STA 6443 DA Algorithms I: Final Exam

Rudy Martinez

12/5/2020

Libraries

```
library(MASS)
library(car)
library(olsrr)
library(DescTools)
library(ResourceSelection)
```

Set Working Directory

```
setwd("/Users/rudymartinez/Desktop/MSDA/Fall 2020/STA 6443_Algorithms I/STAT-Algorithms-1/Final Exam")
```

Read and View Data Structure

```
birthweight = read.csv("birthweight_final.csv", header = TRUE)
birthweight$Black = as.factor(birthweight$Black)
birthweight$Married = as.factor(birthweight$Married)
birthweight$Boy = as.factor(birthweight$Boy)
birthweight$MomSmoke = as.factor(birthweight$MomSmoke)
birthweight$Ed = as.factor(birthweight$Ed)

str(birthweight)
```

```
## 'data.frame':    400 obs. of  10 variables:
##  $ Weight   : int  3657 3317 3232 2948 4564 2977 4394 2977 3855 2977 ...
##  $ Weight_Gr: int  0 1 1 1 0 1 0 1 0 1 ...
##  $ Black    : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ Married  : Factor w/ 2 levels "0","1": 2 1 1 2 2 2 1 2 2 1 ...
##  $ Boy      : Factor w/ 2 levels "0","1": 2 2 2 2 2 1 2 1 1 2 ...
##  $ MomSmoke : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Ed       : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 2 2 1 ...
##  $ MomAge   : int  4 -5 -7 -4 5 -6 -1 -1 1 -2 ...
##  $ MomWtGain: int  2 0 -10 18 27 2 17 -12 10 -2 ...
##  $ Visit    : int  3 3 3 3 3 3 3 3 3 1 ...
```

Exercise 1

Consider to fit a multiple linear regression to model `Weight` using possible explanatory variables; `Black`, `Married`, `Boy`, `MomSmoke`, `Ed`, `MomAge`, `MomWtGain`, and `Visit` (all predictors excluding `Weight_Gr`).

Exercise 1.1

Perform the following four model selection methods and compare their best models. Comment on how they differ or similar in terms of selected variables in the final model. No need to interpret outputs.

Full Model

```
lm.birthweight = lm(Weight ~ Black + Married + Boy + MomSmoke + Ed + MomAge + MomWtGain + Visit, data = birthweight)
```

Stepwise selection with 0.01 p-value criteria for both entry and stay

```
model.stepwise = ols_step_both_p(lm.birthweight, pent = 0.01, prem = 0.01, details = FALSE)
model.stepwise
```

```
##
```

```
##                                Stepwise Selection Summary
## ----------------------------------------------------------------------------------------
##                         Added/                   Adj.
## Step    Variable      Removed     R-Square    R-Square     C(p)         AIC         RMSE
## ----------------------------------------------------------------------------------------
##    1    MomWtGain     addition       0.089       0.087    20.7600    6201.2327    559.8163
##    2    MomSmoke      addition       0.107       0.102    14.6830    6195.4042    555.0626
##    3      Black       addition       0.127       0.120     7.4660    6188.2794    549.4600
## ----------------------------------------
```

Forward selection with 0.01 p-value criteria for entry

```
model.forward = ols_step_forward_p(lm.birthweight, penter = 0.01, details = F)
model.forward
```

```
##
##                              Selection Summary
## ------------------------------------------------------------------------
##         Variable                    Adj.
## Step     Entered     R-Square    R-Square     C(p)         AIC         RMSE
## ------------------------------------------------------------------------
##    1    MomWtGain       0.0893      0.0870    20.7604    6201.2327    559.8163
##    2    MomSmoke        0.1069      0.1024    14.6832    6195.4042    555.0626
##    3    Black           0.1271      0.1205     7.4659    6188.2794    549.4600
## ------------------------------------------------------------------------
```

