4. Complexity and Model Comparison with a hold out set

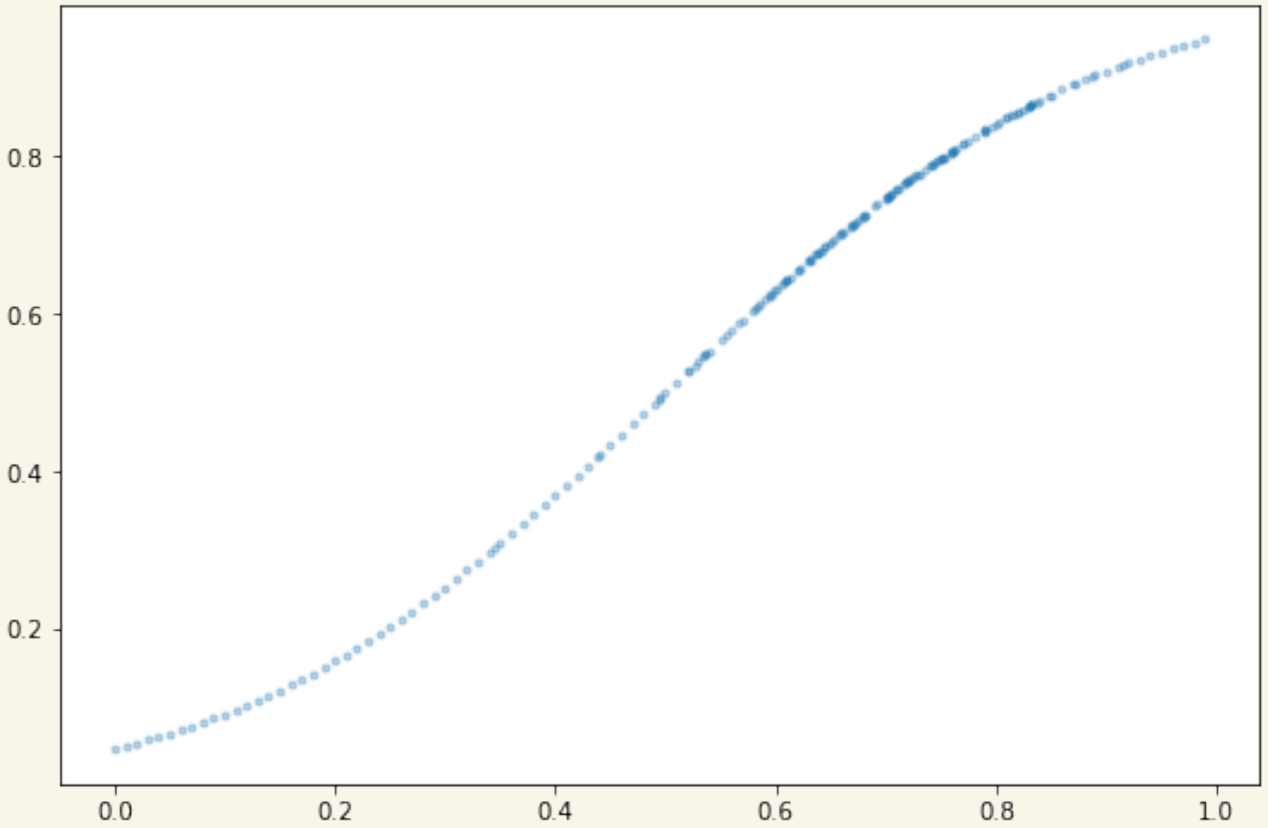
Part of:

The Essence of Learning

All of AI is the estimation of functions.

Univ.AI

Our Example



Consider a very simple scenario, where the probability of voting for Romney is a function only of how religious the population in a county is.

