Association studies and regression

Fall 2016

Sriram Sankararaman
Acknowledgments: Fei Sha, Ameet Talwalkar
Administration

• HW1 will be posted today.
• Due in two weeks.
• Send me an email if you are still not enrolled!
Review of last lecture

- Basics of statistical inference.
  - Parameter estimation, Hypothesis testing, Interval estimation
  - Different types of mistakes
- Multiple hypothesis testing
  - Would like to control: FWER and FDR
  - \( \text{FWER} = P(\text{At least one false positive}) \). Bonferroni procedure controls FWER.
  - \( \text{FDR} = \text{Expected fraction of false discoveries} \). Benjamini-Hochberg procedure controls FDR.
Motivation

Population Association Studies and GWAS

Linear regression
- Univariate solution
- Probabilistic interpretation
- Statistical properties of MLE
- Computational and numerical optimization

Logistic regression
- General setup
- Maximum likelihood estimation
- Gradient descent
- Newton’s method

Application to GWAS
Path to Personalized Genomics

• Basic biology
  • Phenotype is a function of genotype and environment \( i.e., \) learn \( P=f(G,E) \)
  • How much of the phenotype is genetic vs environmental?
  • Find the genetic and environmental factors associated with phenotype
  • Finding drug targets

• Disease prediction
  • Given genetic data, predict the phenotype for an individual
Finding genetic factors that influence a phenotype

- Will be working with a population of individuals
- Genotypes of a population of individuals
  - SNPs (most common)
- Phenotypes in the same set of individuals
  - Binary (disease-healthy), ordinal (number of children or years of schooling), continuous (height, gene expression), heterogeneous and high-dimensional (images, text, videos)
Outline

Motivation

Population Association Studies and GWAS

Linear regression
  Univariate solution
  Probabilistic interpretation
  Statistical properties of MLE
  Computational and numerical optimization

Logistic regression
  General setup
  Maximum likelihood estimation
  Gradient descent
  Newton’s method

Application to GWAS
“Unrelated” population of individuals measured for a phenotype

<table>
<thead>
<tr>
<th>Individual</th>
<th>SNPs</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A C G A A C G G T A A</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C C G G T C G G T C T</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C C T A T G A A A A A A</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>A T G A A G G G T A T</td>
<td>0</td>
</tr>
</tbody>
</table>
Population Association Studies

- Find an association between a SNP and phenotype
- Really want to find a SNP that is causal.

<table>
<thead>
<tr>
<th>Individual</th>
<th>SNPs</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>A C G A A C G G T A A</td>
<td>1 1 0 0</td>
<td></td>
</tr>
</tbody>
</table>
Population Association Studies

- Find an association between a SNP and phenotype
- Really want to find a SNP that is causal.

Simplest form: does the genotype at a single SNP predict phenotype?
Recall some definitions from lecture 1

- **Locus**: position along the chromosome (could be a single base or longer).
- **Allele**: set of variants at a locus
- **Genotype**: sequence of alleles along the loci of an individual

**Individual 1**: (1,CT), (2,GG)
**Individual 2**: (1,TT), (2,GA)
Recall some definitions from lecture 1

- Pick one allele as reference. Other allele is called alternate allele.
- Represent a genotype as the number of copies of the reference allele.
- Each genotype at a single base can be 0/1/2

Locus 1: C is reference

| Individual 1 has genotype 1 | $x_{1,1} = 1$ |
| Individual 2 has genotype 0 | $x_{2,1} = 0$ |
Genome-wide Association Studies (GWAS)

Perform a population association study across the genome

![Graph showing the increase in published GWA reports from 2005 to 2013.](image)
Genome-wide Association Studies (GWAS)

Perform a population association study across the genome
• Simplest form: Univariate regression between SNP and continuous phenotype.

\[ y_i \quad \text{Phenotype for individual } i \]
\[ x_{i,j} \quad \text{Genotype for individual } i \text{ at SNP } j \]
Outline

Motivation

Population Association Studies and GWAS

Linear regression
- Univariate solution
- Probabilistic interpretation
- Statistical properties of MLE
- Computational and numerical optimization

Logistic regression
- General setup
- Maximum likelihood estimation
- Gradient descent
- Newton’s method

Application to GWAS
Goals

- Predict continuous-valued output from inputs.
- Classification refers to binary outputs.
  - Output, response: phenotype
  - Input, covariate, predictor: genotype
  - Common practice to test genotype at a single SNP (univariate regression). We will not make that assumption for now.
GWAS
Single SNP association testing

- Phenotype \approx Mean\ phenoytpe + Effect size \times Genotype + Noise
- Learn parameters that predict the phenotype “well”
  - Squared diffence : (actual – predicted phenotype)^2
Linear regression

- Output: $y \in \mathbb{R}$
- Input: $x \in \mathbb{R}^m$
- Find $f : x \rightarrow y$ that predicts $y$ “well”
- Assume $f(x) = \beta_0 + \beta^T x$.

