UW Biostatistics Working Paper Series

12-26-2007

Model-Robust Bayesian Regression and the Sandwich Estimator

Adam A. Szpiro
University of Washington, aszpiro@u.washington.edu

Kenneth M. Rice
University of Washington, kenrice@u.washington.edu

Thomas Lumley
University of Washington, tlumley@u.washington.edu

Suggested Citation

Szpiro, Adam A.; Rice, Kenneth M.; and Lumley, Thomas, "Model-Robust Bayesian Regression and the Sandwich Estimator" (December 2007). *UW Biostatistics Working Paper Series.* Working Paper 320.
http://biostats.bepress.com/uwbiostat/paper320

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

1 Introduction

Epidemiological studies typically involve data on many hundreds of individuals, and sample sizes of several thousand are not uncommon. The datasets involved are therefore large, but although this makes accurate estimation feasible, standard parametric models almost always fit these data poorly, detracting from the prima facie validity of model-based inference. A modern frequentist approach to this problem is to dispense with the requirement of a correctly-specified parametric model: by stating, non-parametrically, what we want to know about the data-generating mechanism, estimating equations and associated 'robust' or 'sandwich'-based intervals provide accurate large-sample inference, at no more computational effort than fitting a generalized linear model. We investigate an analogous Bayesian approach to model-robust inference. Taking the example of multivariate linear regression, we show that simple Bayesian methods can provide intervals with the same model-robustness as the frequentist approach. We also discuss where the addition of prior information may additionally provide better small-sample properties.

The goal of multivariate linear regression is to find β such that $Y = X\beta$ holds approximately, where Y is an n-vector of dependent variable observations, X is an $n \times m$ matrix of covariate observations, and β is an m-vector of coefficients. Since the early work of Gauss [1] and Legendre [2] it has been understood that the least squares criterion and the associated estimate

$$\hat{\beta}_{ls} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i\beta)^2$$

are appropriate for many applications. This paper is concerned with a Bayesian formulation that is firmly rooted in the idea of minimizing the least squares criterion and that does not rely on assumptions that the data follows an underlying linear relationship, or has constant variance. Our objective is to formulate the problem such that $\hat{\beta}_{ls}$ is the Bayesian point estimate with an appropriate measure of uncertainty that is robust to deviations from the standard modeling assumptions.

A model-based version of Bayesian linear regression can be found in [3]. Regard X as being constructed of n observations of an m-vector x. Assume that Y is derived as n independent samples from the random variable y which is distributed conditionally on x as

$$y|x, \beta, \sigma^2 \sim N(x\beta, \sigma^2) \tag{1}$$

with minimally informative priors on the parameters $p(\beta, \sigma^2) \propto \sigma^{-2}$. It is natural to take as a point estimate the posterior mean

$$\hat{\beta}_{model} = E\beta,$$

and it turns out that this is the least squares estimate $\hat{\beta}_{model} = \hat{\beta}_{ls}$. As a measure of uncertainty it is natural to use the diagonal of the posterior covariance matrix

$$\hat{\sigma}_{model} = \operatorname{diag}\left[\operatorname{Cov}(\beta)\right]^{1/2},$$

which turns out to be the classical frequentist standard error estimate.

The limitations of this model-based formulation are both conceptual and practical. Conceptually, it is unsatisfying to assume that there is a true underlying linear relationship

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

between x and the mean of y and, further, that the random deviation of y around the mean is homoscedastic. It is an extremely rare situation in applications where we believe that a truly linear relationship exists, and that the statistical challenge is simply to find the coefficient. Rather, what we are typically interested in is finding the best approximation to a linear relationship. Furthermore, whatever the relationship between the mean of y and x, we generally do not know if the deviations have constant variance.

On a practical level, we are happy to use the least squares point estimate because it has a familiar interpretation, regardless of model validity. But the uncertainty estimates are a different matter. It is widely held that Bayesian credible intervals should give correct frequentist coverage if the estimation procedure is to be regarded as successful. See Chapter 4 in [3] as well as [4], [5], [6], [7]. However, $\hat{\sigma}_{model}$ is exactly the classical standard error estimate, and violation of either linearity or homoscedasticity leads to asymptotically incorrect frequentist coverage properties. This fact is well known and is easy to verify with simulation studies. An interesting example with a discrete covariate space can be found in [8]. Reference [9] contains a discussion in the Bayesian context with some examples.

A number of Bayesian approaches have been proposed to accommodate heteroscedasticity and nonlinearity. A natural way to deal with heteroscedasticity is to put a more flexible prior on the variance. Several parametric priors are described in [3], [10], and the references therein. More generally, recently-developed non-parametric Bayes methods offer asymptotic convergence of the fitted model to the truth, but (despite the name) this is achieved by use of a fixed class of highly parameterized models, requiring cutting-edge computation, and particular caution in choice of priors [11]. In any case, these methods can result in a different point estimate from the least squares fit. We prefer the least squares paradigm due to its familiar interpretation, regardless of model validity.

Berger has developed a Bayesian theory of robust inference that is separate from concerns about frequentist sampling properties [12]. The focus is on robustness in selecting priors for parameters in a given Bayesian model, but at least conceptually it is clear from the discussion that model choice can be viewed as part of the prior specification. Given the focus on priors for model parameters, the theory in [12] has not been directly applied to our problem. An example application to estimating a normal mean can be found in [13].

Some previous work has elements in common with our approach. The Bayesian method of moments [14] avoids making any assumptions about the variance structure or even the probability distribution of the dependent variable. However, it seems that this approach still requires assuming linearity, and its relationship to traditional Bayesian theory has been subject to debate; see critical comments in [15] and the rejoinder in [16]. Bayesian least squares, described in [17] and [18], emphasizes squared error losses without assuming a true model, but it does not naturally lead to robust solution of the linear regression problem. Finally, reference [19] uses the posterior predictive squared error loss for model selection but does not derive point or interval estimates based on this criterion.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

1.1 Generic Model

Instead of the standard modeling assumptions in equation (1), we assume that y is distributed conditionally on x as

$$y|x, \phi, \sigma^2 \sim N(\phi(x), \sigma^2(x)),$$

with minimally informative priors that make the functions $\phi(\cdot)$ and $\sigma^2(\cdot)$ separately identifiable. We defer being more specific so that we can separately address the discrete and continuous covariate scenarios, but we emphasize that we will assume neither linearity nor homoscedasticity, as priors based on these assumptions would make everything that follows reduce to the model-based version. We begin by describing two model-robust Bayesian formulations. The first gives only point estimates, while the second gives point and interval estimates.

