

Machine Learning & Advanced R



Albert Y. Kim

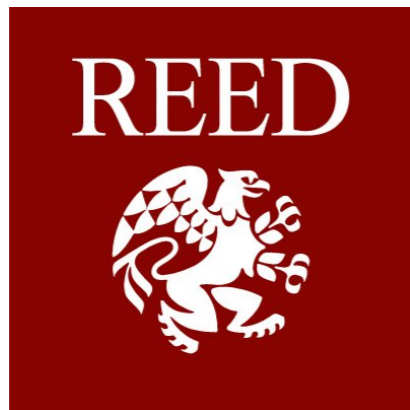
Assistant Professor

[Statistical & Data Sciences](#), Smith College

MassMutual Data Science Summer 2019 Bootcamp

Tuesday 2019/7/22

About me & my background



Middlebury



SMITH COLLEGE

Machine Learning



NETFLIX

Prediction!



NETFLIX



The General Modeling Problem

Machine Learning as Modeling

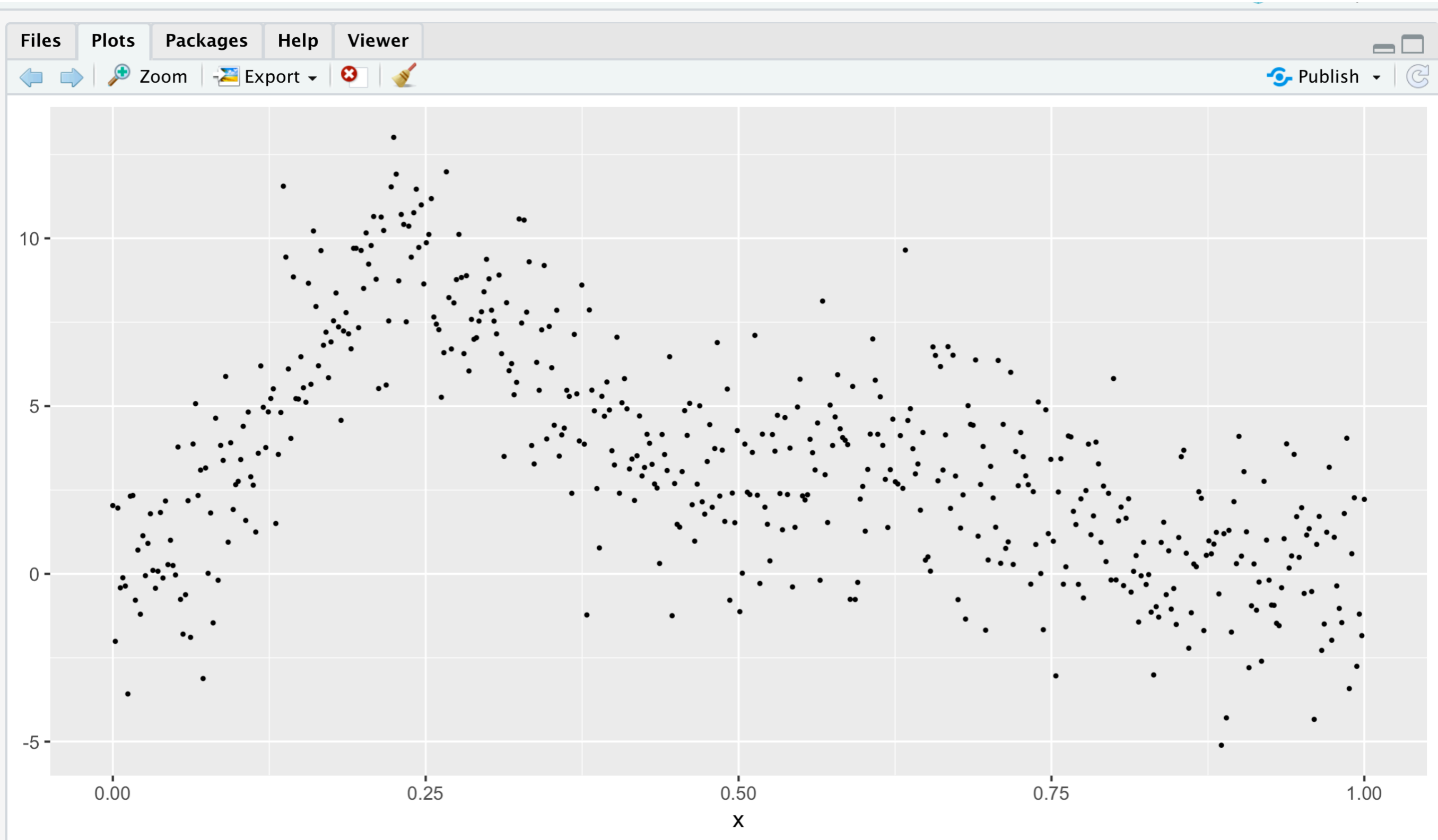
True (Unknown) Model: $y = f(\vec{x}) + \epsilon$

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

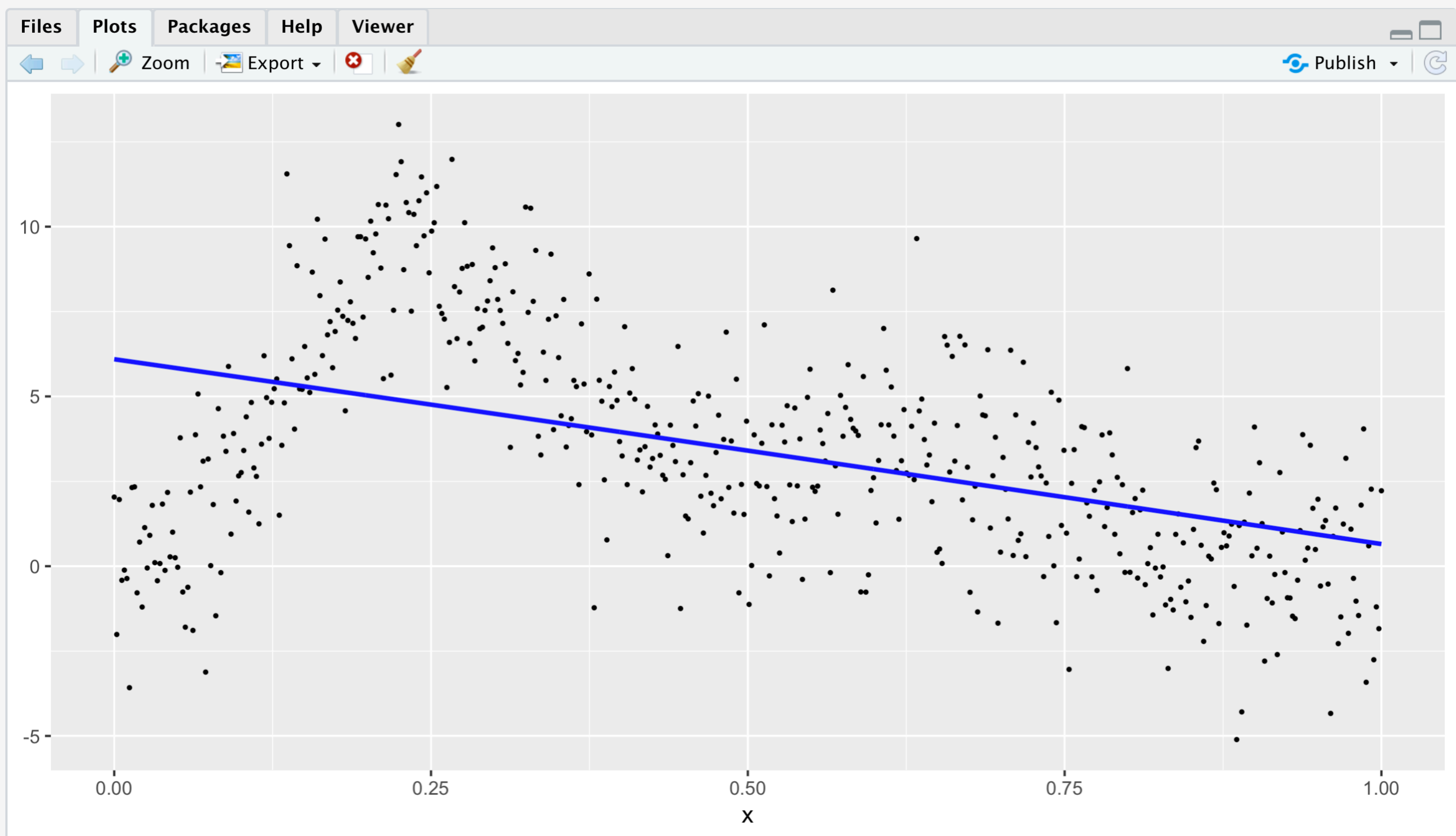
Now to the blackboard for
Chalk Talk #1...