Linear in parameters (hence the name)

$\beta_0$ is called an intercept or bias.

$\beta = (\beta_1, \cdots, \beta_m)$: weights, parameters, or parameter vector

Sometimes $\tilde{\beta} = (\beta_0, \cdots, \beta_m)$ called parameters

What does “well” mean?
Linear regression

- Output: $y \in \mathbb{R}$
- Input: $x \in \mathbb{R}^m$
- Find $f : x \rightarrow y$ that predicts $y$ "well"
- Assume $f(x) = \beta_0 + \beta^T x$.
- Linear in parameters (hence the name)

\( \beta_0 \) is called an intercept or bias.

\( \beta = (\beta_1, \cdots, \beta_m) \): weights, parameters, or parameter vector

Sometimes \( \tilde{\beta} = (\beta_0, \cdots, \beta_m) \) called parameters

What does "well" mean?
Linear regression

Setup

- We have labeled (training) data. \( D = \{(x_i, y_i), i = 1, \ldots, n\} \)
- \( i \): individual
- \( j \): SNP
- \( y_i \): Phenotype of individual \( i \)
- \( x_{i,j} \): Genotype of individual \( i \) at SNP \( j \), \( x_{i,j} \in \{0, 1, 2\} \)
Linear regression

What does “well” mean?

- Minimize the residual sum of squares $RSS$

$$RSS(\tilde{\beta}) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$= \sum_{i=1}^{n} (y_i - (\beta_0 + \beta^T x_i))^2$$

$$= \sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{j=1}^{m} \beta_j x_{i,j}))^2$$

- Ordinary Least Squares estimator

$$\tilde{\beta}^{OLS} = \arg \min_{\tilde{\beta}} RSS(\tilde{\beta})$$
A simple case: $x$ is just one-dimensional ($m=1$)

**Residual sum of squares**

$$RSS(\tilde{\beta}) = \sum_i [y_i - f(x_i)]^2 = \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2$$

We denote $x_i = (x_{i,1})$ (scalar).

Identify stationary points by taking derivative with respect to parameters and setting to zero

$$\frac{\partial RSS(\tilde{\beta})}{\partial \beta_0} = 0 \Rightarrow -2 \sum_i [y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$\frac{\partial RSS(\tilde{\beta})}{\partial \beta_1} = 0 \Rightarrow -2 \sum_i [y_i - (\beta_0 + \beta_1 x_i)]x_i = 0$$
A simple case: $x$ is just one-dimensional ($m=1$)

**Residual sum of squares**

$$RSS(\mathbf{\beta}) = \sum_i [y_i - f(x_i)]^2 = \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2$$

We denote $x_i = (x_{i,1})$ (scalar).

**Identify stationary points by taking derivative with respect to parameters and setting to zero**

$$\frac{\partial RSS(\mathbf{\beta})}{\partial \beta_0} = 0 \Rightarrow -2 \sum_i [y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$\frac{\partial RSS(\mathbf{\beta})}{\partial \beta_1} = 0 \Rightarrow -2 \sum_i [y_i - (\beta_0 + \beta_1 x_i)]x_i = 0$$
Simplify these expressions to get “Normal Equations”

\[ \sum_{i} y_i = n \beta_0 + \beta_1 \sum_{i} x_i \]

\[ \sum_{i} x_i y_i = \beta_0 \sum_{i} x_i + \beta_1 \sum_{i} x_i^2 \]

We have two equations and two unknowns! Do some algebra to get:

\[ \beta_1 = \frac{\sum_{i} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i} (x_i - \bar{x})^2} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x} \]

where \( \bar{x} = \frac{1}{n} \sum_{i} x_i \) and \( \bar{y} = \frac{1}{n} \sum_{i} y_i \).
Simplify these expressions to get “Normal Equations”

\[
\sum_i y_i = n \beta_0 + \beta_1 \sum_i x_i
\]

\[
\sum_i x_i y_i = \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2
\]

We have two equations and two unknowns! Do some algebra to get:

\[
\beta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}
\]

and

\[
\beta_0 = \bar{y} - \beta_1 \bar{x}
\]

where \( \bar{x} = \frac{1}{n} \sum_i x_i \) and \( \bar{y} = \frac{1}{n} \sum_i y_i \).
Simplify these expressions to get “Normal Equations”

\[ \sum_i y_i = n\beta_0 + \beta_1 \sum_i x_i \]

\[ \sum_i x_i y_i = \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 \]

We have two equations and two unknowns! Do some algebra to get:

\[ \beta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \]

and \[ \beta_0 = \bar{y} - \beta_1 \bar{x} \]

where \( \bar{x} = \frac{1}{n} \sum_i x_i \) and \( \bar{y} = \frac{1}{n} \sum_i y_i \).
Why is minimizing RSS sensible?

