Investigation Into Predictors Of Birthweight Among Human Babies

Via Linear Regression Analysis

Submitted to:

Dr. Carl Donovan
School of Mathematics and Statistics
University of St. Andrews

In Partial Fulfillment of the Course:

MT5762

Report Prepared by:

Team Abundance of Onslaught
ID# 150004238
ID# 160015281
ID# 180007719
ID# 190001336
ID# 190025615
ID# 190027153

5 November 2019

Executive Summary

Previous research suggests that the birthweight of human babies plays an important role in the physical and intellectual development (Shiono & Behrman, 1995). Our study goes one step further by developing a reliable predictive model to explain the relationship between the birthweight of babies and a selection of socio-economic and biological characteristics of the parents. Data are appropriately cleaned, and exploratory analysis is performed. An initial model is then fit, after which model assumptions of linearity, normality, independence and homoscedasticity are assessed and verified. The model is then refined, with transformations and interaction terms considered, until final model specifications become apparent. The result is a robust predictive model of birthweight in human babies.

Introduction

The predictive ability of socio-economic and biological factors on baby weight has been subject to regular, and sometimes controversial, studies (Mondal, 2000; Som et al., 2004). Previous studies suggest that the parents' physical properties (i.e., age, body weight, etc.) might have a measurable effect on the birthweight of their babies (Mondal, 2000). Furthermore, it has been proven that factors such as alcohol consumption or smoking by the mother during the pregnancy significantly affect birthweight (Som et al., 2004). In the past, numerous complications in the child's development have been associated with low birth-weight (Shiono & Behrman, 1995). As a result, understanding and identifying potential drivers of birthweight contributes to this field of study.

For instance, although there is a widely-recognized consensus that maternal smoking is highly correlated with low birthweight, controversies arise when trying to properly define the exact effect of maternal smoking on a developing fetus (Parascandola, 2014; Yerushalmy, 1964). A general critique is that scientists conducted studies while having the desirable outcomes in mind, thereby letting self-selection bias affect their results (Parascandola, 2014). As a result, although many different factors have already been taken into consideration in former analyzes, the question remains: *What relationships are there between the birthweight of babies and other measurable variables?*

The purpose of this study is to offer a model to answer this question and to contribute new findings to the ongoing debate. By better understanding which variables are the best predictors of baby weight, we may find information helpful to health officials that will allow for improved quality of care among both mothers and newborn babies. The study analyzes data that has been extracted from the Child Health and Development Studies (CHDS) (Yerushalmy, 1964).

Methods

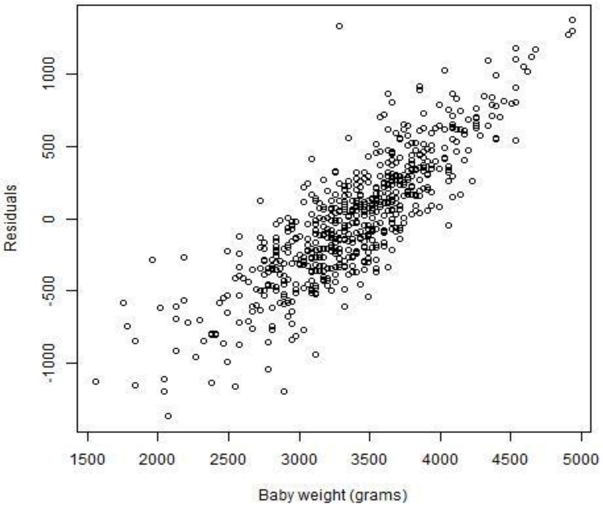
The procedure we utilized to create our model relied on data analysis performed using R statistical software, as well as the open-source GUI, RStudio. Within R, we also installed and used the "ggplot2" package to produce all figures found both in this report and on our Github repository, https://github.com/jwalsky24/AbundanceOfOnsalught. After the data cleaning, initial insights into the data were gained by generating basic summaries and plots. Next, we used bootstrapping as well as the Akaike information criterion (AIC) (Akaike, 1974) to identify the most suitable model for the purpose of this study. The final model selection qualified based not only on statistical measures but logical grounds, as its underlying assumptions remained unviolated and, simply put, the predictors included in the model made sense for a predictive model.

Our team received "babies23.data" featuring 23 fields of information for each of 1,236 newborn babies. This includes the weight of each baby at birth, our response variable, as well as several other variables including information about the child's parents, which will serve as our independent (predictor) variables. After reading the file into R statistical software using the read.table() function, the first task was to convert a selection of responses in the dataset to NAs, so that they could be omitted from our analysis. These were primarily responses where data were not collected, either because the parents were not asked or declined to provide an answer. Second, the height and weight variables were transformed from imperial units to standard metric units. Next, a few variables contained factor levels that we chose to collapse into one, and therefore treat as identical, for the purposes of analysis. Lastly, a single observation that we identified as a problematic outlier because it featured unrealistic values was removed from the dataset.

