Lecture 12: Kernel Regression and Beyond

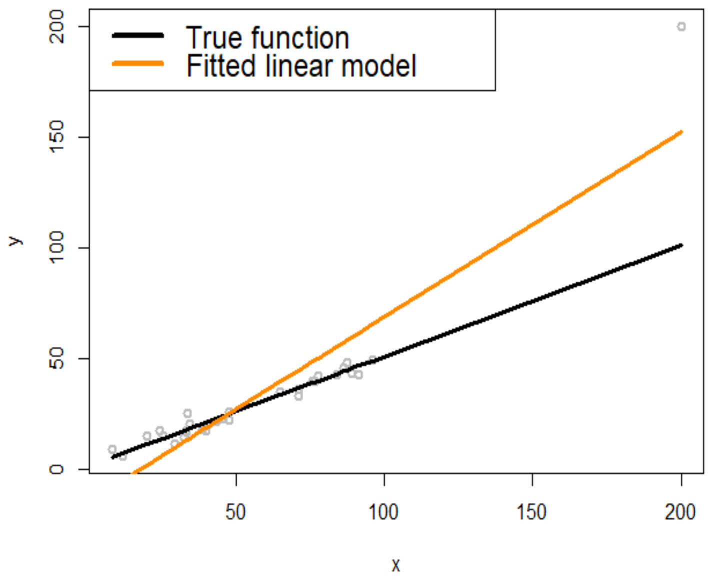
Instructor: Prof. Shuai Huang

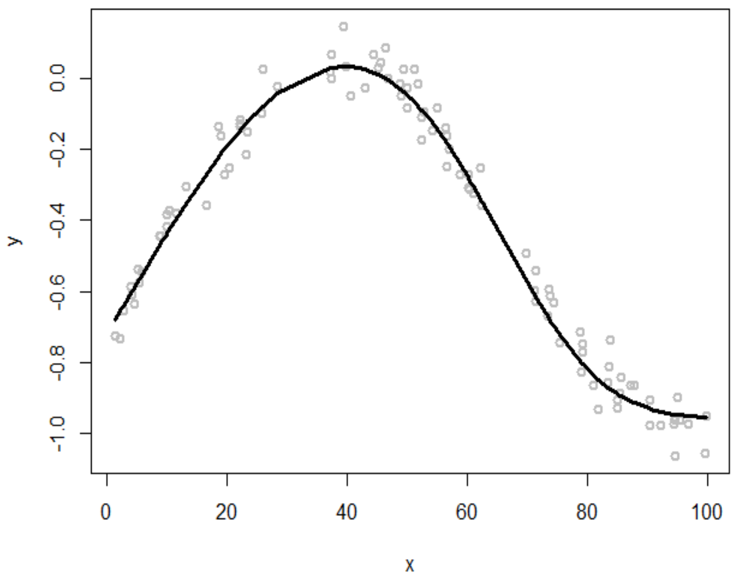
Industrial and Systems Engineering

University of Washington

Limitations of linear regression model

- A global model that generalizes local data points to a central model





How far is it from linear regression to nonlinear regression?

- Let's look at the simple linear regression problem $y = \beta_0 + \beta_1 x$. Let's further simplify it by assuming that we know the mean of y is zero, so is the mean of x. This will lead to the model as $y = \beta_1 x$ and the estimator of β_1 as

$$\beta_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)}{\sum_{i=1}^{n} x_i^2}.$$

- Thus, when we try to make prediction on a new data point with a given x^*, the prediction y^* will be

$$y^* = x^* \frac{\left(\sum_{i=1}^{n} x_i y_i\right)}{\sum_{i=1}^{n} x_i^2}.$$

- This could be further reformed as:

$$y^* = \sum_{i=1}^{n} y_i \frac{x_i}{\sum_{i=1}^{n} x_i^2} x^*,$$

- which is equivalent with

$$y^* = \sum_{i=1}^{n} y_i \frac{x_i x^*}{n S_x^2}.$$

Generalize this insights into development of nonlinear models

- We then pursue a generalized family of model, defined as:

$$y^* = \sum_{n=1}^{N} y_n w(x_n, x^*).$$

- Here, $w(x_n, x^*)$ is the weight that characterizes the similarity between x^* and the existing data points, x_n for $n = 1,2, \dots, N$.
- Roughly speaking, there are two types of similarity metric.
- One is the **K-nearest neighbor (KNN) smoother**:

$$w(x_n, x^*) = \begin{cases} \frac{1}{k}, & if\, x_n\ is\ one\ of\ the\ k\ nearest\ neighbors\ of\ x^* \\ 0, & if\, x_n\ is\ NOT\ in\ the\ k\ nearest\ neighbors\ of\ x^* \end{cases}.$$

