Biostatistics (2014), **0**, 0, *pp.* 1–28
doi:10.1093/biostatistics/samplebibtex

Quantile Regression for Mixed Models

LUKE B. SMITH[1]*, MONTSERRAT FUENTES[1], PENNY GORDON-LARSEN[2], and BRIAN J. REICH[1]

[1]*Department of Statistics, North Carolina State University* Raleigh, North Carolina, 27695-8203, U.S.A.

[2]*Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27516, U.S.A.*

lukebrawleysmith@gmail.com

Summary

Cardiometabolic diseases have substantially increased in China in the past 20 years and blood pressure is a primary modifiable risk factor. Using data from the China Health and Nutrition Survey we examine blood pressure trends in China from 1991 to 2009, with a concentration on age cohorts and urbanicity. Very large values of blood pressure are of interest, so we model the conditional quantile functions of systolic and diastolic blood pressure. This allows the covariate effects in the middle of the distribution to vary from those in the upper tail, the focal point of our analysis. We join the distributions of systolic and diastolic blood pressure using a copula, which permits the relationships between the covariates and the two responses to share information and enables probabilistic statements about systolic and diastolic blood pressure jointly. Our copula maintains the marginal distributions of the group quantile effects while accounting for within-subject dependence, enabling inference at the population and subject levels. We present

*To whom correspondence should be addressed.

a multilevel framework that enables straightforward hypothesis testing for changes in covariate effects across time. Our population level regression effects change across quantile level, year, and blood pressure type, providing a rich environment for inference. To our knowledge, this is the first quantile function model to explicitly model within-subject autocorrelation and is the first quantile function model that accommodates multivariate response. We find that the association between high blood pressure and living in an urban area has evolved from positive to negative, with the strongest changes occurring in the upper tail.

Key words: Bayesian; Blood pressure; Longitudinal; Multivariate; Quantile.

1. Introduction

Globally, cardiovascular disease accounts for approximately 17 million deaths a year, and nearly one third of the total causes of death in 2008 (World Health Organization, 2011). Of these, complications of hypertension account for 9.4 million deaths worldwide every year (Lim *and others*, 2013). Maximum (systolic) blood pressure and minimum (diastolic) blood pressure are physiologically correlated outcomes but are differentially affected by environmental factors (Sesso *and others*, 2000; Franklin *and others*, 2009; Benetos *and others*, 2001; Egan *and others*, 2010; Luepker *and others*, 2012; Choh *and others*, 2011; Chobanian *and others*, 2003). Most studies construct a combined measure using hypertension cutpoints rather than looking across the distribution. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) have differential effects on cardiovascular disease events (Sesso *and others*, 2000; Franklin *and others*, 2009; Benetos *and others*, 2001; Stokes *and others*, 1989), so we model the conditional quantile functions of systolic blood pressure (SBP) and diastolic blood pressure (DBP). This enables inference in the upper tails, the focus of our analysis. We use longitudinal data from the China Health and Nutrition Survey (Popkin *and others*, 2010) to study the impact of urbanicity on those trends.

China provides an outstanding case study given recent and rapid modernization and substantial concomitant environmental change.

Several previous approaches in the longitudinal literature simply ignore the within-subject dependence when estimating the marginal quantile effects. Wang and Zhu (2011) constructed an empirical likelihood under the GEE framework, then adjusted for the dependence in the confidence intervals. For censored data, Wang and Fygenson (2009) ignored the within-subject dependence when estimating the marginal effects and controlled for the within-subject dependence when conducting inference via a rank score test. While these estimators are consistent, ignoring the within-subject dependence for estimation can result in a loss of efficiency and undercoverage.

Another avenue is to introduce dependence via random intercepts, as in Koenker (2004). Waldmann *and others* (2013) and Yue and Rue (2011) assumed asymmetric Laplace errors and included a random subject effect in the location parameter. Presenting separate methodology for marginal and conditional inference, Reich *and others* (2010) accounted for within-cluster dependence via random intercepts and flexibly modeled the density using an infinite mixture of normals. Jung (1996) preserved marginal effects by incorporating correlated errors in a quasi-likelihood model. These models account for within-subject dependence via a location adjustment for each cluster, which may not be sufficiently flexible. Models that incorporate random slopes include Geraci and Bottai (2013), who used numerical integration to average out random effects for marginal inference, and the empirical likelihood of Kim and Yang (2011). The marginal effects of Geraci and Bottai (2013) do not neccesarily maintain their original interpretation after integrating over the random effects. Kim and Yang (2011) permit subject-specific inference for clustered data. While these methods account for dependence, they assume observations within a subject are exchangable. We are interested in inference at the population level of temporally correlated data. Only Jung (1996) incorporates temporally correlated errors within a subject, at the cost of assuming the response is distributed Gaussian.

Collectively these models lack attributes needed for our application. First, we want to conduct inference at multiple quantile levels without assuming our response is distributed Gaussian. The approaches above model one quantile level at a time and can result in "crossing quantiles" (Bondell *and others*, 2010), where for certain values of the predictors the quantile function is decreasing in quantile level. Second, we need to model the autocorrelation within a subject to maintain nominal coverage probabilities. Third, for our application we anticipate that the effect of a covariate on SBP may be similar to its effect on DBP, so we want a bivariate model to facilitate communication across responses.

In this paper we introduce a mixed modeling framework for quantile regression with these necessary attributes. We accomplish these methodological innovations by extending the model of Reich and Smith (2013) to accommodate autocorrelation and multiple responses. In the random component we account for the dependence across time and response via a copula (Nelsen, 1999). This permits the relationships between the covariates and the two responses to share information and enables probabilistic statements about SBP and DBP jointly. Our copula approach maintains the marginal distributions of the group quantile effects while accounting for within-subject dependence, enabling inference at the population and subject levels. Copulas previously utilized in the longitudinal literature (Sun *and others*, 2008; Smith *and others*, 2010) focused on mean inference and do not account for predictors. Copulas have a straightforward connection to quantile function modeling, as both rely on connecting the response to a latent uniformly distributed random variable. Our copula model resembles the usual mixed model (Diggle *and others*, 2002) in that covariates affect both the marginal population distribution via fixed effects and subject specific distributions via random slopes. In the fixed component we allow for different predictor effects across quantile level, response, and year. Our model is centered on the usual Gaussian mixed model and contains it as a special case.

We present a multilevel framework that extends the current Gaussian mixed model to the

quantile regression domain. To our knowledge, this is the first quantile function model for temporally-correlated responses within a subject and the first quantile function model that accommodates a multivariate response. In Section 2 we describe the mixed effect quantile model in the univariate and multivariate cases. In Section 3 we show the results of a simulation study that illustrates the need to account for within-subject dependence in a quantile framework. In Section 4 we analyze hypertension and we conclude in Section 5.

2. Methods

In this section we present our methods for mixed model quantile regression. We first specify the marginal quantile functions in Section 2.1. In Section 2.2 and Section 2.3 we describe different approaches to accommodate within-subject dependence.

2.1 *Marginal Quantile Model*

Denote Y_{ij} as the measurement of SBP on individual $i = 1, 2, ..., N$ at visit $j = 1, 2, ...,$J. This section describes a univariate response, with the multivariate extension in Section 2.3. Let $\mathbf{X}_{ij}$ be a covariate vector of length P for individual i at visit j, and denote the conditional distribution function of Y_{ij} as $F(y|\mathbf{X}_{ij}) = P(Y_{ij} \leqslant y|\mathbf{X}_{ij})$. We specify the distribution of Y_{ij} via its quantile function, defined as $Q(\tau|\mathbf{X}_{ij}) = F^{-1}(\tau|\mathbf{X}_{ij})$, where $\tau \in (0, 1)$ is known as the quantile level. For each response Y_{ij} there exists a latent $U_{ij} \sim$ U(0,1) such that $Y_{ij} = Q(U_{ij}|\mathbf{X}_{ij})$.

