COMS 4771 Lecture 15

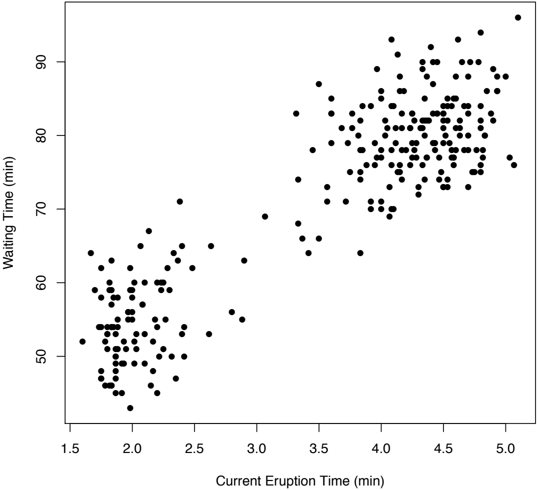
1. Linear regression

Linear regression (introduction)

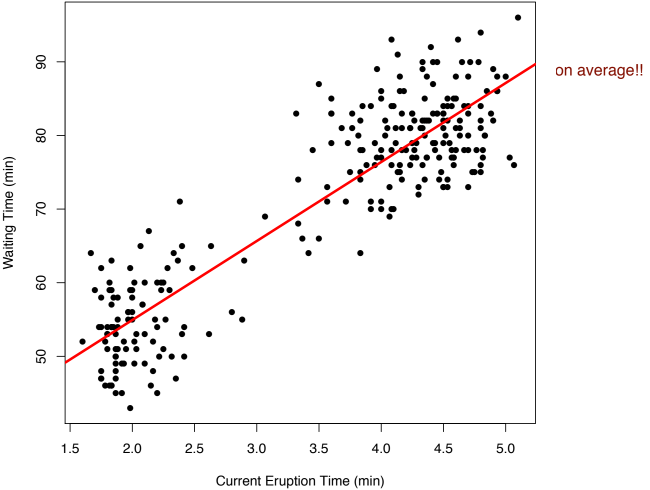
Example: Old Faithful Geyser (Yellowstone)

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

Example: Old Faithful Geyser (Yellowstone)



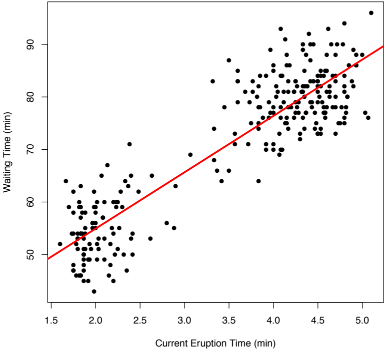
Example: Old Faithful Geyser (Yellowstone)



Example: Old Faithful

Linear regression model

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



Linear regression model in $\mathbb{R}^p$

- Input variables $\boldsymbol{x} := (x_1, x_2, \ldots, x_p)$ ("covariates").
- Output variable y ("response").
- Regression coefficients $\boldsymbol{w} := (w_1, w_2, \ldots, w_p)$, intercept term w_0.

Modeling equation:

$$y = w_0 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \varepsilon$$

where $\varepsilon := y - (w_0 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle)$ is the error term.

Linear regression model in $\mathbb{R}^p$

- Input variables $\boldsymbol{x} := (x_1, x_2, \ldots, x_p)$ ("covariates").
- Output variable y ("response").
- Regression coefficients $\boldsymbol{w} := (w_1, w_2, \ldots, w_p)$, intercept term w_0.

Modeling equation:

$$y = w_0 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \varepsilon$$

where $\varepsilon := y - (w_0 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle)$ is the error term.

