University of Waterloo

Final Project Report

STAT 331 Fall 2018

Group 91:

Gengyao Yuan(20613017)

Shunyu Zhao(20699637)

Contents

1 Summary **2**

2 Model Selection **2**
- 2.1 Brief data overview . . . 2
- 2.2 Transform categorical predictors . . . 2
- 2.3 Drop off NAs and prediction missing data . . . 2
- 2.4 Revisiting the category variables . . . 3
- 2.5 visually inspect data . . . 3
- 2.6 Automated Model Selection . . . 4
  - 2.6.1 min & max model set up . . . 4
  - 2.6.2 display covariates in each model . . . 5
  - 2.6.3 qqplot for residual distribution . . . 6
  - 2.6.4 Press AIC and R^2 . . . 6
- 2.7 Manual Model . . . 7

3 Model Diagnostics **8**
- 3.1 Different types of residual plots . . . 9
- 3.2 Residuals studentlized residuals and standlized residuals . . . 9
- 3.3 PRESS Residuals . . . 9
- 3.4 DFFITS Residuals . . . 9
- 3.5 Comparison of different residual plots . . . 11
- 3.6 Leverage and influence measures . . . 13
- 3.7 outlier . . . 14
- 3.8 Cross-validation . . . 14

4 Discussion **19**
- 4.1 What are the most important factors associated with/influencing birth weight? . . . 19
- 4.2 Low birth weight is considered to be 88 ounces or less. Based on this analysis, would you be able to recommend behavioral changes to parents in order to avoid low birthweight? If so, please carefully formulate your recommendation. . . . 19
- 4.3 Are there any coecients with high p-values retained in the final model? If so, why? . . . 19
- 4.4 Are any of the regression assumptions of the final model violated? If so, which ones? . . . 20
- 4.5 What are the possible deficiencies of the final model? how do these deficienciesnuance your conclusions/recommendations above? . . . 20

5 Appendix **20**

1 Summary

The goal of the STAT 331 final project is to explore the relation of healthy male single-fetus birth weight and some explanatory variables. This report will be divided into 4 main sections:
Summary, which will cover the main purpose of the report and give a brief explanation of how the project will analyze the data. Two candidate models will be produced in the model selection section by using the pre-fitting data diagnostic and automated model selection. Model diagnostics section will perform an in-depth comparison of the two candidates models by comparing different types of residual plots, leverage and influence measures and cross-validation(rPMSE). In the end, there will be a discussion section basing on the result of the most likely linear model we get from the previous sections to talk about several topics such like: ???what is the most important factors associated with/influencing birth weight' After using serial statistical analysis way that we learned in STAT 331 course, we find the fianl model is like lm(formula = wt ~ gestation + parity + meth + mage + mht + mwt + fht + fwt + income + time + number + gestation:income + mwt:income + gestation:mage + gestation:fwt + gestation:mwt + gestation:fht + gestation:mht + fht:income, data = births_clean)

2 Model Selection

2.1 Brief data overview

By view the summary of the data, we notice that there are several illegal data. The domain of "marital"(the mother's marital status) is 1 to 5, but it is clearly showing that there exist 0 in "marital," we replace all the 0 to NA since it is not available data(out of range).
For the categorical predictor "meth"(The self-reported ethnicity of the mother) and "feth"(The self-reported ethnicity of the father), all 0 to 5 is Caucasian meaning that they are in the same group, so we replace the 0-5(Caucasian) to 0, 6(Mexican) to 1, 7(African-American) to 2, 8(Asian) to 3, 9(Mixed) to 4, 10(Other) to 5

2.2 Transform categorical predictors

Since all the categorical predictors should not be treated as continuous variables, although they may look like continuous variables (such like 0,1,2,3,4....), we use "one-hot" encoding scheme to make new factors for them. For example, 'med' means the mother's education, whose domain is 0 to 7, where '0' level means 'elementary school' level, level '1' means 'middle school' level, level '2' means 'high school' .... etc, we just transfer numbers to factors with the same name(for example, number 1 to NEW factor '1'), after successfully transfer all the levels to new factors, dropping all the '0' levels for all categorical predictors by the requirement of one-hot"encoding scheme(we can do that because all predictors have '0' level(meth and feth didn't have, but we already transform them)), and give all the new factors a 0/1 binary variable to show that factor is applied to this data or not.
There is a trick in R code:"as.factor" function, it can automatically transfer variables to new factors, so we applied it on all the categorical predictors(meth, med,feth, fed, marital, smoke, time, number) to factor type instead of continuous variables.

2.3 Drop off NAs and prediction missing data

From the summary report in the previous section, there are lots of NA data points in our data frame. Several methods have been covered in STAT 331 to produce missing data points. However, we use MICE here.
MICE can help us to impute missing values which are drawn from a distribution specifically designed for each missing data points. Don't like replace all NA variables by mean of the data; MICE can also include the 'var' in the data prediction, which can help lessen the bias and make the data close to the original.
Since MCIE is 'prediction' function, thus every time we run this may cause different results, to avoid this, we

always set the seed as 1. And to get a closer result similar to the original data, we set the method type of MICE to 'sample,' which means sample any Random sample from observed values.

The result of anyNA is FALSE shows there is no more NA value in our births_mice data frame.
Since there is no +-INF's data basing on our summary, all the data in data frame now is available and meaningful.

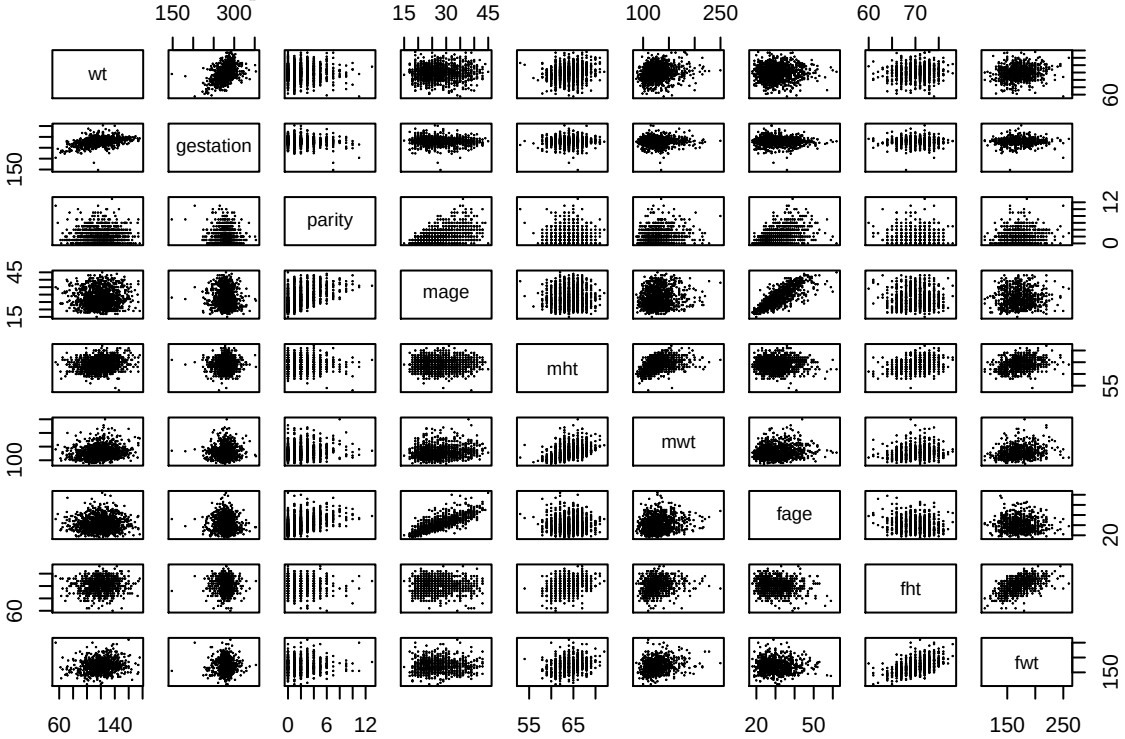
2.4 Revisiting the category variables

After predicting the missing NA data points, it is necessary to revisit categorial variables and factorize the levels of categorical variables into meaningful names; this can be helpful when dueling with interpretation effect of the model.
We also shrink the number of levels for some factors because those levels are significant minorities:
meth & feth: keep 0 as Caucasian, 2 as African-American, shrink all other to other(this change based on the previous shrunk result.) med & fed: keep 1 as middle school, 2 as high school, 3 as high school+trade school, 5 as a college graduate, shrink all other to other marital: keep 1 as married, shrink all the others to other time: keep 0 as never smoke, 1 as still smokes, shrink all the others to other number: keep 0 as never smoked, 1 as (smoke) 1-4 (per day), 2 as 5-9,5 as 20-29, shrink all the others to othera

During we shrinking the varibles, we find that the 'smoke' factor should be exaclty same as the 'time smoke' factor, so we just delete the factor'smoke'

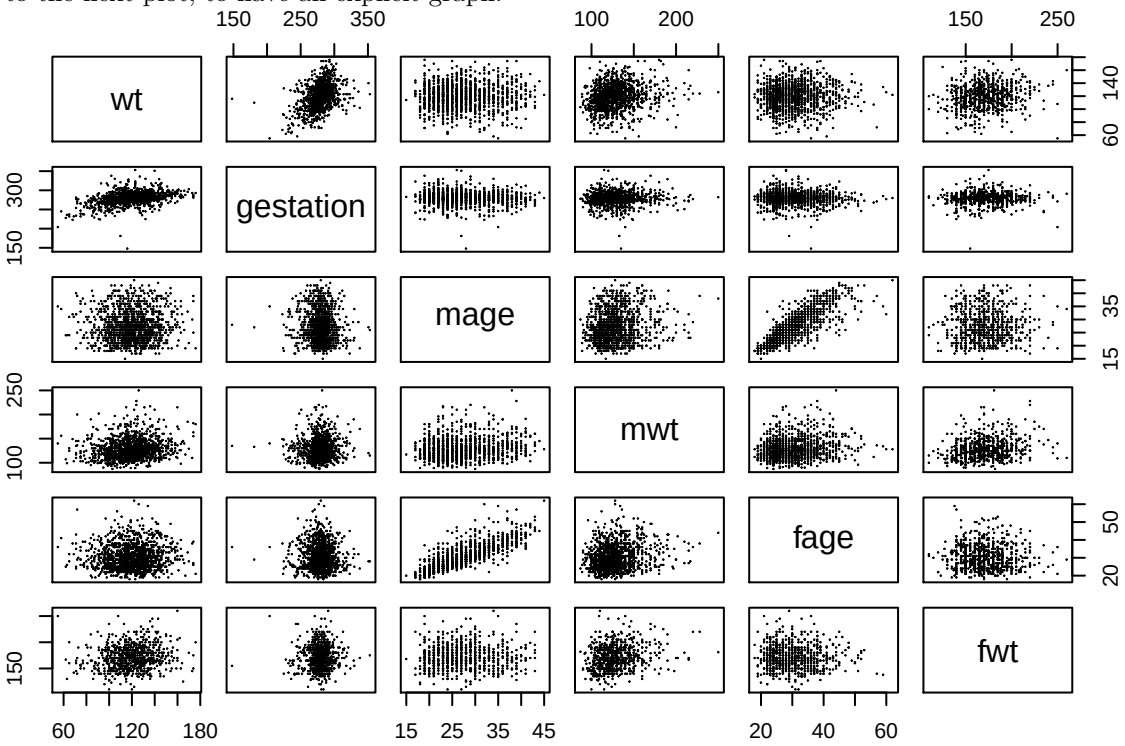
2.5 visually inspect data

By drawing out the pairs plot of the original data, we can have a basic visually inspect to estimation is there is a linear relationship between the variables.



As the reason that we only what to find out if there are linear relationships, we only include 'continuous varibles'(basing on the data definition of the project question) into our graph. The result shows that all the 'continuous varibles' continuously somehow so we don't need to treat any of them as categorical.

Since there are too many variables, we pike up some significant variables that may have a linear relationship to the next plot, to have an explicit graph.



The only clear linear relationship is between wt and gestation. All the other variables should be further discussed.

2.6 Automated Model Selection

2.6.1 min & max model set up

It is necessary to set up a minimum model and maximum model before using the automatic selections. Firstly, we set up the M0 as min with only interaction and Mmax as the maximum that all variables have interactions with each other.

Since the NA chart shows it is obviose most of the coeficients have NA interactions with marital, fed, feth, number, time, meth, med, thus we delete all the interaction with them in our max in order to have a smaller model. We don't add any quadratic terms because basing on the previous pairs plot, there is no graph seems have quadratic relationship.

```
Mmax <- lm(wt ~ (. -marital -fed -feth -number -time - meth -med)^2
           +marital + fed +feth +number +time + meth +med , data = births_clean) # Revised max model
Mstart <- lm(wt ~ ., data = births_clean) #start model
anyNA(coef(Mmax)) # detect coefficients which are NAs
```

```
## [1] FALSE
```

The output of anyNA is FALSE. It shows the Mmax is the minimum model which including as many possible interactions can but can also avoid all NA here.

