Investigation of potential factors affecting birthweight

# Introduction

The purpose of this report is to investigate associations between birthweight of child and six other various potential explanatory variables. Throughout the report, we will build, refine and validate a linear model before using it to later predict the birthweight of 100 individuals from a given test set. In our findings, we concluded that the sex of the child and how often the mother smoked had an insignificant association to birthweight of child whilst conversely, the other explanatory variables which are outlined in exploratory analysis were found to be significant.

# Exploratory analysis

The data set consists of 327 children regarding their birthweight along with the following 6 variables: age of mother; gestation period; sex of child; smoke level during preganacy (levels: ‘No’,‘light’ and ‘heavy’) ; pre-pregnancy weight of mother; rate of growth of child in the first trimester.

Clearly, the variables ‘smoke’ and ‘sex’ should be treated as factors and the remaining as numerical variables.

Below shows a matrix scatter plot to uncover the bivariate relationship between continuous variables.

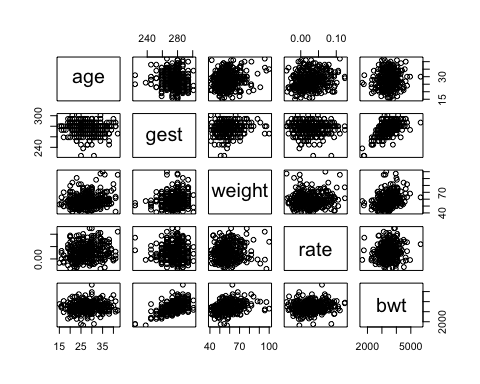


Figure 1: Matrix scatter plot showing pairwise relationship between each of the covariates.

Gestation period and pre-pregnancy weight of the mother are both positively correlated to birthweight of child which would highly suggest including these variables in the linear model would be appropriate. They have a correlation value of 0.54 and 0.27 respectively. There appears a lack of pairwise relationship between the remaining numerical variables (age and rate) against birthweight but later we’ll find out they are significant when other potential explanatory variables are included. No obvious correlation between potential explanatory variables which would prompt us to further investigate if these should be included in the linear model which is explored under modeling section.

We now consider the factor variables.



Figure 2: Boxplots displaying the data structure of the factor variables against birthweight of child.

The gender of the child does not seem to be influential to the birthweight readings as the structure of the boxplot regarding male and female are fairly symmetric. This suggests that including the variable ‘sex’ in the linear model will explain little on what is responsible on affecting birthweight and we’ll see later in modeling that adding this variable increases the R-squared adjusted insignificantly implying its not necessary to include this extra complexity. The average birthweight of a child is very consistent regardless of how much the mother smokes. However, the plot also shows large variances when the mother doesn’t smoke and thus care must be taken as there could be potential data points which are influential (i.e being a high average point and an outlier).

# Modelling

We can use the best subset regression method to find one possible model such that it would involve going through all possible regression models for each number of parameters and then use Mallow’s Cp to determine which model to choose.

## Subset selection object  
## Call: regsubsets.formula(bwt ~ ., data = NewData)  
## 7 Variables (and intercept)  
## Forced in Forced out  
## age FALSE FALSE  
## gest FALSE FALSE  
## sexMale FALSE FALSE  
## smokesLight FALSE FALSE  
## smokesNo FALSE FALSE  
## weight FALSE FALSE  
## rate FALSE FALSE  
## 1 subsets of each size up to 7  
## Selection Algorithm: exhaustive  
## age gest sexMale smokesLight smokesNo weight rate  
## 1 ( 1 ) " " "\*" " " " " " " " " " "   
## 2 ( 1 ) " " "\*" " " " " " " "\*" " "   
## 3 ( 1 ) " " "\*" " " " " " " "\*" "\*"   
## 4 ( 1 ) " " "\*" "\*" " " " " "\*" "\*"   
## 5 ( 1 ) "\*" "\*" "\*" " " " " "\*" "\*"   
## 6 ( 1 ) "\*" "\*" "\*" " " "\*" "\*" "\*"   
## 7 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"

## [1] 31.697938 14.822906 9.860845 7.926588 6.259529 6.503146 8.000000

The model that is most appropriate is when Mallow’s Cp value is approximately equal to the number of parameters in the given model. This means that the best subset regression would output the model which includes all the covariates with the exception that we would have the level ‘light’ removed from the smoke covariate factor. This agrees with the boxplot in exploratory data analysis (EDA) section as the body and whiskers were short implying birthweight for light smokers were very consistent which wouldn’t alter the response much unlike the other 2 levels.

