MORE ABOUT APPROXIMATION

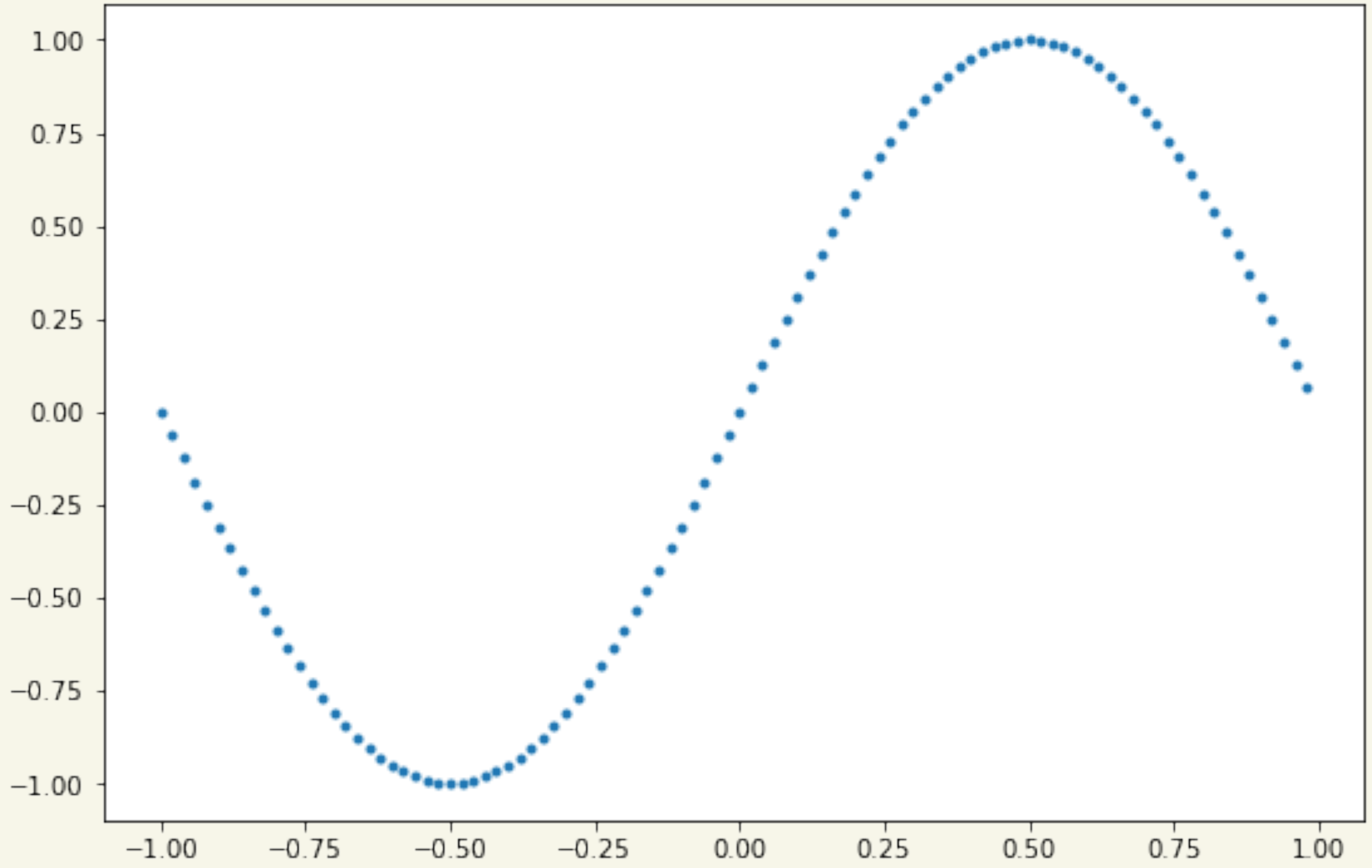
or Learning Without Noise...

Part of:

The Essence of Learning

All of AI is the estimation of functions.

Our target, a sine with no noise.



Constructing a sample from a population

Well usually you are only given a sample. So there is nothing to construct.

But a sample is a set of (x, y) points chosen from the population.

If you had the population you could construct many samples of a smaller size by randomly choosing points into smaller sizedsamples of such points.

This mechanism helps us experiment by **simulation**. Since we have both the population and the samples, we can see what size of samples are needed.

Robustness of fits.

In life you usually only get a sample.

