MITx: Statistics, Computation & Applications

Statistics Refresher

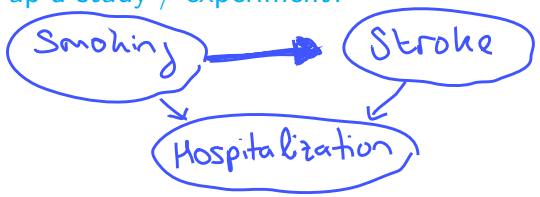
Lecture 1: Observational Studies and Experiments

Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
- Mammography: screening women for breast cancer by X-rays

✶ Does mammography speed up detection by enough to matter?

✶ How would you approach this problem? What is important when setting up a study / experiment?



Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
- Mammography: screening women for breast cancer by X-rays

✶ Does mammography speed up detection by enough to matter?

✶ How would you approach this problem? What is important when setting up a study / experiment?

⇒ Perform a **randomized, controlled, double-blind experiment** to minimize the problem of **confounding**

HIP study: First large-scale randomized controlled experiment on mammography performed in 1960s

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

Reference: D. A. Freedman. *Statistical Models: Theory and Practice*, 2009.

Mammography and breast cancer

- Breast cancer is one of the most common malignancies among women in the United States
- Mammography: screening women for breast cancer by X-rays

✶ Does mammography speed up detection by enough to matter?

✶ How would you approach this problem? What is important when setting up a study / experiment?

⇒ Perform a **randomized, controlled, double-blind experiment** to minimize the problem of **confounding**

HIP study: First large-scale randomized controlled experiment on mammography performed in 1960s

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

Reference: D. A. Freedman. *Statistical Models: Theory and Practice*, 2009.

HIP study

Which rates should be compared to show the efficacy of treatment?

- Seems natural to compare those who accepted screening to those who refused
- But this is an **observational** comparison!
- Becomes clear when comparing the death rates from all other causes
- Instead compare the whole treatment group against the whole control group

✶ **Intention-to-treat analysis**

Hypothesis testing

- Death rate from breast cancer in control group: 0.0020 ($= \frac{63}{31000}$)
- Death rate from breast cancer in treatment group: 0.0013 ($= \frac{39}{31000}$)

Is the difference in death rates between the treatment and control group sufficient to establish that mammography reduces the risk of death from breast cancer?

⇒ Perform a **hypothesis test**

Hypothesis testing

- Death rate from breast cancer in control group: 0.0020 ($= \frac{63}{31000}$)
- Death rate from breast cancer in treatment group: 0.0013 ($= \frac{39}{31000}$)

Is the difference in death rates between the treatment and control group sufficient to establish that mammography reduces the risk of death from breast cancer?

$\Rightarrow$ Perform a **hypothesis test**

Hypothesis testing

1. Determine a **model**:

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi \leq 0.002$ or $\lambda \leq 63$ $\pi \neq 0.002$, $\lambda \neq 63$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

   $T :=$ Number of deaths under H_0:
   $T \sim \text{binomial}(31'000, 0.002)$ or $T \sim \text{Poisson}(63)$

Hypothesis testing

1. Determine a **model**:

   $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$

2. Determine a (mutually exclusive) **null hypothesis** and **alternative**:

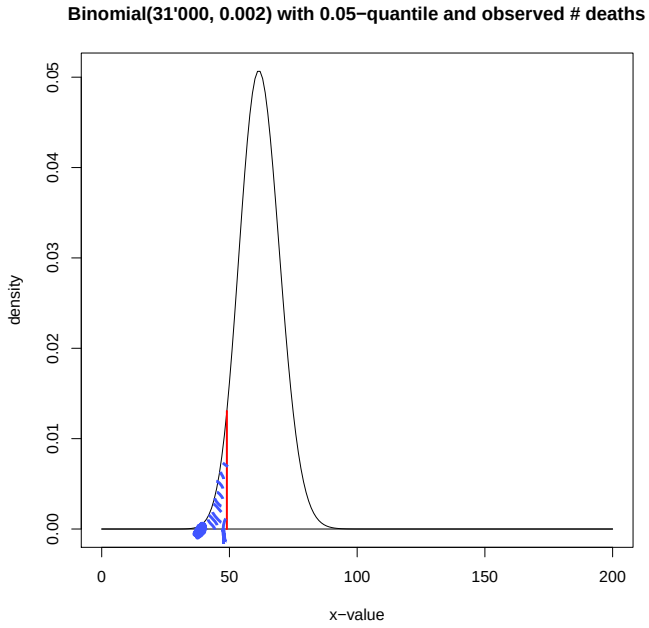
   Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$

3. Determine a **test statistic** (quantity that can differentiate between H_0 and H_A, and whose distribution under H_0 you can compute):

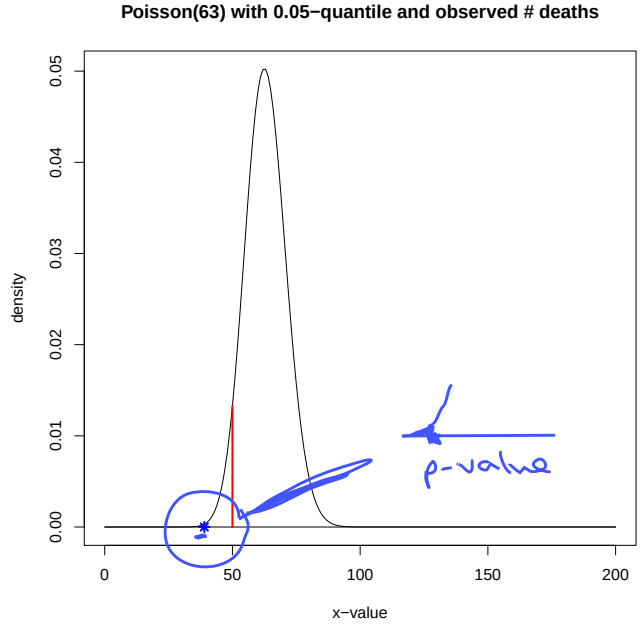
   $T :=$ Number of deaths under H_0:
   $T \sim \text{binomial}(31'000, 0.002)$ or $T \sim \text{Poisson}(63)$

4. Determine a **significance level** (α), i.e. the probability of rejecting H_0 when H_0 is true: e.g. $\alpha = 0.05$

Binomial distribution



Poisson distribution



P-value

- Probability under H_0 to obtain the observed value or a more extreme value of the test statistic

  $\Rightarrow$ p-value is always between 0 and 1!

  For mammography study: p-value is 0.0012 under binomial model and 0.0008 under Poisson model

- Smallest significance level for which H_0 just gets rejected
- Can be used for hypothesis testing: Reject H_0 if p-value $\leq \alpha$
- Quantifies significance of alternative

P-value

- Probability under H_0 to obtain the observed value or a more extreme value of the test statistic

  $\Rightarrow$ p-value is always between 0 and 1!

  For mammography study: p-value is 0.0012 under binomial model and 0.0008 under Poisson model

- Smallest significance level for which H_0 just gets rejected
- Can be used for hypothesis testing: Reject H_0 if p-value $\leq \alpha$
- Quantifies significance of alternative

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

- Signficiance level bounds probability of type I error: $\mathbb{P}(\text{type I error}) \leq \alpha$
- Power $:= 1 - \mathbb{P}(\text{type II error})$

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

- Signficiance level bounds probability of type I error: $\mathbb{P}(\text{type I error}) \leq \alpha$
- Power $:= 1 - \mathbb{P}(\text{type II error})$
- Note that there is a trade-off between the probability of making a type I error and the probability of making a type II error (Why?)
- Note that power of 1-sided test is usually higher than for 2-sided test (Why?)

  $\Rightarrow$ Perform 1-sided test if you are only interested in detecting deviations in one direction

Power

	retain H_0	reject H_0
H_0 true	-	type I error
H_A true	type II error	-

- Signficiance level bounds probability of type I error: $\mathbb{P}(\text{type I error}) \leq \alpha$
- Power $:= 1 - \mathbb{P}(\text{type II error})$
- Note that there is a trade-off between the probability of making a type I error and the probability of making a type II error (Why?)
- Note that power of 1-sided test is usually higher than for 2-sided test (Why?)

  $\Rightarrow$ Perform 1-sided test if you are only interested in detecting deviations in one direction

Recap: Hypothesis testing

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

1. Model: $X_1, \ldots, X_{31'000} \sim \text{Bernoulli}(\pi)$ or $Y \sim \text{Poisson}(\lambda)$
2. Null hypothesis (H_0): $\pi = 0.002$ or $\lambda = 63$
   Alternative (H_A): $\pi < 0.002$ or $\lambda < 63$
3. Test statistic $T =$ Number of deaths under H_0
   $T \sim \text{binomial}(31'000, 0.002)$ or $T \sim \text{Poisson}(63)$
4. Significance level: $\alpha = 0.05$

Any important assumption that we should relax?

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Knowing that 102 subjects died and that number of treatments / controls is 31'000, what is probability that deaths are so unevenly distributed?

Alternative test: assume no knowledge of π_{control}

	breast cancer deaths (rate)	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

Null hypothesis: $\pi_{\text{control}} = \pi_{\text{treatment}}$ Alternative: $\pi_{\text{control}} > \pi_{\text{treatment}}$

Knowing that 102 subjects died and that number of treatments / controls is 31'000, what is probability that deaths are so unevenly distributed?

- Test statistic T: number of deaths among the treated individuals
- Model: **Hypergeometric distribution**:

$$\mathbb{P}_{H_0}(T = 39) = \frac{\binom{31'000}{39}\binom{31'000}{63}}{\binom{62'000}{102}}$$

- p-value $= 0.011 = \sum_{i=0}^{39} \mathbb{P}_{H_0}(T = i)$

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Advantages:

- Does not assume knowledge about the true probability of dying due to breast cancer in the control population

Fisher's exact test

Test on previous slide is known as **Fisher's exact test**

Advantages:

- Does not assume knowledge about the true probability of dying due to breast cancer in the control population

Shortcomings:

- Assumes knowledge of the margins (i.e., row and column sums)
- Alternative is Bernard's test (estimates the margins)
  for more details on this test, see e.g. `http://www.nbi.dk/~petersen/Teaching/Stat2009/Barnard_ExactTest_TwoBinomials.pdf`
- Both tests are difficult to perform on large tables for computational reasons

References

- For a statistics review, including controlled experiments and observational studies (chapters 1 and 2) and hypothesis testing (chapter 26-29):

  D. Freedman, R. Pisani, R. Purves. *Statistics*. 2007.

- For how to perform hypothesis testing in `R` (chapter 4):

  P. Dalgaard. *Introductory Statistics with R*. 2002.

- For observational studies and experiments, including the HIP study (chapter 1):

  D. Freedman. *Statistical Models: Theory and Practice*. 2009.

MITx: Statistics, Computation & Applications

Statistics Refresher

Lecture 2: Hypothesis Testing

Testing the efficacy of a sleeping drug

Testing the efficacy of a sleeping drug

patient	1	2	3	4	5	6	7	8	9	10	mean
drug	6.1	7.0	8.2	7.6	6.5	7.8	6.9	6.7	7.4	5.8	7.00
placebo	5.2	7.9	3.9	4.7	5.3	4.8	4.2	6.1	3.8	6.3	5.22

Question: Does the drug increase hours of sleep enough to matter?

Testing the efficacy of a sleeping drug

patient	1	2	3	4	5	6	7	8	9	10	mean
drug	6.1	7.0	8.2	7.6	6.5	7.8	6.9	6.7	7.4	5.8	7.00
placebo	5.2	7.9	3.9	4.7	5.3	4.8	4.2	6.1	3.8	6.3	5.22

Question: Does the drug increase hours of sleep enough to matter?

- Model: Difference of hours of sleep between drug and placebo

$$X_1, \ldots, X_{10} \sim \mathcal{N}(\mu, \sigma^2)$$

Testing the efficacy of a sleeping drug

patient	1	2	3	4	5	6	7	8	9	10	mean
drug	6.1	7.0	8.2	7.6	6.5	7.8	6.9	6.7	7.4	5.8	7.00
placebo	5.2	7.9	3.9	4.7	5.3	4.8	4.2	6.1	3.8	6.3	5.22

Question: Does the drug increase hours of sleep enough to matter?

- Model: Difference of hours of sleep between drug and placebo

$$X_1, \ldots, X_{10} \sim \mathcal{N}(\mu, \sigma^2)$$

- Null hypothesis (H_0): $\mu = 0$; Alternative (H_A): $\mu > 0$

Testing the efficacy of a sleeping drug

patient	1	2	3	4	5	6	7	8	9	10	mean
drug	6.1	7.0	8.2	7.6	6.5	7.8	6.9	6.7	7.4	5.8	7.00
placebo	5.2	7.9	3.9	4.7	5.3	4.8	4.2	6.1	3.8	6.3	5.22

Question: Does the drug increase hours of sleep enough to matter?

- Model: Difference of hours of sleep between drug and placebo

$$X_1, \ldots, X_{10} \sim \mathcal{N}(\mu, \sigma^2)$$

- Null hypothesis (H_0): $\mu = 0$; Alternative (H_A): $\mu > 0$

- Test statistic: standardized average difference $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$; under H_0:

$$\frac{\bar{X}_n}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Testing the efficacy of a sleeping drug

patient	1	2	3	4	5	6	7	8	9	10	mean
drug	6.1	7.0	8.2	7.6	6.5	7.8	6.9	6.7	7.4	5.8	7.00
placebo	5.2	7.9	3.9	4.7	5.3	4.8	4.2	6.1	3.8	6.3	5.22

Question: Does the drug increase hours of sleep enough to matter?