We also construct another model manually to see if it can better explain the data set. Figure 1 indicates a strong positive correlation between gestation period and birthweight so we first the linear model of birthweight as response and have covariate ‘gest’ as the independent variable.

##   
## Call:  
## lm(formula = bwt ~ gest, data = NewData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -937.13 -291.00 -17.91 266.69 1952.48   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3226.415 578.332 -5.579 5.1e-08 \*\*\*  
## gest 24.230 2.094 11.573 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 426.8 on 325 degrees of freedom  
## Multiple R-squared: 0.2918, Adjusted R-squared: 0.2897   
## F-statistic: 133.9 on 1 and 325 DF, p-value: < 2.2e-16

This simple model has a R-squared value of 0.2918 and thus this covariate alone explains just under one third of the variation in the dependent variable which agrees with the mild positive correlation in the matrix scatter plot. Figure 1 also tells us that there exists a weak positive correlation between the pre-pregnancy weight of the mother and birthweight of the child. So including the ‘weight’ covariate, the linear model becomes

##   
## Call:  
## lm(formula = bwt ~ gest + weight, data = NewData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1108.87 -271.71 -16.48 258.09 1921.24   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3474.425 566.590 -6.132 2.52e-09 \*\*\*  
## gest 22.968 2.062 11.141 < 2e-16 \*\*\*  
## weight 10.228 2.397 4.267 2.60e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 415.9 on 324 degrees of freedom  
## Multiple R-squared: 0.3295, Adjusted R-squared: 0.3254   
## F-statistic: 79.62 on 2 and 324 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Model 1: bwt ~ gest  
## Model 2: bwt ~ gest + weight  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 325 59200907   
## 2 324 56050567 1 3150340 18.21 2.6e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The anova model is useful in comparing nested models has shown significant evidence at the 0.1% level that pre-preganacy weight plays an important role in the association to birthweight. This is further supported by an increase in the adjusted R-squared value from 0.2897 to 0.3254 which implies this current model explains roughly 3% more of the variation in the response compared to that in old model nested inside the new model.

The remaining covariates namely age and rate (i.e. rate of growth of child in the first trimester) sees no correlation to birthweight in the scatterplot. However, we will still consider the model with each covariate added in in turn just to be sure and using F-test (but t-test also suitable due to one parameter difference in the nested models) gives the results below.

## Analysis of Variance Table  
##   
## Model 1: bwt ~ gest + weight  
## Model 2: bwt ~ gest + weight + age  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 324 56050567   
## 2 323 55139978 1 910589 5.3341 0.02154 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table  
##   
## Model 1: bwt ~ gest + weight  
## Model 2: bwt ~ gest + weight + rate  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 324 56050567   
## 2 323 54888563 1 1162004 6.838 0.009342 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Both tests showed that including age and rate in turn suggests evidence at the 5% and 1% level respectively that they had an association to birthweight. This highlights the importance that we shouldn’t conclude a variable is insignificant too early when it was considered in isolation. So now we will include both these covariates into the new model.

##   
## Call:  
## lm(formula = bwt ~ gest + weight + age + rate, data = NewData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1178.58 -257.12 4.69 236.19 1847.84   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3871.007 582.472 -6.646 1.29e-10 \*\*\*  
## gest 23.649 2.051 11.529 < 2e-16 \*\*\*  
## weight 8.962 2.413 3.715 0.00024 \*\*\*  
## age 8.486 4.344 1.953 0.05163 .   
## rate 1662.440 721.551 2.304 0.02186 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 410.4 on 322 degrees of freedom  
## Multiple R-squared: 0.3511, Adjusted R-squared: 0.343   
## F-statistic: 43.56 on 4 and 322 DF, p-value: < 2.2e-16

We would not expect including the covariate ‘sex’ in the linear model would help explain the data set since structure of the boxplots in figure 2 for both gender is almost symmetrical. The summary output below supports this notion as the adjusted R-squared value changes too little for this covariate to be deemed significant.

