Biometrics 000, 1–22 DOI: 000
000 2010

Modeling adverse birth outcomes via latent factor quantile regression

Lane F. Burgette* **and Jerome P. Reiter****

Department of Statistical Science, Duke University

Durham, NC 27708, USA

**email:* lb131@stat.duke.edu

***email:* jerry@stat.duke.edu

Summary:

We describe a Bayesian quantile regression model that uses a factor structure for part of the design matrix. This model is particularly useful when the data comprise numerous indicators of underlying latent factors that analysts wish to include as covariates. We apply the model to a study of birth weights, for which the effects of covariates on the lower quantiles of the response distribution are of interest. The factor structure is appealing in this study because there are many explanatory variables—each with low signal-to-noise ratios—that can be grouped into scientifically meaningful factors. We also show via simulation that the latent factor quantile regression can have greater power to detect covariate effects than standard quantile regression. The methods are available in an R package called `factorQR`.

Key words: Bayesian inference; Gibbs sampling; Low birth weight

1. Introduction

In many applications, analysts are interested in the effects of covariates on one or both of the tails of the response distribution. These effects can be analyzed directly with quantile regression (Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001). In some studies, however, the individual covariates have low signal-to-noise ratios, making it difficult to detect covariate effects on the quantiles. This is particularly likely when the data comprise many correlated covariates and only a modest sample size.

This is the case in our motivating application: to describe the predictors of low birth weight from demographic, environmental, genetic, and psychosocial variables. The data come from the Healthy Pregnancy, Healthy Baby Study (HPHBS), an ongoing prospective cohort study of women in Durham, NC, run by the Southern Center on Environmentally Driven Disparities in Birth Outcomes. The HPHBS comprises dozens of variables recorded on 1288 non-Hispanic black and white mothers. In general, the predictive abilities of the explanatory variables individually are quite weak.

Many of the explanatory variables can be viewed as indicators of underlying latent factors. Arguably, these latent factors are what should be included as predictors of birth outcomes. For example, in the HPHBS there are six moderately correlated variables related to psychosocial health. Taken together, they indicate a clearer measure of a person's mental and social well-being than any one variable alone. Thus, if we wish to assess the association between psychosocial health and pregnancy outcomes, we are better served modeling the latent variable of psychosocial health than the variables individually, which may not show up as significant due to colinearities and low signal-to-noise ratios.

To this end, we present a model for quantile regression that uses a latent factor structure for part of the design matrix. We take a fully Bayesian approach, melding what would be considered a confirmatory factor analysis in the structural equation modeling literature

(Bollen, 1989; Lee, 2007) with a quantile regression model for the response variable (Yu and Moyeed, 2001; Reed and Yu, 2009; Kozumi and Kobayashia, 2009). Using this model, we find that the latent factor of smoking is strongly associated with adverse effects on quantiles of birth weight, whereas a factor for psychosocial health is not. We also present a simulation study illustrating that the latent factor quantile regression can avoid some of the pitfalls of standard procedures for selecting important covariates, in that it can have smaller Type I error rates and greater power than standard approaches when the covariates do in fact correspond to the proposed factor structure. The methods are available in the R package called `factorQR`.

2. Background on adverse birth outcomes

Before describing the model, we give further background on adverse birth outcomes and why a latent factor quantile regression is appropriate for modeling them. Children classified as having a low or very low birth weight (below 2500g and 1500g, respectively) or a premature birth (less than 37 full weeks of gestation at birth) face increased risks of a range of problems (Hack et al., 1995), including those in the physical (Crofts et al., 1998), behavioral (McCormick et al., 1990) and mental (Lorenz et al., 1998) domains. An estimated 8% of births in the U.S. are low weight and 12.8% are preterm, so that the "average" birth in the U.S. is well above the cutoffs (Behrman and Butler, 2007). Thus, if a particular exposure decreases the birth weights of babies who would otherwise have average to high birth weights by some modest amount, it might not be considered deleterious. On the other hand, if an exposure lowers the birth weights of babies who would already be low—by the same modest amount—the exposure would be troubling. Thus, it is appropriate to consider quantile regression over standard linear regression for modeling birth weights.

Researchers have examined numerous variables to explain these adverse birth outcomes. These include environmental variables like air quality, tobacco smoke exposure, and pesticide

exposure (Miranda et al., 2009); psychological stress (Rondo et al., 2003; Orr et al., 2007); and maternal nutrition (Kramer, 1987), among many others. There is considerable disagreement in the literature on the importance of the various proposed explanations of adverse birth outcomes. For example, Rondo et al. (2003) found that maternal stress was associated with both low weight and preterm births, though this conflicts with the findings of Brooke et al. (1989) who found no effect. Peacock et al. (1995) found that stress decreases the length of the pregnancy, but increases birth weight after controlling for gestational age.

Some of this discordance might be attributable to measurement issues. For example, suppose that an analyst is interested in tobacco smoke exposure as an explanatory variable. In practice, it is not clear which single exposure metric of tobacco smoke exposure one should use. Lab assays of cotinine (a metabolite of nicotine) levels in maternal blood or urine are a common measure of tobacco smoke exposure, but cotinine has a half-life of around nine hours in pregnant women (Dukic et al., 2007). Hence, a single cotinine measurement may inaccurately reflect exposure over the course of the pregnancy. Alternatively, self-reported smoking measures can be biased by poor recall and misreporting. Wang et al. (1997) struggle with this exact issue. They find that cotinine levels are an important predictor of lower average birth weights, but that the evidence is "less clear when maternal self-reports of smoking were used" (p. 984). They go on to write that, "The stronger exposure-response relationship for cotinine concentrations suggests that this objective measure more accurately represents the individual differences in smoking behaviour" (p. 984). While this may be true, it seems risky to judge the reliability of competing measurements based on the strength of a relationship that one simultaneously attempts to estimate.

Assuming a factor structure partially resolves this issue. It enables analysts to pool the information from these multiple, imperfect measurements in hopes of more accurately representing the exposure in the quantile regression. Because the method accounts for the

uncertainty between the observed measures and the underlying latent factor, this method is more principled than, for example, taking simple averages or choosing the most significant measure in a multiple regression.

