UW Biostatistics Working Paper Series

5-20-2010

Model-Robust Regression and a Bayesian `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 Regression and a Bayesian `Sandwich' Estimator" (May 2010). *UW Biostatistics Working Paper Series*. Working Paper 338.
http://biostats.bepress.com/uwbiostat/paper338

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

The Annals of Applied Statistics
2010, Vol. 4, No. 4, 2099–2113
DOI: 10.1214/10-AOAS362
© Institute of Mathematical Statistics, 2010

MODEL-ROBUST REGRESSION AND A BAYESIAN "SANDWICH" ESTIMATOR

BY ADAM A. SZPIRO, KENNETH M. RICE AND THOMAS LUMLEY

University of Washington, University of Washington and University of Washington

We present a new Bayesian approach to model-robust linear regression that leads to uncertainty estimates with the same robustness properties as the Huber–White sandwich estimator. The sandwich estimator is known to provide asymptotically correct frequentist inference, even when standard modeling assumptions such as linearity and homoscedasticity in the data-generating mechanism are violated. Our derivation provides a compelling Bayesian justification for using this simple and popular tool, and it also clarifies what is being estimated when the data-generating mechanism is not linear. We demonstrate the applicability of our approach using a simulation study and health care cost data from an evaluation of the Washington State Basic Health Plan.

1. Introduction. The classical theory of uncorrelated linear regression is based on three modeling assumptions: (i) the outcome variable is linearly related to the covariates on average, (ii) random variations from the linear trend are homoscedastic, and (iii) random variations from the linear trend are Normally distributed. Under these assumptions, classical frequentist methods give point estimates and exact probability statements for the sampling distribution of these estimates. Equivalent uncertainty estimates are derived in the Bayesian paradigm, but are stated in terms of the posterior distribution for the unknown slope parameter in the assumed linear model. However, in a typical application none of these modeling assumptions can reasonably be expected to hold in the data-generating mechanism.

We study the relationship between age and average annual outpatient health care costs using data from the evaluation of the Washington State Basic Health Plan. The plan provided subsidized health insurance for low income residents starting in 1989, and the evaluation study included 6918 subjects followed for an average of 22 months (range 1 to 44 months) [Diehr et al. (1993)]. Previous analysis of this data set has shown that the variability is heteroscedastic and not Normally distributed [Lumley et al. (2002)], and it appears from Figure 1 that the relationship deviates from linearity. We are still motivated to estimate a "linear trend" since this appears to be a dominant feature of the data, and while we could consider a transformation to stabilize the variance, this may not be desirable since the primary policy interest is in total or mean dollars not in log-dollars [Diehr et al. (1999)]. We also consider simulated data sets with similar features as illustrated in Figure 2.

Received April 2009; revised April 2010.

Key words and phrases. Bayesian inference, estimating equations, linear regression, robust regression, sandwich estimator.

