Odeya Russo

Profesor Zanontian

STATS 101A

March 24, 2023

STATS 101A Final Project

Introduction

The research question of interest is what are the best predictors for baby weight out of the following factors: gestation period, parity, mother's age, mother's height, mother's weight, and mother's smoking status?

The dataset was found on Kaggle under the name "Pregnancy Data." It was drawn from a study that "considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area" (Debjeet Das).

The response variable is the baby's weight in ounces. The predictor variables are listed below:

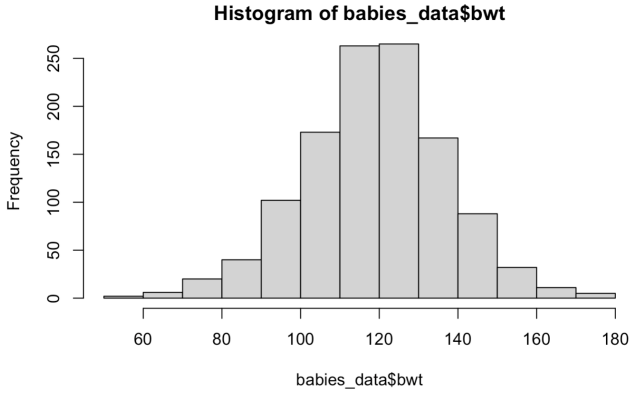
1. Gestation: length of gestation in days
2. Parity: a binary indicator for a first pregnancy (0 = first pregnancy)
3. Age: mother's age in years
4. Height: mother's height in inches
5. Weight: mother's weight in pounds
6. Smoke: a binary indicator for whether the mother smokes (0 = No)

The method I chose to model the relationship between baby weight and predictor variables is a multiple linear regression model. The reason for this choice comes from the simplicity of a linear model and the ability to easily calculate the influence several predictors have on the response variable.

This paper will first begin with a data description of individual variables and an exploration of the relationships between them. Then, it will examine several predictive models and justify the choice for the "best" model. Finally, there will be a short summary and discussion of the limitations and improvements of the analysis.

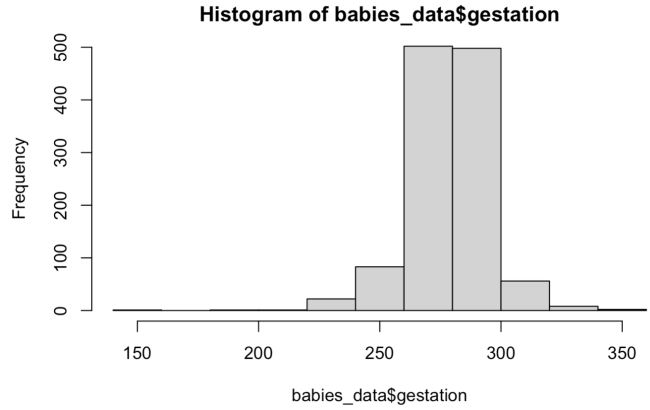
Data Description

The weight of babies in the data has a mean of 119.5 ounces with a standard deviation of 18.32867 ounces and follows a normal distribution. A summary table and histogram of the baby's weight distribution are shown below.



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
55.0	108.0	120.0	119.5	131.0	176.0

The gestation period in the data has a mean of 279.1 days with a standard deviation of 16.01031 days and follows a symmetric unimodal distribution. A summary table and histogram of the gestation period distribution are shown below.

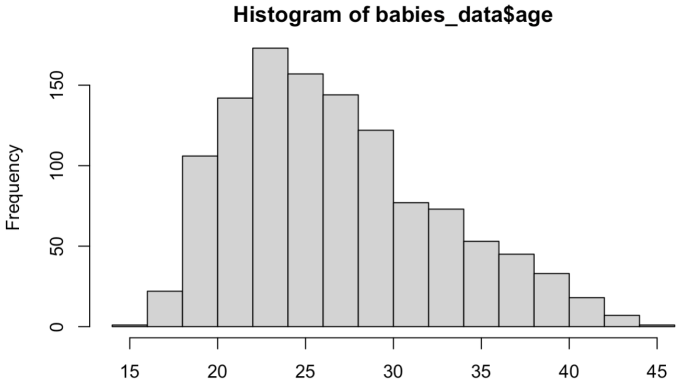


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
148.0	272.0	280.0	279.1	288.0	353.0

For the parity variable in the data, there are 866 babies who were the mother’s first pregnancy and 301 babies who were not the mother’s first pregnancy.

