Modeling adverse birth outcomes via confirmatory 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 confirmatory factor structure for part of the design matrix. This model is appropriate when the covariates are indicators of scientifically-determined latent factors, and it is these latent factors that analysts seek to include as predictors in the quantile regression. We apply the model to a study of birth weights in which the effects of latent variables representing psychosocial health and actual tobacco usage on the lower quantiles of the response distribution are of interest. The models can be fit using an R package called factorQR.

Key words: Bayesian inference; Gibbs sampling; 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). We present a model for quantile regression for settings where the covariates are indicators of scientifically-determined latent factors, and it is these latent factors (rather than individual indicators) that analysts seek to include as predictors.

This is the setting 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. Many of the explanatory variables are viewed by the scientific team as indicators of underlying factors. For example, 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, to assess the association between psychosocial health and birth outcomes, we may be better served modeling the latent variable of psychosocial health than the variables individually, which may be difficult to interpret due to colinearities and low signal-to-noise ratios.

We take a fully Bayesian approach to model specification, 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). The confirmatory factor model is advantageous when investigators can identify latent factors based on scientific knowledge, because it ensures that factors are interpretable and scientifically meaningful; this may not
be the case when using unstructured factor models or principal components analysis (PCA). Such knowledge is implicit in many studies; for example, investigators often gather groups of related covariates with the express intent of capturing and learning about a latent trait. In such cases, the confirmatory factor analysis effectively reflects the intentions of the study design. Of course, there are settings in which a confirmatory factor structure cannot be meaningfully specified before the analysis (e.g., Bai, 2009). Because this is not the case in our motivating application, we do not focus on such models here.

Using the confirmatory factor model, we find that a latent factor representing tobacco smoke exposure is strongly associated with adverse effects on quantiles of birth weight, whereas a factor representing overall psychosocial health is not. We also present a simulation study illustrating the importance of propagating uncertainty from the factor model through inferences. The Bayesian model we present achieves this, whereas a reasonable non-Bayesian strategy — a two-step model that first estimates the factor scores and then regresses the response on the factor scores — underestimates uncertainty. The methods are available in an R package called \texttt{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 some exposure decreases the birth weights of babies who would otherwise have average to high birth weights by some
modest amount, but it does not decrease the birth weights of babies who would otherwise have low birth weights, it might not be considered deleterious. On the other hand, if an exposure lowers the birth weights of babies – by the same modest amount – who already would have low birth weights, but it does not decrease the birth weights of babies who otherwise would have average to high birth weights, the exposure would be troubling. Thus, for modeling birth weights it is appropriate to consider quantile regression over standard linear regression, which implies that the exposure has the same effect across the entire response distribution.

Although the 1500g and 2500g cutoffs are useful benchmarks, we prefer not to discretize birth weights into very low, low, and normal when modeling. Discretizations have scientifically unjustifiable consequences. For example, it is much worse to be born at 1501g than 2499g, yet a discretization would treat both of these as equivalent. Similarly, being born at 2501g is not appreciably better than being born at 2499g. Quantile regression allows us to model non-central aspects of the birth weight distribution while considering the information that discretization masks (Abrevaya and Dahl, 2008; Dunson, 2010).

Researchers have examined numerous variables to explain these adverse birth outcomes. These include environmental variables like tobacco smoke exposure, air quality, and pesticide exposure (Abrevaya, 2006; 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. Using a confirmatory factor structure partially resolves this issue, in that it enables analysts to pool the information from these multiple, imperfect measurements in hopes of more accurately representing the exposure in the quantile regression.

As a related measurement issue, some individual exposures arguably affect birth outcomes through a common biological 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, activate biological processes that 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 offers analysts a way to represent and estimate such underlying constructs in regression models.

3. The confirmatory factor quantile regression model

We now turn to the Bayesian confirmatory 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 \( i \),

\[
\begin{bmatrix}
    x_{i,\omega} \\
    y_i
\end{bmatrix} = \Lambda \omega_i + \begin{bmatrix}
    0 \\
    x_{i,\beta}' \beta
\end{bmatrix} + \begin{bmatrix}
    \varepsilon_{i,-s} \\
    \varepsilon_{i,s}
\end{bmatrix}.
\]  

(1)

Here, \( \Lambda \) is a \( s \times t \) matrix of factor loadings, \( \omega_i \) is a \( t \times 1 \) vector of latent factors, and \( \beta \) is the vector of regression coefficients for \( x_{i,\beta} \). We assume that \( \varepsilon_{i,-s} \sim \text{normal}(0, \Psi) \) 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 \( AL(0, \tau, p) \), where the parameters are the location, scale and skewness, respectively. We have \( p(\varepsilon_{i,s}) = \tau p(1 - p) \exp\{-\tau \rho_p(\varepsilon_{i,s})\} \), where \( \rho_p \) is the tilted absolute value or check function \( \rho_p(x) = 1\{x > 0\} px + 1\{x < 0\}(1 - p)x \). The asymmetric Laplace distribution is commonly used as the basis of Bayesian quantile regression because \( \Pr(\varepsilon_{i,s} \leq 0) = p \).

