STOR 455 Practice Exam Solutions

library(readr)  
library(leaps)  
library(bestglm)  
library(MASS)  
  
BirthWeight <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/BirthWeight.csv")  
abalone\_train <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/abalone\_train.csv")  
abalone\_test <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/abalone\_test.csv")

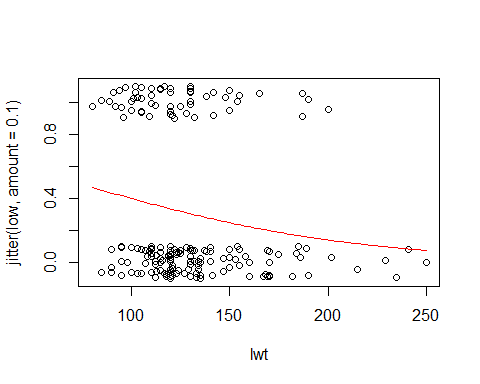
#### Question 1

Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. Behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight. Data were collected at Baystate Medical Center, Springfield, Massachusetts, in 1986 for variables (shown in the table below) that have been shown to be associated with low birth weight in the obstetrical literature.

| Variable | Description |
| --- | --- |
| low | indicator of child’s birth weight less than 2.5 kg. |
| age | mother’s age in years. |
| lwt | mother’s weight in pounds at last menstrual period. |
| race | mother’s race (1 = white, 2 = black, 3 = other). |
| smoke | smoking status during pregnancy. |
| ptl | number of previous premature labours. |
| ht | history of hypertension |
| ui | presence of uterine irritability. |
| ftv | number of physician visits during the first trimester. |
| bwt | child’s birth weight in grams. |

1. Construct and plot a model using the indicator for a child’s low birth weight, *low*, as the response variable, and the mother’s weight in pounds at last menstrual period, *lwt*, and the predictor.

mod1A=glm(low~lwt, data=BirthWeight, family="binomial")  
plot(jitter(low, amount=.1)~lwt, data=BirthWeight)  
  
logit = function(B0, B1, x)  
{  
 exp(B0+B1\*x)/(1+exp(B0+B1\*x))  
}  
  
B0 = summary(mod1A)$coef[1]  
B1 = summary(mod1A)$coef[2]  
  
curve(logit(B0, B1, x),add=TRUE, col="red")



1. Construct a model using the indicator for a child’s low birth weight, *low*, as the response variable, and the mother’s weight in pounds at last menstrual period, age, smoking status during pregnancy, and race as the predictor variables.

mod1B=glm(low~lwt+age+smoke+factor(race), data=BirthWeight, family="binomial")

1. Is there evidence to suggest that the model constructed in part (B) is significantly better than the model constructed in part (A)? Conduct the appropriate hypothesis test. State hypotheses, and provide a conclusion in the context of the data. *6 pts*

Null: The coefficients for the variables age, smoke, and (both dummy) races are 0;  
Alternative: The coefficients for at least one of the variables age, smoke, and race are not 0.  
Statistically significant evidence suggests that at least one of the additional terms has a nonzero coefficient, thus making for a better model than the one with a single predictor.

anova(mod1A, mod1B, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: low ~ lwt  
## Model 2: low ~ lwt + age + smoke + factor(race)  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 187 228.69   
## 2 183 214.58 4 14.113 0.006942 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#You may instead examine the summaries of these models to find the difference between the residual deviances, and compare this G statistic to the chi-squared distribution with 4 df.  
  
summary(mod1A)

##   
## Call:  
## glm(formula = low ~ lwt, family = "binomial", data = BirthWeight)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0951 -0.9022 -0.8018 1.3609 1.9821   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.99831 0.78529 1.271 0.2036   
## lwt -0.01406 0.00617 -2.279 0.0227 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 234.67 on 188 degrees of freedom  
## Residual deviance: 228.69 on 187 degrees of freedom  
## AIC: 232.69  
##   
## Number of Fisher Scoring iterations: 4

summary(mod1B)