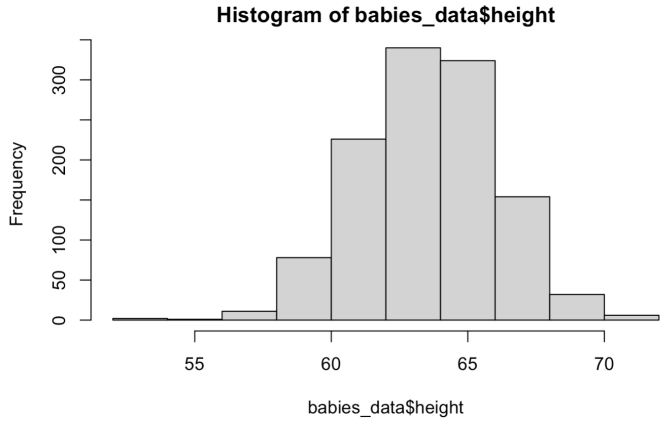
For the smoke variable in the data, there are 715 babies whose mothers did not smoke and 459 babies whose mothers did smoke.

The mother's age in the data has a mean of 27.23 years with a standard deviation of 5.817839 yearsand follows a unimodal right-skewed distribution. A summary table and histogram of the mother's age are shown below.



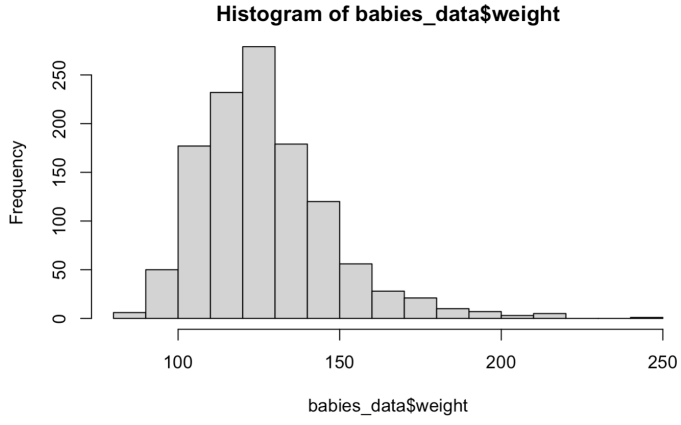

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   23.00   26.00   27.23   31.00   45.00
```

The mother's height in the data has a mean of 64.05 inches with a standard deviation of 2.526102 inches and follows an approximately normal distribution. A summary table and histogram of the mother's height are shown below.



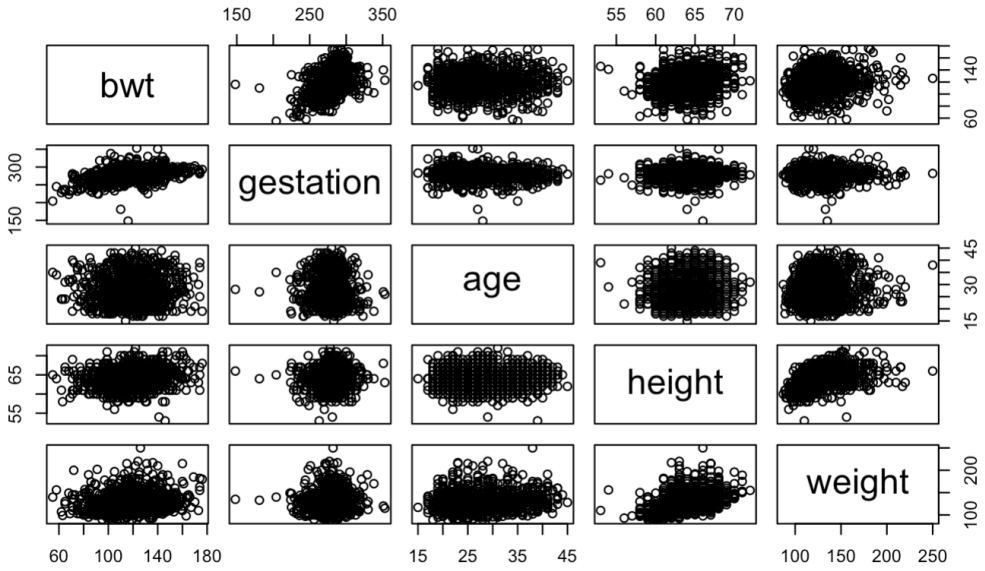

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  53.00   62.00   64.00   64.05   66.00   72.00
```

The mother's weight in the data has a mean of 128.5 pounds with a standard deviation of 20.73428 pounds and follows a unimodal right-skewed distribution. A summary table and histogram of the mother's weight are shown below.




