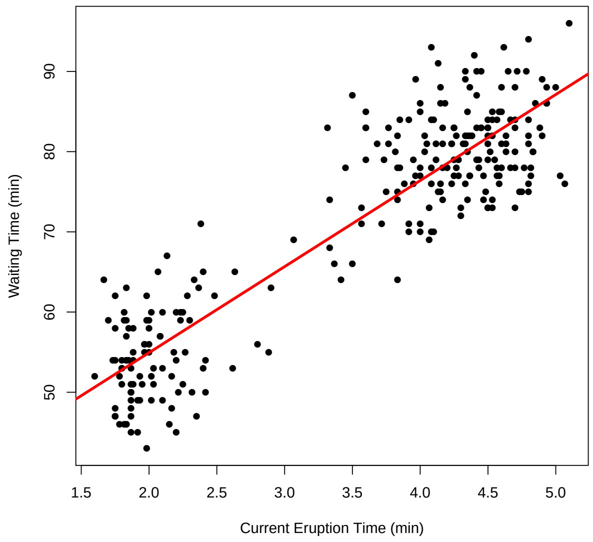
Ordinary least squares

Linear regression

Example: Old Faithful Geyser (Yellowstone)

Time between eruptions seems to be related to duration of previous eruption.

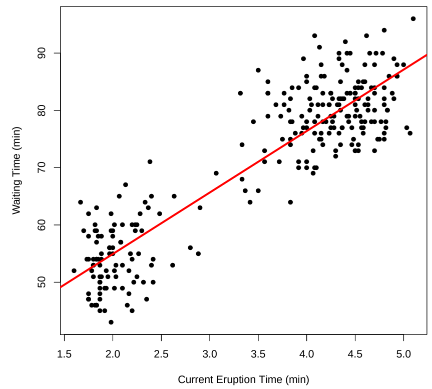
Example: Old Faithful Geyser (Yellowstone)



Example: Old Faithful

Linear regression

$$\text{(wait time)} \;=\; w_0 + \text{(last duration)} \times w_1 + \text{(error)}$$



Multivariate linear regression

Linear regression in $\mathbb{R}^d$

- Input variables $(x_1, x_2, \ldots, x_d)$ (i.e., "covariates", "features").
- Output variable y (i.e., "response", "label").
- Regression coefficients $(w_1, w_2, \ldots, w_d)$, intercept term w_0.

Modeling equation:

$$y \;=\; w_0 + \sum_{j=1}^{d} w_j x_j + \varepsilon$$

where ε is a "noise" or "error" term.

(Statisticians use p instead of d, and β instead of w.)

Ordinary least squares via calculus

Least squares criterion

Data

n pairs of input/output values $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where

$$\boldsymbol{x}_i \;=\; (x_{i,1}, x_{i,2}, \ldots, x_{i,d}), \quad i = 1, 2, \ldots, n.$$

Least squares criterion

Find $(w_0, w_1, \ldots, w_d)$ to minimize sum of squared residuals:

$$\sum_{i=1}^{n} r_i^2$$

where

$$r_i \;:=\; y_i - \left(w_0 + \sum_{j=1}^{d} w_j x_{i,j} \right)$$

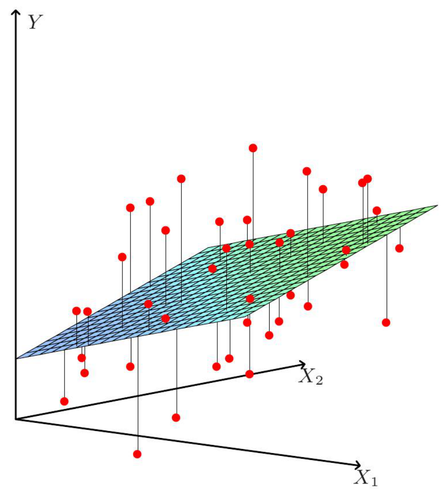
is the i-th residual.

Least squares in pictures

Red dots: data points.

$(w_0, w_1, w_2) \to$ affine hyperplane.

