CS189/CS289A
Introduction to Machine Learning
Lecture 9: Regression

Peter Bartlett

February 17, 2015

Outline

Outline

- Review: Decision theory.

Outline

- Review: Decision theory.
- Empirical risk minimization.

Outline

- Review: Decision theory.
- Empirical risk minimization.
    - Least squares.

Outline

- Review: Decision theory.
- Empirical risk minimization.
    - Least squares.
    - Normal equations.

Outline

- Review: Decision theory.
- Empirical risk minimization.
    - Least squares.
    - Normal equations.
- Linear model with additive Gaussian noise.

Outline

- Review: Decision theory.
- Empirical risk minimization.
    - Least squares.
    - Normal equations.
- Linear model with additive Gaussian noise.
    - Maximum likelihood is least squares.

Outline

- Review: Decision theory.
- Empirical risk minimization.
  - Least squares.
  - Normal equations.
- Linear model with additive Gaussian noise.
  - Maximum likelihood is least squares.
  - Distributions of parameter estimates.

Outline

- **Review: Decision theory.**
- Empirical risk minimization.
    - Least squares.
    - Normal equations.
- Linear model with additive Gaussian noise.
    - Maximum likelihood is least squares.
    - Distributions of parameter estimates.

Review: Quadratic loss

Regression with quadratic loss

Outcomes are in $\mathcal{Y} = \mathbb{R}$.
We consider the quadratic loss function, $\ell(\hat{y}, y) = (\hat{y} - y)^2$.
Risk is expected squared error:

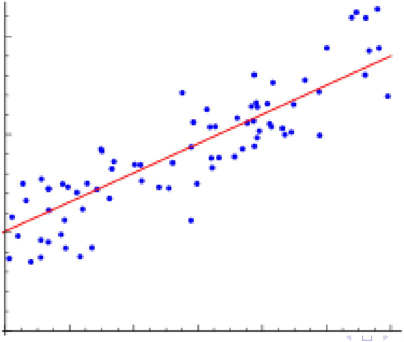
$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2.$$

Review: Quadratic loss

Regression with quadratic loss

Outcomes are in $\mathcal{Y} = \mathbb{R}$.
We consider the quadratic loss function, $\ell(\hat{y}, y) = (\hat{y} - y)^2$.
Risk is expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2.$$



Review: Quadratic loss

Risk:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2$$

Review: Quadratic loss

Risk:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 = \mathbb{E}\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

Review: Quadratic loss

Risk:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 = \mathbb{E}\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

For each X, we minimize the conditional expectation of the loss,

$$\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

Review: Quadratic loss

Risk:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 = \mathbb{E}\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

For each X, we minimize the conditional expectation of the loss,

$$\mathbb{E}\left[(f(X) - Y)^2 | X\right].$$

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

Review: Quadratic loss

Bias-variance decomposition

$$R(f) = \mathbb{E}\underbrace{\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \mathbb{E}\underbrace{\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}}$$

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

Review: Quadratic loss

Bias-variance decomposition

$$R(f) = \mathbb{E}\underbrace{\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \mathbb{E}\underbrace{\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}}$$

$$= \mathbb{E}\left[(f(X) - f^*(X))^2\right] + \mathbb{E}\left[(f^*(X) - Y)^2\right]$$

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

Review: Quadratic loss

Bias-variance decomposition

$$
\begin{aligned}
R(f) &= \underbrace{\mathbb{E}\left[(f(X) - \mathbb{E}[Y|X])^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[(\mathbb{E}[Y|X] - Y)^2\right]}_{\text{variance}} \\
&= \mathbb{E}\left[(f(X) - f^*(X))^2\right] + \mathbb{E}\left[(f^*(X) - Y)^2\right] \\
&= \mathbb{E}\left[(f(X) - f^*(X))^2\right] + R(f^*) .
\end{aligned}
$$

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

Review: Quadratic loss

Bias-variance decomposition

$$
\begin{aligned}
R(f) &= \underbrace{\mathbb{E}\left[\left(f(X) - \mathbb{E}[Y|X]\right)^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y|X] - Y\right)^2\right]}_{\text{variance}} \\
&= \mathbb{E}\left[\left(f(X) - f^*(X)\right)^2\right] + \mathbb{E}\left[\left(f^*(X) - Y\right)^2\right] \\
&= \mathbb{E}\left[\left(f(X) - f^*(X)\right)^2\right] + R\left(f^*\right).
\end{aligned}
$$

$$
R(f) - R^* = \mathbb{E}\left[\left(f(X) - f^*(X)\right)^2\right].
$$

The minimizer is $f^*(X) = \mathbb{E}[Y|X]$.