- Another is the **kernel smoother**:

$$w(x_n, x^*) = \frac{K(x_n, x^*)}{\sum_{n=1}^{N} K(x_n, x^*)}.$$

Some examples of the kernel functions

Kernel function	Mathematical form	Parameters
Linear	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$	$null$
Polynomial	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j + 1)^q$	q
Gaussian radial basis	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma \lVert x_i - x_j \rVert^2}$	$\gamma \geq 0$
Laplace radial basis	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma \lVert x_i - x_j \rVert}$	$\gamma \geq 0$
Hyperbolic tangent	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\boldsymbol{x}_i^T \boldsymbol{x}_j + b)$	b
Sigmoid	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(a\boldsymbol{x}_i^T \boldsymbol{x}_j + b)$	a, b
Bessel function	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \dfrac{bessel_{v+1}^{n}(\sigma \lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert)}{(\lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert)^{-n(v+1)}}$	σ, n, v
ANOVA radial basis	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{k=1}^{n} e^{-\sigma\left(x_i^k - x_j^k\right)}\right)^d$	σ, d

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets

Conditional variance regression model

- **Heteroscedasticity** refers to the phenomenon that the variance of the response variable may also change
- This leads to the following model:

$$y = \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon_{\boldsymbol{x}},$$

- and $\epsilon_{\boldsymbol{x}}$ is the error term that is a normal distribution with varying variance:

$$\epsilon_{\boldsymbol{x}} \sim N(0, \sigma_{\boldsymbol{x}}^2).$$

Parameter estimation ($\boldsymbol{\sigma_x^2}$ is known)

- If we have known the $\sigma_{\boldsymbol{x}}^2$, this will lead to the following scheme for parameter estimation of the unknown regression parameters. The likelihood function is:

$$-\frac{n}{2}\ln 2\pi - \frac{1}{2}\sum_{n=1}^{N}\log\sigma_{\boldsymbol{x}_n}^2 - \frac{1}{2}\sum_{n=1}^{N}\frac{\left(y_n - \boldsymbol{\beta}^T\boldsymbol{x}_n\right)^2}{\sigma_{\boldsymbol{x}_n}^2}.$$

- As we have known $\sigma_{\boldsymbol{x}}^2$, the parameters to be estimated only involve the last part of the likelihood function. Thus, we estimate the parameters that minimize

$$\frac{1}{2}\sum_{n=1}^{N}\frac{\left(y_n - \boldsymbol{\beta}^T\boldsymbol{x}_n\right)^2}{\sigma_{\boldsymbol{x}_n}^2}.$$

- This could be written in the matrix form as

$$\min_{\boldsymbol{\beta}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}),$$

- where $\mathbf{W}$ is a diagonal matrix with its diagonal elements as $\mathbf{W}_{nn} = \frac{1}{\sigma_{\boldsymbol{x}_n}^2}$.
- And we can get that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{Y}$.

Parameter estimation ($\sigma_{\boldsymbol{x}}^2$ is unknown)

We propose the following steps:

- 1. Initialize $\hat{\sigma}_{\boldsymbol{x}_n}^2$ for $n = 1,2, \dots, N$, by any reasonable approach including the random generation of values.
- 2. Build a regression model for the mean of the response variable using the weighted LS estimator. Estimate $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{Y}$ and get $\hat{y}_n = \widehat{\boldsymbol{\beta}}^T\boldsymbol{x}_n$.
- 3. Derive the residuals $\hat{\varepsilon}_n = y_n - \hat{y}_n$.
- 4. Build a regression model, e.g., using the kernel regression which is a nonparametric method, to fit $\hat{\varepsilon}_n^2$ using $\boldsymbol{x}_n$ for $n = 1,2, \dots, N$.
- 5. Predict $\hat{\sigma}_{\boldsymbol{x}_n}^2$ for $n = 1,2, \dots, N$ using the fitted regression model in Step 3.
- 6. Repeat Step 2 – Step 5 until convergence or satisfaction of a stopping criteria (could be a fixed number of iterations or small change of parameters).

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets