Ordinary least squares

Example: Old Faithful Geyser (Yellowstone)

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

Linear regression

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 \(\varepsilon\) is a "noise" or "error" term.

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

Ordinary least squares via calculus

Data

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

\[x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d}), \quad i = 1, 2, \ldots, n.\]

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.

\((w_0, w_1, w_2) \rightarrow \text{affine hyperplane.} \)

Vertical length is error.

Least squares in matrix/vector form

Data in matrix/vector form

\[ A := \begin{bmatrix} \leftarrow x_1^\top \rightarrow \hfill \cr \leftarrow x_2^\top \rightarrow \hfill \cr \vdots \hfill \cr \leftarrow x_n^\top \rightarrow \end{bmatrix}, \quad b := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \]

\( n \times d \) matrix, \( n \times 1 \) 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

\[ r := b - (w_0 1 + A w) = b - \begin{bmatrix} 1 & \text{A} \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix}. \]

Simplification

Standard form of least squares objective function:

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

Simplification: Assume \( A \) already has all-ones vector \( 1 \) as its first column.

So replace \( \begin{bmatrix} 1 & A \end{bmatrix} \) with \( A \), and replace \( (w_0, w) \) with \( w \).

Simplified least squares objective:

\( w \mapsto \| b - A w \|_2^2 \).

Least squares via calculus

Least squares objective is convex function of \( w \).

Suffices to find \( w \) where gradient is zero.

\[ \nabla_w \left\{ \| b - A w \|_2^2 \right\} = 2A^\top (A w - b). \]

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 to a subspace**

Equivalent characterization (when $A^T A$ is invertible):

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

$\Pi \in \mathbb{R}^{n \times n}$ is the orthogonal projection operator for $\text{span}\{a_1, a_2, \ldots, a_d\}$.

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

---

**Closest vector in a subspace**

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

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_d \end{bmatrix}.$$  

**Task:** Find closest vector to $b \in \mathbb{R}^n$ in $\text{span}\{a_1, a_2, \ldots, a_d\}$.

**Solution:** orthogonal projection of $b$ onto $\text{span}\{a_1, a_2, \ldots, a_d\}$

$$\hat{b} := \arg \min_{v \in \text{span}\{a_1, a_2, \ldots, a_d\}} \| b - v \|^2_2.$$  

Every vector in $\text{span}\{a_1, a_2, \ldots, a_d\}$ can be written as $Aw$ for some $w \in \mathbb{R}^d$. Therefore, suffices to find minimizer of

$$w \mapsto \| b - Aw \|^2_2 = \left\| b - \sum_{j=1}^d w_j a_j \right\|^2_2.$$  

---

**Statistical modeling**
Linear regression model

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

Classical statistical model for linear regression:

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

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

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

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

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

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

(Can incorporate “intercept” in the usual way. We’ll omit it here for simplicity.)

Aside: bias/variance

Best representative in terms of expected squared error?

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

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

Bias/variance decomposition

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

$$\mathbb{E}[(Z - \hat{\theta})^2] = (\hat{\theta} - \mu)^2 + \mathbb{E}[(Z - \mu)^2]$$

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

Taking the model seriously

Some people like to interpret the estimated regression coefficients

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

Example:

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

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

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

Maximum likelihood estimator

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

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

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

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

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle w, x_i \rangle)^2 + \text{terms that don’t depend on } w.$$

So $\hat{w}_{\text{ols}}$ (when it exists) is a maximizer of the log-likelihood.
Aside: bias/variance, functional version

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

**Question:** What function \(f: \mathcal{X} \rightarrow \mathbb{R}\) has the smallest expected squared-loss

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

**Answer:**

\[x \mapsto \mathbb{E}[Y \mid X = x].\]

Assuming the model is well-specified

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

**Consequences when \(P\) is well-specified**

- Best predictor of \(Y\) from \(X\) (under squared-loss) is a linear function. This is because

\[
Y \mid X = x \sim N(\langle w_*, x \rangle, \sigma^2),
\]

so

\[
\mathbb{E}[Y \mid X = x] = \langle w_*, x \rangle.
\]

- MLE \((\hat{w}_{\text{ols}})\) is unbiased:

\[
\mathbb{E}[\hat{w}_{\text{ols}}] = w_*.
\]

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 squared-loss is given by

\[
f^*(x) = \mathbb{E}[Y \mid X = x].
\]

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

\[
w_* := \arg \min_{w \in \mathbb{R}^d} \mathbb{E}\left[\left(Y - \langle w, X \rangle\right)^2\right].
\]

(This is uniquely determined if \(\mathbb{E}[XX^\top]\) is invertible!)

Statistical learning
Competing with the best linear predictor

**Goal:** given iid sample \(\{(x_i, y_i)\}_{i=1}^n\) from \(P\), find \(w \in \mathbb{R}^d\) so that excess expected squared-loss

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

approaches 0 as \(n \to \infty\).

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

Ordinary least squares

**Empirical Risk Minimization** (for squared-loss): pick \(w\) to minimize average squared-loss on data, i.e.,

\[
\arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^n \left( y_i - \langle w, x_i \rangle \right)^2.
\]

If the minimizer is unique, it is \(\hat{w}_{\text{ols}}\).

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

Computation

**Naïve computation** 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).
Recap

- Ordinary least squares
  1. Satisfies least squares criterion.
  2. Gives orthoprojection of \( y \) onto column space of \( X \).
  3. MLE for a particular statistical model.
  4. ERM for squared-loss.

- What if \( d > n \)?
  Minimizer of least squares objective is no longer uniquely defined and is often ill-behaved.

  **Solution:** use regularization
  - encourage \( \|w\|^2 \) to be small ("ridge regression") → can kernelize this
  - encourage \( \|w\|_1 \) to be small ("Lasso")
  - encourage \( w \) to be sparse ("sparse regression")