We assume Q is a linear combination of covariates, that is,

$$Q(\tau|\mathbf{X}) = \sum_{p=1}^{P} X_p \beta_p(\tau).$$

The regression parameter $\beta_p(\tau)$ is the effect of the p^{th} covariate on Q evaluated at τ. A one-unit increase in X_p is associated with a $\beta(\tau)$ increase in the τ^{th} population quantile. We refer to $\beta(\tau)$ as a "fixed effect", since this effect applies to the full population.

Similar to Reich and Smith (2013) we project β_p onto a space of M $\geqslant 2$ parametric basis functions $I_1(\tau), ..., I_M(\tau)$ defined by a sequence of knots $0 = \kappa_0 < \kappa_1 < ... < \kappa_M < \kappa_{M+1} = 1$. Let $q_0(\tau)$ be the quantile function of a random variable from a parametric location/scale family with location parameter 0 and scale parameter 1. The basis functions are defined as $I_1(\tau) \equiv 1$, $I_2(\tau) = q_0(\tau)\mathbb{1}_{\tau \leqslant \kappa_1} + q_0(\kappa_1)\mathbb{1}_{\tau > \kappa_1}$, and

$$I_m(\tau) = \begin{cases} 0 & \tau \leqslant \kappa_{m-1} \\ q_0(\tau) - q_0(\kappa_{m-1}) & \kappa_{m-1} \leqslant \tau \leqslant \kappa_m \\ q_0(\kappa_m) - q_0(\kappa_{m-1}) & \tau > \kappa_m \end{cases}$$

for $m > 2$. Our model is of the form $\beta_p(\tau) = \sum_{m=1}^{M} I_m(\tau)\theta_{mp}$ and thus

$$Q(\tau|X_{ij}) = \sum_{p=1}^{P} X_{ijp}\beta_p(\tau) = \sum_{p=1}^{P} X_{ijp} \sum_{m=1}^{M} I_m(\tau)\theta_{mp} \tag{2.1}$$

where θ_{mp} are the regression weights.

We set $\mathrm{X}_{ij1} \equiv 1$ for all i and j for the intercept. To achieve a valid quantile function (i.e. increasing in τ) we map all predictors in the interval $[-1, 1]$, and constrain the regression parameters such that $\theta_{m1} > \sum_{p=2}^{P} |\theta_{mp}|$ for $m > 1$. We model θ_{mp} as a function of a Gaussian random variable $\theta^{\star}_{mp}$. The regression parameter θ_{mp} is set to $\theta^{\star}_{mp}$ if the constraint is satisfied and set to

$$\theta_{mp} = \begin{cases} 0.001 & p = 1 \\ 0 & \text{otherwise} \end{cases}$$

if $\theta^{\star}$ is outside of the constraint space. Details are outlined in Reich and Smith (2013). The latent regression variables $\theta^{\star}_{mp}$ are given Gaussian priors with means μ_{mp} and precisions ι^2_{mp}.

Let $\boldsymbol{\theta}_{m.}$ be the collection of regression parameters associated with basis function m. Denote $\Phi(z)$ and $\Phi^{-1}(\tau)$ as the distribution function and quantile function respectively of a standard normal random variable. When the base quantile function is Gaussian (i.e. $q_0(\tau) = \Phi^{-1}(\tau)$), if $M = 2$ or $\boldsymbol{\theta}_{2.} = ... = \boldsymbol{\theta}_{M.}$ this model simplifies to a Gaussian heteroskedastic regression model, where $Q(\tau) = \mathbf{X}'_{ij}\boldsymbol{\theta}_{m1} + \mathbf{X}'_{ij}\boldsymbol{\theta}_{m2}\Phi^{-1}(\tau)$ and thus $Y_{ij}|\mathbf{X}_{ij} \sim \mathrm{N}(\mathbf{X}'_{ij}\boldsymbol{\theta}_{m1}, (\mathbf{X}'_{ij}\boldsymbol{\theta}_{m2})^2)$. Standard Gaussian linear regression is a special case of the heteroskedastic regression model where $M = 2$ and $\theta_{mp} \equiv 0$ for $m > 1$ and $p > 1$ and $Y_{ij}|\mathbf{X}_{ij} \sim \mathrm{N}(\mathbf{X}'_{ij}\boldsymbol{\theta}_{m1}, \theta^2_{12})$.

2.2 *Mixed Effects Quantile Model*

In this section we extend the standard Gaussian mixed effects model to the quantile regression domain. We utilize Gaussian basis functions ($q_0(\tau) = \Phi^{-1}(\tau)$) for connections to standard mixed models. Recall the canonical Gaussian random effects model (Fitzmaurice *and others*, 2012)

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{E}_i \tag{2.2}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{Z}_i$ is a J by R matrix of random effect covariates, $\boldsymbol{\gamma}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Delta})$ is a vector of length R of random effects specific to unit i, and $\mathbf{E}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Lambda})$ are random errors.

We can rewrite (2.2) in three forms. Conditional on the random effects, $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i, \sigma^2\mathbf{I})$. Marginally over the random effects, $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Psi}_i)$, where $\boldsymbol{\Psi}_i = \mathbf{Z}_i\boldsymbol{\Delta}\mathbf{Z}_i' + \boldsymbol{\Lambda}$. Finally, the marginal quantile function form is $Q(\tau|\mathbf{X}_{ij}) = \mathbf{X}_{ij}'\boldsymbol{\beta} + \psi_{ij}\Phi^{-1}(\tau)$, where ψ_{ij} is the j^{th} diagonal element of $\boldsymbol{\Psi}_i$. Therefore, $Y_{ij} = \mathbf{X}_{ij}'\beta + \psi_{ij}\Phi^{-1}(U_{ij})$, where $U_{ij} \sim$ U(0,1) marginally, with dependence between the U_{ij} within the same subject.

We use the third representation to extend mixed models to the quantile domain by viewing the transformed response as a realization from a potentially correlated Gaussian process. To account for the within-subject dependence, we hierarchically model the latent U_{ij} through a Gaussian copula. Our model is

$$\begin{aligned} Y_{ij} &= \sum_{p=1}^{P} X_{ijp}\beta_p(U_{ij}) \\ U_{ij} &= \Phi(W_{ij}/\sqrt{\psi_{ij}}) \\ \mathbf{W}_i &= \mathbf{Z}_i'\boldsymbol{\gamma}_i + \mathbf{E}_i. \end{aligned} \tag{2.3}$$

The fixed regression effects $\beta(\tau)$ are modeled as in Section 2.1. As in (2.2) the random effects $\boldsymbol{\gamma}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Delta})$ and random errors $\mathbf{E}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Lambda})$.

The copula in (2.3) permits structured dependence in the U_{ij}. This preserves the interpretability of population level quantile effects β_p and accounts for within-subject dependence, enabling simultaneous inference at the population and subject levels. This formulation allows predictors

to have a complex relationship with the response. A covariate can have a different effect in the middle of the distribution relative to the tails. This is represented by the fixed component $\mathbf{X}'\boldsymbol{\beta}(\tau)$, the conditional τ^{th} population quantile, with the same interpretation of covariate effects as in Section 2.1. Further, individuals in a population are allowed to respond differently to the same covariate. This is represented by the random component $\mathbf{Z}_i'\boldsymbol{\gamma}$. A one unit increase in Z_{ijr} is associated with a $\gamma_{ir}/\sqrt{\psi_{ij}}$ increase in the Z-score of individual Y_{ij}.

