Topics in Applied Econometrics for Public Policy

TA Session 3

Sergi Quintana

Barcelona School of Economics

May 2, 2023

Plan for Today

- Nonparametric Series Regression.
- Semiparametric Regression.

Nonparametric Series Regression

As with nonparametric kernel regressoin, the objective is to estimate the conditonal mean of an outcome given the covariates without making assumptions on its functional form.

The difference is that now we will use a collection of terms that apporximate the function, known as basis. The most used are:

- Polynomials.
- Splines.
- B-splines.

Nonparametric Series Regression

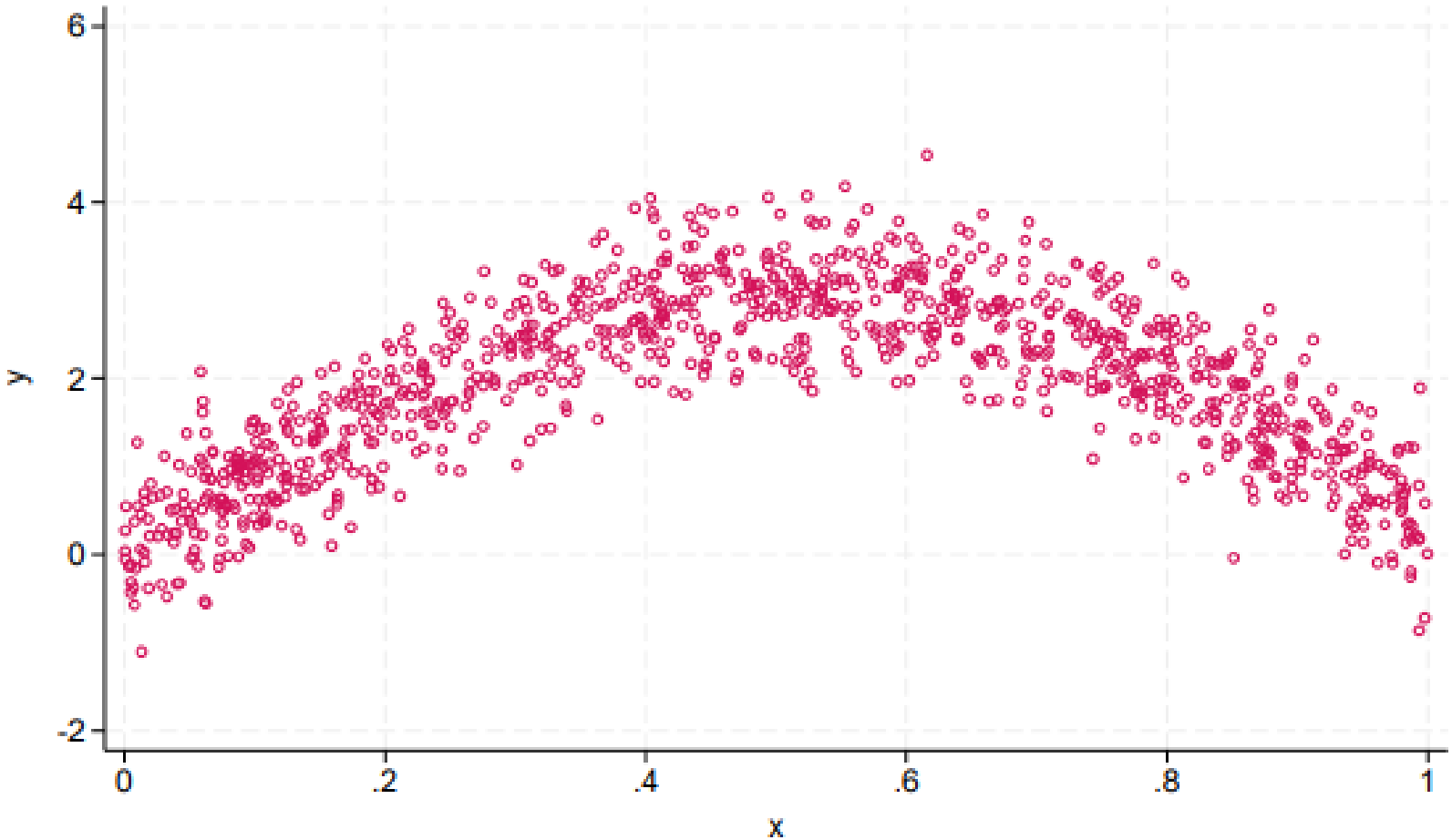
The main two concepts are the **basis** and the **basis function**.