As a related measurement issue, some individual exposures arguably affect birth outcomes through a common causal pathway, so that in actuality they are indicators of an underlying factor. For example, suppose that psychological stress presents differently for many mothers. Some may feel stress because they are socially isolated, some because their pregnancy was unwanted, and others because they feel they are incapable of influencing events that affect their lives (Bandura, 2010). Further, suppose that high levels of psychological stress, however presented, have a negative effect on birth weights. If the indicators are modestly correlated and have low incidence rates marginally, individually they may not be strongly associated with birth weight in analyses, even though their underlying factor is. The factor structure enables analysts to represent such underlying constructs accurately in regression models.

Another reason for discordance is sample size. Most studies with sufficiently detailed data to address complex hypotheses typically have modest sample sizes and, therefore, low power; this makes it challenging to replicate findings. Additionally, analyses often involve many comparisons, thus augmenting the chances of Type 1 errors. This provides further motivation for the factor model: as illustrated with simulations in Section 3, the factor model can improve both power and Type 1 error compared to standard quantile regression.

3. A quantile regression factor model

We now turn to the Bayesian latent factor quantile regression. For each record $i = 1, \ldots, n$, let y_i be the response variable (e.g., birth weight) and $x_i = (x'_{i,\omega}, x'_{i,\beta})'$ be the vector of covariates, where $x_{i,\omega}$ are indicators of the latent factors and $x_{i,\beta}$ are exogenous to the latent

factors. We assume that, for all i,

$$\begin{bmatrix} x_{i,\omega} \\ y_i \end{bmatrix} = \Lambda\omega_i + \begin{bmatrix} 0 \\ x'_{i,\beta}\beta \end{bmatrix} + \begin{bmatrix} \varepsilon_{i,-s} \\ \varepsilon_{i,s} \end{bmatrix}. \tag{1}$$

Here, Λ is a $s \times t$ matrix of factor loadings, ω_i is a $t \times 1$ vector of latent factors and β is the vector of regression coefficients for $x_{i,\beta}$. We assume that $\varepsilon_{i,-s} \overset{\text{iid}}{\sim} \text{normal}(0, \Psi)$ where $\Psi = \text{diag}(\psi_1, \ldots, \psi_{s-1})$.

To model conditional quantiles of y_i, we assume that $\varepsilon_{i,s}$ follow an asymmetric Laplace distribution $\text{AL}(0, \tau, p)$, where the parameters are the location, scale and skewness, respectively (Yu and Zhang, 2005). We have

$$p(\varepsilon_{i,s}) = \tau p(1-p)\exp\{-\tau\rho_p(\varepsilon_{i,s})\}, \tag{2}$$

where ρ_p is the tilted absolute value or check function

$$\rho_p(x) = \begin{cases} px & \text{if } x \geqslant 0 \\ (1-p)|x| & \text{if } x < 0 \end{cases}. \tag{3}$$

The asymmetric Laplace distribution is commonly used as the basis of Bayesian quantile regression because $\Pr(\varepsilon_{i,s} \leqslant 0) = p$. This error distribution also connects the Bayesian analysis to frequentist quantile regression, which proceeds semi-parametrically using ρ_p as a loss function.

The asymmetric Laplace distribution is a convenient choice from a computational standpoint. Reed and Yu (2009) and Kozumi and Kobayashia (2009) independently demonstrated that this distribution can be represented as a mixture of normal distributions. If

$$y_i|w_i, \mu, \tau \sim \text{normal}\left(\frac{1-2p}{p(1-p)}w_i + \mu, \frac{2w_i}{\tau p(1-p)}\right) \tag{4}$$

and $w_i|\tau, \mu$ are independent with an exponential distribution with rate τ, then marginalizing over w_i gives us $y_i|\beta, \tau \sim \text{AL}(\mu, \tau, p)$. Thus, augmenting the parameter space with the latent w_i enables Bayesian inference using the usual normal distribution theory.

In this model, we assume that scientific considerations allow the analyst to specify structure

in the matrix of factor loadings Λ. In particular, we assume

$$\Lambda = \begin{bmatrix} \bigoplus_{j=1}^{t} \boldsymbol{\lambda}_j \\ \Lambda_s \end{bmatrix}, \tag{5}$$

where $\boldsymbol{\lambda}_j$ are column vectors whose first element is fixed at 1 and $\oplus$ indicates the direct sum such that

$$A \oplus B = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}. \tag{6}$$

Λ_s is a t-vector of factor loadings related to the response y; it is analogous to the regression parameters in a standard regression model.

With structural zeros in the Λ matrix, this model is related to but distinct from less structured Bayesian factor models (e.g., Geweke and Zhou, 1996; Aguilar and West, 2000). In a continuum of statistical methods with classical methods designed for $n > p$ at one end and $p \gg n$ factor models at the other extreme (West, 2003), this model is somewhere in the middle. It can significantly reduce the dimension of the predictor space, though it requires the analyst to specify *a priori* which observed variables relate to each latent factor.

Specifying the block diagonal pattern in Λ_{-s} aids interpretation of the factors, especially compared to tools like the singular value decomposition, which often provides factors that are difficult to interpret. For instance, in the HPHBS application, we consider $\boldsymbol{\lambda}_1$ to measure tobacco smoke exposure and $\boldsymbol{\lambda}_2$ to relate to psychosocial health. The factors have this interpretation because investigator-specified zeros in Λ_{-s} ensure that, for instance, only smoking-related measures are tied to $\boldsymbol{\lambda}_1$. In some cases—particularly those with extremely large numbers of explanatory variables like exploratory genetic association studies—it may not be possible to sensibly segment the predictors in this way; we do not focus on such situations.