1.2 Point Estimates

Our first version of a point estimate will be defined as the minimizer of the average posterior predictive squared error $(y^*(x) - x\beta)^2$, where $y^*(x)$ is a random variable with the posterior predictive distribution of y conditional on x. We need to integrate over some measure on x, and the choice of measure is based on the source of covariate observations: *conditional inference* is based on assuming that the set of observed x is fixed, and *population inference* is based on regarding the set of observed x as being sampled at random from an unknown population. The question of whether conditional or population inference is preferable dates to Fisher and Pearson [20] [21]. Some recent comments on the implications can be found in [22] and [23] for linear regression, and in the context of case-control studies in [24] and [25].

For conditional inference, we let $\mathbb{P}(x)$ be the empirical probability measure based on the n fixed observations of x, and we define the estimate

$$\hat{\beta}^*_{cond} = \underset{\beta}{\operatorname{argmin}}\, E \int (y^*(x) - x\beta)^2 \, d\mathbb{P}(x). \tag{2}$$

For population inference, we need to specify a prior for the distribution of x in the population. We will use a Dirichlet type prior, but for now it is enough to assume a parametric form $x \sim P(x|\lambda)$. We define the population estimate

$$\hat{\beta}^*_{pop} = \underset{\beta}{\operatorname{argmin}}\, E \int (y^*(x) - x\beta)^2 \, dP(x|\lambda). \tag{3}$$

The posterior expectation is over random y^* and λ, and $\hat{\beta}^*_{pop}$ can be regarded as minimizing the expected posterior predicted squared error over the random covariate distribution.

We will see that both of these estimates are equal to the least squares fit, at least in the case of discrete covariates. This is satisfying because the formulation does not depend on assuming linearity or homoscedasticity. However, it is not clear how to derive uncertainty estimates. Since the point estimates are not derived as the posterior mean of a parameter, the typical Bayesian approach of using the standard deviation of that parameter is not available.

1.3 Interval Estimates

We now introduce a second formulation based on defining a new random parameter: the least squares fit to $\phi(\cdot)$, the random mean function. We obtain point and uncertainty estimates as the expectation and standard deviation of this parameter. In the conditional inference case,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

we define the parameter

$$\theta^{\dagger}_{cond} = \underset{\beta}{\operatorname{argmin}} \int (\phi(x) - x\beta)^2 \, d\mathbb{P}(x). \tag{4}$$

Each realization of $\theta^{\dagger}_{cond}$ is the least squares fit to a realization from the posterior of $\phi(\cdot)$, ranging over the fixed set of observed x values. This is precisely the quantity we are interested in, since we are looking for a linear approximation to the variation in the mean of y as a function of x. We define

$$\hat{\beta}^{\dagger}_{cond} = E\theta^{\dagger}_{cond} \tag{5}$$

and

$$\hat{\sigma}^{\dagger}_{cond} = \operatorname{diag}\left[\operatorname{Cov}(\theta^{\dagger}_{cond})\right]^{1/2}$$

to be the associated point and uncertainty estimates. Along similar lines, we define a parameter for the population inference case

$$\theta^{\dagger}_{pop} = \underset{\beta}{\operatorname{argmin}} \int (\phi(x) - x\beta)^2 \, dP(x|\lambda), \tag{6}$$

with

$$\hat{\beta}^{\dagger}_{pop} = E\theta^{\dagger}_{pop} \tag{7}$$

and

$$\hat{\sigma}^{\dagger}_{pop} = \operatorname{diag}\left[\operatorname{Cov}(\theta^{\dagger}_{pop})\right]^{1/2}$$

as associated point and uncertainty estimates. The difference from conditional inference is that each realization of $\theta^{\dagger}_{pop}$ is the least squares fit to a realization from the posterior of $\phi(\cdot)$ ranging over a realization from the posterior distribution of x, rather than over fixed observations of x.

1.4 The Sandwich Estimator

We will see in the rest of this paper that the uncertainty estimate for population inference, $\hat{\sigma}^{\dagger}_{pop}$, is asymptotically equivalent to the Huber-White sandwich standard error [26] [27]. This is not surprising since both are obtained in a two step process that decouples uncertainty estimation from the least squares optimization, allowing nonlinearity and heteroscedasticity to affect the uncertainty estimates without changing the point estimates. The connection with the sandwich estimator is important for a number of reasons.

First, it is known that sandwich-based intervals have asymptotically correct coverage even when the assumptions of linearity and homoscedasticity are violated. Thus, we have essentially achieved our objective of defining a Bayesian paradigm for linear regression that is robust to model misspecification both conceptually (since our prior model assumes neither linearity nor homoscedasticity) and in practice (since the resulting intervals have correct frequentist coverage).

Second, Freedman [28] has argued that there are difficulties of interpretation for the sandwich estimator if the mean model is misspecified. The difficulty arises only if one views the sandwich estimator as a way to make maximum likelihood estimation more robust [29],

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper320

and alternative frequentist interpretations are available in terms of estimating equations [30]. Our approach provides an intuitive interpretation of what is being approximated in a Bayesian context by focusing on the best linear fit to a random mean function. In fact, at least for discrete covariates, we obtain a decomposition of the uncertainty into two pieces: one piece accounts for variability induced by randomness of y conditional on x, and the other piece accounts for variability induced by nonlinearity of the true mean model and randomness of x in the population. See equations (12), (13) and (15) in the proof of Theorem 3.

Finally, given the persistence of separate Bayesian and frequentist approaches to inference, it is hoped that identifying features common to both schools will advance the overall development of statistical practice. The sandwich estimator is popular in frequentist analysis, and to our knowledge the present paper is the first Bayesian derivation of equivalent interval estimates.

1.5 Outline of Paper

The plan for the rest of this paper is as follows. In Section 2 we complete specification of priors for the case of a discrete covariate, and we give explicit forms for the associated point and interval estimates. In Section 3 we describe how we apply our approach in the continuous covariate case. In Section 4, we illustrate the discrete and continuous covariate cases with simulation examples. We conclude with a discussion in Section 5.