- Model: Difference of hours of sleep between drug and placebo

$$X_1, \ldots, X_{10} \sim \mathcal{N}(\mu, \sigma^2)$$

- Null hypothesis (H_0): $\mu = 0$; Alternative (H_A): $\mu > 0$

- Test statistic: standardized average difference $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$; under H_0:

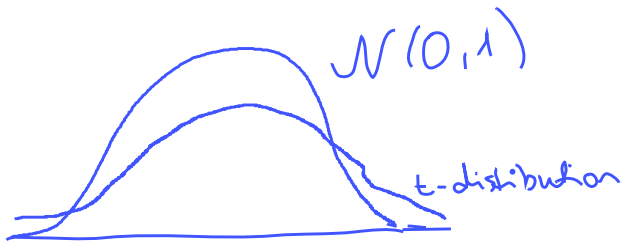
$$\frac{\bar{X}_n}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Note: Shortcoming of this test (**z-test**): assumes σ is known

t-test

- Doesn't assume that the true σ is known
- Uses estimate of σ instead: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$
- Test statistic: $T = \frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}}$; under the null hypothesis:

$$\frac{\bar{X}_n}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1} \qquad \text{(see handout for a derivaton)}$$



t-test

- Doesn't assume that the true σ is known
- Uses estimate of σ instead: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$
- Test statistic: $T = \frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}}$; under the null hypothesis:

$$\frac{\bar{X}_n}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1} \qquad \text{(see handout for a derivaton)}$$

t-distribution: Let $T \sim t_n$. Then

- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$, then $\sum_{i=1}^n X_i^2 \sim \chi^2_n$; $\quad t_n \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_n/n}}$
- $t_n \xrightarrow{n\to\infty} \mathcal{N}(0,1)$
- $\mathbb{E}(T) = 0, \quad \mathrm{Var}(T) = \frac{n}{n-2} > 1$

  $\Rightarrow$ estimating σ introduces uncertainty; more weight in tails

Notes on the t-statistic

t-statistic: $T_n := \dfrac{\bar{X}_n - \mu}{\sqrt{\hat{\sigma}^2/n}}$, where $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$

and $\hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$

and $X_{n+1} - \bar{X}_n \overset{\text{indep}}{\sim} \mathcal{N}(0,1)$

χ^2-distribution, n deg. of freedom: $\sum_{i=1}^n Y_i^2 \sim \chi^2_n$, where $Y_i \sim \mathcal{N}(0,1)$

$= \sum_{i=1}^n Z_i$, where $Z_i \sim \chi^2_1$

t distribution, n deg. of freedom: $\dfrac{Y}{\sqrt{Z/n}} \sim t_n$, where $Y \sim \mathcal{N}(0,1)$ and $Z \sim \chi^2_n$

Claim $T_n \sim t_{n-1}$

Proof:

$$T_n = \frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)}{\sqrt{\frac{1}{n-1}\sqrt{\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2}}}$$

where $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \chi^2_{n-1}$

We need to show: $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \chi^2_{n-1}$

$$\underbrace{\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2}_{\chi^2_n} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n + \bar{X}_n - \mu)^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \frac{n}{\sigma^2}(\bar{X}_n - \mu)^2 + 2\cdot\frac{1}{\sigma^2}(\bar{X}_n - \mu)\sum_{i=1}^n (X_i - \bar{X}_n)$$

$\chi^2_n - \chi^2_1 = \chi^2_{n-1}$

$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \chi^2_{n-1}$

$\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)^2$, where $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, so $\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1$

Recap: Testing the efficacy of a sleeping drug

patient	1	2	3	4	5	6	7	8	9	10
drug	6.1	7.0	8.2	7.6	6.5	7.8	6.9	6.7	7.4	5.8
placebo	5.2	7.9	3.9	4.7	5.3	4.8	4.2	6.1	3.8	6.3

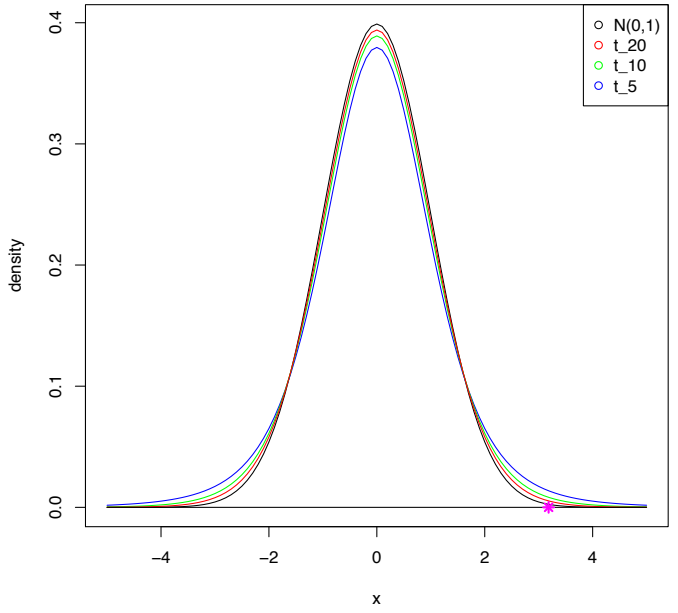
Question: Does the drug increase the length of sleep enough to matter?

- Model: Difference of sleeping time between drug and placebo

$$X_1, \ldots, X_{10} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

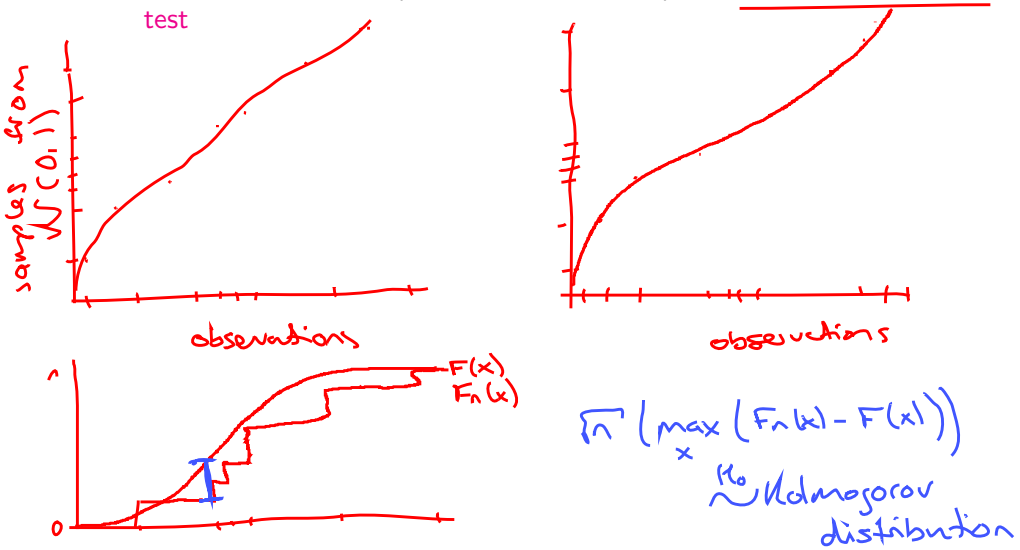
- Null hypothesis (H_0): $\mu = 0$; Alternative (H_A): $\mu > 0$
- z-statistic: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$; assumes σ is known $\quad \bar{X}_n \sim \mathcal{N}(0, \frac{\sigma^2}{n})$
- t-statistic: $\frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$, where $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$

Comparison of normal versus t-distribution



Remarks

- When using a t-test, check assumption of normality!
  - E.g. using a qq-plot (quantile-quantile plot) or a Kolmogorov-Smirnov test



Remarks

- When using a t-test, check assumption of normality!
  - E.g. using a qq-plot (quantile-quantile plot) or a Kolmogorov-Smirnov test
- **Alternative:** Wilcoxon signed rank test
  - Model: $X_1, \ldots X_n \sim F$ symmetric around a mean μ
  - Test statistic: $W = \sum_{i=1}^n \text{sgn}(X_i - \mu) R_i$, where R_i is rank of $|X_i - \mu|$
  - One can show that this test statistic is asymptotically ($n \to \infty$) normally distributed

  $\Rightarrow$ build hypothesis test based on asymptotic distribution

Remarks

- When using a t-test, check assumption of normality!
  - E.g. using a qq-plot (quantile-quantile plot) or a Kolmogorov-Smirnov test
- **Alternative:** Wilcoxon signed rank test
  - Model: $X_1, \dots X_n \sim F$ symmetric around a mean μ
  - Test statistic: $W = \sum_{i=1}^{n} \operatorname{sgn}(X_i - \mu) R_i$, where R_i is rank of $|X_i - \mu|$
  - One can show that this test statistic is asymptotically ($n \to \infty$) normally distributed

    $\Rightarrow$ build hypothesis test based on asymptotic distribution
- Sometimes you might not have paired data: all hypothesis tests discussed in this lecture have unpaired version; as to be expected, unpaired tests are usually less powerful

Confidence interval

Confidence interval

The **confidence interval at level** $1 - \alpha$ is defined as

$$I(X) = \{\mu \mid H_0 \text{ is not rejected at significance level } \alpha\}$$

Confidence interval

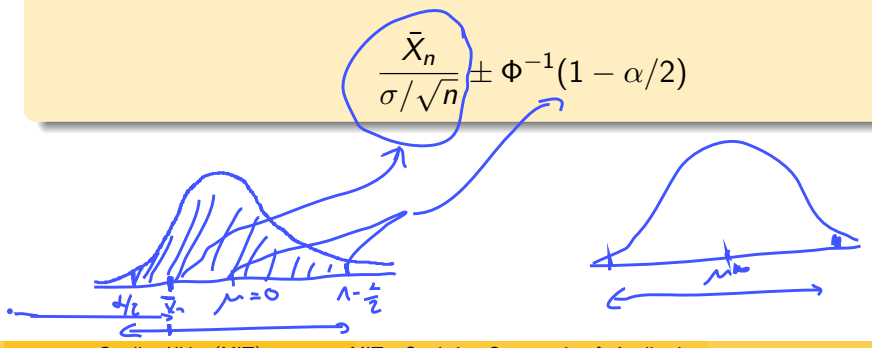
The **confidence interval at level** $1 - \alpha$ is defined as

$$I(X) = \{\mu \mid H_0 \text{ is not rejected at significance level } \alpha\}$$

- $I(X)$ is a random quantity; it depends on the observations
- Often computed based on 2-sided testing and normal approximation

Example: For the sleeping drug example the confidence interval is

$$\frac{\bar{X}_n}{\sigma/\sqrt{n}} \pm \Phi^{-1}(1 - \alpha/2)$$



Confidence interval

The **confidence interval at level** $1-\alpha$ is defined as

$$I(X) = \{\mu \mid H_0 \text{ is not rejected at significance level } \alpha\}$$

- $I(X)$ is a random quantity; it depends on the observations
- Often computed based on 2-sided testing and normal approximation

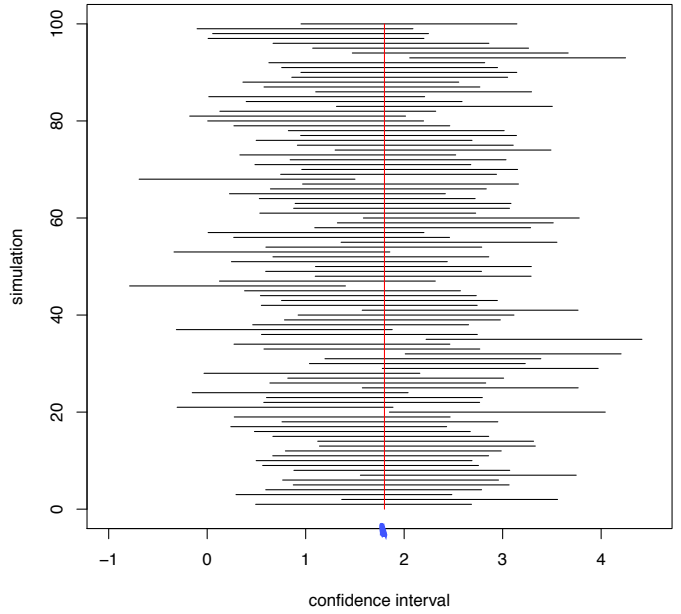
Example: For the sleeping drug example the confidence interval is

$$\frac{\bar{X}_n}{\sigma/\sqrt{n}} \pm \Phi^{-1}(1-\alpha/2)$$

Alternative interpretation of confidence interval: Confidence interval contains true parameter μ with probability $1-\alpha$, i.e.

$$\mathbb{P}_\mu(\mu \in I(X)) = 1-\alpha$$

Confidence interval illustration



General approach: Likelihood ratio test

General approach: Likelihood ratio test

- Model: $X \sim p(x,\theta)$, e.g. $X \sim \text{Binomial}(31'000, \pi)$
- Test: $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_A$, where $\Theta_0 \cap \Theta_A = \emptyset$
  - **Ex:** $H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$ versus $H_A : \pi_{\text{treatment}} \neq \pi_{\text{control}}$

General approach: Likelihood ratio test

- Model: $X \sim p(x, \theta)$, e.g. $X \sim \text{Binomial}(31'000, \pi)$
- Test: $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_A$, where $\Theta_0 \cap \Theta_A = \emptyset$
  - **Ex:** $H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$ versus $H_A : \pi_{\text{treatment}} \neq \pi_{\text{control}}$
- Likelihood ratio: $L(x) = \frac{\max_{\theta \in \Theta_0} p(x;\theta)}{\max_{\theta \in \Theta} p(x;\theta)}$, where $\Theta = \Theta_0 \cup \Theta_A$
  - $p(x;\theta)$ is the probability / density of observing the data x
  - the parameter $\hat{\theta}$ that maximizes $p(x;\theta)$ is called the maximum likelihood estimator (MLE)

General approach: Likelihood ratio test

- Model: $X \sim p(x, \theta)$, e.g. $X \sim \text{Binomial}(31'000, \pi)$
- Test: $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_A$, where $\Theta_0 \cap \Theta_A = \emptyset$
  - **Ex:** $H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$ versus $H_A : \pi_{\text{treatment}} \neq \pi_{\text{control}}$
- Likelihood ratio: $L(x) = \frac{\max_{\theta \in \Theta_0} p(x;\theta)}{\max_{\theta \in \Theta} p(x;\theta)}$, where $\Theta = \Theta_0 \cup \Theta_A$
  - $p(x;\theta)$ is the probability / density of observing the data x
  - the parameter $\hat{\theta}$ that maximizes $p(x;\theta)$ is called the maximum likelihood estimator (MLE)
  - $0 \leq L(x) \leq 1$; $L(x) \ll 1$ if $\theta \in \Theta_A$; $L(x) \approx 1$ if $\theta \in \Theta_0$;

General approach: Likelihood ratio test

- Model: $X \sim p(x, \theta)$, e.g. $X \sim \text{Binomial}(31'000, \pi)$
- Test: $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_A$, where $\Theta_0 \cap \Theta_A = \emptyset$
  - **Ex:** $H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$ versus $H_A : \pi_{\text{treatment}} \neq \pi_{\text{control}}$
- Likelihood ratio: $L(x) = \frac{\max_{\theta \in \Theta_0} p(x;\theta)}{\max_{\theta \in \Theta} p(x;\theta)}$, where $\Theta = \Theta_0 \cup \Theta_A$
  - $p(x;\theta)$ is the probability / density of observing the data x
  - the parameter $\hat{\theta}$ that maximizes $p(x;\theta)$ is called the maximum likelihood estimator (MLE)
  - $0 \leq L(x) \leq 1$; $L(x) \ll 1$ if $\theta \in \Theta_A$; $L(x) \approx 1$ if $\theta \in \Theta_0$;
- Likelihood ratio test: Reject H_0 if $L(x) < \eta$, where η is chosen such that $\mathbb{P}_{H_0}(L(x) \leq \eta) = \alpha$

