Least Squares Estimation: A Linear Algebra Perspective

Notes

April 2, 2025

1 The Problem of Best Approximation

Often in science and engineering, we encounter systems described by linear models that don't perfectly fit observed data. Consider a linear transformation $T : \mathbb{R}^p \to \mathbb{R}^n$, represented by a matrix $X \in \mathbb{R}^{n\times p}$. We might hypothesize a relationship $Y \approx T\beta$ (or $Y \approx X\beta$ in matrix form), where $Y \in \mathbb{R}^n$ is a vector of observations and $\beta \in \mathbb{R}^p$ is a vector of unknown parameters we wish to estimate.

Since the system might be overdetermined ($n > p$) or subject to noise, there might be no β for which $Y = T\beta$ holds exactly. That is, Y may not lie in the image (column space) of T, $\mathrm{Im}(T)$. The core idea of least squares is to find the vector $\hat{\beta}$ such that $T\hat{\beta}$ is the *closest* vector within $\mathrm{Im}(T)$ to the observed vector Y. Closeness is measured using the standard Euclidean distance (or norm).

Definition 1.1 (Least Squares Problem)**.** Given a linear operator $T : \mathbb{R}^p \to \mathbb{R}^n$ (represented by matrix $X \in \mathbb{R}^{n\times p}$) and a vector $Y \in \mathbb{R}^n$, the least squares problem is to find a vector $\hat{\beta} \in \mathbb{R}^p$ that minimizes the squared Euclidean norm of the residual vector $r = Y - T\beta$:

$$\min_{\beta\in\mathbb{R}^p} \|Y - T\beta\|^2$$

2 Derivation of the Least Squares Solution

Let's analyze the objective function $f(\beta) = \|Y - T\beta\|^2$. Using the definition of the norm via the standard inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^n$:

$$\begin{aligned} f(\beta) = \|Y - T\beta\|^2 &= \langle Y - T\beta, Y - T\beta\rangle \\ &= \langle Y, Y\rangle - \langle Y, T\beta\rangle - \langle T\beta, Y\rangle + \langle T\beta, T\beta\rangle \\ &= \|Y\|^2 - 2\langle Y, T\beta\rangle + \|T\beta\|^2 \end{aligned}$$

Now, we introduce the adjoint operator $T^* : \mathbb{R}^n \to \mathbb{R}^p$, which satisfies $\langle v, Tu\rangle = \langle T^*v, u\rangle$ for all $u \in \mathbb{R}^p, v \in \mathbb{R}^n$. In matrix terms, if T is represented by X, then T^* is represented by the transpose X^T (for real vector spaces with the standard inner product). Applying this property:

$$f(\beta) = \|Y\|^2 - 2\langle T^*Y, \beta\rangle + \langle T^*T\beta, \beta\rangle$$

Remark 2.1 (Gradient Identities)*.* Before computing the gradient of $f(\beta)$, let's recall or prove two useful gradient identities involving the standard inner product on $\mathbb{R}^p$. Let $a, \beta \in \mathbb{R}^p$ and $M \in \mathbb{R}^{p\times p}$.

1. **Gradient of a linear function:** $\nabla_\beta\langle a, \beta\rangle = a$.
   Proof: Let $\beta = [\beta_1, \ldots, \beta_p]^T$ and $a = [a_1, \ldots, a_p]^T$. The inner product is $\langle a, \beta\rangle = \sum_{i=1}^p a_i\beta_i$. The gradient is the vector of partial derivatives:

$$\nabla_\beta\langle a, \beta\rangle = \begin{bmatrix} \frac{\partial}{\partial\beta_1}(\sum_i a_i\beta_i) \\ \vdots \\ \frac{\partial}{\partial\beta_p}(\sum_i a_i\beta_i) \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} = a.$$

2. **Gradient of a quadratic form:** $\nabla_\beta \langle M\beta, \beta\rangle = (M + M^T)\beta$.

   Proof: The quadratic form is $\langle M\beta, \beta\rangle = \sum_{i=1}^p (M\beta)_i \beta_i = \sum_{i=1}^p \left(\sum_{j=1}^p M_{ij}\beta_j\right)\beta_i = \sum_{i=1}^p \sum_{j=1}^p M_{ij}\beta_i\beta_j$. We compute the partial derivative with respect to β_k:

$$\begin{aligned}
\frac{\partial}{\partial \beta_k}\left(\sum_{i=1}^p \sum_{j=1}^p M_{ij}\beta_i\beta_j\right) &= \sum_{i=1}^p \sum_{j=1}^p M_{ij}\frac{\partial}{\partial \beta_k}(\beta_i\beta_j) \\
&= \sum_{i=1}^p \sum_{j=1}^p M_{ij}(\delta_{ik}\beta_j + \beta_i\delta_{jk}) \quad \text{(where } \delta_{xy} \text{ is the Kronecker delta)} \\
&= \sum_{j=1}^p M_{kj}\beta_j + \sum_{i=1}^p M_{ik}\beta_i
\end{aligned}$$

   The first term, $\sum_{j=1}^p M_{kj}\beta_j$, is the k-th component of the vector $M\beta$. The second term, $\sum_{i=1}^p M_{ik}\beta_i$, is the k-th component of the vector $M^T\beta$. Therefore, the gradient vector is:

$$\nabla_\beta \langle M\beta, \beta\rangle = M\beta + M^T\beta = (M + M^T)\beta.$$

   Note: If M is symmetric ($M = M^T$), then $\nabla_\beta \langle M\beta, \beta\rangle = 2M\beta$.

To find the minimum of $f(\beta)$, we compute its gradient with respect to β and set it to the zero vector. Using the identities from Remark 2.1, specifically $\nabla_\beta \langle a, \beta\rangle = a$ with $a = T^*Y$, and $\nabla_\beta \langle M\beta, \beta\rangle = (M + M^T)\beta$ with $M = T^*T$:

$$\nabla_\beta f(\beta) = \nabla_\beta(\|Y\|^2 - 2\langle T^*Y, \beta\rangle + \langle T^*T\beta, \beta\rangle)$$

Since $\|Y\|^2$ is constant with respect to β, its gradient is zero.

$$\nabla_\beta f(\beta) = -2\nabla_\beta \langle T^*Y, \beta\rangle + \nabla_\beta \langle (T^*T)\beta, \beta\rangle$$

Applying the identities (and noting that $M = T^*T$ is symmetric, so $M^T = M$):

$$\nabla_\beta f(\beta) = -2(T^*Y) + (T^*T + (T^*T)^T)\beta = -2T^*Y + 2(T^*T)\beta$$

Setting the gradient to zero gives the fundamental equation:

$$-2T^*Y + 2T^*T\beta = 0 \implies T^*T\beta = T^*Y$$

This is known as the **normal equation(s)**.

Theorem 2.2 (Least Squares Solution)**.** *Suppose the linear operator $T : \mathbb{R}^p \to \mathbb{R}^n$ has full column rank, meaning* $\operatorname{rank}(T) = p$ *(or equivalently, the columns of its matrix representation X are linearly independent). Then the operator $T^*T : \mathbb{R}^p \to \mathbb{R}^p$ is invertible. The unique least squares solution $\hat{\beta}$ that minimizes $\|Y - T\beta\|^2$ is given by:*

$$\hat{\beta} = (T^*T)^{-1}T^*Y$$

In matrix notation:

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

Proof. If T has full column rank p, its null space is trivial: $\operatorname{null}(T) = \{0\}$. We show that this implies T^*T is invertible. Suppose $(T^*T)u = 0$ for some $u \in \mathbb{R}^p$. Then $\langle (T^*T)u, u\rangle = \langle 0, u\rangle = 0$. But $\langle (T^*T)u, u\rangle = \langle Tu, Tu\rangle = \|Tu\|^2$. So, $\|Tu\|^2 = 0$, which implies $Tu = 0$. Since $\operatorname{null}(T) = \{0\}$, we must have $u = 0$. Thus, the null space of T^*T is also trivial, $\operatorname{null}(T^*T) = \{0\}$. Since T^*T maps $\mathbb{R}^p$ to $\mathbb{R}^p$ and is injective, it must be invertible.

The normal equation $T^*T\beta = T^*Y$ arises from setting the gradient of the objective function to zero. Since the Hessian of $f(\beta)$ is $2T^*T$, which is positive definite under the full rank assumption (as shown by $\langle (T^*T)u, u\rangle = \|Tu\|^2 > 0$ for $u \neq 0$), the solution to the normal equation corresponds to a unique minimum. Multiplying by the inverse $(T^*T)^{-1}$ gives the unique solution $\hat{\beta}$. □

Remark 2.3 (Geometric Interpretation: Orthogonal Projection). The normal equation $T^*T\hat{\beta} = T^*Y$ can be rewritten as $T^*(Y - T\hat{\beta}) = 0$. This equation holds if and only if the residual vector $r = Y - T\hat{\beta}$ is in the null space of T^*. Recall that the null space of the adjoint, $\operatorname{null}(T^*)$, is the orthogonal complement of the image (or column space) of the original operator T, i.e., $\operatorname{null}(T^*) = (\operatorname{Im} T)^{\perp}$. Thus, the normal equation is equivalent to the condition that the residual $Y - T\hat{\beta}$ must be orthogonal to the subspace $\operatorname{Im}(T)$.

