used as a test set. The data are available at `http://www.ics.uci.edu/~mlearn/MLSummary.html` We used logistic regression for classifying the data, which is essentially equivalent to fitting a neural network with a single hidden node. Ripley (1996) pointed out that there is no gain in fit by using more hidden nodes, so logistic regression is a sufficiently flexible procedure. We find that for each of three model classes (the full model, the best BIC model, and BMA), the Bayesian bootstrap improves predictions, as shown in Table 1, with error rates comparable to bagging (Breiman, 1996a).

6 DISCUSSION

Since Breiman introduced bagging, a number of papers have demonstrated its effectiveness in a variety of contexts. The Bayesian bootstrap leads to similar improvements in prediction rates, with less apparent variability. The approach in this paper is technically equivalent to the weighted likelihood bootstrap (WLB) of Newton and Raftery (1994), which appeared in a different context. They used the WLB as a tool to

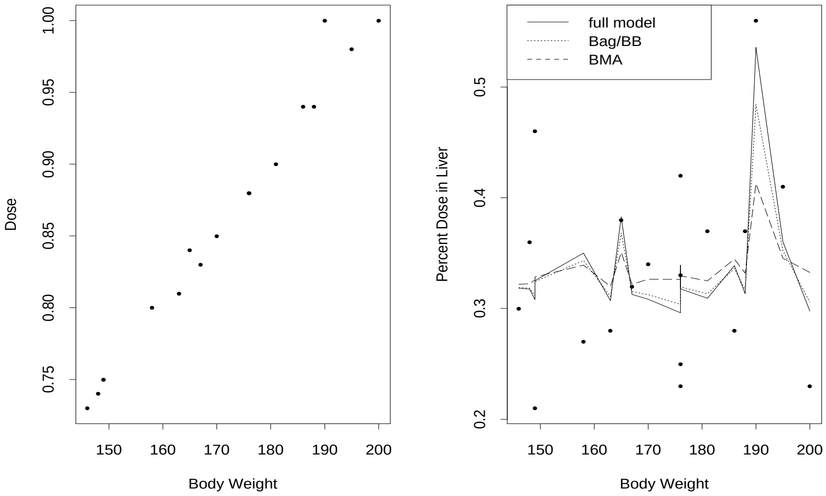

Figure 1: Rat Data: A Point with High Influence and the Resulting Fitted Models

Table 1: Percent Misclassification for the Pima Diabetes Data

Method	% Misclassification
Best single model using BIC	20.2
Bayesian bootstrap on the best BIC model	19.3
Full model	20.2
Bayesian bootstrap on the full model	20.0
Bayesian model averaging	21.1
Bayesian bootstrap and model averaging	20.9

approximate posterior distributions for standard likelihoods, which today can be readily approximated using Markov chain Monte Carlo sampling.

An alternative view of the Bayesian bootstrap is that the data arise from a nonparametric model with distribution F. In this case, $T(F)$ is not necessarily a parameter in the model, but is taken as an interesting summary of the distribution. As nonparametric models, both the bootstrap and Bayesian bootstrap share the problem that they only give positive weight to values of (x, y) that were actually observed. This raises theoretical issues when the bootstrapped quantities are used for prediction at values of x that are not in the original dataset. Other problems with both bootstrap methods are raised by Rubin (1981).

Another problem with using the bootstrap for bagging is that weights may be 0, and bootstrap replicates where X is not of full rank receive positive probability. In these samples, $\hat{\beta}$ is not well defined, although predictions for Y are still defined. Even though these cases have very low probability (and may not appear in samples), they do contribute to the theoretical bootstrap distribution of $T(F)$. This problem is not specific to regression, but occurs in all estimation cases where more than one distinct data point is necessary for well-defined estimates. Further examples of problematic procedures include non-linear and nonparametric regression and estimation of standard deviations or correlations. The use of the Bayesian bootstrap avoids this issue.