Checking Model Assumptions

In order to create a predictive model of baby weight that features the most explanatory power possible, we first examine as to whether a linear model is indeed the most appropriate model. To accomplish this, we first fit a model of baby weight using every available predictor variable. We then check that our initial model meets the following assumptions: Linearity (the relationship must be linear), independence among explanatory variables, normality (errors normally distributed), and homoscedasticity (constant variance).

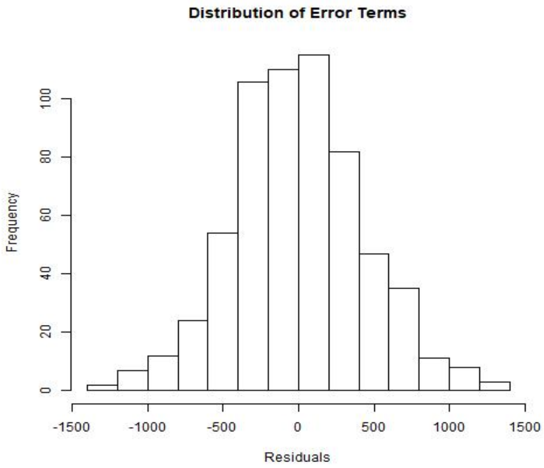
First, we must address a critical question: is the relationship between our response variable and our set of independent variables, in fact, linear? By plotting baby weight on the x axis against the residuals from our model on the y axis below, we see that values appear to increase in a linear pattern from the bottom-left to the top-right. If little or no linearity were present, the values would appear to be much more randomly scattered.



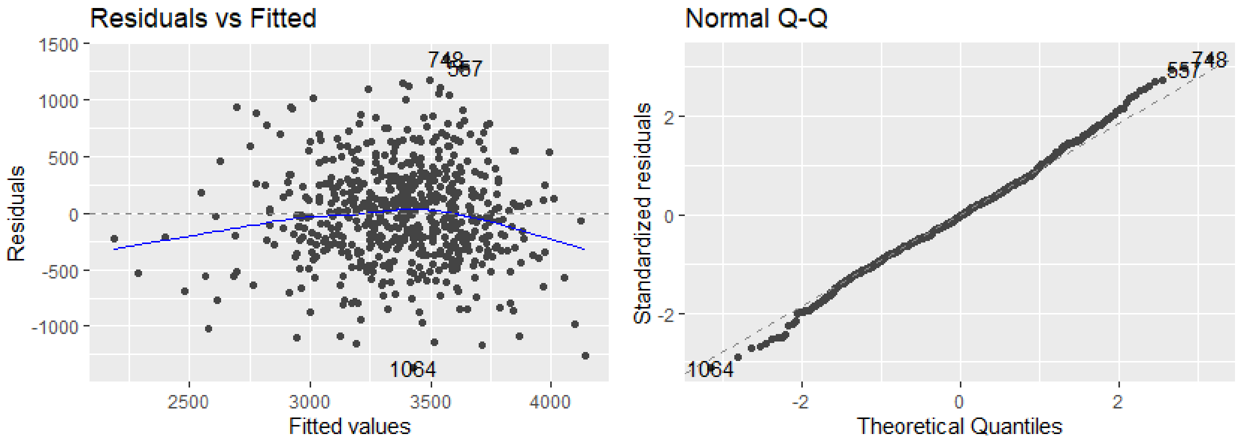
Next, we examine our independent variables for possible conflicting relationships. If some of the variables included in our model are highly correlated, and these relationships are not accounted for, an issue called “collinearity” arises where the variance of one or more of our regression coefficients becomes artificially inflated. In order to prevent this, we first inspect a correlation matrix showing the correlation values (between -1 and 1) of every pair of independent variables to be considered for our final model. We then take note of the seven values either above 0.5 and below -0.5, as these values indicate high positive and negative correlation, respectively.

These highly collinear independent variable pairs are the most likely to cause high variance inflation as measured by the Variance Inflation Factor (VIF). Before finalizing our model, we will want our VIF to be optimized to the greatest extent possible.

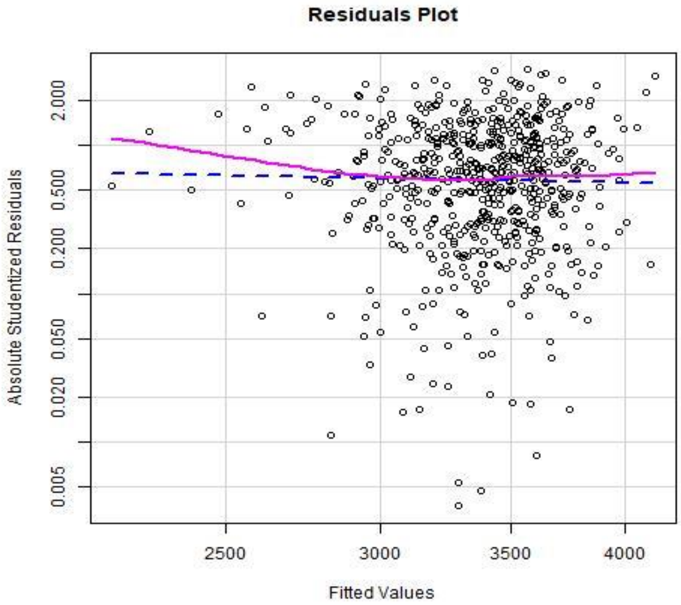
Normality is another assumption we must review before proceeding. This means that the error terms of our model, otherwise known as residuals or noise, must follow an approximately normal distribution. Below you will find a histogram showing the distribution of residuals in our model. The shape resembles that of a normal distribution rather closely.



