MA-686D (Adavanced Data Analysis) Final Project

Relationship between Birth Weight of Babies and Mother's Age, Weight, Health Status, and Other Factors

Sia Bhowmick (165806290)
12/23/2016

Abstract

Low infant birth weight is a matter of concern due to the adverse effects posed by it, such as premature infant death and chronic disease later in life. The primary reasons behind this are the conditions governing the mother's health, lifestyle, and socio-economic status. In this project, a known dataset containing information on mother's condition is analyzed with respect to the weights of their children at birth. The relationship between birth weight and the explanatory variables is modeled by fitting a linear equation to the data. The significance of each predictor is analyzed and interactions between numerical and indicator variables are determined at the 95% significance level. The adequacy of the new model created is measured and effect of the leverage points and the outliers is measured. The final results show the presence of influential points that the reduced model is not adequate. Preliminary tests on the initial linear fit shows lack of covariates in the dataset. However, the linearity and normality assumption of linear regression prevailed.

Introduction

According to the World Health Organization (WHO), low birth weight is defined as weight at birth of less than 2500 grams [1]. This cut-off is based on epidemiological observation that infants weighing less than 2500 grams are more likely to die prematurely [2]. In addition to the concern of premature death, babies weighing less than 2500 grams have harder time fighting infection, eating and gaining weight, maintaining body temperature due to very little body fat. These babies are at a risk of developing neurological and gastrointestinal problems as well. Low birth weight is also closely associated with chronic disease later in life [3].

Low infant birth weight is either caused by birth before 37 weeks of gestation or by constrained fetal growth. Duration of the gestation period depends on factors such as the fetal growth, mother's health and diet, and the physical environment [2]. Mother's nutrition and diet, lifestyle (e.g., tobacco usage), history of previous premature labours, or complications such as hypertension and uterine irritability can hinder growth and development of the fetus, as well as decrease the length of the gestation period. It can be seen that low birth weight is related to babies that are born prematurely or have restricted growth in the womb, or babies that are affected by both of these conditions. Overall, it can be seen that weighing less than 2500 grams at birth is a disadvantage for the baby.

In this project, the relationship between birth weight of babies and mother's age, weight, health condition, ethnicity, and other factors will be analyzed using statistical methods. The goal is to determine the risk factors associated with low infant birth weight and fit the correct response variables to a linear regression model. The data used in this project was acquired from the dataset "birthwt" of the MASS library in R software [4].

Analysis

Data Examination

The dataset ("birthwt") studied in this project is from *Applied Logistic Regression*, written by Hosmer, D.W. and Lemeshow, S. [4]. It was accessed from the MASS library in R software. The data was collected at Baystate Medical Center, Springfield, MA, USA during 1986.

The “birthwt” data frame contains the following columns:

low - indicator of birth weight less than 2.5 kg (1 = yes, 0 = no).

age - mother's age in years.

lwt - mother's weight in pounds at last menstrual period.

race - mother's race (1 = white, 2 = black, 3 = other).

smoke - smoking status during pregnancy (1 = yes, 0 = no).

ptl - number of previous premature labours.

ht - history of hypertension (1 = yes, 0 = no).

ui - presence of uterine irritability (1 = yes, 0 = no).

ftv - number of physician visits during the first trimester.

bwt - birth weight in grams.

Here, **bwt** is the response variable and the list of predictors contain four numerical and five categorical variables. At first glance it can be seen that **low** and **bwt** are categorical and numerical ways of representing infant birth weight, so **low** should be excluded from future observations in order to avoid multicollinearity. The categorical variable **race** has three levels given by *white = 1*, *black = 2*, and *other = 3*, so **race** will contribute to two new variables. Rest of the categorical variables, **smoke**, **ht**, **ui** have two levels and so will not add any additional explanatory variables to the model.

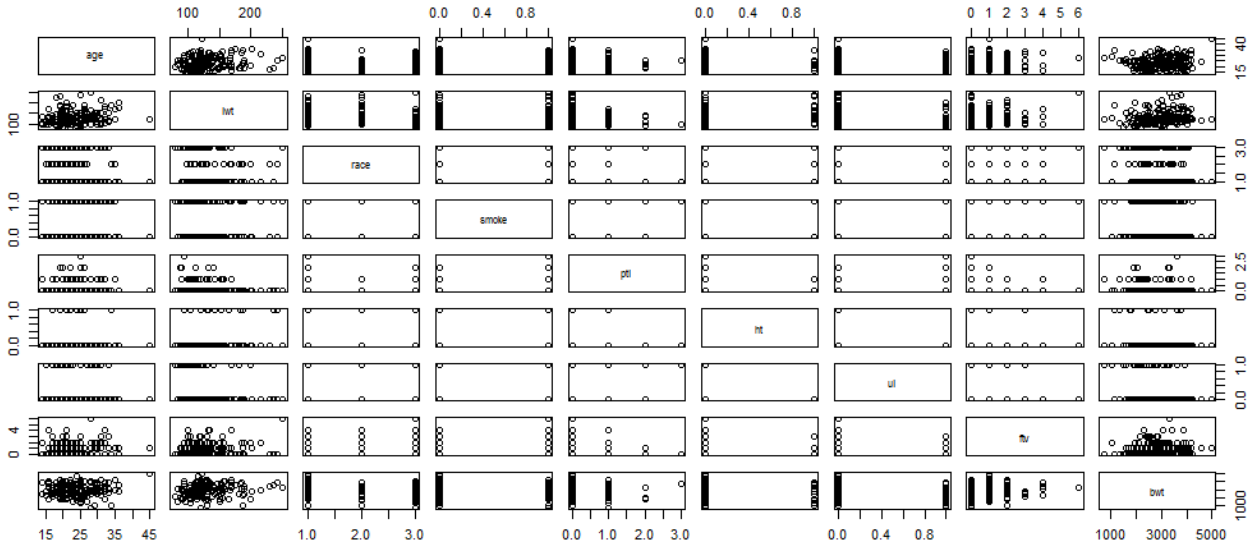

Figure 1: preliminary observations

The figure above shows the linear fit of the response variable with respect to each predictor. This is a way of determining if any of the predictors have a non-linear relationship with the response variable. As it can be see that there is no nonlinearity.