For the CHNS data we anticipate that between-individual variability is strong, which can be estimated through a random intercept inside the copula. Covariates that change across time (e.g. urbanicity) can be used to further capture within-subject variability. For longitudinal data we anticipate serial within-subject correlation, so we model $\boldsymbol{\Lambda} = \mathbf{I} + \lambda\boldsymbol{\Xi}(\alpha)$ as the sum of an identity matrix and a scaled (by positive λ) autoregressive order-1 (AR-1) correlation matrix $\boldsymbol{\Xi}(\alpha)$, where $\boldsymbol{\Xi}(\alpha)[u,v] = \alpha^{-|u-v|}$ with correlation parameter α. The scaling factor λ determines the proportion of variability determined by the temporal signal.

While we have thus far defined our model in terms of Gaussian basis functions, any of the parametric bases described in Reich and Smith (2013) can be utilized to model effects at the population level. Finally, the standard Gaussian mixed model is a special case of (2.3) where $q_0(\tau) = \Phi^{-1}(\tau)$, $M = 2$ and $\theta_{mp} \equiv 0$ for $m > 1$ and $p > 1$. This allows us to center our flexible model on the popular model.

2.3 *Multivariate Mixed Effects Quantile Model*

Here we extend (2.3) to the multivariate domain. We are not concerned with trying to define a multivariate quantile (Chakraborty, 2003). Instead, we want to conduct simultaneous inference on multiple responses. Denote Y_{1ij} and Y_{2ij} as the measurements of SBP and DBP on individual i at time j. We specify different quantile effects for each response (i.e. $\beta_{1p}(\tau_1)$ and $\beta_{2p}(\tau_2)$ for covariate p). Our bivariate model then accounts for dependence between the parameters in these

quantile processes, and in the residual copula model.

Our multivariate mixed quantile model is

$$\begin{aligned} Y_{hij} &= \sum_{p=1}^{P} X_{ijp} \sum_{m=1}^{M} I_m(U_{hij})\theta_{hmp} \\ U_{hij} &= \mathbf{\Phi}(W_{hij}/\sqrt{\psi_{hij}}) \\ \mathbf{W}_i &= \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{E}_i, \end{aligned} \tag{2.4}$$

where now $H = 2$ is the dimension of the response, $\mathbf{W}_i$ is of length JH, and the covariance of $\mathbf{E}_i$ is $\mathbf{\Xi}(\alpha)\otimes\mathbf{\Lambda}+\mathbf{I}$ where $\mathbf{\Xi}(\alpha)[u,v] = \alpha^{-|u-v|}$ with correlation parameter α and $\mathbf{\Lambda}$ is an unstructured $H \times H$ correlation matrix.

This formulation allows the uniform random variables U_{ij} to be interpreted as the individual's percentile relative to the population. That is, an individual may be at the conditional 70^{th} percentile ($U_{1ij} = 0.70$) for SBP and the 75^{th} percentile ($U_{2ij} = 0.75$) for DBP, and the similarity in these percentiles can be exploited.

To borrow strength across the responses we model $(\theta_{1mp}, \theta_{2mp})' \sim BVN(\mu_{mp}\mathbf{1}, \iota^2_{mp}\mathbf{I}_2)$. By shrinking regression effects to a common location, we are able to borrow information across SBP/DBP to estimate covariate effects. This multivariate framework enables statements about joint effects of a predictor, and allows for probabilistic estimates regarding both responses (e.g. the conditional probability an individual has blood pressure higher than 140/90).

We assign μ_{mp} independent normal priors with mean μ_{0mp} and precision ι^2_{0mp}. We give ι_{mp} independent Gamma(a_{mp}, b_{mp}) priors. We designate $\mathbf{\Lambda}$ an inverse Wishart prior with scale matrix $\mathbf{\Lambda}_0$ and ν_0 degrees of freedom. For our application we assign $\mathbf{\Delta}$ a diagonal matrix structure with diagonal elements $\delta_{hr} \overset{\text{iid}}{\sim}$ Gamma(1,1), $h = 1, 2, ..., H, r = 1, 2, ..., R$. In applications with more observations per subject and correlation on the within-subject regression coefficients is easier to detect, more complicated structures for $\mathbf{\Delta}$ could be useful. We use the Metropolis within Gibbs algorithm to sample from the posterior, with details in the Supplementary Materials.

3. Simulation Study

We conducted a simulation study to examine the effect of within-subject dependence on parameter estimation. To construct univariate, auto-correlated responses we generated dependent J-dimensional realizations $W_i \sim N(0, \boldsymbol{\Psi_i})$, where $\boldsymbol{\Psi}_i = \mathbf{Z}_i \boldsymbol{\Delta} \mathbf{Z}_i' + \boldsymbol{\Xi}(\alpha) + \mathbf{I}$ with j^{th} diagonal element ψ_{ij}. The design matrix $\mathbf{Z}_i$ contains an intercept and one continuous predictor $X_{1ij} \overset{\text{iid}}{\sim}$ U(-1,1).

The first factor we examine in the simulation study is the strength of the within-subject dependence. We look at three levels, 0.0,0.5,0.9, of the temporal correlation parameter α, which correspond to no, moderate and strong within-subject dependence, respectively. Our second factor is the strength of the dependence determined by the covariance of the within-subject random effects, $\boldsymbol{\Delta} = \Delta \mathbf{I}_2$. In one setting the variance $\Delta = 0$, corresponding to the coefficients having no effect on the dependence. In the other is a diagonal matrix with nonzero values of $\Delta = 3$, corresponding to roughly 60% of the variance within a subject being explained by covariates.

Given these correlated responses we perform the probability integral transform $U_{ij} = \Phi(W_{ij}/\sqrt{\psi_{ij}})$. The third factor in our study is the marginal distribution given these uniform random variables. The response data are

$$Y_{ij} = 3\Phi^{-1}(U_{ij}) + (X_{1ij} + X_{2ij})(0.5 - U_{ij}) * \mathbb{1}_{U_{ij}<0.5} \tag{1}$$

$$Y_{ij} = (3 + X_{1ij} + X_{2ij}) Q_t(U_{ij}) \tag{2}$$

where Q_t is the quantile function of the student t-distribution with 5 degrees of freedom, $i = 1, 2, ..., N$ individuals, and $j = 1, 2, ..., J$ visits. The covariate X_{2i} is binary with equal probability of -1 and 1 and is constant over time. Design (2) is a heteroskedastic linear model, but design (1) is more complicated, with nonzero effects for only half of the distribution. We generated data at $J = 7$ timepoints for N = 50 and N = 100 individuals. For each level of our design we ran 100 Monte Carlo replications.

We fit our model with and without a copula, and compared our model to the marginal model of Reich *and others* (2010) using 25 approximation terms, denoted "RBW". RBW fits a random intercept but ignores temporal correlation within a subject. We fit 2,3 and 5 basis functions and two different parametric bases (Gaussian and t). Results presented were selected by having the highest log psuedo marginal likelihood (Ibrahim *and others*, 2005). For data type (1) LPML most commonly selected 5 Gaussian basis functions for the independent model without a copula and 5 t-distributed basis functions for the copula model. For data type (2) LPML most commonly selected 2 Student t basis functions for both models. Prior means for $\theta^{\star}_{mp}$ were 1 for $p = 1$ and 0 otherwise, and prior variances were 10. For the copula model we set (scalar) $\Lambda_0 = 1$ and $\nu_0 = 3$, corresponding to a prior mean of 4 and infinite variance for $\mathbf{\Lambda}$. For the Student-t basis functions we gave the shape parameter a normal prior on the log scale with mean log(10) and variance log(10)/2. Averages of coverage probability (CP) of 95% intervals and mean squared error (MSE) of each model evaluated at the quantile levels $\tau = .1, .3, .5, .7, .9$ for the $N = 50$ case are shown in Table 1. All of the conclusions listed below similarly held for the $N = 100$ case, shown in the supplementary material (http://www.biostatistics.oxfordjournals.org).

