Machine Learning Theory

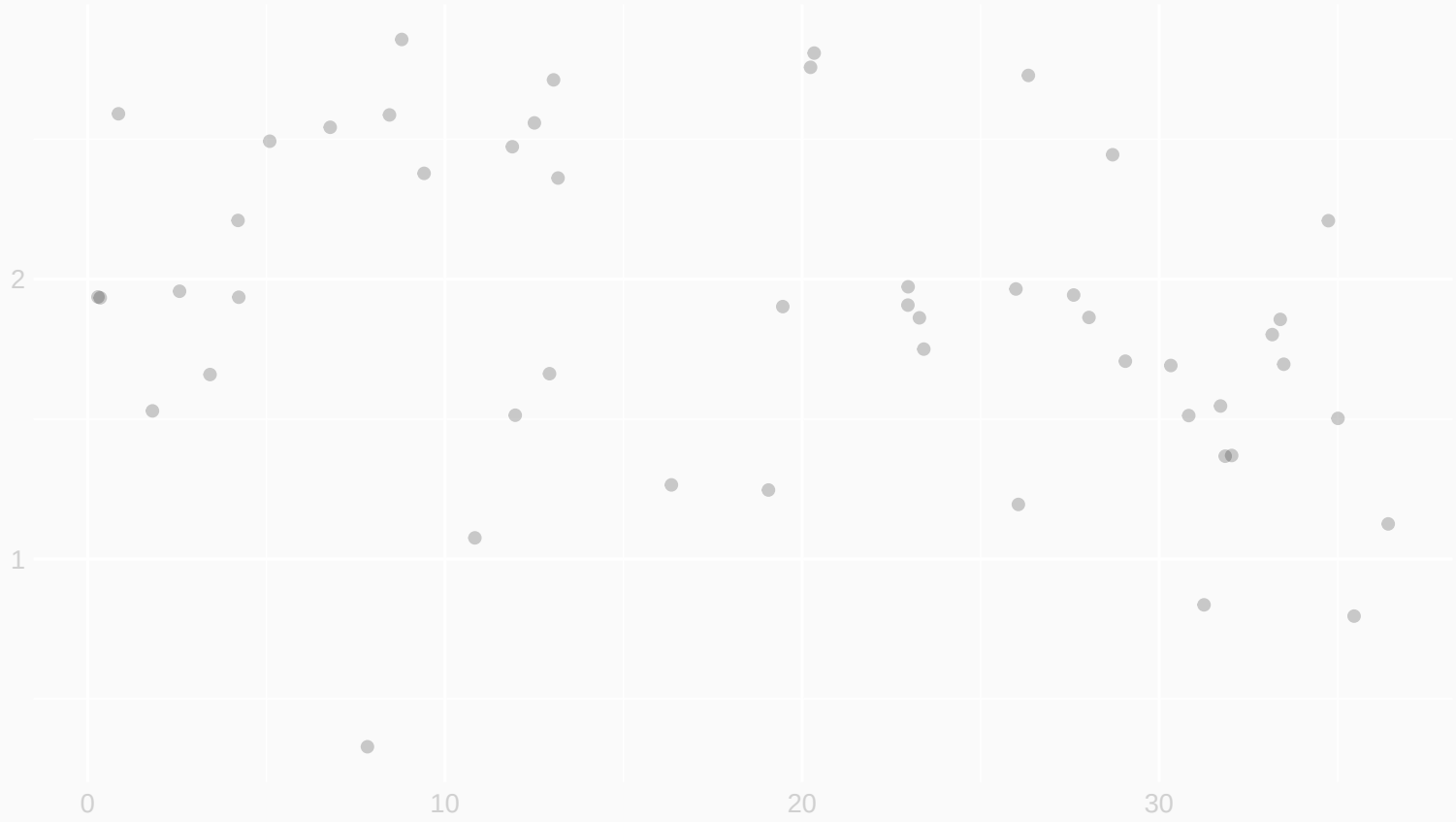
Introduction

David A. Hirshberg

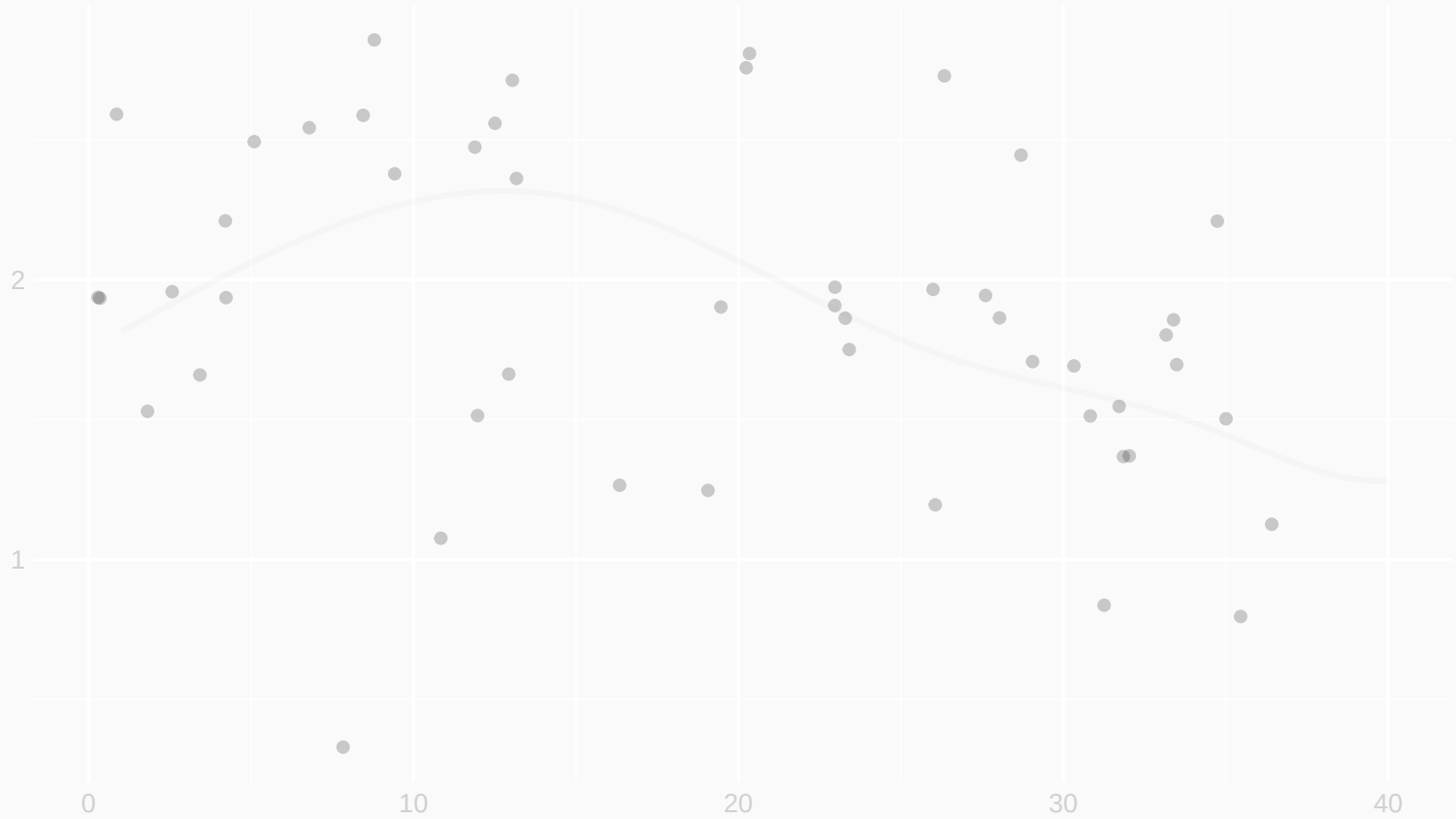
January 31, 2025

Emory University

When we look at data, we're often interested in what's typical. We're looking for some kind of central tendency in a mess.

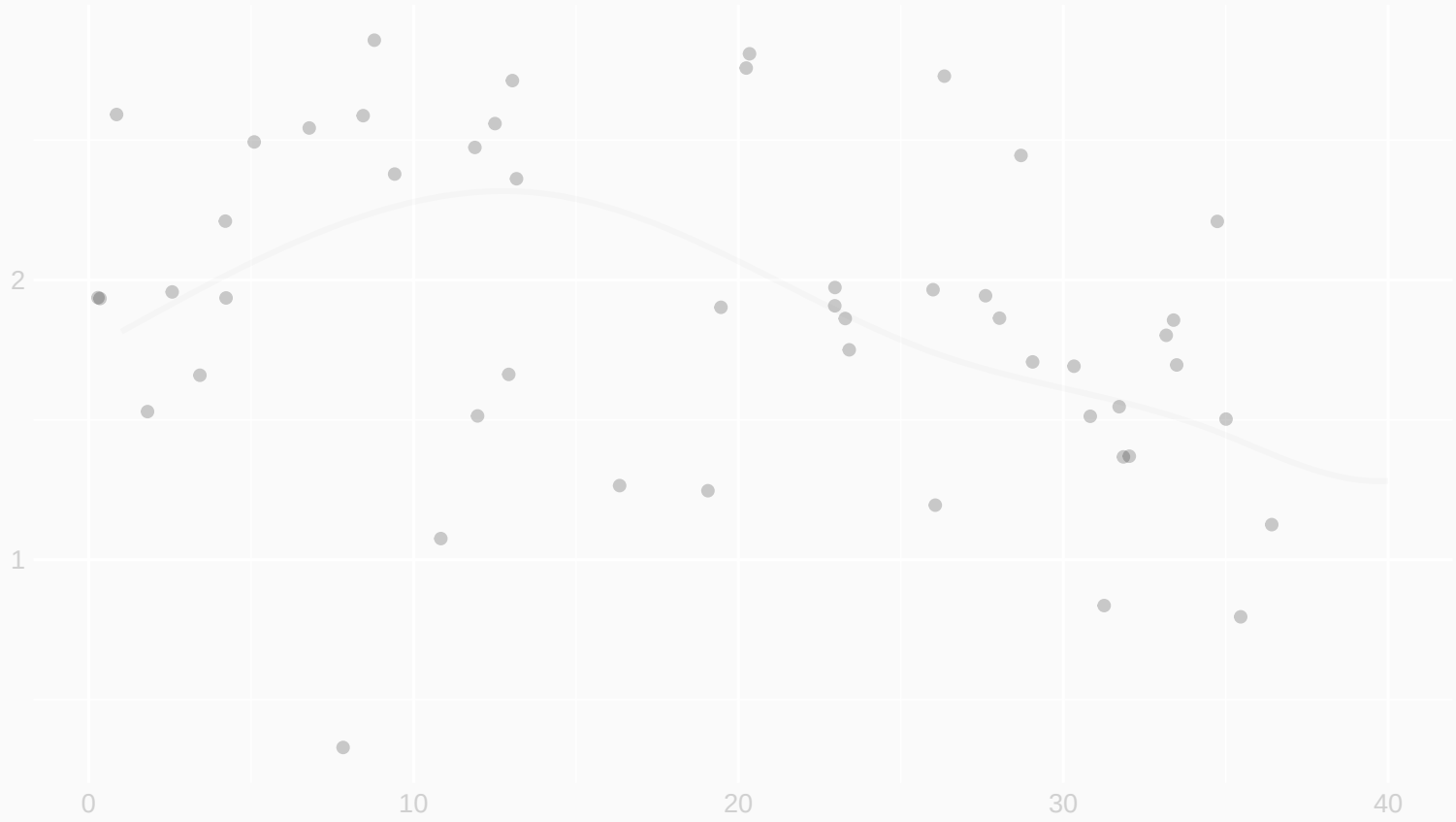


When we draw a curve, it makes the data easier to look at.
You start to see it as a mix of *typical behavior* and *random variation*.



$$Y_i = \underbrace{\mu(X_i)}_{\text{signal}} + \underbrace{\varepsilon_i}_{\text{noise}}$$

But seeing patterns is part of being human.
I could draw other curves here that would feel convincing.



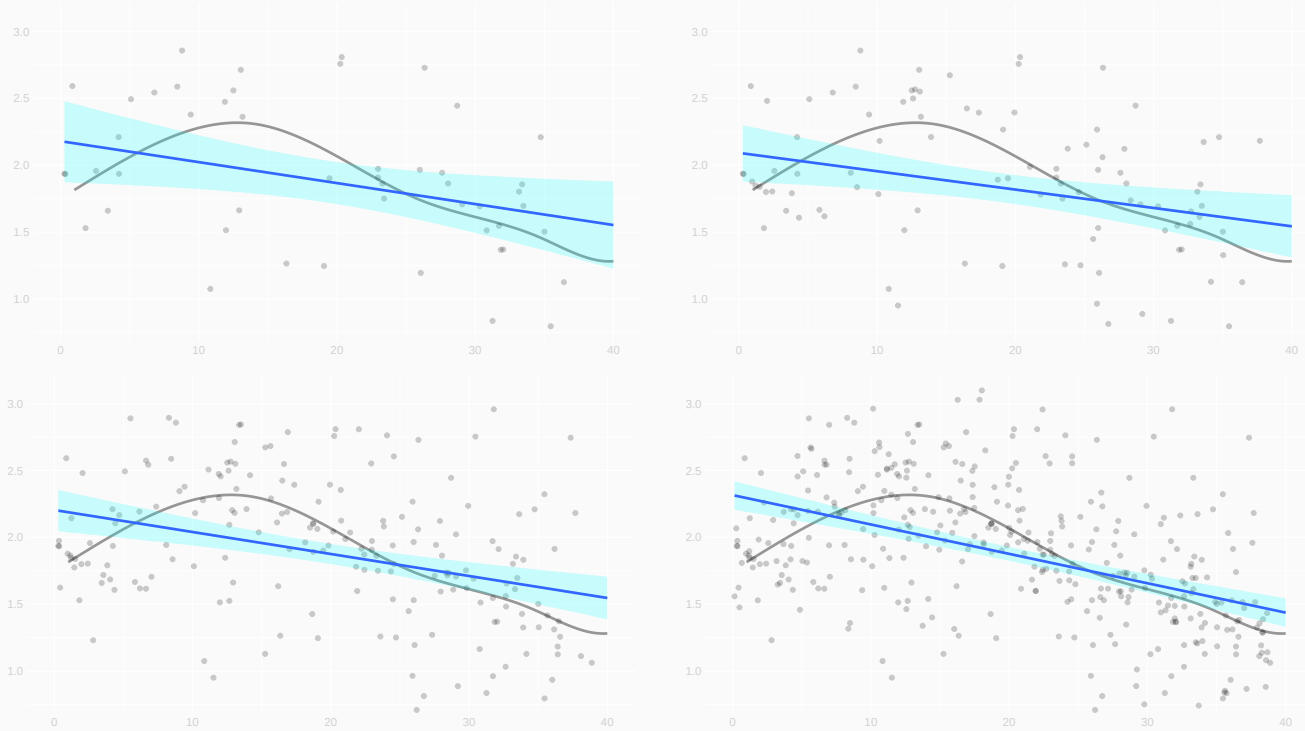
We need help keeping our pattern recognition faculties in check.
Otherwise, we'll convince ourselves of things that aren't true.

That's what statistical theory is for.

Perspectives on Classical Theory

What is it?

Classical asymptotic theory describes what happens as sample size increases when we fit a model with a finite number of parameters, like a line.