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   87.0   114.2   125.0   128.5   139.0   250.0
```

Below is a matrix of scatterplots exploring the relationships between all variables.



The matrix illustrates that the response variable baby weight and predictor variable gestation have a relatively strong positive linear relationship. Baby weight and mother's age appear to have no relationship, as the slope is close to zero. Baby weight and predictor variables mother's height and mother's weight seem somewhat linearly related in the positive direction, with a very large spread of the data points.

Below is a correlation matrix exploring the correlation coefficients between all variables.

	bwt	gestation	parity	age	height	weight	smoke
bwt	1.00000000	0.40754279	-0.043908173	0.026982911	0.203704177	0.15592327	-0.246799515
gestation	0.40754279	1.00000000	0.080916029	-0.053424774	0.070469902	0.02365494	-0.060266842
parity	-0.04390817	0.08091603	1.000000000	-0.351040648	0.043543487	-0.09636209	-0.009598971
age	0.02698291	-0.05342477	-0.351040648	1.000000000	-0.006452846	0.14732211	-0.067771942
height	0.20370418	0.07046990	0.043543487	-0.006452846	1.000000000	0.43528743	0.017506595
weight	0.15592327	0.02365494	-0.096362092	0.147322111	0.435287428	1.00000000	-0.060281396
smoke	-0.24679951	-0.06026684	-0.009598971	-0.067771942	0.017506595	-0.06028140	1.000000000

Baby weight and gestation have the highest correlation among all the combinations of the response variable with the predictor variables, with a correlation coefficient of approximately 0.408. The matrix displays relatively low correlation coefficients between predictor variables, the highest one being 0.435 between the mother's height and the mother's weight.

Results and Interpretation

I first began by fitting a full multiple linear regression model predicting baby weight from all predictor variables gestation, age, parity, height, weight, and smoke.

The summary table for the full model is shown below. It found predictor variables gestation, parity, height, weight, and smoke to be statistically significant. It found age to not be statistically significant. The Adjusted R-squared value of 0.2541 indicates that 25.41 percent of the variation in baby weight can be explained by the full multiple linear regression model.

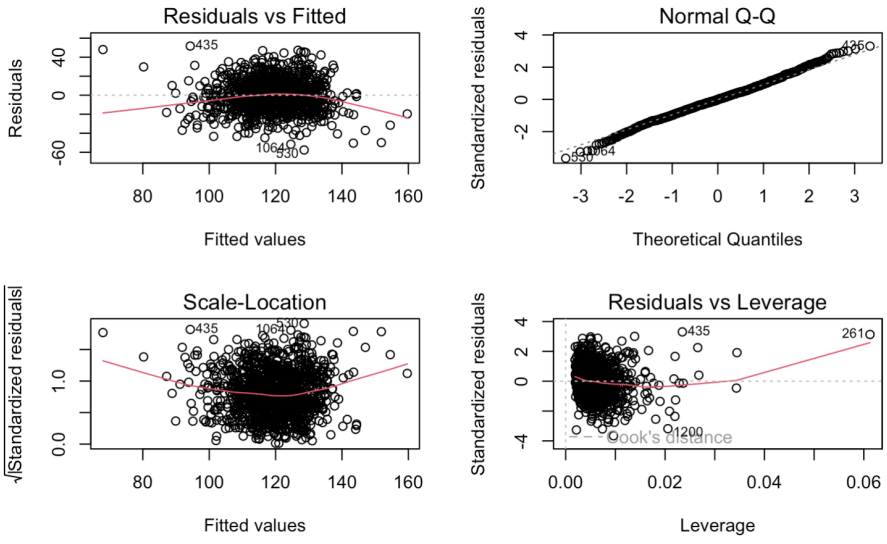
```
Call:
lm(formula = bwt ~ ., data = babies_data)

Residuals:
    Min      1Q  Median      3Q     Max
-57.613 -10.189  -0.135   9.683  51.713