When observations within a subject are independent ($\Delta = 0, \alpha = 0$ case), all models attain the nominal 95% coverage probability for both predictors and data types, except RBW for data type (1). Fitting a copula to independent data seems to have little effect on marginal inference. Increasing α causes undercoverage in the independent model for the continuous predictor X_1. In contrast, the copula and RBW models maintain proper coverage. As within-subject dependence increases, each observation contributes less information about the marginal distribution. This can be seen by the increases in MSE due to increases in α. We compare MSE across the estimators when the covariates do not affect within-subject dependence (i.e. $\Delta = 0$). The copula model is better than RBW with respect to MSE for data type (1). RBW assumes the heteroskedastic model, as in data type (2), yet none of the three models are statistically significantly better with

respect to MSE.

The results change when the subject-level regression coefficients affect dependence (i.e. $\Delta = 3$). RBW and the independent model suffer from poor coverage when the predictors account for dependence in the response. In contrast, the copula model maintains close to nominal coverage. Further, the copula model dominates RBW and the independent model with respect to MSE. The copula model has a statistically significantly lower MSE in roughly half of the cases and is lowest in all cases. In summary, accounting for covariates in the dependence can reduce MSE and preserve coverage.

4. CHNS Analysis

4.1 *Data*

The China Health and Nutrition Survey (CHNS) was designed in 1986 to gauge a range of economic, sociological, demographic and health questions (Popkin *and others*, 2010). The CHNS is a large scale household-based survey drawing from 228 communities which were cluster sampled from 9 provinces. Community structures include villages, townships, urban neighborhoods, and suburban neighborhoods. The communities sampled are designed to be economically and demographically representative of China. Procedures for collecting the data are described in Adair *and others* (2014). We use data collected in 7 waves, starting in 1991 and ending in 2009. We focus on the Shandong province, located in central China, where hypertension rates are elevated (Batis *and others*, 2013). We utilize the urbanicity index of Jones-Smith and Popkin (2010). Rather than dichotomizing communities into urban/rural groups, for each wave Jones-Smith and Popkin (2010) assigned 0-10 scores for each of 12 factors, including population density, economic activity, traditional markets, modern markets, transportation, infrastructure, sanitation, communications, housing, education, diversity, health infrastructure, and social services. Jones-Smith and Popkin (2010) used factor analysis to confirm these factors represent one latent construct.

We have two scientific goals for these data. First, we want to estimate the role of urbanicity in these trends. Second, we want to examine blood pressure trends over time across different age cohorts. We bin individuals into six age groups: age < 18, $18 \leqslant$ age < 30, $30 \leqslant$ age < 40, $40 \leqslant$ age < 50, $50 \leqslant$ age < 60 and age $\geqslant 60$. For individuals with age < 18, blood pressure is very correlated with height and age, rendering uninterpretable comparisons across children and adults. For this reason, most studies focus on children or adults, and we focus on adults in this paper.

The plots in Figure 1 show slight increases over time in blood pressure across both genders and large increases over time in urbanicity. We construct an urbanicity by age interaction effect to look for associations between urbanicity and different age cohorts. As in Attard *and others* (2014) we stratify our analyses by gender. Other covariates include current smoking status (men only due to low female rates) and current pregnancy status (female only). To look for changes across time we include temporal linear trends for all predictors.

Another challenge presented by these data is confounding due to blood pressure medication. Medication artificially supresses blood pressure values. For individuals on medication we ignore the measured values and assume only that they have high blood pressure. Using the method of Reich and Smith (2013) we treat these values as right-censored above the thresholds for high blood pressure, located at 140 for SBP and 90 for DBP.

Our final sample consisted of 1421 females missing 56% of blood pressure measurements and 1248 males missing 55% of blood pressure measurements. Missing household income in year j was imputed using the community average for year j. Missing smoking status was imputed using the value from the previous sampling wave, and assumed to be a nonsmoker in the first wave if missing. Missing pregnancy status was assumed to not be pregnant. With a large number of predictors and so many missing observations, allowing all 14 predictors to change with quantile level is not feasible. In our analysis we have urbanicity change with quantile level and all other

effects be constant with quantile level, that is, we fix $\beta_p(\tau) \equiv \beta_p = \theta_1$ for all τ by setting $\theta_2 = ... = \theta_m = 0$. The interpretations for these effects are equivalent to those in mean regression in that they are allowed to affect the location but not the shape of the response distribution.

We linearly transformed the responses to have mean 0 and standard deviation 1. We assigned $\mu_{m1} \overset{\text{iid}}{\sim}$ N(1,1) priors for the intercept process and $m > 1$ and $\mu_{mp} \overset{\text{iid}}{\sim}$ N(0,1) priors for all other regression parameters. We set $\Lambda \sim$ IW(10, $7\mathbf{\Lambda}_0$) where $\mathbf{\Lambda}_0$ is an $H \times H$ correlation matrix with off-diagonal elements of 0. This corresponds to a prior mean of 1 and variance of 0.4 for the diagonal elements of Λ. This centers the prior distributions of SBP and DBP on a mean zero, unit variance normal distribution that is independent across SBP and DBP. We assigned $\iota^2_{mp} \overset{\text{iid}}{\sim}$ G(1,1) priors. We assigned α a uniform prior on the unit interval. We ran our models for 40,000 MCMC iterations, the first half of which were discarded.

4.2 *Analysis*

We fit 3 different models to compare dependence structures. In model 1 we fit our model without a copula, assuming independence across sampling wave and response. We also fit two copula models. In model 2 the covariance of $\mathbf{W}_i$ is $\mathbf{\Xi}(\alpha) \otimes \mathbf{\Lambda} + \mathbf{I}_{JH}$, where the non-diagonal component is the Kronecker product of an AR-1 correlation matrix and an unstructured 2×2 covariance matrix $\mathbf{\Lambda}$. In model 3 we fit a mean component consisting of a random intercept and an urbanicity effect with the same covariance as model 2, that is, $\mathbf{W}_i = \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{E}_i$. Finally, we fit SBP and DBP jointly and singly for all copula models. For each model we fit M = 2 and M = 4 basis functions. The runs with 4 basis functions had convergence issues, probably due to the large number of missing observations, so we present results for the M = 2 case.

LPML values were -32060, -17681, and -34282 for females for models 1,2, and 3 (-27683, -15576, and -29310 for males). The large values for model 2 indicate strong within-subject correlation that is captured in the covariance. Including subject-specific random effects leads to overfitting.

Figure 2 illustrates the urbanicity random effect $\boldsymbol{\gamma}_{i2}$ on systolic blood pressure across female subjects. These effects are not statistically significant. Nonzero slope effects combined with missing observations can lead to estimating many extreme quantile levels for the first and last sampling waves. In applications with fewer missing observations or more timepoints, random slope effects could be useful.

