STAT 331 Applied Linear Models

Final Project

Shreya Prasad- 20694106
Haripriya Pulyassary - 20654553
Ziyao Tian - 20566818

Summary

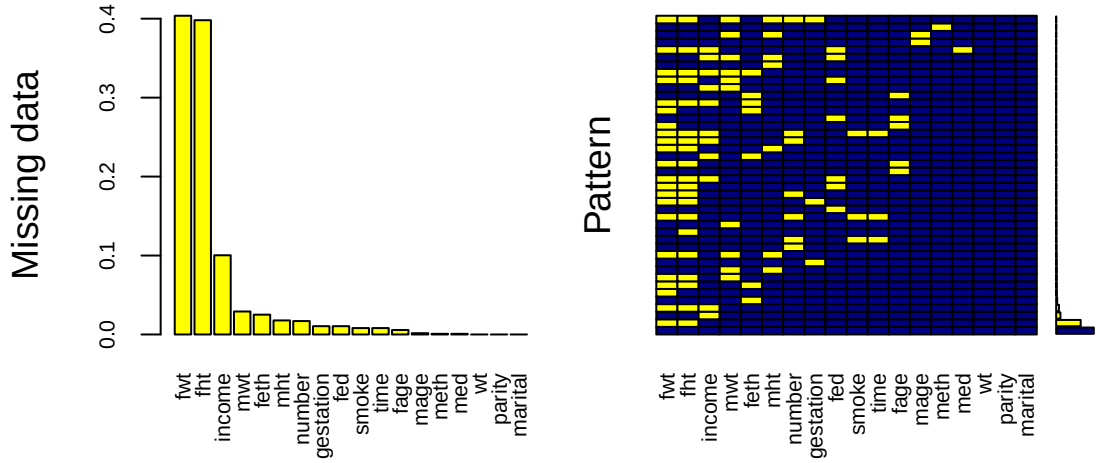
This report summarizes the statistical modelling and analysis results associated with the birth data on 1236 healthy male single-fetus births. The purpose was to explore the relation between birth weight and explanatory variables using linear regression techniques learnt in Stat 331. As a first step, 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

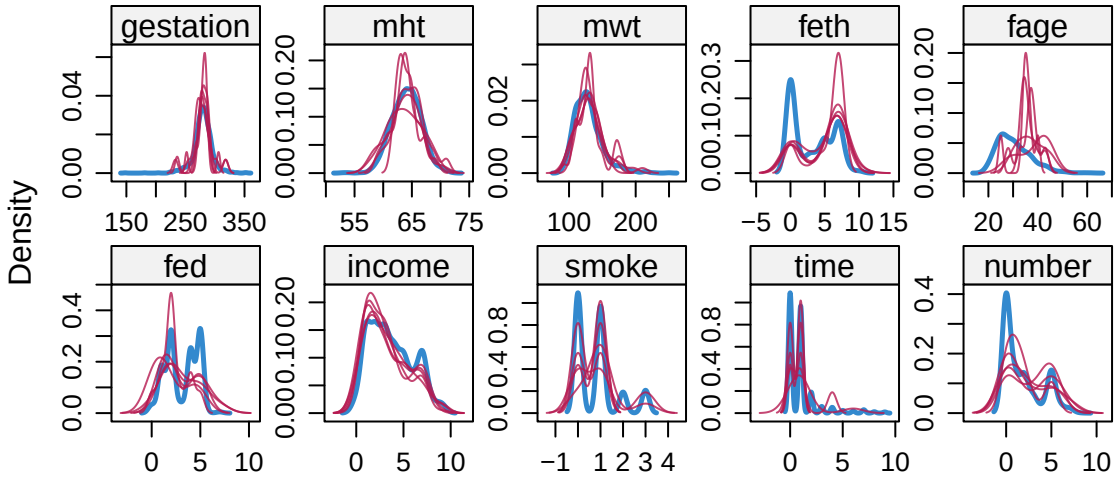
The functions used for data imputation are dependent on the libraries VIM and mice

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.

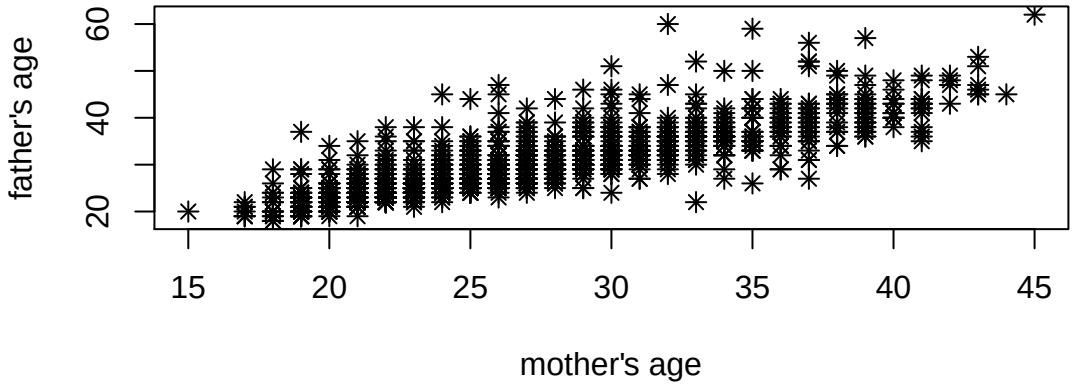
Since the imputed values for income in all imuputed datasets follow a similar distribution, we choose imputed dataset 1 as it has the smallest difference in the average (for all covariates) when compared the observed data.

```
## [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,

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

where n_h is the number of points selected from stratum h, $W_h = \frac{N_h}{N}$ is the proportion of stratum h in the population, and σ_h is the standard deviation of the response variate (in this case, weight) in stratum h. In this case, we do not know the true value, so we must use an estimate for the standard deviation. Since there are multiple categorical variables, optimal allocation has to be used for each categorical variable. Suppose that the desired size of the training set is n, and that the number of categorical variables is r. Optimal allocation was used to sample $\frac{n}{r}$ for each categorical variable $x_1, \ldots, x_r$ (yielding a total training set size of n as desired).

Selection of Candidate Models (Automatic and Manual)

Model 1 (automatic selection)

In the previous stage (pre-fitting diagnostics), we were able to reduce the set of covariates to consider when building the model. The set which we are now considering is: gestation, parity, time, number, smoke, mother's age, mother's weight, mother's height, body-mass index, gestation-parity interaction, gestation-mage interaction, income, and mother's ethnicity. Using this set of variables, we will run forward, backwards, and step-wise selection to obtain 3 different models.

We compare these three models using their respective MSPEs (calculated after training the model on the train set).