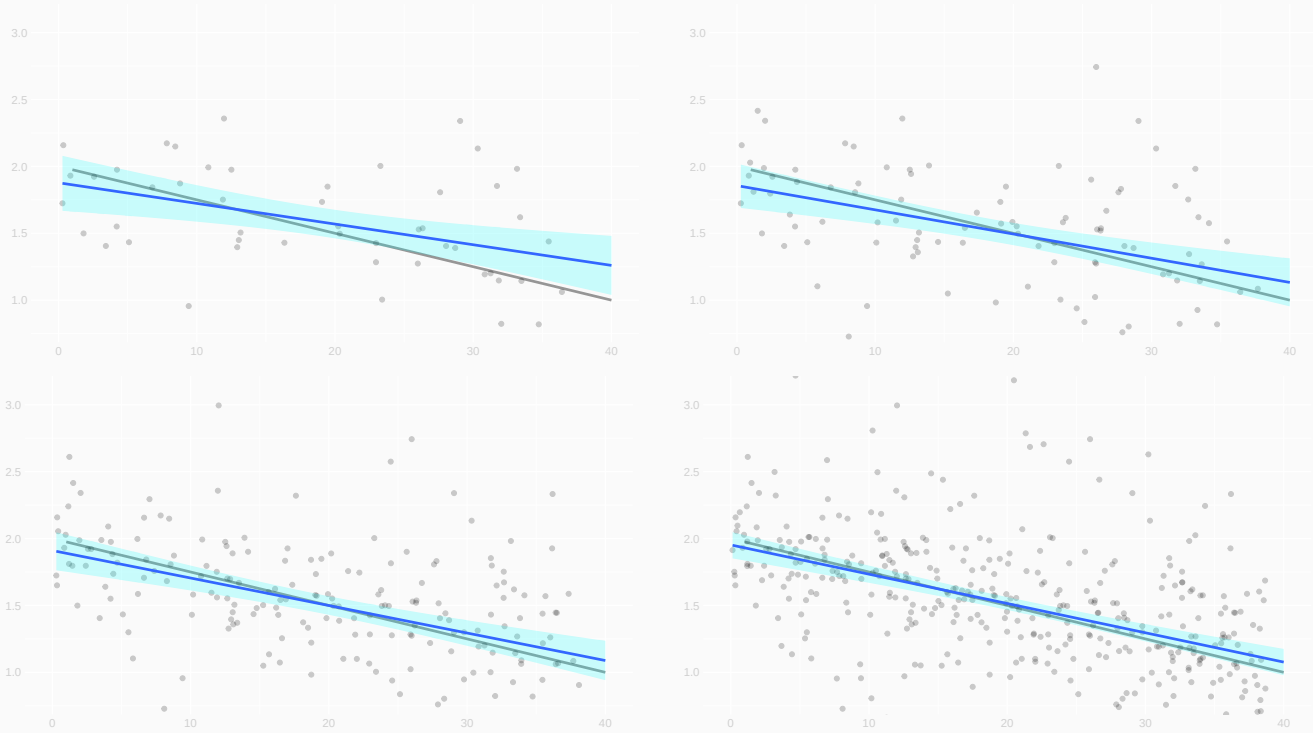
This is what classical asymptotic theory thinks you're doing.

- It tells you how your estimate behaves when sample size is big enough.
- Essentially without specifying how big it'd have to be for this to happen.

The Intro Textbook Perspective

Intro textbooks pile on a bunch of additional assumptions.

They describe what happens when your parameters describe the distribution of the data *completely and correctly*.



This is more their style.

- If you fit a line, the data had better come from a line.
- If you use least squares, you'd better have homoskedastic gaussian noise, too.

If you buy this, data analysis turns out to be pretty textbook.

There's one recipe that works for everything.

- No matter what model you're using, you always know how to fit it. Maximum likelihood is essentially always asymptotically optimal.

$$\hat{p}_{MLE} = p_{\hat{\theta}} \text{ for } \hat{\theta} = \operatorname*{argmax}_{\theta} p_\theta(\text{data})$$

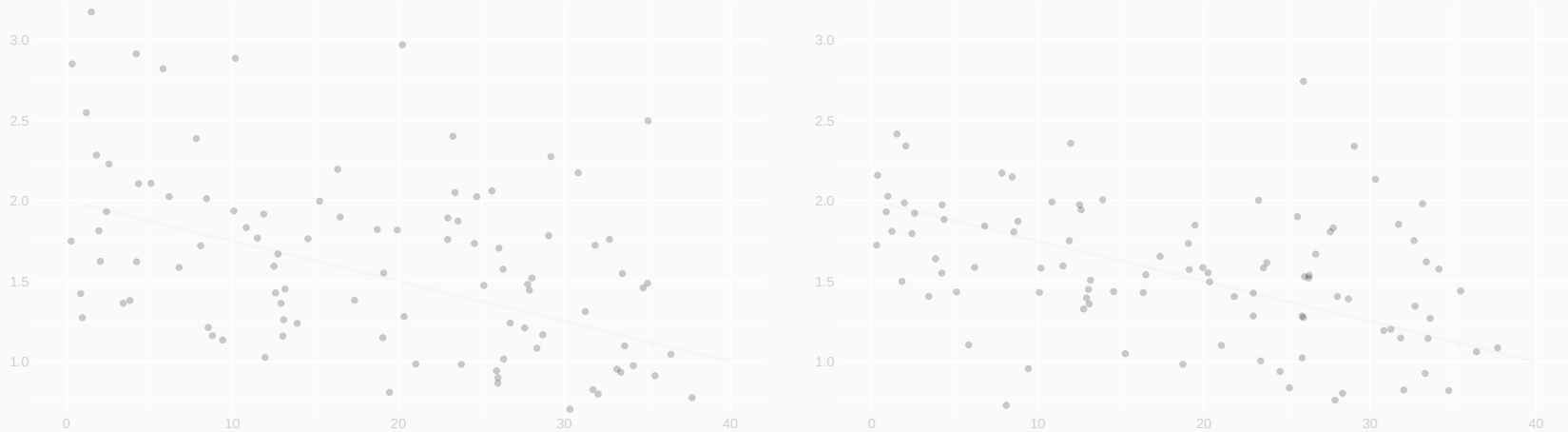
- It doesn't matter what you want to know, either.
- If you're interested in a function of the data's distribution, like an average treatment effect, it's asymptotically optimal to plug the MLE into that function.

$$ATE(\hat{p}_{MLE}) \quad \text{optimally estimates} \quad ATE(p)$$

You’re on your own for the hard part, it turns out

- The MLE depends on aspects of the data’s distribution that you may not care about and probably don’t have much intuition for.
  - Like the distribution of the noise ε_i.
- If you don’t get this right, none of this optimality stuff is true.
- Sometimes you even wind up pretty far from the truth.

That leaves you arguing with your coworkers about nonsense.
Like whether the noise you’ve got is gaussian or laplacian.



Which is which?
Take a guess!

Doing without all this.

It turns out that you don't need the assumptions at all.

If we use any regression model with finitely many parameters
and we fit it via least squares, then as sample size gets large:

1. Its limit is the model's best approximation to Y's conditional mean.

$$\tilde{\mu} = \underset{m \in \text{model}}{\operatorname{argmin}} \, \mathrm{E}\left\{ m(x) - \mathrm{E}[\, Y \mid X = x] \right\}^2 .$$

2. And the difference between our estimate and its limit,
   in any sense you'd care to measure it, is roughly $1/\sqrt{n}$.

$$f(\hat{\mu}) - f(\tilde{\mu}) \approx 1/\sqrt{n}$$

3. And statistical inference is easy. t-statistics are asymptotically standard normal.

$$\sqrt{n}\{f(\hat{\mu}) - f(\tilde{\mu})\}/\sigma \approx N(0, 1)$$

All of this, however, refers to our ability to estimate the model's best approximation.
That's not something that's particularly meaningful without reference to the model.

You can do this for (other) maximum likelihood estimators, too.
You just need a bit more terminology.

This is an *agnostic interpretation*.

- It describes what happens to estimators inspired by some belief
- (essentially, that the model is correct or anyway good enough)
- and it does that without assuming that the belief is true.

It's how statisticians have always thought about simple models.

- The textbook interpretation, in which strong assumptions lead to optimal procedures, helped you think about what to try.
- The agnostic one helped you think about what would happen when you did.

Together, these worked pretty well up until the 60s or so.

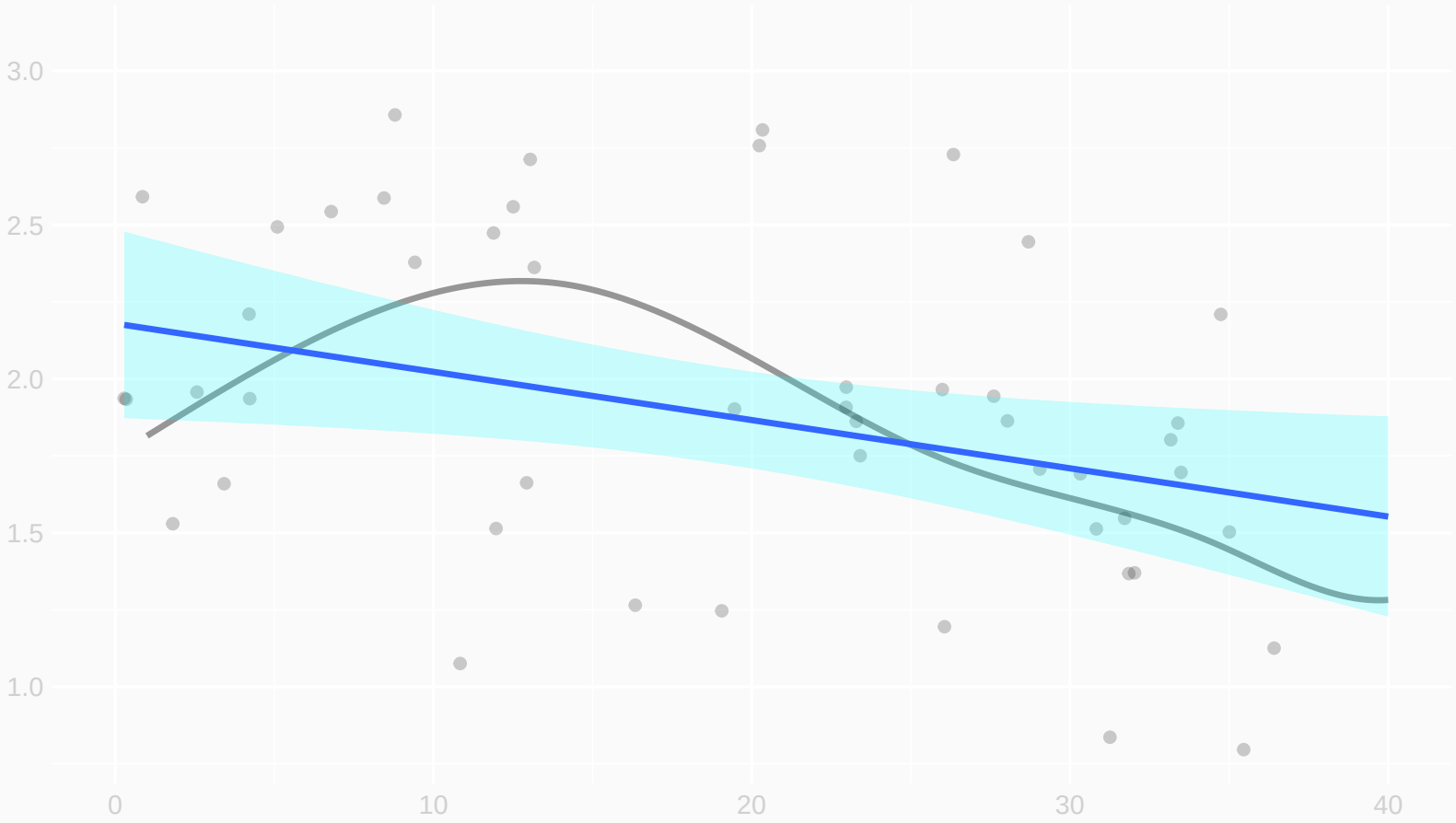
- Models back then had to be simple.
- We had to fit them with pencil and paper.
- So classical asymptotics described what people were actually doing.

But in a modern context, this approach looks pretty weird.

Classical Theory in Action

Let’s interpret some fits!

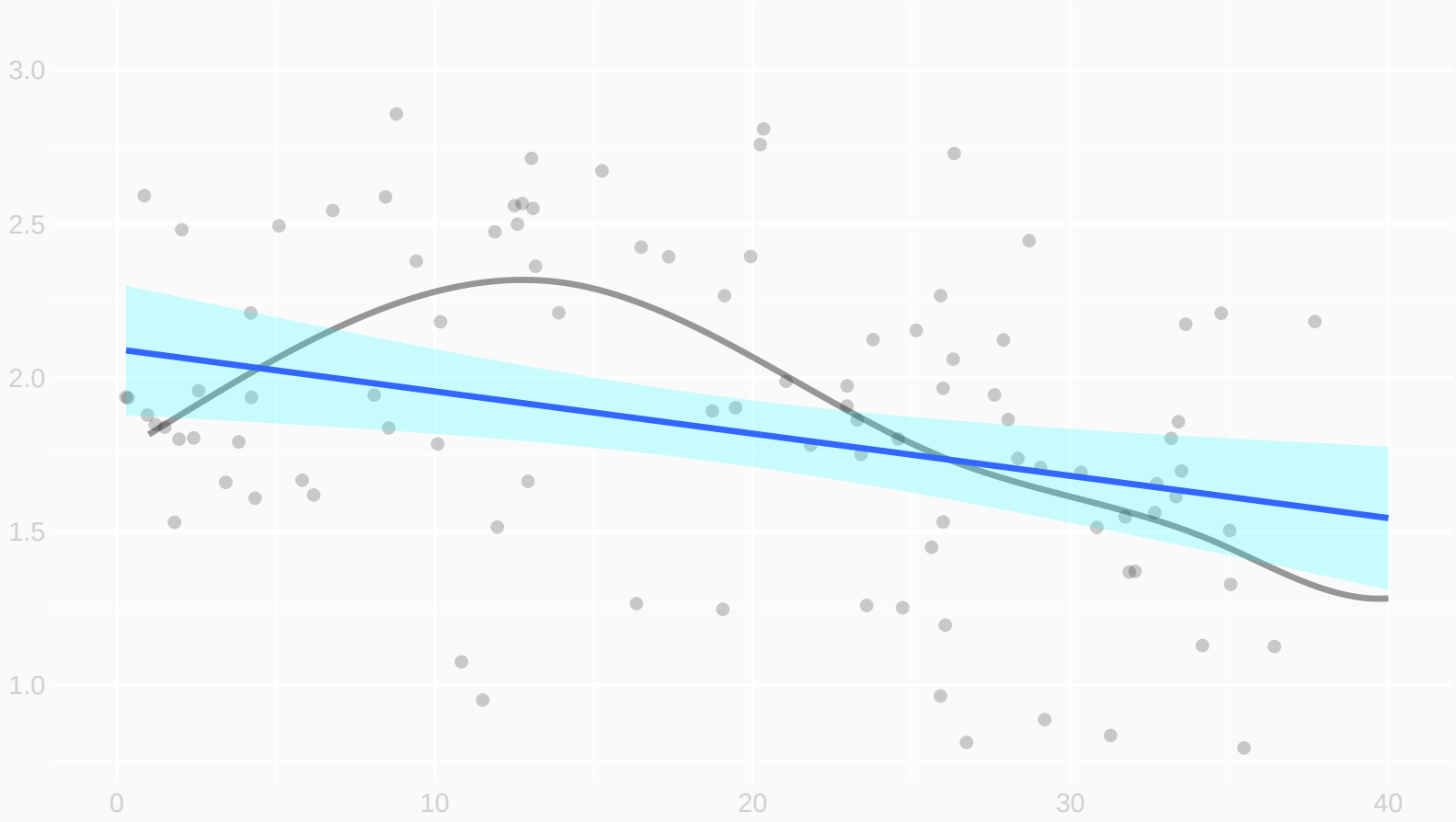
We'll fit a line.



The model we're using, one in which the data follows a linear trend, is wrong.

- If you believe your intro textbook, you're in trouble.
  - Anything you could say would be wrong, too.
- If you take the agnostic perspective, you're not going to be wrong.
  - But you're not going to say anything useful either.