Model Specification

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to the observed data. During this process it is necessary to check the significance of each of the explanatory variables on the response variable. The dataset "birthwt" contains both numeric and indicator variables. So the significance of the interaction terms between these two categories of variables also has to be determined at the 95% level. The final model thus created is tested for adequacy using Mallow's C_p, and adjusted R^2. The dataset is also inspected for possible outliers and leverage to check for influential points.

For the “birthwt” dataset, the fitted multiple linear regression model can be written as:

$$bwt = \beta_0 + \beta_1\, age + \beta_2\, lwt + \beta_3\, race(black) + \beta_4\, race(other) + \beta_5\, smoke + \beta_6\, ptl + \beta_7\, ht + \beta_8\, ui + \beta_9\, ftv + \varepsilon$$

The following assumptions are made for a multiple linear regression model of the form $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$:

1. Linear relationship between the response variable y and the predictors $\underline{x_i}$, where $i = 1, \dots, k$.
2. Constant error variance
3. Error ($\underline{\varepsilon}$) is normally distributed
4. The error (ε_i) from each observation are independent and identically distributed (iid)
5. $X^T X$ is invertible
6. On the Q-Q plot, most of the points stay in on the $45°$ line.

The fitted model is given by:

$$\begin{aligned} fitted\ bwt = {} & 2927.962 - 3.570\ age + 4.354\ lwt - 488.428\ race(black) \\ & - 355.077\ race(other) - 352.045\ smoke - 48.402\ ptl - 592.827\ ht \\ & - 516.081\ ui - 18.058\ ftv \end{aligned}$$

It was calculated using the R command:

```
birth.mod1 <- lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv, data=birthwt)
summary(birth.mod1)
```

This fitted model was constructed using the base categories of no smoking, no history of hypertension, and no presence of uterine irritability for the categorical variables. Similarly, white was used as the base model for mother's race. It can be seen that all the categorical variables have a negative effect on the infant birth weight with respect to the base model.

In order to check if this model violates any of the assumptions stated above, couple of options are available:

1. Residual vs. fitted plot to detect nonlinearity (c.f. Figure 2)
2. Q-Q plot to test normality assumption on the error term (c.f. Figure 3)

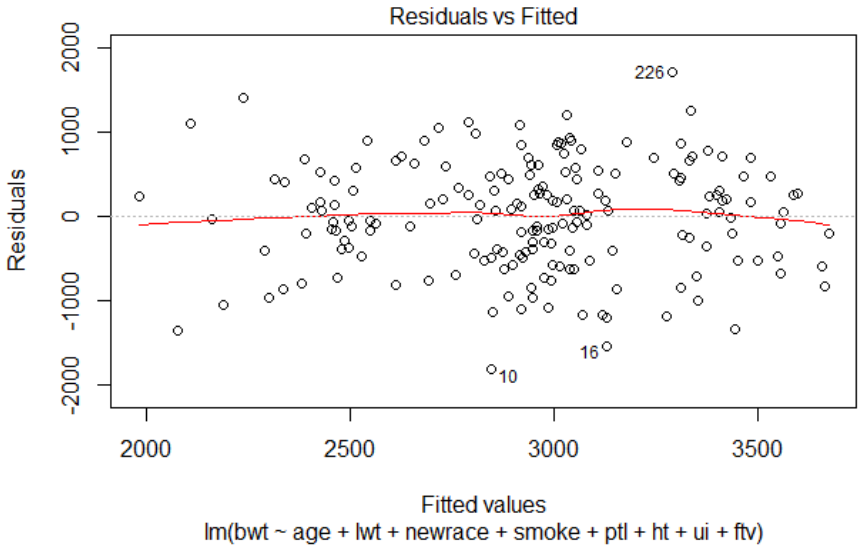

Figure 2 *

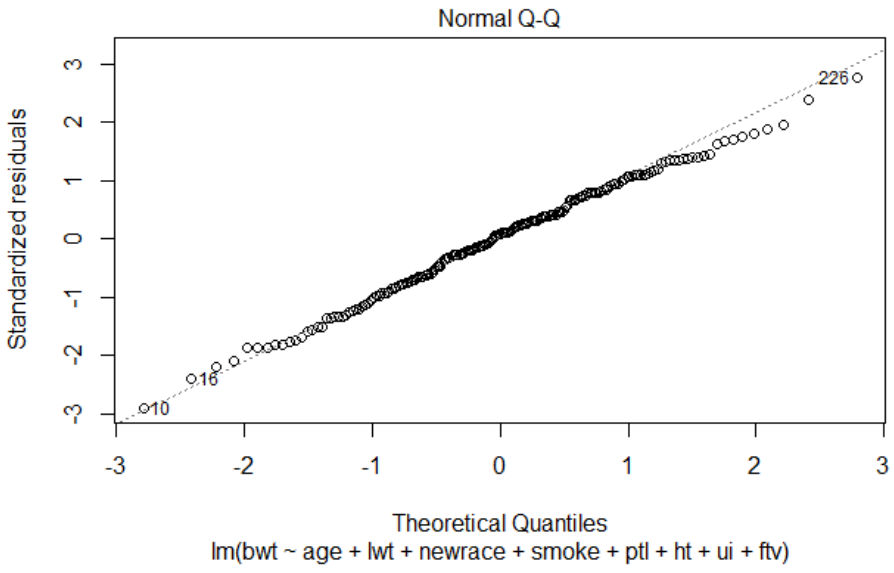

Figure 3 *

*Note that in Figure 2 and Figure 3, *newrace* accounts for two varialbes, so *newrace = newrace 2 + newrace 3*, where *newrace 2* accounts for **race** = black and *newrace 3* accounts for **race** = other (see R-code in Appendix for details)

It can be seen from these two plots that the linearity and iid assumptions have not been violated. The latter can also be confirmed using the Durbin-Watson test, where a value between 0 and 2 shows the regression errors are independent.

