Analyzing Infant Birth Weights through Linear Regression: Impact of Maternal Smoking Status

Jamin Kiukkonen

1 Description

The research hypothesis for this paper was that the more a mother smokes during her pregnancy, the lighter babies they tend to give birth to. The dataset used for this paper was part of a cohort study conducted by Rantakallio et al. The variables of the dataset included birth weight, gender, mother's age, family's social class (4-category scale), number of siblings, and maternal smoking (3-category scale). The variables used in this report for the regression analysis were birth weight, maternal smoking, and mother's age. The mean age of the dataset was 27.5, ranging from 17 to 46. The mean birth weights by maternal smoking status are shown on the table below.

Smoking status	None	Little	Much
Birth weight	3454	3314	3106

Table 1: Mean birth weights by maternal smoking status.

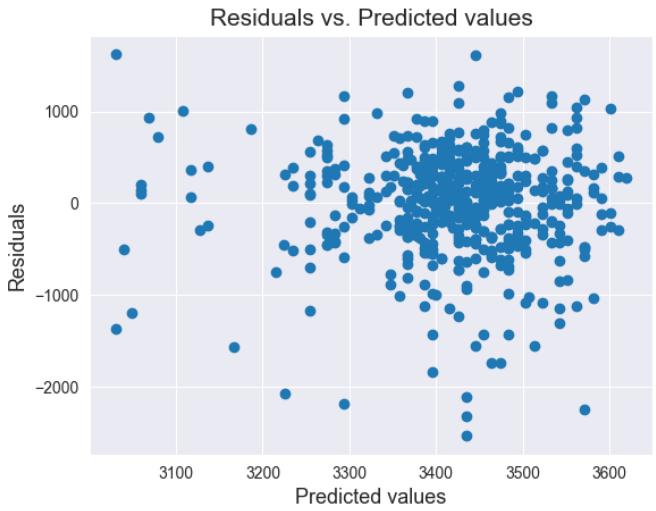
2 Methods

Regression analysis was conducted to answer the question whether mother's who smoke more during pregnancy, give birth to lighter babies. Linear relationship, no multicollinearity, homoscedasticity, and normality of residuals, were assumed. For the regression analysis, birth weight was used as the

response variable, and maternal smoking and mother's age were used as predictor variables. Age was taken as a second predictor variable to evaluate whether mother's age affected the baby's birth weight. The formula of the model used in the regression analysis was as follows

$$Weight = \beta_1\, Smoking[little] + \beta_2\, Smoking[much] + \beta_3\, Smoking[none] + \beta_4\, Age + \epsilon.$$

The information was collected from 500 participants. Out of those 500, 403 did not smoke at all during pregnancy, 78 smoked a little during pregnancy, and 19 smoked much during pregnancy. ANOVA was performed to investigate the differences between the group means of the smoking classes. Assumptions testing revealed that the data is heteroscedastic. This can be illustrated by the graph below

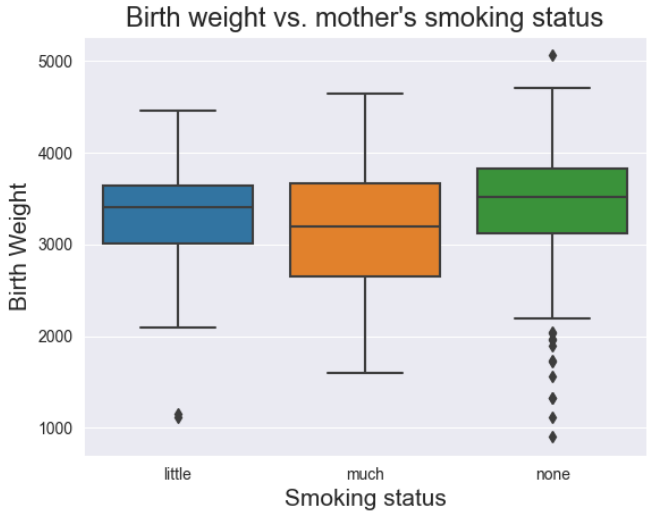


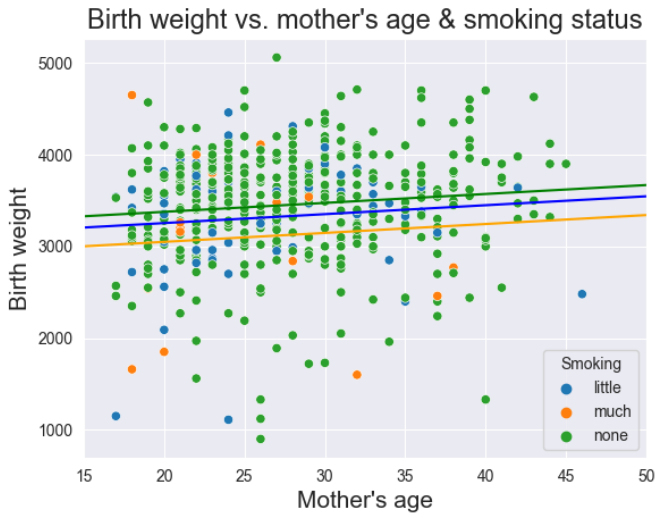
which clearly shows that the variance is higher on the lower end of the predicted birth weights. As for the multicollinearity assumption, a VIF-value of 1 was obtained, which tells us that there is no correlation between the predictor variables.

