STT 810

Homework 6

Due Thursday, December 1 at 11:59:59pm

1. A linear regression model prediction is in the form of $y = 2.18\ x_1 - 4.56x_2$. The model is built using 18 data points. The root mean square error of the residuals is 0.856.

(a) What is the $E(y \mid (x_1,x_2) = (2, -3))$?

(b) What is the probability that $y > 20$, given that $(x_1,x_2) = (2, -3)$?

2. Take the Boston dataset, available in D2L. This data has information about different neighborhoods in Boston, and we will use it to predict the median housing price for the neighborhoods. Here is an explanation of the variables:

CRIM: Per capita crime rate by town

ZN: Proportion of residential land zoned for lots over 25,000 sq. ft

INDUS: Proportion of non-retail business acres per town

CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

NOX: Nitric oxide concentration (parts per 10 million)

RM: Average number of rooms per dwelling

AGE: Proportion of owner-occupied units built prior to 1940

DIS: Weighted distances to five Boston employment centers

RAD: Index of accessibility to radial highways

TAX: Full-value property tax rate per $10,000

PTRATIO: Pupil-teacher ratio by town

B: $1000(Bk - 0.63)^2$, where Bk is the proportion of [people of African American descent] by town

LSTAT: Percentage of lower status of the population

MEDV: Median value of owner-occupied homes in $1000s

(a) Create a pairs plot with ggpairs for the data.
(b) Which other variable correlates the strongest with medv? Note that strongest mean largest absolute value, whether it's positive or negative.
(c) Build a simple linear regression model with that variable as the x, with
    a. Constructing the normal equations $A^TAx = A^Tb$, and solving for the coefficient vector x.

    b. Using lm. Do the coefficients agree?

(d) Next, build a linear regression model with the 2 other variables in addition that correlate most strongly with medv, using lm (so 3 variables total).
    a. How do the adjusted R-squared values compare between the models?
    b. Comment on the significance of the coefficients
    c. Are the coefficients in the correct direction? Be sure to explain.

(e) Now, re-do the regression from (c), except this time
    a. Split the data into training and testing datasets (70/30 split)
    b. Build the model on the training dataset
    c. Use the model to predict values for the test dataset
    d. How do the results compare, in terms of RMSE?

(f) For the linear regression in (b) re-do the linear regressions in the following ways:
    a. Construct a linear model with the variable with the strongest correlation, like in 4(b), but this time do the minimization of RMSE directly, with the optim function. Check to make sure the result matches what you got in the previous homework.
    b. Next construct a linear model of the same form, with mean absolute error as the loss function, and then use optim to find the fit. Compare the model with what you did in the previous part.

(g) Use maximum likelihood to construct the linear regression model from (b). Do the coefficients agree?

(h) For the model in (c), what are the 95% confidence intervals for the parameters, according to the t-values?

(i) For the model in (c), what are the 95% confidence intervals for the parameters, using
    a. regular bootstrapping, and
    b. Bayesian bootstrapping?