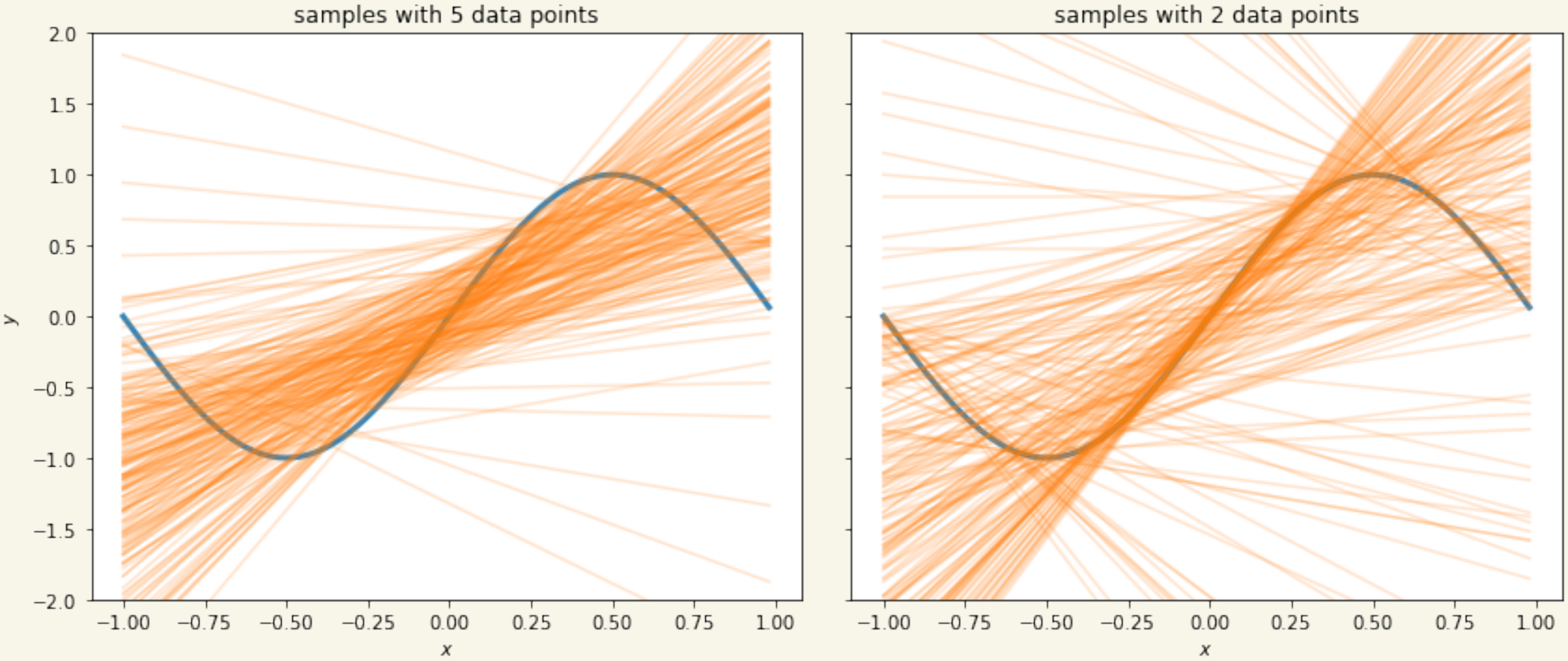
At the end of the day you want to predict on new data which you have not seen.

Equivalently you want to predict on the population.

You want your fits to be robust. If they changed too much from sample to sample, how could you be sure that the results you got on one sample you were generalizing to the population?

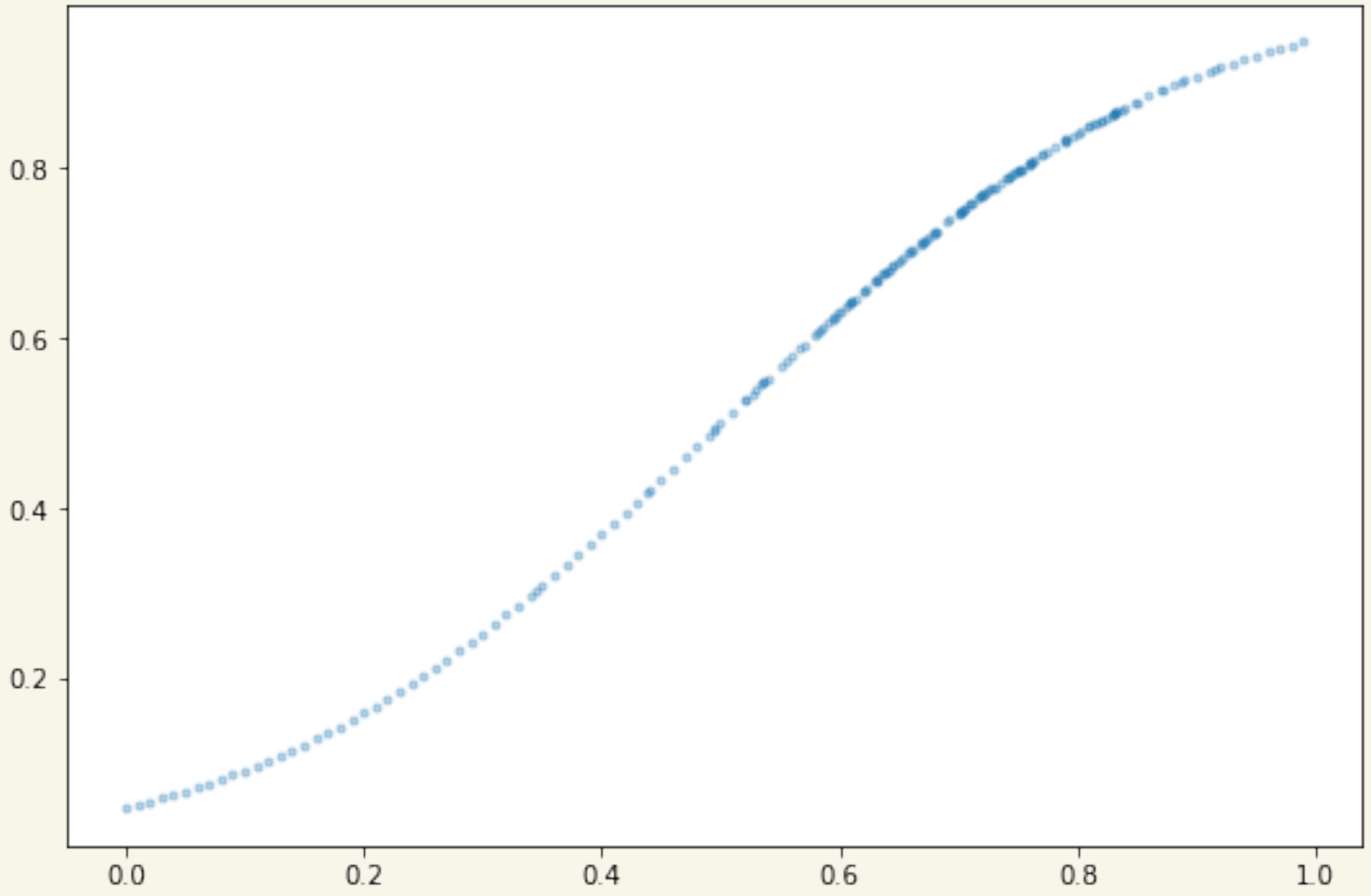
Informally, one condition for this is that your sample is representative of the population.