2.6.2 display covariates in each model

```
# Forward selection
invisible(Mfwd <- step(object = M0,
               scope = list(lower = M0, upper = Mmax),
               direction = "forward", trace = FALSE))
```

```
# Backward elimination selection
Mback <- step(object = Mmax,
              scope = list(lower = M0, upper = Mmax),
              direction = "backward", trace = FALSE)
```

```
# Stepwise selection
Mstep <- step(object = Mstart,
              scope = list(lower = M0, upper = Mmax),
              direction = "both", trace = FALSE)
```

```
##  fwd back step
##   20   33   25
```

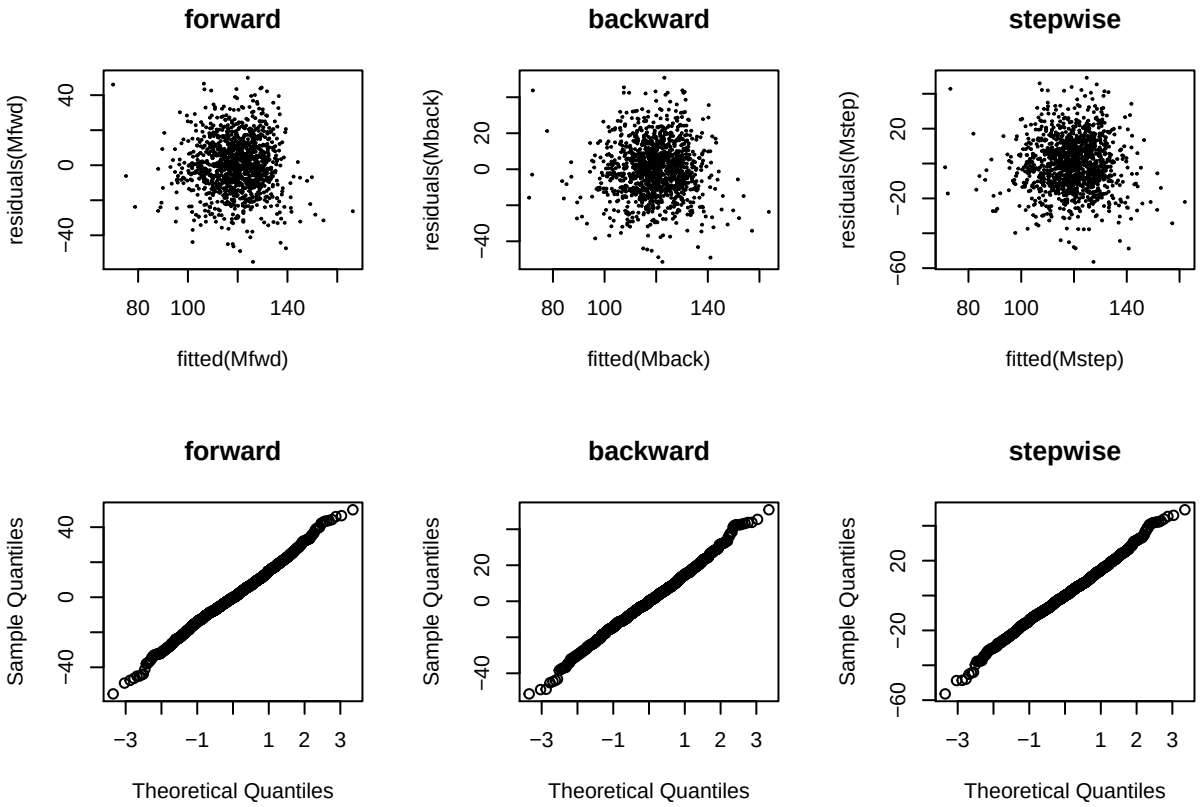
```
## lm(formula = wt ~ gestation + time + mht + meth + parity + number +
##     fwt + mwt + fht + gestation:fwt + gestation:mwt + mht:fwt +
##     gestation:mht + gestation:fht, data = births_clean)
```

```
## lm(formula = wt ~ gestation + parity + mage + mht + mwt + fage +
##     fht + fwt + income + number + time + meth + gestation:mage +
##     gestation:mht + gestation:mwt + gestation:fht + gestation:fwt +
##     gestation:income + parity:mht + mht:fage + mht:fht + mht:fwt +
##     mht:income + mwt:income + fage:fwt + fage:income + fht:income,
##     data = births_clean)
```

```
## lm(formula = wt ~ gestation + parity + meth + mage + mht + mwt +
##     fht + fwt + income + time + number + gestation:income + mwt:income +
##     gestation:mage + gestation:fwt + gestation:mwt + gestation:fht +
##     gestation:mht + fht:income, data = births_clean)
```

The first line of output is the number of parameters for the three models. Parameter for every models are showing after that, following the order: forward selection, backward elimination, and stepwise selection.

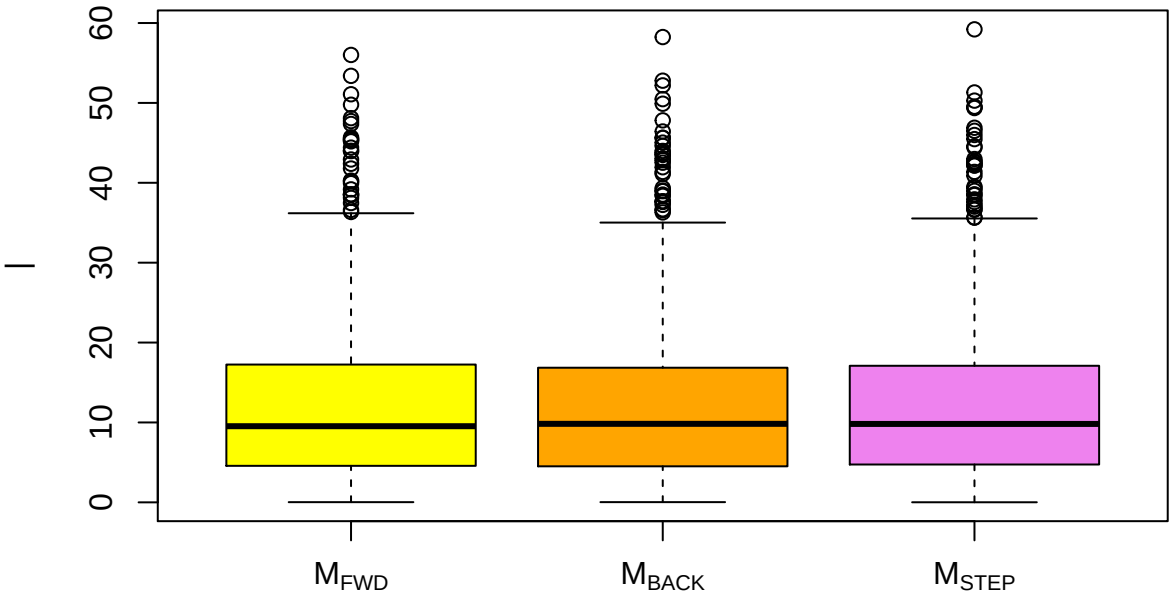
2.6.3 qqplot for residual distribution



From the Residual vs Fitted plot, it is clearly showing that points are randomly distributed around the 0 line. And the points in QQ-plot almost line on the diagonal line. Both of the two graphs prove that the residual distribution follow normal distribution.

2.6.4 Press AIC and R^2

```
##                    M[FWD]      M[BACK]      M[STEP]
## AIC           1.028290e+04 1.026394e+04 1.026355e+04
## PRESS         2.986716e+05 2.955627e+05 2.942003e+05
## R_Squared     3.012018e-01 3.261652e-01 3.176000e-01
## R_adj_Squared 2.902831e-01 3.082411e-01 3.040760e-01
```



Since the press statics equal to the following equation

$$PRESS_i = y_i - y_i = e_i/1 - h_i$$

The 'e' here is the residual error, which means the sum of i PRESSi is the total residual error of the model. Thus the model with lest PRESS is the best model.
Akaike Information Criterion(AIC) equal to

$$AIC = n(1 + log(e^{'}e/n) + log(2pi)) + 2(p + 1)$$

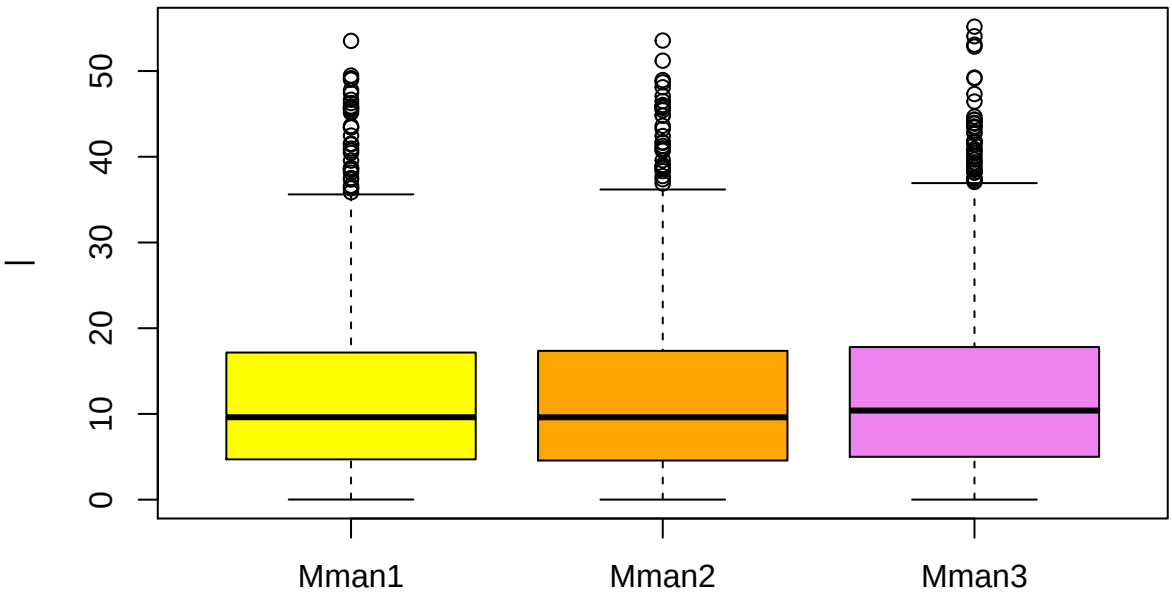
The less AIC means better model with less error because the AIC has the same monotony with 'e'. And by the defination of residual error, R^2 is also less is better. According to the result we get above, Mback(model get by backward elimination) is the best model by automated model section.

2.7 Manual Model

We got 3 manual models from above.The manual model 1 comes from the backward elimination selection. The manual model 2 comes from the stepwise selection. The manual model 3 comes from the pairplots.

```
Mman1 <- lm(formula = wt ~ gestation + parity + meth + mage + mht + mwt +
    fht + fwt + income + time + number,  data = births_clean)
Mman2 <- lm(formula = wt ~ gestation + parity + mage + mht + mwt + fage +
    fht + fwt + income + number + time + meth,  data = births_clean)
Mman3 <- lm(formula = wt ~ gestation + mage + fage  + time + number + meth + feth,
            data = births_clean)
```

```
##                      Mman1        Mman2        Mman3
## AIC           1.029669e+04 1.029822e+04 1.038430e+04
## PRESS         3.006585e+05 3.010658e+05 3.221790e+05
## R_Squared     2.899263e-01 2.901968e-01 2.340586e-01
## R_adj_Squared 2.806062e-01 2.802899e-01 2.259103e-01
```



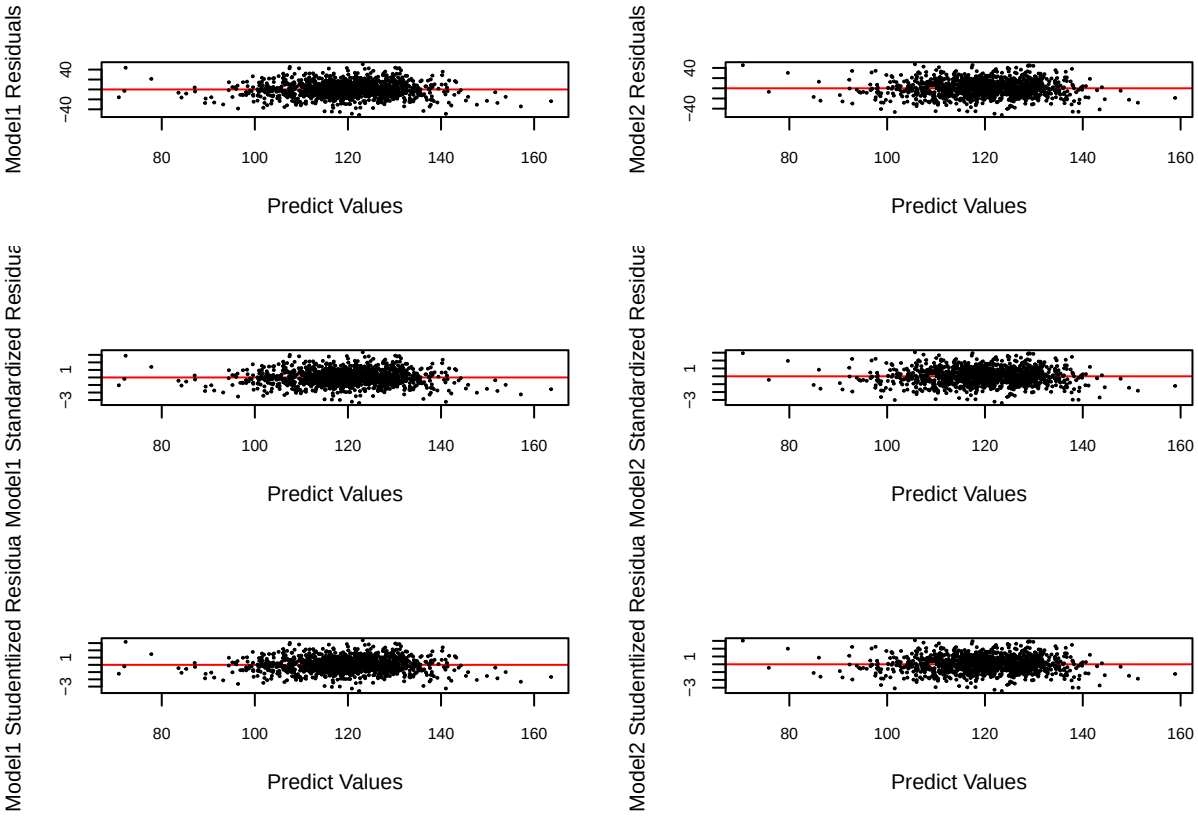
Therefore, by the comparison of AIC, PRESS, R^2 and adjusted R^2, we choose the manual model 1 as our candidate model.