General approach: Likelihood ratio test

- Model: $X \sim p(x, \theta)$, e.g. $X \sim \text{Binomial}(31'000, \pi)$
- Test: $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_A$, where $\Theta_0 \cap \Theta_A = \emptyset$
  - **Ex:** $H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$ versus $H_A : \pi_{\text{treatment}} \neq \pi_{\text{control}}$
- Likelihood ratio: $L(x) = \frac{\max_{\theta \in \Theta_0} p(x;\theta)}{\max_{\theta \in \Theta} p(x;\theta)}$, where $\Theta = \Theta_0 \cup \Theta_A$
  - $p(x;\theta)$ is the probability / density of observing the data x
  - the parameter $\hat{\theta}$ that maximizes $p(x;\theta)$ is called the maximum likelihood estimator (MLE)
  - $0 \leq L(x) \leq 1$; $L(x) \ll 1$ if $\theta \in \Theta_A$; $L(x) \approx 1$ if $\theta \in \Theta_0$;
- Likelihood ratio test: Reject H_0 if $L(x) < \eta$, where η is chosen such that $\mathbb{P}_{H_0}(L(x) \leq \eta) = \alpha$
  - Neyman-Pearson Lemma: Likelihood ratio test is the most powerful among all level α tests for testing $H_0 : \theta = \theta_0$ versus $H_A : \theta = \theta_A$

Asymptotic likelihood ratio test

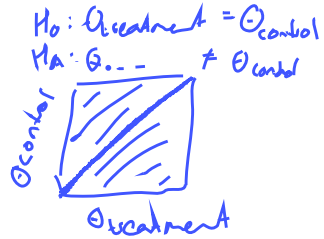
- In general $L(x)$ does not have an easily computable null distribution, i.e., it is difficult to determine η

Asymptotic likelihood ratio test

- In general $L(x)$ does not have an easily computable null distribution, i.e., it is difficult to determine η
- Likelihood ratio statistic: $\Lambda(x) := -2\log(L(x)) = -2\log \frac{\max_{\theta\in\Theta_0} p(x;\theta)}{\max_{\theta\in\Theta} p(x;\theta)}$
  - $0 \le \Lambda(x) < \infty$
  - reject H_0 if $\Lambda(x)$ is too large
- Wilks Theorem: Under H_0,

$$\Lambda(x) \overset{n\to\infty}{\to} \chi^2_d,$$

where $d = \dim(\Theta) - \dim(\Theta_0) > 0$



Asymptotic likelihood ratio test for HIP study

	breast cancer deaths	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

- $H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$ versus $H_A : \pi_{\text{treatment}} \neq \pi_{\text{control}}$
- $\Lambda(x) = -2\log \frac{\max p(x;\,\pi)}{\max p(x;\,\pi_{\text{treatment}},\pi_{\text{control}})}$

Asymptotic likelihood ratio test for HIP study

	breast cancer deaths	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

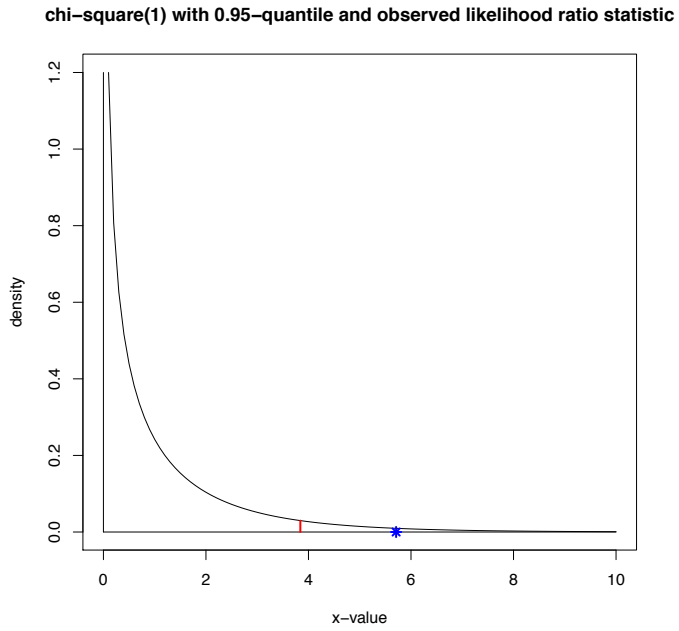
- $H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$ versus $H_A : \pi_{\text{treatment}} \neq \pi_{\text{control}}$
- $\Lambda(x) = -2\log \frac{\max p(x;\,\pi)}{\max p(x;\,\pi_{\text{treatment}},\pi_{\text{control}})}$
- Under H_0 the MLE is $\hat\pi = \frac{102}{62'000}$
- Under H_A the MLEs are $\hat\pi_{\text{treatment}} = \frac{39}{31'000}$ and $\hat\pi_{\text{control}} = \frac{63}{31'000}$
- Then $\Lambda(x) = -2\log \frac{p(x;\,\hat\pi)}{p(x;\,\hat\pi_{\text{treatment}},\hat\pi_{\text{control}})} = \cdots = 5.71$

Asymptotic likelihood ratio test for HIP study

	breast cancer deaths	alive	total
treatment	39 (0.0013)	30'961	31'000
control	63 (0.0020)	30'937	31'000
total	102	61'898	62'000

- $H_0 : \pi_{\text{treatment}} = \pi_{\text{control}}$ versus $H_A : \pi_{\text{treatment}} \neq \pi_{\text{control}}$
- $\Lambda(x) = -2 \log \frac{\max p(x;\, \pi)}{\max p(x;\, \pi_{\text{treatment}}, \pi_{\text{control}})}$
- Under H_0 the MLE is $\hat{\pi} = \frac{102}{62'000}$
- Under H_A the MLEs are $\hat{\pi}_{\text{treatment}} = \frac{39}{31'000}$ and $\hat{\pi}_{\text{control}} = \frac{63}{31'000}$
- Then $\Lambda(x) = -2 \log \frac{p(x;\, \hat{\pi})}{p(x;\, \hat{\pi}_{\text{treatment}}, \hat{\pi}_{\text{control}})} = \cdots = 5.71$
- Under H_0: $\Lambda(x) \overset{n \to \infty}{\to} \chi^2_1$

χ^2-distribution



References

- For a statistics review, including hypothesis testing (chapter 26-29):

  D. Freedman, R. Pisani, R. Purves. *Statistics*. 2007.

- For how to perform hypothesis testing in R (chapter 4):

  P. Dalgaard. *Introductory Statistics with R*. 2002.

MITx:
Statistics, Computation & Applications

Statistics Refresher

Lecture 3: Multiple Hypothesis Testing

Some quotes and research findings

Giovannucci et al., Journal of the National Cancer Institute 87 (1995):

Intake of tomato sauce (p-value of 0.001), tomatoes (p-value of 0.03), and pizza (p-value of 0.05) reduce the risk of prostate cancer;

But for example tomato juice (p-value of 0.67), or cooked spinach (p-value of 0.51), and many other vegetables are not significant.

Some quotes and research findings

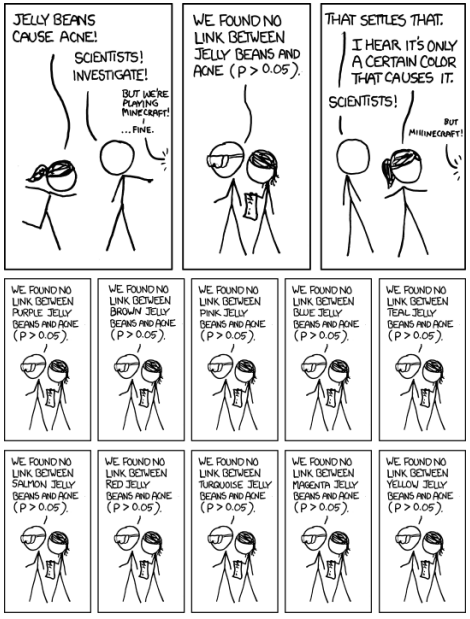
Giovannucci et al., Journal of the National Cancer Institute 87 (1995):

Intake of tomato sauce (p-value of 0.001), tomatoes (p-value of 0.03), and pizza (p-value of 0.05) reduce the risk of prostate cancer;

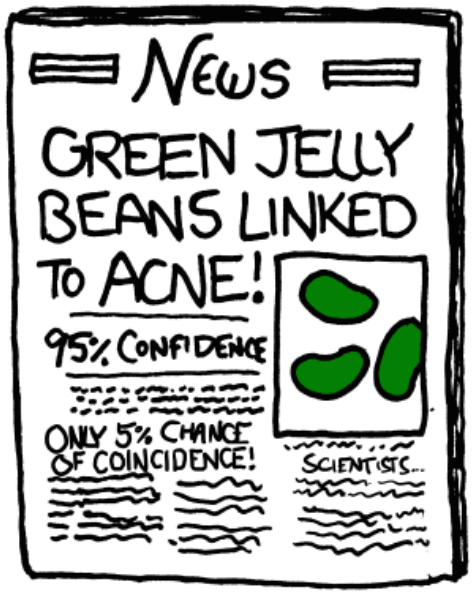
But for example tomato juice (p-value of 0.67), or cooked spinach (p-value of 0.51), and many other vegetables are not significant.

"Orange cars are less likely to have serious damages that are discovered only after the purchase."

Jelly Beans and Acne



Problematic of selective inference



http://imgs.xkcd.com/comics/significant.png

Wonder-syrup

- randomized group of 1000 people
- measure 100 variables before and after taking the syrup: weight, blood pressure, etc.
- perform a paired t-test with a significance level of 5%

Wonder-syrup

- randomized group of 1000 people
- measure 100 variables before and after taking the syrup: weight, blood pressure, etc.
- perform a paired t-test with a significance level of 5%
- $V :=$ # false significant tests: $V \sim \text{Binomial}(100, 0.05)$

$\Rightarrow$ in average 5 out of 100 variables show a significant effect!

Wonder-syrup

- randomized group of 1000 people
- measure 100 variables before and after taking the syrup: weight, blood pressure, etc.
- perform a paired t-test with a significance level of 5%
- $V :=$ # false significant tests: $V \sim \text{Binomial}(100, 0.05)$

$\Rightarrow$ in average 5 out of 100 variables show a significant effect!

Different protection levels

Compute p-values using methods that control:

- family-wise error rate (FWER) $\leq \alpha$, where

$$\text{FWER} = \mathbb{P}(\text{at least one false significant result})$$

- false discovery rate (FDR) $\leq \alpha$, where

$$\text{FDR} = \text{expected fraction of false significant results among all significant results}$$

declared non-sig / significant

H_0 true	a	b	m_0
H_A true	c	d	m_1
	n_0	n_1	m

Corrections for multiple testing

Bonferroni correction:

- Reject H_0 when: $m \cdot p\text{-value} \leq \alpha$
  where m is the total number of hypothesis tests performed
- Bonferroni correction implies $\text{FWER} \leq \alpha$

$$\mathbb{P}(V \geq 1) = \mathbb{P}(V=1) + \mathbb{P}(V=2) + \dots + \mathbb{P}(V=m_0)$$
$$\leq 0 \cdot \mathbb{P}(V=0) + 1\,\mathbb{P}(V=1) + 2\,\mathbb{P}(V=2) + \dots + m_0\,\mathbb{P}(V=m_0)$$
$$= \mathbb{E}[V] \qquad V \sim \text{Binomial}(m_0, \tfrac{\alpha}{m})$$
$$= m_0 \cdot \frac{\alpha}{m}$$
$$\leq \alpha$$

$$\mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V=0) = 1 - (1-\alpha_{ind})^{m_0} \leq 1 - (1-\alpha_{ind})^{m} \leq \alpha$$
$$\Rightarrow \alpha_{ind} = 1 - (1-\alpha)^{1/m}$$

Corrections for multiple testing

Bonferroni correction:

- Reject H_0 when: $m \cdot p\text{-value} \leq \alpha$
  where m is the total number of hypothesis tests performed
- Bonferroni correction implies $\text{FWER} \leq \alpha$

Holm-Bonferroni correction:

- Sort p-values in increasing order: $p_{(1)} \leq \cdots \leq p_{(m)}$
- Reject H_0 when: $(m-i+1)p_{(i)} \leq \alpha$ (more power than Bonferroni)
- Holm-Bonferroni correction implies $\text{FWER} \leq \alpha$

$mp_{(1)} \quad (m-1)p_{(2)} \quad (m-2)p_{(3)} \quad \dots \quad p_{(m)}$

Corrections for multiple testing

Bonferroni correction:

- Reject H_0 when: $m \cdot p\text{-value} \leq \alpha$
  where m is the total number of hypothesis tests performed
- Bonferroni correction implies $\text{FWER} \leq \alpha$

Holm-Bonferroni correction:

- Sort p-values in increasing order: $p_{(1)} \leq \cdots \leq p_{(m)}$
- Reject H_0 when: $(m-i+1)p_{(i)} \leq \alpha$ (more power than Bonferroni)
- Holm-Bonferroni correction implies $\text{FWER} \leq \alpha$

Benjamini-Hochberg correction:

- Sort p-values in increasing order: $p_{(1)} \leq \cdots \leq p_{(m)}$
- Reject H_0 when: $mp_{(i)}/i \leq \alpha$
- Benjamini-Hochberg correction implies $\text{FDR} \leq \alpha$

Commonly accepted practice

- No correction for multiple testing when generating hypotheses (but report number of tests performed)
- $\mathrm{FDR} \leq 10\%$ in exploratory analysis or screening
  - balance between high power and low # of false significant results
- $\mathrm{FWER} \leq 5\%$ in confirmatory analysis
  - food and drug administration (FDA)

References

- Lecture by Yoav Benjamini, THE expert for multiple testing issues:

  `http://simons.berkeley.edu/talks/yoav-benjamini-2013-12-11a`

Data Analysis: Statistical Modeling and Computation in Applications

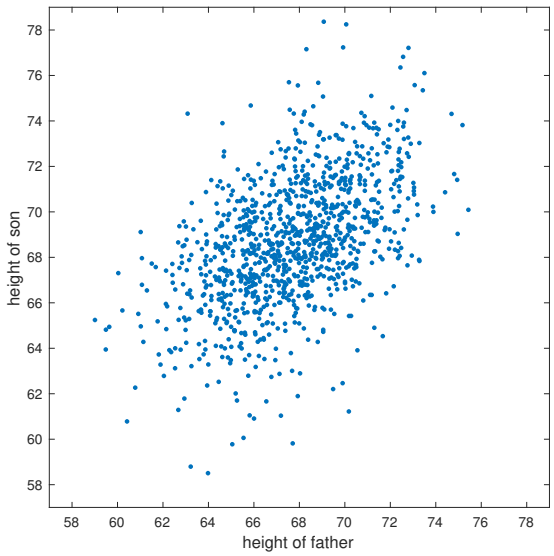
Correlation and Least Squares Regression

Outline

- Correlation
- Regression line
- Evaluation
- Multiple regression
- Computing the estimator
- Variable selection and regularization

Scatter diagram: height of 1078 fathers and their sons

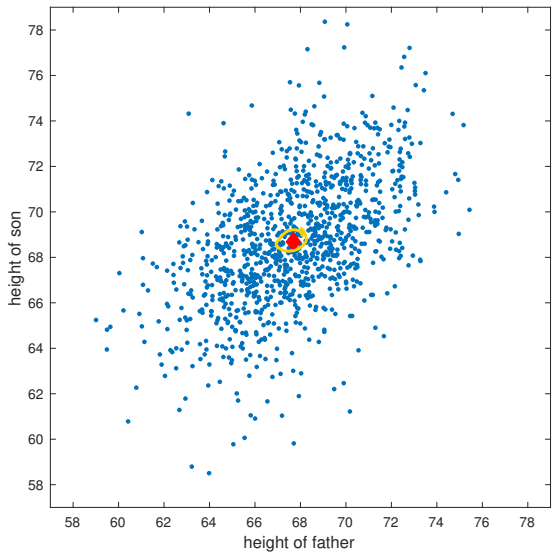
Is there an association?
What kind?



