



Exploration of the Determinants of Baby Birth Weight

Tuesday, November 5

A preliminary investigation and model of the factors contributing to infant birth weight

Executive Summary:

This report explores the impact of certain variables on child birth weight, specifically data drawn from the Child Health and Development Studies (CHDS). To do this, we fit multiple linear models to determine which variables are significant. The linear models were selected based off of an ANOVA (analysis of variance), VIF (variance inflation factor), AIC (Akaike Information Criterion), with the use of the dredge function, durbinWatsonTest, and ncvtTest, as well as bootstrapping, in R (R Core Team, 2019). Furthermore, 5-fold and cross validation is used to determine significance. Through this analysis, we determine that the number of previous pregnancies and a longer gestation period lead to higher infant birth weights. Moreover, as the mother's height and father's weight increase, so does the infant birth weight.

Table of Contents:

Executive Summary	2
Introduction	4
Data Cleaning, Plotting, Summarizing	5
<i>Table 1</i>	
4	
<i>Figure 1</i>	7
Model Selection	7
<i>Table 2</i>	
8	
<i>Table 3</i>	
9	
<i>Table 4: QQ plots of the 4 selected models</i>	
9	
<i>Table 5: Variance Inflation Factor results</i>	
10	
Partial Plots- Using Residuals to check linearity assumption.....	12
<i>Figure 2: Partial Residual plots</i>	12
Cross-Validation.....	
12	
<i>Table 6</i>	
13	
K-fold Cross Validation	13
<i>Table 7</i>	
14	
<i>Table 8: Coefficients of the modeled variables</i>	15
Bootstrapping Confidence Intervals	17
<i>Table 9: 95% Confidence Interval for Model 1</i>	
18	
Discussion of Findings.....	17
Conclusion	18
References	19
Appendices	20
A 1.0) Data Cleaning:	20
A 2.0) Model Selection:	22
A 2.1) Model 2	23

<i>A 3.0) Cross Validation</i>	
24	
<i>A 5.0) 5 Fold Cross Validation</i>	
24	
<i>A 5.0) Bootstrapping</i>	25

Introduction:

The goal of this report is to examine the relationship between infant birth weight and a collection of observed parental characteristics. The data used in this preliminary analysis was collected as part of a larger group of studies, referred to as the Child Health and Development Studies (CHDS). This data was explored by Yerushalmy in his 1964 paper, which found that infants with mothers who smoked were more likely to be born with “low birth weights” but also to be healthier than babies from mothers that did not smoke. The paper concluded that the data showed smoking has a complex effect on infant health, and for this reason, the data deserves further consideration. (Parascandola, 2014). Perhaps the most important takeaway from the controversy sparked by Yerushalmy’s paper is the problematic nature of causal inference, especially in the context of observational studies. This provided the initial justification for this report’s investigation.

Data Cleaning, Plotting, Summarizing:

Due to the large number of variables in the original dataset, the summary of the data was confusing. Many of the variables were discretely coded as observations between 0 and 9 where each number corresponds to some differing level of the variable in question. The variables that are reflected upon in this report are contained in an explanation in Table 1.

id	Identification number
gestation	Length of gestation in days

wt	birth weight in ounces
parity	Total number of previous pregnancies including fetal deaths and stillbirths
race	Mother's race
age	Mother's age in years at termination of pregnancy
ed	Mother's education
ht	Mother's height in inches
wt	Mother's pregnancy weight in pounds à renamed to mothers_weight
drace	Father's race
dage	Father's age
ded	Father's education
dht	Father's height
dwt	Father's weight
marital	Marital status
inc	Family yearly income
smoke	Does mother smoke?
time	If mother quit smoking, how long ago
number	Number of cigarettes smoked per day for past and current smokers

Table 1: Variables initially kept for Analysis

The table above represents the dataset after removal of the irrelevant variables for this analysis. The first variables to omit include *plurality* which simply explains whether or not the fetus was a single fetus (i.e. not a twin), which all of the members of the dataset are. The next removed variable is *date*. The reason behind this decision is due to *date*'s unusual coding. With the key indicating 1096 denotes January 1, 1961, some of the births in question would have been in the 1500s, so we ignore the birth date for this analysis. The next variable we remove is *outcome* which only specifies whether or not a baby lives past 28 days. Since this was the case for all babies in the dataset, this variable was superfluous to our analysis. The *sex* category

assigns *male*, *female*, or *unknown* sex to the infant, but all babies are male, so this provides no use for analysis. Because, these variables provide no additional information or use for the analysis, in conjunction with Occam's Razor, we remove them completely. Finally, the variable *wt* needed to be recoded due to double occurrence in the dataset. While one instance refers to baby birth weight, the other references the mother's weight. To make the two variables distinguishable from one another, the second *wt* variable was renamed *mothers_weight* in the dataset. Several of the variables after the initial data cleaning contain many unknown or non-applicable data points. In order to analyze the dataset properly, the unknown or non-applicable data points are filtered as NA.

For the variables *race* and *education* (for both the mother and father), some variables with different values defined the same criteria. For the mother and father's race, all levels from 0 to 5 define white babies. To simplify the values, we change them all to 5 rather than ranging between 0 and 5. In addition, education levels of 6 and 7 both specify *Trade School, HS unclear*, so we assign them all as 6. The next step in data cleaning is to insert the specific description of each variable level so that it could be easily interpreted in the results. Having an income level of 8 is much less descriptive than knowing the family income is between *\$12,000 and \$14,999*. This leads to a more straight-forward analysis without having to refer back to the description of each variable. Our data cleaning leads to the retainment of the following variables: whether the mother smoked/smokes, whether she had quit, and if so how long ago, the number of cigarettes that she smoked per day, the family's yearly income, the education levels of both mother and father, the race of both the mother and father, as well as the marital status of the mother. Furthermore, the columns in the dataset are renamed to remove confusion.

For data plotting, our analysis focuses on the relationship between many of the variables and how they influence birth weight, initially on mothers who smoke. Initial plotting explores how birth weight is influenced by the mother's and father's age, weight, and height, as well as gestation period. In order to see how the differing number of cigarettes smoked per day may have an impact, we plot the data with color-coding for the different levels of cigarettes smoked per day. We observe no clear signal through these plots. The final step in data cleaning is to write out a new dataset of clean data for the further exploration and analysis.

