School of Mathematics and Statistics

MT5762 Group Project - Linear models, Selection and Inference

A study into the relationship between baby birth-weights and measured variables, with the aim to identify potential drivers of low birth-weight babies.

Module convenor: Carl Donovan

Student IDs: 180029290, 180021390, 180028774, 180016064

Word Count: 3188

Submission deadline: 02/11/2018

Executive Summary

An investigation was undertaken to understand if the birth weight of babies could be related to measured variables captured during the group of studies, Child Health and Development Studies (CHDS). Using the sample data set, findings indicated a linear model could be created using statistical concepts such as AIC and *k*-fold cross validation. However, as the model failed to meet all assumptions required of a linear model it could not reliably, with a strong degree of certainty describe potential drivers of baby weights. Recommendations are therefore given in the discussion and summary sections to support future investigations in their pursuit of a linear model(s) to identify drivers of baby birth-weights.

Contents

1. Introduction

The purpose of this study is to explore and identify relationships between the birth weight of babies, and variables such as, gestation, mother's height and weight, father's height and weight, whether the mother smoked during pregnancy, and if so the number of cigarettes smoked per day. The investigation aims to produce a linear model(s) to identify statistically viable drivers of low birth-weights of babies. Alternatively, to provide recommendations which would help improve similar studies in this area of research. The data utilised was derived from a larger group of studies, the Child Health and Development Studies (CHDS).

2. Data Characteristics

All analysis performed for this investigation used the R Programming Language and was generated using RStudio IDE software application. Figures and tables created during the exploration stage were also mirrored using Microsoft Excel (2013) to support understanding of babies23 data set.

2.1 Data Cleansing

An initial look at the dataset provided (*babies23*) confirms there are n = 1236 observations in total with 23 columns. Full definitions for each column can be found in Appendix I.

Four columns were identified which contained no new or supportive information for the analysis. The columns *pluralty*, *sex* and *outcome* all contain single results, confirming the babies23 data set represents only single, male babies who all survived for at least 28 days. The *id* column was also removed, as it is the identification number for each baby's birth. Therefore, the column contains no predictive value.

To clearly distinguish between covariates which relate to the father and those to the mother, the names have been altered to start either with "*m"* for mother (e.g. *msmoke* = does the mother smoke) or "*d*" for the father (e.g. *ded* = father's education).

2.2 Baby Weight Distribution

To aid the understanding of baby weight distribution, a new column was added to represent low birth weight. This is calculated where baby's weight is less than or equal to 88 ounces (2.5kg) (Pederson 2018). Table 1 confirms babies born with a low birth weight represents 5% of our data.

A histogram (Figure 1) is used to visualise this result and shows the distribution of all male baby weights in the data set. Figure 1 confirms baby weights generally follow a normal distribution. The lower 5% tail represents low birth weight babies.

Low Birth Weight	Number of Observations	Proportion of Data Set (%)
No	1173	0.95
Yes	63	0.05
Total	1236	1.00

Table 1 - Summary of the number of observations, and percentage proportion split by low birth weight. (*Figures rounded to 2 decimal places*)

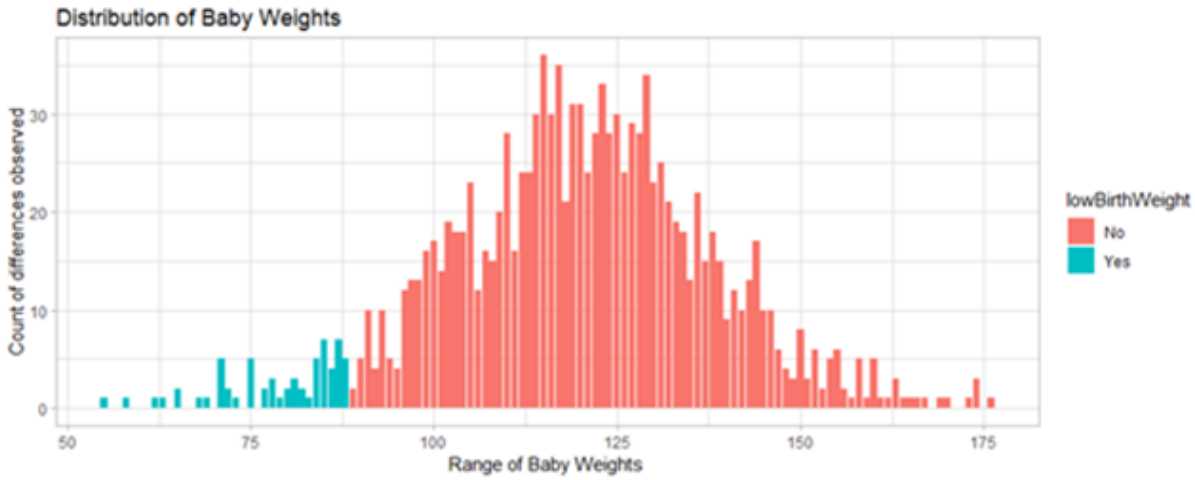


Figure 1 - Distribution of male baby weights (weights measured in ounces), coloured by low birth weight measure.

