1 Geometric Interpretation of OLS

Recall the optimization formulation of OLS,

$\hat{\beta} = \arg \min_{\beta} \| Y - X\beta \|_2^2,$

where the variables are as defined before. The basic tension in the problem above is that in general no exact solution exists to the linear equation

$Y = X\beta;$

otherwise we could use linear algebra to compute $\beta$, and this value would be a minimizer to the optimization problem written above.

Though no exact solution exists to $Y = X\beta$, an interesting question to ask is whether there is some related linear equation for which an exact solution exists. Because the noise is in $Y$ and not $X$, we can imagine that we would like to pick some $\hat{Y}$ such that $\hat{Y} = X\hat{\beta}$ has an exact solution. Recall from linear algebra, that this is equivalent to asking that $\hat{Y} \in \mathcal{R}(X)$ (i.e., $\hat{Y}$ is in the range space of $X$). Now if we think of $\hat{Y}$ as true signal, then we can decompose $Y$ as

$Y = \hat{Y} + \Delta Y,$

where $\Delta Y$ represents orthogonal noise. Because – from Fredholm’s theorem in linear algebra – we know that the range space of $X$ is orthogonal to the null space of $X'$ (i.e., $\mathcal{R}(X) \perp \mathcal{N}(X')$), it must be the case that $\Delta Y \in \mathcal{N}(X')$ since we defined $\hat{Y}$ such that $\hat{Y} \in \mathcal{R}(X)$. As a result, premultiplying $Y = \hat{Y} + \Delta Y$ by $X'$ gives

$X'Y = X'\hat{Y} + X'\Delta Y = X'\hat{Y}.$

The intuition is that premultiplying by $X'$ removes the noise component. And because $\hat{Y} \in \mathcal{R}(X)$ and $\hat{Y} = X\hat{\beta}$, we must have that

$X'Y = X'\hat{Y} = X'X\hat{\beta}.$

Solving this gives $\hat{\beta} = (X'X)^{-1}(X'Y)$, which is our regular equation for the OLS estimate.
2 Local Linear Regression

As seen above, 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

\[ y = \beta_0[x_0] + \sum_{j=1}^{p} \beta_j[x_0] \cdot (x^j - x^j_0) + \epsilon \]

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 \( \beta \) 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 \( K(u) : \mathbb{R} \rightarrow \mathbb{R} \) 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
\[ W_h = \text{diag}(K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h)), \]
then we can rewrite this optimization as
\[
\begin{bmatrix} \hat{\beta}_0[x_0] \\ \beta[x_0] \end{bmatrix} = \arg \min_{\beta_0, \beta} \left\| W_h^{1/2} \left( Y - \begin{bmatrix} \mathbb{1}_n & X_0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} \right) \right\|^2,
\]
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
\[
\begin{bmatrix} \hat{\beta}_0[x_0] \\ \beta[x_0] \end{bmatrix} = \left( \begin{bmatrix} \mathbb{1}_n & X_0 \end{bmatrix}' W_h [\mathbb{1}_n & X_0] \right)^{-1} \left( \begin{bmatrix} \mathbb{1}_n & X_0 \end{bmatrix}' W_h Y \right).
\]

### 3 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 \( \beta \). In particular, we have that
\[
\mathbb{E}( (\hat{\beta} - \beta)^2 ) = \mathbb{E}( (\hat{\beta} - \mathbb{E}(\hat{\beta}) + \mathbb{E}(\hat{\beta}) - \beta)^2 )
\]
\[
= \mathbb{E}( (\mathbb{E}(\hat{\beta}) - \beta)^2 ) + \mathbb{E}( (\hat{\beta} - \mathbb{E}(\hat{\beta}))^2 ) + 2 \mathbb{E}( (\mathbb{E}(\hat{\beta}) - \beta)(\hat{\beta} - \mathbb{E}(\hat{\beta})) )
\]
\[
= \mathbb{E}( (\mathbb{E}(\hat{\beta}) - \beta)^2 ) + \mathbb{E}( (\hat{\beta} - \mathbb{E}(\hat{\beta}))^2 ) + 2(\mathbb{E}(\hat{\beta}) - \beta)(\mathbb{E}(\hat{\beta}) - \mathbb{E}(\hat{\beta}))
\]
\[
= \mathbb{E}( (\mathbb{E}(\hat{\beta}) - \beta)^2 ) + \mathbb{E}( (\hat{\beta} - \mathbb{E}(\hat{\beta}))^2 )
\].

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

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 bias(\( \hat{\beta} \)) = \( \mathbb{E}(\mathbb{E}(\hat{\beta}) - \beta) = 0 \), 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 a model with low bias but extremely high variance.