Backward selection with 0.01 p-value criteria for stay

```
model.backward = ols_step_backward_p(lm.birthweight, prem = 0.01, details = F)
model.backward
```

```
##
##
##                             Elimination Summary
## ------------------------------------------------------------------------
```

```
##          Variable                  Adj.
## Step     Removed     R-Square     R-Square      C(p)        AIC         RMSE
## ----------------------------------------------------------------------------
##    1     Visit          0.1453       0.130    7.1307    6187.8448    546.4642
##    2     MomAge         0.1445       0.1314   5.5158    6186.2385    546.0372
##    3     Married        0.1409       0.130    5.1599    6185.9146    546.4876
##    4     Ed             0.1364       0.1277   5.1841    6185.9688    547.1986
##    5     Boy            0.1271       0.1205   7.4659    6188.2794    549.4600
## ----------------------------------------------------------------------------
```

Adjusted R-squared Criteria

```
model.best.subset = ols_step_best_subset(lm.birthweight)
model.best.subset
```

```
##                       Best Subsets Regression
## --------------------------------------------------------------------
## Model Index    Predictors
## --------------------------------------------------------------------
##      1         MomWtGain
##      2         MomSmoke MomWtGain
##      3         Black MomSmoke MomWtGain
##      4         Black Boy MomSmoke MomWtGain
##      5         Black Boy MomSmoke Ed MomWtGain
##      6         Black Married Boy MomSmoke Ed MomWtGain
##      7         Black Married Boy MomSmoke Ed MomAge MomWtGain
##      8         Black Married Boy MomSmoke Ed MomAge MomWtGain Visit
## --------------------------------------------------------------------
##
##                                                    Subsets Regression Summary
## ------------------------------------------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square     R-Square     R-Square      C(p)         AIC         SBIC          SBC            MSEP             FPE          HSP        APC
## ------------------------------------------------------------------------------------------------------------------------------------------------
##   1        0.0893       0.0870       0.0775     20.7604    