COBRA
COLLECTION OF BIOSTATISTICS
Research Archive

Hosted by The Berkeley Electronic Press

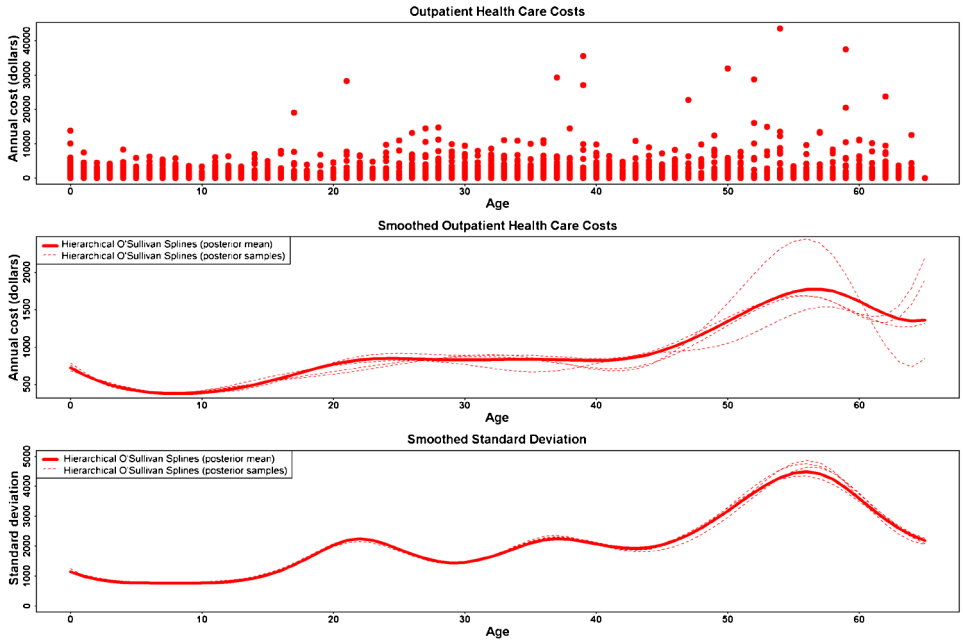

FIG. 1. *Outpatient health care costs from the evaluation of Washington States Basic Health Plan. Top panel: Average annual costs for* 6918 *subjects enrolled in the study. Middle panel: Semi-parametric smoothing estimates of the average annual cost vs. age, fit with Bayesian O'Sullivan splines. Bottom panel: Semi-parametric smoothing estimate of the standard deviation of annual health care costs vs. age, fit with Bayesian O'Sullivan splines. In each of the bottom two panels, the thick red line is the posterior mean of the spline fit, and the thin dashed red lines are example draws from the posterior distribution.*

For the classical theory of linear regression to hold, the Normality assumption is only necessary if we want to derive exact sampling probabilities for the point estimates. In the large sample limit, the central limit theorem alone guarantees that the sampling distribution is asymptotically Normal and that the classical standard error estimates are correct. The linearity and homoscedasticity assumptions, however, are a different matter. If either of these is violated in the data-generating mechanism, then classical standard error estimates are incorrect, even asymptotically. Furthermore, without the assumption of linearity, it is not immediately clear what quantity we are trying to estimate.

A modern frequentist approach to analyzing data that do not conform to classical assumptions is to directly state what we want to know about moments of the data-generating mechanism by way of estimating equations, without making any assumptions about validity of an underlying model. The associated "robust" or "sandwich"-based standard errors provide accurate large sample inference, at no more computational effort than fitting a linear model [Huber (1967); White (1980); Liang and Zeger (1986); Royall (1986)]. The Huber–White sandwich estimator is

Research Archive

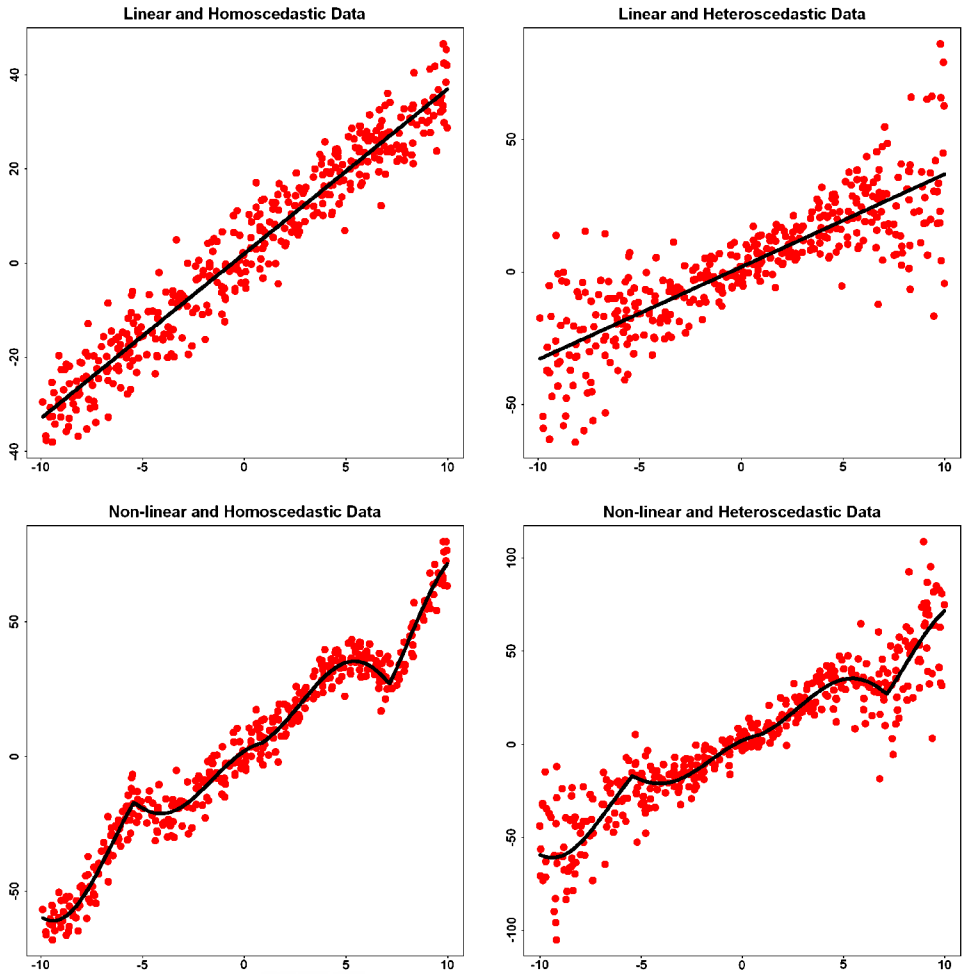

FIG. 2. *Example scatterplots (red dots) and mean functions (black lines) from the four simulation scenarios considered in Section 5 with $n = 400$. The four scatterplots correspond to all possible combination of the linear and nonlinear mean functions and homoscedastic and heteroscedastic variance functions defined in Section 5.*