3 Model Diagnostics

Our candidate models are manual model2 and stepwise model. In this section, we will discuss the different residual plots, leverage and influence measure, outlier and cross-validation in order to find a model which is better.

3.1 Different types of residual plots

3.2 Residuals studentlized residuals and standlized residuals



First, we compared the different residual plots respectively. From the residuals vs predict values, standardized residuals vs predicts values, and studentized residuals vs predicts values plots, we can figure out that. First of all, their means are 0. What is more, they do not have any linear trends so they are independent. Also, generally, they all have constant variances because the error variances do not change a lot (no increasing and decreasing trends)

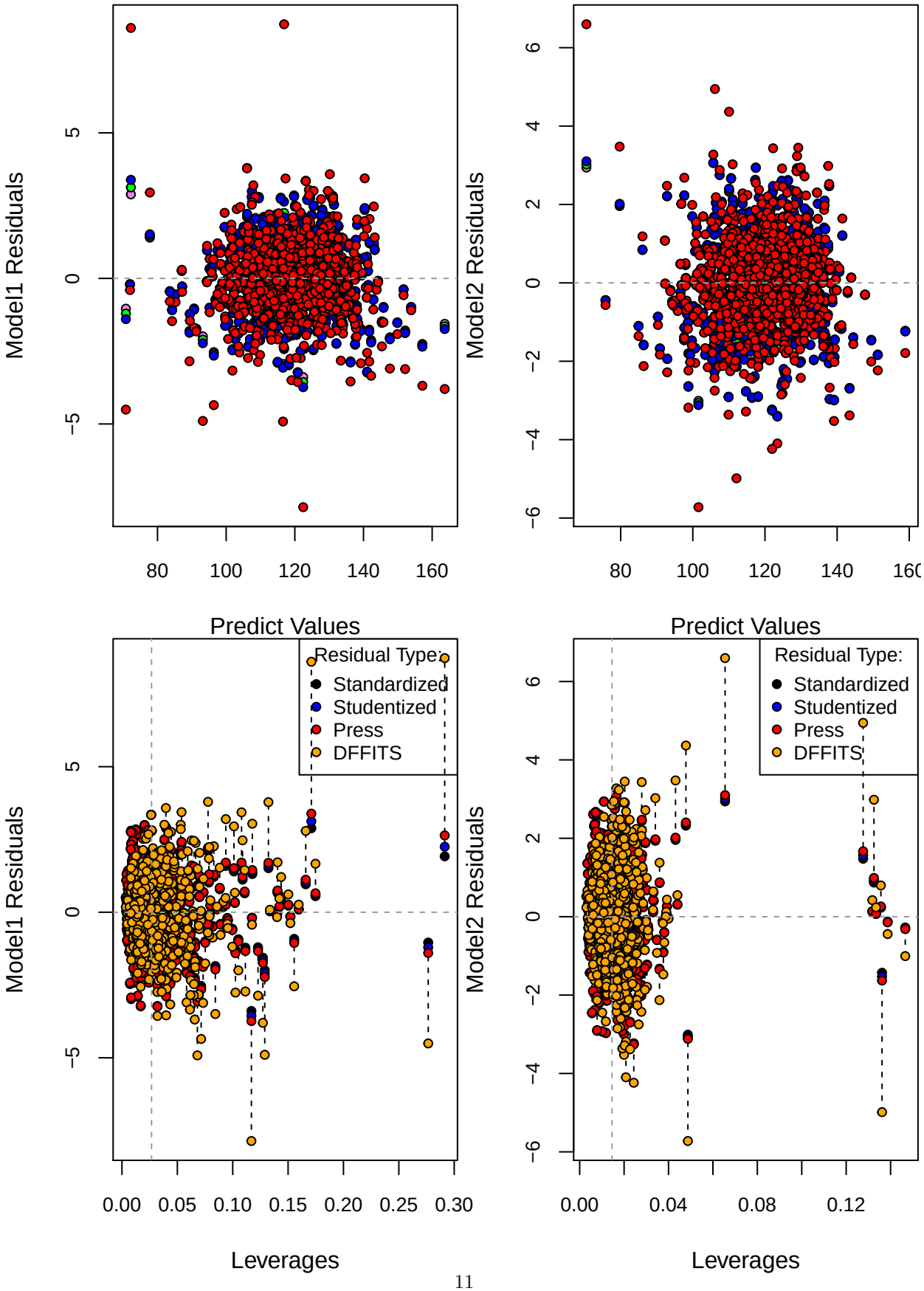
3.3 PRESS Residuals

```
press_model1 <- Re1/(1 - hatvalues(Model1)) #press for model1
press_model2 <- Re2/(1 - hatvalues(Model2)) #press for model2
```

3.4 DFFITS Residuals

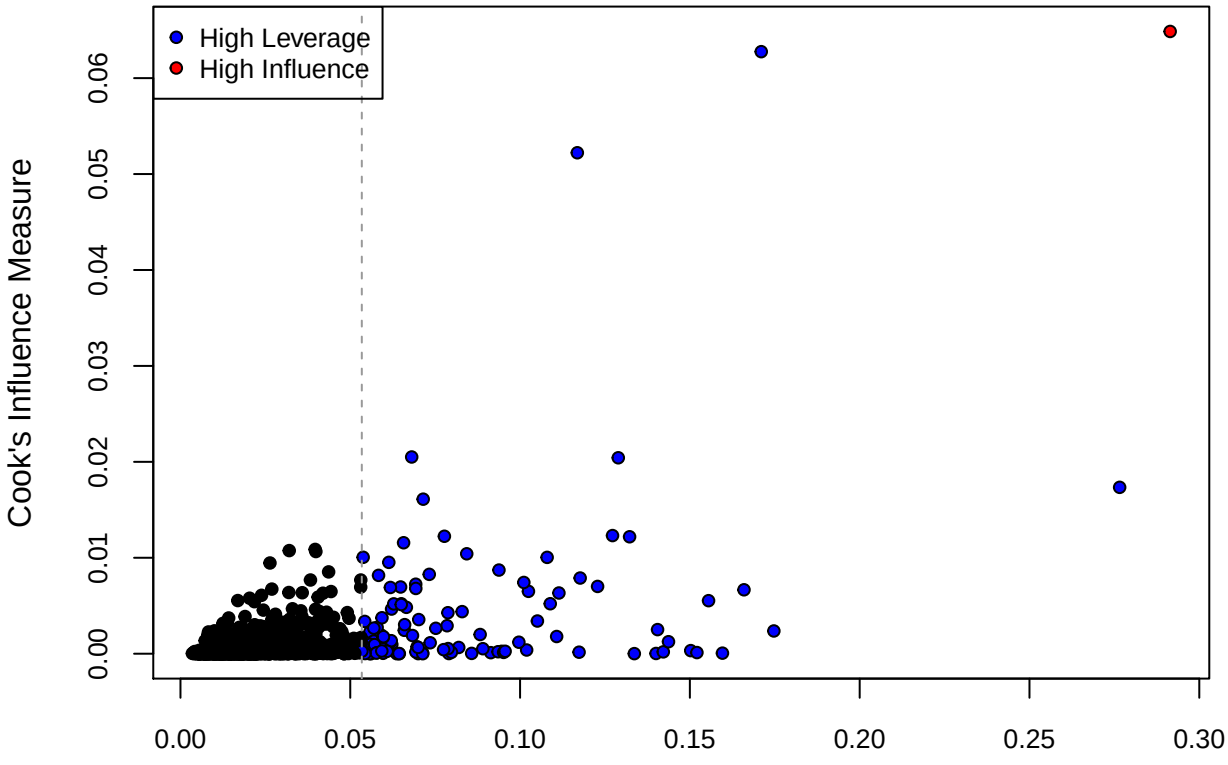
```
dfts1 <- dffits(Model1) #DFFITS for model1
dfts2 <- dffits(Model2) #DFFITS for model2
```

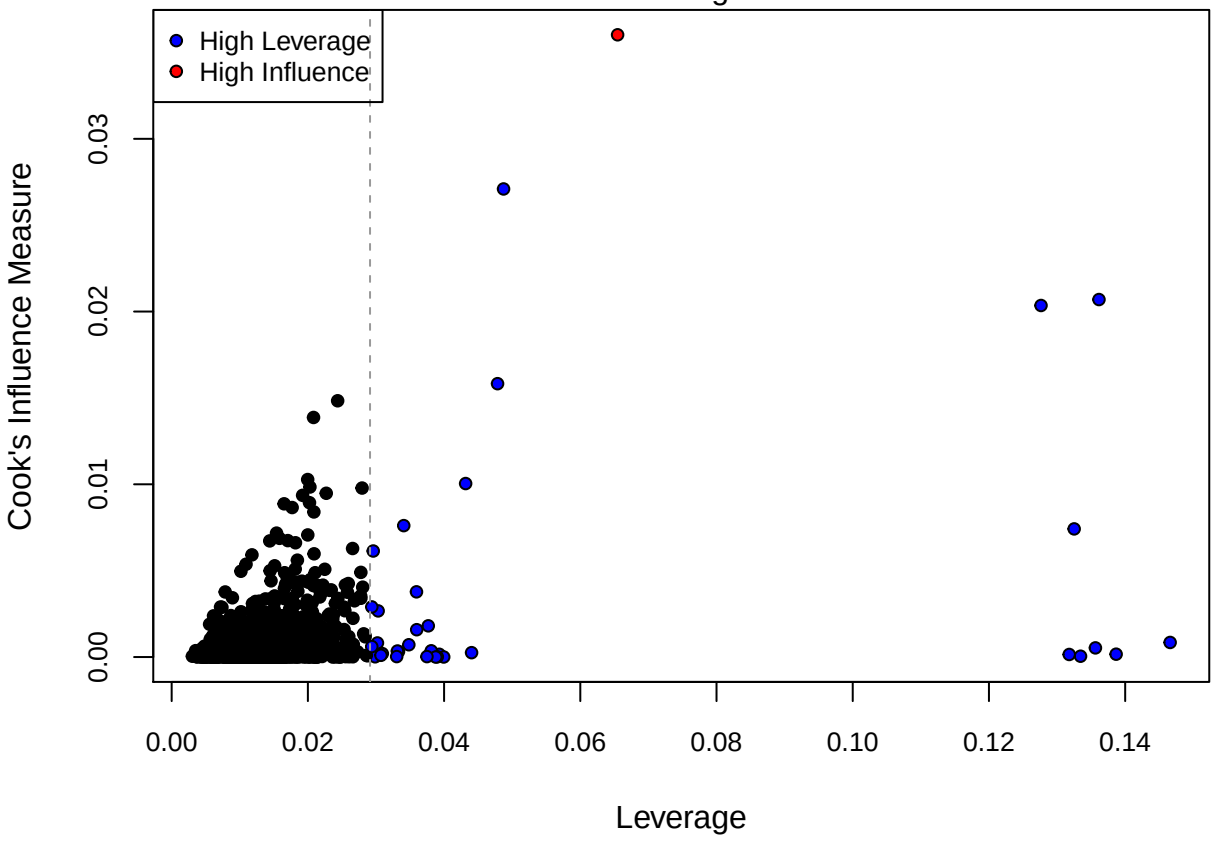
3.5 Comparison of different residual plots



And then, we tried to combine such kinds of residuals in one plot to find the difference between them. In this part, we compared standardized residuals, studentized residuals, press residual and DFFITS. Since they have different scopes, so we make them identical at the average leverage value level hbar = p/n.

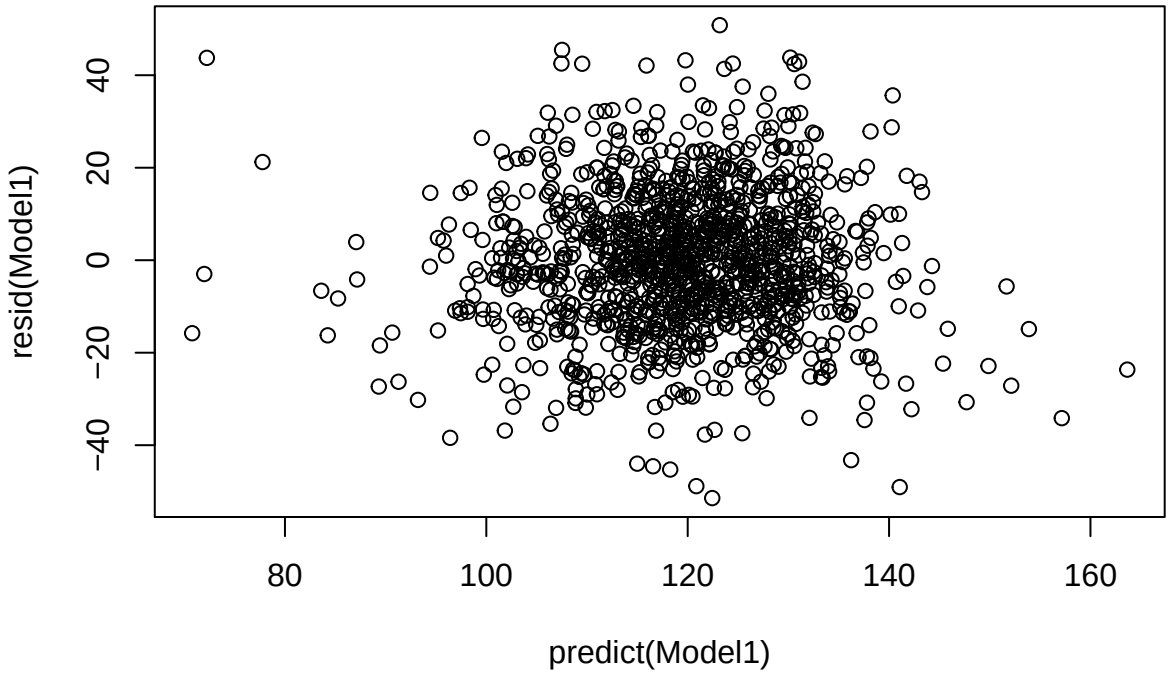
3.6 Leverage and influence measures





In order to figure out the overall influence on cook???s distance against leverage plots, we saw that in Model1 the high influence is as twice as the high leverage, so maybe Model1 has outliers