##   
## Call:  
## lm(formula = bwt ~ gest + weight + age + rate + sex, data = NewData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1134.48 -266.93 -7.93 245.35 1806.94   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3908.593 580.328 -6.735 7.57e-11 \*\*\*  
## gest 23.590 2.043 11.548 < 2e-16 \*\*\*  
## weight 9.182 2.405 3.817 0.000162 \*\*\*  
## age 8.283 4.327 1.914 0.056486 .   
## rate 1661.593 718.493 2.313 0.021376 \*   
## sexMale 87.787 45.353 1.936 0.053790 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 408.7 on 321 degrees of freedom  
## Multiple R-squared: 0.3586, Adjusted R-squared: 0.3486   
## F-statistic: 35.89 on 5 and 321 DF, p-value: < 2.2e-16

Similarly, an F-test to determine whether the factor variable smoke should be included.

## Analysis of Variance Table  
##   
## Model 1: bwt ~ gest + weight + age + rate  
## Model 2: bwt ~ gest + weight + age + rate + smokes  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 322 54245711   
## 2 320 53805046 2 440665 1.3104 0.2712

P-value is 0.2712, so we do not reject H0. No evidence to suggest that how often the mother smoked is associated with birthweight. This agrees with figure 2 such that average birthweight is very similar regardless of the smoking frequency of the mother.

# Model diagnostics

## Loading required package: carData

## [1] 137 290

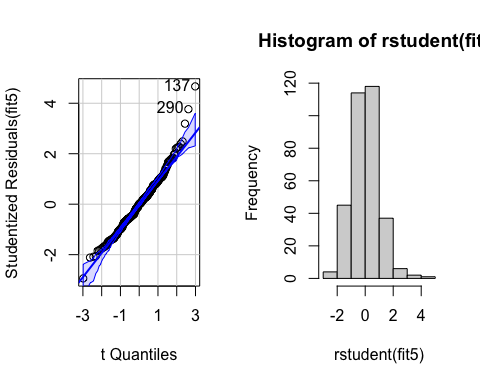


Figure 3: QQ-plot and histogram of the residuals before the transformation

The QQ-plot and the histogram of the residuals suggests a positive skew. We can reduce the skew by transforming the response variable down the ladder of powers.

## [1] 133 137

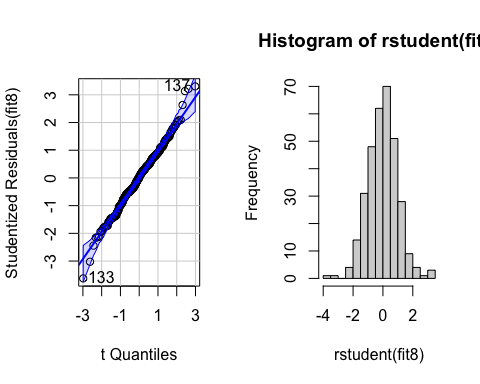


Figure 4: QQ-plot and histogram of the residuals after the transformation

Figure 4 now shows the points in the QQ-plot lying closer to the straight line implying the data is now more closely normally distributed. The histogram also suggests that the errors are normally distributed which infers that this modeling assumption is valid. Now having transformed the linear model means that the data meets more closely to the assumptions of a statistical inference procedure, important for finding reliable conclusions from this model. Now checking the residual plots:

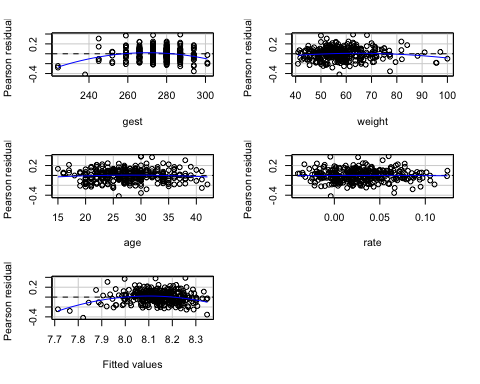


Figure 5: Pearson residual plot for each covariate and also the fitted values

Figure 5 shows non-linearity between the response and the two covariates gest and weight. The relationship could also be non-monotonic so we can try fitting a quadratic model for these covariates to see whether the linear model can approximate the data set more accurately.