Vertical length is error.



Least squares in matrix/vector form

Data in matrix/vector form

$$A := \underbrace{\begin{bmatrix} \longleftarrow & x_1^\top & \longrightarrow \\ \longleftarrow & x_2^\top & \longrightarrow \\ & \vdots & \\ \longleftarrow & x_n^\top & \longrightarrow \end{bmatrix}}_{n \times d \text{ matrix}}, \qquad b := \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1 \text{ vector}},$$

Least squares criterion in matrix/vector form

Find $w_0 \in \mathbb{R}$ and $w := (w_1, w_2, \ldots, w_d) \in \mathbb{R}^d$ to minimize

$$\|r\|_2^2$$

where

$$r := b - (w_0 \mathbf{1} + Aw) = b - \begin{bmatrix} \mathbf{1} & A \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix}.$$

Simplification

Standard form of least squares objective function:

$$(w_0, w) \mapsto \left\| b - \begin{bmatrix} \mathbf{1} & A \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} \right\|_2^2.$$

Simplification: Assume A already has all-ones vector $\mathbf{1}$ as its first column.

So replace $\begin{bmatrix} \mathbf{1} & A \end{bmatrix}$ with A, and replace (w_0, w) with w.
Simplified least squares objective:

$$w \mapsto \|b - Aw\|_2^2.$$

Least squares via calculus

Least squares objective is convex function of w.
Suffices to find w where gradient is zero.

$$\nabla_w \left\{ \|b - Aw\|_2^2 \right\} = 2A^\top (Aw - b).$$

This is zero when

$$(A^\top A) w = A^\top b,$$

a system of linear equations in w (called the "normal equations").

If $A^\top A$ is invertible, the *unique* solution is

$$\hat{w}_{\text{ols}} := (A^\top A)^{-1} A^\top b$$

which we call the **ordinary least squares** solution.

Orthogonal projection a subspace

Closest vector in a subspace

Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the j-th column of matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, so

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_d \end{bmatrix}.$$

Task: Find closest vector to $\boldsymbol{b} \in \mathbb{R}^n$ in $\operatorname{span}\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_d\}$.

Solution: orthogonal projection of $\boldsymbol{b}$ onto $\operatorname{span}\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_d\}$

$$\hat{\boldsymbol{b}} := \operatorname*{arg\,min}_{\boldsymbol{v} \in \operatorname{span}\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_d\}} \|\boldsymbol{b} - \boldsymbol{v}\|_2^2.$$

Every vector in $\operatorname{span}\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_d\}$ can be written as $\boldsymbol{A}\boldsymbol{w}$ for some $\boldsymbol{w} \in \mathbb{R}^d$.

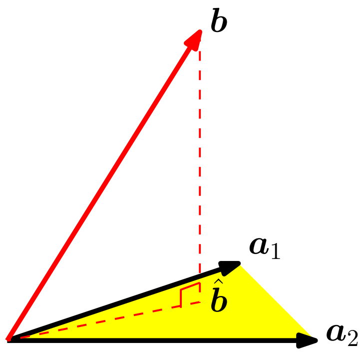
Therefore, suffices to find minimizer of

$$\boldsymbol{w} \mapsto \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{w}\|_2^2 = \left\| \boldsymbol{b} - \sum_{j=1}^d w_j \boldsymbol{a}_j \right\|_2^2.$$

Orthogonal projection to a subspace

Equivalent characterization (when $\boldsymbol{A}^\top \boldsymbol{A}$ is invertible):

$$\hat{\boldsymbol{b}} = \boldsymbol{A}\hat{\boldsymbol{w}}_{\text{ols}} = \underbrace{\boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top}_{\boldsymbol{\Pi}} \boldsymbol{b}.$$



$\boldsymbol{\Pi} \in \mathbb{R}^{n \times n}$ is the orthogonal projection operator for $\operatorname{span}\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_d\}$.

Residual vector $\boldsymbol{r} = \boldsymbol{b} - \hat{\boldsymbol{b}} = \boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{w}}_{\text{ols}}$ is **orthogonal** to all $\boldsymbol{a}_j$.