Using the "dwtest" command in R, the following result was obtained:

```
> # Durbin-Watson Test to see if errors are independent
> dwtest(birth.mod1)

        Durbin-Watson test

data:  birth.mod1
DW = 0.29449, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The D-W test shows that there is no autocorrelation between the regression errors. However, the residual plot in Figure 1 appears clustered, which can only mean that not all possible explanatory variables were included in this dataset. One such variable could be concerning the mother's diet and nutrition. The analysis of the model can now be focussed on the number of explanatory variables to be included in the model as well as any possible interaction terms between the numerical and the categorical variables.

Testing of Hypothesis on the Parameters

As mentioned previously, the fitted linear model (**birth.mod1**) was calculated in R using

```
birth.mod1 <- lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv, data=birthwt)

summary(birth.mod1)
```

where, the output showed the significance of each of the model parameters at 95%.

```
Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht
+ ui +
    ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1825.26  -435.21    55.91   473.46  1701.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2927.962    312.904   9.357  < 2e-16 ***
age           -3.570      9.620  -0.371 0.711012
lwt            4.354      1.736   2.509 0.013007 *
newrace2    -488.428    149.985  -3.257 0.001349 **
newrace3    -355.077    114.753  -3.094 0.002290 **
smoke       -352.045    106.476  -3.306 0.001142 **
ptl          -48.402    101.972  -0.475 0.635607
ht          -592.827    202.321  -2.930 0.003830 **
ui          -516.081    138.885  -3.716 0.000271 ***
ftv          -14.058     46.468  -0.303 0.762598
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 650.3 on 179 degrees of freedom
Multiple R-squared:  0.2427, Adjusted R-squared:  0.2047
F-statistic: 6.376 on 9 and 179 DF,  p-value: 7.891e-08
```

Null hypothesis on the full model is given by that all of the model parameters are equal to zero, and the alternate hypothesis is that at least one of them is not equal to zero. At 95% significance, if the p-value is less than 0.05, the null hypothesis is rejected. In this case, p-value = 7.98×10^{-8}, so the conclusion is that at least one of the model parameters are not equal to zero. This is also verified by the right-most column in the "coefficients" table where significance is denoted by *, **, ***.

Calculating vif using the "vif" command in R showed no presence of multicollinearity as the vif values were less than 10.

