Birth Weights of Babies born in North Carolina

Katherine Hansen

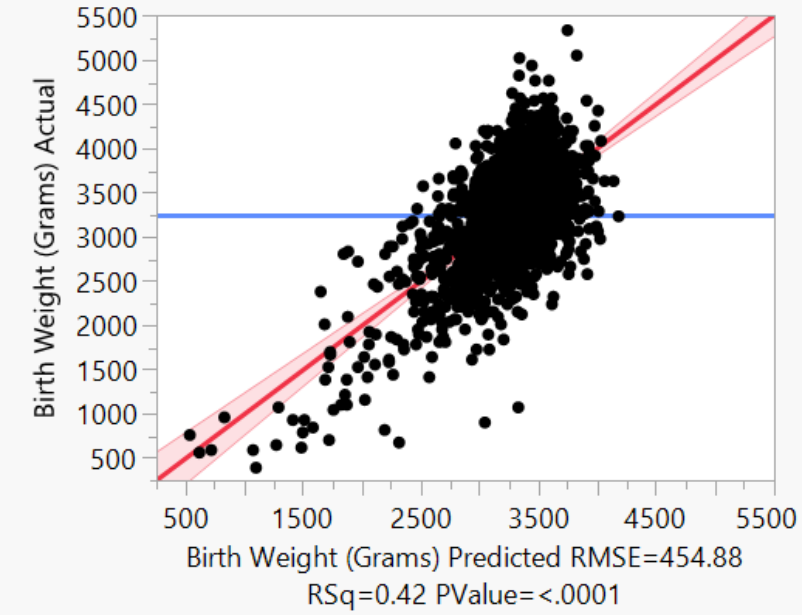
Introduction

Pregnancies are a hugely complicated process that take the better part of a year. The health of the baby is important. Several factors pertaining to the baby's health such as its weight and if they are underweight are measured at birth. I will be using data of a random sample of 2,000 infants born in North Carolina between 2003 and 2007 from the State Center for Health Statistics and the Howard W. Odum Institute for Research in Social Science at UNC at Chapel Hill. I used only the observations that had complete data which left 1931 observations that I then fit a multiple linear regression model to predict the birth weight of a newborn based on several variables. These variables include: the *mother's minority* status as either nonwhite or white, whether the mother is a *smoker*, the *mother's weight gain* during pregnancy in pounds, the number of *children previously* had by the mother, the *plurality of birth* as either single or multiple, the *sex* of the baby, the *gestational age* in weeks, and the *Apgar score* measured five minutes after birth on a scale of 0-10.

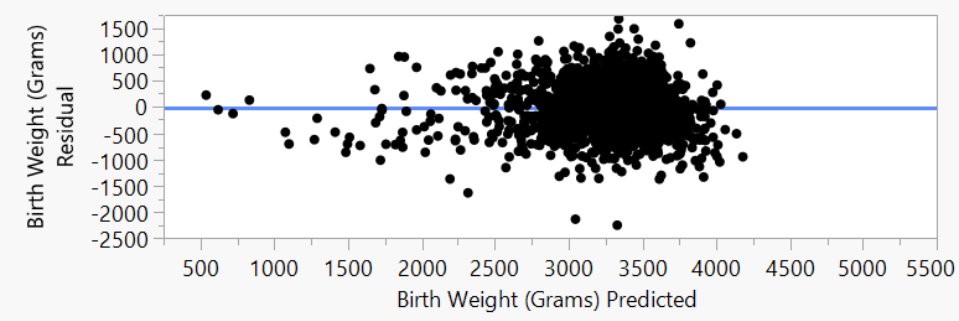
The Model

Initially, I fit a model using all the variables to predict the birthweight which turned out to not be a good fit. While normality and independence are met when checking the model assumptions, linearity seemed violated as shown in Graph 1 by the upwards curve, and looking at the residual by predicted plot shows that the constant variance assumption was also violated given the fanned-out effect observed (Graph 2).

