**CASE STUDY**

Spring 2012

**Modeling the Relationship Between Birth Weight and Maternal Factors**

*Kaymal,*

*Ozcan*

**Table of Contents**

**Part 1**

|  |  |
| --- | --- |
| Introduction | 1 |
| Data Description | 2 |
| Analysis | 4 |
| ANOVA | 4 |
| Regression Analysis | 5 |

**Part 2**

|  |  |
| --- | --- |
| Model Adequacy Checking | 7 |
| Box-Tidwell Transformation | 8 |
| Box-Cox Transformation | 10 |
| Checking for Multicollinearity | 12 |
|  |  |

**Part 3**

|  |  |
| --- | --- |
| Variable Selection | 13 |
| Model Validation | 14 |
| Conclusion | 17 |
|  |  |

**Appendix**

|  |  |
| --- | --- |
| Major Assumptions of CLR Model | A-1 |
| Pairs plot for the model after Box-Tidwell transformation | A-1 |
| The Normal Q-Q Plot of the residuals after Box-Cox Transformation | A-2 |
| Techniques for Dealing with Multicollinearity | A-2 |
|  |  |

**PART 1**

**Introduction**

Every year there are approximately 129,108,390 births in the world[[1]](#footnote-1). That is, in every second almost 4 women give birth to a baby. Some of the most common questions asked before and after the delivery of the child are: What is his/her weight? Is it low?

In this project we will not answer those questions directly; however, we will try to give some answers to the question “What affects the birth weight?” so that one can make a good guess on the weight of a child utilizing specific information about his/her mother. We will derive results of the data using regression analysis, ANOVA and some other techniques to do this.



Photograph: [http://static.guim.co.uk](http://static.guim.co.uk/)

***Figure-1***

It is very hard, in fact almost impossible to analyze every factor that affects the weight of a newborn infant. These would include environmental conditions, genetical affects and maternal issues. Our purpose in this project is to analyze data related to maternal attributes, therefore we will only be using several numerical and categorical data related to mothers and try to learn whether these data -or certain characteristics of a woman- effect the birth weight or not.

But why do we care about the weight of our baby? Birth weight plays a significant role in survival, health, and development of an infant and it is an important indicator of the health.[[2]](#footnote-2)

In the next section we will give more detailed information about the maternal factors and weight data.

**Data Description**

Here is the data which contains observations related to 2500 children that were born in King County/UK in 2001.[[3]](#footnote-3)

***Table-1****: Variables to be used for data analysis.*

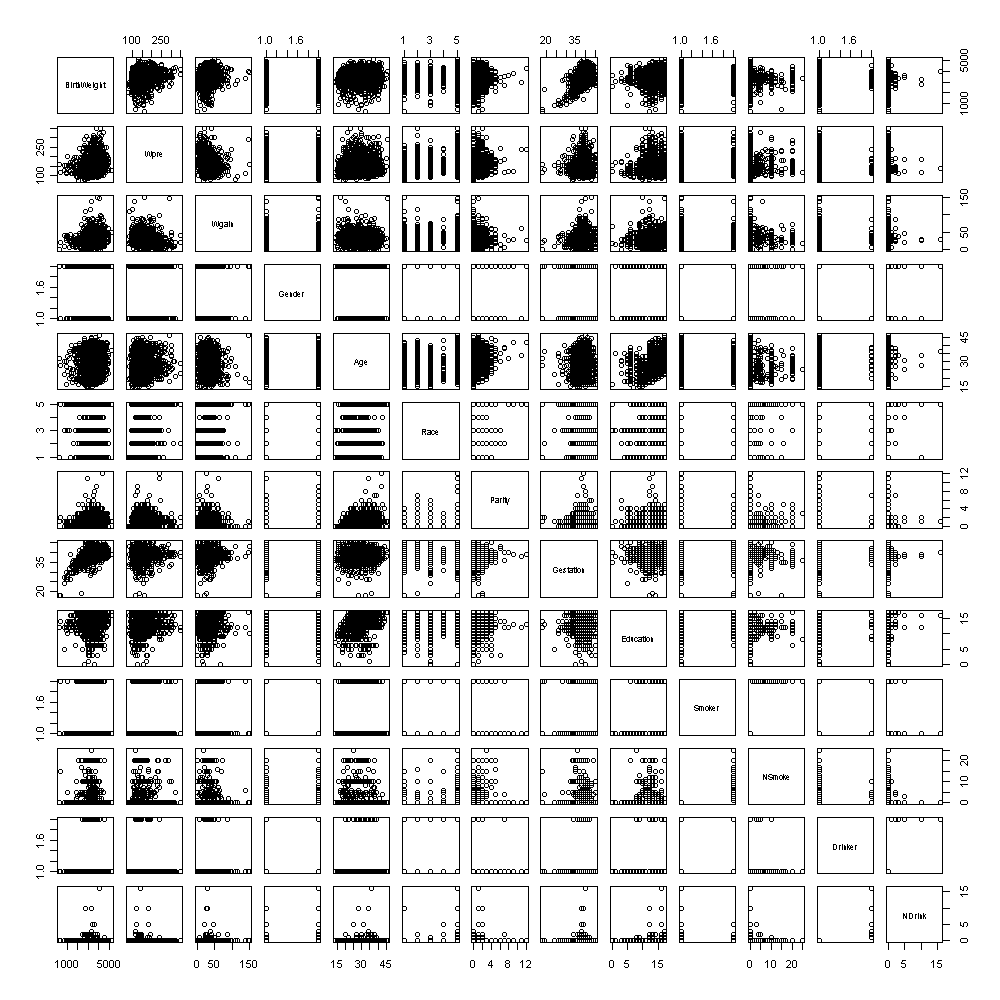
|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Variable** | **Label** | **Type of Variable** | **Description** |
| BirthWeight | y | Continuous - Numeric | Birth weight in grams |
| Wpre | x1 | Continuous - Numeric | Mother's weight in pounds prior to pregnancy |
| Wgain | x2 | Continuous - Numeric | Mother's weight gain in pounds during pregnancy |
| Gender | x3 | Categorical | M = male, F = female baby |
| Age | x4 | Continuous - Numeric | Mother's age in years |
| Race | x5 | Categorical | Race categories (for mother) |
| Parity | x6 | Continuous - Numeric | Number of previous live born infants |
| Education | x7 | Continuous - Numeric | Highest grade completed (add 12 + 1 / year of college) |
| Gestation | x8 | Continuous - Numeric | Weeks from last menses to birth of child |
| Smoker | x9 | Categorical | Y = yes, N = no, U = unknown |
| Nsmoke | x10 | Continuous - Numeric | Number of cigarettes smoked per day during pregnancy |
| Drinker | x11 | Categorical | Y = yes, N = no, U = unknown |
| Ndrink | x12 | Continuous - Numeric | Number of alcoholic drinks per week during pregnancy |

> summary(BwData$BirthWeight)  
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 255 3096 3444 3414 3766 5175

***Figure-2***

The response variable birth-weight is measured on a scale of 0 - 5000. However, the responses for this data set range from a score 255 to 5175, with a mean birth-weight equal to 3414 and a median birth weight 3444. A “pair scatter plot display” of the variables is shown in *Figure-3*.