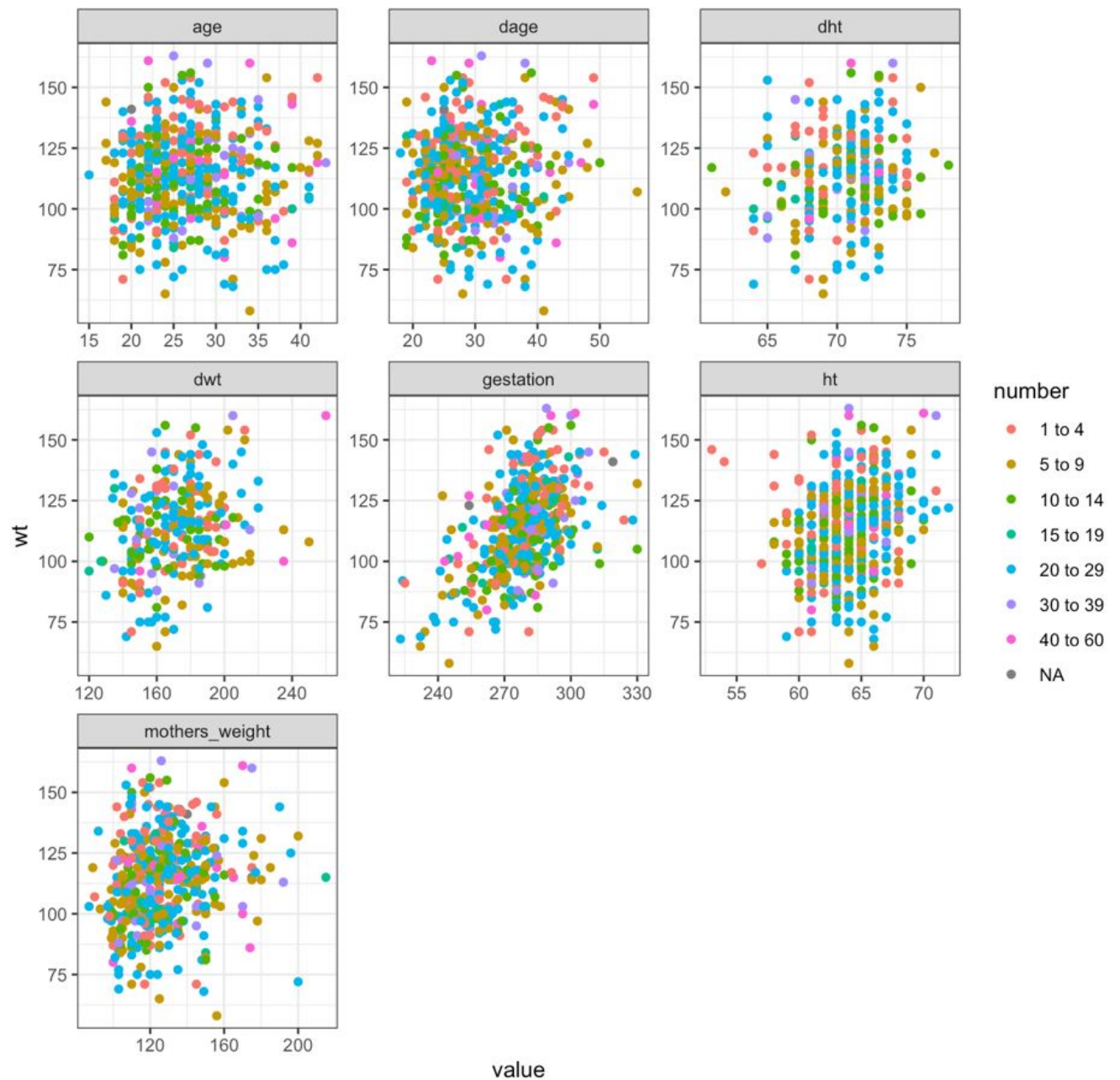


Figure 1

Model Selection:

This part aimed to generate best fit models using existing data, and at the same time, the model should meet all assumptions as well. The model selection procedure includes the

180026223, 190015944, 190019410, 190019859, 190024257, 190028036

following: cleaning of bad data, generating models, deleting any unsuitable variables, selecting the top four models based on best fit, and checking assumptions. For model selection we install the package *MuMin*, which provides ‘*Tools for performing model selection and model averaging*’ and used the *dredge()* function to retain subsets of models. In addition, we also install the *car* package, which are required in order to use *Anova()*, *ncvTest()*, *durbinWatsonTest()* and *vif()*.

After cleaning the dataset of the NA values through *na.omit*, we proceed to the generation of models. By using the basic summary function in R of the model, we omit even more values due to a lack of occurrences, such as variable *number_of_Cigs_per_day*. By removing these data points, we are able to clear the dataset of unnecessary rows of null values and thus simplify the model. We further delete unsuitable variables, and focus not only on deleting the null values, but also removing variables which have high p-values through our analysis of variance, such as variables *marital*, *Family_annual_income*, *fathers_age*, and *mothers_age*. As the p-value increases, the significance of the variable decreases. We remove variables with high p-values to increase efficiency when using the *dredge* function to get subsets of models.

Further model selection leads us to add two interaction terms, *mothers_education:Length_of_Gestation_Days* and *mothers_race:Length_of_Gestation_Days*. This model explores the relationship between gestation days and both the mother’s education and the mother’s race, although separately. It is important to note that any interaction does not necessarily imply a direct relationship between the variables. For instance, if it were found that mother’s education level had an effect on the effect of gestation length, this could be due to factors related to education level, like socioeconomic status. After adding these two interaction terms, we summarize the model, and find that the p-values of these interaction terms are 0.003 and 0.033, so they are both significant in the linear model. These were not the interaction variables we thought were *a priori* logical. At first we expected an interaction between *Length_of_Gestation_Days* and *Time_Since_Mother_Quit*. It was reasoned that the effect of smoking, regardless of how long ago the mother quit, would be mitigated or increased by the amount of time the baby stayed in the womb, or was exposed to the potential effects of the mother’s current or previous smoking habits.

To select the top four models for analysis we use the *dredge()* function to get the dataset of models sorted by AIC value, so we can choose the top four.

In order to test our assumptions we used the *ncvTest()* to check residuals’ constant spread. As we know, the null hypothesis of this test is that ‘*residuals have constant variance*’. While the result shows that all selected models have p-values over 0.05, so we fail to reject the null hypothesis. Results of constant spread test are shown below:

180026223, 190015944, 190019410, 190019859, 190024257, 190028036

	First model	Second model	Third model	Fourth model
P value	0.47521	0.35188	0.64445	0.90684

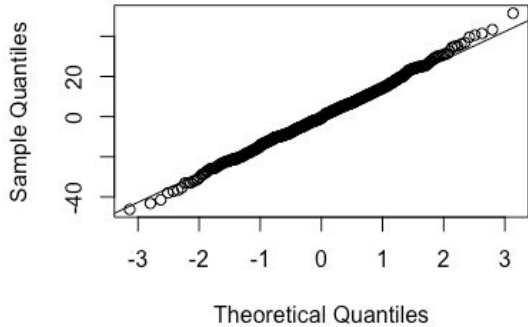
Table 2

In order to check residuals' independence, we make use of the *durbinWatsonTest()*. In this part, the null hypothesis is that '*residuals are independent*'. The outcome of the test shows that the first three models have independent residuals, while the last one does not. Results of independence test are shown below:

	First model	Second model	Third model	Fourth model
P value	0.054	0.07	0.066	0.034

Table 3

Furthermore, we use *shapiro.test()* and *qqplot()* to check residuals' normality. Results of normality test are shown below:

	qq-plot	p-value	Conclusion
First model	<p style="text-align: center;">Normal Q-Q Plot</p> 	0.7971	Normal Distribution

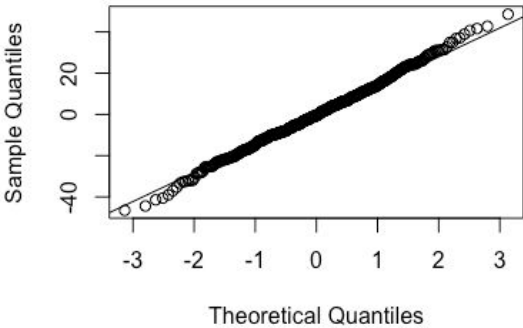
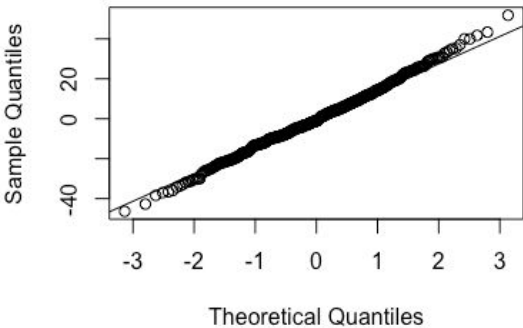
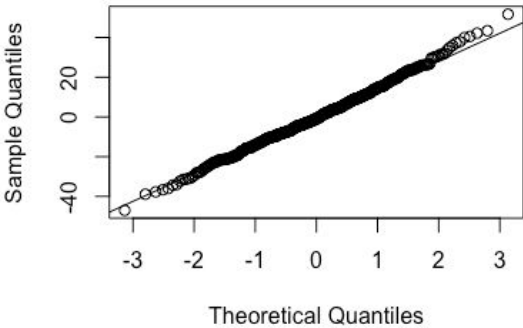
Second model	<p>Normal Q-Q Plot</p> 	0.7252	Normal Distribution
Third model	<p>Normal Q-Q Plot</p> 	0.4678	Normal Distribution
Fourth model	<p>Normal Q-Q Plot</p> 	0.4153	Normal Distribution

Table 4: *QQ plots of the 4 selected models*

To check collinearity we use the *vif()* function, which measures variance inflation factor. The results show that there is no multicollinearity for the remaining variables except for the significant collinearity of the interaction terms in the model. Results of collinearity test are shown below:

	Result		
First model		GVIF	Df
	Length_of_Gestation_Days	1.079677	1
	as.factor(Number_of_previous_pregnancies)	1.371729	10
	mothers_height	1.156650	1
	fathers_race	1.328079	4
	fathers_weight	1.148548	1
	Time_since_mother_quit	1.243678	7
Second model		GVIF	Df
	Length_of_Gestation_Days	1.091716	1
	as.factor(Number_of_previous_pregnancies)	1.445696	10
	mothers_height	1.359097	1
	mothers_weight	1.424060	1
	fathers_race	1.395717	4
	fathers_weight	1.155698	1
Third model		GVIF	Df
	Length_of_Gestation_Days	6.385767e+00	1
	as.factor(Number_of_previous_pregnancies)	2.113813e+00	10
	mothers_education	1.789395e+10	4
	mothers_height	1.189053e+00	1
	fathers_race	1.465840e+00	4
	fathers_weight	1.160877e+00	1
Fourth model		GVIF	Df
	Length_of_Gestation_Days	5.253544e+01	1
	as.factor(Number_of_previous_pregnancies)	2.363235e+00	10
	mothers_race	4.652997e+10	4
	mothers_education	2.280361e+10	4
	mothers_height	1.212395e+00	1
	fathers_weight	1.162779e+00	1
		GVIF	Df
	Time_since_mother_quit	1.546542e+00	7
	Length_of_Gestation_Days:mothers_education	2.307169e+10	4
	Length_of_Gestation_Days:mothers_race	5.107752e+10	4

Table 5: Variance Inflation Factor results

One of the final tests for our best model is plotting the partial residuals. Partial residuals are created when the signal of a covariate is treated as part of the noise of the residual. These partial plots can show how important any covariate is. The larger scatter about the line within the partial plot means a weaker signal from the covariate. The larger residuals should be examined in more detail, as well as non-linear relationships (if the plot is curved, the relationship is likely not linear).

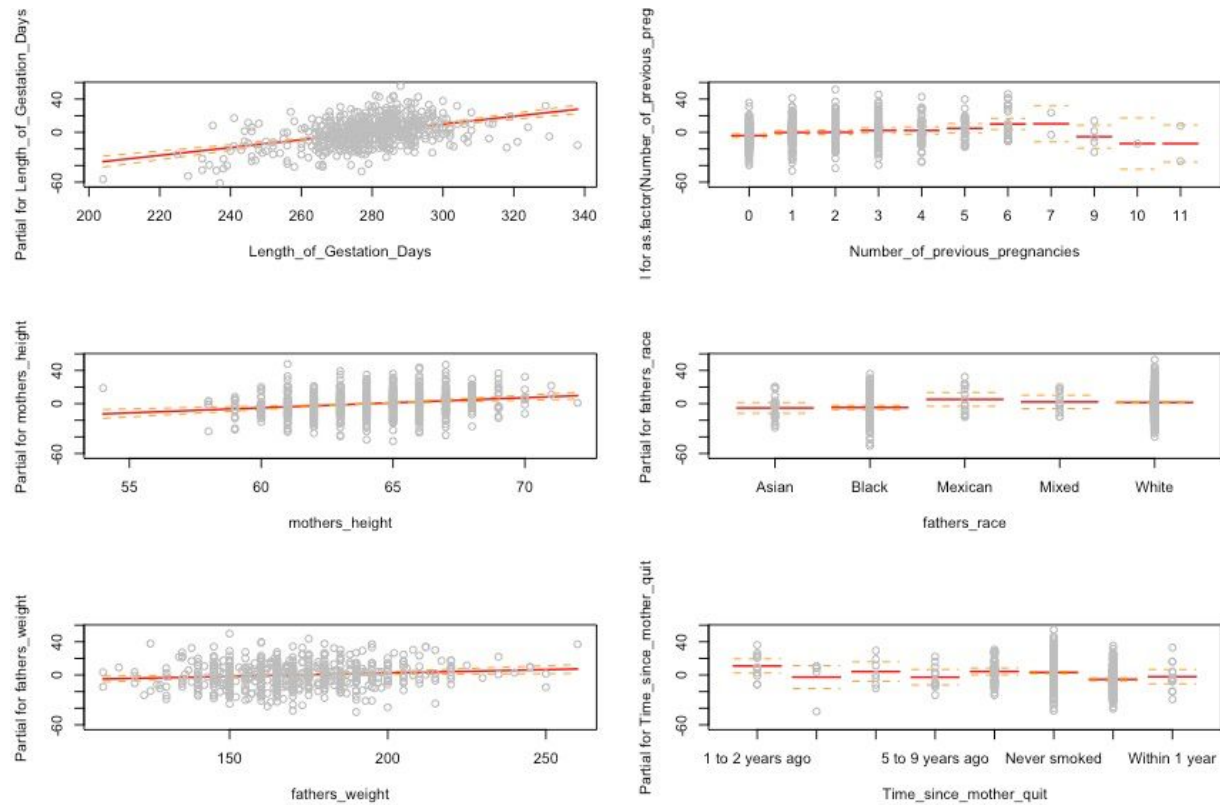


Figure 2: Partial Residual plots

The plots in Figure 2 show no evidence that a linear model is inappropriate for our data. Furthermore, there are no obvious extreme outliers in the data.