6201.2327    5065.9178    6213.2071    125357705.0611    314961.2141    789.4062    0.9199
##   2        0.1069       0.1024        0.091     14.6832    6195.4042    5060.1250    6211.3700    123238567.9566    310405.1540    778.0163    0.9066
##   3        0.1271       0.1205       0.1059      7.4659    6188.2794    5053.1393    6208.2368    120764046.4868    304925.3190    764.3196    0.8906
```

```
##   4        0.1364      0.1277      0.1114     5.1841    6185.9688    5050.9395    6209.9176    119772826.2237    303169.1473    759.9653    0.8854
##   5        0.1409      0.1300      0.1116     5.1599    6185.9146    5050.9720    6213.8549    119462530.2702    303128.3995    759.9203    0.8853
##   6        0.1445      0.1314      0.1103     5.5158    6186.2385    5051.3901    6218.1702    119266462.8033    303374.3231    760.6035    0.8860
##   7        0.1453      0.1300      0.1052     7.1307    6187.8448    5053.0558    6223.7680    119453866.2656    304595.5958    763.7420    0.8896
##   8        0.1456      0.1281      0.1014     9.0000    6189.7111    5054.9736    6229.6258    119720142.0535    306020.7931    767.4022    0.8938
## ----------------------------------------------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

- **Stepwise Selection Model:** 3 Predictors:`MomWtGain`, `MomSmoke`, and `Black`
- **Forward Selection Model:** 3 Predictors: `MomWtGain`, `MomSmoke`, and `Black`
- **Backward Selection Model:** 3 Predictors: `MomWtGain`, `MomSmoke`, and `Black`
- **Best Subsets Approach Model (Adjusted R-Squared):** 6 Predictors: `Black`, `Married`, `Boy`, `MomSmoke`, `Ed`, and `MomWtGain`

Comments: The Stepwise Selection, Forward Selection, and Backward Selection models have the same predictors (`MomWtGain`, `MomSmoke`, and `Black`) - they have these predictors in common with the Best Subsets Approach model. `Married`, `Boy`, and `Ed` are predictors unique to the Best Subsets Approach model.

Exercise 1.2

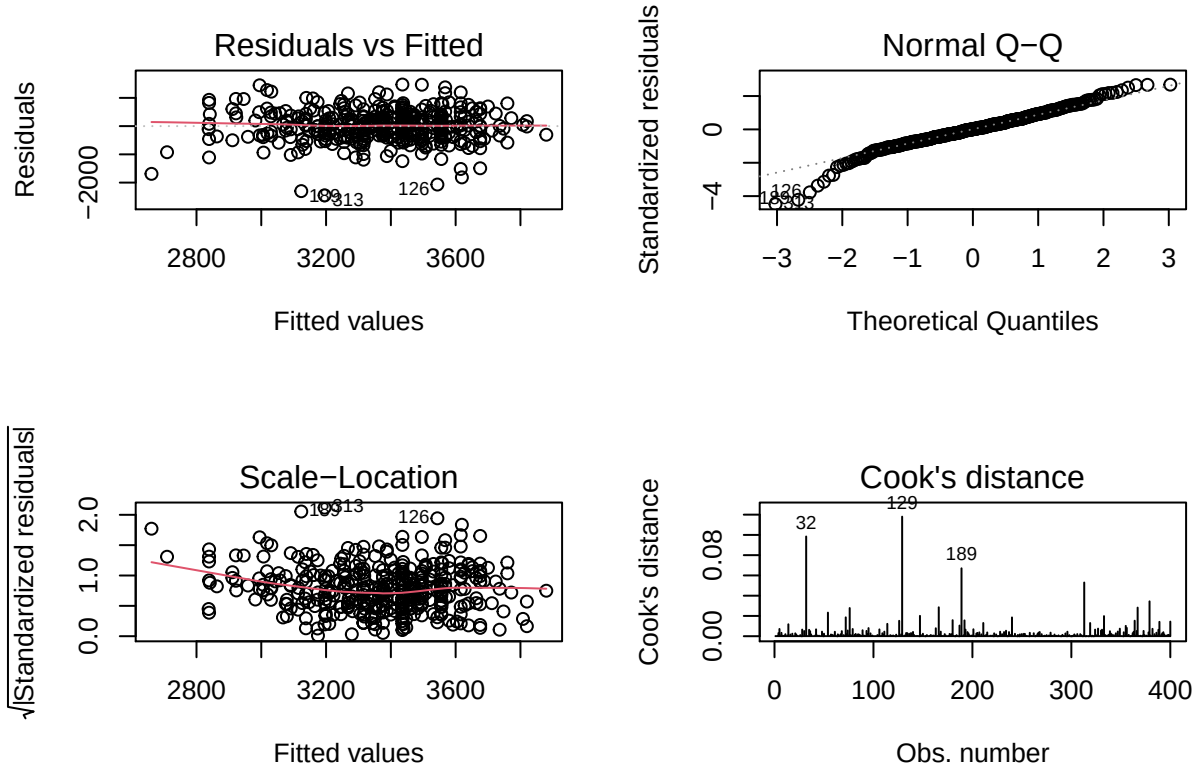
Fit the linear regression with the best model determined by stepwise selection and comment on diagnostics plot. Do not leave observation which has Cook's distance larger than **0.115**. Re-fit the model if necessary. Finally how many observations you use in the final model?

Fit Linear Regression Model (Determined by Stepwise Selection)

