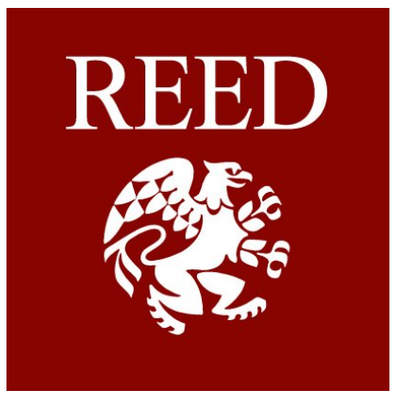
Machine Learning & Advanced R

Albert Y. Kim
Assistant Professor
Statistical & Data Sciences, Smith College

MassMutual Data Science Summer 2019 Bootcamp
Tuesday 2019/7/22

About me & my background



SMITH COLLEGE

Machine Learning

WAYMO

NETFLIX

Prediction!

ITCH FIX

The General Modeling Problem

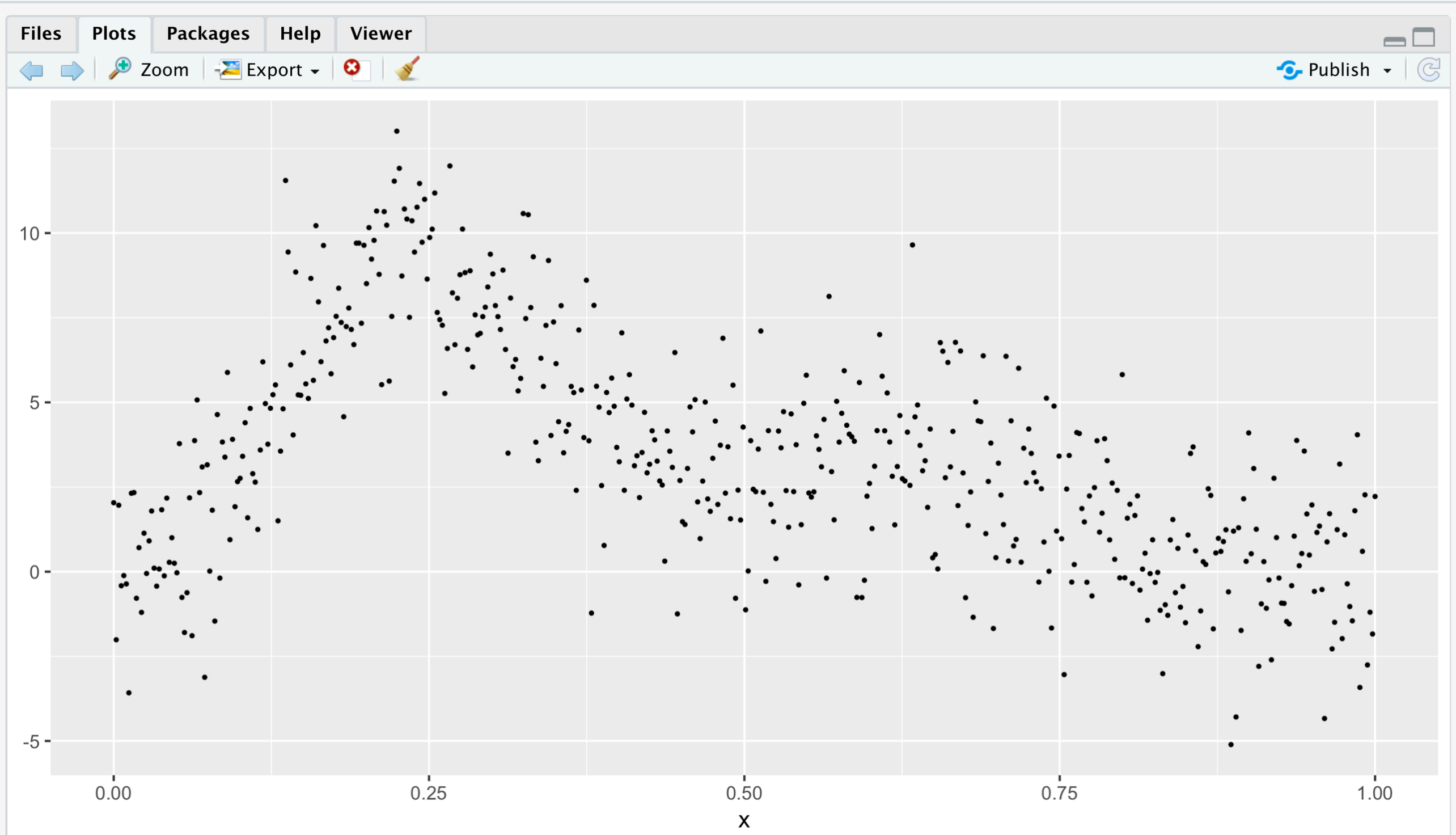
Machine Learning as Modeling

True (Unknown) Model: $y = f(\vec{x}) + \epsilon$

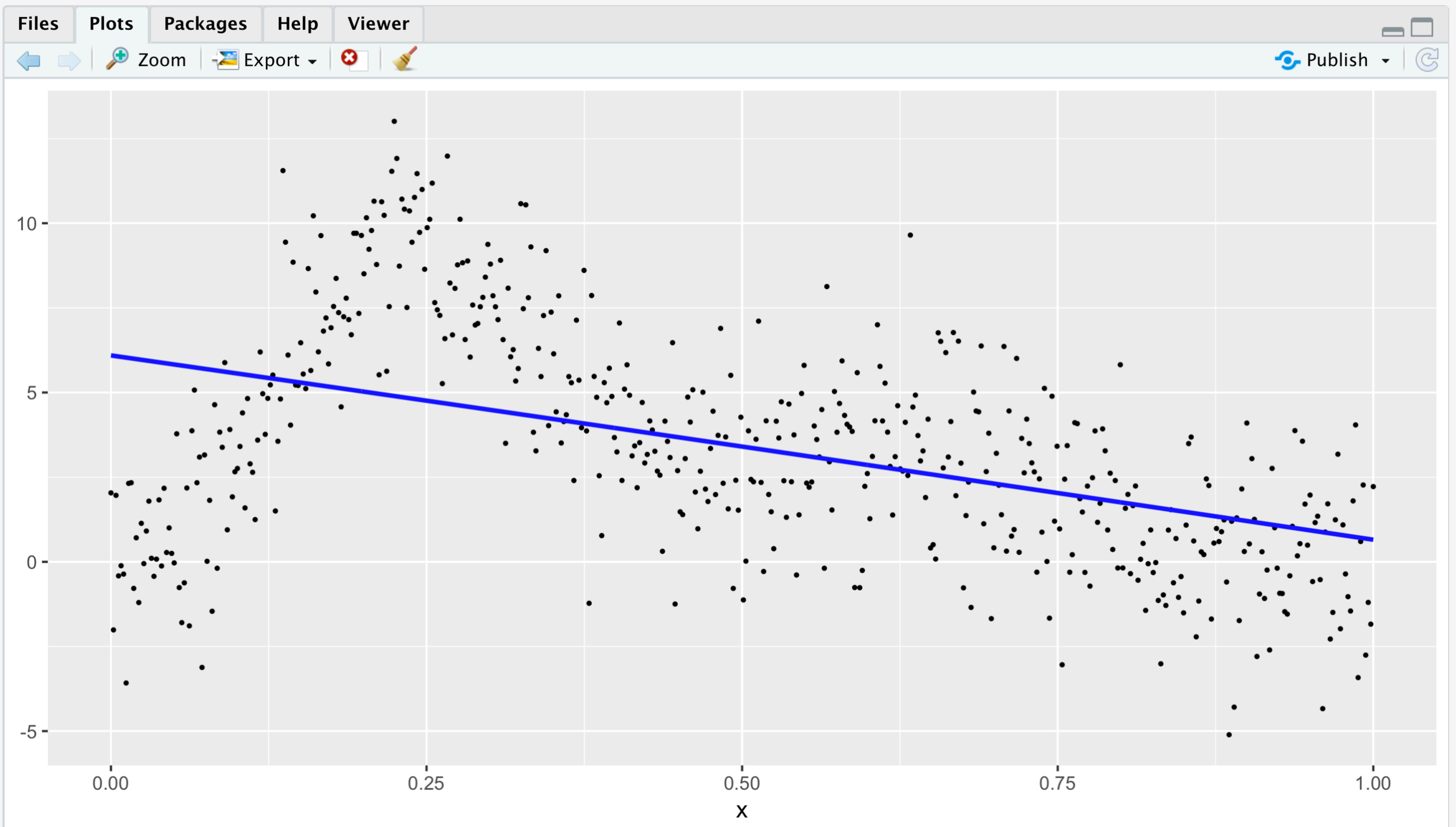
Approximated Model: $\hat{y} = \hat{f}(\vec{x})$

