Homework 2 R markdown

Josh Jarvey

2020-09-09

#### **Intellectual Property:**

These problems are the intellectual property of the instructors and may not be reproduced outside of this course.

#### **Due Date:**

Tuesday, Sep 19, 2017 at 11:59 PM

## Problem 1: Model Assessment

This problem practices application of proper model assessment techniques, with a multiple linear regression model.

Download the data set *Trees.csv* [from Lesson 2 on D2L] and read it into R. Reference with description of the *original* measurements may be found at: <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/trees.html>

## Warning: package 'readr' was built under R version 3.6.2

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:  
## cols(  
## X1 = col\_double(),  
## Volume = col\_double(),  
## Girth = col\_double(),  
## Height = col\_double(),  
## GirthHeight = col\_double(),  
## Girth2 = col\_double(),  
## Girth2Height = col\_double()  
## )

The general goal for this dataset is to predict Volume based on Girth and Height. We will be fitting a predictive model using multiple linear regression. The model is given below:   
Note that there are five predictors, some of which are transformations of the original two variables Girth and Height, for predicting the value of the response variable Volume.

### Question 1 **(2 points)**

Why is *Volume* the most reasonable response variable? *Include real-world reasons (eg. physical practicalities) in your discussion.*

Volume is the most reasonable response variable because as the girth of a tree expands, and the height of a tree grows, it is taking up more area in space. In a 3 dimensional world, this means more volume and therefore a natural relationship should exist between girth/height and volume.

**Text Answer**:

### Questions 2 **(1 point)**

Use multiple linear regression fit the model to the full data set. Identify the coefficient estimates (, , , , ) for the five predictor terms. Recall: The t-statistic tests for the marginal significance of each term. How many of the terms are marginally significant?

#create a regression model using the predictors. None of these predictors are significant  
model = lm(Volume ~ Girth + Height + GirthHeight + Girth2 + Girth2Height, data = trees)  
summary(model)

##   
## Call:  
## lm(formula = Volume ~ Girth + Height + GirthHeight + Girth2 +   
## Girth2Height, data = trees)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1880 -0.7901 -0.0037 1.9306 3.9483   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 48.914179 90.852925 0.538 0.595  
## Girth -8.228180 13.803580 -0.596 0.556  
## Height -0.616152 1.250446 -0.493 0.626  
## GirthHeight 0.103075 0.180291 0.572 0.573  
## Girth2 0.311160 0.536379 0.580 0.567  
## Girth2Height -0.001764 0.006621 -0.266 0.792  
##   
## Residual standard error: 2.659 on 25 degrees of freedom  
## Multiple R-squared: 0.9782, Adjusted R-squared: 0.9738   
## F-statistic: 224.3 on 5 and 25 DF, p-value: < 2.2e-16

**Multiple Choice Answer** **(AUTOGRADED on D2L)**: 0,  
1,  
2,  
3,  
4, or  
5

We now apply k-fold cross-validation to produce honest predictions, using the process outlined in the next several questions.

### Question 3 **(1 point)**

Starting with:

Set R’s seed to 2:

set.seed(2, sample.kind = "Rounding")

## Warning in set.seed(2, sample.kind = "Rounding"): non-uniform 'Rounding' sampler  
## used

and then define define cvgroups (random groups for the cross-validation) using the sample() function.  
Enter your R code below.

**Code Answer**:

# Question 3  
  
 #using the code provided, create a vector with the sequence 1:5, 6x times. There is 1 remaining observation above 30, so combine "1" at the end. This appears to be setting up ~7 folds for CV (one of the groups only has 1 obs)  
groups = c(rep(1:5,6),1)  
 #count the sample size.  
n = nrow(trees)  
 #now, using sample, take a random sample, n times (31 times). Pull that sample from the new groups vector, which has all the numbers 1:5 replicated many times. We can use this newly "randomly shuffled" vector to build our CV groups.  
cvgroups = sample(groups,n)  
  
  
# Questions 4-5  
  
# Question 6

### Question 4 **(1 point)**

Use the 5-fold CV method to assess the model fit. Provide the predicted y-value for the **first** observation:

**Numeric Answer** **(AUTOGRADED on D2L)**:

### Question 5 **(1 point)**

Use the 5-fold CV method to assess the model fit. Provide the predicted y-value for the **second** observation:

**Numeric Answer** **(AUTOGRADED on D2L)**:

### Question 6 **(2 points)**

Calculate and report the based on the 5-fold cross-validation:

# Question 6

**Numeric Answer** **(AUTOGRADED on D2L)**:

### Question 7 **(3 points)**

Enter your R code for computing the measure below.

**Code Answer**:

**Bootstrapping**

We will now use the bootstrap to estimate variability of the coefficients.

### Question 8 **(3 points)**:

Program a function, making use of lm() to fit the linear regression model, that outputs the six coefficient estimates. Set R’s seed to 2:

set.seed(2, sample.kind = "Rounding")

## Warning in set.seed(2, sample.kind = "Rounding"): non-uniform 'Rounding' sampler  
## used

and then use to produce R = 1000 bootstrap estimates for each of , , , , , and .  
Enter your R code below.