- **Basis**: collection of terms that can approximate a smooth function arbitrarily well.
- **Basis function**: uses a subset of these terms to approximate the mean function.

For each chocie of basis (polynomial,splines,B-splines,...) the Stata function *npregress series* will select the basis function for us.

Nonparametric Series Regression

Suppose we want to approximate the conditional mean function of the following data:



Nonparametric Series Regression

We could choose the polynomial basis. Different basis functions are:

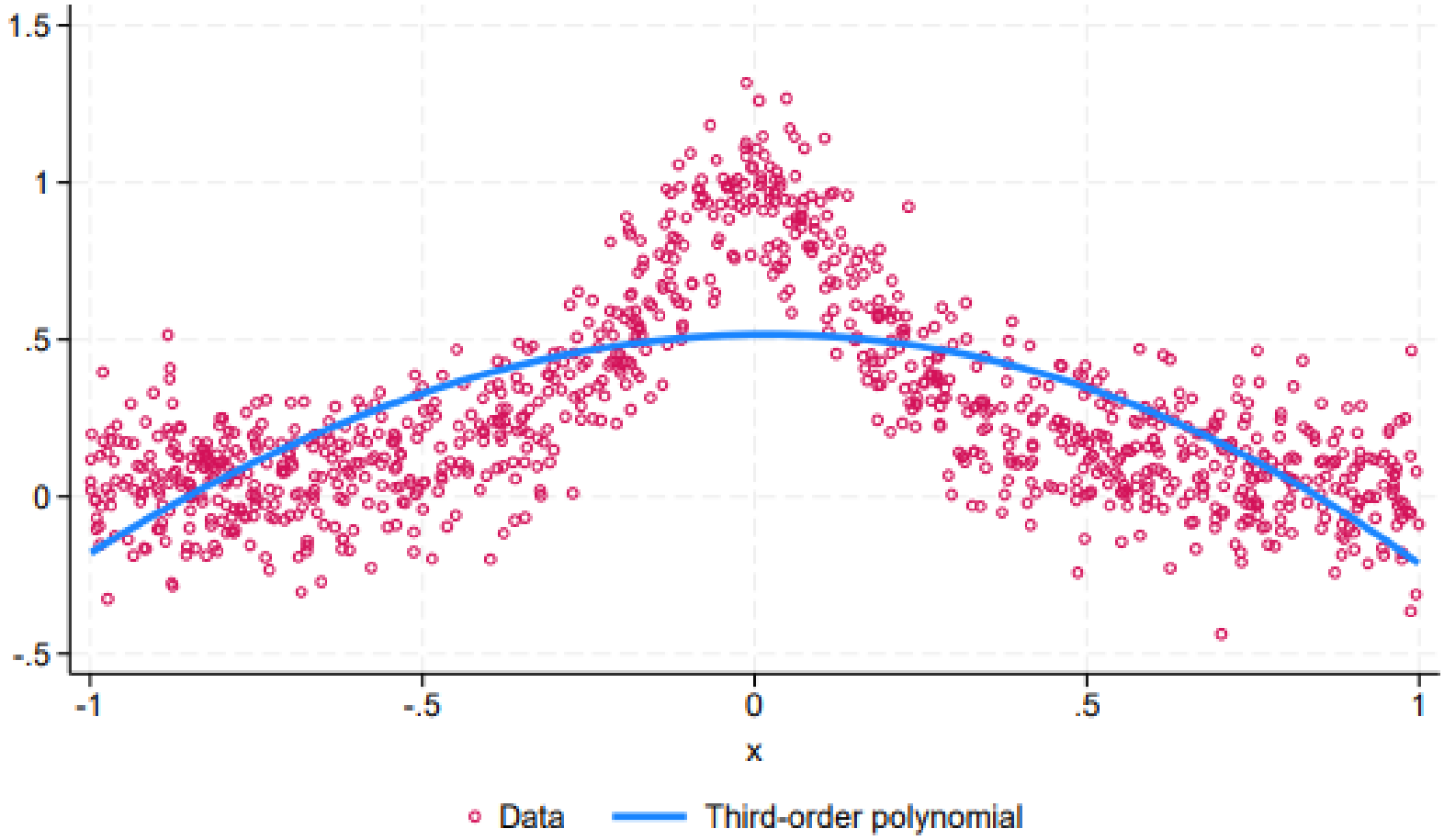
- A linear polynomial, with a constant and x.
- A quadratic polynomial, with a constant, x and x^2.
- More complex polynomials will require a basis function that includes more terms from the polynomial basis.