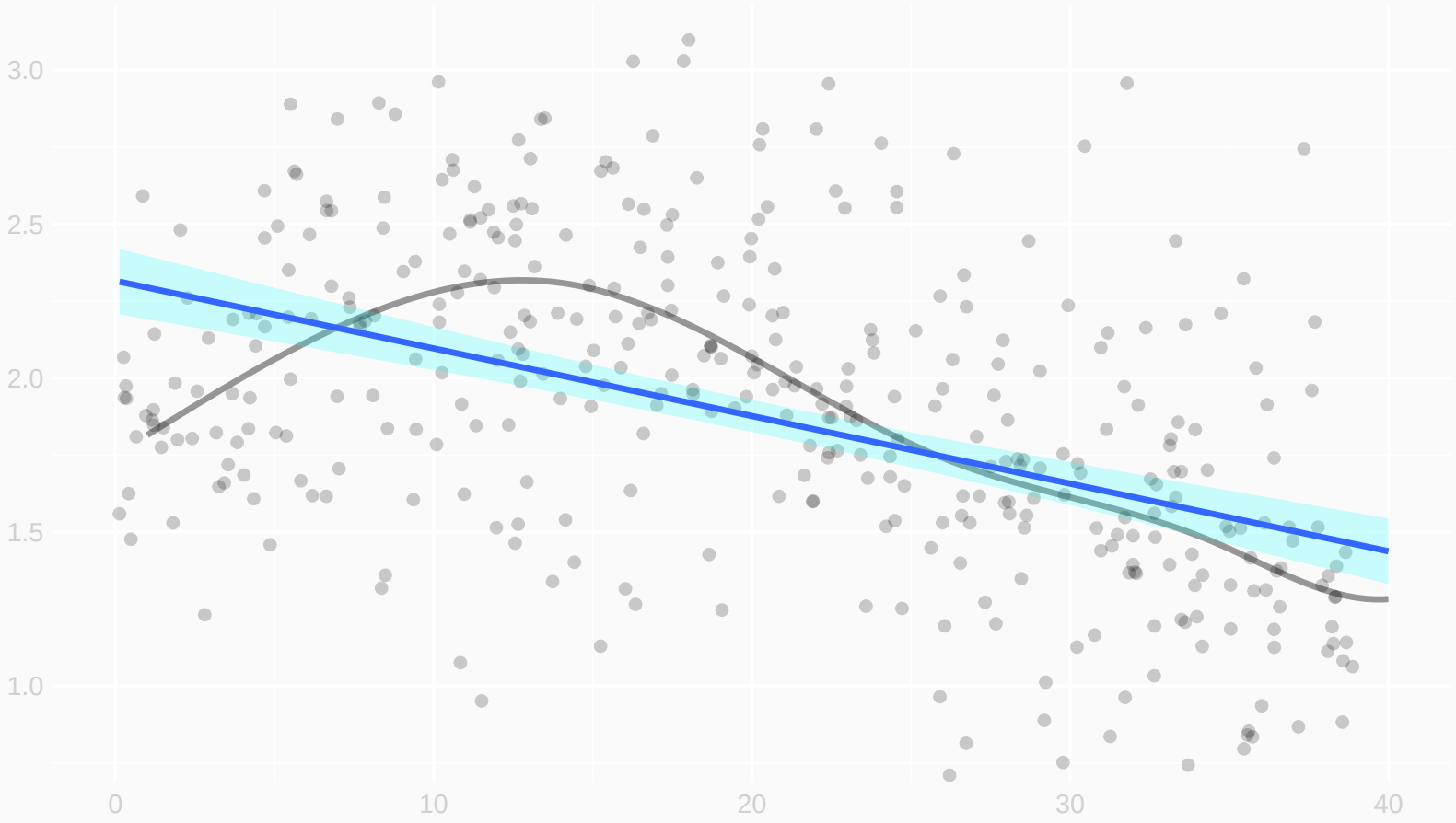
We'll fit a line.



The model we're using, one in which the data follows a linear trend, is wrong.

- As sample size grows, your textbook acts more and more confident in a falsehood.
- And agnostics get more and more precise about a useless technicality.

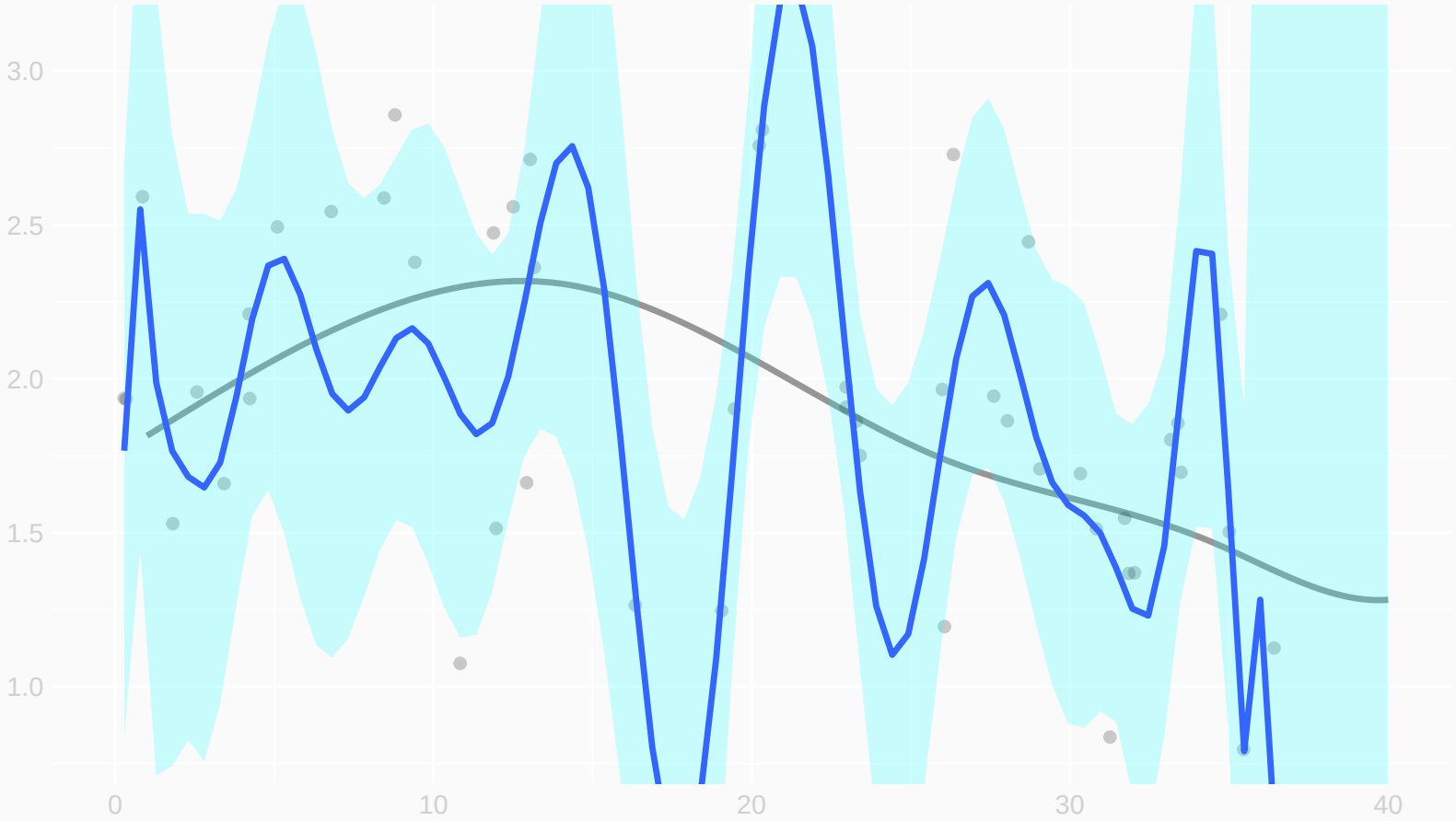
We'll fit a line.



The model we're using, one in which the data follows a linear trend, is wrong.

- When we've tons of data, your textbook will be very precise and very inaccurate.
- And agnostics are very precise, very accurate, and very uninteresting.

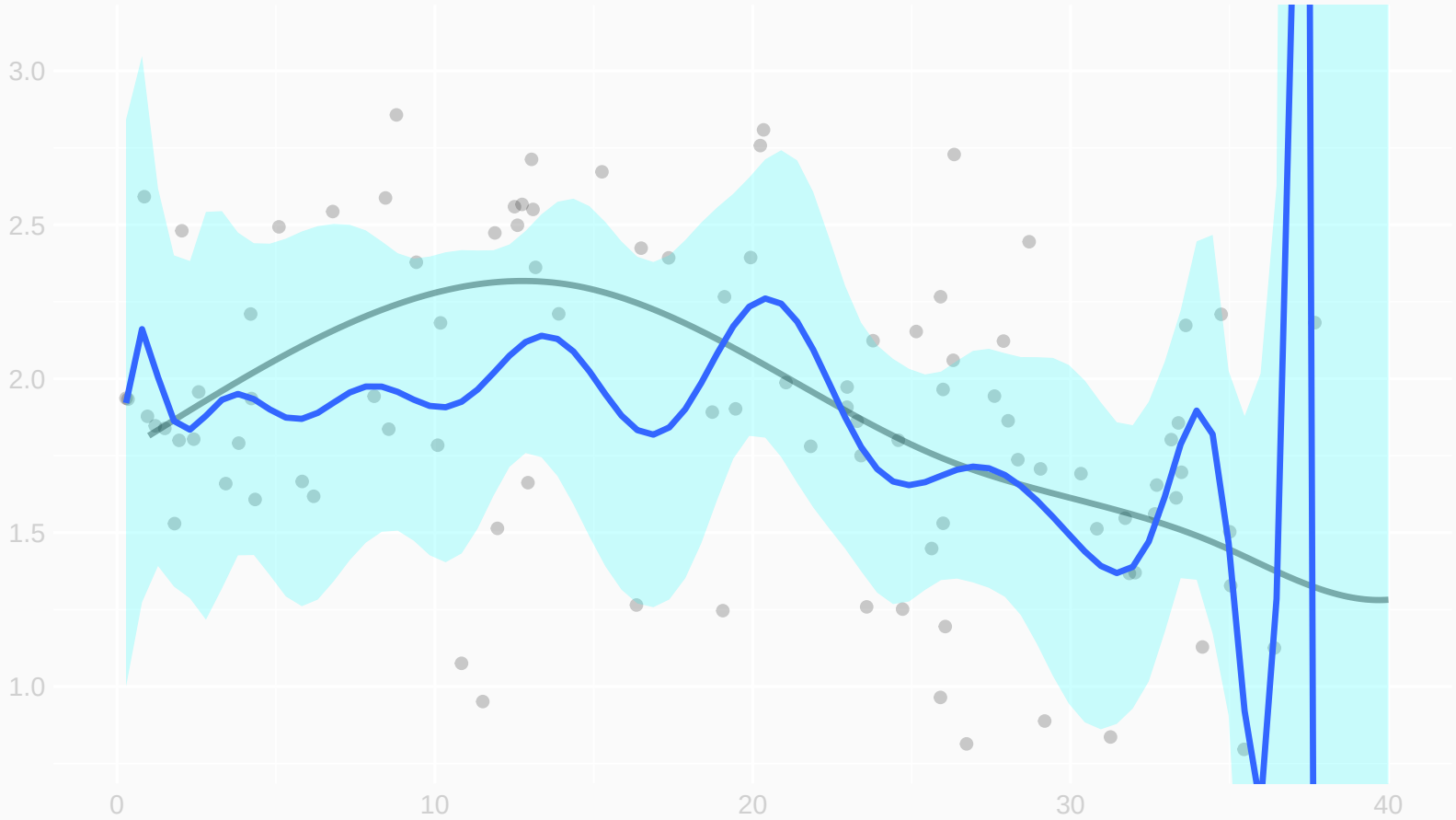
We'll fit a 20th-order polynomial



Now the model we're fitting can approximate the trend well.

- That means your intro textbook won't be meaningfully wrong.
- But it can't say much of anything without a ton of data.
- And neither can the agnostic interpretation.

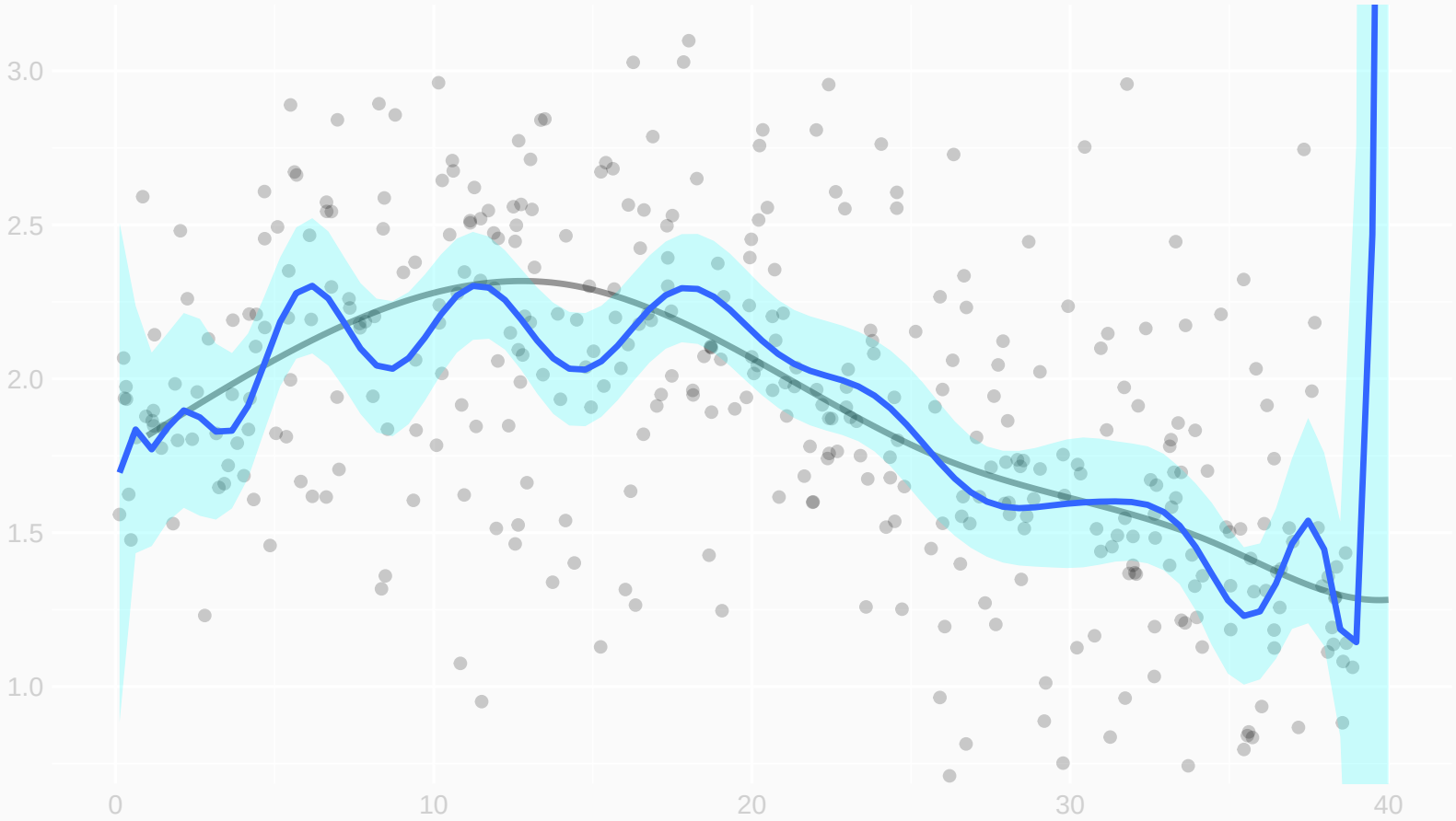
We'll fit a 20th-order polynomial



Now the model we're fitting can approximate the trend well.

- That means your intro textbook won't be meaningfully wrong.
- But it can't say much of anything without a ton of data.
- And neither can the agnostic interpretation.

We'll fit a 20th-order polynomial



Now the model we're fitting can approximate the trend well.

- That means your intro textbook won't be meaningfully wrong.
- But it can't say much of anything without a ton of data.
- And neither can the agnostic interpretation.

What looked weird to you?

What looked weird to you?

- To me, it looked like we were often using models we wouldn't really use.
- That's because our sample size was growing, but our models weren't.
- And technically speaking, that's what classical asymptotics tells us about.

Classical asymptotics is based on a flawed premise

Classical asymptotics describes how our estimates would improve
if we got more data and we kept using the same model.

That's not what we do. We try to pick a model that works at the sample size we have.

- We've got rules of thumb, e.g. 20 observations per parameter.
- We've got model selection heuristics, e.g. AIC.
- We've got cross-validation.
- We've got goodness-of-fit tests.
- And we've got eyes.

Nobody recommends choosing a model without thinking about the sample you have.

Let’s try following the rule of thumb.

