Statistics 512: Final Project
Due April 27, 2018

PLEASE READ CAREFULLY

Introduction

The aim of many statistical techniques is to provide a method for making good predictions. For the purposes of this project, you are given six sets of data, each in two parts: a *training set* with predictors and response with which you are to build your model(s), and a *test set* with just predictor variables. The assignment is to make predictions using whatever techniques or models you think are appropriate. Your grade will be based in part on how close your predictions are to the (hidden) test data responses, which were manufactured using the same processes as those used to produce the training data.

In addition to making good predictions, a good statistical technique will estimate how far off its predictions are expected to be. (We have often seen this value encapsulated in the residual variance (σ^2) term.) The second part of this project involves coming up with a good estimate of the error you expect to make. In this case, you should attempt to predict the sum of squared residuals of your predictions for the test data.

The underlying true models are all "near" the standard linear models we have discussed in class. It is certainly possible that some of the departures we have considered will be included. Don't expect every deviation in every data set, and don't expect that each data set should have exactly one of the deviations.

Theory

The data in this project were generated using models similar to what we considered in this class. For instance, data might be generated using the simple linear model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

for some unknown β_0, β_1, and $\epsilon_i \sim N(0, \sigma^2)$. The training set is a set of points $(\mathbf{X_i}, Y_i)$ for $i = 1, \ldots, n$ ($\mathbf{X_i}$ is possibly multivariate); the test set is simply a set of new predictor values $\mathbf{X_j}$ for $j = n+1, \ldots, N$. For each data set, the values of n and N correspond to the sizes of the training data set and the entire data set. For each $\mathbf{X_j}$, you are expected to produce values $\hat{Y}_{n+1}, \ldots, \hat{Y}_N$, which will be evaluated by their closeness to the true values $Y_{n+1}, \ldots, Y_N$.

Of course, there is a certain amount of variability in the test data set that cannot be accounted for using the training data set. So the predicted values will almost never match up exactly with the test set. As such, it is necessary to quantify how far off those value (in the aggregate) should be. Thus, you should produce a value $S\hat{S}E_{test}$ to estimate

$$SSE_{test} = \sum_{j=n+1}^{N} (\hat{Y}_j - Y_j)^2.$$

Note that for a particular j, $E(\hat{Y}_j - Y_j)^2 = \sigma_j^2\{pred\}$. So the formulae for estimating $\hat{\sigma}^2\{pred\}$ from Sections 2.5 and 6.7 of the book (in particular, equation (6.58)) will probably be useful in making your final estimate.

Mechanics

All data can be found at

`http://www.stat.purdue.edu/~lingsong/teaching/2018spring/final/`

The data sets are such that the last column is the response. (The different data sets are not of the same size and have different numbers of predictor variables.)

Submission

You should submit your prediction files to the blackboard learn.
Make certain your files are clearly labeled, so that there is no ambiguity about which file corresponds to which problem. I have found that a useful rule of thumb for naming files is to use your initials. Thus, my first submitted data set might be named `LZ1.dat`; my second, `LZ2.dat`, etc.

Please submit one file (with test predictions only) per data set. Only files with predictions for **every** test data point will be counted. (So whatever trouble you have, you should fill in *some* prediction for each point.) **Each file should be directly readable as a data file to SAS.** Thus, Word documents are not acceptable; files that have headings *within* the file's contents are not acceptable; and again, files with missing values are not acceptable.

In addition to the prediction files, you should submit a text file that has an estimate of the sum of squared errors ($\sum_{j=n+1}^{N}(Y_j - \hat{Y}_j)^2$) for your predictions. The file should be a simple text file and look something like (as an example)

```
1.25571772
30.91083590
0.22698784
0.00174718585
1.37849871
-0.09877989
```

(OK, so that last value was a joke. You get the idea...)

You should also include a write-up for each data set. The write-up should describe (in a paragraph or so) what approach you used in coming up with your predictions. The discussion should be such that I could reproduce your analysis and thus re-calculate your predictions. **Unless absolutely necessary, do not include SAS code in your write-up.**

You should also include at least one paragraph in your write-up describing your method for estimating the sum of squared residuals (SSE).

There are no restrictions on the file format for the write-up. Hardcopy submission is fine, as long as it arrives to me (or is postmarked) by the due date/time.

Everything (files, papers, whatever) is due by Friday, April 27, 2018 at 11:59pm.

Grading

Evaluation (for each of the six data sets on the website) is based on

1. The reasoning used in determining how to predict. (40%)
2. The accuracy of the predictions. (40%)
3. The reasoning used in determining error estimates. (10%)
4. The accuracy of the error estimates. (10%)

Accuracy is determined as percentage difference from the best possible (or in the case of SSE estimates, the true) values. Each data set is weighted equally.

Last bold note:

This project is to be done with you and your group members. You can consult any books or notes you find appropriate. The person outside your group that you may consult – even with SAS questions – is Dr. Zhang.