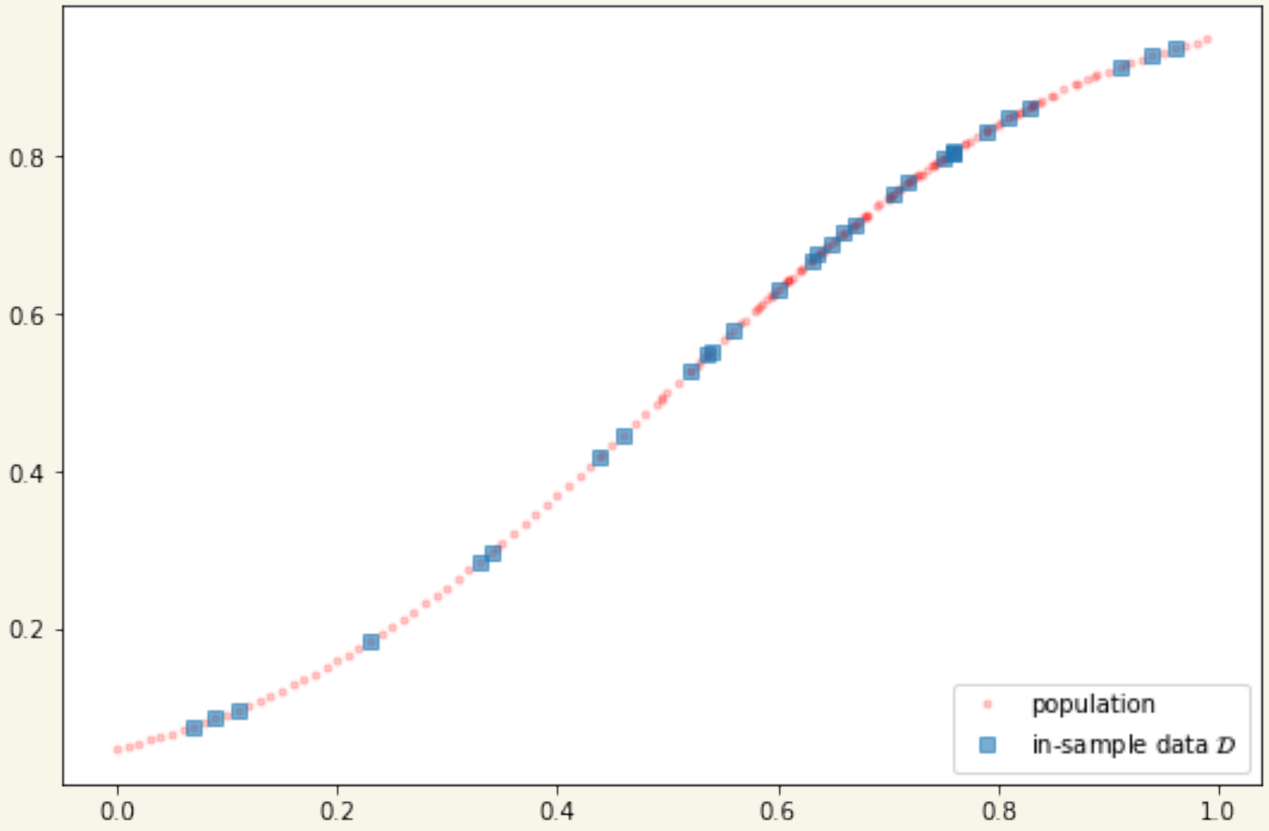
In other words y_i is data that pollsters have taken which tells us their estimate of the fraction of people voting for Romney and x_i is the fraction of religious people in county i.

Let us assume that we have a "population" of 200 counties.