easy to implement with standard software and is widely used in biostatistics. As long as the data have a dominant linear structure, this strategy provides relevant inference for the linear trend and does not depend on detailed modeling of the variance or mean structures.

Finding a Bayesian analogue of estimating equations and the sandwich estimator has been an open problem for some time. In this paper we describe a novel Bayesian framework for linear regression that assumes neither linearity nor homoscedasticity. Even in the absence of a slope parameter, we give a natural definition for the "linear trend" quantity to be estimated and how to measure its uncer-

Research Archive

tainty. We show that in the random covariate setting our Bayesian robust posterior standard deviations are asymptotically equivalent to the commonly used sandwich estimator. Furthermore, with fixed covariates our Bayesian robust uncertainty estimates exhibit better frequentist sampling properties than the sandwich estimator, when the true data-generating mechanism is nonlinear in the covariates.

In Section 2 we set out our notation and define the model-robust Bayesian regression paradigm. In Section 3 we derive our main theoretical results for the case of randomly sampled covariates from a discrete space. In Section 4 we consider extensions to continuous covariates and to a fixed design matrix. We demonstrate the properties of our methodology in a simulation study in Section 5, and in Section 6 we apply it to the annual health care cost data described above. We conclude in Section 7 with a discussion.

2. Notation and definitions.

2.1. *Target of inference*. We consider the familiar situation for multivariate linear regression of having observed an n-vector of outcomes Y and an $n \times m$ matrix of covariate values X, with the stated objective of estimating the "linear relationship" between X and Y. Before determining operationally how to do this, we take care to clarify the quantity of interest in terms of a true (but unknown) data-generating mechanism, without assuming that there is an underlying linear relationship.

We assume that X represents n independent identically distributed observations in $\mathbb{R}^m$ of the m-dimensional covariate random variable x, and that Y represents n corresponding independent observations of the real-valued outcome random variable y. We think of the probability distribution for x as representing the frequency of different covariate values in the population to which we wish to generalize, and the distribution of y conditional on x as the distribution of the outcome for individuals with covariate values x. Suppose that the true joint distribution for x and y admits a density function $\lambda(\cdot)$ for x (with respect to the Lebesgue measure on $\mathbb{R}^m$) such that for any measurable set A

$$P(x \in A) = \int_A \lambda(v)\, dv \tag{1}$$

and a measurable function $\phi(\cdot)$ on $\mathbb{R}^m$ for the mean of y conditional on x such that

$$E(y|x = v) = \phi(v). \tag{2}$$

Throughout, we use v as a dummy variable for x.

Heuristically, we can say that we are interested in the "linear relationship" between x and the true conditional mean of y. If $\phi(\cdot)$ were known to be linear, we would simply be interested in its coefficients. Since we are not assuming that the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper338

true mean function is linear, one possible approach is to define the quantity of interest as the m-vector of minimizing coefficients from the least-squares linear fit

$$\beta = \underset{\alpha}{\operatorname{argmin}} \int \big(\phi(v) - v\alpha\big)^2 \lambda(v)\, dv. \tag{3}$$

We can describe β as the set of m coefficients that minimizes the average squared error over the entire population in approximating the mean value of y by a linear function of x.

The definition of β is essentially a statement about the scientific question of interest, and it is not concerned with the details of random sampling of the observations. We have identified the deterministic function $\phi(\cdot)$ as representing the mean dependence of y on x, and our objective is to approximate this curve by a straight line. We define β as the best linear approximation to the curve $\phi(\cdot)$ by the method of least-squares, an idea that dates to the early work of Gauss (1809), Legendre (1805) and Jacobi (1841). Our goal is inference for β, not for the full function $\phi(\cdot)$.

Freedman (2006) has pointedly described the dangers of fitting a linear model when such a model does not hold and then deriving "robust" standard error estimates for an uninterpretable parameter. Our approach is fundamentally different in that we explicitly recognize that the data-generating mechanism may be nonlinear, and we define β as a quantity of interest that summarizes the linear feature in the data-generating mechanism (this corresponds to the standard definition of β if the data-generating mechanism is linear). While β can be defined mathematically in a very general setting, consistent with the ideas in Freedman (2006), we recommend it as a relevant target of inference only when the data suggest a dominant linear trend.

