Data Analysis On Birthweights Of Children

Jake Denton

11/04/2020

Introduction

In this analysis we look at associations between birth weights of children (in grams) and other measurements. We find that length of gestation period and weight are the most significantly associated variables although each variable, apart from smoking status (the involvement of which in general is discussed separately), may have links with birth weight for this set of data. Despite these findings, the proportion of the variation of birthweight absorbed by any model is less than half.

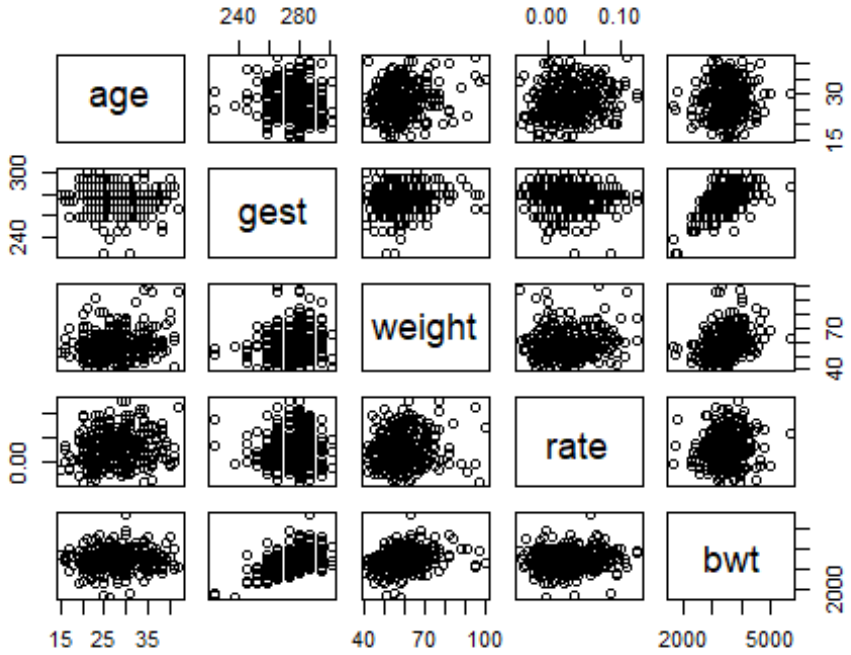
Exploratory Analysis

The data set (split into a training set (to fit a model) and a test set (to test the model)) provided relates to 7 measurements of 427 children. These measurements regard the birthweight of each child, together with other variables. These consist of: age of mother(years), length of gestation period(days), sex of child, the mother's smoking status during pregnancy, the pre-pregnancy weight of the mother(kg), rate of growth of child in the first trimester and the birthweight of the child. Smoking status and sex are factors with 3/2 levels respectively, the remaining variables are numeric measurements. Specifically, smoking status during pregnancy is either 'No' (no smoking at all), 'Light' (some smoking) or 'Heavy' whilst sex is obviously 'Female' or 'Male'. The training set consists of the data for 327 children, the test set contains the remaining 100 children's data.

```
## Loading required package: carData

##
## 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
```

We start by looking at various graphs displaying the data. First of all, let's consider pairwise scatterplots of the numerical measurements.

```r
plot(Births[,c(1,2,5,6,7)]) #produces pairwise scatterplot matrix
```




```r
summary(Births$bwt) #gives us a summary of the birthweight data
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1640    3170    3440    3461    3760    5680
```

```r
sd(Births$bwt) #gives us the standard deviation of the null model
```

```
## [1] 506.3918
```

```r
summary(Births$gest) #gives us summary statistics for gestation period
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     224     273     280     276     280     301
```

