IEOR 265 – Lecture 7
Semiparametric Models

1 Nuisance Parameters

Consider the basic linear model $y_i = x_i'\beta + \epsilon_i$, where ϵ_i is i.i.d. noise with zero noise with finite variance. So far, we have focused on the question of estimating β; but, we could also ask the question whether it is possible to say something about ϵ_i. The reason that we have not addressed this issue is that, because the ϵ_i in this model represent random noise with zero mean, we do not gain any information for the purposes of model prediction (i.e., estimating $\mathbb{E}[y_i|x_i] = x_i'\beta$) by estimating the ϵ_i (or alternatively information about its distribution). However, if we are interested in understanding the uncertainty of our model predictions, then it is valuable to estimate the distribution of ϵ_i.

These ϵ_i are examples of *nuisance parameters*, which are any parameters that are not directly of interest but must be considered in the estimation. (Note that the designation of a parameter as a nuisance parameter is situationally dependent – in some applications, the nuisance parameter is also of interest.) In general, we can have situations in which there are a finite number of nuisance parameters or even an infinite number of nuisance parameters. There is no standard approach to handling nuisance parameters in regression problems. One approach is to estimate the nuisance parameters anyways, but unfortunately it is not always possible to estimate the nuisance parameters. Another approach is to consider the nuisance parameters as "worst-case disturbances" and use minmax estimators, which can be thought of as game-theoretic M-estimators.

1.1 Gaussian Noise in Linear Model

Consider the linear model in the situation where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for some unknown variance σ^2. Recall that the M-estimator was given by

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \Big(-(y_i - x_i'\beta)^2/(2\sigma^2) - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi) \Big).$$

In this case, the nuisance parameter is σ^2. The way this parameter was handled was to observe that the maximizer is independent of σ^2, which allowed us to rewrite the M-estimator as

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} -(y_i - x_i'\beta)^2 = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \arg\min_{\beta} \|Y - X\beta\|_2^2.$$

1.2 Generic Noise in Linear Model

Now consider the linear model in the case where ϵ_i is a generic zero mean distribution, meaning that it is of some unknown distribution. It turns out that we can estimate each ϵ_i in a consistent manner. Suppose that we assume

1. the norm of the x_i is deterministically bounded: $\|x_i\| \leq M$ for a finite $M < \infty$;
2. conditions under which OLS $\hat{\beta}$ is a consistent estimate of β.

Then we can use OLS to estimate the ϵ_i. Define

$$\hat{\epsilon}_i = y_i - x_i'\hat{\beta},$$

and note that

$$\begin{aligned}
|\hat{\epsilon}_i - \epsilon_i| &= |(y_i - x_i'\hat{\beta}) - \epsilon_i| \\
&= |x_i'\beta + \epsilon_i - x_i'\hat{\beta} - \epsilon_i| \\
&= |x_i'(\beta - \hat{\beta})| \\
&\leq \|x_i\| \cdot \|(\beta - \hat{\beta})\|,
\end{aligned}$$

where in the last line we have used the Cauchy-Schwarz inequality. And because of our assumptions, we have that $|\hat{\epsilon}_i - \epsilon_i| = O_p(1/\sqrt{n})$.

Now in turn, our estimates of ϵ_i can be used to estimate other items of interest. For example, we can use our estimates of $\hat{\epsilon}_i$ to estimate population parameters such as variance:

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

This estimator is consistent:

$$\begin{aligned}
|\hat{\sigma^2} - \sigma^2| &= \left| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 - \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - \sigma^2 \right| \\
&\leq \left| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 - \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \right| + \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - \sigma^2 \right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left| \hat{\epsilon}_i^2 - \epsilon_i^2 \right| + \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - \sigma^2 \right|.
\end{aligned}$$

where we have made use of the triangle inequality in the second and third lines. Next note that $|\hat{\epsilon}_i^2 - \epsilon_i^2| = O_p(1/\sqrt{n})$ by a version of the continuous mapping theorem and that $|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - \sigma^2| = O_p(1/\sqrt{n})$ because of the CLT. Thus, we have that $|\hat{\sigma^2} - \sigma^2| = O_p(1/\sqrt{n})$.

2 Partially Linear Model

Consider the following model

$$y_i = x_i'\beta + g(z_i) + \epsilon_i,$$

where $y_i \in \mathbb{R}$, $x_i, \beta \in \mathbb{R}^p$, $z_i \in \mathbb{R}^q$, $g(\cdot)$ is an unknown nonlinear function, and ϵ_i are noise. The data x_i, z_i are i.i.d., and the noise has conditionally zero mean $\mathbb{E}[\epsilon_i | x_i, z_i] = 0$ with unknown and bounded conditional variance $\mathbb{E}[\epsilon_i^2 | x_i, z_i] = \sigma^2(x_i, z_i)$. This is known as a partially linear model because it consists of a (parametric) linear part $x_i'\beta$ and a nonparametric part $g(z_i)$. One can think of the $g(\cdot)$ as an infinite-dimensional nuisance parameter.

3 Single-Index Model

Consider the following model

$$y_i = g(x_i'\beta) + \epsilon_i,$$

where $y_i \in \mathbb{R}$, $x_i, \beta \in \mathbb{R}^p$, $g(\cdot)$ is an unknown nonlinear function, and ϵ_i are noise. The data x_i are i.i.d., and the noise has conditionally zero mean $\mathbb{E}[\epsilon_i | x_i] = 0$. Such single-index models can be used for asset pricing, and here the $g(\cdot)$ can be thought of as an infinite-dimensional nuisance parameter.

4 Nadaraya-Watson Estimation

Consider the nonlinear model $y_i = f(x_i) + \epsilon_i$, where $f(\cdot)$ is an unknown nonlinear function. Suppose that given x_0, we would like to only estimate $f(x_0)$. One estimator that can be used is

$$\hat{\beta}_0[x_0] = \frac{\sum_{i=1}^n K(\|x_i - x_0\|/h) \cdot y_i}{\sum_{i=1}^n K(\|x_i - x_0\|/h)},$$

where $K(\cdot)$ is a kernel function. This estimator is known as the Nadaraya-Watson estimator, and it was one of the earlier techniques developed for nonparametric regression.

4.1 Alternative Characterizations

It turns out that we can characterize this estimator through multiple formulations. The first is as the following M-estimator

$$\hat{\beta}[x_0] = \arg\min_{\beta_0} \|W_h^{1/2}(Y - 1_n\beta_0)\|_2^2 = \arg\min_{\beta_0} \sum_{i=1}^n K(\|x_i - x_0\|/h) \cdot (y_i - \beta_0)^2.$$

A second characterization is as the mean with weights $\{K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h)\}$ of points $\{y_1, \ldots, y_n\}$.

4.2 Small Denominators in Nadaraya-Watson

The denominator of the Nadaraya-Watson estimator is worth examining. Define

$$\hat{g}(x_0) = \frac{1}{nh^p} \sum_{i=1}^{n} K(\|x_i - x_0\|/h),$$

and note that $\hat{g}(x_0)$ is an estimate of the probability density function of x_i at the point x_0. This is known as a kernel density estimate (KDE), and the intuition is that this is a smooth version of a histogram of the x_i.

The denominator of the Nadaraya-Watson estimator is a random variable, and technical problems occur when this denominator is small. This can be visualized graphically. The traditional approach to dealing with this is *trimming*, in which small denominators are eliminated. The trimmed version of the Nadaraya-Watson estimator is

$$\hat{\beta}_0[x_0] = \begin{cases} \frac{\sum_{i=1}^n K(\|x_i - x_0\|/h) \cdot y_i}{\sum_{i=1}^n K(\|x_i - x_0\|/h)}, & \text{if } \sum_{i=1}^n K(\|x_i - x_0\|/h) > \mu \\ 0, & \text{otherwise} \end{cases}.$$

One disadvantage of this approach is that if we think of $\hat{\beta}_0[x_0]$ as a function of x_0, then this function is not differentiable in x_0.

4.3 $L2$-Regularized Nadaraya-Watson Estimator

A new approach is to define the $L2$-regularized Nadaraya-Watson estimator

$$\hat{\beta}_0[x_0] = \frac{\sum_{i=1}^n K(\|x_i - x_0\|/h) \cdot y_i}{\lambda + \sum_{i=1}^n K(\|x_i - x_0\|/h)},$$

where $\lambda > 0$. If the kernel function is differentiable, then the function $\hat{\beta}[x_0]$ is always differentiable in x_0.

The reason for this name is that under the M-estimator interpretation of Nadaraya-Watson estimator, we have that

$$\hat{\beta}[x_0] = \arg\min_{\beta_0} \|W_h^{1/2}(Y - 1_n\beta_0)\|_2^2 + \lambda\|\beta_0\|_2^2 = \arg\min_{\beta_0} \sum_{i=1}^{n} K(\|x_i - x_0\|/h) \cdot (y_i - \beta_0)^2 + \lambda\beta_0^2.$$

Lastly, note that we can also interpret this estimator as the mean with weights

$$\{\lambda, K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h)\}$$

of points $\{0, y_1, \ldots, y_n\}$.