Now to the blackboard for Chalk Talk #1…

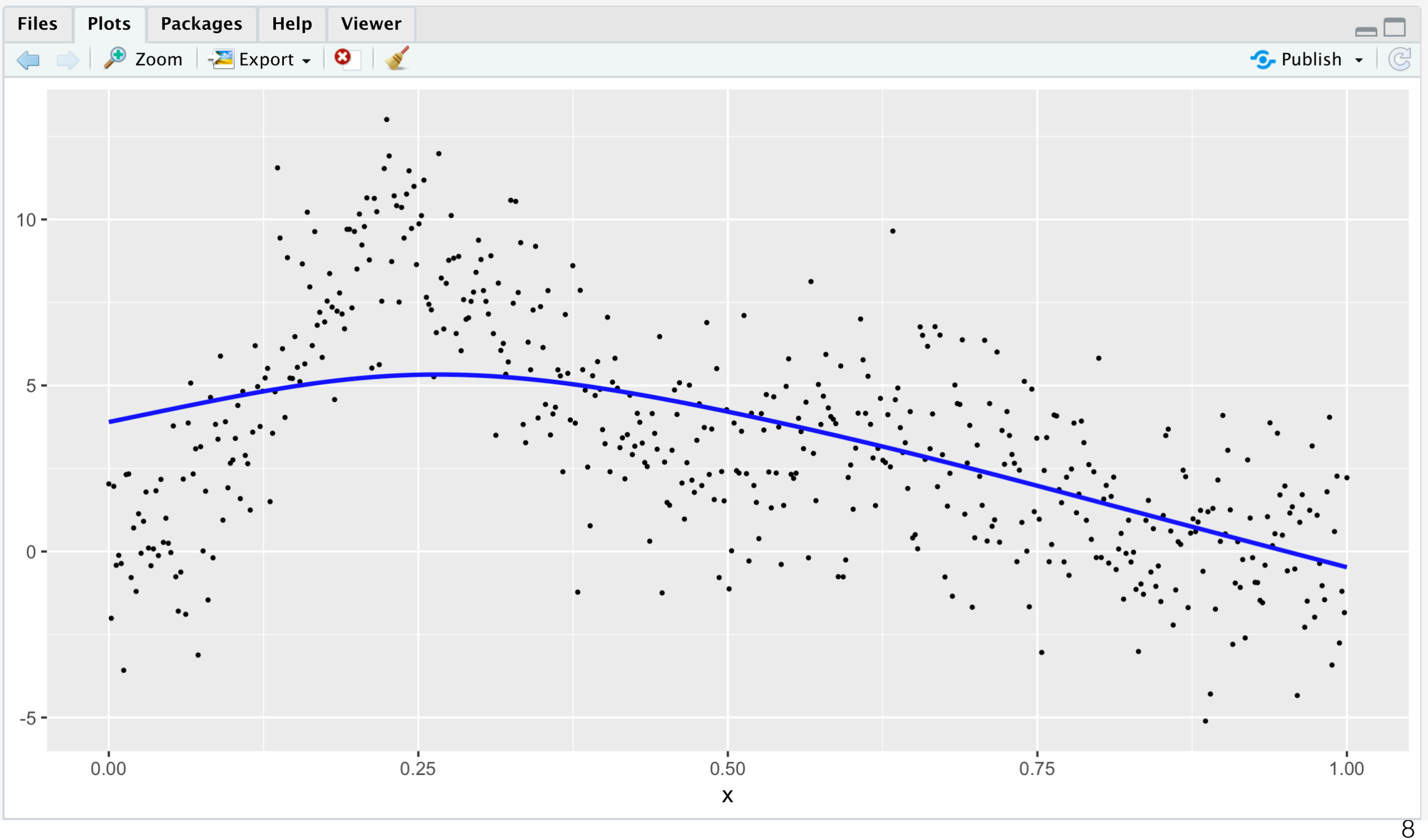
Given Data (x, y) from "unknown" $f(x)$



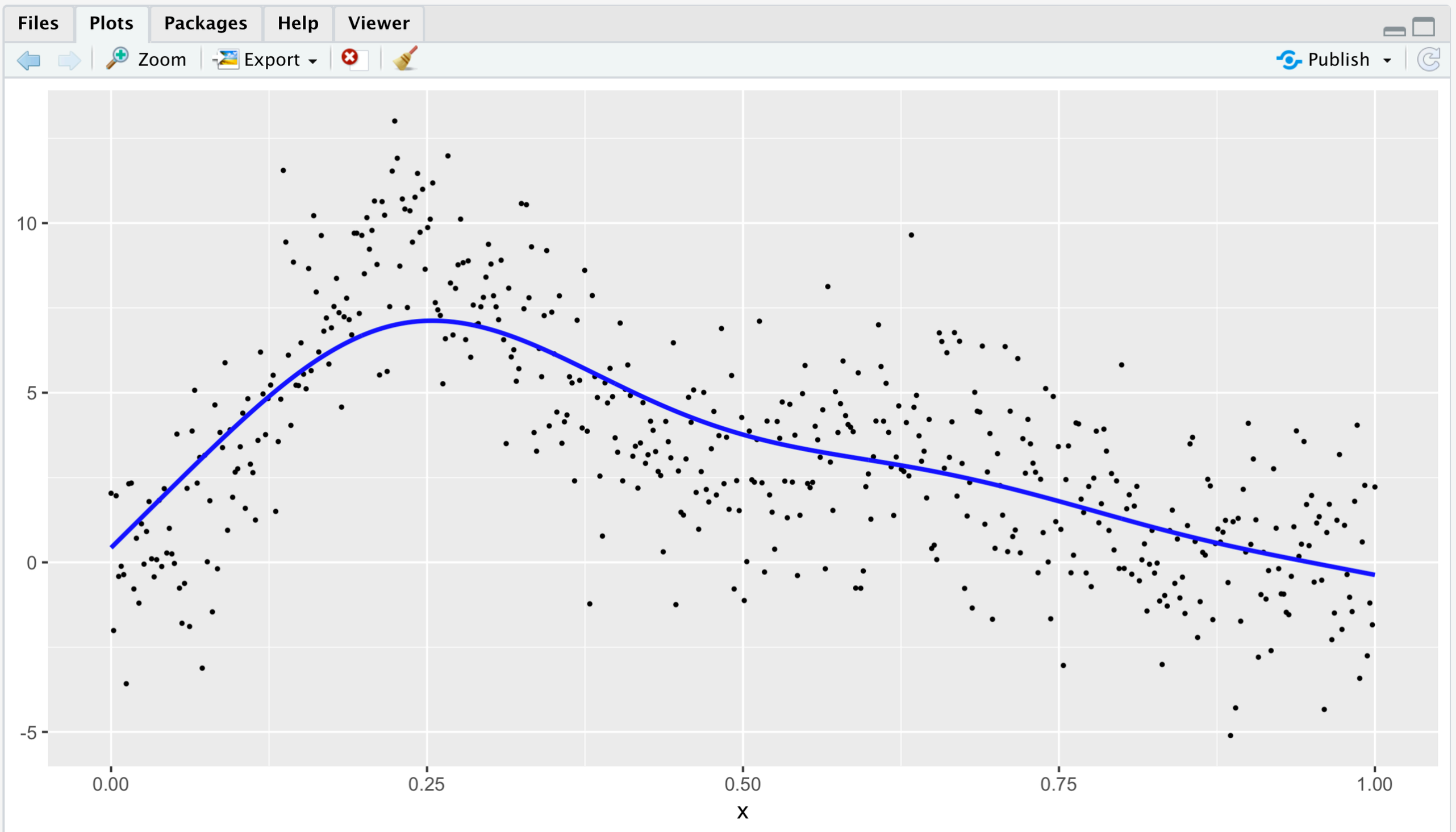
Approximate (i.e. "fit") a Model $\hat{f}(x)$



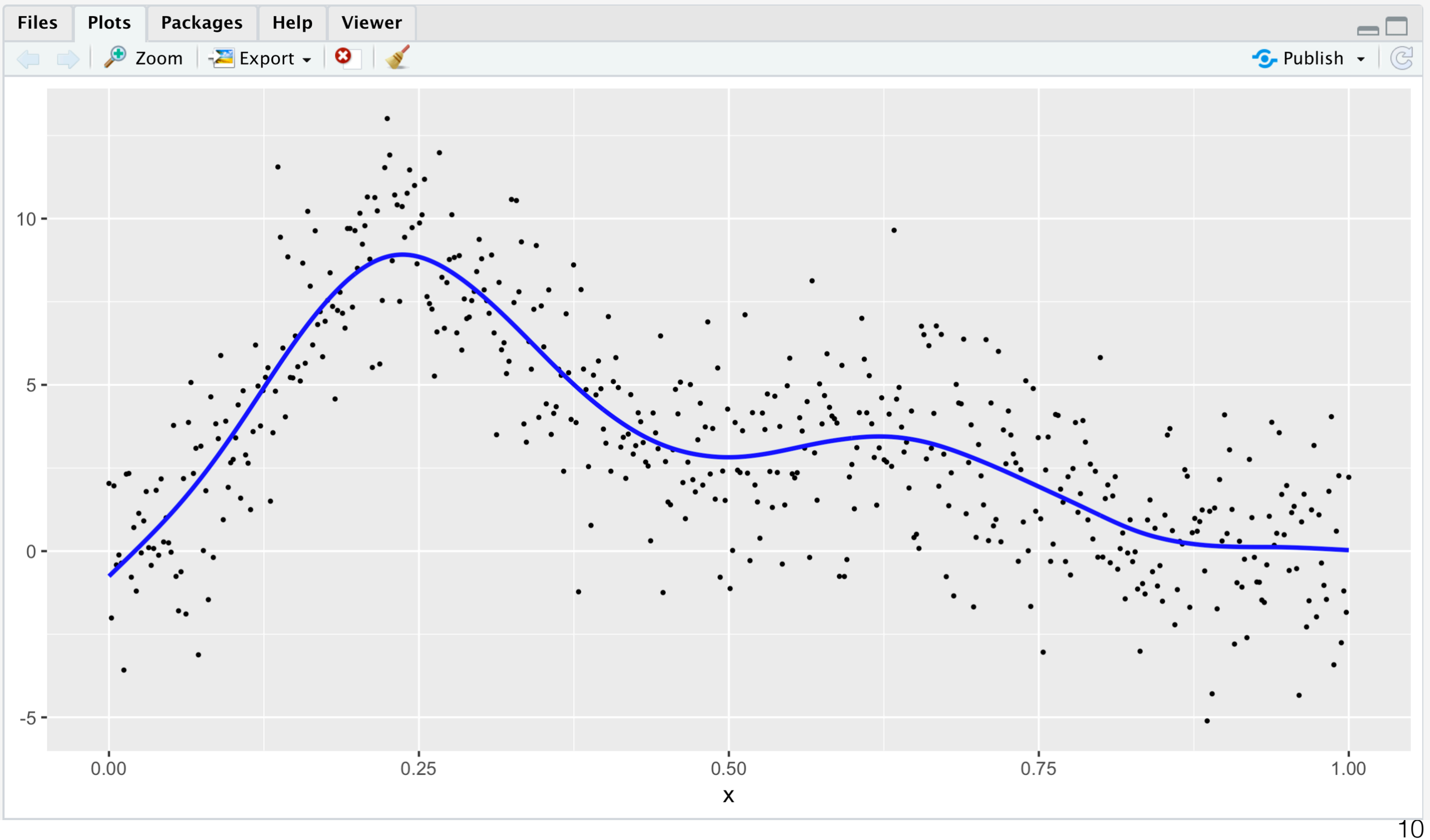
How about this $\hat{y} = \hat{f}(x)$?