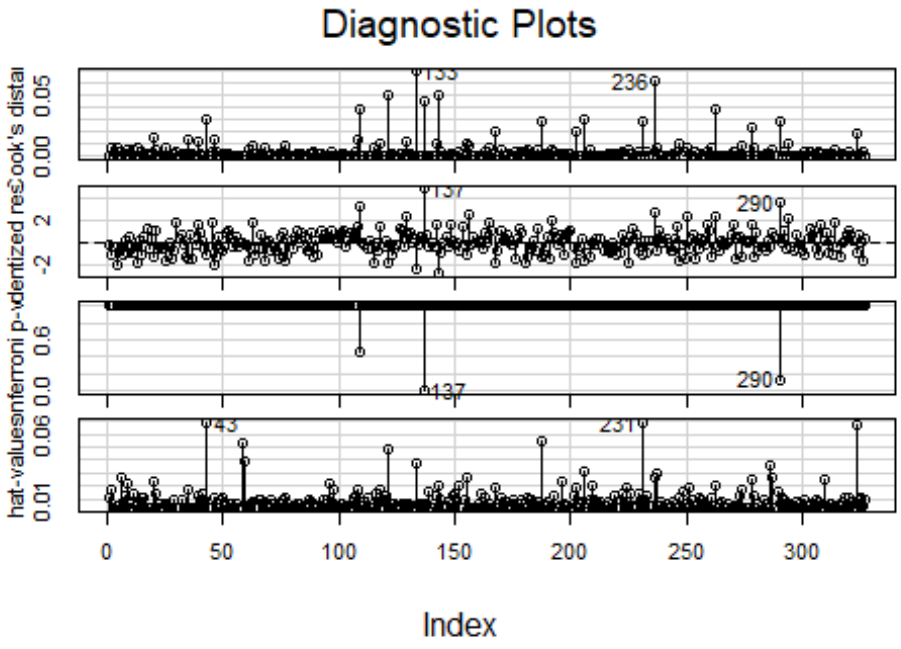
```r
fit1<-lm(bwt~weight+gest,data=Births) #fits a model with weight and gest
sum(with(Births,gest==224)) #counts no. observations with gest==224
```

```
## [1] 2
```

```r
Hat<-hatvalues(fit1) #produces vector of hat values for each observation
Hat[c(43,231)] #tells us the hat values for the observations with gest==224
```

```
##         43        231
## 0.06834957 0.06815245
```

```
D<-cooks.distance(fit1) #produces vector of cook's distances
R<-rstandard(fit1) #produces vector of studentised residuals
R[c(137)] #tells us the residual of the observation with bwt==5680

##      137
## 4.633746

qt(c(.999),df=320) #tells us 99.9th percentile of t distribution with 320
d.o.f

## [1] 3.115895

D[c(43,231)] #shows cook's distances for observations with gest==224

##         43        231
## 0.03043822 0.02942056

influenceIndexPlot(fit1) #produces influence plot for the model above
```



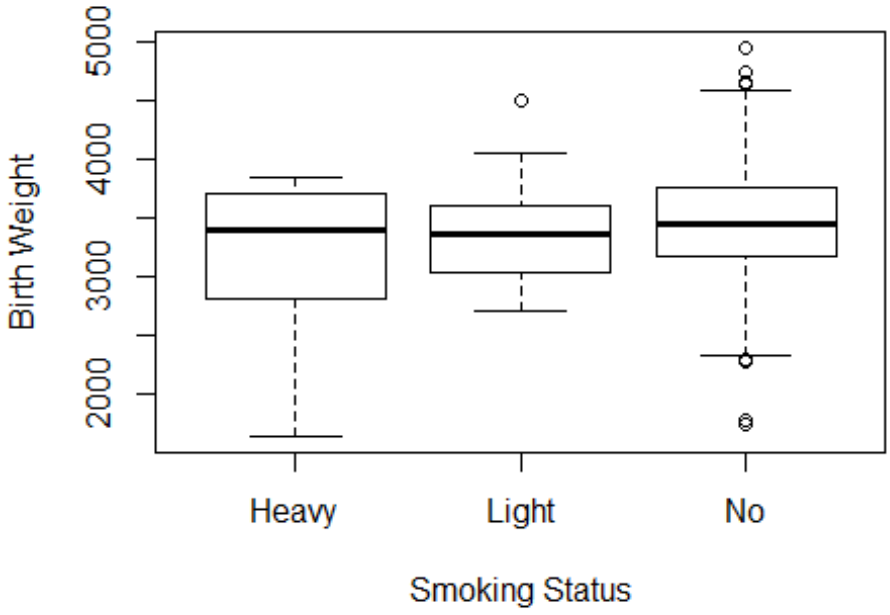

```
BirthTrain<-Births[-c(137),]#removes the outlier from the data set
```

The initial matrix of plots shows that there are clear associations between the length of gestation period, weight of mother and birth weight. A longer gestation period/larger pre-pregnancy weight of mother seem to increase birth weight. The summary statistics for the gestation period highlights that the majority of the observed gestation period values lie in the interval 273-280. However, we find two observations of 224, well below the lower limit of this interval, so we might consider that the observations with a gestation period of 224 are potentially extreme minimum values and so could be high leverage points within the

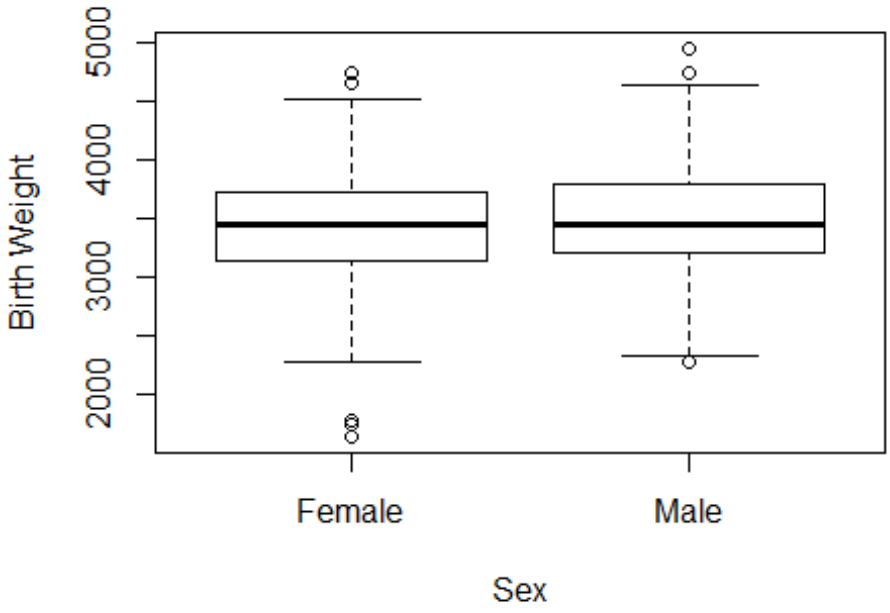
training set. To test whether these points should be included I produced some influence plots based on the model that includes the variables for weight and gestation period. The outputs tell us that these two points have the highest leverage in the training set (hat values approximately 0.068) but still low Cook distances (approximately 0.03), and so these two observations will not be removed.

