Homework 1

STAT 430, Spring 2017

Due: Friday, February 3 by 11:59 PM

Please see the homework instructions document for detailed instructions and some grading notes. Failure to follow instructions will result in point reductions.

Exercise 1

[**12 points**] This question will use data in a file called `hw01-data.csv`. The data contains four predictors `a`, `b`, `c`, `d`, and a response `y`.

Use `set.seed(42)` to control randomization, then randomly split the data into train and test sets using half of the data for each. Next, fit four models using the training data:

- Model 1: `y ~ .`
- Model 2: `y ~ . + I(a ^ 2) + I(b ^ 2) + I(c ^ 2)`
- Model 3: `y ~ . ^ 2 + I(a ^ 2) + I(b ^ 2) + I(c ^ 2)`
- Model 4: `y ~ a * b * c * d * I(a ^ 2) * I(b ^ 2) * I(c ^ 2)`

(a) For each of the models above, report:

- Train RMSE
- Test RMSE
- Number of Parameters, Excluding the Variance

To receive full marks, arrange this information in a well formatted table.

(b) Based on these results do you have evidence that any of these models are overfitting or underfitting? If so, which models?

(c) Find a model that outperforms each of the models above. Report this model's train RMSE, test RMSE, and number of parameters used. **Hint:** If you haven't already, consider some exploratory data analysis. **Hint:** Your instructor's solution uses a model with only six parameters. Yours may have more.

Exercise 2

[**8 points**] For this question we will use the `Boston` data from the `MASS` package. Use `?Boston` to learn more about the data.

```r
library(MASS)
data(Boston)
Boston = as_tibble(Boston)
```

Use `set.seed(314)` to control randomization, then randomly split the data into train and test sets using 456 observations for the training data and the remainder for the testing data. (Roughly 10 percent of the data for the test set.)

Fit a (potentially large) number of **nested** linear models with `medv` as the response. Use train and test RMSE to determine: two models that are probably underfitting, two models that are probably overfitting, and one model between the under and overfitting models. Report (only) these five models as well as their train RMSE, test RMSE, and number of parameters. Note: you may report the models used using their `R` syntax. To receive full marks, arrange this information in a well formatted table.

Exercise 3

[**10 points**] How do outliers and influential points affect prediction? Usually when fitting regression models for explanation, dealing with outliers is a complicated issue. When considering prediction, we can empirically determine what to do.

Continue using the Boston data, training split, and models from Exercise 2. Consider your best model from Exercise 2. Refit this model with each of the following modifications:

- Removing observations from the training data with studentized residuals greater than 2.
- Removing observations from the training data with studentized residuals greater than 3.
- Removing observations from the training data considered influential. That is, with a Cook's distance greater than $\frac{4}{n}$.

(a) Use these four models, including the original model fit to unmodified data, to obtain test RMSE. Summarize these results in a table. Include the number of observations removed for each. Which performs the best? Were you justified modifying the training data?

(b) Using the best of these fitted models, create a 99% **prediction interval** for an observation with the following values:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
0.02763	75.0	2.95	0	0.4280	6.595	21.8	5.4011	3	252	18.3	395.63	4.32