COMPSCIX 415.2 Homework 7

Robert Clements

DUE DATE: Mar 20, 2018 @ 6:15PM

In this assignment. . .

. . . we will work with linear models and training and testing sets.

What to Turn In

For this assignment you have two choices:

You can upload a pdf document (you will have to install latex);
You can upload a standalone html file.

Prerequisite

Basic R Markdown knowledge
R and RStudio
tidyverse and the broom package installed
Access to internet
git and Github

Exercises

Here are the exercises for you to complete.

Go to the House Prices prediction competition on Kaggle here: https://www.kaggle.com/c/house-prices-advanced-regression-tec

Read the Overview Description, Evaluation, and the Data sections.

Download the train.csv and data_description.txt files either from the Kaggle website, or from Canvas (in the homework section). You **do not** need to download the test.csv file, we will be creating our own test set. To download from Kaggle you will probably have to sign up for an account, so I leave that decision up to you.

Skim through the data_description.txt file. There is **a lot** of stuff here, but don't feel the need to read it all.

Next, download the `broom` package.

Answer these questions:

Exercise 1

Load the train.csv dataset into R. How many observations and columns are there?

Exercise 2

Normally at this point you would spend a few days on EDA, but for this homework we will get right to fitting some linear regression models.

Our first step is to randomly split the data into train and test datasets. We will use a 70/30 split. There is an R package that will do the split for you, but let's get some more practice with R and do it ourselves by filling in the blanks in the code below.

```
# load packages
library(tidyverse)
library(broom)

# When taking a random sample, it is often useful to set a seed so that
# your work is reproducible. Setting a seed will guarantee that the same
# random sample will be generated every time, so long as you always set the
# same seed beforehand
set.seed(29283)

# This data already has an Id column which we can make use of.
# Let's create our training set using sample_frac. Fill in the blank.
train_set <- train %>% sample_frac(____)

# let's create our testing set using the Id column. Fill in the blanks.
test_set <- train %>% filter(!(____ %in% ____$Id))
```

Exercise 3

Our target is called `SalePrice`. First, we can fit a simple regression model consisting of only the intercept (the average of `SalePrice`). Fit the model and then use the broom package to

- take a look at the coefficient,
- compare the coefficient to the average value of `SalePrice`, and
- take a look at the R-squared.

Use the code below and fill in the blanks.

```
# Fit a model with intercept only
mod_0 <- lm(SalePrice ~ 1, data = _____)

# Double-check that the average SalePrice is equal to our model's coeffficient
mean(train_set$SalePrice)
tidy(____)

# Check the R-squared
glance(____)
```

Exercise 4

Now fit a linear regression model using `GrLivArea`, `OverallQual`, and `Neighborhood` as the features. Don't forget to look at `data_description.txt` to understand what these variables mean. Ask **yourself** these questions before fitting the model:

- What kind of relationship will these features have with our target?
- Can the relationship be estimated linearly?

- Are these good features, given the problem we are trying to solve?

After fitting the model, output the coefficients and the R-squared using the `broom` package.

Answer these questions:

- How would you interpret the coefficients on `GrLivArea` and `OverallQual`?
- How would you interpret the coefficient on `NeighborhoodBrkSide`?
- Are the features *significant*?
- Are the features *practically significant*?
- Is the model a good fit (to the training set)?

Exercise 5

Evaluate the model on `test_set` using the root mean squared error (RMSE). Use the `predict` function to get the model predictions for the testing set. Recall that RMSE is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{SalePrice}_i - \hat{\text{SalePrice}}_i)^2}$$

Hint: use the `sqrt()` and `mean()` functions:

```
test_predictions <- predict(NAME_OF_YOUR_MODEL_HERE, newdata = test_set)
rmse <- sqrt(mean((___ - ___)^2))
```

Exercise 6 (OPTIONAL - won't be graded)

Feel free to play around with linear regression. Add some other features and see how the model results change. Test the model on `test_set` to compare the RMSE's.

Exercise 7

One downside of the linear model is that it is sensitive to unusual values because the distance incorporates a squared term. Fit a linear model to the simulated data below, and visualise the results. Rerun a few times to generate different simulated datasets. What do you notice about the model?

```
sim1a <- tibble(
  x = rep(1:10, each = 3),
  y = x * 1.5 + 6 + rt(length(x), df = 2)
)
```

Turn in your completed assignment

Commit your changes with the comment "finished assignment 7" and push your R Markdown file and your html or pdf file to Github.

This week you should turn in your assignment by uploading it to Canvas.