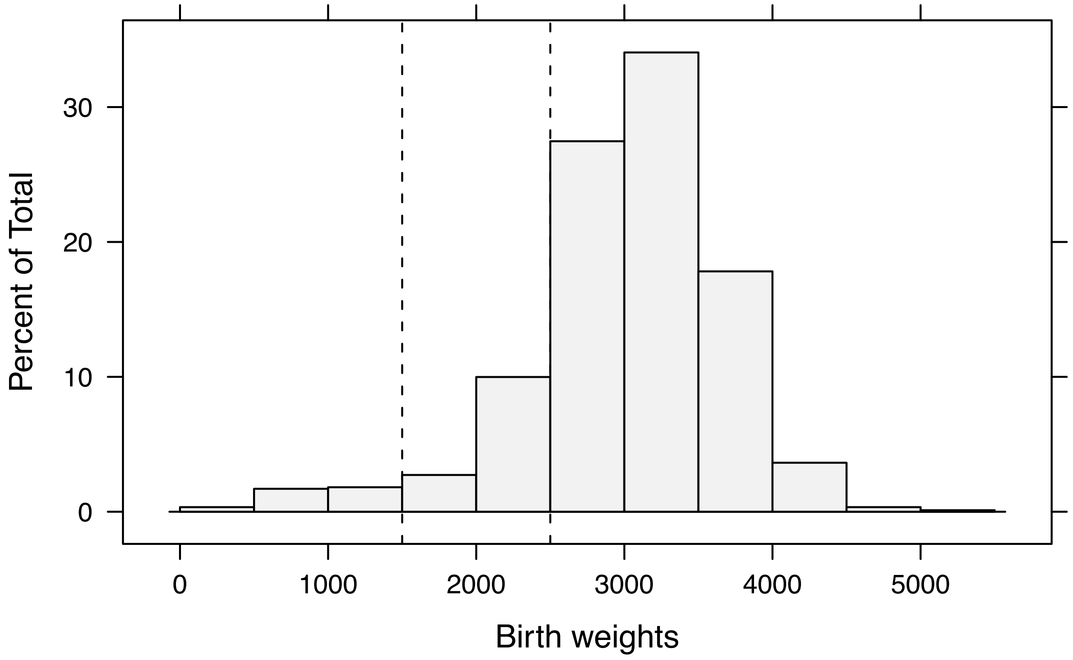
Exploratory quantile regression with many covariates: An application to adverse birth outcomes

June 3, 2011

eAppendix



eFigure 1: Histogram of birth weights with low and very low classifications marked by dashed lines.

Simulation design

The simulation design used to create Figure 2 of the main text is as follows. We consider a quantile regression of the median on ten covariates such that (i) the first three covariates are drawn from a joint normal distribution with pairwise correlation 0.99, (ii) the remaining seven covariates are drawn from a joint normal distribution with pairwise correlation 0.8, and (iii) the true $\beta = (.2, .2, .2, .3, .3, 0, 0, 0, 0, 0)'$ in the quantile regression of some outcome. We generated $n = 500$ observations from this setting, using an asymmetric Laplace distribution (with the 25th percentile at zero) as the error distribution.

Small n, big p simulations

Penalized regression models like lasso have been found to be useful when the number of predictors p exceeds the number of observations n. For example, Chapter 18 of Tibshirani, Hastie and Freedman[1] describes a lasso application with $p = 7129$ predictors and $n = 34$ samples. (See also their references.)

To demonstrate with quantile regression, we simulate 100 normally-distributed observations with 150 available covariates, though only five of the covariates actually impact the outcome variable. We draw the covariates independently from a standard normal distribution. We perform the simulation with three parameter configurations, where the non-zero effect sizes are fixed at 0.25, 0.5, and 1 in magnitude; the signs are chosen randomly. The reported models are quantile regressions of the 20th percentile of the response.

We generate 25 simulated datasets with each parameter configuration. In each dataset, we visually choose the five parameters that are most prominent, i.e., those that rise earliest and fastest in the lasso solution path. Of those five, we record the number that correspond to the true non-zero coefficients. Although choosing a pre-specified number of covariates is not how we envision or recommend analysts using the lasso solution path for exploratory analysis, doing so is convenient for simulations. We illustrate the selection of predictors using two of the simulated datasets in the next sub-section of the supplement.

eTable 1 displays the results from the simulations involving the 25 datasets for the different parameter configurations. When the non-zero β components have magnitude one, the five important predictors are selected in 14 of 25 runs. When the non-zero β components have magnitude 0.5, usually two or three of the important predictors are selected. When the non-zero β have magnitude 0.25, typically one or two of the five are selected.