Rule-of-thumb Polynomial Fits

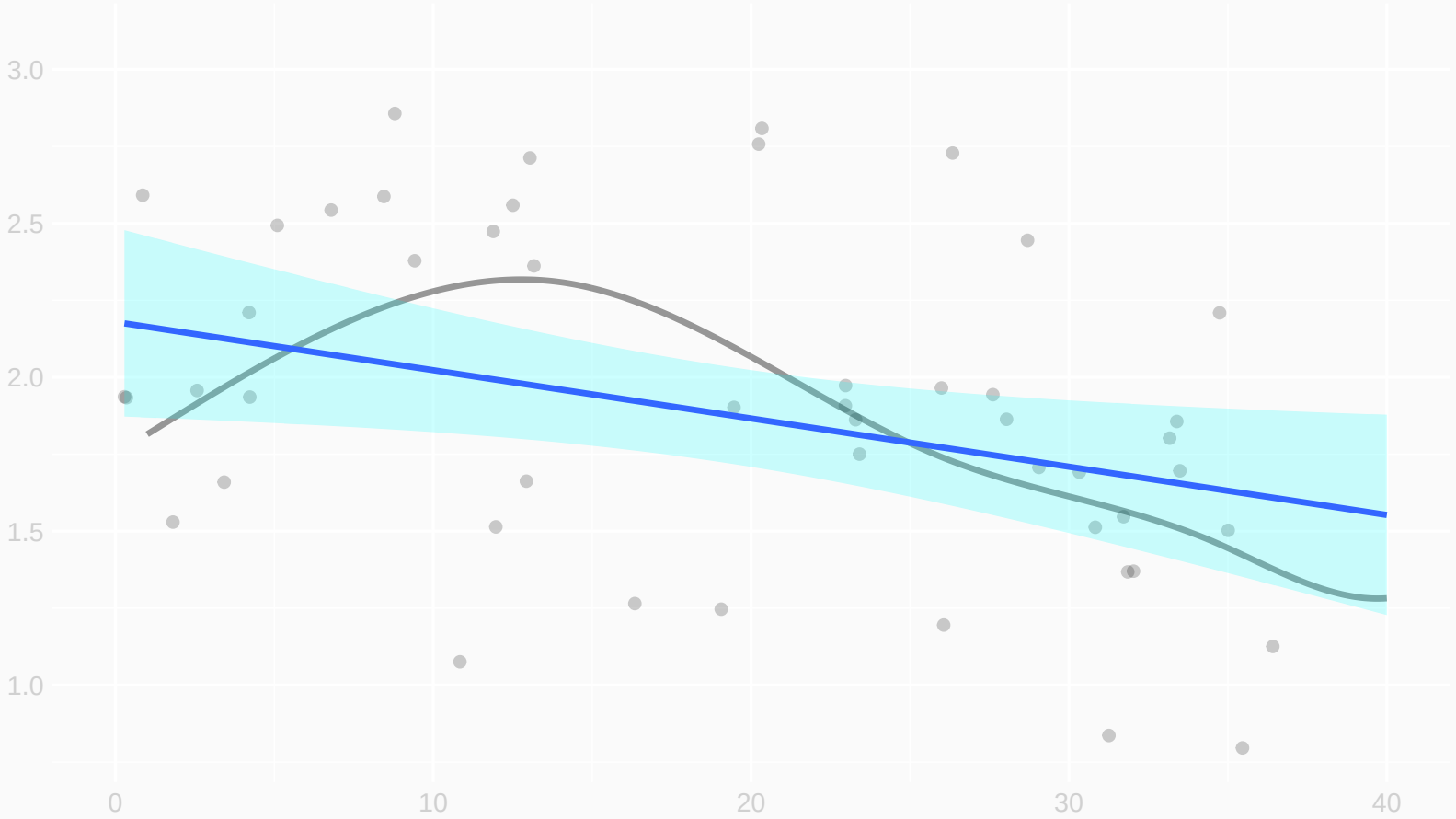


Figure 1: 50 observations, 2 parameters

Rule-of-thumb Polynomial Fits

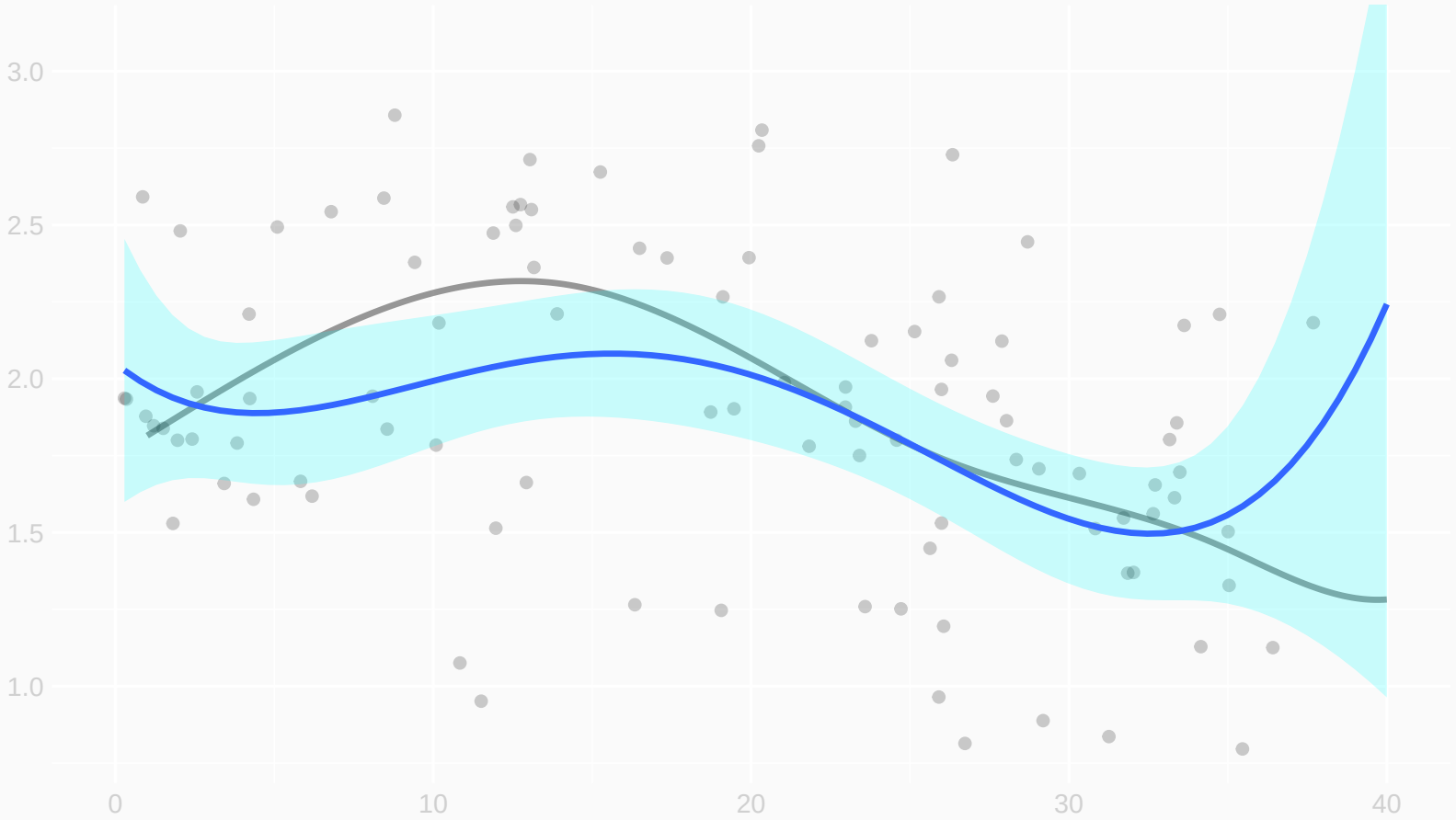


Figure 1: 100 observations, 5 parameters

Rule-of-thumb Polynomial Fits

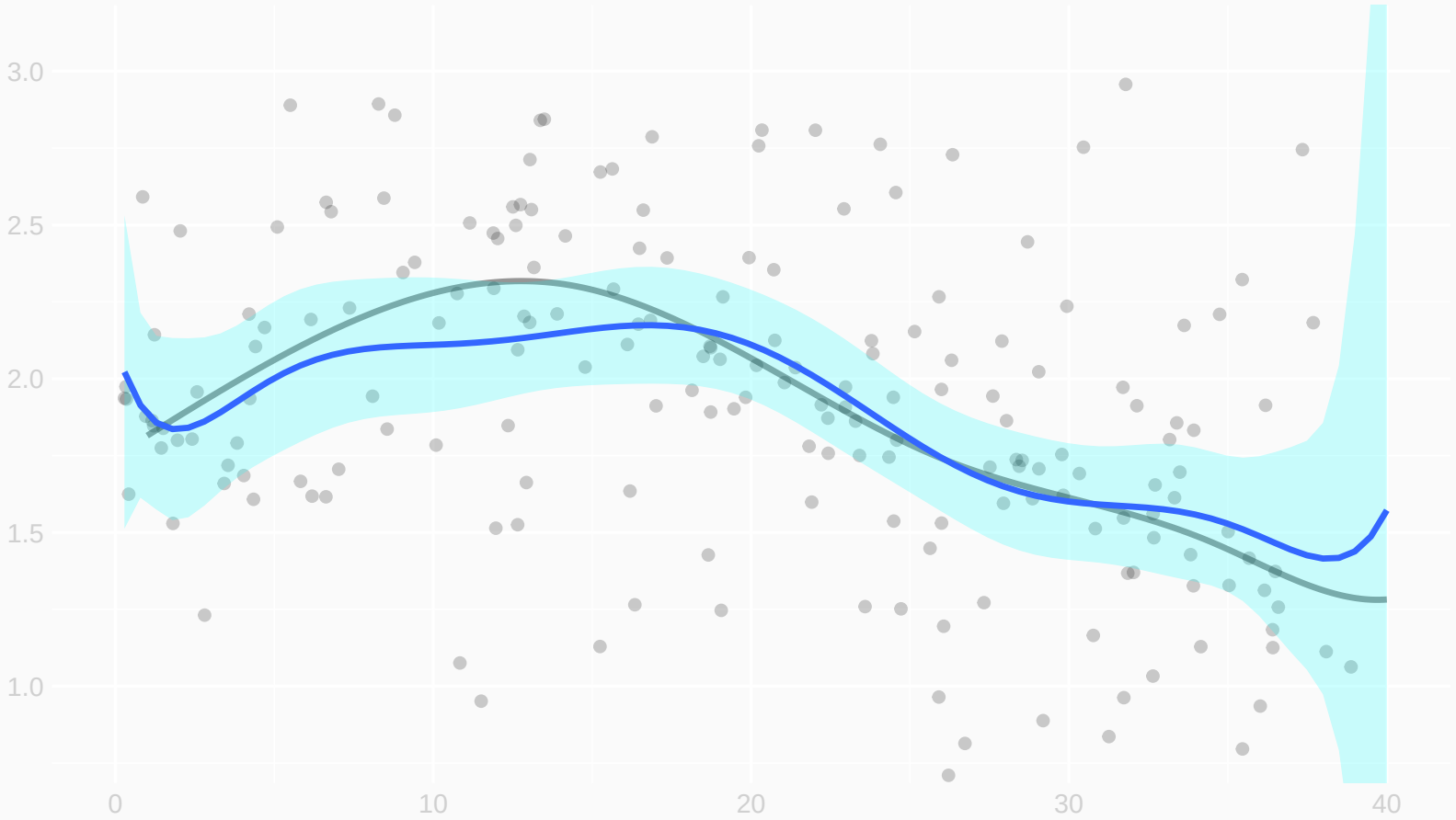


Figure 1: 200 observations, 10 parameters

Rule-of-thumb Polynomial Fits

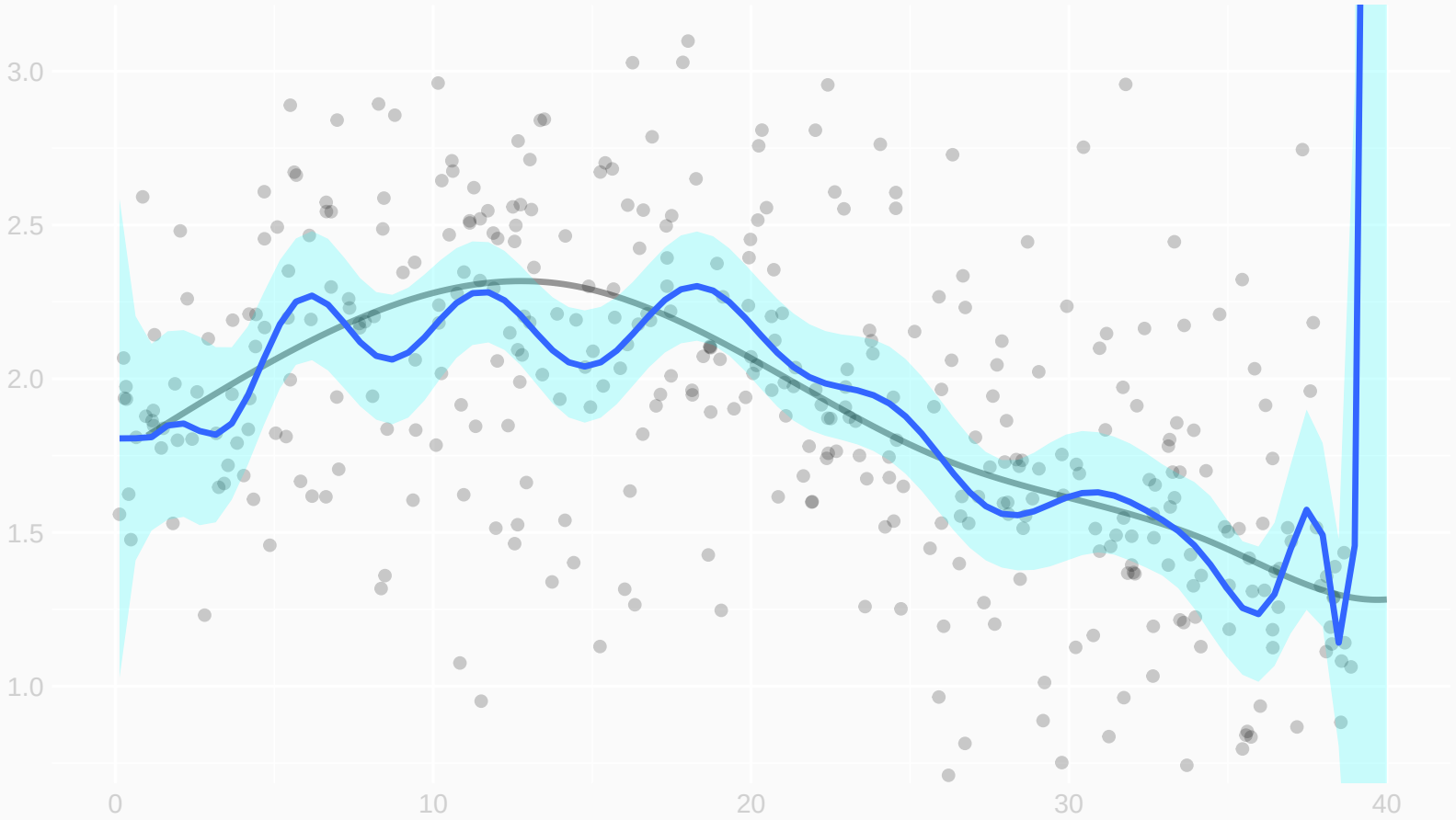


Figure 1: 400 observations, 20 parameters

Our rule of thumb works—more or less

We're fitting the curve we want. But what classical theory predicts isn't happening.

- Our confidence intervals seem to work ok, but they're not really getting narrower.
- The difference from our fitted curve and its limit isn't shrinking like $1/\sqrt{n}$.

That's because we're not doing what classical theory thinks we are.

It's good to have rules of thumb, but there's got to be something behind them.
To understand what's happening, we need theory that matches what we actually do.

Modern Asymptotic Theory

High dimensional theory explains what happens when
our models grow in complexity as our sample size does.

- When our models grow slowly enough, behavior is more or less classical.
- When we follow the rule of thumb, using $p \propto n$ parameters, it's very different.
- Sometimes we use models that are even larger, with $p \gg n$ parameters.
  - This is common in genomics. The genome is just big.
  - Even with the simplest models, e.g. linear ones, you've got $p \gg n$.

Nonparametric theory explains what happens when
we use models that have infinitely many parameters.

- Some of these are easy to think about. Much easier than high order polynomials.
- And they might actually be right, not just good enough at a given sample size.
- That means we don't necessarily have to grow them.
  - Some people like to think of them as models that grow on their own.
  - Often, they think of them as having one parameter per observation.

These two theoretical frameworks aren't much different. The concepts are the same.
We'll talk about both, but focus more on nonparametric stuff.

Classical vs. Modern Theory

Textbook Classical Theory

- There's a recipe that works almost all the time: Maximum Likelihood.
- Everything behaves the same. And inference is often easy.
  - $1/\sqrt{n}$ rates of convergence
  - Negligible bias.
- To estimate summaries like the ATE, we just plug in the MLE. It's optimal.
- Optimal choices depend on uninteresting features of the data, e.g. how the noise is distributed.

Modern Theory

- There's no universal recipe. Usually, some stuff works and some doesn't.
- Things behave very differently. And inference is usually hard.
  - Slow rates of convergence.
  - A real bias/variance trade-off.
- Alternatives to the plug-in (e.g. AIPW, DML) are more reliable.
- We rarely rely on uninteresting features of the data. Estimators that do are considered brittle.

In short, modern theory is more complicated.
But, as you might imagine, reality is complicated.

Context

- The history of this stuff is all about computers. They changed everything.
- Modern theory took off at more or less the same time, around the 60s.
  - The models people used got more complex. And there were more of them.
  - Making sense of it all without a general and accurate theory got too hard.
  - And sample sizes got bigger, so asymptotic approximations could be very accurate.
- By now, it's pretty coherent and it's very precise and accurate in a lot of cases.
- And there are even starting to be good (Ph.D.-level) textbooks.
- But there are a few mysteries left, like what's going on with huge neural nets.

From Parametric Thinking to Nonparametric Thinking

The conceptual shift

We still fit nonparametric models via least squares.

- But we don't think about it as minimizing over the parameters of a model

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - (b_0 + b_1 X_{i1} + \ldots) \}^2$$

- We think about it as minimizing over functions in a set.

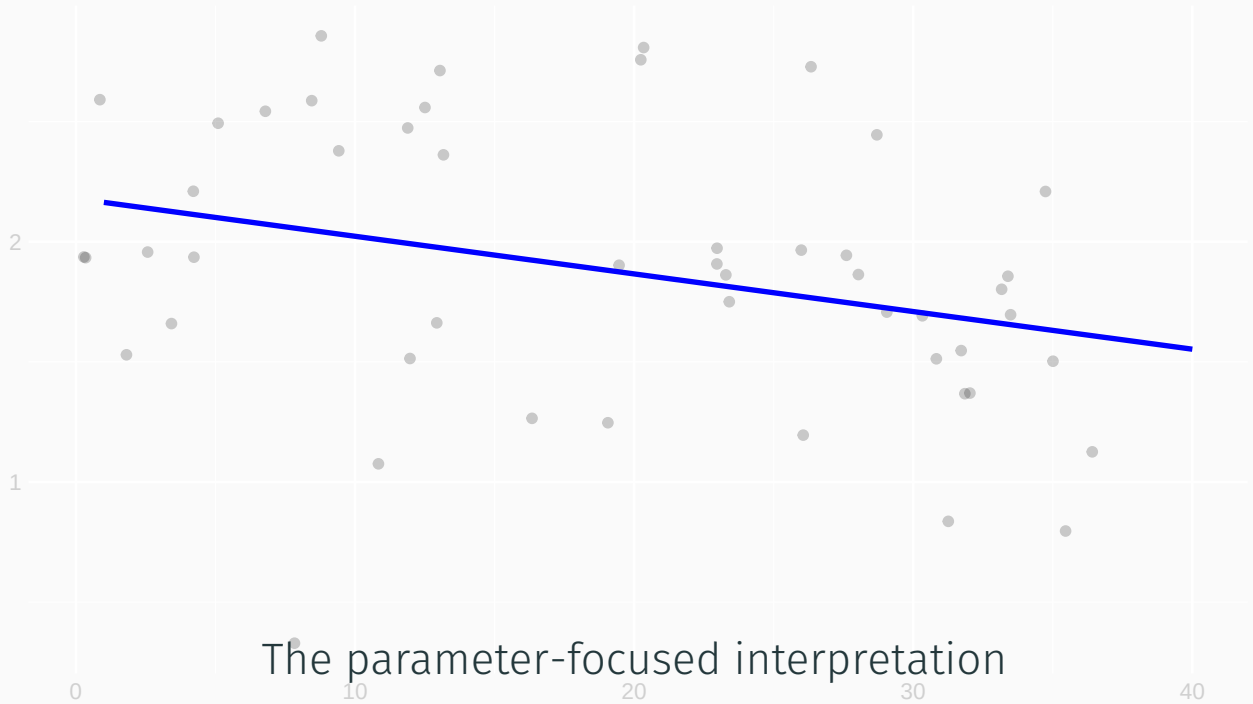
$$\hat{\mu} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m(X_i) \}^2$$

This is just a different way of thinking.

- We can do the same stuff either way.
- But thinking about functions in sets is easier once you get used to it.

As a warm-up, let's reinterpret linear regression.

Linear Regression



The parameter-focused interpretation

$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{where} \quad \hat{\beta} = \underset{b \in \mathbb{R}^2}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - (b_0 + b_1 X_i) \}^2.$$

The function-focused interpretation

$$\hat{\mu}(x) = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m(X_i) \}^2 \quad \text{where} \quad \mathcal{M} = \{ m(x) = b_0 + b_1 x : \beta \in \mathbb{R}^2 \}.$$

Obviously it's the same line either way.
But our interpretation has implications for what we do with it.

What we do with it

Least squares is a tool for estimating Y's conditional mean.

$$\mu(x) = \mathrm{E}[\, Y \mid X = x]$$

If we think we're estimating its slope and intercept,
we might talk about a confidence interval for the slope.

$$\beta_1 \in \hat{\beta}_1 \pm 2\hat{\sigma} \quad \text{with probability} \quad .95$$

This leads to a bit of desperation to think the curve μ *is* a line.

- Otherwise, it doesn't have a slope. Your confidence statement makes no sense.
- You can talk instead about the slope of the best linear approximation to μ.
- That's the agnostic approach. But it's not obvious why you'd care about that.

What we do with it

If we think we're estimating the whole curve μ, we'll talk about distances between the estimate and the curve itself.

$$\int_0^1 \{\hat{\mu}(x) - \mu(x)\}^2 \leq d_n^2$$

Or maybe confidence intervals for summaries of the curve, like its *average slope*.

$$\int_0^1 \mu'(x) \in \int_0^1 \hat{\mu}'(x) \pm 2\hat{\sigma} \quad \text{with probability} \quad .95$$

This you can do without any problem. The curve μ is, after all, a curve.

And we can do this after fitting any model. Not just lines.

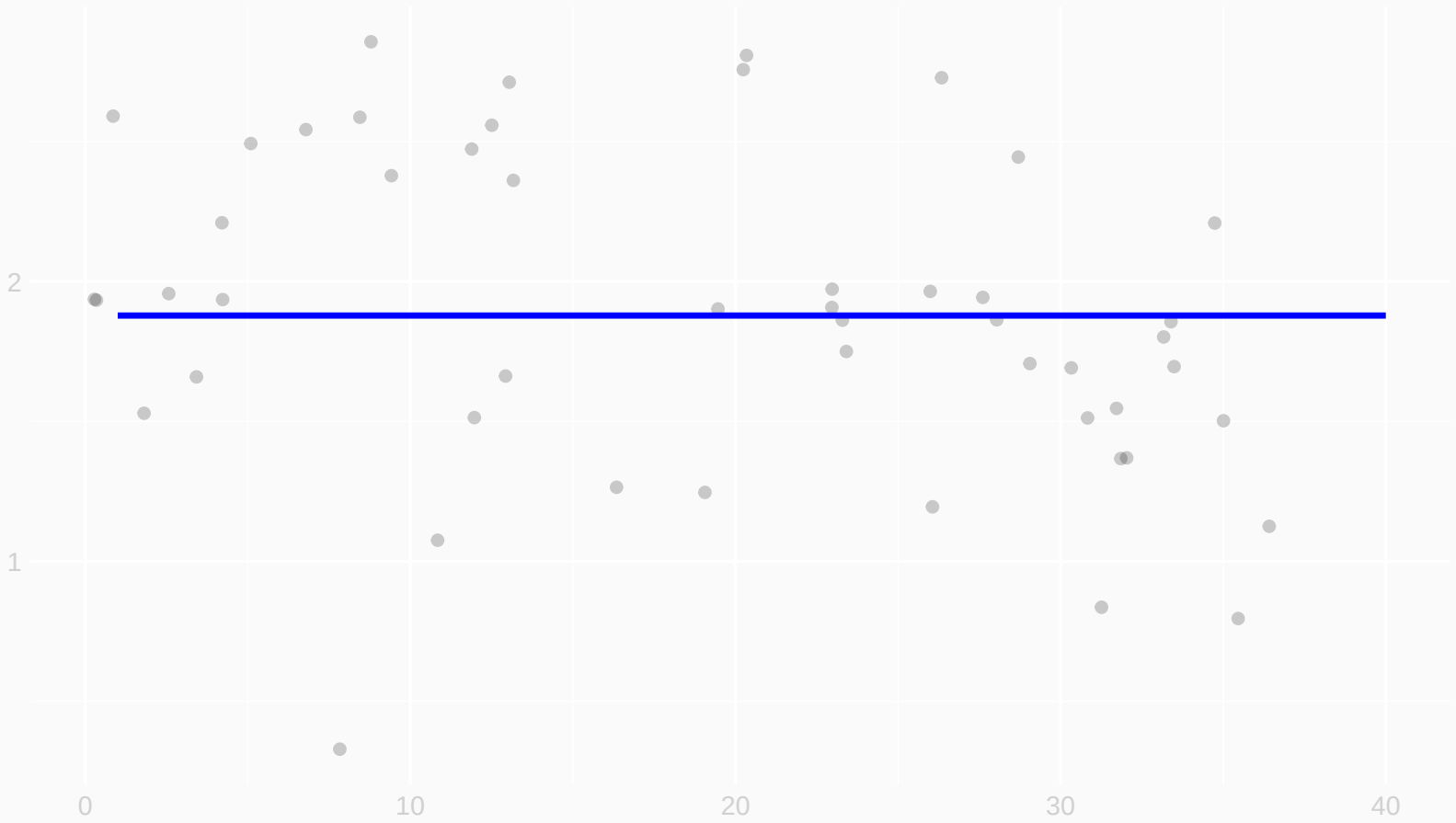
Let’s look at some others models.

We’ll look at examples of two types of model today.

1. Models in which we constrain the *overall shape* of the curve.
2. Models in which we constrain the *wiggliness* of the curve.

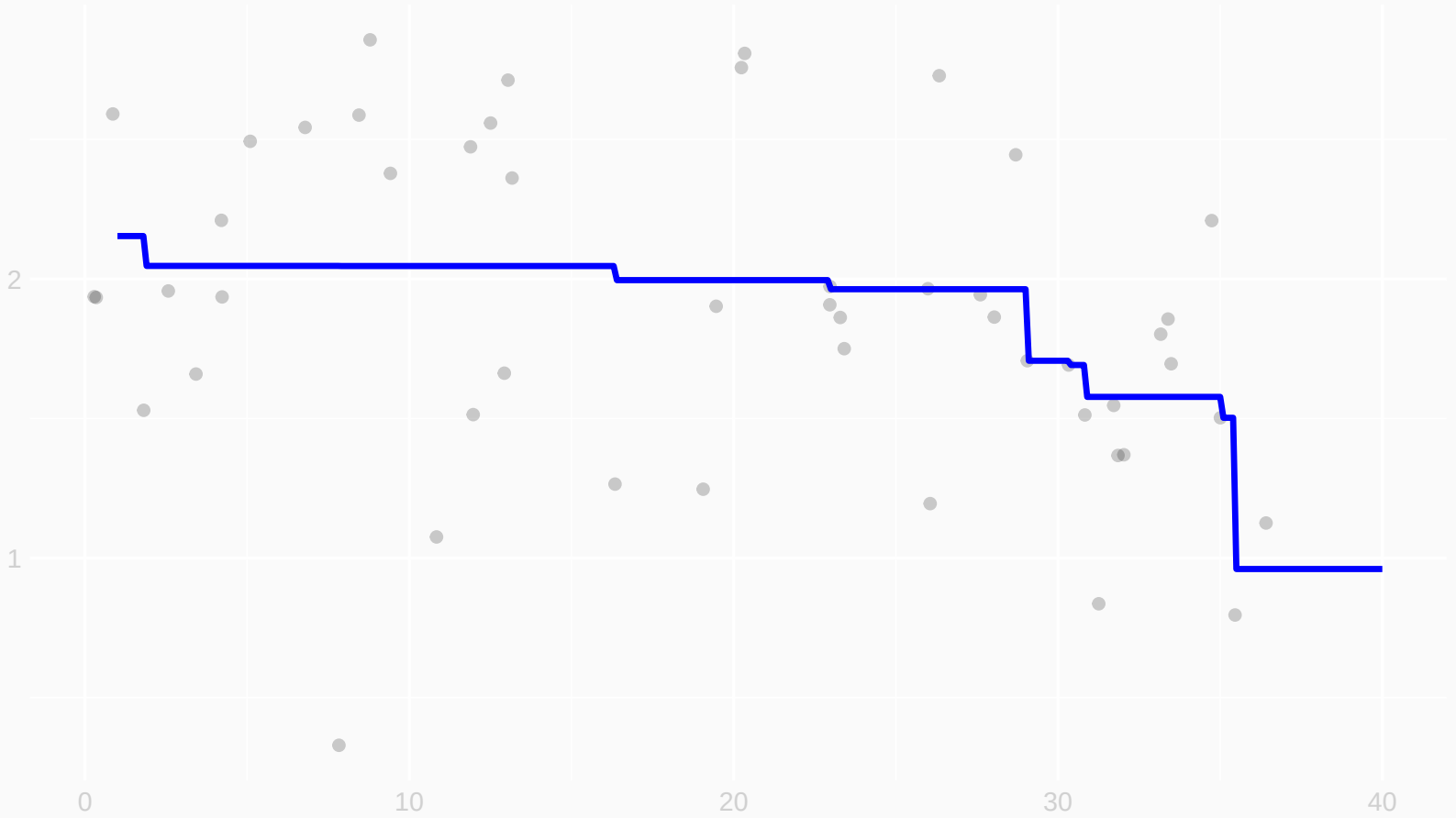
Regression with Shape and Smoothness Constraints

An increasing curve



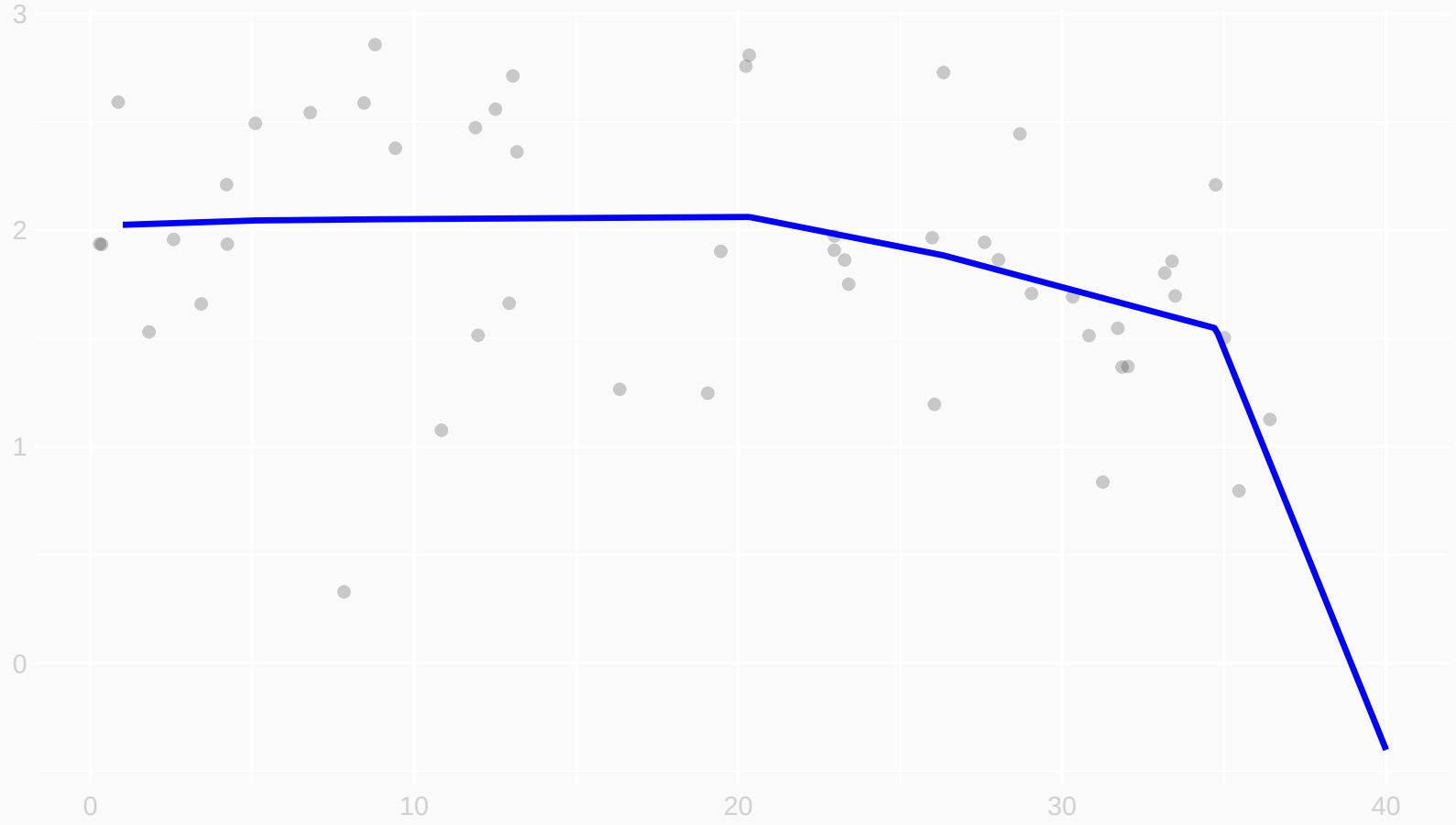
$$\hat{\mu}(x) = \underset{\text{increasing } m}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{ Y_i - m(X_i)\}^2.$$