The pairs plot indicates a possible positive linear relationship between birth weight and weight before pregnancy, weight gained during pregnancy, age, gestation, and parity and a possible negative relationship between birth weight and number of cigarettes smoked per day and number of alcoholic drinks drunk per week, during pregnancy.



***Figure-3:*** *Pairs scatter plot of birth-weight data for the response variable and regression variables.*

**Analysis**

In this section, there will be two types of analysis. First, one-way ANOVA will be used to study the relationship between BirthWeight and Race. The purpose of the ANOVA test is to determine if there is a significant difference between Races in terms of the mean BirthWeight. Second, the ANOVA test will be followed by multiple linear regression analysis.

***ANOVA***

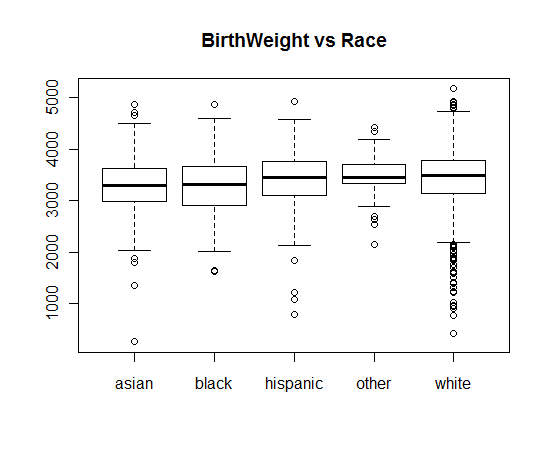
The interest of the ANOVA test is to explore the mean BirthWeight among different races. The one-way ANOVA hypothesis for BirthWeight vs Race are:

Ho = µblack = µhispanic = µwhite = µothers

Ha = At least one pair is different.

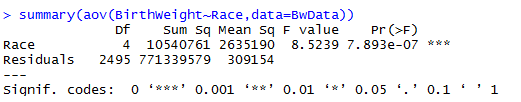
Prior to the ANOVA, we can look at the box plot to evaluate assumptions underlying ANOVA and to see if there are serious outliers.

> boxplot(BwData$BirthWeight~BwData$Race,main="BirthWeight vs Race")



***Figure 4:*** *Box plot of BirthWeight for each of 5 Race category.*

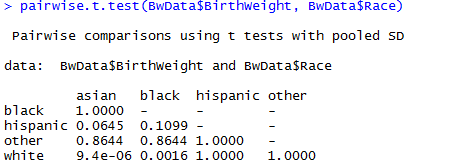
The box plot in Figure-4 points out that there might be a slight difference in average BirthWeight. Nevertheless, as shown above, there is not a serious violation of assumptions that require non-parametric Kruskal-Wallis test.



***Figure 5 :****One-Way ANOVA results for Race.*

The *Figure-4* demonstrates the ANOVA results computed using R. Based on the results above, the null hypothesis is rejected. That is, the average BirthWeight is different at least for one of the races.

In order to determine which race provides different average BirthWeight among others, a pairwise t-test with multiple comparisons is used. The results are shown in *Figure- 5*. The pairwise test indicates that Whites produces different average BirthWeight than the asian and black races.



***Figure-6:*** *Pairwise.t-test for Mean BirthWeight against pair of races.*

***Regression Analysis***

The summary statistics of the main effects model is shown in Figure-7.

Call:

lm(formula = BirthWeight ~ ., data = BwData)

Residuals:

Min 1Q Median 3Q Max

-1985.41 -271.08 -1.88 269.07 1532.57

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2140.6507 159.8823 -13.389 < 2e-16 \*\*\*

Wpre 2.9525 0.2667 11.072 < 2e-16 \*\*\*

Wgain 7.4331 0.6795 10.939 < 2e-16 \*\*\*

GenderM 134.6078 17.7105 7.600 4.15e-14 \*\*\*

Age 3.9201 1.8409 2.130 0.033310 \*

Raceblack -79.7191 41.3236 -1.929 0.053827 .

Racehispanic 121.7720 40.0449 3.041 0.002383 \*\*

Raceother 85.8442 83.4974 1.028 0.304000

Racewhite 87.0110 25.4801 3.415 0.000648 \*\*\*

Parity 54.6660 9.3623 5.839 5.94e-09 \*\*\*

Gestation 116.9856 3.7748 30.991 < 2e-16 \*\*\*

Education 3.8517 4.4162 0.872 0.383200

SmokerY -171.8827 60.6472 -2.834 0.004632 \*\*

NSmoke -3.5334 5.8219 -0.607 0.543966

DrinkerY -92.0545 107.4828 -0.856 0.391827

NDrink 8.0377 24.1204 0.333 0.738989

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 441.8 on 2484 degrees of freedom

Multiple R-squared: 0.3798, Adjusted R-squared: 0.376

F-statistic: 101.4 on 15 and 2484 DF, p-value: < 2.2e-16

***Figure-7:*** *Model with all the regressors included*

In this model, Education, Nsmoke, Drinker, and Ndrink seem to be insignificant, so we will fit a new linear model with only significant regression variables. Here is the model that we will use in further sections of this study.