The posterior means of the off-diagonal elements of the correlation between responses were 0.94 and 0.95 for females and males respectively, with posterior standard deviations around 0.01. Therefore, SBP and DBP at one timepoint within an individual are very strongly correlated, and the posterior distribution of the correlation effects is dominated by the data. The posterior means of the temporal correlation parameter α were 0.12 and 0.10 for females and males respectively, with posterior standard deviations around 0.02. For the univariate fits of blood pressure the posterior means of α ranged from 0.70 to 0.82 with posterior standard deviations around 0.02. The multivariate and temporal correlation seem to be fighting for the same signal. This strong correlation within an individual across response and time is useful when imputing missing values.

For females, the average of the posterior variance of the regression effects at the median $\beta_p(0.5)$ was 3.59 for the multivariate copula model and 4.24 for the multivariate independent model. This 13% increase in posterior variance (5% for males) suggests the independent model may be susceptible to undercoverage. For females, the mean of the posterior variance of the regression effects at the median was 3.62 for the univariate copula model. This 2% decrease in posterior variance for the multivariate model (1% for males) suggests that covariate effects are similar across SBP and DBP. In applications where multivariate observations within an individual were less correlated, we would expect a larger reduction in posterior variance of the effects. In summary, the copula models account for the within-subject dependence and are less susceptible to undercoverage than models that assume independent replications within an individual. The multivariate quantile approach reduces posterior variance by modeling SBP and DBP jointly.

Posterior plots of the intercept process $\beta_1(\tau)$ for the age 40-50 cohort and population urbanicity effects are shown in Figure 3. The intercept process represents the values of our baseline age 40-50 cohort when all predictors are zero, which is a central value after transformation to $[-1, 1]$ for all covariates. For the intercept process, the light 2009 regions differ statistically from the dark 1991 regions for males in the lower tails of SBP and DBP. In contrast to the intercept process, the urbanicity effects change qualitatively from the first to the last sampling wave. In 1991 urban areas had higher blood pressure in the upper tail and lower blood pressure in the lower tail. In 2009 the urbanicity effect is negative or zero for SBP for all quantile levels. In contrast, urban areas are now associated with lower DBP in the upper tail of the distribution.

Estimated location effects are presented in Table 2. Several general associations are apparent. Blood pressure increases with age. There is not strong evidence for an interaction effect between age and urbanicity, indicating that urbanicity has similar effects across age groups. The covariates household income, pregnancy and smoking status have little effect.

5. Discussion

In this paper we have presented novel methods for analysis using mixed models in a quantile regression framework. We conducted a simulation study that illustrates the utility of estimating dependence for quantile regresssion in a longitudinal setting. In our analysis of China Health and Nutrition Survey data in Shandong province, we find strong evidence for within-subject dependence and a gain in information by modeling SBP and DBP jointly. We found that urbanicity is now associated with lower rather than higher blood pressure, especially in the upper tails of the distribution.

We present methods for a Gaussian copula. Gaussian copulas assume independence in the deep tails and assume the same dependence in the lower tail as the upper tail. In this paper we focus on the quantile levels from 0.1 to 0.9. In practice if inference at more extreme quantile levels

is of interest, other copulas should be considered. A more complex copula could be useful in future applications. Another extension is a fully nonparametric approach to the quantile function.

6. Software

Software in the form of R code in the package BSquare is available on the Comprehensive R Archive Network. China Health and Nutrition Survey Data can be found at http://www.cpc.unc.edu/projects/china.

7. Supplementary Material

The results for the $N = 100$ arm are in the supplementary material (http://www.biostatistics.oxfordjournals.org).

7.1 *MCMC*

All prior parameters are independent across basis functions and predictors. Polynomial parameters are also independent. That is, the intercept effect over time is independent of slope effect across time. Let $\theta^{\star}_{dmp}$ be the vector of length $H = 2$ corresponding to the regression coefficients for SBP and DBP for the d^{th} polynomial term, m^{th} basis function and p^{th} predictor. In our model $\theta^{\star}_{dmp} \sim N(\mu_{dmp}, \iota^2_{dmp})$, where ι^2_{dmp} is a precision. Below we suppress the subscripts for d, m, and p. We sample from the posterior distribution for $\theta^{\star}_{dmp}$ using random walk Metropolis, as the posterior distribution does not have a closed form. In the burn in period posterior variances are tuned to have a 30-40% acceptance ratio.

We assign μ a Gaussian prior with mean μ_0 and precision ι^2_0. Conditional on the other pa-

rameters the posterior distribution of μ is of the form

$$\begin{aligned}
[\mu|rest] &\propto \exp\left\{-0.5\iota^2[\boldsymbol{\theta}^\star - \mathbf{1}\mu]'[\boldsymbol{\theta}^\star - \mathbf{1}\mu]\right\}\exp\left\{-0.5(\iota_0^2)[\mu-\mu_0]^2\right\} \\
&\propto \exp\left\{-0.5[\mu^2\left(\iota^2\mathbf{1}'\mathbf{1}+\iota_0^2\right) - 2\mu\left(\iota^2\mathbf{1}'\theta+\iota_0^2\mu_0\right)]\right\} \\
&= \exp\left\{-0.5\left[\mu'\Omega\mu - 2\mu'\omega\right]\right\} \\
&= \exp\left\{-0.5\left[\mu'\Omega\mu - 2\mu'\Omega(\Omega^{-1}\omega)\right]\right\}
\end{aligned}$$

which is the kernel of a normal random variable with mean $\Omega^{-1}\omega$ and variance Ω^{-1} where $\Omega = H\iota^2 + \iota_0^2$ and $\omega = \iota^2\mathbf{1}'\theta + \iota_0^2\mu_0$.

We assign ι^2 a gamma(a,b) prior. Let $\mathbf{x} = \boldsymbol{\theta}^\star - \mathbf{1}\mu$. Conditional on the other parameters the posterior distribution of ι^2 is of the form

$$\begin{aligned}
[\iota^2|rest] &\propto \iota^{H/2}\exp\left\{-0.5\iota^2\mathbf{x}'\mathbf{x}\right\}(\iota^2)^{(a-1)}\exp\left\{-\iota^2/b\right\} \\
&\propto (\iota^2)^{(H/2+a-1)}\exp\left\{-\iota^2\left(0.5\mathbf{x}'\mathbf{x}+1/b\right)\right\} \\
&= (\iota^2)^{(H/2+a-1)}\exp\left\{-\iota^2\left((0.5b\mathbf{x}'\mathbf{x}+1)/b\right)\right\}
\end{aligned}$$

which is the kernel of a gamma random variable with shape $H/2+a$ and scale $(2b/(b\mathbf{x}'\mathbf{x}+2))$.

7.2 *Copula Parameters*

For MCMC we model $\Phi^{-1}(\mathbf{U}_i) = \mathbf{W}_i = \mathbf{D}_i\left[\mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\eta_i} + \mathbf{E}_i\right]$, where $\boldsymbol{\eta_i} \sim \mathrm{N}(0, \boldsymbol{\Xi}(\alpha)\otimes\boldsymbol{\Lambda})$. The regression coefficients $\boldsymbol{\gamma}_i$ have posterior

