IEOR 265 – Lecture 2
Local Linear Regression

1 Local Linear Regression

As seen in the previous lecture, a geometric perspective to regression problems can be quite valuable. Consider a regression model

$$y = f(x) + \epsilon$$

in which $f(\cdot)$ is known to be highly nonlinear but of unknown structure. A nonparametric approach is natural, and one nonparametric method is known as local linear regression (LLR). The idea of this method is that if $f(\cdot)$ has sufficient smoothness (say twice-differentiable), then the model will look linear in small regions of input-space. Suppose that we consider points in input space nearby x_0, then intuitively our model looks like

$$y = \beta_0[x_0] + \sum_{j=1}^{p} \beta_j[x_0] \cdot (x^j - x_0^j) + \epsilon$$

for x near x_0 (e.g., $\|x - x_0\| \leq h$ for some small $h > 0$). The square brackets $[x_0]$ are used to represent the fact that the value of β will vary for different values of x_0.

The idea of a neighborhood of radius h is central to LLR. It is customary in statistics to call this h the *bandwidth*. In this method, we select points within a radius of h from x_0. Furthermore, we can weight the points accordingly so that points closer to x_0 are given more weight than those points further from x_0. To do this, we define a kernel function $K(u) : \mathbb{R} \to \mathbb{R}$ which has the properties

1. Finite Support – $K(u) = 0$ for $|u| \geq 1$;
2. Even Symmetry – $K(u) = K(-u)$;
3. Positive Values – $K(u) > 0$ for $|u| < 1$.

A canonical example is the Epanechnikov kernel

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2), & \text{for } |u| < 1 \\ 0, & \text{otherwise} \end{cases}$$

It turns out that the particular shape of the kernel function is not as important as the bandwidth h. If we choose a large h, then the local linear assumption is not accurate. On the other hand, if we choose a very small h, then the estimate will not be accurate because only a few data points will be considered. It turns out that this tradeoff in the value of h is a manifestation of the bias-variance tradeoff; however, being able to quantify this requires understanding stochastic convergence.

Before we discuss this tradeoff in more detail, we describe the LLR. The idea is to perform a weighted-variant of OLS by using a kernel function and a bandwidth h to provide the weighting. The LLR estimate $\hat{\beta}_0[x_0], \hat{\beta}[x_0]$ is given by the minimizer to the following optimization

$$\begin{bmatrix}\hat{\beta}_0[x_0]\\ \hat{\beta}[x_0]\end{bmatrix} = \arg\min_{\beta_0,\beta} \sum_{i=1}^n K(\|x_i - x_0\|/h)\cdot(y_i - \beta_0 - (x_i - x_0)'\beta)^2.$$

Now if we define a weighting matrix

$$W_h = \operatorname{diag}\left(K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h)\right),$$

then we can rewrite this optimization as

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

where $\mathbb{1}_n$ is a real-valued vector of all ones and of length dimension n and $X_0 = X - x_0'\mathbb{1}_n$. This is identical to the OLS optimization, and so we can use that answer to conclude that

$$\begin{bmatrix}\hat{\beta}_0[x_0]\\ \hat{\beta}[x_0]\end{bmatrix} = (\begin{bmatrix}\mathbb{1}_n & X_0\end{bmatrix}' W_h \begin{bmatrix}\mathbb{1}_n & X_0\end{bmatrix})^{-1}(\begin{bmatrix}\mathbb{1}_n & X_0\end{bmatrix}' W_h Y).$$

2 Bias-Variance Tradeoff

Consider the case of parametric regression with $\beta \in \mathbb{R}$, and suppose that we would like to analyze the expectation of the squared loss of the difference between a estimate $\hat{\beta}$ and the true parameter β. In particular, we have that

$$\begin{aligned}
\mathbb{E}\big((\hat{\beta} - \beta)^2\big) &= \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}) + \mathbb{E}(\hat{\beta}) - \beta)^2\big)\\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big) + 2\mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)(\hat{\beta} - \mathbb{E}(\hat{\beta})\big)\\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big) + 2(\mathbb{E}(\hat{\beta}) - \beta)(\mathbb{E}(\hat{\beta}) - \mathbb{E}(\hat{\beta}))\\
&= \mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big) + \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big).
\end{aligned}$$

The term $\mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta}))^2\big)$ is clearly the variance of the estimate $\hat{\beta}$. The other term $\mathbb{E}\big((\mathbb{E}(\hat{\beta}) - \beta)^2\big)$ measures how far away the "best" estimate is from the true value, and it is common to define $\operatorname{bias}(\hat{\beta}) = \mathbb{E}\big(\mathbb{E}(\hat{\beta}) - \beta\big)$. With this notation, we have that

$$\mathbb{E}\big((\hat{\beta} - \beta)^2\big) = (\operatorname{bias}(\hat{\beta}))^2 + \operatorname{var}(\hat{\beta}).$$

This equation states that the expected estimation error (as measured by the squared loss) is equal to the bias-squared plus the variance, and in fact there is a tradeoff between these two aspects in an estimate.

It is worth making three comments. The first is that if $\text{bias}(\hat{\beta}) = \mathbb{E}\big(\mathbb{E}(\hat{\beta}) - \beta\big) = 0$, then the estimate $\hat{\beta}$ is said to be *unbiased*. Second, this bias-variance tradeoff exists for vector-valued parameters $\beta \in \mathbb{R}^p$, for nonparametric estimates, and other models. Lastly, the term *overfit* is sometimes used to refer to an model with low bias but extremely high variance.

