Proof of Normal Equation of Simple Linear Regression and Multiple Linear Regression

Simple Linear Regression Proof

Goal: We consider the simple linear regression problem, where we want to estimate parameters

$$\beta_0 \quad \text{and} \quad \beta_1$$

so that the model

$$y_i \approx \beta_0 + \beta_1 x_i$$

fits the given data points $\{(x_i, y_i)\}_{i=1}^n$. We define the sum of squared errors (SSE) as

$$\text{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2.$$

1 Finding the Critical Point

To find the critical point, we take partial derivatives of SSE with respect to β_0 and β_1 and set them to zero.

$$\frac{\partial \text{SSE}}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right) = 0,$$
$$\frac{\partial \text{SSE}}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - (\beta_0 + \beta_1 x_i)\right) = 0.$$

Hence, we have the following system of equations:

$$\begin{cases} \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \displaystyle\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0. \end{cases}$$

Solving for β_0 and β_1 leads to the well-known least squares estimators:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \beta_0 = \bar{y} - \beta_1 \bar{x},$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

2 Hessian Matrix and Determinant

To show that this critical point is unique and corresponds to a global minimum, we consider the Hessian matrix of SSE. The Hessian H is the matrix of second partial derivatives:

$$H = \begin{pmatrix} \dfrac{\partial^2 \mathrm{SSE}}{\partial \beta_0^2} & \dfrac{\partial^2 \mathrm{SSE}}{\partial \beta_0 \partial \beta_1} \\ \dfrac{\partial^2 \mathrm{SSE}}{\partial \beta_1 \partial \beta_0} & \dfrac{\partial^2 \mathrm{SSE}}{\partial \beta_1^2} \end{pmatrix}.$$

By direct calculation, we have

$$\frac{\partial^2 \mathrm{SSE}}{\partial \beta_0^2} = 2n, \quad \frac{\partial^2 \mathrm{SSE}}{\partial \beta_0 \partial \beta_1} = \frac{\partial^2 \mathrm{SSE}}{\partial \beta_1 \partial \beta_0} = 2\sum_{i=1}^{n} x_i, \quad \frac{\partial^2 \mathrm{SSE}}{\partial \beta_1^2} = 2\sum_{i=1}^{n} x_i^2.$$

Thus,

$$H = \begin{pmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix}.$$

Determinant of Hessian H

The determinant of H is

$$\det(H) = \begin{vmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{vmatrix} = 4 \begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}.$$

Hence,

$$\det(H) = 4\Big(n\sum_{i=1}^{n} x_i^2 - \big(\sum_{i=1}^{n} x_i\big)^2\Big).$$

One can also see from sample variance identities that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}.$$

If not all x_i are identical, then $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$. This implies

$$n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 > 0,$$

so $\det(H) > 0$. In addition, because each diagonal entry of H is positive and $\det(H) > 0$, H is a positive-definite matrix.

3 Global Minimum

Since $\mathrm{SSE}(\beta_0, \beta_1)$ is a quadratic form in β_0 and β_1, and its Hessian H is positive-definite (as long as the x_i are not all identical), there is exactly one critical point, and it must be a global minimum.

Theorem 1 (Global Minimum of Simple Linear Regression). *Suppose $\{x_i\}$ are not all the same value. Then the unique critical point given by*

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

is the global minimizer of the sum of squared errors $SSE(\beta_0, \beta_1)$.

Multiple Linear Regression via the Best Approximation Theorem

4 Problem Setup

Let X be an $m \times n$ matrix, $\beta \in \mathbb{R}^{n\times 1}$, and $y \in \mathbb{R}^{m\times 1}$. In multiple linear regression (MLR), we seek β that minimizes the Euclidean distance

$$\|X\beta - y\|,$$

where $\|\cdot\|$ denotes the standard Euclidean norm on $\mathbb{R}^m$. Equivalently, we want to solve:

$$\min_{\beta \in \mathbb{R}^n} \|X\beta - y\|.$$

We will use a linear-algebraic approach involving the best approximation theorem and the fundamental theorem of linear algebra.