3.1 *Prior distributions*

We hierarchically specify the following prior distributions for the parameters of (1) as follows.

$$\tau \sim \text{gamma}(c_\tau, d_\tau) \tag{7}$$

$$\psi_k \overset{\text{ind}}{\sim} \text{inv-gamma}(c_k, d_k) \tag{8}$$

$$\beta \sim \text{normal}(a_\beta, B_\beta^{-1}) \tag{9}$$

$$\Lambda_s \sim \text{normal}(a_s, B_s^{-1}) \tag{10}$$

$$\lambda_k|\psi \sim \text{normal}(a_{\lambda,k}, \psi_k/h_{0,k}) \tag{11}$$

$$\Phi \sim \text{inv-Wishart}(R_0, \nu_0) \tag{12}$$

$$v_i|\tau \overset{\text{iid}}{\sim} \text{inv-gamma}(1, \tau) \tag{13}$$

$$\omega_i|\Phi \overset{\text{iid}}{\sim} \text{normal}(0, \Phi) \tag{14}$$

In this specification, v_i is the inverse of the mixing weight w_i in (4). This set of prior distributions combines those of the Bayesian confirmatory factor analysis (Lee, 2007) and the scale mixture representation of the asymmetric Laplace distribution (Reed and Yu, 2009).

We use weakly informative prior distributions in the analyses. In particular, for the gamma-distributed components, we use $c_\tau = d_\tau = 1$ and $c_k = d_k = 1$ for all k. This results in a proper but diffuse distribution that lacks a prior first moment for ψ_k or τ^{-1}. For the factor loadings of interest, Λ_s, the prior distribution is centered at zero with variance 100; β is given the same distribution. For the other factor loadings λ_k, we center the prior distributions at one (a neutral value) and specify $h_{0,k} = 1$ for all k. We center Φ at the identity matrix, and use prior degrees of freedom equal to $t + 1$. In the HPHBS application, we find little sensitivity to making the normal distributions more or less diffuse. In some cases, using a larger ν_0 may speed convergence to regions of high posterior probability.

3.2 *Estimation*

The posterior distributions of parameters in the model can be simulated via Gibbs sampling. Because of the mixture representation of the asymmetric Laplace distribution, the sampling requires draws from standard distributions. The algorithm proceeds as follows:

(1) Sample $\tau|\text{all} \sim \text{gamma}(c_0 + n, \text{rate} = d_0 + \sum_{i=1}^{n} \rho_p(y_i - \Lambda_s \omega_i))$

(2) Sample $v_i|\text{all} \overset{\text{ind}}{\sim} \text{inv-Gaussian}\big(\frac{1}{p(1-p)|y_i - \Lambda_s \omega_i|}, \frac{\tau}{2p(1-p)}\big)$

(3) Sample

$$\omega_i|\text{all} \overset{\text{iid}}{\sim} \text{normal}((\Phi^{-1} + \Lambda'\tilde{\Psi}^{-1}\Lambda)^{-1}\Lambda'\tilde{\Psi}^{-1}\tilde{y}_i, (\Phi^{-1} + \Lambda'\tilde{\Psi}^{-1}\Lambda)^{-1}), \tag{15}$$

where $\tilde{y}_i = (x_i', u_i)'$ with $u_i = y_i - (1-2p)/(p(1-p)v_i)$ and

$$\tilde{\Psi}^{-1} = \text{diag}(\psi_1^{-1}, \ldots, \psi_{s-1}^{-1}, .5\tau p(1-p)v_i)$$

(4) Sample $(\psi_k, \lambda_k)|\text{all}$ for the rows of Λ_{-q} that have free elements:

- Draw from

$$\psi_k^{-1}|x, \omega \sim \text{gamma}(n/2 + c_k, d_k + .5\sum_i (x_{ik} - \omega_{i,j(k)}\hat{\lambda}_k)^2 + .5h_{0,k}(\hat{\lambda}_k - a_{\lambda,k})^2)$$

  where $\hat{\lambda} = \frac{\sum_i x_i\omega_i + a_{\lambda,k}h_{0,k}}{h_{0,k} + \sum_i \omega_i^2}$ and where $j(k)$ refers to the dimension of ω_i that is activated when we take the inner product of ω_i and the kth row of Λ_{-s}

- Draw from $\lambda_k|x, \omega, \psi_k \sim \text{normal}\big(\hat{\lambda}_k, \psi_k/(h_{0,k} + \sum_i \omega_{i,j(k)}^2)\big)$

(5) For the rows of Λ_{-s} that do not have any free elements, sample $\psi_k^{-1} \sim \text{gamma}(c_k, d_k + .5\sum_i (x_{ik} - \omega_{i,j(k)})^2)$

(6) Sample $\Lambda_s|\text{all} \sim \text{normal}(\hat{\Lambda}_s, (.5\tau p(1-p)\Omega'V\Omega + B_s)^{-1})$, where

$$\hat{\Lambda}_s = (.5\tau p(1-p)\Omega'V\Omega + B_s)^{-1}(.5\tau p(1-p)\Omega'Vu + B_s a_s),$$

$V = \text{diag}(\{v_i\})$ and Ω is the $n \times t$ matrix of ω_i. If some covariates do not relate to the latent factors, sample (Λ_s, β) jointly.

(7) Sample $\Phi|\text{all} \sim \text{inverse-Wishart}(\Omega'\Omega + R_0^{-1}, n + \nu_0)$.

In general, the mixing of the Markov chains resulting from this algorithm tends to be quite good. However, for certain initial values in the estimation algorithm, the MCMC can remain for a time in a region where the variance/covariance matrix Φ is very small, with large Λ_s to compensate. In our experience, chains starting in such regions would eventually settle into areas of high posterior probability with more moderate Φ and Λ_s values. We recommend at least one long run (at least 100,000 iterations) to build confidence that the chain is not stuck in some local posterior mode. When the long run reveals no problems, we have found that 2500 burn-in and 5000 stored iterations suffice.

3.3 *Model extensions*

The base model in (1) can be extended in several directions. Here, we describe some that were useful for the HPHBS analysis.

3.3.1 *Dichotomous predictors.* When some of the indicators of the latent factor are dichotomous, the assumption of normality is clearly unrealistic. In this case, analysts can employ a probit model for the Bernoulli indicators. This assumes a latent normal quantity whose sign alone is observed. The scale of such models is not identified; this is taken care of by assuming unit variance for the latent quantity. With a normal prior distribution on a common location parameter for the latent variables, augmenting the parameter space with the latent quantities fits naturally into the sampler described above.