3.7 outlier

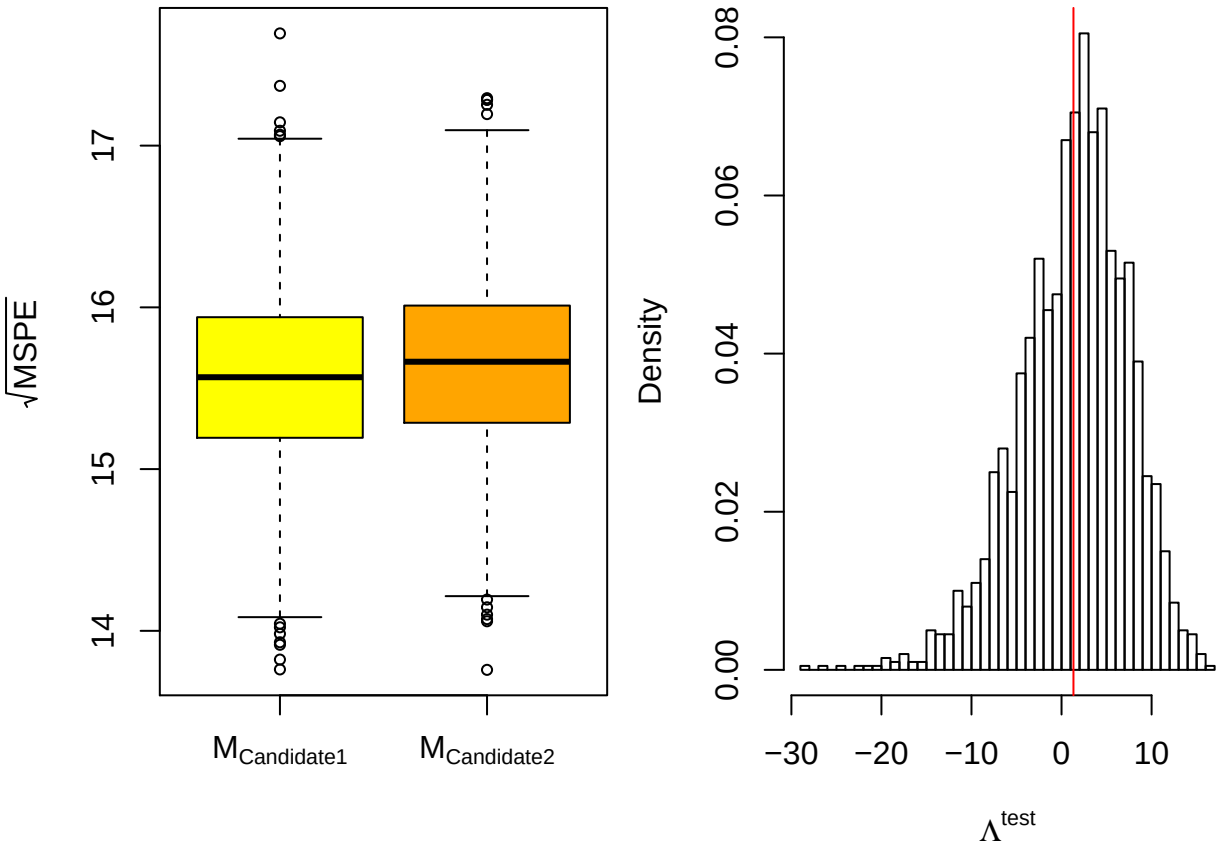



```
## 530
## 530
```

3.8 Cross-validation

```
M1 <- Model1
M2 <- Model2
Mnames_new <- expression(M[Candidate1],M[Candidate2])
# Cross-validation setup
nreps <- 2e3 # number of replications
ntot <- nrow(births_clean) # total number of observations
ntrain <- floor(0.7 * ntot) # size of training set
ntest <- ntot-ntrain # size of test set
mspe1 <- rep(NA, nreps) # sum-of-square errors for each CV replication
mspe2 <- rep(NA, nreps)
logLambda <- rep(NA, nreps) # log-likelihod ratio statistic for each replication
for(ii in 1:nreps) {
  # randomly select training observations
  train.ind <- sample(ntot, ntrain) # training observations
  # refit the models on the subset of training data; ?update for details!
  M1.cv <- update(M1, subset = train.ind)
  M2.cv <- update(M2, subset = train.ind)
  # out-of-sample residuals for both models
  # that is, testing data - predictions with training parameters
  M1.res <- births_clean$wt[-train.ind] -
```

```
            predict(M1.cv, newdata = births_clean[-train.ind,])
  M2.res <- births_clean$wt[-train.ind] -
            predict(M2.cv, newdata = births_clean[-train.ind,])
  # mean-square prediction errors
  mspe1[ii] <- mean(M1.res^2)
  mspe2[ii] <- mean(M2.res^2)
  # out-of-sample likelihood ratio
  M1.sigma <- sqrt(sum(resid(M1.cv)^2)/ntrain) # MLE of sigma
  M2.sigma <- sqrt(sum(resid(M2.cv)^2)/ntrain)
  # since res = y - pred, dnorm(y, pred, sd) = dnorm(res, 0, sd)
  logLambda[ii] <- sum(dnorm(M1.res, mean = 0, sd = M1.sigma, log = TRUE))
  logLambda[ii] <- logLambda[ii] -
                   sum(dnorm(M2.res, mean = 0, sd = M2.sigma, log = TRUE))
}
# plot rMSPE and out-of-sample log
par(mfrow = c(1,2))
par(mar = c(4.5, 4.5, .1, .1))
boxplot(x = list(sqrt(mspe1), sqrt(mspe2)),
        names = Mnames_new, cex = .7,
        ylab = expression(sqrt(MSPE)), col = c("yellow", "orange"))
hist(logLambda, breaks = 50, freq = FALSE,
     xlab = expression(Lambda^{test}),
     main = "", cex = .7)
abline(v = mean(logLambda), col = "red") # average value
```



In this part, we compared the boxplots for root mean square prediction error (rPMSE). According to the graph, it is not difficult to distinguish that the Model1 has a better preference in the comparison result,

because the box on left-side is much lower than the right-side box. Also, by the out-of-sample likelihood ratio statistic plot, we found that the Model1 is better.

Overall, we can say that the final model we select is model1, which is stepwise selection model.

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	4.397e+02	3.699e+02	0.23480
gestation	2.406e-04	9.486e-01	0.99980
parity	1.270e+01	6.172e+00	0.03979
mage	-3.788e+00	1.490e+00	0.01113
mht	-1.137e+01	5.552e+00	0.04071
mwt	1.038e+00	4.197e-01	0.01354
fage	-4.344e+00	2.029e+00	0.03246
fht	-2.234e-01	4.811e+00	0.96297
fwt	-2.695e-01	6.462e-01	0.67670
income	3.886e+00	7.182e+00	0.58850
number1-4	1.114e+01	5.585e+00	0.04634
number5-9	8.652e+00	5.525e+00	0.11761
numberothers	4.413e+00	5.561e+00	0.42764
number20-29	4.667e+00	5.470e+00	0.39369
timestill smokes	-1.496e+01	5.456e+00	0.00619
timeothers	-8.112e+00	5.545e+00	0.14375
methothers	-2.959e-01	1.584e+00	0.85183

```
number20-29           4.667e+00  5.470e+00 0.39369

timestill smokes     -1.496e+01  5.456e+00 0.00619

timeothers           -8.112e+00  5.545e+00 0.14375

methothers           -2.959e-01  1.584e+00 0.85183

methAfrican-American -8.716e+00  1.221e+00 1.62e-12

gestation:mage        1.292e-02  5.271e-03 0.01434

gestation:mht         2.135e-02  1.239e-02 0.08518

gestation:mwt        -3.887e-03  1.495e-03 0.00942

gestation:fht        -2.326e-02  1.081e-02 0.03166

gestation:fwt         4.172e-03  1.308e-03 0.00147

gestation:income      3.418e-02  1.318e-02 0.00960

parity:mht           -1.809e-01  9.656e-02 0.06119

mht:fage              5.925e-02  3.160e-02 0.06101

mht:fht               1.153e-01  6.352e-02 0.06984

mht:fwt              -1.533e-02  8.506e-03 0.07174
```

```
mht:income          -1.398e-01  9.247e-02 0.13075

mwt:income           3.563e-02  1.107e-02 0.00132

fage:fwt             5.021e-03  3.080e-03 0.10328

fage:income         -4.636e-02  2.866e-02 0.10604

fht:income          -1.111e-01  6.753e-02 0.10016
```

4 Discussion

4.1 What are the most important factors associated with/influencing birth weight?

According to the summary of m1, the most important factors are the 'total number of previous pregnancies,' 'African-American mother,' 'the father's weight,' 'the mother's weight,' 'smoke time' and many interactions between 'gestation' with the previous factors. Simply, it seems like the length of gestation, the number of pregnancies times, ethnicity of mother, mother smoke or not and the weight of parents are the most critical factors.

4.2 Low birth weight is considered to be 88 ounces or less. Based on this analysis, would you be able to recommend behavioral changes to parents in order to avoid low birthweight? If so, please carefully formulate your recommendation.

4.3 Are there any coecients with high p-values retained in the final model? If so, why?

YES, the p-value of 'the length of the gestation period' is high, however, its interactions with a total number of previous pregnancies,''African-American mother,' 'the father???s weight,' and 'the mother???s weight,' all has tiny p-values close to 0. This is because 'the length of the gestation period' itself is not important, but its interactions with others are so important for the model.

\subsection{Are there any outlying observations that might be appropriate to remove?

From the fitted values vs. residual values graph in 2.6.3, our final graph is the graph three, since the fitted values are not following any linear relationship, all points are crowded together is acceptable for us. By the definition of outliers, which the points with particular residual means outliers, yes, there are some observations that we need to remove. It is also the same for leverage points for the graph in 2.6.4's boxplot; there are some points with outstanding x, they are leverage points that needed to remove. Since the 2.6.3 graph shows almost all fitted variables follow a line, we can get a conclusion that there is no important variables need to remove.

4.4 Are any of the regression assumptions of the final model violated? If so, which ones?

At first we assume ''the length of the gestation period', 'income' and 'the age' and 'height of parents'are important factors, but basing on the too large p-value, we drop them as last.

4.5 What are the possible deficiencies of the final model? how do these deficienciesnuance your conclusions/recommendations above?

Since there are too many NA data in 'The mother???s height' and 'the total number of previous pregnancies'(over 35%), these NA datapoints may cause serious bias on our model. We try to fix these data points by using MICE founction, if the data of 'The mother???s height' and 'the total number of previous pregnancies' follow normal distribution(the help distribution used in MICE fouction), our prediction will close to original data, otherwise our model will contain some bias.