A decreasing curve



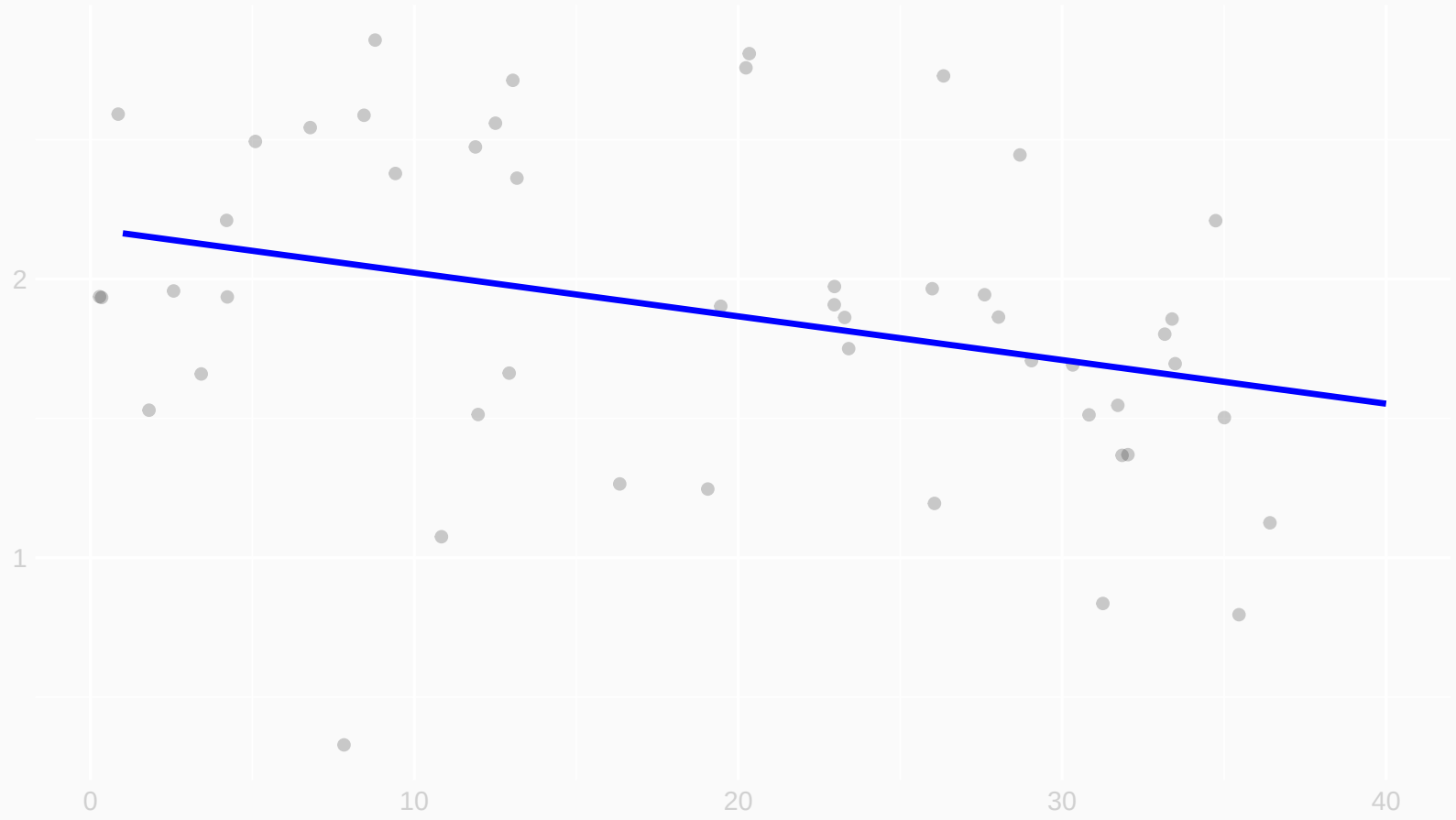
$$\hat{\mu}(x) = \underset{\text{decreasing } m}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m(X_i) \}^2.$$

A concave curve



$$\hat{\mu}(x) = \underset{\substack{m \text{ with} \\ m' \text{ decreasing}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m(X_i) \}^2.$$

A convex curve



$$\hat{\mu}(x) = \underset{\substack{m \text{ with} \\ m' \text{ increasing}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m(X_i) \}^2.$$

Thoughts

- It takes a minute to get used to the idea that you can actually fit these models.
- But it's easy to understand what happens when you do.
- If the shape is right, they fit. If not, they don't.
  - Or they fit part of the data where the shape is right.
  - You don't always have to fit stuff everywhere.
- No fussy tuning parameters to choose, e.g. degree of polynomial, etc.
- And you can choose all kinds of shapes. More options:
  - Bowl-shaped: decreasing then increasing.
  - Mound-shaped: the opposite.
  - S shaped: decreasing then increasing then decreasing.

  And that's just 1D curves. There's more to it in dimension > 1.

The Nonparametric Idea of Model Complexity

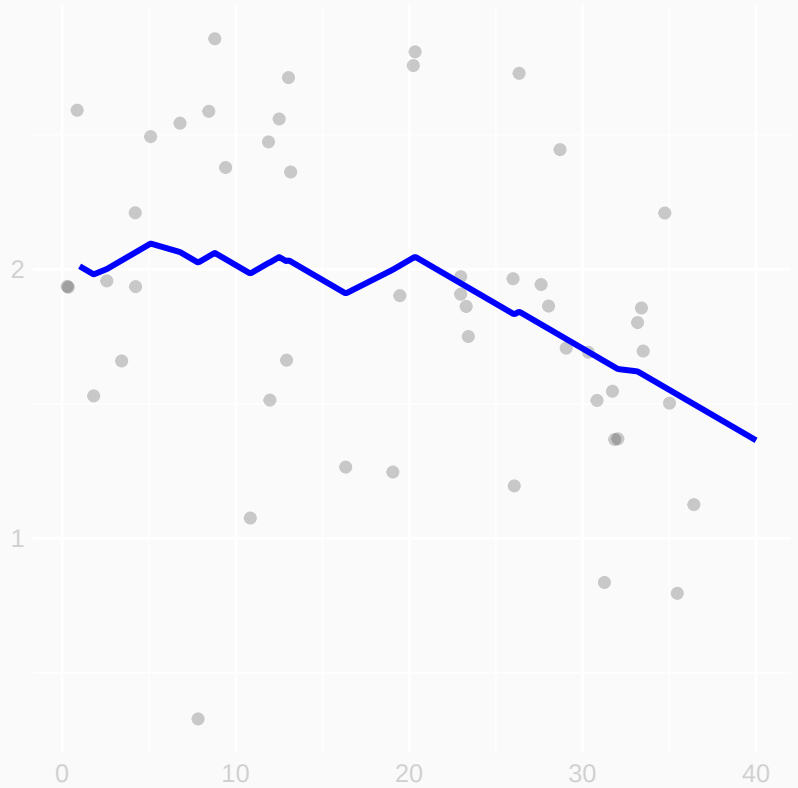
- You can get used to people counting parameters to think about overfitting.
  - We've got our 20 observations/parameter rule of thumb.
  - There are precise model selection criteria that do it, too: AIC, BIC, Adjusted R^2.
- That doesn't work for these models. They have infinitely many parameters.
  - I can't tell you which increasing curve I'm thinking of by writing down some numbers.
- To understand overfitting, we have to go back to first principles.
  - We'll overfit (i.e. fit noise) if there is a curve in your model that can. That's it.
  - A model's ability to fit noise is exactly what determines how fast fits converge.
- These shape-constrained models clearly can't fit noise all that well.
  - Noise jitters. It goes up and down and up again.
  - These go up. Or down. Or maybe up then down.

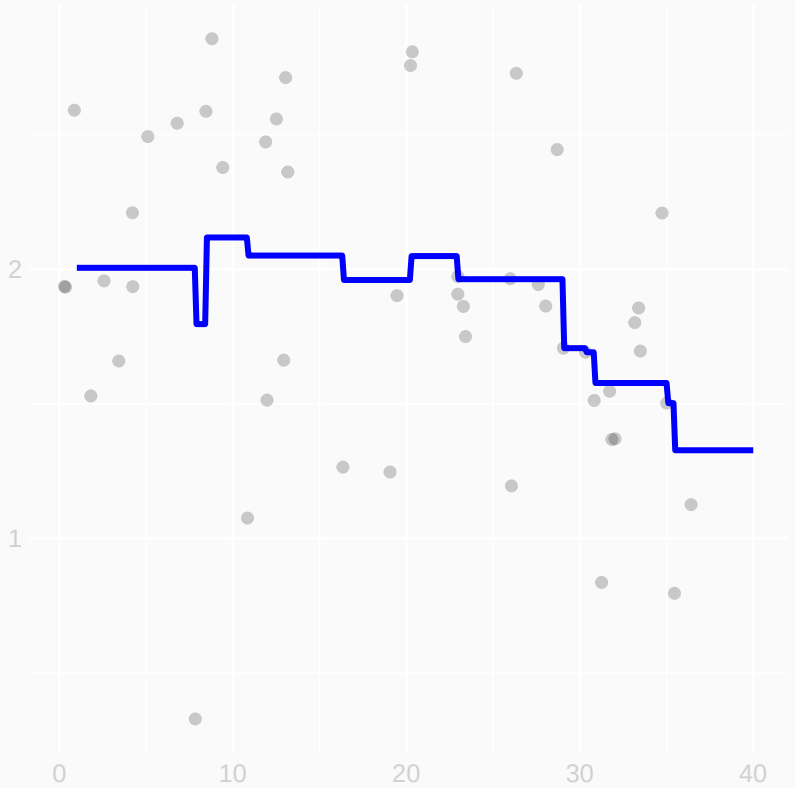
Try it!

Fit this noise by drawing a curve with a simple overall shape.
Increasing, decreasing, convex, concave, bowl-shaped, S-shaped, whatever.

Smoothness Constraints

- Sometimes, we don't know the overall shape of the curve.
- To limit its complexity another way, we use a smoothness constraint.
- That is, we say the curve doesn't wiggle too much or too fast.





$$\hat{\mu}(x) = \underset{\max|m'(x)|\leq B}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2; \quad \hat{\mu}(x) = \underset{\int_0^1|m'(x)|\leq B}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2.$$

About this class

Outline

It's two roughly equal halves.

1. Methods for fitting curves to data.
    - We'll talk about models based on shape and smoothness constraints.
    - And how to implement least squares regression using them.
    - That'll involve writing at lot of convex programs in CVXR.
    - We'll also cover some other stuff, like model selection techniques and the lasso.
2. Theory about how these methods behave.
    - We'll talk about how to understand what a curve fit tells you.
    - Mostly that'll mean what it'll converge to, in what sense, and how fast.
    - That'll help us understand what claims we can justifiably make based on the data.
    - Without this stuff, it's a lot less obvious when you're using a wildly inappropriate method.

We'll also do a few exercises focusing on using this stuff for causal inference.

Check out the syllabus for more detail.

Format

- It'll be about half lecture and half lab. No exams.
- Instead, we'll set aside a few days for review and discussion.
- We'll have problem sets for homework most weeks.
- And a few short essays to reflecting on how the class is going for you.
- These are meant to help us prepare for our review/discussion days.

Next Class

- Bring your laptops.
- We'll write code to fit increasing curves to data.
- This isn't strictly a coding exercise.
- Your computer can't optimize over the set of all increasing curves by itself.
    - You've got to translate stuff into terms it can understand.
    - Then translate what it says back into the terms you want it in.
    - You'll see. It'll be fun.

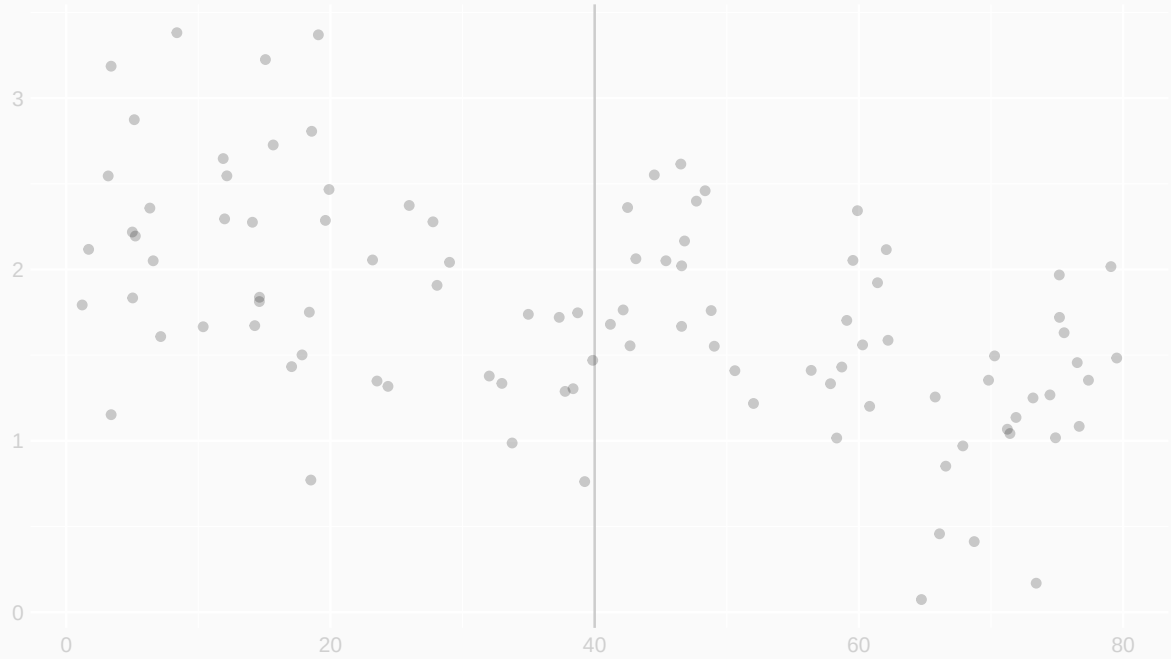
Homework

- For homework, we'll do a warm-up exercise.
- We'll use the package we'll be using in lab, **CVXR**, to fit lines.
- Hopefully this'll make the programming part of the lab a bit quicker.
- It'll also ensure you've got the libraries you need for lab installed.
- If you're having trouble with that, email me. I'll help you sort it out.

Actually using this stuff in Regression Discontinuity Designs

A Fictionalized RDD Example

- Question. Are smaller classes better for 5th graders?
- Data. A state caps class sizes at 40.
  - When there are $x \leq 40$ 5th graders enrolled in a school, they run one class of size x.
  - When there are $x > 40$, they run two classes of average size $x/2$.

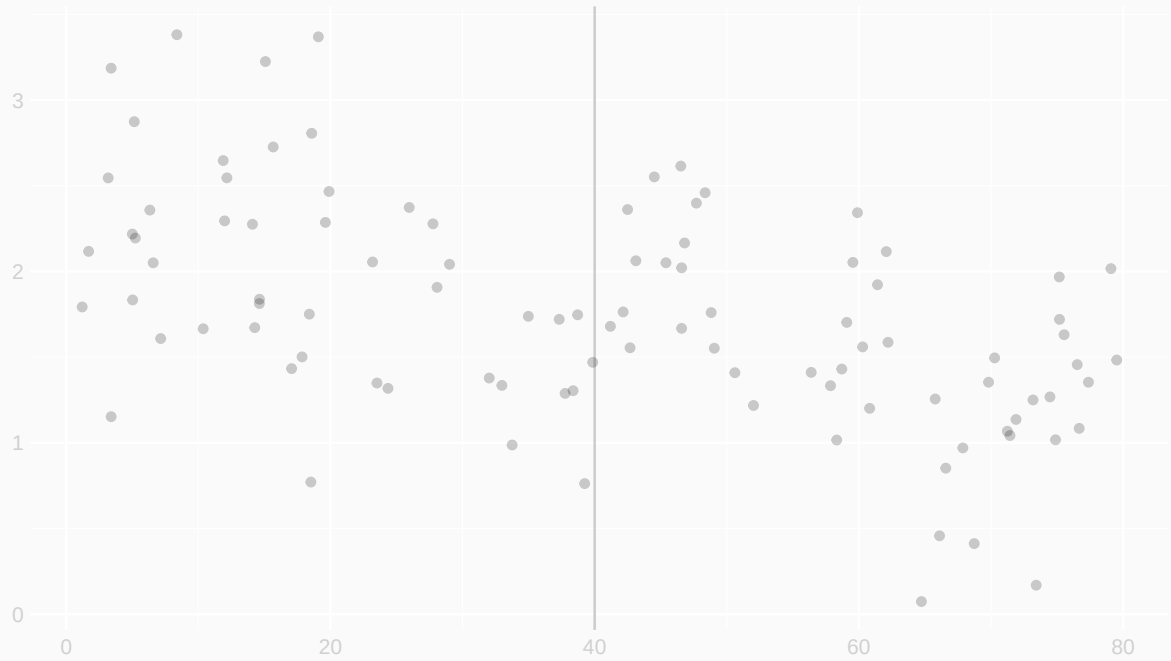


This is a fake-data version of a study of Angrist and Lavy [1999]. The state is Israel.