Linear regression

Consider $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and consider linear (affine) prediction rules

Linear regression

Consider $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and consider linear (affine) prediction rules,

$$F_{lin} := \left\{ x \mapsto x'\beta + \beta_0 : \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R} \right\}.$$

Linear regression

Consider $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and consider linear (affine) prediction rules,

$$F_{lin} := \left\{x \mapsto x'\beta + \beta_0 : \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}\right\}.$$

Two ways to motivate least squares:

Linear regression

Consider $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and consider linear (affine) prediction rules,

$$F_{lin} := \left\{ x \mapsto x'\beta + \beta_0 : \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R} \right\}.$$

Two ways to motivate least squares:

1. Consider the class of linear prediction rules. Minimize *empirical risk* over the class of linear prediction rules.

Linear regression

Consider $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and consider linear (affine) prediction rules,

$$F_{lin} := \left\{x \mapsto x'\beta + \beta_0 : \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}\right\}.$$

Two ways to motivate least squares:

1. Consider the class of linear prediction rules. Minimize *empirical risk* over the class of linear prediction rules.
2. Model the process generating the Y_is as a linear function of the X_is, plus additive Gaussian noise.
   Compute the maximum likelihood estimate for the linear coefficients.

Linear regression

Consider $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, and consider linear (affine) prediction rules,

$$F_{lin} := \left\{ x \mapsto x'\beta + \beta_0 : \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R} \right\}.$$

Two ways to motivate least squares:

1. Consider the class of linear prediction rules. Minimize *empirical risk* over the class of linear prediction rules.
2. Model the process generating the Y_is as a linear function of the X_is, plus additive Gaussian noise.
   Compute the maximum likelihood estimate for the linear coefficients.

In both cases, we arrive at the *normal equations*: the choice of β corresponds to a projection on to a linear sub-space.

Outline

- Review: Decision theory.
- **Empirical risk minimization.**
  - Least squares.
  - Normal equations.
- Linear model with additive Gaussian noise.
  - Maximum likelihood is least squares.
  - Distributions of parameter estimates.

Linear regression: Least squares

Risk and empirical risk

Risk is the expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\,(f(X) - Y)^2 \,.$$

Risk and empirical risk

Risk is the expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 .$$

Empirical risk is the sample average of squared error:

Linear regression: Least squares

Risk and empirical risk

Risk is the expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 .$$

Empirical risk is the sample average of squared error:

$$\hat{R}(f) = \hat{\mathbb{E}}_n \ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - Y_i\right)^2 .$$

Linear regression: Least squares

Risk and empirical risk

Risk is the expected squared error:

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2 .$$

Empirical risk is the sample average of squared error:

$$\hat{R}(f) = \hat{\mathbb{E}}_n \ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - Y_i\right)^2 .$$

Here, $\hat{\mathbb{E}}_n$ means expectation under the *empirical distribution*, which puts mass $1/n$ at each (X_i, Y_i) pair in the sample.

Linear regression: Least squares

We want to choose a linear prediction rule $f \in F_{lin}$ to minimize risk.