```
## lm(formula = wt ~ gestation + smoke + mht + meth + I(gestation *
##     parity) + I(mwt * 703/(mht)^2) + parity + mwt, data = birth.data,
##     subset = train.ind)
## [1] 115024.5
```

```
## lm(formula = wt ~ mage + I(mwt * 703/(mht)^2) + smoke + mht +
##     meth + I(gestation * mage) + I(gestation * parity), data = birth.data,
##     subset = train.ind)
## [1] 109848.6
```

```
## lm(formula = wt ~ parity + time + mage + mht + I(mwt * 703/(mht)^2) +
##     I(gestation * parity) + I(gestation * mage) + meth, data = birth.data,
##     subset = train.ind)
## [1] 115157.1
```

The model obtained from stepwise selection has the lowest MSPE so we will select it as one of our candidate models. We will first begin by analyzing the significance of each covariate in the presence of other covariates. For brevity's sake, only the non-categorical variables are shown below.

```
##                            Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)            2.748489809 17.237324926   0.1594499 8.733635e-01
## mage                  -4.509967490  0.416268998 -10.8342622 2.821334e-25
## I(mwt * 703/(mht)^2)   0.509400476  0.194914428   2.6134570 9.168672e-03
## mht                    1.525102565  0.250294054   6.0932433 1.886433e-09
## I(gestation * mage)    0.016491892  0.001442051  11.4364112 9.598508e-28
## I(gestation * parity)  0.003020828  0.001428050   2.1153519 3.477629e-02
```

For the non-categorical variables, namely mother's age, mother's height, BMI (the body-mass index which is $\left(\frac{703 \times mwt_i}{mht_i^2}\right)$) the gestation-mage interaction variable and the gestation-parity interaction variable, a t-test with the null hypothesis $H_0 : \beta_i = 0$ can be used. The corresponding p-values are all strictly less than 0.05, so at the 0.05 significance level, the null hypothesis can be rejected. Hence, each of mother's age, mother's height, BMI, the gestation-mage interaction variable and the gestation-parity interaction variable are significant in the presence of the other covariates.

For the categorical variables, namely smoke and mother's ethnicity, we must use an F-test to determine if the categorical variables are significant in the presence of the other covariates.

```
## Analysis of Variance Table
##
## Model 1: wt ~ mage + I(mwt * 703/(mht)^2) + mht + meth + I(gestation *
##     mage) + I(gestation * parity)
## Model 2: wt ~ mage + I(mwt * 703/(mht)^2) + smoke + mht + meth + I(gestation *
##     mage) + I(gestation * parity)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    659 170275
## 2    656 162084  3    8191.2 11.051 4.382e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 4.382×10^{-7}, so at the 0.05 significance level, we can conclude that the categorical variable smoke is significant in the presence of the other covariates.

```
## Analysis of Variance Table
##
## Model 1: wt ~ mage + I(mwt * 703/(mht)^2) + mht + smoke + I(gestation *
##     mage) + I(gestation * parity)
## Model 2: wt ~ mage + I(mwt * 703/(mht)^2) + smoke + mht + meth + I(gestation *
##     mage) + I(gestation * parity)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    661 172043
## 2    656 162084  5    9958.7 8.0611 2.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 2.16×10^{-7}, so at the 0.05 significance level, we can conclude that the categorical variable mother's ethnicity is significant in the presence of the other covariates. Thus, in this particular model, all covariates are significant in the presence of each other.

Model 2 (Manual)

The following model was created manually using the evaluation done in the previous stage and intuition. It combines income, smoke and number (which measures the magnitude of the mother's smoking habit) with other biological factors such as gestation, parity, and the mother's height (which was found to be selected frequently by automatic model selection even when different combinations of variables were used). For brevity's sake, only the non-categorical variables are shown below.

```
##                          Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)           -79.0281462 21.39450487 -3.693853 2.395507e-04
## gestation               0.3921665  0.05445977  7.201031 1.665261e-12
## mht                     1.4121580  0.25418283  5.555678 4.042965e-08
## I(gestation * parity)   0.0360373  0.01623026  2.220376 2.673891e-02
## parity                 -9.2947283  4.47561147 -2.076750 3.821870e-02
```

For the non-categorical variables, namely gestation, mother's height, parity and the gestation-parity interaction variable, a t-test with the null hypothesis $H_0 : \beta_i = 0$ can be used. The corresponding p-values are all strictly less than 0.05, so at the 0.05 significance level, the null hypothesis can be rejected. Hence, each of gestation, mother's height, parity, and the gestation-parity interaction covariate are significant in the presence of the other covariates.