2 Discrete Covariate

Let $\xi = (\xi_1, \ldots, \xi_K)$ consist of K non-zero deterministic m-vectors that span $\mathbb{R}^m$, and suppose that the covariate x can take these values. Let n_k be the number of $i = 1, ..., n$ such that $X_i = \xi_k$, where X_i is the ith row of X. We let λ be the diagonal matrix with entries $\lambda_1, \ldots, \lambda_K$ and and take it to be the hyperparameter for the prior distribution of x, such that

$$P(\xi_k|\lambda) = \lambda_k, \quad \sum_{k=1}^{K} \lambda_k = 1.$$

We use an improper Dirichlet prior for λ such that

$$p(\lambda) \propto \prod_{k=1}^{K} \lambda_k^{-1} \quad (0 \text{ if } \sum_{k=1}^{K} \lambda_k \neq 1).$$

The posterior distribution of λ is also Dirichlet with

$$p(\lambda|X) \propto \prod_{k=1}^{K} \lambda_k^{-1+n_k} \quad (0 \text{ if } \sum_{k=1}^{K} \lambda_k \neq 1).$$

One way to simulate values from the posterior is to draw independent gamma variates g_k with shape parameters n_k and unit scale parameters and then set $\lambda_k = g_k/(g_1, \ldots, g_K)$ [31]. There is a connection between the posterior distribution for x and bootstrap resampling [32]. We will comment on this further in Section 5.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

The discrete covariate situation is interesting because we can assume multiple samples at each covariate value, so it is straightforward to compute the posterior distribution for completely unstructured priors on $\phi(\cdot)$ and $\sigma^2(\cdot)$. We introduce vector notation $\phi = (\phi_1, \ldots, \phi_K)$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_K^2)$ with

$$\phi_k = \phi(\xi_k), \quad \sigma_k^2 = \sigma^2(\xi_k),$$

and independent non-informative priors

$$p(\phi_k, \sigma_k^2) \propto \sigma_k^{-2}.$$

We now present a series of three theorems characterizing our model-robust point and uncertainty estimates. All of these results are proved in the appendix. The first theorem establishes that the model-robust point estimates derived by minimizing the posterior predictive squared error recover the least squares point estimate.

Theorem 1 *For a discrete covariate space the estimates $\hat{\beta}^*_{cond}$ and $\hat{\beta}^*_{pop}$ defined by equations (2) and (3) take the form*

$$\hat{\beta}^*_{cond} = \hat{\beta}^*_{pop} = (X^tX)^{-1}X^tY.$$

The second theorem concerns conditional inference based on the mean and standard deviation of the parameter $\theta^\dagger_{cond}$ defined in equation (4). The point estimate is the least squares solution, and the uncertainty estimate has a sandwich form, but with a different covariance matrix than appears in the Huber-White version. The covariance matrix is based only on the variability of y conditional on x. Unlike the Huber-White version, it does not include deviations from the linear fit. This is appropriate because deviation from a linear model does not induce variability in the point estimate if we condition on the observed x.

Theorem 2 *For a discrete covariate space the estimate $\hat{\beta}^\dagger_{cond}$ defined by equation (5) takes the form*

$$\hat{\beta}^\dagger_{cond} = (X^tX)^{-1}X^tY,$$

and assuming there are at least four samples for each covariate value, the corresponding uncertainty estimate has the sandwich form

$$\hat{\sigma}^\dagger_{cond} = \mathrm{diag}\left[(X^tX)^{-1}\left(X^t\Sigma^\dagger X\right)(X^tX)^{-1}\right]^{1/2}$$

where $\Sigma^\dagger$ is the diagonal matrix defined by

$$\Sigma^\dagger_{ij} = \begin{cases} \dfrac{1}{n_k - 3} \displaystyle\sum_{l:X_l=\xi_k} (Y_l - \bar{y}_k)^2 & \textit{if } i = j \textit{ and } X_i = \xi_k \\ \\ 0 & \textit{if } i \neq j \end{cases}$$

and

$$\bar{y}_k = \frac{1}{n_k} \sum_{l:X_l=\xi_k} Y_l.$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

The third theorem concerns the population inference analogue to the previous result. Specifically it contains asymptotic results for estimates based on the mean and standard deviation of the parameter $\theta^{\dagger}_{pop}$ defined in equation (6). The point estimate is again the least squares solution and the uncertainty estimate has a sandwich form, this time with the same covariance matrix as the Huber-White version. It is appropriate that the covariance matrix includes deviations from the linear fit because these can translate into increased variability of the point estimate when we incorporate a random covariate distribution.

Theorem 3 *For a discrete covariate space, assume that y conditional on x has bounded first and second moments. The estimate $\hat{\beta}^{\dagger}_{pop}$ defined by equation (7) takes the asymptotic form*

$$\hat{\beta}^{\dagger}_{pop} - (X^tX)^{-1}X^tY \to 0,$$

and assuming there are at least four samples for each covariate value, the corresponding uncertainty estimate has the asymptotic sandwich form

$$\hat{\sigma}^{\dagger}_{pop} - \mathrm{diag}\left[(X^tX)^{-1}\left(X^t\Sigma X\right)(X^tX)^{-1}\right]^{1/2} = o(n^{-1})$$

where Σ is the diagonal matrix defined by

$$\Sigma_{ij} = \begin{cases} \left(Y_i - X_i(X^tX)^{-1}X^tY\right)^2 & \textit{if } i = j \\ 0 & \textit{otherwise.} \end{cases}$$

The results hold conditionally almost surely for infinite sequences of observations.

The foregoing results show that our model-robust uncertainty estimates incorporate the appropriate sources of variability for conditional and population inference. Consistent with this, we will see in simulation examples in Section 4 that they lead to asymptotically correct intervals, even when the true model is heteroscedastic and/or nonlinear.

3 Continuous Covariate

We turn now to the case of a continuous covariate space. The situation is different because we cannot expect there to be multiple realizations of each covariate value in the sampled set. The problem of estimating $\phi(\cdot)$ and $\sigma^2(\cdot)$ as unconstrained functions is unidentifiable. However, in applied regression settings it is almost always reasonable to assume that these are sufficiently regular to be approximated by semi-parametric smoothing methods. This is a very weak assumption compared to assuming linearity and/or homoscedasticity.