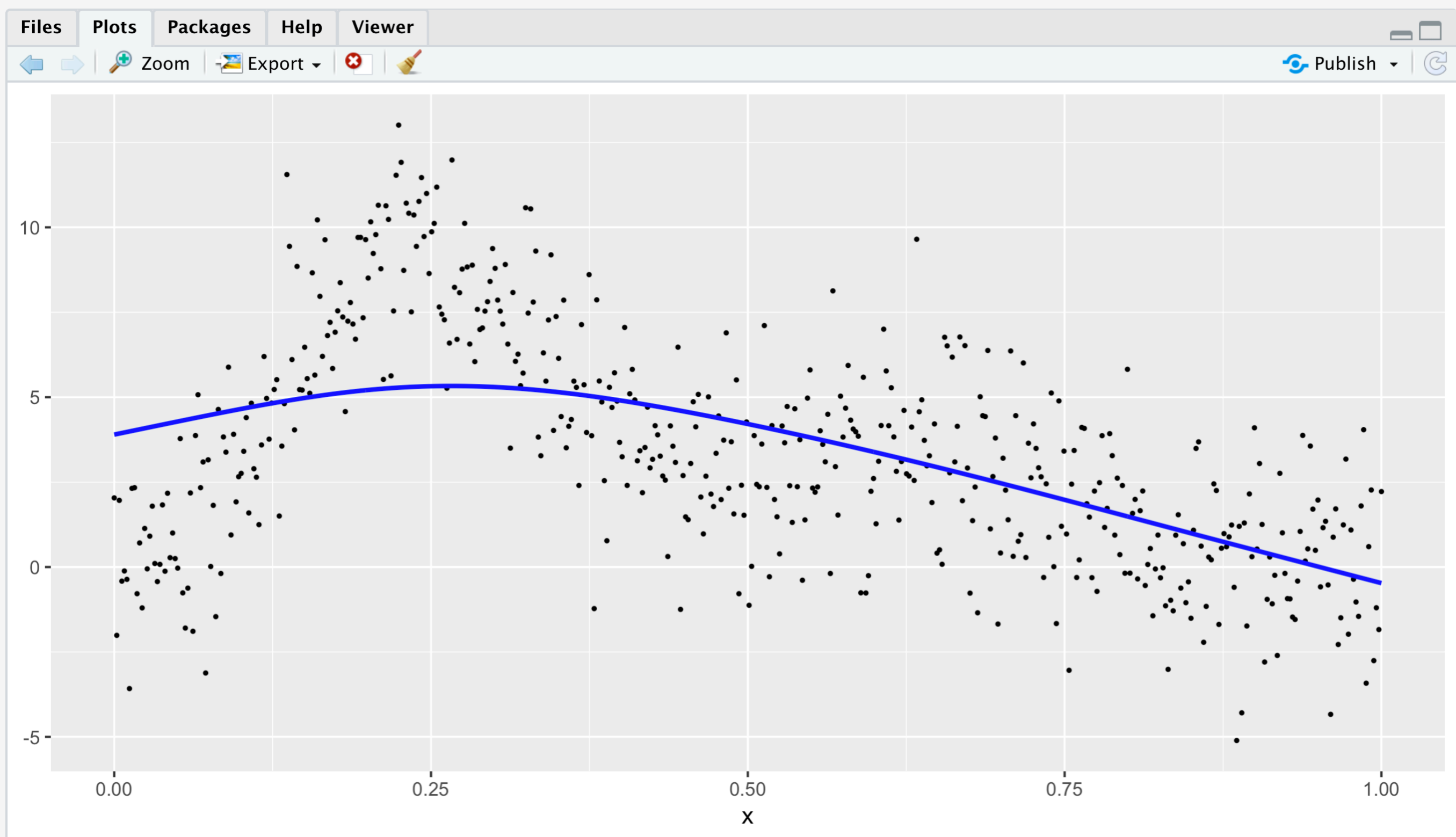
Given Data (x, y) from “unknown” $f(x)$



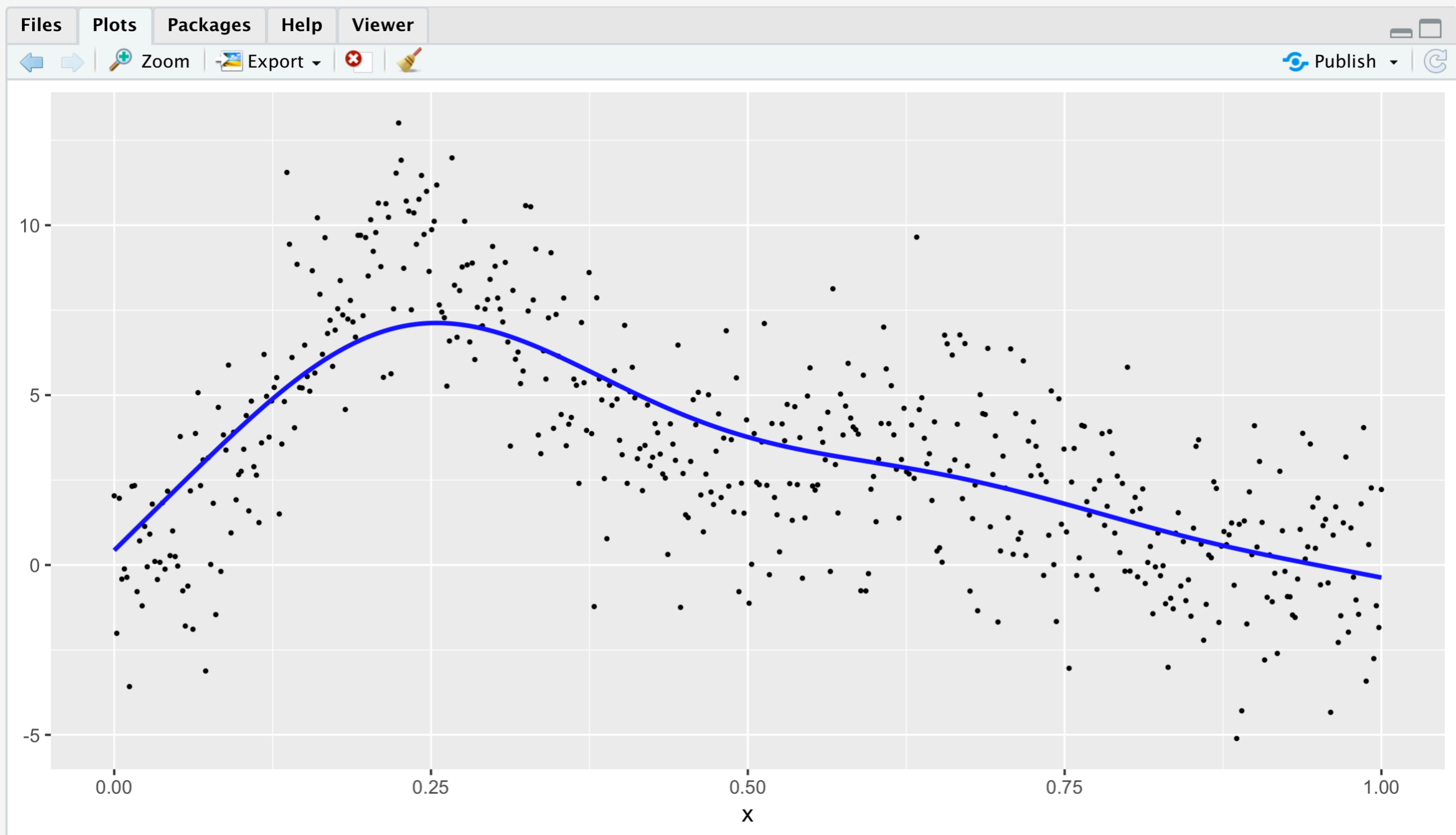
Approximate (i.e. “fit”) a Model $\hat{f}(x)$



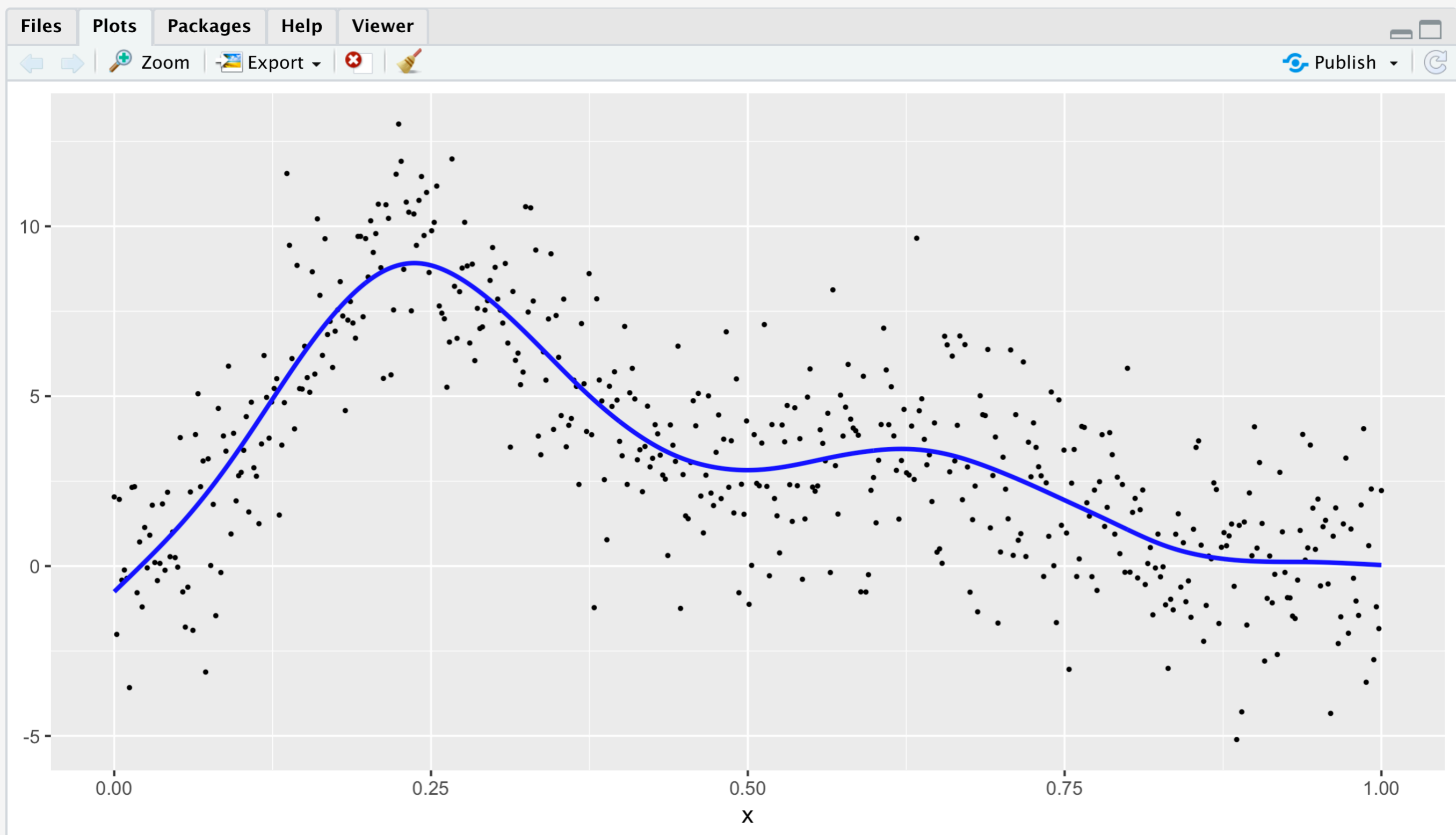
How about this $\hat{y} = \hat{f}(x)$?



