Microeconometrics I – Problem Set 05
Nonparametric and Semiparametric Regressions

Vinícius de Almeida Nery Ferreira

September 19th, 2024

1. DGP

Define the following random variables that define our DGP.

- $\varepsilon_1 \sim \mathcal{N}(0, 1)$;
- $\varepsilon_2 \sim \mathcal{N}(0, 1)$;
- $X = \begin{bmatrix} \tilde{X}_1 \\ \tilde{X}_2 \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$;
- $X_1 = \Phi\left(\tilde{X}_1\right)$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution;
- $X_2 = \Phi\left(\tilde{X}_2\right)$;
- $Y_1 = f(X_1) + \varepsilon_1$, where $f(x) = \sin(\beta_1 \cdot x)$;
- $Y_2 = g(X_1, X_2) = f(X_1) + \beta_2 \cdot X_2 + \varepsilon_2$.

We observe a sample $\{X_1, X_2, Y_1, Y_2\}_{i=1}^N$, where $N = 10,000$. Our target parameters are the function $f(\cdot)$ and the coefficient β_2, where $\beta_1 = 4$ and $\beta_2 = 2$.

2. Plot of f

Using the generated data, we plot the true function $f(\cdot)$ in Figure 1.

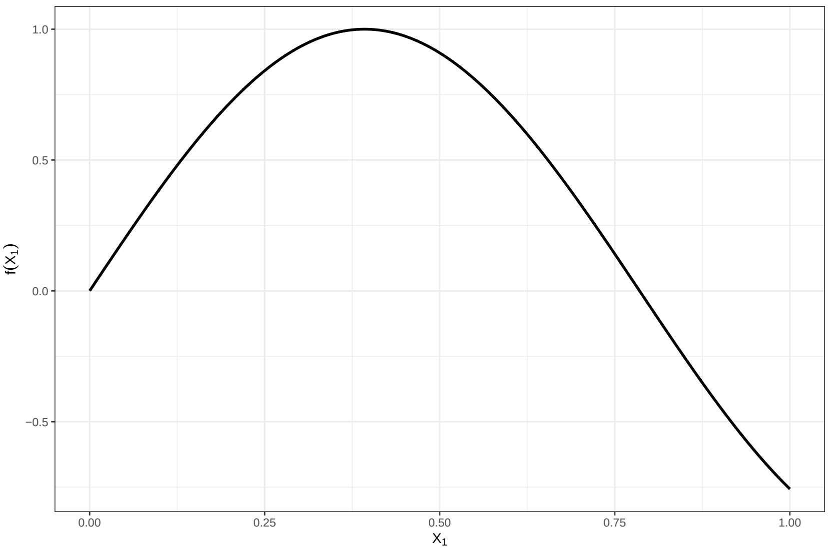

Figure 1: Plot of Function $f(X_1) = \sin(\beta_1 \cdot X_1)$

3 and 4. Estimating f: Nonparametric Local-Linear Regression with X_1 and Y_1

Using X_1 and Y_1, we estimate f by a nonparametric local-linear regression using the `lprobust` function from the `nprobust` **R** package. We keep all parameters as their default, including the number of grid points.

We plot the true function f and the estimated one in Figure 2. We also include the bias-corrected 95%-confidence interval of Calonico, Cattaneo and Ferrell (2018).

Note that the true function is include in the confidence interval at all points X_1. This is expected, since $Y_1 = f(X_1) + \varepsilon_1$, and thus our estimated model is correctly specified.

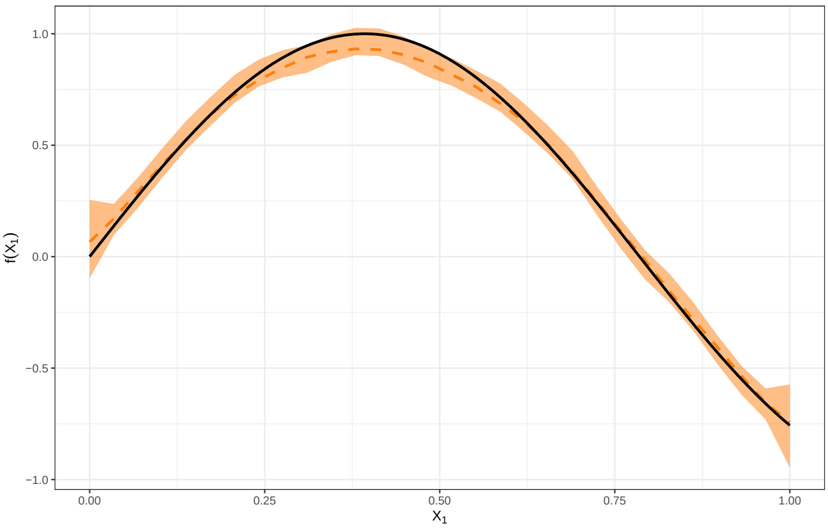

Figure 2: True and Estimated f Using X_1 and Y_1

5 and 6. Estimating f: Nonparametric Local-Linear Regression with X_1 and Y_2

Using X_1 and Y_2, we estimate f by a nonparametric local-linear regression using the `lprobust` function from the `nprobust` **R** package, keeping all parameters as their default. We plot the true function f and the estimated one in Figure 3. We also include the bias-corrected 95%-confidence interval of Calonico, Cattaneo and Ferrell (2018).