2.2. *Bayesian inference*. Since we do not know the true mean function $\phi(\cdot)$ or the true covariate density $\lambda(\cdot)$, we cannot directly calculate β from equation (3), and we need to take advantage of the observations in order to make inference about β. To do this, we embed $\phi(\cdot)$ and $\lambda(\cdot)$ in a flexible Bayesian model in such a way that we can derive posterior distributions for these functions and, thus, derive a posterior distribution for β. The key consideration in constructing the Bayesian model is that it be highly flexible, assuming neither linearity nor homoscedasticity.

We adopt the conditionally Normal model for y,

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

where we have introduced the ancillary unknown variance function $\sigma^2(\cdot)$. To complete the Bayesian model, it remains to specify a prior distribution, with probability measure $\pi(\lambda(\cdot), \phi(\cdot), \sigma^2(\cdot))$, which will be chosen to have a density that can be written

$$p(\lambda(\cdot), \phi(\cdot), \sigma^2(\cdot)) = p_\lambda(\lambda(\cdot))\, p_{\phi,\sigma^2}(\phi(\cdot), \sigma^2(\cdot)). \tag{4}$$

Collection of Biostatistics
Research Archive

We will give specific examples of priors in the remainder of this paper.

Defining priors for the discrete covariate case is relatively straightforward because we can specify a saturated model for the mean and variance functions $\phi(\cdot)$ and $\sigma^2(\cdot)$, and use a Dirichlet distribution for $\lambda(\cdot)$. We derive our main theoretical results for that setting in Section 3. Later, in Section 4, we present a simple and effective approach for extending the method to continuous covariates by using spline-based priors for $\phi(\cdot)$ and $\sigma^2(\cdot)$.

Once we have specified priors in equation (4), standard Bayesian calculus gives a posterior distribution for $\phi(\cdot)$ and $\lambda(\cdot)$,

$$\pi(\lambda(\cdot), \phi(\cdot)|X, Y),$$

and therefore a posterior distribution for the m-dimensional vector β,

$$\pi(\beta|X, Y) = \pi\left(\operatorname*{argmin}_{\alpha} \int \big(\phi(v) - v\alpha\big)^2 \lambda(v)\, dv \Big| X, Y\right). \tag{5}$$

Following common practice, we define a point estimate by taking the posterior mean of β,

$$\hat{\beta}_j = E_\pi(\beta_j|X, Y), \qquad j = 1, \dots, m, \tag{6}$$

and we use its posterior standard deviation as a measure of uncertainty

$$\hat{\sigma}_{\beta_j} = \operatorname{diag}(\operatorname{Cov}_\pi(\beta|X, Y))_j^{1/2}, \qquad j = 1, \dots, m. \tag{7}$$

We can construct approximate moment-based 95% credible intervals with the formulation

$$CI_{95_j} = \hat{\beta}_j \pm 1.96\hat{\sigma}_{\beta_j}, \qquad j = 1, \dots, m.$$

3. Discrete covariates. In this section we complete the specification of the Bayesian model for the discrete covariate case and derive our main theoretical results in that setting. Let $\xi = (\xi_1, \dots, \xi_K)$ consist of K nonzero 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 $i = 1, \dots, n$ such that $X_i = \xi_k$, where X_i is the ith row of X. We let $\lambda(\cdot)$ be a density with mass restricted to $\xi \subset \mathbb{R}^m$, written in the form

$$\lambda(\cdot) = \sum_{k=1}^{K} \lambda_k \delta_{\xi_k}(\cdot),$$

where δ_{ξ_k} is the Dirac delta function with point mass at ξ_k. That is,

$$P\big(x = \xi_k; \lambda(\cdot)\big) = \lambda_k, \qquad \sum_{k=1}^{K} \lambda_k = 1.$$

We use an improper Dirichlet prior for $\lambda(\cdot)$ such that its density can be written

$$p_\lambda(\lambda(\cdot)) \propto \prod_{k=1}^{K} \lambda_k^{-1} \qquad \left(0 \text{ if } \sum_{k=1}^{K} \lambda_k \neq 1\right).$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

The posterior distribution of $\lambda(\cdot)$ is also Dirichlet with density

$$p_{\lambda|X}(\lambda(\cdot)) \propto \prod_{k=1}^{K} \lambda_k^{-1+n_k} \qquad \left(0 \text{ if } \sum_{k=1}^{K} \lambda_k \neq 1\right).$$

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 $\lambda_k = g_k/(g_1, \ldots, g_K)$ [Davison and Hinkley (1997)]. There is also a connection between the posterior distribution for x and bootstrap resampling [Rubin (1981)].