Linear regression: Least squares

We want to choose a linear prediction rule $f \in F_{lin}$ to minimize risk. One approach is to choose the linear prediction rule that minimizes empirical risk:

$$\hat{f} := \arg\min_{f \in F_{lin}} \hat{\mathbb{E}}_n \ell(f(X), Y)$$

Linear regression: Least squares

We want to choose a linear prediction rule $f \in F_{lin}$ to minimize risk. One approach is to choose the linear prediction rule that minimizes empirical risk:

$$\begin{aligned}\hat{f} &:= \arg\min_{f \in F_{lin}} \hat{\mathbb{E}}_n \ell(f(X), Y) \\ &= \arg\min_{f \in F_{lin}} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.\end{aligned}$$

Linear regression: Least squares

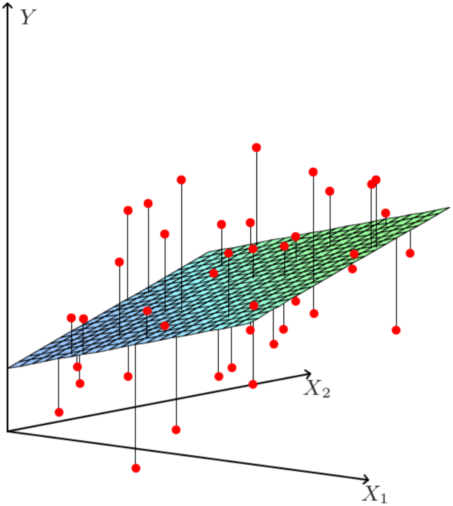


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Linear regression: Least squares

Just as we did when we were considering linear classifiers, we'll simplify notation by bundling the offset term (β_0) into the parameter vector β and assuming that the covariates X_i include a constant 1 component.

Linear regression: Least squares

Just as we did when we were considering linear classifiers, we'll simplify notation by bundling the offset term (β_0) into the parameter vector β and assuming that the covariates X_i include a constant 1 component.
Then $f \in F_{lin}$ is of the form $f(x) = x'\beta$.

Linear regression: Least squares

We wish to find $\hat{f} : x \mapsto x'\hat{\beta}$

Linear regression: Least squares

We wish to find $\hat{f} : x \mapsto x'\hat{\beta}$, where

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (X_i'\beta - Y_i)^2$$

Linear regression: Least squares

We wish to find $\hat{f} : x \mapsto x'\hat{\beta}$, where

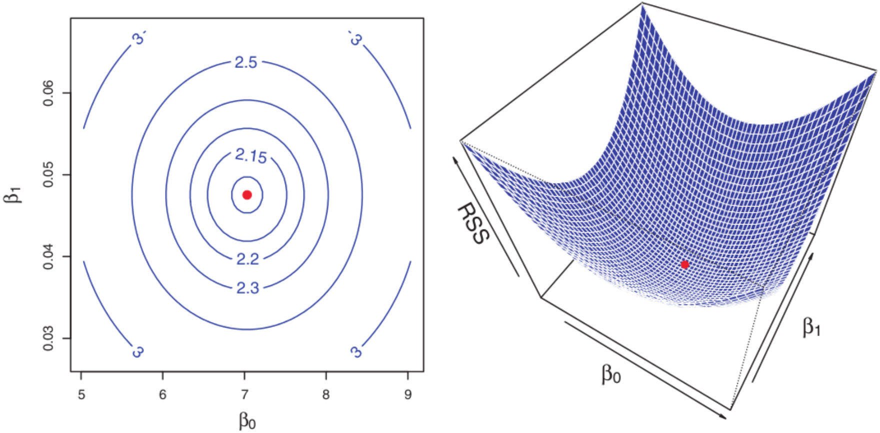
$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (X_i'\beta - Y_i)^2 \\ &= \arg\min_{\beta \in \mathbb{R}^p} \underbrace{\|X\beta - y\|^2}_{\text{RSS}},\end{aligned}$$

where the *design matrix* $X \in \mathbb{R}^{n \times p}$ and response vector $y \in \mathbb{R}^n$ are

$$X = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}, \qquad y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

(Think of $n \gg p$, so X is tall.)

