Exploratory quantile regression with many covariates: An application to adverse birth outcomes

Lane F. Burgette[1], Marie Lynn Miranda[2], and Jerome P. Reiter[1]

[1]Department of Statistical Science, Duke University
[2]Nicholas School of the Environment and Department of Pediatrics, Duke University

October 26, 2010

Abstract

In health sciences research, inference about the conditional mean of a continuous response may not be of primary interest; rather, the effects of covariates on the tails of the response distribution are typically most relevant. Data analysts often use quantile regression to estimate these effects. Specifying which predictors to include in the model can be difficult in large epidemiological studies with extensive available variables, especially when interaction effects are suspected. We introduce two approaches for exploring high-dimensional predictor spaces to identify important predictors in quantile regression settings. These are based on ideas from lasso regression and the elastic net. We apply the approaches to data collected as part of the Healthy Pregnancy, Healthy Baby Study. This is an observational prospective cohort study of adverse birth outcomes in Durham, NC, that includes a wide array of demographic, medical, psychosocial, and environmental variables. The data suggest an interesting interaction effect not previously reported: the lower tails of the birth weight distribution tend to be decreased for mothers who both smoke during pregnancy and have high levels of lead in their blood have lower birth weights than mothers with either risk factor alone.

1 Introduction

In large epidemiological cohort studies, researchers often record data on dozens or even hundreds of variables that are potentially relevant covariates for regression models. When interaction effects are suspected, the number of potentially relevant predictors can exceed the sample size. This creates problems for standard regression modeling approaches, which generally require more observations than predictors. Furthermore, the analyst must specify the predictors to include in the regression model, which can be difficult with a high dimensional covariate space.

Various solutions exist in the literature for handling this problem, including, for example, stepwise regression and its variants,[1] regression trees,[2] random forests,[3] and penalized regression approaches.[4,5] These approaches can be used as exploratory tools to identify sets of important predictors for further study.

Most of these techniques are designed for estimating conditional means of the response variable. Often, however, the effects of covariates on the upper or lower tails of the response distribution are most relevant. These effects can be estimated using quantile regression.[6] For example, quantile regression has been used in epidemiological studies to explore the effect of tobacco use on sleep patterns,[7] the effects of early childhood lead exposure on educational outcomes,[8] and to quantify the degradation of mental acuity in multiple sclerosis patients.[9] While quantile regression enables analysts to focus on percentiles of distributions rather than means, analysts still must decide which covariates and interactions to include in the model.

In this article, we introduce two approaches for exploring the most important predictors in quantile regression in high dimensional settings. The first estimates the regression coefficients subject to the lasso penalty.[4] As we explain in Section 2, this penalty forces many estimated coefficients to equal zero, leaving the most important predictors to have non-zero coefficients. The second implements a quantile regression

version of the elastic net.[10] The elastic net is similar to the lasso in that it penalizes large models, but is more effective at identifying groups of important predictors that are highly correlated with each other.

We apply the two approaches to data from the Healthy Pregnancy, Healthy Baby study, an ongoing, observational prospective cohort study in Durham, NC, focused on the etiology of poor birth outcomes. These adverse outcomes, including low birth weight and preterm birth, have been linked to many problems[11] including blindness,[12] deafness,[13] and behavioral problems.[14] Unfortunately, the causes of adverse birth outcomes are not well understood, although variables cited as being important predictors include smoking,[15] lead exposure,[16] environmental exposures more generally,[11] and psychological stress.[17,18,19]

We focus on non-Hispanic black mothers, resulting at this time in a sample size of 881 births. We seek to build models for birth weight and gestational age, with particular emphasis on low quantiles of these distributions, e.g., the 10th, 20th and 30th percentiles of birth weights and gestational ages. The data comprise 35 covariates, including maternal demographics like age, education and income; maternal medical history variables like measures of hypertension and previous birth outcomes; maternal environmental exposures to cadmium, nicotine, cotinine, mercury, and lead as measured in blood; maternal psychosocial factors like scores on the Interpersonal Support Evaluation List, the CES-D depression scale, the NEO Five Factor Personality Inventory, perceived racism, and availability of social support. We also believe that interactions among these variables may be important predictors of birth outcome quantiles. With main effects and interactions, we have over 600 covariates under consideration.