Further, the summary data for the birthweight variable itself suggests the majority of children weighed between 3 and 4kg at birth. We should thus consider whether the maximum at 5.68 kilograms is an outlier. Again looking at the diagnostic plot we find that this observation (137) has the largest studentised residual (4.634 approximately), and considering the t distribution these residuals come from has 320 degrees of freedom, the 99.9th percentile of which being 3.115895, we should remove this variable from the dataset.

Next we consider the factor variables:



The boxplot for smoking shows that the bulk of the observations for each level of smoking status is similar and their median birthweight values are very close, so there is no clear significant correlation between birthweight/smoking status. However, the heavy smoking data shows a large tail towards the lower birth weights, suggesting heavy smoking may lead to decreased birth weight. Despite this, it is worth noting that there were only 13 observations of “Heavy” smoking out of the 326 observations, and 14 “Light” observations, which I would argue is insufficient to really investigate the true impact of smoking. Investigative plots with smoking and weight/gestation revealed no significant associations so have been omitted.



The boxplot of birth weight against sex reveals there is no significant association between these two variables. We do however see a number of observations of females with much lower birth weights (below 2kg) along with a few heavier male observations (above 4.5kg). Thus sex may have some influence on birth weight, which we will investigate in the modelling process.Investigative plots with sex and weight/gestation period revealed similar relationships to that of sex and birthweight. The majority of the observations sit in the same interval for both sexes for the two variables so there is nothing significant, but females have some observations where the mother has heavier pre-pregnancy weight, and also some observations with much lower gestation periods than the males.

Modelling

We begin by defining the full model and applying the stepwise regression method to obtain an improved model (based on the AIC criterion).

```
library(leaps)
fit<-lm(bwt~.,data=BirthTrain)#fits the full model
fit_step1<-step(fit) #applies stepwise regression from the full model

## Start:  AIC=3908.37
## bwt ~ age + gest + sex + smokes + weight + rate
##
##          Df Sum of Sq      RSS    AIC
## - smokes  2    356280 50314291 3906.7
## <none>                49958011 3908.4
## - sex     1    430347 50388358 3909.2
```

```
## - age     1    561136 50519147 3910.0
## - rate    1    767884 50725894 3911.3
## - weight  1   2250429 52208440 3920.7
## - gest    1  20944968 70902978 4020.5
##
## Step:  AIC=3906.69
## bwt ~ age + gest + sex + weight + rate
##
##          Df Sum of Sq      RSS    AIC
## <none>                50314291 3906.7
## - sex     1    483851 50798143 3907.8
## - age     1    564400 50878691 3908.3
## - rate    1    723174 51037465 3909.3
## - weight  1   2360306 52674597 3919.6
## - gest    1  21293786 71608077 4019.7
```

```
fit0<-lm(bwt~1,data=BirthTrain) #defines null model
fit_step2<-step(fit0, scope=bwt~age+sex+rate+weight+gest+smokes,trace=0)
#applies stepwise regression from the null model, which results in the same
model as above
summary(fit_step2)
```