Data: Pearson K and Lee A. (1903). On the laws of inheritance in man. Biometrika, 2:357-462.
Downloaded from `https://myweb.uiowa.edu/pbreheny/data/pearson.html`

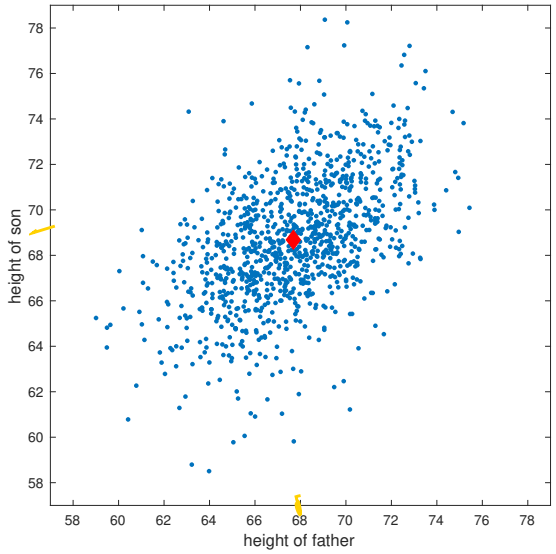
Summarizing the Plot

- average $\bar{x}$, $\bar{y}$



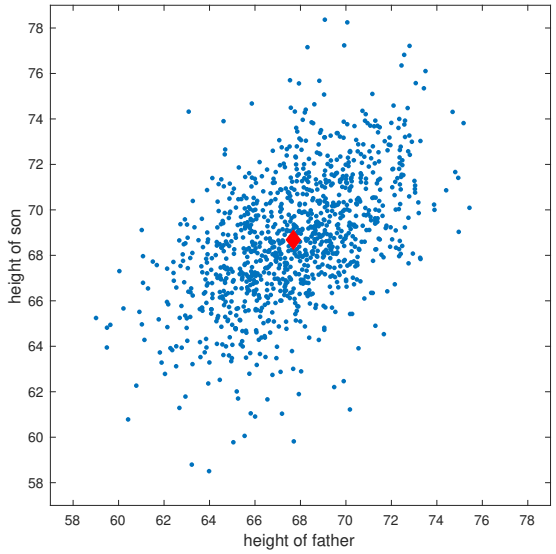
Summarizing the Plot

- average $\bar{x}$, $\bar{y}$
  fathers: $\bar{x} \approx 68$,
  sons: $\bar{y} \approx 69$



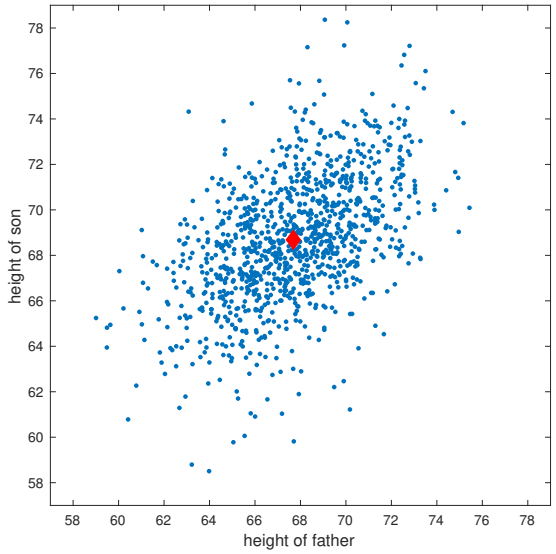
Summarizing the Plot

- average $\bar{x}$, $\bar{y}$
  fathers: $\bar{x} \approx 68$,
  sons: $\bar{y} \approx 69$
- standard deviation
  $s_x = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
  here: $s_x \approx s_y \approx 2.7$



Summarizing the Plot

- average $\bar{x}$, $\bar{y}$
  fathers: $\bar{x} \approx 68$,
  sons: $\bar{y} \approx 69$
- standard deviation
  $s_x = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
  here: $s_x \approx s_y \approx 2.7$
- correlation coefficient
  $r \approx 0.5$



Correlation Coefficient

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

(convert to standard units and take average product)

Correlation Coefficient

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\mathrm{cov}(x,y)}{s_x s_y}$$

(convert to standard units and take average product)

1. symmetric

Correlation Coefficient

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\mathrm{cov}(x,y)}{s_x s_y}$$

(convert to standard units and take average product)

1. symmetric
2. Why standard units?

Correlation Coefficient

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\mathrm{cov}(x,y)}{s_x s_y}$$

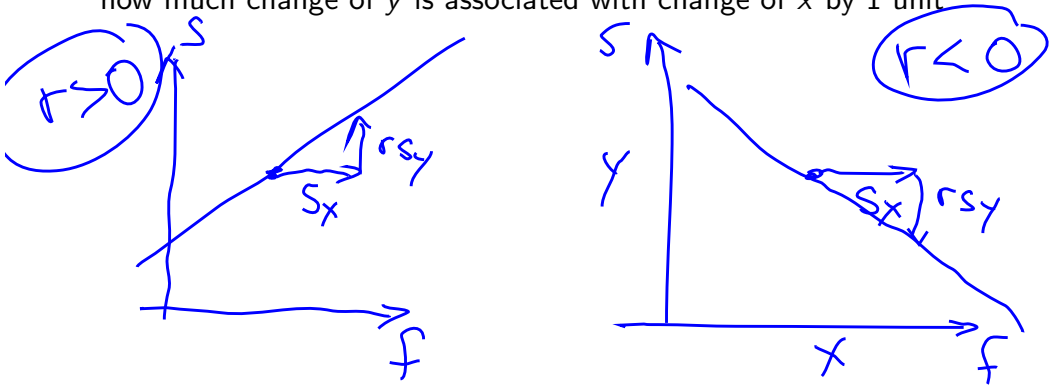
(convert to standard units and take average product)

1. symmetric
2. Why standard units?
   adding or multiplying constants to all x_i or y_i does not change r
3. What does $r \approx 0.5$ mean?

What does the Correlation coefficient mean? (1)

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

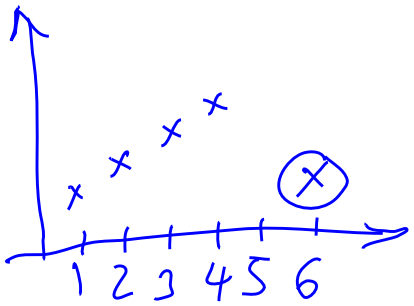
- measures *linear* association between variables:
  how much change of y is associated with change of x by 1 unit



What does the Correlation coefficient mean? (1)

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

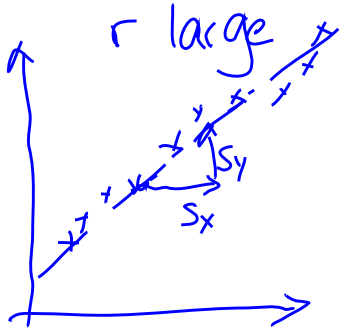
- measures *linear* association between variables:
  how much change of y is associated with change of x by 1 unit

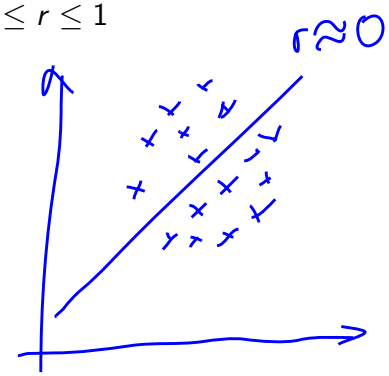


What does the Correlation coefficient mean? (2)

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- measures *clusteredness* along a line: $-1 \le r \le 1$
  sign?





Examples

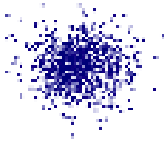
1.

3.

5.

2.

4.

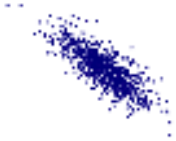


6.

Examples

1. $r = 1$

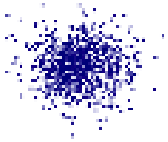
3.



5.

2.

4.



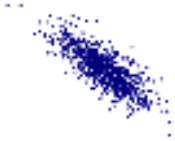
6.

Examples

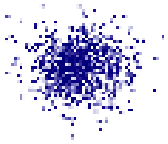
1. $r = 1$

2. $r = -1$

3.



4.



5.

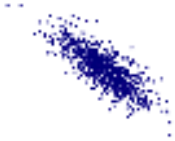
6.

Examples

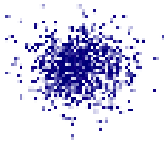
1. $r = 1$

2. $r = -1$

3.



4. $r = 0$



5.

6.

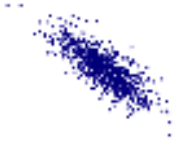
Stefanie Jegelka (and Caroline Uhler) 9 / 9

Examples

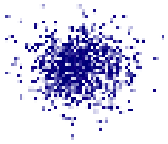
1. $r = 1$

2. $r = -1$

3. $r = -0.8$



4. $r = 0$



5.

6.

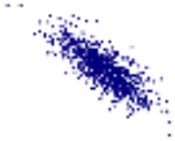
Stefanie Jegelka (and Caroline Uhler) 9 / 9

Examples

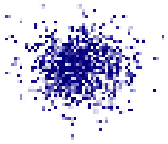
1. $r = 1$

2. $r = -1$

3. $r = -0.8$



4. $r = 0$



5. $r = 0$

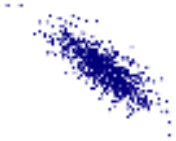
6.

Stefanie Jegelka (and Caroline Uhler) 9 / 9

Examples

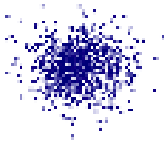
1. $r = 1$

3. $r = -0.8$



5. $r = 0$

2. $r = -1$



4. $r = 0$

6. $r = 0$

Careful with nonlinearities and outliers!

Correlation coefficient: summary

$$r = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- measures *linear* association between variables:
- measures clusteredness along a line
- symmetric (swapping x and y)
- between -1 and 1, and invariant to
  - adding a constant to all x_i or all y_i
  - multiplying to all x_i (all y_i) by a positive constant

Data Analysis: Statistical Modeling and Computation in Applications

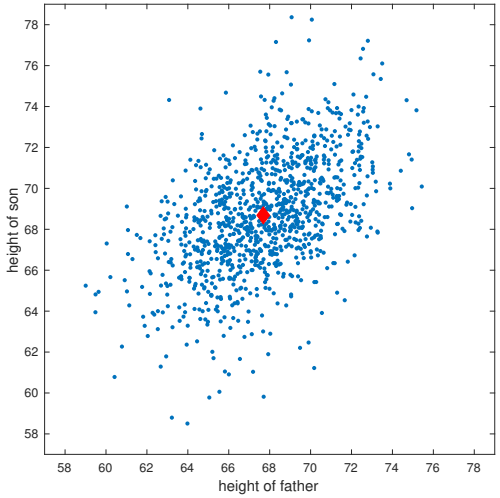
Correlation and Least Squares Regression
Part 2

Outline

Predicting a son's height from the father's height

- fathers: $\bar{x} \approx 68$in, $s_x = 2.7$in
- sons: $\bar{y} \approx 69$in, $s_y = 2.7$in
- $r \approx 0.5$

Suggestion: The sons' average is 1 inch more than the fathers' average. So, if the father's height is 64 inches we expect the son's height to be 65 inches.

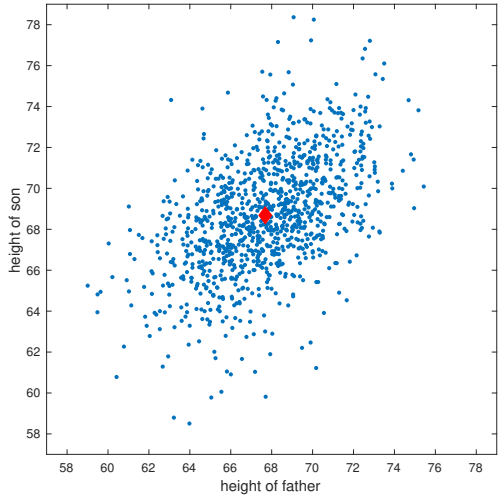


Predicting a son's height from the father's height

- fathers: $\bar{x} \approx 68$in, $s_x = 2.7$in
- sons: $\bar{y} \approx 69$in, $s_y = 2.7$in
- $r \approx 0.5$

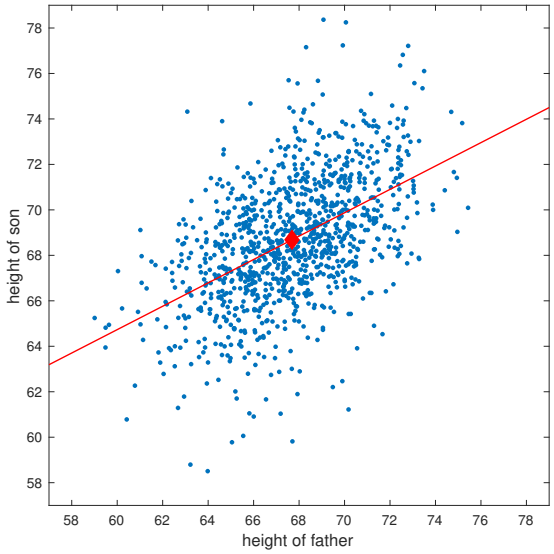
Suggestion: The sons' average is 1 inch more than the fathers' average. So, if the father's height is 64 inches we expect the son's height to be 65 inches.