All aspects of the Healthy Pregnancy, Healthy Baby study, including the analyses presented here, were conducted according to a research protocol approved by

Duke University's Institutional Review Board. The Healthy Pregnancy, Healthy Baby study is embedded within the Southern Center on Environmentally Driven Disparities in Birth Outcomes (SCEDDBO).

2 Methods

Standard linear regressions are of the form

$$y_i = x_i^\top \beta + \varepsilon_i \qquad \text{for} \qquad i = 1, \ldots, n \tag{1}$$

where ε_i are independent errors with mean zero, β is a vector of regression coefficients with length p, and $x_i^\top$ is a row vector of covariates for the ith individual. The linear regression model implies that the distribution of y shifts in its mean but not in its shape for different x_i.

Quantile regression allows for both the mean and shape of the distribution of y to change with x, without assuming a particular error distribution. Hence, it is more flexible than linear regression. Quantile regression uses the same basic model as (1), but assumes that ε_i are independent errors whose τth quantile is equal to zero, i.e., $\Pr(\varepsilon_i \leq 0) = \tau$. As a result, $x_i^\top \beta$ is interpreted as the conditional τth quantile of y given x_i.

As originally presented,[6] quantile regression does not specify a formal probability model. Instead, coefficients are estimated by minimizing the empirical loss function

$$L(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta), \tag{2}$$

where

$$\rho_\tau(a) = \begin{cases} \tau|a| & \text{if } a \geq 0 \\ (1-\tau)|a| & \text{if } a < 0. \end{cases} \tag{3}$$

Techniques exist for estimating standard errors of the estimated coefficients, as well as for testing hypotheses and constructing interval estimates. Koenker and Hallock[20]provide several case studies involving quantile regression, including an application to birth weight data.

Quantile regression has advantages for modeling birth outcomes over other regression approaches. It allows covariates to have different effects on tail values than on the middle of the distribution, whereas linear regression assumes a constant effect across the outcome distribution. It avoids the problems from dichotomizing birth outcomes into low and high values. For example, using a logistic regression on an indicator for low birth weight (less than 2500g) treats a birth at 2499g as fundamentally different from a birth at 2501g; it also treats a birth at 2499g as the same as one at 2000g. Neither of these treatments is scientifically or clinically defensible.

We now present the approaches for exploring the important predictors in quantile regression in high dimensional covariate spaces. We emphasize that these are exploratory techniques and do not provide formal inference. Analysts can use the results from these exploratory techniques to identify useful models.

2.1 Boosted lasso for quantile regression

Most model selection techniques operate by penalizing large models. For example, in standard linear regression, it is popular to select the model with the minimum value of the Akaike Information Criterion (AIC)[21] or the minimum value of the Bayesian Information Criterion (BIC).[22] For AIC, adding an additional predictor penalizes the log likelihood-based criterion value by two; for BIC, the penalty is

increased $\log(n)$. Thus, as the number of predictors p in the model increases, the penalty increases as well.[23]

The lasso method uses a penalty related to the sum of the absolute values of the regression coefficients.[4] Using simplified notation, lasso solutions for β minimize the function

$$\Gamma(\beta; \lambda_1) = L(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j|, \tag{4}$$

for an empirical loss function $L(\beta)$. For large values of λ_1, many components of β are estimated to equal zero. As λ_1 shrinks toward zero, the estimates of β move towards an unpenalized estimate. Equivalently, one can solve (4) by minimizing $L(\beta)$ subject to the restriction $\sum |\beta_j| < t_1$ for a value of t_1 that corresponds to λ_1.[4] We follow the convention of displaying lasso solutions using this representation. Typically, a single value of λ_1 is selected using cross-validation approaches.

Before fitting a lasso model (or using the elastic net), the analyst should transform the covariates to have mean zero and variance one so that the penalties exert their force equally on all components of β. Further, it is standard to work with a centered version of the response so as to penalize the intercept minimally. The analyst also should add any potentially relevant interactions among the covariates to x_i.