```
## Analysis of Variance Table
##
## Model 1: wt ~ gestation + mht + I(gestation * parity) + parity + number +
##     income
```

```
## Model 2: wt ~ gestation + mht + smoke + I(gestation * parity) + parity +
##     number + income
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    649 172235
## 2    647 169747  2    2488.4 4.7424 0.009023 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 0.009, so at the 0.05 significance level, the null hypothesis that smoke is insignificant in the presence of other covariates can be rejected. Hence, the categorical variable smoke is significant in the presence of the other covariates.

```
## Analysis of Variance Table
##
## Model 1: wt ~ gestation + mht + smoke + I(gestation * parity) + parity +
##     income
## Model 2: wt ~ gestation + mht + smoke + I(gestation * parity) + parity +
##     number + income
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    654 172028
## 2    647 169747  7    2281.3 1.2422 0.2772
```

The p value is 0.28, so at the 0.05 significance level, the null hypothesis that number is insignificant in the presence of other covariates cannot be rejected. So we accept the null hypothesis that number is insignificant in the presence of the other covariates

```
## Analysis of Variance Table
##
## Model 1: wt ~ gestation + mht + smoke + I(gestation * parity) + parity +
##     number
## Model 2: wt ~ gestation + mht + smoke + I(gestation * parity) + parity +
##     number + income
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    655 171436
## 2    647 169747  8    1689.2 0.8048 0.5985
```

The p value is 0.60, so at the 0.05 significance level, the null hypothesis that income is insignificant in the presence of other covariates cannot be rejected. So we accept the null hypothesis that number is insignificant in the presence of the other covariates. So we further refine the model by removing income and number.

Refined Model 2

```
##                          Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)           -74.2265183 21.12198303 -3.514183 4.711590e-04
## gestation               0.3852541  0.05363568  7.182794 1.843752e-12
## mht                     1.3660556  0.24823798  5.503008 5.341820e-08
## I(gestation * parity)   0.0382404  0.01599552  2.390694 1.709499e-02
## parity                 -9.8405138  4.41234785 -2.230222 2.606718e-02
```

Since the p-values corresponding to the non-categorical variables are all strictly less than 0.05, at the 0.05 significance level, the null hypothesis can be rejected. Hence, each of gestation, mother's height, parity, and the gestation-parity interaction covariate are significant in the presence of the other covariates in the refined model.

```
## Analysis of Variance Table
##
## Model 1: wt ~ gestation + mht + I(gestation * parity) + parity
```

```
## Model 2: wt ~ gestation + mht + smoke + I(gestation * parity) + parity
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    665 182741
## 2    662 173703  3      9038 11.482 2.402e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 2.402×10^{-7}, so at the 0.05 significance level, the null hypothesis that smoke is insignificant in the presence of other covariates can be rejected. Hence, the categorical variable smoke is significant in the presence of the other covariates.
To summarize, the two models we will proceed to compare are:

Model 1:

```
##           (Intercept)                  mage  I(mwt * 703/(mht)^2)
##           2.748489809          -4.509967490           0.509400476
##              smokenow  smokeuntil pregnancy          smokeused to
##          -7.323507616           0.802088003          -1.041515927
##                   mht             methAsian         methCaucasian
##           1.525102565           3.739051306           7.925141013
##           methMexican             methMixed             methOther
##          19.735897417           7.997043034           7.052603812
##   I(gestation * mage) I(gestation * parity)
##           0.016491892           0.003020828
```

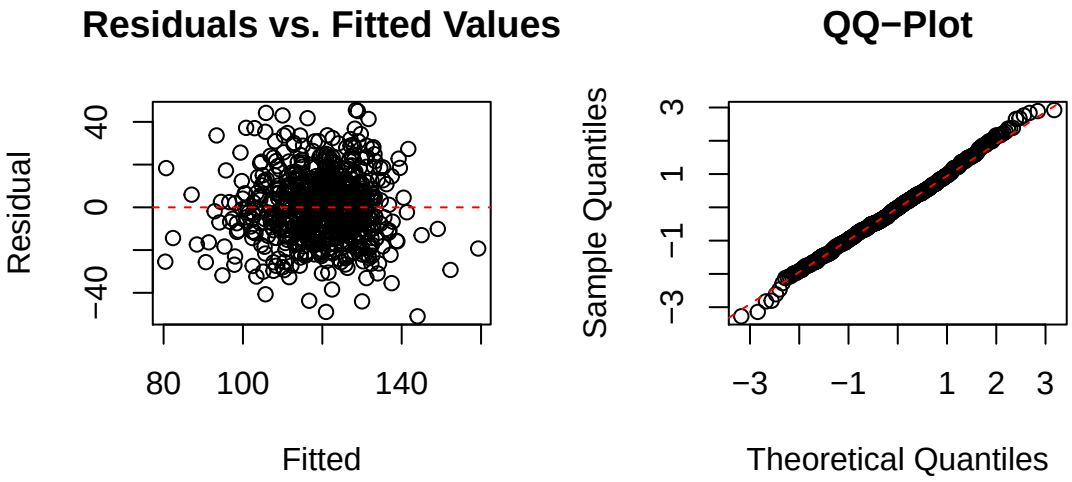
Model 2:

```
##           (Intercept)             gestation                   mht
##           -74.2265183             0.3852541             1.3660556
##              smokenow  smokeuntil pregnancy          smokeused to
##            -7.6680553             0.3658084            -0.8063291
## I(gestation * parity)                parity
##             0.0382404            -9.8405138
```

Model Diagnostics

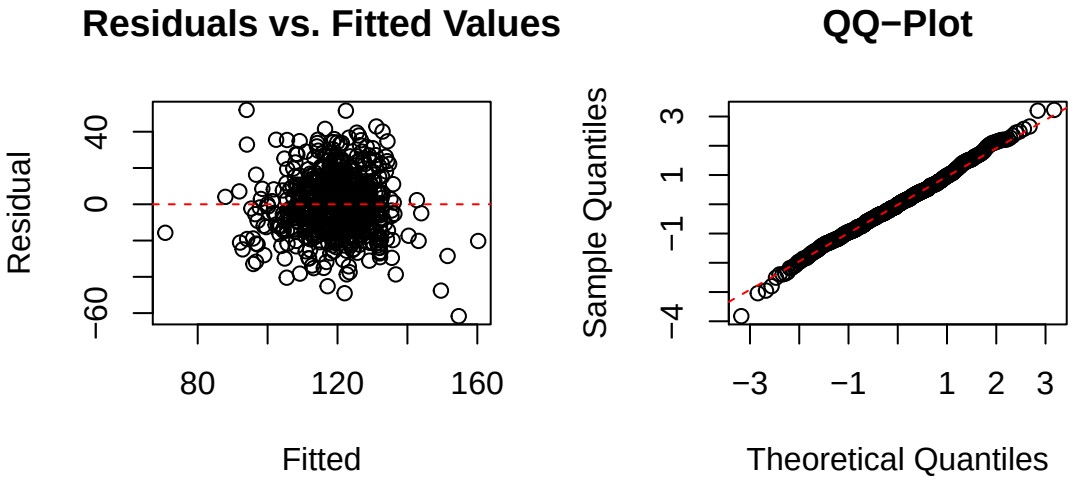
Residual Plots

Model 1 (Selected via Automated Model Selection)



From the residuals vs. fitted values plot, we can see that the conditional mean of the response ($weight_i$) is a linear function. We can also see that the conditional variance of $weight_i$ is constant. From the QQ-plot, we can see that the errors are iid normals. Hence, from these two plots we can conclude that Model 1 satisfies the linear regression model assumptions.

Model 2 (Selected via Manual Model Selection)



From the residuals vs. fitted values plot, we can see that the conditional mean of the response ($weight_i$) is a linear function. We can also see that the conditional variance of $weight_i$ is constant. From the QQ-plot, we can see that the errors are iid normals. Hence, from these two plots we can conclude that Model 2 satisfies the linear regression model assumptions.

K-fold cross-validation

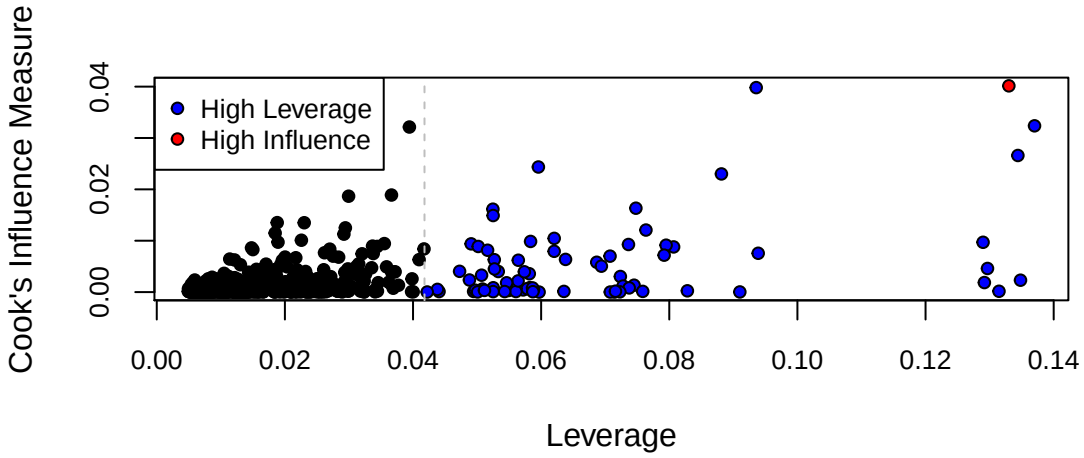
We used the k-fold cross validation to compare the two models(Automatic and Manual). We partitioned the original data into k (10) equal subsets. Then, we trainied each of our models on k-1 folds of the data, which formed the training sets.The accuracy of our models was calculated by validating the predicted results for the kth fold, which formed the test set. The model with a lower value of **MSPE** (Mean Squared Predictive Error) was chosen due to higher predictive accuracy.

```
##   Auto_model Manual_model
##      88067.9      91890.7
```

Looking at the two MSPE values we can say that the model selected through automatic model selection is better since it has a lower value of MSPE on the testing data.

Leverege and Influence measures

Model 1

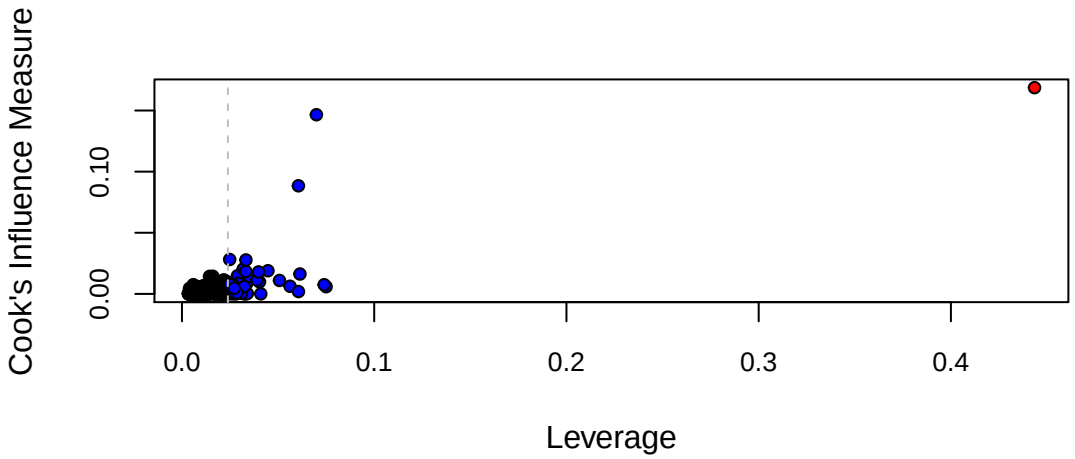


In this graph, we can see that there is one point which has high influence according to Cook's Influence Measure (in red). There are several blue points, which are points that have high leverage. The point with the highest influence is:

```
##      wt gestation parity  time number smoke mage mwt mht             meth
## 150 126       282      8 never  never never   38 250  66 African-American
##          income
## 150 12500-14999
```

So the parity, mother's age and mother's weight are greater than the mean values (1.9,27, 130 respectively) which would explain why this is a high-influence point.

Model 2



In this graph, we can see that there is one point which has high influence according to Cook's Influence Measure (in red). There are several blue points, which are points that have high leverage, however these points are less spread out than in Model 1. The point with the highest influence is:

```
##      wt gestation parity  time number smoke mage mwt mht      meth
## 413 138       288      0 never  never never   19 124  66 Caucasian
##          income
## 413 12500-14999
```

Here, the mother's age is significantly lower than the mean age, which would explain why the point is of greatest influence.

AIC

The AIC estimates the quality of each model, relative to the other model. The larger the difference in AIC, indicates stronger evidence for one model over the other. In this case, we want to compare two models. If $AIC_1 < AIC_2$ then model 1 would be preferred over model 2.

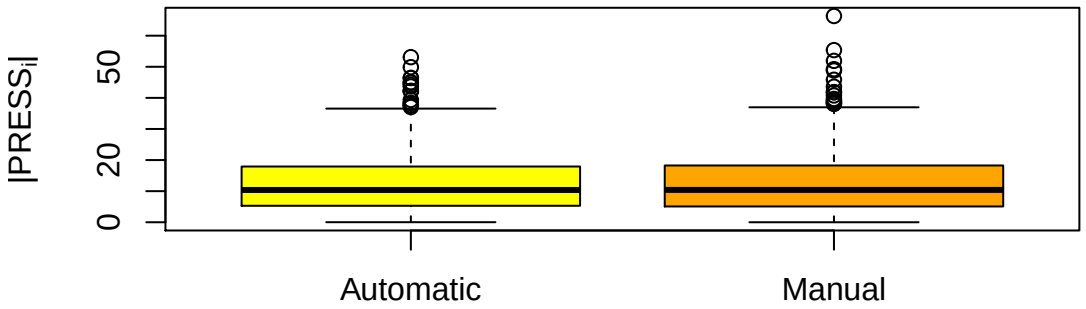
```
## [1] 5608.736
```

```
## [1] 5643.118
```

Since model 1(automatic model) has lower AIC than model 2 (Manually selected), we can say that using model 1 has a greater predictive advantage over model 2.

PRESS Statistic

Press statistic is a form of cross validation to provide a summary measure of the fit of a model to our training set. It should be noted that the training and test set are disjoint. In general, the smaller the PRESS value, the better the model's predictive ability.



We can see that the results for the sum of squared PRESS residuals for the two models are consistent with the cross-validation results, giving a bit of a predictive advantage to the automatic model (Model 1).

Discussion

The final model used following covariates: mother's age, mother's body-mass index, smoke, mother's height, mother's ethnicity, interaction between mother's age and gestation, and interaction between parity and gestation. As determined in the previous section, the model does not seem to violate any of the regression assumptions. There was a point with great influence corresponding to an overweight 38 year old woman. However it would not be wise to exclude it as an "outlier" since there is no justification for why the model would not be able to predict the birth weight of 38-year old mothers.

As discussed in the model selection section, the covariate with the smallest p-value, and hence the most significant (in the presence of the other covariates) was the mother's age. Also, as mentioned in the previous section, for model 1, the point with the highest influence was that with the highest age. It is also very interesting to note that both the gestation-age and the gestation-parity interactions were significant in this model. Both interaction covariates had positive coefficients. Nonetheless, further study would be required to determine the exact extent and nature of these interactions.

Consider two women of the same age, ethnicity, height, weight, parity, and gestation, but one of them smokes now while the other stopped smoking. Then, according to this model, the weight of the baby born to the woman who currently smokes is expected to be lower than the weight of the baby born to the woman who has stopped smoking (since all other factors have been held constant). However, most interestingly, this is true irrespective of when the mother stopped smoking (the coefficient + intercept is positive for both "used to smoke" and "smoked until pregnancy").

It should also be noted that the BMI is a significant covariate and the coefficient is positive. Thus, advice for prospective parents to reduce the probability of having an underweight baby would be to ensure that the mother has a healthy body-mass index, and to also stop smoking. The model suggests that it is better to quit smoking before pregnancy, however quitting at any point is still helpful.

Future Direction

As noted earlier, a significant portion of the father's data was missing. Due to this, the models did not use information from the father, which is a serious shortcoming since a child's genetic information is determined by both parents. Should this investigation be continued, data which is not missing the height/weight of fathers should be collected. Strategies used to reduce non-response bias could be employed as well (e.g. re-surveying a subset of the non-response group). A limitation was that there were only two socio-economic factors in the variable set (ethnicity and income) and one of them, income, did not have great predictive power. This could be due to the tendency that an individual has to lie about their income if they either make significantly less or significantly more than the average person.

Appendix

Setup for producing the data frame