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.41085   14.34657  -5.605 2.60e-08 ***
gestation     0.44398    0.02910  15.258  < 2e-16 ***
parity       -3.32720    1.12895  -2.947  0.00327 **
age          -0.00895    0.08582  -0.104  0.91696
height        1.15402    0.20502   5.629 2.27e-08 ***
weight        0.05017    0.02524   1.987  0.04711 *
smoke        -8.40073    0.95382  -8.807  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.83 on 1167 degrees of freedom
Multiple R-squared:  0.258,     Adjusted R-squared:  0.2541
F-statistic: 67.61 on 6 and 1167 DF,  p-value: < 2.2e-16
```

The diagnostic plots for the full model are shown below.



As the diagnostic plots show, the full model fits the data really well. The residual and standardized residual plots show a good random scatter and roughly a mean of zero, indicating that the errors have constant variance. The normal QQ plot follows a relatively straight line, indicating the normality of the error terms. The leverage vs residual plot shows 59 potential leverage points that fall outside of the (-4, 4) boundary or have leverages greater than 0.0119.

As previously seen, the correlation matrix does not present high correlations between predictor variables. Additionally, the VIFs for all predictor variables are all below 5. Therefore, the full model displays no issues with multicollinearity.

```
gestation    parity       age    height    weight     smoke
 1.016006  1.155657  1.167015  1.255641  1.282295  1.014995
```

Since all the model assumptions are satisfied and the diagnostic plots present no significant issues, no transformations are necessary for the variables in this dataset.

Because of the project requirements, I performed a log transformation on the response variable on my full model. Since the ranges of the predictor and response variables are significantly less than one order of magnitude, a log transform will likely not help the data. This is consistent with my findings as the log transformation did not improve the diagnostic plots and presented issues with constant variance and normality of error terms. The summary table and diagnostic plots are listed in the appendix.

To find the best combination of predictor variables for predicting baby weight, variable selection was performed. The backward elimination using the p-values approach, the backward elimination using AIC approach, and the forward selection using AIC approach arrived at the same conclusion: bwt ~ gestation + smoke + height + parity + weight is the optimal model. The backward elimination using BIC approach and forward selection using BIC approach determined that bwt ~ gestation + parity + height + smoke is the optimal model. The tables of all approaches are included in the appendix.

To determine the “best” model with the best combination of predictors, I performed a partial F-test comparing model 1: bwt ~ gestation + parity + height + smoke with model 2: bwt ~ gestation + smoke + height + parity + weight. The results are shown below.

```
Analysis of Variance Table

Model 1: bwt ~ gestation + smoke + height + parity
Model 2: bwt ~ gestation + smoke + height + parity + weight
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1169 293404
2   1168 292412  1    992.37 3.9639 0.04672 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value of 0.04672 is less than the significance level of 0.05, we reject the null hypothesis in favor of the alternative hypothesis. The evidence suggests that model 2 is the one that fits our data best.

Therefore, the "best" predictive model for predicting baby weight in this dataset is model 2: bwt ~ gestation + smoke + height + parity + weight. I arrived at this conclusion because the diagnostic plots confirmed the model assumptions, and because of the results of variable selection.