3 Results

Visualizing the dataset revealed that indeed mother's who smoke more during pregnancy, tend to get somewhat lighter babies. This can be illus-

trated by a boxplot, and a scatterplot with the best-fit lines produced by the regression model.





As we can see, the boxplot shows that mother's who smoke more tend to get somewhat lighter babies. It is, however, worth noting that there are many outliers in the non-smoking group. The linear regression model did a poor job predicting the birth weights, with a R-squared score of 0.028, and a RSE score of 593. Yet, the model still predicts that the more one smokes, the

lighter baby one will get. The regression coefficients and the corresponding confidence intervals are shown on the table below.

	Coef	Lower limit	Upper limit
Smoking[none]	3182.1638	2944.520	3419.808
Smoking[little]	3059.5412	2806.592	3312.491
Smoking[much]	2854.7747	2513.160	3196.389
Age	9.7332	1.480	17.986

Table 2: Regression coefficients and confidence intervals.

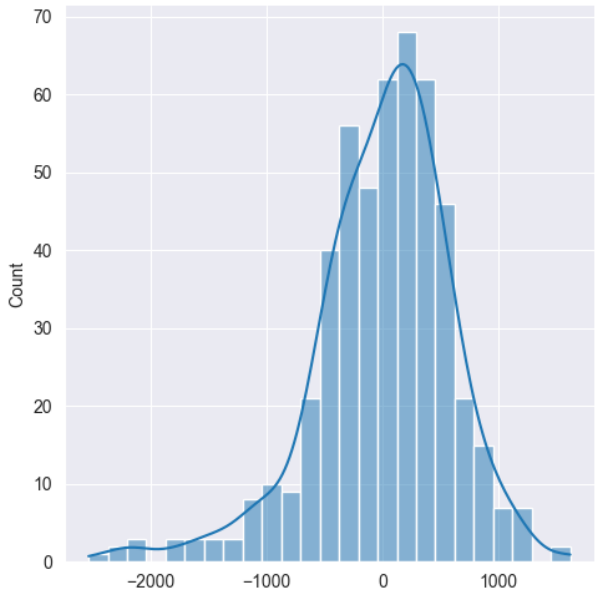
Our model tells us that a mother who is classified in the "none" smoking class, is expected to have a baby who's weight is around 3182 grams. A mother who is classified in the "little" smoking class is expected to have a baby who's weight is around 3060 grams, and a mother that of "much" smoking class is expected to have a baby who's weight is around 2855 grams. The age coefficient, which is the slope of our model, shows us that a mother who gets a one unit (year) increase in age, is expected to have a 9.7 grams heavier baby. Analysis of variance was conducted to evaluate whether the predicted group means differed significantly from each other. This is illustrated by the table below

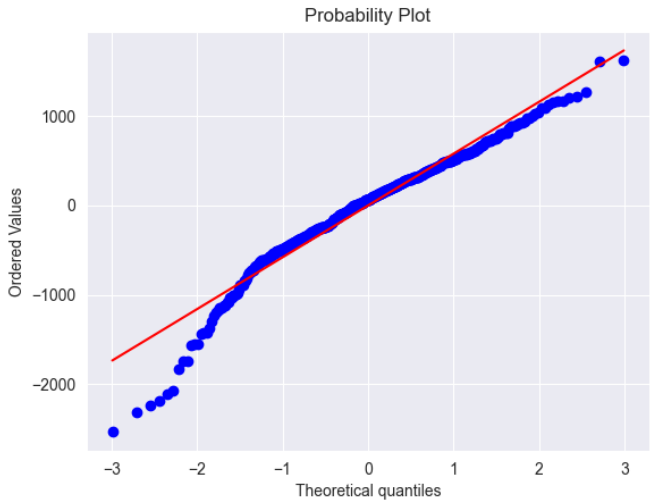
	sum_{sq}	df	F	PR($> F$)
Smoking	2.477254e+08	3.0	234.979266	7.827668e-95
Age	1.886749e+06	1.0	5.369013	2.090390e-02
Residual	1.743016e+08	496.0	NaN	NaN

Table 3: Analysis of variance

which shows that the group means seem to differ significantly, and that there is also an association between mother's age and expected birth weight.

The residual analysis revealed that the residuals are normally distributed, with a slight skewness to the left. The histogram of residuals and the probability plot shown below imply that the residuals are indeed centered around zero, which indicates that they are normally distributed. The mean value of the residuals is 0.0003, which also supports the claim that they are coming from a normal distribution.





4 Conclusion

In conclusion, The regression model with the given independent variables did not predict the birth weight of a baby well. By the support of ANOVA, maternal smoking was found to be associated with the difference

in mean birth weights among the different smoking classes. Also mother's age was found to be associated with the birth weight of a baby. In the future, more sophisticated models might be needed to predict the response variable. Assessment of the linear relationship between the response variable and predictor variables need to be evaluated more thoroughly in the future.