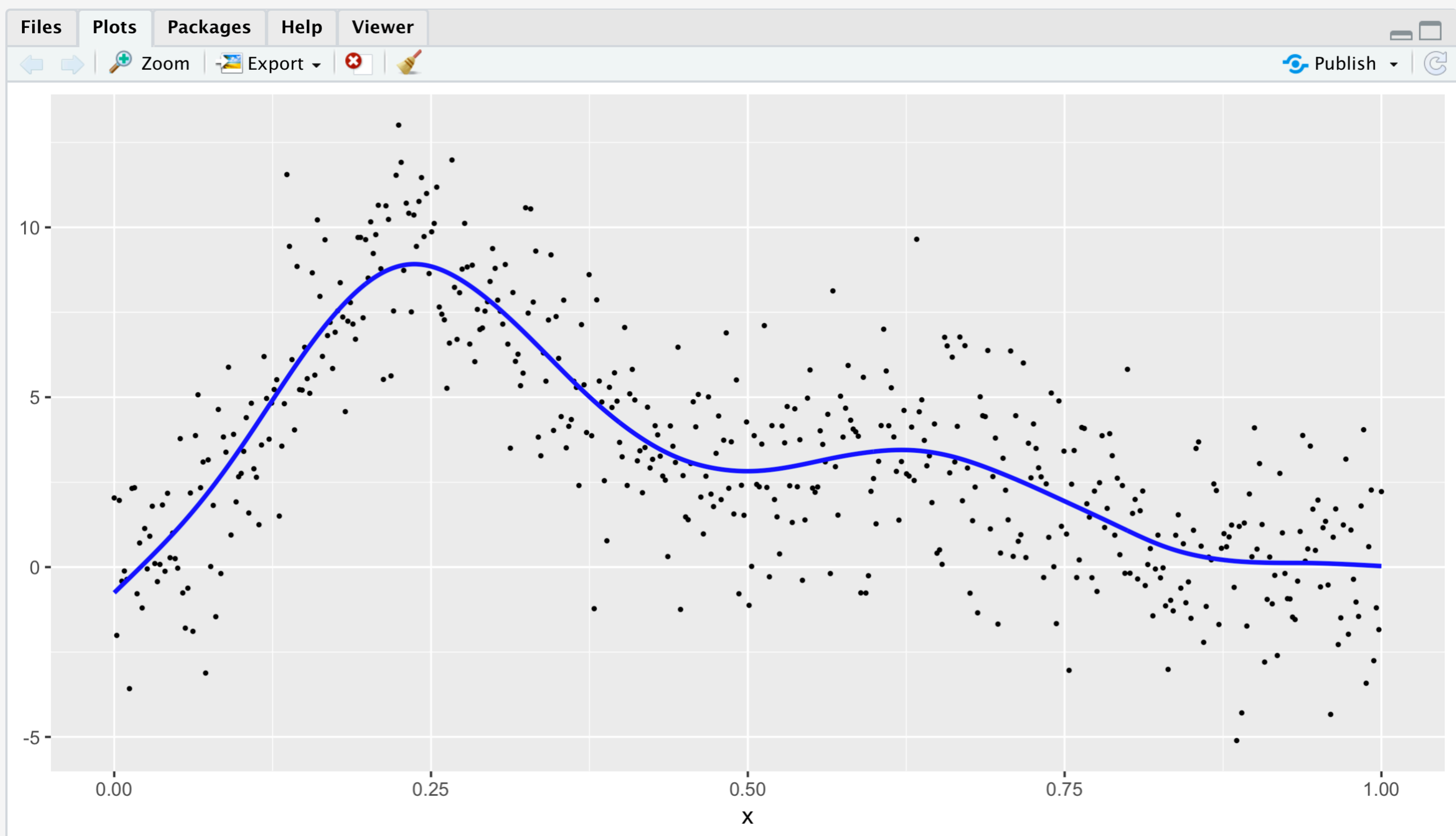
How about this $\hat{f}(x)$?



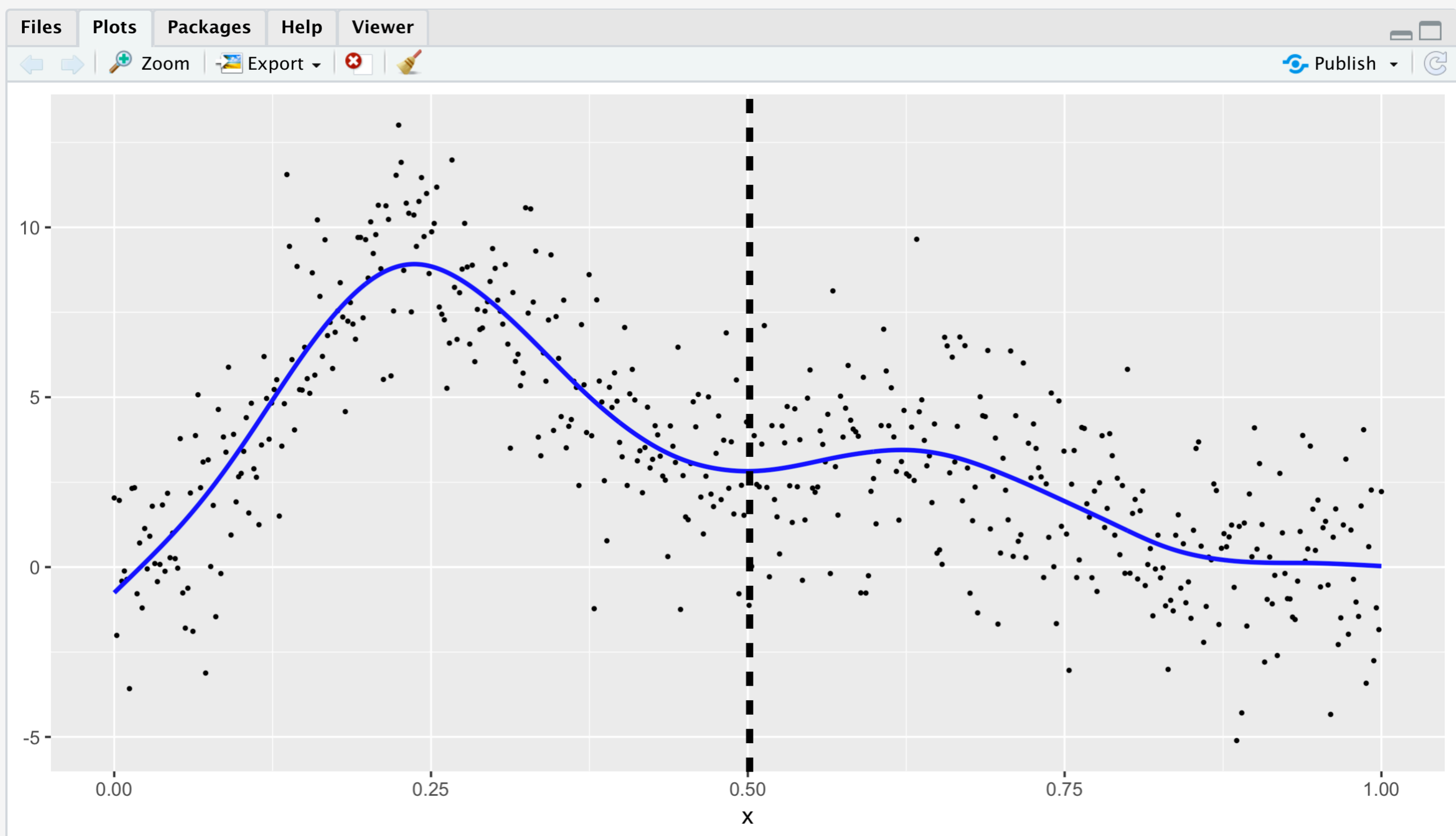
How about this $\hat{f}(x)$?



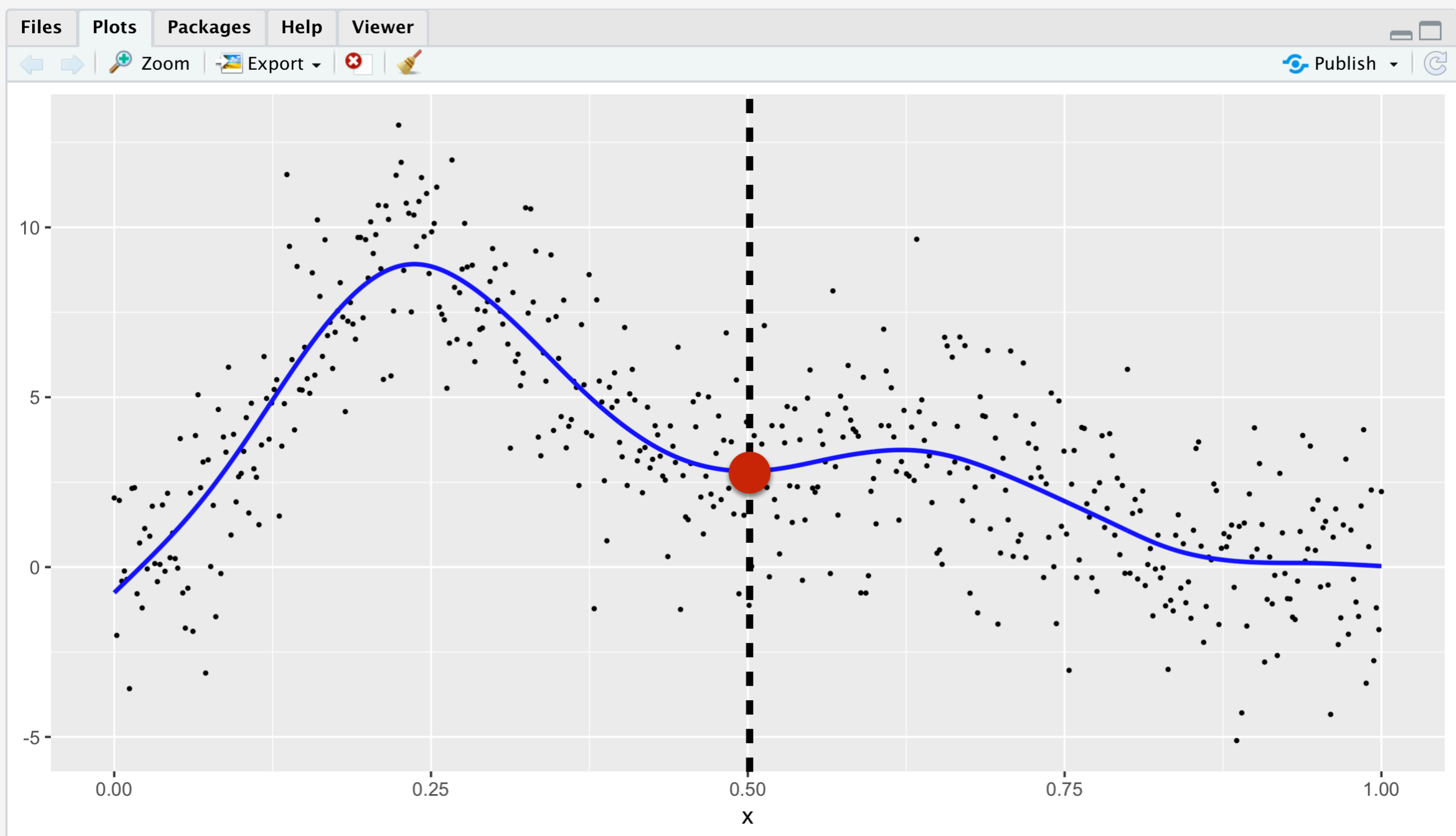
What does this $\hat{f}(x)$ predict for $x = 0.5$?