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

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

and independent noninformative prior densities such that

$$p_{\phi,\sigma^2}(\phi(\cdot), \sigma^2(\cdot)) = \prod_{k=1}^{K} p_{\phi_k,\sigma_k^2}(\phi_k, \sigma_k^2)$$

and

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

It turns out that for $n_k \geq 4$, ϕ_k has a posterior t-distribution with easily computable mean and scale parameters. Formulas based on the data X and Y are given in the Online Supplement [Szpiro, Rice and Lumley (2010)].

Our main theoretical result is contained in the following theorem, which is proved in the Online Supplement [Szpiro, Rice and Lumley (2010)]. It states that, asymptotically, the posterior mean point estimate derived in equation (6) is the least squares fit to the data X and Y, and that the posterior standard deviation from equation (7) has the sandwich form. The term "sandwich" refers to the algebraic formation in equation (8), where colloquially the $(X^t X)^{-1}$ terms are the "bread," and $(X^t \Sigma X)$ is the "meat."

THEOREM 1. *For a discrete covariate space, assume that y conditional on x has bounded first and second moments. The m-dimensional estimate $\hat{\beta}$ defined by equation (6) takes the asymptotic form*

$$\hat{\beta} - (X^t X)^{-1} X^t Y \to 0,$$

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

$$\hat{\sigma}_\beta - \operatorname{diag}[(X^t X)^{-1}(X^t \Sigma X)(X^t X)^{-1}]^{1/2} = o(n^{-1}), \tag{8}$$

COBRA
Collection of Biostatistics
Research Archive

where Σ *is the diagonal matrix defined by*

$$\Sigma_{ij} = \begin{cases} (Y_i - X_i(X^tX)^{-1}X^tY)^2, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

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

4. Extensions.

4.1. *Continuous covariates.* We consider extending our approach to a continuous covariate space. The situation is different from discrete covariates 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, using semi-parametric smoothing methods. This is a very weak assumption compared to assuming linearity and/or homoscedasticity. 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.

We restrict to scalar x in a model with an intercept, and approximate $\phi(\cdot)$ and $\log\sigma(\cdot)$ with penalized O'Sullivan splines using a method based on Wand and Ormerod (2008), extended to allow for heteroscedasticity. We pick Q knots spread uniformly over the potential range of x and set

$$\phi(v; u) = \alpha_0 + \alpha_1 v + \sum_{q=1}^{Q} u_q B_q(v),$$

$$\log\sigma(v; w) = \gamma_0 + \gamma_1 v + \sum_{q=1}^{Q} w_q B_q(v),$$

where the $B_q(\cdot)$ 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 w involves some transformations and amounts to the following. Define the matrix Z to incorporate an appropriate penalty term as in Section 4 of Wand and Ormerod (2008) and let

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

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

with independent priors $a_q \sim N(0, \sigma_a^2)$ and $b_q \sim N(0, 0.1)$ and hyperparameter distributed as $(\sigma_a^2)^{-1} \sim \text{Gamma}(0.1, 0.1)$. It is straightforward to simulate

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

from the posterior distributions using WinBUGS software [Lunn et al. (2000); Crainiceanu et al. (2005)]. For a prior on the covariate x we use the limiting case of a Dirichlet process that gives rise to the same posterior Dirichlet distribution as we had for discrete covariates [Gasparini (1995)].

An analogous result to Theorem 1 can be expected to hold under mild regularity conditions on the true mean and standard deviation functions $\phi(\cdot)$ and $\sigma^2(\cdot)$ in the data-generating mechanism. We do not state such a result here, but we provide supporting evidence from a simulation study in Section 5.

4.2. *Fixed design matrix.* Our development up to now explicitly treats X and Y as being jointly sampled from a random population. The fact that we obtain an equivalent estimator to the sandwich form suggests that the sandwich estimator also corresponds to the random X setting. This is easily seen from equation (9) since the variance estimate Σ in the "meat" involves residuals from a linear model and is bounded away from zero if the data-generating mechanism is nonlinear, even if the observations Y are deterministic conditional on X.