No! Correlation Coefficient...



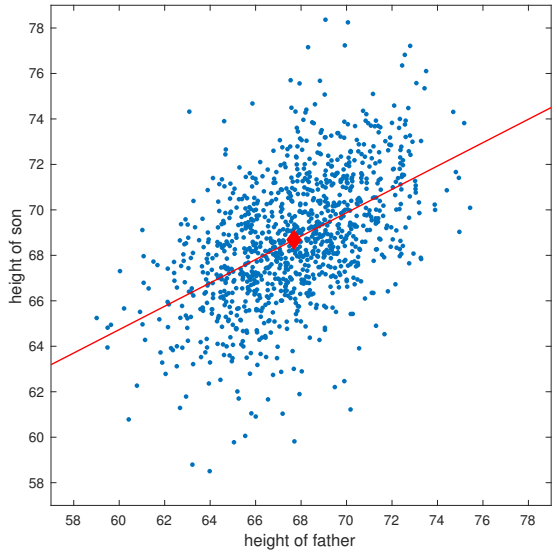
Stefanie Jegelka (and Caroline Uhler) 3 / 10

Regression line: what does it mean?



Regression line: what does it mean?

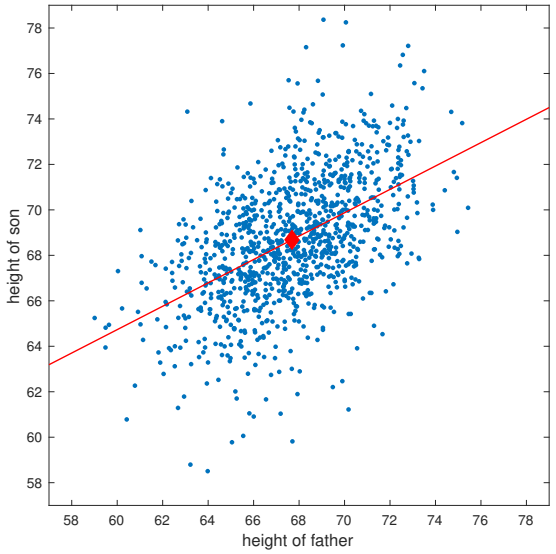
1. Increase of 1 std dev in x associated with increase of r std dev in y.



Stefanie Jegelka (and Caroline Uhler) 4 / 10

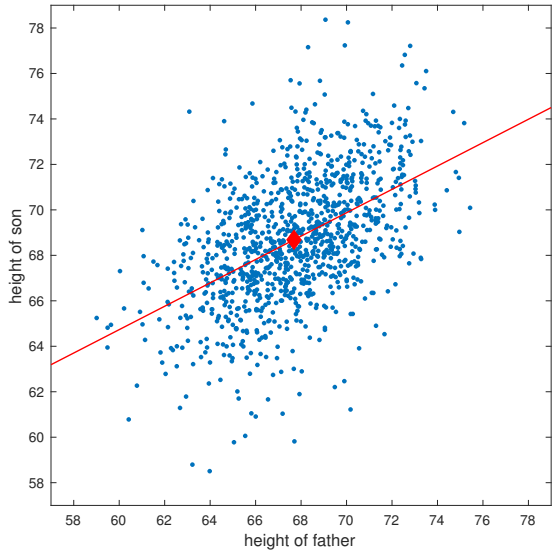
Regression line: what does it mean?

1. Increase of 1 std dev in x associated with increase of r std dev in y.
2. Interpolating conditional averages of y given x



Regression line: what does it mean?

1. Increase of 1 std dev in x associated with increase of r std dev in y.
2. Interpolating conditional averages of y given x
3. Solution to least squares



Stefanie Jegelka (and Caroline Uhler) 4 / 10

Regression Line for y on x

model:

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

- fit to *minimize RMS error* (Gauss)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\beta_0 + \beta_1 x_i - y_i)^2}$$

Regression Line for y on x

model:

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0 \quad \text{with } \hat{\beta}_1 = r\frac{s_y}{s_x}$$

- fit to *minimize RMS error* (Gauss)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\beta_0 + \beta_1 x_i - y_i)^2}$$

- RMS error is $\sqrt{1-r^2}s_y$

Stefanie Jegelka (and Caroline Uhler) 5 / 10

Regression Line for y on x

model:

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0 \quad \text{with } \hat{\beta}_1 = r\frac{s_y}{s_x}$$

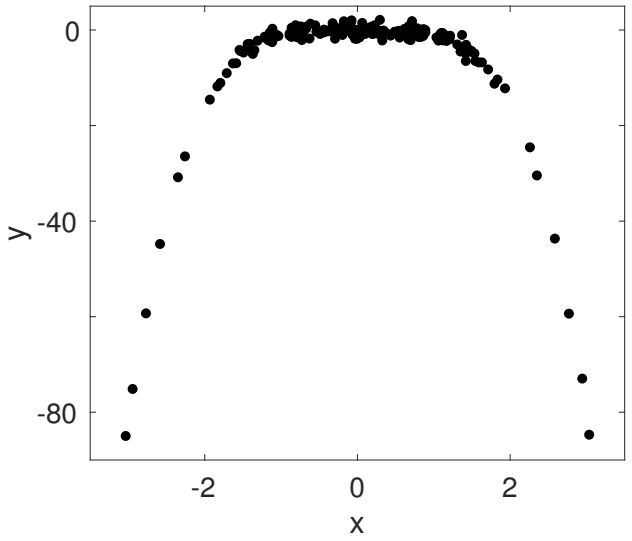
- fit to *minimize RMS error* (Gauss)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\beta_0 + \beta_1 x_i - y_i)^2}$$

- RMS error is $\sqrt{1-r^2}s_y$
- not the same as the regression line of x on y

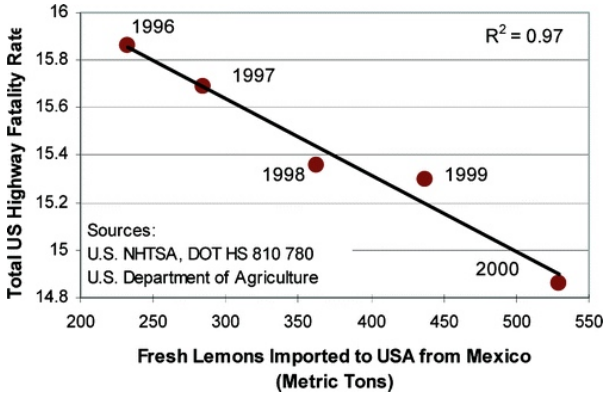
3 words of caution (1)

- Only measures a linear relationship.



3 words of caution (2)

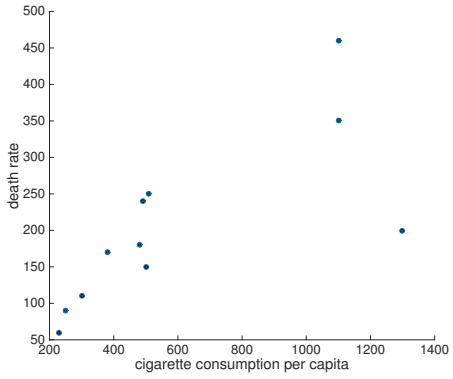
- Correlation is not equal to causation.



3 words of caution (3)

Country	*Cigarette consumption*	*Deaths per million*
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1,100	350
Great Britain	1,100	460
Iceland	230	60
Netherlands	490	240
Norway	250	90
Sweden	300	110
Switzerland	510	250
U.S.	1,300	200

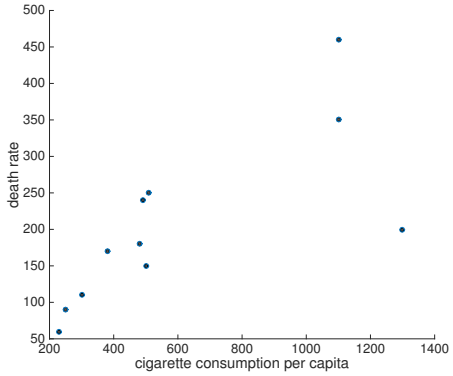
$r \approx 0.74$



(Source: Freedman, Pisani, Purves. Statistics)

3 words of caution (3)

Country	Cigarette consumption	Deaths per million
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1,100	350
Great Britain	1,100	460
Iceland	230	60
Netherlands	490	240
Norway	250	90
Sweden	300	110
Switzerland	510	250
U.S.	1,300	200



$r \approx 0.74$

Ecological correlations tend to overstate the strength of an association for individuals.

(Source: Freedman, Pisani, Purves. Statistics)

Ecological Correlation

Summary: Regression line

- interpolates conditional averages of y given x
- solves least squares problem
- slope: rs_y/s_x
- caution: linear relationship, and not impling causality
- caution: ecological correlations

Data Analysis: Statistical Modeling and Computation in Applications

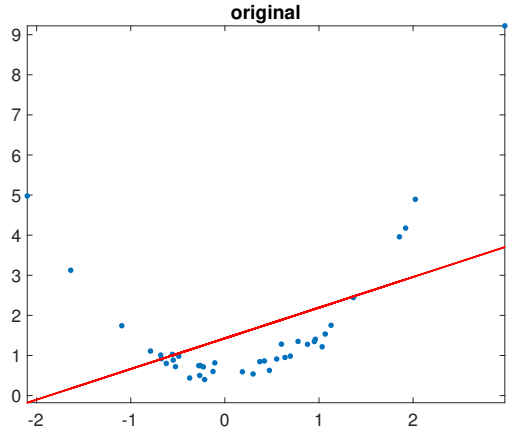
Correlation and Least Squares Regression
Part 3

Outline

- Correlation
- Regression line
- Evaluation
- Multiple regression
- Computing the estimator
- Variable selection and regularization

How do we evaluate our regression line?

- We fit a model. Does it make sense?



Does the model make sense?

Assumptions:

- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Does the model make sense?

Assumptions:

- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Genneral idea: Plot the residuals $e_i = y_i - \hat{y}_i$:

Does the model make sense?

Assumptions:

- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Genneral idea: Plot the residuals $e_i = y_i - \hat{y}_i$:

- should show no pattern (e.g. due to nonlinear association)
- points regularly scattered around 0

Does the model make sense?

Assumptions:

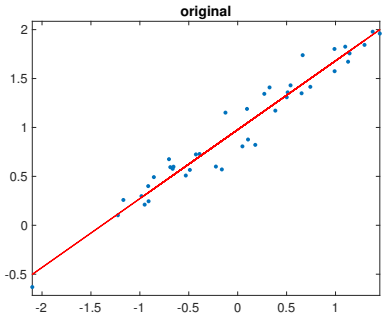
- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Genneral idea: Plot the residuals $e_i = y_i - \hat{y}_i$:

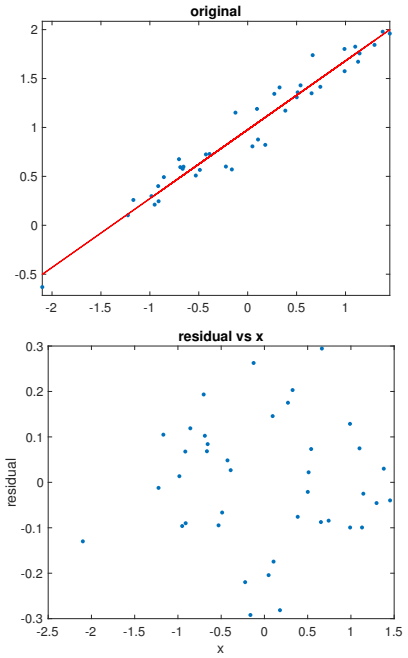
- should show no pattern (e.g. due to nonlinear association)
- points regularly scattered around 0

Variable transformations can help, e.g. $\log(y)$, $\sqrt{y}$, $\sqrt{x}$, $\log(x)$, x^2