##   
## Call:  
## glm(formula = low ~ lwt + age + smoke + factor(race), family = "binomial",   
## data = BirthWeight)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5173 -0.9065 -0.5865 1.3035 2.0401   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.332452 1.107672 0.300 0.76407   
## lwt -0.012526 0.006386 -1.961 0.04982 \*   
## age -0.022478 0.034170 -0.658 0.51065   
## smoke 1.054439 0.380000 2.775 0.00552 \*\*  
## factor(race)2 1.231671 0.517152 2.382 0.01724 \*   
## factor(race)3 0.943263 0.416232 2.266 0.02344 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 234.67 on 188 degrees of freedom  
## Residual deviance: 214.58 on 183 degrees of freedom  
## AIC: 226.58  
##   
## Number of Fisher Scoring iterations: 4

G = summary(mod1A)$deviance - summary(mod1B)$deviance  
1 - pchisq(G, 4)

## [1] 0.006941683

1. Use one of the model selection procedures covered in class to determine the best model to predict the indicator for a child’s low birth weight, *low*.

The “best” models produced each way look a bit different, since AIC and BIC values are not directly comparable. Each method has positives and negatives. The models are quite different! BIC penalizes model complexity more heavily, hence the “best” models have fewer terms.

#With bestglm  
  
# Must factor race so it is considered categorical  
  
BirthWeight$race = as.factor(BirthWeight$race)  
  
# Must move low to last column and remove bwt.   
  
# bwt is the baby's birth weight, which will directly correspond  
# to low and cause an error. When predicting if a baby has low   
# birth weight, you won;t know their birth weight first  
  
BirthWeight\_forbestglm = BirthWeight[,c(2:9, 1)]  
head(BirthWeight\_forbestglm)

## # A tibble: 6 x 9  
## age lwt race smoke ptl ht ui ftv low  
## <dbl> <dbl> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 19 182 2 0 0 0 1 0 0  
## 2 33 155 3 0 0 0 0 3 0  
## 3 20 105 1 1 0 0 0 1 0  
## 4 21 108 1 1 0 0 1 2 0  
## 5 18 107 1 1 0 0 1 0 0  
## 6 21 124 3 0 0 0 0 0 0

#This line is sometimes need to restore structure to the dataframe  
BirthWeight\_forbestglm = as.data.frame(BirthWeight\_forbestglm)  
  
bestglm1D = bestglm(BirthWeight\_forbestglm, family=binomial)

## Morgan-Tatar search since family is non-gaussian.

## Note: factors present with more than 2 levels.

bestglm1D$BestModels

## age lwt race smoke ptl ht ui ftv Criterion  
## 1 FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE 231.6256  
## 2 FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE 231.6890  
## 3 FALSE TRUE FALSE FALSE FALSE TRUE TRUE FALSE 232.3381  
## 4 FALSE TRUE FALSE TRUE FALSE TRUE FALSE FALSE 232.5830  
## 5 FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE 233.1344

#With stepAIC  
  
BirthWeight\_foretepAIC = BirthWeight[1:9]  
  
# I factored the variable race first above.   
# You could have factored it inside of the glm() function.  
  
mod1D = glm(low~., data=BirthWeight\_foretepAIC, family="binomial")  
none = glm(low~1, data=BirthWeight\_foretepAIC, family="binomial")  
  
# You could have dine this with backwards or forwards as well.  
  
stepAIC(none, scope = list(upper = mod1D), trace=0)

##   
## Call: glm(formula = low ~ ptl + lwt + ht + race + smoke + ui, family = "binomial",   
## data = BirthWeight\_foretepAIC)  
##   
## Coefficients:  
## (Intercept) ptl lwt ht race2 race3   
## -0.08655 0.50321 -0.01591 1.85504 1.32572 0.89708   
## smoke ui   
## 0.93873 0.78570   
##   
## Degrees of Freedom: 188 Total (i.e. Null); 181 Residual  
## Null Deviance: 234.7   
## Residual Deviance: 202 AIC: 218

#### Question 2

Abalone are marine gastropod molluscs, which means they are marine snails.The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

| Variable | Description |
| --- | --- |
| Sex | M, F |
| Length | longest shell measurement in mm |
| Diameter | perpendicular to length in mm |
| Height | with meat in shell in mm |
| Whole weight | whole abalone in g |
| Shucked weight | weight of meat in g |
| Viscera weight | gut weight (after bleeding) in g |
| Shell weight | after being dried in g |
| Rings | number of rings |

1. Construct a model to predict abalones’ age (using *rings* as the response) with the lowest Mallow’s Cp using any/all of the variables in the *abalone train* dataset. Do not use transformations, or second or greater order terms, or perform an analysis of the residuals.