To adapt the lasso to exploratory quantile regression, we set $L(\beta)$ to be the loss function in (2). Rather than use one specific t_1, we investigate the solutions of (4) for a wide range of t_1, beginning with zero and ending at a large value. We find the lasso solutions via a boosting algorithm;[24] see the online supplement for details of the algorithm.

The solution paths, i.e., plots of estimated β versus t_1, can be used to identify important variables in the quantile regression. In particular, we search for groups of variables whose estimates take on non-zero values early in the solution path. To

illustrate with simulated data, consider the right panel of Figure 1, which shows the lasso solution paths for a median regression with ten variables, five of which have non-zero true coefficients. The estimates corresponding to four variables quickly take on non-zero values as we move from left to right on the plot, suggesting that these are important variables. The other estimated coefficients take much longer to rise above zero, suggesting that they are not as important. Hence, if seeking to fit a parsimonious median regression, we would select the first four variables as predictors as a starting point for quantile regression modeling, adding other predictors based on scientific considerations.

It is possible to use cross-validation to select one value of t_1, considering the fitted model at that point in the solution path to be a final model. As a general technique, cross-validation involves randomly breaking the data into, say, ten groups of roughly equal size. For each group, we fit the lasso quantile regression on the other 90% of records, predict the outcomes for the 10% of the records in the group, and repeat the process to make out-of-sample predictions for each of the ten groups. We measure predictive errors using the quantile loss function (2), choosing the t_1 that yields the smallest average predictive loss.

For our exploratory analyses, however, we do not pursue cross-validation. We do not aim to use statistical considerations alone in order to select a single "best" model. Rather, we prefer to use the penalized regression as a ranking process in terms of groups of variables. We then combine scientific considerations with the information gleaned from the exploratory fits in the formal model building process. Even though epidemiological theory may not dictate exact models, it can give us significant guidance.

Although this illustrative example had a modest number of predictors, the lasso works in settings with more predictors than observations. We simply follow the

Figure 1: Lasso and elastic net solution paths for a quantile regression of the 50th percentile of simulated data. Three of the covariates are very strongly correlated and have an associated true regression parameter of 0.2 (solid lines). The other covariates are less strongly correlated with each other. Two of these have a true β value of 0.3 (dashed lines) and the rest have a true value of zero (dotted lines).

solution path until the penalty is weak enough that our estimate becomes equal to the unpenalized estimate, at which point the algorithm stops. Indeed, the lasso penalization has been applied in genetic studies where there are vastly more covariates than observations.[25,26] With a large number of predictors, we typically are interested in models that have significantly fewer non-zero β elements than observations, so we can stop the algorithm early in order to save computational time.

2.2 Boosted elastic net for quantile regression

As an exploratory tool, lasso can be sub-optimal when there are highly correlated groups of predictors, as it will tend to grant only one of them a non-zero β coefficient. In such settings, the elastic net[10] can be used identify groups of important predictors. This penalizes the empirical loss function by adding a term of $\lambda_2 \sum \beta^2$ to the lasso criterion. Thus, we seek the value of β that minimizes

$$\Gamma_{\text{EN}}(\beta; \lambda_1, \lambda_2) = L(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{5}$$

for given λ_1 and λ_2. As with the lasso, large values of λ_1 encourage many zeros in the solution for β, and small values of λ_1 do not. On the other hand, when λ_2 is large relative to λ_1, many components of β are encouraged to take on small but non-zero values, since the penalty disfavors large coefficients. Values of λ_1 and λ_2 are usually set by cross-validation.

To see why the elastic net penalty encourages groups of correlated predictors to

enter the model simultaneously relative to the lasso penalty, consider two covariates X_1 and X_2 that are nearly identical for all subjects. The lasso penalty cannot easily distinguish between a solution with $(\beta_1, \beta_2) = (0, \phi)$ and $(\beta_1, \beta_2) = (\phi/2, \phi/2)$ for some $\phi > 0$, so that the solution may present only one non-zero predictor. The elastic net penalty with $\lambda_2 > 0$ prefers $(\phi/2, \phi/2)$, so that both predictors take on non-zero values at similar points in the solution path.

