Modern Regression

Jean-Luc Bouchot

School of Mathematics and Statistics
Beijing Institute of Technology
jlbouchot@bit.edu.cn

Spring 2021

1. Multiple linear regression

Outline

Remark 2.1

Remember that we are now interested in the following problem: given d predictors $x = x_1, \cdots, x_d$ and a dependent variable y, find the $d+1$ coefficients $\beta \in \mathbb{R}^{d+1}$ in the linear regression:

$$y = f_\beta(x) = \beta_0 + \sum_{k=1}^{d} \beta_k x_k.$$

This is called the **Multiple Linear Regression** model.

Example 2.1

Let y denotes the selling price of a house. Define x_1 the number of bedroom, x_2 the number of bathrooms, x_3 the area, x_4 its age, etc ...

Proposition 2.1

The MLR model defined above is indeed linear.

Remark 2.2

One may define degree d polynomial features. The model can still be considered linear.

Remark 2.3

We recall some definitions:

- $X \in \mathbb{R}^{n \times D}$ denotes the data matrix (or sometimes, in some other contexts, sensing matrix or design matrix)
- Each row of the data matrix corresponds to one individual / sample $\mathbf{x}_i \in \mathbb{R}^{1 \times D}$, for $1 \leq i \leq n$. Remember D might be d if we have no bias and d unique features, or $d+1$ if it includes the intercept.
- Each column of the data matrix corresponds to a specific feature.

Remark 2.4

Some setup:

- A noisy measurement process: $y = f_\beta(\mathbf{x}) + \varepsilon$ with ε being a R.V. with a specified distribution (typically, normally distributed)
- Some measurements:

$$y_i = f_\beta(\mathbf{x}_i) + \varepsilon_i = \mathbf{x}_i \beta + \varepsilon_i, \quad \forall 1 \leq i \leq n.$$

- Define the vector of measurements: $\mathbf{y} = [y_1, \cdots, y_n]^T$.
- Fit the parameters according to a least squares criteria:

$$L(X, \mathbf{y}, \beta) = \|X\beta - \mathbf{y}\|_2^2.$$

Proposition 2.2

Let $X \in \mathbb{R}^{n \times D}$ be the data matrix with $n \geq D$ and full rank. Let $\mathbf{y} \in \mathbb{R}^n$ be given. The least squares estimator is given by

$$\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}.$$

The associated predictions are given by

$$\widehat{\mathbf{y}} = X\widehat{\beta} = X(X^T X)^{-1} X^T \mathbf{y}.$$

Remark 2.5

Some important points:

- As with the SLR, the estimated coefficients are expressed as linear combinations of the measured vector $\mathbf{y}$. This is a linear estimator!
- $\widehat{\beta}$ is obtained by solving the **normal equations**!
- The matrix $H = X(X^T X)^{-1} X^T$ such that $\widehat{\mathbf{y}} = H\mathbf{y}$ is called the **Hat Matrix**.

Proposition 2.3

The regression residuals is obtained as

$$\mathbf{e} = (I - H)\,\mathbf{y}.$$

Example 2.2

A delivery company is looking at estimating the time it takes for the delivery rounds. They consider two predictors: x the number of parcels, and y the distance walked by the driver. The have measured the following data:

Time	# parcels	Distance
16.68	7	560
11.50	3	220
12.03	3	340
14.88	4	80
13.75	6	150
18.11	7	330

What are the most influential variables?

Definition 2.1

As for the SLR case, we may define

- *The* **error sum of squares** *or Residual sum of squares, as*

$$SSR = \sum_{i=1}^{n} (\widehat{y_i} - y_i)^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2.$$

- *The* **residual mean square**, *or standard error of regression, as*

$$MSE = \frac{SSR}{n - D}.$$

Proposition 2.4

The standard error of regression is an unbiased estimator of σ^2.