Call:

lm(formula = BirthWeight ~ Wpre + Wgain + Gender + Age + Race +

Parity + Gestation + Smoker, data = BwData)

Residuals:

Min 1Q Median 3Q Max

-1978.54 -269.44 -3.01 270.31 1523.83

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2116.0334 156.7372 -13.501 < 2e-16 \*\*\*

Wpre 2.9411 0.2661 11.051 < 2e-16 \*\*\*

Wgain 7.4292 0.6786 10.948 < 2e-16 \*\*\*

GenderM 134.8410 17.6974 7.619 3.60e-14 \*\*\*

Age 4.6296 1.6082 2.879 0.004026 \*\*

Raceblack -81.9918 41.1230 -1.994 0.046281 \*

Racehispanic 110.6054 38.2070 2.895 0.003826 \*\*

Raceother 83.0130 83.2222 0.997 0.318626

Racewhite 86.6270 25.4408 3.405 0.000672 \*\*\*

Parity 52.3062 9.0490 5.780 8.39e-09 \*\*\*

Gestation 117.3327 3.7617 31.192 < 2e-16 \*\*\*

SmokerY -208.0431 35.7075 -5.826 6.40e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 441.7 on 2488 degrees of freedom

Multiple R-squared: 0.3793, Adjusted R-squared: 0.3766

F-statistic: 138.2 on 11 and 2488 DF, p-value: < 2.2e-16

***Figure-8:*** *Model with a subset of regressors included*

The R output related to the regression analysis (Figure-8) indicates that the null hypothesis in the ANOVA for regression is rejected. At least one of the regression coefficients is different than zero. The R2 and R2adjusted values appear to be small, but close together, which shows that approximately 38% of the variation can be explained by the regression variables. This can happen either because important regression variables are missing or unnecessary variables have been included. However, it is common to have such lower R2 values in studies related to human characteristics. The fitted model is:

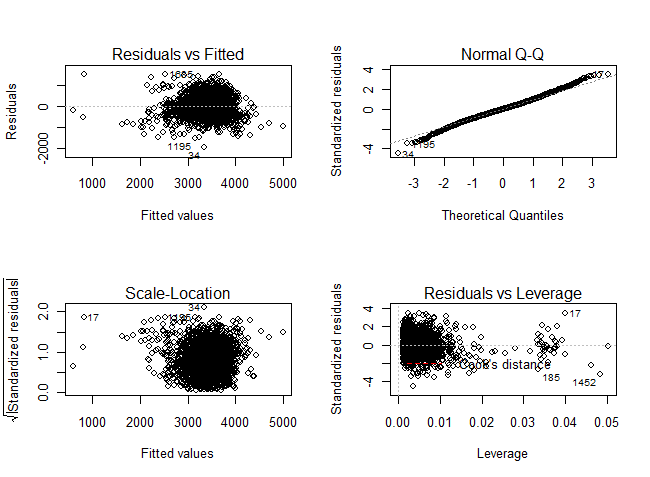
The fitted regression coefficients support the scatter plot given in *Figure-3*. For example, the regression variable smoker has a negative slope; if the mother is a smoker, this will reduce the average birth weight value.

**PART-2**

**Model Adequacy Checking**

In this part of the case study, we are going to assess the adequacy of our model and make necessary changes. We will use the model given in Figure-8 to get better results.

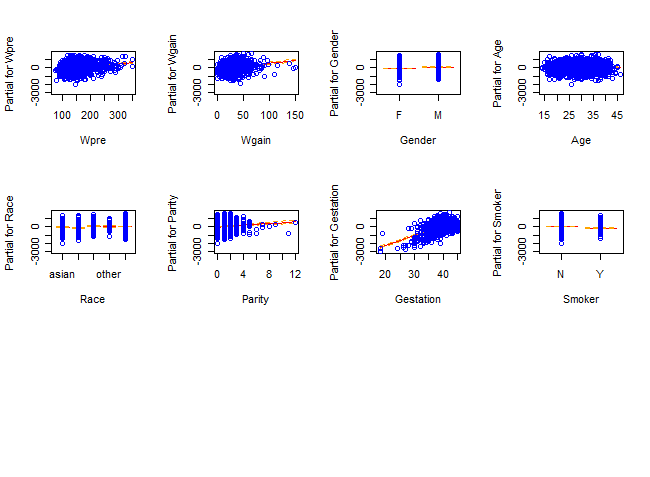
In Figure-3 (Part-1, pg.3), we can see nonlinear relationships between the response variable and some of the regressors. Additionally, the residual standard error is high and R2 and R2Adjusted values are low (Figure-8). Now let’s look at the Residual Plots for a thorough analysis of the model and check the model assumptions.



***Figure-9:*** *Residual Plots for subsets of regressors included*

The residuals appear to be almost normally distributed (verifed by fat pencil test), with a mean of zero. They also seem to be scattered randomly with no indication of a pattern. So we have uncorrolated errors. However, a seemingly double bow in “Residuals vs Fitted” plot indicates that the variance of the residuals might not be constant and we might have residuals that are heteroscedastic.

Wpre, Wgain, and Gestation plots in Figure-10 show that extra terms may be required for the main effects model. This is seen because of the slight curvature of the residuals. Also the variance assosiated with the “other“ race appears to be smaller than the races Asian, Black, Hispanic, and White. These suggest that variance stabilization might help.



***Figure-10:*** *Partial Residual Plots for Plots for subsets of regressors included*

The model assumptions 1 and 3 mentioned in Table-A.1 (pg.13) could not be satisfied properly by the main effects model. Adding interactions to the model did not improve the summary statistics, so we decided to apply power transformations to the model.

***Box-Tidwell Transformation***

The model relationship between Wgain, Wpre, Parity, Gestation, and BirthWeight seem to exhibit slıght curvature. Thus, we will utilize Box-Tidwell procedure to see whether power transformation on these regressors is appropriate or not.