3.3.2 *Manifest/latent variable interactions.* Often scientific hypotheses relate to interactions involving the latent factors in the models. For example, in the HPHBS we are interested in interactions between a latent tobacco exposure factor and manifest (i.e., observed) genetic measures. For such manifest/latent interactions, one can express y_i in (1) as

$$y_i = \Lambda_s \omega_i + \Lambda_{s,\text{int}}(x_{i,\text{int}} \odot \omega_{i,\text{int}}) + x_{i,\beta}\beta + \varepsilon_{is}, \tag{16}$$

where $\Lambda_{s,\text{int}}$ is a row vector of interaction effects, $x_{i,\text{int}}$ is a vector of manifest variables related to the interaction, $\omega_{i,\text{int}}$ is a vector of (perhaps repeated) elements of ω_i, and $\odot$ indicates the Hadamard (component-wise) product. Variables in $x_{i,\text{int}}$ can be repeated in $x_{i,\beta}$.

Sampling proceeds in roughly the same manner as in the base model without interactions, albeit with minor modifications. Let $\Omega^* = [\Omega, \Omega_{\text{int}} \odot X_{\text{int}}]$, where Ω_{int} and X_{int} contain $\omega_{i,\text{int}}$ and $x_{i,\text{int}}$ in their rows, respectively. In the draw of Λ_s in Step 6 of the algorithm, Ω is replaced by Ω^*, which yields a joint draw of Λ_s and $\Lambda_{s,\text{int}}$. In Steps 1, 2 and 3, the sth row of Λ (Λ_s) is replaced by $\Lambda_s + \Lambda_{s,\text{int}} \odot x_{i,\text{int}}$.

3.3.3 *Latent variable interactions.* Analysts might be interested in estimating interactions between two (or more) of the latent factors. For instance, rather than modeling $y_i = \Lambda_s \omega_i + \varepsilon_i$, one might have $y_i = \Lambda_s \omega_i + \omega_{i,1}\omega_{i,2}\lambda_{1,2} + \varepsilon_i$. Conditional on the latent factors, the estimation strategy is not complicated, just as linear regression with interaction terms is not materially different from one that only contains main effects. However, the interaction complicates the sampling of ω_i somewhat, as they are no longer conditionally conjugate.

Analysts can use a Metropolis-Hastings step to sample the ω_i. We have found that a proposal distribution based on the Gibbs transition that does not account for the interaction works sufficiently well. We perturb the current ω_i value by adding a normally distributed draw with mean zero and covariance $(\Phi^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}$. When the effect of the interaction is weak, this proposal should be very competitive. We accept the proposed value of ω_i with probability

$$\min\left\{1, \frac{p(\omega_i^{\text{prop}}|y_i,\theta)}{p(\omega_i^{\text{curr}}|y_i,\theta)}\right\} \tag{17}$$

where "curr" and "prop" stand for current and proposed values, respectively, and

$$p(\omega_i|y_i,\theta) \propto \exp\{-.5\omega_i'\Phi^{-1}\omega_i - .5(x_i - \Lambda_{-s})'\Psi^{-1}(x_i - \Lambda_{-s}) - .5(y_i - \mu_{y_i})^2/\sigma_y^2\} \tag{18}$$

where μ_y and σ_y^2 are the mean and variance in (4).

Since all ω_i are conditionally independent of one another, only one observation affects the

acceptance probabilities. This means that the conditional density of ω_i tends to be relatively flat, and that the contribution of the interaction term should not dominate the distribution, making it relatively easy to make competitive Metropolis proposals.

4. Simulation results of power and Type 1 error

4.1 *Factor models and power*

One primary motivation for using the factor model is the desire to make the most of data characterized by a relatively modest sample size and weak signals. In this section, we present a simulation study of the statistical power of the latent factor quantile regression. We proceed by comparing the estimates from the quantile factor regression model to those of standard quantile regressions that ignore the factor structure. We simulate data from the factor model, so it should be no surprise that the factor model outperforms the standard quantile regressions; the question is whether the standard regressions adequately uncover the underlying structure.

We simulate data using a one factor ($t = 1$) model with

$$\Lambda = [1, 0.5, 1.5, 0.5, 1.5, 0.2]', \tag{19}$$

and Ψ equal to the identity matrix. This gives us a theoretical variance in the explanatory variables of

$$\begin{bmatrix} 2 & 0.5 & 1.5 & 0.5 & 1.5 \\ 0.5 & 1.25 & 0.75 & 0.25 & 0.75 \\ 1.7 & 0.75 & 3.25 & 0.75 & 2.25 \\ 0.5 & 0.25 & 0.75 & 1.25 & 0.75 \\ 1.5 & 0.75 & 2.25 & 0.75 & 3.25 \end{bmatrix}. \tag{20}$$

We model the 20th percentile, which corresponds to the true asymmetric Laplace error distribution.

To carry out the simulation, we consider sample sizes between 250 and 1,000. We fit

the data using the (correct) factor model, and frequentist and Bayesian quantile regression models that do not model the observed covariates as being related to an underlying factor. For the Bayesian analyses, we assign a prior variance of 10 (centered at zero) to the regression coefficients. We run the Bayesian estimation algorithms for 10,000 iterations, and discard the first 5,000 as burn-in. Rarely, the Bayesian methods gave results that indicated that convergence was not reached (see the end of section 3.2). We re-ran the estimation procedure in such cases.

We summarize the results in Table 1. Even with the smaller sample sizes, the factor model nearly always estimates the regression coefficient of interest to be different from zero. On the other hand, the standard quantile regressions often result in confidence (or credible) intervals that overlap zero for all coefficients. Although this does not imply that hypothesis tests would select a null model, the lack of marginally significant variables at the very least makes interpretation more difficult.

With a sample size of 750 or 1000, the standard quantile regressions typically find at least one variable with an associated 95% interval that does not cover zero. However, among the 800 datasets that we simulated, in only one case were all five intervals completely above zero for the frequentist regression, and this never happened for the standard Bayesian quantile regression. Even with 1000 observations, the most common outcome was two intervals that did not cover zero for both of the non-factor approaches. Thus, if scientific considerations suggest an underlying factor structure, ignoring it can harm analysts' ability to detect signals in the data.

[Table 1 about here.]