**Probabilistic interpretation**

- Noisy observation model

\[ Y = \beta_0 + \beta_1 X + \epsilon \]

where \( \epsilon \sim \mathcal{N}(0, \sigma^2) \) is a Gaussian random variable

- Likelihood of one training sample \((x_i, y_i)\)

\[
p(y_i|x_i, (\beta_0, \beta_1, \sigma^2)) = \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2}}
\]
Why is minimizing RSS sensible?

Probabilistic interpretation

- Noisy observation model

\[ Y = \beta_0 + \beta_1 X + \epsilon \]

where \( \epsilon \sim \mathcal{N}(0, \sigma^2) \) is a Gaussian random variable

- Likelihood of one training sample \((x_i, y_i)\)

\[
p(y_i|x_i, (\beta_0, \beta_1, \sigma^2)) = \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{[y_i-(\beta_0+\beta_1 x_i)]^2}{2\sigma^2}}
\]
Probabilistic interpretation

Log-likelihood of the training data $\mathcal{D}$ (assuming i.i.d)

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) \equiv \log P(\mathcal{D}|(\beta_0, \beta_1, \sigma^2))$$
Probabilistic interpretation

Log-likelihood of the training data $\mathcal{D}$ (assuming i.i.d)

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) \equiv \log P(\mathcal{D}|(\beta_0, \beta_1, \sigma^2))$$

$$= \log \prod_{i=1}^{n} p(y_i|x_i, (\beta_0, \beta_1, \sigma^2)) = \sum_i \log p(y_i|x_i, (\beta_0, \beta_1, \sigma^2))$$
Probabilistic interpretation

Log-likelihood of the training data \( \mathcal{D} \) (assuming i.i.d)

\[ \mathcal{L}\mathcal{L}(\beta_0, \beta_1, \sigma^2) \equiv \log P(\mathcal{D}|(\beta_0, \beta_1, \sigma^2)) \]

\[ = \log \prod_{i=1}^{n} p(y_i|x_i, (\beta_0, \beta_1, \sigma^2)) = \sum_{i} \log p(y_i|x_i, (\beta_0, \beta_1, \sigma^2)) \]

\[ = \sum_{i} \left\{ -\frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} - \log \sqrt{2\pi\sigma^2} \right\} \]
Probabilistic interpretation

Log-likelihood of the training data $\mathcal{D}$ (assuming i.i.d)

$$\mathcal{L}\mathcal{L}(\beta_0, \beta_1, \sigma^2) \equiv \log P(\mathcal{D}|(\beta_0, \beta_1, \sigma^2))$$

$$= \log \prod_{i=1}^{n} p(y_i|x_i, (\beta_0, \beta_1, \sigma^2)) = \sum_{i} \log p(y_i|x_i, (\beta_0, \beta_1, \sigma^2))$$

$$= \sum_{i} \left\{ -\frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} - \log \sqrt{2\pi\sigma^2} \right\}$$

$$= -\frac{1}{2\sigma^2} \sum_{i} [y_i - (\beta_0 + \beta_1 x_i)]^2 - \frac{n}{2} \log \sigma^2 - \frac{n}{2} \log(2\pi)$$
Probabilistic interpretation

Log-likelihood of the training data $\mathcal{D}$ (assuming i.i.d)

$$\mathcal{L}_\mathcal{L}(\beta_0, \beta_1, \sigma^2) \equiv \log P(\mathcal{D}|(\beta_0, \beta_1, \sigma^2))$$

$$= \log \prod_{i=1}^{n} p(y_i|x_i, (\beta_0, \beta_1, \sigma^2)) = \sum_{i} \log p(y_i|x_i, (\beta_0, \beta_1, \sigma^2))$$

$$= \sum_{i} \left\{ - \frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} - \log \sqrt{2\pi\sigma^2} \right\}$$

$$= -\frac{1}{2\sigma^2} \sum_{i} [y_i - (\beta_0 + \beta_1 x_i)]^2 - \frac{n}{2} \log \sigma^2 - \frac{n}{2} \log(2\pi)$$

$$= -\frac{1}{2} \left\{ \frac{1}{\sigma^2} \sum_{i} [y_i - (\beta_0 + \beta_1 x_i)]^2 + n \log \sigma^2 \right\} + \text{const}$$

Relationship between minimizing RSS and maximizing the log-likelihood?
Maximum likelihood estimation

Estimating $\sigma$, $\beta_0$ and $\beta_1$ can be done in two steps

- Maximize over $\beta_0$ and $\beta_1$

$$\arg\max_{\beta_0, \beta_1} LL(\beta_0, \beta_1, \sigma^2)$$

$$\Leftrightarrow \arg\min_{\beta_0, \beta_1} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 \leftarrow \text{That is } RSS(\tilde{\beta})!$$

