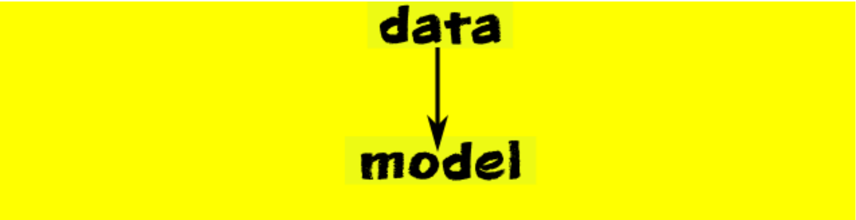
Model Selection

Shibendra Kumar Singh (BS2015)

1 Introduction

In most analysis scenario, we have to work with data and the first step is to make a model for it.



One can design lots of model with the available data and choose the best one among them which fits the data the most.
Example: Polynomial of different degrees:
For instance, one may choose the model of the form: $y = a + b.e^x + \epsilon$ or$log(y) = a + bx + \epsilon$.
Both the above model may be appealing, but these are not the same.

In a different note, we could have a *homoscedastic model* (i.e., a $Gauss - Markov$ setup) or a *heteroscedastic model*. But we know that *homoscedastic* is not a good assumption. So we may want to compare this model with the *heteroscedastic model*.
The need for *heteroscedastic model* might be less convincing in the beginning. However in reality, there are many such models useful and one needs to compare between them.
In all the cases we would have different types of models and we need to choose the one which would fit the data set well.

BEWARE:
The following are the two commonly held mistakes. Let's discuss one by one.

1. **Comparing models, NOT test of hypothesis:** We are trying to compare between different models, not use test of hypothesis to conclude one is better than the other.
   In case of test of hypothesis, we were given some particular models, which we already know that is true.
   We only try to answer questions that go dipper into the values of μ or σ^2. We need to explore because the best model on training data might perform poor on the test data.

2. **Comparing models, not methods:** We are not comparing whether one should take mean or rather work with the median. We only compare models and find which one among them gives a better fit.