```
# This script produces birth.data, which is the data to be used for
# the regression model, and train.inds, the indices for the train-set
# and num.train which is the size of the training set
# and MSPE function for calculating MSPE for a given model (uses train.ind)
births <- read.csv("chds_births.csv")
meth.names <- c('Caucasian','Caucasian','Caucasian','Caucasian',
'Caucasian','Caucasian','Mexican', 'African-American',
'Asian', 'Mixed', 'Other')
med.names <- c('elementary', 'middle', 'hs', 'hs + trade',
'hs + college', 'college', 'trade', 'unclear')
feth.names <- c('Caucasian','Caucasian','Caucasian','Caucasian',
'Caucasian','Cauasian', 'Mexican', 'African-American',
'Asian', 'Mixed', 'Other')
fed.names <- c('elementary', 'middle', 'hs', 'hs + trade',
'hs + college', 'college', 'trade', 'unclear')
marital.names <- c(NA, 'married', 'separated', 'divorced', 'widowed', 'never married')
income.names <- c('<2500', '2500-4999', '5000-7499', '7500-9999',
'10000-12499', '12500-14999', '15000-17499', '20000-22499', '>22500')
smoke.names <- c('never', 'now', 'until pregnancy', 'used to')
time.names <- c('never', 'still smokes', 'during pregnancy',
'less than a year', '1-2yrs', '2-3yrs', '3-4yrs', '5-9yrs',
'10+yrs', 'quit - unknown then')
number.names <- c('never', '1-4', '5-9', '10-14', '15-19',
'20-29', '30-39', '40-60', '>60', 'smoked, amount unknown')
births$meth <- meth.names[births$meth + 1]
births$feth<- feth.names[births$feth + 1]
births$fed <- fed.names[births$fed + 1]
births$marital <- marital.names[births$marital+1]
births$income <- income.names[fdata$income + 1]
births$smoke <- smoke.names[births$smoke + 1]
births$time <- time.names[births$time + 1]
births$number <- number.names[births$number + 1]
keeps <- c("wt", "gestation", "parity", "time", "number", "smoke", "mage",
"mwt", "mht", "meth","income")
cat.var <- c("smoke", "number", "time", "income")
birth.data <- births[keeps]
birth.data <- na.omit(birth.data)
#Initial Models
ntot <- dim(birth.data)[1]
ntrain <- 1000
train.ind <- c(NA)
for (c in cat.var) {
  train.ind.curr <- OPTALLOC(birth.data, c, "wt", ntrain/length(cat.var), 23430)
  train.ind <- unique(c(train.ind, train.ind.curr))
}
train.ind <- na.omit(train.ind)
num.train <- length(train.ind)
```

```r
MSPE <- function(M, train) {
  print(M$call)
  print(sum((birth.data$wt[-train] - predict(M, newdata = birth.data[-train,]))^2))
}
```

Missing Data and Imputation for income covariate

```r
#To run this code, we need the following libraries
#* mice
#* VIM
#Reading in the data
fdata <- read.csv("chds_births.csv")
head(fdata)

#Calculat the number of missing valuesin each column
na_count <- sapply(fdata, function(y) sum(length(which(is.na(y)))))
na.count <- data.frame(na_count)
na.count

#Check the columns that have more than 10% of the data missing
count <- sapply(fdata, function(y) length(y))
na_percent <- (na_count/count)*100
na_percent <- data.frame(na_percent)
na_percent

library(VIM)
```

```r
#Plot to visualize missing data
#miss_plot <- aggr(fdata, col=c('navyblue','yellow'),
  #                   numbers=TRUE, sortVars=TRUE,
  #                  labels=names(fdata), cex.axis=.7,
  #                gap=3, ylab=c("Missing data","Pattern"))

#Deleting father's weight and height due to missing data
fdata$fht <- NULL
fdata$fwt <- NULL
colnames(fdata)
head(fdata)

library(mice)
#The md.pattern function allows us to get a better understanding of the pattern of missing data
#md.pattern(fdata)
```

```r
#Imputting the data with "pmm" method as referenced from "https://stefvanbuuren.name/mice/"
imp <- mice(fdata, m = 5, maxit = 50, meth = 'pmm', seed = 500)

#This shows the imputation for each of the 5 iterations
head(complete(imp))
#Looking at one of the complete datasets #4
head(complete(imp,2))
summary(imp)
```

```r
#Plot of density functions of imputed data overlayed on the observed values for each of the data.
#densityplot(imp)
#blue - observed obseravtions
#magenta - imputed values

imp_1 <- data.frame(complete(imp,1))
imp_2 <- data.frame(complete(imp,2))
imp_3 <- data.frame(complete(imp,3))
imp_4 <- data.frame(complete(imp,4))
imp_5 <- data.frame(complete(imp,5))

cm1 <- sapply(fdata, mean, na.rm = T) - sapply(imp_1, mean)
cm2 <- sapply(fdata, mean, na.rm = T) - sapply(imp_2, mean)
cm3 <- sapply(fdata, mean, na.rm = T) - sapply(imp_3, mean)
cm4 <- sapply(fdata, mean, na.rm = T) - sapply(imp_4, mean)
cm5 <- sapply(fdata, mean, na.rm = T) - sapply(imp_5, mean)
sum_imp_data <- c(sum(cm1), sum(cm2), sum(cm3), sum(cm4), sum(cm5))
abs(sum_imp_data)

#We choose to include the imputed income data from imputed dataset 1 into our
#original dataset and then carry #our model diagnostics and selection from there.
#Get sample 1 income values
imputed_income <- imp_1$income
fdata$income <- imputed_income
```

Optimal Allocation