We conclude that the boosted lasso method is helpful for identifying important predictors in large p, small n applications, with its efficacy improving with an increased signal-to-noise ratio. However, selecting a pre-specified number of predictors can miss important covariates.

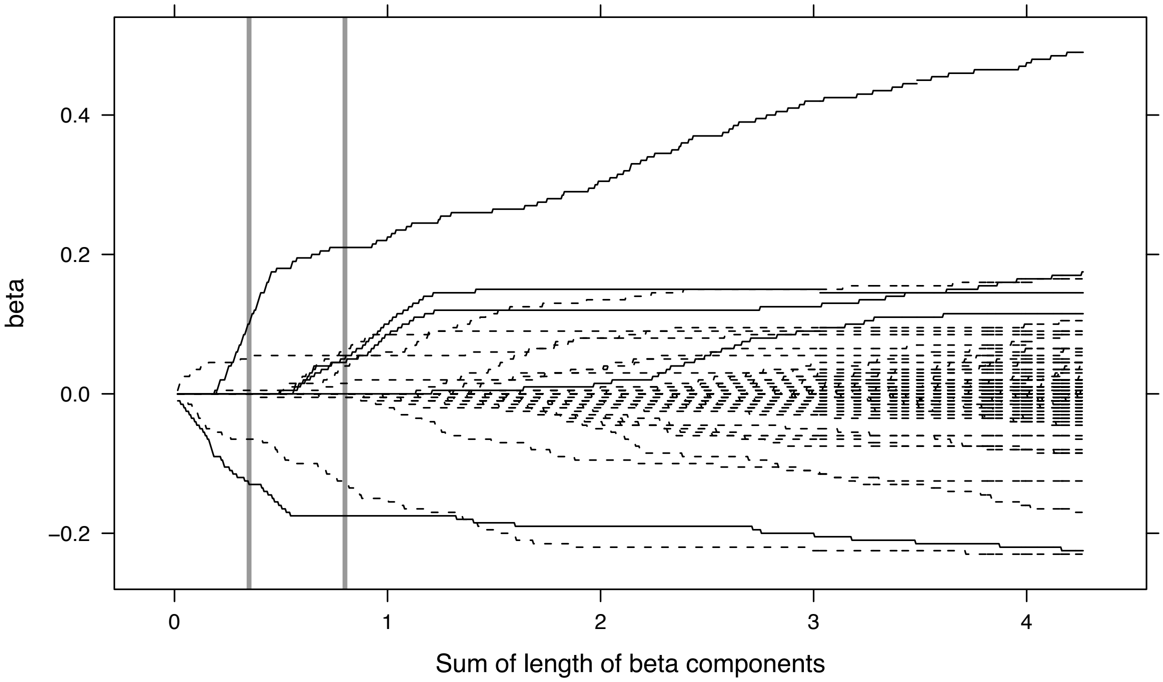
	Number correct					
β magnitude	5	4	3	2	1	0
1	14	9	2	0	0	0
0.5	0	4	10	9	2	0
0.25	0	0	2	6	14	3

eTable 1: Results from small n, big p simulations. In each case, there are five nonzero coefficients out of 150. The table records the number of times that the most prominent variables from a lasso correspond to a given number of the true nonzero coefficients.

How many variables should we consider?

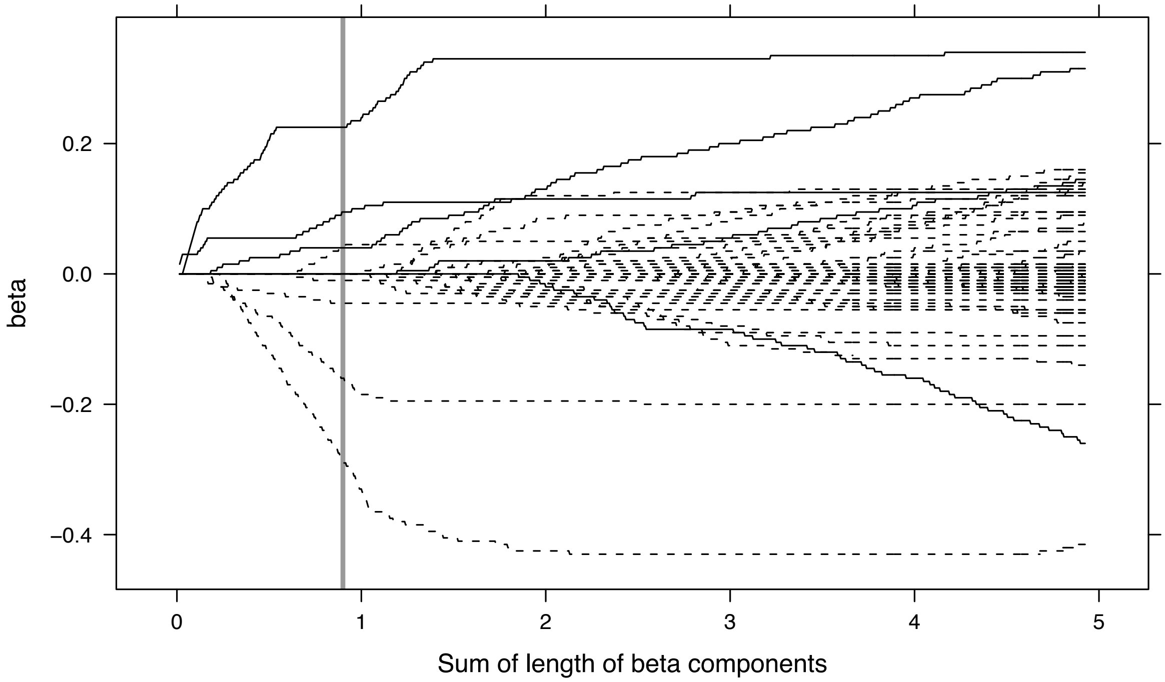
To further illustrate the use of lasso solution paths, we analyze one of the 25 runs that are summarized in the middle row of eTable 1. eFigure 2 displays the lasso solution path. A group of four predictors clearly stands out, moving decisively away from zero by the first vertical line. These would be the starting point for formal model building. This group includes two of the five truly important variables. We also would consider the next group of five variables that take on nonzero values and move quickly away from the $\beta = 0$ line, which captures another two of the truly important variables. After that, the nonzero β quickly proliferate, so that we would stop using the solution path at this point. We note, however, that analysts can follow the solution path further if they are willing to entertain the complicated models that result.

eFigure 3 displays the solution path from another dataset from the middle row of eTable 1. Here, seven β components quickly take on nonzero values; we would flag them as being important for formal model building. This group includes three of the five truly important variables. After the vertical line, the sparsity quickly degrades so we would not flag more variables on the basis of the exploratory lasso fit alone. Because two of the important variables would be difficult to detect in the exploratory analysis, we should use scientific knowledge where available when building a final model.



eFigure 2: Solution path for lasso quantile regression of simulated data. The most prominent variables take on nonzero values to the left of the first vertical, gray line. We would consider the second group to be the variables that have moved decisively away from zero by the second vertical line. Truly important variables correspond to solid lines.

To summarize, we consider the covariates whose associated parameters rise early and quickly in the solution path. The definition of "early" will vary with the application and the analyst. Some may feel comfortable building — and interpreting — models with scores of predictors, in which case they may follow the solution paths deeply. Further, if the signals are strong and the sample size is large, the data may support large models. In the end, the goal of the exploratory analysis of the solution path is to suggest a manageable number of covariates that may usefully supplement the models that scientific beliefs suggest.



eFigure 3: Solution path for lasso quantile regression of simulated data. The most prominent variables take on nonzero values to the left of the first vertical, gray line. Truly important variables correspond to solid lines.

Algorithm for the boosted lasso

Efficient lasso algorithms have been derived for standard linear models. These enable calculation of the lasso estimates of β over the entire solution path at essentially the same computational expense as that of a single ordinary least squares fit.[2] These computational tricks do not apply to quantile regression. We therefore use a statistical method called boosting that can approximate the path taken as the penalty λ_1 is reduced to zero. The method and its theoretical validity as an approximation to lasso solution paths is described by Zhao and Yu.[3]

The idea is as follows. We start with $\beta = 0$, which is the lasso solution for $\lambda_1 = \infty$. We then add or subtract a fixed, small amount δ to one dimension of β such that $L(\beta)$ is reduced as much as possible. We relax λ_1 toward zero to allow this step. This process of adding $\pm\delta$ to the most favorable dimension of β and relaxing the penalty is repeated, though at each step the algorithm considers steps back toward zero for the components of β that are non-zero at the current iteration. At some point, no further reduction of $L(\beta)$ is possible, and this corresponds to the lasso solution for $\lambda_1 = 0$, stopping the algorithm. In summary, boosted lasso can be thought of as a fractional forward selection/backward deletion algorithm that approximates the true lasso solution path. This approximation is improved by choosing a smaller value of δ. Although the solution paths in the figures may appear smooth, they are actually composed of horizontal line segments with small jumps of size δ.

Although this is a direct application of the boosting algorithm of Zhao and Yu[3] to the quantile regression empirical loss function, we are not aware of other exploratory analyses that taken this approach. R code that implements this technique is available from the authors.