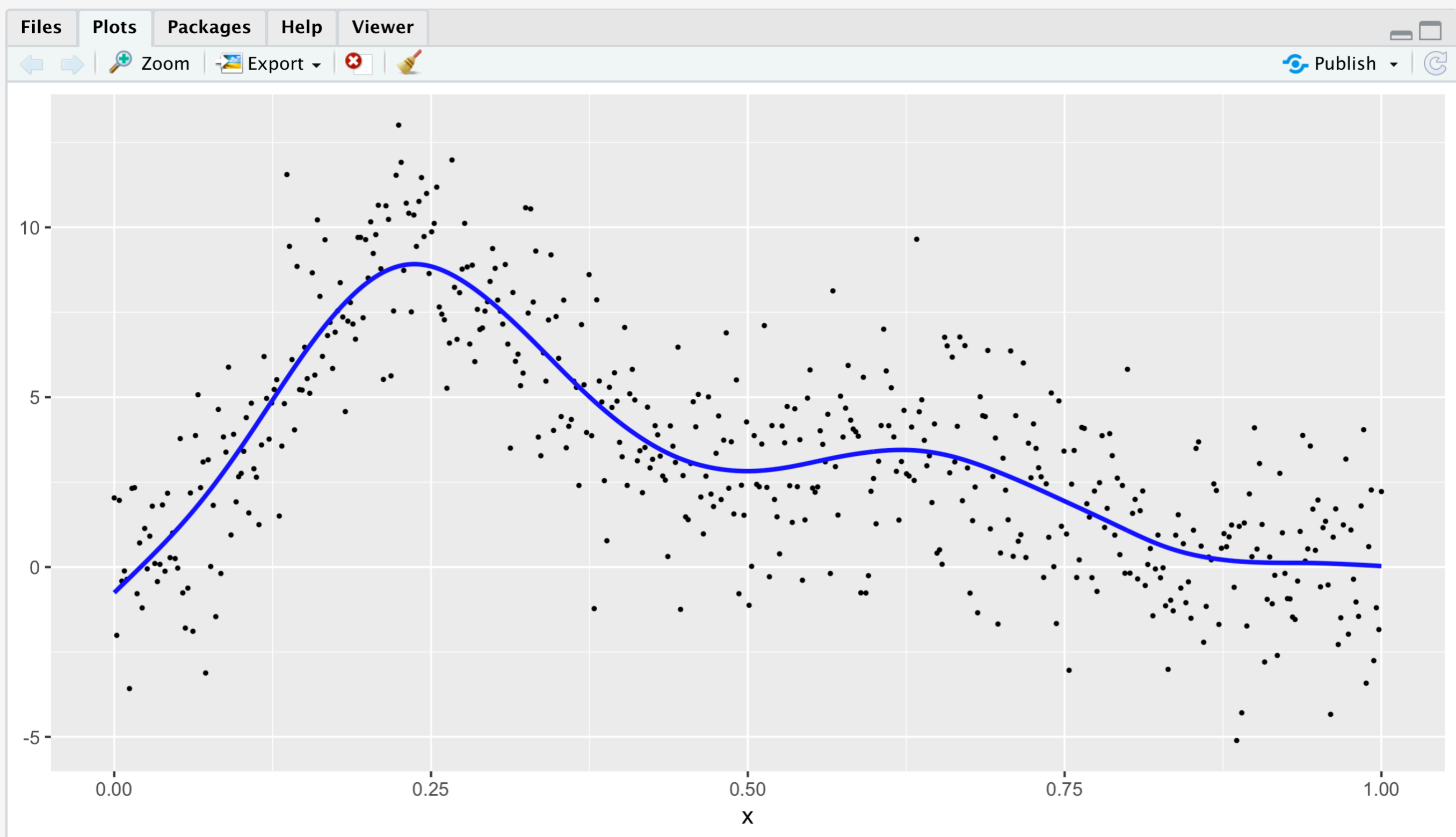
What does this $\hat{f}(x)$ predict for $x = 0.5$?



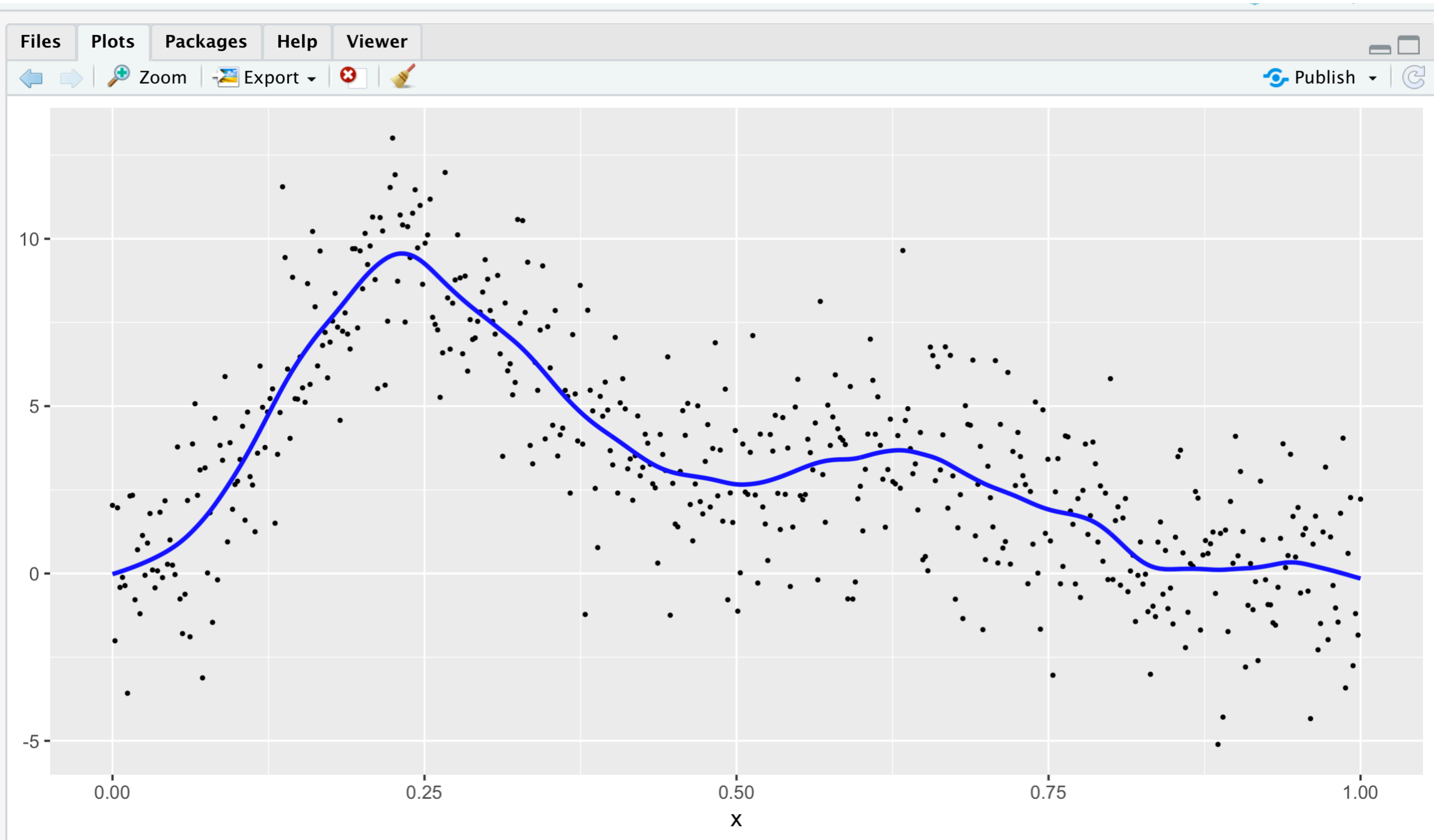
What does this $\hat{f}(x)$ predict for $x = 0.5$?



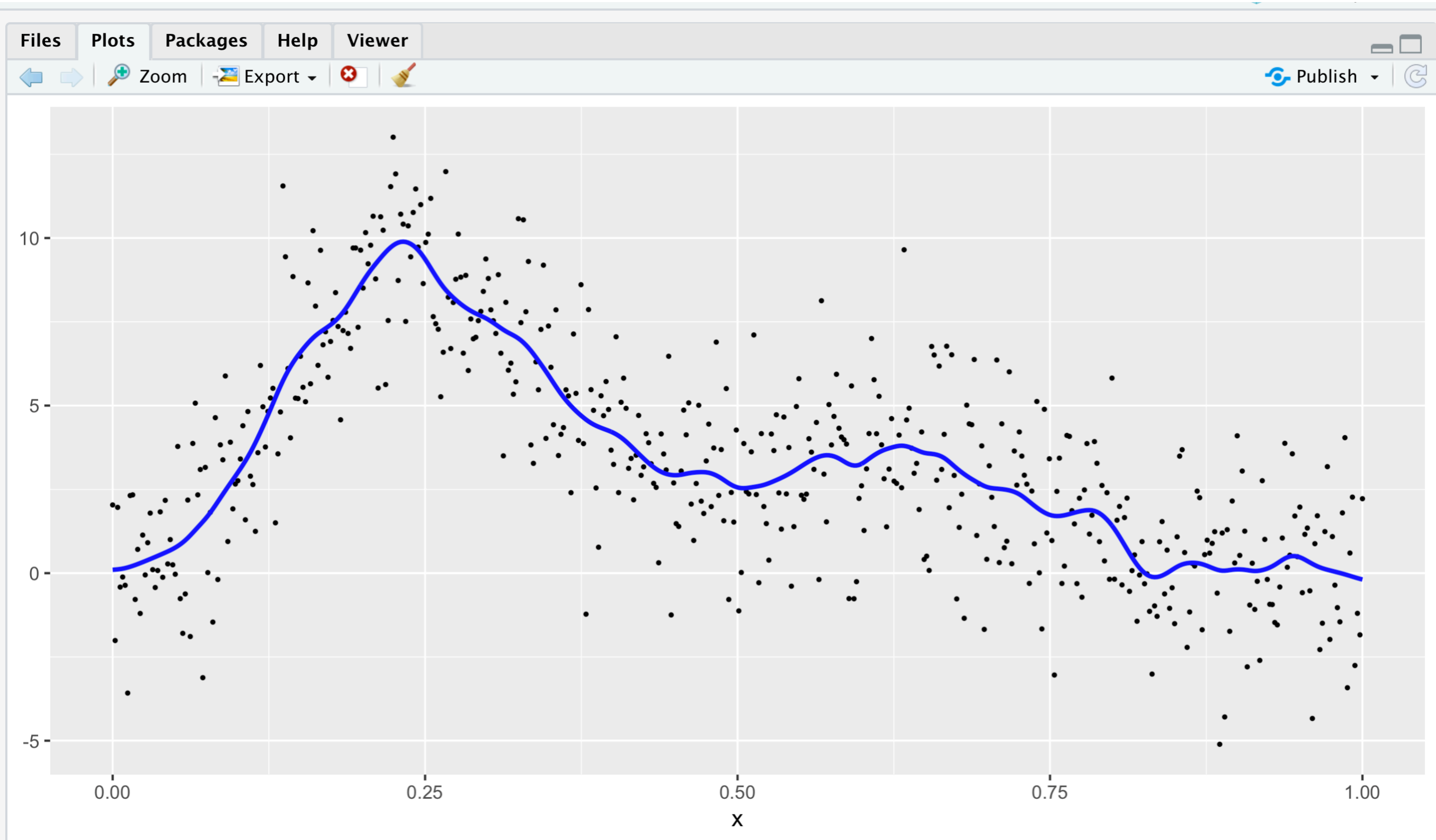
Ok, great. But instead of this $\hat{f}(x)$...