To confirm this normality assessment, we use the autoplot() function from the ‘ggfortify’ package to produce the plots shown below. The graph on the left shows a scatterplot with fitted predictor values represented on the x axis and residuals represented on the y axis. The graph on the right is a plot of quantiles. The approximately random scatter of values in the “Residuals vs Fitted” plot, paired with the close adherence of values to the diagonal line in the Q-Q plot, confirm that our error terms are normally distributed for the purposes of this study.



Our last model assumption to be checked is that the error terms of our model have approximately constant variance across the range of fitted values. This can be performed by plotting our fitted response values on the x axis against our error terms on the y axis. The resulting plot for our model can be seen here:



While variance is slightly higher for some small outlying fitted values, the purple line is mostly flat where most of the data are shown (x > ~2800), validating our assumption that variance is approximately constant as baby weights increase.

Results

With preliminary model assumptions now validated, we start to refine our model. We first inspect a summary of our “full” model, i.e. a model that considers every predictor variable in our dataset. We then construct “slimmer” model containing only those predictor variables whose contribution to the “full” model was statistically significant – in this case, gestation period, mother’s height, father’s weight and number of cigarettes smoked. While this model turns out to be a decent fit to our data, three potential improvements are then considered: square-root transformation, removal of a critical outlier, and inclusion of interaction terms.

First, a model is constructed that fits our “slimmer” model not to baby weight, but to the *square-root of* baby weight. Unfortunately, while this does produce an improved AIC value, using this square-root transformation makes the covariates less significant or not significant. As a result, we chose not to use this or any transformation in our final model.

Next, we look closer at extreme values in our dataset, with unusually short/long gestation periods and unusually low/high birthweights being the focus. Two observations, one in row 261 and one in row 979, stand out as observations whose values may have been recorded wrong. Upon closer inspection, while values appearing in row 979 are realistic in the context of our data, row 261 shows a gestation period of 148 days. As of 2018, the shortest gestation period ever for a surviving human baby was 22 weeks, or 154 days (Associated Press, 2018) so this value cannot be correct for a surviving baby. We therefore chose to omit this observation from the dataset used to build our final model.

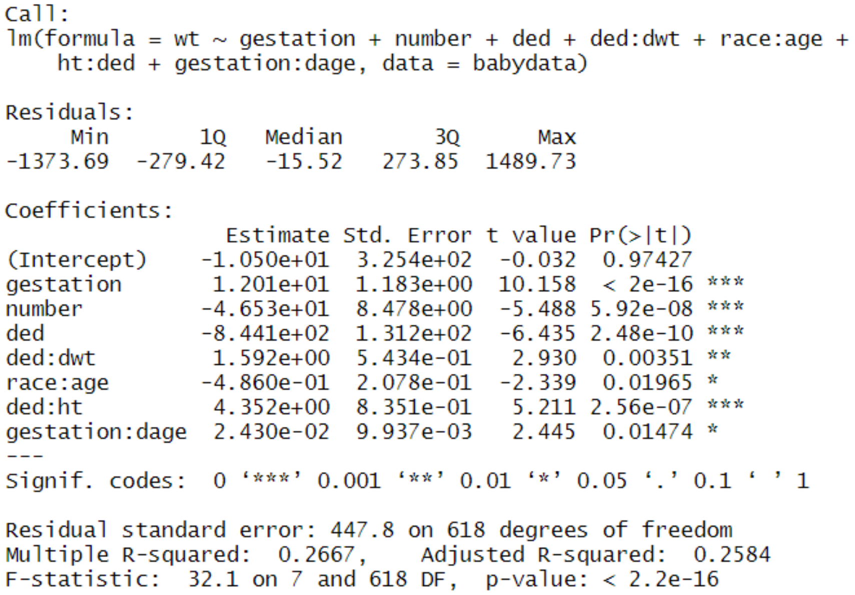
Finally, we consider whether to include interaction terms in our model. For any given pair of independent variables, we may include in our model, we want to consider the possibility that neither predictor on its own has a significant effect on birthweight, but that the combination of two may. To investigate this, we construct a model containing all possible interaction terms, then eliminate any statistically insignificant terms. We then add any interaction terms that remain to our “slimmer” model from above, then fit the model again, once again eliminating any statistically insignificant terms.