- Maximize over $s = \sigma^2$ (we could estimate $\sigma$ directly)

$$\frac{\partial LL(\beta_0, \beta_1, s)}{\partial s} = -\frac{1}{2} \left\{ -\frac{1}{s^2} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 + n \frac{1}{s} \right\} = 0$$

$$\rightarrow \sigma^* = s^* = \frac{1}{n} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 = \frac{RSS(\tilde{\beta})}{n}$$
Maximum likelihood estimation

Estimating \( \sigma, \beta_0 \) and \( \beta_1 \) can be done in two steps

- Maximize over \( \beta_0 \) and \( \beta_1 \)

\[
\arg \max_{\beta_0, \beta_1} LL(\beta_0, \beta_1, \sigma^2)
\]
\[
\Leftrightarrow \arg \min_{\beta_0, \beta_1} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 \leftarrow \text{That is } RSS(\tilde{\beta})!
\]

- Maximize over \( s = \sigma^2 \) (we could estimate \( \sigma \) directly)

\[
\frac{\partial LL(\beta_0, \beta_1, s)}{\partial s} = -\frac{1}{2} \left\{ -\frac{1}{s^2} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 + n \frac{1}{s} \right\} = 0
\]
\[
\rightarrow \sigma^* = s^* = \frac{1}{n} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 = \frac{RSS(\tilde{\beta})}{n}
\]
Maximum likelihood estimation

Estimating \( \sigma, \beta_0 \) and \( \beta_1 \) can be done in two steps

- Maximize over \( \beta_0 \) and \( \beta_1 \)

\[
\arg \max_{\beta_0, \beta_1} \mathcal{L}(\beta_0, \beta_1, \sigma^2) \quad \Leftrightarrow \quad \arg \min_{\beta_0, \beta_1} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 \quad \text{← That is RSS}(\tilde{\beta})!
\]

- Maximize over \( s = \sigma^2 \) (we could estimate \( \sigma \) directly)

\[
\frac{\partial \mathcal{L}(\beta_0, \beta_1, s)}{\partial s} = -\frac{1}{2} \left\{ -\frac{1}{s^2} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 + n \frac{1}{s} \right\} = 0
\]

\[
\rightarrow \sigma^* = s^* = \frac{1}{n} \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 = \frac{RSS(\tilde{\beta})}{n}
\]
How does this probabilistic interpretation help us?

- It gives a solid footing to our intuition: minimizing $\text{RSS}(\tilde{\beta})$ is a sensible thing based on reasonable modeling assumptions.
- Estimating $\sigma^*$ tells us how much noise there could be in our predictions. For example, it allows us to place confidence intervals around our predictions.
Linear regression when $\mathbf{x}$ is $m$-dimensional

Probabilistic model

\[
\begin{align*}
  y_i &= \tilde{\beta}^T \tilde{x}_i + \epsilon_i \\
  \epsilon_i &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2)
\end{align*}
\]

where we have redefined some variables (by augmenting)

\[
\begin{align*}
  \tilde{x} &\leftarrow [1 \ x_1 \ x_2 \ \ldots \ x_m]^T, \quad \tilde{\beta} &\leftarrow [\beta_0 \ \beta_1 \ \beta_2 \ \ldots \ \beta_m]^T
\end{align*}
\]

The likelihood of the parameters $\theta = (\tilde{\beta}, \sigma^2)$

\[
\begin{align*}
  \mathcal{L}(\theta) &= \prod_{i=1}^{n} \Pr(y_i | \tilde{x}_i, \theta) \\
  &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \tilde{\beta}^T \tilde{x}_i)^2}{2\sigma^2} \right)
\end{align*}
\]

Choose parameters to maximize the likelihood.
\[ \hat{\theta} = \arg \max_{\theta} \log \mathcal{L}(\theta) \]

\[ \mathcal{L}(\theta) \equiv \log \mathcal{L}(\theta) \]

\[ = \sum_{i=1}^{n} \log \Pr(y_i|x_i, \theta) \]

\[ = -\sum_{i=1}^{n} \frac{(y_i - \tilde{\beta}^T \tilde{x}_i)^2}{2\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2) \]

\[ = -\frac{1}{2\sigma^2} RSS(\tilde{\beta}) - \frac{n}{2} \log(2\pi\sigma^2) \]

Maximizing the likelihood is equivalent to minimizing the RSS, i.e., least squares.
Linear regression
Computing the MLE

\[ \frac{\partial \mathcal{L}}{\partial \sigma^2}(\hat{\theta}) = 0 \]
\[ \nabla \tilde{\beta} \mathcal{L}(\hat{\theta}) = 0 \]
Computing the MLE

\[ \frac{\partial \mathcal{L}}{\partial \sigma^2}(\hat{\theta}) = 0 \]

\[ \frac{\partial}{\partial \sigma^2} \left( -\frac{1}{2\sigma^2} \text{RSS}(\tilde{\beta}) - \frac{n}{2} \log(2\pi\sigma^2) \right) = 0 \]

\[ \frac{1}{2(\sigma^2)^2} \text{RSS}(\tilde{\beta}) - \frac{n}{2\sigma^2} = 0 \]