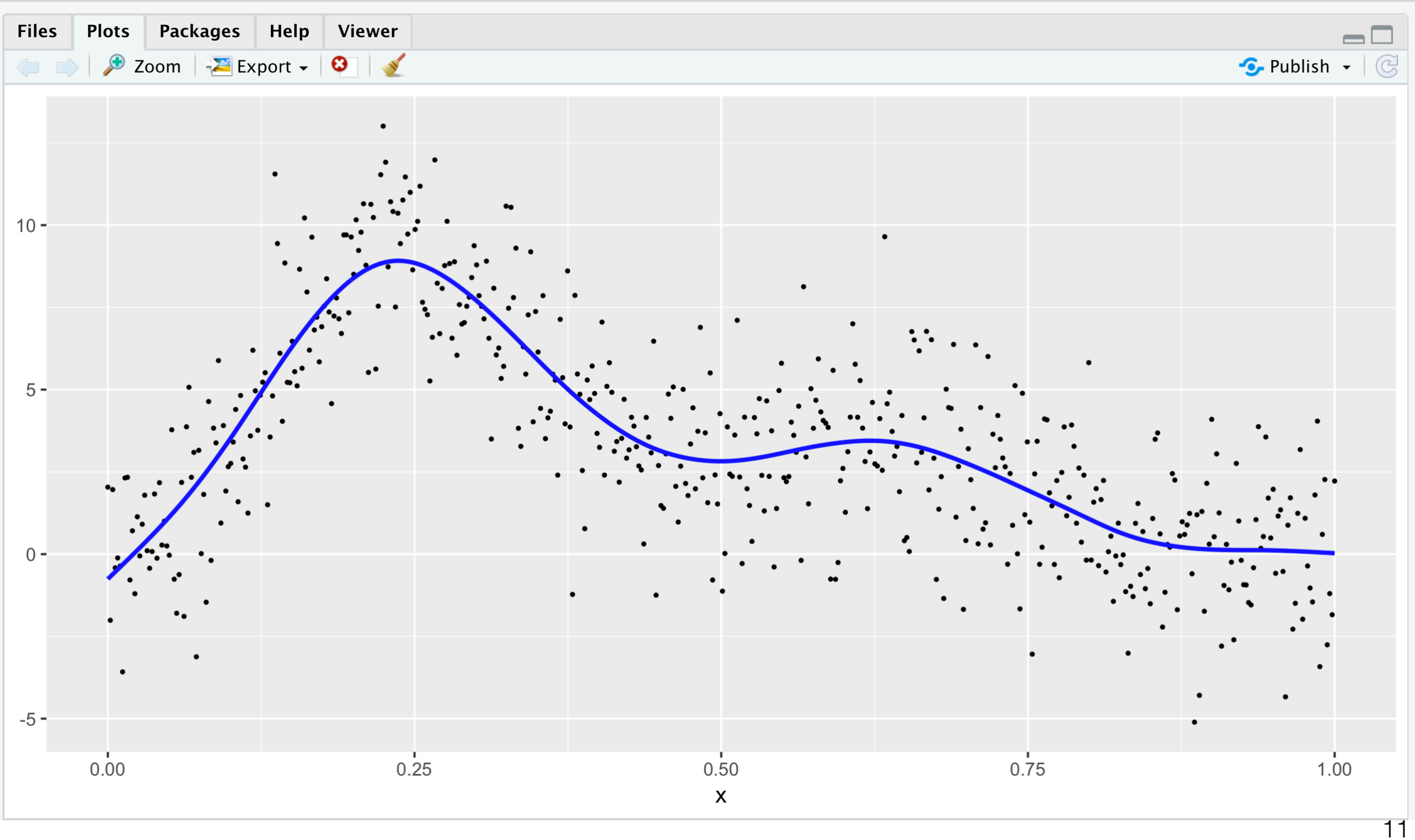
How about this $\hat{f}(x)$?



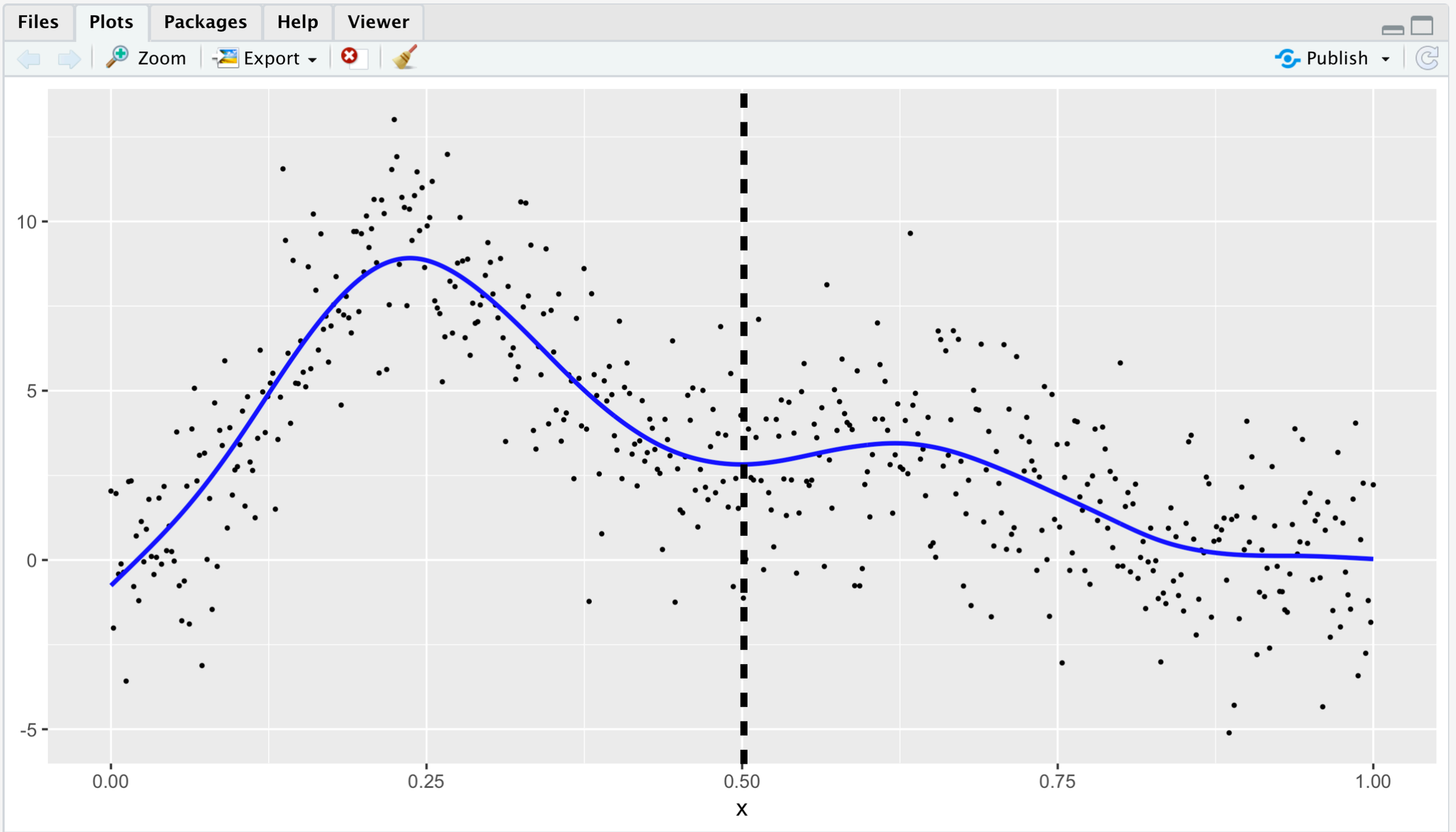
How about this $\hat{f}(x)$?



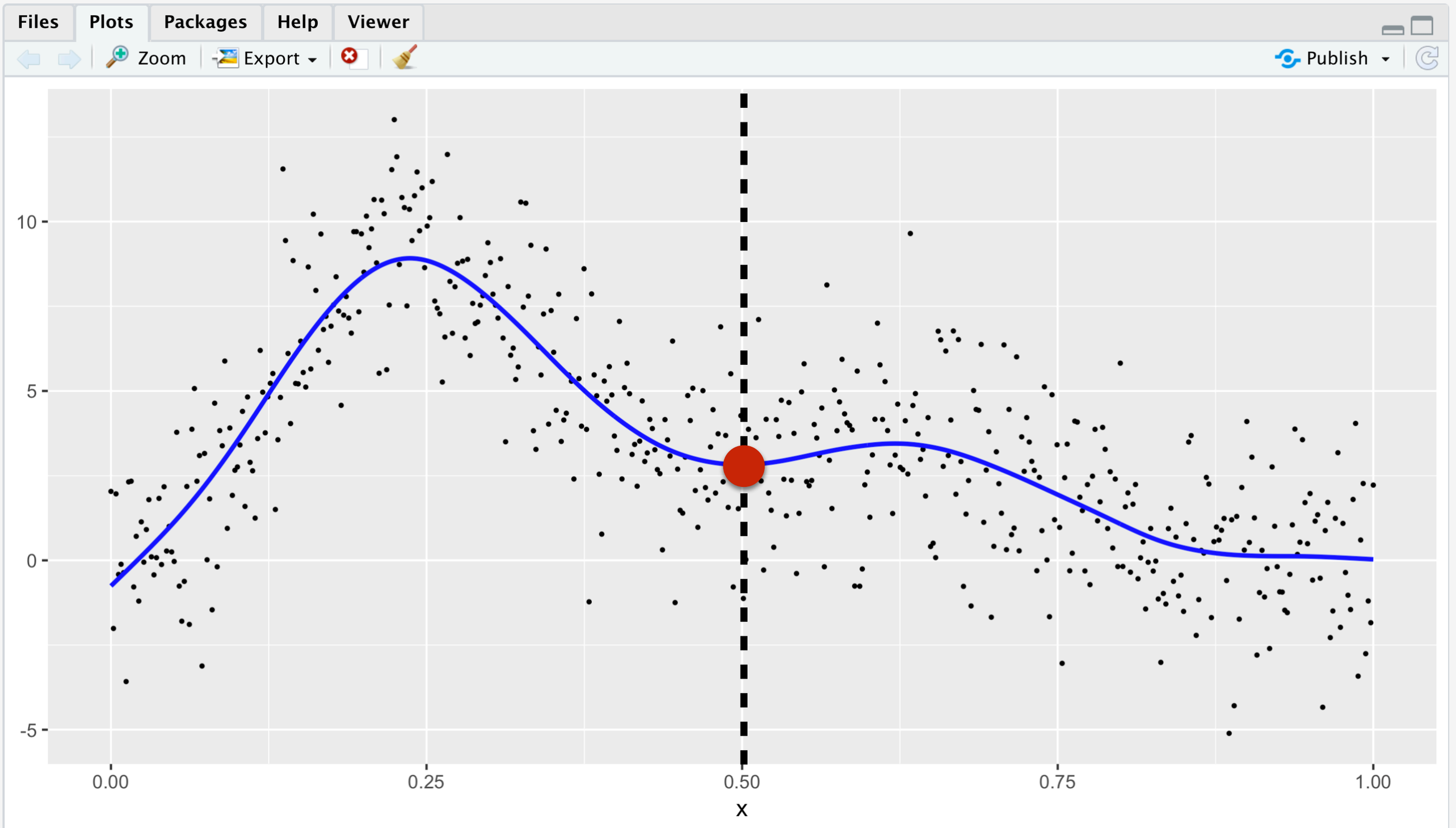
What does this $\hat{f}(x)$ predict for $x = 0.5$?