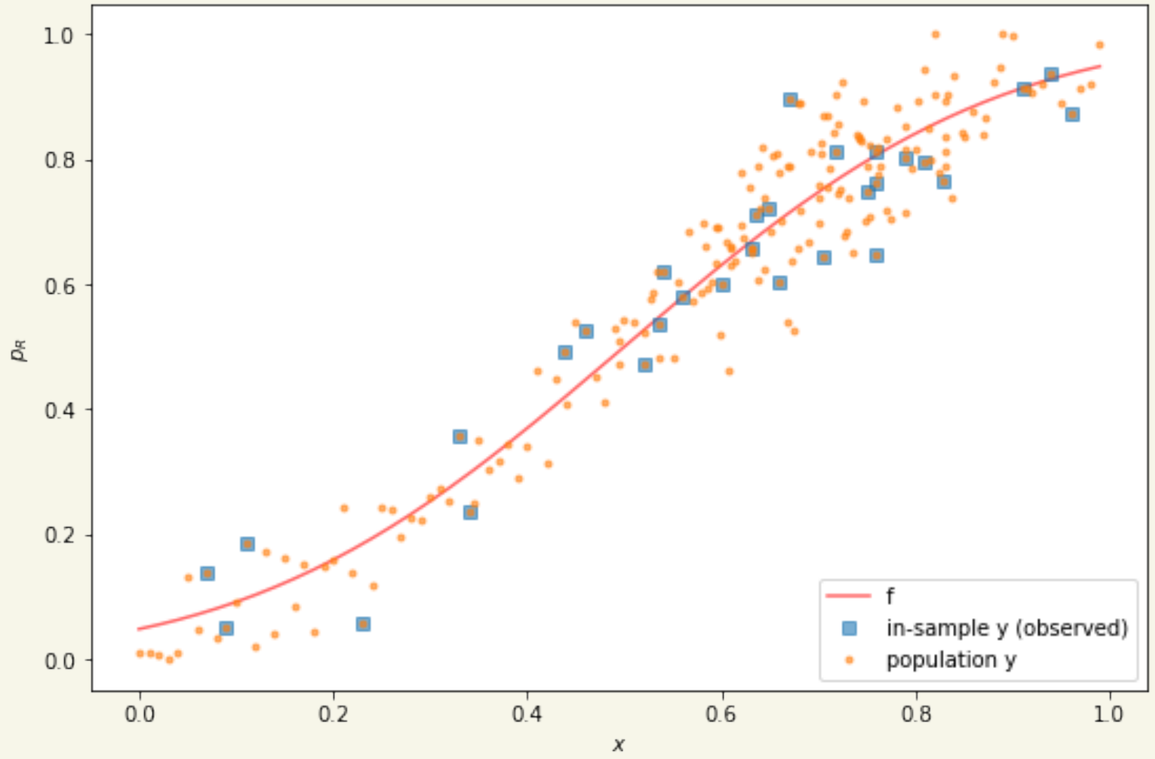
A sample from this population

Lets assume that out of this population of 200 points we are given a sample $\mathcal{D}$ of 30 data points. Such data is called **in-sample data**. Contrastingly, the entire population of data points is also called **out-of-sample data**.

Now pretend the red dots are taken away, and you are left with the blue squares. Our job look at different hypotheses and find the best one amongst them.



A noisy population (and thus sample)



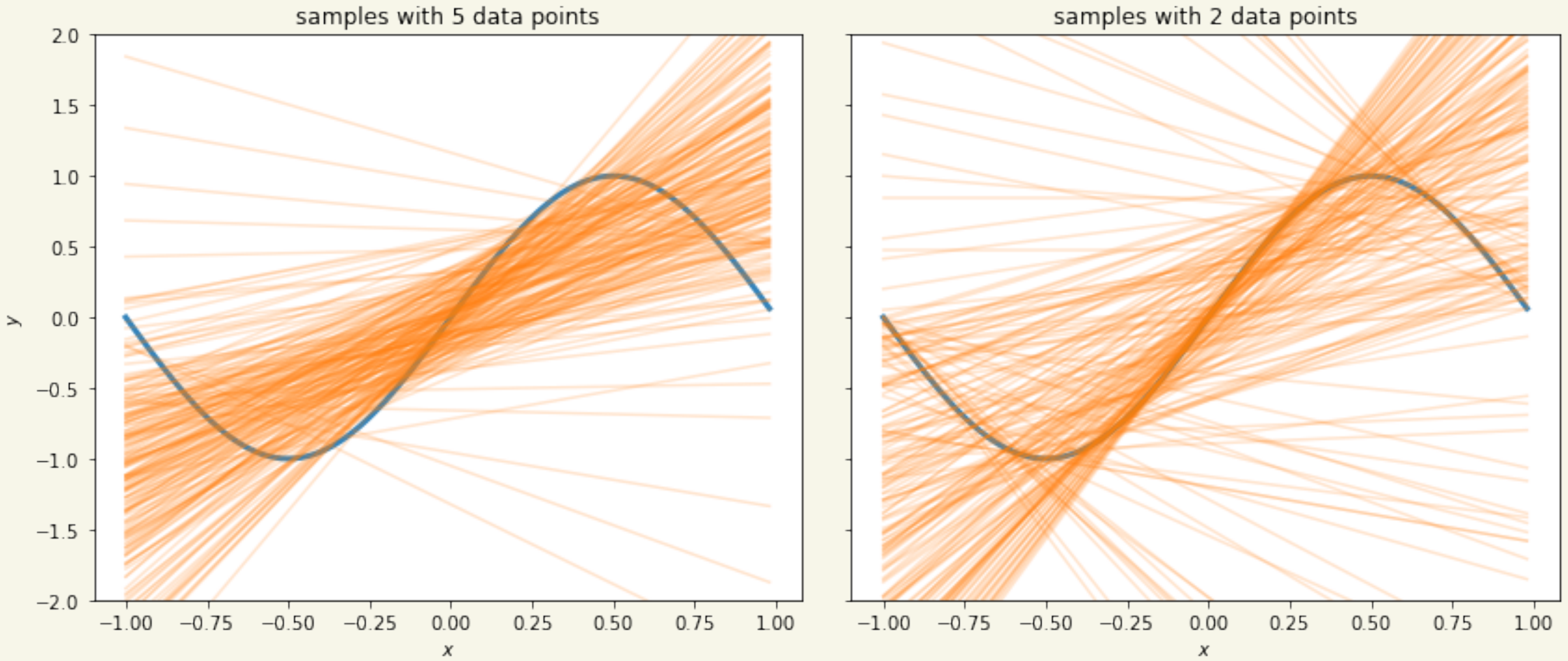
Now let us assume there is some noise in our measurements..

