P8111: Linear Regression Models

Final Project

Part 1

Anthony Santistevan

uni: acs2244

May 12, 2014

Low birth weight has long been known to be associated with numerous health outcomes in childhood health and has more recently been associated with health outcomes in adult life. The goal of this analysis is to identify relationships between demographic and biological covariates and birth weight in order to better understand possible contributing factors to low birth weight.

The initial data set consisted of nineteen covariates ranging from biological information about the infant (e.g. gestational age, head circumference, body length) to demographic information on the parents (e.g. race, family income, smoking status) on a total of 4342 infants. The response variable was birth weight measured in grams. Two covariates present (previous number of low birth weights and number of prior small for gestational age babies) were omitted from the analysis as they had a value of zero for all subjects. Parity was also present however not used in the analysis as there were only three subjects with non-zero values. One subject was removed from the analysis because the value for the covariate of menarche was 0 (the mother was reportedly 0 years old at onset of menarche) which is biologically implausible and was likely a coding error. One of the covariates (mother's weight gain) was perfectly collinear with two of the other covariates (mother's pre-pregnancy weight and mother's weight at delivery) thus only mother's weight gain was used in the analysis. The final analysis was conducted on the remaining $N = 4341$ subjects and the 14 covariates used in the analysis are summarized in Table 1.

Several models were explored including multiple linear regressions, LASSO models, and generalized additive models with varying subsets of the covariates. Models were fit in R version 3.1.0 using software packages *glmnet* (LASSO) and *mgcv* (additive models). In the multiple linear and LASSO regressions, polynomial terms for several continuous predictors (menarche, gestational age, pre-pregnancy BMI, mother's age) were explored and equivalently smooth terms for these covariates were explored in the additive models. Regression assumptions were checked by visual inspection of Q-Q plots, plots of residuals against fitted values, and plots of Cooke's distance. The tails of the distribution for the error terms deviated somewhat from normality (approximately 8 subjects deviated strongly). These subjects were included in the analysis as they were not found to be highly influential. Residual plots showed indication of greater variance in the residuals for smaller fitted values, however various Cox transformations did not remedy this so analyses were run directly on birth weight. Inclusion criteria for covariates was based on a mixture of AIC values and predictive performance; if inclusion of the covariate increased the AIC and/or did not change predictive performance, the covariate was not included in the model. Predictive performance was assessed using cross validation in which one-fifth of the data set was removed and kept as a test set. Models were fit on the remaining data (training set) and then the sum of the squared errors were computed

between the predicted values and the actual values of the test set based on predictions from the trained models. This procedure was repeated 100 times to identify variability in the relative performance between models. Models taking non-linearity into account for gestational age and weight gain performed best.

The selected model is an additive model with smooth terms for gestational age and weight gain during pregnancy (Figure 1). Parameter estimates along with test statistics and p-values are summarized in Table 2. There was a significant relationship between the smooth parameter for gestational age and birth weight ($F_{3.991,8} = 14.96, p < 2e-16$) when controlling for the other covariates in the model. The smooth term for gestational age (Figure 1a) suggests that a gestation period less than approximately 39 weeks has a net negative effect on birth weight and that this effect reduces as gestation approaches 39 weeks. Between gestation ages of 39 and 42 weeks there is a positive effect of gestation age on infant weight. This effect increases as gestation length increases and reaches a maximum at approximately 42 weeks. There does not appear to be any further increase in birth weight beyond gestation of 42 weeks. There is a significant relationship ($F_{4.114,8} = 11.35, p < 2e-16$) between mothers' weight gain and birth weight (Figure 1b) when controlling for the other covariates which suggests that birth weights are lower among infants whose mothers' weight gain during pregnancy was less than 22 pounds; this effect is much larger for mothers who lost a large amount of weight during pregnancy and is reduced as weight gain approaches 22 pounds. For women who gained between 22 and 50 pounds during pregnancy, there is a positive relationship between mothers' weight gain and birth weight and these mothers had lager babies on average. Mother's weight gain beyond 50 pounds appears to change this relationship and is associated with a decrease in birth weight, although these mothers still had larger babies on average.