2.3 Notable relationships between variables

The relationship of the two variables *msmoke* and *mtime* was explored in further detail. The mothers smoking status is categorised by *msmoke*, while *mtime* addresses if the mother quit, how long ago did she quit. When *msmoke* is "until current pregnancy", the response for how long ago she quit (*mtime*) is "during current pregnancy". Collinearity was highlighted between the two variables. Quitting during current pregnancy is ambiguous, and so it was agreed to remove *mtime* from the analysis. The ambiguity is highlighted by the red point in Figure 2.

A second collinear relationship was identified between *msmoke* and *mnumber*. In this instance, *msmoke* was removed from the analysis due to *mnumber* yielding the better predictive value during the screening of the models.

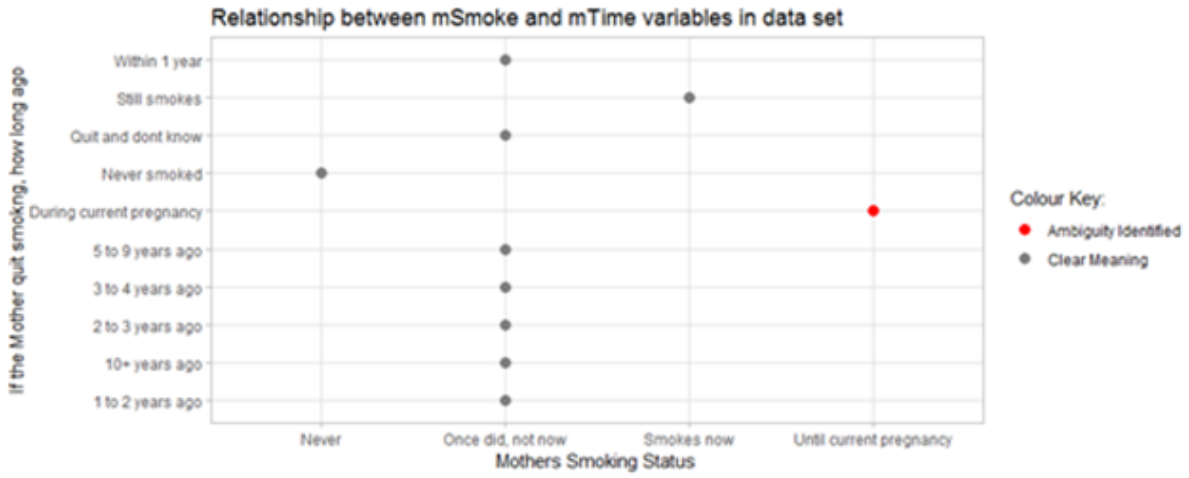


Figure 2 - Relationship between mother's smoking status and how long ago mother quit (if she did quit).

3. Methodology

The identification of predictive variables used to model baby weights can be approached in many ways. This investigation explores one first order interaction effect, two manual computational techniques (p-value and adjusted R-squared criteria), and one automated method. The automated method used is stepwise regression selection Akaike's Information Criterion (AIC).

3.1 First order interaction model

As a starting point, the study interprets one fitted interaction effect based off prior beliefs (*priori logical).* The predictor variables *msmoke* and *gestation* were explored against the response variable *wt* (the baby weights). Smoking is commonly held to be linked to lower baby birth weight, and logically so is a shorter gestation period. To investigate this, a new column, *mEverSmoked*, was made which indicates whether the mother has ever smoked. A linear model was fitted with the interaction of these covariates to predict the baby weight. Hence, Model 1 is;

$$wt \sim gestation:mEverSmoked$$

3.2 p-value

The p-value backward stepwise comparison is used to screen for the most suitable model selection modification. Using the 'backward' method the initial model/starting point contains all covariables which are tested. The p-values are obtained, and the least statistically significant variable is excluded. This process is repeated until only statistically significant p-values remain. The null hypothesis for these tests is that the (1), the slope of the linear association for that

covariable, is zero. Only variables which reject this null hypothesis are retained. These variables may be used to predict the response variable, and hence are incorporated in the model.

Predictor variable elimination using p-value criteria yields the following linear model (Model 2);

$$wt \sim gestation + mparity + mht + drace + mnumber$$

An important note of the p-value criteria for model selection is that it is regarded as a 'biased' method since the next best model depends on the previous step. This could lead to biased results and unexplored model possibilities.

3.3 Adjusted R-square

A second linear model was fitted using backwards elimination. However, this time adjusted R-squared values were used to exclude predictors. Adjusted R-squared takes into account the number of predictors included in the model and is optimised when all the predictors add value. The penalty adds fidelity and parsimony to the data. Predictors were removed one at a time from the model, until the value of adjusted R-squared was as high as possible.