Before declaring this model to be our final model, we will need to confirm once again that our model meets the assumptions of a linear model, this time using statistical tests. Using a

Shapiro-Wilk test, a Durbin-Watson test and a Non-Constant Variance test, we fail to reject the null hypotheses that the data underlying our model are normal (p-value = 0.068), independent (p-value = 0.09), and have residuals with constant variance (p-value = 0.06), respectively. Having effectively "passed" these assumptions tests, we now declare this model to be our final model.

Final Model

Below is a summary of our final predictive model of baby weight. The model features seven coefficients, of which four are interaction terms, with 625 – 7 = 618 degrees of freedom. The result is a multiple R squared value of 0.2667, which means that our final model can explain 26.67% of the variability in birthweights in our dataset.



```
Call:
lm(formula = wt ~ gestation + number + ded + ded:dwt + race:age +
    ht:ded + gestation:dage, data = babydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-1373.69  -279.42   -15.52   273.85  1489.73 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -1.050e+01  3.254e+02  -0.032  0.97427    
gestation       1.201e+01  1.183e+00  10.158  < 2e-16 ***
number         -4.653e+01  8.478e+00  -5.488 5.92e-08 ***
ded            -8.441e+02  1.312e+02  -6.435 2.48e-10 ***
ded:dwt         1.592e+00  5.434e-01   2.930  0.00351 ** 
race:age       -4.860e-01  2.078e-01  -2.339  0.01965 *  
ded:ht          4.352e+00  8.351e-01   5.211 2.56e-07 ***
gestation:dage  2.430e-02  9.937e-03   2.445  0.01474 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 447.8 on 618 degrees of freedom
Multiple R-squared:  0.2667,	Adjusted R-squared:  0.2584 
F-statistic:  32.1 on 7 and 618 DF,  p-value: < 2.2e-16
```

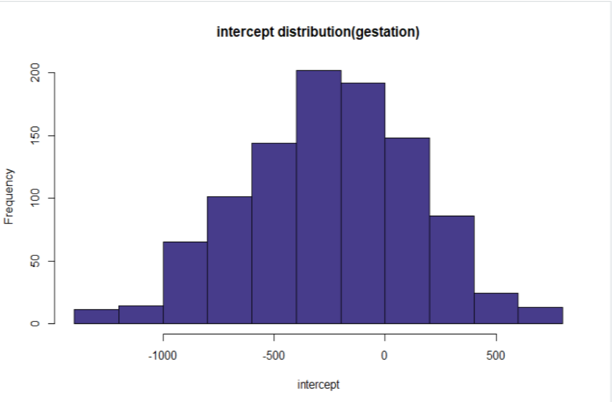
A description of the terms in our final model can be found here:

- gestation: time from conception to birth (in days)
- number: for current and past smokers, number of cigarettes smoked by mother per day (“0” if mother does not smoke)
- ded: father’s education level
- ded:dwt: interaction term between father’s education level and father’s weight (in pounds)
- race:age: interaction term between mother’s race and mother’s age (in years)
- ded:ht: interaction term between father’s education level and mother’s height
- gestation:dage: interaction term between gestation period (in days) and father’s age (in years)

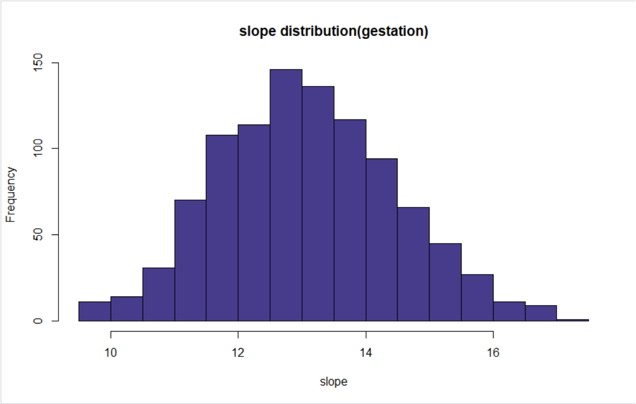
Bootstrapping

Bootstrapping is an algorithm based on random sampling with replacement that is used to estimate standard error, confidence intervals and deviation of a model (Efron & Tibshirani, 1993). In our report, we use non-parametric bootstrapping, which, instead of parametric bootstrapping, makes no assumptions of how the observations are distributed. We have resampled our data a thousand times and then performed interference on these resampled data.

First, we examined the model where the birthweight depends on the gestation period. The distribution of the intercept of this model is shown in the following histogram:

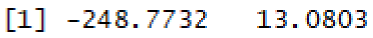


Based on the following plot of the slope of this model we find a positive relationship between birthweight and gestation.



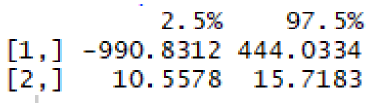
To support this result, we calculate the mean of the intercept and slope, and also output the confidence intervals.

