Assignment 6 - Model Building

your name goes here

Due: Wednesday, March 7, 2018, noon, to Sakai

Summary

Primarily from the topics in Chapter 9 of your text, this homework assignment gives you practice making decisions about which variables to include in a model.

Assignment

1. Two authors wrote as follows: "Our research utilized a multiple regression model. Two of the predictor variables important in our theory turned out the be highly correlated in our data set. This made it difficult to assess the individual effects of each of these variables separately. We retained both variables in our model, however, because the high coefficient of multiple determination makes this difficulty unimportant." Comment.

2. In forward stepwise regression. what advantage is there in using a relatively small α-to-enter value for adding variables? What advantage is there in using a large α-to-enter value?

3. Prepare a flowchart of each of the following selection methods: (a) stepwise (forward-backward) regression. (b) forward (forward only) selection. (c) backward (backward only) elimination. (Feel free to do this with pencil.)

4. First we split the dataset into test and training sets. On the training data - using BIC, AIC, C_p, R^2 or adjusted R^2 - we perform best subset (looking at all possible subsets of the explanatory variables), forward selection, and backward selection on a single data set. For each approach, we obtain $m + 1$ models, containing $0, 1, 2, \ldots, m$ coefficients (that is, we get $m + 1$ different forward models, etc.). Explain your answers:

   (a) Which of the three models with k predictors has the smallest training SSE? (The training SSE is the sum of squares error of the training observations predicted using the model created using the training observations.)

   (b) Which of the three models with k predictors has the smallest test SSE? (The test SSE is the sum of squares error of the test observations predicted using the model created using the training observations.)

   (c) True (always True) or False (not always True):

      i. The predictors in the k-variable model identified by forward selection are a subset of the predictors in the $(k + 1)$-variable model identified by forward selection.

      ii. The predictors in the k-variable model identified by backward selection are a subset of the predictors in the $(k + 1)$-variable model identified by backward selection.

      iii. The predictors in the k-variable model identified by backward selection are a subset of the predictors in the $(k + 1)$-variable model identified by forward selection.

iv. The predictors in the k-variable model identified by forward selection are a subset of the predictors in the $(k+1)$-variable model identified by backward selection.

v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k+1)$-variable model identified by best subset selection.

5. (Following up on HW5...) Data were collected on the volume of users on the Northampton Rail Trail in Florence, Massachusetts. Variables in the data set include the number of crossings on a particular day (measured by a sensor near the intersection with Chestnut Street, volume), the average of the min and max temperature in degrees Fahrenheit for that day (avgtemp), and a dichotomous indicator of whether the day was a weekday or a weekend/holiday (weekday).

```
require(mosaic); require(dplyr)
require(mosaicData)
data(RailTrail)
RailTrail = mutate(RailTrail, daytype = ifelse(weekday==1, "Weekday", "Wkend/Holiday"))
```

Consider the following full (additive) linear model predicting the volume on the Northampton Rail Trail.

```
summary(lm(volume ~ hightemp + lowtemp + cloudcover + precip,
           data=RailTrail))$coef

##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)   35.308293  59.795831  0.5904809 5.564350e-01
## hightemp       6.571283   1.153119  5.6987019 1.700272e-07
## lowtemp       -1.289582   1.386987 -0.9297725 3.551219e-01
## cloudcover    -7.500899   3.850869 -1.9478456 5.473396e-02
## precip      -100.616367  42.064479 -2.3919556 1.896411e-02
```

(a) Calculate the coefficient of partial determination for each value below. Explain what each coefficient measures / interpret your results. Here is the sentence from class:

The coefficient of partial determination measures the marginal contribution of one X variable when all others are already included in the model.

$$R^2_{Y1|2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$$

$R^2_{Y1|2}$ measures the proportionate reduction in "the variation in Y remaining after X_2 is included in the model" that is gained by also including X_1 in the model.

i. $R^2_{Yprecip}$

ii. $R^2_{Ycloudcover,precip}$ (not conditional)

iii. $R^2_{Ycloudcover|precip}$

iv. $R^2_{Ylowtemp|precip,hightemp}$

v. $R^2_{Yhightemp,lowtemp|precip}$

6. In this exercise, we will predict the number of applications received using the other variables in the College data set.

```
require(ISLR); require(rms)
require(dplyr); require(leaps)
data(College)
```

Split the data set into a training set ($\approx 2/3$ of the observations) and a test set ($\approx 1/3$ of the observations). The split is done for you in the following code:

```
set.seed(47)   # feel free to change this number if you want to
col.subset <- sample(c(TRUE, FALSE), nrow(College), replace=TRUE, prob=c(1/3,2/3))
col.tst <- College[col.subset,]
col.trn <-College[!col.subset,]
dim(col.tst)

## [1] 282  18

dim(col.trn)

## [1] 495  18
```

An analysis using best subsets and SSE

Below, `nvmax=3`. You will need to change that argument so as to use all the variables in the dataset. Go through the output below to make sure you understand it.

An asterisk indicates that a given variable is included in the corresponding model. For instance, this output indicates that the best two-variable model contains only `Accept` and `Top10perc`. By default, `regsubsets()` only reports results up to the best eight-variable model. But the `nvmax` option can be used in order to return as many variables as are desired.