This error distribution also connects the Bayesian analysis to standard frequentist quantile regression, which proceeds semiparametrically using \( \rho_p \) as a loss function; we have found these two methods yield similar point estimates. More flexible alternatives include empirical likelihood (Schennach, 2005; Lancaster and Jun, 2010), substitution likelihood (Dunson and Taylor, 2005), and nonparametric mixture (Taddy and Kottas, 2010) approaches.

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
\[ y_i | w_i, \mu, \tau \sim \text{normal}\left( \frac{1 - 2p}{p(1 - p)} w_i + \mu, \frac{2w_i}{\tau p(1 - p)} \right) \]
(2)
and \( w_i | \tau, \mu \) are independent with an exponential distribution with rate \( \tau \), 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 \( \Lambda \). In particular, we assume
\[ \Lambda = \left[ \bigoplus_{j=1}^t \lambda_j \right], \]
(3)
where \( \lambda_j \) are column vectors whose first element is fixed at 1 and \( \oplus \) indicates the direct sum that concatenates the operands into a block diagonal matrix. \( \Lambda_s \) 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 \( \Lambda \) 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 \textit{a priori} which observed variables relate to each latent factor.

Specifying the block diagonal pattern in \( \Lambda_{-s} \) facilitates interpretation of the factors. For instance, in the HPHBS application, \( \lambda_1 \) measures tobacco smoke exposure and \( \lambda_2 \) relates to psychosocial health. The factors have this interpretation because investigator-specified zeros in \( \Lambda_{-s} \) ensure that, for instance, only smoking-related measures are tied to \( \lambda_1 \).

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. Although we do not focus on such situations, one could append a matrix of unstructured factor loadings to the $\Lambda$ matrix proposed in this model, and potentially even learn the number of such extra factors (Lopes and West, 2004).

The joint model of the covariates and outcomes implies a particular distribution of $x$ on $y$, e.g., an environmental exposure conditional on a birth outcome. We do not assign a causal interpretation to such conditional distributions. Rather, we interpret the model as saying that $x_{i,ω}$ and $y_i$ are conditionally independent given the latent factors. Knowing a birth outcome may influence our beliefs about a mother’s unknown exposure to an environmental pollutant (because it may sharpen our beliefs about the latent factor), but this is a result of correlation rather than causation. This interpretation is in line with the rich literature concerning applied Bayesian factor models (e.g., West, 2003; Merl et al., 2009), and a similar interpretation applies to measurement error models (e.g., Berry et al., 2002).

3.1 Prior distributions

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

\[
\begin{align*}
\tau & \sim \text{gamma}(c_\tau, d_\tau) \\
\psi_k & \sim \text{inv-gamma}(c_k, d_k) \\
\beta & \sim \text{normal}(a_\beta, B_\beta^{-1}) \\
\Lambda_s & \sim \text{normal}(a_s, B_s^{-1}) \\
\lambda_k | \psi & \sim \text{normal}(a_{\lambda,k}, \psi_k/h_{0,k}) \\
\Phi & \sim \text{inv-Wishart}(R_0, \nu_0) \\
v_i | \tau & \sim \text{inv-gamma}(1, \tau) \\
\omega_i | \Phi & \sim \text{normal}(0, \Phi)
\end{align*}
\]

In this specification, $v_i$ is the inverse of the mixing weight $w_i$ in (2). 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). Details on fitting the model are given in the online supplement.

We use weakly informative prior distributions in the applied analyses. In particular, for the gamma-distributed components, we use \( c_r = d_r = 1 \) and \( c_k = d_k = 1 \) for all \( k \). This results in a proper but diffuse distribution that lacks a prior first moment for \( \psi_k \) or \( \tau^{-1} \). For the factor loadings of interest, \( \Lambda_s \), the prior distribution is centered at zero with variance 100; \( \beta \) is given the same distribution. For the other factor loadings \( \lambda_k \), we center the prior distributions at one (a neutral value) and specify \( h_{0,k} = 1 \) for all \( k \). We center \( \Phi \) 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 \( \nu_0 \) may speed convergence to regions of high posterior probability.

