# Linear Regression

Hanxiao Du

## Contents



## 1 The 1-D case

Our goal is to learn a mapping $y = f(x)$, where $x$ and $y$ are both real-valued scalars (i.e. $x \in \mathbb{R},\ y \in \mathbb{R}$). Take $f$ to be an linear function (actually an affine function) of the form:

$$y = wx + b \tag{1}$$

where $w$ is a weight and $b$ is a bias.
We wish to estimate $w$ and $b$ from the $N$ training pairs $\{(x_i, y_i)\}_{i=1}^N$. Then, once we have to estimates of $w$ and $b$, we can compute $y$ for some new $x$.
We would like to find the parameters (i.e. $w$ and $b$) such that minimize the residual errors (i.e. $y_i - f(x_i) = y_i - (wx_i + b)$).
The most commonly-used way to estimate the parameters is by least-square regression/estimation or ordinary least square (OLS) estimation. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares. We define an energy function (a.k.a. objective function):

$$E(w, b) = \sum_{i=1}^{N} (y_i - (wx_i + b))^2 \tag{2}$$

This is merely the summation of the squared residual errors. Here, the reasons of using squared residual instead of summing all residuals directly or summing all the absolute values of the residuals are:

1. It is mathematically easier to find the derivatives of the objective function so that the optimization process is much easier.

2. By Gauss-Markov theorem, assuming the Gauss-Markov assumptions, the OLS estimator has the lowest sampling variance within the class of linear unbiased estimators (most efficient). The Gauss-Markov assumptions are:

- The residuals $\epsilon_i$ have mean zero, i.e. $\mathbb{E}[\epsilon_i] = 0$
- The residuals are homoscedastic, i.e. they all have the same finite variance:

$$Var(\epsilon_i) = \sigma^2 < \infty$$

- Distinct error terms are uncorrelated, i.e. $Cov(\epsilon_i, \epsilon_j) = 0 \ \forall i \neq j$.

Proof of Gauss-Markov theorem can be found in Section 4.1

3. The OLS estimator is equivalent to the maximum likelihood estimator (MLE) under the normality assumption for the error terms.

Thus we would like to find $\hat{w}_{OLS}$ and $\hat{b}_{OLS}$ that minimize $E$. That is:

$$\hat{b}_{OLS} = \underset{b}{\operatorname{argmin}} \, E(w, b) \tag{3}$$

$$\hat{w}_{OLS} = \underset{w}{\operatorname{argmin}} \, E(w, \hat{b}_{OLS}) \tag{4}$$

Here, we can use the derivative test to find the global extrema.
Solve for $\hat{b}_{OLS}$:

$$\frac{\partial E}{\partial b} = \frac{\partial \sum_{i=1}^{N} (y_i - (wx_i + b))^2}{\partial b} \tag{5}$$

$$= \sum_{i=1}^{N} -2(y_i - (wx_i + b)) = 0 \tag{6}$$

$$\implies \sum_{i=1}^{n} y_i - w \sum_{i=1}^{n} x_i - Nb = 0 \tag{7}$$

$$\implies \hat{b}_{OLS} = \frac{\sum_{i=1}^{n} y_i - w \sum_{i=1}^{n} x_i}{N} \tag{8}$$

$$\implies \hat{b}_{OLS} = \bar{y} - w\bar{x} \tag{9}$$

where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{N}$ is the average of $y_i$'s, and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$ is the average of $x_i$'s.
Notice that we did not check if $\frac{\partial^2 E}{\partial b^2} > 0$ to ensure that $\hat{b}_{OLS}$ minimizes $E$, this is because we know that $E$ is a quadratic function of $b$ and we can conclude that $\hat{b}_{OLS}$ minimizes $E$ by just inspecting $E$. The same principle applied in finding $\hat{w}_{OLS}$.
Solve for $\hat{w}_{OLS}$:

$$\frac{\partial E}{\partial w} = \frac{\partial \sum_{i=1}^{N}(y_i - (wx_i + \hat{b}_{OLS}))^2}{\partial w} \tag{10}$$

$$= \frac{\partial \sum_{i=1}^{N}(y_i - (wx_i + \bar{y} - w\bar{x}))^2}{\partial w} \tag{11}$$

$$= \frac{\partial \sum_{i=1}^{N}(y_i - \bar{y} - w(x_i - \bar{x}))^2}{\partial w} \tag{12}$$

$$= -2\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y} - w(x_i - \bar{x})) = 0 \tag{13}$$

$$\implies \sum_{i=1}^{N} x_i y_i - \bar{y}\sum_{i=1}^{N} x_i - \bar{x}\sum_{i=1}^{N} y_i + \bar{x}\bar{y} - w\sum_{i=1}^{N}(x_i - \bar{x})^2 = 0 \tag{14}$$

$$\implies \hat{w}_{OLS} = \frac{\sum_{i=1}^{N}(x_i y_i - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{15}$$

$$\implies \hat{w}_{OLS} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{16}$$