```
vif(birth.mod1)
     age      lwt newrace2 newrace3    smoke      ptl       ht       ui
  1.1551   1.2521   1.1927   1.3466   1.2070   1.1250   1.0877   1.0879
     ftv
  1.0771
```

Following this, seven different models were constructed to account for interaction terms between the numerical and categorical variables and between one numerical variable with respect to another.

The structures of these models and the hypothesis test results on the parameters are discussed below for two of the categorical-numerical cases and two of the numerical-numerical cases. Details from R-output are provided in the Appendix.

birth.mod2 - *Interaction of* ***smoke*** *with respect to the numerical variables*

bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv + smoke:age + smoke:lwt + smoke:ptl + smoke:ftv

The results showed that only the interaction term between **smoke** and **age** was significant with a coefficient of -46.354.

Hypothesis test result for parameters of the second model:

p-value = 3.028×10^{-7}, which is less than 0.05, so at least one of the model parameters are significant.

birth.mod3 - *Interaction of* ***race(black)*** *and* ***race(other)*** *with respect to the numerical variables*

bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv + newrace:age + newrace:lwt + newrace:ptl + newrace:ftv

The results showed that none of the interaction terms with respect to **race(black)** and **race(other)** was significant at 95%.

Hypothesis test result for parameters of the third model:

p-value = 1.787×10^{-5}, which is less than 0.05, so at least one of the model parameters are significant.

birth.mod6 - *Interaction of* ***age*** *with respect to the numerical variables*

bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv + age:lwt + age:ptl + age:ftv

The results showed that only the interaction term between **ftv** and **age** was significant with a coefficient of 16.993.

Hypothesis test result for parameters of the sixth model:

p-value = 1.046×10^{-7}, which is less than 0.05, so at least one of the model parameters are significant.

birth.mod7 - *Interaction of* ***lwt*** *with respect to the numerical variables*

bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv + lwt:ptl + lwt:ftv

The results showed that none of the interaction terms are significant.

Hypothesis test result for parameters of the seventh model:

p-value = 4.771×10^{-7}, which is less than 0.05, so at least one of the model parameters are significant.

Based on the significance of each parameter in **birth.mod1** and the interaction terms in models **birth.mod2** - **birth.mod8**, the following model was constructed.

semifinal.mod <- lm(bwt~ lwt + newrace + smoke + ht + ui + age:smoke + age:ht + age:ftv , data=birthwt)

From the model summary (c.f. R-output in Appendix), only **lwt**, **race(black)**, **race(other)**, and **ui** were observed to be significant at 95%. The test of hypothesis showed that the p-value at 3.815×10^{-9} was less than 0.05.

The final model then became:

$$fitted\ bwt = 2677.357 + 3.763\ lwt - 449.125\ race(black) - 229.074\ race(other) - 528.047\ ui$$

This model shows that presence of uterine irritability negatively affects the child's birth weight. The race of the mother also has a negative effect on the birth weight, compared to when the mother is white.

Testing of Adequacy

Mallow's Cp of the initial model (**birth.mod1**) is calculated as follows:

$$C_p = \frac{SSE(p)}{\hat{\sigma}^2_{full}} - [n - 2(k+1)]$$

where, k is the number of regressors in the reduced model and $p = k + 1$ is the number of parameters in the reduced model.

In order to check adequacy of the full model, its C_p should be calculated. The $\hat{\sigma}^2$ in Mallow's Cp is the residual mean square after regression on the complete set of K regressors and can be estimated by mean square error.

$$C_{p_{(full)}} = \frac{SSE(p)}{\hat{\sigma}^2_{full}} - [n - 2(k+1)]$$
$$C_{p_{(full)}} = \frac{\hat{\sigma}^2_{full}\ (n-k-1)}{\hat{\sigma}^2_{full}} - [n - 2(k+1)]$$
$$C_{p_{(full)}} = 2p - k - 1$$
$$C_{p_{(full)}} = 2(10) - 9 + 1 = 10$$

Ideally, $C_p \approx k + 1$, but for the full model $C_p = k + 1$ (always), therefore, C_p cannot be used to evaluate the full model.

From the R-output, $R^2_{adj} = 0.2047$

Mallow's Cp of the initial model (**final.mod**) is calculated as follows:

$$C_p = \frac{SSE_{red}}{\hat{\sigma}^2_{full}} - [n - 2(k_{red} + 1)]$$

$$C_p = \frac{\hat{\sigma}^2_{red}\,(n - k_{red} - 1)}{\hat{\sigma}^2_{full}} - [n - 2(k_{red} + 1)]$$

$$C_p = \frac{(679.5)^2\,(189 - 4 - 1)}{(650.3)^2} - [189 - 2(4 + 1)] = 21.895$$

From the R-output, $R^2_{adj} = 0.1317$

Leverage and Outliers

The residual vs. fitted and the Q-Q plots (Figure 2 and Figure 3, respectively) show the 10th, 16th, and 226th observations (birth weight: 1021, 1588, and 4990 grams, respectively) as outliers. The next step is to find out the leverage for each of these outliers and measure their influence on the linear model.

The leverage of an observation at $\underline{x}_i$ is defined by its $(X^TX)^{-1}$ norm: $\underline{x}_i^T (X^TX)^{-1} \underline{x}_i$, which is really the ii^{th} element of the hat matrix ($H = X(X^TX)^{-1}X\hat{}T$). Therefore the leverage of the i^{th} observation is $h_{ii} = \underline{x}_i^T (X^TX)^{-1} \underline{x}_i$. If $h_{ii} > \frac{2\,(k+1)}{n}$, the observation has high leverage.

Leverage of the outliers was calculated in R using "`lm.influence(birth.mod1)$hat`". The results for 9 explanatory variables and 189 data points are tabulated below.

	10th observation	**16th observation**	**226th observation**
$\boldsymbol{h_{ii}}^*$	0.066	0.111	0.003
Measure of leverage $\frac{\boldsymbol{2\,(k+1)}}{\boldsymbol{n}}$	0.106		

Table 1: Leverage of the outliers

* values given to three decimal places.

From Table 1 it can be seen that the 16th observation has high leverage.

Cook's distance:

To determine if the i^{th} observation is influential, the linear model is fitted without that point and the difference between least square estimates ($\underline{\hat{\beta}}$) of the full data set is compared with that $\underline{\hat{\beta}}_{(i)}$ of the data set without the i^{th} observation. In order to decide if this difference is big or small, the Cook's distance is calculated.

Cook's distance is given by the standard formula,

$$D_i = \frac{1}{(k+1)MSE}\left(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)}\right)^T (X^T X)^{-1}\left(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)}\right)$$

where, X is the design matrix, k is the number of explanatory variables, and MSE is the mean sum of square error given by $SSE/(n-k-1)$.

The following simplified version of the formula for Cook's distance will be used to calculate the influence of the 10th, 16th, and 226th observations.