Linear regression: Least squares



Linear regression: Least squares

Defining

$$RSS(\beta) = \frac{1}{2} \|X\beta - y\|^2$$

Linear regression: Least squares

Defining

$$\begin{aligned} RSS(\beta) &= \frac{1}{2}\left\|X\beta - y\right\|^2 \\ &= \frac{1}{2}\left(X\beta - y\right)'\left(X\beta - y\right) \end{aligned}$$

Linear regression: Least squares

Defining

$$\begin{aligned} RSS(\beta) &= \frac{1}{2}\left\|X\beta - y\right\|^2 \\ &= \frac{1}{2}\left(X\beta - y\right)'\left(X\beta - y\right) \\ &= \frac{1}{2}\beta' X' X\beta - y' X\beta + \frac{1}{2} y' y, \end{aligned}$$

Linear regression: Least squares

Defining

$$
\begin{aligned}
RSS(\beta) &= \frac{1}{2} \|X\beta - y\|^2 \\
&= \frac{1}{2} (X\beta - y)' (X\beta - y) \\
&= \frac{1}{2}\beta' X' X\beta - y' X\beta + \frac{1}{2} y' y,
\end{aligned}
$$

we can differentiate wrt β:

$$
\nabla_\beta RSS(\beta) = X'X\beta - X'y, \qquad \nabla^2_\beta RSS(\beta) = X'X.
$$

Linear regression: Least squares

Defining

$$
\begin{aligned}
RSS(\beta) &= \frac{1}{2}\left\|X\beta - y\right\|^2 \\
&= \frac{1}{2}\left(X\beta - y\right)'\left(X\beta - y\right) \\
&= \frac{1}{2}\beta' X' X\beta - y' X\beta + \frac{1}{2}y' y,
\end{aligned}
$$

we can differentiate wrt β:

$$
\nabla_\beta RSS(\beta) = X'X\beta - X'y, \qquad \nabla^2_\beta RSS(\beta) = X'X.
$$

Now, $X'X \succeq 0$

Linear regression: Least squares

Defining

$$\begin{aligned} RSS(\beta) &= \frac{1}{2}\|X\beta - y\|^2 \\ &= \frac{1}{2}(X\beta - y)'(X\beta - y) \\ &= \frac{1}{2}\beta' X' X \beta - y' X \beta + \frac{1}{2} y' y, \end{aligned}$$

we can differentiate wrt β:

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y, \qquad \nabla^2_\beta RSS(\beta) = X'X.$$

Now, $X'X \succeq 0$, so setting $\nabla_\beta RSS(\beta) = 0$ gives a minimum of RSS

Linear regression: Least squares

Defining

$$\begin{aligned} RSS(\beta) &= \frac{1}{2}\|X\beta - y\|^2 \\ &= \frac{1}{2}(X\beta - y)'(X\beta - y) \\ &= \frac{1}{2}\beta' X' X\beta - y' X\beta + \frac{1}{2} y' y, \end{aligned}$$

we can differentiate wrt β:

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y, \qquad \nabla_\beta^2 RSS(\beta) = X'X.$$

Now, $X'X \succeq 0$, so setting $\nabla_\beta RSS(\beta) = 0$ gives a minimum of RSS when

$$X'X\beta = X'y.$$

Linear regression: Least squares

Normal equations

$$X'X\beta = X'y.$$

Linear regression: Least squares

Normal equations

$$X'X\beta = X'y.$$

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Linear regression: Least squares

A projection viewpoint

We are aiming to find β to minimize $\|y - X\beta\|$.

Linear regression: Least squares

A projection viewpoint

We are aiming to find β to minimize $\|y - X\beta\|$.
Writing

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix},$$

Linear regression: Least squares

A projection viewpoint

We are aiming to find β to minimize $\|y - X\beta\|$.
Writing

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix},$$

we have

$$y - X\beta = y - \sum_{j=1}^{p} \beta_j x_j.$$

Linear regression: Least squares

A projection viewpoint

We are aiming to find β to minimize $\|y - X\beta\|$.
Writing

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix},$$

we have

$$y - X\beta = y - \sum_{j=1}^{p} \beta_j x_j.$$

That is, we want to find a linear combination of the columns $x_j \in \mathbb{R}^n$ of X that minimizes Euclidean distance to $y \in \mathbb{R}^n$.