Any minimally informative smoothing prior for $\phi(\cdot)$ and $\log \sigma(\cdot)$ would be an appropriate analogue to the priors from the discrete covariate case. A review of some relevant methods for univariate and multivariate problems can be found in [33]. We describe a particular choice of spline prior that we implement in our examples and leave the general issue of choosing optimal smoothing priors for future work.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

We restrict to scalar x in a model with an intercept and approximate $\phi(x)$ and $\log \sigma(x)$ with penalized O'Sullivan splines using a method based on [34], extended to allow for heteroscedasticity. The basic idea is that we pick Q knots spread uniformly over the potential range of x and set

$$\begin{aligned} \phi(x;u) &= \alpha_0 + \alpha_1 x + \sum_{q=1}^{Q} u_q B_q(x) \\ \log \sigma(x;v) &= \gamma_0 + \gamma_1 x + \sum_{q=1}^{Q} v_q B_q(x), \end{aligned}$$

where the $B_q(x)$ are B-spline basis functions defined by the knot locations, with independent priors $\alpha_i \sim N(0, 10^6)$, $\gamma_i \sim N(0, 10^6)$. The specification of priors for u and v involves some transformations and amounts to the following. Define the matrix Z to incorporate an appropriate penalty term as in Section 4 of [34] and let

$$\phi(x_i;a) = \alpha_0 + \alpha_1 x + \sum_{q=1}^{Q} a_q Z_{iq}$$

$$\log \sigma(x_i;b) = \gamma_0 + \gamma_1 x + \sum_{q=1}^{Q} b_q Z_{iq}$$

with independent priors $a_q \sim N(0, \sigma_a^2)$ and $b_q \sim N(0, \sigma_b^2)$ and hyperparameters distributed as $(\sigma_a^2)^{-1} \sim \text{Gamma}(0.1, 0.1)$ and $(\sigma_b^2)^{-1} \sim \text{Gamma}(0.1, 0.1)$. It is straightforward to simulate from the posterior distributions using WinBUGS software [35] [36].

For a prior on the covariate we use the limiting case of a Dirichlet process that gives rise to the same posterior Dirichlet distribution as we had for discrete covariates [37].

4 Simulation Examples

4.1 Discrete Covariate

We consider an example of a discrete covariate with $m = 3$ that can take the $K = 10$ values

$$\begin{aligned} \xi &= (1) \times (-10, -5, 0, 5, 10) \times (1, 2) \\ &= (1, -10, 1), \ldots, (1, 10, 1), (1, -10, 2), \ldots, (1, 10, 2) \end{aligned}$$

and does so in the population with probabilities

$$p(\xi) = (0.05, 0.05, 0.1, \ldots, 0.1, 0.15, 0.15).$$

We estimate the multivariate parameter $(\beta_0, \beta_1, \beta_2)$ in our simulations, but we single out the β_1 component for discussion and abuse notation below by omitting the subscript.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper320

We consider several cases for the true distribution of y given x. We specify a linear response

$$f_{lin}(x) = 1 + 3.5x_1 + 2x_2$$

for which the true conditional and population least squares fits are

$$\beta_{lin,cond} = \beta_{lin,pop} = 3.50.$$

We also specify a nonlinear response relationship

$$f_{nonlin}(x) = 1 + 3.5x_1 + 2x_2 + x_1^2 + x_2^2$$

for which the true population least squares fit has the known value

$$\beta_{nonlin,pop} = 4.303,$$

whereas the conditional least squares fit $\beta_{nonlin,cond}$ varies depending on the sampled set of covariate values. We also consider an equal variance model $\sigma^2_{equal} = 1$ and an unequal variance model $\sigma^2_{unequal} = (1+x_1^2/25)^2$. Factorial combination gives four possible true models. For each of the four models we generate 1000 random realizations of X and Y with $n = 100, 200, 400$.

First assuming we are interested in conditional inference, we calculate our robust Bayesian conditional point estimate $\hat{\beta}^{\dagger}_{cond}$ and 95% credible intervals based on $\hat{\sigma}^{\dagger}_{cond}$. In Table 1 we compare frequentist mean bias, average interval width, and coverage to the model-based and Huber-White sandwich-based intervals. In the case of a linear response, the Huber-White and robust Bayesian intervals provide 95% coverage for equal and unequal variance, and the model-based intervals are correct for the equal variance situation but are too narrow in the case of unequal variance. Although we expect the asymptotic coverage probabilities to be exactly 95%, the observed values deviate slightly due to finite sample sizes and Monte Carlo error in the simulations. When the underlying response is nonlinear, the model-based and Huber-White intervals are conservative with 100% coverage because they incorporate deviation from the linear trend into the standard error estimates. This is not a problem with the robust Bayesian intervals because they base the interval widths only on local variability of the response at each value of the covariate.

In the case of population inference shown in Table 2, the situation is somewhat different. The model-based intervals are correct for the linear equal variance situation but are too narrow in the case of unequal variance or nonlinear response. We have shown that the robust Bayesian intervals are asymptotically equivalent to Huber-White intervals for marginal inference, and in the simulation study both provide correct 95% coverage for all underlying models. It is correct to include deviation from a linear trend in the interval widths because this correctly accounts for the fact that the line being approximated depends on the covariate values, with the degree of dependence increasing for more nonlinearity in the response.

COBRA
Collection of Biostatistics
Research Archive

4.2 Continuous Covariate

We consider a second set of examples with a single continuous covariate uniformly distributed in the interval $[-10, 10]$. As in the discrete covariate example we evaluate performance for

Hosted by The Berkeley Electronic Press

four true distributions of y given x. These are obtained by taking combinations of the linear response

$$f_{lin}(x) = 2 + 3.5x$$

and the nonlinear response

$$f_{nonlin}(x) = 2 + 3.5x + x^2,$$

as well as the equal variance model $\sigma^2_{equal} = 1$ and unequal variance model $\sigma^2_{unequal} = (1 + x^2/25)^2$. For each of the four models we generate 200 random realizations of X and Y with $n = 100, 200, 400$. Results are given in Table 3 for conditional inference and Table 4 for population inference.

The conclusions are identical to the discrete covariate situation. Model-based intervals fail by being either conservative or anti-conservative in all situation except for a linear response with equal variance. Our Bayesian robust intervals give asymptotically correct coverage for all cases, and for population inference they are equivalent to the Huber-White sandwich estimator. For conditional inference, the Huber-White sandwich estimator is conservative if the true mean model is nonlinear.

5 Discussion

The main contribution of this paper is a simple Bayesian framework for linear regression that recovers the least squares solution with uncertainty estimates that correctly account for heteroscedasticity and nonlinearity. Our model-robust intervals can be constructed for conditional or population inference and have good frequentist coverage properties in both situations. The population inference estimates are equivalent to the Huber-White sandwich estimator. It bear emphasis that in establishing this equivalence, we have decomposed the sandwich estimator into pieces that separately account for random variability of y given x and nonlinearity in the true mean response; see equations (12), (13), and (15).

