T15-03 Executive summary

Christopher Saad, Derek Ng, Stuart Morrison, Yifan Jin

This version was compiled on November 4, 2019

Abstract. This report undertakes analysis on the effect of prenatal risk factors on the outturn birth weight of an infant. We apply data measured in a not-for-profit hospital in Massachusetts in 1986 that recorded the birth weight of 189 infants, as well as if the mother exhibited certain risk factors, including whether she smoked or had had a premature labour before (Hosmer, David W., et al., 2013). We apply multiple linear regression models, fitted using AIC and CV based goodness-of-fit criteria, to determine the unilateral effect of each of the risk factors. We find that each of the risk factors we examined had a statistically significant effect of reducing the birth weight of the infants in our sample. In our discussion we highlight the limitations of our analysis, ie, the potential bias in our sample caused by the non-profit nature of the hospital at which the data was recorded

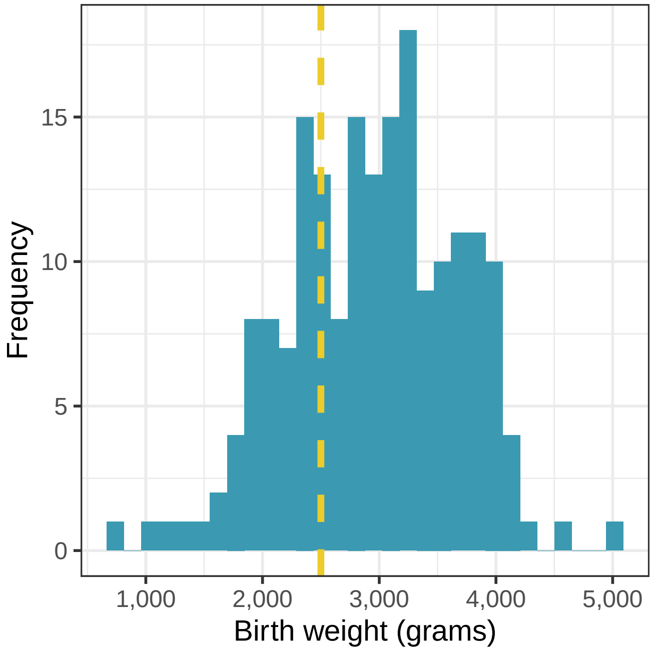
Introduction. This report undertakes analysis on the effect of prenatal risk factors on the birth weight of an infant. Low birth weight is associated with higher infant morality rates and birth defect rates and so, it is import to understand the drivers of low birth weight and the effect of various risk factors.

In this report, we consider the effect of several risk factors, ie, the race of the mother, whether the mother smokes, whether the mother suffers from hypertension and whether the mother has had uterine irritability before. We apply a multiple linear regression framework to understand how the risk over several of this factors at once may coalesce into the low birth weight of a child.

Data set. The data set we use in our analysis was originally collected by physicians in 1986 to examine the effects of risk factors on the birth weight of an infant (Hosmer, David W., et al., 2013). It is included in the `MASS` package in `R`.

There are 10 variables included in the data and the key dependent variable for this analysis is `birthwt` - a variable measuring the infant's weight at birth. The other variables in the data show the risk factors we examine, ie: + the racial background of the mother + whether the mother is a smoker + whether the mother has had a premature labour before + whether the mother experiences hypertension + whether the mother experiences physical irritability

Examining the distribution of infant birth weights in our data shows that 31.2% of the infants in our data had a clinically low birth weight, ie, below 2,500 grams.



Multiple cross sections of demographics are represented in the data. The figure in appendix 1 shows the distribution of a selection of key variables in the data set.

Analysis. Firstly, we used a backwards selection AIC regression model, AIC is an indicator model quality, the less AIC the better the model. For backward selection, we firstly run a full regression model which contains all the explanatory variables in our dataset. Then, we try to drop variables to lower the AIC, by running through this algorithm, our AIC regression contains four explanatory variables: race, mother_weight, history of hypertension, smoking and uterine irritability.

Secondly, to confirm our model is appropriate, we use backward selection with p-value. Again, we firstly run a full regression model. Then we try to drop the variable in this regression model with the largest p-value. Next, we run a new regression without this variable and try to drop another variable with largest p-value in this new regression model. By running this process again and again, we drop all statistically insignificant variables i.e. the p-value is greater than 0.05. Surprisingly, the second method produces the same result as the first result. This supports our model selection process.

Assumptions. Before we proceed to analyse our models results we must first ensure that all our assumptions are satisfied. Our multiple linear regression model requires that all residuals ε_i are $iid\ N(0, \sigma^2)$ and that there is a linear relationship between y and x. This can be simplified to the following 4 assumptions: 1. Linearity - the relationship between Y and x is linear 2. Independence - all the errors are independent of each other 3. Homoskedasticity - the errors have constant variance 4. Normality - the errors follow a normal distribution

By looking at appendix 2 we can determine if our assumptions are met. Since there's no obvious pattern in the residual vs fitted values plot (e.g. no smiley face of frowny face) other than the approximately horizontal line it doesn't appear that we have misspecified the model as being linear. the residuals don't appear to be fanning out or changing their variability over the range of the fitted values so the constant error variance assumption is met, and thus the Homoskedasticity assumption is met. Finally in the QQ plot, the points are reasonably close to the diagonal line. The top 8 or so points are not quite on the line, but it's not a severe enough departure to cause too much concern. Thus the normality assumption is at least approximately satisfied.

Therefore given all our assumptions are met, our multiple linear regression model can be reliably analysed.

Results. Our final model is:

$$\widehat{Birthweight} \sim mother.weight+race+smoke+hyptertension+ute$$

Our full model has 8 variables and an unadjusted R-squared of 0.2467. In our new simplified model, we dropped 3 variables and this obtained an R-squared of 0.2404 which is only slightly smaller than the full model despite having less variables.

The in-sample performance of the final model provided an R-squared that shows that approximately 24% of the total variability in birth_weight is explained by the explanatory variables. The most statistically significant regressors at the 1% level were found to be race, smoke, hypertension and uterine irritability.

We used 10-fold cross validation to measure out of sample performance. From the output we can see that for the full model has a RMSE of 676.93 and MAE of 542.10. The simplified model has a RMSE of 658.87 and MAE of 534.58. Thus we can see that the simplified model outperforms the full model.

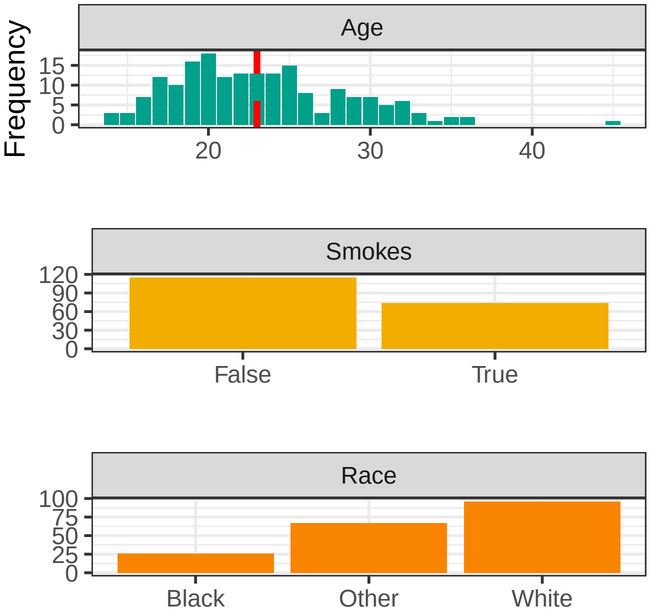
Now to interpret the constant and coefficients: (See Appendix 3)

Making all the explanatory variables equal to zero, on average the predicted birth weight is 2837.26 grams. Holding all other variables constant, a 1 pound increase in the mother's weight, on average would have a predicted increase in birth_weight by 4.24 grams. *Holding all other variables constant, if the mother's race is black, on average the birth weight is predicted to be lower 475.06 grams than if the mother was white.* Holding all other variables constant, if the mother's race is other (not black or white), on average the birth weight is predicted to be lower by 348.15 grams than if the mother was white. *Holding all other variables constant, if the mother smoked during pregnancy, on average the birth weight is predicted to decrease by 356.32 grams.* Holding all other variables constant, if the mother has a history of hypertension, on average the birth weight is predicted to decrease by 585.19 grams. *Holding all other variables constant, if the mother has uterine irritations, on average the birth weight is predicted to decrease by 525.52grams.

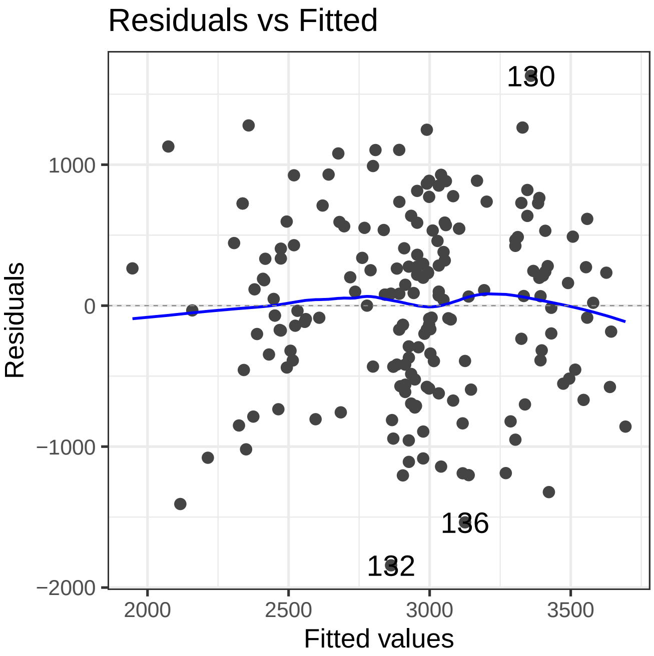
Discussion and conclusion. In conclusion we examined how certain risk factors effect the infant's birth weight by applying a multiple linear regression model to delineate the effect of each of the factors. We found that each of the risk factors we examined have a statistically significant effect on an infant's birth weight, putting that infant at risk.

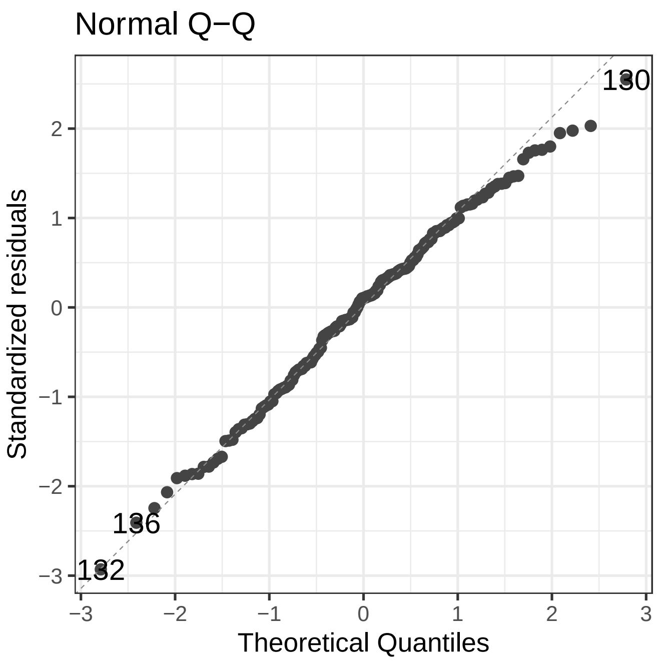
However there are some potential limitations. Most significantly being the data itself being recorded at a charity health centre, where more disadvantage and lower income mothers are expected. Thus the data may not be an accurate representation of the general population. Furthermore it is difficult to say with absolute certainty that each risk factor is independent of one another. For example is a mother that smokes also at risk of being over weight or vice versa. None the less we believe our model to be accurate and the risk factors certain.

Appendix 1.



Appendix 2.





```
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 170, 169, 170, 170, 169, 171, ...
Resampling results:

  RMSE      Rsquared   MAE
  658.8707  0.2283474  534.5831

Tuning parameter 'intercept' was held constant at a value of TRU
```

Appendix 3.

Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	2837.26	243.68	11.64	0.00
mother_weight	4.24	1.68	2.53	0.01
raceBlack	-475.06	145.60	-3.26	0.00
raceOther	-348.15	112.36	-3.10	0.00
smokesTrue	-356.32	103.44	-3.44	0.00
hypertension	-585.19	199.64	-2.93	0.00
uterine_irr	-525.52	134.68	-3.90	0.00

Full Model

```
Linear Regression

189 samples
  8 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 169, 170, 171, 169, 171, 171, ...
Resampling results:

  RMSE      Rsquared   MAE
  676.9286  0.2007824  542.098

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Simple Model

```
Linear Regression

189 samples
  5 predictor
```