In Stata, *npregress series* will select the basis function for us if we choose the polynomial basis.

It will be selected optimally, optimizing the tradeoff of bias and variance.

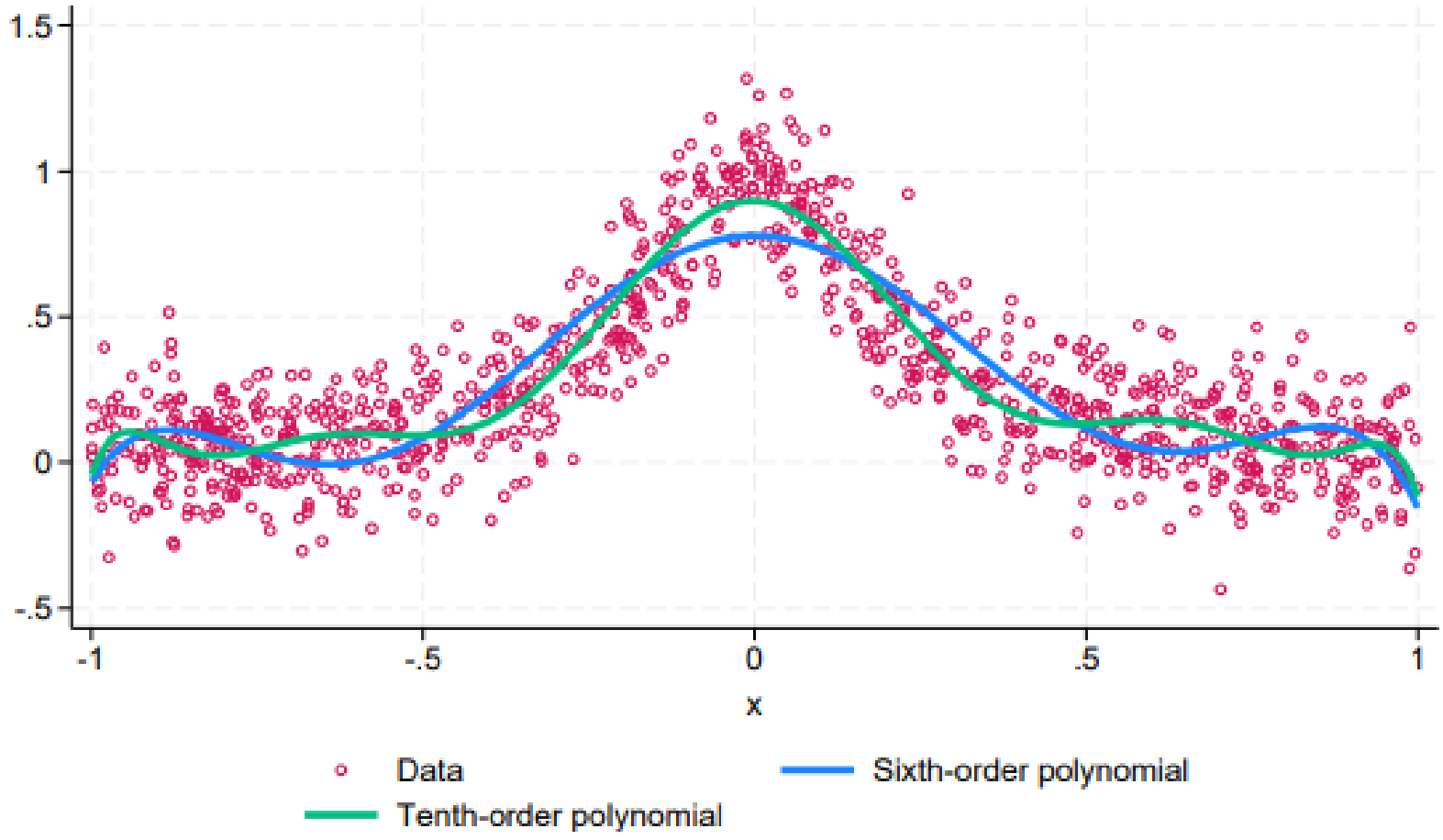
Nonparametric Series Regression. Runge's Phenomenon

Polynomials are the most intuitive basis but not the preferred.
They are bad at interpolating.



Nonparametric Series Regression. Runge's Phenomenon

From the previous graph it looks that we need a higher order polynomial.



Nonparametric Series Regression. Runge's Phenomenon

The prediction improves in the middle but becomes more variable at the edges.

This is because at the boundaries of the support of the covariates as you increase the order of the polynomial, the polynomial approximation oscillates frequently, even when the true function does not behave this way.