5 Appendix

```
library(mice)

# data input
chds_births <- read.csv(file = "chds_births.csv")
summary(chds_births) # first look of the data
```

```
##        wt          gestation         parity            meth
##  Min.   : 55.0   Min.   :148.0   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:108.8   1st Qu.:272.0   1st Qu.: 0.000   1st Qu.: 0.000
##  Median :120.0   Median :280.0   Median : 1.000   Median : 3.000
##  Mean   :119.6   Mean   :279.3   Mean   : 1.932   Mean   : 3.129
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.: 3.000   3rd Qu.: 7.000
##  Max.   :176.0   Max.   :353.0   Max.   :13.000   Max.   :10.000
##                  NA's   :13                       NA's   :1
##       mage            med             mht            mwt
##  Min.   :15.00   Min.   :0.000   Min.   :53.00   Min.   : 87.0
##  1st Qu.:23.00   1st Qu.:2.000   1st Qu.:62.00   1st Qu.:114.8
##  Median :26.00   Median :2.000   Median :64.00   Median :125.0
##  Mean   :27.26   Mean   :2.917   Mean   :64.05   Mean   :128.6
##  3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:66.00   3rd Qu.:139.0
##  Max.   :45.00   Max.   :7.000   Max.   :72.00   Max.   :250.0
##  NA's   :2       NA's   :1       NA's   :22      NA's   :36
##       feth            fage            fed             fht
##  Min.   : 0.000   Min.   :18.00   Min.   :0.000   Min.   :60.0
##  1st Qu.: 0.000   1st Qu.:25.00   1st Qu.:2.000   1st Qu.:68.0
##  Median : 3.000   Median :29.00   Median :4.000   Median :71.0
##  Mean   : 3.154   Mean   :30.35   Mean   :3.127   Mean   :70.2
##  3rd Qu.: 7.000   3rd Qu.:34.00   3rd Qu.:5.000   3rd Qu.:72.0
##  Max.   :10.000   Max.   :62.00   Max.   :7.000   Max.   :78.0
##  NA's   :31       NA's   :7       NA's   :13      NA's   :492
##       fwt           marital          income          smoke
##  Min.   :110.0   Min.   :0.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:155.0   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:0.0000
##  Median :170.0   Median :1.000   Median :3.000   Median :1.0000
```

```
##  Mean   :171.2   Mean   :1.038   Mean   :3.701   Mean   :0.8018
##  3rd Qu.:185.0   3rd Qu.:1.000   3rd Qu.:5.000   3rd Qu.:1.0000
##  Max.   :260.0   Max.   :5.000   Max.   :9.000   Max.   :3.0000
##  NA's   :499                     NA's   :124     NA's   :10
##       time            number
##  Min.   :0.0000   Min.   :0.00
##  1st Qu.:0.0000   1st Qu.:0.00
##  Median :1.0000   Median :1.00
##  Mean   :0.9625   Mean   :1.76
##  3rd Qu.:1.0000   3rd Qu.:3.00
##  Max.   :9.0000   Max.   :8.00
##  NA's   :10       NA's   :21
```

```r
# standardize the category variable marital
chds_births$marital[!chds_births$marital %in% c(1:5)] = NA

# standardize the category variable meth
chds_births$meth[chds_births$meth %in% c(0,1,2,3,4,5)] = 0
chds_births$meth[chds_births$meth %in% c(6)] = 1
chds_births$meth[chds_births$meth %in% c(7)] = 2
chds_births$meth[chds_births$meth %in% c(8)] = 3
chds_births$meth[chds_births$meth %in% c(9)] = 4
chds_births$meth[chds_births$meth %in% c(10)] = 5

# standardize the category variable feth
chds_births$feth[chds_births$feth %in% c(0,1,2,3,4,5)] = 0
chds_births$feth[chds_births$feth %in% c(6)] = 1
chds_births$feth[chds_births$feth %in% c(7)] = 2
chds_births$feth[chds_births$feth %in% c(8)] = 3
chds_births$feth[chds_births$feth %in% c(9)] = 4
chds_births$feth[chds_births$feth %in% c(10)] = 5

# change all category variables to factor instead of continues variables
chds_births$meth <- as.factor(chds_births$meth)
chds_births$med <- as.factor(chds_births$med)
chds_births$feth <- as.factor(chds_births$feth)
chds_births$fed <- as.factor(chds_births$fed)
chds_births$marital <- as.factor(chds_births$marital)
chds_births$smoke <- as.factor(chds_births$smoke)
chds_births$time <- as.factor(chds_births$time)
chds_births$number <- as.factor(chds_births$number)

# remove the NA by mice which method is sample
births_mice <- mice(chds_births, method = "sample", seed = 1)
```

```
##
##  iter imp variable
##    1   1  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##    1   2  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##    1   3  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##    1   4  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##    1   5  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##    2   1  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##    2   2  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##    2   3  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
```

```
##   2   4  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   2   5  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   3   1  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   3   2  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   3   3  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   3   4  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   3   5  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   4   1  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   4   2  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   4   3  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   4   4  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   4   5  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   5   1  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   5   2  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   5   3  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   5   4  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
##   5   5  gestation  meth  mage  med  mht  mwt  feth  fage  fed  fht  fwt  marital  income  smoke  ti
```

```
## Warning: Number of logged events: 160
```

```
births_clean <- complete(births_mice)

anyNA(births_clean) #check the NA after removement
```

```
## [1] FALSE
```

```
 # since we operate by levels of the factor, for
 # factor '0' it will be position '1' in levels

# category variable: meth
levels(births_clean$meth)
```

```
## [1] "0" "1" "2" "3" "4" "5"
```

```
levels(births_clean$meth)[1] <- "Caucasian"
levels(births_clean$meth)[3] <- "African-American"
levels(births_clean$meth)[c(2,4,5,6)]<- "others"

# category variable: med
levels(births_clean$med)[2] <- "middle-school"
levels(births_clean$med)[3] <- "high-school"
levels(births_clean$med)[5] <- "high-school + some college"
levels(births_clean$med)[6] <- "college graduate"
levels(births_clean$med)[c(1,4,7,8)]<- "others"

# category variable: feth
levels(births_clean$feth)
```

```
## [1] "0" "1" "2" "3" "4" "5"
```

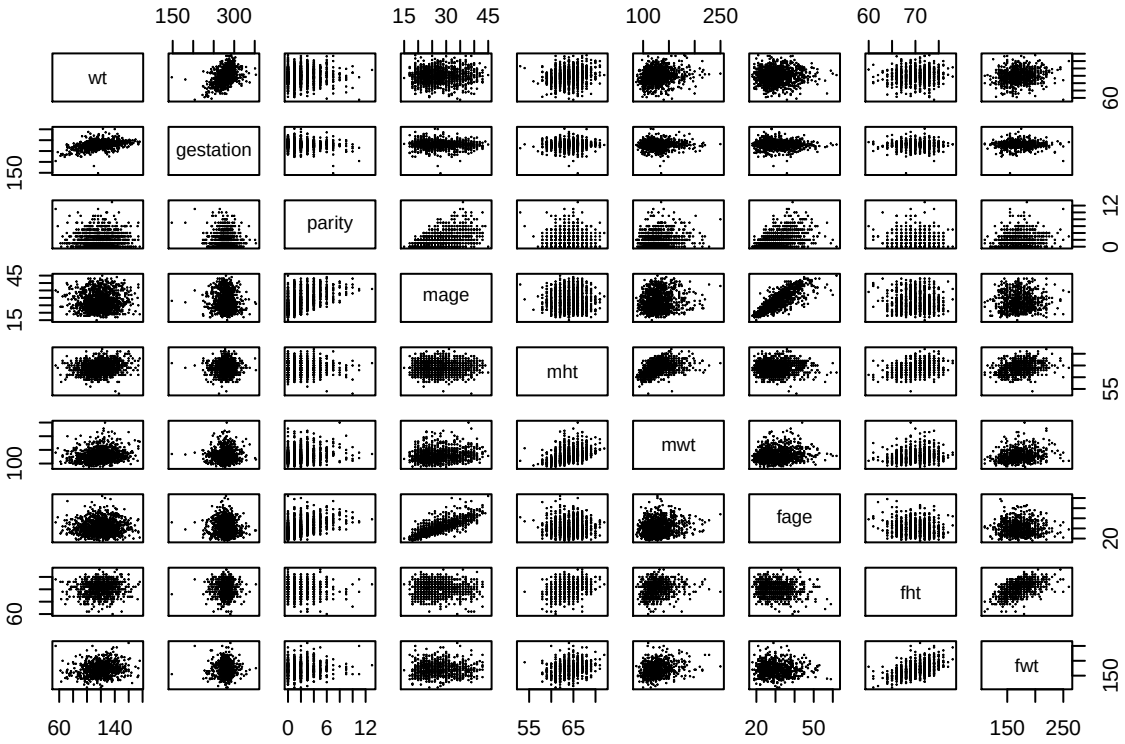
```
levels(births_clean$feth)[1] <- "Caucasian"
levels(births_clean$feth)[3] <- "African-American"
levels(births_clean$feth)[c(2,4,5,6)]<- "others"

# category variable: fed
levels(births_clean$fed)[2] <- "middle-school"
levels(births_clean$fed)[3] <- "high-school"
```

```r
levels(births_clean$fed)[5] <- "high-school + some college"
levels(births_clean$fed)[6] <- "college graduate"
levels(births_clean$fed)[c(1,4,7,8)]<- "others"

# category variable: marital
levels(births_clean$marital)[1] <- "married"
levels(births_clean$marital)[c(2,3,4,5)] <- "others"

# delete smoke
births_clean$smoke = NULL

# category variable: time
levels(births_clean$time)[1] <- "never smoked"
levels(births_clean$time)[2] <- "still smokes"
levels(births_clean$time)[c(3,4,5,6,7,8,9,10)] <- "others"

# category variable: number
levels(births_clean$number)[1] <- "never smoked"
levels(births_clean$number)[2] <- "1-4"
levels(births_clean$number)[3] <- "5-9"
levels(births_clean$number)[6] <- "20-29"
levels(births_clean$number)[c(4,5,7,8,9,10)] <- "others"

#summary of new data
summary(births_clean)
```

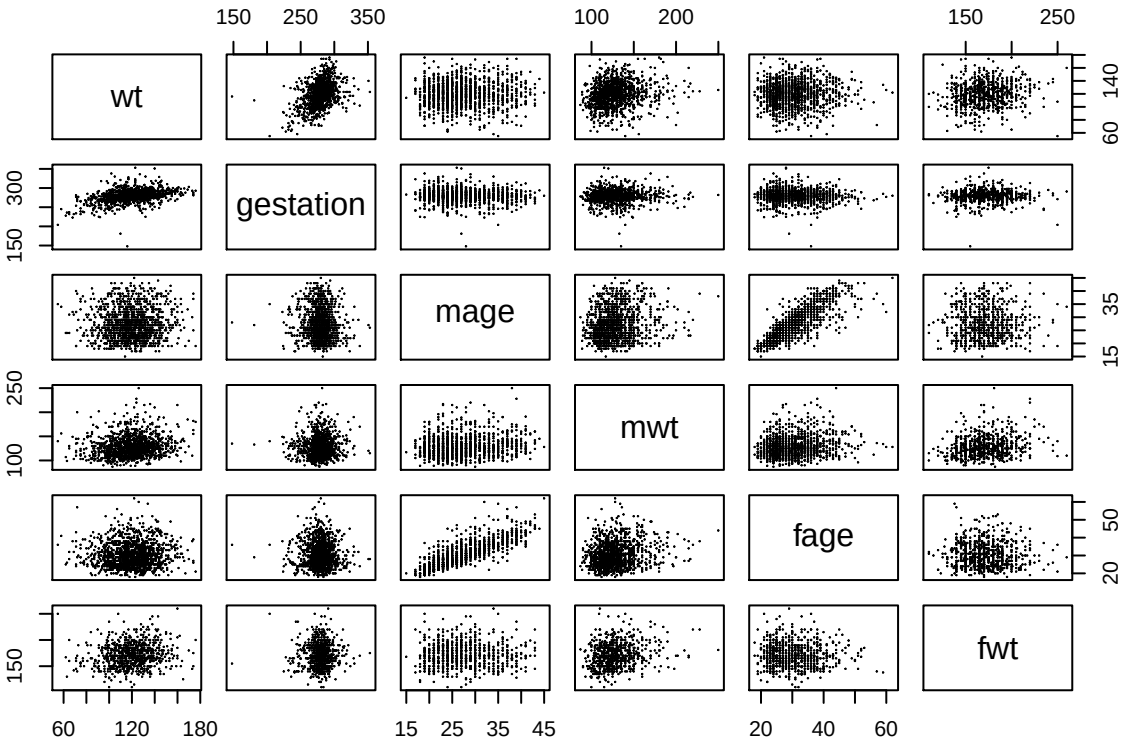
```
##        wt          gestation         parity                    meth
##  Min.   : 55.0   Min.   :148.0   Min.   : 0.000   Caucasian       :871
##  1st Qu.:108.8   1st Qu.:272.0   1st Qu.: 0.000   others          :121
##  Median :120.0   Median :280.0   Median : 1.000   African-American:244
##  Mean   :119.6   Mean   :279.3   Mean   : 1.932
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.: 3.000
##  Max.   :176.0   Max.   :353.0   Max.   :13.000
##       mage                                med           mht
##  Min.   :15.00   others                     : 91   Min.   :53.00
##  1st Qu.:23.00   middle-school              :183   1st Qu.:62.00
##  Median :26.00   high-school                :444   Median :64.00
##  Mean   :27.25   high-school + some college:299   Mean   :64.04
##  3rd Qu.:31.00   college graduate           :219   3rd Qu.:66.00
##  Max.   :45.00                                     Max.   :72.00
##       mwt                      feth          fage
##  Min.   : 87.0   Caucasian       :881   Min.   :18.00
##  1st Qu.:114.0   others          : 97   1st Qu.:25.00
##  Median :125.0   African-American:258   Median :29.00
##  Mean   :128.6                          Mean   :30.35
##  3rd Qu.:139.0                          3rd Qu.:34.00
##  Max.   :250.0                          Max.   :62.00
##                          fed           fht             fwt
##  others                     : 78   Min.   :60.00   Min.   :110.0
##  middle-school              :195   1st Qu.:68.00   1st Qu.:155.0
##  high-school                :345   Median :71.00   Median :170.0
##  high-school + some college:268   Mean   :70.23   Mean   :170.9
##  college graduate           :350   3rd Qu.:72.00   3rd Qu.:185.0
##                                    Max.   :78.00   Max.   :260.0
```

```
##     marital         income                time              number
##  married:1210   Min.   :0.000   never smoked:548   never smoked:550
##  others :  26   1st Qu.:2.000   still smokes:489   1-4          :160
##                 Median :3.000   others       :199   5-9          :168
##                 Mean   :3.667                      others       :157
##                 3rd Qu.:5.000                      20-29        :201
##                 Max.   :9.000
```

```
# pairplot for continues data
pairs(~wt + gestation + parity + mage + mht + mwt + fage + fht + fwt,cex = .05,  data = chds_births)
```




```
pairs(~wt + gestation + mage + mwt + fage + fwt,cex = .05,  data = chds_births) # detailed pairplots
```




```
M0 <- lm(wt ~ 1, data = births_clean) # initial model
Mmax <- lm(wt ~ (.)^2 , data = births_clean) # full model
Mstart <- lm(wt ~ ., data = births_clean) #start model

# detect coefficients which are NAs
beta.max <- coef(Mmax)
names(beta.max)[is.na(beta.max)]
```

```
##  [1] "medhigh-school + some college:number20-29"
##  [2] "medcollege graduate:number20-29"
##  [3] "mht:number20-29"
##  [4] "mwt:number20-29"
##  [5] "fethAfrican-American:maritalothers"
##  [6] "fethothers:number20-29"
##  [7] "fethAfrican-American:number20-29"
##  [8] "fage:number20-29"
##  [9] "fedcollege graduate:maritalothers"
## [10] "fedmiddle-school:number20-29"
## [11] "fedhigh-school:number20-29"
## [12] "fedhigh-school + some college:number20-29"
## [13] "fedcollege graduate:number20-29"
## [14] "fht:number20-29"
## [15] "fwt:number20-29"
## [16] "maritalothers:number1-4"
## [17] "maritalothers:number20-29"
## [18] "income:number20-29"
## [19] "timeothers:number1-4"
## [20] "timeothers:number5-9"
## [21] "timeothers:numberothers"
## [22] "timestill smokes:number20-29"
## [23] "timeothers:number20-29"
```

```r
anyNA(coef(Mmax))
```

```
## [1] TRUE
```

```r
Mmax <- lm(wt ~ (. -marital -fed -feth -number -time - meth -med)^2
           +marital + fed +feth +number +time + meth +med , data = births_clean) # Revised max model
Mstart <- lm(wt ~ ., data = births_clean) #start model
anyNA(coef(Mmax)) # detect coefficients which are NAs
```

```
## [1] FALSE
```

```r
# Forward selection
invisible(Mfwd <- step(object = M0,
                scope = list(lower = M0, upper = Mmax),
                direction = "forward", trace = FALSE))

# Backward elimination selection
Mback <- step(object = Mmax,
              scope = list(lower = M0, upper = Mmax),
              direction = "backward", trace = FALSE)

# Stepwise selection
Mstep <- step(object = Mstart,
              scope = list(lower = M0, upper = Mmax),
              direction = "both", trace = FALSE)

c(fwd = length(coef(Mfwd)), back = length(coef(Mback)), step = length(coef(Mstep)))
```

```
##  fwd back step
##   20   33   25
```

```r
Mfwd$call
```

```
## lm(formula = wt ~ gestation + time + mht + meth + parity + number +
##     fwt + mwt + fht + gestation:fwt + gestation:mwt + mht:fwt +
##     gestation:mht + gestation:fht, data = births_clean)
```

```r
Mback$call
```