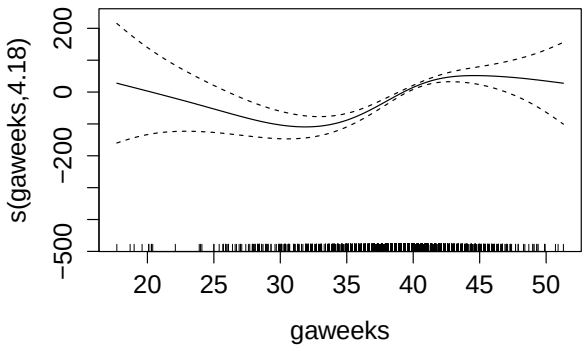
The parametric effects of the model show that girls were heavier than boys on average ($\hat{\beta} = 27.749, p = 0.001$). Head circumference ($\hat{\beta} = 130.039, p < 2e-16$) and baby length ($\hat{\beta} = 73.861, p < 2e-16$) significantly increased with birth weight. Mother's height and pre-pregnancy BMI were significantly positively associated with birth weight ($\hat{\beta} = 12.256, p = 1.02e-13$; $\hat{\beta} = 8.536, p = 5.99e-10$, respectively). An omnibus test for mother's race showed at least two races differed in birth weight ($F_{3,4341} = 62.898, p < 2e-16$). Children with black mothers and children with Puerto Rican mothers had significantly lower birth weights than children with white mothers ($\hat{\beta} = -135.720, p < 6e-16$; $\hat{\beta} = -99.017, p = 9.03e-7$, Bonferroni corrected, respectively). There was no statistically significant difference between infants whose mothers' were Asian and those whose mother's were white ($\hat{\beta} = -75.255, p = 0.2232$, Bonferonni corrected). The number of cigarettes smoke per day while pregnant was negatively associated with birth weight ($\hat{\beta} = -4.928, p < 2e-16$) while controlling for other covariates. Family income did not have a statistically significant association with birth weight when controlling for the other covariates ($\hat{\beta} = 0.296, p = 0.0893$). Results are summarized in Table 2.

Table 1: Summary Statistics. Data in the table summarize the covariates and response variable used in the final model.

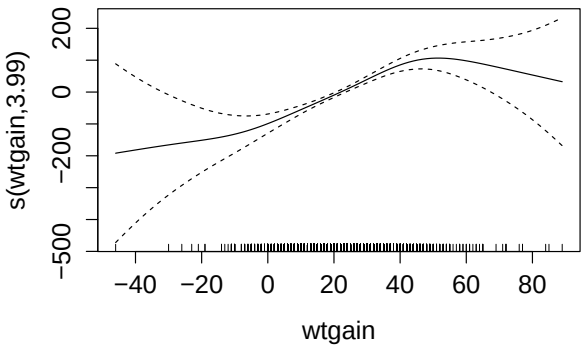
	Male	2230 (51.4)	
	Female	2112 (48.6)	
Frequency of Race (%)		Mother	Father
	White	2147 (49.4)	2123 (49)
	Black	1909 (43.9)	1910 (44)
	Asian	43 (1)	46 (1)
	Puerto Rican	243 (5.7)	248 (5.7)
	Other	0 (0)	14 (0.3)
Malformation			
	Present	15 (0.3)	
	Absent	4326 (99.7)	
	Median	1st Quartile	3rd Quartile
Birth Weight (g)	3132	2807	3459
Head Circumference (cm)	34	33	35
Length (cm)	50	48	51
Gestation Age (weeks)	39.9	38.3	41.1
Mother's Weight Gain (lbs)	22	15	28
Mother's Pre-Pregnancy BMI	21.03	19.53	22.91
Mother's Age (years)	20	18	22
Mother's Height (inches)	63	62	65
Mother's Age at Menarche	12	12	13
Cigarettes/day	0	0	5
Family Income (hundreds/month)	35	25	65

Table 2: Additive Model. Linear coefficients and significance of predictors of birth weight. *Approximate p-value.