Graph 1. Initial Plot of Actual vs. Predicted Values

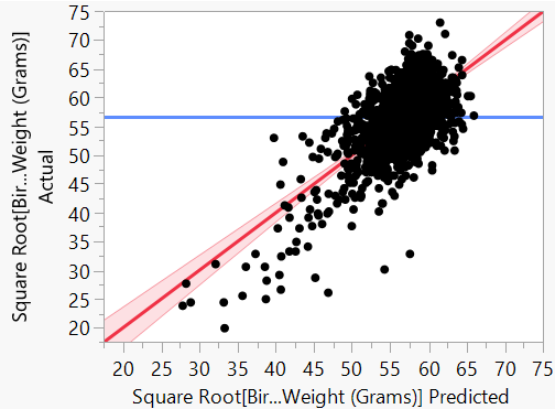


Graph 2. Initial Plot of Residuals vs. Predicted Values

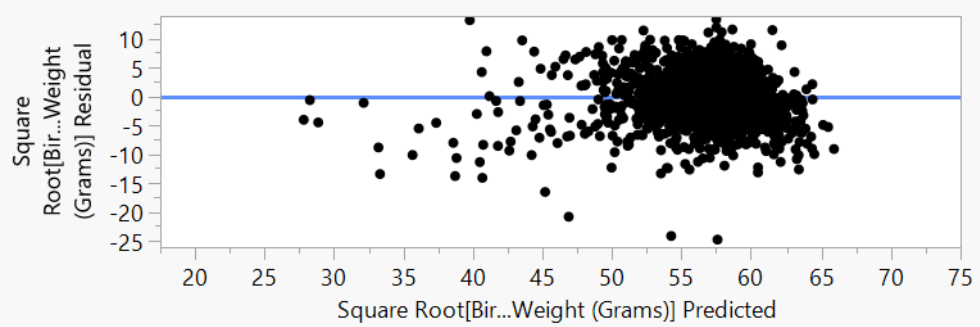


Since both linearity and constant variance were not met, I decided to use the square root of the response, birth weight, to remedy these. For each observation, I took the square root of the birth weight and then used those as the response and refit the line (Graph 3). Now linearity is met seen in Graph 4 where the majority of points are distributed equally around the y = 0 line. And the constant variance is also improved seeing as most of the points are within a horizontal band also seen in Graph 4.

Graph 3. Refit line



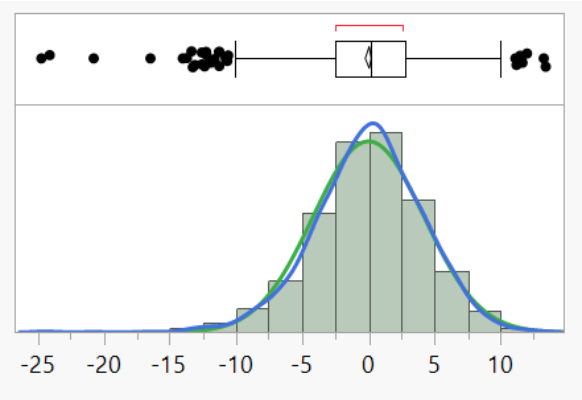
Graph 4. Residual Plot



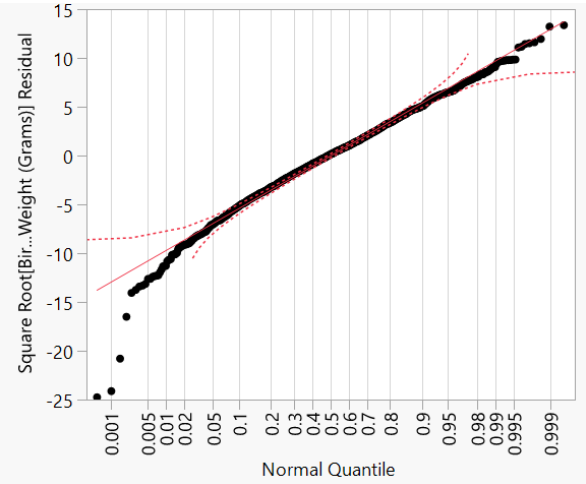
Still, independence is assumed because the data were not collected in a specific order, and the normality assumption met as the residuals follow a normal distribution though slightly skewed

left, Graph 5, and the normal quantile plot, Graph 6, is fairly straight and again we can see the small skew in the tail.