\[ \hat{\sigma}^2 = \frac{\text{RSS}(\tilde{\beta})}{n} \]
Linear regression

Computing the MLE

\[ \nabla_{\tilde{\beta}} \mathcal{L}(\hat{\theta}) = 0 \]

\[ \nabla_{\tilde{\beta}} \left( -\frac{1}{2\sigma^2} RSS(\tilde{\beta}) - \frac{n}{2} \log(2\pi\sigma^2) \right) = 0 \]

\[ \nabla_{\tilde{\beta}} RSS(\tilde{\beta}) = 0 \]
**RSS**($\tilde{\beta}$) in matrix form

$$RSS(\tilde{\beta}) = \sum_i [y_i - (\beta_0 + \sum_j \beta_j x_{i,j})]^2 = \sum_i [y_i - \tilde{\beta}^T \tilde{x}_i]^2$$

which leads to

$$RSS(\tilde{\beta}) = \sum_i (y_i - \tilde{\beta}^T \tilde{x}_i)(y_i - \tilde{x}_i^T \tilde{\beta}) = \sum_i \left\{ \tilde{\beta}^T \tilde{x}_i \tilde{x}_i^T \tilde{\beta} - 2y_i \tilde{x}_i^T \tilde{\beta} + \text{const.} \right\} = \left\{ \tilde{\beta}^T \left( \sum_i \tilde{x}_i \tilde{x}_i^T \right) \tilde{\beta} - 2 \left( \sum_i y_i \tilde{x}_i^T \right) \tilde{\beta} \right\} + \text{const.}$$
**RSS($\tilde{\beta}$) in matrix form**

$$RSS(\tilde{\beta}) = \sum_i [y_i - (\beta_0 + \sum_j \beta_j x_{i,j})]^2 = \sum_i [y_i - \tilde{\beta}^T \tilde{x}_i]^2$$

which leads to

$$RSS(\tilde{\beta}) = \sum_i (y_i - \tilde{\beta}^T \tilde{x}_i)(y_i - \tilde{x}_i^T \tilde{\beta})$$

$$= \sum_i \left\{ \tilde{\beta}^T \tilde{x}_i \tilde{x}_i^T \tilde{\beta} - 2 y_i \tilde{x}_i^T \tilde{\beta} + \text{const.} \right\}$$

$$= \left\{ \tilde{\beta}^T \left( \sum_i \tilde{x}_i \tilde{x}_i^T \right) \tilde{\beta} - 2 \left( \sum_i y_i \tilde{x}_i^T \right) \tilde{\beta} \right\} + \text{const.}$$
**RSS(\(\tilde{\beta}\)) in matrix form**

\[
\text{RSS}(\tilde{\beta}) = \sum_i \left[ y_i - (\beta_0 + \sum_j \beta_j x_{i,j}) \right]^2 = \sum_i \left[ y_i - \tilde{\beta}^T \tilde{x}_i \right]^2
\]

which leads to

\[
\text{RSS}(\tilde{\beta}) = \sum_i (y_i - \tilde{\beta}^T \tilde{x}_i)(y_i - \tilde{x}_i^T \tilde{\beta})
\]

\[
= \sum_i \left\{ \tilde{\beta}^T \tilde{x}_i \tilde{x}_i^T \tilde{\beta} - 2y_i \tilde{x}_i^T \tilde{\beta} + \text{const.} \right\}
\]

\[
= \left\{ \tilde{\beta}^T \left( \sum_i \tilde{x}_i \tilde{x}_i^T \right) \tilde{\beta} - 2 \left( \sum_i y_i \tilde{x}_i^T \right) \tilde{\beta} \right\} + \text{const.}
\]
### Linear regression

**RSS(\(\tilde{\beta}\)) in matrix form**

\[
RSS(\tilde{\beta}) = \sum_i [y_i - (\beta_0 + \sum_j \beta_j x_{i,j})]^2 = \sum_i [y_i - \tilde{\beta}^T \tilde{x}_i]^2
\]

which leads to

\[
RSS(\tilde{\beta}) = \sum_i (y_i - \tilde{\beta}^T \tilde{x}_i)(y_i - \tilde{x}_i^T \tilde{\beta})
\]

\[
= \sum_i \left\{ \tilde{\beta}^T \tilde{x}_i \tilde{x}_i^T \tilde{\beta} - 2y_i \tilde{x}_i^T \tilde{\beta} + \text{const.} \right\}
\]

\[
= \left\{ \tilde{\beta}^T \left( \sum_i \tilde{x}_i \tilde{x}_i^T \right) \tilde{\beta} - 2 \left( \sum_i y_i \tilde{x}_i^T \right) \tilde{\beta} \right\} + \text{const.}
\]
**RSS(\(\tilde{\beta}\)) in new notations**