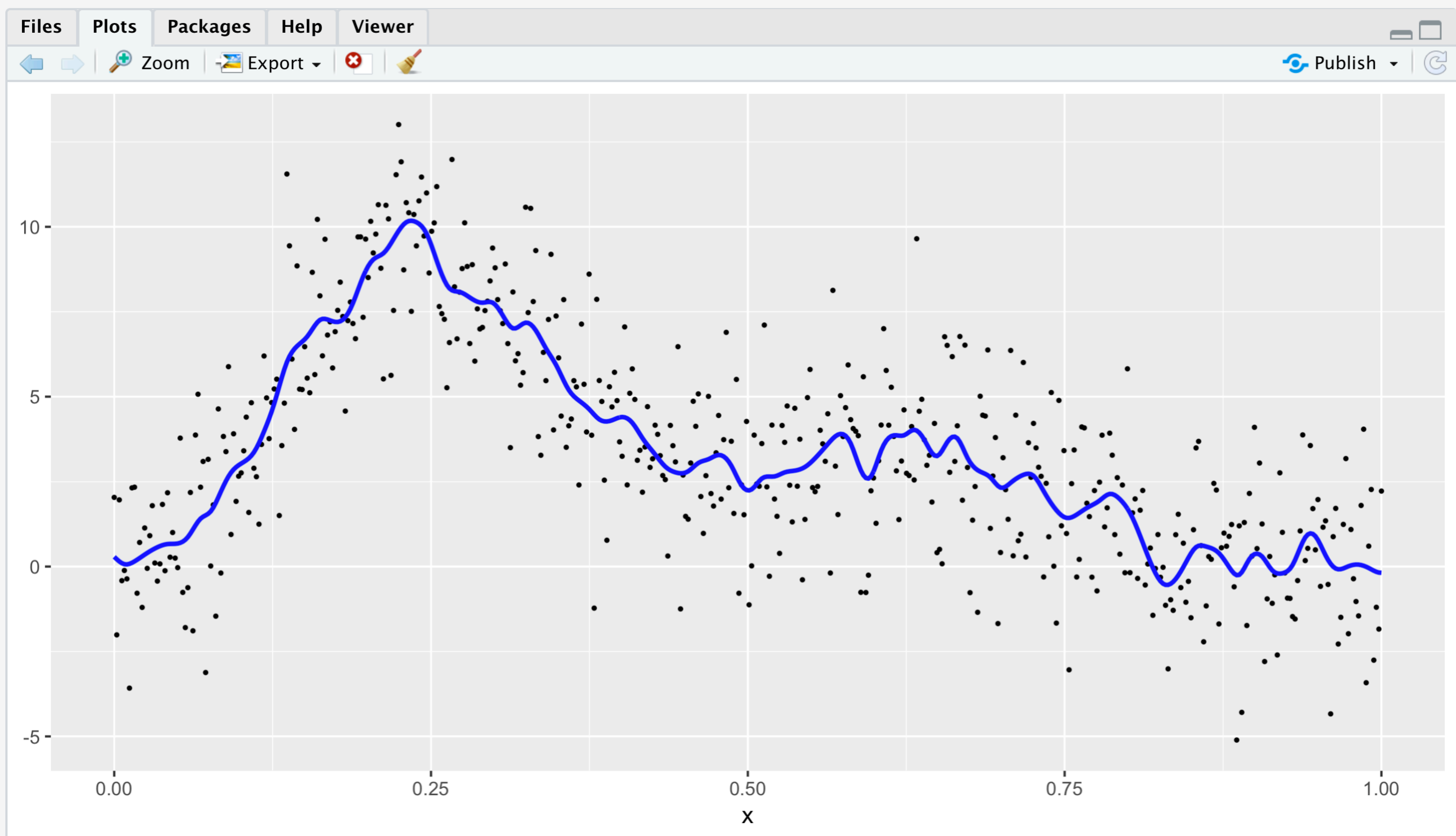
How about this $\hat{f}(x)$?



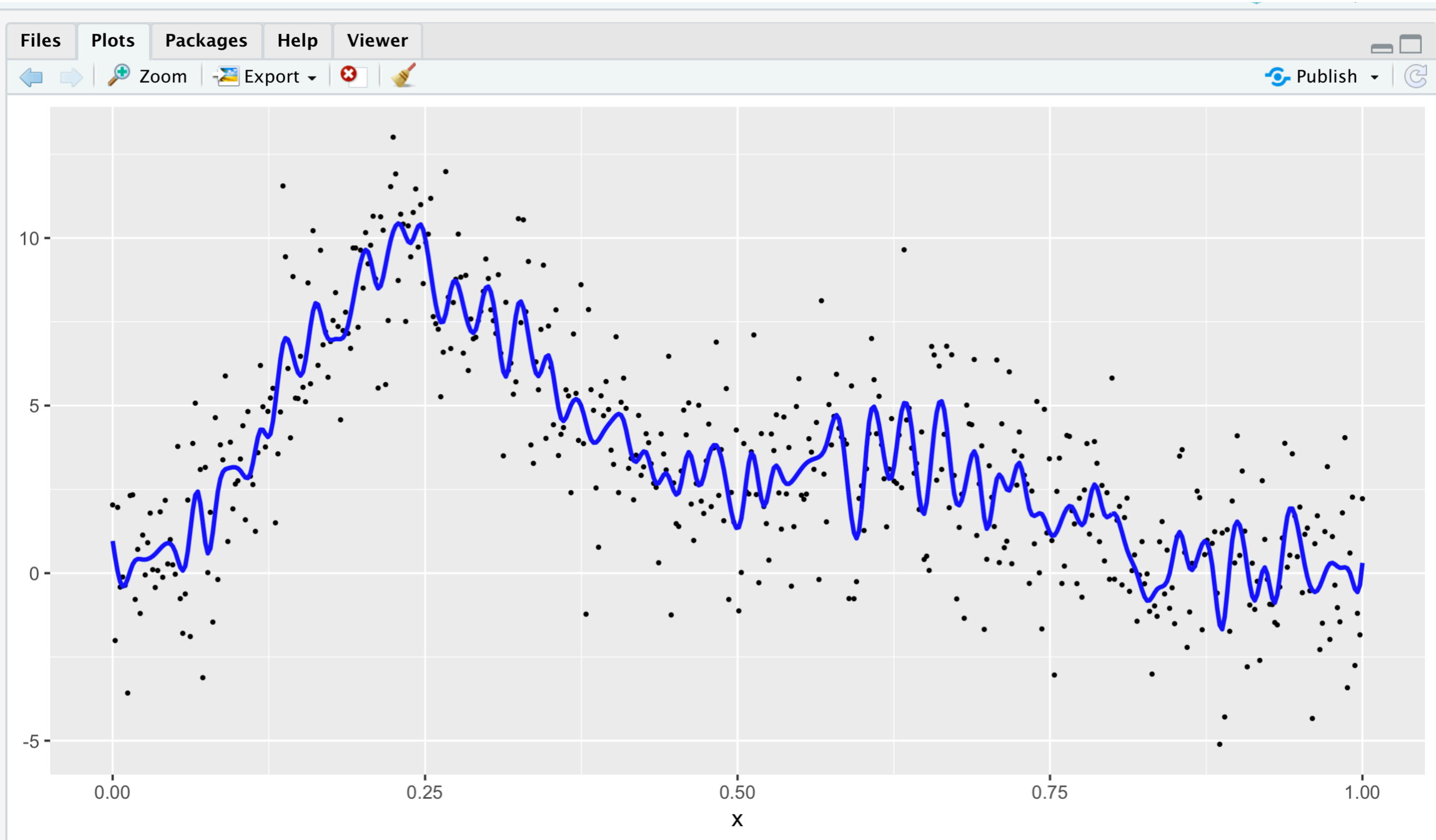
How about this $\hat{f}(x)$?