Least squares criterion

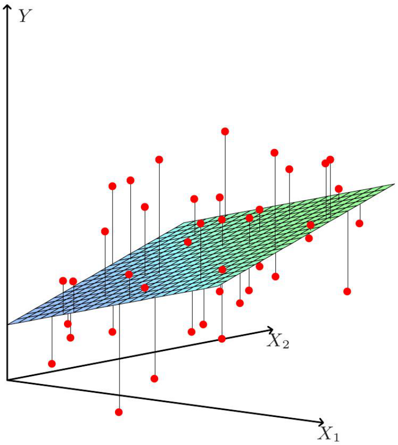
Given pairs of input/output values, find $(w_0, \boldsymbol{w})$ to minimize ε^2 (on average)

Least squares in pictures

Red dots: data points.

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

Vertical length is error.



Least squares criterion

Given training data

$$\boldsymbol{X} = \begin{bmatrix} — & \boldsymbol{x}^{(1)\top} & — \\ — & \boldsymbol{x}^{(2)\top} & — \\ & \vdots & \\ — & \boldsymbol{x}^{(n)\top} & — \end{bmatrix} \in \mathbb{R}^{n \times p}, \qquad \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix},$$

find $w_0 \in \mathbb{R}$ and $\boldsymbol{w} \in \mathbb{R}^p$ to minimize

$$f_{\text{ls}}(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^{n} \Big(y^{(i)} - \big(w_0 + \langle \boldsymbol{x}^{(i)}, \boldsymbol{w} \rangle \big) \Big)^2 = \frac{1}{n} \left\| \boldsymbol{y} - \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix} \right\|_2^2 .$$

Text

add extra 1 column in the matrix X

Simplification

Replace $\boldsymbol{X}$ with $\begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix}$ and $\boldsymbol{w}$ with $(w_0, \boldsymbol{w})$, so least squares criterion is more simply written as

$$f_{\text{ls}}(\boldsymbol{w}) = \frac{1}{n} \| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} \|_2^2 .$$

Least squares via calculus

Least squares criterion is convex function of $\boldsymbol{w}$;
so suffices to find $\boldsymbol{w}$ where gradient is zero.

Least squares via calculus

Least squares criterion is convex function of $\boldsymbol{w}$;
so suffices to find $\boldsymbol{w}$ where gradient is zero.

Take gradient with respect to $\boldsymbol{w}$:

$$\nabla_{\boldsymbol{w}} f_{\mathrm{ls}}(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \right\} = \frac{2}{n} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}).$$

Least squares via calculus

Least squares criterion is convex function of $\boldsymbol{w}$;
so suffices to find $\boldsymbol{w}$ where gradient is zero.

Take gradient with respect to $\boldsymbol{w}$:

$$\nabla_{\boldsymbol{w}} f_{\mathrm{ls}}(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \right\} = \frac{2}{n} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}).$$

This is zero when

$$(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y},$$

a linear system of equations in $\boldsymbol{w}$ ("normal equations").

Least squares criterion is convex function of $\boldsymbol{w}$;
so suffices to find $\boldsymbol{w}$ where gradient is zero.

Take gradient with respect to $\boldsymbol{w}$:

$$\nabla_{\boldsymbol{w}} f_{\mathrm{ls}}(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \right\} = \boxed{\frac{2}{n} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}).}$$

This is zero when

$$(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y},$$

a linear system of equations in $\boldsymbol{w}$ ("normal equations").

If $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, solution is

$$\hat{\boldsymbol{w}}_{\mathrm{ols}} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

("ordinary least squares").

Column view of least squares

Least squares criterion

Let $\boldsymbol{x}_j \in \mathbb{R}^n$ be the j-th column of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, so

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_p \end{bmatrix}.$$

Least squares criterion

Let $\boldsymbol{x}_j \in \mathbb{R}^n$ be the j-th column of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, so

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_p \end{bmatrix}.$$

Find linear combination $\sum_{j=1}^{p} w_j \boldsymbol{x}_j$ of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ so as to minimize

$$f_{\mathrm{ls}}(\boldsymbol{w}) \;=\; \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \;=\; \frac{1}{n}\left\|\boldsymbol{y} - \sum_{j=1}^{p} w_j \boldsymbol{x}_j\right\|_2^2.$$

Least squares criterion

Let $\boldsymbol{x}_j \in \mathbb{R}^n$ be the j-th column of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, so

consist some sub-space

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boxed{\boldsymbol{x}_p} \end{bmatrix}$$

Find linear combination $\sum_{j=1}^{p} w_j \boldsymbol{x}_j$ of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ so as to minimize

column view

$$f_{\text{ls}}(\boldsymbol{w}) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boxed{\boldsymbol{w}\|_2^2} = \frac{1}{n}\left\|\boldsymbol{y} - \sum_{j=1}^{p} w_j \boldsymbol{x}_j\right\|_2^2.$$

linear combination!