Selection through adjusted R-squared methodology produces Model 3;

$$wt \sim gestation + mparity + mht + drace + ded + dwt + mnumber$$

3.4 AIC stepwise comparison

To apply this method a linear model was created which contains all explanatory variables in the data set. The "*step"* function in R was utilised from the "*car"* package. This iteratively steps through AIC theory and returns Model 4 as the covariates which yield the lowest score. The theory for AIC stepwise comparison is formulated as:

$$AIC= -2(k) + 2p$$

where the k is the log-likelihood, and p shows the number of parameters in the model. The AIC stepwise comparison method seeks to identify a model that minimizes this value. AIC stepwise comparison uses a penalised likelihood for model selection.

The resulting model (Model 4) is;

$$wt \sim gestation + mparity + mht + drace + dwt + mnumber$$

A noteworthy detail of fitting using AIC is the algorithmic inclusion/exclusion of covariates can generate a model which might be sensitive to noise. AIC optimises for fit, not for interpretability. Moreover, if automated, the intermediate models are not checked for validity.

4. Methodology of Model Selection

To select the 'best' model AIC, BIC and k-fold cross-validation techniques were considered. The model selection criteria are detailed below.

4.1 AIC and BIC

In general, AIC is considered the first model selection criterion that should be used to select the 'best' model (Fabozzi et al 2014). AIC takes into consideration the trade-off between making a model more complex by requiring more parameters, and the penalty imposed by the additional parameters. Bayesian Information Criterion (BIC) imposes a greater penalty for the number of parameters than AIC. For both the AIC and BIC selection criteria, the 'best' model has the lowest output relative to the other models. BIC is more useful when selecting a correct model. However, AIC is more appropriate when finding the best model for predicting future observations (Chakrabarti and Ghosh 2011).

4.2 k-fold cross-validation

Cross validation is a popular strategy for model selection. The method splits data to estimate the risk associated with each model. The original sample is split into *k* equal subsets. One subset is set aside and is referred to as the validation sample. The remaining *k-1* subsets are referred to as the training sample and are used to train the fitted model. The process of cross-validation then takes place *k* times. This includes splitting the training sample into further *k* subsets. Each subset undergoes *k* iterations, with each of the *k* subsets used exactly once as a validation data set, while the others are used to train the model for that iteration. The results returned are averaged to provide one estimation. Cross validation avoids overfitting, as the training and validation samples are kept independent when the data is *iid* normal (Arlot and Celisse 2010).

Typically, the k-fold cross-validation is performed using k = 5 or k = 10. Empirical testing has shown these values to produce test error rate estimates which do not include disproportionately high bias or variances (James 2013). This investigation used 5-fold cross-validation (k = 5) to identify the 'best' model.

4.3 Final Model Selection

Table 2 shows the results of the model selection criteria. The table highlights that models 2 and 4 (p-value and AIC) are notably better than the other two. 5-fold cross-validation scores are very similar between the two. Model 4 (AIC) was selected as the 'best' model as it scored the lowest AIC result of 5003.58 and is as follows;

$$wt \sim gestation + mparity + mht + drace + dwt + mnumber$$

Statistical reasonings suggest that these six covariables have the strongest relationship to the response variable, birth weight of babies.

Model	AIC	BIC	Cross Validation (CV) Type	CV - RMSE	CV - R-squared	CV - MAE
Model 1 First order interaction	5067.36	5084.94	k = 5	16.70	0.20	13.02
Model 2 p-value	5005.49	5080.21	k = 5	15.89	0.28	12.52
Model 3 adjusted R-squared	5005.45	5150.49	k = 5	16.19	0.26	12.80
Model 4 AIC	5003.58	5082.70	k = 5	15.92	0.28	12.50

Table 2 - Linear model appraisal table, summaries AIC, BIC and 5-fold cross validation techniques. (*Figures rounded to 2 decimal places*)

5. Model Diagnostics

Before using the 'best' model to make predictions, the following assumptions of linear models are examined;

- Constant Spread
- Linearity
- Collinearity
- Normality
- Independence

Results for which are detailed below.

5.1 Constant Spread Assumption

Tests for non-constant errors were conducted using the Breusche-Pagan test. Where the Null Hypothesis (Ho) states that the errors are homoscedastic/have constant variance. Since the p-value result is greater than 0.05 the null hypothesis cannot be rejected, thus residuals are homoscedastic and constant spread assumption holds true (Figure 3). Findings are detailed in Appendix II.

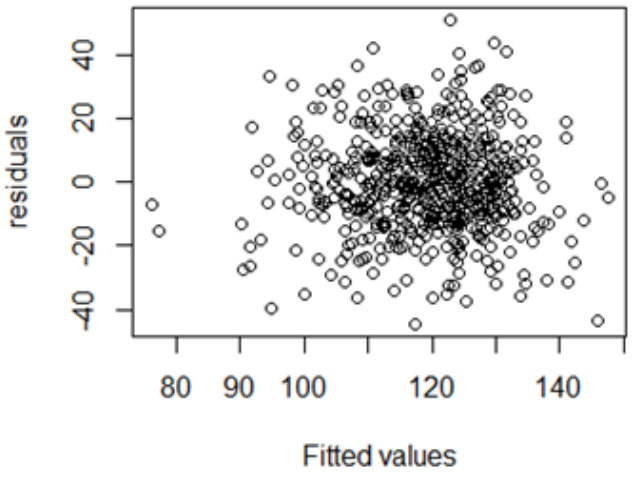


Figure 3 – A visualisation of the constant spread of the errors.

5.2 Assumption of Linearity

To assess linearity for each covariate, partial residual plots were produced which confirmed assumption of linearity holds true (Figures can be found in Appendix II).

5.3 Diagnosing collinearity

This was assessed using Variance Inflation Factors (VIFs) to measure the instability of parameter estimates due to dependencies between covariates. No covariates were highlighted; hence collinearity assumption holds true.

5.4 Assumption of Independence

Tests for serial correlation of residuals were completed using Durbin-Watson test (Ho: the errors are serially uncorrelated). Since the p-value result is less than 0.05 the null hypothesis is rejected. The alternative hypothesis is accepted that the residuals are autocorrelated. Therefore, assumption of independence does not hold true.

5.5 Assumption of Normality

Shapiro-Wilks test for normality was performed (Ho: the residuals are normally distributed). The p-value result is greater than 0.05, and so the null hypothesis fails to be rejected (See Appendix II). This is complimented by the Quantile-Quantile (QQ) plot in Figure 4 which shows that residuals roughly follow a straight line. Hence, assumption of normality holds true.

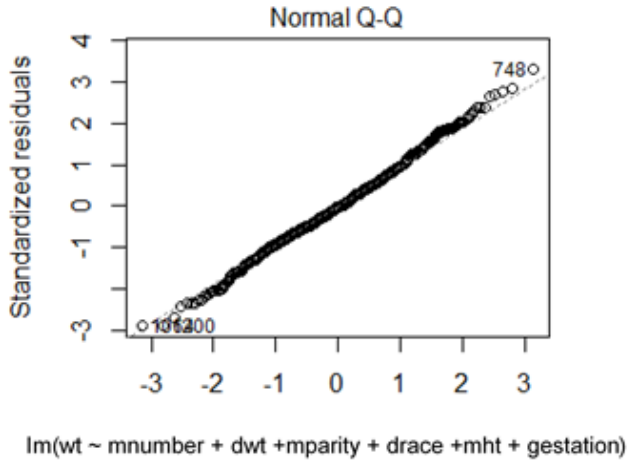


Figure 4 – Quantile-quantile (QQ) plot for the 'best' model.

6. Bootstrapping the 'best' model

Non-parametric bootstrap technique was used to obtain upper and lower 2.5% confidence intervals on parameter estimates from the 'best' model (Figure 5). The advantage of conducting the bootstrap, instead of relying on the parametric confidence intervals, is that the parameter estimates do not need to follow a normal distribution. The bootstrap confidence intervals have been constructed with 999 bootstrap resamples. The results are then compared with the parametric confidence intervals of the 'best' model.

Confidence intervals indicate that coefficients for *mnumber* '40-60', '5-9', 30-39, 'Never', 'smoked but not now', as well as '*drace* Black' all contain zero value as a possible result (see Figure 5). Therefore, the null hypothesis cannot be rejected for these coefficients. That is to say, they do not support variance in baby weight variable.

7. Discussion

In order to address the concerns and interpretations appropriately, both a qualitative and a quantitative interpretation of the current investigation follow.

7.1 Qualitative Interpretation

Firstly, the data has been gathered using observational methodology. The variables are not consciously altered to identify how they impact the birth-weight of babies, as it would be unethical

Coefficients	Confidence Intervals 2.5%	97.5%
Intercept	-159.14	-78.11
mnumber10-14	-14.97	-3.75
mnumber15-19	-17.91	-6.51
mnumber20-29	-15.31	-6.57
mnumber 30-39	17.48	-3.24
mnumber 40-60	-24.18	14.19
mnumber 5-9	-10.14	0.13
mnumber never	-5.52	2.56
mnumber smoke but don't now	-13.18	4.98
dwt	-0.008	0.12
mparity	0.22	1.72
draceBlack	-5.42	8.81
draceMex	3.31	23.25
draceMixed	2.14	20.56
dracewhite	1.63	14.54
mht	0.82	1.84
gestation	0.38	0.59

Figure 5 – Empirical confidence intervals for the 'best' model obtained from bootstrapping.

(*Figures rounded to 2 decimal places*)

and implausible to conduct an experimental study which may impact the health of a newborn baby.

The precise process of data collection is unknown and therefore could be prone to human error. Especially if, for example the mother's weight is self-measured or assessed.

The assumption of independence does not hold true for the 'best' model. Therefore, relationships between measured variables and the birth weight of babies may be non-linear, for example a quadratic relationship.

Predicting baby weight is not a simple problem. The authors believe there are more explanatory variables not captured in the current data set that could offer greater predictive power. An example could be to include the birth weight of mothers and fathers, as birth weight may be hereditary and biologically linked.

