Summary

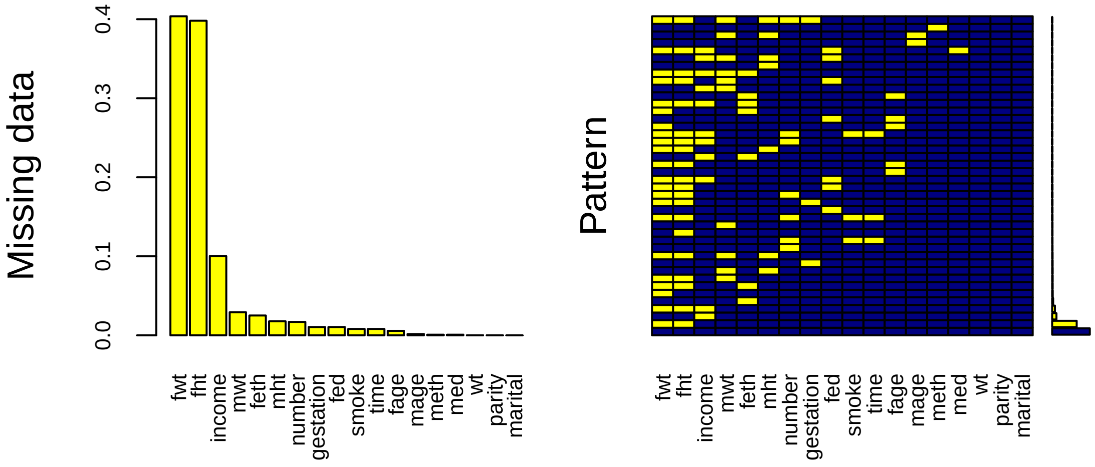
This report summarizes the statistical modelling and analysis results associated with the birth data on 1236 healthy male single-fetus births. The purpose was to explore the relation between birth weight and explanatory variables using linear regression techniques learnt in Stat 331. As a first step, we performed preliminary diagnostics on the data and refined the set of regressors. Subsequently, we imputed data for the "income" covariate using the "mice" library. Automatic Model Selection was then performed.
Stratified random sampling with optimal allocation eliminated the possibility of missing categories in the training set. Once the training set was determined, forwards, backwards, and stepwise selection was used to determine a candidate model. The second candidate model was created manually and refined by removing insignificant covariates. Various tests including AIC, PRESS statistic, K-fold cross validation were performed for model diagnostics. It was found that the automatic model selection, with the lower value of Mean Squared Predictive Error, had a predictive advantage over the manually selected model.

Pre-fitting Diagnostics

Missing Data and Imputation

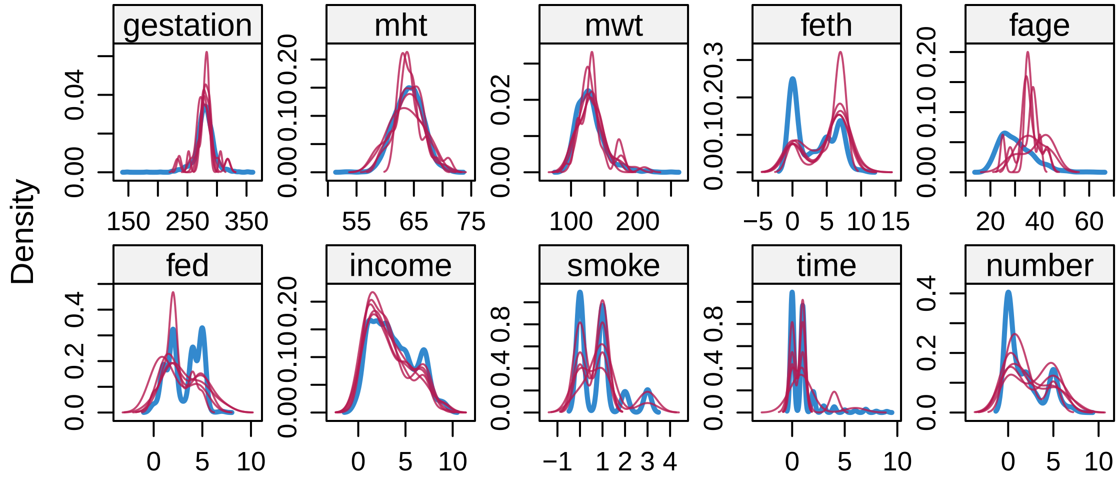
The functions used for data imputation are dependent on the libraries VIM and mice

Visualizing the missing data



Looking at the above charts, we might want to remove or impute values into the following features of data, fwt: Fathers Weight, fht: Fathers Height, income. Since imputing may not work for fathers height and weight (greater than 30% NAs), we will remove those features from our data. Logically, it makes sense for the baby's weight to not depend significantly on these features.