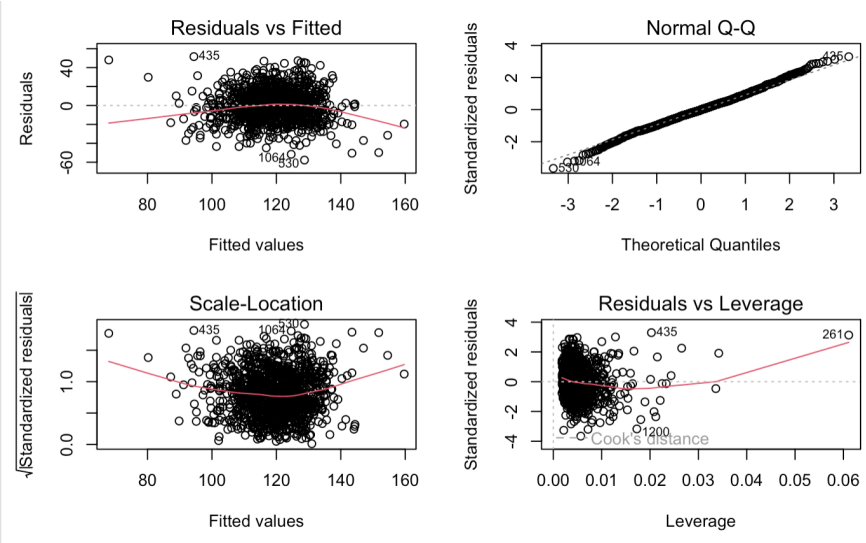
The summary table and diagnostic plots for the "best" predictive model are shown below.

```
Call:
lm(formula = bwt ~ gestation + smoke + height + parity + weight,
    data = babies_data)

Residuals:
    Min      1Q  Median      3Q     Max
-57.716 -10.150  -0.159   9.689  51.620

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.71321   14.04465  -5.747 1.16e-08 ***
gestation     0.44408    0.02907  15.276  < 2e-16 ***
smoke        -8.39390    0.95117  -8.825  < 2e-16 ***
height        1.15497    0.20473   5.641 2.11e-08 ***
parity       -3.28762    1.06281  -3.093  0.00203 **
weight        0.04983    0.02503   1.991  0.04672 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.82 on 1168 degrees of freedom
Multiple R-squared:  0.2579,    Adjusted R-squared:  0.2548
F-statistic:  81.2 on 5 and 1168 DF,  p-value: < 2.2e-16
```



Again, the diagnostic plots indicate that the "best" model fits the data well. The residual and standardized residual plots show a good random scatter and roughly a mean of zero, indicating that the errors have constant variance. The normal QQ plot follows a relatively straight line, indicating the normality of the error terms. The leverage vs residual plot shows 59 potential leverage points that fall outside of the (-4, 4) boundary or have leverages greater than 0.0119, which is a reasonable number considering the dataset contains 1174 observations.

The "best" predictive model found has led to multiple findings. In all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area, gestation period, parity, mother's height, mother's weight, and mother's smoking status have a statistically significant effect on baby weight while the mother's age does not. Variables of gestation and smoke appear to be the most statistically significant, indicating a very strong association with baby weight in these specific pregnancies. Based on the regression coefficients, the mother's smoking status has the largest estimated effect on baby weight, followed by parity. When examining the slopes of the predictor variables, it appears that gestation, height, and weight have a positive relationship with baby weight. Interestingly, smoke and parity appear to have an inverse relationship with baby weight. Lastly, 25.48 percent of the variation in the baby weight of all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area is explained by the full multiple linear regression model. The model presents a relatively low Adjusted R-squared value, however, is not surprising for such a dataset exploring human behaviors.

Discussion

This paper explored the influence of predictor variables gestation period, parity, mother's age, mother's height, mother's weight, and mother's smoking status on baby weight. After exploring the individual distributions and relationships between all variables, I fit a multiple linear regression model to predict baby weight. Through examining diagnostic plots and variable selection, it was found that the "best" predictive model is bwt ~ gestation + smoke + height + parity + weight. In other words, predictor variables gestation, smoke, height, parity, and weight were found to be statistically significant in having an association with baby weight in all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay.

Medical studies confirm the majority of my findings. A study done in Rural Karnataka, India found a "significant association between the birth weight of the baby and the maternal age, maternal education, per capita income of the family, time of antenatal registration, number of antenatal visits, physical work during pregnancy, height, and weight in pregnancy" (Metgud). This is consistent with my findings of the statistically significant predictor variables mother's height and mother's weight, and their positive regression coefficients. Another study done in São Paulo state, Brazil discovered that "smoking during pregnancy is associated with lower birth weight in full-term infants" (Kataoka). Again, this study supports my findings, as smoke was one of the most statistically significant predictor variables with a negative correlation coefficient. Interestingly, several studies have discovered that first babies are more likely to weigh less than their siblings (HealthWise Staff). This is inconsistent with our findings, as parity was found to have a negative regression coefficient indicating an inverse relationship with baby weight.

The limitations of this study lie in the accuracy of the dataset and the limited predictor variables. Because this study was not a random sample and only considered pregnancies in a certain time period in a specific area, I cannot generalize my findings about the factors associated with a baby's weight to the general population. Additionally, simply too few predictor variables are included in this dataset to predict a baby's weight accurately. There are many other factors at play, for example, the baby's health status, the economic status of the family, the stress levels of the mother, etc.