Cross Validation:

When a model is fitted, we only have information about how it performs for the dataset it was fitted against, e.g. the training dataset. The model does not provide any information on the accuracy of its predictions on new data.

To assess the predictive performance of the models, we test the model against unseen data. One common method to estimate the predictive ability of a model is by using cross-validation (Zhang & Yang, 2015). It is a data resampling method used to evaluate the generalization ability of a predictive model (Berrar, 2019) by providing an estimate for the performance of a model on new data.

Cross validation is based on splitting the data: one set of the data, also referred to as the ‘training’ set, is used for fitting the model, while the remaining data, the ‘validation’ set, is used for evaluating the performance of the model (Arlot & Celisse, 2010).

For this report we use the mean squared error to measure the predictive ability of the models. The mean squared error (MSE) is the average of the squared differences between the predicted value and the actual value. It is non-negative and the closer it is to 0, the better the predictive ability of the model is. The steps for cross validation are as follows:

1. The data is split into two parts at random.
2. One of the parts is reserved as the validation sample.
3. The remaining part of the dataset is used to fit or train the model.
4. We use the fitted model on the validation sample to obtain predicted outputs.
5. Take note of the mean square error of the model on the sample.

For the top two models, the above steps are taken to validate the data. The model that has the lower MSE value is then chosen as the final model. For the data splitting, 20% of the data is reserved as a validation set while 80% of the data is used as a training set for fitting of the model.

The results of the validation are in the table below.

Model	MSE
model 1	272.33
model 2	276.10

Table 6: MSE results from cross validation (using 20% validation sample)

Therefore, model 1 (where the birth weight depends on: the number of previous pregnancies, the mother's height, the father's race, the father's weight, and the time since the mother quit smoking) is the better model due to its lower MSE.

K-fold Cross Validation:

Another method of cross-validation, the k-fold cross-validation, is when the data is randomly split into k samples that are approximately the same size. The procedure for k-fold cross validation is similar to the one stated above, except that the data is now split into k samples and the validation steps are performed k times. The steps for k-fold cross validation are outlined as follows:

1. The data is split into K approximately sized samples.
2. One of the samples is reserved as the validation sample.
3. The remaining K-1 samples are used to fit or 'train' the model.

4. The model fitted is then used on the validation sample to obtain predicted outputs.
5. The MSE of the model on the validation sample is taken note of.
6. Steps 2 to 5 are repeated until every split sample has served as the validation sample (i.e. K times).

The prediction ability of the model is then just the average of all the MSE recorded.

A k-fold cross validation provides a better estimation of the fit of the model as each data point is used to ‘train’ the model k-1 times and validate the model once. The larger the k, in general, the more unbiased the fit of the model, but it also then has a larger variance (Rodriguez, Pérez, & Lozano, 2010). Typically, k is chosen to be between 5 and 10.

We perform a 5-fold cross validation on the two best models to confirm our conclusion from the cross validation results in the previous section.

The emphasis here is on the variable *Number_of_previous_pregnancies* which was used as a factor type variable during model selection, i.e. each distinct number is recorded as its own category. As the k-fold cross validation involves splitting the data in K samples, fitting the model and then using it on the respective samples, we would have to ensure that data points for each category of the *Number_of_previous_pregnancies* variable exists in each split sample. Before performing the 5-fold cross validation, we check that there are at least 5 data points for each category (e.g. there are at least 5 data points with 9 pregnancies, etc). If there are categories that have less than 5 data points, we remove all data points for these categories, as they cannot be split into each of the 5 samples and may cause a misleading fit or fit a model that is unable to validate certain data points. Upon checking, *Number_of_previous_pregnancies* of numbers 10 and 7 have less than 5 data points, and thus we remove all data points relating to them. We then perform the 5-fold cross validation using the pre-processed data on the top two models.

The results of the 5-fold cross validation are presented in the table below.

Model	MSE (5-fold c.v.)
model 1	251.69
model 2	252.68

Table 7: MSE results from 5-fold cross validation

The result above reaffirms the conclusion of the cross validation performed previously, which is that model 1 is the better model.

Cross validation was performed on the top two models from the model selection process to assess the predictive ability of the models on new data. 20% of the data was reserved as the validation sample, while the models were fitted on the remaining 80% of the data. The mean squared error (MSE) was used as an estimation of the predictive performance of the models, with a better-fitting model having a lower MSE. Based on the results, it was found that model 1 (where the birth weight depends on the number of previous pregnancies, the mother's height, the father's race, the father's weight, and the time since the mother quit smoking) is the better model as it has a lower MSE.

A 5-fold cross validation was also performed on the top two models. It is one of the cross validation methods which provides a better estimation of the fit of the models. The results of the 5-fold cross validation reinforces the conclusion that model 1 is the better-fitted model.

Model 1 Description

Model 1					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-102.196	20.8191	-4.909	1.20E-06	***
Length of Gestation Days	0.47037	0.04432	10.61	<2e-16	***
Number of Previous Pregnancies: 1	3.87216	1.83368	2.112	0.03516	*
Number of Previous Pregnancies: 2	4.04355	1.98273	2.039	0.04188	*
Number of Previous Pregnancies: 3	6.18063	2.22252	2.781	0.0056	**
Number of Previous Pregnancies: 4	6.17093	2.66123	2.319	0.02077	*
Number of Previous Pregnancies: 5	8.77849	3.27633	2.679	0.00759	**
Number of Previous Pregnancies: 6	14.02579	3.68033	3.811	0.00015	***
Number of Previous Pregnancies: 7	14.30361	11.0522	1.294	0.19614	
Number of Previous Pregnancies: 9	-1.27363	7.07877	-0.18	0.85728	
Number of Previous Pregnancies: 10	-9.57343	9.19023	-1.042	0.298	
mothers height	1.21731	0.26555	4.584	5.63E-06	***
Fathers Race: Black	0.51382	3.6508	0.141	0.88813	
Father's Race: Mexican	10.32308	5.29448	1.95	0.0517	.
Fathers Race: Mixed	7.2869	5.31779	1.37	0.17115	
Fathers Race: White	6.6433	3.44819	1.927	0.05454	.
Father's Weight	0.08065	0.03005	2.684	0.0075	**
Time since mother quit: 2 to 3 years ago	-13.3719	8.1437	-1.642	0.10116	
Time since mother quit: 3 to 4 years ago	-6.71229	7.24774	-0.926	0.35478	
Time since mother quit: 5 to 9 years ago	-13.4897	6.36192	-2.12	0.03442	*
Time since mother quit: During current pregnancy	-6.58358	4.81333	-1.368	0.17193	
Time since mother quit: Never smoked	-7.77924	4.40265	-1.767	0.07778	.
Time since mother quit: Still smokes	-16.1642	4.42804	-3.65	0.00029	***
Time since mother quit: Within 1 year	-12.8956	6.1428	-2.099	0.03624	*
	0 '***'	0.001 '***'	0.01 '**'	0.05 '.'	0.1 ''

