Linear regression through Optimization

Chris Cornwell

September 11, 2025

Outline

Outline

Extending the linear regression procedure

If you have data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_P, y_P)$ where, for some $N > 1$ the input $\mathbf{x}_i$ is a vector in $\mathbb{R}^N$, can you still do linear regression?

Extending the linear regression procedure

If you have data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_P, y_P)$ where, for some $N > 1$ the input $\mathbf{x}_i$ is a vector in $\mathbb{R}^N$, can you still do linear regression?

For each i with $1 \leq i \leq P$, say that $\mathbf{x}_i = \begin{bmatrix} x_{1,i} \\ x_{2,i} \\ \vdots \\ x_{N,i} \end{bmatrix}$. Define a $P \times (N+1)$ matrix A so that the i^{th} row is $[1, x_{1,i}, x_{2,i}, \ldots, x_{N,i}]$.

Extending the linear regression procedure

If you have data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_P, y_P)$ where, for some $N > 1$ the input $\mathbf{x}_i$ is a vector in $\mathbb{R}^N$, can you still do linear regression?

For each i with $1 \leq i \leq P$, say that $\mathbf{x}_i = \begin{bmatrix} x_{1,i} \\ x_{2,i} \\ \vdots \\ x_{N,i} \end{bmatrix}$. Define a $P \times (N+1)$ matrix A so that the i^{th} row is $[1, x_{1,i}, x_{2,i}, \ldots, x_{N,i}]$.

In other words, set $\tilde{\mathbf{x}}_i$ equal to the vector in $\mathbb{R}^{N+1}$ with 1 as its first component and $\mathbf{x}_i$ for the remaining N components. Then

$$A = \begin{bmatrix} — & \tilde{\mathbf{x}}_1^T & — \\ — & \tilde{\mathbf{x}}_2^T & — \\ & \vdots & \\ — & \tilde{\mathbf{x}}_P^T & — \end{bmatrix}.$$

Extending the linear regression procedure

A solution $\tilde{\mathbf{w}}^\star = \left[b^\star, w_1^\star, \ldots, w_N^\star\right]^T$ to the normal equations $A^T A\tilde{\mathbf{w}} = A^T\mathbf{y}$ (with A as above) gives coefficients for a linear model for the data.[1]

[1]The matrix $A^T A$ is $(N+1) \times (N+1)$ and $A^T\mathbf{y}$ is a vector in $\mathbb{R}^{N+1}$.
[2]Generalization of a plane in $\mathbb{R}^3$.

Extending the linear regression procedure

A solution $\tilde{\mathbf{w}}^\star = \left[b^\star, w_1^\star, \ldots, w_N^\star\right]^T$ to the normal equations $A^T A\tilde{\mathbf{w}} = A^T\mathbf{y}$ (with A as above) gives coefficients for a linear model for the data.[1]

Write $\mathbf{w}^\star \in \mathbb{R}^N$ for the non-constant coefficient vector $\left[w_1^\star, w_2^\star, \ldots, w_N^\star\right]^T$. The affine linear model on the data, in N variables, is

$$f_{\tilde{\mathbf{w}}^\star}(\mathbf{x}) = b^\star + w_1^\star x_1 + \ldots + w_N^\star x_N = b^\star + \mathbf{x}^T\mathbf{w}^\star.$$

[1]The matrix $A^T A$ is $(N+1) \times (N+1)$ and $A^T\mathbf{y}$ is a vector in $\mathbb{R}^{N+1}$.
[2]Generalization of a plane in $\mathbb{R}^3$.

Extending the linear regression procedure

A solution $\tilde{\mathbf{w}}^\star = \left[b^\star, w_1^\star, \ldots, w_N^\star\right]^T$ to the normal equations $A^T A\tilde{\mathbf{w}} = A^T\mathbf{y}$ (with A as above) gives coefficients for a linear model for the data.[1]

Write $\mathbf{w}^\star \in \mathbb{R}^N$ for the non-constant coefficient vector $\left[w_1^\star, w_2^\star, \ldots, w_N^\star\right]^T$. The affine linear model on the data, in N variables, is

$$f_{\tilde{\mathbf{w}}^\star}(\mathbf{x}) = b^\star + w_1^\star x_1 + \ldots + w_N^\star x_N = b^\star + \mathbf{x}^T\mathbf{w}^\star.$$

(1) $f_{\tilde{\mathbf{w}}^\star}(\mathbf{x})$ is affine linear, meaning the difference in two function values is a dot product on the difference of the input. Specifically, $f_{\tilde{\mathbf{w}}^\star}(\mathbf{x}) - f_{\tilde{\mathbf{w}}^\star}(\mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T\mathbf{w}^\star$.

[1] The matrix $A^T A$ is $(N+1) \times (N+1)$ and $A^T\mathbf{y}$ is a vector in $\mathbb{R}^{N+1}$.
[2] Generalization of a plane in $\mathbb{R}^3$.

Extending the linear regression procedure

A solution $\tilde{\mathbf{w}}^\star = \left[b^\star, w_1^\star, \ldots, w_N^\star\right]^T$ to the normal equations $A^T A\tilde{\mathbf{w}} = A^T\mathbf{y}$ (with A as above) gives coefficients for a linear model for the data.[1]

Write $\mathbf{w}^\star \in \mathbb{R}^N$ for the non-constant coefficient vector $\left[w_1^\star, w_2^\star, \ldots, w_N^\star\right]^T$. The affine linear model on the data, in N variables, is

$$f_{\tilde{\mathbf{w}}^\star}(\mathbf{x}) = b^\star + w_1^\star x_1 + \ldots + w_N^\star x_N = b^\star + \mathbf{x}^T\mathbf{w}^\star.$$

(1) $f_{\tilde{\mathbf{w}}^\star}(\mathbf{x})$ is affine linear, meaning the difference in two function values is a dot product on the difference of the input. Specifically, $f_{\tilde{\mathbf{w}}^\star}(\mathbf{x}) - f_{\tilde{\mathbf{w}}^\star}(\mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T\mathbf{w}^\star$.

(2) The graph of $f_{\tilde{\mathbf{w}}^\star}(\mathbf{x})$ is a hyperplane[2] in $\mathbb{R}^{N+1}$ with normal vector $\left[1, -w_1^\star, \ldots, -w_N^\star\right]^T$.

---

[1]The matrix $A^T A$ is $(N+1) \times (N+1)$ and $A^T\mathbf{y}$ is a vector in $\mathbb{R}^{N+1}$.
[2]Generalization of a plane in $\mathbb{R}^3$.

Example

A concrete example: using the data in `'Advertising.csv'`, found in the DataSets folder of the course site.

This contains data on amounts spent (in thousands of dollars) on TV, Radio, and Newspaper advertising in 200 different markets, as well as the amounts sold in each market (in thousands of units).

Example

A concrete example: using the data in `'Advertising.csv'`, found in the DataSets folder of the course site.
This contains data on amounts spent (in thousands of dollars) on TV, Radio, and Newspaper advertising in 200 different markets, as well as the amounts sold in each market (in thousands of units).

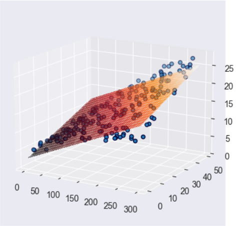
To perform linear regression, with independent variables `'TV'`, `'Radio'`, and `'Newspaper'` and setting y equal to `'Sales'`, one may follow the procedure above. The matrix A described is 200×4 ($N = 3$).

Example

A concrete example: using the data in `'Advertising.csv'`, found in the DataSets folder of the course site.

This contains data on amounts spent (in thousands of dollars) on TV, Radio, and Newspaper advertising in 200 different markets, as well as the amounts sold in each market (in thousands of units).

To perform linear regression, with independent variables `'TV'`, `'Radio'`, and `'Newspaper'` and setting y equal to `'Sales'`, one may follow the procedure above. The matrix A described is 200×4 ($N = 3$). The resulting coefficients are (approximately)



3D projection of Advertising model

$$[b^\star, w_1^\star, w_2^\star, w_3^\star] = [2.9389, 0.0458, 0.1885, -0.001].$$

Outline

Generalizing Linear regression to multiple variables

Measuring the fitness of linear models

Least Squares loss function

On given data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_P, y_P)$, the measure that is commonly used for how well a linear regression model, $f_{b,\mathbf{w}}(\mathbf{x}) = b + \mathbf{x}^T\mathbf{w}$, fits the data is the Least Squares loss (cost) function.

- *This is P times the Mean Squared Error.*

Following notation from textbook, this loss function is

$$g(b, \mathbf{w}) = \sum_{p=1}^{P} \left(f_{b,\mathbf{w}}(\mathbf{x}_p) - y_p\right)^2 .$$

Meaning of Least Squares loss

For $1 \leq p \leq P$, since $f_{b,\mathbf{w}}(\mathbf{x}_p) = \hat{y}_p$, the quantity $\left|f_{b,\mathbf{w}}(\mathbf{x}_p) - y_p\right|$ is the vertical distance from (x_p, y_p) to the point predicted by the linear model, $(x_p, \hat{y}_p)$.

Additionally, the length of the vector $\mathbf{y} - \hat{\mathbf{y}}$ (which is the distance from $\mathbf{y}$ to the point determined by $\tilde{\mathbf{w}}$, in the column space of our feature matrix) is equal to

$$\sqrt{(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \ldots + (\hat{y}_P - y_P)^2} = \sqrt{g(b, \mathbf{w})}.$$

We see that minimizing $g(b, \mathbf{w})$ is the same as minimizing that distance, which will give us the $\hat{\mathbf{y}}$ in the column space that makes $\mathbf{y} - \hat{\mathbf{y}}$ be orthogonal to the column space.

Minimizing the Least Squares loss

The data $\left\{\left(\mathbf{x}_p, y_p\right)\right\}_{p=1}^{P}$ is fixed. How do we solve the problem

$$\underset{b,\mathbf{w}}{\text{minimize}} \quad g\left(b, \mathbf{w}\right)?$$

We can use methods from calculus, specifically the first order condition – that we want all partial derivatives equal to zero.

- Note, what are the variables of the function g? They are the parameters $b, w_1, w_2, \ldots, w_N$.

Minimizing the Least Squares loss

The data $\left\{\left(\mathbf{x}_p, y_p\right)\right\}_{p=1}^{P}$ is fixed. How do we solve the problem

$$\underset{b,\mathbf{w}}{\text{minimize}} \; g\left(b, \mathbf{w}\right)?$$

We can use methods from calculus, specifically the first order condition – that we want all partial derivatives equal to zero.

- Note, what are the variables of the function g? They are the parameters $b, w_1, w_2, \ldots, w_N$.

Next: a review of calculus minimization techniques.