A crucial point to note is the current data only relates to male babies. This is indicated by the *sex* column of the dataset, and thus limits any predictive model to a biased perspective of the real-world environment (more than one gender exists).

Details such as geographical, temporal origin and relative population demographics are unknown at this stage. Considerations towards confounding variables like these are critical to robustly generate inferences to a wider population from a representative sample. Furthermore, consulting subject matter experts to aid and support data set understanding and interpretations of known but rare covariate relationships to baby weight, could help to result in more informed analysis.

7.2 Quantitative Interpretation

The 'best' model has six covariates; *gestation, mparity, mht, drace, dwt and mnumber*.

Mathematical notation in a linear model form is;

$$wt = b0 + (b1 * gestation) + (b2 * mparity) + (b3 * mht) + (b4 * drace) + (b5 * dwt) + (b6 * mnumber)$$

Appendix IV presents a thorough interpretation of all parameter estimates shown in Table 3.

Example interpretations include:

- A 1 ounce increase in weight of an infant tends to be correlated with an increase of 0.96 in total number of pregnancies (*mparity*), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with a decrease of 12.17 units for 15-19 number of cigarettes smoked per day for past and current smokers compared with the baseline of 1-4 cigarettes smoked per day (*mnumber15-19*), ceteris paribus.

Coefficients	Estimates	Std. Error	t-value	p-value	Significance
Intercept	-116.25	19.93	-5.83	9.09e-09	***
mnumber10-14	-9.06	3.11	-2.92	3.69e-03	**
mnumber15-19	-12.17	5.25	-2.32	0.02	*
mnumber20-29	-11.06	2.39	-4.62	4.64e-06	***
mnumber30-39	-10.53	4.10	-2.57	0.01	*
mnumber40-60	-5.10	7.26	-0.70	0.48	-
mnumber5-9	-4.90	2.59	-1.89	0.06	.
mnumberNever	-1.42	2.03	-0.70	0.48	
mnumbersmoke but don't now	-4.79	6.63	-0.72	0.47	
dwt	0.06	0.03	1.95	0.05	.
mparity	0.97	0.35	2.78	5.61e-03	**
draceBlack	1.64	3.70	0.44	0.66	
draceMex	12.72	4.97	2.56	0.01	*
draceMixed	11.64	5.34	2.18	0.03	*
dracewhite	8.31	3.49	2.38	0.02	*
mht	1.32	0.26	5.01	7.19e-07	***
gestation	0.49	0.04	11.29	2e-16	***

Table 3 - A summary of the coefficient estimates of the 'best' model, those noted with *'s indicate statistical significance to the response variable. (*Figures rounded to 2 decimal places*)

The above interpretations of the statistical output appear sound from both an analytical and practical standpoint, the baby birth-weight can be predicted using the parameter estimates and state linear associations. Another two interpretation examples include:

- A 1 ounce increase in weight of an infant tends to be correlated with an increase of 12.72 units in father's race from Mexican (*draceMex)* background compared with the baseline of Asian background, ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with an increase of 8.31 units in father's race from White (*dracewhite*) background compared with the baseline of Asian background, ceteris paribus.

The quantitative interpretation of the examples above are valid as the 'best' model has found '*drace'* to be a statistically significant predictive variable for baby birth-weights. However, this interpretation leaves us none the wiser as ethnicity cannot be quantified in units. Moreover, this is an observational study, and so variables may sway heavily considering variations in the gathered sample. Variables may also be affected by related external factors, such as geographical or socio-economic factors, which may not have been captured in the current investigation. In summary, the practicality of the 'best' model faces limitations with regards to its quantitative methodology and population representativeness, but also to some extent by the difference between statistical and actual values.

8. Summary

The main objective of this study was to produce a linear model(s) to identify statistically viable drivers of low birth-weights of babies. While a 'best' model was selected using statistical methodologies (including AIC and 5-fold cross-validation), which indicated possible strong relationships between six covariables and babies birth weight;

$$wt \sim gestation + mparity + mht + drace + dwt + mnumber$$

The model failed to meet all assumptions required for a linear model. Therefore, it could not be used to describe potential drivers of low birth-weight babies reliably with a strong degree of statistical certainty.

Recommendations to enable future investigations include:

- Considering more than one gender.
- Documenting considerations for confounding factors that may be the possible cause of any indications detected. Also agree upon a cohesive strategy to minimise wherever possible (such as geographical and temporal origin).
- Applying a stratified random sampling approach to capture a representative sample of a specified population, in a non-discriminatory way
- Carefully consider the included covariates, to minimise collinearity wherever possible (potentially with the help of subject matter experts to aid and support). Alternatively, if similar covariates are necessary and cannot be avoided, clear documentation which justifies the purpose of variables behind this selection should be included.

Recommendations are not limited to the above but would help produce a linear model(s) to identify statistically viable drivers of low birth-weights babies.

9. References

Arlot, S. and Celisse, A., (2010) 'A survey of cross-validation procedures for model selection', Statistics surveys. The author, under a Creative Commons Attribution License, 4, pp. 40–79.