Linear regression: Least squares

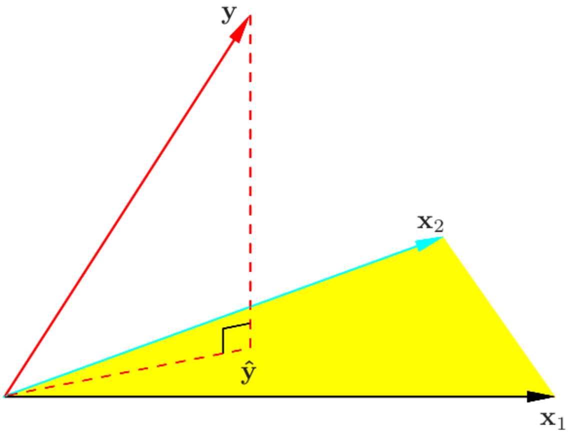


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

Linear regression: Least squares

Projection Theorem

The optimal approximation $\hat{y}$ in the space spanned by the columns x_j of X has an error $y - \hat{y}$ that is orthogonal to that column space.

Linear regression: Least squares

Projection Theorem

The optimal approximation $\hat{y}$ in the space spanned by the columns x_j of X has an error $y - \hat{y}$ that is orthogonal to that column space.
That is,

$$(y - \hat{y})'X = 0 \qquad \Leftrightarrow \qquad X'(y - X\beta) = 0 \qquad \Leftrightarrow \qquad X'y = X'X\beta.$$

Linear regression: Least squares

Projection Theorem

The optimal approximation $\hat{y}$ in the space spanned by the columns x_j of X has an error $y - \hat{y}$ that is orthogonal to that column space.
That is,

$$(y - \hat{y})'X = 0 \qquad \Leftrightarrow \qquad X'(y - X\beta) = 0 \qquad \Leftrightarrow \qquad X'y = X'X\beta.$$

Normal equations

$$X'X\beta = X'y.$$

Linear regression: Least squares

Projection Theorem

The optimal approximation $\hat{y}$ in the space spanned by the columns x_j of X has an error $y - \hat{y}$ that is orthogonal to that column space.
That is,

$$(y - \hat{y})'X = 0 \quad \Leftrightarrow \quad X'(y - X\beta) = 0 \quad \Leftrightarrow \quad X'y = X'X\beta.$$

Normal equations

$$X'X\beta = X'y.$$

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Linear regression: Least squares

Risk versus empirical risk

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2,$$

Linear regression: Least squares

Risk versus empirical risk

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2,$$

$$\hat{R}(f) = \hat{\mathbb{E}}_n\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - Y_i\right)^2.$$

Linear regression: Least squares

Risk versus empirical risk

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2,$$

$$\hat{R}(f) = \hat{\mathbb{E}}_n\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - Y_i\right)^2.$$

When is the risk of the empirical risk minimizer $\hat{f}$ close to the minimal risk?

Linear regression: Least squares

Risk versus empirical risk

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2,$$

$$\hat{R}(f) = \hat{\mathbb{E}}_n\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

When is the risk of the empirical risk minimizer $\hat{f}$ close to the minimal risk?

It suffices if

Linear regression: Least squares

Risk versus empirical risk

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\,(f(X) - Y)^2\,,$$

$$\hat{R}(f) = \hat{\mathbb{E}}_n\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2\,.$$

When is the risk of the empirical risk minimizer $\hat{f}$ close to the minimal risk?

It suffices if

1. the X come from a compact set,

Linear regression: Least squares

Risk versus empirical risk

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\left(f(X) - Y\right)^2,$$

$$\hat{R}(f) = \hat{\mathbb{E}}_n\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

When is the risk of the empirical risk minimizer $\hat{f}$ close to the minimal risk?

It suffices if

1. the X come from a compact set,
2. the Y_is have tails that are not too heavy (e.g., sub-Gaussian),

Linear regression: Least squares

Risk versus empirical risk

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\,(f(X) - Y)^2\,,$$

$$\hat{R}(f) = \hat{\mathbb{E}}_n\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2\,.$$

When is the risk of the empirical risk minimizer $\hat{f}$ close to the minimal risk?