- It has been simplified to make our discussion easier.
- Real schools sometimes had more than 80 5th-graders enrolled.
- And they didn't follow this cap perfectly.

A Fictionalized RDD Example

- Question. Are smaller classes better for 5th graders?
- Data. A state caps class sizes at 40.
  - When there are $x \leq 40$ 5th graders enrolled in a school, they run one class of size x.
  - When there are $x > 40$, they run two classes of average size $x/2$.



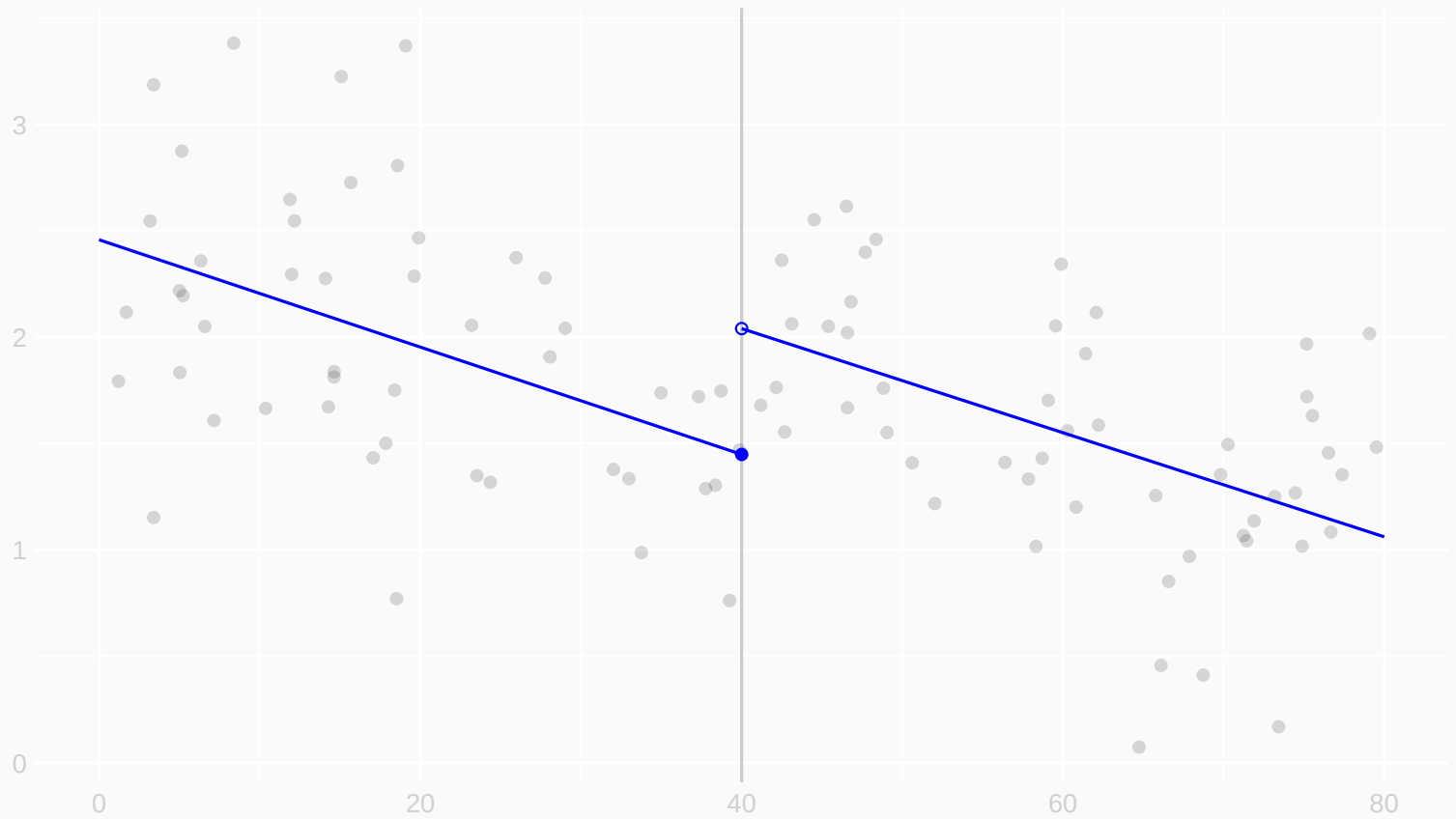
We can use this to estimate the average effect of having 20 vs. 40 students/class.

$$\text{effect} = \mu(40+) - \mu(40-) \qquad \text{where}$$

$$\mu(x) = \mathrm{E}[\text{avg. test score}_i \mid \text{enrolled 5th graders}_i = x].$$

All we have to do is estimate $\mu(x)$ just to the left and to the right of 40.

The Simple Approach

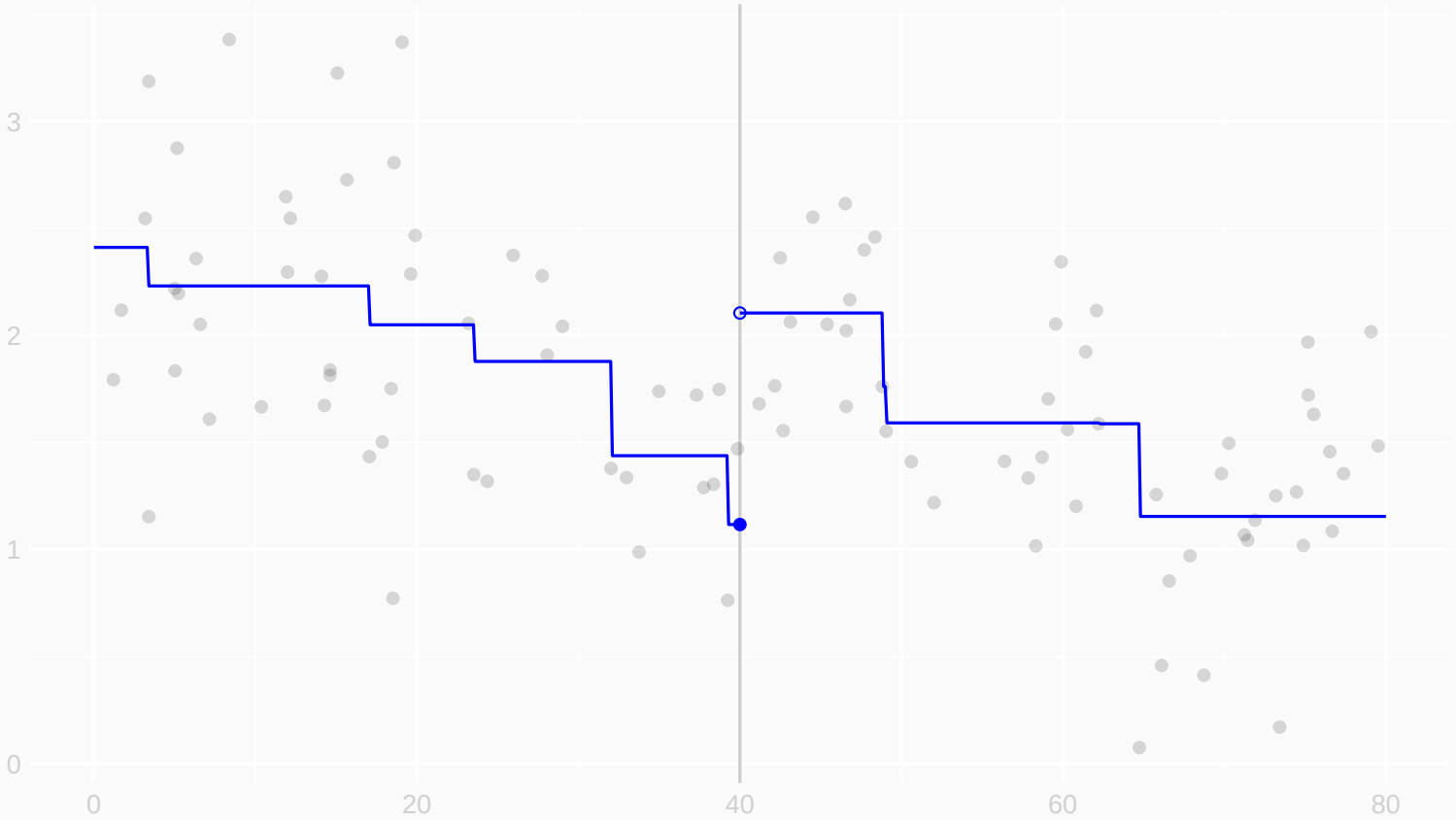


$$\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 0.59$$

where

$$\hat{\mu}_{\text{left}} = \underset{m \in \mathcal{M}}{\text{argmin}} \sum_{i:X_i \leq 40} \{Y_i - m(X_i)\}^2 \quad \text{and} \quad \hat{\mu}_{\text{right}} = \underset{m \in \mathcal{M}}{\text{argmin}} \sum_{i:X_i > 40} \{Y_i - m(X_i)\}^2.$$

The Simple Approach

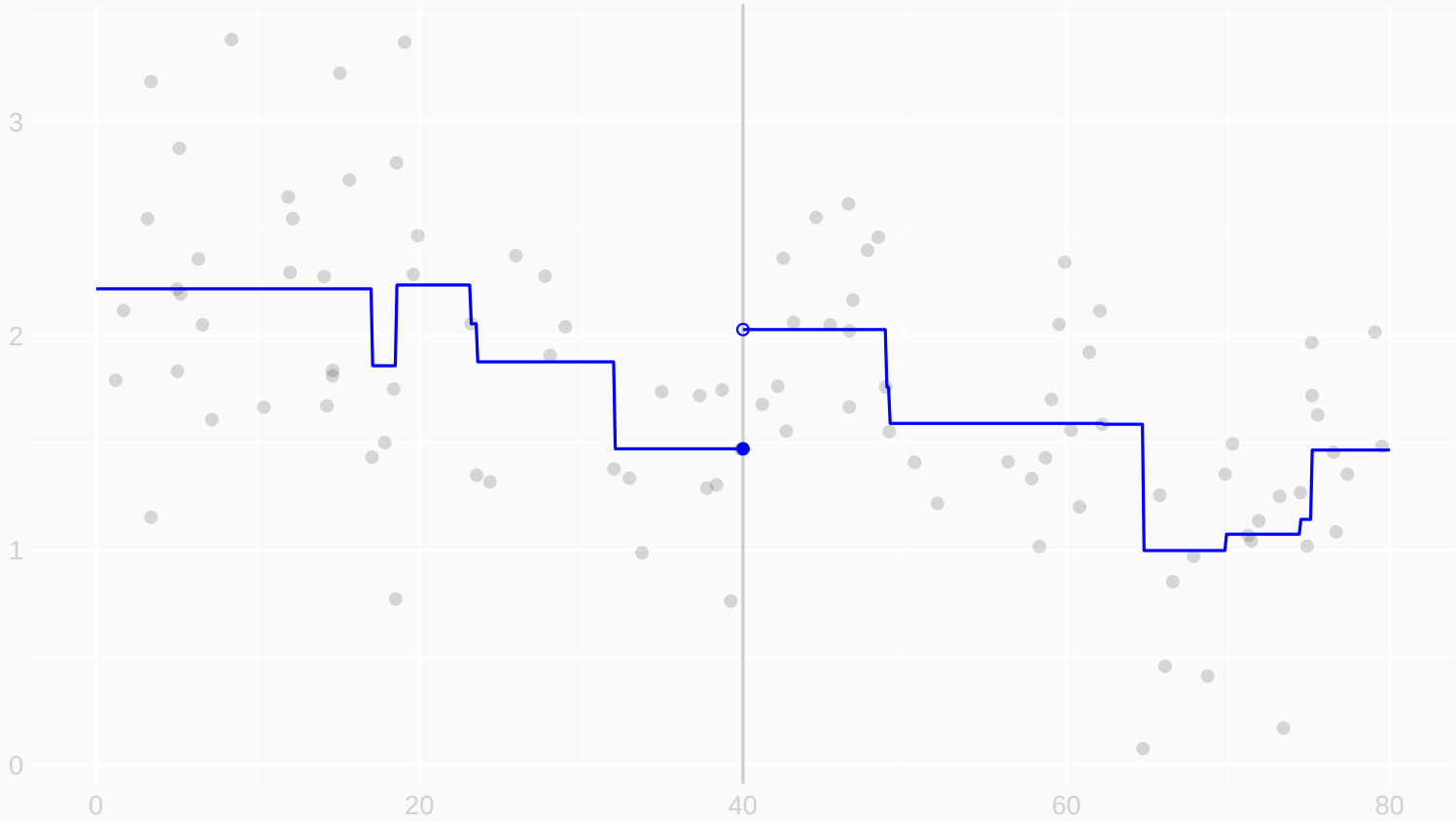


$$\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 0.99$$

where

$$\hat{\mu}_{\text{left}} = \underset{m \in \mathcal{M}}{\text{argmin}} \sum_{i:X_i \leq 40} \{Y_i - m(X_i)\}^2 \quad \text{and} \quad \hat{\mu}_{\text{right}} = \underset{m \in \mathcal{M}}{\text{argmin}} \sum_{i:X_i > 40} \{Y_i - m(X_i)\}^2.$$

The Simple Approach



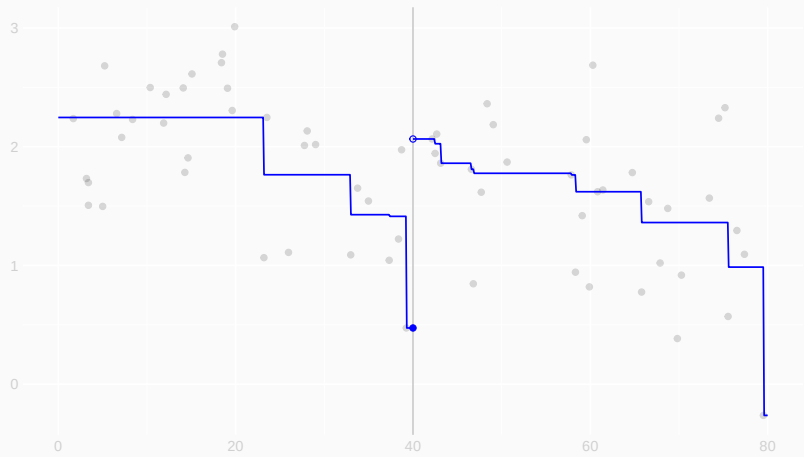
$$\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 0.56$$

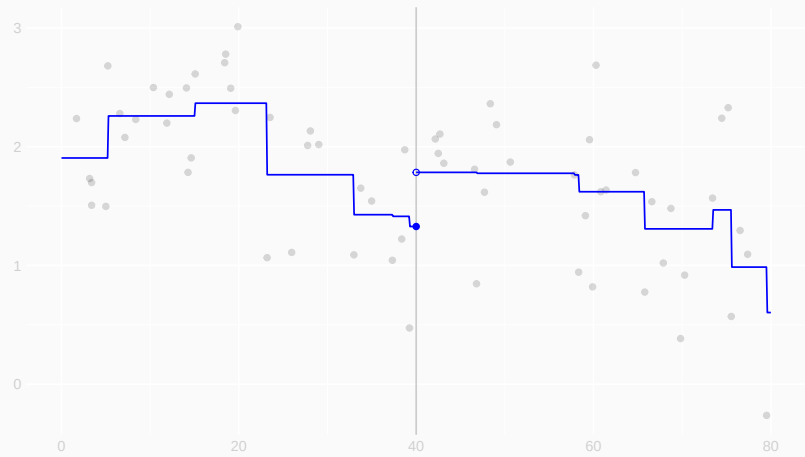
where

$$\hat{\mu}_{\text{left}} = \underset{m \in \mathcal{M}}{\text{argmin}} \sum_{i:X_i \leq 40} \{Y_i - m(X_i)\}^2 \quad \text{and} \quad \hat{\mu}_{\text{right}} = \underset{m \in \mathcal{M}}{\text{argmin}} \sum_{i:X_i > 40} \{Y_i - m(X_i)\}^2.$$

RDD is hard

If we'd used 2/3 of our data, our estimates would differ much more.