Predictions must be robust

- predictions cannot be like our 20th order polynomials above
- high sensitivity to sample is a very bad sign.

There are two ways to fix this

- get larger samples, so that they are more representative (not always possible)
- fit simpler models
- the complexity of a model you can fit depends on the size of the dataset you have


samples with 5 data points
samples with 2 data points
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
x

MODEL COMPARISON

Why? We want a model thats complex enough but not so complex that it overfits.

- we want to choose which Hypothesis set is best
- and we know it should be the one that minimizes risk on the sample

$$A : R_{\mathcal{D}}(g) \ \ smallest\, on\, \mathcal{H}$$

But I can make sample risk go to zero by choosing a polynomial with as many coefficients as there are points in my sample. This is called interpolation.

But it will be overfit hugely, and the error on the population at large will be huge.

Population error R_{out} should be small

$$B : R_{out}(g) \approx R_{\mathcal{D}}(g)$$

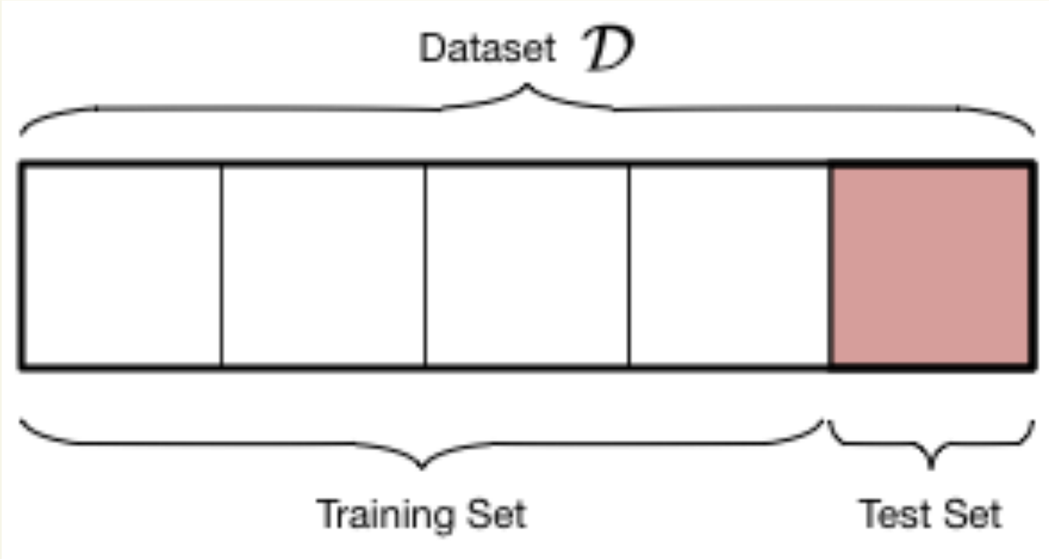
B: Population, or out-of-sample risk is WELL estimated by in-sample risk, and is thus small