```
lm.birthweight.stepwise.1 = lm(Weight ~ MomWtGain + MomSmoke + Black, data = birthweight)
```

Diagnostics Plot

```
par(mfrow=c(2,2))
plot(lm.birthweight.stepwise.1, which=1:4)
```

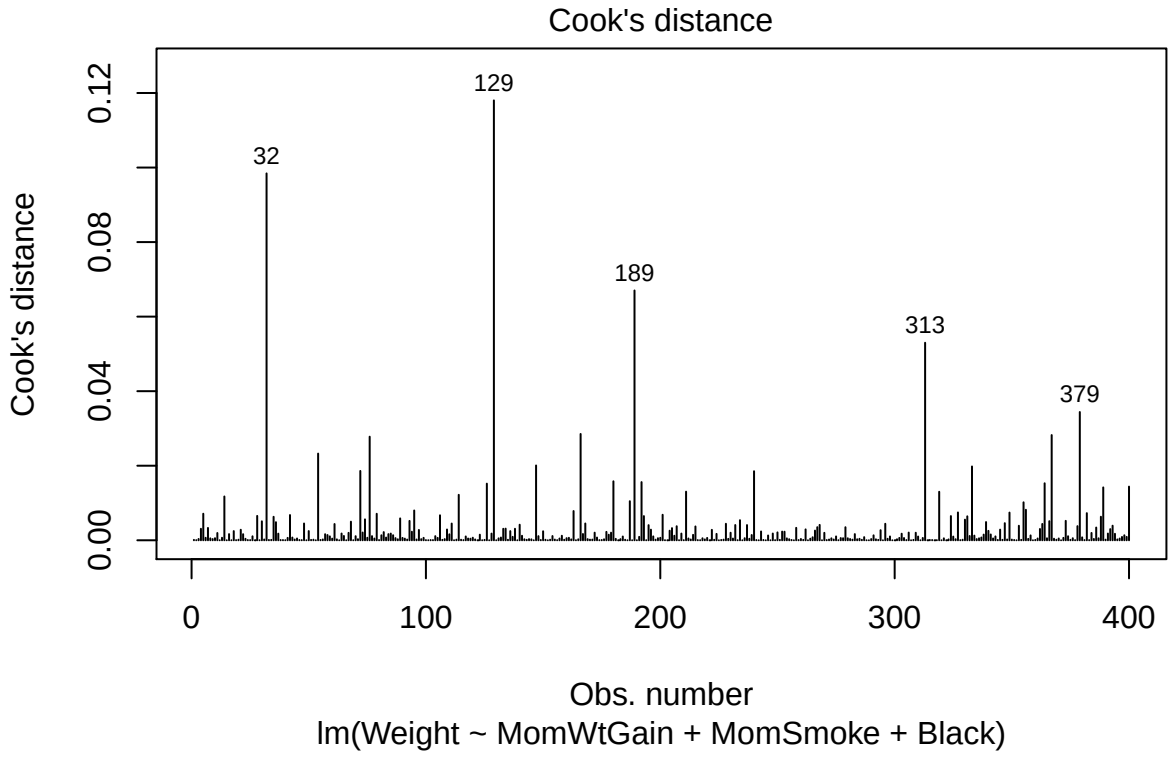


- **Normality Check:** Looking at the **Normal Q-Q Plot**, we see that many of the points fall along the line for the majority of the graph. However, looking at the points in the extremities of the graph, they appear to curve off the line. This indicates that an assumption of normality is **not** reasonable. This is reinforced by looking at the **sqrt(Standardized Residuals) Plot**. Because a considerable number of observations fall above 1.5 along the Y-axis, an assumption of normality is **not** reasonable.

- **Equal Variance Check:** Looking at the **Standardized Residuals Plot**, we see that there **is** a pattern in the residual plot. This supports heteroscedasticity.

Cook's Distance

```
plot(lm.birthweight.stepwise.1, which = 4, id.n = 5)
```



```
inf.id.1 = which(cooks.distance(lm.birthweight.stepwise.1) > 0.115)
inf.id.1
```

```
## 129
## 129
```

- **Issues Identified in Model Diagnostics:** Observation 129 has a Cook's distance larger than 0.115; therefore, the model will be refitted without this influential point.

Refitted Model

```
lm.birthweight.stepwise.2 = lm(Weight ~ MomWtGain + MomSmoke + Black, data = birthweight[-inf.id.1,])
```

- **Observations:** The Final Model will have 399 observations.

Exercise 1.3

How much of the variation in `Weight` is explained by the final model?

```
summary(lm.birthweight.stepwise.2)$r.squared
```

```
## [1] 0.1366329
```

- **R-Squared:** Based on the above R-Squared, 13.66% of the variation in `Weight` can be explained by the model; therefore, the predictive power of this model is low.

Exercise 1.4

Interpret the relationship between predictor variables (in the final model) and Weight value specifically.