**Code Answer**:

# Question 8   
  
# Questions 9-14

### Questions 9-14 **(6 points, 1 each)**:

Use your bootstrap estimates to estimate the standard error, , for each of i = 0, 1, 2, 3, 4, 5.

**Numeric Answer**   
**(AUTOGRADED on D2L)**:

### Question 15 **(2 points)**:

The standard errors estimated from usual linear regression methods are shown in the R output below:

How do these values compare to the standard errors computed in the previous set of questions?

**Multiple Choice Answer** **(AUTOGRADED on D2L)**: one of

1. The estimates from usual linear regression methods are **greater**.
2. The estimates from usual linear regression methods are **less**.
3. The two sets of estimates are about the **same**.

## Problem 2 - Model Selection

This problem practices application of proper model selection techniques, with a multiple linear regression model. We will continue working with the predictive model using multiple linear regression. However, we will now consider selection between 6 possible models:

Model 1:

Model 2:

Model 3:

Model 4:

Model 5:

Model 6:

### Questions 16-17 **(2 points, 1 each)**:

Use LOOCV (note n = 31) method to calculate for each of Models 1-6. Report the for Models 1 and 2.

**Numeric Answer**   
**(AUTOGRADED on D2L)**:  
For Model 1, =  
For Model 2, =  
(use code space in next question)

### Question 18 **(4 points)**:

Enter your R code for computing the measure for Model 6 below.

**Possible Answer**:

#Q18  
  
#Q16  
  
#Q17

### Question 19 **(1 point)**:

Which model would you select based on the values of for LOOCV?

**Multiple Choice Answer** **(AUTOGRADED on D2L)**: one of  
Model 1,  
Model 2,  
Model 3,  
Model 4,  
Model 5, or  
Model 6

### Question 20 **(1 point)**:

Explain why you chose the model selected in the previous question.

**Text Answer**:

### Questions 21-22 **(2 points, 1 each)**:

Using the same split of the data into five sets as you performed in Problem 1, use 5-fold cross-validation method to calculate for each of Models 1-6. Report the for Models 1 and 2.

#Q21  
  
#Q22

**Numeric Answer**   
**(AUTOGRADED on D2L)**:  
For Model 1, =  
For Model 2, =  
(use code space above)

### Question 23 **(1 point)**:

Which model would you select based on the values of for 5-fold CV?

**Multiple Choice Answer** **(AUTOGRADED on D2L)**: one of  
Model 1,  
Model 2,  
Model 3,  
Model 4,  
Model 5, or  
Model 6

### Question 24 **(2 points)**:

Considering the form of the model that was selected by cross-validation, why does this model make sense from a practical standpoint?

**Text Answer**:

## Problem 3 - Model Assessment & Selection with KNN

This problem practices application of proper model assessment and selection techniques, with the kNN model.

**Important**: Use the FNN library for fitting K-nearest neighbors, to obtain consistent answers.

In this problem, you will once again use the K-nearest neighbors approach to analyze the gas mileage of cars. You will use the **Auto** data set from the ISLR package, along with the two new variables, **weight.std** and **year.std** (standardized values of the weight and year), that you created in Homework 1: K-Nearest Neighbors.

### Question 25 **(3 points)**:

**Model assessment** Starting with:

Set R’s seed to 2:

set.seed(2, sample.kind = "Rounding")

## Warning in set.seed(2, sample.kind = "Rounding"): non-uniform 'Rounding' sampler  
## used

and use sample() to divide the data into ten sets. Then use 10-fold cross-validation method to calculate for 1-nearest neighbor regression. Remember to re-standardize each training set inside the cross-validation. Report the value.

**Numeric Answer**   
**(AUTOGRADED on D2L)**:  
 =   
(use code space in next question)

### Question 26 **(4 points)**:

Enter your R code for computing the measure below.

**Code Answer**:

### Question 27 **(1 point)**:

In general, how should the value compare to the value of MSE (computed by reusing the same data used to fit the model)?

**Multiple Choice Answer** **(AUTOGRADED on D2L)**: one of ,  
, or

### Question 28 **(3 points)**:

Consider models 1-30 as the k-nearest neighbors regression for values of k from 1 to 30. Using the same split of the data into ten sets as you performed in the Model assessment section, use 10-fold cross-validation method to calculate CV(10) for each of Models 1-30; remember to re-standardize each training set inside the cross-validation. Make a plot of the CV(10) as a function of k. Upload your plot to the Quiz question.

\*\*Plot upload:

### Question 29 **(2 points)**:

Which k (number of nearest neighbors) would you select based on the values of for 10-fold CV?

**Numeric (Integer) Answer**   
**(AUTOGRADED on D2L)**:

### Question 30 **(2 points)**:

Explain why you chose the k value specified in the previous question. *Comment on both model predictive ability and model complexity.*

**Text Answer**: