Predictive modelling of weight at birth

Flynn Entwistle, Jerry Shum, Lara Pierce, Shaun Lim, Eric Huang

This version was compiled on February 26, 2024

We sought to analyse the significance of 8 variables as predictors of weight at birth in a multiple regression model, using observations collected from Baystate Medical Centre, MA (n=189). A supplementary investigation also took place to compare out of sample performance between models trained on the original presentation of categorical variables, and an alternative dataset where factor levels with few observations had been merged. 6 of the studied variables were found to be significant predictors of birth weight, with merged factor levels in the case of few observations providing the best out of sample performance.

Introduction

Low birth weight expresses comorbidity with a broad range of health complications among infants, shown to be prevalent in 15-20% of newborn mortality cases worldwide, and a predictor for stunting and other metabolic risk factors in later life (McCormick, 1985). But despite its effects being well documented in the current scientific literature, predictors of low birth weight during pregnancy remain scarcely considered. Addressing this deficit, we sought to find the most significant predictors of low birth weight. Multiple regression models were constructed for this purpose, before choosing one which optimised for out of sample performance to increase the how well our model generalised to future observations. Future investigations should consider how well the findings of this study generalise to predict birth weight within a more diverse sample - between different hospitals and regions.

Discussion of data. The `birthwt` dataset obtained from the `MASS` package contains observations from ($n = 189$) mothers collected in 1986 at Baystate Medical Centre in Springfield, Massachusetts. The dataset contained 10 variables in total. For the purposes of this study, we took infant birth weight as the dependent variable. `low` provided no useful information in predicting birth weight, so was dropped from the dataset. This left 4 numerical and 4 categorical independent variables which we could investigate as predictors of birth weight. Studied variables were renamed and transformed to metric units for legibility and ease of comparison. Table A1 contains a tabular summary of the dataset.

Analysis

Preliminary transformations. From our EDA, we knew some predictors to individually not satisfy our assumptions with respect to the dependent variable.

```
#  Warning: The dot-dot notation (`..rr.label..`) was deprecated in ggplot2 3.4.0.
#  i Please use `after_stat(rr.label)` instead.
#  This warning is displayed once every 8 hours.
#  Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
#  generated.
```

Figure 1 shows the discrete numerical variables *number of premature labours* and *first trimester physician visits* to each have too few intervals to confidently suggest a linear relationship.

No mathematical transformations would be appropriate here. But, rather than dropping these variables from the model, their few unique values allowed us to reasonably transform them into categorical variables as per Figure 2

```
birthwt$Premature_labours %<>% factor()
birthwt$First_tri_physician_visits %<>% factor()
```

Additionally, it was found that when applying a log transformation to *weight at last menstruation*, linearity and homoscedascity assumptions were more confidently satisfied against the dependent variable, with a more uniform spread of residuals above and below 0 when plotting against fitted values.

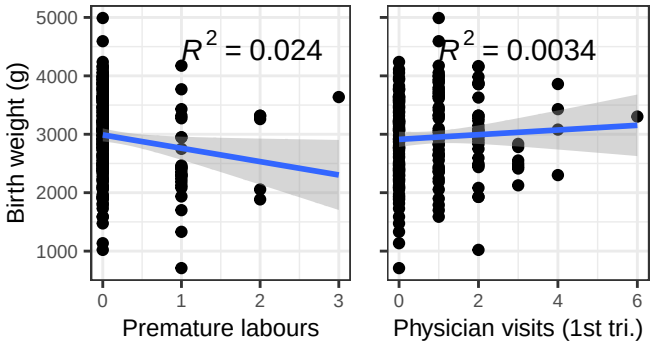

Fig. 1. Premature labours and first trimester physician visits against birth weight (numerical)