Means of intercept and slope:



```
[1] -248.7732   13.0803
```

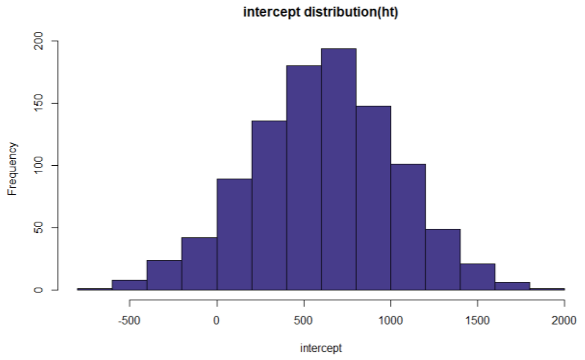
Confidence intervals of intercept and slope:



```
           2.5%     97.5%
[1,] -990.8312 444.0334
[2,]   10.5578  15.7183
```

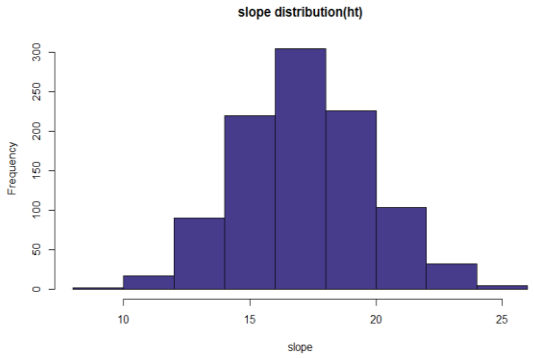
As shown, the mean of intercept is -248.77 and its confidence interval of 95% is given from -990.83 to 444.03. The mean of the slope is 13.08 and the respective confidence interval is given from 10.56 to 15.72.

Now we look at how the birthweight depends on the height of the mother. The following figure shows how the intercept is distributed in this model:



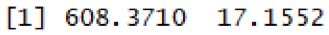
We conclude that in this case the intercept fluctuates more strongly than in the model of birthweight depending on the gestation period.

The histogram of slope distribution is shown here:



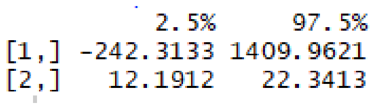
Again, we find a positive relationship supported by calculated means and confidence intervals.

Means of intercept and slope:



```
[1] 608.3710  17.1552
```

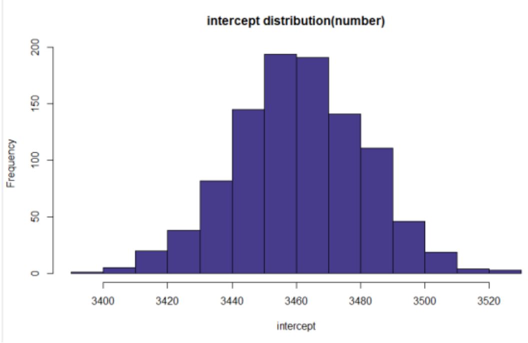
Confidence intervals of intercept and slope:

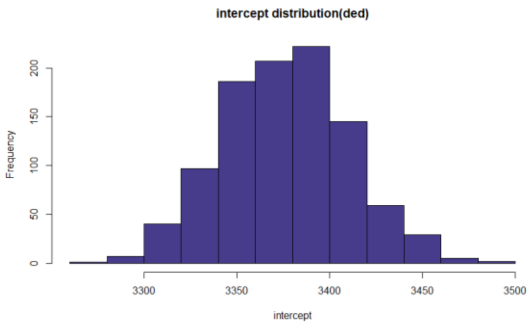


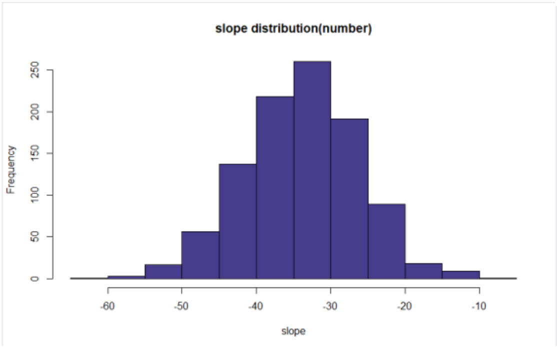
```
          2.5%      97.5%
[1,] -242.3133 1409.9621
[2,]   12.1912   22.3413
```

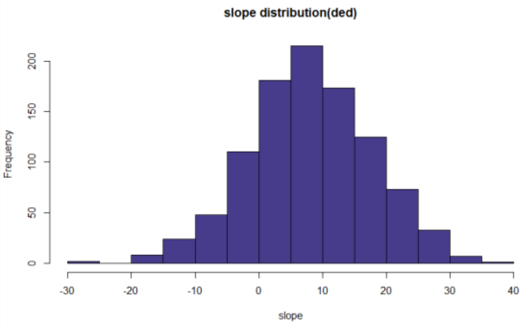
In this case the mean of the intercept is 608.37 and its confidence interval is given from -242.31 to -1409.96. The mean of the slope is 17.16 with a confidence interval from 12.19 to 22.34.

The same as above is done for the models where birthweight only depends on the number of cigarettes smoked by the mother or on the educational level of the father. Hence, we plot the intercept and the slope of the models of wt (the birthweight) related to number (number of cigarettes) and of wt (the birthweight) related to ded (educational level of the father).



