CPE 490 590: Machine Learning for Engineering Applications

06 Advanced Regression

Rahul Bhadani

Electrical & Computer Engineering, The University of Alabama in Huntsville

Outline

Polynomial Regression

Polynomial Regression

Regression Model

$$y_i = w_0 + w_1 x_i + w_2 x_i^2 + \cdots + w_k x_i^k + \epsilon_i \tag{1}$$

In this model, we only have single feature.

Vectorizing Polynomial Regression

Regression Model

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots x_1^k \\ 1 & x_2 & x_2^2 & \cdots x_2^k \\ \cdots & \cdots & \cdots & \cdots\cdots \\ 1 & x_n & x_n^2 & \cdots x_n^k \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \Rightarrow \mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon} \quad (2)$$

$\mathbf{X}$ is called the Vandermonde matrix (this is different from $\mathbf{X}$ obtained in the matrix form of the MLR model).

Least Square Solution

To find, $\mathbf{w}$,

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$
$$\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{y}$$

Then, the ordinary least square estimate provides the solution as

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Polynomial regression can be expanded to multiple features.

Fitted Values and Residuals

The residual term is

$$\mathbf{e}_{n\times 1} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{w}}$$

The vector of the fitted value $\hat{\mathbf{y}}$ can be expressed in terms of the hat matrix $\mathbf{H}$ as

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{H}\mathbf{y}$$

where

$$\mathbf{H}_{n\times n} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

Then, we could also write $\mathbf{e}_{n\times 1} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.

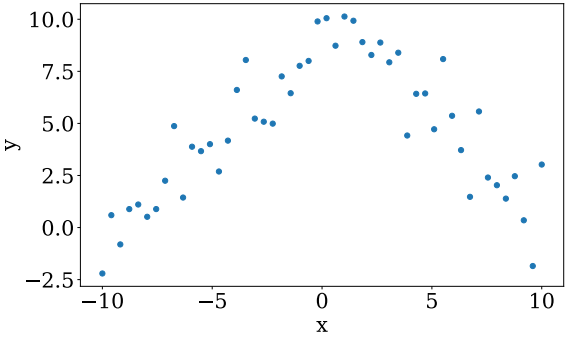
Polynomial Model with Multiple Features

Polynomial Regression can also expand to multiple features:

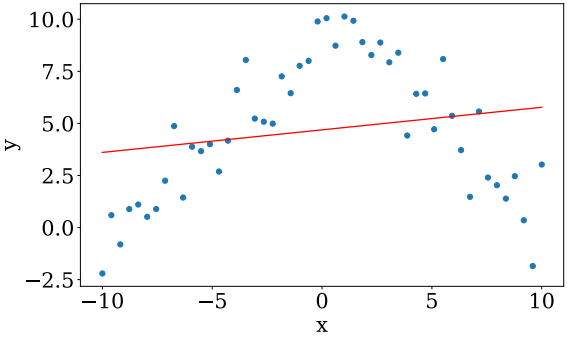
$$y_i = w_0 + w_1 \times area + w_2 \times (area)^2 + w_3 \times num\_rooms + w_4 \times yard\_sqft$$

Locally Weighted Kernel Regression

Intuition



Intuition



Intuition

Clearly, here are two sets of data points.

Intuition

We may benefit from fitting two kinds of regression lines.
That can be done using a locally-weighted regression line.

Locally Weighted Kernel Regression

Implementing Locally Weighted Regression

1. Provide a smoothing parameter τ.
2. Provide a degree of polynomial fit (if you are doing polynomial fit).
3. Select some query points.

Cost Function for Locally Weight Kernel Regression

Cost Function

$$J(\mathbf{w}) = \frac{1}{2n}(\mathbf{y} - \mathbf{x}\mathbf{w})^\top \boldsymbol{\kappa}(\mathbf{y} - \mathbf{x}\mathbf{w})$$

where

$$\boldsymbol{\kappa} = \begin{bmatrix} \kappa_1 & 0 & \cdots & 0 \\ 0 & \kappa_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \kappa_n \end{bmatrix}$$

is a **kernel function**.

Estimator

Closed Form of Estimation

$$\mathbf{w} = \frac{1}{n}(\mathbf{x}\kappa\mathbf{x})^{-1}\mathbf{x}^{\top}\mathbf{w}\mathbf{y} \tag{3}$$

This method is also called **Lowess (locally weighted regression scatter plot smoothing)** or **Loess**.