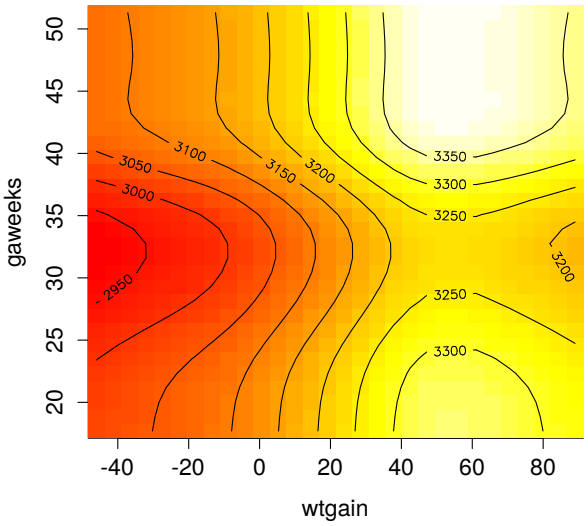
Smooth terms	edf	Ref.df	F	p-value*
Mother's weight gain (lbs)	3.191	8	14.96	< 2e-16
Gestational age (weeks)	4.114	8	11.35	< 2e-16
Parametric coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-5838.5409	141.7917	-41.177	< 2e-16
Sex (Female)	27.7485	8.4343	3.290	0.00101
Head Circumference (cm)	130.0385	3.4446	37.751	< 2e-16
Baby length (cm)	73.8611	2.0204	36.557	< 2e-16
Mother's height (in)	12.2560	1.6422	7.463	1.02e-13
Pre-pregnancy BMI	8.5363	1.3758	6.205	5.99e-10
Mother's race: Black	-135.7200	9.9234	-13.677	< 2e-16
Mother's race: Asian	-75.2549	42.1730	-1.784	0.07442
Mother's race: Puerto Rican	-99.0166	19.2991	-5.131	3.01e-07
Cigarettes/day	-4.9279	0.5835	-8.445	< 2e-16
Family Income	0.2960	0.1742	1.699	0.08933



(a) Gestational Age (weeks)



(b) Weight Gain (lbs)



(c) Contour of Gestation and Weight Gain

Figure 1: Smooth Functions from Additive Model. (a) Smooth function for gestational age (b) smooth function for weight gain (c) contour plot of joint effects of gestational age and weight gain. Black lines indicate constant birth weight (grams)

Part 2.

Quantile regression aims to extend the familiar procedure of modeling the conditional mean of some variable given a set of covariates (e.g. ordinary least squared regression, *OLS*) to the modeling of a conditional *quantile* given a set of covariates. In OLS, parameter estimates ($\hat{\beta}$) for the slopes of the covariates are obtained by minimizing the residual sum of squares (RSS) over the set of parameters β. In quantile regression, letting τ be the quantile of interest, the $\hat{\beta}_\tau$'s are estimated by minimizing the sum of the tilted absolute value function of the differences between the observed response variables (y_i, $i = 1, 2, \ldots, n$) and the conditional quantile function [1]. This essentially weights the effects of the points above and below the quantile of interest equally in the minimization process. In the univariate case with predictor x_i and letting $\tau = 0.5$, this is equivalent to

$$\min_{\beta} \sum_{i=1}^{n} |y_i - x_i\beta|$$

and yields estimates for the conditional median. Because this function is non-differentiable, a linear programming approach must be employed for the minimization [1].

This technique is innately robust to the effects of outliers as they have little effect on the conditional median. This effect is demonstrated in Figure 2a in which the introduction of a single outlying observation has a large effect on the slope of the best fit line in OLS (dotted) however the median regression line (solid) is unaffected and captures the structure of the data more truthfully. Note that the OLS estimate completely fails to capture the conditional mean for the lower and upper values of x, whereas the median regression captures the conditional means quite well.

To compare the differences between parameter estimates from a OLS and a median regression, I simulated 10,000 datasets of 100 data points following the form

$$y_i = 2x_i + \epsilon$$

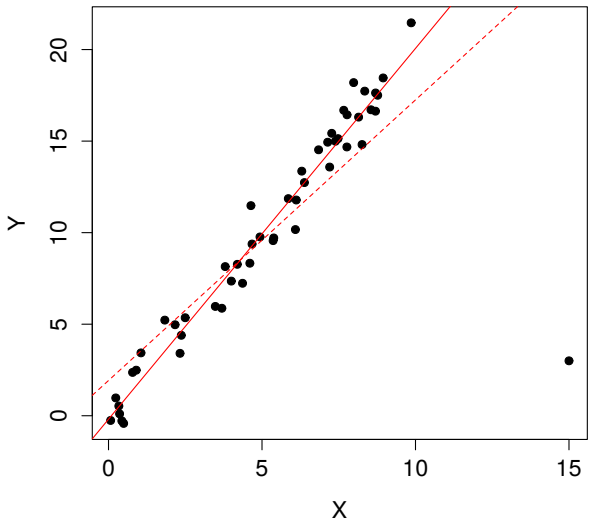
where $\epsilon \sim N(0, 1)$ and $x \sim \text{Unif}[0, 10]$. Results from the analysis are in Figure 2b and show that while both parameters are unbiased towards the true population parameter $\beta = 2$, there is less variance in the estimate in the OLS distribution. This suggests that in the absence of outliers, OLS (red curve) performs superior to median regression due to lower variance in the estimates and will provide tighter confidence intervals about the estimates. To assess differences between the two forms of regression in the presence of an outlier, I performed the same simulation above with the introduction of one outlying point $\{x_o, y_o\}$ where $x_o \sim N(15, 3)$ and $y_o \sim N(3, 1)$ in each simulated data set. Results of the simulation are in Figure 2c and show that the parameter estimate for