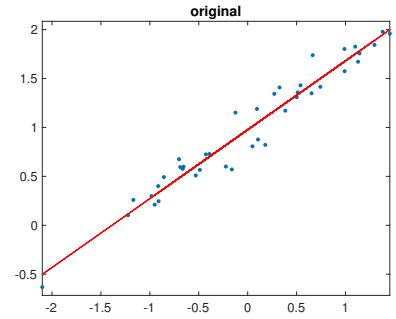
Example 1: assumptions hold

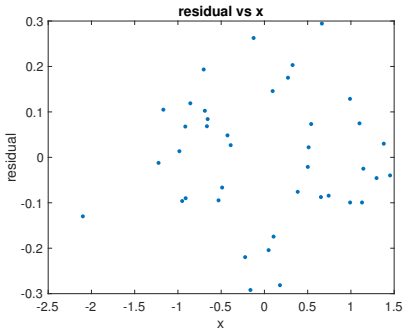


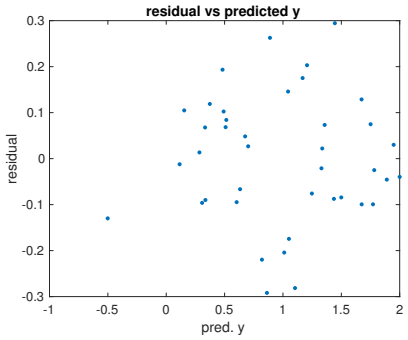
Example 1: assumptions hold



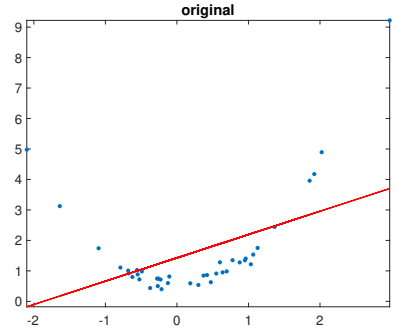
Example 1: assumptions hold



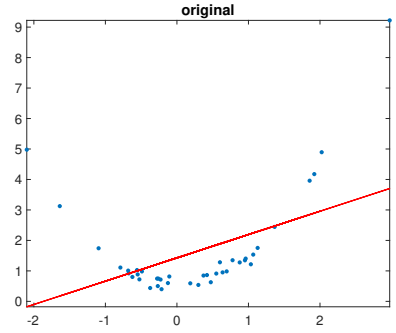


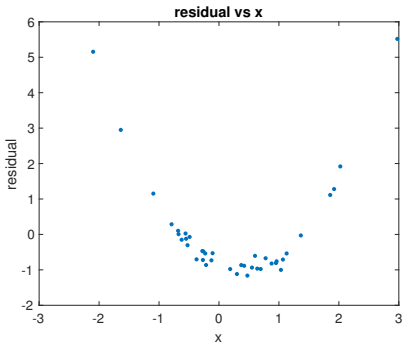


Example 2

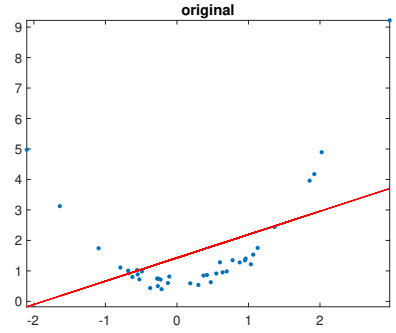


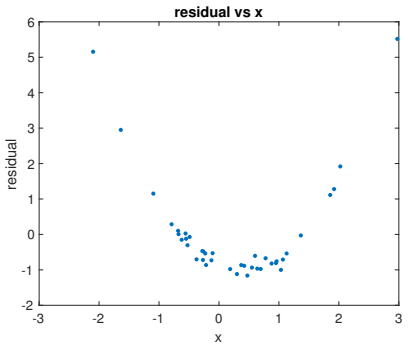
Example 2

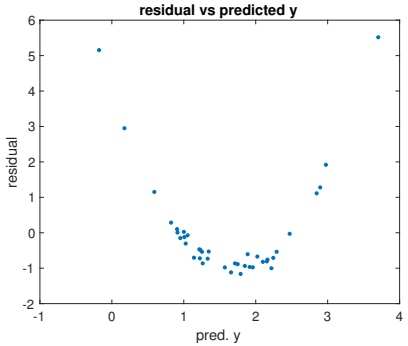




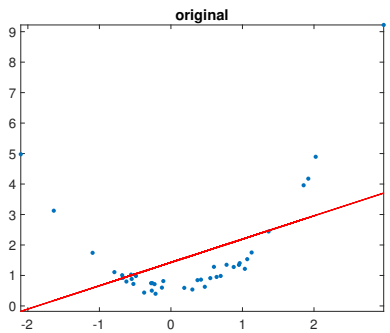
Example 2

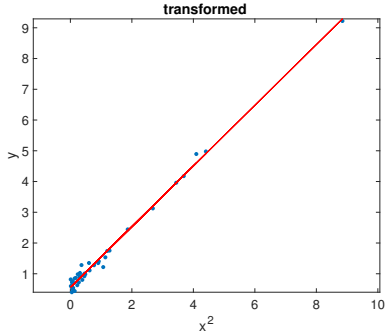




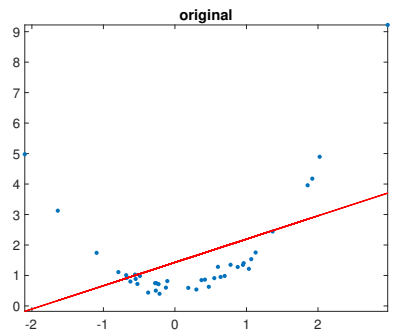


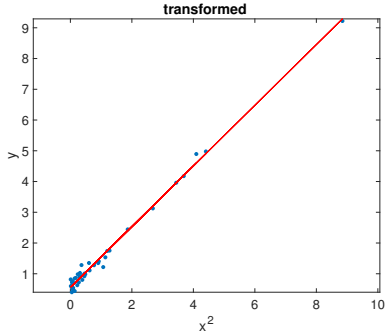
Example 2: transformation x^2

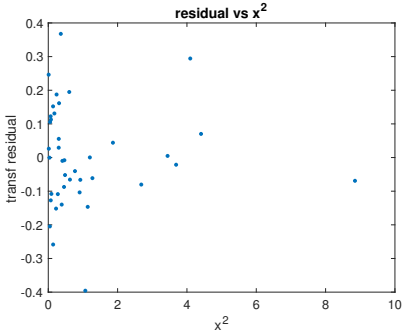




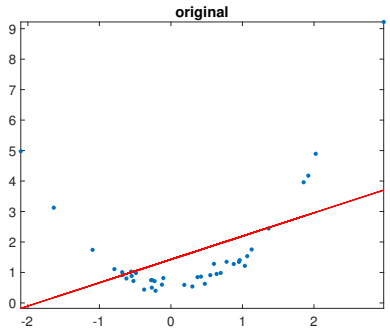
Example 2: transformation x^2

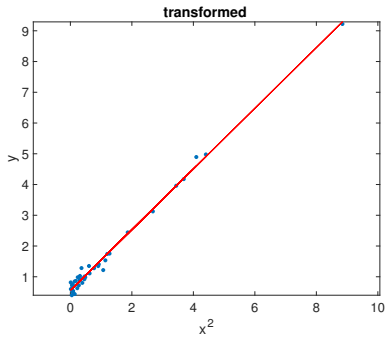


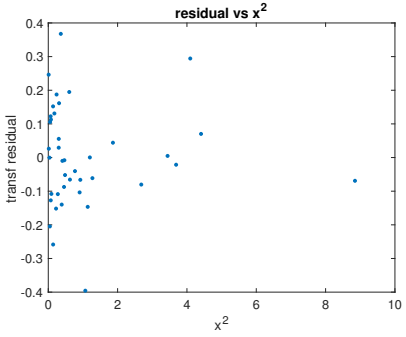


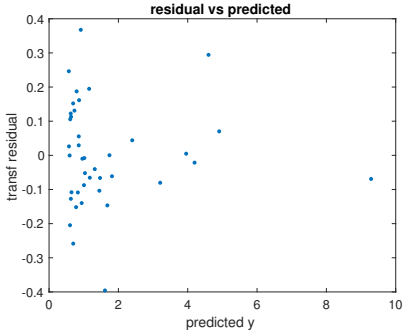


Example 2: transformation x^2

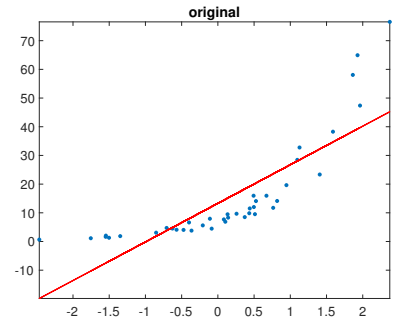




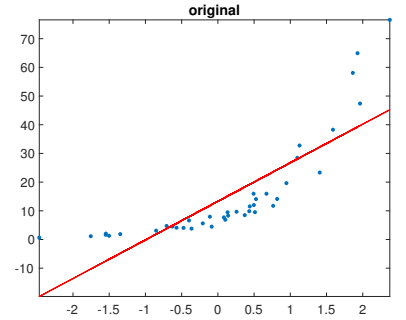


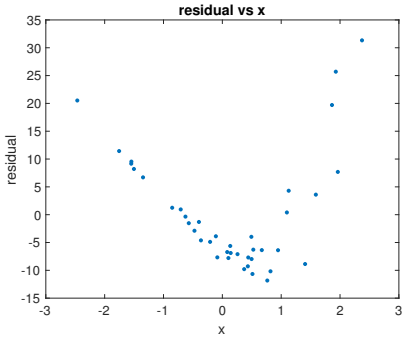


Example 3

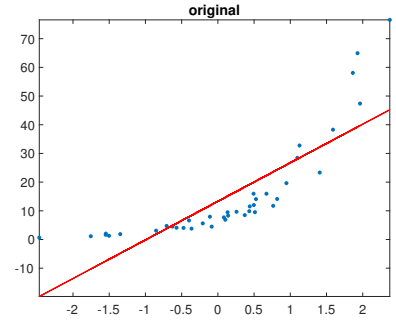


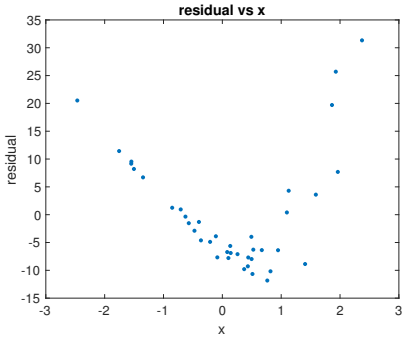
Example 3

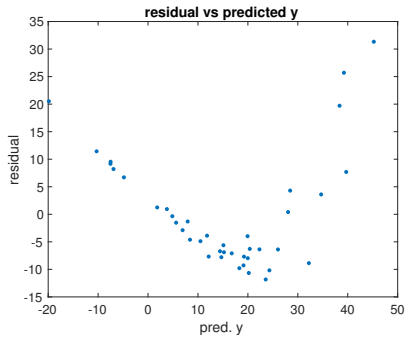




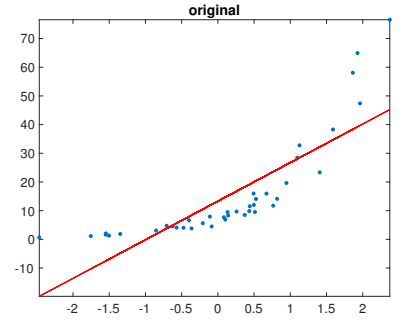
Example 3

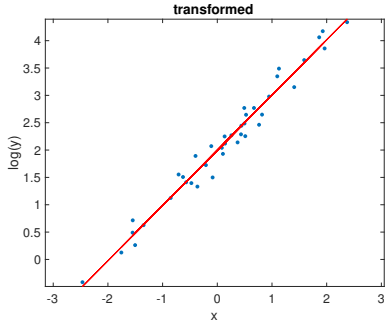




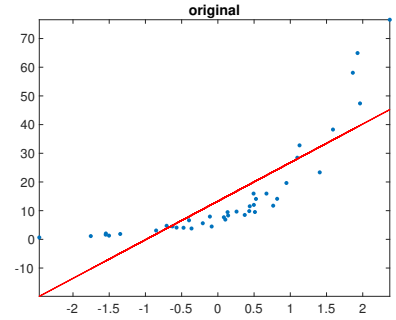


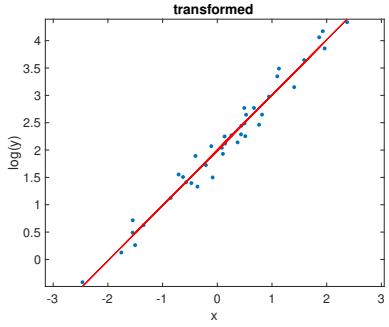
Example 3: transformation $\log(y)$

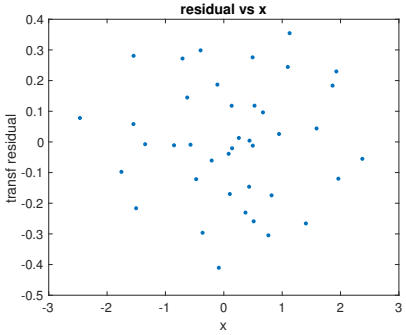




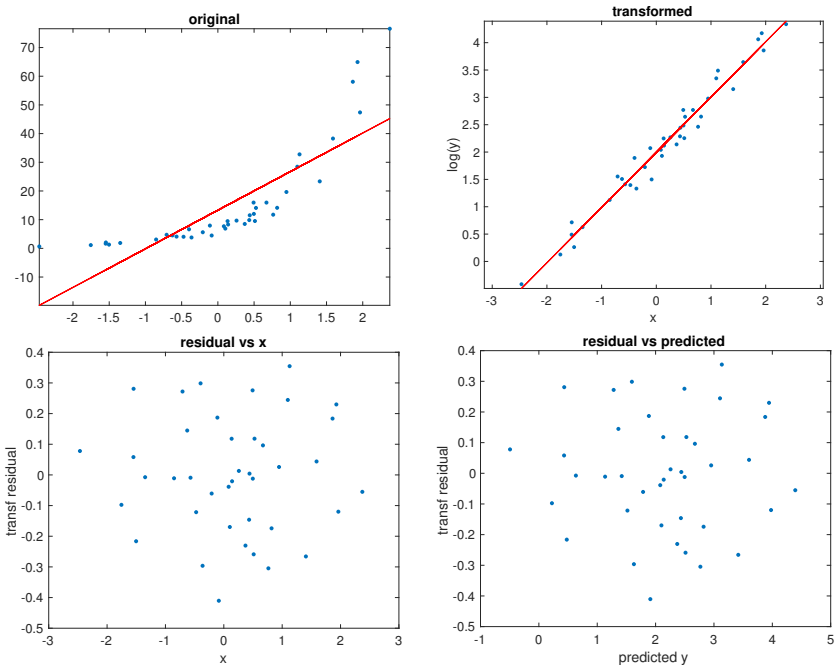
Example 3: transformation $\log(y)$







Example 3: transformation $\log(y)$



Does the model make sense?