$$\begin{aligned}
[\boldsymbol{\gamma}_i|rest] &\propto \exp\left\{-0.5[\boldsymbol{W}_i - \boldsymbol{D}_i\left(\boldsymbol{Z}_i\boldsymbol{\gamma}_i+\boldsymbol{\eta}_i\right)]'\boldsymbol{D}_i^{-2}[\boldsymbol{W}_i - \boldsymbol{D}_i\left(\boldsymbol{Z}_i\boldsymbol{\gamma}_i+\boldsymbol{\eta}_i\right)]\right\}\exp\left\{-0.5\boldsymbol{\gamma}_i'\boldsymbol{\Delta}^{-1}\boldsymbol{\gamma}_i\right\} \\
&\propto \exp\left\{-0.5\boldsymbol{\gamma}_i'\left(\boldsymbol{Z}_i'\boldsymbol{D}_i\boldsymbol{D}_i^{-1}\boldsymbol{D}_i^{-1}\boldsymbol{D}_i\boldsymbol{Z}_i + \boldsymbol{\Delta}^{-1}\right)\boldsymbol{\gamma}_i + \boldsymbol{\gamma}_i\boldsymbol{Z}_i'\boldsymbol{D}_i\boldsymbol{D}_i^{-2}\left(\boldsymbol{W}_i - \boldsymbol{D}_i\boldsymbol{\eta}_i\right)\right\} \\
&= \exp\left\{-0.5\boldsymbol{\gamma}_i'\left(\mathbf{Z}_i'\mathbf{Z}_i + \boldsymbol{\Delta}^{-1}\right)\boldsymbol{\gamma}_i + \boldsymbol{\gamma}_i\boldsymbol{Z}_i'\left(\mathbf{D}_i^{-1}\mathbf{W}_i - \boldsymbol{\eta}_i\right)\right\}
\end{aligned}$$

which is the kernel of a multivariate normal random variable with mean $\Omega^{-1}\omega$ and variance Ω^{-1} where $\Omega = \boldsymbol{Z}_i'\boldsymbol{Z}_i + \boldsymbol{\Delta}^{-1}$ and $\omega = \boldsymbol{Z}_i'\left(\mathbf{D}_i^{-1}\mathbf{W}_i - \boldsymbol{\eta}_i\right)$.

The regression coefficients $\boldsymbol{\eta}_i$ have posterior

$$\begin{aligned}
[\boldsymbol{\eta}_i|rest] &\propto \exp\left\{-0.5[\boldsymbol{W}_i - \boldsymbol{D}_i\left(\boldsymbol{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\eta}_i\right)]'\boldsymbol{D}_i^{-2}[\boldsymbol{W}_i - \boldsymbol{D}_i\left(\boldsymbol{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\eta}_i\right)]\right\} \\
&\quad * \exp\left\{-0.5\boldsymbol{\eta}_i'\boldsymbol{\Psi}^{-1}\boldsymbol{\eta}_i\right\} \\
&\quad \propto \exp\left\{-0.5\boldsymbol{\eta}_i'\left(\boldsymbol{D}_i\boldsymbol{D}_i^{-1}\boldsymbol{D}_i^{-1}\boldsymbol{D}_i + \boldsymbol{\Psi}^{-1}\right)\boldsymbol{\eta}_i + \boldsymbol{\eta}_i'\boldsymbol{D}_i\boldsymbol{D}_i^{-2}\left(\boldsymbol{W}_i - \boldsymbol{D}_i\boldsymbol{Z}_i\boldsymbol{\gamma}_i\right)\right\} \\
&\quad = \exp\left\{-0.5\boldsymbol{\eta}_i'\left(\mathbf{I} + \boldsymbol{\Psi}^{-1}\right)\boldsymbol{\eta}_i + \boldsymbol{\eta}_i'\left(\boldsymbol{D}_i^{-1}\boldsymbol{W}_i - \boldsymbol{Z}_i\boldsymbol{\gamma}_i\right)\right\}
\end{aligned}$$

which is the kernel of a multivariate normal random variable with mean $\Omega^{-1}\omega$ and variance Ω^{-1} where $\Omega = \mathbf{I} + \boldsymbol{\Psi}^{-1}$ and $\omega = \boldsymbol{D}_i^{-1}\boldsymbol{W}_i - \boldsymbol{Z}_i\boldsymbol{\gamma}_i$.

The elements of $\boldsymbol{\Delta}$ and $\boldsymbol{\Lambda}$ are singly updated using random walk Metropolis. The correlation parameter α is sampled using independent Metropolis updates.

Acknowledgments

The authors are grateful for support from NSF grant DMS-1107046, NIH grant SR01ES014843, as well as NHLBI (R01-HL108427). The China Health and Nutrition Survey is funded by NICHD (R01-HD30880), although no direct support was received from grant for this analysis. We also are grateful to the Carolina Population Center (R24 HD050924) for general support. *Conflict of Interest*: None declared.

References

Adair, LS, Gordon-Larsen, P, Du, SF, Zhang, B and Popkin, BM. (2014). The emergence of cardiometabolic disease risk in chinese children and adults: consequences of changes in diet, physical activity and obesity. *Obesity Reviews* **15**(S1), 49–59.

Attard, Samantha M., Herring, Amy H., Zhang, Bing, Du, Shufa, Popkin, Barry M.

and Gordon-Larsen, Penny. (2014). Differential changes in systolic and diastolic blood pressure and urbanization: a 20-year multilevel analysis in modernizing china. *Submitted* **1**(1), 1 – 105.

Batis, Carolina, Gordon-Larsen, Penny, Cole, Stephen R, Du, Shufa, Zhang, Bing and Popkin, Barry. (2013). Sodium intake from various time frames and incident hypertension among chinese adults. *Epidemiology* **24**(3), 410–418.

Benetos, Athanase, Thomas, Frederique, Safar, Michel E, Bean, Kathryn E and Guize, Louis. (2001). Should diastolic and systolic blood pressure be considered for cardiovascular risk evaluation: a study in middle-aged men and women. *Journal of the American College of Cardiology* **37**(1), 163–168.

Bondell, Howard D, Reich, Brian J and Wang, Huixia. (2010). Noncrossing quantile regression curve estimation. *Biometrika* **97**(4), 825–838.

Chakraborty, Biman. (2003). On multivariate quantile regression. *Journal of statistical planning and inference* **110**(1), 109–132.

Chobanian, Aram V, Bakris, George L, Black, Henry R, Cushman, William C, Green, Lee A, Izzo Jr, Joseph L, Jones, Daniel W, Materson, Barry J, Oparil, Suzanne, Wright Jr, Jackson T *and others*. (2003). The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: the jnc 7 report. *Jama* **289**(19), 2560–2571.

Choh, Audrey C, Nahhas, Ramzi W, Lee, Miryoung, Choi, Youn Su, Chumlea, William C, Duren, Dana L, Sherwood, Richard J, Towne, Bradford, Siervogel, Roger M, Demerath, Ellen W *and others*. (2011). Secular trends in blood pressure during early-to-middle adulthood: the fels longitudinal study. *Journal of hypertension* **29**(5), 838.

Diggle, Peter, Heagerty, Patrick, Liang, Kung-Yee and Zeger, Scott. (2002). *Analysis of longitudinal data*. Oxford University Press.

Egan, Brent M, Zhao, Yumin and Axon, R Neal. (2010). Us trends in prevalence, awareness, treatment, and control of hypertension, 1988-2008. *Jama* **303**(20), 2043–2050.

Fitzmaurice, Garrett M, Laird, Nan M and Ware, James H. (2012). *Applied longitudinal analysis*, Volume 998. John Wiley & Sons.

Franklin, Stanley S, Lopez, Victor A, Wong, Nathan D, Mitchell, Gary F, Larson, Martin G, Vasan, Ramachandran S and Levy, Daniel. (2009). Single versus combined blood pressure components and risk for cardiovascular disease the framingham heart study. *Circulation* **119**(2), 243–250.

Geraci, Marco and Bottai, Matteo. (2013). Linear quantile mixed models. *Statistics and Computing*, 1–19.

Ibrahim, Joseph G, Chen, Ming-Hui and Sinha, Debajyoti. (2005). *Bayesian survival analysis*. Wiley Online Library.