We create 200 samples of size 5 and size 2 each and fit straight lines to them. What do you conclude?



Univ.AI

Another Population 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.

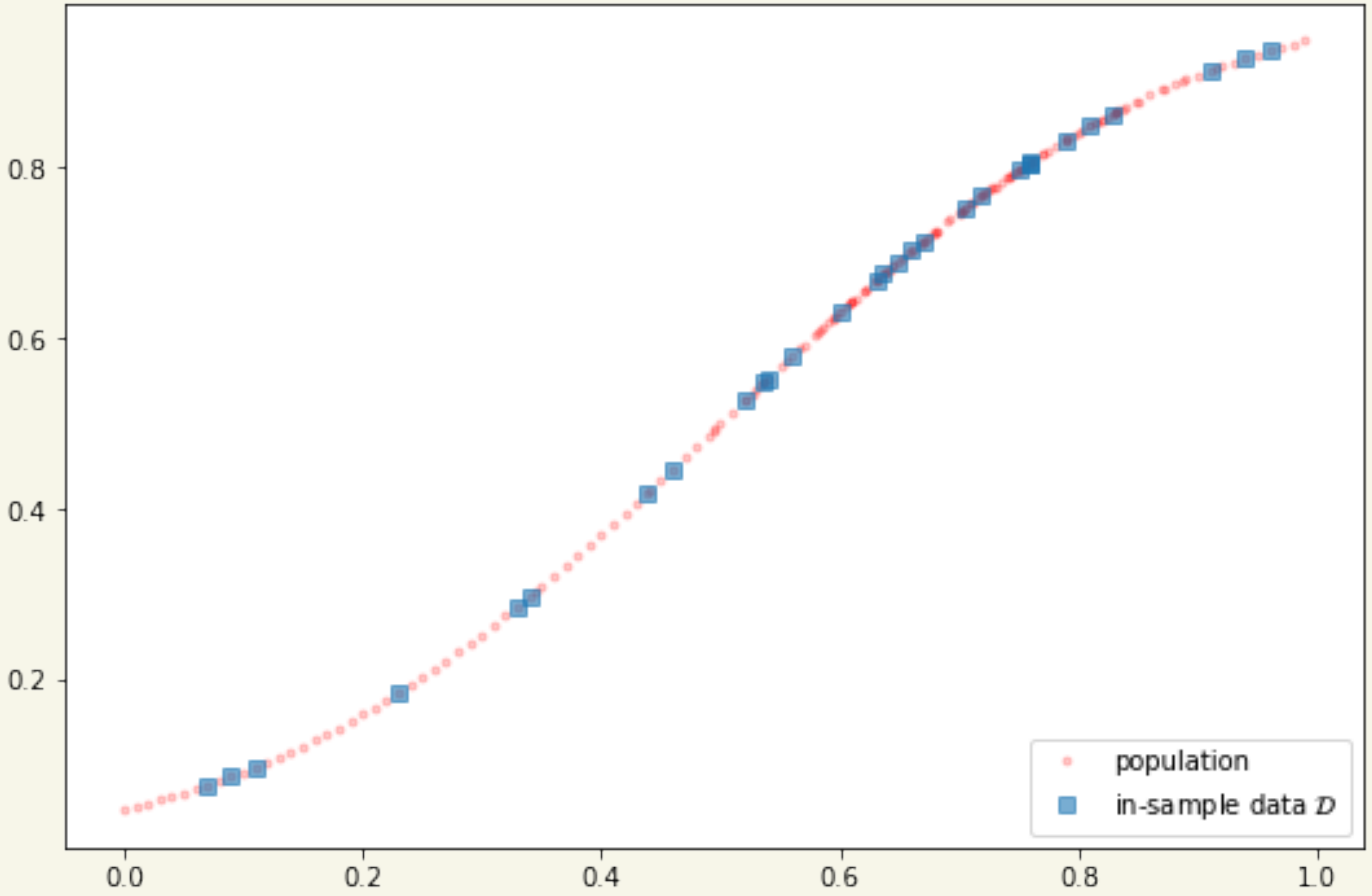
In other words y_i is data that pollsters have taken which tells us their estimate of the fraction of people voting for Romney and x_i is the fraction of religious people in county i.