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)MSE}\frac{h_{ii}}{(1-h_{ii})^2}$$

where, y_i is the birth weight of the i^{th} observation and $\hat{y}_i$ is its fitted value. For the initial linear model, $k = 9$ and $MSE = (650.3)^2$ (c.f. R-code in appendix). An observation has high influence if $D_i > 1$.

Therefore, Cook's distance for 10th, 16th, and 226th observations are:

	10th observation	**16th observation**	**226th observation**
$\boldsymbol{y_i}$	1021	1599	4990
$\boldsymbol{\hat{y}_i}$	2846.258	3129.599	3288.800
$\boldsymbol{D_i}$	0.0591	0.0183	0.0959

where, the fitted value was calculated using the "fitted" command in R.

Table 2: Cook's distance

As it can be seen from the table above, none of the outliers are influential as per Cook's distance calculation.

DFFITS:

DFFITS is used to determine the effect of the i^{th} observation on j^{th} fitted value $\hat{y}_j$. The standardized formula is given by:

$$DFFITS = \frac{y_j - \hat{y}_{j(i)}}{\sqrt{MSE_{(i)}\,(1-h_{ii})}}$$

where, $\hat{y}_{j(i)}$ is the fitted value without the i^{th} observation. Using the "dffits" command in R,

	10th observation	**16th observation**	**226th observation**
$\boldsymbol{DFFITS}$	-0.786	-0.434	0.998
Measure of influence $\boldsymbol{2\sqrt{\frac{(k+1)}{n}}}$	0.460		

Table 3: DFFITS

- If the value of $DFFITS$ is positive (negative), the effect of the i^{th} observation is to increase (decrease) the estimate of y_j.
- The i^{th} observation is influential on the j^{th} fitted value if the absolute value of $FFITS > 2\sqrt{\frac{(k+1)}{n}}$.

In this case, absolute value of $DFFITS > 0.460$, and so the 10th, 16th, and 226th observations are influential on the j^{th} fitted value. In addition, the 10th and 16th observations decrease the estimate of birth weight and the 226th observation increases it.

DFBETAS:

DFBETAS is used to determine the effect of the i^{th} observation on the estimate $\widehat{\beta_J}$. The standardized formula is given by:

$$DFBETAS_{j,(i)} = \frac{\widehat{\beta_J} - \widehat{\beta_{J_{(i)}}}}{\sqrt{MSE_{(i)}\ (X^T X)^{-1}_{jj}}}$$

Using the "dfbetas" command in R,

β_j	$DFBETAS_{j,(i)}$		
	10th observation	16th observation	226th observation
Intercept	0.0458	0.201	-0.365
age	-0.193	-0.152	0.905
lwt	0.0536	-0.171	-0.255
race(black)	0.115	0.00329	0.0416
race(other)	0.254	-0.245	-0.143
smoke	0.242	0.0247	-0.156
ptl	0.184	0.0502	-0.155
ht	-0.0504	0.102	0.00754
ui	-0.609	0.0584	-0.009
ftv	-0.209	0.178	-0.149

Table 4: DFBETAS

- Positive $DFBETAS_{j,(i)}$ values show that the i^{th} observation increases the estimate of β_j.
- Negative $DFBETAS_{j,(i)}$ values show that the i^{th} observation decreases the estimate of β_j.
- The observation is influential if the absolute value of $FBETAS_{j,(i)} > 2/\sqrt{n}$.

In this case, for 189 data points $\left|DFBETAS_{j,(i)}\right| > 0.145$. So, for example, the 16th and 226th observations decrease the estimate of **lwt** but the 10th observation increases it. However, the 16th and 226th observations are influential and not the 10th one.

Implementation of Nonparametric Estimation

Parametric regression techniques, such as multiple linear regression, pose a lot of restrictions on the functional relationship between the explanatory and response variables. Compared to these methods, non-parametric techniques are more flexible. A type of non-parametric regression called the local polynomial regression use a formula of the following form:

$$y = m(x) + \varepsilon$$

where, the goal is to obtain an estimate $\widehat{m}(x; x_0)$ of $m(x)$. In local polynomial regression, the entire dataset is used to estimate a single point x_0 . Using Taylor series expansion, the local polynomial regression model is written as:

$$m(x) = m(x_0) + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \cdots + \beta_q(x - x_0)^q + \varepsilon$$

Using the Kernel Weights formula,

$$KW_i = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - x_0}{h}\right)^2\right]$$

where, h is the bandwidth, the objective function is given by:

$$\sum_{i=1}^{n} KW_i\left[y_i - \left(\beta_1(x - x_0) + \beta_2(x - x_0)^2 + \cdots + \beta_q(x - x_0)^q\right)\right]$$

The error can be minimized using this following objective function with respect to x_0 and β_j's, and hence the estimator of $m(x_0)$ can be obtained. A new sets of weight is generated for every point in the sample which results in new least square estimators for every y_i.

Results

- Preliminary observations on the model demonstrated linear relationship between the response variable and each predictor.
- The multiple linear regression assumption of linear relationship between covariates and response variables was not violated, as well the assumption that the error terms are independent and identically distributed.
- At each stage of building the model, the null hypothesis on the model parameters was rejected at 95% significant level using the p-value.
- Lack of multicollinearity for the initial and final models was determined using vif calculations.
- The final reduced model was obtained by maintaining significance of each the parameters at 95%, however, the Mallow's Cp calculations showed that the model is not adequate.
- According to calculations done under *Leverage and Outliers*:
  - Only the 16^{th} observation has high leverage
  - Cook's distance showed that none of the outliers are influential
  - DFFITS calculation showed that the 10^{th}, 16^{th}, and 226^{th} observations are influential on the j^{th} fitted value
  - DFBETAS calculation showed a mixed trend
  - So overall, it can be said that the 10^{th}, 16^{th}, and 226^{th} observations are influential.

Discussion and Recommendations

The data analysis showed presence of influential observations. These observations were used throughout the process of determining the final model, which did not turn out to be adequate as per the high Mallow's Cp value and low R^2_{adj} value. The final model could be improved by eliminating the influential observations in the very beginning. Alternatively, a local polynomial regression model could be implemented instead of the widely used multiple linear regression model. This dataset could also include other important covariates which contribute to decreasing infant birth weight.

References

1. World Health Organization, *International statistical classification of diseases and related health problems*, tenth revision, World Health Organization, Geneva, 1992.
2. Kramer, M.S., "Determinants of Low Birth Weight: Methodological assessment and meta-analysis", *Bulletin of the World Health Organization*, vol. 65, no. 5, 1987, pp. 663–737.
3. Barker, D.J.P. (ed.), Fetal and infant origins of disease, BMJ Books, London, 1992.
4. Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression.* New York: Wiley.

Appendix (R-code and output)

```
> library(MASS)
> data(birthwt)
> attach(birthwt)
The following objects are masked from birthwt (pos = 4):

    age, bwt, ftv, ht, low, lwt, ptl, race, smoke, ui

> library(DAAG)
Error in loadNamespace(j <- i[[1L]], c(lib.loc, .libPaths()), versionCheck = vI[[j]]) :
  there is no package called 'RColorBrewer'
Error: package or namespace load failed for 'DAAG'
>
> # low (indicator) represents the same thing as bwt(numerical), so we don't consider low in the
model.
> birthwt$low<- NULL
>
> # for black = 2 and other = 3 in race
> newrace <- factor(race)
>
> pairs(birthwt)
>
> # preliminary model
> birth.mod1 <- lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv
+                  , data=birthwt)
> summary(birth.mod1)

Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht + ui +
    ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1825.26  -435.21    55.91   473.46  1701.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2927.962    312.904   9.357  < 2e-16 ***
age           -3.570      9.620  -0.371 0.711012
lwt            4.354      1.736   2.509 0.013007 *
newrace2    -488.428    149.985  -3.257 0.001349 **
newrace3    -355.077    114.753  -3.094 0.002290 **
smoke       -352.045    106.476  -3.306 0.001142 **
ptl          -48.402    101.972  -0.475 0.635607
ht          -592.827    202.321  -2.930 0.003830 **
ui          -516.081    138.885  -3.716 0.000271 ***
ftv          -14.058     46.468  -0.303 0.762598
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 650.3 on 179 degrees of freedom
Multiple R-squared:  0.2427, Adjusted R-squared:  0.2047
F-statistic: 6.376 on 9 and 179 DF,  p-value: 7.891e-08

> options(max.print=1000000)
> dfbetas(birth.mod1)

      (Intercept)           age           lwt      newrace2      newrace3         smoke
ptl
```

```
226 -3.654085e-01  9.050149e-01 -2.552825e-01  4.157626e-02 -0.1433483212 -0.1555820521 -
0.1547486467
10   4.579341e-02 -1.930279e-01  5.633451e-02  1.153405e-01  0.2536761428  0.2419141562
0.1839969703
16   2.009095e-01 -1.520890e-01 -1.712891e-01  3.292925e-03 -0.2453139602  0.0246712831
0.0502433597
               ht            ui           ftv
226  0.0075406264 -9.091848e-03 -0.1494038053
10  -0.0504369318 -6.088849e-01 -0.2090046797
16   0.1021340491  5.838767e-02  0.1779752640

 [...]

> vif(birth.mod1)
     age      lwt newrace2 newrace3    smoke      ptl       ht       ui
  1.1551   1.2521   1.1927   1.3466   1.2070   1.1250   1.0877   1.0879
     ftv
  1.0771
>
> # 10th observation is at position 132
>
> lm.influence(birth.mod1)$hat[132] # hat value
        10
0.06555205
> dffits(birth.mod1)[132]
        10
-0.7855851
> dfbetas(birth.mod1)[132]
[1] 0.04579341
> fitted(birth.mod1)[132] # fitted value
      10
2846.258
>
> # 16th observation is at position 136
>
> lm.influence(birth.mod1)$hat[136] # hat value
        16
0.03063097
> dffits(birth.mod1)[136]
        16
-0.4338771
> dfbetas(birth.mod1)[136]
[1] 0.2009095
> fitted(birth.mod1)[136] # fitted value
      16
3129.599
>
> # 226th observation is at position 130
>
> lm.influence(birth.mod1)$hat[130] # hat value
     226
0.110768
> dffits(birth.mod1)[130]
      226
0.9980329
> dfbetas(birth.mod1)[130]
[1] -0.3654085
> fitted(birth.mod1)[130] # fitted value
   226
3288.8
>
> # residual vs. fitted
> plot(birth.mod1, which=1)
>
> # normal Q-Q plot
```

```
> plot(birth.mod1, which=2)
>
> # Durbin-Watson Test to see if errors are independent
> dwtest(birth.mod1)

        Durbin-Watson test

data:  birth.mod1
DW = 0.29449, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
>
> # interaction wrt smoke
> birth.mod2 <- lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv
+                  + smoke:age + smoke:lwt + smoke:ptl + smoke:ftv , data=birthwt)
> summary(birth.mod2)

Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht + ui +
    ftv + smoke:age + smoke:lwt + smoke:ptl + smoke:ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1848.14  -428.88    66.27   461.81  1362.98

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2544.011    404.159   6.295 2.40e-09 ***
age           12.869     12.087   1.065  0.28847
lwt            4.229      2.347   1.802  0.07330 .
newrace2    -408.340    153.447  -2.661  0.00851 **
newrace3    -313.266    119.216  -2.628  0.00936 **
smoke        697.848    606.482   1.151  0.25145
ptl         -116.918    165.358  -0.707  0.48047
ht          -582.229    205.262  -2.837  0.00510 **
ui          -565.330    140.970  -4.010 8.97e-05 ***
ftv          -16.244     64.774  -0.251  0.80227
age:smoke    -46.354     20.402  -2.272  0.02430 *
lwt:smoke     -0.212      3.371  -0.063  0.94991
smoke:ptl    147.453    209.129   0.705  0.48170
smoke:ftv     41.160     96.010   0.429  0.66867
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 647.6 on 175 degrees of freedom
Multiple R-squared:  0.2658, Adjusted R-squared:  0.2113
F-statistic: 4.873 on 13 and 175 DF,  p-value: 3.028e-07

>
> # interaction wrt newrace
> birth.mod3 <- lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv
+                  + newrace:age + newrace:lwt + newrace:ptl + newrace:ftv, data=birthwt)
> summary(birth.mod3)

Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht + ui +
    ftv + newrace:age + newrace:lwt + newrace:ptl + newrace:ftv,
    data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1814.94  -450.46    10.47   452.64  1543.91