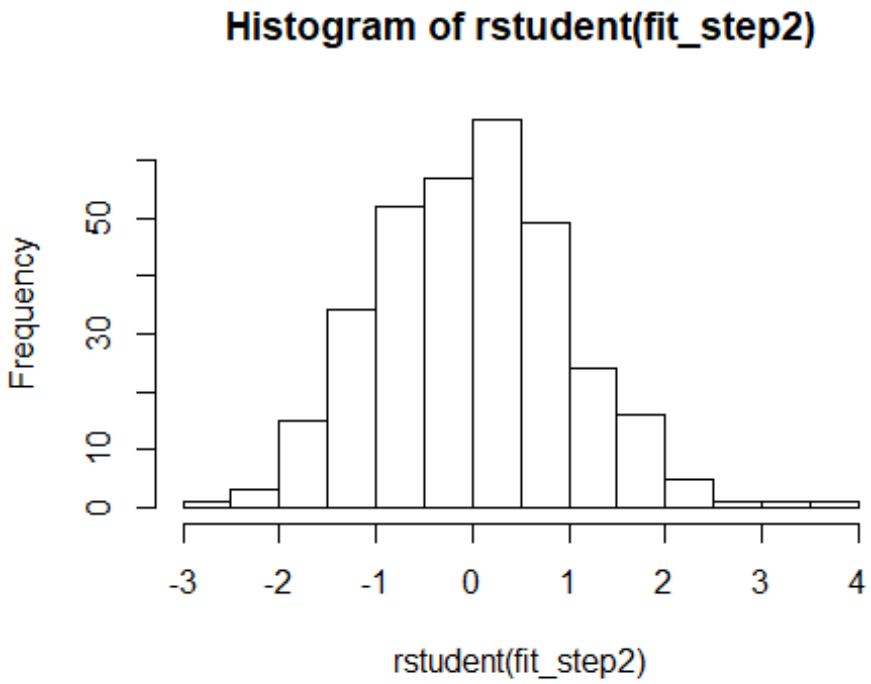
```
##
## Call:
## lm(formula = bwt ~ gest + weight + rate + age + sex, data = BirthTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1116.06  -260.75     0.29   238.93  1469.02
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3750.768    564.084  -6.649 1.27e-10 ***
## gest           23.098      1.985  11.637  < 2e-16 ***
## weight          9.042      2.334   3.874  0.00013 ***
## rate         1496.956    698.005   2.145  0.03274 *
## age             7.955      4.199   1.895  0.05904 .
## sexMale        77.293     44.061   1.754  0.08035 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 396.5 on 320 degrees of freedom
## Multiple R-squared:  0.3603, Adjusted R-squared:  0.3503
## F-statistic: 36.05 on 5 and 320 DF,  p-value: < 2.2e-16
```

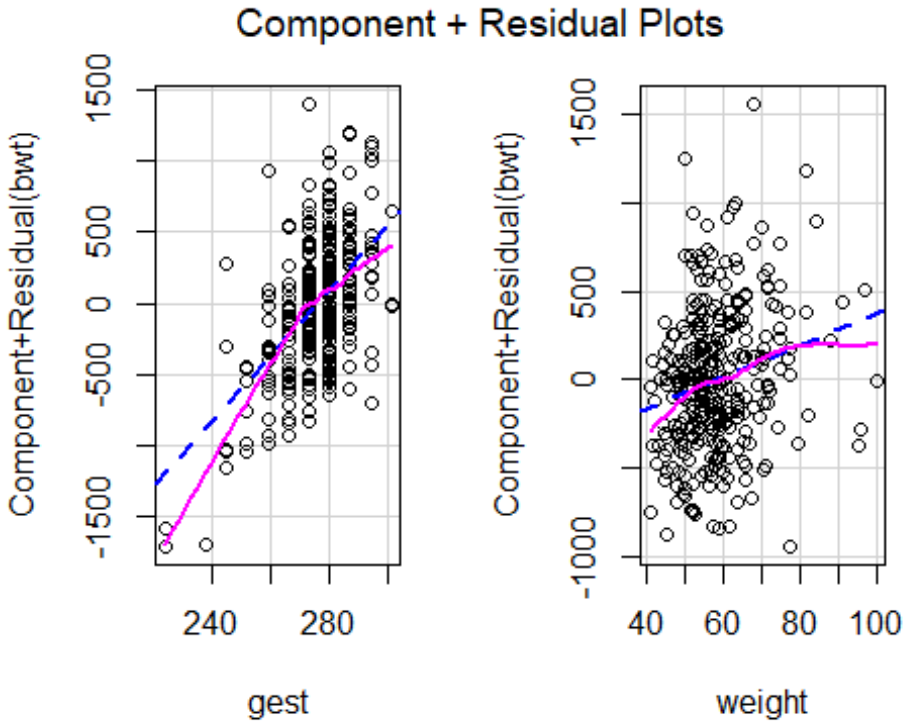
First of all it is worth noting from the step regression process that the variable for smoking status is the first to be removed but its exclusion only reduced the AIC value by 1.7. This suggests the full model is actually a reasonable model which we will see more evidence of later in this report. Also, applying the stepwise regression from the null model reveals that the first variable added is the length of gestation period (reduces AIC by over 100) followed

by weight, after which the addition of other variables only decreases AIC slightly (less than 5 so they have limited effect). This shows that the most heavily correlated variables are the length of gestation period and weight, as suspected from the scatterplots we first looked at. The summary statistics for the model obtained reveals an adjusted R squared of 0.3503 and an R squared of 0.3603 showing that the model produced only absorbs just over a third of the proportion of variation in birthweight. The estimate for the standard deviation is 396.5g (quite large, means that the model can miss the true value in the training set by 1kg). The t values show that sex and age aren't significant after controlling for the other variables at the 95% level. As a result we will conduct an analysis of variance test to see if it's worth including these variables in the model.