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



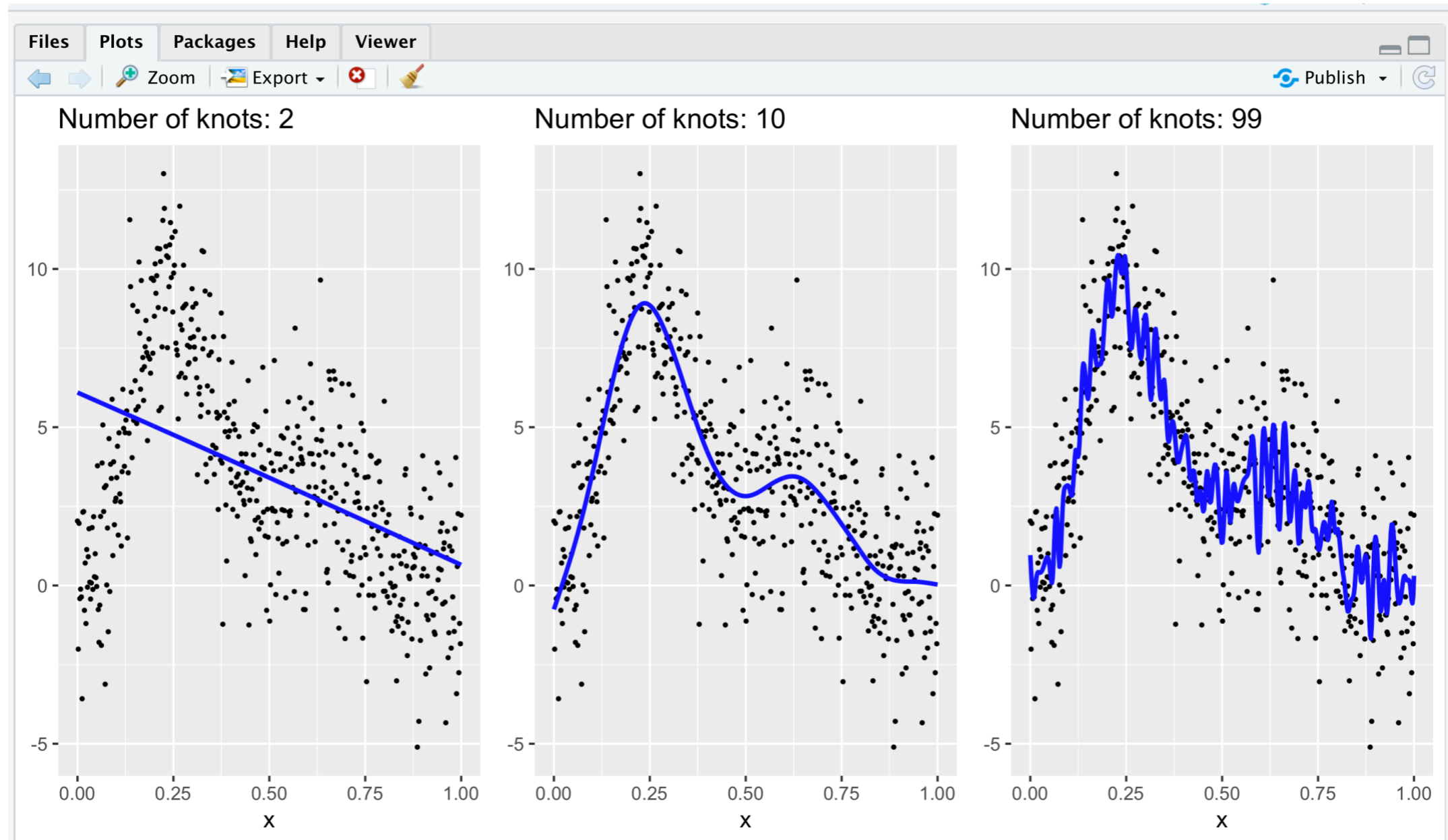
How about this $\hat{f}(x)$?



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 “wobble” is the **complexity of the model**
- Occam’s Razor

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



Underfit!

“Just right!”

Overfit!

What is Machine Learning?

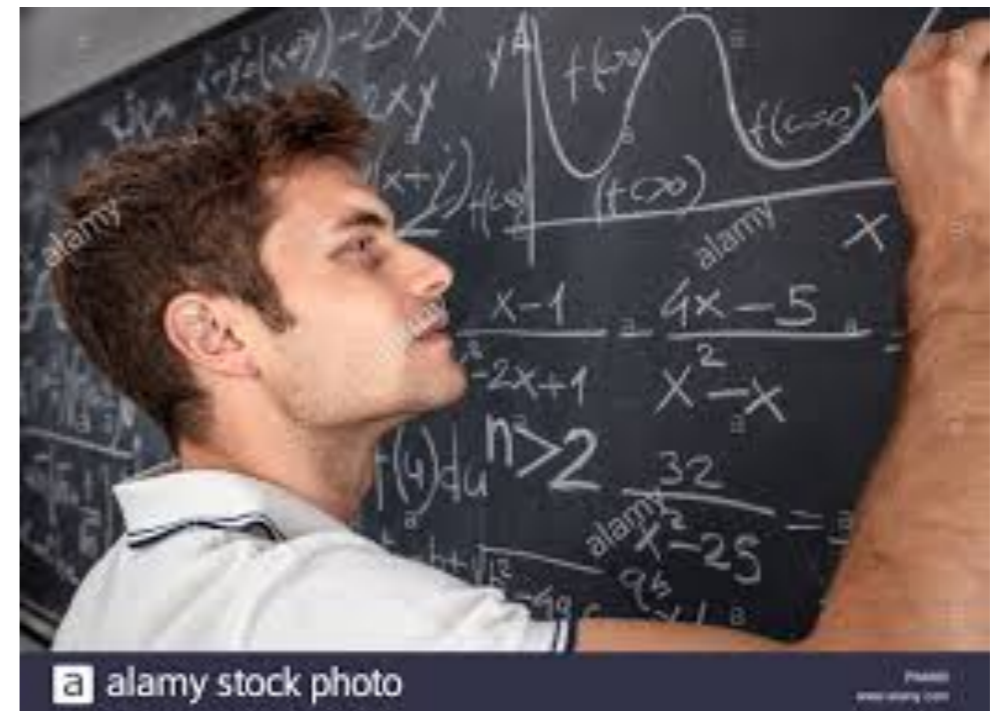
Machine Learning

- **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](https://www.kaggle.com):

The screenshot shows a Kaggle competition page for 'Google Analytics Customer Revenue Prediction'. The competition is a 'Featured Prediction Competition' with a prize money of \$45,000. It is hosted by RStudio and has 1,104 teams participating, with a month to go. The page includes a navigation menu with 'Overview', 'Data', 'Kernels', 'Discussion', 'Leaderboard', 'Rules', and 'Team'. The 'Leaderboard' tab is active, showing a 'Public Leaderboard' with a note that it is calculated with all test data. There are links for 'Raw Data' and 'Refresh'. A legend indicates medal colors: 'In the money' (green), 'Gold' (orange), 'Silver' (grey), and 'Bronze' (brown). The leaderboard table has columns for '#', 'Δ1w', 'Team Name', 'Kernel', 'Team Members', 'Score', 'Entries', and 'Last'. The top three teams are: 1. Marwen Sallem (In the money, Root Mean Squared Error score), 2. Paulo Pinto (Gold, Perfect Score), and 3. Its Me (Silver, Perfect Score). A red box highlights the 'Root Mean Squared Error' text in the score column for the first team.

Featured Prediction Competition

Google Analytics Customer Revenue Prediction

Predict how much GStore customers will spend

\$45,000 Prize Money

RStudio · 1,104 teams · a month to go

Overview Data Kernels Discussion **Leaderboard** Rules Team

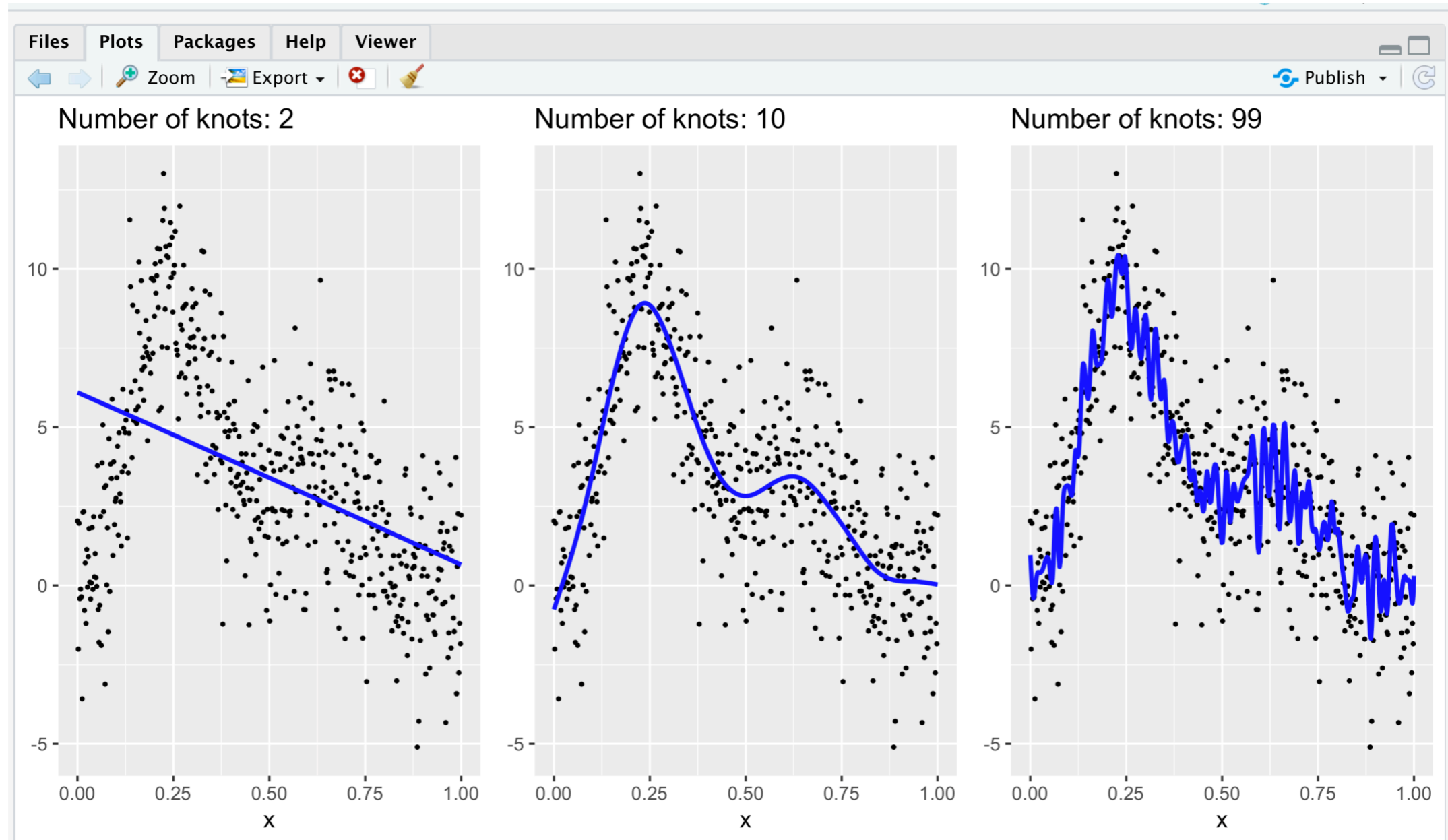
Public Leaderboard Private Leaderboard

This leaderboard is calculated with all of the test data. [Raw Data](#) [Refresh](#)

In the money Gold Silver Bronze

#	Δ1w	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	—	Marwen Sallem			Root Mean Squared Error	2	2mo
2	—	Paulo Pinto	</> 1line Perfect Score		0.0000	13	1mo
3	—	Its Me			0.0000	6	2mo

Issue of underfitting vs overfitting?