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

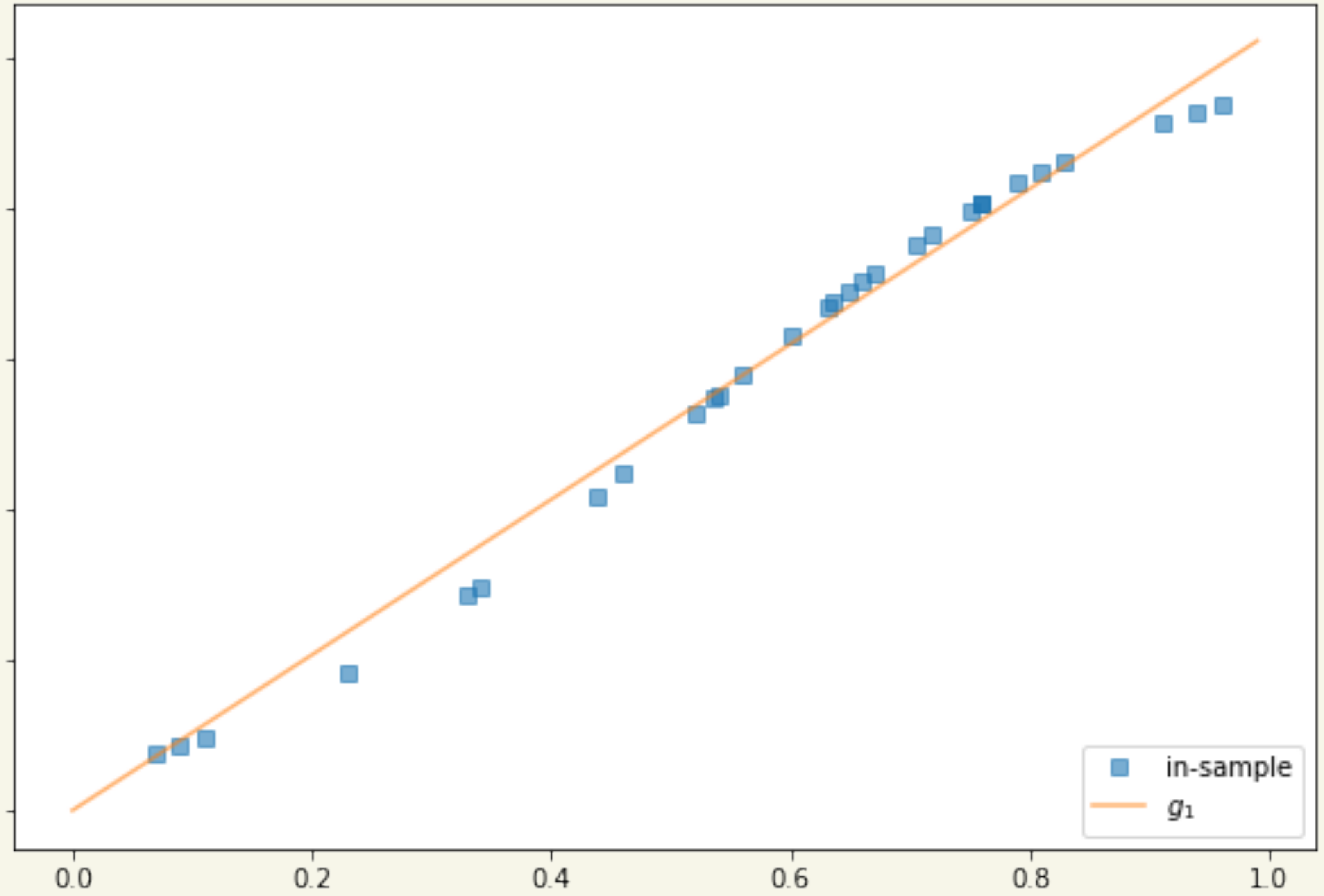
A sample from this population

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.