To adapt elastic net to exploratory quantile regression, we set $L(\beta)$ to be the loss function in (2). Rather than select one value of (λ_1, λ_2), we investigate the solutions for a range of values to identify important predictors. To facilitate the exploration, we replace λ_2 with $\lambda_1\lambda_2^*$, where λ_2^* is a positive constant. For a fixed λ_2^*, we fit the solution path as λ_1 shrinks toward zero. As with the lasso solutions, we look for groups of coefficients that move away from zero early and fast in the solution path. We again use boosting to fit the solution paths given λ_2^*; see the online supplement.

We examine the solution paths for several values of λ_2^*, searching for a value of λ_2^* large enough to allow more variables to enter the model early in the solution path (compared to the lasso fit), but small enough that some parsimony is maintained early in the solution path. Appropriate values of λ_2^* depend on the data. We recommend starting with $\lambda_2^* = 1$ and moving up or down by factors of five or ten until these goals are achieved. It may be helpful to inspect intermediate values of λ_2^*, but our experience is that the fits are not extremely sensitive to λ_2^*, so that examining a few values typically suffices.

To illustrate, consider Figure 2, which shows the fits for λ_2^* between zero and 1000. The fit with $\lambda_2^* = 1$ is nearly unchanged from the lasso ($\lambda_2^* = 0$) fit. When λ_2^* is 10 or 100, a fifth variable enters the model early, thus appearing as an important predictor. When we increase λ_2^* to 500 or 1000, it becomes difficult to discern exactly which variables are likely to be most important since most of them take on non-

zero values early in the solution path; we argue that these fits are not useful for an exploratory analysis. Thus, in this simulated dataset, fits with λ_2^* equal to 10 or 100 satisfy our goals.

Figure 2: Detail of elastic net solution paths for a quantile regression of the 50th percentile of simulated data. Using the same data as Figure 1, we fit the elastic net with $\lambda_2^* \in (0, 1, 10, 100, 500, 1000)$, where $\lambda_2^* = 0$ is equivalent to the lasso.

2.3 A framework for penalized exploratory quantile regression

The lasso and elastic net quantile regressions can be combined to improve exploratory analyses for quantile regression. We recommend that analysts take the following steps.

First, the analyst fits the lasso regression solution path at the quantile of interest. The analyst isolates variables with coefficients that take on non-zero values early in the solution path. These variables are the starting point for formal model building. The analyst can rank the importance of the remaining variables in the order in which their corresponding coefficients take on non-zero values. The analyst may choose to include some of these additional variables in the formal model based on scientific considerations and formal hypothesis testing.

Second, as a check on the lasso results, the analyst fits the elastic net at several values of λ_2^*, as outlined in Section 2.2. Here, the analyst is looking in particular for variables that appear early in the chosen elastic net fit, but are not prominent in the lasso. These variables can be considered for inclusion based on scientific merits and formal hypothesis tests. The elastic net is especially useful in settings where interaction effects are suspected, since interaction variables are often collinear with main effects terms, and thus may be excluded by the lasso.

This two step process is an exploratory framework designed to highlight promi-

nent groups of predictors. Like all exploratory model building techniques, it should supplement, and not replace, scientific rationales for building the regression models. Analysts should add or subtract variables in accordance with such rationales. Nonetheless, in high dimensional settings with many predictors, exploratory quantile regression tools can help analysts find important predictors that might otherwise be difficult to uncover, as we now illustrate on some real rather than simulated data.

3 Results

Using the 881 births in the Healthy Pregnancy, Healthy Baby (HPHB) study in Durham, NC, we seek quantile regressions for the 10th, 20th and 30th percentiles of babies' weights and gestational ages at birth. We include in x the main effects of all predictors (using indicator variables for multi-category covariates), squared terms for continuous predictors, and all two-way interactions. Summary statistics for the covariates are in Table 1 and a histogram of birth weights is provided in the online supplement.