Table 8: Coefficients of the modeled variables

The above table shows the coefficient estimates of the best model determined by our model selection analysis. It is important to note that there are two types of variables in our model: continuous numerical, and categorical variables. The categorical variables include

Number of Previous Pregnancies, *Father's Race*, and *Time Since Quitting*. The intercept includes one factor level for each categorical variable. In the case of our model, the intercept represents the intercept when *Number of Previous Pregnancies* is 0, *Father's Race* is Asian, and *Time Since Mother Quit* is 1-2 years.

While observing the results of the linear model, it is important to recognize the difference between statistical and practical significance. Statistical significance, as measured by the p-value of our coefficients reflects the probability that the differences observed in baby weight as determined by the predictor variable arose due to sampling variability alone. Practical significance refers to the practical impact any predictor variable has on the outcome being observed, baby weight in this case. Just because a variable is statistically significant does not make it practically significant.

A few interesting observations can be made from the summary of Model 1. It appears that as the length of gestation days increases, so does the weight of the baby. This makes sense considering that a baby's weight is related to how long it has been developing. As the height of the mother increases, so does the predicted weight of the baby. This is also not surprising, given there is often a link between parental height and child size. A similar relationship exists between the weight of the baby and the father's weight.

The categorical variables yielded some peculiar results. Interestingly, the relationship between number of previous pregnancies and weight is not constant. As the number of previous pregnancies increases, the baby weight also seems to increase, at least up until 7 previous pregnancies. Once the number of previous pregnancies exceeds 7, more previous pregnancies seems to reduce the weight of the baby. Our model also indicates that the race of the baby's father plays a role in determining the baby's weight. This could be explained similarly to the effect of the father's weight. Perhaps there is a link between race and weight or another physical characteristic. Finally, the results of the 'Time Since Mother Quit' variable are surprising. Relative to the mother quitting 1-2 years ago, the baby's weight is lower if the mother never smoked. This seems to contradict the expected relationship between smoking and baby weight. Furthermore, the baby's weight according to the model will likely be lower if the mother quit longer than 1-2 years ago. Once again this completely defies our expectation. The only expected result from this categorical variable is that the weight of the baby is expected to be lower if the mother still smokes and smokes during current pregnancy. These surprising and contradictory findings warrant further investigation.

Bootstrapping Confidence Intervals:

We use a bootstrapping method to determine the confidence intervals for our model coefficients. This method treats a sample (dataset) as the overall population and uses random

sampling to determine the accuracy of our coefficient estimates. Since statistics are often calculated on the basis of a single sample, this method is useful if the theoretical distribution is unknown. Since the original dataset is used to sample with replacement, any given row of data can be selected multiple times.

95% Confidence Intervals for Coefficients	2.50%	97.50%
(Intercept)	-142.4135069	-56.09800226
Length of Gestation Days	0.34734814	0.57489104
Number of Previous Pregnancies 1	0.42610826	7.30580203
Number of Previous Pregnancies 2	0.53690644	7.8740638
Number of Previous Pregnancies 3	1.91082102	11.12800829
Number of Previous Pregnancies 4	0.54176537	11.88847035
Number of Previous Pregnancies 5	3.34584299	14.31621639
Number of Previous Pregnancies 6	5.14939581	23.64154405
Number of Previous Pregnancies 7	-3.64352605	31.72133661
Number of Previous Pregnancies 9	-16.00357532	10.76231842
Number of Previous Pregnancies 10	-14.4243618	-5.17410166
Number of Previous Pregnancies 11	-38.16462317	14.61754429
Mothers Height	0.71261753	1.77254959
Father's Race Black	-6.17040386	7.5296009
Father's Race Mexican	0.44827563	20.75599112
Father's Race Mixed	-1.15612994	15.84546208
Father's Race White	0.53792728	12.77850418
Father's Weight	0.01596916	0.1421258
Time Since Mother Quit 2 to 3 Years Ago	-37.06992239	4.28974502
Time Since Mother Quit 3 to 4 Years Ago	-21.61503472	9.04437268
Time Since Mother Quit 5 to 9 Years Ago	-24.88558645	-0.88065919
Time Since Mother Quit During Current Pregnancy	-15.93955806	3.57174531
Time Since Mother Quit Never Smoked	-16.75861808	1.50548649
Time Since Mother Quit Still Smokes	-24.54314851	-6.67823766
Time Since Mother Quit Within 1 Year	-25.62932807	-0.02223374

Table 9: 95% Confidence Interval for Model 1

A number of samples are drawn in order to equal the number of data points from the original dataset. This process was repeated 1000 times. Each of the 1000 new sample datasets are then used to generate coefficients for the linear model. The bootstrapping method therefore creates 1000 estimates for each coefficient and these estimates in turn make it possible to determine a confidence interval for the coefficients that does not rely solely on the assumption of normality for the data.

Conclusion:

After cleaning the data and determining the best model fit, it was found that baby weight is best predicted by *Length of Gestation*, *Number of Previous Pregnancies*, *Mother's Height*, *Father's Race*, *Father's Weight*, and *Time Since Mother Quit Smoking*. While it is not entirely surprising that *Mother's Height*, *Father's Weight*, and *Length of Gestation* are associated with increases in baby weight, our model yielded surprising results for *Number of Previous Pregnancies*, and *Time Since Mother Quit Smoking*. For the most part, our model shows the opposite relationship

between smoking and baby weight than what was expected. These results warrant further analysis of the data and perhaps greater study with a different data set. These preliminary findings exemplify the problem of attributing causality based on observational data.

References:

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4, 40-79. doi:10.1214/09-SS054

Berrar, D. (2019). Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542-545). Academic Press.
doi:<https://doi.org/10.1016/B978-0-12-809633-8.20349-X>

Parascandola M. (2014). Commentary: Smoking, birthweight and mortality: Jacob Yerushalmy on self-selection and the pitfalls of causal inference. *International journal of epidemiology*, 43(5), 1373–1377. doi:10.1093/ije/dyu163

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/>.