This is known as **Runge's phenomenon**. Increasing the complexity of the polynomial does not improve the approximation.

Nonparametric Series Regression. Splines.

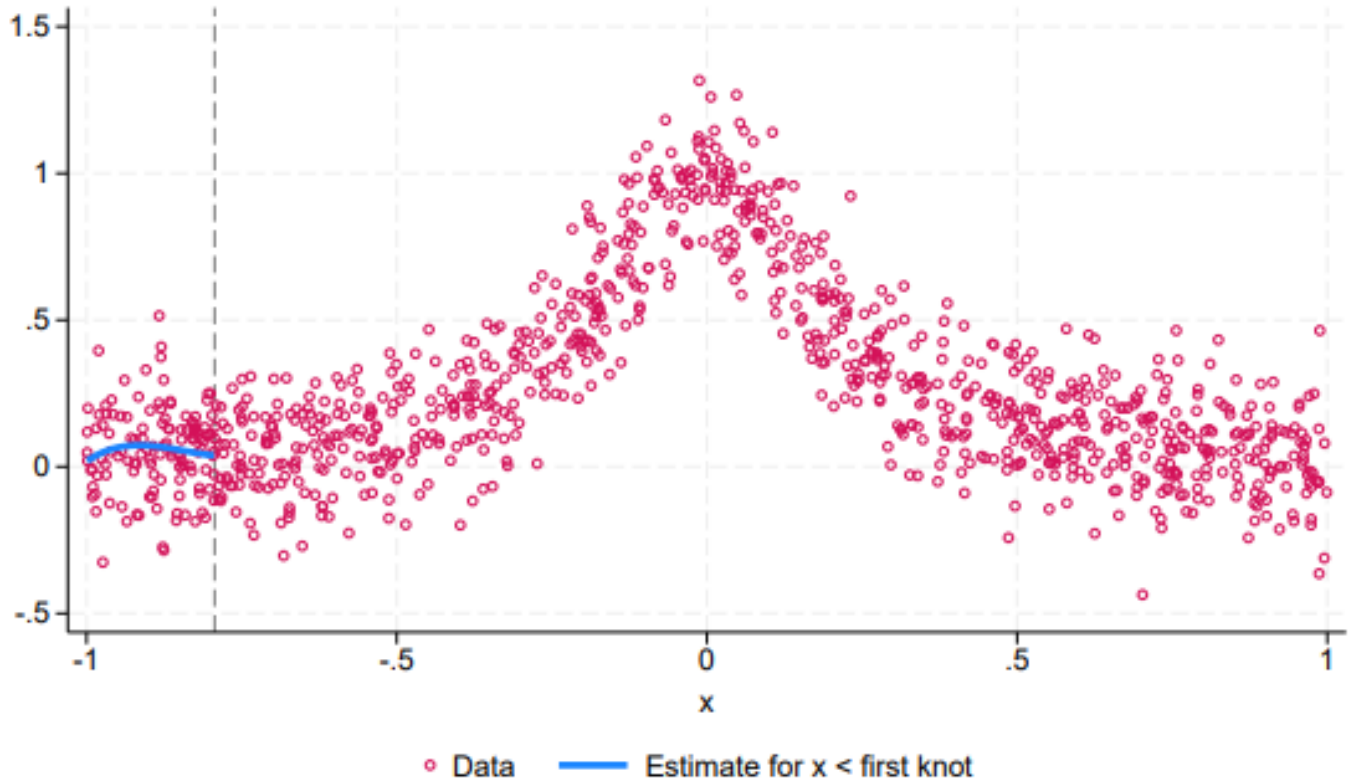
- Splines and B-splines solve the interpolating problem.
- Splines continuously connect a set of low-order polynomials to create a basis to approximate a smooth function.
- The idea is to break the support of x into subregions.
- Now we can allow for a different low-order polynomial in each subregion, so that it correctly approximates the data at that subregion.
- Furthermore, we can force the polynomials in neighboring regions to be continuously connected.
- The subregion boundaries are known as the knot points because they are where the different polynomials are tied together.

Nonparametric Series Regression. Splines

We could for example fit a third order polynomial at a specific subregion:

$$\hat{\mathbb{E}}(y_i|x_i \leq t_1) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3$$

where t_1 is the knot denoting the boundary of the subregion.

