Underfitting and overfitting

As you have been learning, a multiple regression model is built using sample data from the population of interest with the goal of applying the model to unseen data from the population and getting reliable results. **Underfitting** and **overfitting** are two obstacles that the multiple regression model must mitigate so it can be applicable. In this reading, you will gain a general understanding of underfitting and get a closer look at overfitting.

The two ways a model can be unreliable

Underfitting

In the case of underfitting, a multiple regression model fails to capture the underlying pattern in the outcome variable. An underfitting model has a low R-squared value.

A model can underfit the data for a variety of reasons. The independent variables in the model might not have a strong relationship with the outcome variable. In this situation, different or additional predictors are needed. It could be the case that the sample dataset is too small, and this prevents the model from being able to learn the relationship between the predictors and the outcome. Using more sample data to build the model might reduce the problem of underfitting.

Consider the example of a multiple regression model that predicts the resale price of a pre-owned car. This model has two predictors: the color of the car and the year it was manufactured. The model's R-squared value is quite low. This indicates that the model is underfitting because the current predictors do not have a strong relationship with the car's resale price. There are likely other important predictors missing from the multiple regression model, like the mileage on the car or the make of the car.

There are additional reasons that a multiple regression model might underfit the data, and the methods used to reduce this obstacle depend on the specific context. Because an underfitting multiple regression model is not able to capture the relationship between predictors and outcome in the sample data, this model will also not be able to produce reliable results when it is used on unseen data from the population.

The difference between training data and test data

Before you learn more about overfitting, it is important to cover a step data scientists take before building a multiple regression model. They divide the sample data into two categories called **training data** and **test data**. Training data is used to build the model, and test data is used to evaluate the model's performance after it has been built. Splitting the sample data in this way is also called **holdout sampling**,

with the holdout sample being the test data. Holdout sampling allows data scientists to evaluate how a model performs on data it has not experienced yet.

The holdout sample might also be called the **validation data**. Regardless, the general idea remains the same: this is the data that is used to evaluate the model.

Data scientists obtain the training and test data by randomly splitting the sample dataset so that each record exclusively belongs to one of the two categories. This way, some records are used as the training data and other records are used as the test data.

Overfitting

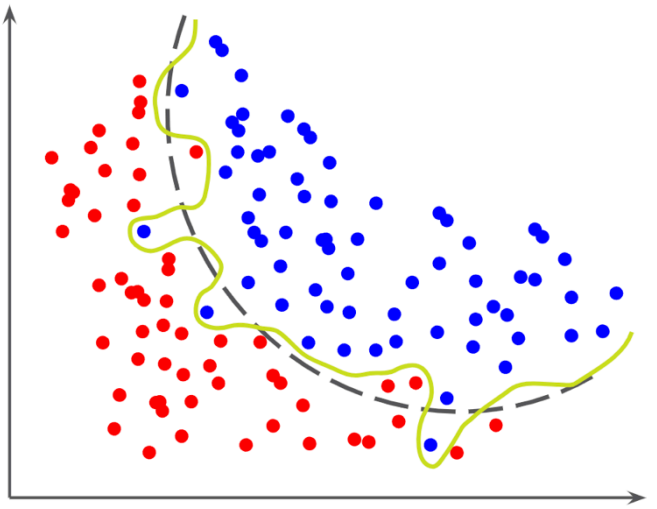
Underfitting causes a multiple regression model to perform poorly on the training data, which indicates that the model performance on test data will also be substandard. In contrast, overfitting causes a model to perform well on training data, but its performance is considerably worse when evaluated using the unseen test data. That's why data scientists compare model performance on training data versus test data to identify overfitting.

Why is there a discrepancy between an overfitting model's performance on training data versus test data?

An overfitting model fits the observed or training data too specifically, making the model unable to generate suitable estimates for the general population. This multiple regression model has captured the **signal** (i.e. the relationship between the predictors and the outcome variable) *and* the **noise** (i.e. the randomness in the dataset that is not part of that relationship). You cannot use an overfitting model to draw conclusions for the population because this model **only** applies to the data used to build it.

In the plot below, the dashed black line represents an optimal multiple regression model that performs well in distinguishing between the red and blue dots without overfitting the data. In contrast, the squiggly yellow line represents a model that overfits the data. Although this line might even do a slightly better job of separating the blue dots from the red ones, it is too specific to this data and will not perform well on another sample from the same population. In contrast, the black line will be continuously reliable in distinguishing between the two colors.

Overfitting versus a “Just Right” Fit



Why does overfitting result in a higher R-squared value?

Earlier you learned that R-squared is a goodness of fit measure because it tells you the proportion of variance in the outcome variable that is captured by the independent variables in the multiple regression model. However, as you add more independent variables to a model, the associated R-squared value will increase regardless of whether or not those predictors have a strong relationship with the outcome variable.

In the example of the multiple regression model that predicts car resale price, you could continue to add more independent variables to the model, such as the number of letters in the name of the person selling the car and the favorite food of the person who bought the car (if you had this data, of course). These predictors are very unlikely to

have a relationship with the resale price of the car, but if you add them to your multiple regression model, the R-squared value would still increase. Although this could lead you to think that the model with more predictors is performing better, the inflated R-squared value is a false sign of improvement.

Generally, R-squared will continue to increase with more predictors because the model will become overly specific to the data it was built on even if the predictors do not have a strong relationship with the outcome variable. This is why a high R-squared value is not enough by itself to indicate that the model will perform well and might instead be a sign of overfitting.

When to use adjusted R-squared instead

Along with the R-squared value, a multiple regression model also has an associated **adjusted R-squared value**. The adjusted R-squared penalizes the addition of more independent variables to the multiple regression model. Additionally, the adjusted R-squared only captures the proportion of variation explained by the independent variables that show a significant relationship with the outcome variable. These differences prevent the adjusted R-squared value from becoming inflated like the R-squared value.

When comparing between multiple regression models with varying numbers of predictors, you might find that models with more predictors have a higher R-squared value. This could be a result of overfitting. To avoid selecting an overfitting model with an inflated R-squared, use the adjusted R-squared metric to select the optimal model.

Bias versus variance

A model that underfits the sample data is described as having a high **bias** whereas a model that does not perform well on new data is described as having high **variance**. In data science, there is a phenomenon known as the **bias versus variance tradeoff**. This tradeoff is a dilemma that data scientists face when building any machine learning model because an ideal model should have low bias and low variance. This is another way of saying that it should neither underfit nor overfit. However, as you try to lower bias, variance inevitably increases and vice versa.

This is why you can never fully resolve the problems of underfitting and overfitting. Instead, focus on reducing these problems in your multiple regression model as much as possible.

Key takeaways

Both underfitting and overfitting are obstacles to building a

reliable multiple regression model. Although underfitting can be identified by model performance on training data, you must evaluate both training and test performance to identify overfitting. Because overfitting will result in an inflated R-squared value, use the adjusted R-squared value when comparing among multiple regression models with varying numbers of predictors. Although you cannot fully eliminate underfitting and overfitting from the model because of the bias versus variance tradeoff, you can significantly reduce these problems after they have been identified.

Resources for more information

- A detailed description of underfitting and how to mitigate it
- Scikit-learn library documentation for the train_test_split function
- A blog discussing multiple, adjusted, and predicted R-squared values