Chakrabarti, A. and Ghosh, J. K., (2011) 'AIC, BIC and Recent Advances in Model Selection', in Bandyopadhyay, P. S. and Forster, M. R. B. T.-P. of S. (eds) Handbook of the Philosophy of Science. Amsterdam: North-Holland, pp. 583–605.

Fabozzi, F.J., Focardi, S.M., Rachev, S.T. and Arshanapalli, B.G., (2014). *The basics of financial econometrics: Tools, concepts, and asset management applications*. John Wiley & Sons.

James, G., Witten, D., Hastie, T. and Tibshirani, R., (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

Pederson, L. M., (2018) 'Guidelines for low birth weight: a literature review comparing national guidelines in Lao PDR with WHO guidelines', *FEATURE: Manifesto à la mode (p. 2)*, p. 28.

R: A Language and Environment for Statistical Computing, R Core Team, R Foundation for Statistical Computing., (2018). Vienna, Austria https://www.R-project.org/

RStudio Team. (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA http://www.rstudio.com/

Appendix I - **Variables in data file**

1. id - identification number
2. pluralty - 5= single fetus
3. outcome - 1= live birth that survived at least 28 days
4. date - birth date where 1096 = January1,1961
5. gestation - length of gestation in days
6. sex - infant's sex 1=male 2=female 9=unknown
7. wt - birth weight in ounces (999 unknown)
8. mparity - total number of previous pregnancies including
fetal deaths and still births, 99=unknown
9. mrace - mother's race 0-5=white 6=mex 7=black 8=asian 9=mixed 99=unknown
10. mage - mother's age in years at termination of pregnancy, 99=unknown
11. med - mother's education 0= less than 8th grade,
1 = 8th -12th grade - did not graduate,
2= HS graduate--no other schooling, 3= HS+trade,
4=HS+some college 5= College graduate, 6&7 Trade school HS unclear, 9=unknown
12. mht - mother's height in inches to the last completed inch
99=unknown
13. mwt - mother prepregnancy wt in pounds, 999=unknown
14. drace - father's race, coding same as mother's race.
15. dage - father's age, coding same as mother's age.
16. ded - father's education, coding same as mother's education.
17. dht - father's height, coding same as for mother's height
18. dwt - father's weight coding same as for mother's weight
19. marital 1=married, 2= legally separated, 3= divorced,
4=widowed, 5=never married
20. inc - family yearly income in $2500 increments 0 = under 2500,
1=2500-4999, ..., 8= 12,500-14,999, 9=15000+,
98=unknown, 99=not asked
21. msmoke - does mother smoke? 0=never, 1= smokes now,
2=until current pregnancy, 3=once did, not now, 9=unknown
22. mtime - If mother quit, how long ago? 0=never smoked, 1=still smokes,
2=during current preg, 3=within 1 yr, 4= 1 to 2 years ago,
5= 2 to 3 yr ago, 6= 3 to 4 yrs ago, 7=5 to 9yrs ago,
8=10+yrs ago, 9=quit and don't know, 98=unknown, 99=not asked
23. mnumber - number of cigs smoked per day for past and current smokers
0=never, 1=1-4, 2=5-9, 3=10-14, 4=15-19, 5=20-29, 6=30-39, 7=40-60, 8=60+, 9=smoke but don't know, 98=unknown, 99=not asked

Appendix II - Figures and tables from diagnosing assumptions of the 'best' model

Table 1. Non-constant Variance Test

Non-constant Variance Score Test
Variance formula: ~ fitted.values , Chi square = 0.01369702, Df = 1, p-value = 0.90683