> library("car")

> boxTidwell(BirthWeight~I(Wgain+0.00001)+I(Parity+0.00001)+Wpre+Gestation,

+ ~Gender+Age+Smoker+Race, max.iter=20,data=BwData)

Score Statistic p-value MLE of lambda

I(Wgain + 1e-05) -4.090206 0.0000431 0.5559401

I(Parity + 1e-05) -4.748441 0.0000020 0.2851038

Wpre -3.405825 0.0006596 -1.3609540

Gestation -5.792218 0.0000000 -0.0449861

iterations = 12

***Figure-11****: Lambda values produced by**Box-Tidwell function*

The lambda values related to the regressors Wgain, Wpre, Parity, and Gestation are calculated in presence of all other regressors. We also added 0.00001 to the values of Wgain and Parity since some of the observations have value of zero in the dataset. As a result, we came up with the lambda values shown in Figure-11. The transformed model is:

> BwData.lm3<-lm(BirthWeight~Wpre+I(1/Wpre)+Wgain+I(sqrt(1e-05+Wgain))+Gender

+ +Age+Race+Parity+I(log(1e-05+Parity))+Gestation+I(log(Gestation))

+ +Smoker, data=BwData)

> summary(BwData.lm3)

Call:

lm(formula = BirthWeight ~ Wpre + I(1/Wpre) + Wgain + I(sqrt(1e-05 +

Wgain)) + Gender + Age + Race + Parity + I(log(1e-05 + Parity)) +

Gestation + I(log(Gestation)) + Smoker, data = BwData)

Residuals:

Min 1Q Median 3Q Max

-2016.61 -266.11 -5.66 259.60 2329.20

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.503e+04 2.862e+03 -5.252 1.63e-07 \*\*\*

Wpre -2.403e-02 8.068e-01 -0.030 0.9762

I(1/Wpre) -8.096e+04 1.943e+04 -4.166 3.20e-05 \*\*\*

Wgain -3.214e-01 2.479e+00 -0.130 0.8969

I(sqrt(1e-05 + Wgain)) 8.784e+01 2.708e+01 3.243 0.0012 \*\*

GenderM 1.355e+02 1.746e+01 7.762 1.21e-14 \*\*\*

Age 3.211e+00 1.601e+00 2.006 0.0450 \*

Raceblack -9.995e+01 4.097e+01 -2.439 0.0148 \*

Racehispanic 9.050e+01 3.810e+01 2.376 0.0176 \*

Raceother 4.559e+01 8.235e+01 0.554 0.5799

Racewhite 5.824e+01 2.569e+01 2.267 0.0235 \*

Parity 9.916e+00 1.332e+01 0.744 0.4568

I(log(1e-05 + Parity)) 9.854e+00 2.336e+00 4.219 2.54e-05 \*\*\*

Gestation -3.182e+01 3.088e+01 -1.030 0.3029

I(log(Gestation)) 5.367e+03 1.107e+03 4.848 1.32e-06 \*\*\*

SmokerY -1.801e+02 3.539e+01 -5.089 3.86e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 435.5 on 2484 degrees of freedom

Multiple R-squared: 0.3975, Adjusted R-squared: 0.3939

F-statistic: 109.3 on 15 and 2484 DF, p-value: < 2.2e-16

***Figure-12:*** *Summery statistics of the transformed model*

The summary statistics of the transformed model (Figure-12) improved slightly when compared to the summary statistics of main effects model.

> anova(BwData.lm2,BwData.lm3)

Analysis of Variance Table

Model 1: BirthWeight ~ Wpre + Wgain + Gender + Age + Race + Parity + Gestation + Smoker

Model 2: BirthWeight ~ Wpre + I(1/Wpre) + Wgain + I(sqrt(1e-05 + Wgain)) +

Gender + Age + Race + Parity + I(log(1e-05 + Parity)) + Gestation +

I(log(Gestation)) + Smoker

Res.Df RSS Df Sum of Sq F Pr(>F)

1 2488 485310457

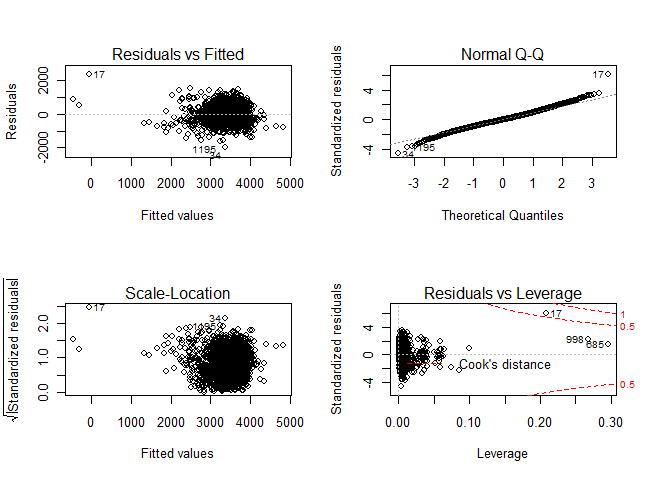
2 2484 471054135 4 14256322 18.794 3.1e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

***Figure-13 :*** *Exra-sum-of-squares method results to evaluate the transformed model*

Extra-sum-of-squares method shown in Figure-13 also indicates that the addition of the 1/Wpre, sqrt(Wgain), ln(Parity), and ln(Gestation) terms are significant (reject the null hypothesis that coefficients are different than zero).

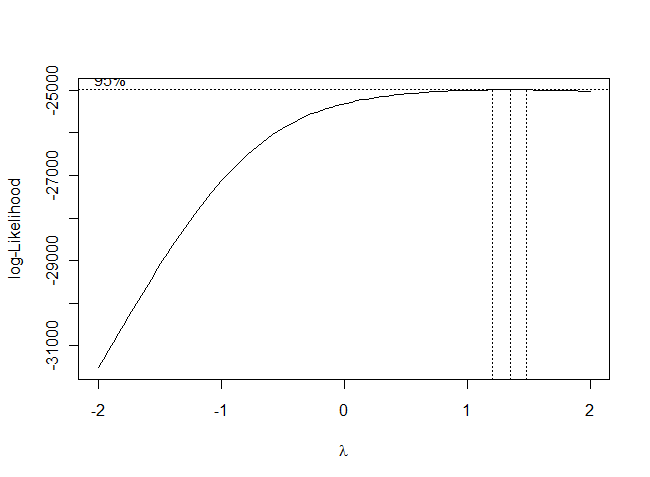


***Figure-14:*** *Residual Plots for transformed model (BwData.lm3)*

When we look at the “Residual vs Fitted” graph in Figure-14, there appears to be a linear relationship between the response and the regressors. The residuals appear to be normally distributed (Q-Q plot satisfies the fat pencil test better than the Q-Q plot of the previous model), with a mean of zero. The variance of the residuals appear to be scattered randomly. “Residuals vs Fitted” plot indicates that there might be a constant variance. Additionaly, there are no points of which Cook’s Distance is greater than one.

