Module 8 Lab Submission

Ben Tankus

For this exploration, we will simulate some data and then use the `regsubsets()` function to select models.

The data we simulate below will have some predictor variables (`X1`, `X2`, and `X3`) that contribute to the response, and some other predictor variables (`W1`, `W2`, and `W3`) that do not contribute to the response. The goal is to explore how well these methods of model selection do at identifying the correct variables.

```
n <- 30

set.seed(12345)
X1 <- rnorm(n)
X2 <- rnorm(n)
X3 <- 0.5*X1 + 0.5*X2 + rnorm(n, 0, 0.5)

W1 <- rnorm(n)
W2 <- rnorm(n)
W3 <- 0.4*X1 + 0.3*X2 + rnorm(n, 0, 0.4)


Y <- 1 + 0.3*X1 + 0.3*X2 + 0.5*X3 + rnorm(n)

lab8Data <- data.frame(Y, X1, X2, X3, W1, W2, W3)
```

Note that the true model is

$$Y_i = 1 + 0.3X1_i + 0.3X2_i + 0.5X3_i + \epsilon_i$$

so our hope is that we would select `X1`, `X2`, and `X3`, and not `W1`, `W2`, or `W3`.

1. **Use the `regsubsets()` function to perform best subset selection, modeling `Y` as a function of all the other variables in the data set `lab8Data`. Name the resulting object `lab8.regfit.best`**

```
lab8.regfit.best <-  regsubsets(Y ~ ., data = lab8Data)
```

2. **Use the `summary()` function to summarize the `lab8.regfit.best` object that you got from the previous step, and store the summary object as `lab8.reg.best.summary`. What are the RSS values for the best subsets of each size (1 - 6)?**

```
lab8.reg.best.summary <- summary(lab8.regfit.best)
lab8.reg.best.summary
```

```
## Subset selection object
## Call: regsubsets.formula(Y ~ ., data = lab8Data)
## 6 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## X3     FALSE      FALSE
## W1     FALSE      FALSE
## W2     FALSE      FALSE
## W3     FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          X1  X2  X3  W1  W2  W3
## 1  ( 1 ) " " " " "*" " " " " " "
## 2  ( 1 ) " " " " "*" " " "*" " "
## 3  ( 1 ) "*" " " "*" " " "*" " "
## 4  ( 1 ) "*" "*" "*" " " "*" " "
## 5  ( 1 ) "*" "*" "*" " " "*" "*"
## 6  ( 1 ) "*" "*" "*" "*" "*" "*"
```

```
print("")
```

```
## [1] ""
```

```
print("RSS Values for subsets of each size, 1-6 respectively")
```

```
## [1] "RSS Values for subsets of each size, 1-6 respectively"
```

```
lab8.reg.best.summary$rss
```

```
## [1] 15.85931 15.01322 14.42506 14.30083 14.28906 14.28876
```

3. **Use the `glm()` function to fit the best model with three predictors (in this case, `X1`, `X3`, and `W2`) and then use the `cv.glm()` function to find the LOOCV error estimates for this model. Store the object resulting from `cv.glm()` as `lab8.cv.err`. What is the value of the first element of the `delta` component of this object?**

```
fitGLM <- glm(Y ~ X1 + X3 + W2, data= lab8Data)
lab8.cv.err <- cv.glm(lab8Data, fitGLM)

print('First value of the delta component is:')
```

```
## [1] "First value of the delta component is:"
```

```
lab8.cv.err$delta[1]
```

```
## [1] 0.6057914
```

4. **Repeat the above step, but with the true predictors in the model instead (`X1`, `X2`, and `X3`). How do the delta values compare for the 'best' 3-predictor model vs. the true model?**

```
fitGLM2 <- glm(Y ~ X1 + X3 + X2 , data= lab8Data)
lab8.cv.err2 <- cv.glm(lab8Data, fitGLM2)

fitGLMFull <- glm(Y ~ . , data= lab8Data)
lab8.cv.errFull <- cv.glm(lab8Data, fitGLMFull)

print('With W term (3 predictor):')
```

```
## [1] "With W term (3 predictor):"
```

```
lab8.cv.err$delta
```

```
## [1] 0.6057914 0.6035681
```

```
print('Only x terms (True model):')
```

```
## [1] "Only x terms (True model):"
```

```
lab8.cv.err2$delta
```

```
## [1] 0.6035685 0.6018996
```

```
print('Full model (Xs and Ws):')
```

```
## [1] "Full model (Xs and Ws):"
```

```
lab8.cv.errFull$delta
```

```
## [1] 0.7959964 0.7898601
```

The delta variables are very similar and are identical up to two decimal places between the best predictor model and the true model.