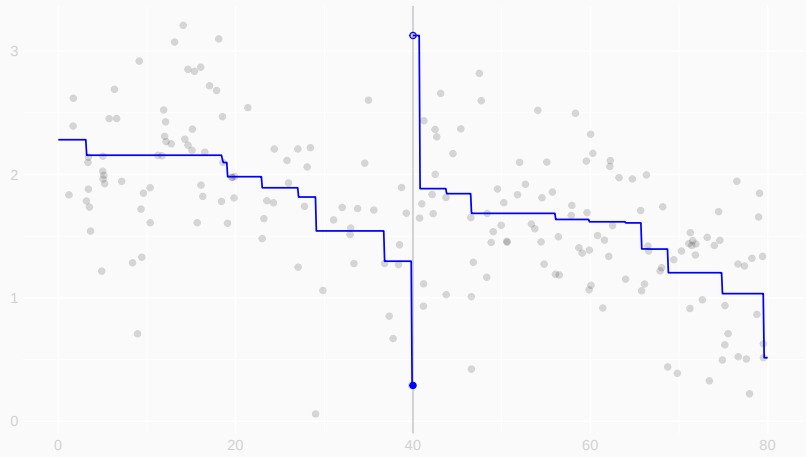


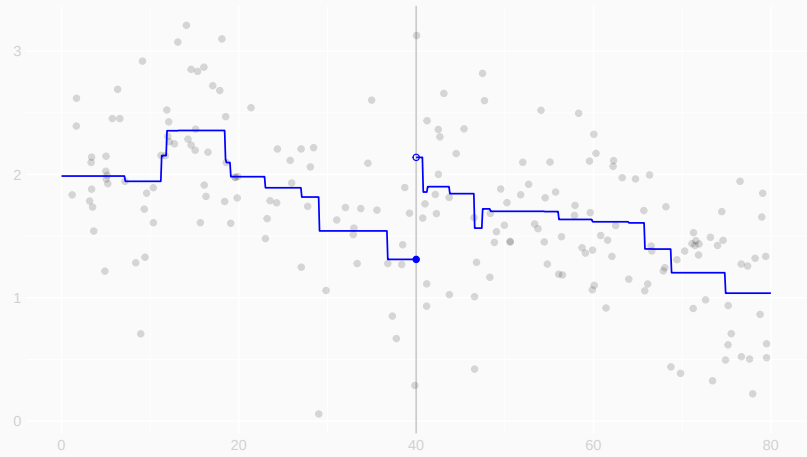
$\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 1.59$ $\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 0.46$

Why is it hard?

RDD is hard

It's even more extreme when we use twice as much data.



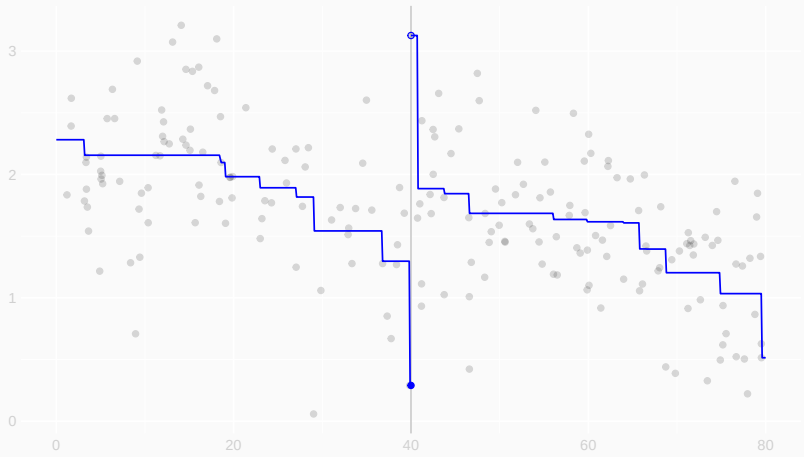


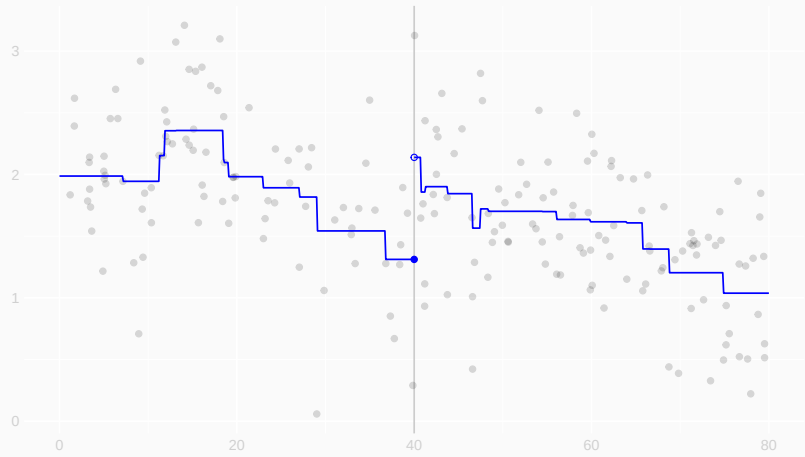
$\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 2.84$ $\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 0.83$

Why is it hard?

RDD is hard

It's even more extreme when we use twice as much data.





$\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 2.84$ $\widehat{\text{effect}} = \hat{\mu}_{\text{right}}(40) - \hat{\mu}_{\text{left}}(40) = 0.83$

Why is it hard?

- We're using our fitted curve at or past the edge of the data.
- It's far more sensitive to our choice of model out there.
- There's lots of debate about how to do it right. Your choices matter.
- It's important to understand where your estimate is really coming from.

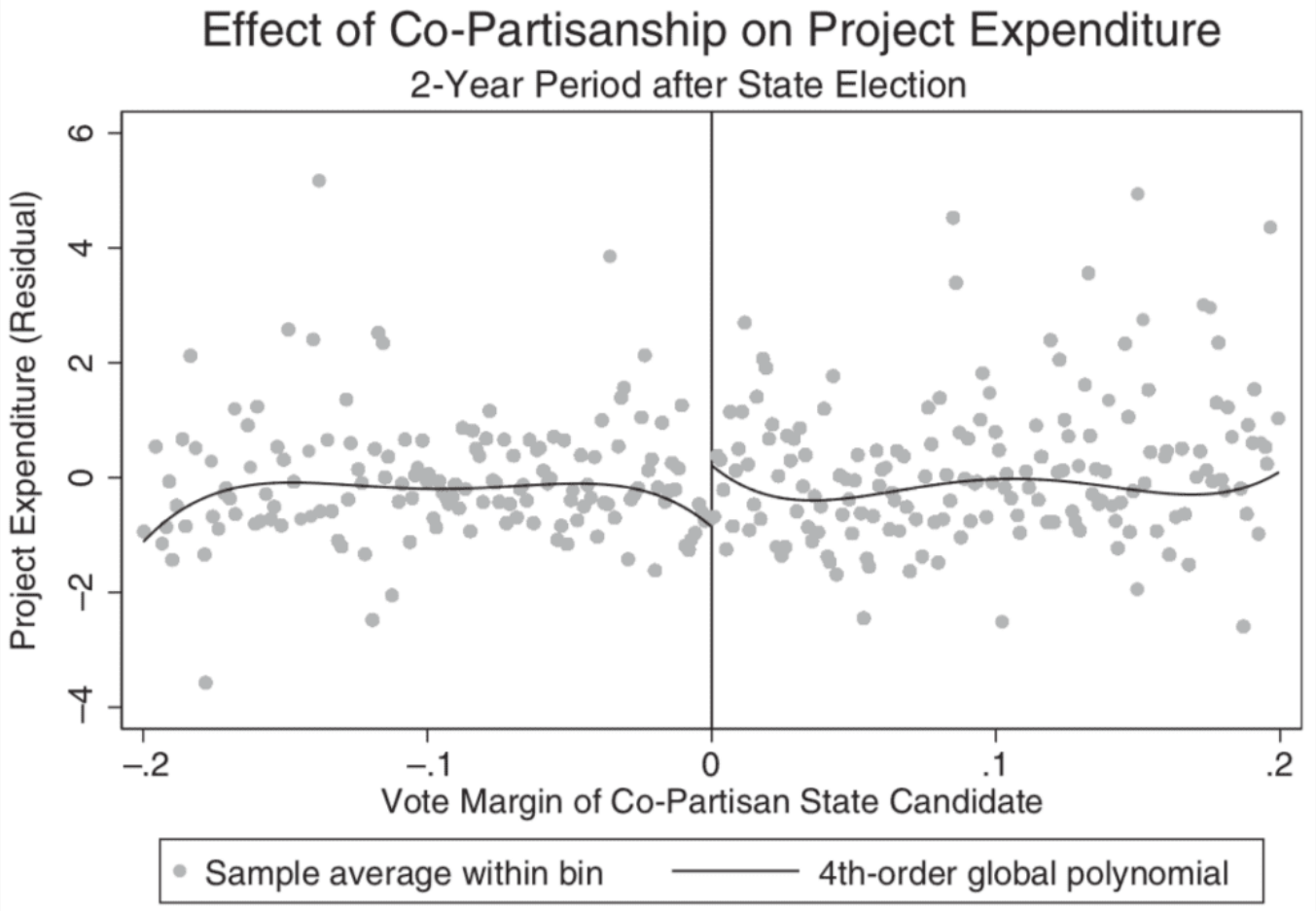
Understanding RDD is going to take everything we'll learn.

- If there's time and interest, we'll come back to it at the end of the semester.
- For now, let's enjoy some bad RDD estimates.

Bad RDD estimates

borrowed from Andrew Gelman's blog

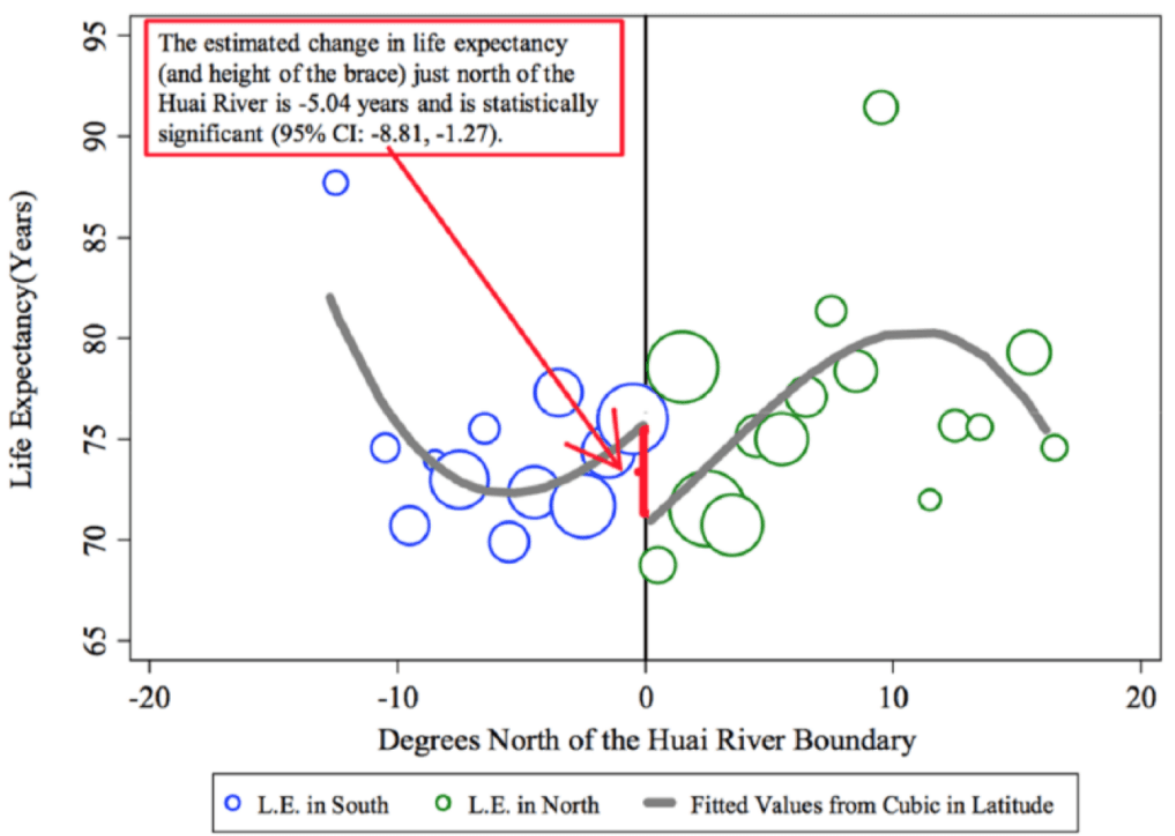
Pork happens in India



Fitting quartic polynomials, someone argued that members of Indian parliament preferentially allocated funds to districts represented by members of their party.

See post on Andrew Gelman's blog.

Coal pollution kills you in China



The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

Fitting cubics polynomials, someone argued that a policy of providing free coal in northern China created enough localized pollution to cost 5 years of life on average.

See post on Andrew Gelman's blog.

Losing elections kills you in the US

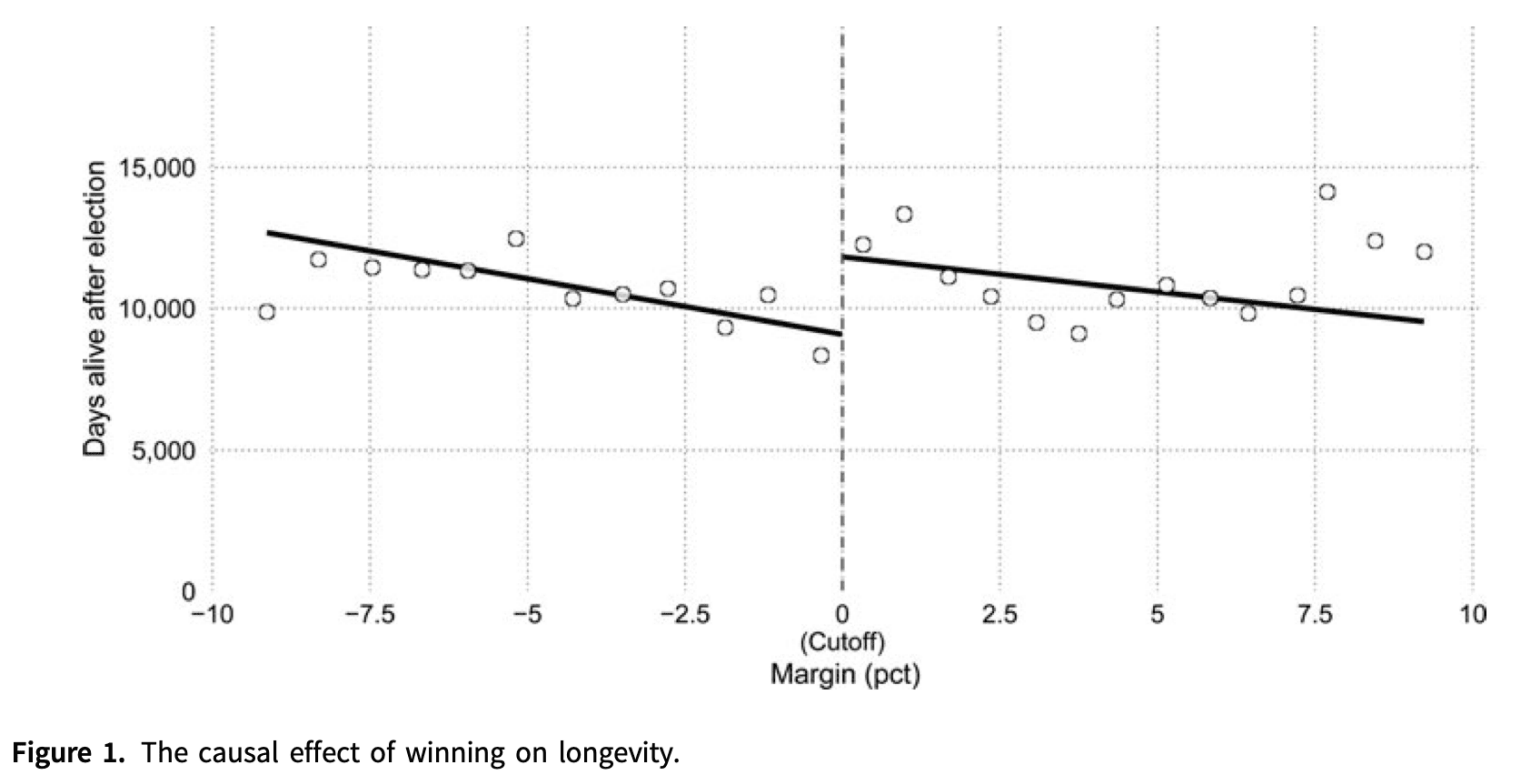


Figure 1. The causal effect of winning on longevity.

Fitting lines, someone argued that among candidates for governor, winning the election somehow netted them 5-10 extra years of life.

See post on Andrew Gelman's blog.

You can do better

It's funny that this happens so often.

- Convention dictates you include these pictures.
- And anyone looking at them can see that there's little or no evidence of an effect.

Advice

1. Definitely make sure you look at the fit on the data.
2. Think very carefully about what the data does tell you and can tell you.

You don't need to take this class to do that.

- But it helps to have the right concepts when you're thinking.
- That's what this class is about.

References

Joshua D Angrist and Victor Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly journal of economics*, 114(2): 533–575, 1999.