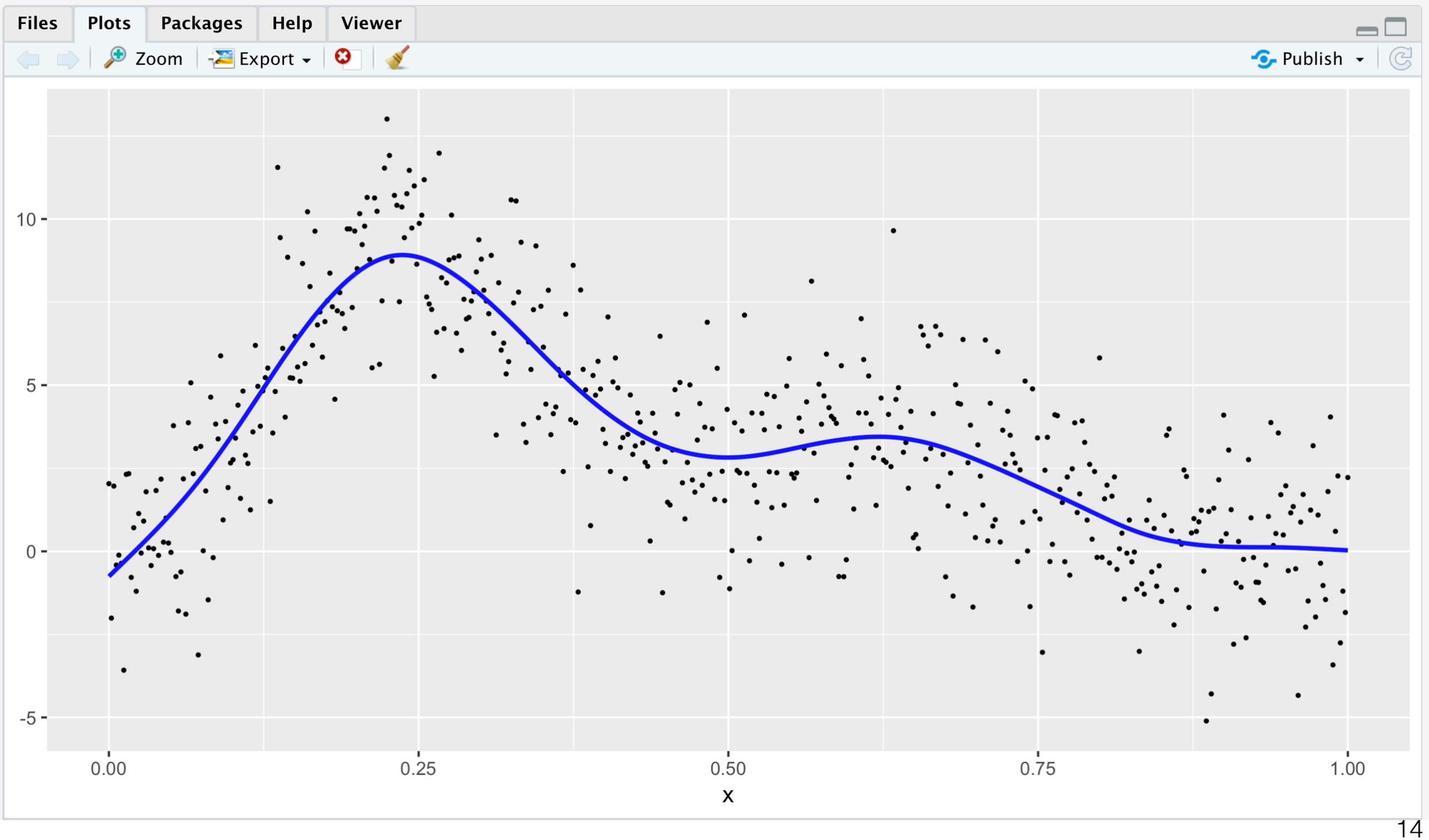
What does this $\hat{f}(x)$ predict for $x = 0.5$?



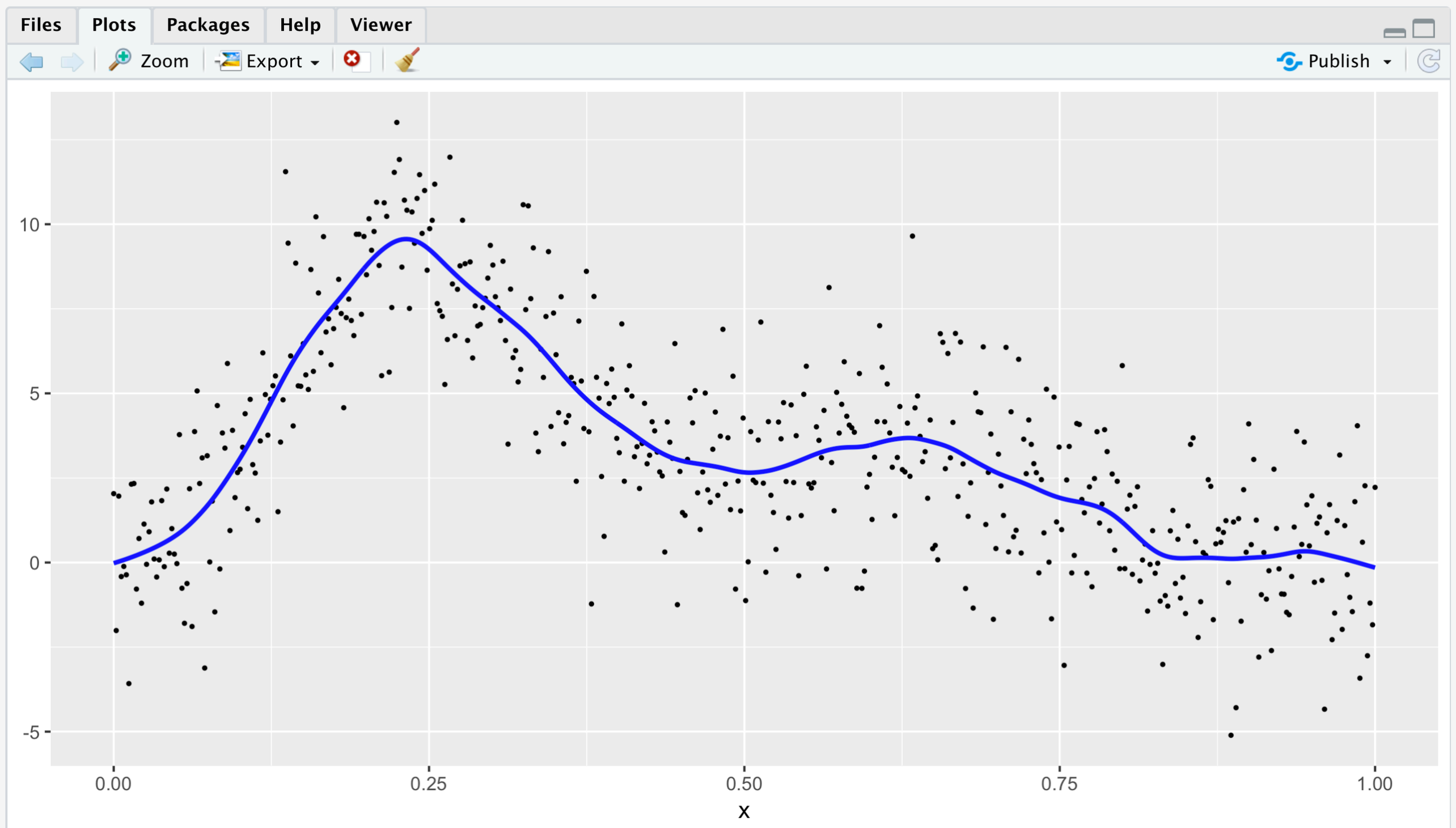
What does this $\hat{f}(x)$ predict for $x = 0.5$?



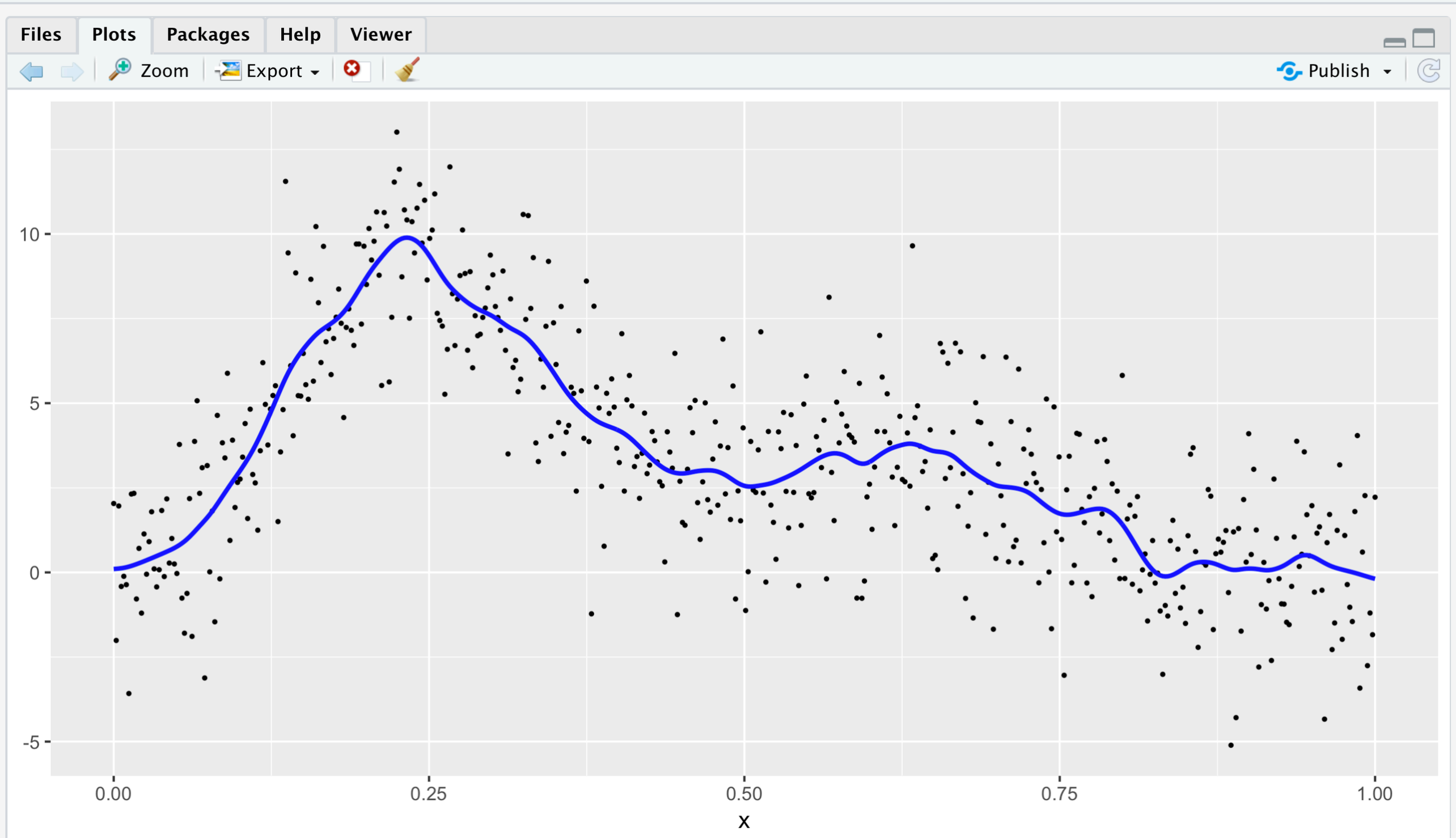
Ok, great. But instead of this $\hat{f}(x)$ …