To improve the study, one will obtain a more representative sample of the general population containing information about baby weight and associated factors. A random sample would allow the findings to be generalized to a larger population. Moreover, the dataset should include more predictor variables. To better predict the weight of babies, one should explore the associations between baby weight and many different factors covering both aspects of nature and nurture.

Appendix:

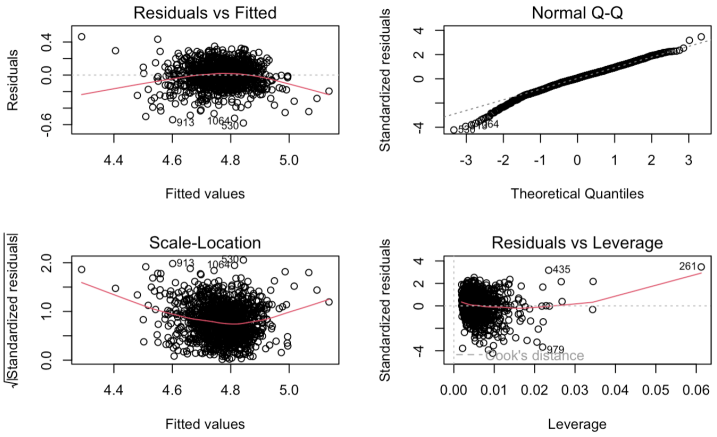
Log Transformation Summary Table and Diagnostic Plots

```
Call:
lm(formula = log(bwt) ~ ., data = babies_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58060 -0.07772  0.00674  0.08640  0.46403

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.0087835  0.1252532  24.022  < 2e-16 ***
gestation    0.0041043  0.0002540  16.156  < 2e-16 ***
parity      -0.0286183  0.0098563  -2.904  0.00376 **
age         -0.0003686  0.0007493  -0.492  0.62280
height       0.0095616  0.0017899   5.342 1.11e-07 ***
weight       0.0003897  0.0002204   1.768  0.07729 .
smoke       -0.0733856  0.0083274  -8.813  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1382 on 1167 degrees of freedom
Multiple R-squared:  0.2679,    Adjusted R-squared:  0.2641
F-statistic: 71.16 on 6 and 1167 DF,  p-value: < 2.2e-16
```



Backward elimination using AIC

```
Start:  AIC=6491.82
bwt ~ gestation + parity + age + height + weight + smoke

            Df Sum of Sq    RSS    AIC
- age        1         3 292412 6489.8
<none>                   292409 6491.8
- weight     1       990 293399 6493.8
- parity     1      2176 294586 6498.5
- height     1      7939 300348 6521.3
- smoke      1     19437 311846 6565.4
- gestation  1     58334 350744 6703.4

Step:  AIC=6489.83
bwt ~ gestation + parity + height + weight + smoke

            Df Sum of Sq    RSS    AIC
<none>                   292412 6489.8
- weight     1       992 293404 6491.8
- parity     1      2396 294808 6497.4
- height     1      7968 300380 6519.4
- smoke      1     19497 311909 6563.6
- gestation  1     58421 350833 6701.7
```

Backwards elimination using BIC

```
Start:  AIC=6527.3
bwt ~ gestation + parity + age + height + weight + smoke

            Df Sum of Sq    RSS    AIC
- age        1         3 292412 6520.2
- weight     1       990 293399 6524.2
<none>                   292409 6527.3
- parity     1      2176 294586 6528.9
- height     1      7939 300348 6551.7
- smoke      1     19437 311846 6595.8
- gestation  1     58334 350744 6733.8

Step:  AIC=6520.24
bwt ~ gestation + parity + height + weight + smoke

            Df Sum of Sq    RSS    AIC
- weight     1       992 293404 6517.2
<none>                   292412 6520.2
- parity     1      2396 294808 6522.8
- height     1      7968 300380 6544.7
- smoke      1     19497 311909 6589.0
- gestation  1     58421 350833 6727.0

Step:  AIC=6517.15
bwt ~ gestation + parity + height + smoke

            Df Sum of Sq    RSS    AIC
<none>                   293404 6517.2
- parity     1      2857 296261 6521.5
- height     1     13261 306665 6562.0
- smoke      1     20306 313710 6588.6
- gestation  1     58383 351787 6723.1
```