Rodriguez, J. D., Pérez, A., & Lozano, J. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 569 - 575. Retrieved October 2019

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1), 95-112. Retrieved October 2019, from <https://doi.org/10.1016/j.jeconom.2015.02.006>

Appendices:

Appendix 1.0) Data Cleaning

```
library(dplyr)
library(tidyverse)
library(crunch)
library(ggthemes)
library(dplyr)
library(lubridate)
library(ggplot2)
library(ggthemes)
library(doParallel)
library(parallel)

babies <- read.csv("babies23.data", sep = "")
#babies <- read.csv("babies.csv")
babies

#Remove the plurality, outcome, date, and sex columns from the dataframe
babies <- subset(babies, select = -c(plurality, outcome, date, sex))

#Rename mother's weight because R assigns wt.1 when there are 2 "wt" columns (wt is the birth weight of the baby)
babies <- babies %>% rename(mothers_weight = wt.1)

#Filter all the null values
babies %>% filter(wt == 999)
babies %>% filter(gestation == 999)
babies %>% filter(race == 99 | race == 10)
babies %>% filter(inc == 98 | inc == 99)
babies %>% filter(age == 99)
babies %>% filter(ed == 9)
babies %>% filter(ht == 99)
babies %>% filter(smoke == 9)
babies %>% filter(time == 98 | time == 99 | time == 9)
babies %>% filter(drace == 99 | drace == 10)
babies %>% filter(dage == 99)
babies %>% filter(ded == 9)
babies %>% filter(dht == 99)
babies %>% filter(number == 98 | number == 99 | number == 9)

babies %>% filter(dwt == 999)
babies %>% filter(mothers_weight == 999)

#Replace unknown values with NA
babies$wt[babies$wt == 999] = NA
babies$gestation[babies$gestation == 999] = NA
babies$race[babies$race == 99] = NA
babies$race[babies$race == 10] = NA
babies$inc[babies$inc == 98] = NA
babies$inc[babies$inc == 99] = NA
babies$age[babies$age == 99] = NA
babies$ed[babies$ed == 9] = NA
babies$ht[babies$ht == 99] = NA
babies$smoke[babies$smoke == 9] = NA
babies$time[babies$time == 98] = NA
babies$time[babies$time == 99] = NA
babies$time[babies$time == 9] = NA
babies$number[babies$number == 98] = NA
babies$number[babies$number == 99] = NA
babies$number[babies$number == 9] = NA
babies$drace[babies$drace == 99] = NA
babies$drace[babies$drace == 10] = NA
babies$dage[babies$dage == 99] = NA
babies$ded[babies$ded == 9] = NA
babies$dht[babies$dht == 99] = NA
babies$dwt[babies$dwt == 999] = NA
babies$mothers_weight[babies$mothers_weight == 999] = NA
```

```

#Set all mother's and father's race from 0-5 as 5, because they are all white
#Set all Mother's and Father's education for "Trade school, HS unclear" as 6 because 6 and 7 are interchangeable
#Group all of the number of previous pregnancies greater than 10 as 10 because of the few data samples
babies$race[babies$race %in% 0:5] = 5
babies$drace[babies$drace %in% 0:5] = 5
babies$ed[babies$ed %in% 6:7] = 6
babies$deds[babies$deds %in% 6:7] = 6
babies$parity[babies$parity %in% 10:13]=10

#Write out variables as factors to ease confusion for what each number means
babies$smoke <- factor(babies$smoke,
  levels = c(0, 1, 2, 3),
  labels = c("Never", "Smokes Now", "Smoked Until Current Pregnancy", "Once smoked, doesn't now"))
babies$time <- factor(babies$time,
  levels = c(0, 1, 2, 3, 4, 5, 6, 7),
  labels = c("Never smoked", "Still smokes", "During current pregnancy", "Within 1 year", "1 to 2 years ago",
    "2 to 3 years ago", "3 to 4 years ago", "5 to 9 years ago"))
babies$inc <- factor(babies$inc,
  levels = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9),
  labels = c("Under 2500", "2500 to 4999", "5000 to 6250", "6250 to 7500", "7500 to 8750", "8750 to 10000",
    "10,000 to 11250", "11250 to 12500", "12500 to 14999", "Over 15000" ))
babies$number <- factor(babies$number,
  levels = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9),
  labels = c("0", "1 to 4", "5 to 9", "10 to 14", "15 to 19", "20 to 29", "30 to 39", "40 to 60", "Over 60",
    "smoke but dont know"))
babies$ed <- factor(babies$ed,
  levels = c(1, 2, 3, 4, 5, 6),
  labels = c("8th to 12th grade and did not graduate", "HS graduate but no other schooling",
    "HS and trade", "HS and some college", "College graduate", "Trade school HS unclear"))
babies$deds <- factor(babies$deds,
  levels = c(1, 2, 3, 4, 5, 6),
  labels = c("8th to 12th grade and did not graduate", "HS graduate but no other schooling",
    "HS and trade", "HS and some college", "College graduate", "Trade school HS unclear"))
babies$race <- factor(babies$race,
  levels = c(5, 6, 7, 8, 9),

babies$drace <- factor(babies$drace,
  levels = c(5, 6, 7, 8, 9),
  labels = c("White", "Mexican", "Black", "Asian", "Mixed"))
babies$marital <- factor(babies$marital,
  levels = c(1, 2, 3, 4, 5),
  labels = c("Married", "Legally separated", "Divorced", "Widowed", "Never married"))

#Plotting :For mothers who smoked, plotting the quantity of cigarettes smoked per day vs baby birth weight
smokes_now <- babies %>% filter(smoke == "Smokes Now")

smokes_now %>%
  gather(age, dage, ht, dht, gestation, dwt, mothers_weight, key = "param", value = "value") %>%
  ggplot(aes(x = value, y = wt, colour = number)) +
  geom_smooth() + facet_wrap(~param, scales = "free") + theme_bw()

#Rename headers to further explain what each column means
babies <- babies %>% rename(Birth_Weight = wt)
babies <- babies %>% rename(mothers_education = ed)
babies <- babies %>% rename(mothers_race = race)
babies <- babies %>% rename(mothers_age = age)
babies <- babies %>% rename(mothers_height = ht)
babies <- babies %>% rename(fathers_education = ded)
babies <- babies %>% rename(fathers_height = dht)
babies <- babies %>% rename(fathers_weight = dwt)
babies <- babies %>% rename(fathers_race = drace)
babies <- babies %>% rename(fathers_age = dage)
babies <- babies %>% rename(Family_annual_income = inc)
babies <- babies %>% rename(number_of_Cigs_per_day = number)
babies <- babies %>% rename(Number_of_previous_pregnancies = parity)
babies <- babies %>% rename(Time_since_mother_quit = time)
babies <- babies %>% rename(Length_of_Gestation_Days = gestation)