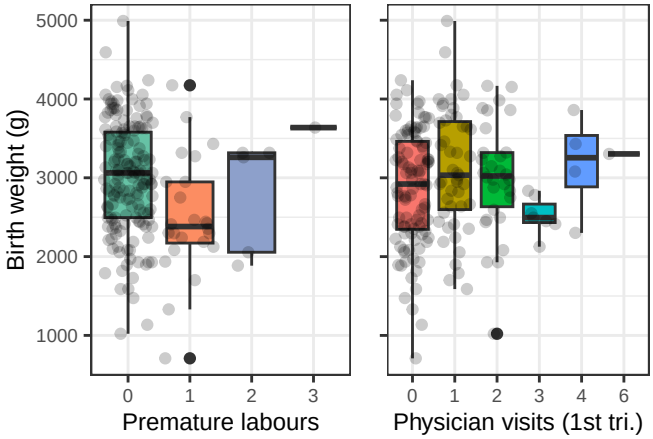

Fig. 2. Premature labours and first trimester physician visits against birth weight (factored)

```
birthwt %<>% mutate(log_weight_at_last_menstruation
          = log(Weight_at_last_menstruation))
```

Alternative dataset with merged `ptl` and `ftv` levels. Factorising premature labours and first trimester physician visits allowed our linearity assumptions to be met, but created some levels with very few observations ($n \leq 5$).

We hypothesised that this could be a source of overfitting - unnecessarily complicating our model with extra predictors whilst providing unreliable information about the dependent variable at these levels.

To test for this, we created an alternative dataset where levels with few observations in these variables had been merged together as per Table 1.

Premature labours	Original factor levels	Merged factor levels
0	159	159
1	24	–
2	5	–
3	1	–
>=1	–	30

1st tri. physician visits	Original factor levels	Merged factor levels
0	100	100
1	47	47
2	30	30
3	7	–
4	4	–
6	1	–
>= 3	–	12

Table 1. Level merging - Premature labours (left) and First trimester physician visits (right)

These will be analysed in comparison with models trained on the original presentation of these factors.

Post-transformation assumption checking. After transformation, we constructed full models for both datasets incorporating all variables as predictors of birth weight.

```
M1.original <- lm(Birth_weight ~ ., data=birthwt)
M1.merged <- lm(Birth_weight ~ ., data=birthwt.mrgd.lvls)
```

Assumptions were checked for each model:

- By experimental design, the observations are naturally independent, so our **independence** assumption is satisfied.
- Moreover, we have a sufficient number of observations to rely on the central limit theorem to satisfy our **normality** assumption.

To check our remaining assumptions, we constructed residuals against fitted values plots for each:

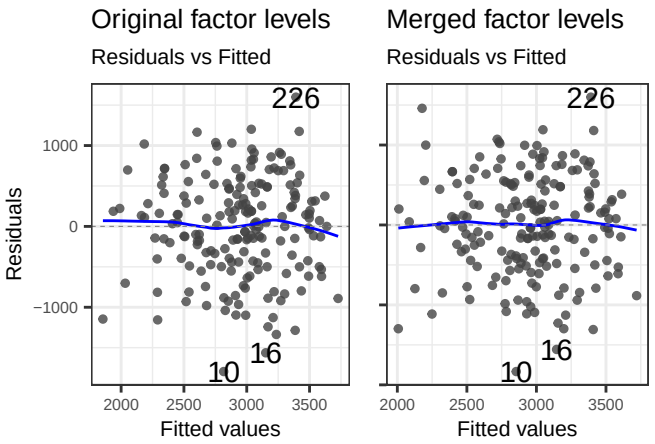

Fig. 3. Residuals vs Fitted values plot (Full model, post transformations)

Figure 3 shows no obvious non-linear pattern in the residuals in either plot, with residuals approximately equally distributed about 0, so our **linearity** and **homoscedascity** assumptions were similarly satisfied.

Model selection. With our assumptions satisfied, we began stepwise variable selection to construct models for each dataset, including only the most significant predictors.