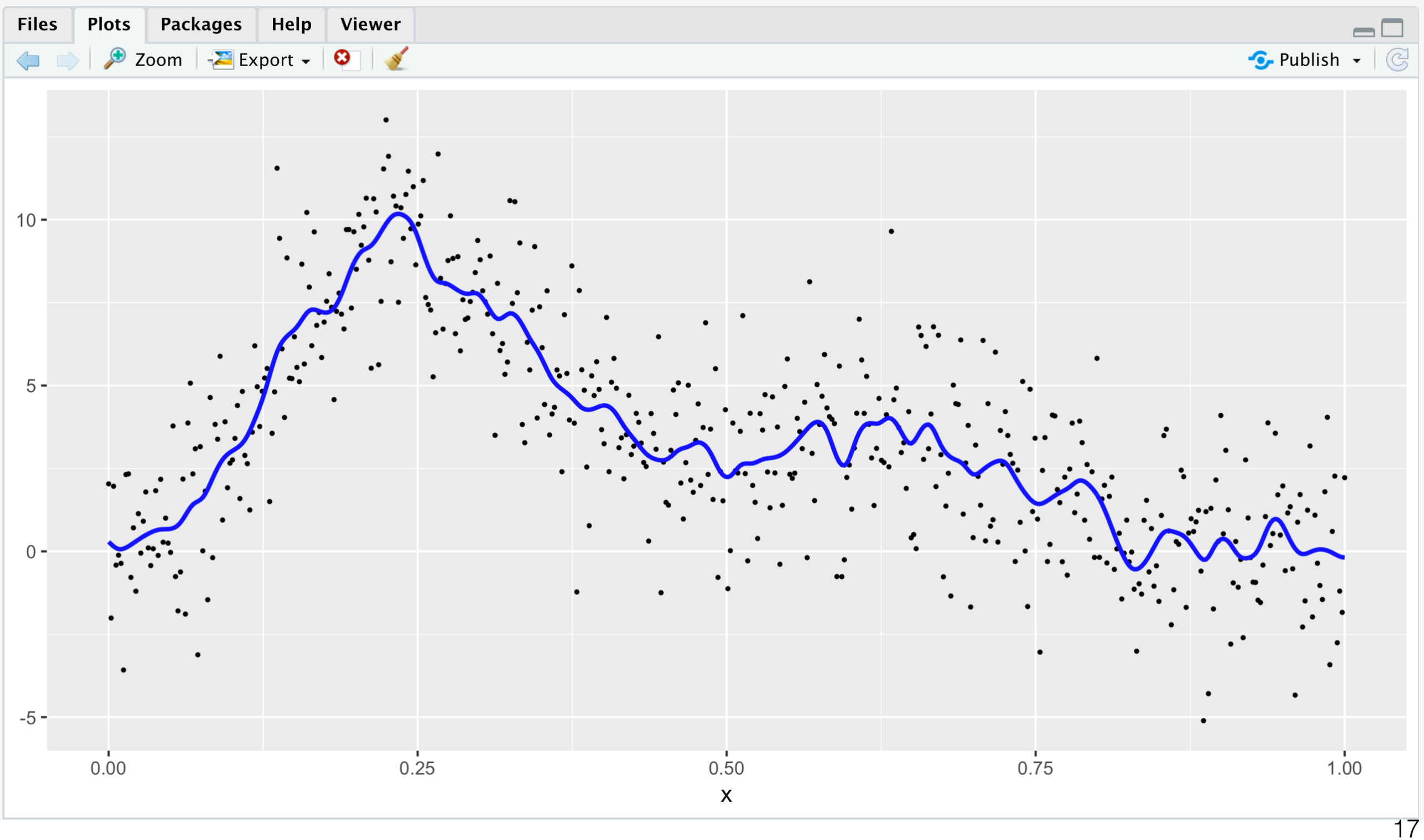
How about this $\hat{f}(x)$?



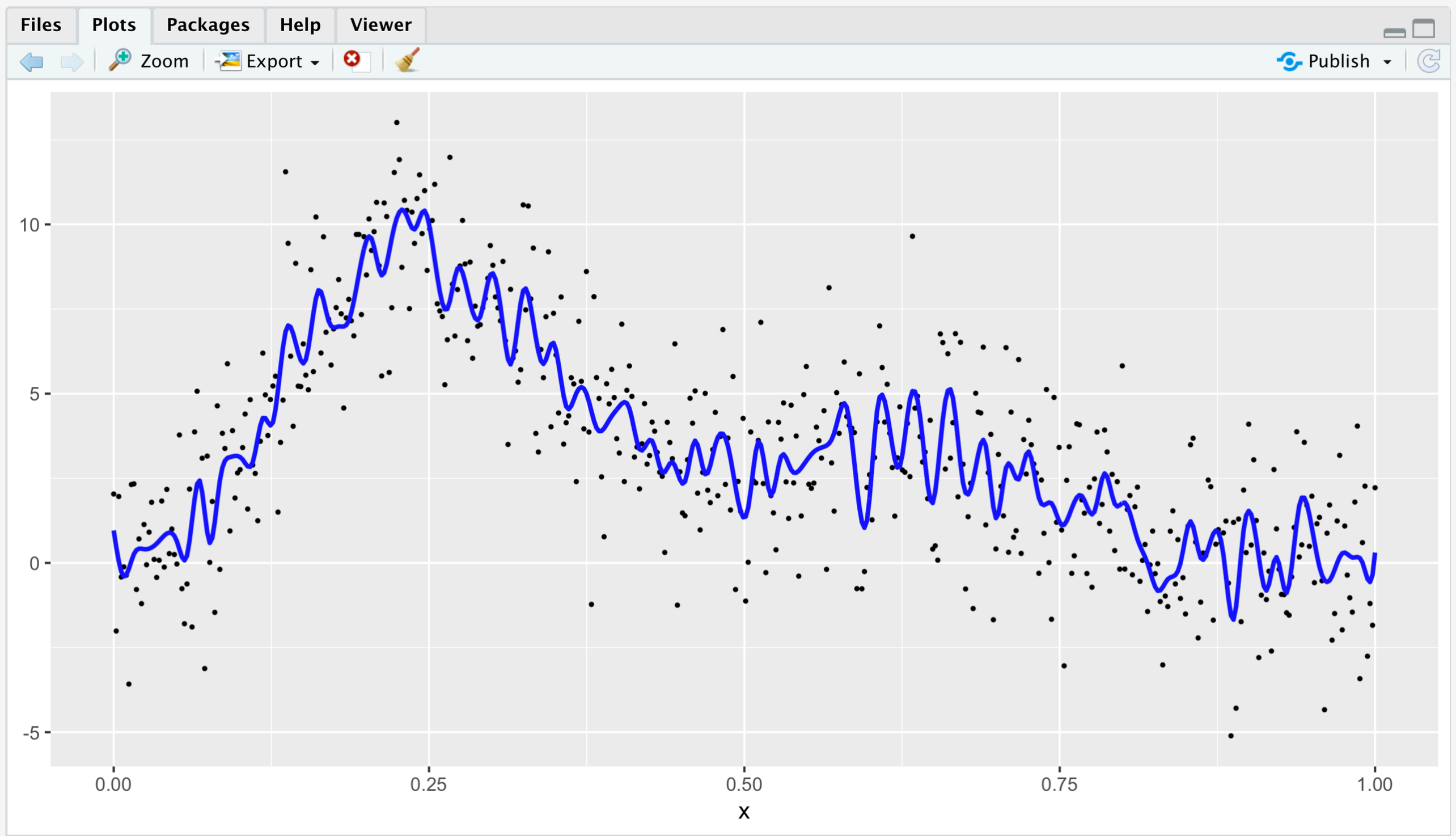
How about this $\hat{f}(x)$?



How about this $\hat{f}(x)$?



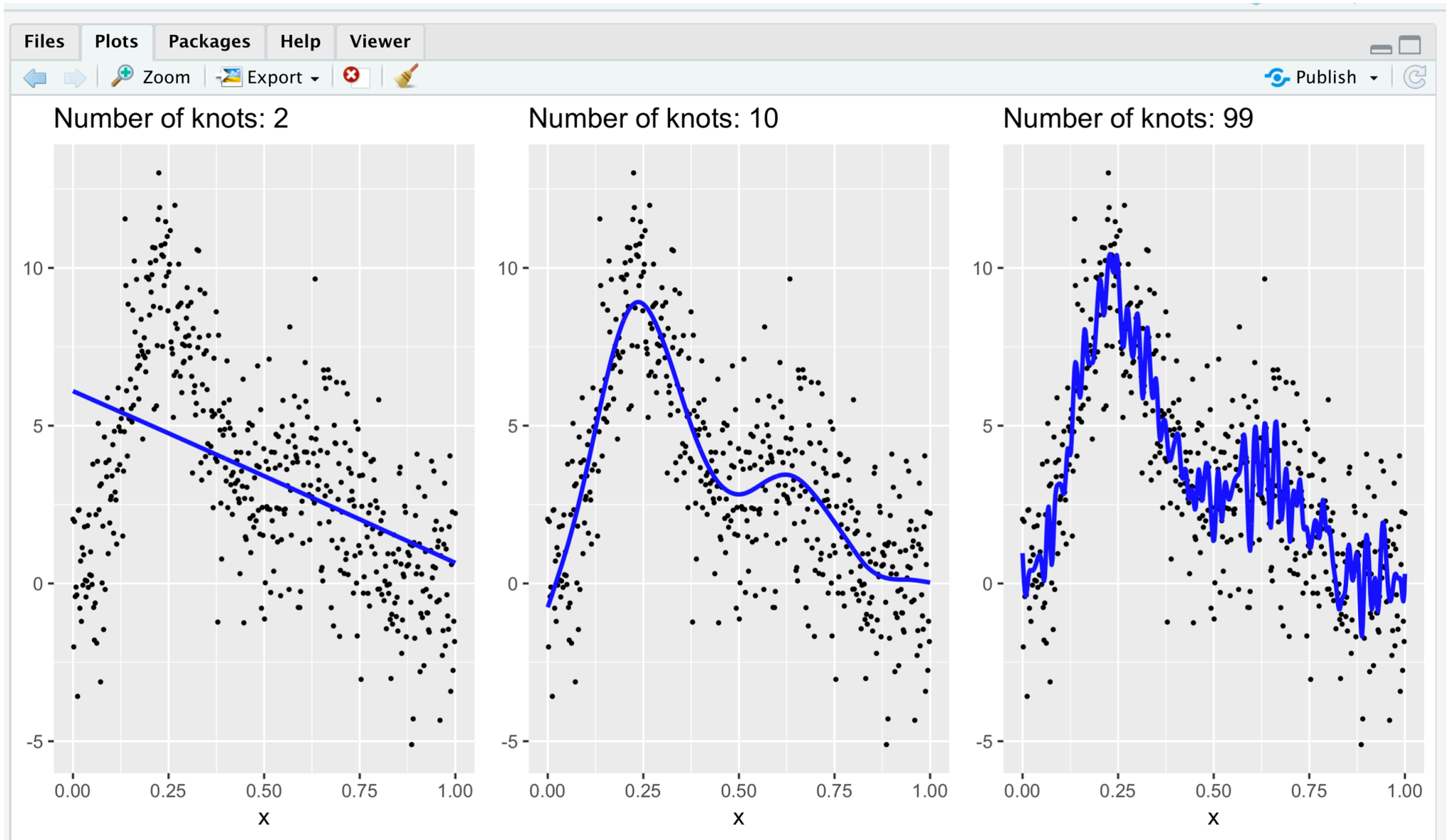
How about this $\hat{f}(x)$?



Model Fitting Method: (Cubic) Splines

- Splines fit the blue curve $\hat{f}(x)$ that **minimizes** the (squared) vertical distances between all:
  - predicted points $\hat{y} = \hat{f}(x)$ and
  - observed points y
- Amount of “wiggle” is the **complexity of the model**
- Occam’s Razor

Three Different $\hat{f}(x)$



Underfit!

"Just right!"

Overfit!

What is Machine Learning?

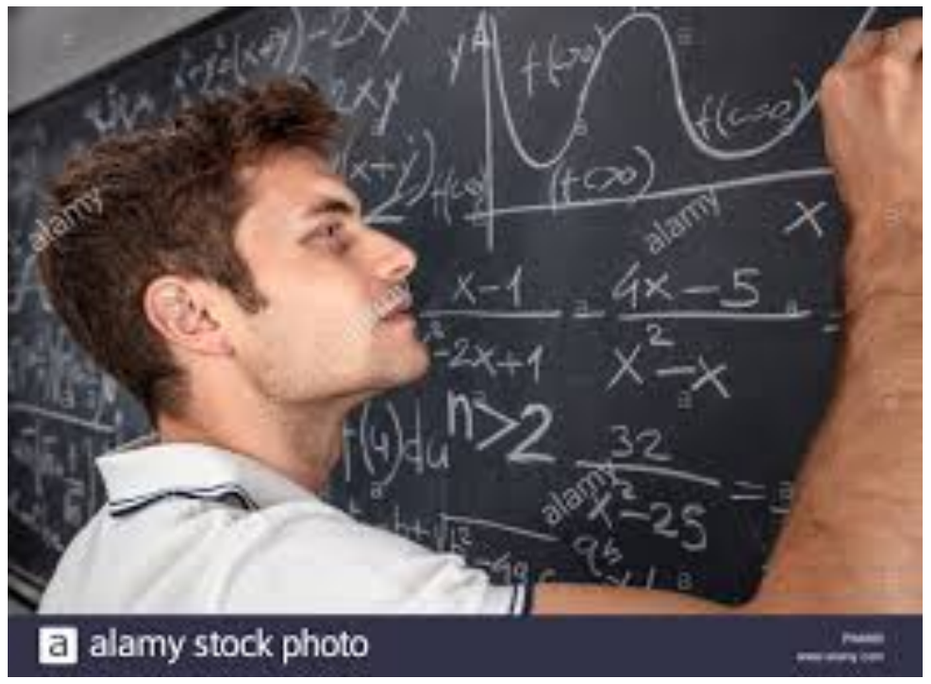
Machine Learning

- **Goal of Modeling:** Fit models $\hat{f}(x)$ that best approximate the true (unknown) model $f(x)$
- **Goal of Machine Learning:** Fit models that best "predict" the outcome variable

Model Assessment Metric

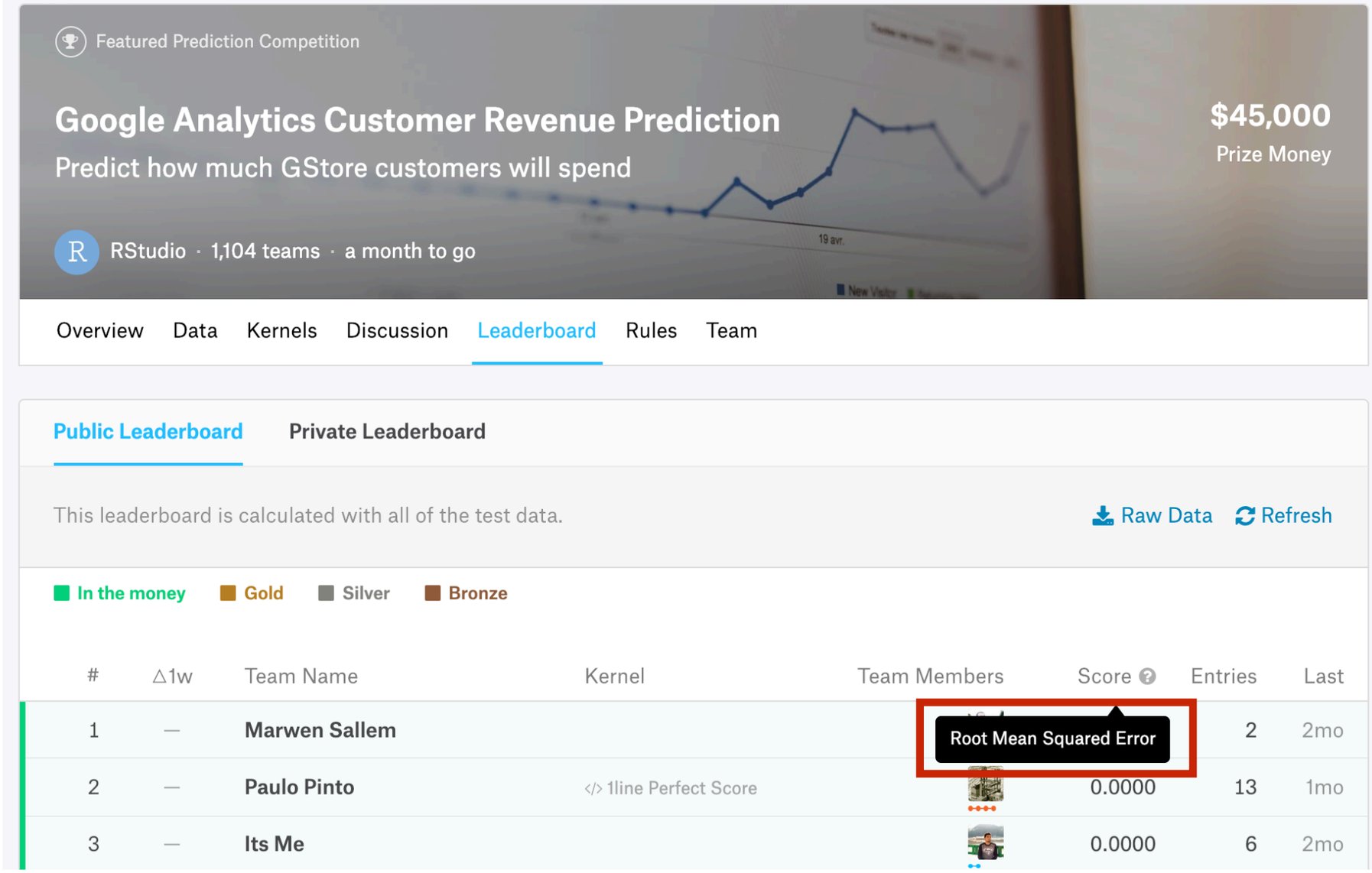
- Question: “How good is our model?”
- Answer: Metrics like the **Mean Square(d) Error**