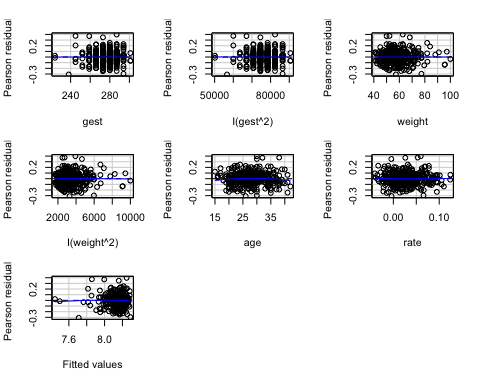


Figure 6: Pearson residual plot after adding higher order modelling terms

Figure 6 now shows that the line of best fit of the residuals is almost a horizontal line. This is an improvement from the previous model in figure 5 where we now have residuals that are more symmetrically distributed about 0 with no clear patterns of residuals. Note that these are still healthy diagnostic plots despite the data appearing to be unbalanced to the left side regarding the covariate weight and to the right side for fitted values.

In figure 1, there were a few plots regarding the concerning covariates we have now in our model which lie relatively far away from the overall trend. This could suggest that we may have extreme data points in our data set which may alter the accuracy of the model and could violate our modeling assumptions.

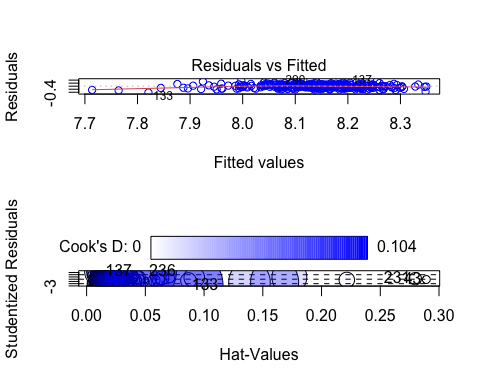


Figure 7: Circle area is proportial to Cook’s Distance

## StudRes Hat CookD  
## 43 -0.1464629 0.28919541 0.001250630  
## 133 -2.8226084 0.08548248 0.104120009  
## 137 3.5293165 0.01200103 0.020867444  
## 231 0.2342124 0.27981516 0.003053752  
## 236 3.3798158 0.04927073 0.081902972

Figure 7 does not show any large Cook’s distances which is generally considered to be greater than one. However, in comparison to other Cook’s distances, the 133rd observation appears to be significantly larger which would imply that it could potentially be a mildly-leveraged point. This is further supported in the upper plot of figure 7 where the residual for this observation shows a large residual relative to other data points implying a potential outlier. After removing this point, the new residual plots looks as follows.

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

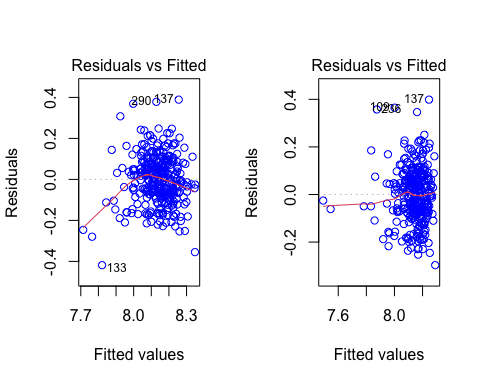


Figure 8: A fitted residual plots showing the before and after of removing an observation

Figure 8 now shows the changes in the line of best fit after removing the 133rd observation. As we can see, the red line now follows more closely to a horizontal line suggesting now that the linearity assumption holds reasonably well.

