**How important is Parsimony versus Accuracy?**

There are two main approaches in Statistics when it comes to model building. One is to come up with a model that is simple to interpret and explains the relationship between $latex X$ and $latex Y$ well. The other one is to build a model that yields accurate predictions regardless of the form of $latex X$ and the complexity of the model. In a perfect world, we would like to produce a model that is simple, interpretable and has the highest predictive power. In reality, however, this is not attainable. In this blog post, I will compare the two approaches towards model building and will explain in what situations parsimony or accuracy are preferred.



**Definition and Pitfalls of Parsimony and Accuracy**

When we want to create a parsimonious model, we are interested in explaining how some underlying factors that are measured by some variables $latex X$, cause an underlying effect measured by the variable $latex Y$. In short, we want to understand the relationship between $latex X$ and $latex Y$. This is achieved by a model that is interpretable. Hence, a model that uses no more parameters than necessary to explain the relationship well. On the other hand, when one is seeking a model for prediction purposes, we are not interested in the relationship of $latex X$ and $latex Y$. Instead, we want a model that predicts well and yields high accuracy, regardless of the complexity of the model. Therefore, we can treat the model as a black box without knowing the relationship between $latex X$ and $latex Y$.

There are a few pitfalls when building models for the purpose of parsimony and accuracy. This is underfitting and overfitting respectively. So, when building a very simple model, there is the risk that we fail to capture the true signal; these models underfit the data and we fail the relationship between $latex X$ and $latex Y$. On the other hand, when we want to build a model for prediction purposes, we want to increase the accuracy of the prediction. This is often achieved with very complex models. However, the more complex a model, the more we run into the risk of capturing the noise from our training data. So, we have a almost perfect fit for our training data but for new observations, our model predicts poorly. This phenomenon is called overfitting.

**Parsimony vs. Accuracy**

* **When to Choose Between Parsimony and Accuracy**

Now, after we know the definitions of a model that is parsimonious, hence created to explain, and a model created for predictions, hence very accurate, we will now explain which method is preferred in certain situations. The goal of a model depends on the research question and on what a particular person or business is interested in. So, suppose that $latex x\_1, \dots,x\_p$ are characteristics of a patient’s blood sample. $latex Y$ is a variable encoding a patient’s risk for severe adverse reaction to a particular drug. In this particular example, we want to predict $latex Y$ using $latex X$ with the highest accuracy possible. This is because we do not want to give patients the wrong drug which makes them sick. Therefore, in this example we can treat $latex \hat{f}(x)$ as a black box and the only interest is to give the right drug to patients, regardless of the complexity of the model. When we want to build a model to explain the relationship of how $latex X$ affects $latex Y$, we want to come up with a parsimonious model. When a burger company wants to know how much increase of advertising budget increases sales, then we seek a model that explains well. So, the burger company is not interested in predicting sales for next week, it is rather interested in knowing if they should spend more on advertising or not. In this scenario, we cannot treat $latex \hat{f}(x)$ as a black box. Therefore, the form of $latex \hat{f}(x)$ must be simple and interpretable, hence parsimonious.

* **How Netflix Handled Parsimony vs. Accuracy**

For our next example of parsimony versus accuracy, we choose to look at Netflix as a case study in order to outline the trade off of parsimony and accuracy. In 2006, Netflix announced a Netflix price, which was worth 1 million dollars to whoever improved their recommendation system by 10%. The goal was to reduce the mean squared error from 0.9525 to 0.8527 or less. One year into the competition, the mean squared error was improved by 8.43% and a model with a final combination of 107 algorithms was implemented in the Netflix recommendation system. An additional two years after that, the magic 10% mark was cracked. However, Netflix analysed some of the methods used for the model that reduced the mean squared error below 0.8527 and concluded that the additional accuracy gains did no seem to justify the engineering effort needed to bring the methods into a production environment. Hence, Netflix went with the more parsimonious model which was less accurate. In other words, we can say that whether or not we increase complexity for additional accuracy is not a data science decision, it is a business decision.

**Conclusion**

To reiterate, the importance of parsimony versus accuracy is unanswerable and depends on the goal and resources of an individual or business. However, we have answered in which situations parsimony and accuracy are desired. The need for parsimony and interpretability explains, for example, why logistic regression is preferred to discriminant analysis because the coefficients can be uniquely estimated and have the meaning of log-odds. On the other hand, if we are interested in predictions and discriminant analysis yields a higher accuracy than logistic regression, then prediction does not need a deep understanding of what is observed. Therefore, whether to go with a parsimonious model or one that is complex and predicts well is a choice of interest. There is no superior model and we need both parsimony and accuracy in the study of statistics.