Let's see how this relates to orthogonal projection explicitly. Let $\{u_1, \ldots, u_p\}$ be an orthonormal basis for the subspace $\operatorname{Im}(T) \subseteq \mathbb{R}^n$. The condition that $Y - T\hat{\beta}$ is orthogonal to $\operatorname{Im}(T)$ means it must be orthogonal to every basis vector:

$$\langle Y - T\hat{\beta}, u_i\rangle = 0 \quad \text{for all } i = 1, \ldots, p$$

This implies that

$$\langle Y, u_i\rangle = \langle T\hat{\beta}, u_i\rangle \quad \text{for all } i = 1, \ldots, p$$

Now, consider the vector $T\hat{\beta}$. Since it lies in the subspace $\operatorname{Im}(T)$, it can be expressed as a linear combination of the orthonormal basis vectors using its coordinates with respect to that basis:

$$T\hat{\beta} = \sum_{i=1}^{p} \langle T\hat{\beta}, u_i\rangle u_i$$

Substituting the result from the orthogonality condition ($\langle T\hat{\beta}, u_i\rangle = \langle Y, u_i\rangle$):

$$T\hat{\beta} = \sum_{i=1}^{p} \langle Y, u_i\rangle u_i$$

This is precisely the standard formula for the **orthogonal projection** of the vector Y onto the subspace spanned by the orthonormal basis $\{u_1, \ldots, u_p\}$, which is $\operatorname{Im}(T)$. We denote this projection as $P_{\operatorname{Im}(T)}Y$.

Therefore, the least squares solution $\hat{\beta}$ is the vector such that $T\hat{\beta}$ is the orthogonal projection of Y onto the image of T. This confirms our initial intuition: the vector in $\operatorname{Im}(T)$ closest to Y is its orthogonal projection onto that subspace.

3 Application: Simple Linear Regression

Let's apply this framework to the familiar problem of simple linear regression. We model a relationship $y_i \approx \beta_0 + \beta_1 x_i$ for $i = 1, \ldots, n$.

We can set this up in our vector/matrix framework: Let $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$. Let $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \in \mathbb{R}^2$. Let $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n\times 2}$. Our model is $Y \approx X\beta$. We assume X has full column rank, which requires $n \geq 2$ and that not all x_i values are the same.

The least squares estimate $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$ is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$. (Note that even though we use T^* abstractly, its matrix representation remains X^T).

Let's compute the components:

- $X^T X = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$
- $X^T Y = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$

The determinant of $X^T X$ is $\det(X^T X) = n \sum x_i^2 - (\sum x_i)^2$. The inverse is $(X^T X)^{-1} = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$.

Now, we find $\hat{\beta}$:

$$\begin{aligned} \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} &= (X^T X)^{-1} X^T Y \\ &= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\ &= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} (\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i) \\ n \sum x_i y_i - (\sum x_i)(\sum y_i) \end{bmatrix} \end{aligned}$$

Extracting the components gives the standard formulas:

$$\begin{aligned} \hat{\beta}_1 &= \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} \\ \hat{\beta}_0 &= \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n \sum x_i^2 - (\sum x_i)^2} \end{aligned}$$

Remark 3.1 (Connection to Statistics). These formulas become more intuitive when related to sample statistics. Let $\bar{x} = \frac{1}{n} \sum x_i$ and $\bar{y} = \frac{1}{n} \sum y_i$ be the sample means. Algebraic manipulation (or recognizing centered forms) shows:

- $n \sum x_i^2 - (\sum x_i)^2 = n \sum (x_i - \bar{x})^2$
- $n \sum x_i y_i - (\sum x_i)(\sum y_i) = n \sum (x_i - \bar{x})(y_i - \bar{y})$

Substituting these into the formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x, y)}{\text{Sample Variance}(x)}$$

For $\hat{\beta}_0$, we can simplify its formula or use the property that the regression line passes through the point of means $(\bar{x}, \bar{y})$: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. This gives:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

These are the familiar estimators for the slope and intercept in simple linear regression, derived elegantly from the general principle of minimizing squared error via orthogonal projection in vector spaces. If we view (x_i, y_i) as samples from a bivariate distribution (X, Y), these sample statistics estimate the population parameters: $\hat{\beta}_1 \to \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$ and $\hat{\beta}_0 \to \text{E}[Y] - \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \text{E}[X]$, assuming $\text{Var}(X) \neq 0$.