A desirable feature of our approach is that it can easily be modified to explicitly treat the fixed X scenario. To do this, we simply replace the random density for X in equation (5) with a deterministic density corresponding to the actual sampled values

$$\lambda_{\text{fixed}}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(\cdot),$$

where δ_{X_i} is the Dirac delta function with point mass at X_i. Then we proceed exactly as in Section 2.2 to define the quantity of interest

$$\beta_{\text{fixed}} = \underset{\alpha}{\operatorname{argmin}} \int \big(\phi(v) - v\alpha\big)^2 \lambda_{\text{fixed}}(v)\, dv. \tag{10}$$

The point estimate for fixed X inference is

$$\hat{\beta}_{\text{fixed}} = E_\pi(\beta_{\text{fixed}} | X, Y), \tag{11}$$

and the corresponding measure of uncertainty is

$$\hat{\sigma}_{\beta,\text{fixed}} = \operatorname{diag}(\operatorname{Cov}_\pi(\beta_{\text{fixed}} | X, Y)^{1/2}). \tag{12}$$

Notice that the only difference between the definitions of β and β_{fixed} is that in equation (5) the density $\lambda(\cdot)$ is random while in equation (10) the corresponding density is a deterministic function of the fixed X values.

For the discrete covariate setting we obtain the following result, which is proved in the Online Supplement [Szpiro, Rice and Lumley (2010)].

THEOREM 2. *For a discrete covariate space the m-dimensional estimate $\hat{\beta}_{\text{fixed}}$ defined by equation (11) takes the form*

$$\hat{\beta}_{\text{fixed}} = (X^t X)^{-1} X^t Y,$$

COBRA
BERKELEY ELECTRONIC PRESS
Collection of Biostatistics
Research Archive

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

$$\hat{\sigma}_{\beta,\text{fixed}} = \text{diag}[(X^t X)^{-1}(X^t \Sigma^\dagger X)(X^t X)^{-1}]^{1/2},$$

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

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

and

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

Since the matrix $\Sigma^\dagger$ in the "meat" only includes variation of Y around its mean, conditional on X, this form of the sandwich estimator appropriately describes sampling variability for fixed X, even if the data-generating mechanism is nonlinear.

5. Simulations. We consider examples with a single continuous covariate uniformly distributed in the interval $[-10, 10]$, and we evaluate performance for four true distributions of y given x. These are obtained by taking combinations of the linear response

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

and the nonlinear response

$$f_{\text{nonlin}}(x) = 2 + 3.5x\big(1 + |\cos(x/2 - 2)|\big)$$

as well as the equal variance model $\sigma^2_{\text{equal}} = 5$ and unequal variance model $\sigma^2_{\text{unequal}} = (5 + x^2/5)$. Example scatterplots of data from each of the four data-generating models, along with the corresponding mean response functions, are shown in Figure 2.

For each of the four models we generate 1000 random realizations of X and Y with $n = 400, 800$. Results are given in Table 1 for inference based on random X. The model-based intervals (i.e., standard Bayesian or frequentist linear regression) fail to give approximate 95% coverage by being anti-conservative in all situations except for a linear response with equal variance. Our Bayesian robust intervals give approximately correct 95% coverage for all cases, just like the sandwich intervals.

We repeat the simulation, treating the observed design matrix X as fixed. The results are shown in Table 2. As expected, for a linear data-generating mechanism the results are essentially the same as for random X inference. The model-based intervals are correct only for the equal variance case, while the sandwich and Bayesian robust intervals give correct coverage for unequal variance as well. If the

Research Archive

TABLE 1
Frequentist properties of estimates for continuous covariate (random X)

			$n = 400$			$n = 800$		
			Bias	**Width**	**Coverage**	**Bias**	**Width**	**Coverage**
Linear	Equal variance	Model based	0.001	0.170	0.938	−0.001	0.120	0.956
		Sandwich	0.001	0.170	0.940	−0.001	0.120	0.955
		Bayes robust	0.002	0.177	0.943	0.000	0.123	0.959
	Unequal variance	Model based	0.001	0.445	0.859	−0.002	0.314	0.863
		Sandwich	0.001	0.601	0.948	−0.002	0.426	0.957
		Bayes robust	0.002	0.607	0.955	−0.001	0.428	0.956
Nonlinear	Equal variance	Model based	0.001	0.262	0.929	−0.001	0.185	0.921
		Sandwich	0.001	0.298	0.959	−0.001	0.211	0.955
		Bayes robust	0.009	0.289	0.950	0.003	0.207	0.950
	Unequal variance	Model based	0.002	0.487	0.859	−0.003	0.345	0.865
		Sandwich	0.002	0.648	0.959	−0.003	0.460	0.952
		Bayes robust	−0.030	0.657	0.951	−0.019	0.460	0.944

data-generating mechanism is nonlinear and homoscedastic, then the model-based intervals and sandwich intervals are conservative since they implicitly account for random sampling of X, while the Bayesian robust intervals give approximately nominal 95% coverage. If the data-generating mechanism is both nonlinear and heteroscedastic, the sandwich intervals are still slightly conservative when com-

TABLE 2
Frequentist properties of estimates for continuous covariate (fixed X)