[Table 2 about here.]

4.2 *Factor models and multiple comparisons*

In addition to facilitating detection of effects that may go unnoticed, factor models also can reduce exposure to the issue of multiple comparisons. For example, with five measures of tobacco smoke exposure, it would be easy to inflate the chance of a Type 1 error by choosing the measure that yields the most significant effect. Yet, with relatively high correlations amongst the predictors, a Bonferroni-type correction is an unappealing option as it is likely to be overly conservative.

To demonstrate this lower risk of making a Type 1 error when accounting for the factor structure, we repeat the previous simulations, this time with $\Lambda_s = 0$. For this simulation, we count the percent of fits that have credible or confidence intervals that do not overlap zero. (See Table 2.) In this situation, we find that the factor model indeed does produce such intervals much less often. Both the Bayesian factor and non-factor models use a prior for the regression coefficients centered at zero with variance 10. The effect of this weak prior along with the correctly specified likelihood explain the superior performance of the standard Bayesian quantile regression compared to the frequentist version (the default behavior of the `rq` function in the `quantreg` package (Koenker, 2009)), which appears somewhat anti-conservative in this case.

5. Analysis of the HPHBS

We now turn to the analysis of the HPHBS data. Among the eligible cases (singleton births in Durham, NC), 1009 mothers are non-Hispanic black women and 279 are non-Hispanic white women. Among these cases, 203 of the births qualify as preterm births; 184 of the births are below the 2500g cutoff for the low birth weight classification; and, 38 are below the 1500g very low birth weight cutoff. We focus on only birth weight as the outcome variable here.

As with many observational studies, there is a modest amount of missing data spread over many variables. We address this though multiple imputation (Rubin, 1987) of ten completed datasets using sequential regression trees; see Burgette and Reiter (2010) for details. To estimate the posterior distributions of the parameters in the Bayesian quantile regressions, we run separate Gibbs samplers on each of ten completed datasets. After convergence of each sampler, we mix the resulting parameter draws to form a sample from the posterior distribution; see (Gelman et al., 2004, p. 520) and Zhou and Reiter (2010) for justification of this approach.

5.1 *Factor main effects: smoking and psychosocial health*

We begin with an analysis of birth weight that includes factors related to smoking and psychological and social well-being. For the smoking latent factor, we use laboratory measures of blood levels of nicotine and cotinine, along with self-reported measures of tobacco use at the time of delivery, exposure to environmental tobacco smoke, and an indicator of whether the mother is a current or former smoker, taken earlier in the pregnancy.

The psychosocial health factor is conceptually broader, though mothers with high values would seem to be well-adjusted and exposed to low stress. The indicators of the factor include the total score on the Center for Epidemiologic Studies Depression Scale (CES-D), which summarizes 20 aspects of depressive feelings and actions (Radloff, 1977); an indicator of whether the pregnancy was unwanted rather than wanted or merely mistimed; a measure of paternal support; a measure of perceived self-efficacy; and a score on the Interpersonal Support Evaluation List (ISEL) test, which measures the availability of social support (Cohen et al., 1985). See the online supplementary material for further descriptions of these variables. We standardize all of these quantities (other than dichotomous variables), and switch signs as needed so that "higher is better" for each indicator variable. Thus, a mother with a low

score on this factor might be pregnant with an unwanted baby, feel unable to control the course of her own life, have little support and feel depressed.

Pointing each manifest variable in the same direction is desirable because we center the prior distribution for the free elements in Λ_{-q} at one; making the switch in scale makes the prior-induced regularization more consistent. However, negative factor loadings are allowable. In the fitted model, both sets of the factors have estimated loadings in Λ_{-s} that are positive and effectively bounded well away from zero. This suggests that the chosen manifest variables capture meaningful latent quantities.

In addition to the factors, we include a set of standard controls. These include the mother's age, race, marital status (married or cohabiting versus separated) and the sex of the child, since boys are known to be heavier on average.

[Figure 1 about here.]

Figures 1 and 2 graphically summarize the results. The factor loading connecting the response with the tobacco exposure factor is estimated to be significantly below zero at the 50th percentile. The effect is slightly less strong on the lower percentiles of the response distribution. In comparison, the results of a frequentist median regression indicate that none of the tobacco-related variables have significant t-values at the 95% level when they are all included in the model; the point estimates associated with nicotine and cotinine are even positive. On the other hand, when we include nicotine or cotinine as the only tobacco-related measure, the associated estimate is negative and quite significant. Hence, the factor structure simplifies interpretation of the results and does not require us to choose from the available measures of smoking exposure.

The estimated loadings for the psychosocial factor are centered firmly on zero, indicating no significant effect at any of the response quantiles. Even in our failure to find an interesting effect, we see the advantage of the factor model. If we do not assume the latent structure,

the individual point estimates cover a range of positive and negative estimates, but with credible intervals that all cover zero. It can be difficult to distill these varying estimates into a statement of whether there is even weak evidence for the effects of measures in this domain. The factor model is less ambiguous in this regard: it simply does not support the hypothesis that the psychosocial variables are associated with low birth weights in these data.

[Figure 2 about here.]

Among the standard controls in Figure 2, the results are in accord with what has been reported elsewhere. Male babies tend to be heavier, though this effect decays at the low end of the response distribution. Black babies are significantly lower in weight, especially at the low end of the response distribution. Additionally, there is some evidence that women who are married or cohabiting have better birth outcomes.

5.2 *Including interactions: a genetic analysis*

A specific hypothesis of interest is the presence of an interactive effect between the interleukin 6 (IL-6) gene (or more precisely the protein that is encoded by that gene) and tobacco smoke exposure. This is plausible because there is evidence that the IL-6 protein is a key player in the inflammatory response pathway (Kaplanski et al., 2003), and that tobacco exposure can induce this type of inflammation (Van der Vaart et al., 2004). Similar gene/tobacco interactive effects have been implicated previously in the study of low birth weights (Wang et al., 2002).