Back to the blackboard for Chalk Talk #2…

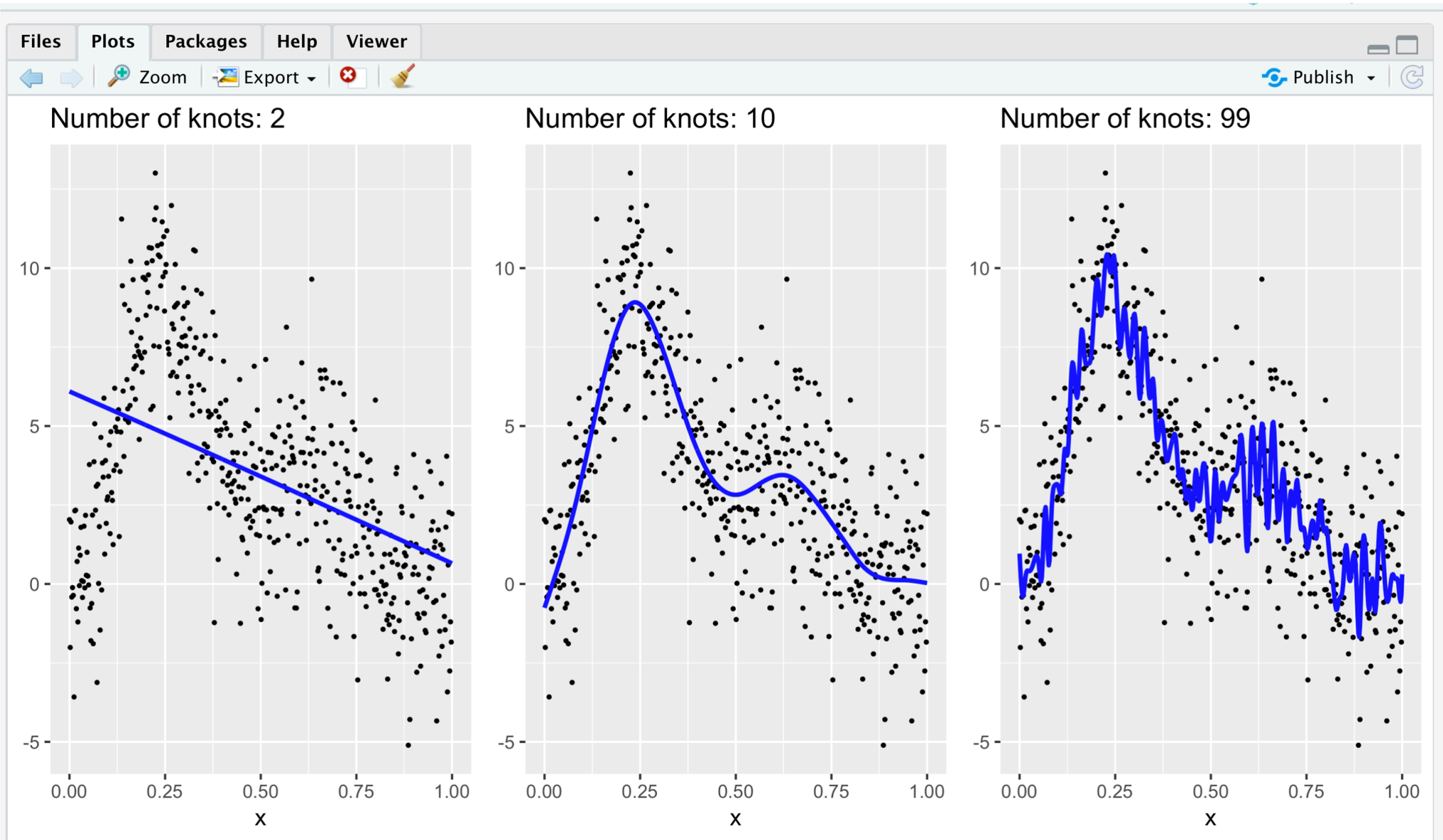


Mean Squared Error

On Machine Learning predictive modeling competition site [Kaggle]():



Issue of underfitting vs overfitting?



Underfit!

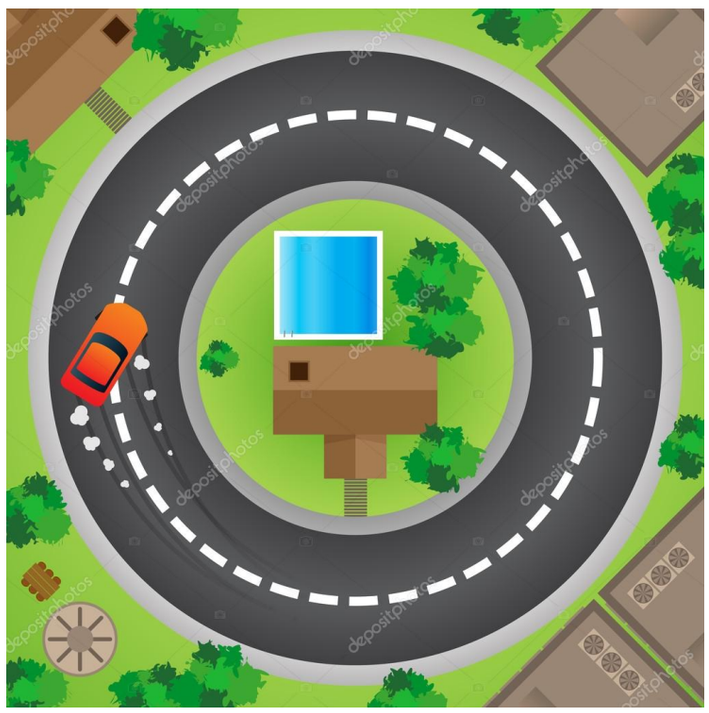
"Just right!"

Overfit!

Validation Set Approach

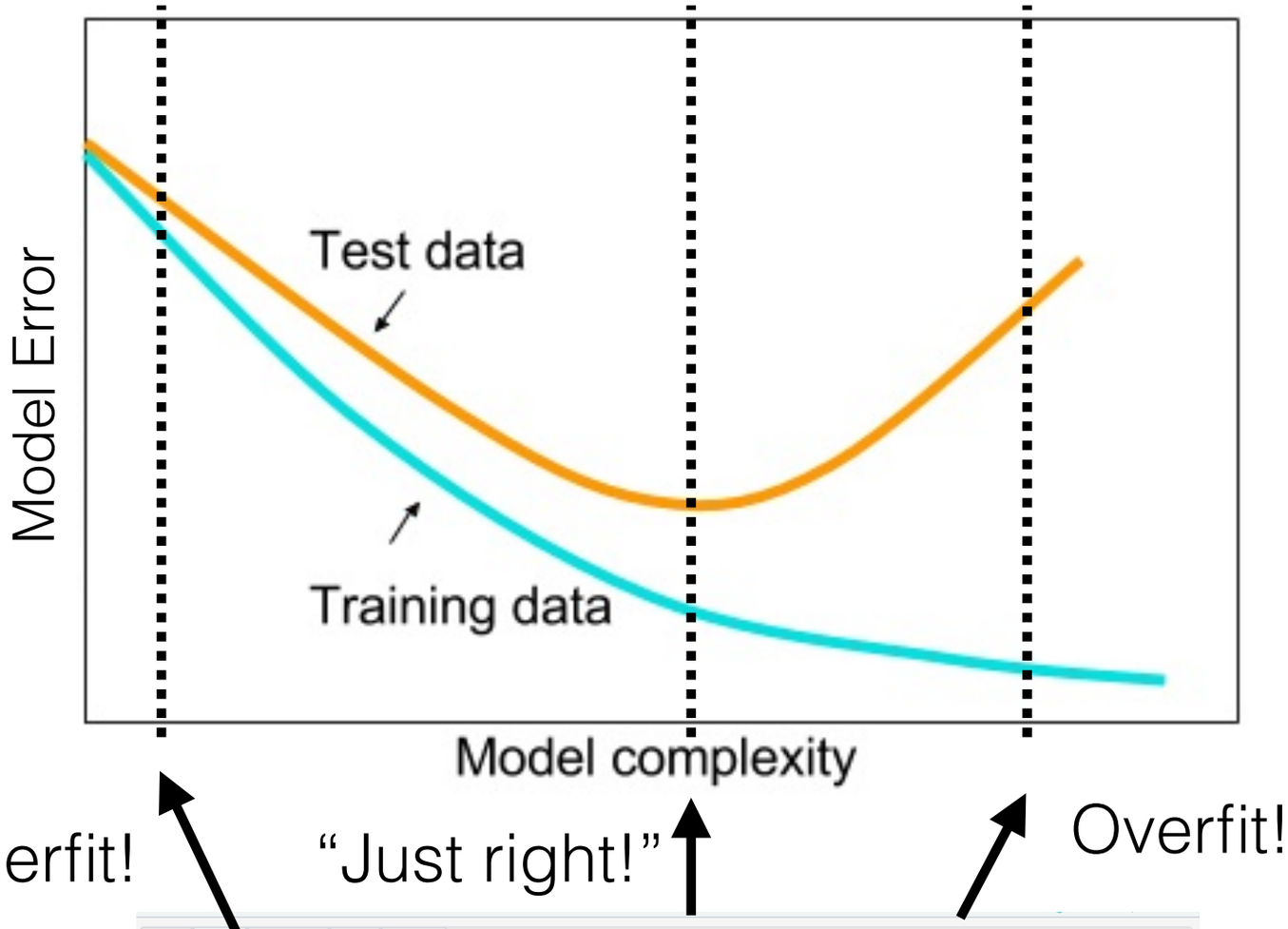
Fit/train model on ***training*** data

Assess model on independent ***test*** data

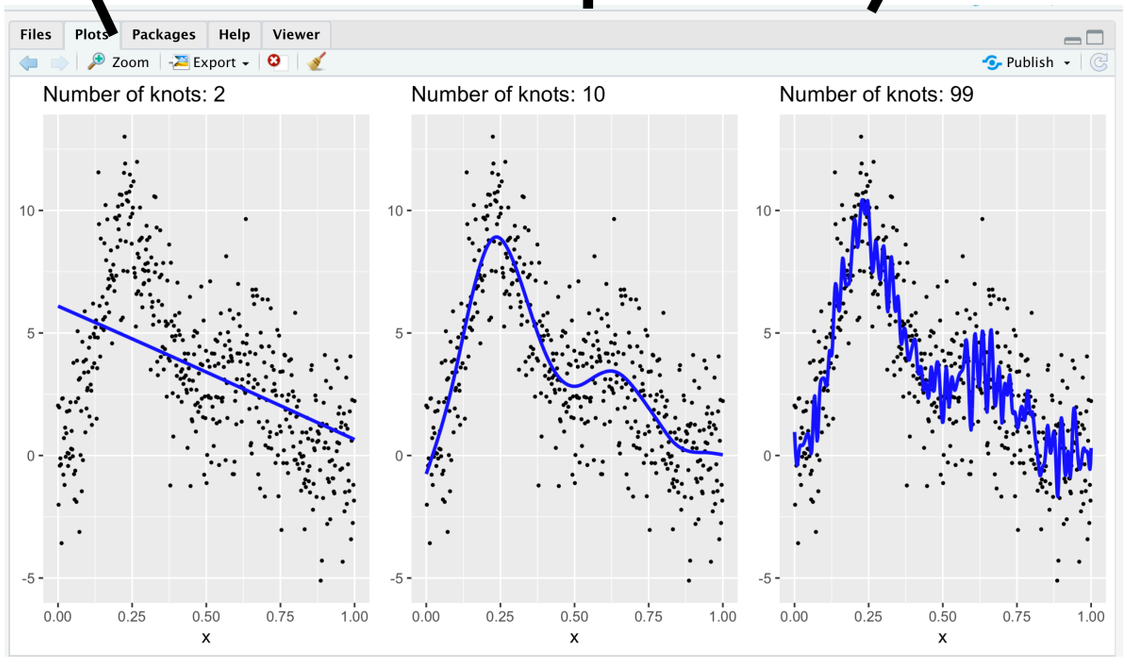


Typical Model Performance

1. Fit/train your model on training data
2. Assess your model error on either:
   a) the same training data or
   b) independent test data