Algorithm for the boosted elastic net

To fit the elastic net quantile regression model, we again rely on boosting. The algorithm is subtly different from the lasso case (requiring the generalized boosting approach of Zhao and Yu[3]), but it still consists of iteratively adding or subtracting steps of size δ from the dimension of β that is most favorable in terms of minimizing the elastic net criterion $\Gamma_{\text{EN}}(\beta)$, along with relaxing the penalty when necessary. R code for the boosted elastic net is also available from the authors.

Because the quantile regression loss function is convex but not strictly convex, it is possible that the boosted lasso algorithm will get stuck in local minima. Zhao and Yu[3] did not find this to be a problem when using boosting to fit penalized support vector machine models that use a "hinge loss" that is not strictly convex. Similarly, we did not encounter this issue in our applied analyses. If it does occur—as it did in an elastic net fit of simulated example that we considered—one component of β will bounce back and forth between two values. While this problem may make boosting unsuitable to fit a particular model, it will be obvious to the analyst, so it is not the type of problem that is likely to lead to a faulty analysis.

Additional models considered

The model we report in Table 2 of the article is the product of formal model comparison using terms suggested by the exploratory analysis. We started with a quantile regression model that includes age, (age)^2; gender of the baby; an indicator of whether it is the mother's first pregnancy; blood lead and an indicator of tobacco use during the pregnancy and their interaction; an interaction between maternal age and tobacco; and, indicators of the state of the parents' relationship, the ISEL appraisal score and their interaction.

We consider formal hypothesis tests (using the 20th percentile fits) starting with this large model. A hypothesis test between the full Model B (eTable **??**) and Model C (eTable **??**) gave a P-value greater than 0.5. Similarly, a test between the reported model (Table 2) and Model C gave a P-value in excess of 0.5. A test that considers removing the lead/smoking interaction gives a P-value of 0.067, and the test that considers removing lead, tobacco and their interaction gives $P = 0.035$.

	10th percentile		20th percentile	
	Estimate	Interval	Estimate	Interval
Intercept	2221	(1621, 2820)	2594	(2188, 3000))
Age (centered)	-3	(-33, 26)	-1	(-17, 14)
Age^2 (centered)	-3	(-7,1)	-3	(-5, -1)
Male	-20	(-244, 203)	21	(-100, 143)
2nd baby or later	7	(-273, 287)	69	(-64, 203)
Blood lead	210	(10, 411)	113	(-2, 228)
Tobacco use during pregnancy	144	(-145, 428)	-56	(-218, 107)
Lead/Tobacco	-398	(-1169, 372)	-186	(-382, 11)
Age/Tobacco	15	(-39, 69)	2	(-26, 30)
Relationship: visiting	-360	(-1429, 709)	-237	(-882, 407)
Relationship: not visiting	-85	(-1183, 1013)	-61	(-717, 594)
ISEL Appraisal	1	(-53, 55)	3	(-33, 39)
Visiting/ISEL-A	15	(-52, 162)	28	(-33, 89)
Not visiting/ISEL-A	7	(-98, 112)	12	(-54, 78)

eTable 2: Model B: Quantile regression of 10th and 20th percentiles of birth weight, using prominent interactions from the exploratory analysis. The relationship variable uses married or cohabiting as a baseline.

	10th percentile		20th percentile	
	Estimate	Interval	Estimate	Interval
Intercept	2280	(1955, 2605)	2645	(2492, 2799))
Age (centered)	-6	(-39, 27)	-1	(-17, 16)
Age^2 (centered)	-3	(-7,1)	-3	(-5, -1)
Male	-19	(-245, 207)	38	(-92, 168)
2nd baby or later	-28	(-276, 220)	51	(-90, 168)
Blood lead	217	(38, 395)	111	(-7, 229)
Tobacco use during pregnancy	121	(-161, 404)	-93	(-256, 71)
Lead/Tobacco	-400	(-1129, 330)	-170	(-374, 33)
Age/Tobacco	14	(-39, 69)	-1	(-33, 30)

eTable 3: Model C: Quantile regression of 10th and 20th percentiles of birth weight. This model removes the relationship and ISEL Appraisal variables from Model B.

IRB statement

All aspects of the HPHBS, including the analyses presented here, were conducted according to a research protocol approved by Duke University's Institutional Review Board. HPHBS is embedded within the Southern Center on Environmentally Driven Disparities in Birth Outcomes (SCEDDBO).

References

1. T Hastie, R Tibshirani, JH Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer; 2009.
2. B Efron, T Hastie, I Johnstone, R Tibshirani. Least angle regression. *Ann Stat.* 2004;32:407–499.
3. P Zhao, B Yu. Boosted lasso. In: H Lui, R Stine, L Auslender, chairs. *Workshop on Feature Selection for Data Mining.* Newport Beach, CA: SIAM; 2005:35–44.