Using the mice library, we perform imputation for the income covariate.



Blue denotes the observed values and magenta denotes the imputed values. From the density plot, we can say that imputed values are indeed plausible for the income covariate.

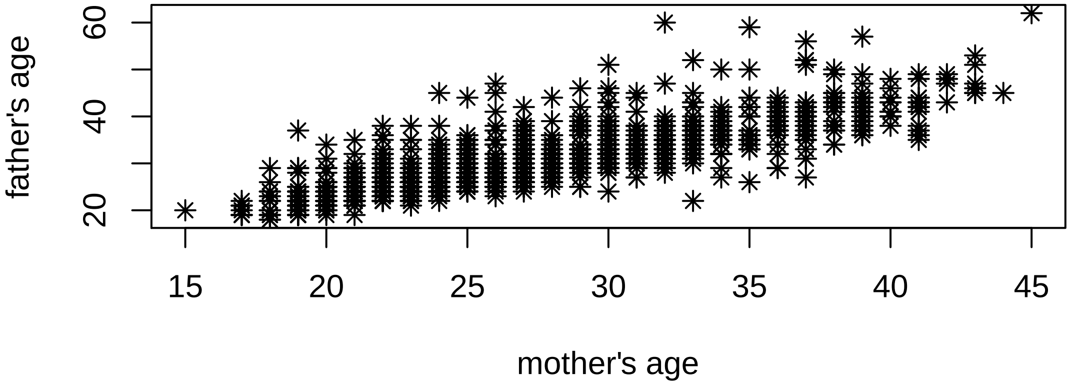
Since the imputed values for income in all imuputed datasets follow a similar distribution, we choose imputed dataset 1 as it has the smallest difference in the average (for all covariates) when compared the observed data.

```
## [1] 0.01060839 0.23552749 0.07856956 0.08018768 0.02598056
```

We choose to include the imputed income data from imputed dataset 1 into our original dataset and then perform model diagnostics, selection.

Refined Variable Set

Since approximately 40% of the father's height and father's weight data is missing, these two covariates will be excluded.In order to create candidate models for futher examination, the current variable set must first be refined significantly. Consider the following plot:



From this plot, a linear relationship between the mother's age and the father's age is very obvious. More concretely, in a linear model where mother's age is the only covariate used to predict the father's age, the mother's age has a p-value of 1.263×10^{-299}. So father's age has a very strong positive correlation with mother's age. Hence, including father's age along with mother's age gives redundant information. The mother's age is more important to include, as there may be interactions between other biological factors of the mother (such as gestation). Due to these reasons, father's age was also excluded from the refined set of variables to consider.

By studying the scatter plots between pairs of variables (omitted for the sake of brevity), the following list of variables were selected: gestation, parity, time, number, smoke, mother's age, mother's weight, mother's height, income, and mother's ethnicity. In addition to this, an interaction effect between gestation and parity was added, as well as an interaction effect between gestation and mother's age.

Instead of a standard interaction variable between the mother's weight and height (of the form $weight \times height$), the mother's body-mass index was used. This is an interaction variable between the weight and height of the mother, but is defined as

$$\frac{weight \times 703}{(height)^2}$$

This variable gives insight into whether the mother is a healthy weight, overweight, or underweight. We initially tried centering the body-mass index variable at 30, which is the threshold for obesity, however it was found that this reduced the predictive power of the model.

The VIF for the covariates are presented below.

```
## gestation    parity      meth      mage       med       mht       mwt
##  1.068271  1.663366  1.975567  3.786943  1.736444  1.477580  1.401956
##      feth      fage       fed       fht       fwt   marital    income
##  1.947246  3.433770  1.653082  1.671355  1.513025  1.038491  1.231330
##     smoke      time    number
##  6.711304  5.665772  1.631473
```

As shown by the table above, all of the VIF covariates are relatively low, and hence they are not colinear.

Model Selection

Training Set

Automatic Model Selection requires the dataset to be partitioned into a training set and a testing set. However, there are several categorical variables in this dataset. If, for a given categorical variable x_i, with levels $1, \ldots, k$, only levels $1, \ldots, j$, $j < k$ occur in the training set, then there will be no β_k corresponding to $I(x_i = k)$. Hence, if $x_i^* = k$, then $y(x^*)$ will be undefined. To avoid this error, stratified random sampling must be used. In stratified random sampling, data points are randomly selected from each strata. The strata in this case are the levels for the categorical variable. *Optimal Allocation* was used to determine how many data points should be selected for each strata. In optimal allocation,

$$n_h = \frac{W_h \sigma_h}{W_1 \sigma_1 + \cdots + W_H \sigma_H} n$$