GSCM 410/510 – Summer 2021
Solutions Week 4

Directions

It is OK to work together on the Homework. Of course, make sure that everyone working together on an assignment is learning the material. Each person still turns in their own homework even if exactly the same as that of other group members. Learning this material well enough to do the homework problems is how you succeed in this course.

Short-Answer Problems

These concepts can appear on the short-answer part of the tests. As part of this homework, answer the following questions, usually just several sentences that include the definition.

1. *Briefly explain how multiple regression enhances the two primary purposes of regression analysis?*

Predictor variables (features) that are relevant (correlate with y), and provide unique information (don't correlate with the other X's), lead to a) better, more accurate prediction, and b) a better understanding of how the variables are related to each other.

2. *What criteria that a potential feature (predictor variable) should satisfy before added to a model?*

Predictor variables (features) should be relevant (correlate with y), and provide unique information (do not correlate with the other X's).

3. *What is the purpose and benefit of feature selection (i.e., select predictor variables for a model).*

Not all potential features are relevant (correlate with y) and unique (do not correlate with other X's). As such, they contribute little, or even detract, from forecasting efficiency and model interpretability. Particularly for large data sets, they can also add potential machine time for the computations. As such, best to rid the model of irrelevant features.

4. *One potential issue with multiple regression is collinearity. Describe the problem and how it can be addressed.*

Collinearity means that predictor variables (features) correlate substantially with each other. Collinearity increases the standard errors of the estimated collinear slope coefficients as the estimation algorithm cannot readily separate their effects (holding the others constant). Informal detection is to inspect the feature correlation matrix for high correlations. More formal is to regress each feature on all the others, to detect which ones can be explained in

terms of the other features and perhaps, then, not needed in the model, a feature (de)selection technique.

5. How does a correlation matrix, perhaps in the form of a heat map, facilitate feature selection?

The heat map is a visualization of a correlation matrix. An informal, but useful, feature selection technique is to delete some collinear features. The heat map can not only provide the correlations, but also color codes each according to the magnitude of each correlation. Such color-coding assists in identifying large correlations.

6. For a given set of customers, almost all weigh between 110 and 300 lbs. One customer, an outlier, reports a weight of 460 lbs. What is the basis for dropping the customer from the analysis? How does such an action change the reported results?

Either the data value is mis-entered, or, if correct, any generalizations would not properly apply to these people, and perhaps bias the model for the vast majority of customers. In practice, experiment with different deletion thresholds as perhaps a better model can be constructed by focusing, for example, on only people with weights between 100 and 300 lbs. With otherwise so much variability, the model may perform more poorly for everyone, instead of better performance for the vast majority of customers who weigh between 100 and 300 lbs.

7. Distinguish between training data and testing data

A core concept of machine learning is that forecasting efficiency cannot be evaluated on the data on which the model trained, i.e., the data from which the model coefficients are estimated. That evaluation can only occur by observing the errors of applying the model to new data, which is the actual forecasting situation.

8. Why can a model not be properly evaluated on its training data?

Every data set sampled from a population differs from any other data set sampled from the same population. Every sample reflects the underlying population values, but every corresponding sample value, such as the mean, does not equal the corresponding population value. So fitting sample data from which the model trained fits random sampling error as well as true, stable population characteristics. A model can fit training data perfectly, but have no useful ability to forecast on new (testing) data.

9. Define overfitting.

Overfitting is when a model is too complex, where the extra complexity takes advantage of random sampling fluctuations in the training data to increase fit.

10. What is the problem overfitting presents in evaluating model forecasting performance?

An overfit model fits the training data well, but has poor generalization to actual forecasting, that is, to new data (e.g., the testing data). The good fit to the training data is irrelevant.

11. *How can ¥the analyst determine if a model is overfit?*
    Compare the fit of the model from the training data to the testing data. If there is a big decrease, the model is overfit to the training data.
12. *Define underfitting and discuss the problem it presents for model development.*
    Underfitting means the model is too simple to capture all the information in the training data that is not random variation, but reflects stable aspects of the underlying population.
13. *What does it mean to state that "A model should be made as simple as possible, but not simpler."?*
    Make the model as complex as it can be to capture the relevant information in the training data to avoid underfitting without so much complexity the model overfits.
14. *What is a hold-out sample, and what is its purpose?*
    A hold-out sample is the testing data, data on which the model was not trained (fit). Its purpose is to evaluate the forecasting efficacy of the model in a true forecasting situation of which the model is "unaware" of the value of y, but the analyst is aware, so can evaluate the error directly.
15. *How does k-fold cross-validation extend the concept of a hold-out sample?*
    With k-fold validation there are k hold-out samples and so k cross-validations.
16. *Why is k-fold cross-validation preferred to just splitting the original data into training and testing data, a train/test split?*
    Instead of just one arbitrary, usually randomly selected hold-out sample, there are k hold-out samples. Any one train/test split may result, by chance, in a weird test sample or training sample. With k different such training/test splits the average performance of the model across all the k-folds yields a more stable estimate of the forecasting efficacy, such as with MSE or s_e.