```
summary(lm.birthweight.stepwise.2)
```

```
##
## Call:
## lm(formula = Weight ~ MomWtGain + MomSmoke + Black, data = birthweight[-inf.id.1,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2427.02  -309.20     2.98   315.40  1472.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3434.252     32.078 107.059  < 2e-16 ***
## MomWtGain     13.112      2.113   6.204 1.39e-09 ***
## MomSmoke1   -238.923     76.251  -3.133  0.00186 **
## Black1      -198.519     78.022  -2.544  0.01133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 542.2 on 395 degrees of freedom
## Multiple R-squared:  0.1366, Adjusted R-squared:  0.1301
## F-statistic: 20.84 on 3 and 395 DF,  p-value: 1.493e-12
```

Model Significance: The model p-value of 1.493e-12 is below the significance level of 0.05; therefore, we reject the null and conclude that the multiple linear regression model is useful to explain the behavior of `Weight`.

Individual Term Significance

- A p-value of 0.01133 is below the significance level of 0.05; therefore, `Black` and `Weight` have a significant linear relationship.
- A p-value of 0.00186 is below the significance level of 0.05; therefore, `MomSmoke` and `Weight` have a significant linear relationship.
- A p-value of 1.39e-09 is below the significance level of 0.05; therefore, `MomWtGain` and `Weight` have a significant linear relationship.
- **Interpretation:** On average, Black women who smoke give birth to infants that weigh less than women who are not Black and don't smoke when `MomWtGain` (Mother's weight gain during pregnancy) is the same.

Exercise 2

Now we consider fitting a logistic regression for low birthweight (`Weight_Gr`=1). Again consider `Black`, `Married`, `Boy`, `MomSmoke`, `Ed`, `MomAge`, `MomWtGain`, and `Visit` as possible explanatory variables.

Exercise 2.1

Perform following model selection methods and compare their best models. Comment how they differ or similar in terms of selected variables (Stepwise Selection with AIC and BIC Criteria).

Fit Logistic Regression Model

```
glm.null.2 = glm(Weight_Gr ~ 1, data = birthweight, family = "binomial")

glm.full.2 = glm(Weight_Gr ~ Black + Married + Boy + MomSmoke + Ed + MomAge + MomWtGain + Visit,
                 data = birthweight, family = "binomial")
```

Stepwise Selection with AIC Criteria

```
step.model.2.aic = step(glm.null.2, scope = list(upper = glm.full.2),
                                        direction = "both", test = "Chisq", trace = F)

summary(step.model.2.aic)
```

```
##
## Call:
## glm(formula = Weight_Gr ~ MomWtGain + MomSmoke + MomAge + Boy +
##     Ed, family = "binomial", data = birthweight)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9790  -1.0470  -0.6085   1.0966   2.0012
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.240486   0.188075   1.279  0.20101
## MomWtGain   -0.038047   0.008471  -4.492 7.07e-06 ***
## MomSmoke1    0.818590   0.310227   2.639  0.00832 **
## MomAge      -0.044444   0.019040  -2.334  0.01959 *
## Boy1        -0.407560   0.212600  -1.917  0.05523 .
```

```
## Ed1         -0.366259   0.217910  -1.681  0.09280 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.43  on 399  degrees of freedom
## Residual deviance: 510.15  on 394  degrees of freedom
## AIC: 522.15
##
## Number of Fisher Scoring iterations: 4
```

Stepwise Selection with BIC Criteria

```
step.model.2.bic = step(glm.null.2, scope = list(upper = glm.full.2),
                                        direction = "both", test = "Chisq", trace = F, k=log(nrow(birthweight)))