It suffices if

1. the X come from a compact set,
2. the Y_is have tails that are not too heavy (e.g., sub-Gaussian),
3. $\|\hat{\theta}\|$ is not too large, and

Linear regression: Least squares

Risk versus empirical risk

$$R(f) = \mathbb{E}\ell(f(X), Y) = \mathbb{E}\,(f(X) - Y)^2\,,$$

$$\hat{R}(f) = \hat{\mathbb{E}}_n\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2\,.$$

When is the risk of the empirical risk minimizer $\hat{f}$ close to the minimal risk?

It suffices if

1. the X come from a compact set,
2. the Y_is have tails that are not too heavy (e.g., sub-Gaussian),
3. $\|\hat{\theta}\|$ is not too large, and
4. $n \gg p$.

Outline

- Review: Decision theory.
- Empirical risk minimization.
  - Least squares.
  - Normal equations.
- **Linear model with additive Gaussian noise.**
  - Maximum likelihood is least squares.
  - Distributions of parameter estimates.

Linear regression: probability models

Linear model

Linear regression: probability models

Linear model

Model the conditional distribution of Y given $X = x$ as

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Linear regression: probability models

Linear model

Model the conditional distribution of Y given $X = x$ as

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

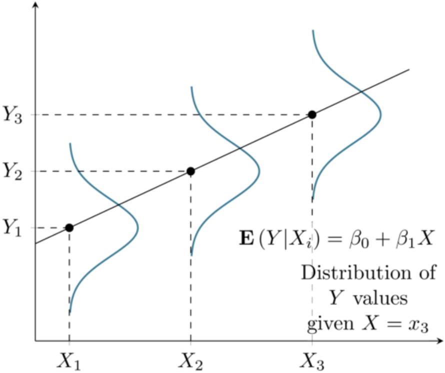
Linear regression: probability models

Linear model

Model the conditional distribution of Y given $X = x$ as

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.



Linear models

How to estimate β?

How to estimate β?

Maximum likelihood

Linear models

How to estimate β?

Maximum likelihood

Conditional likelihood:

$$L(\beta) = \prod_{i=1}^{n} p(Y_i | X_i, \beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - X_i' \beta)^2 \right).$$

Linear models

How to estimate β?

Maximum likelihood

Conditional likelihood:

$$L(\beta) = \prod_{i=1}^{n} p(Y_i|X_i, \beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i'\beta)^2\right).$$

Log likelihood:

$$\ell(\beta) = (\text{function of } \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i'\beta)^2.$$

Linear models

How to estimate β?

Maximum likelihood

Conditional likelihood:

$$L(\beta) = \prod_{i=1}^{n} p(Y_i|X_i, \beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - X_i'\beta)^2 \right).$$

Log likelihood:

$$\ell(\beta) = (\text{function of } \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - X_i'\beta)^2.$$

Maximum likelihood is least squares.

Linear models

Bias and variance of $\hat{\beta}$

Linear models

Bias and variance of $\hat{\beta}$

Fix X.

Linear models

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

Linear models

Bias and variance of $\hat\beta$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$\mathbb{E}\hat\beta$$

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$\mathbb{E}\hat{\beta} = \mathbb{E}\left[(X'X)^{-1}X'y\right]$$

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$\begin{aligned}\mathbb{E}\hat{\beta} &= \mathbb{E}\left[(X'X)^{-1}X'y\right] \\ &= (X'X)^{-1}X'\mathbb{E}y\end{aligned}$$

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$\begin{aligned}
\mathbb{E}\hat{\beta} &= \mathbb{E}\left[(X'X)^{-1}X'y\right] \\
&= (X'X)^{-1}X'\mathbb{E}y \\
&= (X'X)^{-1}X'X\beta
\end{aligned}$$

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$\begin{aligned}
\mathbb{E}\hat{\beta} &= \mathbb{E}\left[(X'X)^{-1}X'y\right] \\
&= (X'X)^{-1}X'\mathbb{E}y \\
&= (X'X)^{-1}X'X\beta \\
&= \beta.
\end{aligned}$$