Statistical modeling

Linear regression model

(Below, $\boldsymbol{X}$ is a random vector in $\mathbb{R}^d$, and Y is a real-valued random variable.)

Classical statistical model for linear regression:

$$\mathcal{P} := \left\{ P_{\boldsymbol{w},\sigma^2} : \boldsymbol{w} \in \mathbb{R}^d,\ \sigma^2 > 0 \right\}$$

where each $P_{\boldsymbol{w},\sigma^2}$ specifies distribution of $Y \mid \boldsymbol{X} = \boldsymbol{x}$ for each $\boldsymbol{x} \in \mathbb{R}^d$:

$$(\boldsymbol{X}, Y) \sim P_{\boldsymbol{w},\sigma^2} \iff Y \mid \boldsymbol{X} = \boldsymbol{x} \ \sim\ \mathrm{N}\left(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, \sigma^2\right).$$

$P_{\boldsymbol{w},\sigma^2}$ usually described by

$$Y = \langle \boldsymbol{w}, \boldsymbol{X} \rangle + \sigma Z,$$

where $\boldsymbol{X}$ and Z are independent, and $Z \sim \mathrm{N}(0,1)$.

(Can also incorporate "intercept" term; we omit it here for simplicity.)

Maximum likelihood estimator

Question: What is MLE for $\boldsymbol{w}$ given $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ (regarded as iid sample)?

Answer: The ordinary least squares *estimator* (when it exists)!

Log-likelihood of $\boldsymbol{w}$ given $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$:

$$\sum_{i=1}^n \ln \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle\right)^2}{2\sigma^2} \right) \right\}$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^n \left(y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right)^2 + \text{terms that don't depend on } \boldsymbol{w}.$$

So $\hat{\boldsymbol{w}}_{\text{ols}}$ (when it exists) is a maximizer of the log-likelihood.

Taking the model seriously

Some people like to interpret the estimated regression coefficients

$$\hat{\boldsymbol{w}}_{\text{ols}} = (\hat{w}_{\text{ols},1}, \hat{w}_{\text{ols},2}, \dots, \hat{w}_{\text{ols},d}).$$

Example:

$\hat{w}_{\text{ols},j} = 0 \longrightarrow$ variable X_j has negligible effect on Y.

$|\hat{w}_{\text{ols},j}| \gg 0 \longrightarrow$ variable X_j has significant effect on Y.

Hypothesis tests typically consider $P_{\boldsymbol{0},\sigma^2} \in \mathcal{P}$ as the null distribution.

Aside: bias/variance

Best representative in terms of expected square loss?

Given a collection of numbers $z_1, z_2, \dots, z_n \in \mathbb{R}$, what number θ minimizes the average squared-differences:

$$\frac{1}{n}\sum_{i=1}^n (z_i - \theta)^2\ ?$$

Bias/variance decomposition

For any random variable Z and any number $\theta \in \mathbb{R}$,

$$\mathbb{E}\left[(Z-\theta)^2\right] = \underbrace{(\theta-\mu)^2}_{\text{squared bias}} + \underbrace{\mathbb{E}\left[(Z-\mu)^2\right]}_{\text{variance}}$$

where $\mu := \mathbb{E}(Z)$.

Aside: bias/variance, functional version

Consider an arbitrary random pair (X, Y) with values in $\mathcal{X} \times \mathbb{R}$.

Question: What function $f \colon \mathcal{X} \to \mathbb{R}$ has the smallest expected squared loss

$$\mathbb{E}\left[\left(Y - f(X)\right)^2\right] \;=\; \mathbb{E}\left[\mathbb{E}\left[\left(Y - f(X)\right)^2 \mid X\right]\right]\,?$$

Answer:

$$x \;\mapsto\; \mathbb{E}\left[Y \mid X = x\right].$$

Assuming the model is well-specified

Definition

The linear regression model $\mathcal{P}$ is **well-specified** if distribution of $(\boldsymbol{X}, Y)$ is given by $P_{\boldsymbol{w}_\star, \sigma^2}$ for some $P_{\boldsymbol{w}_\star, \sigma^2} \in \mathcal{P}$.