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2909.8068   444.7734   6.542 6.76e-10 ***
age             4.5122    13.1965   0.342 0.732826
lwt             2.9132     2.5057   1.163 0.246610
newrace2     -164.2889   854.8885  -0.192 0.847833
newrace3     -437.1704   683.5007  -0.640 0.523285
smoke        -339.8628   112.3291  -3.026 0.002865 **
ptl            24.5761   135.5035   0.181 0.856293
ht           -532.5286   214.5497  -2.482 0.014027 *
ui           -533.2638   144.7542  -3.684 0.000308 ***
ftv           -25.0880    72.8484  -0.344 0.730979
age:newrace2  -28.4525    32.7047  -0.870 0.385530
age:newrace3  -14.7934    23.2997  -0.635 0.526329
lwt:newrace2    2.3764     4.3394   0.548 0.584655
lwt:newrace3    3.7751     4.4399   0.850 0.396370
newrace2:ptl  -99.0955   417.6440  -0.237 0.812729
newrace3:ptl -190.4686   217.3177  -0.876 0.382014
newrace2:ftv   -1.6401   146.6033  -0.011 0.991087
newrace3:ftv   -0.8418   108.5843  -0.008 0.993824
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 660.1 on 171 degrees of freedom
Multiple R-squared:  0.2546, Adjusted R-squared:  0.1805
F-statistic: 3.436 on 17 and 171 DF,  p-value: 1.787e-05

>
> # interaction wrt ht
> birth.mod4 <- lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv
+                  + ht:age + ht:lwt + ht:ptl + ht:ftv, data=birthwt)
> summary(birth.mod4)

Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht + ui +
    ftv + ht:age + ht:lwt + ht:ptl + ht:ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1832.32  -418.06    42.85   458.40  1583.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2945.501    326.783   9.014 3.41e-16 ***
age            2.134      9.847   0.217 0.828648
lwt            3.056      1.857   1.645 0.101671
newrace2    -444.859    151.785  -2.931 0.003832 **
newrace3    -318.778    115.001  -2.772 0.006175 **
smoke       -368.128    106.854  -3.445 0.000714 ***
ptl          -32.007    102.986  -0.311 0.756333
ht           451.979   1282.660   0.352 0.724980
ui          -528.954    137.687  -3.842 0.000171 ***
ftv          -11.208     47.387  -0.237 0.813311
age:ht       -98.663     47.973  -2.057 0.041205 *
lwt:ht         8.086      4.866   1.662 0.098377 .
ptl:ht      -137.877    560.685  -0.246 0.806041
ht:ftv         2.731    230.354   0.012 0.990556
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 643.4 on 175 degrees of freedom
Multiple R-squared:  0.2754, Adjusted R-squared:  0.2216
F-statistic: 5.117 on 13 and 175 DF,  p-value: 1.148e-07

> # interaction wrt ui
> birth.mod5 <- lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv
```

```
+                  + ui:age + ui:lwt + ui:ptl + ui:ftv, data=birthwt)
> summary(birth.mod5)

Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht + ui +
    ftv + ui:age + ui:lwt + ui:ptl + ui:ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1587.94  -429.57    34.64   477.42  1620.92

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2819.0193   322.1064   8.752 1.73e-15 ***
age           -0.2237    10.2557  -0.022 0.982622
lwt            4.6233     1.8607   2.485 0.013904 *
newrace2    -416.3005   154.1273  -2.701 0.007592 **
newrace3    -330.5376   115.8149  -2.854 0.004838 **
smoke       -371.8584   109.3735  -3.400 0.000835 ***
ptl         -182.0194   132.2206  -1.377 0.170384
ht          -601.4291   203.2368  -2.959 0.003511 **
ui           652.7871  1060.6200   0.615 0.539039
ftv           -8.5363    48.6067  -0.176 0.860795
age:ui       -25.6436    35.3040  -0.726 0.468585
lwt:ui        -6.1981     5.2077  -1.190 0.235591
ptl:ui       351.5295   208.1585   1.689 0.093047 .
ui:ftv        41.4262   182.2309   0.227 0.820434
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 649 on 175 degrees of freedom
Multiple R-squared:  0.2626, Adjusted R-squared:  0.2078
F-statistic: 4.793 on 13 and 175 DF,  p-value: 4.172e-07


>
> vif(birth.mod5)
     age      lwt newrace2 newrace3    smoke      ptl       ht       ui
  1.3179   1.4448   1.2645   1.3770   1.2786   1.8989   1.1019  63.6920
     ftv   age:ui   lwt:ui   ptl:ui   ui:ftv
  1.1831  36.7570  22.9770   2.2737   2.5451
>
> # interaction of age wrt the numeric terms
> birth.mod6 <-lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv
+                 + age:lwt + age:ptl + age:ftv , data=birthwt)
> summary(birth.mod6)

Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht + ui +
    ftv + age:lwt + age:ptl + age:ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1875.78  -442.79    38.91   423.11  1613.17

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1978.6897  1002.9474   1.973 0.050076 .
age           36.5028    41.1737   0.887 0.376529
lwt           14.4953     7.2817   1.991 0.048067 *
newrace2    -461.1709   150.5754  -3.063 0.002538 **
newrace3    -349.6886   114.5788  -3.052 0.002626 **
smoke       -366.5877   106.2107  -3.452 0.000698 ***
ptl         -259.8066   572.6531  -0.454 0.650611
ht          -596.8288   201.4494  -2.963 0.003472 **
ui          -547.9240   140.1208  -3.910 0.000131 ***
```