Forward selection using AIC

```
Start:  AIC=6830.08
bwt ~ 1

            Df Sum of Sq    RSS    AIC
+ gestation  1     65450 328608 6618.8
+ smoke      1     24002 370056 6758.3
+ height     1     16352 377706 6782.3
+ weight     1      9580 384477 6803.2
+ parity     1       760 393298 6829.8
<none>                   394058 6830.1
+ age        1       287 393771 6831.2

Step:  AIC=6618.84
bwt ~ gestation

         Df Sum of Sq    RSS    AIC
+ smoke   1   19533.4 309075 6548.9
+ height  1   12126.1 316482 6576.7
+ weight  1    8437.0 320171 6590.3
+ parity  1    2344.7 326264 6612.4
+ age     1     939.4 327669 6617.5
<none>                328608 6618.8

Step:  AIC=6548.9
bwt ~ gestation + smoke

         Df Sum of Sq    RSS    AIC
+ height  1   12814.1 296261 6501.2
+ weight  1    7015.1 302060 6523.9
+ parity  1    2409.5 306665 6541.7
<none>                309075 6548.9
+ age     1     430.5 308644 6549.3

Step:  AIC=6501.19
bwt ~ gestation + smoke + height

         Df Sum of Sq    RSS    AIC
+ parity  1   2856.55 293404 6491.8
+ weight  1   1453.37 294808 6497.4
<none>                296261 6501.2
+ age     1    435.89 295825 6501.5

Step:  AIC=6491.81
bwt ~ gestation + smoke + height + parity

         Df Sum of Sq    RSS    AIC
+ weight  1    992.37 292412 6489.8
<none>                293404 6491.8
+ age     1      5.43 293399 6493.8

Step:  AIC=6489.83
bwt ~ gestation + smoke + height + parity + weight

       Df Sum of Sq    RSS    AIC
<none>              292412 6489.8
+ age   1    2.7253 292409 6491.8
```

Forward selection using BIC

```
Start:  AIC=6835.15
bwt ~ 1

            Df Sum of Sq    RSS    AIC
+ gestation  1     65450 328608 6629.0
+ smoke      1     24002 370056 6768.4
+ height     1     16352 377706 6792.5
+ weight     1      9580 384477 6813.3
<none>                   394058 6835.1
+ parity     1       760 393298 6839.9
+ age        1       287 393771 6841.4

Step:  AIC=6628.98
bwt ~ gestation

         Df Sum of Sq    RSS    AIC
+ smoke   1   19533.4 309075 6564.1
+ height  1   12126.1 316482 6591.9
+ weight  1    8437.0 320171 6605.5
+ parity  1    2344.7 326264 6627.6
<none>                328608 6629.0
+ age     1     939.4 327669 6632.7

Step:  AIC=6564.1
bwt ~ gestation + smoke

         Df Sum of Sq    RSS    AIC
+ height  1   12814.1 296261 6521.5
+ weight  1    7015.1 302060 6544.2
+ parity  1    2409.5 306665 6562.0
<none>                309075 6564.1
+ age     1     430.5 308644 6569.5

Step:  AIC=6521.46
bwt ~ gestation + smoke + height

         Df Sum of Sq    RSS    AIC
+ parity  1   2856.55 293404 6517.2
<none>                296261 6521.5
+ weight  1   1453.37 294808 6522.8
+ age     1    435.89 295825 6526.8

Step:  AIC=6517.15
bwt ~ gestation + smoke + height + parity

         Df Sum of Sq    RSS    AIC
<none>                293404 6517.2
+ weight  1    992.37 292412 6520.2
+ age     1      5.43 293399 6524.2
```

Sources

Das, Debjeet. “Pregnancy Data.” *Kaggle*, 7 Mar. 2023, https://www.kaggle.com/datasets/debjeetdas/babies-birth-weight.

Kataoka, Mariana Caricati, et al. “Smoking during Pregnancy and Harm Reduction in Birth Weight: A Cross-Sectional Study - BMC Pregnancy and Childbirth.” *BioMed Central*, BioMed Central, 12 Mar. 2018, https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-018-1694-4#citeas.

Metgud, Chandra S, et al. “Factors Affecting Birth Weight of a Newborn--a Community Based Study in Rural Karnataka, India.” *PloS One*, U.S. National Library of Medicine, 2012, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3390317/.

“Physical Growth in Newborns.” *MyHealth.Alberta.ca Government of Alberta Personal Health Portal*, https://myhealth.alberta.ca/Health/Pages/conditions.aspx?hwid=te6295.