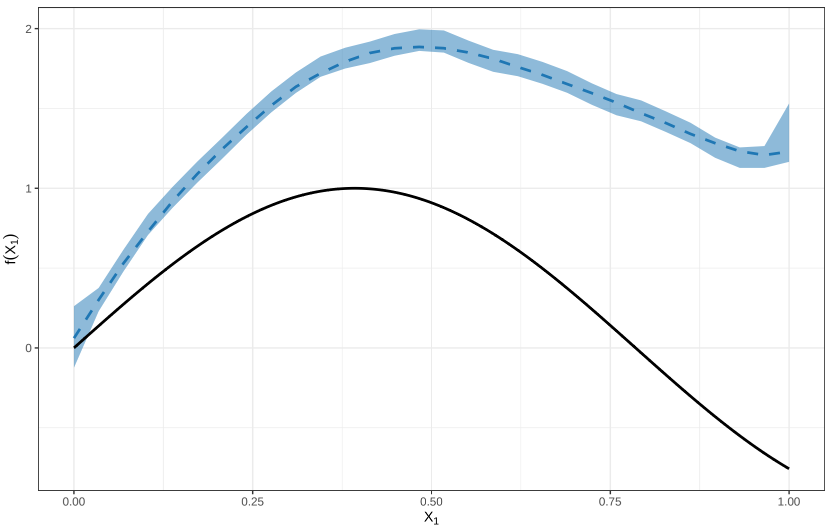

Figure 3: True and Estimated f Using X_2 and Y_1

As opposed to the previous item, the estimated and true functions f are very different. This because $Y_2 = g(X_1, X_2) = f(X_1)+\beta_2 \cdot X_2+\varepsilon_2$, and so our simple nonparametric model of $Y_2 = f(X_1)+\eta_2$ is misspecified, as we don't include the parametric part $\beta_2 \cdot X_2$.

7-9. Estimating f: Semiparametric Local-Linear Regression with X_1, X_2 and Y_2

To estimate the semiparametric model

$$Y_2 = g(X_1, X_2) = f(X_1) + \beta_2 \cdot X_2 + \varepsilon_2 \quad \text{where} \quad f(x) = \sin(\beta_1 \cdot x)$$

we will follow the steps written in *Estimating Model 3* in the Lecture Notes, which we describe below.

1. Fit a nonparametric local linear regression of X_2 on X_1. The evaluation grid contains all $10,000$ points of X_1 in increasing order. Let $\hat{X}_{2i}$ be the fitted values.
2. Calculate the residuals $\hat{U}_i = X_{2i} - \hat{X}_{2i}$ for all $i \in \{1, ..., N\}$.
3. Fit a nonparametric local linear regression of Y_2 on X_1, again with all points of X_1 in increasing order. Let $\hat{Y}_{2i}$ be the fitted values.
4. Calculate the residuals $\hat{U}_{Y,i} = Y_{2i} - \hat{Y}_{2i}$ for all $i \in \{1, ..., N\}$.
5. Using the Frisch-Waugh-Lovell (FWL) intuition, we regress $\hat{U}_{Y,i}$ on $\hat{U}_i$ using a parametric linear regression to estimate β_2. In this way, we are "filtering out" all parts of Y_2 and X_2 that are explained by X_1, and so we have that $\hat{\beta}_2 \xrightarrow{p} \beta_2$.
6. Residualize the outcome from its parametric part, *i.e.*, calculate $\tilde{Y}_{2i} = Y_{2i} - \hat{\beta}_2 \cdot X_{2i}\ \forall i \in \{1, ..., N\}$.
7. Estimate $f(X_1) = Y_2 - \beta_2 \cdot X_2 - \varepsilon_2 = \text{plim}(\tilde{Y}_{2i}) - \varepsilon_2$ using a nonparametric local-linear regression of $\tilde{Y}_{2i}$ on X_{1i}. In this step, we used the default number of grid points of `lprobust`.

We plot the true function f and the estimated one in Figure 3. We also include the bias-corrected 95%-confidence interval of Calonico, Cattaneo and Ferrell (2018). As now we are using the correct model estimation procedure, these objects are very close, with the true value of the funnction always being included in the 95%-confidence interval.

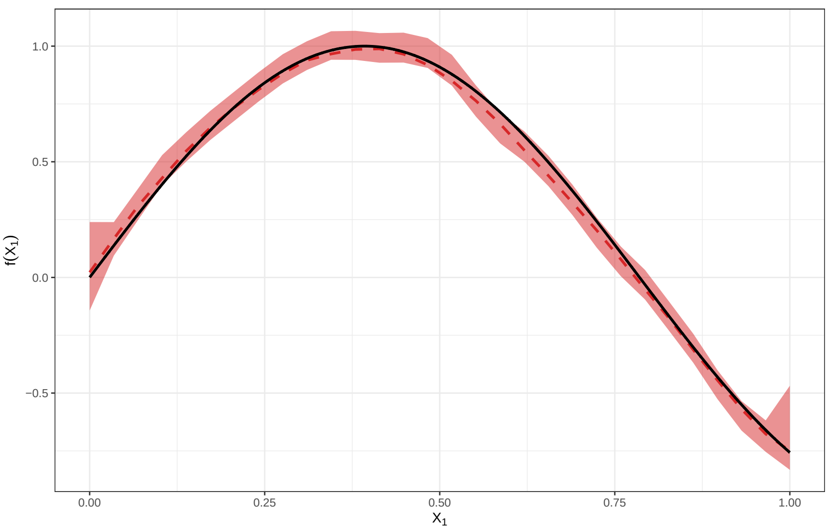

Figure 4: True and Estimated f Using X_1, X_2 and Y_1

Our estimated $\hat{\beta}_2$ in step 5 was 1.971, which is very close to the true value of $\beta_2 = 2$. As mentioned in step 5, these objects should be close to each other due to the nonparametric version of the FWL Theorem: we residualized X_2 and Y_2 from X_1, and so $\hat{\beta}_2 \xrightarrow{p} \beta_2$.