			$n = 400$			$n = 800$		
			Bias	**Width**	**Coverage**	**Bias**	**Width**	**Coverage**
Linear	Equal variance	Model based	0.001	0.170	0.938	−0.001	0.120	0.956
		Sandwich	0.001	0.170	0.940	−0.001	0.120	0.955
		Bayes robust	0.002	0.173	0.941	0.000	0.121	0.954
	Unequal variance	Model based	0.001	0.445	0.859	−0.002	0.314	0.863
		Sandwich	0.001	0.601	0.948	−0.002	0.426	0.957
		Bayes robust	0.002	0.607	0.951	−0.001	0.425	0.953
Nonlinear	Equal variance	Model based	0.001	0.262	0.986	−0.001	0.185	0.998
		Sandwich	0.001	0.298	0.999	−0.001	0.211	1.000
		Bayes robust	0.009	0.187	0.959	0.003	0.128	0.963
	Unequal variance	Model based	0.001	0.487	0.893	−0.002	0.345	0.888
		Sandwich	0.001	0.648	0.961	−0.002	0.460	0.968
		Bayes robust	−0.030	0.629	0.947	−0.018	0.437	0.953

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

TABLE 3
Linear regression of average annual outpatient health care cost data from the evaluation of the Washington State Basic Health Plan

	Discrete X		**Continous X**	
	$\hat{\beta}$	$\hat{\sigma}_\beta$	$\hat{\beta}$	$\hat{\sigma}_\beta$
Model-based	16.1	1.25	16.1	1.25
Sandwich	16.1	1.67	16.1	1.67
Bayes robust (random X)	16.1	1.72	15.9	1.71
Bayes robust (fixed X)	16.1	1.70	15.9	1.70

pared to the Bayesian robust intervals, but both give approximately nominal 95% coverage. In this situation the model-based intervals are anti-conservative.

Overall, our Bayesian robust intervals give approximately nominal 95% coverage in all situations. The model-based and sandwich-based intervals can fail by being either conservative or anti-conservative, depending on details of the data-generating mechanism and the distinction between random and fixed X sampling.

6. Health care cost data. We illustrate our methods using data from the evaluation of the Washington State Basic Health Plan, as described earlier in Section 1 and in more detail by Diehr et al. (1993). We use the variable "cost of outpatient care" as the outcome and assess its "linear relationship" with age. The data are shown in the top panel of Figure 1, and O'Sullivan spline fits (see Section 4.1) to the mean and standard deviation as functions of age are shown in the bottom two panels. The thick red lines are the posterior means of the Bayesian spline fits, and the thin dashed red lines are example draws from the posterior distributions.

We can regard the age covariate as either discrete or continuous, and to illustrate our methodology, we do the analysis both ways. Results are shown in Table 3. The difference in average annual outpatient health care costs associated with a one year difference in age is estimated to be 16.1 dollars, with a model-based standard error of 1.25 dollars. As expected, in light of the heteroscedasticity, the sandwich form gives a larger standard error estimate of 1.67 dollars. The uncertainty estimates from our Bayesian robust estimators range from 1.70 dollars to 1.72 dollars, agreeing very closely with the sandwich values. The point estimate is nearly identical to the least squares fit (15.9 dollars) when we model age as continuous in the Bayesian robust approach; the slight difference is probably due to approximations involved in fitting the spline model. The Bayesian robust standard deviations are nearly identical when we model X as random or fixed, indicating that random variations in average costs conditional on age contribute more to the uncertainty than does nonlinearity in the data-generating mechanism.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper338

7. Discussion. The main contribution of this paper is a model-robust Bayesian framework for linear regression that gives uncertainty estimates equivalent to the sandwich form for random covariate sampling and with superior sampling properties for a fixed design matrix. In both situations, our estimates correctly account for heteroscedasticity and nonlinearity, in the sense of giving asymptotically valid frequentist sampling properties. The idea is to describe the data-generating mechanism nonparametrically, and then to define a functional of the true data-generating mechanism as the quantity of interest for inference. Once this quantity is defined, we follow common Bayesian practice and derive a point estimate as its posterior mean and an uncertainty estimate as its posterior standard deviation. In the case considered here, the quantity of interest is the least-squares linear fit to the (potentially nonlinear) mean of the outcome random variable y conditional on the covariate random variable x.

Our conceptual framework is powerful because it provides a general definition of linear regression in a model-agnostic framework. We can move seamlessly between classical model-based inference and robust sandwich-based inference simply by using different priors for $\phi(\cdot)$ and $\sigma(\cdot)$. If subjective prior information is available, this can also be included without any modification to the methodology. Regardless of what information is encoded in the priors, our target of inference remains the same and has an explicit interpretation in terms of the trend in the data-generating mechanism.