We have used a Dirichlet prior for the covariate space in population inference so that our estimate of the covariate distribution is essentially the sampling distribution for the observed values. There is a connection with the Bayesian bootstrap [32], but we apply resampling only to the covariates and not to either the dependent variable or the estimated residuals. It is known that bootstrap procedures and the sandwich estimator are related and are asymptotically equivalent for a broad class of regression problems [38][39]. Thus, in some sense our results are not surprising, but they reflect a fundamentally different approach to modeling and are not subsumed in bootstrap ideas. The appropriate role for bootstrap methods has been the subject of much discussion; in a Bayesian context, see for example reference [40].

In the continuous covariate case, we use splines to approximate the mean and variance functions. This is necessary because the mean and variance are not separately identifiable from a single sample at each covariate value. This can be regarded as a weakness in our approach, but it also suggests an opportunity to improve on the small-sample performance of the sandwich estimator by incorporating additional prior information. Our spline model can work in any situation where the true mean and variance are smooth functions of the covariates. This is a very reasonable assumption for applied problems. In fact, the implicit

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

assumption of the sandwich estimator that the variance function has no structure whatsoever seems overly permissive. By using properly calibrated splines or other semi-parametric priors, it should be possible to improve upon the small-sample performance by borrowing information from nearby covariate values. This approach appears particularly promising in the context of generalized estimating equations, where there may be many samples but too few clusters to accurately estimate a completely unstructured covariance matrix.

A Proofs of theorems

We begin with some observations and notation for the posterior of ϕ and the predictive distribution y^*. Results for the posterior variance of ϕ are conditional on $n_k \geq 4$ for all k. Conditioning on the hyperparameters we have

$$y|(x = \xi_k, \phi_k, \sigma_k^2) \sim N\left(\phi_k, \sigma_k^2\right).$$

It is known [41] that the posterior ϕ can be decomposed into its deterministic and random components

$$\phi = \bar{y} + \varepsilon \tag{8}$$

such that

$$\bar{y}_k = \bar{y}(\xi_k) = \frac{1}{n_k} \sum_{l:X_l=\xi_k} Y_l,$$

and the $\varepsilon_k = \varepsilon(\xi_k)$ are independent zero mean t-distributed random variables with $n_k - 1$ degrees of freedom and variances

$$\mathrm{Var}(\varepsilon_k) = \frac{1}{n_k(n_k-3)} \sum_{l:X_l=\xi_k} (Y_l - \bar{y}_k)^2. \tag{9}$$

We let Φ be the n-vector defined by $\Phi_i = \phi(X_i)$, with posterior mean

$$E\Phi = \overline{Y} = \left(\bar{y}(X_1), \ldots, \bar{y}(X_n)\right). \tag{10}$$

For the posterior predictive process, we let y^* be the K-vector with entries $y_k^* = y^*(\xi_k)$ and then we can write

$$y_k^*|\phi_k, \sigma_k^2 = \phi_k + \eta_k, \quad \eta_k|\sigma_k^2 \sim N(0, \sigma_k^2) \tag{11}$$

with the η_k independent of each other.

Denote $E_{x|\lambda}$ as expectation with respect to the Dirichlet measure $P(x|\lambda)$, and $\mathbb{E}_x$ as expectation with respect to the empirical measure $\mathbb{P}(x)$.

Proof of Theorem 1. It follows from equations (2) and (11) that

$$\begin{aligned}
\hat{\beta}^*_{cond} &= \underset{\beta}{\operatorname{argmin}}\, E_{y^*}\left(\mathbb{E}_x\left[(y^*(x) - x\beta)^2\right]\right) \\
&= \mathbb{E}_x\left[x^t x\right]^{-1} E_{y^*}\left(\mathbb{E}_x\left[x^t y^*(x)\right]\right) \\
&= \left(X^t X\right)^{-1} X^t \overline{Y} \\
&= \left(X^t X\right)^{-1} X^t Y.
\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

We exploit the repeated structures of X and $\overline{Y}$ to obtain the fourth line. For the population inference version

$$\begin{aligned}\hat{\beta}^*_{pop} &= E_\lambda \left(E_{x|\lambda}\left[x^t x\right]\right)^{-1} E_{y^*,\lambda}\left(E_{x|\lambda}\left[x^t y^*(x)\right]\right) \\ &= E_\lambda[\xi^t \lambda \xi]^{-1} E_{y^*,\lambda}\left(\xi \lambda y^*\right) \\ &= \left(X^t X\right)^{-1} X^t Y.\end{aligned}$$

In the last line, we use equation (11) to get the expected value of y^*, and we exploit the property of the posterior Dirichlet distribution $E\lambda_k = n_k/n$. ■

Proof of Theorem 2. Similarly to the first set of calculations in the proof of Theorem 1, we use the definition of $\theta^\dagger_{cond}$ in equation (4) to get

$$\theta^\dagger_{cond} = (X^t X)^{-1} X^t \Phi.$$

We use the expected value of Φ from equation (10) and the repeated structures of X and $\overline{Y}$ to obtain

$$E\theta^\dagger_{cond} = \left(X^t X\right)^{-1} X^t Y.$$

This establishes the first equality in the theorem. The second equality follows by using equation (9) to calculate the variance of $\theta^\dagger_{cond}$ and rearranging terms so the covariance matrix in the sandwich is diagonal. ■

For the asymptotic results in Theorem 3 we need the following lemma, which is a version of the exchangeable central limit theorem. An equivalent formulation of the Dirichlet posterior weights has a vector $\tilde{\lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n)$ corresponding to probabilities of resampling each of the observed $(x_1, \ldots, x_n)$.