The low values of each R squared and high standard deviation suggests we should maybe look at other plots such as QQ-plots, histograms and residual plots in order to find which transformations of variables may benefit the model.

```
## Analysis of Variance Table
##
## Model 1: bwt ~ gest + weight + rate
## Model 2: bwt ~ gest + weight + rate + age + sex
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1    322 51387541
## 2    320 50314291  2   1073250 3.4129 0.03415 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





The analysis of variance test has a p-value of 0.034 and so we would reject the null hypothesis (that age and sex aren't correlated with birth weight) at the 95% level and so keep the model given by the stepwise regression. We keep both variables despite the analysis of variance suggesting it could only be one variable having the significant effect

(this being the alternative hypothesis) since their p values are less than 0.1 and removing either only decreases the adjusted R squared upon investigation.

A QQ-plot displays a few observations just outside the acceptance region for the normal distribution assumption, but overall this assumption is a good one and the data follows a straight line as required so this plot has been omitted. On the other hand the histogram has a positive skew in the studentised residuals. From this feature we might want to try taking logarithms/square roots of the birthweights. The component plus residual plots for gestation period/pre-pregnancy weight show non-monotone relationships, and so we could also add quadratic terms in these variables and see how it affects the model. We'll investigate the addition of these quadratic terms first.

```
fitquad<-
lm(bwt~gest+I(gest^2)+weight+rate+age+I(weight^2)+smokes+sex,data=BirthTrain)
#defines full model with quadratic terms included
fitquad_step<-step(fitquad,trace=0) #applies stepwise regression (starting
from null gives same result)
summary(fitquad_step)$coef #provides us with coefficients/t-values for model
above

##                   Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -2.538410e+04 6.593738e+03 -3.849729 0.0001428760
## gest         1.765856e+02 4.867160e+01  3.628103 0.0003323494
## I(gest^2)   -2.841000e-01 8.996053e-02 -3.158052 0.0017401714
## weight       4.086776e+01 1.718238e+01  2.378469 0.0179730934
## rate         1.385413e+03 6.887861e+02  2.011383 0.0451257529
## age          8.146921e+00 4.142450e+00  1.966692 0.0500853226
## I(weight^2) -2.485272e-01 1.325459e-01 -1.875027 0.0617024849
```

The model now contains all variables apart from the sex and smoking status, but again it is worth noting how little the AIC value changes as variables are included/excluded (largest change to AIC is by 2.18 when smoking removed in initial step, this is a difference of 0.056%). The value of adjusted R squared has increased to 0.369 so the quadratic terms have benefitted the model, although not by much (estimate for the standard deviation is 390.8, so this decreased slightly). It must also be noted that age and the quadratic term in weight aren't significant after controlling for other variables at the 95% level, however their p values are <0.1 so they would be included if one was considering a 90% test. This makes the decision of inclusion a bit more borderline, so an analysis of variance is conducted below:

```
fitquad2<-lm(bwt~gest+I(gest^2)+weight+rate,data=BirthTrain)
fitquad3<-lm(bwt~gest+I(gest^2)+weight+rate+age,data=BirthTrain)
anova(fitquad2,fitquad_step) #conducts anova test to compare the two models

## Analysis of Variance Table
##
## Model 1: bwt ~ gest + I(gest^2) + weight + rate
## Model 2: bwt ~ gest + I(gest^2) + weight + rate + age + I(weight^2)
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1    321 49794358
```

```
## 2    319 48716000  2   1078358 3.5306 0.03044 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

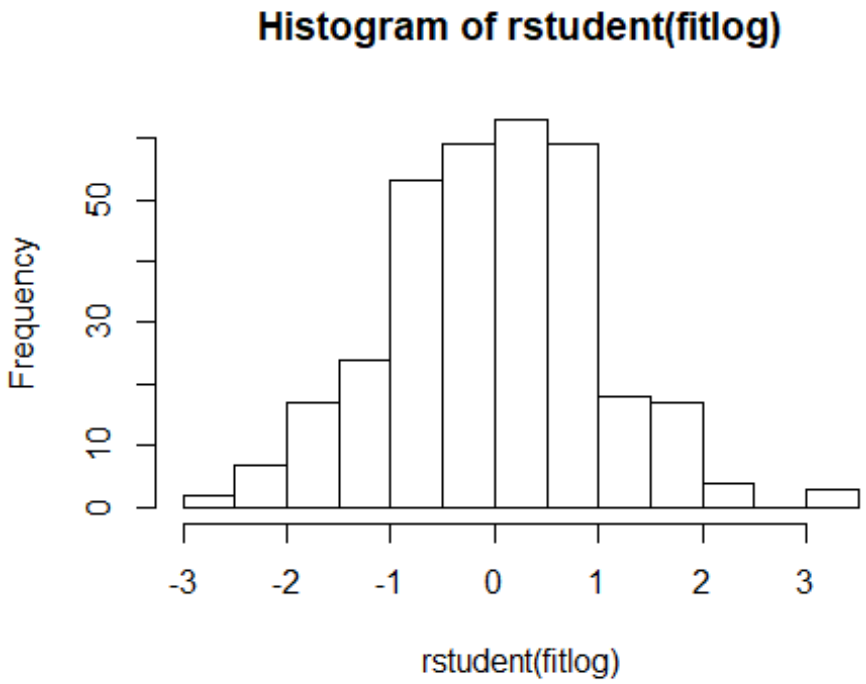
This reveals that 0.01<p-value<0.05 and so I should keep at least one of the variables in the model as there is evidence to suggest some association between them and birth weight (would reject the null hypothesis that they have no effect on birth weight at the 95% level). In fact, we'll consider the previous model with some other quadratic models that include/exclude these terms (defined in the above code chunk) and compare their predictions as the analysis of variance simply proves either one or both of the variables has an effect at the 95% level.