Choosing a Kernel Function

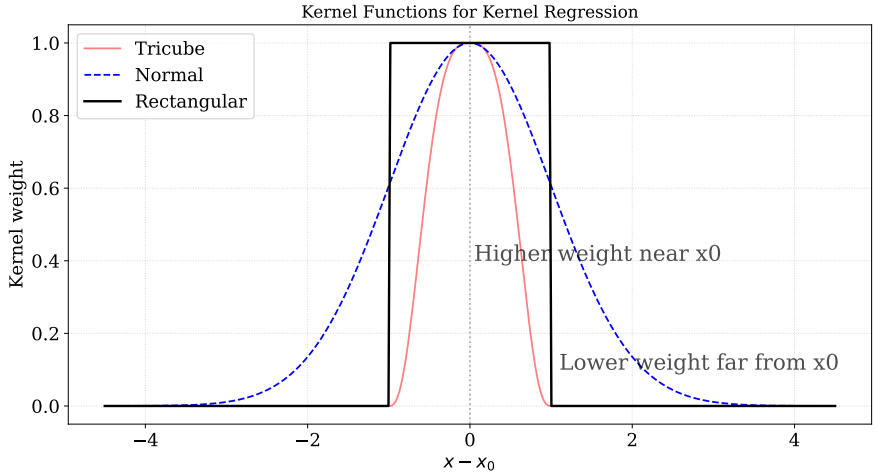
The basic idea of kernel regression is that in estimating the model f, when we want to fit a regression line in the vicnity of x_0, it is desirable to give greater weight to observations that are close to the pivot x_0.

Choosing a Kernel Function

We first stanardize the predictor variable x as $z = \dfrac{x - x_0}{h}$ where x_0 is pivot point, and h is called kernel bandwidth (which is a design choice).

We need a kernel function $\kappa(z)$ that attaches greatest weight to observations that are close to the focal x_0 and then falls off symmetrically and smoothly as $|z|$ grows.

Some Kernel Functions



Gaussian Kernel

One popular kernel function for the LWR is Gaussian function (also known as radial basis function (RBF)):
The radial kernel (Radial Basis Function (RBF)) is a Gaussian function, a popular measure of similarity between two data points, and it decreases with the distance ranging from 0 to 1. 1 means two points are identical.
RBF between two points x and x' is defined as

$$\kappa(x, x_0) = \exp\left(-\frac{||x - x_0||^2}{2h^2}\right)$$

where h is the bandwidth of the kernel function.

The Tricube Kernel

$$\kappa(z) = \begin{cases} (1 - |z|^3)^3 & |z| < 1 \\ 0 & |z| \geq 1 \end{cases}$$

The Rectangular Kernel

$$\kappa(z) = \begin{cases} 1 & |z| < 1 \\ 0 & |z| \geq 1 \end{cases}$$

Steps in Implementing Loess

- Provide smoothing value, h
- Provide degree of $\mathbf{w}$ for the polynomial fit
- Select pivot point(s) x_0.

Additional Reading and Reference

Chapter 18, Applied Regression Analysis and Generalized Linear Models, John Fox, Third Edition, SAGE Publications, Inc.

Classwork

- For Loess, if we choose a different pivot point x_0, what computation can be reused?
- What do we need to store vs what we can compute for a new pivot point?

Model Diagnostics and Assumption Checking

Assumptions in Linear Regression

1. Linearity
2. Additivity
3. Normality of the Error Terms
4. No Multicollinearity among Predictors
5. No Autocorrelation of the Error Terms
6. Homoscedasticity

Residual Plot Analysis

- $e = \hat{y} - y$ vs x is called residual plot.
- Residual plot provides a valuable information about assumption validation.

Linearity Assumption in Linear Regression

Simple Linear Regression assumes that there is linear relationship between the predictors (e.g. independent variables or features) and the response variable (e.g. dependent variable or label).

How to detect the violation of linearity?

- Coefficient of Determination
- Plot the data ☺

Additivity Assumption in Linear Regression

Additivity assumes that the impact of different predictors is additive. For example, consider

$$y = w_0 + w_1 x_1 + w_2 x_2 + \epsilon$$

Additivity says that changing a predictor by some amount will contribute to change in the response in similar proportion.