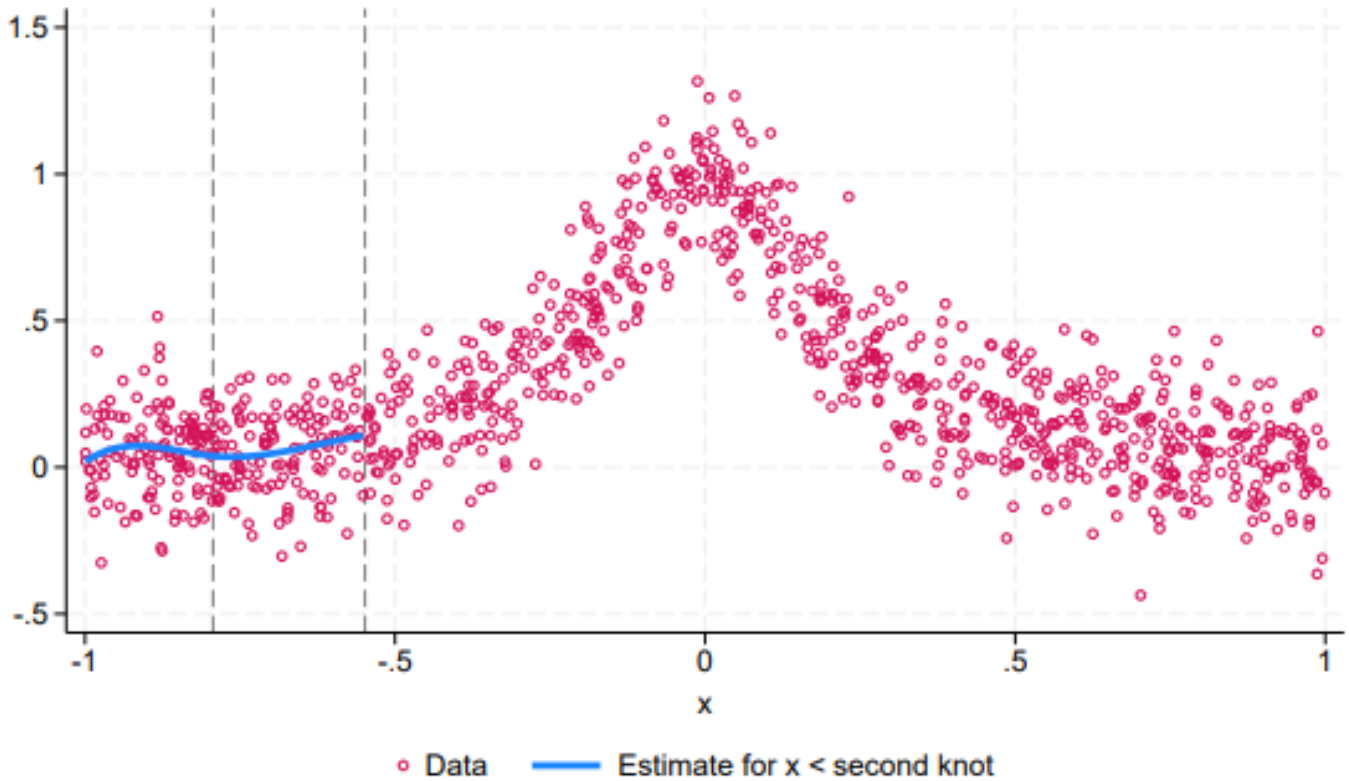


Nonparametric Series Regression. Splines

We can continue to draw the blue line by fitting a polynomial to the next subregion, and forcing the polynomials to be connected:

$$\hat{\mathbb{E}}(y_i|x_i \leq t_2) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3 + \hat{\beta}_4 (x_i - t_1)^3 (x_i > t_1)$$

Graphically:

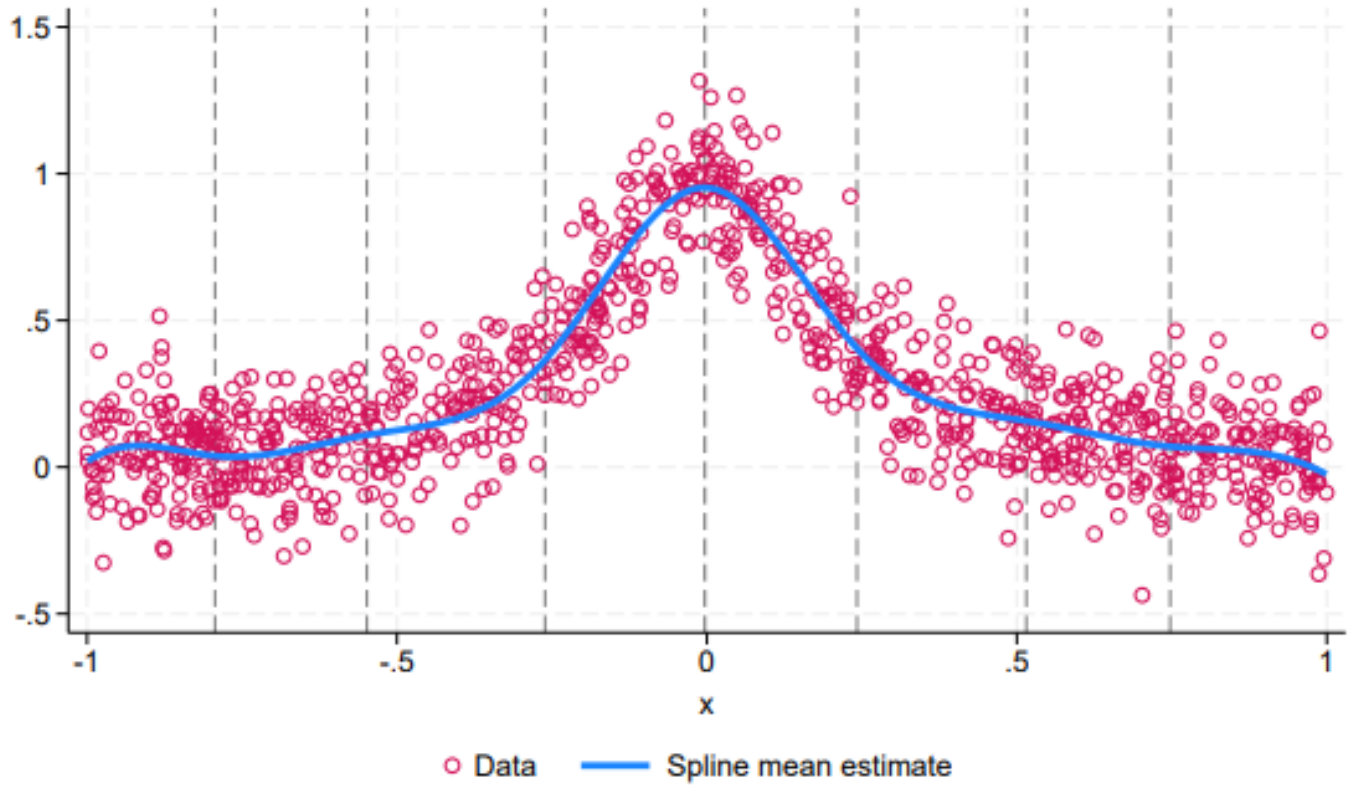


Nonparametric Series Regression. Splines

If we continue for all the support of x, we get:

$$\hat{\mathbb{E}}(y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3 + \sum_{j=1}^{J} \beta_{j+3} \max(x_i - t_j, 0)^3$$

where J is the total amount of knots. Graphically:



Nonparametric Series Regression. Splines

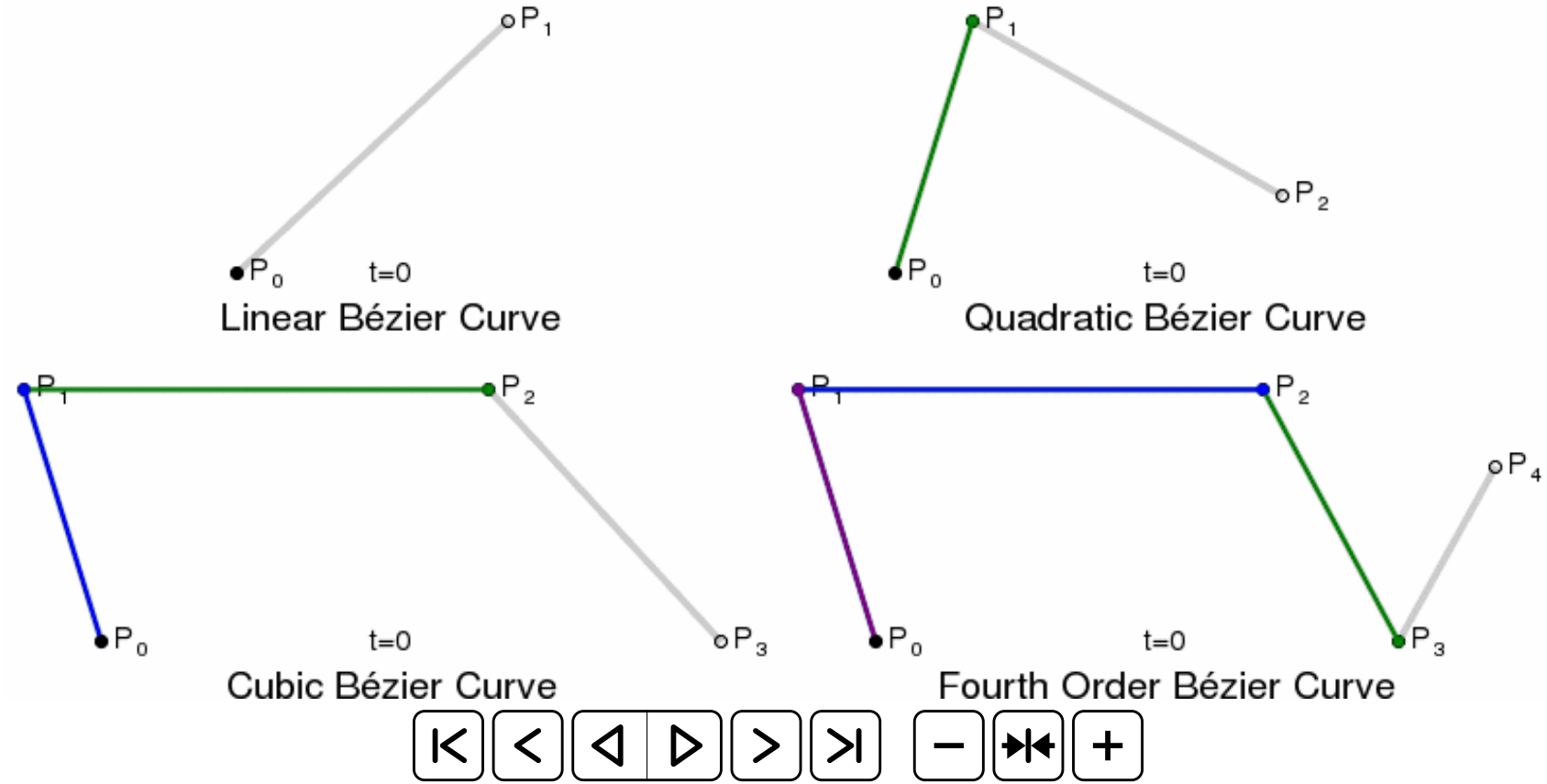
Polynomial splines are preferred to a polynomial basis because they are better at approximation. However, piecewise polynomial splines also have some issues:

- They are highly collinear.
- They are numerically unstable.

B-splines avoid this problem, so each term that goes into the conditional mean approximation is orthogonal.