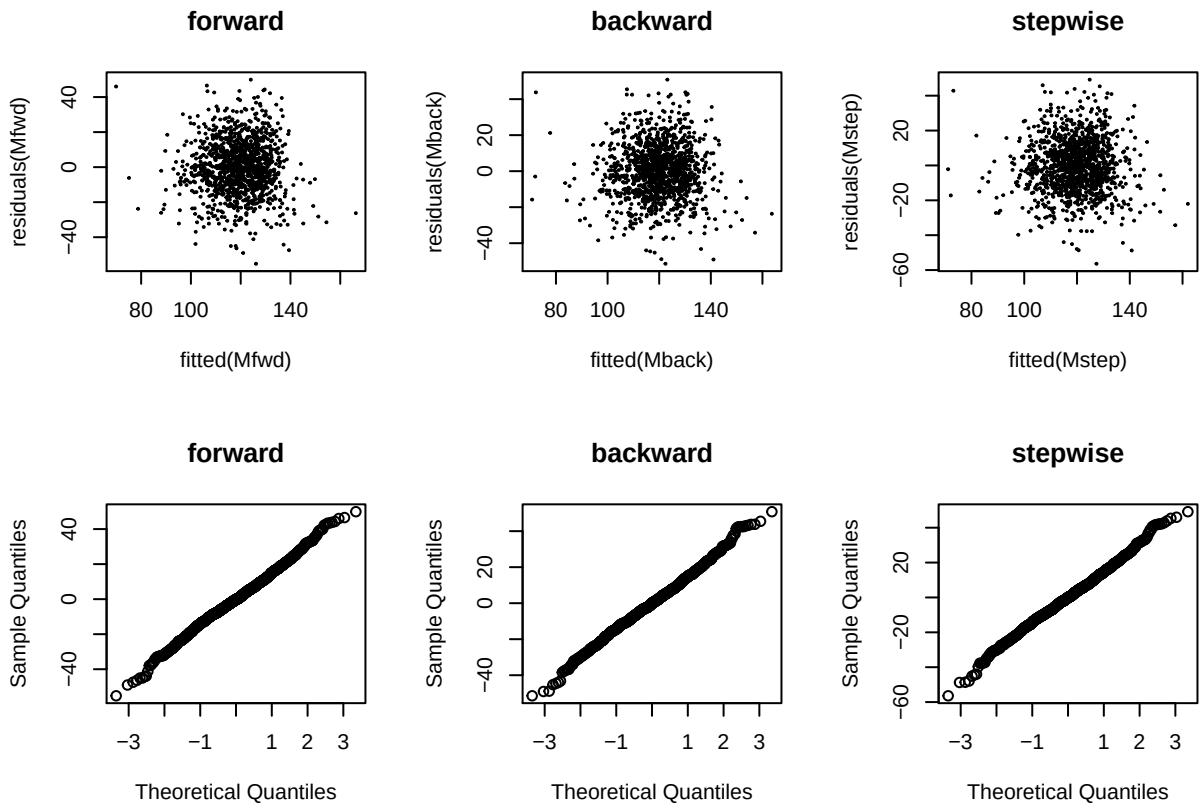
```
## lm(formula = wt ~ gestation + parity + mage + mht + mwt + fage +
##     fht + fwt + income + number + time + meth + gestation:mage +
##     gestation:mht + gestation:mwt + gestation:fht + gestation:fwt +
##     gestation:income + parity:mht + mht:fage + mht:fht + mht:fwt +
##     mht:income + mwt:income + fage:fwt + fage:income + fht:income,
##     data = births_clean)
```

```r
Mstep$call
```

```
## lm(formula = wt ~ gestation + parity + meth + mage + mht + mwt +
##     fht + fwt + income + time + number + gestation:income + mwt:income +
##     gestation:mage + gestation:fwt + gestation:mwt + gestation:fht +
##     gestation:mht + fht:income, data = births_clean)
```

```r
par(mfrow=c(2,3))
plot(fitted(Mfwd), residuals(Mfwd), main="forward", cex = .2)
plot(fitted(Mback), residuals(Mback), main="backward", cex = .2)
plot(fitted(Mstep), residuals(Mstep), main="stepwise", cex =.2)
qqnorm(residuals(Mfwd), main="forward")
```

```r
qqnorm(residuals(Mback), main="backward")
qqnorm(residuals(Mstep), main="stepwise")
```




```r
M1 <- Mfwd
M2 <- Mback
M3 <- Mstep
Mnames <- expression(M[FWD], M[BACK], M[STEP])

# press for 3 automated models
press1 <- resid(M1)/(1-hatvalues(M1))
press2 <- resid(M2)/(1-hatvalues(M2))
press3 <- resid(M3)/(1-hatvalues(M3))
PRESS = c(sum(press1^2), sum(press2^2), sum(press3^2))

# R^2 for 3 automated models
r_square1 <- summary(Mfwd)$r.squared
r_square2 <- summary(Mback)$r.squared
r_square3 <- summary(Mstep)$r.squared
R_Squared <- c(r_square1,r_square2,r_square3)

# Adjusted R^2 for 3 automated models
r_adj_square1 <- summary(M1)$adj.r.squared
r_adj_square2 <- summary(M2)$adj.r.squared
r_adj_square3 <- summary(M3)$adj.r.squared
R_adj_Squared <- c(r_adj_square1,r_adj_square2,r_adj_square3)

# AIC for 3 automated models
AIC1 <- AIC(M1)
```

```r
AIC2 <- AIC(M2)
AIC3 <- AIC(M3)
AIC = c(AIC1,AIC2,AIC3)

# display results
disp <- rbind(AIC,PRESS,R_Squared,R_adj_Squared)
colnames(disp) <- Mnames
disp
```

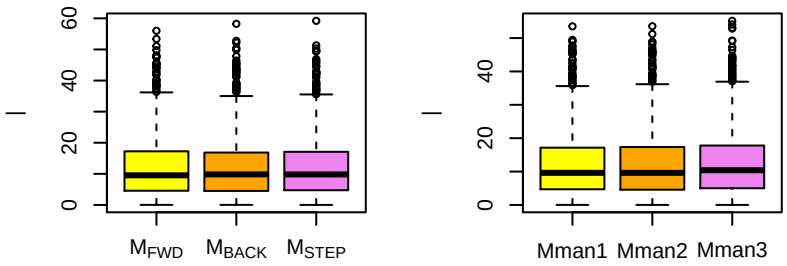
```
##                      M[FWD]      M[BACK]      M[STEP]
## AIC           1.028290e+04 1.026394e+04 1.026355e+04
## PRESS         2.986716e+05 2.955627e+05 2.942003e+05
## R_Squared     3.012018e-01 3.261652e-01 3.176000e-01
## R_adj_Squared 2.902831e-01 3.082411e-01 3.040760e-01
```

```r
#plot PRESS statistics
boxplot(x = list(abs(press1),abs(press2),abs(press3)), names = Mnames,
        ylab = expression("|", PRESS[i], "|"),col = c("yellow","orange","violet"))

Mman1 <- lm(formula = wt ~ gestation + parity + meth + mage + mht + mwt +
    fht + fwt + income + time + number,  data = births_clean)
Mman2 <- lm(formula = wt ~ gestation + parity + mage + mht + mwt + fage +
    fht + fwt + income + number + time + meth,  data = births_clean)
Mman3 <- lm(formula = wt ~ gestation + mage + fage  + time + number + meth + feth,
            data = births_clean)

Mnames <- expression(Mman1, Mman2, Mman3)
# press for 3 manual models
press1 <- resid(Mman1)/(1-hatvalues(Mman1))
press2 <- resid(Mman2)/(1-hatvalues(Mman2))
press3 <- resid(Mman3)/(1-hatvalues(Mman3))
PRESS = c(sum(press1^2), sum(press2^2), sum(press3^2))

# R^2 for 3 manual models
r_square1 <- summary(Mman1)$r.squared
r_square2 <- summary(Mman2)$r.squared
r_square3 <- summary(Mman3)$r.squared
R_Squared <- c(r_square1,r_square2,r_square3)

# Adjusted R^2 for 3 manual models
r_adj_square1 <- summary(Mman1)$adj.r.squared
r_adj_square2 <- summary(Mman2)$adj.r.squared
r_adj_square3 <- summary(Mman3)$adj.r.squared
R_adj_Squared <- c(r_adj_square1,r_adj_square2,r_adj_square3)


# AIC for 3 manual models
AIC1 <- AIC(Mman1)
AIC2 <- AIC(Mman2)
AIC3 <- AIC(Mman3)
AIC = c(AIC1,AIC2,AIC3)

# display results
disp <- rbind(AIC,PRESS,R_Squared,R_adj_Squared)
```

```
colnames(disp) <- Mnames
disp
```

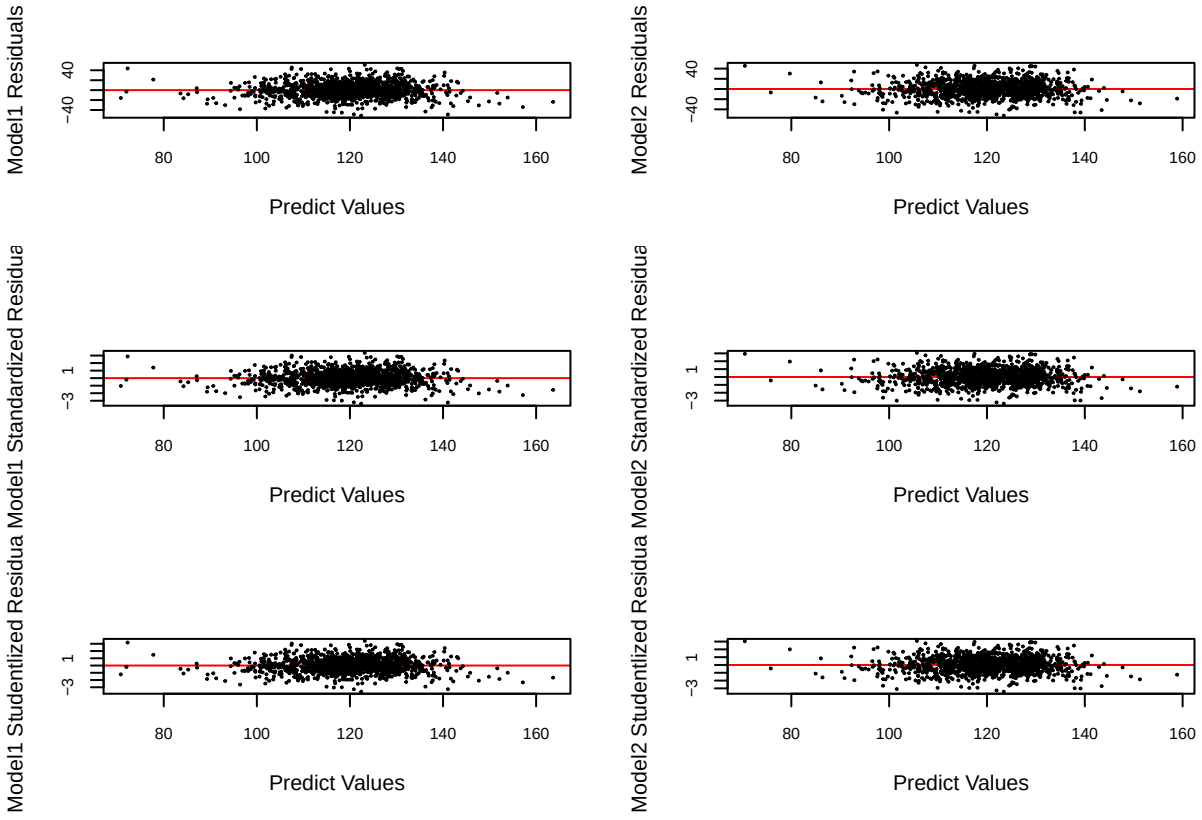
```
##                       Mman1        Mman2        Mman3
## AIC            1.029669e+04 1.029822e+04 1.038430e+04
## PRESS          3.006585e+05 3.010658e+05 3.221790e+05
## R_Squared      2.899263e-01 2.901968e-01 2.340586e-01
## R_adj_Squared 2.806062e-01 2.802899e-01 2.259103e-01
```

```
#plot PRESS statistics
boxplot(x = list(abs(press1),abs(press2),abs(press3)), names = Mnames,
        ylab = expression("|", PRESS[i], "|"),col = c("yellow","orange","violet"))

Model1 <- Mback
Model2 <- Mman2

h1 <- hatvalues(Model1) # hat matrix for model1
h2 <- hatvalues(Model2) # hat matrix for model2

y1.hat <- predict(Model1) # predicted values for model1
y2.hat <- predict(Model2) # predicted values for model2

Re1 <- residuals(Model1) # Residual for model1
Re2 <- residuals(Model2) # Residual for model2

StanRe1 <- Re1/sigma(Model1) # Standard Residual for model1
StanRe2 <- Re2/sigma(Model2) # Standard Residual for model2

StudRe1 <- StanRe1 / sqrt(1-hatvalues(Model1)) # Studentized Residual for model1
StudRe2 <- StanRe2 / sqrt(1-hatvalues(Model2)) # Studentized Residual for model2


par(mfrow=c(3,2))
```



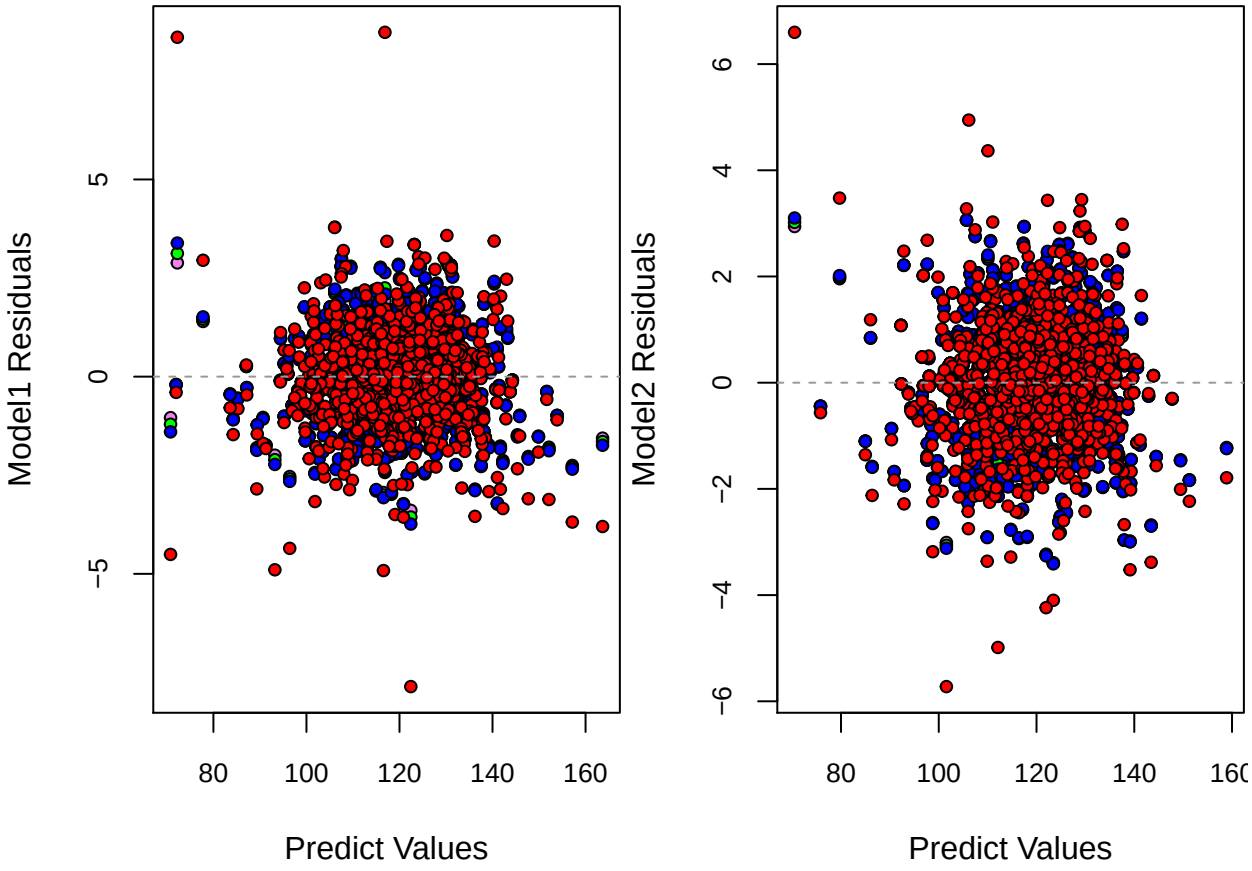

```
## Residual plots
plot(predict(Model1), Re1, xlab = "Predict Values", ylab = "Model1 Residuals", cex.axis = .8, cex = .2,
     abline(h = mean(Re1), col = "red"))
plot(predict(Model2), Re2, xlab = "Predict Values", ylab = "Model2 Residuals", cex.axis = .8, cex = .2,
     abline(h = mean(Re2), col = "red"))

## Standlized residual plots
plot(predict(Model1), StanRe1, xlab = "Predict Values", ylab = "Model1 Standardized Residuals", cex.axis
     cex = .2, abline(h = mean(StanRe1), col = "red"))
plot(predict(Model2), StanRe2, xlab = "Predict Values", ylab = "Model2 Standardized Residuals", cex.axis
     cex = .2, abline(h = mean(StanRe2), col = "red"))
```

```r
## Studentlized residuals plots
plot(predict(Model1), StudRe1, xlab = "Predict Values", ylab = "Model1 Studentlized Residuals", cex.axis
     cex = .2, abline(h = mean(StudRe1), col = "red"))
plot(predict(Model2), StudRe2, xlab = "Predict Values", ylab = "Model2 Studentlized Residuals", cex.axis
     cex = .2, abline(h = mean(StudRe2), col = "red"))
```



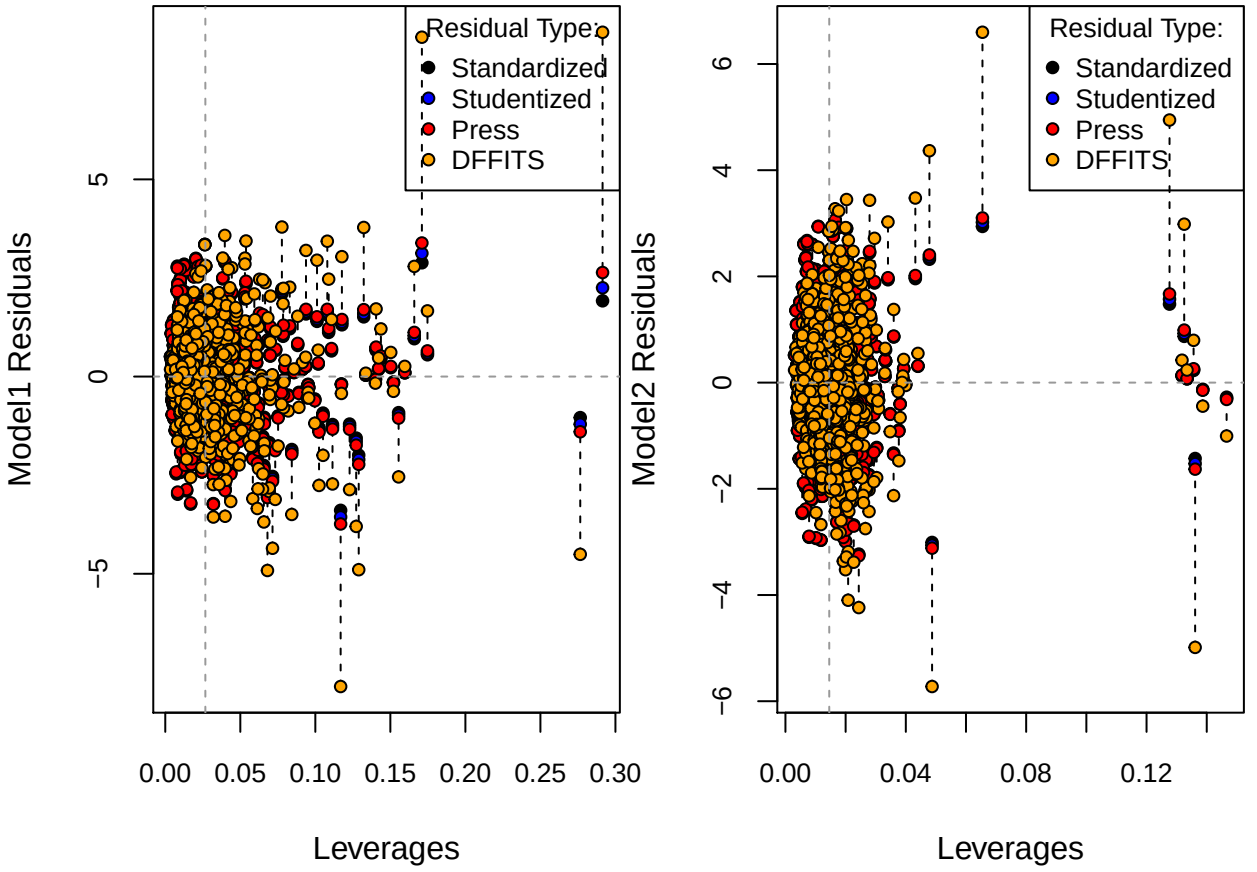

```r
press_model1 <- Re1/(1 - hatvalues(Model1)) #press for model1
press_model2 <- Re2/(1 - hatvalues(Model2)) #press for model2

dfts1 <- dffits(Model1) #DFFITS for model1
dfts2 <- dffits(Model2) #DFFITS for model2

# standlize each of these
p1 <- length(coef(Model1))
n1 <- nobs(Model1)
hbar1 <- p1/n1
StudRe1.stan <- StudRe1 * sqrt(1-hbar1)
press_model1.stan <- press_model1*(1-hbar1)/sigma(Model1)
dfts1.stan <- dfts1*(1-hbar1)/sqrt(hbar1)


# plots all
par(mfrow = c(1,2), mar = c(4,4,.1,.1))
plot(predict(Model1), rep(0, length(predict(Model1))),
     type = "n",
     ylim = range(StanRe1,StudRe1.stan,dfts1.stan,press_model1.stan),
     xlab = "Predict Values",
     ylab = "Model1 Residuals",
```

```
     cex.axis = .8)
segments(x0 = h1,
        y0 = pmin(StanRe1, StudRe1.stan, press_model1.stan, dfts1.stan),
        y1 = pmax(StanRe1, StudRe1.stan, press_model1.stan, dfts1.stan),
        lty = 2)
points(predict(Model1), StanRe1, pch = 21, bg = "violet", cex = .8)
points(predict(Model1), StudRe1.stan, pch = 21, bg = "green", cex = .8)
points(predict(Model1), press_model1.stan, pch = 21, bg = "blue", cex = .8)
points(predict(Model1), dfts1.stan, pch = 21, bg = "red", cex = .8)
abline(h = 0, col = "grey60", lty =2) #horizontal line

## model2
# standlize each of these
p2 <- length(coef(Model2))
n2 <- nobs(Model2)
hbar2 <- p2/n2
StudRe2.stan <- StudRe2 * sqrt(1-hbar2)
press_model2.stan <- press_model2*(1-hbar2)/sigma(Model2)
dfts2.stan <- dfts2*(1-hbar2)/sqrt(hbar2)

# plots all
plot(predict(Model2), rep(0, length(predict(Model2))),
     type = "n",
     ylim = range(StanRe2,StudRe2.stan,dfts2.stan,press_model2.stan),
     xlab = "Predict Values",
     ylab = "Model2 Residuals",
     cex.axis = .8)
segments(x0 = h2,
        y0 = pmin(StanRe2, StudRe2.stan, press_model2.stan, dfts2.stan),
        y1 = pmax(StanRe2, StudRe2.stan, press_model2.stan, dfts2.stan),
        lty = 2)
points(predict(Model2), StanRe2, pch = 21, bg = "violet", cex = .8)
points(predict(Model2), StudRe2.stan, pch = 21, bg = "green", cex = .8)
points(predict(Model2), press_model2.stan, pch = 21, bg = "blue", cex = .8)
points(predict(Model2), dfts2.stan, pch = 21, bg = "red", cex = .8)
abline(h = 0, col = "grey60", lty =2) #horizontal line
```