Graph 5. Residual Distribution



Graph 6. Normal Quantile Plot of Residuals



Next, I investigated possible outliers, firstly with respect to the predictors. I considered values twice as large as the mean leverage value calculated by the number of parameters divided by the number of observations. This came out to be 0.009 and while some observations had caught my attention none of the observations surpassed the mean leverage value. To check for outliers in respect to the response, the square root of birth weight, I used Bonferroni's correction. This resulted in observations 723, 1866, and 846 being deemed outliers. Then to test whether these outliers are influential or not, I computed the Cook's distance for each observation. I deemed observations influential if they exceeded the 50^{th} percentile of the F-distribution [F(9,1922)=0.93]. Using this cutoff, no outlier was influential.

I conducted a multiple linear regression to predict the square root of birth weight(g) of a baby born in North Carolina based on the mother's minority status, whether the mother is a smoker, how much weight the mother gained during pregnancy(lbs.), whether the mother has had children previously, whether a single baby or multiple babies were born, the baby's sex, the

baby's gestational age(weeks), and the Apgar score at five minutes. The results indicated these predictors explained 46.1% of the total variation in the response ($R^2 = 0.461$, $F(8,1922) = 1.51$, $p < .0001$). All predictors were significant predictors of birth weight as shown in Table 1. The variable coefficients should be interpreted as the average change after controlling for all the other variables. Also remember that the response is the square root of birth weight, not a baby's birth weight directly. So, the square root birth weight of a baby born to a Nonwhite woman is 1.60 less than if the mother was White. The baby of a mother who smokes lowers the response by 2.25 when compared to a non-smoking mother. For every ten pounds a mother gains during pregnancy, the response increases by 0.50; and for every one increase in children the mother has had previously had, the square root birth weight increases by 0.44. The response is 5.56 lower when the mother births multiple babies than if only a single child is born. The square root birth weight of females is 0.87 less than that of males. For every additional week in gestation, the response increases by 1.11. Lastly, a one-unit increase of the Apgar score at five minutes results in a 0.86 increase of the square root birth weight.

Table 1. Multiple Linear Regression Results

Variable	Parameter estimate	Standard error	T-stat	p-value	Lower 95% CI	Upper 95% CI
Mother Minority[Nonwhite]	-1.60	0.21	-7.64	<.0001	-2.02	-1.19
Mother Smoker[Y]	-2.25	0.30	-7.54	<.0001	-2.83	-1.66
Mother Weight Gain	0.05	0.01	7.50	<.0001	0.04	0.07
Mother Previous Children	0.44	0.08	5.46	<.0001	0.28	0.60
Plurality of Birth[Multiple]	-5.56	0.56	-9.90	<.0001	-6.66	-4.46
Sex[Female]	-0.87	0.19	-4.55	<.0001	-1.25	-0.50
Gestational Age	1.11	0.04	27.37	<.0001	1.03	1.19
Apgar Score Five Min	0.86	0.15	5.84	<.0001	0.57	1.15