***Box-Cox Transformation***

In this section, we will perform a transformation on y and try to have a better stabilization of the variance.



***Figure-15:*** *Box-Cox Power Transformation Plot*

Transformation power, lambda, suggested by the box-cox method is 1.25 which is close to one. We do not expect an improvement in the summary statistics, but we will transform y with this power.

***Figure-16****: Summary statistics of box-cox transformation*

> tBirthWeight<-(BwData$BirthWeight)^1.25

> BwData.lm4<-lm(tBirthWeight~BwData$Wpre+BwData$Wgain+BwData$Gender

+ +BwData$Age+BwData$Race+BwData$Parity+BwData$Gestation

+ +BwData$Smoker)

> summary(BwData.lm4)

Residuals:

Min 1Q Median 3Q Max

-17218.1 -2633.7 -132.9 2538.3 15353.6

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -24420.615 1484.296 -16.453 < 2e-16 \*\*\*

BwData$Wpre 28.282 2.520 11.222 < 2e-16 \*\*\*

BwData$Wgain 70.779 6.426 11.014 < 2e-16 \*\*\*

BwData$GenderM 1292.346 167.594 7.711 1.79e-14 \*\*\*

BwData$Age 44.599 15.229 2.929 0.003437 \*\*

BwData$Raceblack -747.928 389.433 -1.921 0.054902 .

BwData$Racehispanic 1070.110 361.819 2.958 0.003130 \*\*

BwData$Raceother 772.363 788.111 0.980 0.327173

BwData$Racewhite 834.821 240.923 3.465 0.000539 \*\*\*

BwData$Parity 492.428 85.694 5.746 1.02e-08 \*\*\*

BwData$Gestation 1063.432 35.623 29.852 < 2e-16 \*\*\*

BwData$SmokerY -1984.227 338.149 -5.868 5.00e-09 \*\*\*

---

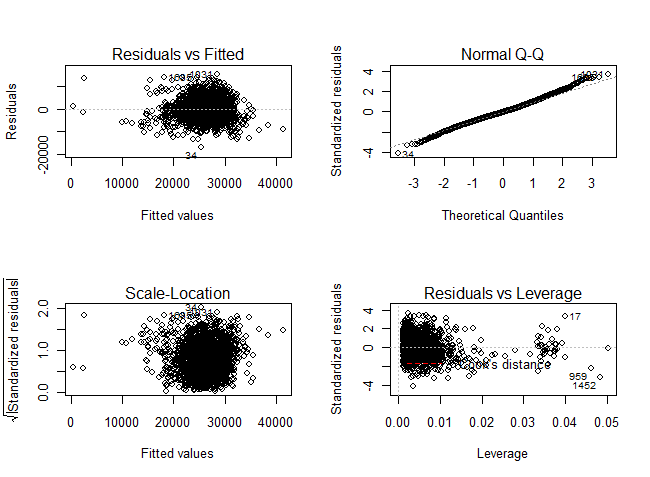
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4182 on 2488 degrees of freedom

Multiple R-squared: 0.3666, Adjusted R-squared: 0.3638

F-statistic: 130.9 on 11 and 2488 DF, p-value: < 2.2e-16

The summary statistic didn’t improve in this model. The normality assumption of residuals are almost violated in the tails as shown in Figure-A.2. The residuals seem to be scattered randomly with no indication of a pattern as in the main effects model. So we have uncorrolated errors. The variance of the residuals also seems to be the same as the variance of residuals in our main effects model.



***Figure-17:*** *Residual Plots for transformed model (BwData.lm4)*

After performing x and y transformations on our main effects model, we conclude that while the summary statistics improved slightly with box-tidwell transformation, the box-cox method did not improve the summary statistics at all.

***Checking for Multicollinearity***

The model assumptions 1-5 are examined for the transformed model. Now let’s check the model assumption 6 for multicollinearity via the variance inflation factors (VIF).

***Figure-18:*** *VIF’s of transformed model*

> vif(BwData.lm3)

GVIF Df GVIF^(1/(2\*Df))

Wpre 10.261116 1 3.203298

I(1/Wpre) 10.299135 1 3.209227

Wgain 14.560078 1 3.815767

I(sqrt(1e-05 + Wgain)) 14.937003 1 3.864842

Gender 1.003512 1 1.001755

Age 1.216316 1 1.102867

Race 1.256692 4 1.028972

Parity 2.538909 1 1.593395

I(log(1e-05 + Parity)) 2.507848 1 1.583619

Gestation 70.991790 1 8.425663

I(log(Gestation)) 71.159500 1 8.435609

Smoker 1.074668 1 1.036662

As a general rule of thumb, VIF’s greater than 10 are sign of severe or serious multicollinearity. We have VIF’s greater than 10 in our transformed model, so there is reason to suspect multicollinearity. There are four major techniques to deal with this problem (Table-A.2). We’ll try variable selection method and see if we can get lower VIF values in the next part.

**PART 3**

**Variable Selection**

We use stepAIC function from the MASS package to perform the variable selection.

> summary(BwData.AIC.lm)

Call:

lm(formula = BirthWeight ~ I(1/Wpre) + I(sqrt(1e-05 + Wgain)) +

Gender + Age + Race + I(log(1e-05 + Parity)) + I(log(Gestation)) +

Smoker, data = BwData)

Residuals:

Min 1Q Median 3Q Max

-2013.06 -266.11 -2.33 259.83 2156.54

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -12119.687 490.490 -24.709 < 2e-16 \*\*\*

I(1/Wpre) -80389.969 6416.009 -12.530 < 2e-16 \*\*\*

I(sqrt(1e-05 + Wgain)) 84.588 7.270 11.635 < 2e-16 \*\*\*

GenderM 135.572 17.443 7.772 1.12e-14 \*\*\*

Age 3.341 1.581 2.112 0.0348 \*

Raceblack -97.851 40.552 -2.413 0.0159 \*

Racehispanic 91.851 37.779 2.431 0.0151 \*

Raceother 47.954 82.186 0.583 0.5596

Racewhite 59.104 25.428 2.324 0.0202 \*

I(log(1e-05 + Parity)) 11.179 1.563 7.153 1.11e-12 \*\*\*

I(log(Gestation)) 4236.691 133.093 31.833 < 2e-16 \*\*\*

SmokerY -179.723 35.021 -5.132 3.09e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