5 Key Linear-Algebraic Facts

5.1 Column Space and Null Space

Recall that:

$$\text{col}(X) = \{Xv : v \in \mathbb{R}^n\}, \quad \text{null}(X^T) = \{w \in \mathbb{R}^m : X^T w = 0\}.$$

The *Fundamental Theorem of Linear Algebra* states that:

$$\mathbb{R}^m = \text{col}(X) \oplus \text{null}(X^T),$$

i.e., every vector $v \in \mathbb{R}^m$ can be written uniquely as a sum of something in $\text{col}(X)$ and something in $\text{null}(X^T)$. Furthermore,

$$\text{col}(X)^\perp = \text{null}(X^T),$$

meaning the subspace $\text{col}(X)$ is orthogonal to $\text{null}(X^T)$.

Proof. Sketch of why $[\mathbb{R}^m = \text{col}(X) \oplus \text{null}(X^T)]$

- Let $v \in \text{col}(X)$. Then $v = Xc$ for some $c \in \mathbb{R}^n$.
- Let $w \in \text{null}(X^T)$. Then $X^T w = 0$.

- Note that
$$\langle v, w\rangle = \langle Xc, w\rangle = c^T(X^T w) = c^T(0) = 0,$$
which shows that any vector in $\mathrm{col}(X)$ is orthogonal to any vector in $\mathrm{null}(X^T)$.

- The rank-nullity theorem implies
$$\dim(\mathrm{col}(X)) + \dim(\mathrm{null}(X^T)) = m,$$
so these two spaces form a direct sum decomposition of $\mathbb{R}^m$.

- Orthogonality further implies $\mathrm{col}(X)^\perp = \mathrm{null}(X^T)$.

□

6 Best Approximation Approach to MLR

We want to minimize
$$\|X\beta - y\|.$$
From the fundamental theorem, we can decompose any $y \in \mathbb{R}^m$ uniquely as
$$y = y_c + y_n, \quad \text{where } y_c \in \mathrm{col}(X) \text{ and } y_n \in \mathrm{null}(X^T).$$
Hence,
$$\|X\beta - y\| = \|X\beta - (y_c + y_n)\| = \|(X\beta - y_c) - y_n\|.$$
Observe that $X\beta$ always lies in $\mathrm{col}(X)$. So if we seek the "best approximation" of y by vectors in $\mathrm{col}(X)$, the optimal choice is to match the part of y that lies in $\mathrm{col}(X)$, i.e.,
$$X\beta = y_c.$$
Then the difference $X\beta - y_c$ is $\mathbf{0}$, so
$$\|X\beta - y\| = \|0 - y_n\| = \|y_n\|,$$
which is the smallest possible distance we can achieve (because y_n is orthogonal to $\mathrm{col}(X)$ and cannot be "canceled" by any choice of $X\beta$).

6.1 Normal Equations

The vector y_c is the orthogonal projection of y onto $\mathrm{col}(X)$. Mathematically, we enforce orthogonality to the null space of X^T:
$$y - X\beta \in \mathrm{null}(X^T) \quad \Longleftrightarrow \quad X^T(y - X\beta) = 0.$$
Thus we get the *normal equations*:
$$X^T X\beta = X^T y.$$
If $X^T X$ is invertible (for example, if X has full column rank), then we solve for β:
$$\beta = (X^T X)^{-1} X^T y.$$
This β is the unique least-squares solution that minimizes $\|X\beta - y\|$, and hence it is the *best approximation* of y by columns of X.

Theorem 2 (Uniqueness and Existence)**.** *If X is an $m \times n$ matrix of full column rank (i.e., rank n), then $X^T X$ is invertible, and the solution $\beta = (X^T X)^{-1} X^T y$ is the unique least-squares solution for* $\min_\beta \|X\beta - y\|$.

7 Summary

Using the decomposition $\mathbb{R}^m = \mathrm{col}(X) \oplus \mathrm{null}(X^T)$ and the orthogonality principle, we can interpret the least-squares solution as the orthogonal projection of y onto $\mathrm{col}(X)$. Consequently, the normal equations

$$X^T(y - X\beta) = 0 \quad \Longleftrightarrow \quad X^T y = X^T X\beta$$

guarantee that $y - X\beta$ lies in $\mathrm{null}(X^T)$. Solving for β yields

$$\beta = (X^T X)^{-1} X^T y,$$

which exactly matches the final formula shown in the referenced image:

$$X^\top y = X^\top X\,\beta \quad \Longrightarrow \quad \beta = (X^\top X)^{-1} X^\top y.$$