MSDS 6372: Unit 2 HW 2

Conceptual questions.

1. TRUE or FALSE. Using model fitting metrics based on the training data set as a criterion to assess a model under or overfitting is good idea.

FALSE

1. TRUE or FALSE. The prediction error from a model applied on a test set is 100. This implies that the model is predicting horribly.

FALSE

1. State 2 Advantages of Multiple Linear Regression and 2 Disadvantages.

Advantages- 1. Interpretation of how the predictors relate to the response

2. When model correctly specifies the truth, predictions will perform better than other methods

3. Useful in smaller data sets where more complex algorithm will tend to overfit

Disadvantages- 1. Time consuming in terms of model building complexity (Assumptions)

2. Will under fit when relationship between response and predictors is complex

3. Multicollinearity and transformations make interpretation more difficult

1. Do the predictors of a multiple linear regression model have to be normally distributed?

This is a common misconception. The residuals after modeling the predictors must be normally distributed. There is no assumption on the predictors specifically. Think about it, if this were true, then we can never include a categorical predictor since they are clearly not normally distribued.

1. Suppose we have the following predictive model fits. Can you sketch out what the predictions look like graphically? The point of this exercise is to illustrate that model complexity can be achieved with interaction terms. Its just up to the modeler to actually do it.

where and

The sketch of this should have two unique lines with there own individual slopes and intercepts. Males should have intercept of 5 and slope of 1. Females have intercept of 10 and slope of 3.

where and

The sketch of this should have two lines with there own individual intercepts but the same slope. Males should have intercept of 5 and slope of 1. Females have intercept of 10 and slope of 1.

where and

This is an extension of the one above it. Here we will have 4 lines all with different intercepts but the same slope. It may be helpful to plug in numbers for the betas to visualize it.

where and

This creates two unique quadratic functions, one for females and one for males.

Computation Assignment

One of the major advantages of LASSO regression is that the estimate of the regression coefficients (betas) are allowed to be biased whereas the OLS estimates are forced to be unbiased. This highlights the commonly referred to “variance / bias trade-off”. Since MSE = Variance + Bias2, it is easy to see that for the OLS unbiased estimates that the MSE(betas) = Variance(betas) since the Bias is zero. However, with the biased LASSO estimates, one is often able to reduce the variance of the estimate of the betas at the cost of introducing a little bias. Often the reduction in the variance is greater than the increase in the squared bias and we see a reduction in the MSE of the betas.

In short, this means that our estimate of the regression equation with the smaller MSE has greater probability of being close to the equation with the real betas: the real trend. This means that if we cross validate our model on a test set (that maintains the same trend as the training set but with different noise), the model with the smaller MSE has greater probability of capturing more of the true trend. Statistically, this will be reflected in statistics such as the ASE (Test) (Average Squared Error for the model trained on the training set and used to fit the test set) and R squared of the test set (basically any goodness of fit statistic that is **with respect to the model and the test set**.)

We will use the ASE (Test) statistic to provide evidence of preferred models for this data set: LASSO or OLS.

Find the HW 2 Sas Code in the HW 2 folder. This code has the SAT data set as well as some code to divide the data set into a training and test set.

The assignment for this week is simple. Simply run the code and make an argument / discussion as to which model / estimates (LASSO or OLS) will provide better predictive ability. Why do you think the result plays out the way it does? Perhaps a look at the data visually?

Make sure and copy and paste all relevant output to support your decision. Don’t overthink this. The answer can reference a single statistic.

Extra Effort for R users: Try to replicate this in R. The example code I provided will be helpful here but you will probably have to sort some pieces out. Note: testMSE in my R script is analogous (not necessarily equivalent) to the ASE in SAS.

SOLUTION:

|  |  |
| --- | --- |
| LASSO FIT | OLS FIT |
|  |  |
|  |  |

In this particular example it doesn’t appear that the LASSO estimates provide better performance with respect to ASE from the test set (LASSO ASE (TEST) = 651 v. OLS ASE (TEST) = 251. For reference, the LASSO and OLS estimates are provided in the second row.