Consequences when $\mathcal{P}$ is well-specified

- Best predictor of Y from $\boldsymbol{X}$ (under square loss) is a linear function. This is because
$$Y \mid \boldsymbol{X} = \boldsymbol{x} \;\sim\; \mathrm{N}(\langle \boldsymbol{w}_\star, \boldsymbol{x} \rangle, \sigma^2)\,,$$
so
$$\mathbb{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}\right] \;=\; \langle \boldsymbol{w}_\star, \boldsymbol{x} \rangle\,.$$
- MLE ($\hat{\boldsymbol{w}}_{\mathrm{ols}}$) is unbiased (again, assuming it exists):
$$\mathbb{E}[\hat{\boldsymbol{w}}_{\mathrm{ols}}] \;=\; \boldsymbol{w}_\star\,.$$

Statistical learning

Statistical learning for regression

- Probability distribution P over $\mathcal{X} \times \mathbb{R}$; let $(X, Y) \sim P$.
- Think of P as being comprised of two parts.
  1. Marginal distribution of X (a distribution over $\mathcal{X}$).
  2. Conditional distribution of Y given $X = x$, for each $x \in \mathcal{X}$.
- The predictor with smallest expected square loss is given by
$$f^*(x) \;=\; \mathbb{E}\left[Y \mid X = x\right].$$

If $\mathcal{X} = \mathbb{R}^d$ ($\boldsymbol{X}$ is random vector in $\mathbb{R}^d$): best *linear* predictor is $\boldsymbol{x} \mapsto \langle \boldsymbol{w}_\star, \boldsymbol{x} \rangle$, where

$$\boldsymbol{w}_\star \;:=\; \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^d} \mathbb{E}\left[\left(Y - \langle \boldsymbol{w}, \boldsymbol{X} \rangle\right)^2\right].$$

(This is uniquely determined if $\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right]$ is invertible!)

Competing with the best linear predictor

Goal: given iid sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from P, find $\boldsymbol{w} \in \mathbb{R}^d$ so that excess expected square loss

$$\mathbb{E}\left[\left(Y - \langle \boldsymbol{w}, \boldsymbol{X} \rangle\right)^2\right] - \mathbb{E}\left[\left(Y - \langle \boldsymbol{w}_\star, \boldsymbol{X} \rangle\right)^2\right]$$

approaches 0 as $n \to \infty$.

Note: no assumption like $Y = \langle \boldsymbol{w}_\star, \boldsymbol{X} \rangle + \sigma Z$ for $Z \sim \mathrm{N}(0,1)$.
Conditional expectation function $\boldsymbol{x} \mapsto \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ could be non-linear!

Ordinary least squares

Empirical Risk Minimization (for square loss): pick $\boldsymbol{w}$ to minimize average square loss on data, i.e.,

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{i=1}^n \left(y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right)^2 .$$

If the minimizer is unique, it is $\hat{\boldsymbol{w}}_{\mathrm{ols}}$.

- Convex optimization problem that happens to have "closed form" solution.
- ERM solution is not unique unless $\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$ is invertible.
- **Predictive performance**:
  - $n < d$: Could be rubbish.
  - $n \geq d$: Excess expected square loss decreases at a rate of $O\left(\frac{d}{n}\right)$ (under some general conditions).

Computation

Computation of ordinary least squares

Naïve computation (based on solving normal equations) takes $O(nd^2)$ time.

Hopes for speeding things up:

- Exploit sparsity or other structure in data matrix.
- Use iterative methods (e.g., *conjugate gradient* method).

Key takeaways

1. Four different ways to arrive at OLS
   1.1 Satisfies least squares criterion.
   1.2 Gives orthoprojection of $\boldsymbol{y}$ onto column space of $\boldsymbol{X}$.
   1.3 MLE for a particular statistical model, an unbiased estimator.
   1.4 ERM for square loss.
2. How to compute MLE via solving normal equations.
3. When does OLS exist?