β under OLS (red curve) are highly variable and biased away from the true parameter $\beta = 2$. The parameter estimates for β under the median model however are much less variable and are close to the true parameter $\beta = 2$. From this, it is evident that median regression is robust to the presence of outliers and yields consistent results whereas OLS is highly sensitive to the presence of outliers.

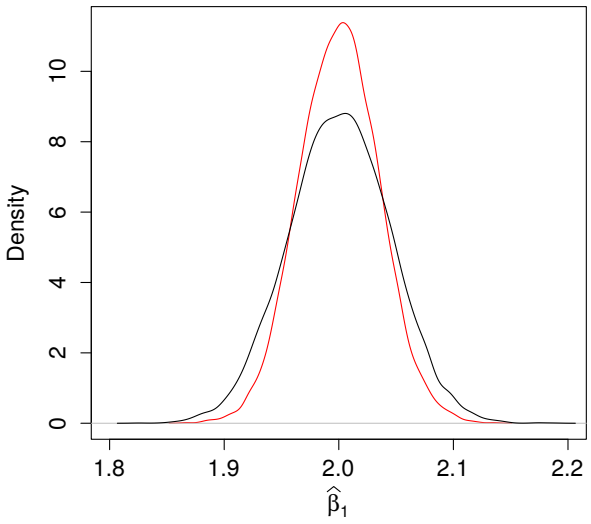
Using the *rqss()* function in the *quantreg* package in R version 3.1.0, I fit the median regression equivalent of the additive model from Part 1 of this report. A comparison of the parametric parameter estimates from both regressions is shown in Table 3. Parameter estimates were quite similar in general with only the slope for mother's race: Asian and baby length appearing largely different. The larger negative effect on mother's being Asian in median regression suggest there may have been some high leverage points which masked this association in the regular additive model. Additionally, the higher positive relationship for baby length may suggest that there were outliers on the lower end of the distribution for baby length.

One benefit of quantile regression is that one is no longer limited to exploring the association of covariates to only the central tendency of a response. In this framework, one may specify *any particular* quantile of interest and study the effects of covariates on the response in the selected quantile(s). That is, researchers may be interested in the association between covariates on the lower quartile of a distribution; by using quantile regression, they may accomplish this and highlight differences between how other percentiles are associated with the same covariates. This is a quite powerful approach because researchers may learn more about the association between covariates and a response across the entire *distribution* of responses, rather than limiting themselves to the mean/median. I utilized this approach to re-analyze the data from Part 1 of this report in the 10th, 25th, 50th, 75th, and 90th percentiles.

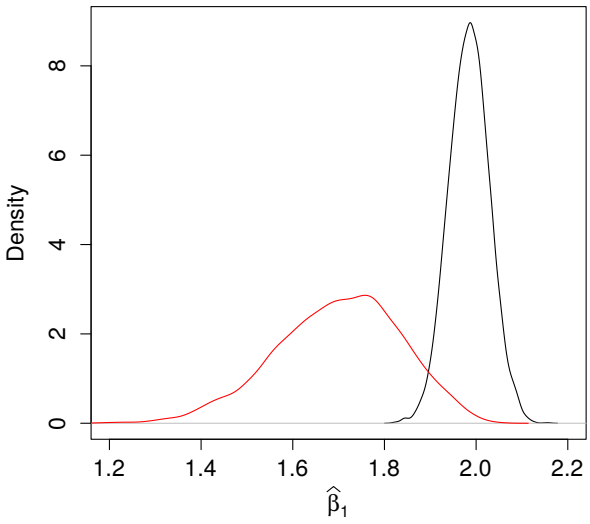
Figure 3 compares the regression coefficients for each of the parametric terms in the quantile additive models with their associated 95% confidence intervals. The parameter estimate for each of the parametric terms from the additive model in Part 1 are overlaid with their 95% confidence intervals for comparison. There are clear differential effects across the quantiles in several of the covariates. Sex differences in birth weight appear to reduce as the quantiles increase. Head length appears to have a larger effect on birth weight as the quantiles increase. Mother's height has a substantially larger effect on birth weight in higher quantiles. Lastly, the number of cigarettes per day smoked while pregnant appears to have a larger negative effect on the higher quantiles.



