Assignment Question 1d

Tuesday, 31 March 2020 10:20 PM

The mean squared error (MSE) is a measure of model accuracy. It is calculated by using the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

Where fhat(xi) is the prediction of y for the ith observation and yi is the true ith observation. Essentially MSE measures how close the model fits to the actual data by averaging the distance (or error) between the true and predicted values. MSE can be measured on a training set of data (the data used to build the fitted model) or a test set (data from the true model that has not been exposed to the fitted model).

Ultimately, we are interested in minimising the test MSE as we want a useful model that can accurately predict the response variable from unseen data. For example, if I wanted to predict business sales using other information such as the season or the state of the economy, I am not concerned with how the model predicts business sales last week or last month. I am instead interested in how the model predicts sales for next week or next month as this may help me make future decisions such as preparing inventory.

Figure 2 is an example of a flexible model that would have a very low training MSE, as the model is fitted very closely to the test data. The inflexible linear model in figure 1 would have a higher training MSE as there is more distance between the model and the observations than in Figure 2. As the level of flexibility increases, the model fits the training observations more tightly and the training MSE decreases.

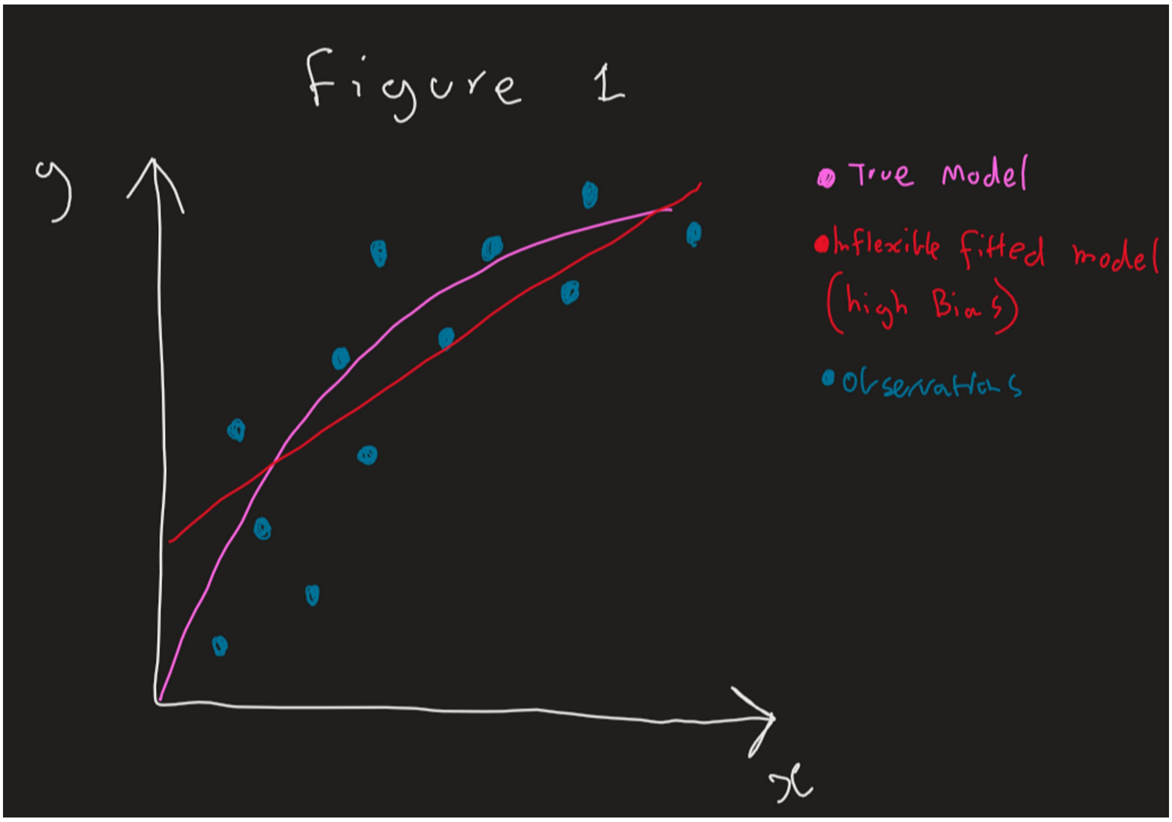
In order to create a model that imitates the true model, it is clear we need a balanced model, that is, a model that is more flexible then figure 1 so that it captures the non-linearity but less flexible then figure 2 so that it is not as sensitive and does not pick up small patterns that may not exist in the test data. The balancing of flexibility to minimise the test MSE is a proven mathematical property given by:

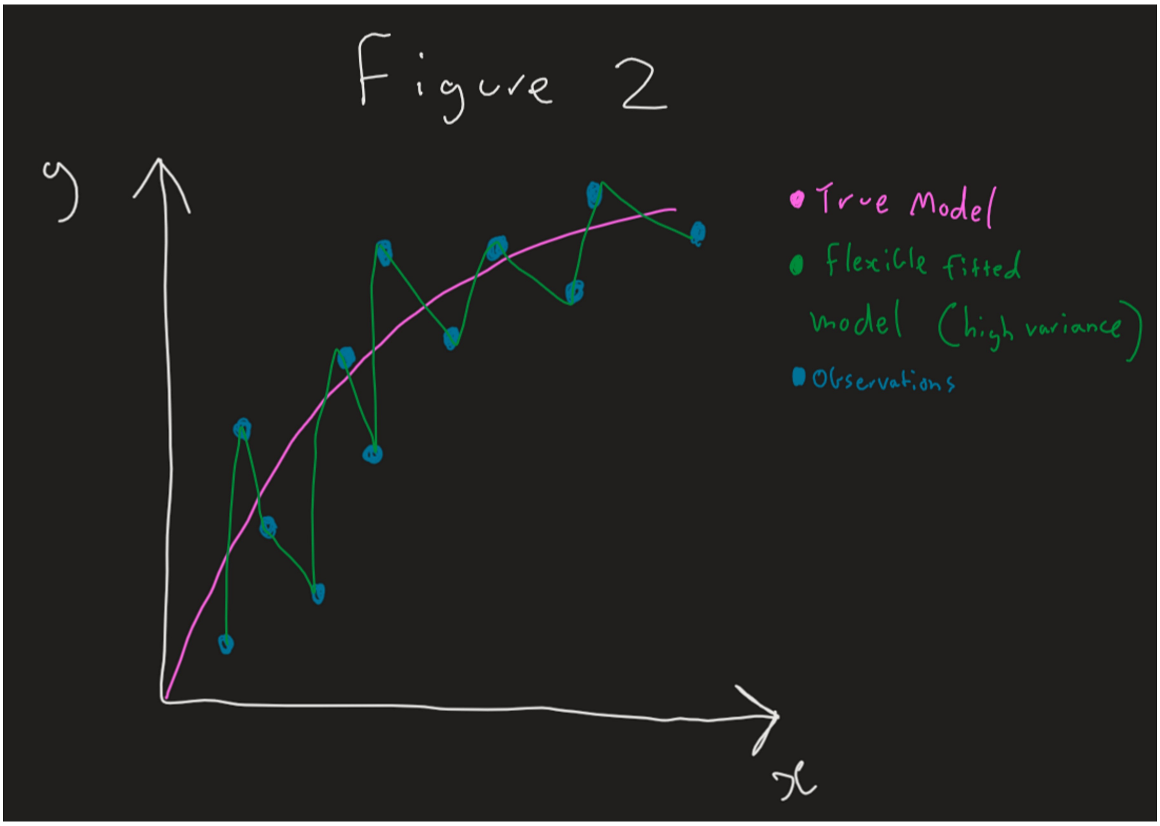
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

This equation tells us that the expected test MSE can be broken down into three components; the variance of fhat(x0), the squared bias of fhat(x0) and the irreducible variance of the error term.

The expected test MSE refers to the average test MSE that we would obtain if the fitted model were estimated repeatedly on a large number of training data sets. The overall expected test MSE is found by averaging the expectation over all possible x0 values in the test data.

To minimise the test MSE, we must try to minimise both the variance and bias in our model. Variance refers to the magnitude in which the fitted model changes when estimated using a different data set. Flexible models such of that in Figure 2 have higher variance as the model will likely change dramatically to fit the new observations tightly. On the other hand, an inflexible model such as Figure 1 would likely have low variance as the linear shape would only vary slightly to best fit the new observations. Bias refers to the error introduced by estimating a potentially complicated relationship with an overly simple model. Inflexible models such as the model shown in Figure 1 generally have more bias then flexible models such as the model in Figure 2. In our example, Figure 1 has high bias as it has failed to account for the non-linear pattern of the true model.





References:
http://www.cs.umd.edu/class/summer2019/cmsc320/files/Lec22_June27_2019.pdf
https://stats.stackexchange.com/questions/379319/how-is-test-mse-being-calculated-here
Casella, G., Fienberg, S., & Olkin, I. (2013). An Introduction to Statistical Learning. In Springer Texts in Statistics. https://doi.org/10.1016/j.peva.2007.06.006