3 Types of Stochastic Convergence

There are several types of stochastic convergence, and they can be thought of as direct analogs of *convergence of measures*. Here, we will only be concerned with two basic types of stochastic convergence that are commonly used in the theory of regression.

3.1 Convergence in Distribution

A sequence of random variables $X_1, X_2, \ldots$ converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(u) = F_X(u),$$

for every point u at which $F_X(u)$ is continuous, where $F_{X_n}(u)$ is the distribution function for X_n and $F_X(u)$ is the distribution function for X. This type of convergence is denoted $X_n \xrightarrow{d} X$.

3.2 Convergence in Probability

A sequence of random variables $X_1, X_2, \ldots$ converges in probability to a random variable X if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0.$$

This type of convergence is denoted $X_n \xrightarrow{p} X$ or as $X_n - X \xrightarrow{p} 0$. There are a few additional notations for denoting convergence in distribution, and these are similar to big-O notation that is used in mathematics. We define the following little-o_p notation

$$X_n = o_p(a_n) \Leftrightarrow X_n/a_n \xrightarrow{p} 0.$$

There is a similar big-O_p notation that denotes stochastic boundedness. We say that $X_n = O_p(a_n)$ if for any $\epsilon > 0$ there finite $M > 0$ such that

$$\mathbb{P}(|X_n/a_n| > M) < \epsilon, \forall n.$$

3.3 Relationships Between Modes of Convergence

There are several important points to note:

- Convergence in probability implies convergence in distribution.
- Convergence in distribution does not always imply convergence in probability.
- If X_n converges in distribution to a constant x_0, then X_n also converges in probability to x_0.

4 Concentration About the Mean

Consider an infinite sequence of (mutually) independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots$, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample average. There are a number of results that show that the sample average $\overline{X}_n$ is "close" to the mean of the distribution. The intuition is that the sample average can only deviate "far" from the mean if the random variables act in concert to pull the average in one direction, but the probability that the random variables pull in the same direction is small because of their independence.

4.1 Weak Law of Large Numbers

If the random variables X_i have a finite first moment $\mathbb{E}|X_i| < \infty$, then $\overline{X}_n \xrightarrow{p} \mu$ where $\mu = \mathbb{E}(X_i)$. In words — the weak law of large numbers states that if i.i.d. random variables have a finite first moment, then their sample average converges in probability to the mean of the distribution. Note that we could also write this result as $\overline{X}_n - \mu = o_p(1)$.

4.2 Central Limit Theorem

A more precise statement of the convergence of sample averages is given by the following theorem: If the random variables X_i with mean $\mathbb{E}(X_i) = \mu$ have a finite variance $\text{Var}(X_i) = \sigma^2 < \infty$, then

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\mathcal{N}(0, \sigma^2)$ is the distribution of a Gaussian random variable with mean 0 and variance σ^2. This is a more precise statement because it describes the distribution of the sample average when it is appropriately scaled by $\sqrt{n}$. This scaling is important because otherwise $\overline{X}_n \xrightarrow{d} \mu$ by the weak law of large numbers (and the fact that convergence in probability implies convergence in distribution).

5 Extensions of Central Limit Theorem

There are a number of extensions of the Central Limit Theorem, and these are important for understanding the convergence properties of regression estimators, which can be quite complicated.

5.1 Continuous Mapping Theorem

Suppose that $g(u)$ is a continuous function. The continuous mapping theorem states that convergence of a sequence of random variables $\{X_n\}$ to a limiting random variable X is preserved under continuous mappings. More specifically:

- If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.
- If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.

5.2 Slutsky's Theorem

If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} y_0$, where y_0 is a constant, then

- $X_n + Y_n \xrightarrow{d} X + y_0$;
- $Y_n X_n \xrightarrow{d} y_0 X$;

There is a technical point to note about this theorem: Convergence of Y_n to a constant is a subtle but important feature for this result, because the theorem will not generally hold when Y_n converges to a non-constant random variable. Consider the example where $X_n \sim \mathcal{N}(0,1)$ and $Y_n = X_n$, then $X_n + Y_n = \mathcal{N}(0,4)$ which does not converge in distribution to $\mathcal{N}(0,1) + \mathcal{N}(0,1) = \mathcal{N}(0,2)$.

5.3 Delta Method

Using the Continuous Mapping Theorem and Slutsky's Theorem, we can now state an extension of the Central Limit Theorem. Consider an infinite sequence of random variables $\{X_n\}$ that satisfies

$$\sqrt{n}[X_n - \theta] \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where θ and σ^2 are finite constants. If $g(u)$ is continuously differentiable at θ and $g'(\theta) \neq 0$, then

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{d} g'(\theta)\mathcal{N}(0, \sigma^2).$$

This result is derived using the Lagrange form of Taylor's Theorem, and can hence be thought of as a Taylor polynomial expansion version of the Central Limit Theorem. There are higher-order versions of the Delta Method that we do not discuss here. For instance, the Second-Order Delta Method applies when $g'(u) = 0$ and $g''(u) \neq 0$.