HW 8
STAT 672, Summer 2015

1) (0 points) Do Exercise 1 on p. 259 of \textit{ISL}.

2) For this problem, you can merely submit the things that I specifically request in the various parts, or you can submit some of your work in addition to the answers, but if you do that be sure to \textbf{highlight the specific things I request in yellow}.

Use the following lines of code to read in a data set and and divide it into a training set and a test set. Be sure to set the seed of $R$'s random number generator to 123 right before you create the \texttt{train} vector. This will insure that everyone will split the data into the same two parts.

\begin{verbatim}
lbw = read.table("http://mason.gmu.edu/~csutton/lbw.txt", header=TRUE)
set.seed(123)
train = sample(189,100,replace=FALSE)
lbw.train=lbw[train,]
lbw.test=lbw[-train,]
\end{verbatim}

This code will first read in a data set of 189 cases. Then a subset of size 100 is “randomly” chosen from \{1,2,\ldots,189\} and stored in the vector \texttt{train}. The next two lines put 100 cases into a data frame called \texttt{lbw.train}, and put the other 89 cases into a data frame called \texttt{lbw.test}. In order to check things enter

\begin{verbatim}
dim(lbw.train)
head(lbw.train)
dim(lbw.test)
head(lbw.test)
\end{verbatim}

You should see that the dimension of \texttt{lbw.train} is 100 by 11, and that the first 3 values of \texttt{ID} are 142, 30, and 169. You should also see that the dimension of \texttt{lbw.test} is 89 by 11, and that the first 3 values of \texttt{ID} are 86, 87, and 89.

For the purpose of this assignment, let’s pretend that all we have are the 100 cases of the training data set, and also that we only have \texttt{AGE} (age of mother), \texttt{LWT} (weight of mother at last menstrual period), \texttt{SMOKE} (smoking status during pregnancy, coded 1 for yes, and 0 for no), \texttt{PTL} (number of times mother has had premature labor in the past), \texttt{HT} (an indicator of hypertension, coded 1 for yes, and 0 for no), \texttt{UI} (an indicator of uterine irritability, coded 1 for yes, and 0 for no), and \texttt{FTV} (number of physician visits during the first trimester), as predictors, in addition to the response variable \texttt{BWT} (birth weight, in grams). It can be noted that the data sets also include another predictor, \texttt{RACE} (that we won’t use), as well as columns for \texttt{ID} and an indicator of low birth weight, \texttt{LOW}. In order to make our data sets easier to use, eliminate the columns we won’t use by entering the following into $R$.

\begin{verbatim}
lbw.train=lbw.train[,c(3,4,6,7,8,9,10,11)]
lbw.test=lbw.test[,c(3,4,6,7,8,9,10,11)]
\end{verbatim}

Having done all of the things indicated above, check the dimensions of the training and test sets (to make sure they are 100 by 8 and 89 by 8), load the \texttt{boot} library and the \texttt{leaps} library, and set the seed of the random number generator again (just in case something called for below uses a random number):

\begin{verbatim}
dim(lbw.train)
dim(lbw.test)
library(boot)
library(leaps)
set.seed(123)
\end{verbatim}

(You may have to first install the \texttt{leaps} library/package, if you haven’t already done so.)

Find the best subset of predictors for models having from 1 to 7 predictors, look to see what predictors the best model of each “size” includes, and look at the adjusted $R^2$ value for each of the models. This can be done in $R$ by entering the following:

\begin{verbatim}
regfit.best=regsubsets(BWT~.,data=lbw.train)
best.sum=summary(regfit.best)
best.sum
best.sum$adjr2
\end{verbatim}
(a) (3 points) What predictors are used in the model which corresponds to the highest adjusted $R^2$ value?

(b) (3 points) What predictors are used in the model which corresponds to the lowest $C_p$ value? (Note that the very bottom of p. 245 gives the names of the quantities that belong to a summary object resulting from a best subsets procedure. So to see the $C_p$ values for the seven models, enter `best.sum$cp`.)

(c) (3 points) What predictors are used in the model which corresponds to the lowest BIC value?

(d) (1 point) The $C_p$ values supplied by R are not those given by (6.2) on p. 211 of ISL. Rather, they come from using the alternative expression given in the footnote on p. 211, with $d$ being the number of parameters, and not the number of predictors. The Adjusted $R^2$ values supplied by R correspond to to (6.4) on p. 212 of ISL, with $d$ being either the number of predictors or the number of parameters. Which is it? Provide justification for your conclusion, based on values obtained from `best.sum`.

LOOCV can be used to estimate the test MSPE for the best simple regression model as follows:

```r
glm.fit1=glm(BWT~UI,data=lbw.train)
cv.err1=cv.glm(lbw.train,glm.fit1)
cv.err1$delta[2]
```

(e) (2 points) What is the LOOCV estimate of test MSPE for the model fit using the best pair of predictors?

(f) (2 points) What is the LOOCV estimate of test MSPE for the model fit using the best three predictors?

(g) (2 points) What is the LOOCV estimate of test MSPE for the model fit using the best four predictors?

(h) (4 points) Now, using the data in `lbw.test` in the role of a validation set, do something very similar to what is done on the bottom half of p. 248 and the top half of p. 249 of ISL to estimate the test MSPE for each of the seven models in the sequence of models obtained by the best subsets approach earlier. Which predictors are used in the model yielding the lowest of the test MSPE estimates obtained, and what is the value of the corresponding test MSPE estimate? (Note: We can view this part of the exercise in two ways. We can view the MSPE estimates obtained for this part as giving us an unbiased assessment of how the models built from the training sample perform with regard to making new predictions; thus giving us a way to see if Adjusted $R^2$, $C_p$, or BIC led to the best choice of a model. (None of them do, according to the MSPE estimates obtained for this part.) But having only 89 cases in the test set limits the accuracy of any such conclusions. An alternative way of viewing this part is to suppose that we started with all 189 cases, and set aside 89 of them to serve as a validation set to assess the prediction accuracy of a variety of models we build from the other 100 cases; and in the end we’d use all 189 cases to fit the model which appears to be best.)