Nonparametric Regression

- Section 5.6 gives a rank-based procedure for estimating a regression function when the function is unknown and nonlinear BUT known to be monotonic.

- Here we will examine a distribution-free method of estimating a very general type of regression function.

- In nonparametric regression, we assume very little about the functional form of the regression function.

- We assume the model:
  \[ E(Y | X = x) = f(x) \]
  where \( f(\cdot) \) is unknown but is typically assumed to be a smooth and continuous function.

- We also assume independence for the residuals
  \[ Y_i - f(X_i), \quad i = 1, \ldots, n \]

**Goal:** Estimate the mean response function \( f(\cdot) \).

Advantages of Nonparametric Regression

- Useful when we cannot know the relationship between \( Y \) and \( X \)
- More flexible type of regression model
- Can account for unusual behavior in the data
- Less likely to have bias resulting from wrong model being chosen
Disadvantages of Nonparametric Regression

- Not as easy to interpret
- No easy way to describe relationship between $Y$ and $X$ with a formula (must be done with a graph)
- Inference is not as straightforward

**Note:** Nonparametric regression is sometimes called scatter plot smoothing.

**Kernel Regression**

- The idea behind kernel regression is to estimate $f(x)$ at each value $x^*$ along the horizontal axis.

- At each value $x^*$, the estimate $\hat{f}(x^*)$ is simply an average of the $y$-values of the observations near $x^*$.
- Consider a “window” of points centered at $x^*$:

```
\begin{tikzpicture}
    \draw[->] (0,0) -- (5,0) node[right] {$x$};
    \draw[->] (0,0) -- (0,5) node[above] {$y$};
    \draw (2,2) rectangle (4,4);
    \node at (3,3) {The value of $\hat{f}(x^*)$ is the average of the $y$-values inside this window.};
\end{tikzpicture}
```
• The width of this window is called the **bandwidth**.

• At each different \( x^* \), the window of points moved to the left or right (moving average).

• Better idea: Use weighted average of \( y \)-values, with more weight on points near \( x^* \).

• This can be done using a **weighting** function known as a **kernel**.

• Then, for any \( x^* \),

\[
\hat{f}_\lambda(x^*) = \frac{1}{n} \sum_{i=1}^{n} w_i y_i
\]

where the weights \( w_i = \frac{1}{\lambda} K\left( \frac{x^* - x_i}{\lambda} \right) \)

\( K(\cdot) \) is a kernel function, which typically is a **density** function symmetric about 0.

\( \lambda = \) bandwidth, which controls the **smoothness** of the estimate of \( f(x) \).

**Possible choices of kernel:**

- Uniform (box) kernel: Gives all points in window **equal weight**; gives all points outside window **no weight**.

- Normal kernel: Gives points near \( x^* \) more weight; gives points far from \( x^* \) less weight.

- Epanechnikov kernel: \( K(x) = \begin{cases} 0.75 (1-x^2) & \text{for } |x|<1 \\ 0 & \text{otherwise} \end{cases} \)

**Compromise:** Gives points closer to \( x^* \) more weight, but gives points very far from \( x^* \) no weight.
Note: The Nadaraya-Watson estimator

\[ \hat{f}_\lambda(x) = \sum_{i=1}^{n} \frac{\omega_i}{\sum_{j=1}^{n} \omega_j} y_i \]

is a modification that assures that the weights for the \( Y_i \)'s will sum to one.

- The choice of **bandwidth** \( \lambda \) is of more practical importance than the choice of kernel.

- The bandwidth controls how many data values are used to compute \( f(x^*) \) at each \( x^* \).

Large \( \lambda \rightarrow \) many data values used at each estimation

\[ \Rightarrow \text{low variability of estimate, smoother-looking curve} \]

Small \( \lambda \rightarrow \) fewer data values used at each estimation

\[ \Rightarrow \text{high variability of estimate, wiggly-looking curve} \]
• Choosing $\lambda$ too large results in an estimate that **oversmooths** the true nature of the relationship between $Y$ and $X$.

• Choosing $\lambda$ too small results in an estimate that follows the "noise" in the data too closely.

• Often the best choice of $\lambda$ is made through visual inspection (pick the roughest estimate that does not fluctuate implausibly?).

• Automatic bandwidth selection methods such as **cross-validation** are also available – this chooses the $\lambda$ that minimizes a mean squared prediction error.

**Example:** We have data on the horsepower ($X$) and gas mileage ($Y$, in miles per gallon) of 82 cars, from Heavenrich et al. (1991).

• On computer: The R function `ksmooth` performs kernel regression (see web page for examples with various kernel functions and bandwidths).

  - **Best choice** appears to be the normal kernel with `bw = 60`.

  - Other smoothing functions may produce better results.