Model Selection

Prof. Dennis Sun

Data 401
Where We Left Off

We’ve discussed how to fit models of the form

\[ Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_d x_{id} + \epsilon_i. \]

Note that \( d \) is possibly much larger than the number of columns in our data frame:

1. **Categorical Variables**: The column `Siding` has different levels: vinyl, metal, wood, brick, stucco, and stone. This automatically gets expanded into 5 variables.

2. **Basis Expansions**: We can expand the column `Bedrooms` into variables `Bedrooms, Bedrooms^2, Bedrooms^3, ....`

So far, we’ve assumed that the exact model is known in advance.
The Problem of Model Selection

What if we are considering multiple candidate models?

For example:
- Should our model be linear in the number of bedrooms? Quadratic? Cubic?
- Should we include siding in the model or not?

We want to use the data to inform our choice of model! This problem is known as **model selection**.

Model selection consists of two questions:

1. **Given two models**, how do we determine which one is “better”?
2. **How do we efficiently search through the space of possible models?**