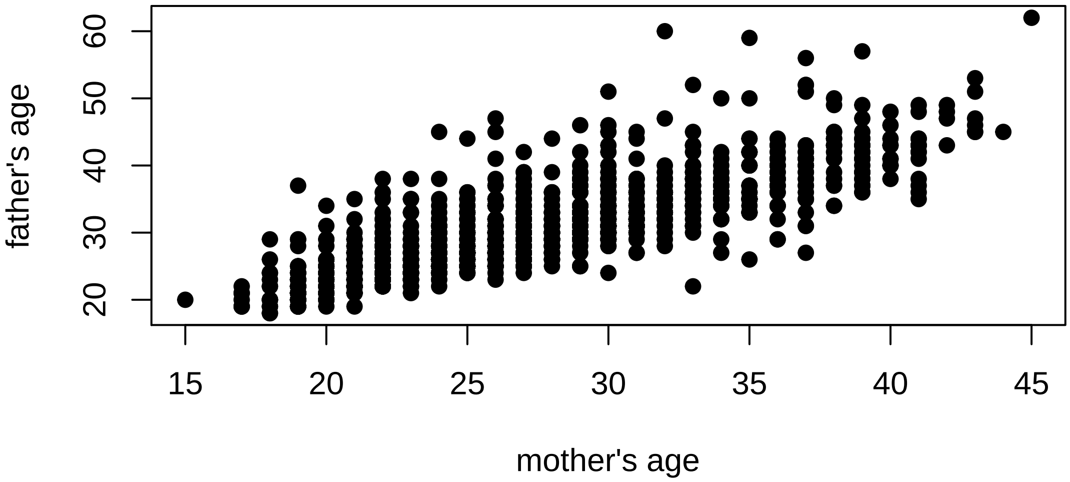
Stat 331 Final Project

2018-12-1

Model Selection

Refined Variable Set

As aforementioned, the father's height and the father's weight have been removed from the dataset. In order to create candidate models for futher examination, the current variable set must first be refined significantly. Consider the following plot:



```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 4.0143530 0.53616590  7.487147  1.341011e-13
## mage        0.9674098 0.01926662 50.211706 1.263000e-299
```

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

$$\frac{motherweight \times 703}{(mother'sheight)^2}$$

. 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