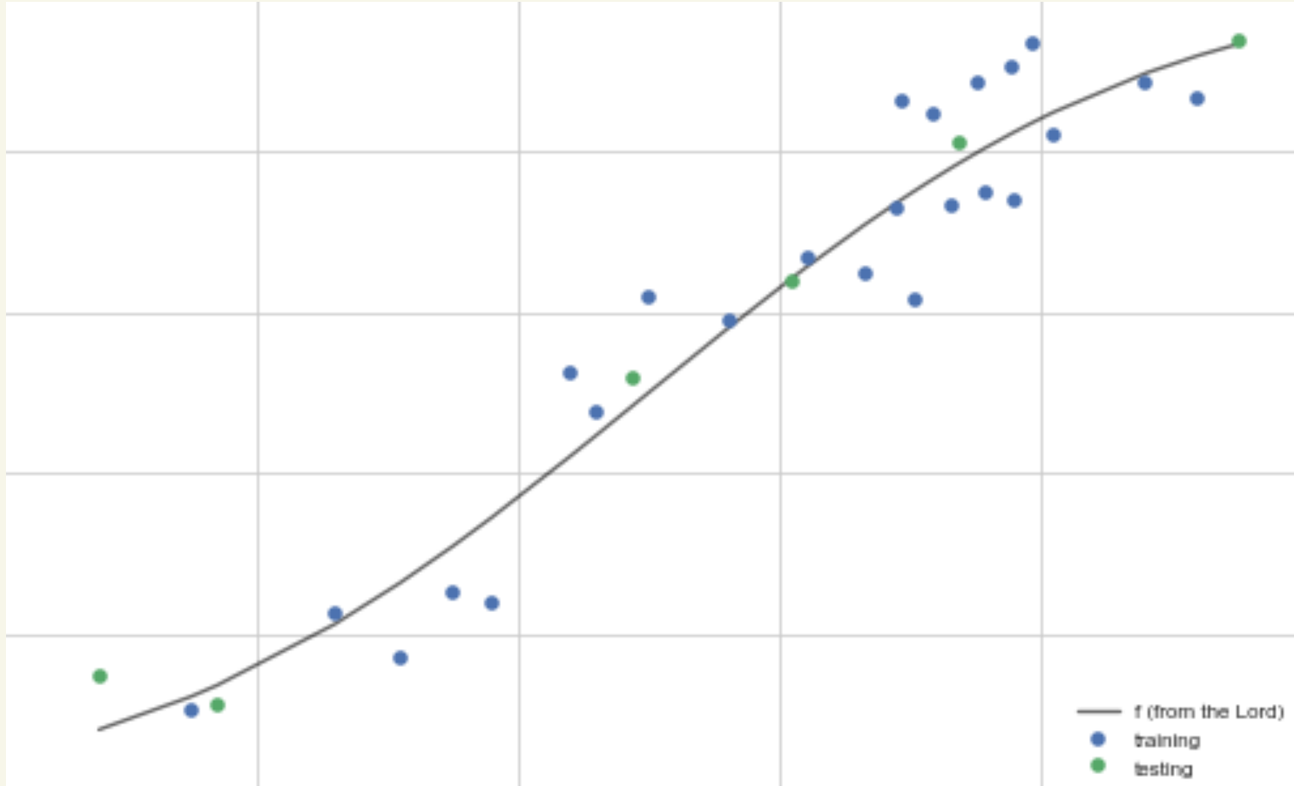
SO

- The sample must be representative of the population!
- The model should not be too complex, or a small perturbation in coverage going from sample to population will sink us!

How do we estimate population error R_{out} ?