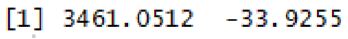






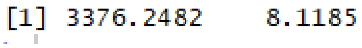
In both cases the intercept is a great positive number. We see that the slope in the model wt ~ number is negative, therefore we have a negative relationship in this case, meaning that the birthweight decreases as the number of cigarettes per day increases. On the other hand, we see that the slope of wt ~ ded varies from -30 to 40, hence we cannot conclude a significant relationship between the weight of newborns and the academic level of the father. For further support we calculate the means and confidence intervals in all cases.

Means of intercept (number) and slope (number):



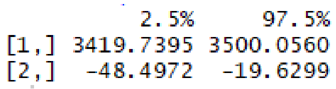
```
[1] 3461.0512  -33.9255
```

Means of intercept (ded) and slope (ded):

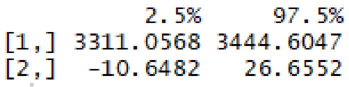


```
[1] 3376.2482    8.1185
```

Confidence intervals of intercept (number), slope (number), intercept (ded) and slope (ded):



```
           2.5%     97.5%
[1,] 3419.7395 3500.0560
[2,]  -48.4972  -19.6299
```



```
           2.5%     97.5%
[1,] 3311.0568 3444.6047
[2,]  -10.6482   26.6552
```

The mean of the intercept in wt ~ number is 3461.0512 and the mean of the intercept in wt ~ ded is 3376.25. The confidence intervals of 95% are given by [3419.74, 3500.06] and [3311.06, 3444.60].

The mean of the slope in wt ~ number is -33.93, while the mean of the slope in wt ~ ded is 8.12. The confidence intervals are given by [-48.50, -19.63] and [-10.65, 26.66], respectively.

Now we look at the interaction models. The first interaction model is given by wt ~ ded:dwt. Hence, we examine the coherent effect of the education and the weight of the father on the birthweight of the child. The plots below show the intercept and the slope of this model:

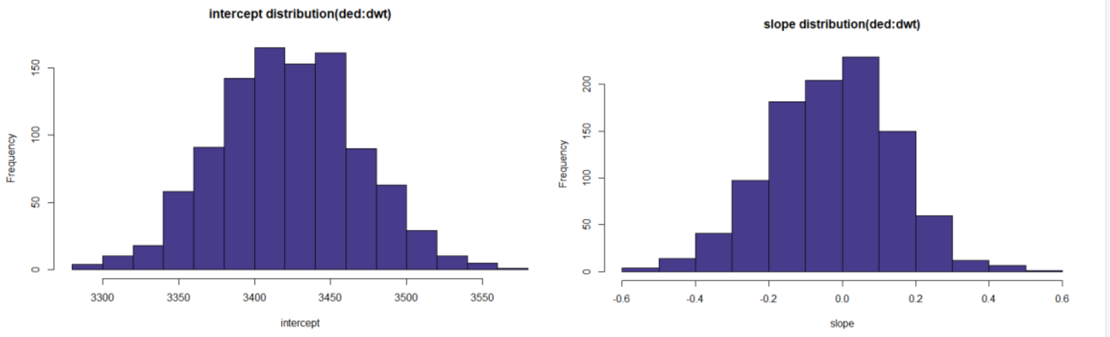
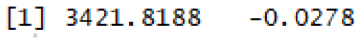


Figure 1.17

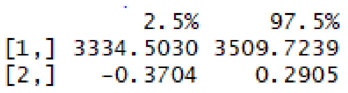
Both the intercept and the slope are not as variable as in the models before. Again, we have a great intercept and a slope that does not imply a definite relationship between wt and ded:dwt. As seen below, we get means of 3421.82 for the intercept and 0.0278 for the slope.

Means of intercept and slope:



```
[1] 3421.8188   -0.0278
```

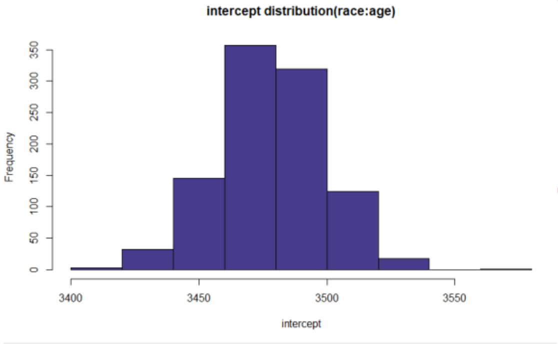
Confidence intervals of intercept and slope:

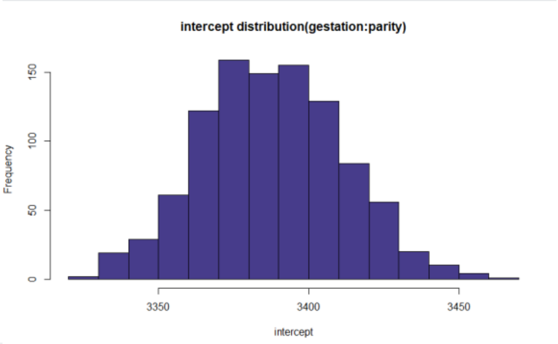


```
          2.5%     97.5%
[1,] 3334.5030 3509.7239
[2,]   -0.3704    0.2905
```