With lasso quantile regressions, we find that measures of tobacco smoke exposure take on primary importance, particularly for birth weight. For instance, when modeling the 10th percentile of birth weight, four of the nine first covariates to enter the model are interactions that include a tobacco measure (see Figure 3); at the 20th percentile, four of the first seven are tobacco-related; at the 30th percentile, five of the first six are interactions involving tobacco. These results are consistent with the literature that extensively documents the impact of smoking on birthweight. Across the quantiles, a lead/tobacco interaction is consistently flagged.

Other important variables from the lasso 10th percentile regression for birth weight include the mother's age selected as a squared term and in an interaction with environmental tobacco smoke exposure. The former suggests that the associ-

ation between age and birth weight is a U-shaped curve, which is again consistent with the literature.[27] In addition, the ISEL appraisal score, which is a measure of the availability of someone with whom to discuss problems, was selected in an interaction with a variable indicating perceived social standing and an interaction with having a "visiting" relationship with the father of the child. At the 20th percentile, prominent variables (other than those involving tobacco) include an interaction between negative paternal support and NEO-conscientiousness (a measure of organization and persistence) and an interaction between negative paternal support and the perceived self-efficacy score, which measures a woman's belief that she can steer her own life's trajectory.[28] Qualitatively, the results are similar at the 30th percentile of birth weight. Smoking measures are prominent, and environmental measures other than a tobacco/lead interaction are slow to enter the model.

In the presented results, we focus on birth weight as the response because lasso fits of gestational age reveal fewer useful explanatory variables, with less consistency across the response quantiles. This may be evidence that the determinants of gestational age are not well-captured by our covariates.

Figure 3: Lasso and elastic net solution paths for a quantile regression of the 10th percentile of birth weight with blood pressure measures removed. Tobacco-related variables and interactions including tobacco-related variables are solid lines; all others are dashed. The elastic net fit uses $\lambda_2^* = 0.05$.

Moving to the elastic net, we find that the fits are essentially identical to the lasso fit when $\lambda_2^* = 0.01$. When $\lambda_2^* = 0.1$, dozens of variables take on non-zero values very early in the fitting process, which makes identifying the important variables difficult. Hence, we focus on the $\lambda_2^* = 0.05$ elastic net. This fit generally selects the same variables as lasso, but it also identifies other potentially important covariates: interactions between blood lead levels and mother's self-reported tobacco use at the start of the pregnancy, and between blood lead levels and mother's tobacco use

during the pregnancy, are among the early variables with non-zero coefficients.

Using the results of the exploratory analyses, we turn to estimating regressions for the 10th, 20th and 30th percentiles of birth weight. Although not selected in the quantile regressions, we include some standard controls known to be predictors of birth weight, including sex of the baby and an indicator for first birth.[27,29] Since the mother's age appeared important as a squared term, we include age and $(\text{age})^2$. We also consider the tobacco use at prenatal intake/lead interaction that was the first variable to appear in our 10th percentile lasso and elastic net fits, and add in the main effects to comply with the hierarchy principle. We also fit models including parental relationship status, ISEL appraisal scores, and their interaction; however, statistically insignificant F-tests of these effects suggested that they could be removed from the models without much sacrifice. We similarly remove the maternal age/tobacco interaction after an F-test.

The results are summarized in Table 2. While age does not approach significance in any of the quantiles in Table 1, age squared is consistently negative and significant. Because age squared is centered, this suggests a U-shaped effect of maternal age on birth weight; i.e., women at the bottom and top end of the age distribution will tend to have lower birth weights. It is not until the 30th percentile that infant sex nearly reaches significance, and parity (first or higher order birth) never reaches significance in the quantiles presented. Since infant sex and parity are well-established predictors of birth weight in traditional estimates of the mean regression analysis, we might reasonably conclude that at lower quantiles, other processes are overwhelming the effects that dominate at the mean.