For each dataset, forward and backward selection procedures produced the same models, with only `mother age`, and `first trimester physician visits` dropped from the model. This choice of selected predictors was further verified by an exhaustive search - finding the same models for each dataset for the same number of predictors.

The normality and independence assumptions for these newly constructed models remained unchanged from the full models. Residual vs fitted values plots (Figure A1) were constructed to reassess our linearity and homoscedascity assumptions, which were found to still be satisfied.

Results

	Full model (Orig. lvls)	Refined model (Orig. lvls)	Full model (Merged lvls)	Refined model (Merged lvls)
Observations	189	189	189	189
R^2 / R^2 adjusted	0.292 / 0.230	0.275 / 0.239	0.257 / 0.211	0.250 / 0.221
AIC	2995.971	2988.232	2996.909	2990.702

Table 2. R^2 and AIC of each model

In-sample performance. Table 2 shows the coefficient of determination, R^2, was highest in the model with the most predictors, and lowest in the model with the least predictors. This was to be expected, as adding more predictors to a model will always have a non-increasing effect on the residual sum of squares.

Accordingly, the full model with the most number of predictors explained the greatest proportion of variation within our results, giving it the best in-sample performance.

However, this statistic alone tells us little about how the models perform on unseen data, which the AIC attempts to account for by penalising for more predictors, discouraging overfitting and hinting at better out-of-sample performance.

We had significant reductions in AIC between each full model and its refined alternative, indicating our refined models had better fit per added predictor. Between the two refined models, the alternative with original factor levels of premature labours had a marginally better AIC by ~2 points - making them approximately equally well fitting.

An expanded version of Table 2 including model coefficients can be found in the appendix, as Table A2.

Validating out of sample performance. To gauge out of sample performance, we performed 10-fold cross-validation with 1000 repeats, with the results shown in Figure 4. Significant reductions in Root Mean Squared Error and Mean Absolute Error that in our refined models, after dropping the least significant predictors. Our models trained on the data with merged factor levels also had better out of sample performance than those trained on the original factor levels.

Final model and interpretation. In choosing our final model, we prioritised out of sample performance to allow to optimise for generalisability to future observations, whilst also balancing for reasonable in-sample performance. Accordingly, we chose the refined model with merged factor levels - having the best out of sample performance at minimal cost to R^2:

$$\widehat{\text{Birth weight}} = 610.22 + 132.66(\text{Race}_{\text{Other}}) + 460.01(\text{Race}_{\text{White}}) - 316.87(\text{Smoke}_{\text{Yes}}) - 211.68(\text{PTL}_{>=1}) - 562.41(\text{HT}_{\text{Yes}}) - 483.45(\text{UI}_{\text{Yes}}) + 572.37(\log(\text{LMW}))$$

Holding all other predictors constant, mothers of "other" and "white" race had positive correlations with birth weight in comparison to black mothers, whereas histories of smoking, premature labour/s, hypertension, and uterine irritation were negatively correlated with birth weight.

On average, a one percent increase in weight at last menstruation would be expected to result in a 5.7 gram increase in weight at birth.

Discussion and conclusion. At the cost of avoiding overfitting, we merged factor levels in the number of premature labour & number of first trimester physician visits, resulting in a loss of information. This contributes to a limitation of our analysis, since our model is blind to any differences from observations larger than our unique categories. This possibly introduces bias and limits the model's predictive capacity.

Our data is only collected from one US medical centre, which is not a representative sample of the wider population. This induces selection bias, and the observations may not generalise well to other populations.

To overcome both of these limitations, more data collection is required. A random sample with more observations would provide us with more information, possibly reducing the necessity of merging factor levels. It would also reduce selection bias, and provide better overall model predictions.

In summary, there was a significant improvement in accuracy and a decrease in errors for out of sample performance when we used our refined model with merged factor levels. This was sufficiently balanced, at the worthwhile expense of a slightly smaller R^2 value. We also found that the variables of mother's age and number of physician visits were not very informative predictors for birth weight and hence could be dropped from the regression model entirely.