3.2 Model extensions

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

3.2.1 Dichotomous and ordered categorical factor indicators. 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 variable \( z_i \) 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 our proposed Gibbs sampler.

In the case of a variable with \( R \) ordered categories, this approach easily generalizes to an ordered probit model (Albert and Chib, 1993). We again introduce a unit variance normal latent variable \( z_i \) along with a vector of cutoffs \( \gamma = [\gamma_1, \ldots, \gamma_R]' \), with \( \gamma_1 = -\infty \) and \( \gamma_2 = 0 \)
for identification. Our observable is then \( x_i = \max\{j : \gamma_j < z_i\} \). It is standard to assume the improper uniform prior \( p(\gamma_3, \ldots, \gamma_R) \propto 1\{0 < \gamma_3 < \ldots < \gamma_R\} \).

The ordered probit model also can be useful when one of the factor indicators is a continuous variable but cannot be transformed to approximate normality. In this case, one can discretize the variable, and model it with the ordered probit specification.

3.2.2 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

\[
y_i = \Lambda_s \omega_i + \Lambda_{s,\text{int}} (x_{i,\text{int}} \odot \omega_{\text{int}}) + x_{i,\beta}\beta + \varepsilon_i,
\]

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, \( \omega_{\text{int}} \) is a vector of (perhaps repeated) elements of \( \omega_i \), and \( \odot \) indicates the Hadamard (component-wise) product. Variables in \( x_{i,\text{int}} \) can be repeated in \( x_{i,\beta} \).

3.2.3 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 \( \omega_i \) somewhat, as they are no longer conditionally conjugate. We suggest a Metropolis-Hastings approach in the online supplement.

4. Simulation studies

As an alternative to the Bayesian model, one could utilize the scientifically-determined latent factors via a simpler two-step approach of (i) fitting a frequentist confirmatory factor model (e.g., with the \texttt{sem} package in R), and (ii) performing frequentist quantile regression on the
fitted factor scores (e.g., with the rq function in the quantreg package in R). In this section, we compare the repeated sampling properties of the Bayesian model with those from this simpler approach.

To reduce the computational burden associated with repeated application of the Bayesian estimation, we focus on the one factor \((t = 1)\) model. We use \(\Lambda = [1, 0.5, 0.5, 0.2]'\) and \(\Psi = \text{diag}([1, 1.5, .5])\) for a model of the 20th percentile with unit variance for the latent \(\omega_i\). For sample sizes of 250, 500, 750, and 1000, we repeat the simulation 1000 times. We focus on \(\Lambda_s\), the regression parameter of interest, and compare coverage, bias, and root mean square error (RMSE) for the joint Bayesian model and the two-step model.

Figure 1 presents the results of the simulation studies. Biases are uniformly small for the two methods, and, as one would expect, RMSEs decrease with an increasing sample size. The 95% confidence intervals for the two-step method cover the truth in only around 90% of the runs. In contrast, the estimated coverage rate for the 95% credible intervals for the Bayesian model are close to the nominal 95%. The Bayesian model has better frequentist coverage properties because it properly accounts for uncertainty in the factor structure.

[Figure 1 about here.]

Another alternative is to ignore the scientific knowledge about the factors in the quantile regression model. For example, one could run PCA on the covariates and include the components in the quantile regression. Or, to reduce dimensions, one could employ penalized methods such as lasso and ridge regression. It is difficult to compare the repeated sampling properties of such approaches with those of the Bayesian model in (1), because the true values of the coefficients — and indeed the functional form of the predictors themselves — for each are different. Nonetheless, there are clear qualitative differences between confirmatory factor models and unstructured approaches. For example, consider a lasso regression in which two psychosocial indicators’ coefficients have a sign consistent with scientific expectations, two
are estimated to be zero, and one has an unexpected sign. It is not clear how the analyst should interpret this result when the latent factor is the meaningful quantity. Similarly, there is no guarantee that automatic dimension-reducing techniques like PCA will produce scientifically meaningful factors. Indeed, as we discuss in Section 5, the factors identified by PCA applied to the HPHBS data are not easy to interpret.

