Final Exam Programming

Jason Louwagie

12/9/2020

#### Use significance levels of .05 unless the instructions state otherwise.

# Execise 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).

### (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.

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

### Forward selection with 0.01 p-value criteria for entry

### Backward selection with 0.01 p-value criteria for stay

### Adjusted R-squared criteria

lm1 <- lm(Weight ~ Black + Married + Boy + MomSmoke + Ed + MomAge + MomWtGain, data = birthweight)  
  
# stepwise selection  
model.stepwise<-ols\_step\_both\_p(lm1, 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 21.6870 6201.2327 559.8163   
## 2 MomSmoke addition 0.107 0.102 15.5910 6195.4042 555.0626   
## 3 Black addition 0.127 0.120 8.3540 6188.2794 549.4600   
## -----------------------------------------------------------------------------------------

# forward selection  
model.forward<-ols\_step\_forward\_p(lm1, 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 21.6867 6201.2327 559.8163   
## 2 MomSmoke 0.1069 0.1024 15.5915 6195.4042 555.0626   
## 3 Black 0.1271 0.1205 8.3537 6188.2794 549.4600   
## -----------------------------------------------------------------------------

# backward selection  
model.backward<-ols\_step\_backward\_p(lm1, prem = 0.01, details = F)  
model.backward

##   
##   
## Elimination Summary   
## ---------------------------------------------------------------------------  
## Variable Adj.   
## Step Removed R-Square R-Square C(p) AIC RMSE   
## ---------------------------------------------------------------------------  
## 1 MomAge 0.1445 0.1314 6.3860 6186.2385 546.0372   
## 2 Married 0.1409 0.130 6.0337 6185.9146 546.4876   
## 3 Ed 0.1364 0.1277 6.0624 6185.9688 547.1986   
## 4 Boy 0.1271 0.1205 8.3537 6188.2794 549.4600   
## ---------------------------------------------------------------------------

model.best.subset<-ols\_step\_best\_subset(lm1)  
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   
## -------------------------------------------------------------  
##   
## 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 21.6867 6201.2327 5065.9089 6213.2071 125357705.0611 314961.2141 789.4062 0.9199   
## 2 0.1069 0.1024 0.091 15.5915 6195.4042 5060.1118 6211.3700 123238567.9566 310405.1540 778.0163 0.9066   
## 3 0.1271 0.1205 0.1059 8.3537 6188.2794 5053.1215 6208.2368 120764046.4868 304925.3190 764.3196 0.8906   
## 4 0.1364 0.1277 0.1114 6.0624 6185.9688 5050.9172 6209.9176 119772826.2237 303169.1473 759.9653 0.8854   
## 5 0.1409 0.1300 0.1116 6.0337 6185.9146 5050.9450 6213.8549 119462530.2702 303128.3995 759.9203 0.8853   
## 6 0.1445 0.1314 0.1103 6.3860 6186.2385 5051.3586 6218.1702 119266462.8033 303374.3231 760.6035 0.8860   
## 7 0.1453 0.1300 0.1052 8.0000 6187.8448 5053.0197 6223.7680 119453866.2656 304595.5958 763.7420 0.8896   
## --------------------------------------------------------------------------------------------------------------------------------------------------  
## 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

#### Our final model for stepwise selection is lm(Weight ~ MomWtGain + MomSmoke + Black, data = birthweight)

#### Our final model for forward selection is lm(Weight ~ MomWtGain + MomSmoke + Black, data = birthweight)

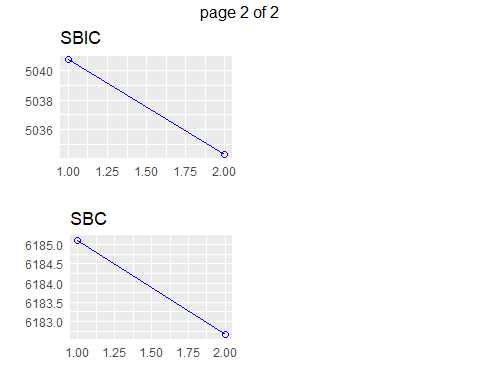
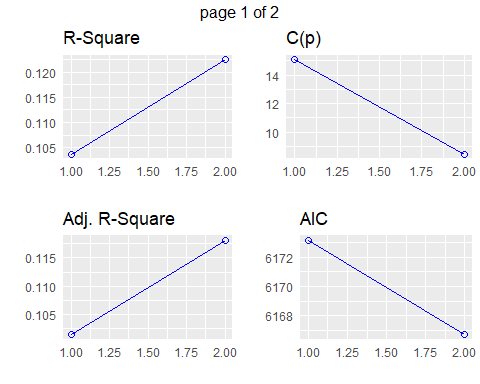
#### Our final model for backward selection is lm(Weight ~ MomWtGain + MomSmoke + Black + Weight\_Gr + Visit, data = birthweight)

#### Our final model for Adj-R Squared selection is lm(Weight ~ MomWtGain + MomSmoke + Black + Married + Boy + Ed, data = birthweight)

#### As we look at the final models for each selection process, we can see that all final models have MomWtGain, MomSmoke, Black. We can also observe that stepwise and forward selection have the same best models. We should also take note that while these two model have the same best model, the best model for Backward and Adjusted R-Sqaure contain 5 and 6 variables, respectively.