```r
OPTALLOC <- function(df, stratum, y, n, seed){
  #
  # OPTALLOC returns the training set selected via.
  # stratified random sampling with optimal allocation
  # for a given categorical variable "stratum"
  #   Input
  #     df: a dataframe
  #     stratum: the categorical variable whose levels
  #               are the strata
  #     y: the response covariate name
  #     n: the desired size of the training set
  #     seed: the seed for the random number generator
  #
  set.seed(seed) # set seed
  N <- length(df[,stratum]) #population size

  #initialize vectors
  vars <- c(rep (NA, length(unique(df[stratum]))))
  Wh <- c(rep(NA, length(unique(df[stratum]))))
  nh <- c(rep(NA, length(unique(df[stratum]))))
  counter <- 1

  #group by strata
  for (x in split(birth.data, birth.data[stratum])) {
```

```
    if (is.na(var(x[1][,y]))) {
      vars[counter] <- 0
    } else {
      vars[counter] <- var(x[1][,y])
    }
    Wh[counter] <- length(x[1][,y])/N
    nh[counter] <- sqrt(vars[counter])*(length(x[1][,y])/N)
    if(length(x[1]) > 0 && nh[counter] ==0 ){
      nh[counter] <- 1
    }
    counter <- counter + 1
  }

  #calculate nh
  den <- sum(nh)
  nh <- round(nh/den * n)

  #stratified sampling
  counter <- 1
  train.inds = c(NA)
  for (x in split(birth.data, birth.data[stratum])) {
    train.inds = c(train.inds, sample(as.numeric(rownames(x)), nh[counter]))
    counter <- counter + 1
  }

  return(train.inds)
}
```

Automatic Model Selection

```
# Selection of Candidate Models (Automatic and Manual)

#source("setup.R")
set.seed(6024)
MSPE <- function(M, train) {
  print(M$call)
  print(sum((birth.data$wt[-train] - predict(M, newdata = birth.data[-train,]))^2))
}
```

```
#Initial Models
M0 <- lm(wt ~ 1, data = birth.data, subset = train.ind)
Mmax <- lm(wt ~ gestation + parity + time + number+ smoke + mage + mwt + mht
           + I(mwt*703/(mht)^2) + I(gestation*parity) + I(gestation*mage)
           + income + meth, data=birth.data, subset = train.ind)
Mstart <- lm(wt ~ gestation + mage + I(mwt*703/(mht)^2) + smoke + mht + mwt
             + income, data=birth.data, subset = train.ind)
ntot <- dim(birth.data)[1]
ntrain <- num.train
# forward selection
Mfwd <- step(object = M0, # starting point model
scope = list(lower = M0, upper = Mmax), # smallest and largest model
```

```
direction = "forward",
trace = FALSE) # trace prints out information
# backward selection
Mback <- step(object = Mmax, # starting point model
scope = list(lower = M0, upper = Mmax),
direction = "backward", trace = FALSE)
# stepwise selection (both directions)
Mstep <- step(object = Mstart,
scope = list(lower = M0, upper = Mmax),
direction = "both", trace = FALSE)
```

```
#MSPE(Mfwd, train.ind)
#MSPE(Mstep, train.ind)
#MSPE(Mback, train.ind)
```

```
summary(Mstep)$coefficients[non.cat.inds,]
```

```
#smoke
Mstep.smoke.red <- lm(formula = wt ~ mage + I(mwt * 703/(mht)^2)+ mht
                  + meth + I(gestation * mage) + I(gestation * parity),
                  data= birth.data, subset = train.ind)
```

```
anova(Mstep.smoke.red, Mstep)
```

```
#mother's ethnicity
Mstep.meth.red <- lm(formula = wt ~ mage + I(mwt * 703/(mht)^2)+ mht
                  + smoke + I(gestation * mage) + I(gestation * parity),
                  data= birth.data, subset = train.ind)
```

```
anova(Mstep.meth.red, Mstep)
```

```
Mstep.new <- lm(formula = wt ~ mage + I(mwt * 703/(mht)^2) + smoke + mht +
meth + I(gestation * mage) + I(gestation * parity) + parity, data = birth.data,
subset = train.ind)
non.cat.inds <- c(14, 15)
```

```
summary(Mstep.new)$coefficients[non.cat.inds,]
```

```
non.cat.inds <- c(1, 2, 3, 7, 8)
```

```
M.manual <- lm(wt ~ gestation + mht + smoke + I(gestation*parity) + parity + number
               + income, data = birth.data, subset = train.ind)
summary(M.manual)$coefficients[non.cat.inds, ]
```

```
#smoke
M.manual.smoke.red <-  lm(wt ~ gestation + mht  + I(gestation*parity) + parity +
number + income, data = birth.data, subset=train.ind)
```

```r
anova(M.manual.smoke.red, M.manual)
```

```r
#number
M.manual.number.red <-  lm(wt ~ gestation + mht +smoke + I(gestation*parity) + parity +
income, data = birth.data, subset=train.ind)
```

```r
anova(M.manual.number.red, M.manual)
```

```r
#income
M.manual.income.red <-  lm(wt ~ gestation + mht +smoke + I(gestation*parity) +
  parity + number, data = birth.data, subset=train.ind)
anova(M.manual.income.red, M.manual)
M.2 <- lm(wt ~ gestation + mht + smoke + I(gestation*parity) + parity, data = birth.data,
subset = train.ind)
summary(M.2)$coefficients[non.cat.inds,]
```

```r
M.2.smoke.red <-  lm(wt ~ gestation + mht  + I(gestation*parity) + parity,
                     data = birth.data, subset=train.ind)
anova(M.2.smoke.red, M.2)
Mstep$call
M.2$call
```

Model Comparison