5 Partially Linear Model

Recall the following partially linear model

$$y_i = x_i'\beta + g(z_i) + \epsilon_i = f(x_i, z_i; \beta) + \epsilon_i,$$

where $y_i \in \mathbb{R}$, $x_i, \beta \in \mathbb{R}^p$, $z_i \in \mathbb{R}^q$, $g(\cdot)$ is an unknown nonlinear function, and ϵ_i are noise. The data x_i, z_i are i.i.d., and the noise has conditionally zero mean $\mathbb{E}[\epsilon_i | x_i, z_i] = 0$ with unknown and bounded conditional variance $\mathbb{E}[\epsilon_i^2 | x_i, z_i] = \sigma^2(x_i, z_i)$. This model is known as a partially linear model because it consists of a (parametric) linear part $x_i'\beta$ and a nonparametric part $g(z_i)$. One can think of the $g(\cdot)$ as an infinite-dimensional nuisance parameter, but in some situations this function can be of interest.

5.1 Nonparametric Approach

Suppose we were to compute a LLR of this model at an arbitrary point x_0, z_0 within the support of the x_i,z_i:

$$\begin{bmatrix} \hat{\beta}_0[x_0, z_0] \\ \hat{\beta}[x_0, z_0] \\ \hat{\eta}[x_0, z_0] \end{bmatrix} = \arg\min_{\beta_0, \beta, \eta} \left\| W_h^{1/2} \left(Y - \begin{bmatrix} 1_n & X_0 & Z_0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \\ \eta \end{bmatrix} \right) \right\|_2^2,$$

where $X_0 = X - x_0' 1_n$, $Z_0 = Z - z_0' 1_n$, and

$$W_h = \text{diag}\left(K\left(\frac{1}{h} \left\| \begin{bmatrix} x_1 \\ z_1 \end{bmatrix} - \begin{bmatrix} x_0 \\ z_0 \end{bmatrix} \right\| \right), \ldots, K\left(\frac{1}{h} \left\| \begin{bmatrix} x_n \\ z_n \end{bmatrix} - \begin{bmatrix} x_0 \\ z_0 \end{bmatrix} \right\| \right) \right).$$

By noting that $\nabla_x f = \beta$, one estimate of the parametric coefficients is $\hat{\beta} = \hat{\beta}[x_0, z_0]$. That is, in principle, we can use a purely nonparametric approach to estimate the parameters of this partially linear model. However, the rate of convergence will be $O_p(n^{-2/(p+q+4)})$. This is much slower than the parametric rate $O_p(1/\sqrt{n})$.

5.2 Semiparametric Approach

Ideally, our estimates of β should converge at the parametric rate $O_p(1/\sqrt{n})$, but the $g(z_i)$ term causes difficulties in being able to achieve this. But if we could somehow subtract out this term, then we would be able to estimate β at the parametric rate. This is the intuition behind the semiparametric approach. Observe that

$$\mathbb{E}[y_i | z_i] = \mathbb{E}[x_i'\beta + g(z_i) + \epsilon_i | z_i] = \mathbb{E}[x_i | z_i]'\beta + g(z_i),$$

and so

$$y_i - \mathbb{E}[y_i | z_i] = (x_i'\beta + g(z_i) + \epsilon_i) - \mathbb{E}[x_i | z_i]'\beta - g(z_i) = (x_i - \mathbb{E}[x_i | z_i])'\beta + \epsilon_i.$$

Now if we define

$$\hat{Y} = \begin{bmatrix} \mathbb{E}[y_1|z_1] \\ \vdots \\ \mathbb{E}[y_n|z_n] \end{bmatrix}$$

and

$$\hat{X} = \begin{bmatrix} \mathbb{E}[x_1|z_1]' \\ \vdots \\ \mathbb{E}[x_n|z_n]' \end{bmatrix}$$

then we can define an estimator

$$\hat{\beta} = \arg\min_{\beta} \|(Y - \hat{Y}) - (X - \hat{X})\beta\|_2^2 = ((X - \hat{X})'(X - \hat{X}))^{-1}((X - \hat{X})'(Y - \hat{Y})).$$

The only question is how can we compute $\mathbb{E}[x_i|z_i]$ and $\mathbb{E}[y_i|z_i]$? It turns out that if we compute those values with the trimmed version of the Nadaraya-Watson estimator, then the estimate $\hat{\beta}$ converges at the parametric rate under reasonable technical conditions. Intuitively, we would expect that we could alternatively use the $L2$-regularized Nadaraya-Watson estimator, but this has not yet been proven to be the case.