Our estimation approach transparently distinguishes between the cases where the observed covariates are regarded as random and where they are regarded as a fixed design matrix. We obtain good frequentist coverage properties in both situations, and our estimates are equivalent to the sandwich form when the covariates are treated as random. For the fixed design matrix setting, our Bayesian robust intervals can provide notably better sampling properties than the sandwich estimator in the situation where the true data-generating mechanism has a mean that is nonlinear in the covariates. This is true asymptotically, since the sandwich estimator overestimates the standard errors in this situation by confusing the part of the residuals that results from nonlinearity in the data-generating mechanism (which does not vary across samples and should not contribute to standard error estimates) with the random component in the residuals (which varies across samples and should be accounted for in estimating standard errors). This result can be seen in our simulation examples in Section 5 and by comparing the expressions for standard errors in Theorems 1 and 2. Elsewhere, we have derived similar results for a fixed design matrix by employing a Bayesian decision theoretic formalism [Rice, Lumley and Szpiro (2008)].

In the continuous covariate case, we use splines to approximate the mean and variance functions for y conditional on x. This is necessary because the mean and variance are not separately identifiable from a single sample at each covariate value. It can be regarded as a weakness in our approach, but it also suggests an

SCORA
Scholarship Commons
Research Archive

Hosted by The Berkeley Electronic Press

opportunity to improve on the small-sample performance of the sandwich estimator by incorporating additional prior information. Our use of splines will work in any situation where the true mean and variance are smooth functions of the covariates. This smoothness is a very reasonable assumption for applied problems. In fact, the implicit 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.

Acknowledgments. The authors would like to thank the editor and two anonymous referees for a number of very helpful suggestions.

SUPPLEMENTARY MATERIAL

Proofs of theorems in "Model robust regression and a Bayesian 'sandwich' estimator" (Szpiro, Rice, and Lumley) (DOI: 10.1214/10-AOAS362SUPP; .pdf). We provide proofs of the theorems stated in the paper "Model robust regression and a Bayesian 'sandwich' estimator" by Adam A. Szpiro, Kenneth M. Rice and Thomas Lumley.

REFERENCES

CRAINICEANU, C., RUPPERT, D. and WAND, M. P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *J. Statist. Soft.* **14** 1–24.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Applications*. Univ. Cambridge Press. MR1478673

DIEHR, P., MADDEN, C., MARTIN, D. P., PATRICK, D. L. and MAYERS, M. (1993). Who enrolled in a state program for the uninsured: Was there adverse selection? *Medical Care* **31** 1093–1105.

DIEHR, P., YANEZ, D., ASH, A. and HORNBROOK, M. (1999). Methods for analyzing health care utilization and costs. *Annual Review of Public Health* **20** 125–144.

FREEDMAN, D. A. (2006). On the so-called 'Huber sandwich estimator' and 'robust standard errors.' *Amer. Statist.* **60** 299–302. MR2291297

GASPARINI, M. (1995). Exact multivariate Bayesian bootstrap distributions of moments. *Ann. Statist.* **23** 762–768. MR1345198

GAUSS, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*. Friedrich Perthes and I.H. Besser, Hamburg.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability* **1** 221–233. MR0216620

JACOBI, C. G. J. (1841). De formatione et proprietatibus determinantum. *J. Reine Angew. Math.* **22** 285–318.

LEGENDRE, A. M. (1805). *Nouvelles Methodes Pour la Determination des Orbites des Cometes*. F. Didot, Paris.

BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper338

LIANG, K. Y. and ZEGER, S. A. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

LUMLEY, T., DIEHR, P., EMERSON, S. and CHEN, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health* **23** 151–169.

LUNN, D. J., THOMAS, A. and BEST, N. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statist. Comput.* **10** 325–337.

RICE, K., LUMLEY, T. and SZPIRO, A. (2008). Trading bias for precision: Decision theory for intervals and sets. UW Biostatistics Working Paper Series, Working Paper 336.

ROYALL, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *Int. Statist. Rev.* **54** 221–226. MR0962937

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134. MR0600538

SZPIRO, A. A., RICE, K. M. and LUMLEY, T. (2010). Supplement to "Model robust regression and a Bayesian 'sandwich' estimator." DOI: 10.1214/10-AOAS362SUPP.

WAND, M. P. and ORMEROD, J. T. (2008). On semiparametric regression with O'Sullivan penalised splines. *Aust. N.Z. J. Statist.* **50** 179–198. MR2431193

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF WASHINGTON
BOX 357232
SEATTLE, WASHINGTON 98195
USA

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press