Now to improve the model even further we'll consider taking the logarithm of the response variable birth weight. We repeat the initial modelling steps with the transformed variable.

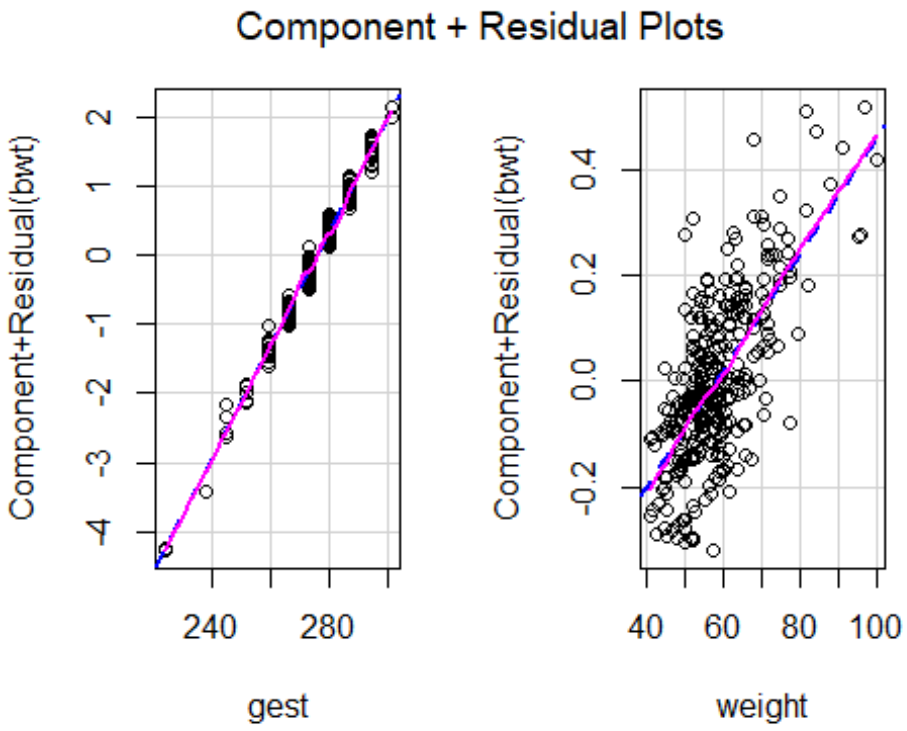
```
BirthTrain$bwt<-log(BirthTrain$bwt) #applies transformation to birth weight
fitlog0<-lm(bwt~1,data=BirthTrain)  #defines null model
fitlog<-step(fitlog0,
scope=bwt~age+sex+rate+weight+I(weight^2)+gest+I(gest^2)+smokes,trace=0)
#applies step regression from the null
summary(fitlog)$coef
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -4.472399e+00 1.912304e+00 -2.338749 1.996558e-02
## gest         8.190126e-02 1.411565e-02  5.802161 1.577990e-08
## I(gest^2)   -1.374501e-04 2.609019e-05 -5.268268 2.539473e-07
## weight       1.100834e-02 4.983202e-03  2.209090 2.787871e-02
## rate         4.195622e-01 1.997605e-01  2.100326 3.648496e-02
## age          2.613819e-03 1.201386e-03  2.175670 3.031368e-02
## I(weight^2) -6.690987e-05 3.844072e-05 -1.740599 8.271818e-02
```

```
hist(rstudent(fitlog))
```