Jones-Smith, Jessica C and Popkin, Barry M. (2010). Understanding community context and adult health changes in china: development of an urbanicity scale. *Social Science & Medicine* **71**(8), 1436–1446.

Jung, Sin-Ho. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association* **91**(433), 251–257.

Kim, MiOk and Yang, Yunwen. (2011). Semiparametric approach to a random effects quantile regression model. *Journal of the American Statistical Association* **106**(496), 1405–1417.

Koenker, Roger. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91**(1), 74–89.

Lim, Stephen S, Vos, Theo, Flaxman, Abraham D, Danaei, Goodarz, Shibuya, Kenji, Adair-Rohani, Heather, AlMazroa, Mohammad A, Amann, Markus, Anderson, H Ross, Andrews, Kathryn G *and others*. (2013). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet* **380**(9859), 2224–2260.

Luepker, Russell V, Steffen, Lyn M, Jacobs, David R, Zhou, Xia and Blackburn, Henry. (2012). Trends in blood pressure and hypertension detection, treatment and control 1980-2009: The minnesota heart survey. *Circulation*, CIRCULATIONAHA–112.

Nelsen, Roger B. (1999). *An introduction to copulas*. Springer.

Popkin, Barry M, Du, Shufa, Zhai, Fengying and Zhang, Bing. (2010). Cohort profile: The china health and nutrition surveymonitoring and understanding socio-economic and health change in china, 1989–2011. *International journal of epidemiology* **39**(6), 1435–1440.

Reich, Brian J, Bondell, Howard D and Wang, Huixia J. (2010*a*). Flexible bayesian quantile regression for independent and clustered data. *Biostatistics* **11**(2), 337–352.

Reich, Brian J, Bondell, Howard D and Wang, Huixia J. (2010*b*). Flexible bayesian quantile regression for independent and clustered data. *Biostatistics* **11**(2), 337–352.

Reich, Brian J and Smith, Luke B. (2013). Bayesian quantile regression for censored data. *Biometrics* **69**(3), 651–660.

Sesso, Howard D, Stampfer, Meir J, Rosner, Bernard, Hennekens, Charles H, Gaziano, J Michael, Manson, JoAnn E and Glynn, Robert J. (2000). Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in men. *Hypertension* **36**(5), 801–807.

Smith, Michael S, Min, Aleksey, Almeida, Carlos and Czado, Claudia. (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association* **105**(492), 1467–1479.

Stokes, J 3rd, Kannel, William B, Wolf, Philip A, D'Agostino, Ralph B and Cupples, L Adrienne. (1989). Blood pressure as a risk factor for cardiovascular disease. the framingham study–30 years of follow-up. *Hypertension* **13**(5 Suppl), I13.

Sun, Jiafeng, Frees, Edward W and Rosenberg, Marjorie A. (2008). Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics* **42**(2), 817–830.

Waldmann, Elisabeth, Kneib, Thomas, Yue, Yu Ryan, Lang, Stefan and Flexeder, Claudia. (2013). Bayesian semiparametric additive quantile regression. *Statistical Modelling* **13**(3), 223–252.

Wang, Huixia Judy and Fygenson, Mendel. (2009). Inference for censored quantile regression models in longitudinal studies. *The Annals of Statistics*, 756–781.

Wang, Huixia Judy and Zhu, Zhongyi. (2011). Empirical likelihood for quantile regression models with longitudinal data. *Journal of Statistical Planning and Inference* **141**(4), 1603 – 1615.

World Health Organization, Geneva. (2011). Causes of death 2008: data sources and methods.

Yue, Yu Ryan and Rue, Håvard. (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis* **55**(1), 84–96.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]

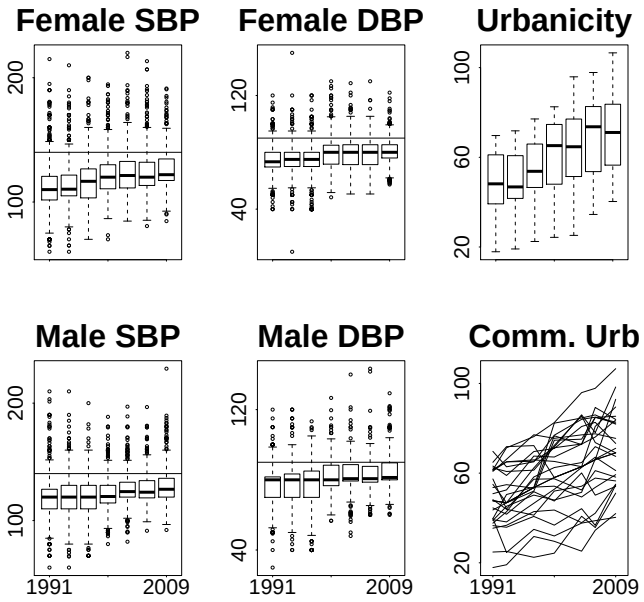

Fig. 1. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) by gender and urbanicity scores across time. Blood pressure measurements are in millimeters of mercury (mmHg). Urbanicity is a composite score measuring 12 features of the community environment (Jones-Smith and Popkin, 2010). Horizontal lines represent thresholds for high blood pressure, located at 140 mmHg and 90 mmHg for systolic and diastolic blood pressure respectively.

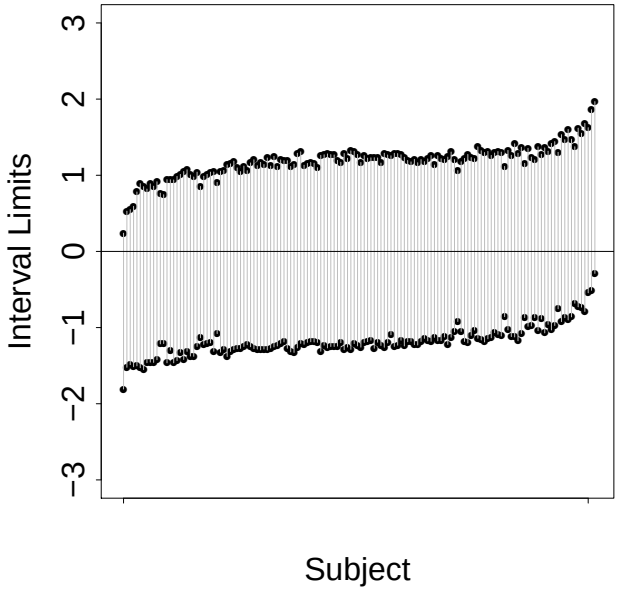

Fig. 2. Plots of posterior credible sets of urbanicity random effects on females for systolic blood pressure. For visual clarity posterior credible sets are ordered by posterior median and every 10th subject is shown.

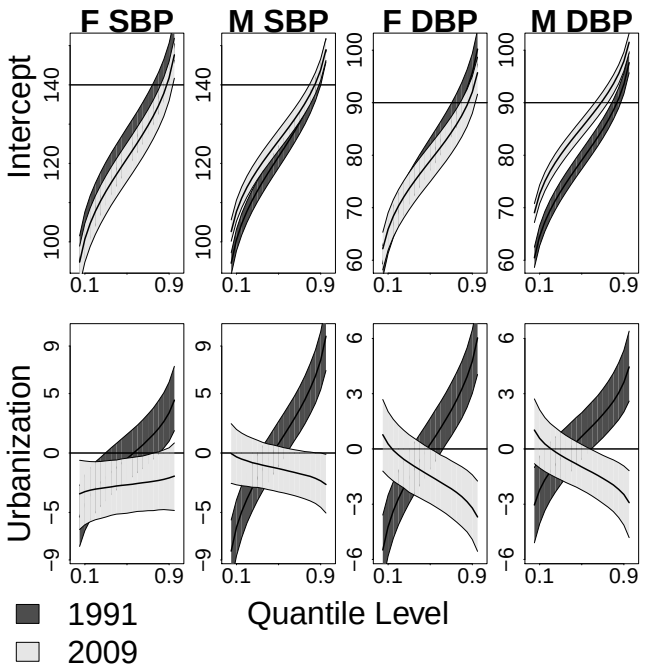

Fig. 3. Plots of the intercept process for the age 40-50 cohort and population urbanicity effects by gender and blood pressure type. The intercept process is the distribution of the response when all covariates are 0. The urbanicity plots are the effects of a one standard deviation increase in urbanicity on the τ^{th} quantile of blood pressure. Dark regions correspond to 1991 estimates, while light regions correspond to 2009 estimates.