Assumptions:

- *linear* relationship $Y = \beta_1 X + \beta_0 + \epsilon$
- errors ϵ_i, ϵ_j are mean zero, independent, and Gaussian

Genneral idea: Plot the residuals $e_i = y_i - \hat{y}_i$:

- should show no pattern (e.g. due to nonlinear association)
- points regularly scattered around 0

Variable transformations can help, e.g. $\log(y)$, $\sqrt{y}$, $\sqrt{x}$, $\log(x)$, x^2

Data Analysis: Statistical Modeling and Computation in Applications

Correlation and Least Squares Regression
Part 4

Outline

Multiple regression

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

y_i x_{i1} x_{i2}

Multiple regression

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

Multiple regression

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

$41 = \beta_0 + 190\beta_1 + 67\beta_2 + \epsilon_1$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

y X β ϵ

Multiple regression

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

$$N \begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

y

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$
  - $\mathbf{X}$ design matrix: $N \times p$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$
  - $\mathbf{X}$ design matrix: $N \times p$
  - $\boldsymbol{\beta}$ parameters: $p \times 1$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$
  - $\mathbf{X}$ design matrix: $N \times p$
  - $\boldsymbol{\beta}$ parameters: $p \times 1$
  - ϵ: random error / disturbances
    ϵ_i are iid, $\mathbb{E}[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

Examples of multiple regression

- **Simple linear regression**:

$$p = 2,\ X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ y_i = \beta_0 + \beta_1 x_1$$

Examples of multiple regression

- **Simple linear regression**:

$$p = 2,\ X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ \ y_i = \beta_0 + \beta_1 x_1$$

- **Quadratic (polynomial) regression**:

$$p = 3,\ X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & X_N^2 \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix},\ \ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2$$

Examples of multiple regression

- **Effect on groups**. Consider an example where we have data obtained on different days. The effect of the days can be modeled as

$$y_i = \underbrace{\beta_0}_{\text{day 1}} + \underbrace{\beta_1}_{\text{day 2}} + \underbrace{\beta_2}_{\text{day 3}} + \epsilon_i$$

Examples of multiple regression

- **Effect on groups**. Consider an example where we have data obtained on different days. The effect of the days can be modeled as

$$y_i = \underbrace{\beta_0}_{\text{day 1}} + \underbrace{\beta_1}_{\text{day 2}} + \underbrace{\beta_2}_{\text{day 3}} + \epsilon_i$$

$$p = 3,\quad X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Multiple regression

ozone	radiation	temp
41	190	67
36	118	72
12	149	74
18	313	62

- Model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- vector form: $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
- Matrix-vector form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$
  - $\mathbf{y}$ dependent / response variable: $N \times 1$
  - $\mathbf{X}$ design matrix: $N \times p$
  - $\boldsymbol{\beta}$ parameters: $p \times 1$
  - ϵ: random error / disturbances
    ϵ_i are iid, $\mathbb{E}[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2$

$$\begin{pmatrix} 41 \\ 36 \\ 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 1 & 190 & 67 \\ 1 & 118 & 72 \\ 1 & 149 & 74 \\ 1 & 313 & 62 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

$$y = X \beta + \epsilon$$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- setting derivative to zero gives *normal equations*

$$\mathbf{X}^\top\mathbf{X}\hat{\beta} = \mathbf{X}^\top\mathbf{y}$$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- setting derivative to zero gives *normal equations*

$$\mathbf{X}^\top\mathbf{X}\hat{\beta} = \mathbf{X}^\top\mathbf{y}$$

- if $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- setting derivative to zero gives *normal equations*

$$\mathbf{X}^\top\mathbf{X}\hat{\beta} = \mathbf{X}^\top\mathbf{y}$$

- if $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
- fitted values: $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}\mathbf{y}$

"hat matrix"

Deriving the normal equations

- least squares objective:

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Deriving the normal equations

- least squares objective:

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- set gradient to zero. *Gradient* is the vector of partial derivatives:

Deriving the normal equations

- least squares objective:

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- set gradient to zero. *Gradient* is the vector of partial derivatives:

$$\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial f}{\partial \beta_0} \\ \frac{\partial f}{\partial \beta_1} \\ \vdots \\ \frac{\partial f}{\partial \beta_{p-1}} \end{pmatrix} \overset{!}{=} 0$$

If $\boldsymbol{\beta}$ is $p \times 1$, then $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$ is $p \times 1$.

Partial derivative

- example: 1 data point, $p = 2$:

$$f(\boldsymbol{\beta}) = (y_1 - x_{11}\beta_1 - \beta_0)^2$$

Partial derivative

- example: 1 data point, $p = 2$:

$$f(\boldsymbol{\beta}) = (y_1 - x_{11}\beta_1 - \beta_0)^2$$

- derivative:

$$\frac{\partial f}{\partial \beta_1} = -2x_{11}(y_1 - x_{11}\beta_1 - \beta_0)$$

Partial derivative

- example: 1 data point, $p = 2$:

$$f(\boldsymbol{\beta}) = (y_1 - x_{11}\beta_1 - \beta_0)^2$$

- derivative:

$$\frac{\partial f}{\partial \beta_1} = -2x_{11}(y_1 - x_{11}\beta_1 - \beta_0)$$

- similarly:

$$\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Data Analysis: Statistical Modeling and Computation in Applications

Correlation and Least Squares Regression
Part 5

Outline

- Correlation
- Regression line
- Evaluation
- Multiple regression
- Computing the estimator
- Variable selection and regularization

Ordinary Least Squares estimator (OLS)

- model: $y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$
- fitted values: $\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2$
  or $\hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$
- least squares:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\beta)^2 = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2$$

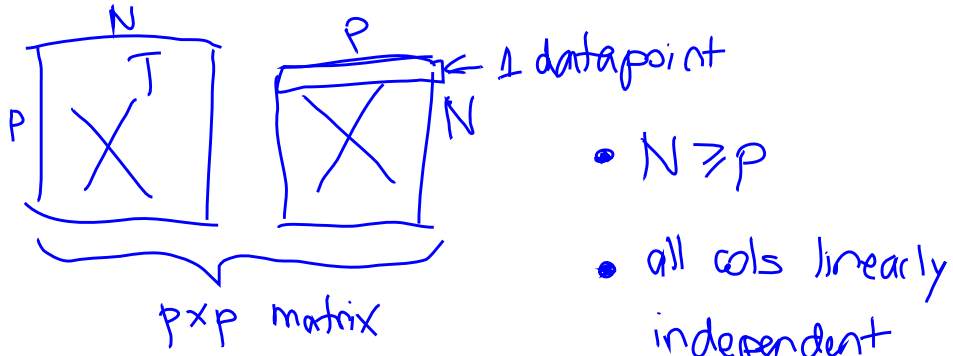
- setting derivative to zero gives *normal equations*

$$\mathbf{X}^\top\mathbf{X}\hat{\beta} = \mathbf{X}^\top\mathbf{y}$$

- if $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$

When is $\mathbf{X}^\top\mathbf{X}$ invertible?

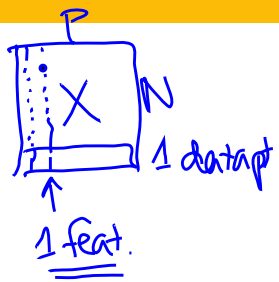
- if $\mathbf{X}^\top\mathbf{X}$ has *full rank*:



- $N \geq p$
- all cols linearly independent

When is $\mathbf{X}^\top\mathbf{X}$ invertible?

- if $\mathbf{X}^\top\mathbf{X}$ has *full rank*:
- $N \geq p$



$\beta_0 + 2\beta_1 = 5$

$N=1$
$p=2$

β_1 β_2

When is $\mathbf{X}^\top \mathbf{X}$ invertible?

- if $\mathbf{X}^\top \mathbf{X}$ has *full rank*:
- $N \geq p$
- all columns of $\mathbf{X}$ linearly independent

If $p > N \ldots$

Regularize!

If $p > N \ldots$

Regularize!

- ℓ_2 **penalty**: minimize

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \underbrace{\|\boldsymbol{\beta}\|_2^2}_{\sum_{j=0}^{p-1} \beta_j^2}$$

penalizes large values of β_j
always unique $\hat{\boldsymbol{\beta}}$.

If $p > N \ldots$

Regularize!

- ℓ_2 **penalty**: minimize

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \underbrace{\|\boldsymbol{\beta}\|_2^2}_{\sum_{j=0}^{p-1} \beta_j^2}$$

penalizes large values of β_j
always unique $\hat{\boldsymbol{\beta}}$.

- ℓ_1 **penalty (Lasso)**: minimize

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \underbrace{\|\boldsymbol{\beta}\|_1}_{\sum_{j=0}^{p-1} |\beta_j|}$$

prefers *sparse* β (few nonzero coordinates)

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!

- Recall: model and estimator:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = \sigma^2 \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!
- Recall: model and estimator:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = \sigma^2 \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

- **OLS is (conditionally) unbiased**: $\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$.

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!
- Recall: model and estimator:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = \sigma^2 \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

- **OLS is (conditionally) unbiased**: $\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$.
- **Gaussianity:** If $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, model correct and $\mathbf{X}$ fixed, then $\hat{\boldsymbol{\beta}}$ is normal: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$

Model selection: Which variables to include in the model?

$\beta_j = 0$ would mean I exclude variable j from the prediction.

- **Idea:** β_j is a random variable. Do a t-test!
- Recall: model and estimator:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = \sigma^2 \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

- **OLS is (conditionally) unbiased**: $\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$.
- **Gaussianity:** If $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, model correct and $\mathbf{X}$ fixed, then $\hat{\boldsymbol{\beta}}$ is normal: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$
- *t-test to test $\beta_j = 0$ vs. $\beta_j \neq 0$*: estimate σ^2 as $\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i-1}^{N}(y_i - \hat{y}_i)^2$, then $(N-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p}$.

Backward Model Selection

Which variables should I include in my model?
$\beta_j = 0$ would mean I exclude variable j from the prediction.

Backward Model Selection

Which variables should I include in my model?
$\beta_j = 0$ would mean I exclude variable j from the prediction.

- Fit a model that uses all variables:
  $y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_{p-1} x_{i,p-1} + \epsilon_i$

Backward Model Selection

Which variables should I include in my model?
$\beta_j = 0$ would mean I exclude variable j from the prediction.

- Fit a model that uses all variables:
  $y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_{p-1} x_{i,p-1} + \epsilon_i$
- Use the t-test to determine variables that are not significant. Of those, remove the one with the largest p-value. Re-fit and repeat until all variables have significant p-values.

References

- D. Freedman, R. Pisani, R. Purves. *Statistics*. 2007. Part III.
- D. Freedman. *Statistical Models – Theory and Practice*. 2009. Chapters 2–4.

Data Analysis: Statistical Modeling and Computation in Applications

Gradient Descent

Outline

Recall: Ordinary least Squares Estimator (OLS)

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} (y_i - \mathbf{x}_i \boldsymbol{w})^2$$

- Today: general method for finding minima

Notation: to avoid confusion with "typical" notation in optimization vs statistics, for today's lecture, we replaces β by $\boldsymbol{w}$: $\boldsymbol{w}$ are the parameters to optimize.

Recall: Ordinary least Squares Estimator (OLS)

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} (y_i - \mathbf{x}_i \boldsymbol{w})^2$$

- Today: general method for finding minima

- Recall:
  setting derivative to zero gives *normal equations* and closed form $\hat{\boldsymbol{w}}$

Notation: to avoid confusion with "typical" notation in optimization vs statistics, for today's lecture, we replaces β by $\boldsymbol{w}$: $\boldsymbol{w}$ are the parameters to optimize.

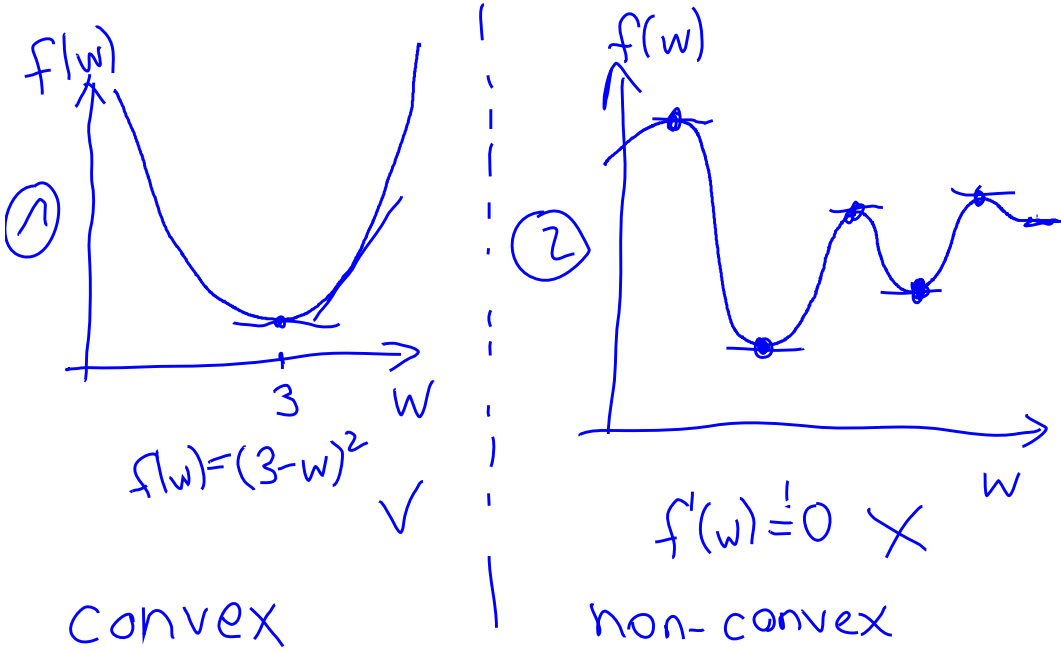
Recall: Ordinary least Squares Estimator (OLS)

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} (y_i - \mathbf{x}_i \boldsymbol{w})^2$$

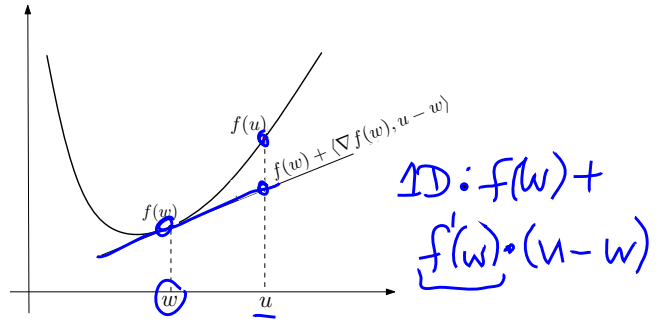
- Today: general method for finding minima
- Recall:
  setting derivative to zero gives *normal equations* and closed form $\hat{\boldsymbol{w}}$
- *When does setting the derivative to zero give the minimum?*

Notation: to avoid confusion with "typical" notation in optimization vs statistics, for today's lecture, we replaces β by $\boldsymbol{w}$: $\boldsymbol{w}$ are the parameters to optimize.

Intuition



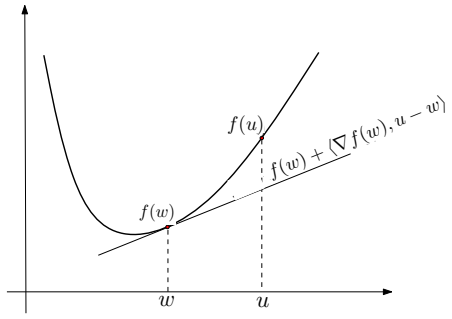
Convexity: "bowl-shapedness"



Function f is *convex* if at each point, the gradient gives a linear lower bound, i.e., for all u, w:

$$f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle.$$

Convexity: "bowl-shapedness"

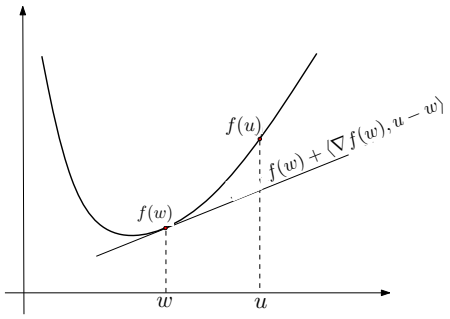


Function f is *convex* if at each point, the gradient gives a linear lower bound, i.e., for all u, w:

$$f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle.$$

- if $\nabla f(w) = 0$ (local property), then w is a *global minimum* (global): $f(u) \geq f(w) + \langle 0, u - w \rangle$ for all u. Not for non-convex functions!