(a)



(b)



(c)

Figure 2: Effects of Outliers on Coefficient Estimates in OLS and Median Regression. (a) Example of effects of an outlier in a single data set. Solid line in median regression fit and dotted line is OLS fit. (b) Density of slope estimates of 10,000 simulated data sets from OLS (red) and median regression (black) (c) Density of slope estimates of 10,000 simulated data sets in the presence of an outlier from OLS (red) and median regression (black)

Table 3: Comparison of parametric estimates from standard additive model and median additive model.

	Additive Model	Median Additive Model
(Intercept)	-5839.99	-6216.41
Sex (Female)	27.79	31.66
Head Circumference (cm)	130.02	124.45
Baby length (cm)	73.88	81.29
Mother height (in)	12.27	12.25
Mother race: Black	-135.87	-121.74
Mother race: Asian	-75.02	-130.09
Mother race: Puerto Rican	-99.35	-83.88
Mother's pre-pregnancy BMI	8.55	8.65
Cigarettes/day	-4.93	-4.81
Family income	0.30	0.20

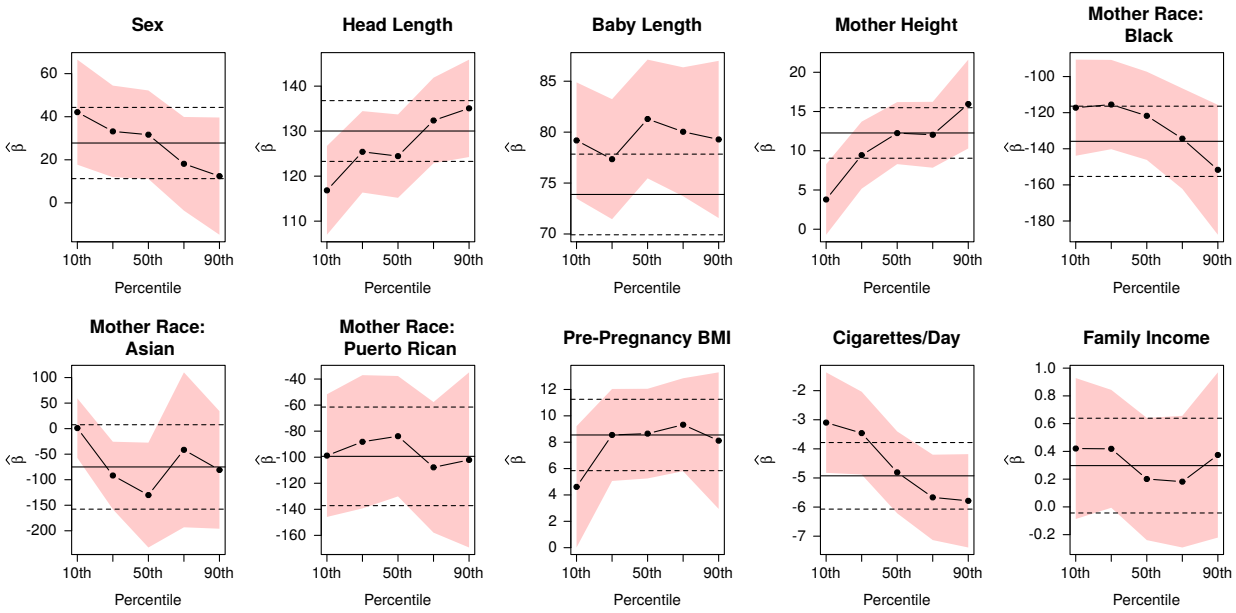


Figure 3: Quantile Parameter Estimates from Birth Weight Model. Bottom axes: 10th, 25th, 50th, 75th, and 90th percentiles. Left axis: parameter estimates. Black points indicate point estimates from the respective quantile regressions and red area indicate 95% confidence intervals. Black horizontal solid lines are the point estimates from the mean model used in Part 1 and dotted horizontal lines are their 95% confidence intervals.

References

[1] Koenker, R., Hallock, K. Quantile Regression. Journal of Economic Perspective, Vol. 15, Number 4, Fall 2001, pp 143-156