Eculian projection

Approximation of $\boldsymbol{y}$ via ordinary least squares (assuming $\boldsymbol{X}^\top \boldsymbol{X}$ invertible):

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}_{\text{ols}} = \sum_{j=1}^{p} \hat{w}_{\text{ols},j} \boldsymbol{x}_j.$$

dissect by row!

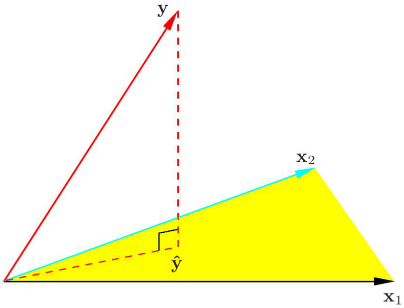
suppose W(ols) has already been got!

Column view of least squares

$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}_{\text{ols}}$ is Text the orthogonal projection of $\boldsymbol{y}$ onto $\operatorname{span}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$:

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

equal to W(ols) to meet this condition.



when is the smallest???
The distance is the smallest!!!

Projection is the shortest distance~~
Thus we should adjust W(ols) to achieve the projection

project all y on the span!

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

Statistical learning perspective

Linear regression model

- Let P be a distribution over $\mathbb{R}^p \times \mathbb{R}$, and $(\boldsymbol{x}, y) \sim P$.

  Define

  $$\boldsymbol{w}_\star := \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \mathbb{E}\left[\Big(y - \langle \boldsymbol{x}, \boldsymbol{w} \rangle\Big)^2\right] = \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \mathbb{E}\left[\Big(\mathbb{E}(y|\boldsymbol{x}) - \langle \boldsymbol{x}, \boldsymbol{w} \rangle\Big)^2\right].$$

  (Best linear approximation of *conditional expectation function* $\mathbb{E}(y|\boldsymbol{x})$.)

Linear regression model

- Let P be a distribution over $\mathbb{R}^p \times \mathbb{R}$, and $(\boldsymbol{x}, y) \sim P$.

  y's value is also depneds on vector x's value

  Define

  the optimal one,
  not perfect is accepte.

$$\boldsymbol{w}_\star := \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^p} \mathbb{E}\left[\Big(y - \langle \boldsymbol{x}, \boldsymbol{w}\rangle\Big)^2\right] = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^p} \mathbb{E}\left[\Big(\mathbb{E}(y|\boldsymbol{x}) - \langle \boldsymbol{x}, \boldsymbol{w}\rangle\Big)^2\right].$$

  (Best linear approximation of *conditional expectation function* $\mathbb{E}(y|\boldsymbol{x})$.)

- **Goal**: given i.i.d. sample S from P, find $\boldsymbol{w} \in \mathbb{R}^p$ so that execess mean squared error

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

  is small (and $\to 0$ as sample size $\to \infty$).

  try to get close to the most optimal one!

Ordinary least squares

Ordinary least squares picks $\boldsymbol{w}$ to minimize empirical mean squared error based on i.i.d. sample S:

$$\hat{\boldsymbol{w}}_{\mathrm{ols}} := \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \sum_{(\boldsymbol{x}, y) \in S} \Big(y - \langle \boldsymbol{x}, \boldsymbol{w} \rangle \Big)^2 .$$

Ordinary least squares

Ordinary least squares picks $\boldsymbol{w}$ to minimize empirical mean squared error based on i.i.d. sample S:

$$\hat{\boldsymbol{w}}_{\text{ols}} := \underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg\min} \sum_{(\boldsymbol{x},y) \in S} \Big(y - \langle \boldsymbol{x}, \boldsymbol{w} \rangle \Big)^2 .$$

- Not uniquely defined unless $\sum_{(\boldsymbol{x},y) \in S} \boldsymbol{x}\boldsymbol{x}^\top$ is invertible.