summary(step.model.2.bic)
```

```
##
## Call:
## glm(formula = Weight_Gr ~ MomWtGain + MomSmoke + MomAge, family = "binomial",
##     data = birthweight)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.016  -1.073  -0.669   1.103   2.000
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.132541   0.112817  -1.175  0.24006
## MomWtGain   -0.036819   0.008389  -4.389 1.14e-05 ***
## MomSmoke1    0.865786   0.309944   2.793  0.00522 **
## MomAge      -0.048266   0.018730  -2.577  0.00997 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.43  on 399  degrees of freedom
## Residual deviance: 516.39  on 396  degrees of freedom
## AIC: 524.39
##
## Number of Fisher Scoring iterations: 4
```

AIC and BIC Comparison

AIC Significant Predictors: Based on the Stepwise Selection with AIC results, `MomWtGain`, `MomSmoke`, and `MomAge` are significant predictors as they have a p-value below the significance level of 0.05.

BIC Significant Predictors: Based on the Stepwise Selection with BIC results, `MomSmoke`, `MomAge`, and `MomWtGain` are significant predictors as they have a p-value below the significance level of 0.05.

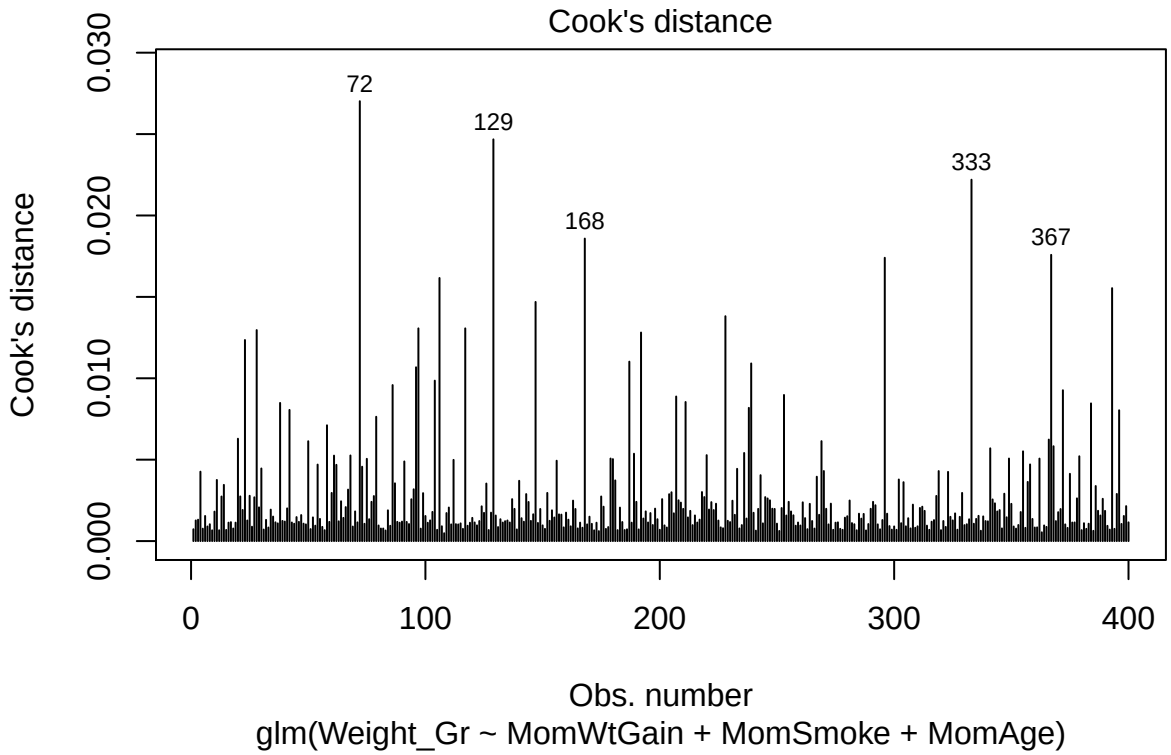
- **Observation:** Both AIC and BIC criteria yielded the same significant predictors. However, the AIC criteria model includes two additional variables in comparison to the BIC criteria model - `Ed` and `Boy`.

Exercise 2.2

Fit the logistic regression with the best model determined by stepwise selection with BIC criteria. Do not leave observation which has cook's d larger than **0.1**. Re-fit the model if necessary. Finally how many observations you use in the final model?

Cook's Distance (Influential Points)

```
plot(step.model.2.bic, which = 4, id.n = 5)
```




```
inf.id.2 = which(cooks.distance(step.model.2.bic) > 0.1)
inf.id.2
```

```
## named integer(0)
```

- **Issues Identified in Model Diagnostics:** There are no observations with a Cook's distance larger than 0.1.
- **How many observations are used in the final model?** All 400 observations are used in the final model as there were no influential points.

Exercise 2.3

Based on your final model, interpret the explicit relationship between response and predictors using Odds Ratio.