**Design matrix and target vector**

\[
\tilde{X} = \begin{pmatrix}
\tilde{x}_1^T \\
\vdots \\
\tilde{x}_n^T
\end{pmatrix} \in \mathbb{R}^{n \times (m+1)}, \quad y = \begin{pmatrix}
y_1 \\
\vdots \\
y_n
\end{pmatrix}
\]

**Compact expression**

\[
\begin{align*}
RSS(\tilde{\beta}) &= \| \tilde{X} \tilde{\beta} - y \|^2_2 \\
&= (\tilde{X} \tilde{\beta} - y)^T (\tilde{X} \tilde{\beta} - y) \\
&= (\tilde{\beta}^T \tilde{X}^T - y^T) (\tilde{X} \tilde{\beta} - y) \\
&= \left\{ \tilde{\beta}^T \tilde{X}^T \tilde{X} \tilde{\beta} - 2 (\tilde{X}^T y)^T \tilde{\beta} \right\} + \text{const}
\end{align*}
\]
**Design matrix and target vector**

\[
\tilde{X} = \begin{pmatrix}
\tilde{x}_1^T \\
\vdots \\
\tilde{x}_n^T
\end{pmatrix} \in \mathbb{R}^{n \times (m+1)}, \quad y = \begin{pmatrix}
y_1 \\
\vdots \\
y_n
\end{pmatrix}
\]

**Compact expression**

\[
RSS(\tilde{\beta}) = \| \tilde{X} \tilde{\beta} - y \|^2_2
\]
\[
= (\tilde{X} \tilde{\beta} - y)^T \tilde{X} \tilde{\beta} - y
\]
\[
= (\tilde{\beta}^T \tilde{X}^T - y^T) \tilde{X} \tilde{\beta} - y
\]
\[
= \left\{ \tilde{\beta}^T \tilde{X}^T \tilde{X} \tilde{\beta} - 2 (\tilde{X}^T y)^T \tilde{\beta} \right\} + \text{const}
\]
Solution in matrix form

Compact expression

\[
RSS(\tilde{\beta}) = \| \tilde{X} \tilde{\beta} - y \|^2 = \left\{ \tilde{\beta}^T \tilde{X}^T \tilde{X} \tilde{\beta} - 2 \left( \tilde{X}^T y \right)^T \tilde{\beta} \right\} + \text{const}
\]

Gradients of Linear and Quadratic Functions

- \( \nabla_x b^T x = b \)
- \( \nabla_x x^T Ax = 2Ax \) (symmetric \( A \))

Normal equation

\[
\nabla_{\tilde{\beta}} RSS(\tilde{\beta}) \propto 2\tilde{X}^T \tilde{X} \tilde{\beta} - 2\tilde{X}^T y = 0
\]

This leads to the ordinary least squares (OLS) solution

\[
\hat{\beta} = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T y
\]
Solution in matrix form

Compact expression

\[
RSS(\tilde{\beta}) = \|\tilde{X}\tilde{\beta} - y\|_2^2 = \left\{\tilde{\beta}^T \tilde{X}^T \tilde{X} \tilde{\beta} - 2 (\tilde{X}^T y)^T \tilde{\beta}\right\} + \text{const}
\]

Gradients of Linear and Quadratic Functions

- \(\nabla_x b^T x = b\)
- \(\nabla_x x^T Ax = 2Ax\) (symmetric \(A\))

Normal equation

\[
\nabla_{\tilde{\beta}} RSS(\tilde{\beta}) \propto 2\tilde{X}^T \tilde{X} \tilde{\beta} - 2\tilde{X}^T y = 0
\]

This leads to the ordinary least squares (OLS) solution

\[
\hat{\beta} = \left(\tilde{X}^T \tilde{X}\right)^{-1} \tilde{X}^T y
\]
Solution in matrix form

Compact expression

\[ \text{RSS}(\tilde{\beta}) = \| \tilde{X} \tilde{\beta} - y \|_2^2 = \left\{ \tilde{\beta}^T \tilde{X}^T \tilde{X} \tilde{\beta} - 2 \left( \tilde{X}^T y \right)^T \tilde{\beta} \right\} + \text{const} \]

Gradients of Linear and Quadratic Functions

- \( \nabla_x b^T x = b \)
- \( \nabla_x x^T A x = 2A x \) (symmetric \( A \))

Normal equation

\[ \nabla_{\tilde{\beta}} \text{RSS}(\tilde{\beta}) \propto 2 \tilde{X}^T \tilde{X} \tilde{\beta} - 2 \tilde{X}^T y = 0 \]

This leads to the ordinary least squares (OLS) solution

\[ \hat{\beta} = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T y \]
Remember the $\sigma^2$ parameter

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n}$$
Mini-Summary

• Linear regression is the linear combination of features \( f : \mathbf{x} \rightarrow y \), with
\[
f(x) = \beta_0 + \sum_j \beta_j x_j = \beta_0 + \mathbf{\beta}^T \mathbf{x}
\]
• If we minimize residual sum of squares as our learning objective, we get a closed-form solution of parameters
• Probabilistic interpretation: maximum likelihood if assuming residual is Gaussian distributed
• MLE
\[
\hat{\beta} = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \mathbf{y}
\]
\[
\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n}
\]
Linear regression
Statistical properties of the Least squares estimator

\[ \mathbb{E} \left[ \hat{\beta} | \tilde{X} \right] = \tilde{\beta} \]
\[ \text{Cov} \left[ \hat{\beta} | \tilde{X} \right] = \sigma^2 (\tilde{X}^T \tilde{X})^{-1} \]

Assumptions:
- Linear model is correct
- Exogeneous covariates. \[ \mathbb{E} \left[ \epsilon | \tilde{X} \right] = 0 \]
- Uncorrelated, homoskedastic errors. \[ \text{Cov} \left[ \epsilon | \tilde{X} \right] = \sigma^2 I_n \]
- Does not assume normally distributed errors.
- OLS is best linear unbiased estimator (Gauss-Markov theorem) i.e., has minimum variance among all linear unbiased estimators.
OLS is unbiased

\[ y = \tilde{X}\tilde{\beta} + \epsilon \]

\[ \mathbb{E} \left[ \hat{\beta} \middle| \tilde{X} \right] = \mathbb{E} \left[ (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y \middle| \tilde{X} \right] \]
OLS is unbiased

\[ y = \tilde{X}\tilde{\beta} + \epsilon \]

\[
E \left[ \hat{\beta} \middle| \tilde{X} \right] = E \left[ \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T y \middle| \tilde{X} \right] = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T E \left[ y \middle| \tilde{X} \right]
\]
OLS is unbiased

\[ y = \tilde{X}\tilde{\beta} + \epsilon \]

\[
\mathbb{E} \left[ \hat{\beta} \middle| \tilde{X} \right] = \mathbb{E} \left[ \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T y \middle| \tilde{X} \right] = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \mathbb{E} \left[ y \middle| \tilde{X} \right] = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \mathbb{E} \left[ \left( \tilde{X}\tilde{\beta} + \epsilon \right) \middle| \tilde{X} \right]
\]
OLS is unbiased

\[ y = \tilde{X}\tilde{\beta} + \epsilon \]

\[
\mathbb{E} \left[ \tilde{\beta} \mid \tilde{X} \right] = \mathbb{E} \left[ (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y \mid \tilde{X} \right] \\
= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \mathbb{E} \left[ y \mid \tilde{X} \right] \\
= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \mathbb{E} \left[ (\tilde{X}\tilde{\beta} + \epsilon) \mid \tilde{X} \right] \\
= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \left( \mathbb{E} \left[ \tilde{X}\tilde{\beta} \mid \tilde{X} \right] + \mathbb{E} \left[ \epsilon \mid \tilde{X} \right] \right)
OLS is unbiased

\[ y = \mathbf{\tilde{X}} \mathbf{\tilde{\beta}} + \epsilon \]

\[
\mathbb{E} \left[ \hat{\beta} \mid \tilde{X} \right] = \mathbb{E} \left[ \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T y \mid \tilde{X} \right] \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \mathbb{E} \left[ y \mid \tilde{X} \right] \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \mathbb{E} \left[ \left( \tilde{X} \mathbf{\tilde{\beta}} + \epsilon \right) \mid \tilde{X} \right] \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \left( \mathbb{E} \left[ \tilde{X} \mathbf{\tilde{\beta}} \mid \tilde{X} \right] + \mathbb{E} \left[ \epsilon \mid \tilde{X} \right] \right) \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \left( \mathbf{\tilde{X}} \mathbf{\tilde{\beta}} + \mathbb{E} \left[ \epsilon \mid \tilde{X} \right] \right)
OLS is unbiased

\[ y = \tilde{X}\tilde{\beta} + \epsilon \]

\[
\mathbb{E} \left[ \hat{\beta} \middle| \tilde{X} \right] = \mathbb{E} \left[ \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T y \middle| \tilde{X} \right] \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \mathbb{E} \left[ y \middle| \tilde{X} \right] \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \mathbb{E} \left[ \left( \tilde{X}\tilde{\beta} + \epsilon \right) \middle| \tilde{X} \right] \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \left( \mathbb{E} \left[ \tilde{X}\tilde{\beta} \middle| \tilde{X} \right] + \mathbb{E} \left[ \epsilon \middle| \tilde{X} \right] \right) \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \left( \tilde{X}\tilde{\beta} + \mathbb{E} \left[ \epsilon \middle| \tilde{X} \right] \right) \\
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \tilde{X}\tilde{\beta}
OLS is unbiased

\[ y = \tilde{X}\tilde{\beta} + \epsilon \]

\[
E \left[ \tilde{\beta} \bigg| \tilde{X} \right] = E \left[ \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T y \bigg| \tilde{X} \right]
\]

\[
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T E \left[ y \big| \tilde{X} \right]
\]

\[
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T E \left[ \left( \tilde{X}\tilde{\beta} + \epsilon \right) \bigg| \tilde{X} \right]
\]

\[
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \left( E \left[ \tilde{X}\tilde{\beta} \bigg| \tilde{X} \right] + E \left[ \epsilon \bigg| \tilde{X} \right] \right)
\]

\[
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \left( \tilde{X}\tilde{\beta} + E \left[ \epsilon \bigg| \tilde{X} \right] \right)
\]

\[
= \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X}^T \tilde{\beta}
\]

\[
= \tilde{\beta}
\]
Exercises:
Use similar argument to show that

\[ \text{Cov} \left[ \hat{\beta} | X \right] = \sigma^2 (X^T X)^{-1} \]