Lemma 1 *Let $\{a_{nj}\}$ be a bounded triangular array of constants such that*

$$\frac{1}{n}\sum_{j=1}^{n}(a_{nj} - \bar{a}_n)^2 \to \sigma^2,$$

where $\bar{a}_n = \frac{1}{n}\sum_{j=1}^{n} a_{nj}$. Then

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(a_{nj}\tilde{\lambda}_j - \bar{a}_n\right) \xrightarrow{d} N(0, \sigma^2).$$

Proof. The result is a special case of Lemma 4.6 in [42]. ■

Proof of Theorem 3. We condition on an infinite sequence of observations of x and y and in everything that follows implicitly index by n. By the law of large numbers, the $\bar{y}(x)$ are

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

uniformly bounded for all n. We begin by analyzing $\hat{\beta}^{\dagger}_{pop}$,

$$\begin{aligned}
\hat{\beta}^{\dagger}_{pop} &= E_{\phi,\lambda}\left\{E_{x|\lambda}[x^t x]^{-1} E_{x|\phi,\lambda}[x^t \phi(x)]\right\} \\
&= E_{\lambda}\left\{E_{x|\lambda}[x^t x]^{-1} E_{x|\lambda}[x^t \bar{y}(x)]\right\} + E_{\varepsilon,\lambda}\left\{E_{x|\lambda}[x^t x]^{-1} E_{x|\varepsilon,\lambda}[x^t \varepsilon(x)]\right\} \\
&= E_{\lambda}\left\{E_{x|\lambda}[x^t x]^{-1} E_{x|\lambda}[x^t \bar{y}(x)]\right\}.
\end{aligned}$$

The last line uses the fact that $\varepsilon(x)$ has zero mean. Convergence of $\hat{\beta}^{\dagger}_{pop} - (X^t X)^{-1} X^t Y$ to zero follows from the mean values of the Dirichlet weights and the continuous mapping theorem, since $\bar{y}(x)$ is uniformly bounded.

To calculate the variance of $\theta^{\dagger}_{pop}$, we note that

$$\begin{aligned}
\mathrm{Var}(\theta^{\dagger}_{pop}) &= \mathrm{Var}_{\phi,\lambda}\left\{E_{x|\lambda}[x^t x]^{-1} E_{x|\phi,\lambda}[x^t \phi(x)]\right\} \\
&= \mathrm{Var}_{\varepsilon,\lambda}\left\{E_{x|\lambda}[x^t x]^{-1} E_{x|\varepsilon,\lambda}[x^t\left(\bar{y}(x) + \varepsilon(x)\right)]\right\} \qquad (12) \\
&= \mathrm{Var}_{\lambda}\left\{E_{x|\lambda}[x^t x]^{-1} E_{x|\lambda}[x^t \bar{y}(x)]\right\} + \mathrm{Var}_{\varepsilon,\lambda}\left\{E_{x|\lambda}[x^t x]^{-1} E_{x|\varepsilon,\lambda}[x^t \varepsilon(x)]\right\}.
\end{aligned}$$

The first two lines follow from equations (6) and (8). To verify the third line, note that the terms involving $\bar{y}$ and ε are uncorrelated since conditional on λ, $\bar{y}$ is deterministic and ε has mean zero.

We calculate sandwich forms for the two variances on the right hand side of (12) and complete the proof by comparing the sum of the respective covariance matrices to Σ. First we show that

$$\mathrm{Var}_{\lambda}\left[E_{x|\lambda}[x^t x]^{-1} E_{x|\lambda}[x^t \bar{y}(x)]\right] - (X^t X)^{-1}\left(X^t \tilde{\Sigma} X\right)(X^t X)^{-1} = o(n^{-1}) \qquad (13)$$

with $\tilde{\Sigma}$ defined by

$$\tilde{\Sigma}_{ij} = \begin{cases} \left(\overline{Y}_i - X_i (X^t X)^{-1} X^t \overline{Y}\right)^2 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

This is equivalent to showing that

$$\mathrm{Var}_{\lambda}\left[E_{x|\lambda}[x^t x]^{-1} E_{x|\lambda}[x^t \psi(x)]\right] - (X^t X)^{-1}\left(X^t \Sigma' X\right)(X^t X)^{-1} = o(n^{-1}) \qquad (14)$$

with Σ' defined by

$$\Sigma'_{ij} = \begin{cases} \left(\Psi_i - X_i (X^t X)^{-1} X^t \Psi\right)^2 & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases}$$

where $\psi(x)$ is the true mean of y conditional on x, and Ψ is the n-vector defined by $\Psi_i = \psi(X_i)$. The equivalence follows from the law of large numbers because $\tilde{\Sigma}$ and Σ'

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

are asymptotically the same, and the first terms in (13) and (14) can be seen to have the same limit by applying Lemma 1 and the bootstrap delta method [38] to each. The same delta method argument establishes that the first term in (14) is asymptotically equivalent to the sampling variance for the linear regression problem with fixed response $\psi(x)$. The second term in (14) is just the Huber-White sandwich estimator for that problem. Since the sandwich estimator is aymptotically consistent for the sampling variance, equation (14) follows.

Next we note that since $\varepsilon(x)$ has mean zero, the second term on the right hand side of (12) can be written

$$\mathrm{Var}_{\varepsilon,\lambda}\left[E_{x|\lambda}[x^t x]^{-1} E_{x|\varepsilon,\lambda}[x^t \varepsilon(x)]\right] = E_\lambda\left\{\mathrm{Var}_\varepsilon\left(E_{x|\lambda}[x^t x]^{-1} E_{x|\varepsilon,\lambda}[x^t \varepsilon(x)]\right)\right\}.$$

It follows from equation (9), moment properties of the posterior Dirichlet weights given in [32], and the continuous mapping theorem that

$$E_\lambda\left\{\mathrm{Var}_\varepsilon\left(E_{x|\lambda}[x^t x]^{-1} E_{x|\varepsilon,\lambda}[x^t \varepsilon(x)]\right)\right\} - (X^t X)^{-1}\left(X^t \Sigma^\dagger X\right)(X^t X)^{-1} = o(n^{-1}), \qquad (15)$$

where $\Sigma^\dagger$ is the diagonal matrix defined previously by

$$\Sigma^\dagger_{ij} = \begin{cases} \dfrac{1}{n_k - 3} \displaystyle\sum_{l:X_l=\xi_k} (Y_l - \bar{y}_k)^2 & \text{if } i = j \text{ and } X_i = \xi_k \\ \\ 0 & \text{if } i \neq j. \end{cases}$$

To finish the proof, we use that by elementary calculations

$$\sum_{i:X_i=\xi_k} \Sigma_{ii} = \sum_{i:X_i=\xi_k} \left(\tilde{\Sigma}_{ii} + \Sigma^\dagger_{ii}\right)$$

holds for each $k = 1, \ldots, K$, up to degrees of freedom corrections. ■

References

[1] C. F. Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum.* 1809.

[2] A. M. Legendre, *Nouvelles Methodes Pour la Determination des Orbites des Cometes.* 1805.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis, 2nd Edition.* Chapman & Hall/CRC, 2004.

[4] T. J. Sweeting, "Coverage probability bias, objective Bayes and the likelihood principle," *Biometrika*, vol. 88, no. 3, pp. 657–675, 2001.