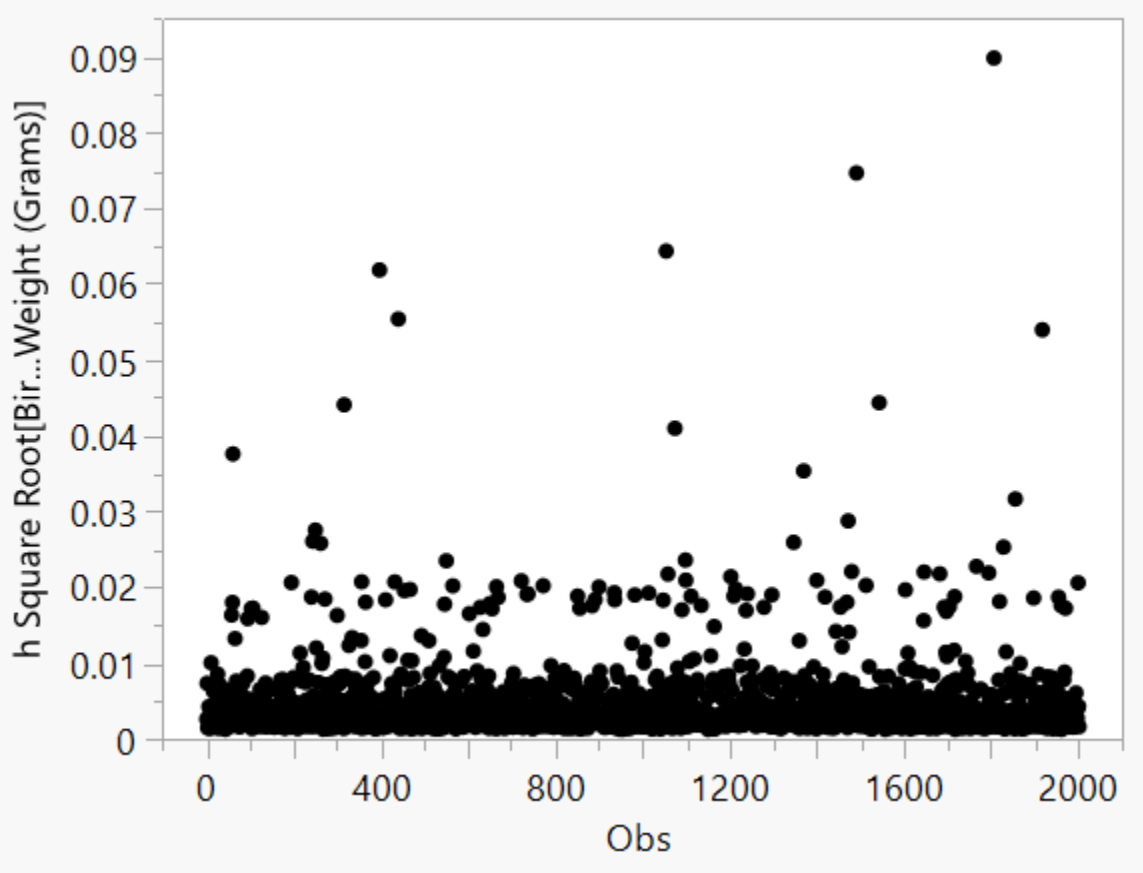
Conclusion

The results of my analysis are statistically significant and state that about 46.1% of babies' birth weight can be explained by the *mother's minority* status, whether the mother is a *smoker*, the *mother's weight gain* during pregnancy, the number of *children previously* had by the mother, the *plurality of birth*, the *sex* of the baby, the *gestational age* , and the *Apgar score* measured after five minutes. A limitation of this analysis is that the data was only from North Carolina, meaning that these results are only generalizable to babies born in North Carolina, and this is not a causal relationship. We cannot control who smokes or who does not ethically, and so we cannot run a casual experiment to test the direct effect of traits like smoking. Another point is that these variables only explained about less than half of the variation in the birth weight, so

there are other variables to consider outside of this data. It would be interesting to investigate other factors.

Appendix:

Outliers and Influence



	rstudent	unadjusted p-value	Bonferroni p
723	-5.952133	3.1378e-09	6.0686e-06
1866	-5.818402	6.9436e-09	1.3429e-05
846	-5.000611	6.2341e-07	1.2057e-03

```
> qf(.5,9,1922)
[1] 0.9273053
```

Model Summaries

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	28870.365	3608.80	205.0857
Error	1922	33820.514	17.60	Prob > F
C. Total	1930	62690.879		<.0001*

Summary of Fit

RSquare	0.460519
RSquare Adj	0.458274
Root Mean Square Error	4.194821
Mean of Response	56.68821
Observations (or Sum Wgts)	1931

References

Data Source: Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill. 2009. North Carolina Vital