Table 1. Coverage probability (CP) and mean squared error (MSE) for the N = 50 arm of the simulation study. Nominal coverage probability is 95%. We compare treating the data as independent within a subject ("Ind"), fitting with a copula ("Cop"), and the random effects model of (Reich *and others*, 2010) ("RBW"). Coverage and MSE were evaluated at and averaged over the quantile levels $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. For datatype = 1, MSE values are less than depicted values by a factor of 10. Estimators whose MSE were statistically significantly different than the copula model are indicated by *.

$\Delta = 0$, Datatype = 1												
	Coverage						MSE					
	$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 0.9$		$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 0.9$	
	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2
Ind	0.91	0.95	0.88	0.94	0.73	0.94	0.05	0.09	0.07	0.10	0.11	0.11
Cop	0.95	0.96	0.95	0.97	0.94	0.96	0.05	0.10	0.06	0.10	0.09	0.10
RBW	0.83	0.90	0.86	0.89	0.88	0.87	0.11*	0.16*	0.13*	0.15*	0.16*	0.14*
$\Delta = 0$, Datatype = 2												
	Coverage						MSE					
	$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 0.9$		$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 0.9$	
	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2
Ind	0.94	0.97	0.89	0.94	0.76	0.95	0.06	0.09	0.09	0.11	0.15	0.13
Cop	0.98	0.98	0.95	0.98	0.93	0.97	0.07	0.11	0.10	0.13	0.16	0.13
RBW	0.96	0.98	0.95	0.96	0.93	0.95	0.07	0.10	0.10	0.11	0.14	0.10
$\Delta = 3$, Datatype = 1												
	Coverage						MSE					
	$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 0.9$		$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 0.9$	
	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2
Ind	0.61	0.76	0.58	0.78	0.56	0.72	0.17	0.24	0.19	0.25*	0.23	0.24*
Cop	0.92	0.91	0.91	0.90	0.89	0.91	0.14	0.18	0.15	0.17	0.18	0.17
RBW	0.85	0.70	0.85	0.69	0.86	0.67	0.23*	0.26*	0.24	0.26*	0.25	0.24*
$\Delta = 3$, Datatype = 2												
	Coverage						MSE					
	$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 0.9$		$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 0.9$	
	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2
Ind	0.64	0.76	0.60	0.78	0.57	0.74	0.26	0.26	0.28	0.25	0.32*	0.29*
Cop	0.90	0.90	0.89	0.89	0.91	0.92	0.21	0.21	0.20	0.21	0.19	0.17
RBW	0.86	0.80	0.84	0.79	0.83	0.77	0.26	0.21	0.26	0.22	0.31*	0.24

Table 2. Posterior parameter estimates and 95% credible intervals for location effects. Mean effects include age cohort (with baseline group aged 40-50), urbanicity by age cohort interaction (indicated by U *), household income (HHI), current pregnancy status, and smoking.

Female Effects								
Predictor	SBP 1991		SBP 2009		DBP 1991		DBP 2009	
$18 < \text{Age} < 30$	-5.5	(-6.9,-3.9)	-4.2	(-5.9,-2.7)	-2.8	(-3.7,-2.0)	-3.2	(-4.3,-2.2)
$30 < \text{Age} < 40$	-4.5	(-5.9,-3.3)	-1.5	(-2.8,-0.0)	-2.6	(-3.5,-1.8)	-0.2	(-1.1,0.7)
$50 < \text{Age} < 60$	2.7	(1.0,4.3)	3.2	(1.8,4.6)	1.3	(0.1,2.3)	1.4	(0.5,2.3)
$60 < \text{Age}$	7.7	(6.0,9.6)	5.5	(4.1,6.9)	3.1	(2.0,4.4)	1.9	(0.9,2.7)
U * $(18 < \text{Age} < 30)$	-0.9	(-5.7,3.6)	1.1	(-4.1,7.1)	1.2	(-2.0,4.2)	3.6	(0.3,7.4)
U * $(30 < \text{Age} < 40)$	-4.5	(-10.1,1.5)	0.8	(-4.1,6.2)	-0.8	(-4.4,3.1)	-0.0	(-3.0,3.1)
U * $(50 < \text{Age} < 60)$	4.2	(-2.8,10.9)	0.4	(-4.9,5.5)	2.2	(-2.5,7.0)	2.0	(-1.2,5.1)
U * $(60 < \text{Age})$	2.7	(-3.2,8.2)	6.6	(1.6,11.4)	1.9	(-2.1,5.6)	1.3	(-1.7,4.7)
HHI	4.0	(-0.7,8.2)	-2.0	(-3.6,-0.1)	2.2	(-0.6,4.8)	-0.8	(-1.9,0.4)
Pregnant	-1.1	(-4.5,2.9)	2.0	(-1.8,4.8)	-1.0	(-2.8,1.6)	-0.3	(-3.2,1.9)
Male Effects								
Predictor	SBP 1991		SBP 2009		DBP 1991		DBP 2009	
$18 < \text{Age} < 30$	-1.9	(-3.6,-0.5)	-4.0	(-5.7,-2.5)	-1.6	(-2.5,-0.5)	-3.3	(-4.2,-2.4)
$30 < \text{Age} < 40$	-1.3	(-3.0,0.0)	-2.1	(-3.3,-0.7)	-1.0	(-2.0,0.1)	-1.2	(-2.1,-0.4)
$50 < \text{Age} < 60$	4.7	(3.4,6.5)	0.8	(-0.4,2.1)	2.1	(1.1,3.3)	0.1	(-0.7,0.9)
$60 < \text{Age}$	8.3	(6.2,10.1)	3.2	(1.9,4.5)	3.6	(2.4,4.8)	0.3	(-0.5,1.2)
U * $(18 < \text{Age} < 30)$	-0.2	(-5.8,4.7)	-0.5	(-5.8,4.9)	-1.6	(-5.2,1.8)	3.1	(-0.4,6.6)
U * $(30 < \text{Age} < 40)$	-5.2	(-10.1,-0.6)	-1.2	(-6.1,3.5)	-4.0	(-7.6,-0.5)	1.5	(-1.7,4.4)
U * $(50 < \text{Age} < 60)$	5.7	(-0.1,11.5)	0.6	(-4.5,6.4)	0.9	(-2.9,5.3)	0.9	(-2.6,4.4)
U * $(60 < \text{Age})$	7.0	(1.6,13.6)	2.4	(-2.5,7.4)	1.0	(-2.9,5.1)	-0.0	(-3.2,3.1)
HHI	2.6	(-0.5,5.3)	-1.8	(-3.4,-0.2)	0.9	(-1.0,3.0)	-0.7	(-1.8,0.3)
Smoke	-0.1	(-0.9,0.9)	0.5	(-0.2,1.4)	-0.1	(-0.7,0.5)	0.0	(-0.5,0.5)