In the HPHBS data, we have single nucleotide polymorphism (SNP) information at three loci in the IL-6 gene. It is not known *a priori* how the SNP markers relate to IL-6 expression, merely that they capture genetic diversity in this important gene. The SNP observations are unordered, with three levels at each site, so we use two dummy variables to encode the information with the most common variant as the base category. We interact these dummy variables with the latent tobacco factor, which is modeled using the manifest indicators

described in the previous subsection. We keep the same two-factor form as in the previous section, simply adding on the SNP/tobacco interaction.

In short, the genetic information described by the SNP observations do not describe a meaningful amount of the variability in birth weight outcomes at the 20th and 50th percentiles. All of the 90% credible intervals comfortably cover zero (Table 3). Even though the factor model failed to detect a meaningful interaction in these genetic analyses, as well as several others not reported here, the factor model is useful in two respects. First, by allowing for a latent tobacco factor, we make use of all of the available information on tobacco exposure, which should capture more information than any single manifest variable in the data. Thus, if there is an effect, we would expect that the factor model would give us a better chance of detecting it than a single, noisier measurement would. Second, the factor structure eliminates an aspect of the problem of multiple comparisons. If the interaction between one tobacco-related measure and a SNP indicator were significant, but the relationship between the same SNP and another tobacco measure were not, it is not entirely clear what our conclusion should be.

[Table 3 about here.]

5.3 *Model checking*

Checking model fit is an important step in the modeling process (Gelman et al., 1996). In this study, we use a method suggested by Lee and Neocleous (2010), which compares predicted quantiles to the observed data. Specifically, we examine the quantity

$$T_p = \frac{1}{n}\sum_{i=1}^{n} I\{y_i \leqslant \hat{Q}_p(i|\Omega, \beta, \Lambda_s, x_{i,\beta})\}, \tag{21}$$

where $\hat{Q}_p(i|\cdot)$ is the predicted pth quantile for the ith record in the dataset. Because we have multiple draws of the parameters that define these predictions, we average T_p across the stored MCMC draws. If the quantile regression model of the pth quantile fits well, we expect $T_p \approx p$, so we focus on plots of T_p versus p.

From the dashed line in Figure 3, we can see that T_p is close to p across the range of quantiles for our Bayesian factor model, though T_p does tend to be lower than p. We compare this to a standard frequentist quantile regression that does not assume the factor structure, so that x_β and x_ω are included as standard explanatory variables. Examining the dotted line in the figure, we can see that T_p is much smaller than p for larger values of p. This may be evidence that the standard quantile regression is over-fitting the data for large values of p since, for example, nearly one third of the observations are larger than the corresponding fitted, conditional 90th percentiles.

[Figure 3 about here.]

6. Summary

We developed an approach for Bayesian quantile regression when some covariates are indicators of underlying common factors. We applied this model in a study of the predictors of birth weights. The results suggested that smoking during pregnancy is associated with decreased birth weight, even at the lower end of the response distribution. This is in accord with the meta-analysis of Shah and Bracken (2000). However, the results did not suggest a significant effect of psychosocial factors on birth weights. Of course, we could be missing important confounders that mask effects in the study, as is the case with any observational study. Nonetheless, we hope that the methodology presented here is useful in other analyses of birth outcomes and related contexts.

Received September 2010.

Supplementary Materials

Web Appendices referenced in Section 5.1 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgements

The authors wish to thank Allison Ashley-Koch, David Dunson, Melanie Garrett, Pamela Maxson and Marie Lynn Miranda for helpful suggestions. This research was funded by Environmental Protection Agency grant R833293.

References

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics* **18,** 338–357.

Bandura, A. (2010). Self-efficacy. In *Corsini Encyclopedia of Psychology*. Wiley Online Library.

Behrman, R. and Butler, A. (2007). *Preterm Birth: Causes, Consequences, and Prevention*. National Academies Press, Washington, DC.

Bollen, K. (1989). *Structural equations with latent variables*. Wiley-Interscience, New York.

Brooke, O., Anderson, H., Bland, J., Peacock, J., and Stewart, C. (1989). Effects on birth weight of smoking, alcohol, caffeine, socioeconomic factors, and psychosocial stress. *British Medical Journal* **298,** 795–801.

Burgette, L. and Reiter, J. (2010). Multiple imputation via sequential regression trees. *American Journal of Epidemiology* forthcoming, doi: 10.1093/aje/kwq260.

Cohen, S., Mermelstein, R., Kamarck, T., Hoberman, H., Sarason, I., and Sarason, B. (1985). Social support: Theory, research and applications. *The Hague, Netherlands: Martinus Nijhoff* pages 73–94.

Crofts, B., King, R., and Johnson, A. (1998). The contribution of low birth weight to severe vision loss in a geographically defined population. *British Journal of Ophthalmology* **82,** 9–13.

Dukic, V., Niessner, M., Benowitz, N., Hans, S., and Wakschlag, L. (2007). Modeling

the relationship of cotinine and self-reported measures of maternal smoking during pregnancy: A deterministic approach. *Nicotine & Tobacco Research* **9,** 453–465.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis.* Chapman and Hall/CRC, Boca Raton, FL.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6,** 733–759.

Geweke, J. and Zhou, G. (1996). Measuring the price of the arbitrage pricing theory. *Review of Financial Studies* **9,** 557–587.

Hack, M., Klein, N., and Taylor, H. (1995). Long-term developmental outcomes of low birth weight infants. *The Future of Children* **5,** 176–196.

Kaplanski, G., Marin, V., Montero-Julian, F., Mantovani, A., and Farnarier, C. (2003). IL-6: A regulator of the transition from neutrophil to monocyte recruitment during inflammation. *Trends in Immunology* **24,** 25–29.

Koenker, R. (2009). *quantreg: Quantile regression.* R package version 4.44.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica* **46,** 33–50.

Koenker, R. and Hallock, K. (2001). Quantile regression. *Journal of Economic Perspectives* **15,** 143–156.

Kozumi, H. and Kobayashia, G. (2009). Gibbs sampling methods for Bayesian quantile regression. *Kobe University Technical Report* **2009-2,** 1–15.

Kramer, M. (1987). Determinants of low birth weight: Methodological assessment and meta-analysis. *Bulletin of the World Health Organization* **65,** 663–737.

Lee, D. and Neocleous, T. (2010). Bayesian quantile regression for count data with application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* forthcoming.