```


Appendix 2.0) Model Selection

```
# import this package to use dredge() to get subsets of models
library(MuMIn)
# import this package to use Anova(), ncvtTest(), durbinWatsonTest() and vif()
library(car)
# Import package effects to use for partial residual plotting for interaction variables
library("effects")

# read .csv data
raw_data <- read.csv("BabiesData.csv")
# omit NA value
processed_data <- na.omit(raw_data)
# number_of_data <- ceiling(nrow(processed_data) * 0.8)
# set.seed(1111)
# sequence_data <- sample(nrow(processed_data), number_of_data)
# model_data <- processed_data[sequence_data, ]
# test_data <- processed_data[-sequence_data, ]

# generate linear model
raw_model <- lm(formula = Birth_Weight ~ Length_of_Gestation_Days + as.factor(Number_of_previous_pregnancies) +
  mothers_race + mothers_age + mothers_education + mothers_height +
  mothers_weight + fathers_race + fathers_age + fathers_education +
  fathers_height + fathers_weight + marital + Family_annual_income +
  smoke + Time_since_mother_quit + number_of_Cigs_per_day,
  data = processed_data, na.action = "na.fail")

summary(raw_model)
Anova(raw_model)

# delete variable smoke because it is highly related to other variables
# delete variables which have very high p-value
raw_model <- update(raw_model, .~. - smoke - marital - Family_annual_income - fathers_age - mothers_age)
# delete variable number_of_Cigs_per_day because there are NAs in summary
# we can see from summary(raw_model), it shows that 'Coefficients: (1 not defined because of singularities)'
# NAs in the coefficients table means there are too few subjects to assess the influence of number_of_Cigs_per_days
# we can simplify this model by removing it
raw_model <- update(raw_model, .~. - number_of_Cigs_per_day)
summary(raw_model)
Anova(raw_model)

# raw_model <- dredge(raw_model)
# raw_model <- head(raw_model, n=10)

# try to add interactions in the raw_model
second_model <- lm(formula = Birth_Weight ~ Length_of_Gestation_Days + as.factor(Number_of_previous_pregnancies) +
  mothers_race + mothers_education + mothers_height + mothers_weight +
  fathers_race + fathers_education + fathers_height + fathers_weight +
  Time_since_mother_quit + mothers_education:Length_of_Gestation_Days +
  mothers_race:Length_of_Gestation_Days, data = processed_data, na.action = "na.fail")

summary(second_model)
Anova(second_model)

# using dredge() to find all possible subsets
# select top4 models with AIC value
new_model <- dredge(second_model)
models <- head(new_model, n=4)

model_1 <- models[1]
model_1 <- lm(formula = Birth_Weight ~ Length_of_Gestation_Days + as.factor(Number_of_previous_pregnancies) +
  mothers_height + fathers_race + fathers_weight + Time_since_mother_quit,
  data = processed_data, na.action = "na.fail")
summary(model_1)

model_2 <- models[2]
model_2 <- lm(formula = Birth_Weight ~ Length_of_Gestation_Days + as.factor(Number_of_previous_pregnancies) +
  mothers_height + mothers_weight + fathers_race + fathers_weight + Time_since_mother_quit,
  data = processed_data, na.action = "na.fail")

model_3 <- models[3]
model_3 <- lm(formula = Birth_Weight ~ Length_of_Gestation_Days + as.factor(Number_of_previous_pregnancies) +
  mothers_education + mothers_height + fathers_race + fathers_weight +
  Time_since_mother_quit + mothers_education:Length_of_Gestation_Days,
  data = processed_data, na.action = "na.fail")

model_4 <- models[4]
model_4 <- lm(formula = Birth_Weight ~ Length_of_Gestation_Days + as.factor(Number_of_previous_pregnancies) +
  mothers_race + mothers_education + mothers_height + fathers_weight +
  Time_since_mother_quit + mothers_education:Length_of_Gestation_Days +
  mothers_race:Length_of_Gestation_Days, data = processed_data, na.action = "na.fail")
```

```

# Checking model assumptions
# Error Distribution
# p-value is 0.7971, fail to deny H0, so residuals of model obey normal distribution
qqnorm(resid(model_1))
qqline(resid(model_1))
shapiro.test(resid(model_1))
hist(resid(model_1))
# p-value is 0.7252, fail to deny H0, so residuals of model obey normal distribution
qqnorm(resid(model_2))
qqline(resid(model_2))
shapiro.test(resid(model_2))
hist(resid(model_2))
# p-value is 0.4678, fail to deny H0, so residuals of model obey normal distribution
qqnorm(resid(model_3))
qqline(resid(model_3))
shapiro.test(resid(model_3))
hist(resid(model_3))
# p-value is 0.4153, fail to deny H0, so residuals of model obey normal distribution
qqnorm(resid(model_4))
qqline(resid(model_4))
shapiro.test(resid(model_4))
hist(resid(model_4))

# constant spread
# p-value is 0.47521, fail to deny H0, so residuals have constant variance
ncvTest(model_1)
# p-value is 0.35188, fail to deny H0, so residuals have constant variance
ncvTest(model_2)
# p-value is 0.64445, fail to deny H0, so residuals have constant variance
ncvTest(model_3)
# p-value is 0.90684, fail to deny H0, so residuals have constant variance
ncvTest(model_4)

# Independence
# p-value is 0.06, fail to deny H0, so residuals are independent
durbinWatsonTest(model_1)
# p-value is 0.06, fail to deny H0, so residuals are independent
durbinWatsonTest(model_2)
# p-value is 0.07, fail to deny H0, so residuals are independent
durbinWatsonTest(model_3)
# p-value is 0.038, deny H0, so residuals are not independent
durbinWatsonTest(model_4)

```

Appendix 2.1) Model selection: Model 2

Model 2					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-96.4442	21.3002	-4.528	7.29E-06	***
Length of Gestation Days	0.46443	0.04455	10.43	<2e-16	***
Number of Previous Pregnancies: 1	3.74265	1.83557	2.039	0.04193	*
Number of Previous Pregnancies: 2	3.74918	1.99532	1.879	0.06077	.
Number of Previous Pregnancies: 3	5.71934	2.25114	2.541	0.01134	*
Number of Previous Pregnancies: 4	5.95845	2.66512	2.236	0.02577	*
Number of Previous Pregnancies: 5	8.58795	3.27805	2.62	0.00904	**
Number of Previous Pregnancies: 6	13.53179	3.69909	3.658	0.00028	***
Number of Previous Pregnancies: 7	13.31731	11.0739	1.203	0.22965	
Number of Previous Pregnancies: 9	-2.52461	7.14393	-0.353	0.72393	
Number of Previous Pregnancies: 10	-9.85233	9.18797	-1.072	0.28405	
Mothers Height	1.07681	0.28776	3.742	0.0002	***
Mothers Weight	0.04718	0.03734	1.263	0.20694	
Fathers Race: Black	-0.12388	3.6836	-0.034	0.97318	
Fathers Race: Mexican	10.13563	5.29373	1.915	0.05605	.
Father's Race: Mixed	6.23395	5.37989	1.159	0.24706	
Father's Race: White	6.33453	3.455	1.833	0.06727	.
Father's Weight	0.07769	0.03013	2.579	0.01017	*
Time since mother quit: 2 to 3 years ago	-13.0762	8.14271	-1.606	0.10887	
Time since mother quit: 3 to 4 years ago	-7.16172	7.2526	-0.987	0.32384	
Time since mother quit: 5 to 9 years ago	-14.1016	6.37694	-2.211	0.02742	*
Time since mother quit: During current pregnancy	-6.50475	4.81116	-1.352	0.17692	
Time since mother quit: Never smoked	-7.8954	4.40125	-1.794	0.07337	.
Time since mother quit: Still smokes	-16.1388	4.42573	-3.647	0.00029	***
Time since mother quit: Within 1 year	-12.9844	6.13992	-2.115	0.03489	*
	0 '***'	0.001 '***'	0.01 '**'	0.05 '.'	0.1 ''