# model with regsubsets  
  
regsubsets2A=regsubsets(rings~., data=abalone\_train)  
  
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/ShowSubsets.R")  
ShowSubsets(regsubsets2A)

## sexM length diameter height weight\_whole weight\_shucked weight\_viscera  
## 1 ( 1 )   
## 2 ( 1 ) \* \*   
## 3 ( 1 ) \* \* \*  
## 4 ( 1 ) \* \* \* \*  
## 5 ( 1 ) \* \* \* \* \*  
## 6 ( 1 ) \* \* \* \* \*  
## 7 ( 1 ) \* \* \* \* \* \*  
## 8 ( 1 ) \* \* \* \* \* \* \*  
## weight\_shell Rsq adjRsq Cp  
## 1 ( 1 ) \* 19.40 19.29 227.63  
## 2 ( 1 ) 35.60 35.41 44.03  
## 3 ( 1 ) 37.49 37.22 24.31  
## 4 ( 1 ) 39.05 38.70 8.45  
## 5 ( 1 ) 39.53 39.10 4.96  
## 6 ( 1 ) \* 39.69 39.17 5.11  
## 7 ( 1 ) \* 39.70 39.09 7.03  
## 8 ( 1 ) \* 39.70 39.00 9.00

mod2A = lm(rings~length+height+weight\_whole+weight\_shucked+weight\_viscera, data=abalone\_train)  
  
# model with step - produces the same best model  
  
none2 = lm(rings~1, data=abalone\_train)  
full = lm(rings~., data=abalone\_train)  
MSE = (summary(full)$sigma)^2  
  
step(none2,scope=list(upper=full),scale=MSE, trace=0)

##   
## Call:  
## lm(formula = rings ~ weight\_shucked + weight\_whole + height +   
## weight\_viscera + length, data = abalone\_train)  
##   
## Coefficients:  
## (Intercept) weight\_shucked weight\_whole height weight\_viscera   
## 4.331 -24.898 14.057 20.404 -14.205   
## length   
## 5.742

#model with forward - produces slightly different model  
  
step(none2,scope=list(upper=full),scale=MSE, direction="forward", trace=0)

##   
## Call:  
## lm(formula = rings ~ weight\_shell + weight\_shucked + weight\_whole +   
## height + weight\_viscera + length, data = abalone\_train)  
##   
## Coefficients:  
## (Intercept) weight\_shell weight\_shucked weight\_whole height   
## 4.534 3.942 -22.957 11.906 18.939   
## weight\_viscera length   
## -12.338 5.383

#model with backward - produces the same best model  
  
step(full,scale=MSE, trace=0)

##   
## Call:  
## lm(formula = rings ~ length + height + weight\_whole + weight\_shucked +   
## weight\_viscera, data = abalone\_train)  
##   
## Coefficients:  
## (Intercept) length height weight\_whole weight\_shucked   
## 4.331 5.742 20.404 14.057 -24.898   
## weight\_viscera   
## -14.205

1. A second dataset, abalone\_test, contains additional data for 500 more abalone. Use this dataset, and your model constructed in part (A), to perform a cross validation analysis of your model. Calcuate and comment on the cross-validation correlation, shrinkage, and analysis of holdout residuals. Does the model constructed in part (A) appear to be similarly effective for predicting the number of rings for abalone?

Holdout residual mean relatively close to zero (close is relative)  
Holdout standard deviation is very similar to the standard error of the regression line for the original model constructed from the training data.  
The shape of the holdout residulas is approximately normally distributed, but might indicate a slight bias and the center seems to be shifted left.  
Shrinkage is near 0.10, which isn’t as small as it could be, but suggests that the model predicts the new data similarly as well as the old.

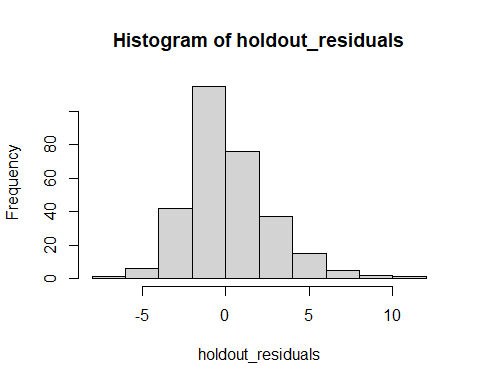
fit = predict(mod2A, abalone\_test)  
  
holdout\_residuals = abalone\_test$rings - fit  
  
mean(holdout\_residuals)