Underfit!

“Just right!”

Overfit!

Validation Set Approach

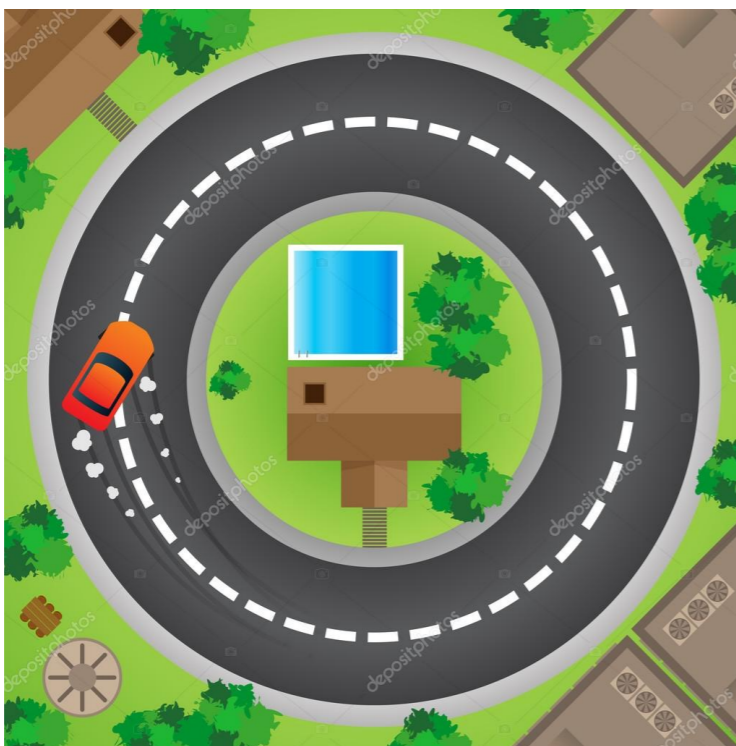


Split your data into:



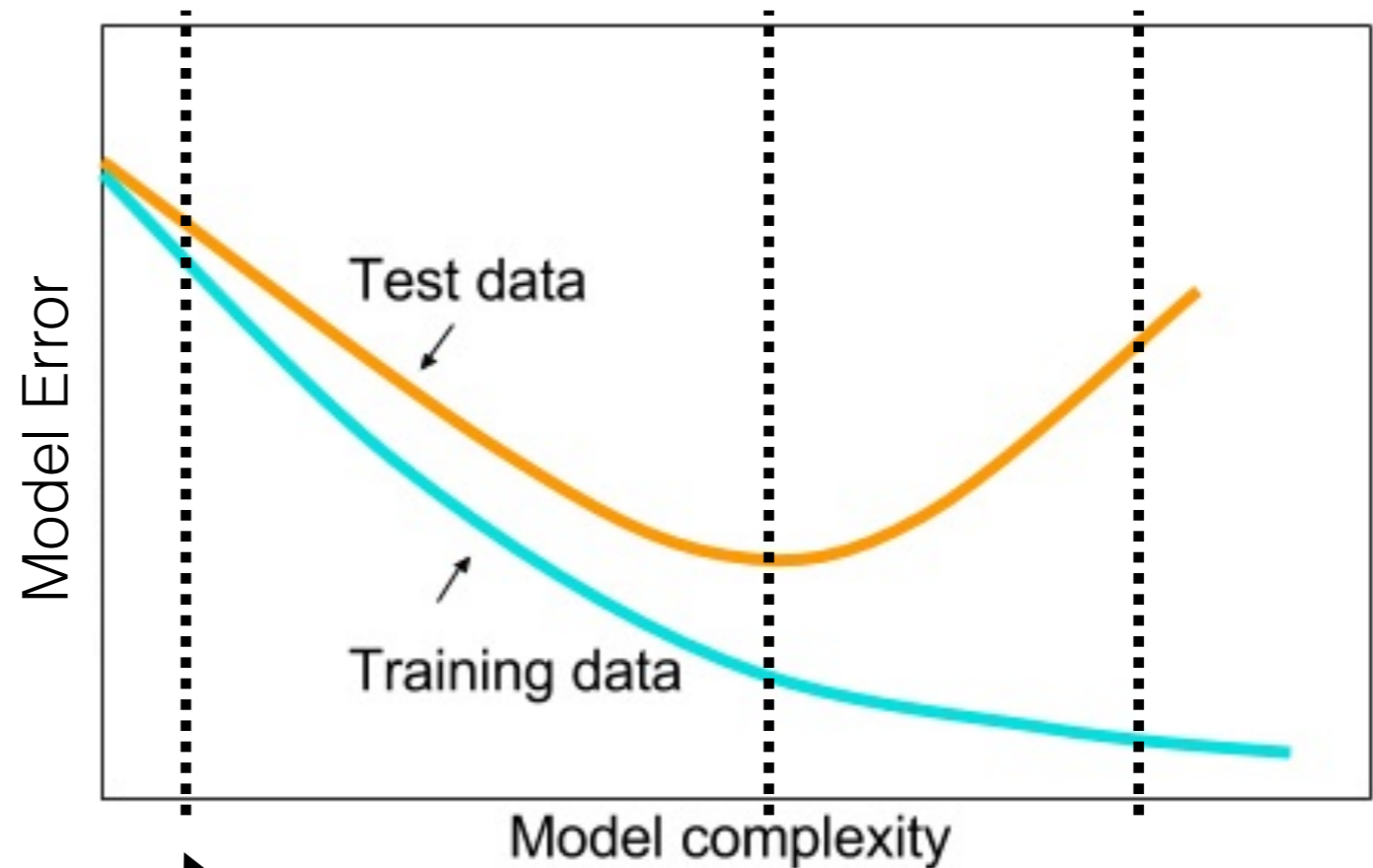
Fit/train model on *training* data

Assess model on independent *test* data



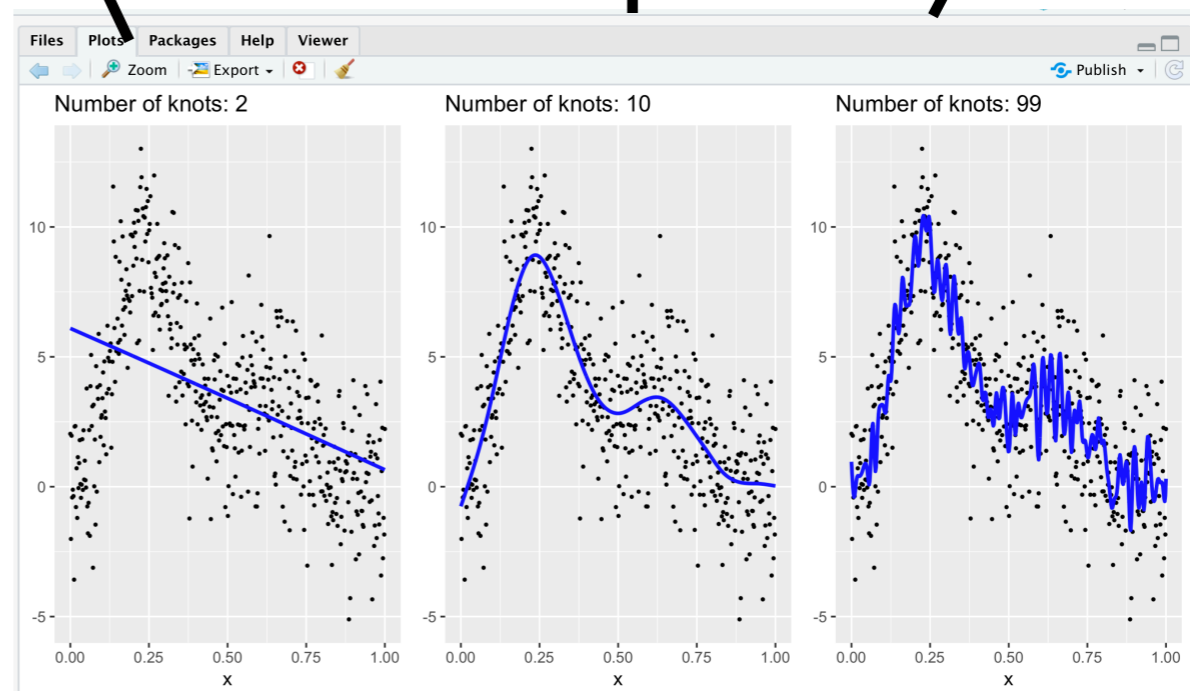
Typical Model Performance

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



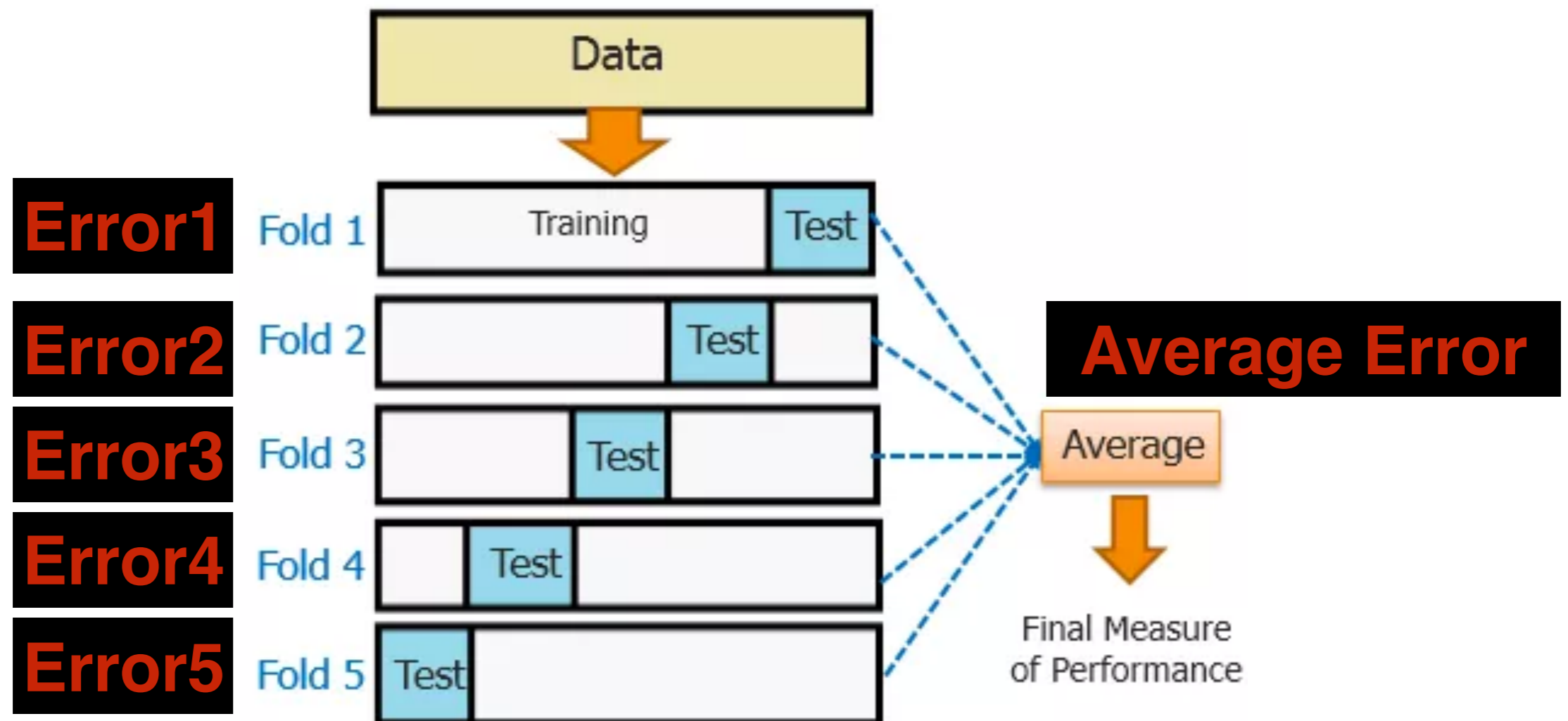
Underfit! “Just right!” Overfit!