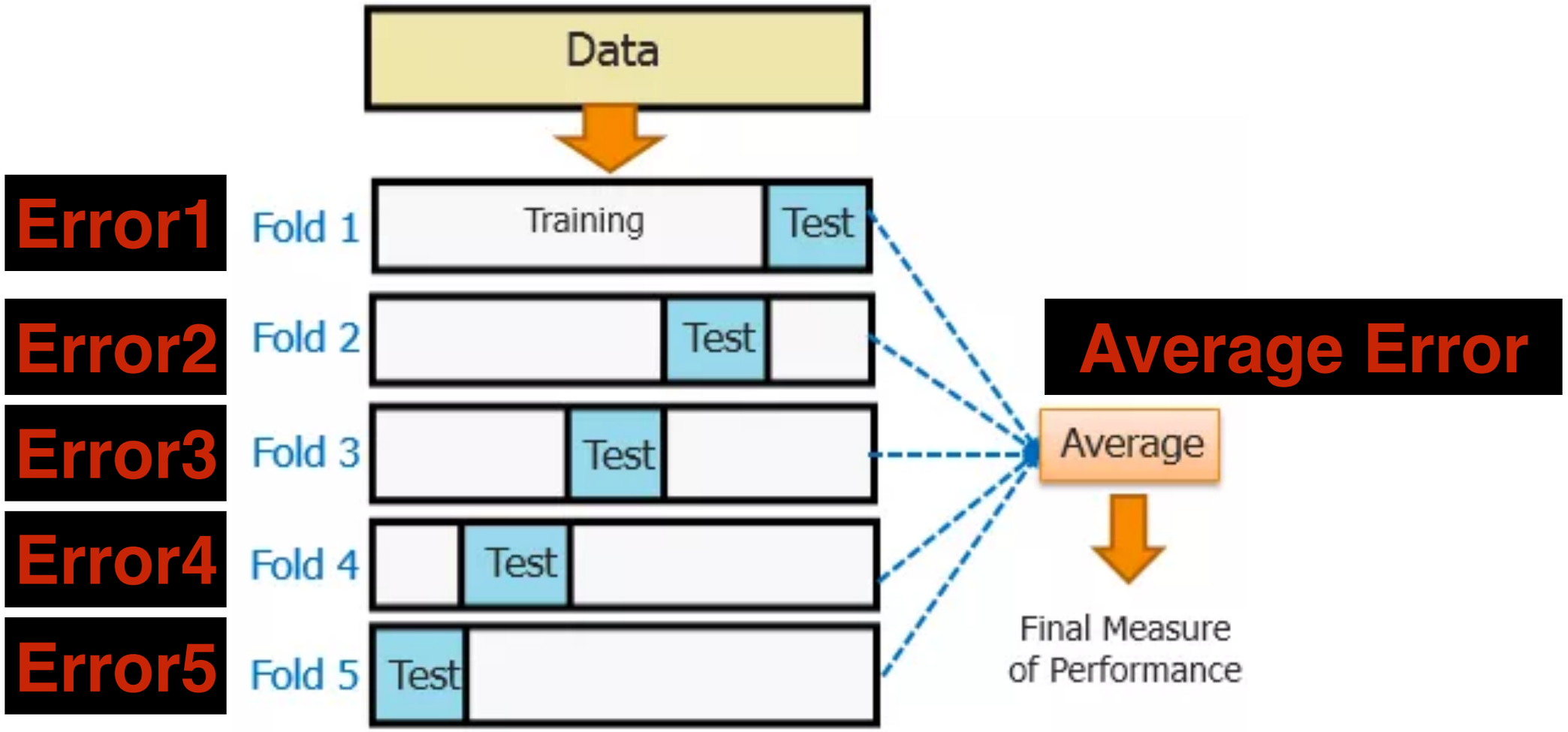


Recall for splines, the # of knots controls the **model complexity**



Generalization: 5-Fold Crossvalidation

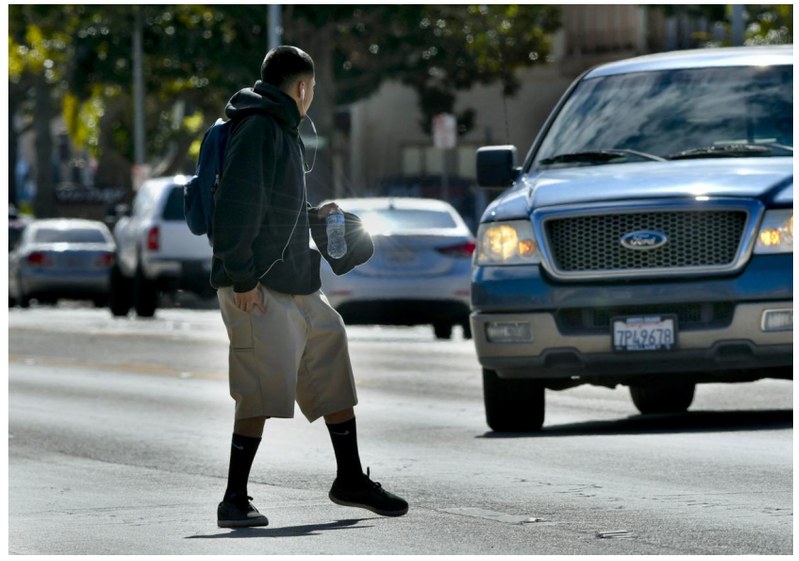
Repeat validation training/test set split 5 times:



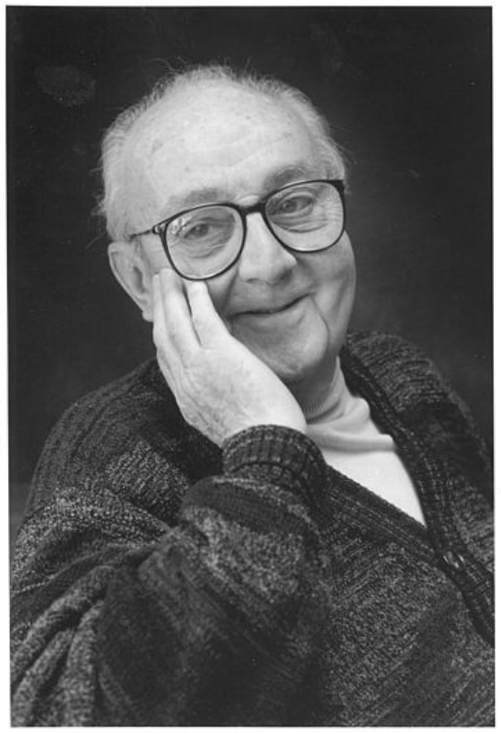
Concluding Thoughts

$\hat{f}(x)$ Model Fitting

Modeling is not as objective as you think

Scenario:	What they think is an “appropriate” model...	… might not be the same for these folks:
		 
		

To Close: Two Quotes on Modeling

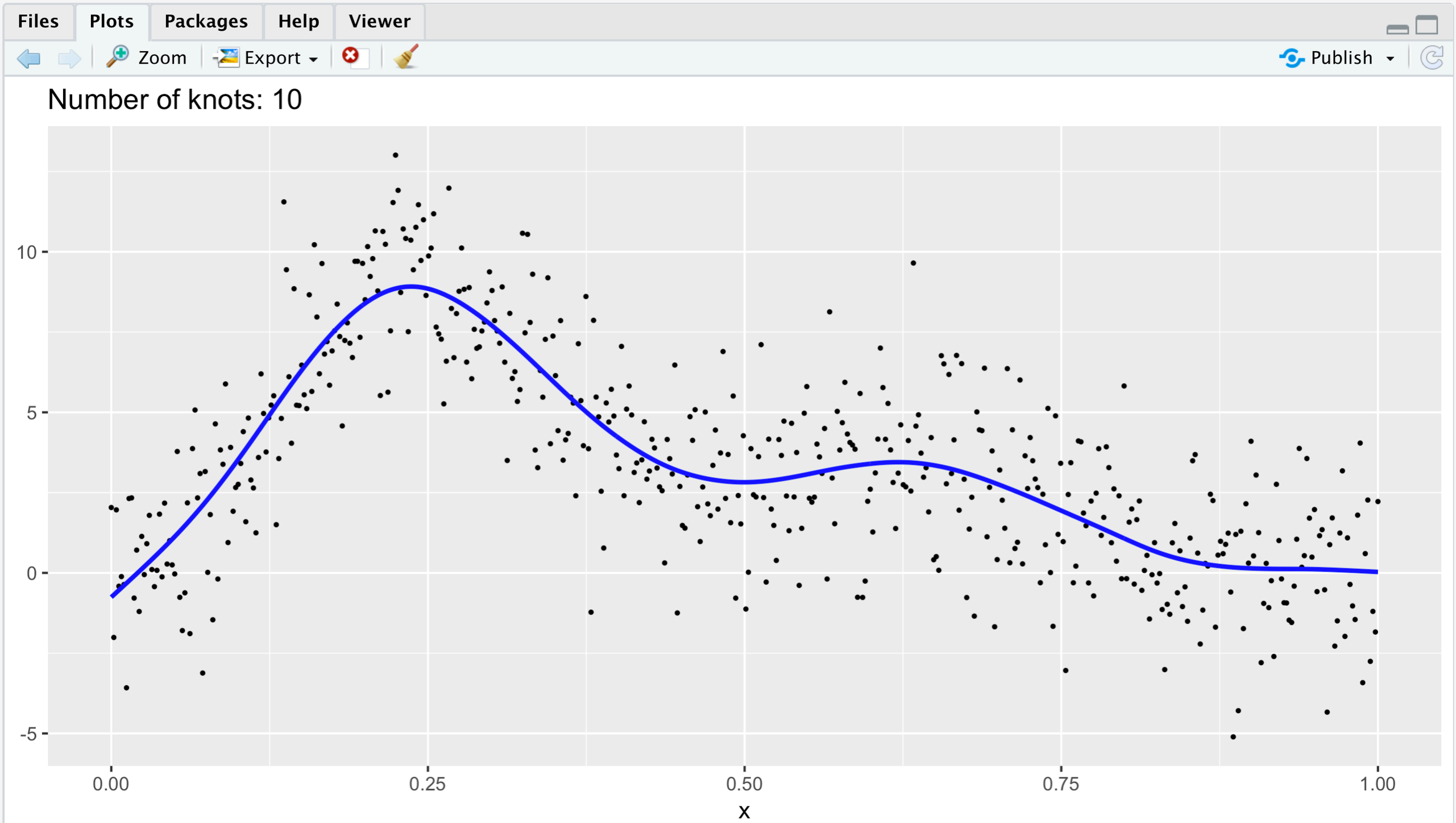


"All models are wrong,
but some are useful."
George Box

"WTF is up with your
$\hat{f}(x)$?" @rudeboybert

Lastly: A “Wizard of Oz” Reveal…

Our approximated $\hat{f}(x)$ …



… was pretty close to the *true* model:

$$f(x) = 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10}$$