The coefficients on the environmental variables (the continuous blood lead measurement, tobacco use at prenatal care intake, and their interaction) are especially interesting. The combined main and interaction effects of lead and tobacco are es-

sentially a wash at the 10th percentile. At the 20th percentile, women who both use tobacco and who have lead exposure that is one standard deviation above the mean have a net decrease in birth weight of 174 grams. Note, however, that at the 10th and 20th percentiles, the coefficient estimates are not significant. At the 30th percentile, the combined main and interaction effects are associated with a 168 decrement in birth weight. Note that here the interaction effect is significant, and the two main effects approach significance.

4 Conclusion

With health outcomes like birth weight, we are often interested in the effects of various covariates across the distribution, or may be especially interested in explaining what is happening on the tails of the distribution, rather than at the mean. Here we presented a penalized quantile regression framework that allows us to explore the lower aspects of the birth weight distribution with the many available covariates and their interactions.

The exploratory quantile regression analyses presented in this paper revealed several important characteristics of our data. We note that the effect of standard covariates like infant sex and maternal parity do not exert much influence on the lower quantiles of the distribution — in contrast to traditional estimates of the mean, which demonstrate them to be important explanatory variables. With a larger dataset, we might construct narrower sets of quantiles and perhaps tease out when variables transition from influence to non-influence, and *vice versa*.

The coefficients on the environmental variables (blood lead, tobacco use, and their interaction) are especially interesting. The recent meta-analysis of Navas *et al.*[30] has found that "evidence is sufficient to infer a causal relationship between lead exposure and high blood pressure" (p. 478). At the same time, we know

that hypertension is associated with poorer birth outcomes.[31] We also know the anomalous but persistent result that smoking during pregnancy reduces the risk for preeclampsia[32,33,34] (although among preeclamptics, smoking increases the risk for perinatal mortality and fetal growth restriction). Both preeclampsia and maternal hypertension are associated with lower birth weights. Because hypertension can potentially be either the cause of the result of problems in pregnancy (especially if it progresses to preeclampsia), this exploratory analysis suggests that a nuanced treatment of the lead-hypertension-smoking nexus may be critical to understanding adverse birth outcomes. Such modeling is part of our ongoing research agenda.

The exploratory quantile regression framework proposed in this paper represents a relatively quick and simple method for exploring non-central features of the response distribution. Other penalized quantile regression methods have been proposed,[35,36] but they can lead to significant computational challenges. Boosting, however, is a simple and flexible method for fitting these models. As quantile regression continues to gain popularity, we feel that such exploratory techniques will be an important component of the epidemiologist's toolbox.

References

1. R.B. Bendel and A.A. Afifi. Comparison of stopping rules in forward stepwise regression. *Journal of the American Statistical Association*, 72(357):46–53, 1977.
2. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, FL, 1984.
3. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
4. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
5. A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthog-

onal problems. *Technometrics*, 42(1):80–86, 2000.

6. R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.
7. L. Zhang, J. Samet, B. Caffo, and N.M. Punjabi. Cigarette smoking and nocturnal sleep architecture. *American Journal of Epidemiology*, 164(6):529–537, 2006.
8. M.L. Miranda, D. Kim, J. Reiter, M.A. Overstreet Galeano, and P. Maxson. Environmental contributors to the achievement gap. *Neurotoxicology*, 30(6):1019–1024, 2009.
9. R.A. Marrie, N.V. Dawson, and A. Garland. Quantile regression and restricted cubic splines are useful for exploring relationships between continuous variables. *Journal of Clinical Epidemiology*, 62(5):511–517, 2009.
10. H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.
11. M.L. Miranda, P. Maxson, and S. Edwards. Environmental contributions to disparities in pregnancy outcomes. *Epidemiologic Reviews*, 31(1):67–83, 2009.
12. B.J. Crofts, R. King, and A. Johnson. The contribution of low birth weight to severe vision loss in a geographically defined population. *British Journal of Ophthalmology*, 82(1):9–13, 1998.
13. J.M. Lorenz, D.E. Wooliever, J.R. Jetton, and N. Paneth. A quantitative review of mortality and developmental disability in extremely premature newborns. *Archives of Pediatrics and Adolescent Medicine*, 152(5):425–435, 1998.
14. G. Ross, E.G. Lipper, and P.A.M. Auld. Social competence and behavior problems in premature children at school age. *Pediatrics*, 86(3):391–397, 1990.
15. D.A. Savitz, N. Dole, J.W. Terry Jr, H. Zhou, and J.M. Thorp Jr. Smoking and pregnancy outcome among African-American and white women in central

North Carolina. *Epidemiology*, 12(6):636642, 2001.