```r
# Automatic
M.auto <- lm(formula = wt ~ mage + I(mwt * 703/(mht)^2) + smoke + mht + meth +
I(gestation * mage) + I(gestation * parity), data = birth.data,
subset = train.ind)
#Manual Selection
M.manual <- lm(formula = wt ~ gestation + mht + smoke + I(gestation * parity)
+ parity, data = birth.data, subset = train.ind)
zres <- residuals(M.auto)
sig.hat <- sqrt(sum(zres^2)/(length(zres)-2))
zres <- zres/sig.hat
```

Residual Plots

Model 1 (Selected via. Automated Model Selection)

```r
#residuals vs. fitted values
par(mfrow = c(1,2))
#plot(predict(M.auto), residuals(M.auto), main="Residuals vs. Fitted Values")
#abline(h = 0, col = "red", lty = 2) #horizontal line
#standardize residuals
#qqnorm(zres, main = "QQ-Plot")
#qqline(zres, col='red', lty = 2)
```

Model 2 (Selected via. Automated Model Selection)

```
zres <- residuals(M.manual)
sig.hat <- sqrt(sum(zres^2)/(length(zres)-2))
zres <- zres/sig.hat
```

```
#residuals vs. fitted values
#par(mfrow = c(1,2))
#plot(predict(M.manual), residuals(M.manual), main="Residuals vs. Fitted Values")
#abline(h = 0, col = "red", lty = 2) #horizontal line
#standardize residuals
#qqnorm(zres, main = "QQ-Plot")
#qqline(zres, col='red', lty = 2)
```

K fold cross validation

```
set.seed(2)
require(caret)
```

```
## Loading required package: caret
```

```
## Loading required package: ggplot2
```

```
#We use the caret library to perform K fold cross validation
#Create the indices for the k folds, with 2/3rd of the data #being used as the training set
ind= createDataPartition(birth.data$wt, p = 2/3, list = FALSE )
#Using the train.ind from the above function, we can calculate the training and testing set
trainDF <- birth.data[ind, ]
testDF <- birth.data[-ind, ]
#The ControlParameters specify the cross fold validation method for 5 folds
ControlParameters <- trainControl(method = "cv",
number = 10, savePredictions = TRUE, classProbs = TRUE)
#Cross Validation for the first model, which was selected through automatic model selection.
#Training the model on the training set using the specified control parameters
M.auto_train <- train(wt ~ parity + time + mage
+ mht + I(mwt * 703/(mht)^2) + I(gestation * mage) +  meth,
data = trainDF, method = "lm",trControl  = ControlParameters, na.action = na.omit)
#Predictions on the test data from the automatic model
M.auto_predictions <- predict(M.auto_train, testDF)
#Observed values of weight for the test data indices
observed_data_test <-  birth.data[-ind,]$wt
#Calculate the MPSE
M.auto_mspe <- sum((M.auto_predictions - observed_data_test)^2)
#For Manual model
#Cross Validation for the second model, which was manually selected
#We can use the same control parameters
#Training the model on the training set using the specified control parameters
M.manual_train <- train(wt ~ gestation + mht + smoke + I(gestation*parity) +
parity + number +income, data = trainDF,
```

```
method = "lm",trControl  = ControlParameters, na.action = na.omit)
#Predictions on the test data from the manual data
M.manual_predictions <- predict(M.manual_train, testDF)
#Calculate the MPSE
M.manual_mspe <- sum((M.manual_predictions - observed_data_test)^2)
#Display nicely
signif(c(Auto_model = M.auto_mspe, Manual_model = M.manual_mspe))
```

Leverege and Influence measures

Model 1

```
Leverage <- hat(model.matrix(M.auto))
h <- hatvalues(M.auto)
n<-nobs(M.auto)
#cook's distance vs. leverage
D <- cooks.distance(M.auto)
infl.ind <- which.max(D)
hbar <- length(coef(M.auto))/n
lev.ind <- h > 2*hbar
clrs <- rep("black", len=n)
clrs[lev.ind] <- "blue"
clrs[infl.ind] <- "red"
par(mfrow = c(1, 1))
cex <- .8
```

```
#plot(h, D, xlab = "Leverage", ylab="Cook's Influence Measure", pch=21, bg=clrs,
 #    cex=cex, cex.axis = cex)
#abline(v=2*hbar, col="grey", lty=2)
#legend("topleft", legend=c("High Leverage", "High Influence"), pch = 21,
#pt.bg = c("blue", "red"), cex=cex, pt.cex = cex)
```

```
print(birth.data[infl.ind,])
```

Model 2

```
Leverage <- hat(model.matrix(M.manual))
h <- hatvalues(M.manual)
n<-nobs(M.manual)
#cook's distance vs. leverage
D <- cooks.distance(M.manual)
infl.ind <- which.max(D)
hbar <- length(coef(M.manual))/n
lev.ind <- h > 2*hbar
clrs <- rep("black", len=n)
clrs[lev.ind] <- "blue"
clrs[infl.ind] <- "red"
par(mfrow = c(1, 1))
cex <- .8
```

```
#plot(h, D, xlab = "Leverage", ylab="Cook's Influence Measure", pch=21, bg=clrs, #cex=cex, cex.axis = c
#abline(v=2*hbar, col="grey", lty=2)
```

```
print(birth.data[infl.ind,])
```

AIC

```
# models to compare
M1 <- M.auto
M2 <- M.manual
AIC(M1)
AIC(M2)
```

PRESS STATISTIC

```
# models to compare
M1 <- M.auto
M2 <- M.manual
# PRESS statistics
press1 <- resid(M1)/(1-hatvalues(M1)) # M1
press2 <- resid(M2)/(1-hatvalues(M2)) # M2
# plot PRESS statistics
#boxplot(x = list(abs(press1), abs(press2)), names = c("Automatic", "Manual")
#,ylab = expression(group("|", PRESS[i], "|")),
#col = c("yellow", "orange"))
```