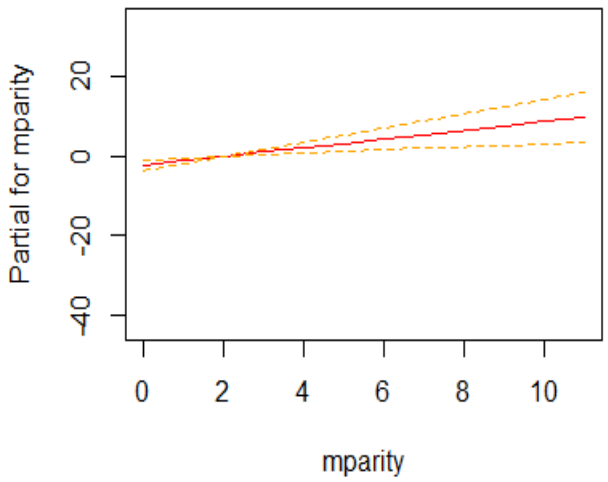
Table 2. Durbin-Watson Test

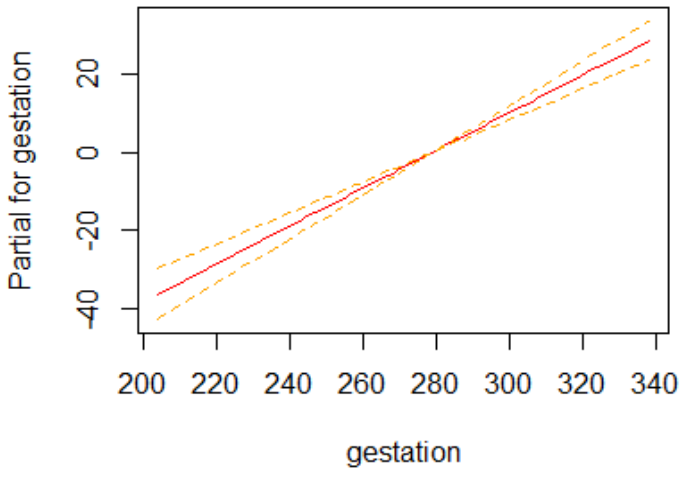
La	Autocorrelation	D-W Statistics	P-Value
1	0.084144863	1.833952	0.036

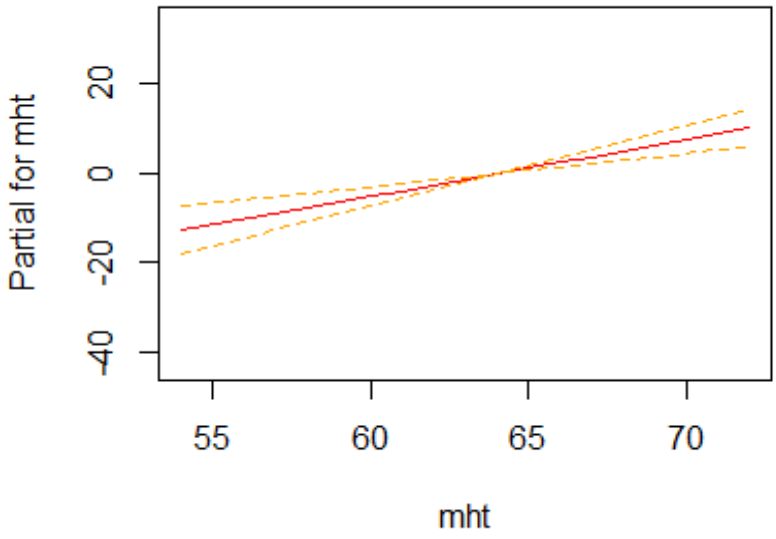
Table 3. Shapiro-Wilk Normality Test

Shapiro-Wilk Normality Test		
Data: Final Model	W = 0.99696	P-value = 0.3253

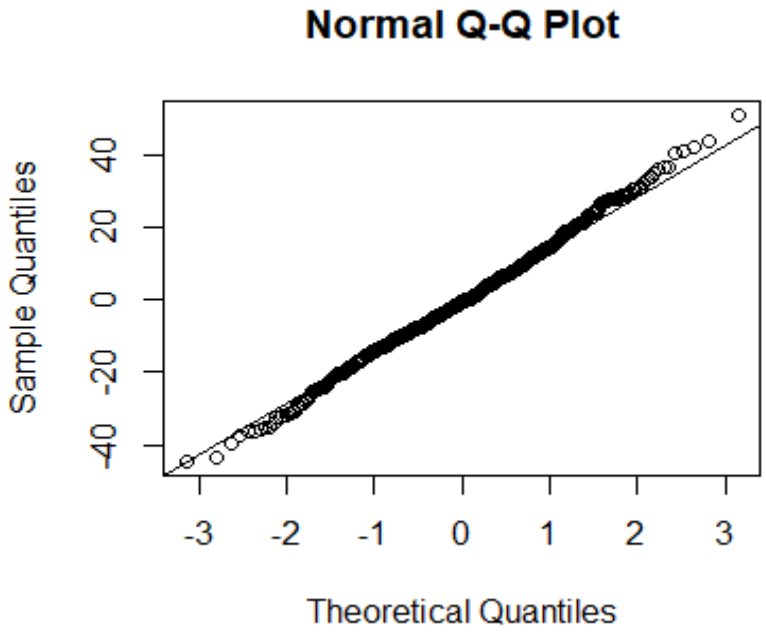
Partial Regression Plots for the 'best' model:



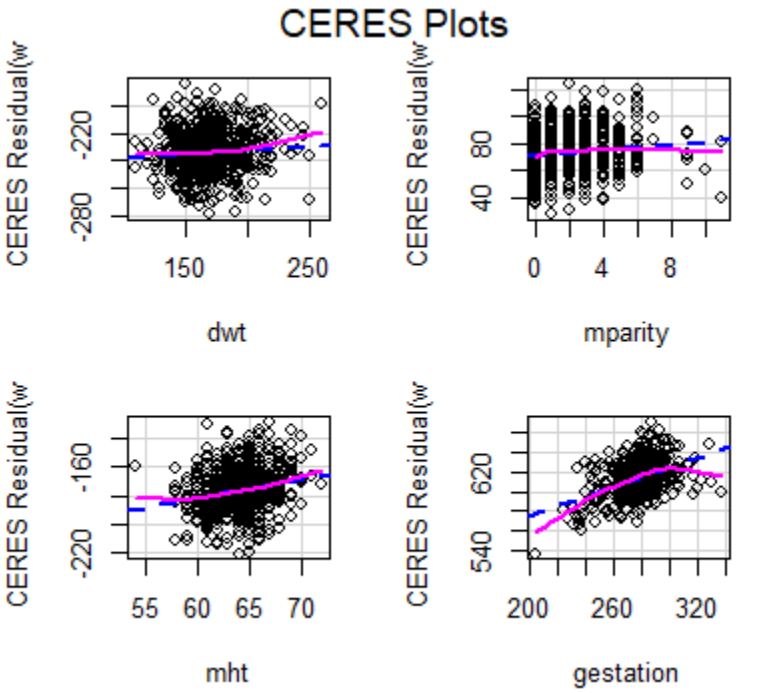




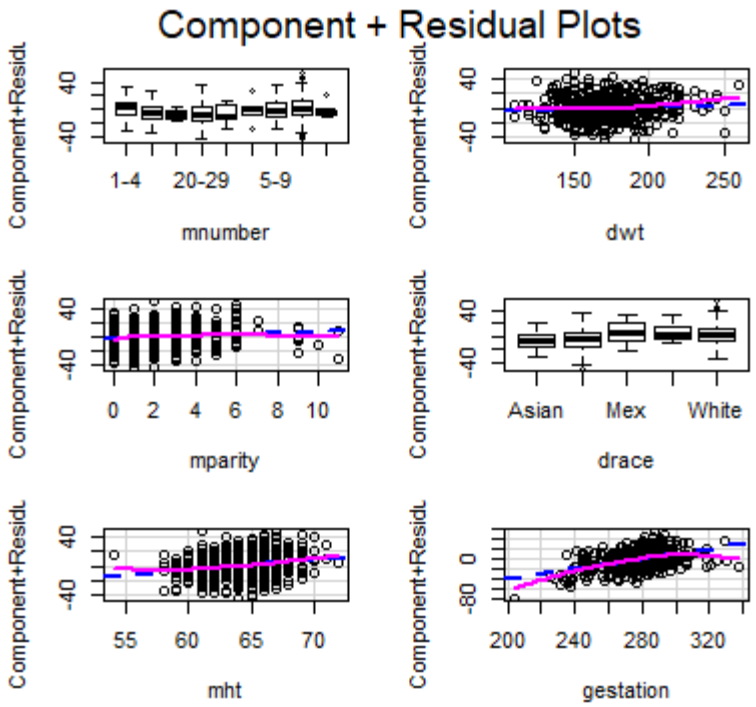
Q-Q Plot for Durban-Watson Test:



CERES Plots for the best model:



Component + Residual Plots of the 'best' model:



Appendix III - Summary of parameter estimates

The estimates of all coefficients can be interpreted as:

- A 1 ounce increase in weight of an infant tends to be associated with an increase of 0.48927 in length of gestation, ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with an increase of 0.96521 in total number of pregnancies (mparity), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be associated with an increase of 1.32096 of mother's height in inches (mht), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be associated with an increase of 0.05883 pounds in father's weight (dwt), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with an increase of 1.63659 units of father's race from Black background compared with the baseline of Asian background (draceBlack), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with an increase of 12.71662 units in father's race from Mexican background compared with the baseline of Asian background (draceMex), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with an increase of 11.63911 units in father's race from Mixed background compared with the baseline of Asian background (draceMixed), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with an increase of 8.31428 units in father's race from White background compared with the baseline of Asian background (dracewhite), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with a decrease of 9.05501 units for 10-14 number of cigarettes smoked per day for past and current smokers compared with the baseline of 1-4 cigarettes smoked per day (mnumber10-14), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with a decrease of 12.17145 units for 15-19 number of cigarettes smoked per day for past and current smokers compared with the baseline of 1-4 cigarettes smoked per day (mnumber15-19), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with a decrease of 11.06015 units for 20-29 number of cigarettes smoked per day for past and current

smokers compared with the baseline of 1-4 cigarettes smoked per day (mnumber20-29), ceteris paribus.

- A 1 ounce increase in weight of an infant tends to be correlated with a decrease of 10.53040 units for 30-39 number of cigarettes smoked per day for past and current smokers compared with the baseline of 1-4 cigarettes smoked per day (mnumber30-39), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with a decrease of 5.10201 units for 40-60 number of cigarettes smoked per day for past and current smokers compared with the baseline of 1-4 cigarettes smoked per day (mnumber40-60), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with a decrease of 4.90493 units for 5-9 number of cigarettes smoked per day for past and current smokers compared with the baseline of 1-4 cigarettes smoked per day (mnumber5-9), ceteris paribus.
- A 1 ounce increase in weight of an infant tends to be correlated with a decrease of 4.78863 units for number of cigarettes smoked per day but not now compared with the baseline of 1-4 cigarettes smoked per day (mnumber smoke but don’t now), ceteris paribus.

Appendix IV – Effect plots of some covariates of the 'best' model