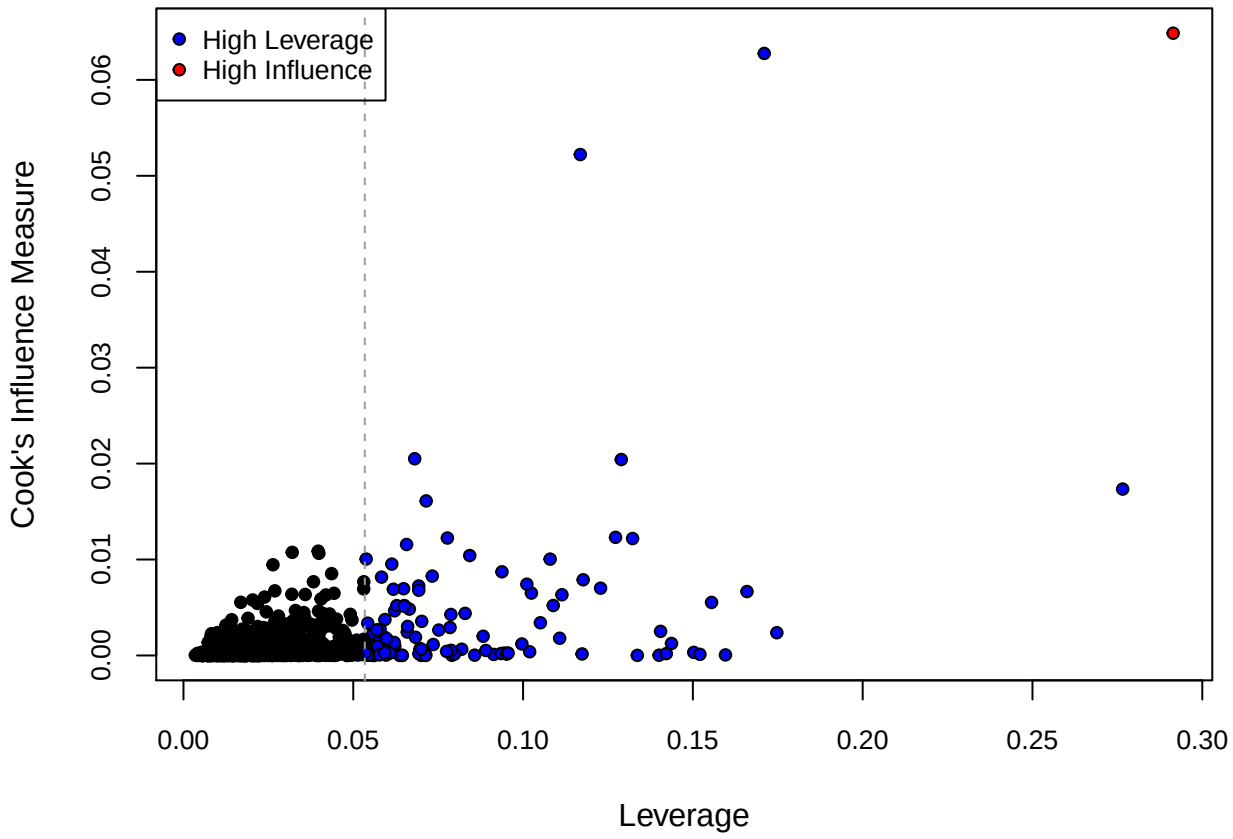

```
# plots Residuals vs Leverages
# model1
plot(h1, rep(0, length(y1.hat)), type = "n", cex.axis = .8,
     ylim = range(StanRe1, StudRe1.stan, press_model1.stan, dfts1.stan),
     xlab = "Leverages", ylab = "Model1 Residuals")
segments(x0 = h1,
        y0 = pmin(StanRe1, StudRe1.stan, press_model1.stan, dfts1.stan),
        y1 = pmax(StanRe1, StudRe1.stan, press_model1.stan, dfts1.stan),
        lty = 2)
points(h1, StanRe1, pch = 21, bg = "black", cex =.8)
points(h1, StudRe1.stan, pch = 21, bg = "blue", cex =.8)
points(h1,press_model1.stan, pch = 21, bg = "red", cex = .8)
points(h1,dfts1.stan, pch = 21, bg = "orange", cex = .8)
abline(v = hbar1, col = "grey60", lty = 2)
abline(h = 0, col = "grey60", lty =2) #horizontal line
legend("topright",legend = c("Standardized", "Studentized","Press","DFFITS"), pch = 21, pt.bg = c("blac

# model2
plot(h2, rep(0, length(y2.hat)), type = "n", cex.axis = .8,
     ylim = range(StanRe2, StudRe2.stan, press_model2.stan, dfts2.stan),
     xlab = "Leverages", ylab = "Model2 Residuals")
segments(x0 = h2,
        y0 = pmin(StanRe2, StudRe2.stan, press_model2.stan, dfts2.stan),
        y1 = pmax(StanRe2, StudRe2.stan, press_model2.stan, dfts2.stan),
        lty = 2)
points(h2, StanRe2, pch = 21, bg = "black", cex =.8)
```

```
points(h2, StudRe2.stan, pch = 21, bg = "blue", cex =.8)
points(h2,press_model2.stan, pch = 21, bg = "red", cex = .8)
points(h2,dfts2.stan, pch = 21, bg = "orange", cex = .8)
abline(v = hbar2, col = "grey60", lty = 2)
abline(h = 0, col = "grey60", lty =2) #horizontal line
legend("topright",legend = c("Standardized", "Studentized","Press","DFFITS"), pch = 21, pt.bg = c("blac
```

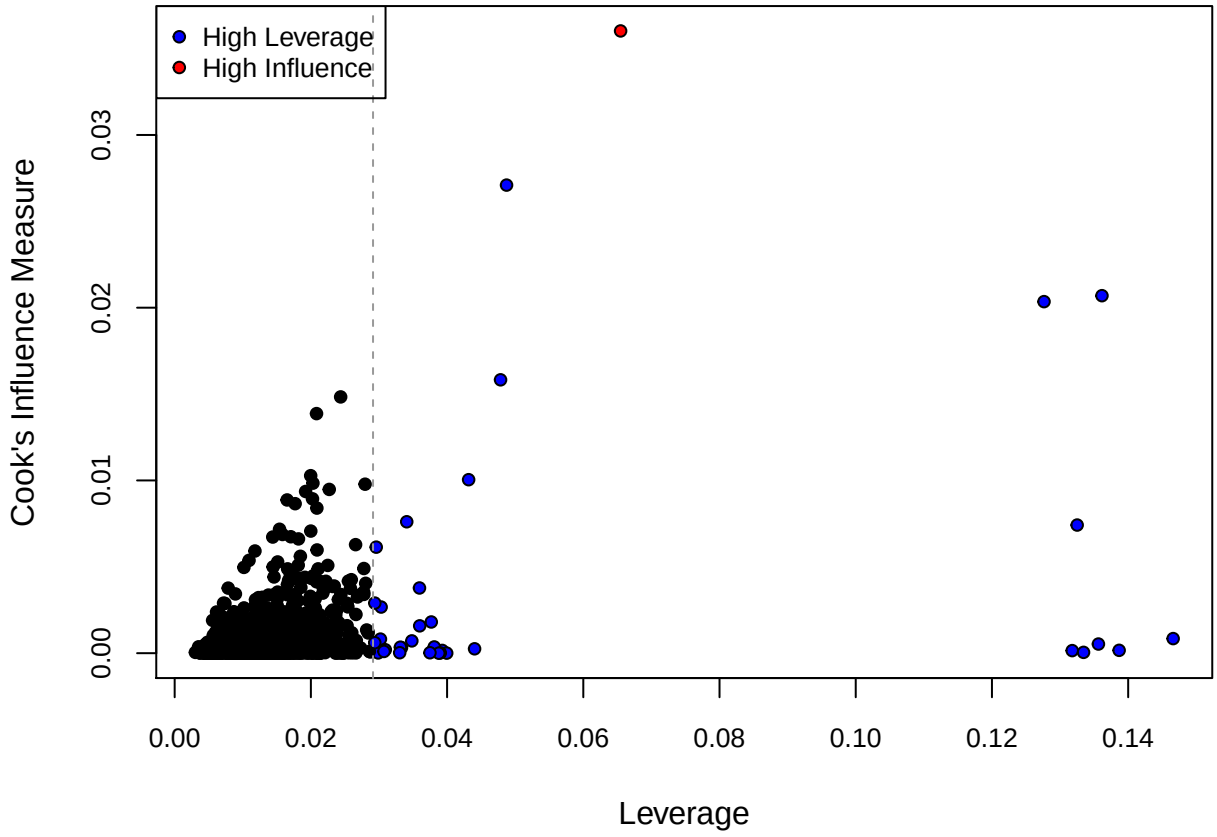



```
# compute leverage

h1 <- hatvalues(Model1)
h2 <- hatvalues(Model2)

D1 <- cooks.distance(Model1)
D2 <- cooks.distance(Model2)

infl1.ind <- which.max(D1)
infl2.ind <- which.max(D2)

lev1.ind <- h1 > 2*hbar1
lev2.ind <- h2 > 2*hbar2

clrs <- rep("black", len = nobs(Model1))
clrs[lev1.ind] <- "blue"
clrs[infl1.ind] <-"red"
par(mfrow = c(1,1), mar = c(4,4,1,1))
cex <- .8
```

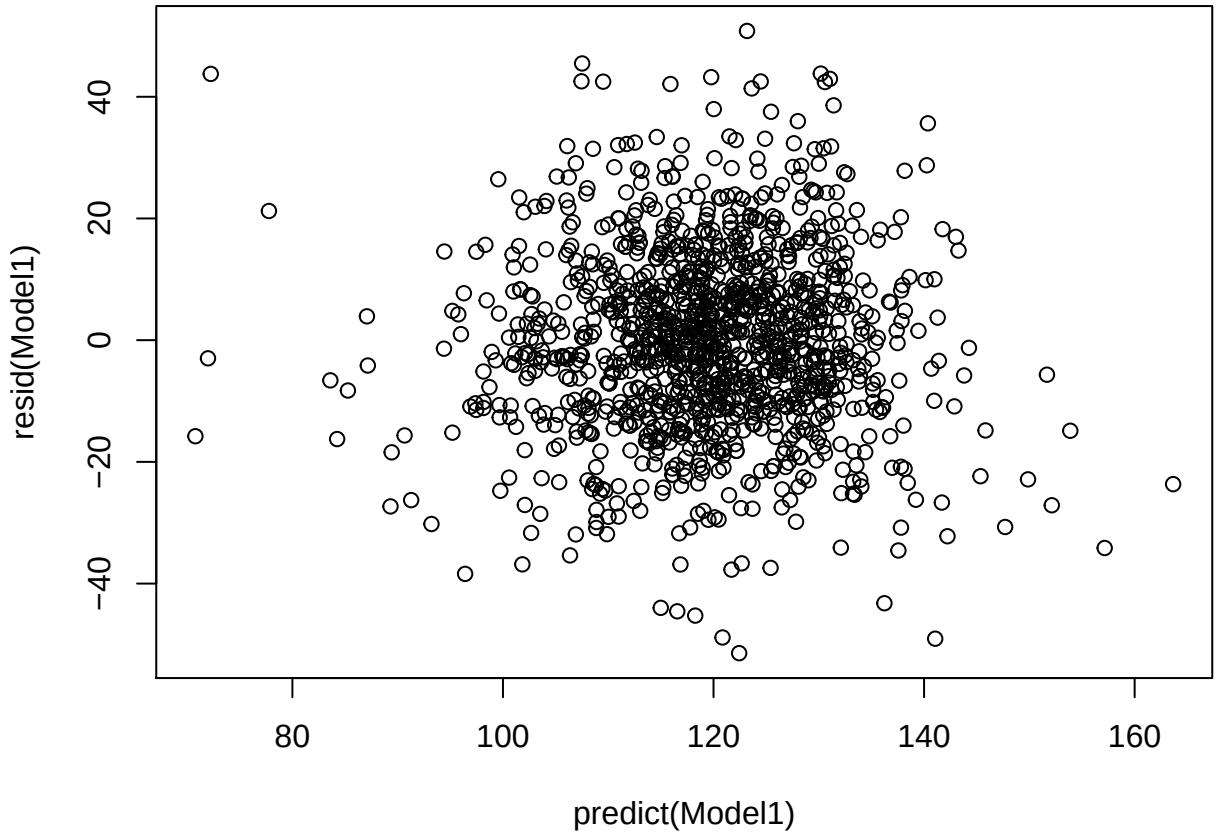
```
plot(h1,D1,xlab = "Leverage", ylab = "Cook's Influence Measure", pch = 21, bg = clrs, cex = cex, cex.axi
abline(v = 2*hbar1, col = "grey60", lty = 2)
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21, pt.bg = c("blue", "red"),cex
```




```
clrs <- rep("black", len = nobs(Model1))
clrs[lev2.ind] <- "blue"
clrs[infl2.ind] <-"red"
par(mfrow = c(1,1), mar = c(4,4,1,1))
cex <- .8
plot(h2,D2,xlab = "Leverage", ylab = "Cook's Influence Measure", pch = 21, bg = clrs, cex = cex, cex.axi
abline(v = 2*hbar2, col = "grey60", lty = 2)
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21, pt.bg = c("blue", "red"),cex
```




```
plot(predict(Model1), resid(Model1))
```




```
entry1 <- which.max(abs(resid(Model1)))

M1 <- Model1
M2 <- Model2
Mnames_new <- expression(M[Candidate1],M[Candidate2])
# Cross-validation setup
nreps <- 2e3 # number of replications
ntot <- nrow(births_clean) # total number of observations
ntrain <- floor(0.7 * ntot) # size of training set
ntest <- ntot-ntrain # size of test set
mspe1 <- rep(NA, nreps) # sum-of-square errors for each CV replication
mspe2 <- rep(NA, nreps)
logLambda <- rep(NA, nreps) # log-likelihod ratio statistic for each replication
for(ii in 1:nreps) {
  # randomly select training observations
  train.ind <- sample(ntot, ntrain) # training observations
  # refit the models on the subset of training data; ?update for details!
  M1.cv <- update(M1, subset = train.ind)
  M2.cv <- update(M2, subset = train.ind)
  # out-of-sample residuals for both models
  # that is, testing data - predictions with training parameters
  M1.res <- births_clean$wt[-train.ind] -
            predict(M1.cv, newdata = births_clean[-train.ind,])
  M2.res <- births_clean$wt[-train.ind] -
            predict(M2.cv, newdata = births_clean[-train.ind,])
  # mean-square prediction errors
  mspe1[ii] <- mean(M1.res^2)
```

```r
  mspe2[ii] <- mean(M2.res^2)
  # out-of-sample likelihood ratio
  M1.sigma <- sqrt(sum(resid(M1.cv)^2)/ntrain) # MLE of sigma
  M2.sigma <- sqrt(sum(resid(M2.cv)^2)/ntrain)
  # since res = y - pred, dnorm(y, pred, sd) = dnorm(res, 0, sd)
  logLambda[ii] <- sum(dnorm(M1.res, mean = 0, sd = M1.sigma, log = TRUE))
  logLambda[ii] <- logLambda[ii] -
                   sum(dnorm(M2.res, mean = 0, sd = M2.sigma, log = TRUE))
}
# plot rMSPE and out-of-sample log
par(mfrow = c(1,2))
par(mar = c(4.5, 4.5, .1, .1))
boxplot(x = list(sqrt(mspe1), sqrt(mspe2)),
        names = Mnames_new, cex = .7,
        ylab = expression(sqrt(MSPE)), col = c("yellow", "orange"))
hist(logLambda, breaks = 50, freq = FALSE,
     xlab = expression(Lambda^{test}),
     main = "", cex = .7)
abline(v = mean(logLambda), col = "red") # average value
```