```
round(exp(step.model.2.bic$coefficients),3)
```

```
## (Intercept)   MomWtGain   MomSmoke1      MomAge
##       0.876       0.964       2.377       0.953
```

Odds Ratio Interpretation:

- The odds of a woman delivering a low birthweight infant decrease by a factor of exp(-0.048266) = 0.953 with a one unit increase in `MomAge` when all other predictors are held constant.
- The odds of a woman delivering a low birthweight infant decrease by a factor of exp(-0.036819) = 0.964 with a one unit increase in `MomWtGain` when all other predictors are held constant.
- The odds that a woman delivers a low birthweight infant is exp(0.865786) = 2.377 times for the `MomSmoke` 1 group (Smoking Mom) compared to group 0 (Non-Smoking Mom.

Exercise 2.4

Which woman has the high chance to deliver a low birthweight infant?

Based on the Odds Ratio results, A younger woman who weighs less and smokes has a high chance of delivering a low birthweight infant.

Exercise 2.5

What is the sample proportion of low birthweight infant in dataset?

```
fit.prob = predict(step.model.2.bic, type = "response")

sample.prop = mean(birthweight$Weight_Gr)
sample.prop
```

```
## [1] 0.4925
```

- The **sample proportion** of low birthweight infants is 0.4925.

Exercise 2.6

Perform classification with probability cut-off set as sample proportion you answer in (5). What is misclassification rate?

Classification with Sample Proportion Threshold

```
pred.class.2 = ifelse(fit.prob > sample.prop, 1, 0)
head(cbind(birthweight[c("MomSmoke", "MomAge", "MomWtGain", "Weight_Gr")], fit.prob, pred.class.2),10)
```

```
##    MomSmoke MomAge MomWtGain Weight_Gr  fit.prob pred.class.2
## 1         0      4         2         0 0.4014941            0
## 2         0     -5         0         1 0.5271706            1
## 3         0     -7       -10         1 0.6395722            1
## 4         0     -4        18         1 0.3538380            0
## 5         0      5        27         0 0.2029473            0
## 6         0     -6         2         1 0.5208426            1
## 7         0     -1        17         0 0.3295566            0
## 8         0     -1       -12         1 0.5884467            1
## 9         0      1        10         0 0.3660980            0
## 10        1     -2        -2         1 0.7116506            1
```

Misclassification Rate

```
mean(birthweight$Weight_Gr != pred.class.2)
```

```
## [1] 0.355
```

- The **misclassification rate** is 0.355.

Exercise 2.7

Comment on Goodness of fit test and make a conclusion

Hosmer-Lemeshow Test

```
hoslem.test(step.model.2.bic$y, fitted(step.model.2.bic), g=10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  step.model.2.bic$y, fitted(step.model.2.bic)
## X-squared = 9.2068, df = 8, p-value = 0.3252
```

- **Goodness of Fit:** The Hosmer-Lemeshow Test yielded a p-value of 0.3252 which is above the significance level of 0.05. We **do not** reject the null; therefore, the **model is adequate**.

Exercise 3

Compare results from Exercise 1-2 and comment on different or similar conclusions from each analysis. Low birthweight is a risk factor that can lead to infant mortality. If you want to implement a low-birthweight prevention program, what would you suggest to pregnant women?

Exercise 1 Results

Results: On average, Black women who smoke give birth to infants that weigh less than women who are not Black and don't smoke when `MomWtGain` (Mother's weight gain during pregnancy) is the same.

Exercise 2 Results

Results: A younger woman who weighs less and smokes has a high chance of delivering a low birthweight infant.

What would you suggest to pregnant women?

Based upon these results, I would suggest for pregnant women to avoid smoking and to have a healthy weight gain during their pregnancy.