Bias: First try straight lines

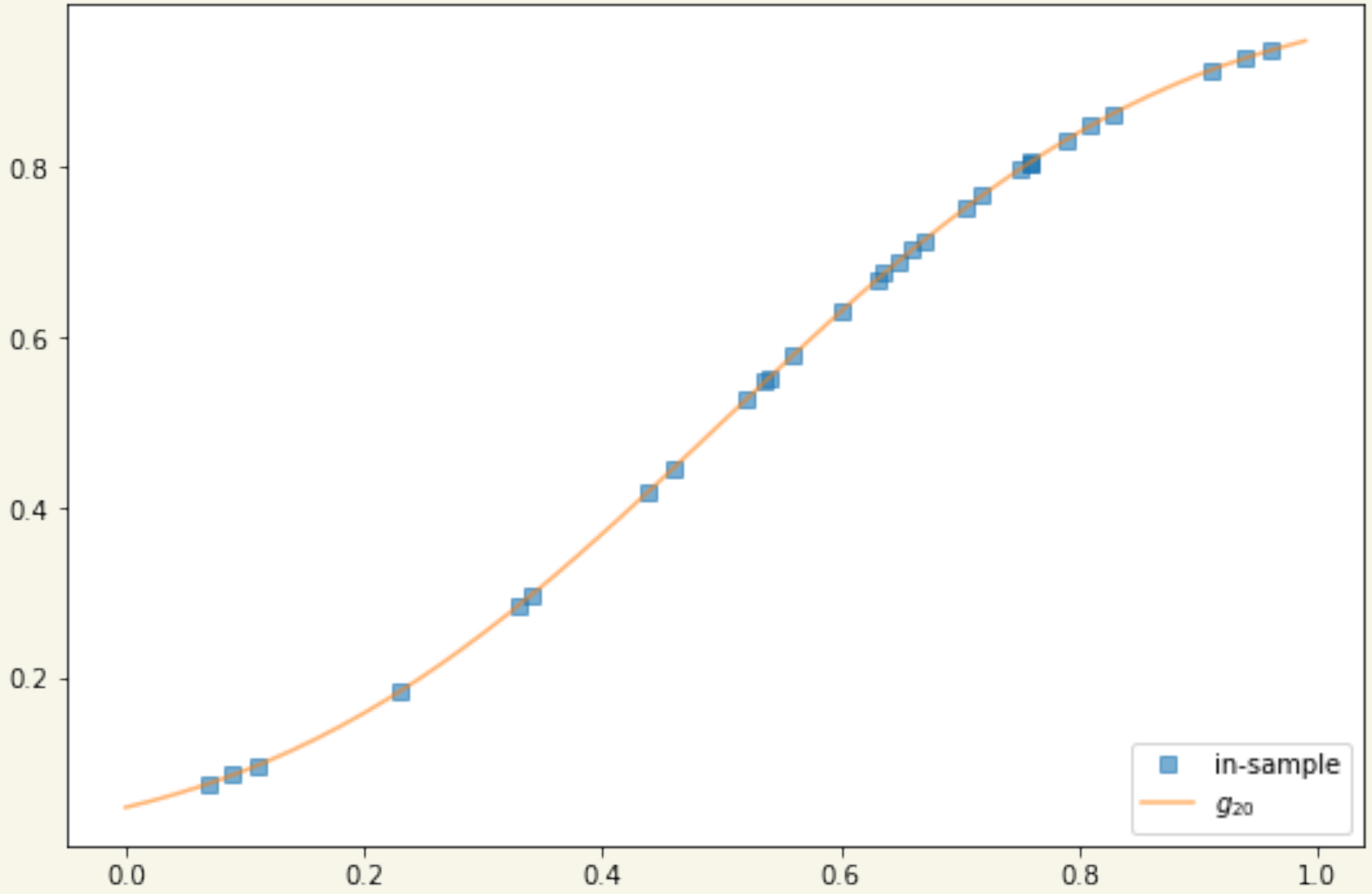


Fit with straight lines. That is hypothesis set $\mathcal{H}_1$:

$h_1(x) = a_0 + a_1 x$ and find the best fit $g_1(x)$

This is **Bias or Mis-specification Error** obtained by using $g1(x)$ in place of the actual $f(x)$.

Now try a 20th order polynomial



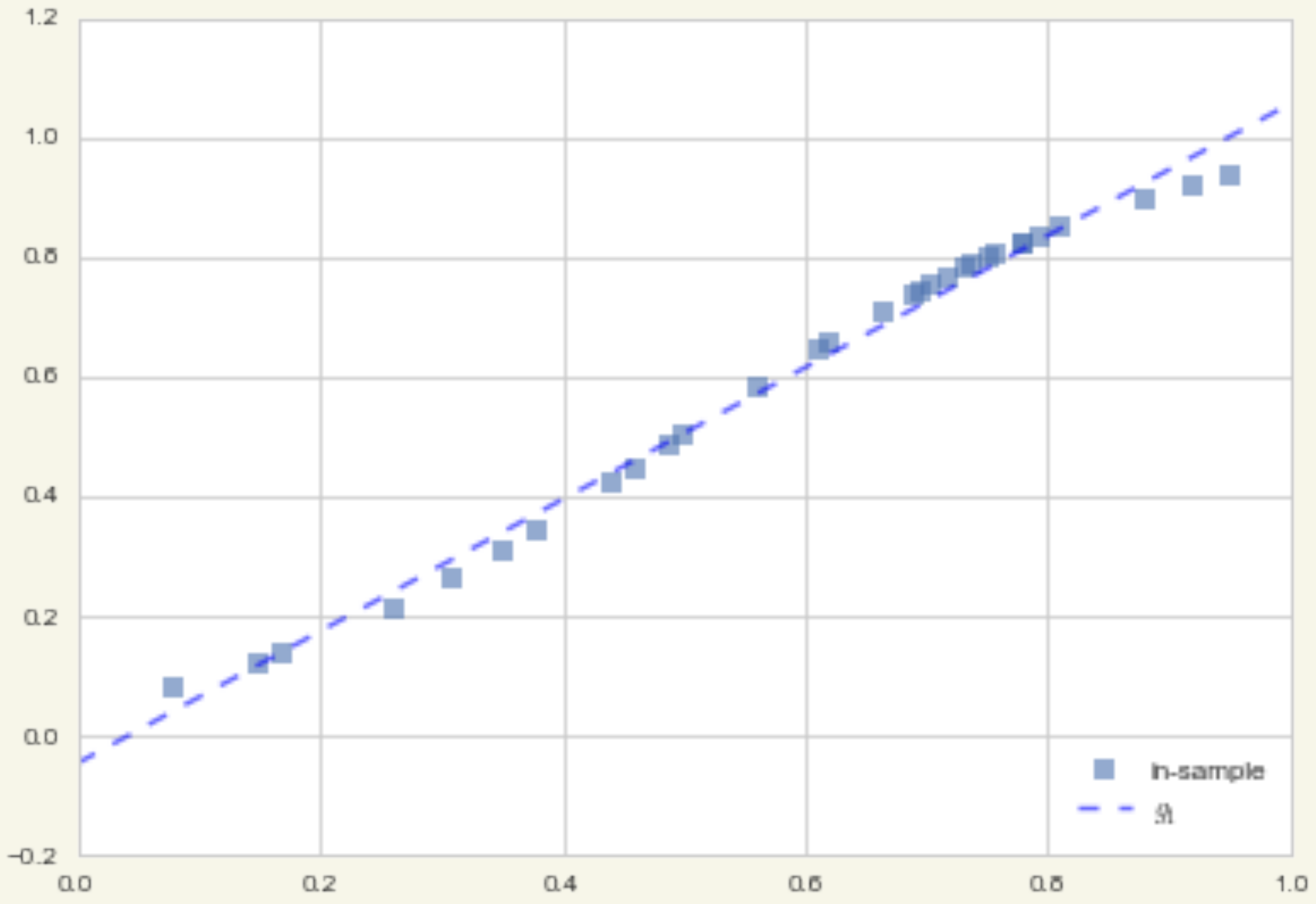
it is clear that the space of straight lines $\mathcal{H}_1$ does not capture the curving in the data. So let us consider the more complex hypothesis space $\mathcal{H}_{20}$, the set of all 20th order polynomials $h_{20}(x)$:

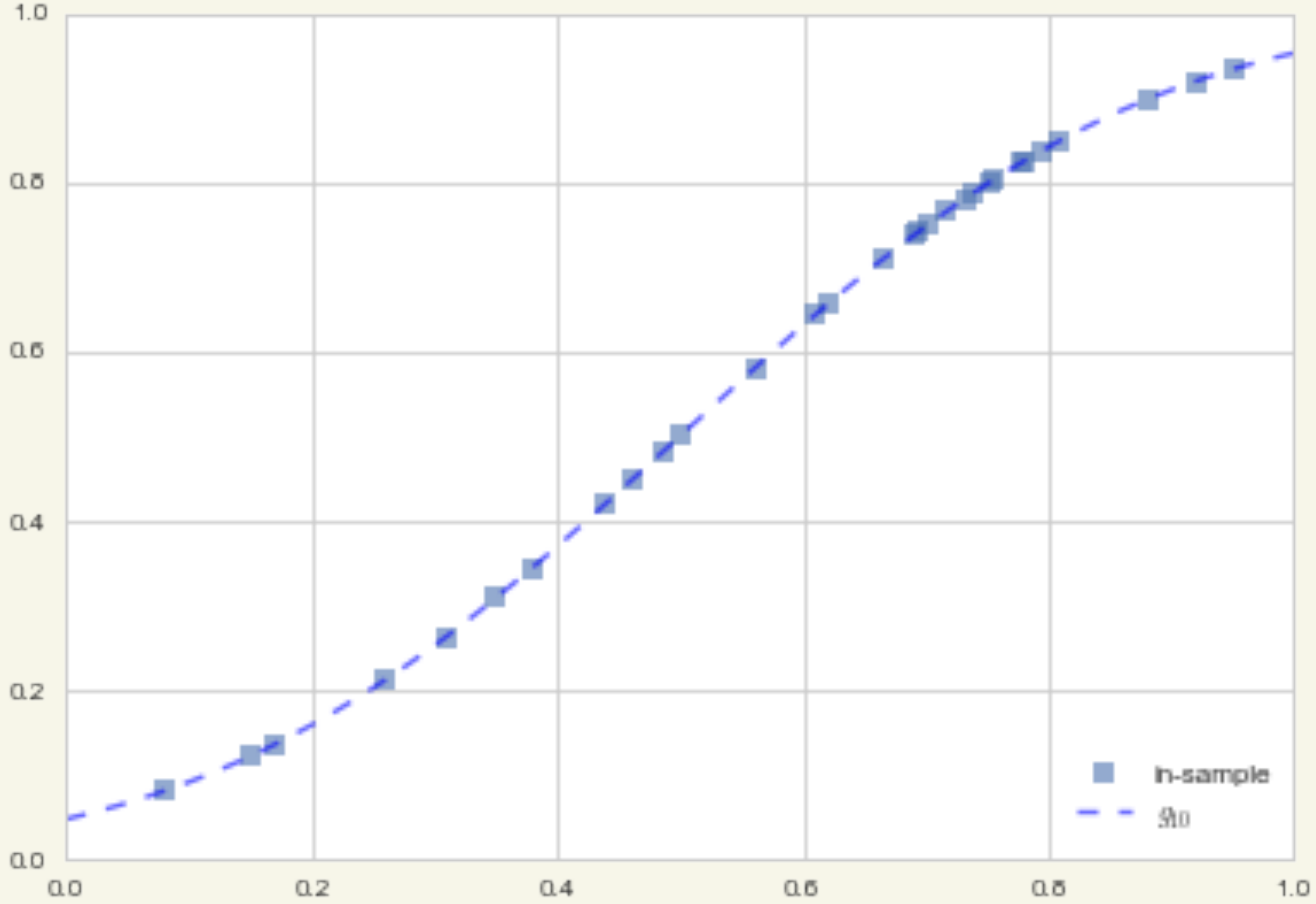
$$h_{20}(x) = \sum_{i=0}^{20} a_i x^i \,.$$

The best fit $g_{20}(x)$ seems to be great!