[5] R. E. Kass, "Kinds of Bayesians (comment on articles by Berger and Goldstein)," *Bayesian Analysis*, vol. 1, no. 3, pp. 437–440, 2006.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

[6] L. Wasserman, "Frequentist Bayes is objective (comment on articles by Berger and Goldstein)," *Bayesian Analysis*, vol. 1, no. 3, pp. 451–456, 2006.

[7] J. O. Berger, V. D. Oliveira, and B. Sanso, "Objective Bayesian analysis of spatially correlated data," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1361–1374, 2001.

[8] J. A. Jacquez, F. J. Mather, and C. R. Crawford, "Linear regression with non-constant, unknown error variances: Sampling experiments with least squares, weighted least squares and maximum likelihood estimators," *Biometrics*, vol. 24, no. 3, pp. 607–626, 1968.

[9] D. S. Leslie, R. Kohn, and D. J. Nott, "A general approach to heteroscedastic linear regression," *Statistics and Computing (to appear).*

[10] W. J. Boscardin and A. Gelman, "Bayesian regression with parametric models for uncertainty," *Advances in Econometrics*, vol. 11A, pp. 87–109, 1996.

[11] J. K. Ghosh and R. V. Ramamoorthi, *Bayesian Nonparametrics.* Springer, 2003.

[12] J. O. Berger, "The robust Bayesian viewpoint," in *Robustness of Bayesian Analysis*, pp. 63–124, 1984.

[13] T.-H. Fan and J. O. Berger, "Robust Bayesian displays for standard inferences concerning a normal mean," *Computational Statistics and Data Analysis*, vol. 33, pp. 381–399, 2000.

[14] A. Zellner, "Bayesian method of moments (BMOM) analysis of mean and regression models," in *Modelling and Prediction* (J. C. Lee, ed.), pp. 61–74, Springer, 1996.

[15] S. Geisser and T. Seidenfeld, "Remarks on the 'Bayesian' method of moments," *Journal of Applied Statistics*, vol. 26, no. 1, pp. 97–101, 1999.

[16] A. Zellner, "Remarks on the 'critique' of the Bayesian method of moments," *Journal of Applied Statistics*, vol. 28, no. 6, pp. 775–778, 2001.

[17] H. D. Brunk, "Bayesian least squares estimates of univariate regression functions," *Communications in Statistics*, vol. A9, no. 11, pp. 1101–2236, 1980.

[18] E. P. Smouse, "A note on Bayesian least squares inference for finite population models," *Journal of the American Statistical Association*, vol. 79, no. 386, pp. 390–392, 1984.

[19] A. E. Gelfand and S. K. Ghosh, "Model choice: A minimum posterior predictive loss approach," *Biometrika*, vol. 85, no. 1, pp. 1–11, 1998.

[20] J. Aldrich, "Fisher and regression," *Statistical Science*, vol. 20, no. 4, pp. 401–417, 2005.

[21] S. E. Fienberg, "A brief history of statistics in three and one-half chapters: A review essay," *Statistical Science*, vol. 7, no. 2, pp. 208–225, 1992.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

[22] R. Tibshirani, “Discussion: Jackknife, bootstrap and other resampling methods in regression analysis,” *Annals of Statistics*, vol. 14, no. 4, pp. 1335–1339, 1986.

[23] C. F. J. Wu, “Rejoinder: Jackknife, bootstrap and other resampling methods in regression analysis,” *Annals of Statistics*, vol. 14, no. 4, pp. 1343–1350, 1986.

[24] A. Scott and C. Wild, “On the robustness of weighted methods for fitting models to case-control data,” *Journal of the Royal Statistical Society B*, vol. 64, no. 2, pp. 207–219, 2002.

[25] N. E. Breslow, “Statistics in epidemiology: the case-control study,” *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 14–28, 1996.

[26] P. J. Huber, “The behavior of maximum likelihood estimates under nonstandard conditions,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability*, vol. 1, pp. 221–233, 1967.

[27] H. White, “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity,” *Econometrica*, vol. 48, no. 4, pp. 817–838, 1980.

[28] D. A. Freedman, “On the so-called ’huber sandwich estimator’ and ’robust standard errors’,” *The American Statistician*, vol. 60, no. 4, pp. 299–302, 2006.

[29] R. M. Royall, “Model robust confidence intervals using maximum likelihood estimators,” *International Statistical Review*, vol. 54, no. 2, pp. 221–226, 1986.

[30] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.* Chapman and Hall/CRC, 2006.

[31] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications.* University of Cambridge Press, 1997.

[32] D. B. Rubin, “The bayesian bootstrap,” *Annals of Statistics*, vol. 9, no. 1, pp. 130–134, 1981.

[33] P. Muller and F. A. Quintana, “Nonparametric Bayesian data analysis,” *Statistical Science*, vol. 19, no. 1, pp. 95–110, 2004.

[34] M. P. Wand and J. T. Ormerod, “On semiparametric regression with O’Sullivan penalised splines,” *Australian and New Zealand Journal of Statistics (to appear).*

[35] D. J. Lunn, A. Thomas, and N. Best, “Winbugs – a Bayesian modelling framework: concepts, structure, and extensibility,” *Statistics and Computing*, vol. 10, pp. 325–337, 2000.

[36] C. Cariniceanu, D. Ruppert, and M. P. Wand, “Bayesian analysis for penalized spline regression using winbugs,” *Journal of Statistical Software*, vol. 14, no. 14, pp. 1–24, 2005.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

[37] M. Gasparini, "Exact multivariate bayesian bootstrap distributions of moments," *Annals of Statistics*, vol. 23, no. 3, pp. 762–768, 1995.

[38] A. W. van der Vaart, *Asymptotic Statistics.* University of Cambridge Press, 1998.

[39] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes.* Springer, 1996.

[40] M. J. Schervish, "Comment on: Bootstrap: More than a stab in the dark?," *Statistical Science*, vol. 9, no. 3, pp. 408–410, 1994.

[41] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis.* Wiley, 1992.