The output (`$outmat`) gives the best model for each of 1, 2, 3 variables. We also see that the three variable model is best for all the criteria: Cp, BIC, adjusted R^2.

```
col.best <- regsubsets(Apps ~., data=col.trn, nvmax=3)
col.best.sum <- summary(col.best)
names(col.best.sum)

## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"

col.best.sum$outmat

##          PrivateYes Accept Enroll Top10perc Top25perc F.Undergrad
## 1  ( 1 ) " "        "*"    " "    " "       " "       " "
## 2  ( 1 ) " "        "*"    " "    "*"       " "       " "
## 3  ( 1 ) " "        "*"    " "    "*"       "*"       " "
##          P.Undergrad Outstate Room.Board Books Personal PhD Terminal
## 1  ( 1 ) " "          " "      " "        " "   " "      " " " "
## 2  ( 1 ) " "          " "      " "        " "   " "      " " " "
## 3  ( 1 ) " "          " "      " "        " "   " "      " " " "
##          S.F.Ratio perc.alumni Expend Grad.Rate
## 1  ( 1 ) " "       " "         " "    " "
## 2  ( 1 ) " "       " "         " "    " "
## 3  ( 1 ) " "       " "         " "    " "

which.min(col.best.sum$cp)

## [1] 3
```

```r
which.max(col.best.sum$adjr2)
```

```
## [1] 3
```

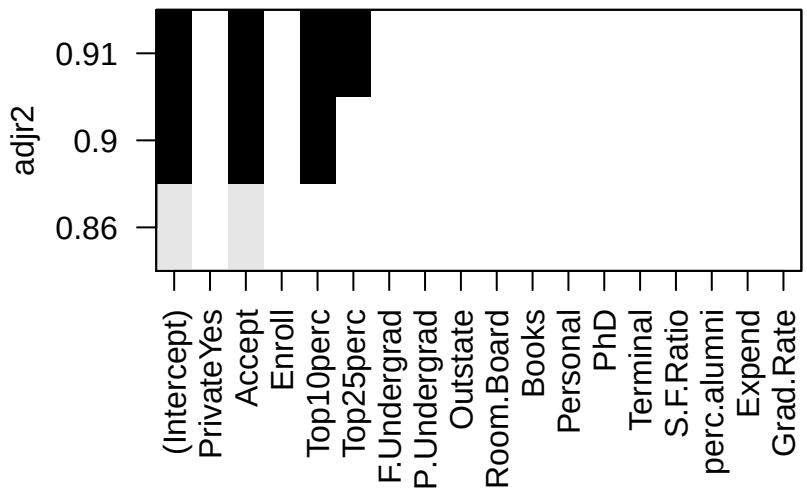
```r
which.min(col.best.sum$bic)
```

```
## [1] 3
```

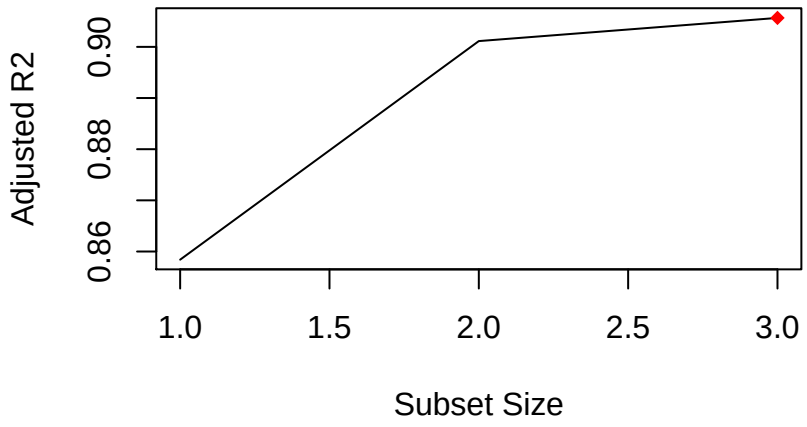
```r
best.num <-  which.max(col.best.sum$adjr2)  # you might choose a different criteria


coef(col.best, best.num)
```

```
## (Intercept)      Accept   Top10perc   Top25perc
## -199.989190    1.421544   72.132516  -29.156976
```

```r
plot(col.best, scale="adjr2")  # scale options are bic, Cp, adjr2, r2
```




```r
plot(col.best.sum$adjr2, xlab = "Subset Size", ylab = "Adjusted R2",  type = "l")
points(best.num, col.best.sum$adjr2[best.num], pch = 18, col = "red")
```



(a) Using the entire dataset, look at some of the relationships between the variables. You might consider using a pairs plot (but pairs on the entire dataset probably isn't readable/interpretable). (`?pairs`) Comment on the exploratory relationships.

(b) Using only the training data, find the best model (out of all possible subsets) - **allowing for as many variables as needed** (no interactions) using Cp, BIC, and adjusted R^2. (Note: you may come up with 1 model, you may come up with 3 different models.)

i. Print the coefficient estimates.

ii. Provide 3 plots: the Adjusted R^2 as a function of subset size, Cp as a function of subset size, and BIC as a function of subset size. Indicate on the plot where the best model is obtained.

iii. For each of the three (or fewer) models, run lm. **Comment** on the significance of the variables chosen by the criteria.

iv. Give a few sentences on which model you would present to the client (and why).

(c) Using nested F-tests with stepwise models, re-assess the college data (only the training data).

i. Using forward selection, which model is selected? (Variables must be added into the model one at a time using `add1`.)

ii. Using backward selection, which model is selected? (Variables must be removed from the model one at a time using `drop1`.)

iii. Which model would you present to the client? Why?

(d) With the two models from above(one best subsets using SSE, the other stepwise using F-tests), apply the model to get the SSE for the test data. Which model gives smaller SSE for test data? Is that what you expected? Why?

[R hint: use the predict function where newdata=col.tst. Then take those predictions and find the sum of squares: `sum((newpreds - col.tst$Apps)^2)`]