Rubin's Bayesian bootstrap can be viewed as a limiting case of the nonparametric distribution for F using a Dirichlet process prior (Gasparini, 1995). If the prior for F is a Dirichlet process with parameter α ($DP(\alpha)$), then the posterior distribution for F is again a Dirichlet process $DP(\alpha+\sum_{i=1}^{n}\delta_{Z_i})$ (Ferguson, 1973). In the limit as $\alpha(\mathcal{R}^{p+1})$ goes to 0, the posterior distribution is a $DP(\sum_{i=1}^{n}\delta_{Z_i})$, with mean equal to the empirical cumulative distribution function. It is this limiting noninformative case that is equivalent to Rubin's Bayesian bootstrap (Gasparini, 1995).

Fully non-parametric or semi-parametric Bayesian

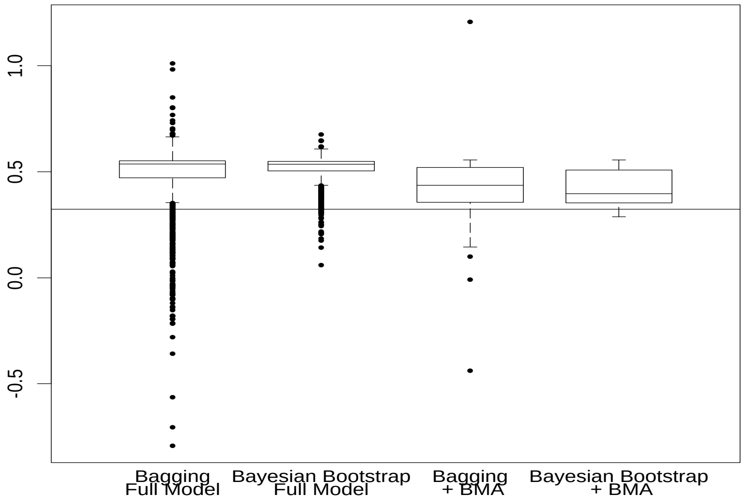

Figure 2: Distribution of Estimated Means for Case 3 in the Rat Liver Data Set. The Horizontal Line is the Overall Mean without Case 3.

models that can adapt to nonlinearities and are robust to model mis-specification are other useful alternatives to bagging or BB. While these are typically more computationally intensive than either form of bootstrapping, they may resolve theoretical problems with bootstrapping noted by Rubin that are inherited by both bagging and BB. As computing environments improve, these approaches may see wider use.

Acknowledgments

This research was partially supported by NSF grants DMS 9733013 and DMS 9873275. We would like to thank Steve MacEachern for helpful discussions.

References

Breiman, L. (1996a). "Bagging Predictors." *Machine Learning*, 26, 2, 123–140.

— (1996b). "Heuristics of Instability and Stabilization in Model Selection." *The Annals of Statistics*, 24, 2350–2383.

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics*, 1, 2, 209–230.

Gasparini, M. (1995). "Exact Multivariate Bayesian Bootstrap Distributions of Moments." *Annals of Statistics*, 23, 762–768.

Hoeting, J. A., Madigan, D., Raftery, A., and Volinsky, C. T. (1999). "Bayesian Model Averaging: A Tutorial (with discussion)." *Statistical Science*, 14, 4, 382–417.

Liang, K.-Y. and Zeger, S. L. (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, 73, 13–22.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (Second Edition)*. Chapman & Hall.

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap (with discussion)." *Journal of the Royal Statistical Society B*, 56, 3–48.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rubin, D. B. (1981). "The Bayesian Bootstrap." *Annals of Statistics*, 9, 130–134.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, 6, 2, 461–464.

Smith, J., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus." In *Proceedings of the Symposium on Computer Applications and Medical Care*, 261–265. IEEE Computer Society Press.

Weisberg, S. (1985). *Applied Linear Regression*. 2nd ed. New York: John Wiley & Sons.