Notice that for this equation, the $b$ in $E$ is dependent to $w$, thus we need to express $b$ in terms of $w$ in order to find the derivative of $w$ properly.

# 2 Multidimensional Inputs

For a $D$-dimensional linear regression, we can express the linear model in terms of matrices:

$$f(X) = X\beta \tag{17}$$

where $\beta$ is a $(D+1) \times 1$ matrix (or a column vector) of parameters in the form

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix} \tag{18}$$

$\beta_0$ is the intercept term and $\beta_i \; \forall i \in [D]$ is the coefficient of the i-th attribute/feature. $X$ is a $N \times (D+1)$ design matrix of the form:

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(D)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(D)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N^{(1)} & x_N^{(2)} & \cdots & x_N^{(D)} \end{bmatrix} \tag{19}$$

where $x_i^{(j)} \; \forall i \in [N] \; \forall j \in [D]$ denotes the j-th attribute of the i-th instance. Notice that all the values in the first column of $X$ are all 1's, this is because when we do the matrix multiplication $X\beta$, the intercept term $\beta_0$ is added to each column of the product. i.e. this is a dummy column for the intercept term.

The least-squares objective function is then $E(\beta) = ||y - X\beta||_2^2$ if we treat $y$ and $X\beta$ as column vectors:

$$X\beta = \begin{bmatrix} \sum_{j=1}^{D} \beta_j x_1^{(j)} + \beta_0 \\ \sum_{j=1}^{D} \beta_j x_2^{(j)} + \beta_0 \\ \vdots \\ \sum_{j=1}^{D} \beta_j x_N^{(j)} + \beta_0 \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{20}$$

Notice that

$$y^T X\beta = (y^T X\beta)^T = (X\beta)^T y = \beta^T X^T y \tag{21}$$

since $y^T X\beta$ is just a $1 \times 1$ matrix, thus symmetric. Therefore, the squared Euclidean norm

$$\begin{aligned} E(\beta) &= ||y - X\beta||_2^2 && (22)\\ &= (y - X\beta)^T (y - X\beta) && (23)\\ &= y^T y - y^T X\beta - (X\beta)^T y + (X\beta)^T X\beta && (24)\\ &= \beta^T X^T X\beta - 2\beta^T X^T y + y^T y \text{ by (21)} && (25) \end{aligned}$$

The derivatives of matrices are calculated by scalar-by-matrix derivative. This is a derivative formula sheet:
Matrix_derivatives_cribsheet.pdf
Now, solve for $\hat{\beta}_{OLS}$ by using the derivative test:

$$\begin{aligned} \frac{\partial E}{\partial \beta} &= 2X^T X\beta - 2X^T y = 0 && (26)\\ \implies \hat{\beta}_{OLS} &= (X^T X)^{-1} X^T y && (27) \end{aligned}$$

Or equivalently, $\hat{\beta}_{OLS} = X^+ y$, where $X^+ = (X^T X)^{-1} X^T$ is called the pseudo-inverse of $X$. This is also a projection matrix.

# 3 Multidimensional Outputs

In the most general case, both the inputs and outputs may be multidimensional.

# 4 Appendix

## 4.1 Gauss-Markov Theorem

**Theorem 1** (Gauss-Markov theorem)**.** *Suppose we have a linear model $y = X\beta + \epsilon$, where $\epsilon$ is a column vector containing all residuals $\epsilon_i$'s.*
*Under the Gauss-Markov assumptions,*

1. *The residuals $\epsilon_i$ have mean zero, i.e. $\mathbb{E}[\epsilon_i] = 0$ or $\mathbb{E}[\epsilon] = \vec{0}$.*

2. *They are homoscedastic, i.e. they all have the same finite variance:*

$$Var(\epsilon_i) = \sigma^2 < \infty$$

3. *Distinct error terms are uncorrelated, i.e.* $Cov(\epsilon_i, \epsilon_j) = 0 \ \forall i \neq j$.

*or equivalently to assumption 2 and 3,* $Var(\epsilon) = \sigma^2 I$ *where* $\sigma^2 < \infty$,
*we have that the ordinary least squares (OLS) of the parameter:*

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$$

*is the BLUE (Best Linear Unbiased Estimator), i.e. it is the most efficient estimator (has lowest variance) among all linear unbiased estimators.*

*Proof.*

$$\begin{aligned}
\hat{\beta}_{OLS} &= (X^T X)^{-1} X^T y && (28)\\
&= (X^T X)^{-1} X^T (X\beta + \epsilon) && (29)\\
&= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \epsilon && (30)\\
&= \beta + (X^T X)^{-1} X^T \epsilon && (31)
\end{aligned}$$

Thus, the unbiasedness of $\hat{\beta}_{OLS}$ gives:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_{OLS}] &= \mathbb{E}[\beta + (X^T X)^{-1} X^T \epsilon] && (32)\\
&= \beta + (X^T X)^{-1} X^T \mathbb{E}[\epsilon] && (33)\\
&= \beta, \text{ since by assumption 1, } \mathbb{E}[\epsilon] = \vec{0}. && (34)
\end{aligned}$$

Also by assumption 1,

$$Var(\epsilon) = \mathbb{E}[(\epsilon - \mathbb{E}[\epsilon])(\epsilon - \mathbb{E}[\epsilon])^T] = \mathbb{E}[\epsilon\epsilon^T] = \sigma^2 I \tag{35}$$

Thus,

$$Var(y) = Var(X\beta + \epsilon) = Var(\epsilon) = \sigma^2 I \tag{36}$$

Now, we calculate the variance of the $\hat{\beta}_{OLS}$:

$$\begin{aligned}
Var(\hat{\beta}_{OLS}) &= \mathbb{E}[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)^T] && (37)\\
&= \mathbb{E}[((X^T X)^{-1} X^T \epsilon)((X^T X)^{-1} X^T \epsilon)^T] && (38)\\
&= \mathbb{E}[(X^T X)^{-1} X^T \epsilon\epsilon^T X (X^T X)^{-1}] && (39)\\
&= (X^T X)^{-1} X^T \mathbb{E}[\epsilon\epsilon^T] X (X^T X)^{-1} && (40)\\
&= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \text{ by (35)} && (41)\\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} && (42)\\
&= \sigma^2 (X^T X)^{-1} && (43)
\end{aligned}$$

Now, for any linear unbiased estimator $\hat{\beta}$ for $\beta$, it must can be expressed in linear form:

$$\hat{\beta} = Cy$$

for some constant matrix $C$.
Also, let $D$ be the matrix such that $C = (X^T X)^{-1} X^T + D$.
Additionally, by the unbiasedness assumption of the estimator $\hat{\beta}$:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[Cy] && (44)\\
&= \mathbb{E}[((X^T X)^{-1} X^T + D)(X\beta + \epsilon)] && (45)\\
&= \beta + DX\beta = \beta && (46)\\
\implies DX\beta = \underset{\sim}{0} &\implies DX = \underset{\sim}{0} && (47)
\end{aligned}$$

where $\underset{\sim}{0}$ is the zero matrix.
Another fact from linear algebra:
Finally, calculate the variance of $\hat{\beta}$:

$$\begin{aligned}
Var(\hat{\beta}) &= Var[Cy] && (48)\\
&= CVar(y)C^T && (49)\\
&= \sigma^2 CC^T \text{ by (36)} && (50)\\
&= \sigma^2[(X^T X)^{-1} X^T + D][X(X^T X)^{-1} + D^T] && (51)\\
&= \sigma^2[(X^T X)^{-1} X^T X (X^T X)^{-1} + (X^T X)^{-1}(DX)^t] && (52)\\
&+ DX(X^T X)^{-1} + DD^T] && (53)\\
&= \sigma^2 (X^T X)^{-1} + \sigma^2 DD^T \text{ by (47)} && (54)\\
&= Var(\hat{\beta}_{OLS}) + \sigma^2 DD^T && (55)
\end{aligned}$$

notice that $\sigma^2 > 0$ and $DD^T$ is definite symmetric.
Therefore, $\hat{\beta}_{OLS}$ has the lowest variance among all linear unbiased estimator so that it is the BLUE (Best Linear Unbiased Estimator).

■