```
ftv         -425.2102   211.4515  -2.011 0.045861 *
age:lwt       -0.4276     0.2949  -1.450 0.148829
age:ptl        9.0444    23.6340   0.383 0.702415
age:ftv       16.9931     8.5505   1.987 0.048431 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 645 on 176 degrees of freedom
Multiple R-squared:  0.2676, Adjusted R-squared:  0.2177
F-statistic: 5.359 on 12 and 176 DF,  p-value: 1.046e-07

>
> vif(birth.mod6)
     age      lwt newrace2 newrace3    smoke      ptl       ht       ui
 21.5100  22.4070   1.2221   1.3649   1.2210  36.0700   1.0963   1.1257
     ftv  age:lwt  age:ptl  age:ftv
 22.6730  48.7540  36.3110  25.1890
>
> # interaction of lwt wrt remainder of the numeric terms
> birth.mod7 <-lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv
+                 + lwt:ptl + lwt:ftv , data=birthwt)
> summary(birth.mod7)

Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht + ui +
    ftv + lwt:ptl + lwt:ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1819.61  -428.44    59.03   463.66  1699.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2985.4161   368.0567   8.111 8.07e-14 ***
age           -3.5869     9.6804  -0.371  0.71143
lwt            3.9424     2.2010   1.791  0.07497 .
newrace2    -489.2554   150.7837  -3.245  0.00141 **
newrace3    -362.0769   119.5940  -3.028  0.00283 **
smoke       -354.3871   109.7043  -3.230  0.00147 **
ptl         -262.2399   490.7878  -0.534  0.59379
ht          -585.2626   205.9336  -2.842  0.00501 **
ui          -519.7760   139.8142  -3.718  0.00027 ***
ftv          -39.2879   154.0147  -0.255  0.79895
lwt:ptl        1.8531     4.1584   0.446  0.65641
lwt:ftv        0.1733     1.0103   0.172  0.86402
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 653.6 on 177 degrees of freedom
Multiple R-squared:  0.2437, Adjusted R-squared:  0.1967
F-statistic: 5.184 on 11 and 177 DF,  p-value: 4.771e-07

>
> # interaction of ptl wrt remainder of the numeric terms
> birth.mod8 <-lm(bwt~ age + lwt + newrace + smoke + ptl + ht + ui + ftv
+                 + ptl:ftv , data=birthwt)
> summary(birth.mod8)

Call:
lm(formula = bwt ~ age + lwt + newrace + smoke + ptl + ht + ui +
    ftv + ptl:ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1823.83  -433.29    55.64   467.88  1702.18
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2928.355    313.851   9.330  < 2e-16 ***
age           -3.589      9.653  -0.372 0.710454
lwt            4.359      1.742   2.502 0.013267 *
newrace2    -488.986    150.708  -3.245 0.001405 **
newrace3    -355.387    115.196  -3.085 0.002360 **
smoke       -351.262    107.613  -3.264 0.001317 **
ptl          -52.096    120.276  -0.433 0.665440
ht          -593.010    202.911  -2.923 0.003923 **
ui          -515.771    139.375  -3.701 0.000287 ***
ftv          -15.159     50.278  -0.302 0.763374
ptl:ftv        6.532    111.974   0.058 0.953547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 652.1 on 178 degrees of freedom
Multiple R-squared:  0.2428, Adjusted R-squared:  0.2002
F-statistic: 5.706 on 10 and 178 DF,  p-value: 2.089e-07

>
> semifinal.mod <- lm(bwt~ lwt + newrace + smoke + ht + ui + age:smoke + age:ht + age:ftv ,
data=birthwt)
> summary(semifinal.mod)

Call:
lm(formula = bwt ~ lwt + newrace + smoke + ht + ui + age:smoke +
    age:ht + age:ftv, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1821.79  -431.62    51.41   455.56  1629.07

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2778.760    241.835  11.490  < 2e-16 ***
lwt            4.457      1.681   2.651  0.00875 **
newrace2    -426.627    145.863  -2.925  0.00389 **
newrace3    -318.968    112.044  -2.847  0.00493 **
smoke        274.202    388.770   0.705  0.48154
ht          1130.243   1055.807   1.071  0.28584
ui          -559.149    134.554  -4.156 5.02e-05 ***
smoke:age    -26.983     16.220  -1.664  0.09795 .
ht:age       -75.376     45.375  -1.661  0.09843 .
age:ftv        0.755      1.809   0.417  0.67690
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 638.1 on 179 degrees of freedom
Multiple R-squared:  0.2709, Adjusted R-squared:  0.2343
F-statistic: 7.392 on 9 and 179 DF,  p-value: 3.815e-09

>
> vif(semifinal.mod)
lwt  newrace2  newrace3     smoke        ht        ui smoke:age
   1.2204    1.1717    1.3334   16.7140   30.7670    1.0606   16.5200
   ht:age   age:ftv
  30.9410    1.1517
>
> # final model: should be written as: final.mod <- lm(bwt~ lwt + newrace2 + newrace3 + ui,
data=birthwt)
> final.mod <- lm(bwt~ lwt + newrace + ui, data=birthwt)
> summary(final.mod)
```

```
Call:
lm(formula = bwt ~ lwt + newrace + ui, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-1796.85  -480.50    38.32   465.68  1849.83

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2677.357    239.335  11.187  < 2e-16 ***
lwt            3.763      1.708   2.202 0.028881 *
newrace2    -449.125    152.286  -2.949 0.003599 **
newrace3    -229.074    110.118  -2.080 0.038886 *
ui          -528.047    140.846  -3.749 0.000238 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 679.5 on 184 degrees of freedom
Multiple R-squared:  0.1502, Adjusted R-squared:  0.1317
F-statistic:  8.13 on 4 and 184 DF,  p-value: 4.657e-06

>
> vif(final.mod)
     lwt newrace2 newrace3       ui
  1.1114   1.1263   1.1358   1.0248
```