Lee, S. (2007). *Structural equation modeling: A Bayesian approach.* Wiley, Hoboken, New

Jersey.

Lorenz, J., Wooliever, D., Jetton, J., and Paneth, N. (1998). A quantitative review of mortality and developmental disability in extremely premature newborns. *Archives of Pediatrics and Adolescent medicine* **152,** 425–435.

McCormick, M., Gortmaker, S., and Sobol, A. (1990). Very low birth weight children: Behavior problems and school difficulty in a national sample. *Journal of Pediatrics* **117,** 687–693.

Miranda, M., Maxson, P., and Edwards, S. (2009). Environmental contributions to disparities in pregnancy outcomes. *Epidemiologic Reviews* **31,** 67–83.

Orr, S. T., Reiter, J. P., Blazer, D. G., and James, S. A. (2007). Maternal prenatal pregnancy-related anxiety and spontaneous preterm birth in Baltimore, Maryland. *Psychosomatic Medicine* **69,** 566–570.

Peacock, J., Bland, J., and Anderson, H. (1995). Preterm delivery: Effects of socioeconomic factors, psychological stress, smoking, alcohol, and caffeine. *British Medical Journal* **311,** 531–535.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* **1,** 385–401.

Reed, C. and Yu, K. (2009). A partially collapsed Gibbs sampler for Bayesian quantile regression. *Brunel University Technical Report* pages 1–20.

Rondo, P., Ferreira, R., Nogueira, F., Ribeiro, M., Lobert, H., and Artes, R. (2003). Maternal psychological stress and distress as predictors of low birth weight, prematurity and intrauterine growth retardation. *European Journal of Clinical Nutrition* **57,** 266–272.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons Inc.

Shah, N. and Bracken, M. (2000). A systematic review and meta-analysis of prospective studies on the association between maternal cigarette smoking and preterm delivery.

American Journal of Obstetrics and Gynecology **182,** 465–472.

Van der Vaart, H., Postma, D., Timens, W., and Ten Hacken, N. (2004). Acute effects of cigarette smoke on inflammation and oxidative stress: A review. *Thorax* **59,** 713–721.

Wang, X., Tager, I., Van Vunakis, H., Speizer, F., and Hanrahan, J. (1997). Maternal smoking during pregnancy, urine cotinine concentrations, and birth outcomes. A prospective cohort study. *International Journal of Epidemiology* **26,** 978–988.

Wang, X., Zuckerman, B., Pearson, C., Kaufman, G., Chen, C., Wang, G., Niu, T., Wise, P., Bauchner, H., and Xu, X. (2002). Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. *Journal of the American Medical Association* **287,** 195–202.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. *Bayesian Statistics* **7,** 723–732.

Yu, K. and Moyeed, R. (2001). Bayesian quantile regression. *Statistics and Probability Letters* **54,** 437–447.

Yu, K. and Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics-Theory and Methods* **34,** 1867–1879.

Zhou, X. and Reiter, J. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician* **64,** 159–163.

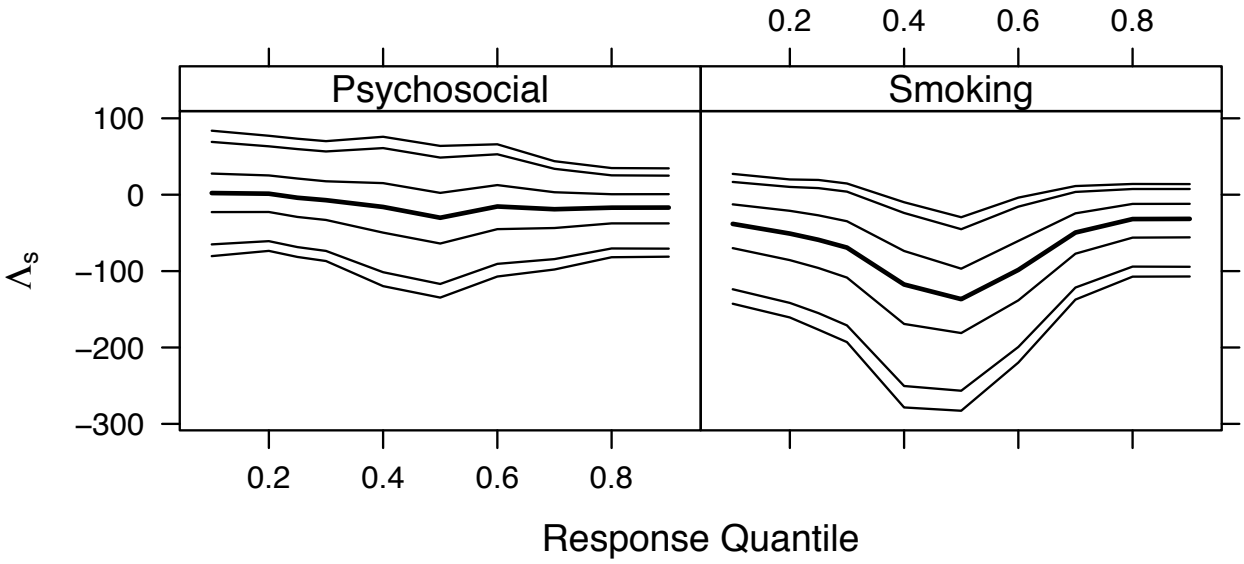

Figure 1. Middle 95%, 90% and 50% confidence bands and median for response factor loadings Λ_s. The tobacco loading includes five measures of tobacco-smoke related exposure, and the psychosocial health factor relates to six measures of well-being. The coefficient estimates correspond to a change of two standard deviations in the latent factor scale.

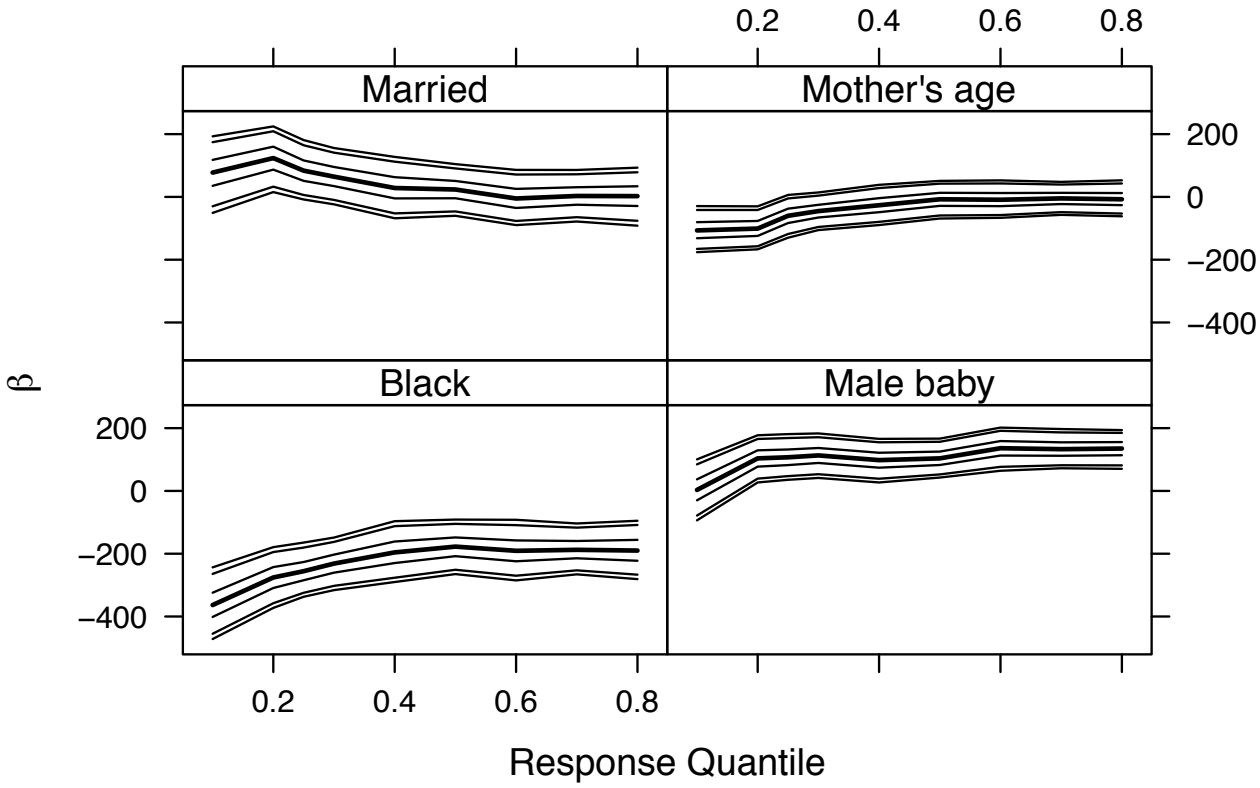

Figure 2. Middle 95%, 90% and 50% confidence bands and median for standard control variables across various response quantiles. The mother's age is measured in decades.

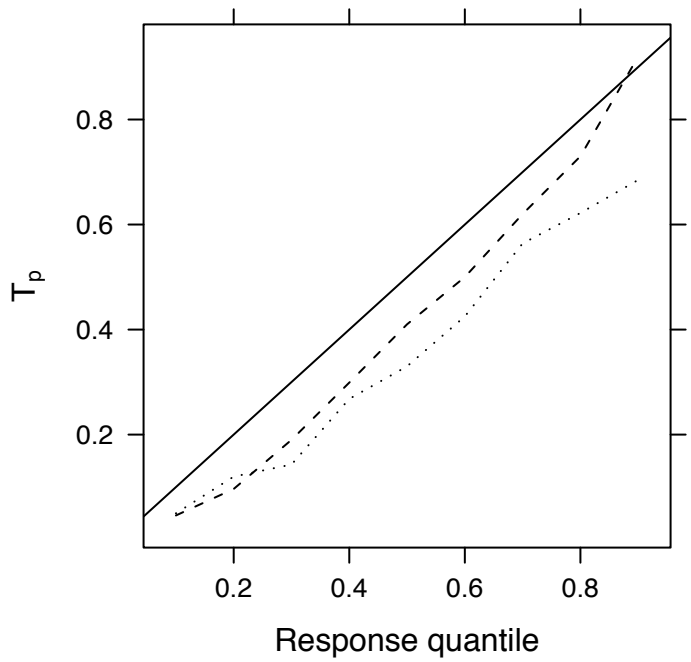

Figure 3. Percent of observations less than the fitted quantile (T_p) as a function of the response quantile, for the Bayesian factor model (dashed line) and standard frequentist quantile regression (dotted line), with the $y = x$ line for reference (solid line).

Table 1

Simulation results of estimated statistical power. The first column is the estimated power for the factor model. The next two columns are from the standard Bayesian quantile regressions, with the left column estimating the probability of having at least one variable's 95% credible interval not overlap 0; the right column is average number of such variables. The last two columns repeat this, except with 95% frequentist confidence intervals. Each cell represents the results from 200 simulated datasets.

		Bayes		Frequentist	
N	Factor	$\geqslant 1$ significant	Number significant	$\geqslant 1$	Number
250	0.95	0.54	0.64	0.73	0.97
500	1	0.72	1	0.84	1.27
750	1	0.91	1.32	0.97	1.7
1000	1	0.96	1.59	0.995	1.96

Table 2

Estimated risks of committing Type 1 errors. The columns display the percent of fitted models that have 95% credible intervals or confidence intervals that do not cover zero, which is the true value. Each cell represents the results from 200 simulated datasets.

N	Factor	Non-factor Bayes	Frequentist
250	6.5	25	44
500	3.5	24.5	42
1000	5.5	25	42.5

Table 3

Posterior 90% credible regions for IL-6 SNP/tobacco interactions. Estimates are in grams, for a change in two standard deviations in the latent tobacco scale.

		Dummy 1		Dummy 2	
SNP	Response percentile	5th	95th	5th	95th
1	0.2	-160	113	-187	84
1	0.5	-135	138	-205	58
2	0.2	-243	111	-135	155
2	0.5	-172	240	-121	292
3	0.2	-137	105	-109	131
3	0.5	-104	122	-78	136