# Prediction

## Model 1

##   
## Call:  
## lm(formula = log(bwt) ~ gest + I(gest^2) + weight + I(weight^2) +   
## age + rate, data = NewData\_minus\_one)  
##   
## Coefficients:  
## (Intercept) gest I(gest^2) weight I(weight^2) age   
## -3.113e+00 7.174e-02 -1.191e-04 1.236e-02 -7.679e-05 2.614e-03   
## rate   
## 4.224e-01

## Model 2

##   
## Call:  
## lm(formula = bwt ~ age + gest + sex + smokes + weight + rate,   
## data = Births)  
##   
## Coefficients:  
## (Intercept) age gest sexMale smokesLight smokesNo   
## -4010.687 8.262 23.434 83.385 111.809 167.604   
## weight rate   
## 8.979 1711.712

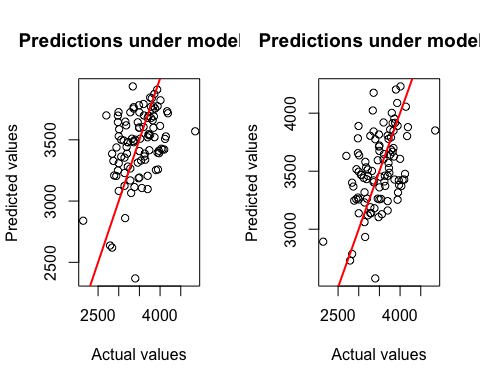


Figure 9: Predicted vs observed values under 2 different models

Model 1 has a slightly better fit as points on the plot lie closer to line of best fit (red line) implying that fitted values approximate birthweight from the new data set marginally better than model 2. Model 1 does a reasonable job at predicting birthweight but there are two individuals where this model under-predicts the birthweight (as can also be seen in model 2) in which these two points lie far right to line of best fit. It could potentially be that information about the covariates for these individuals were inputted incorrectly which led to this. Also, to say model 1 is better would be subjective as these two plots look fairly similar so one could argue that both these models are indistinguishable.

Below are the predictions of 100 individual birthweights after being given information on the explanatory variables for each individual from the test set.

## 1 2 3 4 5 6 7 8   
## 3874.283 3186.672 3586.214 3386.844 3501.016 2620.031 3774.187 3699.093   
## 9 10 11 12 13 14 15 16   
## 3204.683 3327.459 3253.161 3520.465 3490.210 3678.031 3106.466 3362.073   
## 17 18 19 20 21 22 23 24   
## 3421.043 3415.043 3725.920 3618.080 3258.004 3124.245 3354.551 3658.951   
## 25 26 27 28 29 30 31 32   
## 3082.807 3509.275 3773.086 3677.681 3494.352 3844.969 3480.648 3536.672   
## 33 34 35 36 37 38 39 40   
## 3308.450 3313.816 3716.654 3697.648 3480.124 3733.034 3281.619 3839.279   
## 41 42 43 44 45 46 47 48   
## 3600.811 3206.178 3114.341 3387.499 3394.197 3598.248 3480.989 3420.979   
## 49 50 51 52 53 54 55 56   
## 3408.016 3545.952 3386.422 3569.444 3427.208 3509.177 3507.059 3604.905   
## 57 58 59 60 61 62 63 64   
## 3529.463 3766.278 3527.648 3613.256 3724.624 3723.871 3539.041 3694.433   
## 65 66 67 68 69 70 71 72   
## 3792.272 3065.652 3745.589 3440.854 2368.801 3909.748 3266.042 3782.118   
## 73 74 75 76 77 78 79 80   
## 3330.821 3363.897 3935.897 3755.419 3648.010 3192.167 2839.615 3097.755   
## 81 82 83 84 85 86 87 88   
## 3769.947 3598.113 3500.269 3249.932 3698.604 3228.803 3754.843 3337.550   
## 89 90 91 92 93 94 95 96   
## 3521.260 3643.040 2637.337 3437.836 3208.632 3410.178 3691.719 3644.300   
## 97 98 99 100   
## 3390.475 2860.811 3822.570 3168.480

# Summary for medical professionals

After investigating the data set, we found immediately that birthweight was mildly positively correlated to gestation periods and weakly positively correlated to the pre-preganacy weight of the mother. These variables were found to be very statistically significant in the linear model and is plausible to suggest they have strong associations to birthweight.

The rate of growth of child in first semester say “R” and age of the mother say “A” were found to be less significant but also found to have an important associations to birthweight of child. Despite these variables being insignificant to birthweight when they are considered in isolation, they became significant after other important variables were accounted for in the linear model, namely gestation period and pre-pregnancy weight of mother. This could suggest that (R) and (A) may have associations to gestation periods and pre-pregnancy weight as they had no association in isolation to birthweight. We would recommend further investigations led by subject matter experts to determine whether the above holds true.