Univ.AI

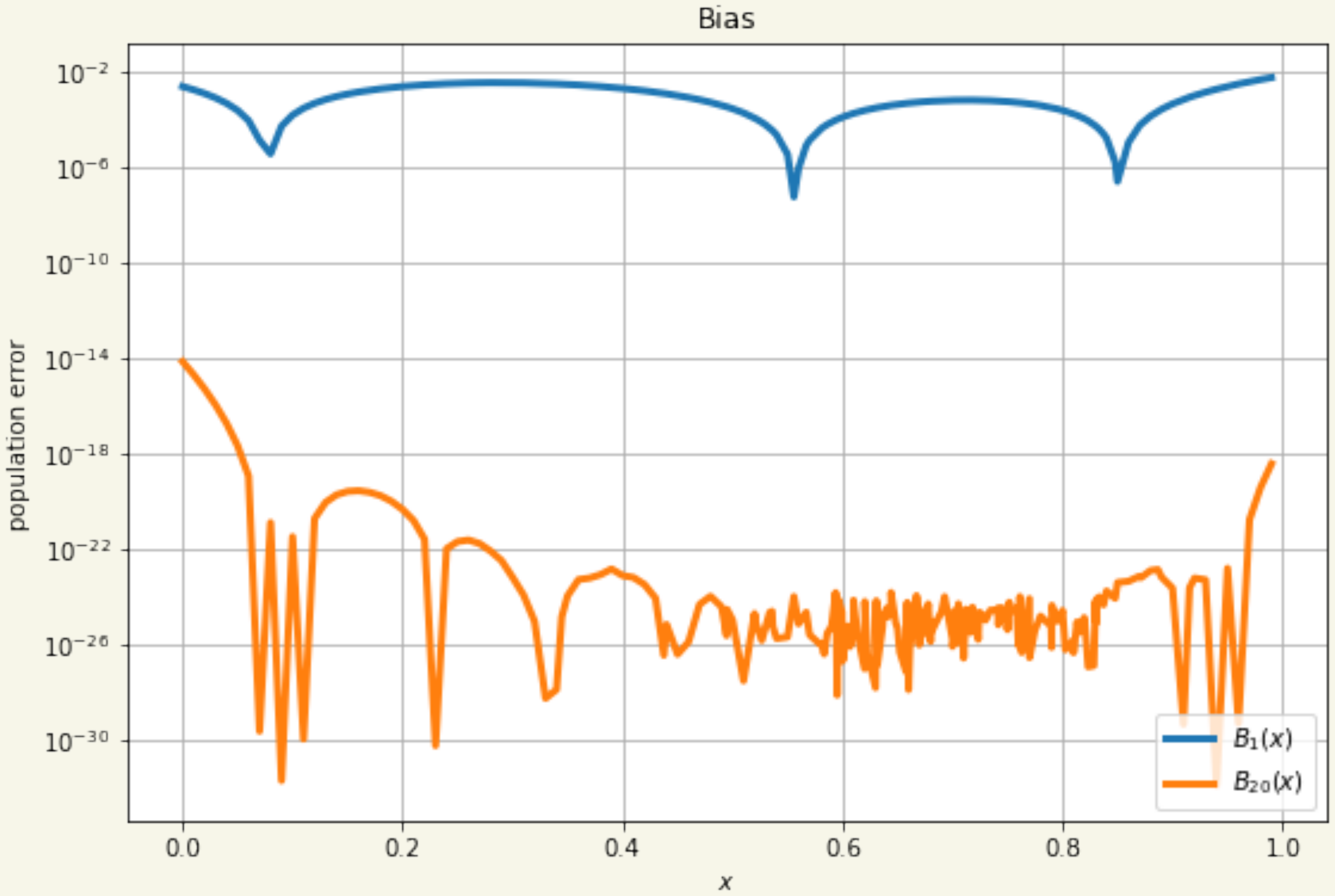
A sample of 30 points of data. Which fit is better? Line in $\mathcal{H}_1$ or curve in $\mathcal{H}_{20}$?





Plotting the squared error

This shows that the point-wise squared error on the entire domain of the function is lower for $g_{20}(x)$ and it thus seems to be a better fit. In other words the bias is lower for $g_{20}(x)$ compared to $g_1(x)$.



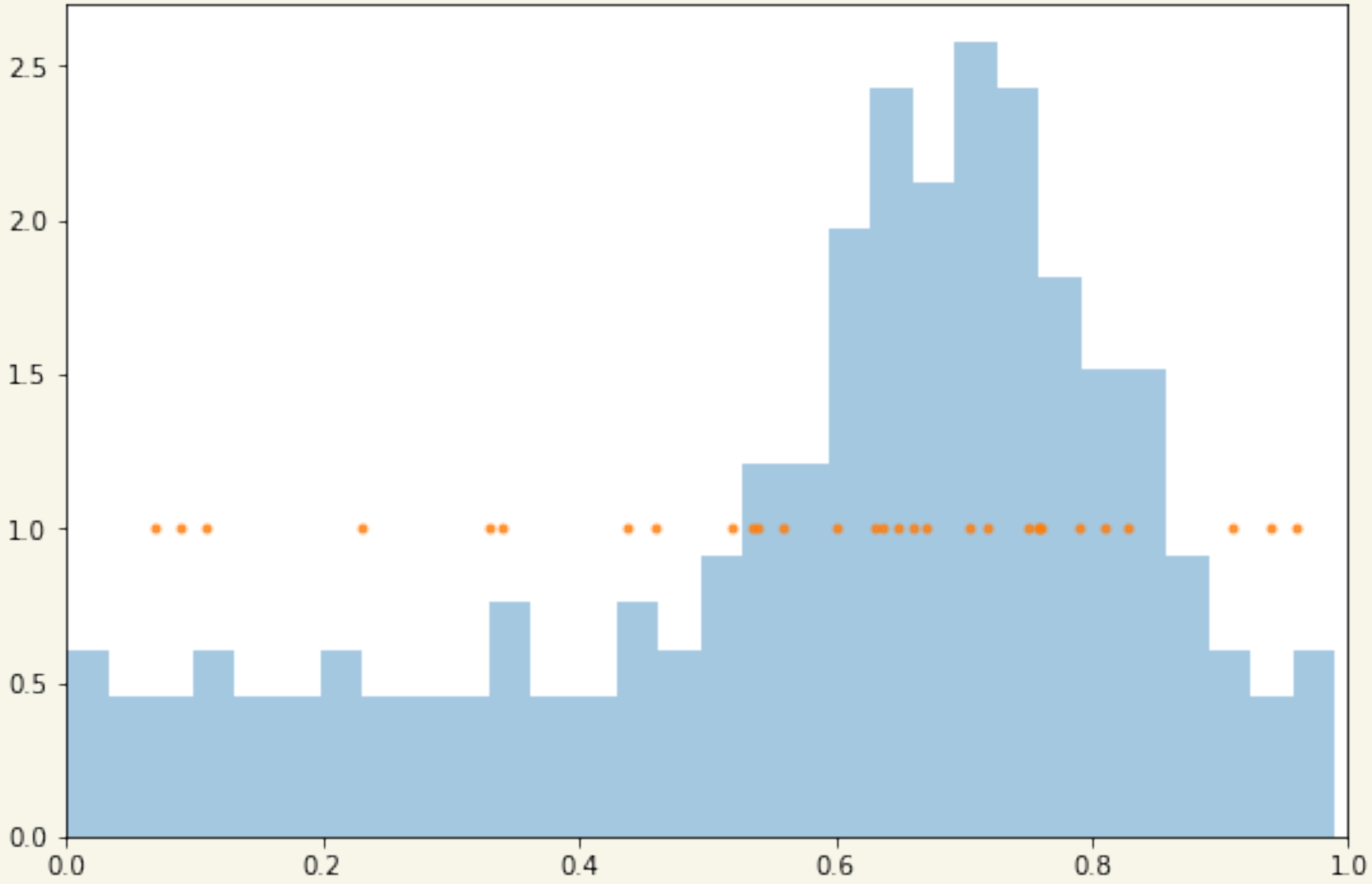
How do we generalize from sample to population?

