Predictive Analytics
Regression and Classification

Lecture 1 : Part 4

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2019

Regression Model

- Given a vector of inputs $\mathbf{X}_{n\times p} = ((X_{ij}))$, we predict the output $\mathbf{y}$ via model

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}.$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

- $\mathbf{X}_{n\times p}$ known as **design matrix** typically are considered as deterministic and $n > p$.

Model Assumptions

- Given a vector of inputs $\mathbf{X}_{n\times p} = ((X_{ij}))$, we predict the output $\mathbf{y}$ via model

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}.$$

- $\mathbf{X}_{n\times p}$ known as **design matrix** typically are considered as deterministic and $n > p$.

- ϵ, (also known as **error** / **residuals**) for all i are random variables, $i = 1, 2, \cdots, n$
  1. $\mathbb{E}(\epsilon_i) = 0, \forall \; i$
  2. $\mathbb{V}ar(\epsilon_i) = \mathbb{E}(\epsilon_i^2) = \sigma^2, \forall \; i$ Homoscedasticity
  3. $\mathbb{C}ov(\epsilon_i, \epsilon_j) = \mathbb{E}(\epsilon_i\epsilon_j) = 0, \forall \; i \neq j$ Independence

Model Assumptions in Matrix Notation

- Given a vector of inputs $\mathbf{X}_{n\times p} = ((X_{ij}))$, we predict the output $\mathbf{y}$ via model

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}.$$

- $\mathbf{X}_{n\times p}$ known as **design matrix** typically are considered as deterministic.

- ϵ, (also known as **error** / **residuals**) for all i are random variables, $i = 1, 2, \cdots, n$
  1. $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}_n$
  2. $\mathbb{C}ov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$

Implication of the Assumptions

- Assumption:
  1. $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}_n$
  2. $\mathbb{C}ov(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$
- It induces distribution on $\mathbf{y}$, such that

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta} + \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta}$$

  and

$$\mathbb{C}ov(\mathbf{y}) = \mathbb{C}ov(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$$

- Note that we have not made any distributional assumption on $\boldsymbol{\epsilon}$ yet.
- We will introduce that assumption little later.

Implication of the Assumptions

- What is the expected value of $c\mathbf{y}$? If c is a constant.

Result 1 We know

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta},$$

then

$$\mathbb{E}(c\mathbf{y}) = c\mathbf{X}\boldsymbol{\beta}.$$

- Now consider the ordinary least square estimator (OLS) estimator of $\boldsymbol{\beta}$?

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

Result 2 OLS estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Implication of the Assumptions

- Suppose we are interested in some linear combination of the regression corefficients, like $f(\boldsymbol{\beta}) = c^T \boldsymbol{\beta}$.

Result 3 Then the unbiased estimatior of $c^T \boldsymbol{\beta}$ is $c^T \hat{\boldsymbol{\beta}}$, i.e.,

$$\mathbb{E}(c^T \hat{\boldsymbol{\beta}}) = c^T \boldsymbol{\beta},$$

- Suppose $c = x_0$ is a test point. Then we are interested in prediction $f(x_0) = x_0^T \beta$ are of this form.

Gauss Markov Theorem

- If we have any other linear estimator $\tilde{\theta} = a^T \mathbf{y}$ is unbiased for $c^T \boldsymbol{\beta}$, that is
$$\mathbb{E}(a^T \mathbf{y}) = c^T \boldsymbol{\beta},$$
then
$$\mathbb{V}ar(c^T \hat{\boldsymbol{\beta}}) \leq \mathbb{V}ar(a^T \mathbf{y})$$
- Proof is home work problem.

Note OLS estimates of the parameters $\boldsymbol{\beta}$ have the smallest variance among all linear unbiased estimates.

Notes on Gauss Markov Theorem

- Consider the mean squared error (MSE) of an estimator $\tilde{\theta}$ in estimating θ:

$$\begin{aligned} MSE(\tilde{\theta}) &= \mathbb{E}(\tilde{\theta} - \theta)^2 \\ &= \mathbb{V}ar(\tilde{\theta}) + [\mathbb{E}(\tilde{\theta}) - \theta]^2 \\ &= \mathbb{V}ar(\tilde{\theta}) + [bias]^2 \end{aligned}$$

- The Gauss-Markov theorem implies that the least squares estimator has the smallest MSE of all linear estimators with no bias.

- However, there may well exist a biased estimator with smaller MSE. For example: (i) Ridge estimator or (ii) James-Stein shrinkage estimator of β trade a little bias for reduction of variance and its MSE are lowere than the OLS estimator.

Why Mean Square Error?

- **MSE** is directly related to **prediction accuracy**.
- Consider the prediction of the new response at input x_0

$$y_0 = f(x_0) + \epsilon_0.$$

- The expected prediction error of an estimate $\hat{f}(x_0) = x_0^T\hat{\beta}$ is

$$\begin{aligned}\mathbb{E}(y_0 - \hat{f}(x_0))^2 &= \sigma^2 + \mathbb{E}(x_0^T\hat{\beta} - f(x_0))^2 \\ &= \sigma^2 + MSE(x_0^T\hat{\beta})\end{aligned}$$

- Expected prediction error and MSE differ only by the constant σ^2.

In the next part...

- We will discuss the some examples...