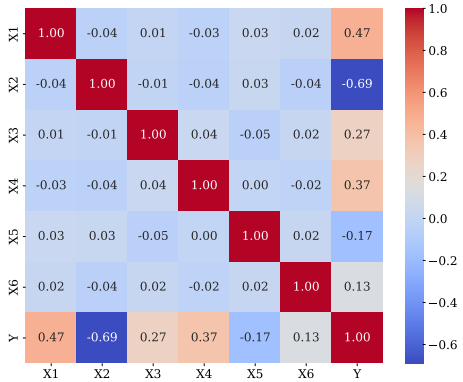
Example: Let $\tilde{x}_2 = x_2 + 1$, then

$$\begin{aligned}
\tilde{y} &= w_0 + w_1 x_1 + w_2 \tilde{x}_2 + \epsilon \\
&= w_0 + w_1 x_1 + w_2 (x_2 + 1) + \epsilon \\
&= (w_0 + w_1 x_1 + w_2 x_2 + \epsilon) + w_2 \\
&= y + w_2
\end{aligned}$$

Hence, we see incremental change in x_2 causes incremental change in y.

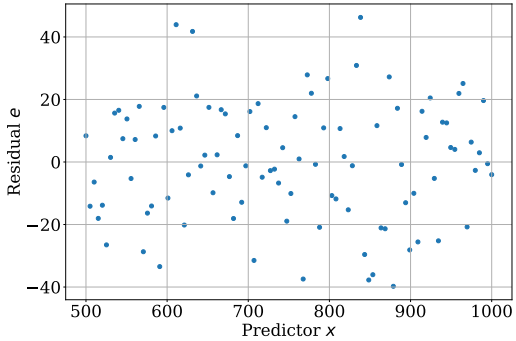
Additivity Violation

Additivity violation occur when there is interaction term in the true model or a higher order term(.e.g. $x_1 x_2$, or x_2^2) that we didn't consider. We can plot correlation matrix to check for additivity violation.



Normality Assumptions

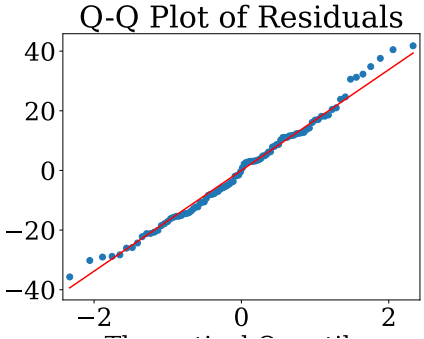
- We assume that error are zero mean constant variance Gaussian distribution.
- Scatter plot of residual, and density estimation plot should exhibit any deviation from normality.



Follows the normality assumption

QQ-plot for Normality

- A Q-Q plot (Quantile-Quantile plot) can help in assessing normality assumption.
- It compares the quantiles of the residuals from the regression model with the quantiles of a theoretical normal distribution.

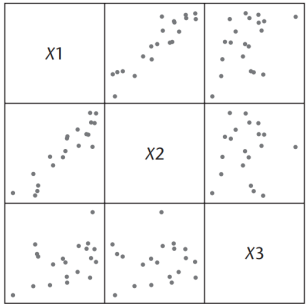


Follows the normality assumption

Multicollinearity

- In case of multiple regression, if predictor variables are correlated, multicollinearity among them is said to exist.
- Plot the scatterplot matrix to check for multicollinearity.

(a) Scatter Plot Matrix of X Variables



(b) Correlation Matrix of X Variables

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.0 & .924 & .458 \\ .924 & 1.0 & .085 \\ .458 & .085 & 1.0 \end{bmatrix}$$

x_1 and x_2 are highly correlated.

Reading: Chapter 7, Applied Linear Regression Models, Kutner, Nachtsheim, and Neter

Autocorrelation

One of the standard assumptions in the regression model is that the error terms ϵ_i and ϵ_j, associated with the ith and jth observations, are uncorrelated.

- Correlation in the error terms suggests that there is additional information in the data that has not been exploited in the current model.
- When the observations have a natural sequential order, the correlation is referred to as **autocorrelation**. One example may be that the data is a timeseries data.
- Or, Adjacent residuals tend to be similar in both temporal and spatial dimensions.