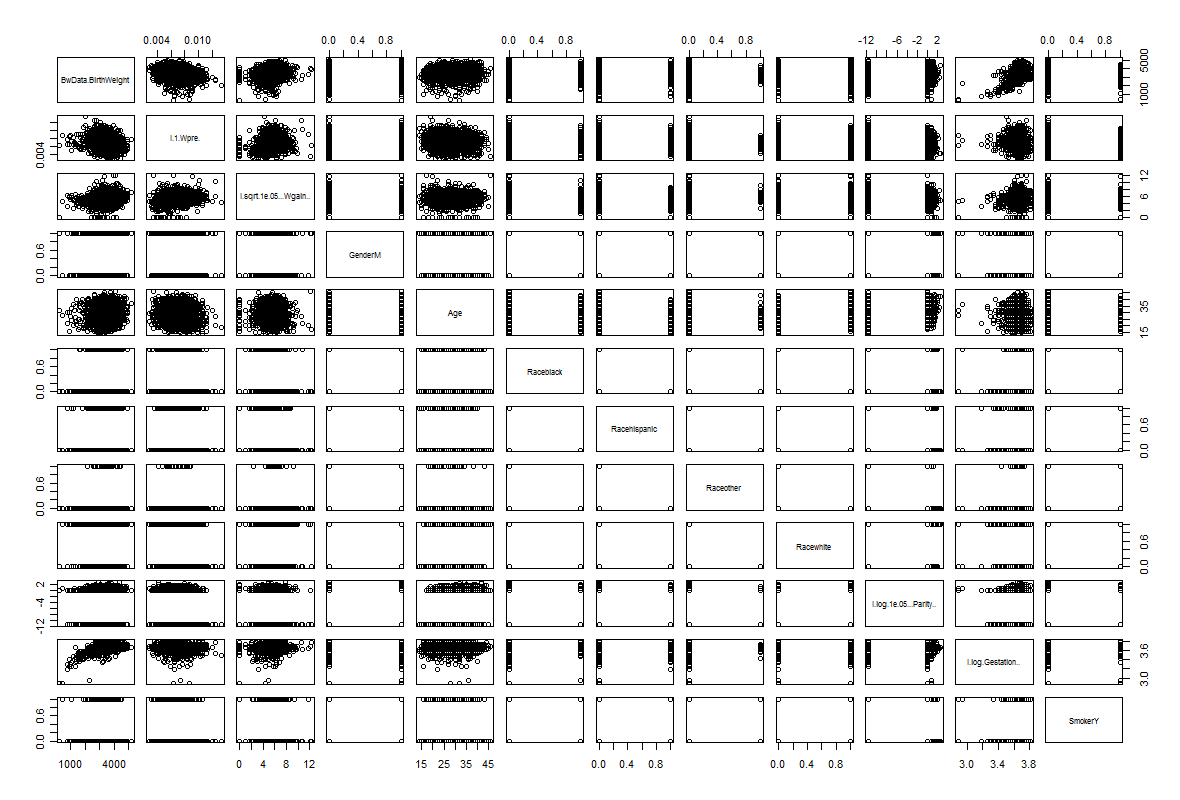
Residual standard error: 435.3 on 2488 degrees of freedom

Multiple R-squared: 0.3971, Adjusted R-squared: 0.3945

F-statistic: 149 on 11 and 2488 DF, p-value: < 2.2e-16

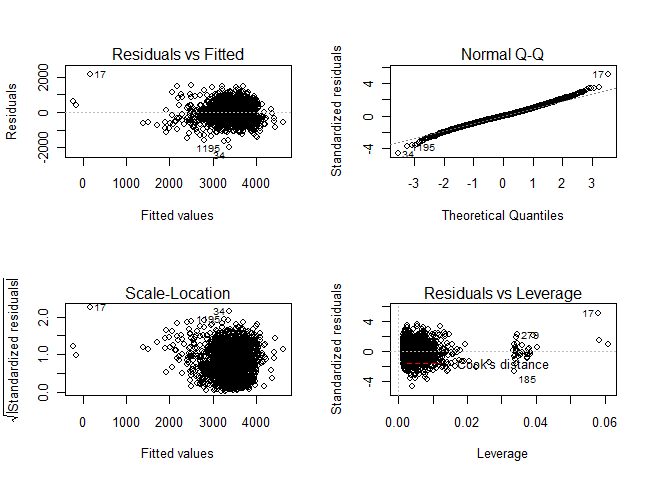
***Figure-19:*** *Summary Statistics of the model after stepwise elimination*

The final model is:

****

***Figure-20:*** *Pairs scatter plot of the final model.*

Now let’s go through assumptions 1-6 and check for adequacy again. When we look at the pairs plot given in Figure-20, there doesn’t seem to be a nonlinear relationship between the response and regressors.



***Figure-21:*** *Residual Plots for final model*

The residuals appear to be normally distributed (Q-Q plot satisfies the fat pencil test) with a mean of zero. When we look at the “Residual vs Fitted” graph in Figure-21, the variance of the residuals seem to be constant. The residuals seem to be scattered randomly without any pattern. So we have uncorrelated errors. Additionally, there isn’t any point which has cook’s distance greater than 1.

> vif(BwData.AIC.lm3)

GVIF Df GVIF^(1/(2\*Df))

I(1/Wpre) 1.123923 1 1.060153

I(sqrt(1e-05 + Wgain)) 1.077438 1 1.037997

Gender 1.002577 1 1.001288

Age 1.188267 1 1.090077

Race 1.209058 4 1.024014

I(log(1e-05 + Parity)) 1.123987 1 1.060182

I(log(Gestation)) 1.029281 1 1.014535

Smoker 1.053571 1 1.026436

***Figure-22:*** *Variation Inflation Factors of the final model*

All the VIF’s are smaller than 10, so there is no reason to suspect multicollinearity.

**Model Validation**

The final model seems to satisfy all of the CLR model assumption 1 through 6. According to some suggestions, to perform a cross-validation we need to have *n* ≥ 2*p*+25 where *n* is the number of observation and *p* is number of parameters . Since we have a data set with 2500 observations, we can utilize this technique for our model.

First, we randomly select 90% of the observations from the original data set to form a training data set, and the rest of the data is used to form a test data set. Second, we refit the model with training data set. We call this model “training model”.

***Figure-23:*** *The summary statistics of training model*

> BwData.train.lm<-lm(BirthWeight~I(1/Wpre) + I(sqrt(1e-05 + Wgain)) +

+ Gender + Age + Race + I(log(1e-05 + Parity)) + I(log(Gestation)) +

+ Smoker, data = BwData.training)

> summary(BwData.train.lm)

Call:

lm(formula = BirthWeight ~ I(1/Wpre) + I(sqrt(1e-05 + Wgain)) +

Gender + Age + Race + I(log(1e-05 + Parity)) + I(log(Gestation)) +

Smoker, data = BwData.training)

Residuals:

Min 1Q Median 3Q Max

-2006.72 -264.01 0.02 263.22 2108.20

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -11890.439 520.322 -22.852 < 2e-16 \*\*\*

I(1/Wpre) -79042.213 6788.984 -11.643 < 2e-16 \*\*\*

I(sqrt(1e-05 + Wgain)) 84.380 7.765 10.867 < 2e-16 \*\*\*

GenderM 127.010 18.574 6.838 1.03e-11 \*\*\*

Age 3.946 1.694 2.329 0.0199 \*

Raceblack -95.202 43.439 -2.192 0.0285 \*

Racehispanic 101.096 40.366 2.505 0.0123 \*

Raceother 57.081 83.460 0.684 0.4941

Racewhite 67.622 27.108 2.495 0.0127 \*

I(log(1e-05 + Parity)) 11.731 1.671 7.021 2.92e-12 \*\*\*

I(log(Gestation)) 4167.694 141.085 29.540 < 2e-16 \*\*\*

SmokerY -182.686 36.840 -4.959 7.62e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 439.5 on 2238 degrees of freedom

Multiple R-squared: 0.3909, Adjusted R-squared: 0.3879

F-statistic: 130.6 on 11 and 2238 DF, p-value: < 2.2e-16

The summary statistics of the training model are given in Figure-23. Ideally, the estimated coefficients of the final regression model should be stable. That is, the coefficients should remain almost unchanged if small changes are made on the data.

***Table-2:*** *Comparison between the coefficients of the final model and the coefficients of the training model*

|  |  |  |  |
| --- | --- | --- | --- |
| **Regressors** | **Coefficients**  **(Final Model)** | **Coefficients**  **(Training Model)** | **Difference**  **%** |
| (Intercept) | -12119.687 | -11890.439 | 1.891534 |
| I(1/Wpre) | -80389.969 | -79042.213 | 1.6765226 |
| I(sqrt(1e-05 + Wgain)) | 84.588 | 84.380 | 0.24589776 |
| GenderM | 135.572 | 127.010 | 6.31546337 |
| Age | 3.341 | 3.946 | 18.1083508 |
| Raceblack | -97.851 | -95.202 | 2.70717724 |
| Racehispanic | 91.851 | 101.096 | 10.0652143 |
| Raceother | 47.954 | 57.081 | 19.0328231 |
| Racewhite | 59.104 | 67.622 | 14.4118841 |
| I(log(1e-05 + Parity)) | 11.179 | 11.731 | 4.93782986 |
| I(log(Gestation)) | 4236.691 | 4167.694 | 1.6285587 |
| SmokerY | -179.723 | -182.686 | 1.6486482 |

The comparison between the coefficients of final model and the coefficients of training model is given in Table-2. The table shows that the coefficients of the final model remain almost unchanged and the coefficients have same signs and reasonable magnitudes.

Now, we are ready to use the training model to make predictions on the testing data. These predictions will help us in the validation process.

> predictest<-predict(BwData.train.lm, newdata=BwData.test)

> predictest[1:10]

1 2 3 4 5 6 7 8 9

3560.715 3190.289 3497.161 3883.560 3833.100 3514.839 3637.997 3502.932 3070.169

10

3464.708

***Figure-24:*** *Predictions on test data (10 predictions are given as a sample)*

After making predictions using the R code shown in Figure-24, we can now calculate the average squared prediction error (ASPE), which is a good measure of comparison, using the formula below. For *g* new observations:

***Figure-25:*** *Calculation of residuals and ASPE*

> resid<-(BwData.test$BirthWeight-predictest)

> resid[1:10]

1 2 3 4 5 6 7 8

209.2845 -213.2894 -180.1613 -254.5598 -800.1000 310.1612 955.0032 -213.9319

9 10

-150.1688 -136.7078

> ASPE<-sum((resid^2))/250

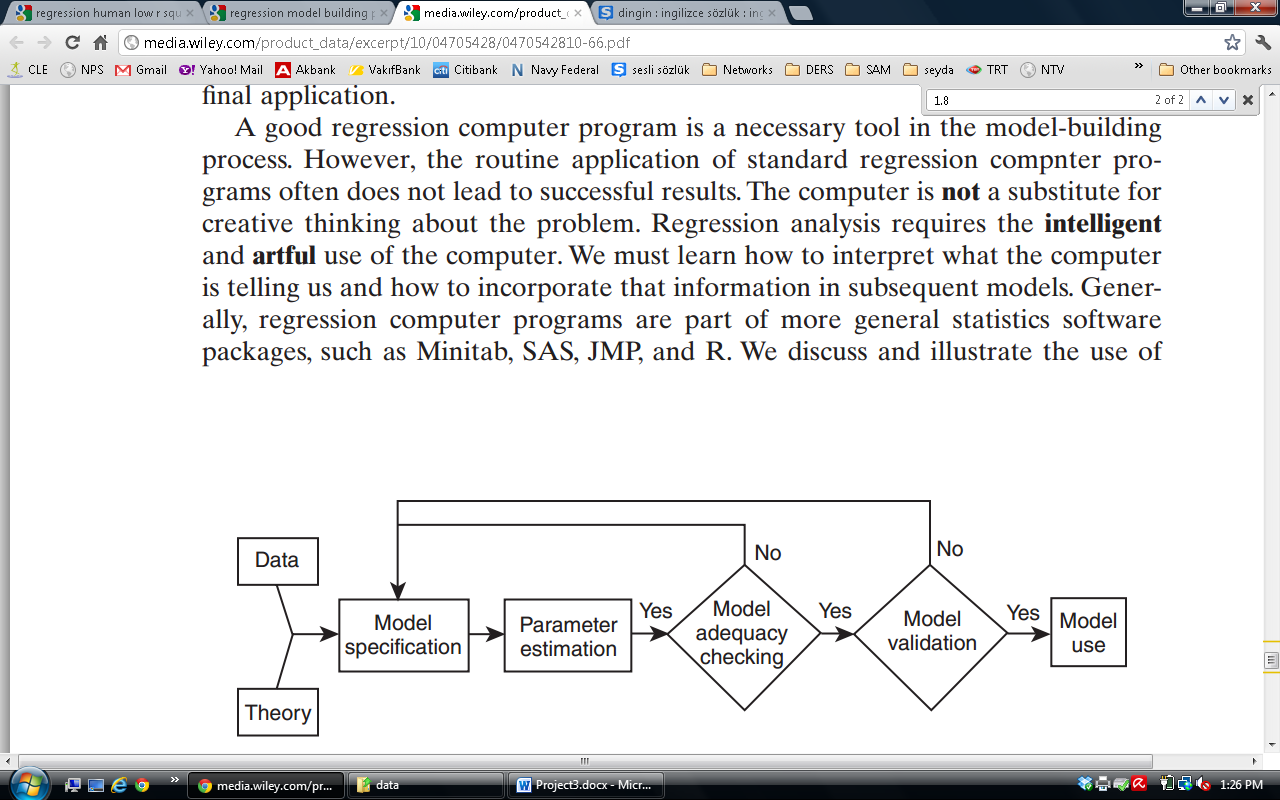
> sqrt(ASPE)