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



Daniela Witten

@daniela_witten

Follow



Perfect gym for a statistician

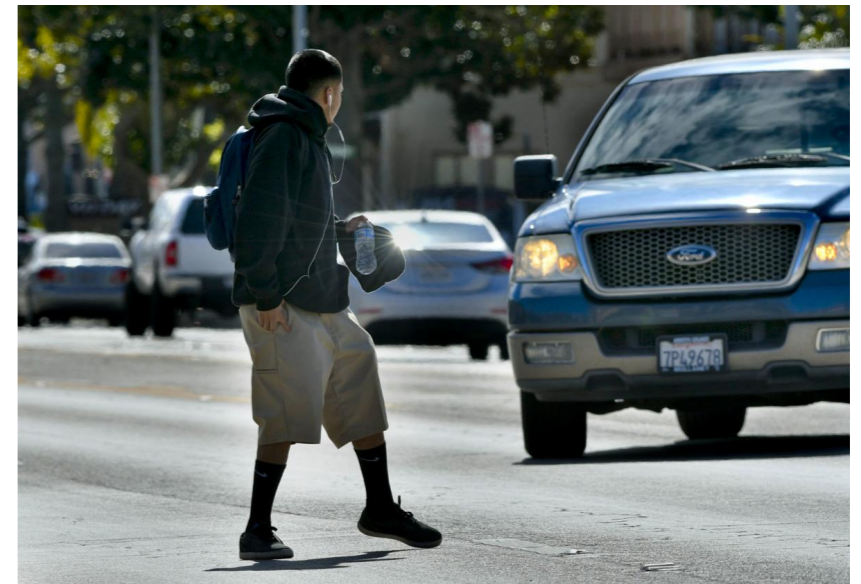
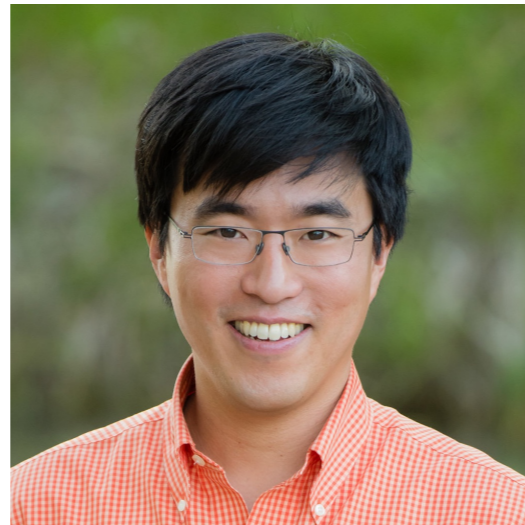


Modeling is not as objective as you think

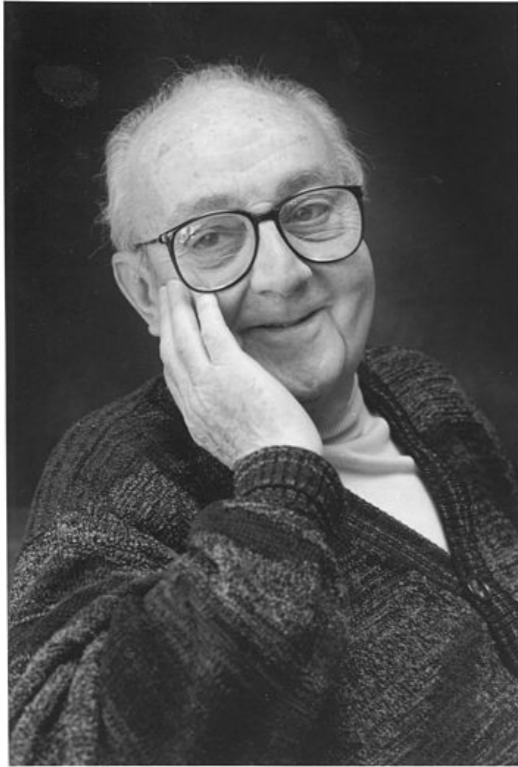
Scenario:

What they think is an
“appropriate” model...

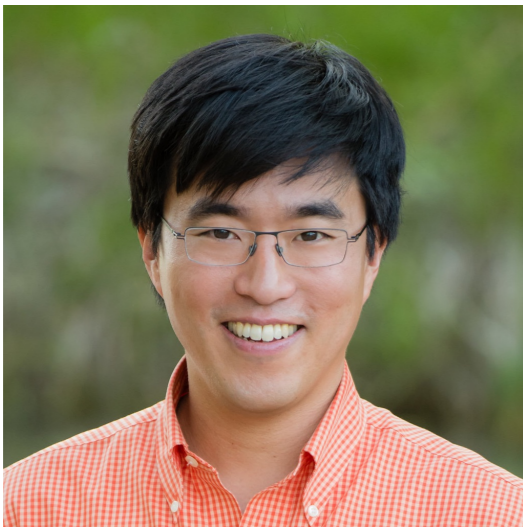
... might not be the
same for these folks:



To Close: Two Quotes on Modeling



“All models are wrong,
but some are useful.”
George Box

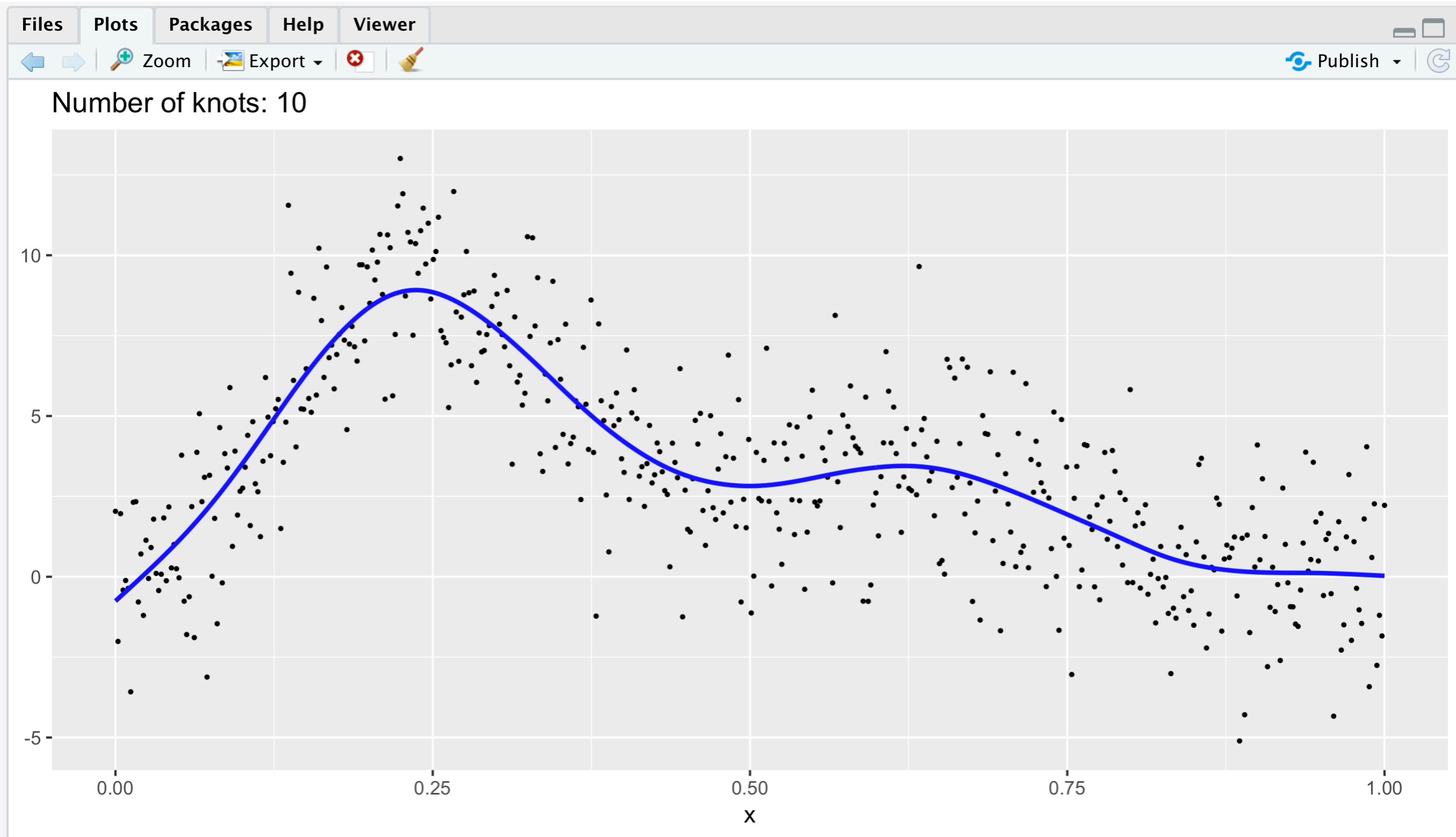


“WTF is up with your
 $\hat{f}(x)$?” @rudeboybert

Lastly: A “Wizard of Oz” Reveal...



Our approximated $\hat{f}(x)$...



... was pretty close to the *true* model:

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