Reference: Chapter 9, Regression Analysis By Example Using R Ali S. Hadi, Samprit Chatterjee

Example: Lack of independence

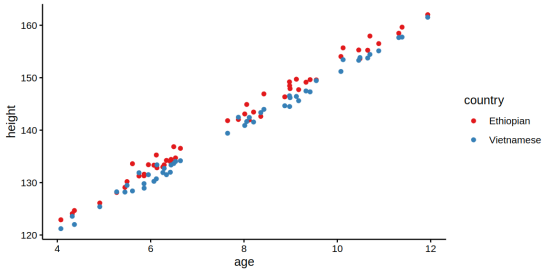


Figure 7.6: Data on age and height in children from two countries.

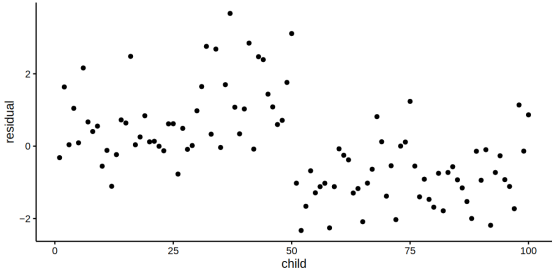


Figure 7.7: Residual plot after regressing height on age.

Two groups stand out

Two groups stand out

Reference: Analysing Data Using Linear Models. (2021). SM van den Berg

Example: Deviation from Linearity

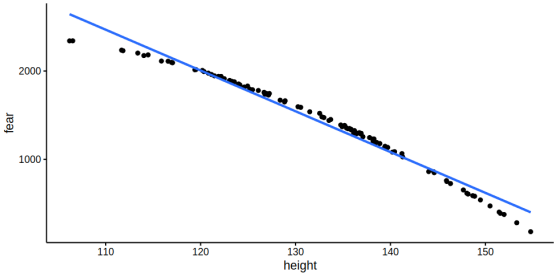


Figure 7.13: Least squares regression line for fear of snakes on height in 100 children.

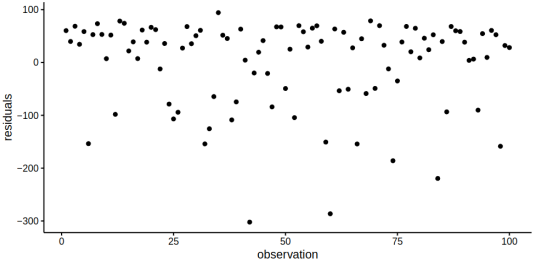


Figure 7.14: Residual plot after regressing fear of snakes on height.

Deviation from Linearity

Residual Plot not symmetric, showing deviation from Linearity

Reference: Analysing Data Using Linear Models. (2021). SM van den Berg

Example: Deviation from Linearity

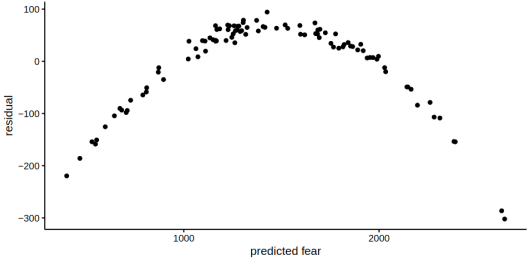


Figure 7.15: Residual plot after regressing fear of snakes on height.

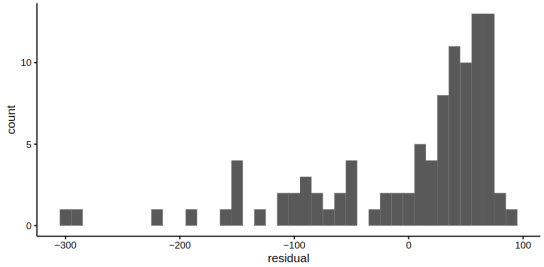


Figure 7.16: Histogram of the residuals after regressing fear of snakes on height.

Deviation from Linearity

Residual Histogram not Gaussian

Reference: Analysing Data Using Linear Models. (2021). SM van den Berg

Remedial Measures

- Lack of Linearity: Use nonlinear regression
- Nonconstancy of Error Variance: Use weighted least square method (or Loess), sometimes transformation are effective too.
- Nonindependence of Error Terms: Autocorrelation may be present. Use transformed variables such as $y(t)_{tr} = y(t) - \rho y(t-1)$
- Nonnormality of Error Terms: Transform data. E.g. $y_{tr} = \sqrt{y}$, $x = \log(x)$, etc.

Reading: Chapter 3, Chapter 12, Applied Linear Regression Models, Kutner, Nachtsheim, and Neter