[42] J. Praestgaard and J. A. Wellner, "Exchangeably weighted bootstraps of the general empirical process," *Annals of Probability*, vol. 21, no. 4, pp. 2053–2086, 1993.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

			n = 100			n = 200			n = 400		
			Bias	Width	Coverage	Bias	Width	Coverage	Bias	Width	Coverage
Linear ($\beta_{true} = 3.5$)	Equal variance	Model Based	0.000	0.057	0.949	0.000	0.040	0.950	0.000	0.028	9.949
		Huber-White	0.000	0.056	0.948	0.000	0.040	0.941	0.000	0.028	0.948
		Bayes Robust	0.000	0.068	0.981	0.000	0.043	0.960	0.000	0.029	0.951
	Unequal variance	Model Based	−0.002	0.194	0.885	0.000	0.138	0.855	0.000	0.098	0.866
		Huber-White	−0.002	0.250	0.954	0.000	0.180	0.945	0.000	0.128	0.948
		Bayes Robust	−0.002	0.306	0.979	0.000	0.196	0.959	0.000	0.133	0.955
Nonlinear (β_{true} variable)	Equal variance	Model Based	−0.001	2.36	1.000	0.000	1.670	1.000	0.000	1.179	1.000
		Huber-White	−0.001	2.68	1.000	0.000	1.922	1.000	0.000	1.356	1.000
		Bayes Robust	−0.001	0.068	0.978	0.000	0.043	0.957	0.000	0.029	0.958
	Unequal variance	Model Based	0.006	2.367	1.000	0.000	1.68	1.000	−0.001	1.183	1.000
		Huber-White	0.006	2.683	1.000	0.000	1.933	1.000	−0.001	1.362	1.000
		Bayes Robust	0.006	0.304	0.981	0.000	0.195	0.954	−0.001	0.133	0.949

Table 1: Frequentist Properties of Conditional Estimates for Discrete Covariate

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

			n = 100			n = 200			n = 400		
			Bias	Width	Coverage	Bias	Width	Coverage	Bias	Width	Coverage
Linear ($\beta_{true} = 3.50$)	Equal variance	Model Based	0.000	0.057	0.949	0.000	0.040	0.950	0.000	0.028	0.949
		Huber-White	0.000	0.056	0.948	0.000	0.040	0.941	0.000	0.028	0.948
		Bayes Robust	0.000	0.072	0.984	0.000	0.045	0.967	0.000	0.030	0.952
	Unequal variance	Model Based	−0.002	0.194	0.885	0.000	0.138	0.855	0.000	0.098	0.866
		Huber-White	−0.002	0.250	0.954	0.000	0.180	0.945	0.000	0.128	0.948
		Bayes Robust	−0.002	0.320	0.983	0.000	0.200	0.963	0.000	0.135	0.959
Nonlinear ($\beta_{true} = 4.303$)	Equal variance	Model Based	−0.113	2.362	0.944	0.022	1.670	0.907	−0.005	1.179	0.914
		Huber-White	−0.113	2.681	0.964	0.022	1.922	0.956	−0.005	1.356	0.955
		Bayes Robust	−0.090	2.668	0.963	0.035	1.918	0.950	0.001	1.355	0.953
	Unequal variance	Model Based	−0.138	2.367	0.936	0.012	1.676	0.893	−0.010	1.183	0.918
		Huber-White	−0.138	2.683	0.967	0.012	1.934	0.942	−0.010	1.362	0.949
		Bayes Robust	−0.115	2.680	0.963	0.025	1.932	0.940	−0.004	1.361	0.946

Table 2: Frequentist Properties of Population Estimates for Discrete Covariate

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

			n = 100			n = 200			n = 400		
			Bias	Width	Coverage	Bias	Width	Coverage	Bias	Width	Coverage
Linear ($\beta_{true} = 3.50$)	Equal variance	Model Based	0.000	0.018	0.925	0.000	0.012	0.945	0.000	0.009	0.930
		Huber-White	0.000	0.017	0.915	0.000	0.012	0.935	0.000	0.009	0.940
		Bayes Robust	0.000	0.018	0.935	0.000	0.012	0.945	0.000	0.009	0.935
	Unequal variance	Model Based	0.000	0.046	0.830	-0.001	0.032	0.815	0.002	0.023	0.855
		Huber-White	0.000	0.061	0.920	-0.001	0.043	0.935	0.002	0.031	0.930
		Bayes Robust	0.000	0.060	0.920	-0.001	0.042	0.920	0.002	0.030	0.925
Nonlinear (β_{true} variable)	Equal variance	Model Based	0.000	0.515	1.000	0.000	0.365	1.000	0.000	0.258	1.000
		Huber-White	0.000	0.639	1.000	0.000	0.456	1.000	0.000	0.322	1.000
		Bayes Robust	0.000	0.030	0.940	0.000	0.015	0.945	0.000	0.009	0.940
	Unequal variance	Model Based	0.000	0.517	1.000	-0.001	0.367	1.000	0.002	0.259	1.000
		Huber-White	0.000	0.642	1.000	-0.001	0.458	1.000	0.002	0.324	1.000
		Bayes Robust	0.000	0.069	0.950	-0.001	0.046	0.940	0.002	0.031	0.935

Table 3: Frequentist Properties of Conditional Estimates for Continuous Covariate

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

			n = 100			n = 200			n = 400		
			Bias	Width	Coverage	Bias	Width	Coverage	Bias	Width	Coverage
Linear ($\beta_{true} = 3.50$)	Equal variance	Model Based	0.000	0.018	0.925	0.000	0.012	0.945	0.000	0.009	0.930
		Huber-White	0.000	0.017	0.915	0.000	0.012	0.935	0.000	0.009	0.940
		Bayes Robust	0.000	0.019	0.940	0.000	0.013	0.945	0.000	0.009	0.945
	Unequal variance	Model Based	0.000	0.046	0.830	−0.001	0.032	0.815	0.002	0.023	0.855
		Huber-White	0.000	0.061	0.920	−0.001	0.043	0.935	0.002	0.031	0.930
		Bayes Robust	0.002	0.061	0.935	0.001	0.043	0.930	0.002	0.030	0.945
Nonlinear ($\beta_{true} = 3.50$)	Equal variance	Model Based	−0.094	0.515	0.875	−0.076	0.365	0.850	−0.044	0.258	0.845
		Huber-White	−0.094	0.639	0.935	−0.076	0.456	0.945	−0.044	0.322	0.935
		Bayes Robust	−0.109	0.634	0.920	−0.080	0.454	0.945	−0.043	0.321	0.935
	Unequal variance	Model Based	−0.095	0.517	0.880	−0.076	0.367	0.840	−0.043	0.259	0.855
		Huber-White	−0.095	0.642	0.935	−0.076	0.458	0.940	−0.043	0.324	0.930
		Bayes Robust	−0.091	0.633	0.925	−0.076	0.454	0.945	−0.041	0.322	0.930

Table 4: Frequentist Properties of Population Estimates for Continuous Covariate