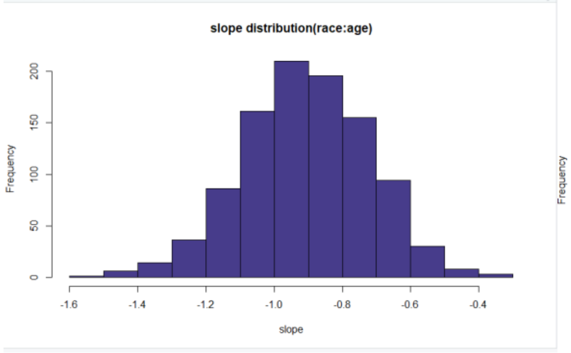
The confidence intervals of 95% are given by [3334.51, 3509.72] and [-0.37, 0.29].

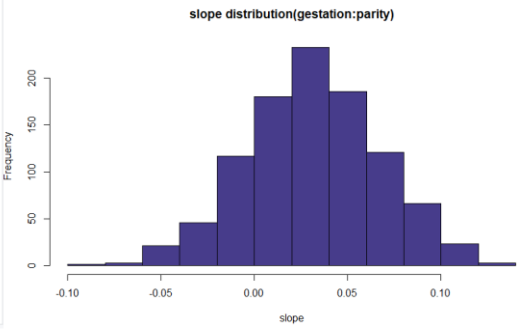
We do the same for two other interaction models. The first one reflects the interaction effect of the mother's race and age on the birthweight and the second one shows how the birthweight depends on the gestation period and the total number of previous pregnancies including fetal deaths and still births (here given by the parameter parity).



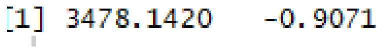


From the above diagrams we observe high values and intercept in both models and from the below plots we conclude a negative relationship of wt and race:age and a indefinite relationship between wt and gestation:parity.



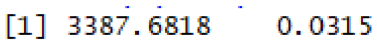


Means of intercept (race:age) and slope (race:age):



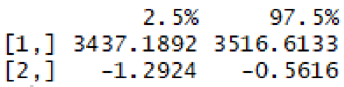
```
[1] 3478.1420   -0.9071
```

Means of intercept (gestation:parity) and slope (gestation:parity):

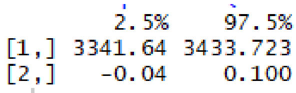


```
[1] 3387.6818    0.0315
```

Confidence intervals of intercept (race:age), slope (race:age), intercept (gestation:parity) and slope (gestation:parity):



```
         2.5%     97.5%
[1,] 3437.1892 3516.6133
[2,]   -1.2924   -0.5616
```



```
        2.5%    97.5%
[1,] 3341.64 3433.723
[2,]   -0.04    0.100
```

For wt ~ race:age the mean of the intercept is 3478.14 and the mean of the slope is -0.9071. The confidence intervals are [3437.19, 3516.61] and [-1.29, -0.56]. In the model wt ~ gestation:parity the mean of the intercept is 3387.68 and the mean of the slope is 0.032. The confidence intervals are from 3341.64 to 3.44 and from -0.04 to 0.10, respectively.

Error Calculation from MSE and Cross-Validation Sets

We randomly split all 1236 observations of our initial dataset into two separate sets: First, a training set of 866 observations (about 70% of all observations), which was used to specify our two final proposed models. Second, a cross-validation test set of 370 observations (about 30% of all observations), which we will later use to cross-validate our proposed models against novel data to estimate the out-of-sample error. Our previous analysis and model selection yielded two final models. We include one final model which contains outliers and one final alternative model which does not contain any outliers. In the analysis following, we want to test these models for statistical significance and the order of magnitude of their error.

First, we employ the mean-squared error as a first measure of error of our two final models. The mean squared error (MSE) is a measure of discrepancy between the estimated and the true underlying value, thus, a measure of an estimator's quality (Lehmann & Casella, 1998). It is quasi the variance of a linear regression model. For the alternative model including outliers, we estimate an MSE of 188860. For the final model excluding outliers, we estimate an MSE of 188318.6. We find that the model excluding outliers has a slightly lower MSE. Therefore, this model has a slightly lower error.

Second, we want to test the out-of-sample error of our two final models. This is the error of our model predictions when tested with new data, which have not been used for model specification. We employ the cross-validation data set of 370 observations from earlier and calculate the out-of-sample error for both final models. For our first model including outliers, we calculate an out-of-sample error of 2033.08. For our second model excluding outliers, we calculate an out-of-sample error of 1996.87. Again, we find that the second model produces a slightly lower out-of-sample error. We can therefore conclude that we have successfully tested the errors of our models using MSE and the out-of-sample error on a cross-validation set.

Study Limitations and Recommendations

While our final model has impressive predictive power when it comes to how much a baby will weigh at birth, we unfortunately cannot say that any of the predictors featured in the model *cause* a baby to be born at a given weight. This is particularly relevant because these data we were given to analyze do not come from an experiment. Rather, they come from an observational study, where researchers collected data on newborn babies and their parents but did not introduce an intervention that might cause birthweight to fluctuate (when all other variables held equal). If we did seek to prove causality however, we would have to conduct one experiment for each of the explanatory variables. Each influencing parameter would be assigned to a group as a treatment while all other factors would be held equal, e.g. the mothers would have to smoke 10 cigarettes a day during their pregnancies which would equal one specific combination of the parameters smoke, time and number. In addition, we would also have to have a control group (no treatment = usual smoking habit) and a placebo group (placebo treatment = cigarette without nicotine). While studies like this are not only completely untenable on ethical ground, they are also extremely costly and highly elaborate. So, even though we cannot establish causality, an observational study is the only opportunity in this case.

While the final model developed by our team is a reasonably good fit to our data, we were hindered by the fact that our sample size was limited, leading to a high margin of error. The original "babies23.data" file contained n=1236 observations, but after excluding NA values, the dataset we fit our model to contained only n=625. With only a few hundred observations, the likelihood increases that the sample mean of each of our variables will misrepresent the population means they are meant to estimate. This decreases our ability to identify outliers that skew the data. A similar study conducted with a larger sample size would benefit from more precise mean and variance estimates. The subsequent model of birthweight would be a closer fit to the data, and outliers would be easier to identify. As a result, further studies in this area of research ought to emphasize collection a high sample size.

There is also the possibility that by excluding observations that contained one or more NA values, we inadvertently introduced bias into the model. For example, upon closer inspection, both the "father's height" and "father's weight" variables each featured just under five hundred NA values, and all of these NAs were excluded from the dataset used to build the final model. It is therefore possible that our dataset has a sampling bias against babies born to single mothers, as this would be a possible explanation for the father's information not being collected. Furthermore, over one hundred of the original n=1236 observations were excluded because income was not reported. We assume that an individual with a lower income is less likely to want to report their income to study administrators than individuals with a higher income. If this assumption is correct, the data used to construct our model may have been biased towards those with a higher income.

Lastly, while we did have nineteen variables eligible for inclusion into our model, our predictive model would likely perform better if a wider array of information was available to us. For example, information about any health problems suffered by either of the parents might have produced a model with less error. In addition, while the smoking habits of the babies' parents are captured in our dataset, information on their level of consumption of alcohol and other drugs might be informative here.

Summary and Conclusion

The intention behind our analysis was to provide novel insight to the ongoing debate on the effects of variables on birthweight of babies, that were previously unknown. We proposed multiple possible models including different potential variables, that might influence birthweight. We then used a bootstrapping approach to select the best model and calculated the error on two measures (MSE and out-of-sample error). Our final model employs seven final variables. Following our discussion, this model can be used in a limited way to explain deviations in the birthweight of babies. Although the interpretability of our model is limited, it contributes to the ongoing debate and provides ground for further investigations.

Bibliography

- ➢ Aguirre-Urreta M I, Rönkkö M (2018) Statistical inference with PLSc using bootstrap confidence intervals[J]. MIS Quarterly, 42(3): 1001-1020.
- ➢ Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723. doi:10.1109/TAC.1974.1100705
- ➢ Associated Press (2018). Earliest surviving preemie to remain in hospital. Retrieved November 5, 2019, from http://www.nbcnews.com/id/17237979/ns/health-childrens_health/t/earliest-surviving-preemie-remain-hospital.
- ➢ Efron & Tibshirani（1993) An introduction to the bootstrap Chapman & Hall.
- ➢ Lehmann, E. L., & Casella, G. (1998). Theory of Point Estimation (2nd ed.). New York: Springer.
- ➢ Mondal, B. (2000). Risk factors for low birth weight in Nepali infants. The Indian Journal of Pediatrics, 67(7), 477–482.
- ➢ 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 (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- ➢ Shiono, P. H., & Behrman, R. E. (1995). Low birth weight: Analysis and recommendations. The Future of Children, 5(1), 4–18. doi: 10.2307/1602504
- ➢ Som, S., Pal, M., Adak, D. K., Gharami, A. K., Bharati, S., & Bharati, P. (2004). Effect of socio-economic and biological variables on birth weight in Madhya Pradesh, India. Mal J Nutr, 10(2), 159–171.
- ➢ Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

- Wickham, François, Henry and Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr
- Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
- Wild & Seber (2000). Chance Encounters-A First Course in Data Analysis and Inference. Kingsport, Wiley and sons Inc.
- Yerushalmy, J. (1964). Mother's cigarette smoking and survival of infant. American Journal of Obstetrics and Gynecology, 88(4), 505–518. doi: https://doi.org/10.1016/0002-9378(64)90509-5