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$\begin{aligned}\mathbb{E}\hat{\beta} &= \mathbb{E}\left[(X'X)^{-1}X'y\right] \\ &= (X'X)^{-1}X'\mathbb{E}y \\ &= (X'X)^{-1}X'X\beta \\ &= \beta.\end{aligned}$$

$$\mathrm{Cov}(\hat{\beta})$$

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$\begin{aligned}
\mathbb{E}\hat{\beta} &= \mathbb{E}\left[(X'X)^{-1}X'y\right] \\
&= (X'X)^{-1}X'\mathbb{E}y \\
&= (X'X)^{-1}X'X\beta \\
&= \beta. \\
\mathrm{Cov}(\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right]
\end{aligned}$$

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$\begin{aligned}
\mathbb{E}\hat{\beta} &= \mathbb{E}\left[(X'X)^{-1}X'y\right] \\
&= (X'X)^{-1}X'\mathbb{E}y \\
&= (X'X)^{-1}X'X\beta \\
&= \beta. \\
\mathrm{Cov}(\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] \\
&= \mathbb{E}\left[\left((X'X)^{-1}X'y-\beta\right)\left((X'X)^{-1}X'y-\beta\right)'\right]
\end{aligned}$$

Bias and variance of $\hat{\beta}$

Fix X. Provided $\mathbb{E}y = X\beta$ and $\mathrm{Cov}(y) = \sigma^2 I$,

$$
\begin{aligned}
\mathbb{E}\hat{\beta} &= \mathbb{E}\left[(X'X)^{-1}X'y\right] \\
&= (X'X)^{-1}X'\mathbb{E}y \\
&= (X'X)^{-1}X'X\beta \\
&= \beta. \\
\mathrm{Cov}(\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] \\
&= \mathbb{E}\left[\left((X'X)^{-1}X'y-\beta\right)\left((X'X)^{-1}X'y-\beta\right)'\right] \\
&\ \vdots \\
&= \sigma^2(X'X)^{-1}.
\end{aligned}
$$

Linear models

Statistical tests

Linear models

Statistical tests

If the data is generated by a linear model with $\mathcal{N}(0, \sigma^2)$ noise,

Statistical tests

If the data is generated by a linear model with $\mathcal{N}(0, \sigma^2)$ noise, then:

Statistical tests

If the data is generated by a linear model with $\mathcal{N}(0, \sigma^2)$ noise, then:

- We can compute distributions of parameter estimates:

$$\hat{\beta} \sim \mathcal{N}(\beta, (X'X)^{-1}\sigma^2)$$

Linear models

Statistical tests

If the data is generated by a linear model with $\mathcal{N}(0, \sigma^2)$ noise, then:

- We can compute distributions of parameter estimates:

$$\hat{\beta} \sim \mathcal{N}(\beta, (X'X)^{-1}\sigma^2)$$

- We can calculate approximate confidence sets for the parameters: the standardized coefficient is

$$z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{v_j}},$$

which is normal (here, v_j is the jth diagonal entry of $(X'X)^{-1}$).

Linear models

Statistical tests

If the data is generated by a linear model with $\mathcal{N}(0, \sigma^2)$ noise, then:

- We can compute distributions of parameter estimates:

$$\hat{\beta} \sim \mathcal{N}(\beta, (X'X)^{-1}\sigma^2)$$

- We can calculate approximate confidence sets for the parameters: the standardized coefficient is

$$z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{v_j}},$$

  which is normal (here, v_j is the jth diagonal entry of $(X'X)^{-1}$).

- In particular, we can design tests for non-zero values of parameters.

Linear models

Statistical tests

If the data is generated by a linear model with $\mathcal{N}(0, \sigma^2)$ noise, then:

- We can compute distributions of parameter estimates:

$$\hat{\beta} \sim \mathcal{N}(\beta, (X'X)^{-1}\sigma^2)$$

- We can calculate approximate confidence sets for the parameters: the standardized coefficient is

$$z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{v_j}},$$

  which is normal (here, v_j is the jth diagonal entry of $(X'X)^{-1}$).
- In particular, we can design tests for non-zero values of parameters.

(see text)

Outline

- Review: Decision theory.
- Empirical risk minimization.
  - Least squares.
  - Normal equations.
- Linear model with additive Gaussian noise.
  - Maximum likelihood is least squares.
  - Distributions of parameter estimates.