[1] 396.8837

We need to compare the value of ASPE to the residual mean square (MSResidual). Closer values are a good indication of a valid model. The calculations of the ASPE is shown in Figure-25. The residual standard error of the final model is 435.3, and the square root of ASPE is 396.8837. The difference between those amounts to 38.4163 which is approximately 8% of the residual standard error of the final model.

**CONCLUSION**

In this study, we analyze the various factors that affect the birth weight. However, the human nature is very complicated and it is really hard to make good predictions on human related areas.



***Figure-26:*** *Regression Model Building Process***[[4]](#footnote-4)**

Following “Regression Model Building Process” shown in Figure-26, we come up with the final fitted model shown below:

The final model is the weighted combination of significant regressors with linear transformations on some variables to meet the least squares criterions. Of all the 12 factors shown in Table-1 (pg.2) it is interesting to see only 8 factors in the model. These factors are:

* Mother's weight prior to pregnancy
* Mother's weight gain during pregnancy
* Gender of the baby
* Age
* Race
* Parity
* Gestation
* Whether mother smoke or not.

The most important factor in this model is gestation. The longer the gestation period the heavier the baby. Similarly, a mother's weight gain during pregnancy, age and parity have positive effects to the birthweight, while mother's weight prior to pregnancy and whether a mother smokes or not have negative impacts.

The resulting model helps us to make estimations on birth weight. For example, the birth weight of a white mother’s baby will be 156.955 grams more than a black mother’s baby if we keep all the other factors constant. Or, if the baby is a boy, then his weight will be higher.

Surpsingly, whether a mother drinks alcohol or not, and the number of drinks a mother takes during a week period do not have a significant effect on birth weight. Education level does not also have an significant impact on birth weight of the baby.

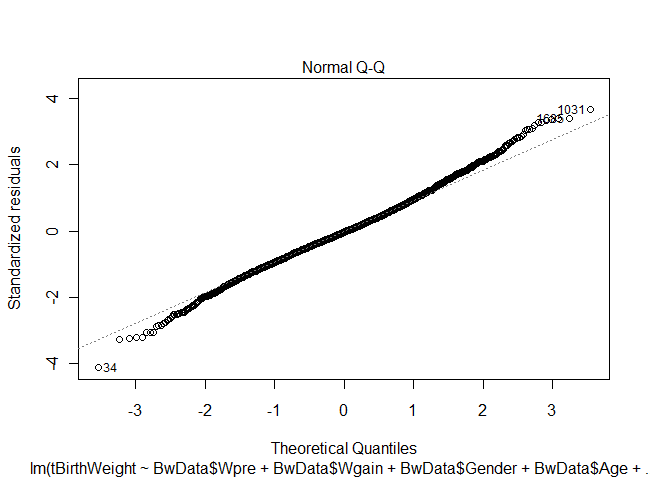
To conclude, we tried to answer the question “What affects the birth weight?” in this study. There are many factors that can increse or decrese the birthweight, and we created a model that can estimate this value using the data set at hand.

**APPENDIX**

**1. Major Assumptions of CLR Model:**

|  |  |
| --- | --- |
| ***Table-A.1*** *: Major Assumptions of CLR Model* | |
| **Number** | **Modeling Assumption** |
| 1 | The relationship between the response y and the regressors is linear |
| 2 | The error term Ɛ has zero mean |
| 3 | The error term Ɛ has constant variance(σ2) |
| 4 | The errors are uncorrelated |
| 5 | The errors are normally distributed |
| 6 | The regressors are independent |
| **2. Pairs plot for the model after Box-Tidwell transformation**  C:\Users\Turgut\Dropbox\Dersler\Data\caseStudy\Part3\parisbt.png  ***Figure-A.1 :*** *Pairs plot for the model after Box-Tidwell transformation* | |

**3. The Normal Q-Q Plot of the residuals after Box-Cox Transformation**



***Figure-A.2 :*** *The normal Q-Q plot of the residuals after box-cox transformation.*

**3. Techniques for Dealing with Multicollinearity**

***Table-A.2*** *: Techniques for Deaing with Multicollinearity*

|  |  |
| --- | --- |
|  | **What to Do** |
| 1 | Collect more data (and in the right places) |
| 2 | Based on knowledge of regressors, remove regressors or combine regressors |
| 3 | Use principle components to let the data help you decide which linear combinations of regressors to use |
| 4 | Use variable selection strategies perform variable selection. |

1. http://hypertextbook.com/facts/2004/VanessaChambers.shtml [↑](#footnote-ref-1)
2. http://www.education.vic.gov.au/healthwellbeing/childyouth/catalogue/sections/birthweight-ind1.htm [↑](#footnote-ref-2)
3. <http://www.maths.bris.ac.uk/~mahsb/birthweight.data> [↑](#footnote-ref-3)
4. Introduction to Linear Regression Analysis (Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining), 5th Edition [↑](#footnote-ref-4)