Ordinary least squares

Ordinary least squares picks $\boldsymbol{w}$ to minimize empirical mean squared error based on i.i.d. sample S:

$$\hat{\boldsymbol{w}}_{\text{ols}} := \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \sum_{(\boldsymbol{x}, y) \in S} \Big(y - \langle \boldsymbol{x}, \boldsymbol{w} \rangle \Big)^2.$$

- Not uniquely defined unless $\boxed{\sum_{(\boldsymbol{x},y)\in S} \boldsymbol{x}\boldsymbol{x}^\top \text{ is invertible.}}$

there is a requirement over the size of sample n

- **Predictive performance**: (with $n = |S|$)
  - $n < p$: Could be rubbish.
  - $n \geq p$: Excess mean squared error decreases at a rate of $O\left(\frac{p}{n}\right)$ (under some general conditions).

Statistical estimation perspective

Maximum likelihood interpretation

Suppose the distribution P of $(\boldsymbol{x}, y)$ is such that, conditioned on $\boldsymbol{x}$,

$$y|\boldsymbol{x} \sim \mathcal{N}\big(\langle \boldsymbol{x}, \boldsymbol{w}_\star \rangle, \sigma^2\big).$$

Maximum likelihood interpretation

Suppose the distribution P of $(\boldsymbol{x}, y)$ is such that, conditioned on $\boldsymbol{x}$,

$$y|\boldsymbol{x} \sim \mathcal{N}\big(\langle \boldsymbol{x}, \boldsymbol{w}_\star \rangle, \sigma^2\big).$$

- **Question**: Given i.i.d. sample S from P, what is the MLE for $\boldsymbol{w}_\star$?

Maximum likelihood interpretation

w* is assumed to
associate with u of the
distributation

Suppose the distribution P of $(\boldsymbol{x}, y)$ is such that, conditioned on $\boldsymbol{x}$,

$$y|\boldsymbol{x} \sim \mathcal{N}\big(\langle \boldsymbol{x}, \boldsymbol{w}_\star \rangle, \sigma^2\big).$$

note the the assumption of linear
regression, fix x
y condition on x is a Gussian Distribution

- **Question**: Given i.i.d. sample S from P, what is the MLE for $\boldsymbol{w}_\star$?
- **Answer**: The ordinary least squares estimator.

  Log-likelihood of $\boldsymbol{w}$ given S:

$$\sum_{(\boldsymbol{x},y)\in S} \ln\left\{\exp\left(-\frac{1}{2}\Big(y - \langle \boldsymbol{x}, \boldsymbol{w} \rangle\Big)^2\right)\right\}$$

  (plus terms that don't depend on $\boldsymbol{w}$).

  $\longrightarrow$ maximizing likelihood $\equiv$ minimizing empirical mean squared error.

Aside: logistic regression

Suppose P is a distribution over $\mathbb{R}^p \times \{0,1\}$, and $(\boldsymbol{x}, y) \sim P$ satisfies

$$\Pr(y = 1 | \boldsymbol{x}) = \frac{1}{1 + \exp(-\langle \boldsymbol{x}, \boldsymbol{w}_\star \rangle)}.$$

Aside: logistic regression

Suppose P is a distribution over $\mathbb{R}^p \times \{0,1\}$, and $(\boldsymbol{x}, y) \sim P$ satisfies

$$\Pr(y = 1 | \boldsymbol{x}) = \frac{1}{1 + \exp(-\langle \boldsymbol{x}, \boldsymbol{w}_\star \rangle)}.$$

- **Question**: Given i.i.d. sample S from P, what is the MLE for $\boldsymbol{w}_\star$?