## [1] 0.1286006

sd(holdout\_residuals)

## [1] 2.596266

hist(holdout\_residuals)



summary(mod2A)

##   
## Call:  
## lm(formula = rings ~ length + height + weight\_whole + weight\_shucked +   
## weight\_viscera, data = abalone\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8210 -1.5879 -0.3408 1.0211 11.9233   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.3314 0.9636 4.495 8.16e-06 \*\*\*  
## length 5.7423 2.4497 2.344 0.019357 \*   
## height 20.4039 5.7006 3.579 0.000369 \*\*\*  
## weight\_whole 14.0574 1.1463 12.263 < 2e-16 \*\*\*  
## weight\_shucked -24.8985 1.5680 -15.879 < 2e-16 \*\*\*  
## weight\_viscera -14.2047 2.8975 -4.902 1.18e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.439 on 694 degrees of freedom  
## Multiple R-squared: 0.3953, Adjusted R-squared: 0.391   
## F-statistic: 90.74 on 5 and 694 DF, p-value: < 2.2e-16

cv\_corr = cor(fit, abalone\_test$rings)  
  
summary(mod2A)$r.squared - cv\_corr^2

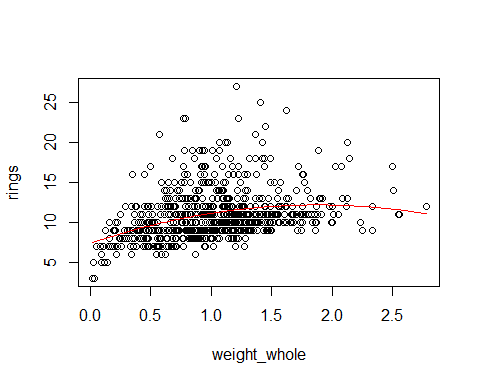
## [1] 0.09549159

1. Linearity is an issue in any abalone model that uses the various measures of weight to predict the number of rings. Would a polynomial model be more appropriate? Contruct a quadratic model using *rings* as the response and the *weight whole* as the predictor. Plot the data and the curve on the same axes. Use the *abalone train* dataset.

mod2C = lm(rings~weight\_whole+I(weight\_whole^2), data=abalone\_train)  
  
summary(mod2C)

##   
## Call:  
## lm(formula = rings ~ weight\_whole + I(weight\_whole^2), data = abalone\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8056 -1.9640 -0.8028 1.0129 15.4238   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.3778 0.4598 16.045 < 2e-16 \*\*\*  
## weight\_whole 5.1079 0.8353 6.115 1.61e-09 \*\*\*  
## I(weight\_whole^2) -1.3569 0.3569 -3.802 0.000156 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.952 on 697 degrees of freedom  
## Multiple R-squared: 0.1104, Adjusted R-squared: 0.1079   
## F-statistic: 43.26 on 2 and 697 DF, p-value: < 2.2e-16

a = summary(mod2C)$coef[3,1]  
b = summary(mod2C)$coef[2,1]  
c = summary(mod2C)$coef[1,1]  
  
plot(rings~weight\_whole, data=abalone\_train)  
curve(a\*x^2 + b\*x + c, add=TRUE, col="red")



1. Consider a model that uses the log( *rings* ) as the response variable. The predictor variables for the model are *diameter*, *length*, *sex*, and the interactions between *sex* and each other predictor variable. Perform a hypothesis test to determine if the model including the interaction terms is significantly better than a model including the same variables but without the interactions. Include the hypotheses and conclusion.

null: coefficients for the interaction terms are 0;  
alternative: the coefficients for at least one interaction term is nonzero.  
Since the p-value is small, there is significant evidence to suggest that at least one of the interaction terms have a nonzero coefficient.

mod2D1 = lm(log(rings)~length+diameter+sex+length\*sex+diameter\*sex, data=abalone\_train)  
  
mod2D2 = lm(log(rings)~length+diameter+sex, data=abalone\_train)  
  
anova(mod2D2, mod2D1)

## Analysis of Variance Table  
##   
## Model 1: log(rings) ~ length + diameter + sex  
## Model 2: log(rings) ~ length + diameter + sex + length \* sex + diameter \*   
## sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 696 40.884   
## 2 694 40.153 2 0.73099 6.3172 0.001911 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1