16. T. Gonzalez-Cossio, K.E. Peterson, L.H. Sanin, E. Fishbein, E. Palazuelos, A. Aro, M. Hernandez-Avila, and H. Hu. Decrease in birth weight in relation to maternal bone-lead burden. *Pediatrics*, 100(5):856–862, 1997.
17. R.W. Newton and L.P. Hunt. Psychosocial stress in pregnancy and its relation to low birth weight. *British Medical Journal*, 288(6425):1191–1194, 1984.
18. S.T. Orr, J.P. Reiter, D.G. Blazer, and S.A. James. Maternal prenatal pregnancy-related anxiety and spontaneous preterm birth in Baltimore, Maryland. *Psychosomatic Medicine*, 69:566–570, 2007.
19. S.T. Orr, D.G. Blazer, S.A. James, and J.P. Reiter. Depressive symptoms and indicators of maternal health status during pregnancy. *Journal of Womens Health*, 16:535–542, 2007.
20. R. Koenker and K.F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, 2001.
21. H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, eds. B.N. Petrox and F. Caski, Budapest: Academiai Kiado. 267–281, 1973.
22. G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.
23. K.P. Burnham and D.R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, 2004.
24. P. Zhao and B. Yu. Boosted lasso. *Feature Selection for Data Mining*, 35–44, 2005.
25. W. Shi, K. Lee, and G. Wahba. Detecting disease-causing genes by lasso-patternsearch algorithm. In *BMC proceedings*, 1:S60, 2007.

26. S. Xu. An expectationmaximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity*, 105:483–499, 2010.

27. G.K. Swamy, S. Edwards, A. Gelfand, S.A. James, and M.L. Miranda. Maternal age, birth order and race: Differential effects on birthweight. *Journal of Epidemiology and Community Health*, forthcoming:1–23, 2010.

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

29. P. Magnus, K. Berg, and T. Bjerkedal. The association of parity and birth weight: Testing the sensitization hypothesis. *Early Human Development*, 12(1):49–54, 1985.

30. A. Navas-Acien, E. Guallar, E.K. Silbergeld, and S.J. Rothenberg. Lead exposure and cardiovascular disease—A systematic review. *Environmental Health Perspectives*, 115(3):472–482, 2007.

31. M.L. Miranda, G.K. Swamy, S. Edwards, P. Maxson, A. Gelfand, and S. James. Disparities in Maternal Hypertension and Pregnancy Outcomes: Evidence from North Carolina, 1994–2003. *Public Health Reports*, 125(4):579–587, 2010.

32. S. Cnattingius, J.L. Mills, J. Yuen, O. Eriksson, and H. Salonen. The paradoxical effect of smoking in preeclamptic pregnancies: Smoking reduces the incidence but increases the rates of perinatal mortality, abruptio placentae, and intrauterine growth restriction. *American Journal of Obstetrics and Gynecology*, 177(1):156–161, 1997.

33. A. Castles, E.K. Adams, C.L. Melvin, C. Kelsch, and M.L. Boulton. Effects of smoking during pregnancy: Five meta-analyses. *American Journal of Preventive Medicine*, 16(3):208–215, 1999.

34. K.Y. Lain, R.W. Powers, M.A. Krohn, R.B. Ness, W.R. Crombleholme, and J.M. Roberts. Urinary cotinine concentration confirms the reduced risk of preeclamp-

sia with tobacco exposure. *American Journal of Obstetrics and Gynecology*, 181(5):1192–1196, 1999.

35. Y. Li and J. Zhu. L_1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185, 2008.

36. Q. Li, R. Xi, and N. Lin. Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3):533–556, 2010.