5. Analysis of the HPHBS data

We now turn to the analysis of the HPHBS data, where we focus on modeling birth weights. Among the eligible cases (singleton births in Durham, NC), 1009 mothers are non-Hispanic black women and 279 are non-Hispanic white women. In these data, the marginal 10th and 20th percentiles are 2320g and 2670g, respectively. Therefore, exposures that lower the corresponding conditional quantiles can pull birth weights into ranges where health risks increase.

As with many observational studies, there is a modest amount of missing data spread over many variables. We address this through 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 p. 520 of Gelman et al. (2004) and Zhou and Reiter (2010) for justification of this approach.

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]
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. We also include a set of standard controls: 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.

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 perceived stress; 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 $\Lambda_{-q}$ 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 $\Lambda_{-s}$ that are positive and effectively bounded well away from zero. This suggests that the chosen manifest variables capture meaningful latent quantities.

Figures 2 through 4 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 appears 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.

Associations between the birth weights and the tobacco exposure factor can be further displayed with plots of the estimated cumulative distribution function (cdf) of birth weights at different tobacco exposure levels. The estimated cdf is derived using the fitted values from sets of latent factor quantile regressions. Figure 3 displays the estimated distribution as a function of high and low values of the tobacco exposure factor. These distributions are for female babies born to 26-year-old, unmarried, black mothers with an average psychosocial factor score. The figure also indicates that the estimated effect of tobacco exposure is largest around the 50th percentile, with the effect degrading somewhat in the tails of the birth weight distribution.

The estimated loadings for the psychosocial factor are centered firmly on zero, indicating no significant effect at any of the response quantiles. Even though there is no significant effect, using the factor model offers an advantage in interpretation. 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.

Among the standard controls in Figure 4, 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.

For comparison, we also ignored the scientifically-determined confirmatory factor structure and instead used PCA on the covariates that were used as indicators of the factor structure. The first principal component has loading magnitudes between 0.2 and 0.4 for each of the variables except the measure of whether the pregnancy was intended, which has a loading of 0.14. The second and third principle components also load strongly across the functional groupings. Thus, the PCA does not separately identify the tobacco and psychosocial indicators of scientific interest to the investigators. Arguably, it results in factors that are difficult to interpret.

[Table 1 about here.]

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 (Kaplan 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 does 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 1). Although the factor model did not detect a meaningful interaction in these genetic analyses, as well as several others not reported here, the factor model is advantageous 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.

[Figure 5 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

\[
T_p = \frac{1}{n} \sum_{i=1}^{n} I\{y_i \leq \hat{Q}_p(i|\Omega, \beta, \Lambda_s, x_i, \beta)\},
\]

where \( \hat{Q}_p(i|\cdot) \) is the predicted \( p \)th quantile for the \( i \)th 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 \( p \)th 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 5, we can see that \( T_p \) is close to \( p \) across the range of quantiles for the 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 \( x_\beta \) and \( x_\omega \) 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.

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 in contexts where quantile regression on scientifically-determined latent factors is of interest.

Received September 2010; revised March and April 2011.

Supplementary Materials
Web Appendices referenced in Sections 3.1, 3.2 and 5.1 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgements
The authors thank Allison Ashley-Koch, David Dunson, Melanie Garrett, Pamela Maxson, Marie Lynn Miranda, an Editor, an Associate Editor, and an anonymous referee for helpful comments. This research was funded by Environmental Protection Agency grant R833293.

References


 Confirmatory factor quantile regression


Figure 1. Bias, root mean squared error (RMSE), and coverage probabilities of nominally 95% intervals, averaged over 1000 repetitions at sample sizes between 250 and 1000. The solid line is from our proposed model and the dashed line uses a two-step approach that does not propagate uncertainty from the factor model to the response model.
Figure 2. Middle 95%, 90% and 50% confidence bands and median for response factor loadings $\Lambda_s$. 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 3. Fitted quantiles of birth weight for low (solid line) and high (dashed line) tobacco smoke exposure. Here, low and high tobacco exposures are defined to be the 10th and 90th percentiles of the latent tobacco scale. These results are for female babies born to 26 year-old, unmarried, black women, with the psychosocial score set to the mean.
Figure 4. 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 5. 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
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.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Response percentile</th>
<th>Dummy 1</th>
<th>Dummy 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5th</td>
<td>95th</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>-160</td>
<td>113</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>-135</td>
<td>138</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>-243</td>
<td>111</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>-172</td>
<td>240</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>-137</td>
<td>105</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>-104</td>
<td>122</td>
</tr>
</tbody>
</table>