### (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?

lm2 <- lm(Weight ~ MomWtGain + MomSmoke + Black, data = birthweight)  
  
cook <- which(cooks.distance(lm2)>.115)  
  
lm2 <- lm(Weight ~ MomWtGain + MomSmoke + Black, data = birthweight[-cook,])  
  
model.stepwise2 <- ols\_step\_both\_p(lm2, pent = 0.01, prem = 0.01, details = FALSE)  
  
plot(model.stepwise2)



#### From the plots, the only issue present is that the R-Squared and the Adj-R Sqaure plots do not level off as they increase.

### (3) How much of the variation in Weight is explained by the final model?

summary(lm2)

##   
## Call:  
## lm(formula = Weight ~ MomWtGain + MomSmoke + Black, data = birthweight[-cook,   
## ])  
##   
## 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

#### From our final model, we can see an R-Squared output of 0.1301. This means the model describes 13.01% of the variation in Weight.

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

#### Individual T-Test for predictors

##### H0: Bx = 0 (No linear relationship)

##### Ha: Bx != 0 (Liner Relationship)

#### From the summary, we can see that MomWtGain and MomSmoke have p-values that are less than our significance level of .01; however, the variable Black has a p-value that is larger than .01. Due to this, for MomWtGain and MomSmoke we reject the null hypothesis, and for Black we fail to reject the null hypothesis. MomWtGain and MomSmoke do not equal zero and have a linear relationship, while Black equals zero and does not have a linear relationship.

# 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.

### (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 criteria

### Stepwise selection with BIC criteria

glm.f.null <- glm(Weight\_Gr~1, data = birthweight, family = "binomial")  
glm.f.full <- glm(Weight\_Gr ~ .-Weight, data = birthweight, family = "binomial")  
  
step.models.AIC<-step(glm.f.null, scope = list(upper=glm.f.full),  
 direction="both",test="Chisq", trace = F)   
  
step.models.BIC<-step(glm.f.null, scope = list(upper=glm.f.full),  
 direction="both",test="Chisq", trace = F, k=log(nrow(birthweight)))  
  
summary(step.models.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

summary(step.models.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

#### From our summaries, we can observe that the AIC final model is slightly different from the BIC final model. The AIC final model is glm(Weight\_Gr ~ MomWtGain + MomSmoke + MomAge + Boy + Ed, family = “binomial”, data = birthweight). Our BIC final model is glm(Weight\_Gr ~ MomWtGain + MomSmoke + MomAge, family = “binomial”, data = birthweight).

#### From these two final models, we can make note that both models contain the predictors MomWtGain, MomSmoke, and MomAge. We can also make note that the AIC final model has more predictors, those predictors being Boy and Ed.

### (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?

Q2 <- glm(Weight\_Gr ~ MomWtGain + MomSmoke + MomAge, family = "binomial", data = birthweight)  
  
cook2 <- which(cooks.distance(Q2)>.1)  
  
nobs(Q2)

## [1] 400

#### Looking at our cook’s distance formula, we can see that there are no influential points and refitting the model is not neccesary. Oue final model will have 400 observations.

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

round(exp(Q2$coefficients),2)

## (Intercept) MomWtGain MomSmoke1 MomAge   
## 0.88 0.96 2.38 0.95

summary(step.models.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

#### From out results we can observe, that the odds of low birthweight changes by a factor of exp(-.0368) = .96 with one unit increase of MomWtGain, by a factor of exp(.866) = 2.38 with one unit increase of MomSmoke, by a factor of exp(-.0483) = .95 with one unit increase of MomAge.

### (4) Which woman has the high chance to deliver a low birthweight infant? For example, answer will be like “a married, high-educated, and older woman has the high chance to deliver a low birthweight infant.”

round(exp(Q2$coefficients),2)

## (Intercept) MomWtGain MomSmoke1 MomAge   
## 0.88 0.96 2.38 0.95

#### From our results, we can conclude that a woman who has lost weight during preganacy, who smokes, and who is younger has a higher chance to deliver a low birthweight infant.

### (5) What is the sample proportion of low birthweight infant in dataset?

sample.prop <- mean(birthweight$Weight\_Gr)  
  
sample.prop

## [1] 0.4925

#### From our results, we can see that the sample proportion of low birthweight is .4925

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

fit.prob <- predict(step.models.BIC, type = "response")  
pred.class <- ifelse(fit.prob > sample.prop, 1, 0)  
  
mean(birthweight$Weight\_Gr != pred.class)

## [1] 0.355

#### We can see that our misclassification rate is .355 or 35.5%.

### (7) Comment on Goodness of fit test and make a conclusion

### For Lemeshow Goodness-Of-Fit

#### H0: Model is adequate

#### H1: Model is not adequate

hoslem.test(step.models.BIC$y, fitted(step.models.BIC), g=10)

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

#### From the Hosmer and Lemeshow test, we can observe a p-value that is greater than our significance level of .05. Due to this, we fail to reject the null hypothesis. The model is adequate.

# Exercise 3

## Compare results from Exercise 1-2 and comment on different or similar conclusions from each analysis.

#### By observing the results of exercises 1 and 2, we can take note of some similarities and differences. We can observe that both final models contain the predictors MomWt and MomSmoke. We can also see that each final model has 3 predictors in their final models. Some differences we can observe are that exercise 1 has 2 influential points while exercise 2 has zero influential points. The third predictor in exercise 1 is Black, while the third predictor in exercise 2 is MomAge.

## Low birthweight is a risk factor that can lead infant mortality. If you want to implement a low-birthweight prevention program, what would you suggest to pregnant women?

#### If I were to implement a low-birthweight prevention program, I would suggest pregnant women to do thier best to maintain thier weight during pregnancy and to not smoke.