What we'd like to do is **make good predictions**. In the language of cost, what we are really after is to minimize the cost **out-of-sample**, on the **population** at large. But this presents us with a conundrum: *how can we minimize the risk on points we havent yet seen*?

(we have seen the population in our simulation here, but remember that in the real world we dont get to see the population)

This is why we (a) minimize the risk on the set of points that we have to find g and then (b) hope that once we have found our best model g, our risk does not particularly change out-of-sample, or when using a different set of points. And that the rish/error remains low!

Intuitively, to do this, we need to ask ourselves:



How representative is our sample?

For example, if we want to use g, our estimand of f to predict for large x, or more religious counties, we would need a good sampling of points x closer to 1. And we wont do well if we try and predict low-religiousness counties from a sample of high-religiousness ones.

Our points seem to follow our (god given) histogram well.

Statement of the Learning Problem

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

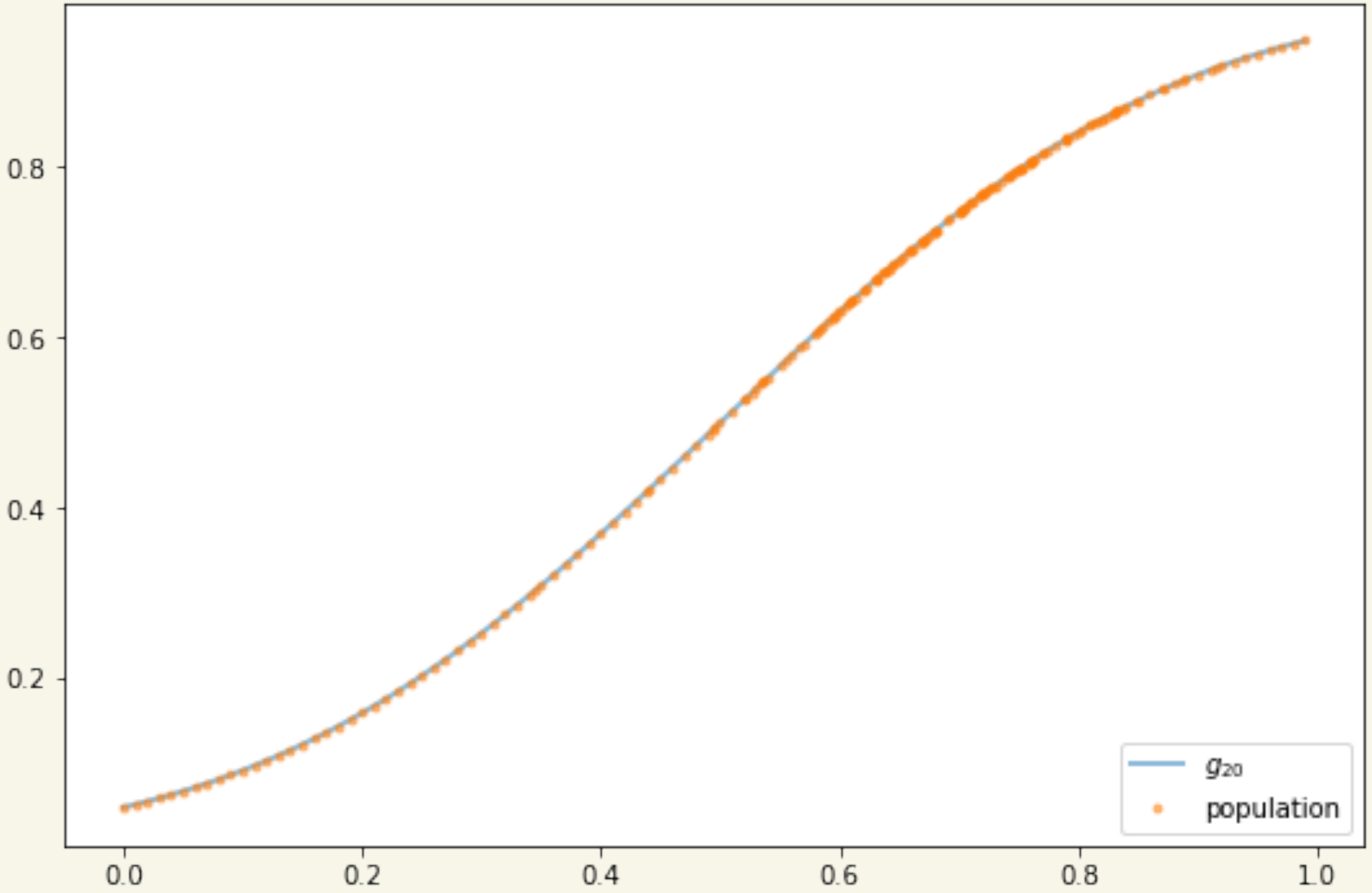
A: In-sample risk is small

B: Population, or out-of-sample risk is WELL estimated by in-sample risk.

The sample must be representative of the population!
Thus the out of sample risk is also small.

HOW DO WE ENSURE THIS? (see next chapter)

Prediction on population: good!



Univ.AI