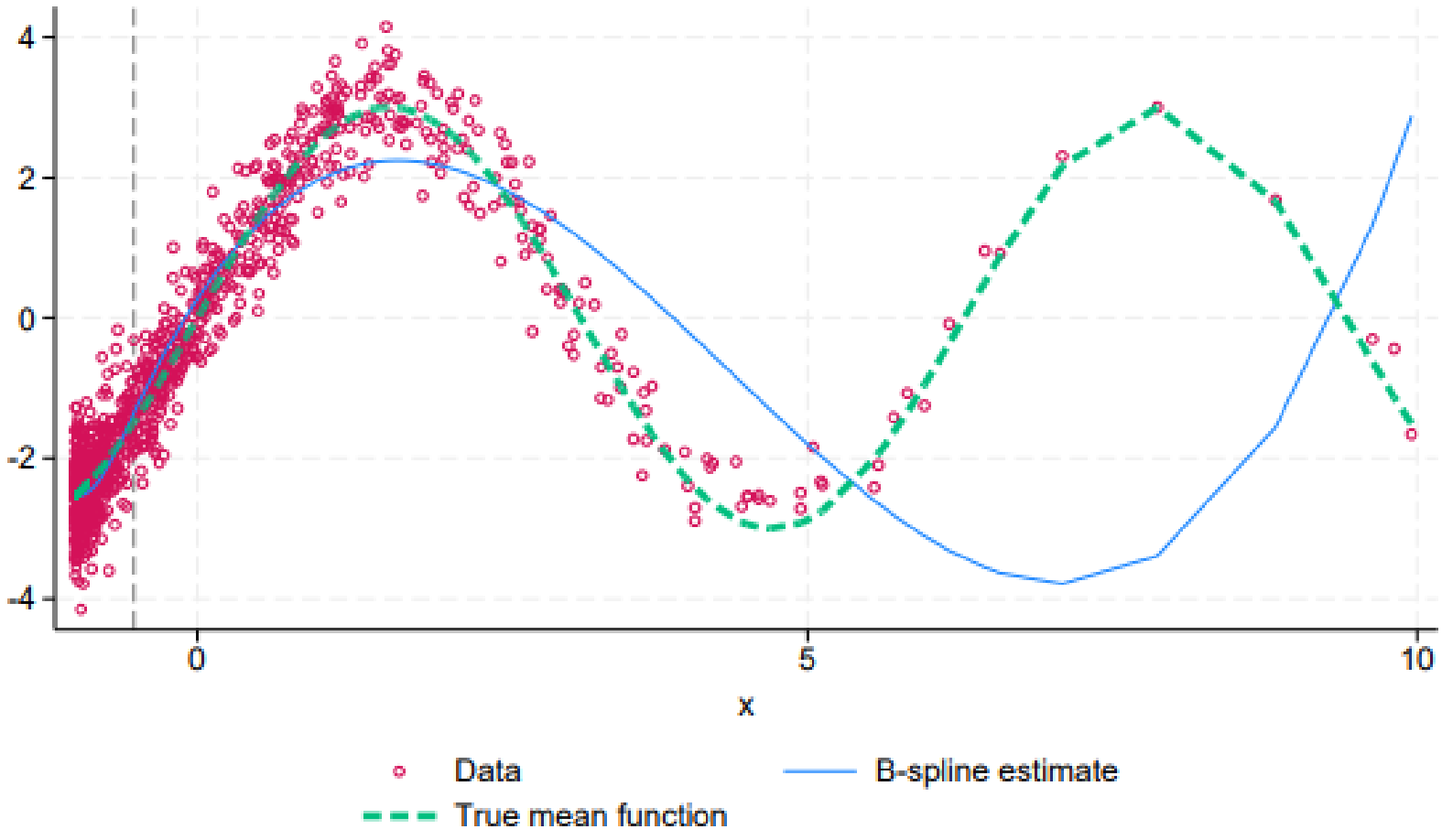
It is for this reason that B-splines are the default basis for *npregress series*.

Nonparametric Series Regression. Bézier splines



Nonparametric Series Regression. Limitations

When there are no enough data points at each subrange estimation is compromised.



Nonparametric Series Regression. Model Selection

Once we choose a polynomial basis we still need to define some characteristics:

- The order of the polynomial.
- The knots of the splines.

We can manually choose them or use criteria such as cross-validation.

Nonparametric Series Regression.

Stata implementation with *npregress series*.

Semiparametric Regression

It sets some structure to the data.

It combines a parametric component with a nonparametric component.

Generally it is used to fit a parametric model in which the functional form of some covariates is not known.

There are many semiparametric models, the most used are:

- Partially linear models.
- Single index models.

Partially Linear Model

It specifies the conditional mean to be linear in some component and nonlinear in others:

$$\mathbb{E}[y|x,z] = x'\beta + \lambda(z)$$

where $\lambda()$ is a unknown function. It has two advantages:

- Allows for any form to the function $\lambda()$.
- $\hat{\beta}$ is $\sqrt{n}$-consistent.

The estimation method will differ depending if we use splines or kernels for the nonparametric term.

Partially Linear Model

The easiest way is to parametrize $\lambda()$ with a basis like for example a polynomial.

We can do:

- Fractional polynomials.
- Spline regression.
- Generalized additive model.
- Robinson's difference estimation.

Robinson's Estimation

Robinson's proposed an estimation algorithm based on the fact that taking conditional expectations on z we get:

$$E[y|z] = E[x|z]'\beta + \lambda(z)$$

Which implies:

$$y - E[y|z] = (x - E[x|z])'\beta + u \tag{1}$$

The algorithm is then:

1. Regress y on z using some nonparametric method.
2. Regress x on z using some nonparametric method.
3. Estimate β using the previous regression model.
4. Recover $\lambda(z)$ using $\hat{\beta}$.

Single Index Model

It specifies the conditional mean to be an unknown scalar function of a linear combination of the regressors:

$$E[y|x] = g(x`\beta) \tag{2}$$

where the function $g()$ is unknown.

Notice that the $g()$ function will only be identified up to a location and scale since $g(x`\beta)$ is equivalent to $g^*(a + bx'\beta)$.

Therefore, we will set the coefficient of one variable to 1. Notice that rescaling the vector β by a constant and a similar rescaling of the function $g()$ by the inverse of the constant will produce the same regression.

Semiparametric Models.

Stata aplications.