Appendix 3.0) Validation

```
library(dplyr)
#we use cleaned up data 'processed_data'
raw_data <- read.csv("BabiesData.csv")
processed_data <- na.omit(raw_data)
babies_df <- processed_data

#according to the result of AIC we choose top two models as model_1 and model_2
model1 <- Birth_Weight ~ Length_of_Gestation_Days + as.factor(Number_of_previous_pregnancies) +
  mothers_height + fathers_race + fathers_weight + Time_since_mother_quit

str(babies_df)

model2 <- Birth_Weight ~ Length_of_Gestation_Days + as.factor(Number_of_previous_pregnancies) +
  mothers_height + mothers_weight + fathers_race + fathers_weight + Time_since_mother_quit

# VALIDATION ON 20% OF DATA
set.seed(111)
sample <- sample(seq_len(nrow(babies_df)),size = floor(0.8*nrow(babies_df)),replace=F)
training_sample <- babies_df[sample,]
validation_sample <- babies_df[-sample,]

##train model on training sample and predict using validation sample
#MODEL_1
train_model1 <- lm(formula=model1,data=training_sample) #train model_1
pred_model1 <- train_model1 %>% predict(validation_sample) #predict by model_1
mse_model1 <- mean((validation_sample$Birth_Weight - pred_model1)^2) #calculate mse for model1

#MODEL_2
train_model2 <- lm(model2,data=training_sample) #fit model2
pred_model2 <- train_model2 %>% predict(validation_sample) #validate fitted model2
mse_model2 <- mean((validation_sample$Birth_Weight - pred_model2)^2) #calculate mse for model2
```

Appendix 4.0) 5 Fold Cross Validation

```
#(STEP-BY-STEP, i.e. splitting data-fitting model-predicting-calculate mse)
library(caret)
k=5
mse_five_fold <- NULL
#if number of data points in each level for Number_of_Pregnancies <5, remove
subset(babies_df,Number_of_previous_pregnancies==11)
subset(babies_df,Number_of_previous_pregnancies==10)
subset(babies_df,Number_of_previous_pregnancies==9)
subset(babies_df,Number_of_previous_pregnancies==7)
a <- subset(babies_df, Number_of_previous_pregnancies <=6)
a <- rbind(a,subset(babies_df,Number_of_previous_pregnancies==9))

k_folds_cv <- function(k,model_to_fit){
  #split the data into 5 'folds'; assign one fold as validation set, and the remaining as training set
  folds <- createFolds(a$Birth_Weight,k=k,list=T,returnTrain = T)
  for (i in 1:k){
    #fit model to the training set
    train_model <-lm(model_to_fit,data=a[folds[[i]],],)
    #use fitted model to predict values on validation set
    pred <- predict(object = train_model, newdata=a[-folds[[i]],])
    #calculate and store mse values
    mse_five_fold[i] <- mean((a[-folds[[i]],]$Birth_Weight-pred)^2)
  }
  print(mse_five_fold) #outputs mse for each cv (where each 'fold' was assigned the sample)
  mean(mse_five_fold) #outputs average of the mse values
}

#run k-fold cv function from above on both models to compare mse values
set.seed(111)
k_folds_cv(5,model1)
k_folds_cv(5,model2)

best_model <- model1
#END#
```


Appendix 5.0) Bootstrapping

```
# read in cleaned babies data file
raw_data<-read.csv("BabiesData.csv")
#Omit all NA values from the cleaned babies data file
BabiesData<-na.omit(raw_data)
BabiesData$Number_of_previous_pregnancies<-as.factor(BabiesData$Number_of_previous_pregnancies)
# Detect number of cores
nCores<-detectCores()
# Define cluster you make
myClust<-makeCluster(nCores-1,type="FORK")
# Bootstrap confidence intervals for all covariates and parameters
lmBoot <- function(inputData, nBoot,regmodel){
  # Create variable myformula to use in regression model later
  myformula<-as.formula(regmodel)
  # Create variable called input data
  nrowdata<-nrow(inputData)
  # Create a list for compiling Bootstraps
  bootList<-list()
  # Create a for loop to run multiple bootstraps
  for(i in 1:nBoot){
    # Create a Sample Index
    sampleIndex <- sample(1:nrowdata, nrowdata, replace = T)
    # Sample the dataset
    bootData<- inputData[sampleIndex,]
    #Compile all bootstrap samples into a list called bootData
    bootList[[i]]<-bootData
  }

  # Run lm model on all bootstrap samples to give 'nBoot' coefficient estimates
  ourBootReg<- parLapply(myClust,bootList,function(itemFromList){coef(lm(myformula,data=itemFromList))})
  # Compile all coefficient estimates into one dataframe
  dataframeOfBootCoefs<-plyr::ldply(ourBootReg,rbind)
  #return(ourBootReg)
  # take central 95% of each columns estimates (simulated sampling dists)
  parApply(myClust,dataframeOfBootCoefs,2,quantile,probs=c(0.025,0.975),na.rm=TRUE)
}

# Run lmBoot function on our best model

lmBoot(BabiesData,100,Birth_Weight ~ Length_of_Gestation_Days + mothers_height+fathers_race+fathers_weight + Time_since_mother_quit)
lmBoot(BabiesData,100,Birth_Weight ~ Time_since_mother_quit)
lmBoot(BabiesData,100,Birth_Weight ~ Number_of_previous_pregnancies)
lmBoot(BabiesData,100,Birth_Weight ~ Length_of_Gestation_Days + mothers_height + fathers_race + fathers_weight)
lmBoot(BabiesData,100,Birth_Weight ~ Length_of_Gestation_Days + Number_of_previous_pregnancies + mothers_height +
  fathers_race + fathers_weight + Time_since_mother_quit)
# Stop cluster used by lmBoot
stopCluster(myClust)
```