Hold out some data



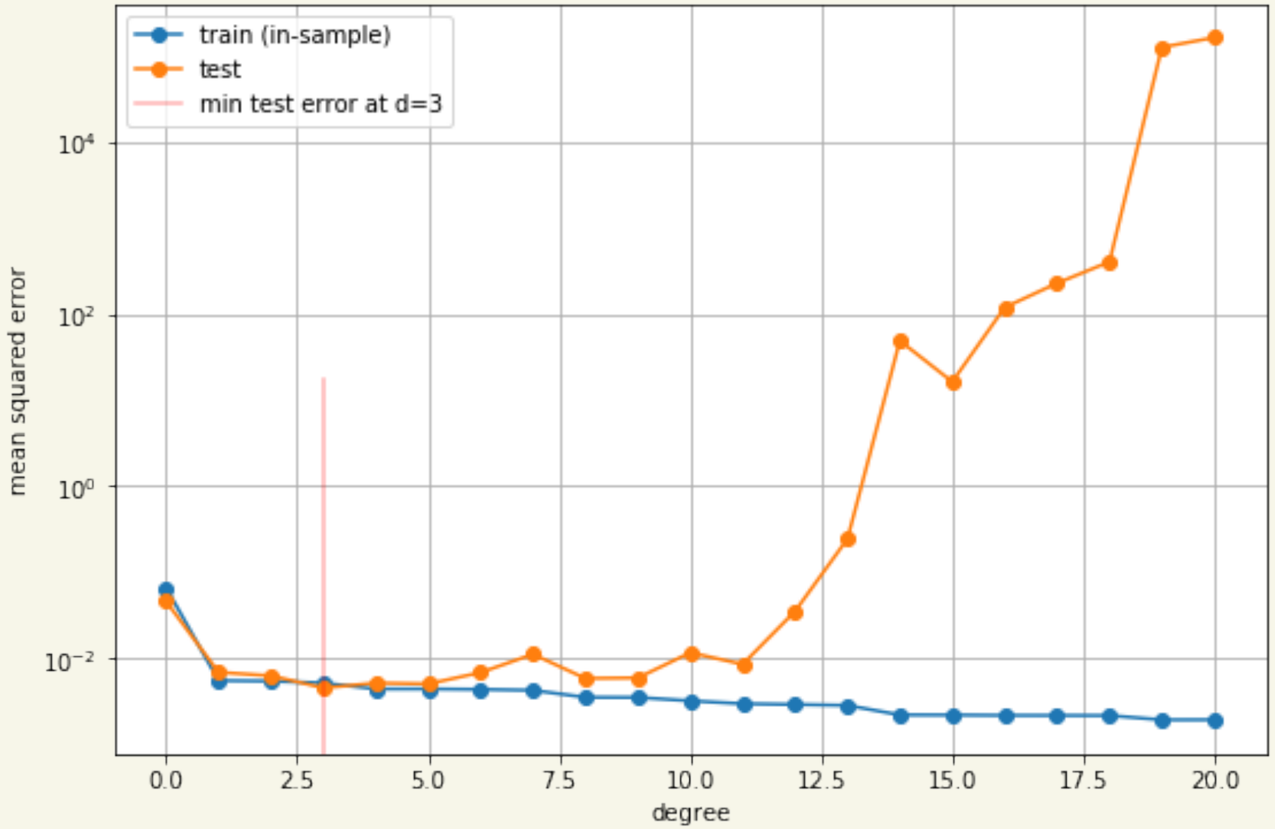


Held-out set is a proxy for the population

- You are now fitting on the set thats the subtraction of the held-out set from the sample. We will now call this the **training set**
- One can show that the error on the held-out set estimates the population error. This set is also called a **validation** set (or sometimes a **test set**).
- This is not surprising as its just another sample.
- The point here is that it will have as much similarity to the population as the rest of your sample..
- ..But also that the diversity in the held out set will act as a perturbation which will penalize overfit models

Evaluate error on held-out set

1. For each polynomial degree d, we minimize the training risk (we fit a model for EACH hypothesis space: $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{20}, \mathcal{H}_{21}, \ldots$)and find the coefficients which give us the best fit model
2. We evaluate the error or risk on the validation set
3. Once the validation set risk starts going up we are overfitting
4. We balance $R_{out}(g) \approx R_{\mathcal{D}}(g)$ and $A : R_{\mathcal{D}}(g)\ smallest\ on\ \mathcal{H}$ by choosing a model with the lowest validation risk



Complexity Plot