ASIDE: LOGISTIC REGRESSION

Suppose P is a distribution over $\mathbb{R}^p \times \{0, 1\}$, and $(\boldsymbol{x}, y) \sim P$ satisfies

$$\Pr(y = 1|\boldsymbol{x}) = \frac{1}{1 + \exp(-\langle \boldsymbol{x}, \boldsymbol{w}_\star \rangle)}.$$

- **Question**: Given i.i.d. sample S from P, what is the MLE for $\boldsymbol{w}_\star$?
- **Answer**: The empirical minimizer of logistic loss

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^p} \boxed{\frac{1}{n} \sum_{(\boldsymbol{x}, y) \in S} \ell_{\log}(y \langle \boldsymbol{x}, \boldsymbol{w} \rangle).}$$

note the difference of minimizer, usually we calculate the classifier in the form of pr(1|w), but at here we should compute w*.

TAKING $\boldsymbol{w}_\star$ SERIOUSLY

Sometimes people seriously interpret the estimated regression coefficients.

Taking $\boldsymbol{w}_\star$ seriously

Sometimes people seriously interpret the estimated regression coefficients.

Example:

$$\hat{w}_i = 0 \quad \longrightarrow \quad \text{variable } x_i \text{ has negligible effect on } y \text{ in } \boldsymbol{w}_\star.$$
$$|\hat{w}_i| \gg 0 \quad \longrightarrow \quad \text{variable } x_i \text{ has significant effect on } y \text{ in } \boldsymbol{w}_\star.$$

TAKING $\boldsymbol{w}_\star$ SERIOUSLY

Sometimes people seriously interpret the estimated regression coefficients.

Example:

$$\begin{aligned} \hat{w}_i = 0 \quad &\longrightarrow \quad \text{variable } x_i \text{ has negligible effect on } y \text{ in } \boldsymbol{w}_\star. \\ |\hat{w}_i| \gg 0 \quad &\longrightarrow \quad \text{variable } x_i \text{ has significant effect on } y \text{ in } \boldsymbol{w}_\star. \end{aligned}$$

Hypothesis tests for this usually assume $y|\boldsymbol{x} \sim \mathcal{N}(\langle \boldsymbol{x}, \boldsymbol{w}_\star \rangle, \sigma^2)$.

Recap

- **Ordinary least squares**:
    1. Affine hyperplane that minimizes least squares criterion.
    2. Approximates $\boldsymbol{y}$ as linear combination of columns of $\boldsymbol{X}$.
    3. In statistical learning, excess mean squared error is $O(p/n)$.
    4. Maximum likelihood estimator under a Gaussian assumption.

RECAP

- **Ordinary least squares**:
  1. Affine hyperplane that minimizes least squares criterion.
  2. Approximates $\boldsymbol{y}$ as linear combination of columns of $\boldsymbol{X}$.
  3. In statistical learning, excess mean squared error is $O(p/n)$.
  4. Maximum likelihood estimator under a Gaussian assumption.
- What if $p > n$?

Recap

- **Ordinary least squares**:
  1. Affine hyperplane that minimizes least squares criterion.
  2. Approximates $\boldsymbol{y}$ as linear combination of columns of $\boldsymbol{X}$.
  3. In statistical learning, excess mean squared error is $O(p/n)$.
  4. Maximum likelihood estimator under a Gaussian assumption.
- What if $p > n$?

  Ordinary least squares no longer uniquely defined, and often ill-behaved.

Recap

- **Ordinary least squares**:
  1. Affine hyperplane that minimizes least squares criterion.
  2. Approximates $\boldsymbol{y}$ as linear combination of columns of $\boldsymbol{X}$.
  3. In statistical learning, excess mean squared error is $O(p/n)$.
  4. Maximum likelihood estimator under a Gaussian assumption.
- What if $p > n$?

  Ordinary least squares no longer uniquely defined, and often ill-behaved.

  Use regularization:
  - force $\|\boldsymbol{w}\|_2^2$ to be small ("ridge regression") $\rightarrow$ can kernelize this
  - force $\|\boldsymbol{w}\|_1$ to be small ("Lasso")
  - force $\boldsymbol{w}$ to be sparse ("sparse regression")