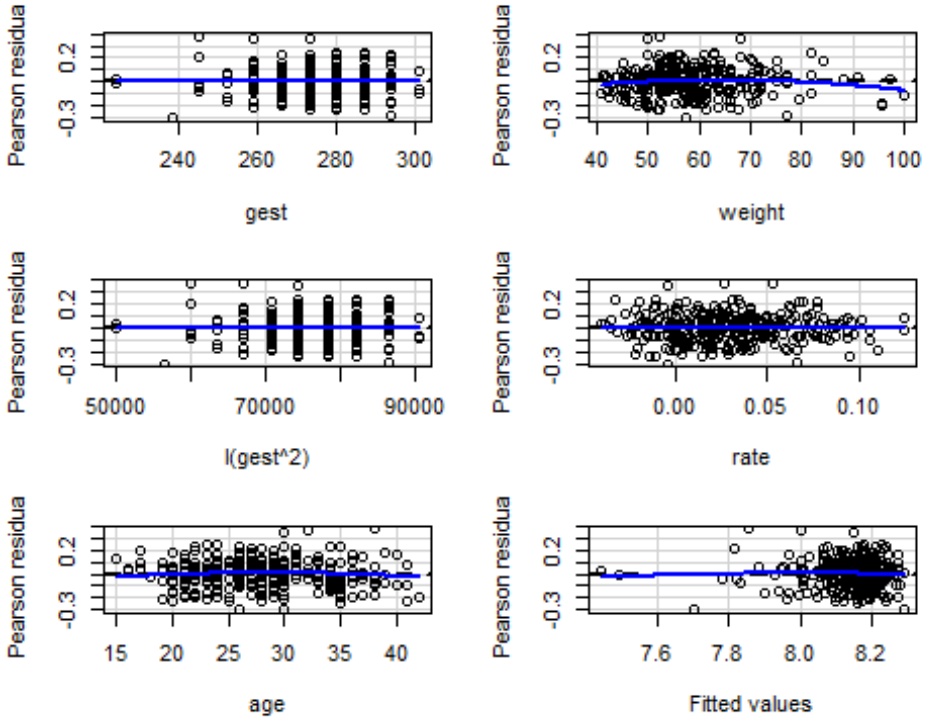



```
crPlots(fitlog,terms=~gest+weight)
```



This model involving the transformed birth weight again includes every term apart from the mother's smoking status during pregnancy, and the sex of the baby, although it is worth noting that the AIC value is not increased greatly by inclusion of the sex or smokes variables in the final step, showing again the full model fits the data reasonably well.

The plots show the effect of the transformations on birth weight, weight of mother and gestation period. The transformation led to benefits in the histogram as the data is made slightly more symmetric (although clearly some larger residuals still) whilst the component plus residual plots are monotone straight lines that follow the suggested blue dash lines nicely, especially for the gestation period variable. Furthermore, the adjusted R squared has increased to 0.4338. However, similar to before, we find that the quadratic term in weight is borderline in that it has a p-value between 0.05 and 0.1 after controlling for other variables, plus removing it reduces adjusted R squared slightly to 0.4302. As a result we could consider the predictions given by either model.



The residual plots for this model are almost ideal straight lines, with the exception of weight which is slightly curved, if we really wanted to nail down perfect residual plots we could consider adding a cubic weight instead of the quadratic and seeing if this helps.

In summary we have found that the variables with significant associations with birth weight include: length of gestation period, pre-pregnancy weight of mother, rate of growth in the first trimester, and age of the mother. Additionally, the quadratic term for gestation period is heavily associated with the response variable. Smoking status was not included in any of the final models whilst the sex of the baby was only included before any transformations were applied, and so for this training dataset there is insufficient evidence

to show strong associations between these two variables and birth weight. Nevertheless, each R squared (normal/adjusted) is less than 0.5 even for the most well-fitted model, suggesting other variables not accounted for in the data could have an effect on baby weight since these variables don't absorb the majority of its variation.

Interpreting The Model

```
summary(fitfinal)$coef
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) -4.2528409320 1.914181e+00 -2.221755 2.699905e-02
## gest         0.0822343454 1.415904e-02  5.807904 1.525885e-08
## weight       0.0024124766 6.684266e-04  3.609187 3.563489e-04
## I(gest^2)   -0.0001379991 2.617088e-05 -5.273003 2.475422e-07
## rate         0.4351937259 2.001904e-01  2.173899 3.044537e-02
## age          0.0025182398 1.203930e-03  2.091683 3.725490e-02
```

The final model (which applies to logarithm of birth weight, and supports evidence given by other models) suggests that the following things contribute to increased birth weight: Longer gestation period, larger pre-pregnancy birth weight of mother, larger rate of growth in the first trimester, and older ages of mother.

Intuition tells us that the longer a foetus remains in the mother, the more nutrition it receives from the mother which would aid growth/increase weight. Also, if the mother weighs more pre-pregnancy, its likely she'll receive good nourishment throughout pregnancy which would again allow the foetus to receive substrates for growth. The weight measurement could be interpreted as a quantitative estimate of the nutrition the baby may receive, or a suggestion of the size of the parents which could have a positive effect on growth. A high rate of growth in the first trimester is more obvious, as if the foetus continues a high rate of growth it is bound to be heavier upon being born. The age factor is a bit more dubious and further experiments with more data sets may aid the interpretation of this variable.

