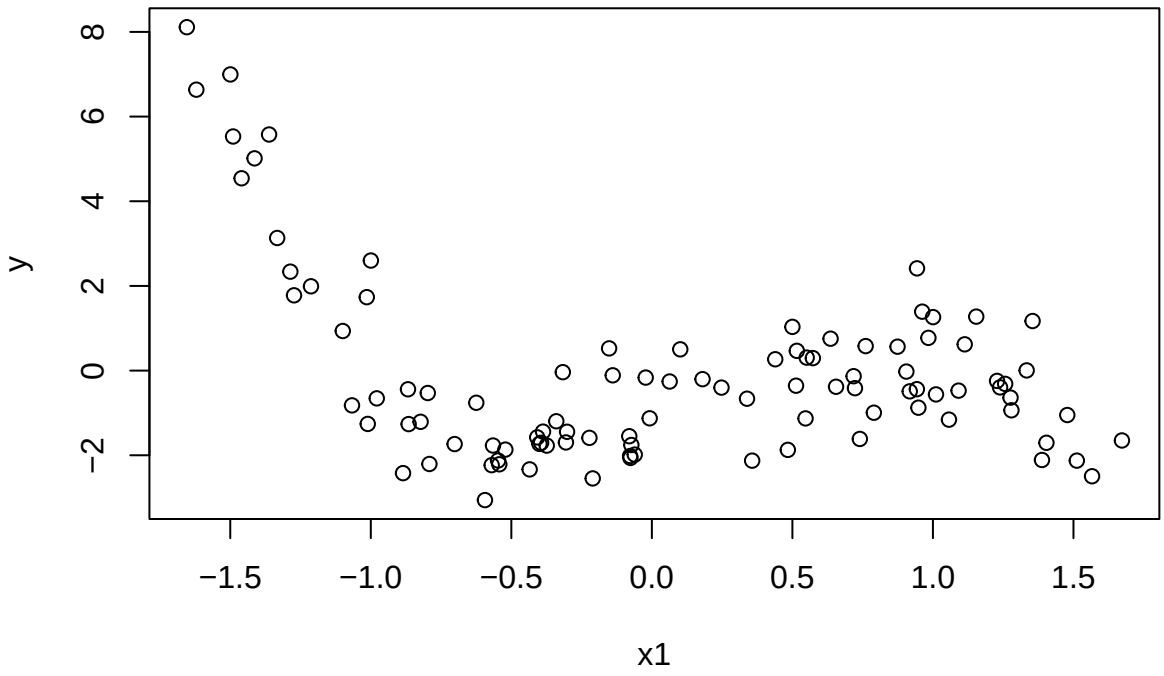
Homework: Ch 06

STAT 4510/7510

Due Thursday, March 24, 11:59 pm

In this HW, we will generate simulated data, and use this data to perform model selection.

```
set.seed(1)
x1 <- runif(100, -1.7, 1.7)
x2 <- x1^2; x3 <- x1^3; x4 <- x1^4; x5 <- x1^5
x6 <- x1^6; x7 <- x1^7; x8 <- x1^8; x9 <- x1^9
x10 <- x1^10
y <- -1.3 + 2*x1 + 1.5*x2 - 2*x3 + rnorm(100)
data_df <- data.frame(y, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10)
data_mat <- as.matrix(data_df)
# We prepare the data set in two different objects: data_df (data frame), data_mat (matrix)
plot(x1, y)
```



Problem 1

(a) According to the code generating the data, which predictor variables, among $x_1, \cdots, x_{10}$, are desired to be found to be associated with the response variable by model selection methods?

(b) Use the function `regsubsets()` in the library `leaps` to perform best subset selection (using `data_df`). Show the summary of the outcome. Which variables are included in the best 1-predictor, 3-predictor, and 5-predictor models?

(c) What is the best model obtained according to C_p, BIC, and adjusted R^2? Show plots for each criterion to provide evidence for your answer.

(d) Using 'coef()', report the coefficients of the best models obtained by C_p, BIC, and adjusted R^2, respectively.

(e) Find the best model using forward stepwise selection. At this time, use only BIC to determine the best model. Which variables are included in the model?

(e) Find the best model using backward stepwise selection. At this time, use only C_p to determine the best model. Which variables are included in the model?

(f) Compare the best model you obtained from best subset selection, forward stepwise, and backward stepwise methods with the true underlying model. Briefly describe advantages and disadvantages of those methods based on what you observed from the outcome.

Problem 2

Now using the data `data_mat`, we will perform the lasso regression model. Prepare a grid of values of λ by running the following code.

```
grid=10^seq(10,-2,length=100)
```

(a) Split the data into training (70%) and testing (30%) sets using the seed number 1. You can simply prepare `train` and `test` vectors that include row indices for each set.

(b) Create `lasso.mod` by fitting lasso regression using the training set. Then use 10-fold CV to find the best λ with the seed number 2. Plot the outcome of cross validation. Which value of λ is the best?

(c) Make predictions for the test set using the fitted model `lasso.mod` with the best λ. Compute the test MSE.

(d) Refit the model on the full data set with the best λ. Extract the regression coefficients estimates. Compare the outcome with the true model.

Problem 3

Now using the data `data_df`, perform the Principal component regression (PCR) model.

(a) Perform model training for PCR using the training set. Use cross validation to determine the number of principal components to be used, with the seed number `3`. Show the summary of fit and validation plot. How many PC is the best?

(b) Test the model with the test set using the best number of PC. Compute test MSE.

(c) Refit the PCR on the full data set with the best number of PC. Show the summary of the fit. How much of the variability of predictors is explained by the PCs? How much of the variability of response variable is explained by the PC regression?