Linear Regression

Hongkyu Kim

1 Introduction

Suppose we are given a dataset $\mathcal{D} = \{(x_i, y_i) : x_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i = 1, 2, \cdots, m\}$, and we want to predict y value for the new input value x. One of the most simple way to do so is to start with the linear regression. The core assumption of the linear regression is a linear relation between response (y_i) and predictors (x_i).

$$y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_p x^{(p)}, \tag{1}$$

where $x = (x^{(1)}, x^{(2)}, \cdots, x^{(p)})$. However, it is purely an assumption. In real data, there is inevitably some errors, and in many cases response and predictors are related in a non-linear fashion. Still the linear regression is a good starting point, and may give decent results when there are not many data at hand.

2 Linear model and assumptions

2.1 Model description

The standard linear regression model is

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \cdots + \beta_p x_i^{(p)} + \varepsilon_i, \text{ for } i = 1, 2, \cdots, m \tag{2}$$

with

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ for } i = 1, 2, \cdots, m \text{ and unknown } \sigma^2 > 0. \tag{3}$$

We can write the model in matrix form if we extend predicts as follows.

$$x = (x^{(1)}, x^{(2)}, \cdots, x^{(p)}) \quad \rightarrow \quad x = (1, x^{(1)}, x^{(2)}, \cdots, x^{(p)}). \tag{4}$$

$$Y = X\beta + \varepsilon, \tag{5}$$

where

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(p)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(p)} \\ \vdots & & \ddots & & \vdots \\ 1 & x_m^{(1)} & x_m^{(2)} & \cdots & x_m^{(p)} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_m). \tag{6}$$

2.2 Assumptions

The below are the assumptions of the linear regression model.

- Response and predictors have indeed a linear relationship.
- Predictor values (X) are deterministic.
- Errors ε's are normally distributed with mean 0, and (unknown) variance σ^2, and ε_i's are uncorrelated.
- ε's are independent with β. (ε's are unexplicable errors of the model)
- $m \geq p+1$ and $(X^T X)^{-1}$ exists, or X^{-1} exists, or columns of X are linearly independent.

3 Least squares estimator

Our main interest is twofold: 1) the value of β, and 2) how significant they are. To answer these, we need to have a notion of an "error" of a data point from the linear model. Given the coefficient vector β, we define the residual as

$$r_i = y_i - (\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \cdots + \beta_p x_i^{(p)}), \tag{7}$$

and the "error" of this data point (x_i, y_i) as r_i^2. The total "error" of the data is just the sum of all r_i's, i.e., the residual sum of squares (RSS). Now we find $\hat{\beta}$ that minimizes the RSS, and it is called the **least squares estimator** for an obvious reason.

$$RSS(\beta) = \sum_{i=1}^{m} r_i^2 = ||Y - X\beta||_2^2, \tag{8}$$

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||_2^2. \tag{9}$$

Note that RSS(β) can be expressed as a product of matrices, i.e.,

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta). \tag{10}$$

Using the product rule of differentiation, and the fact that

$$\frac{\partial}{\partial \beta}(X\beta)^T = X^T \text{ and } \frac{\partial}{\partial \beta^T}(X\beta) = X, \tag{11}$$

we get

$$\frac{\partial}{\partial \beta} RSS(\beta) = -2X^T(Y - X\beta) \quad \text{and} \quad \frac{\partial}{\partial \beta^T \partial \beta} RSS(\beta) = 2X^T X. \tag{12}$$

Since we assume that X has full column rank (i.e., columns are linearly independent), $2X^T X$ is positive definite, and therefore $RSS(\beta)$ achieves the global minimum at which $\frac{\partial}{\partial \beta} RSS(\beta) = 0$. In other words,

$$-2X^T(Y - X\hat{\beta}) = 0 \quad \rightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T Y. \tag{13}$$

If we are given a new input x_0, we can estimate the corresponding response as

$$\hat{y_0} = x_0^T \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_0^{(1)} + \cdots + \hat{\beta}_p x_0^{(p)}. \tag{14}$$

4 Inference about coefficients

One of the assumptions of the linear regression states about the distribution of errors ε that they are normally distributed with mean $\mathbf{0}$ and covariance $\sigma^2 I_m$. We can use this assumption to get the distribution of $\hat{\beta}$.

$$\hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) = \beta + (X^T X)^{-1} X^T \varepsilon, \tag{15}$$

hence,

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}). \tag{16}$$

Especially, we have

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 (X^T X)^{-1}_{jj}) \text{ for } j = 0, 1, \cdots, p \tag{17}$$

with indexing of $(X^T X)^{-1}$ starts with 0. We can now test whether a certain coefficient is 0 or not with this distribution.

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0. \tag{18}$$

With the null hypothesis H_0,

$$\frac{\hat{\beta}_j}{\sigma \sqrt{(X^T X)^{-1}_{jj}}} \sim \mathcal{N}(0, 1). \tag{19}$$

However, the true variance of the error ε is unknown, so we have to use an estimator instead. An estimator for the variance $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{1}{m - p - 1} \sum_{i=1}^{m} (y_i - \hat{y_i})^2 = \frac{1}{m - p - 1} ||Y - X\hat{\beta}||_2^2. \tag{20}$$

It is known that

$$\frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}}} \sim t_{m-p-1} \tag{21}$$

where t_{m-p-1} follows the t-distribution with $(m - p - 1)$ degrees of freedom. p value of H_0 can be calculated as

$$p = \Pr\left(T \geq \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}}} \right) \tag{22}$$

where T follows t-distribution with $(m - p - 1)$ degrees of freedom.
We can even get a $(1 - 2\alpha)$ confidence interval for β_j:

$$\left(\hat{\beta}_j - \tau_{(m-p-1),(1-\alpha)} \hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}}, \hat{\beta}_j + \tau_{(m-p-1),(1-\alpha)} \hat{\sigma} \sqrt{(X^T X)^{-1}_{jj}} \right) \tag{23}$$

with $\tau_{(m-p-1),(1-\alpha)}$ being the $(1 - \alpha)$ quantile of t-distribution with $(m - p - 1)$ degrees of freedom.

5 More about least squares estimator

Residual sum of squares is not the only "error" one can impose. Actually, there are many other possible "errors" and estimators. Nonetheless, least squares estimator is popular because of its nice properties. For example, it is easy to get the estimator just by differentiating the RSS, and there is a nice closed-form formula. Furthermore, there is Gauss-Markov theorem that implies that Least squares estimator of β has the "smallest" covariance among all linear unbiased estimator of β. Therefore, each parameter has the lowest mean squared error (MSE).

$$\begin{aligned} \mathrm{MSE}[\tilde{\theta}] &= \mathbb{E}[(\theta - \tilde{\theta})^2] && (24)\\ &= \mathrm{Var}(\tilde{\theta}) + (\mathbb{E}\tilde{\theta} - \theta)^2 && (25)\\ &= \mathrm{Var}(\tilde{\theta}) + \mathrm{bias}^2(\tilde{\theta}). && (26) \end{aligned}$$

As one can expect, least squares estimator is not always the best estimator. MSE can be reduced further with a biased estimator by significantly reducing the variance of the estimator.

MSE is also closely related to prediction accuracy. Suppose we have a new input x_0, and corresponding true response $y_0 = f(x_0) + \varepsilon_0$. Then the expected prediction error is

$$\begin{aligned} \mathbb{E}[(\tilde{f}(x_0) - y_0)^2] &= \mathbb{E}[(\tilde{f}(x_0) - f(x_0) - \varepsilon_0)^2] && (27)\\ &= \mathbb{E}[\varepsilon_0^2] + \mathbb{E}[(\tilde{f}(x_0) - f(x_0))^2] && (28)\\ &= \sigma^2 + \mathrm{MSE}[\tilde{f}(x_0)] && (29) \end{aligned}$$

with $f(x_0) = x_0^T\beta$ and $\tilde{f}(x_0) = x_0^T\tilde{\beta}$. Therefore, the prediction accuracy has constant difference (σ^2) with the MSE of the estimated value.