We can say more if we assume the errors are normally distributed

\[ \hat{\beta} \sim \mathcal{N}(\tilde{\beta}, \sigma^2 (X^T X)^{-1}) \]
Computational complexity

Bottleneck of computing the solution?

\[ \hat{\beta} = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X} y \]

Matrix multiply of \( \tilde{X}^T \tilde{X} \in \mathbb{R}^{(m+1) \times (m+1)} \)

Inverting the matrix \( \tilde{X}^T \tilde{X} \)

How many operations do we need?

- \( O(nm^2) \) for matrix multiplication
- \( O(m^3) \) (e.g., using Gauss-Jordan elimination) or \( O(m^{2.373}) \) (recent theoretical advances) for matrix inversion
- Impractical for very large \( m \) or \( n \)
Computational complexity

Bottleneck of computing the solution?

\[ \hat{\beta} = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X} y \]

Matrix multiply of \( \tilde{X}^T \tilde{X} \in \mathbb{R}^{(m+1) \times (m+1)} \)
Inverting the matrix \( \tilde{X}^T \tilde{X} \)

How many operations do we need?

- \( O(nm^2) \) for matrix multiplication
- \( O(m^3) \) (e.g., using Gauss-Jordan elimination) or \( O(m^{2.373}) \) (recent theoretical advances) for matrix inversion
- Impractical for very large \( m \) or \( n \)
Computational complexity

Bottleneck of computing the solution?

\[ \hat{\beta} = \left( \tilde{X}^T \tilde{X} \right)^{-1} \tilde{X} y \]

Matrix multiply of \( \tilde{X}^T \tilde{X} \in \mathbb{R}^{(m+1) \times (m+1)} \)
Inverting the matrix \( \tilde{X}^T \tilde{X} \)

How many operations do we need?

\begin{itemize}
  \item \( O(nm^2) \) for matrix multiplication
  \item \( O(m^3) \) (e.g., using Gauss-Jordan elimination) or \( O(m^{2.373}) \) (recent theoretical advances) for matrix inversion
  \item Impractical for very large \( m \) or \( n \)
\end{itemize}
Alternative method: using numerical optimization

- Use gradient descent (more details when we talk about logistic regression next)
• Assuming the probabilistic model is correct, the MLE is unbiased.
• Computing the MLE can be done by solving normal equations or by numerical optimization.
Prediction
Given a new input $x_*$, our best guess of $y$: $y_* = \hat{\beta}_0 + \hat{\beta}^T x_*$
Hypothesis testing
Test if $\beta_j = 0$ (Wald test)

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 \left[(X^TX)^{-1}\right]_{j,j})$$

Under $H_0 : \beta_j = 0$

$$\hat{\beta}_j \sim \mathcal{N}(0, \sigma^2 \left[(X^TX)^{-1}\right]_{j,j})$$

$$\frac{\hat{\beta}_j}{\hat{\sigma}_j} \approx \mathcal{N}(0, 1)$$

Here $\hat{\sigma}_j = \hat{\sigma}^2 \left[(X^TX)^{-1}\right]_{j,j}$
(Generalized) Likelihood Ratio test. Remember lecture 2.

\[ \Lambda \equiv \frac{L(\hat{\beta}_0)}{L(\hat{\beta})} \]

\[-2 \log \Lambda = 2 \left[ \log L(\hat{\beta}) - \log L(\hat{\beta}_0) \right] \]

\[-2 \log \Lambda \to \chi^2_1 \]

\(\hat{\beta}_0: \) MLE with \(\beta_j = 0\)
Do the observed statistics match the model assumptions?

- Residuals $e_i = y_i - \hat{\beta}^T x_i$
- If the model assumptions hold, residuals must be independent and normally distributed with equal variance.
- How do we check this?
Are residuals normally distributed (or more generally, distributed according to a known distribution) ?

Q-Q plot
  - Plot empirical quantiles vs theoretical quantiles.
  - Close to diagonal implies empirical distribution is close to theoretical distribution.
Linear regression
Model diagnostics

- Q-Q plot

![Q-Q plot for normally distributed errors](image1)

Normally distributed errors

![Q-Q plot for Student-t distributed errors](image2)

Student-t distributed errors