One variable which is included in the dataset which is counter-intuitively excluded from each model is smoking status. It has long been known that smoking even a small amount during pregnancy reduces blood flow to the foetus. Since the blood system is responsible for transportation of food and nutrients, it is easy to see why this would slow down growth rate and lead to lower birth weight. It was found to be insignificant in the modelling process but it is worth noting (as was done in the exploratory analysis), that there were less than 30 observations of any smoking during pregnancy in the training set of data (set of 327 observations). This relatively low number of observations may display some effect of smoking on birthweight (some observations where smoking was heavy were noticeably lower within the boxplot) but that effect is more than likely drowned out by the randomness of the observations where there was no smoking. In fact, considering a model of birthweight with smoking status alone, we'd find that heavy smoking is significant at the 90% level, this is shown below and tells us that not smoking adds an estimated 273 grams to the birth weight. This explains why the stepwise regression shows that the AIC value is not massively changed when smoking status is removed.

```
BirthTrain$bwt<-exp(BirthTrain$bwt) #Reverses the transformation
fitsmoke<-lm(bwt~smokes, data=BirthTrain)
summary(fitsmoke)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 3194.6154   136.0250 23.485499 7.635019e-72
## smokesLight  210.3846   188.9020  1.113723 2.662260e-01
## smokesNo     273.0803   138.9506  1.965304 5.023560e-02
```

As suggested in the summary of the previous section, the variables which were measured may not fully explain the data. A few examples of other variables that could be considered are: birth order (is this the mother's first child, some studies show that the second child can weigh more than the first), multiple births (it's not mentioned whether any of the births in the data set were twins) and alcohol abuse. All of these are factors which could potentially affect a child's birth weight.

Model Predictions

```
##                                    MSE
## Full                          156503.0
## No Quadratic Terms            156062.6
## Quadratic Terms               160022.9
## Quadratic Terms-I(weight^2)   153023.4
## Final(includes log)           155264.6
## Log model+I(weight^2)         162206.6
```

The table above displays the mean-squared error associated with a number of the best models considered in the modelling part of this report. In particular, it tells us that none of the models are very good at making predictions of the test data when compared with the full model since the magnitude of the mean-squared error for the full model does not decrease significantly upon predicting with any model. In fact, inclusion/exclusion of single variables causes large fluctuations in the mean-squared error, as can be seen in the quadratic model where removing the quadratic in weight decreases the mean squared error by 7000. Noticeably, some models predicted the test data less well than the simple full model (our quadratic model and log model). The model that obtained the best predictions of the test data was the quadratic model which excluded smoking, sex and the quadratic weight variables.

To investigate the predictions of some of the better models in comparison to the full model we'll look at some summary data regarding the errors of their predictions in the test data.

```
fullerror<-fullpredictions-TestResponses #defines vector of errors for full
model
summary(fullerror)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1020.49  -221.16    22.16    18.86   286.71   927.51
```

```
standardmodelerror<-steppredictions-TestResponses #defines vector of errors
for normal model (no transformations applied to variables)
summary(standardmodelerror)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1027.92  -217.20    26.71    18.05   278.66   919.36

quaderror<-quadpredictions2-TestResponses #defines vector of errors for best
quadratic model
summary(quaderror)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1178.530  -222.591    -3.298     2.161   282.587   977.748

logerror<-finalpredictions-TestResponses #defines vector of errors for best
log model
summary(logerror)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1218.14  -270.12   -26.61   -24.16   259.12   966.55
```

The summary data for the full model shows that the predictions were larger on average than the true values of the birthweights, with a mean error of 18.86 grams. This property of overestimating the true values is shared by the standard model (no transformed variables involved) and the quadratic model. Each of these models had the majority of error values between -220 and 280 grams, and so this shows that the full model has similar predicting power to these other models which fit the training set better. The quadratic model (which had the lowest mean square error) has a mean error of only 2 grams, so it is clear that this model should be used if any further predictions are required.

On the other hand, the logarithmic model underestimated the true birthweights from the test set by 24 grams on average with most errors lying in the interval -270 to 260 grams, a slight shift in the intervals from the other models. Each model has enormous minimums and maximums of errors which are similar in magnitude for all, which suggests that there could be more outliers in the data to be investigated, and that there are other explanatory variables to consider as discussed previously.

Summary

We found strong evidence to suggest that babies with pregnancies that involved greater length of gestation period and pre-pregnancy weight of mother along with higher rate of growth in the first trimester are heavier at birth. There is also some evidence that older age of mother increases birth weight after accounting for the other variables.

Sex was not found to be significant in most of the potential models for birth weight after controlling for the other variables. In the model where sex was involved, male babies tended to be slightly heavier at birth. Smoking status of the mother was also insignificant and was not involved in any of the final models, which was discussed as potentially due to the lack of data where mothers did smoke during pregnancy.

Lastly, we conclude that although it is clear that the most plausible explanation for the birth weights for this set of data involves the gestation period and pre-pregnancy weight of mother, other variables not included would have been helpful in forming a model that accurately predicts all the test birth weights. Further, any model investigated did not

improve on the full model a great deal or absorb a large proportion of the variation of birthweights, and so if other data sets were considered we could well find that all these variables are partly involved in birth weight.