Convexity: "bowl-shapedness"

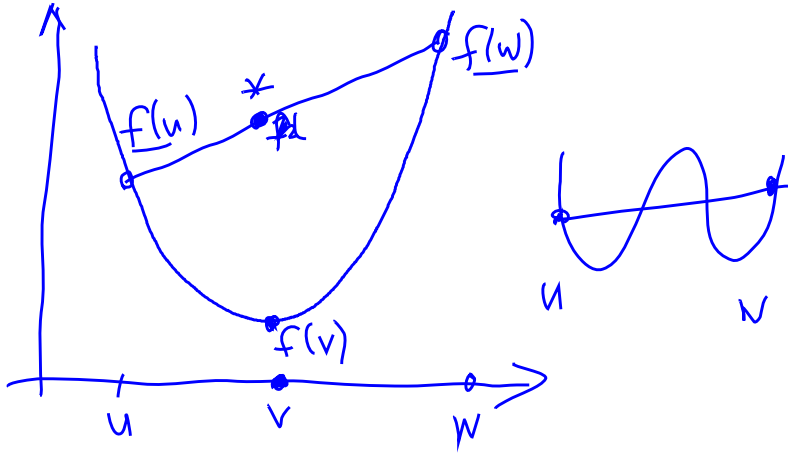


Function f is *convex* if at each point, the gradient gives a linear lower bound, i.e., for all u, w:

$$f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle.$$

- if $\nabla f(w) = 0$ (local property), then w is a *global minimum* (global): $f(u) \geq f(w) + \langle 0, u - w \rangle$ for all u. Not for non-convex functions!
- Hence, for convex f, all we need to find is a point with $\nabla f(w) = 0$.

Convexity: "chord across bowl"



- *"chord across bowl"*: f is convex if for all u, w and $0 \leq \lambda \leq 1$:

$$f\big(\underbrace{\lambda w + (1-\lambda)u}_{v}\big) \leq \underbrace{\lambda f(w) + (1-\lambda) f(u)}_{*}$$

Convexity: 3 criteria

1. "linear lower bound"
2. "chord across bowl"

Convexity: 3 criteria

1. "linear lower bound"
2. "chord across bowl"
3. if f is twice differentiable: f is convex if for all w $\nabla^2 f(w)$ is positive semidefinite (in 1D: $f''(w) \geq 0$).

A is psd if $v^T A v \geq 0$ $\forall v$

Examples of convex functions

scalar	**vector**	
$f(w) = bw$	$f(w) = b^\top w$	for constant b

Examples of convex functions

scalar	**vector**	
$f(w) = bw$	$f(w) = b^\top w$	for constant b
$f(w) = aw^2$	$f(w) = w^\top \mathbf{A} w$	for constant $a \geq 0$, $\mathbf{A}$ positive semidefinite

Stefanie Jegelka (and Caroline Uhler) 8 / 31

Examples of convex functions

scalar	**vector**	
$f(w) = bw$	$f(w) = b^\top w$	for constant b
$f(w) = aw^2$	$f(w) = w^\top \mathbf{A} w$	for constant $a \geq 0$, $\mathbf{A}$ positive semidefinite
$f(w) = \exp(bw)$	$f(w) = \exp(b^\top w)$	for constant b

Stefanie Jegelka (and Caroline Uhler) 8 / 31

Examples of convex functions

scalar	**vector**	
$f(w) = bw$	$f(w) = b^\top w$	for constant b
$f(w) = aw^2$	$f(w) = w^\top \mathbf{A} w$	for constant $a \geq 0$, $\mathbf{A}$ positive semidefinite
$f(w) = \exp(bw)$	$f(w) = \exp(b^\top w)$	for constant b

If $f(w)$ and $g(w)$ are convex functions, then $f(w) + g(w)$ is convex.
$\Rightarrow$ least squares loss is convex!

Stefanie Jegelka (and Caroline Uhler) 8 / 31

Outline

- Convex functions
- Gradient descent: main scheme
- Direction
- Step size
- Convergence
- Stochastic Gradient Descent

Optimization

- sometimes, solving $\nabla f(w) = 0$ is difficult directly:
  - nonlinear equations
  - matrix inversion expensive for large matrices ...

Optimization

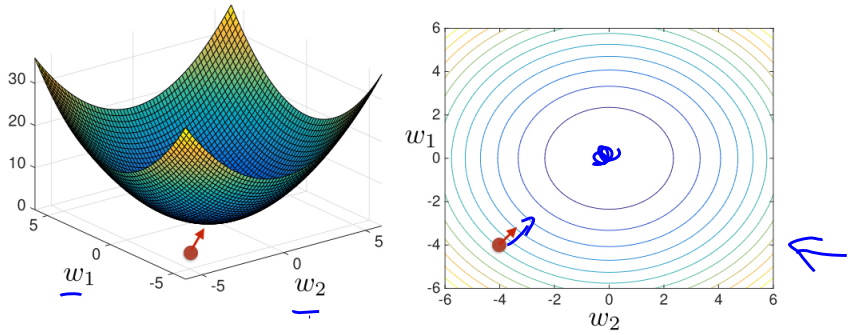
- sometimes, solving $\nabla f(w) = 0$ is difficult directly:
  - nonlinear equations
  - matrix inversion expensive for large matrices ...
- **Optimization**: essential ingredient of modern data science.
  find
$$\min_{w \in \mathbb{R}^p} f(w)$$
  Examples: minimizing (regularized) loss/error, maximizing likelihood, ...

Optimization

- sometimes, solving $\nabla f(w) = 0$ is difficult directly:
  - nonlinear equations
  - matrix inversion expensive for large matrices ...
- **Optimization**: essential ingredient of modern data science.
  find
$$\min_{w \in \mathbb{R}^p} f(w)$$
  Examples: minimizing (regularized) loss/error, maximizing likelihood, ...
- typically iteratively

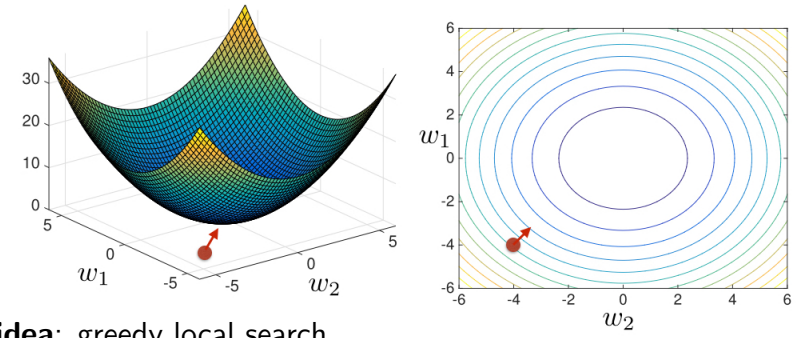
Optimization: Gradient Descent

$$\min_{w \in \mathbb{R}^p} f(w)$$



Optimization: Gradient Descent

$$\min_{w \in \mathbb{R}^p} f(w)$$

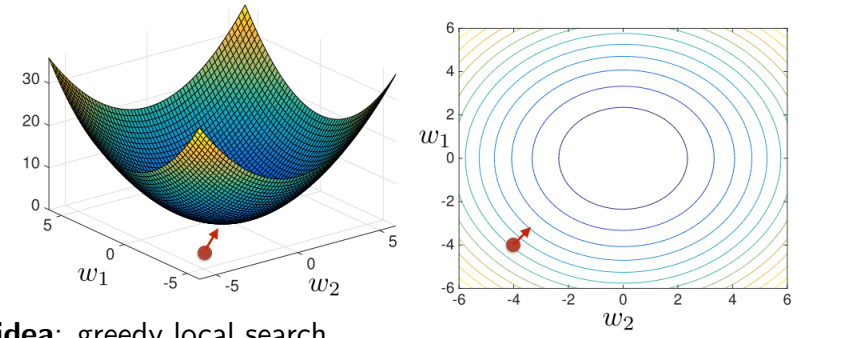


Basic idea: greedy local search

- Start with an arbitrary guess w^0.
- In each iteration, move in direction of "progress" (determined *locally*)

Optimization: Gradient Descent

$$\min_{w \in \mathbb{R}^p} f(w)$$



Basic idea: greedy local search

- Start with an arbitrary guess w^0.
- In each iteration, move in direction of "progress" (determined *locally*)

- if done right, finds point w with $\nabla f(w) \approx 0$.
  convex f: global minimum.
  non-convex f: local minimum, local maximum or saddle point

General scheme

start with some w^0
for $t = 0, 1, 2, \ldots$

$$w^{t+1} \leftarrow w^t + \underbrace{\alpha_t}_{\text{step size}} \underbrace{d^t}_{\text{direction}}$$

$w_i^{t+1} \leftarrow w_i^t + \alpha d_i^t$

General scheme

start with some w^0
for $t = 0, 1, 2, \ldots$

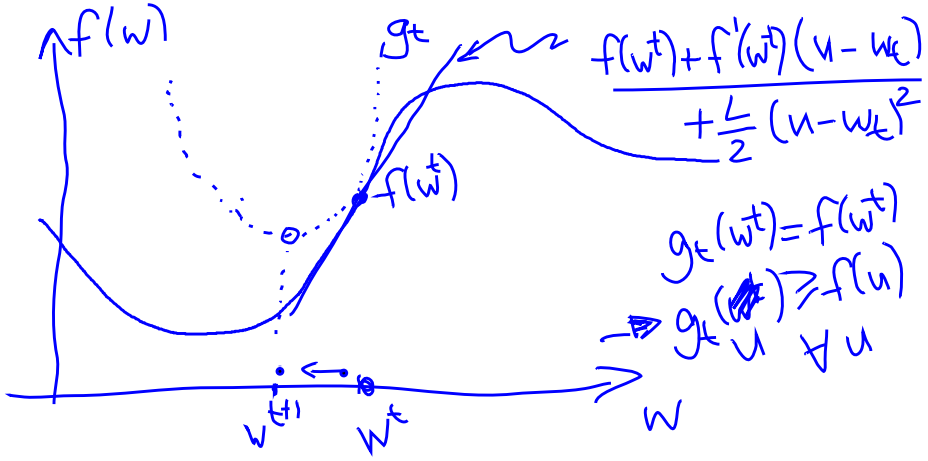
$$w^{t+1} \leftarrow w^t + \underbrace{\alpha_t}_{\text{step size}} \underbrace{d^t}_{\text{direction}}$$

Questions:

1. which direction d^t?
2. step size α_t?
3. how many iterations?

Deriving GD: Intuition

Solving $\nabla f(w) = 0$ is difficult in general, but easy for "nice" quadratic functions: Approximate f locally with quadratic function g_t, minimize g_t.



Deriving GD: Quadratic Upper bounds

- **Want:** (1) $g_t(w) \geq f(w)\ \forall w$ and (2) $g(w^t) = f(w^t)$.

Deriving GD: Quadratic Upper bounds

- **Want:** (1) $g_t(w) \geq f(w)\ \forall w$ and (2) $g(w^t) = f(w^t)$.
- Use *Taylor expansion* to get $g_t(w)$: in 1D

Deriving GD: Quadratic Upper bounds

- **Want:** (1) $g_t(w) \geq f(w)\ \forall w$ and (2) $g(w^t) = f(w^t)$.
- Use *Taylor expansion* to get $g_t(w)$: in 1D

$$f(u) = f(w^t) + f'(w^t)(u - w^t) + \frac{1}{2}f''(w^t)(u - w^t)^2 + \dots$$

Deriving GD: Quadratic Upper bounds

- **Want:** (1) $g_t(w) \geq f(w)\ \forall w$ and (2) $g(w^t) = f(w^t)$.
- Use *Taylor expansion* to get $g_t(w)$: in 1D

$$f(u) = f(w^t) + f'(w^t)(u - w^t) + \frac{1}{2}f''(w^t)(u - w^t)^2 + \dots$$

$$\geq g_t(u) = f(w^t) + f'(w^t)(u - w^t) + \frac{L}{2}(u - w^t)^2$$

- Choose L large enough to satisfy (1).

Deriving GD: Quadratic Upper bounds

- **Want:** (1) $g_t(w) \geq f(w)\ \forall w$ and (2) $g(w^t) = f(w^t)$.
- Use *Taylor expansion* to get $g_t(w)$: in 1D

$$f(u) = f(w^t) + f'(w^t)(u - w^t) + \frac{1}{2}f''(w^t)(u - w^t)^2 + \dots$$

$$g_t(u) = f(w^t) + f'(w^t)(u - w^t) + \frac{L}{2}(u - w^t)^2$$

- Choose L large enough to satisfy (1).
- Setting $g_t'(u) = 0$ gives minimizer of g:

$$u_t = w^t - \frac{1}{L}f'(w^t) = w^{t+1}$$

$$u_t = w^t - \frac{1}{L}\nabla_w f(w^t) = w^{t+1}$$

Deriving GD: Quadratic Upper bounds

- **Want:** (1) $g_t(w) \geq f(w)\ \forall w$ and (2) $g(w^t) = f(w^t)$.
- Use *Taylor expansion* to get $g_t(w)$: in 1D

$$f(u) = f(w^t) + f'(w^t)(u - w^t) + \frac{1}{2}f''(w^t)(u - w^t)^2 + \dots$$

$$g_t(u) = f(w^t) + f'(w^t)(u - w^t) + \frac{L}{2}(u - w^t)^2$$

- Choose L large enough to satisfy (1).
- Setting $g_t'(u) = 0$ gives minimizer of g:

$$u_t = w^t - \frac{1}{L}f'(w^t)$$

Set $w^{t+1} = u_t = w^t - \frac{1}{L}f'(w^t)$, i.e., $d_t = -f'(w^t)$, $\alpha = 1/L$.

Example: least squares regression

- We obtain iteration: $w^{t+1} \leftarrow w^t - \alpha_t \nabla f(w^t)$.

Example: least squares regression

- We obtain iteration: $w^{t+1} \leftarrow w^t - \alpha_t \nabla f(w^t)$.
- Gradient for squared loss:

$$\nabla_w \Big(\sum_{i=1}^N (y_i - x_i \cdot w)^2 \Big) = \sum_{i=1}^N \nabla_w (y_i - x_i \cdot w)^2$$

Example: least squares regression

- We obtain iteration: $w^{t+1} \leftarrow w^t - \alpha_t \nabla f(w^t)$.
- Gradient for squared loss:

$$\nabla_w \Big(\sum_{i=1}^N (y_i - x_i \cdot w)^2 \Big) = \sum_{i=1}^N \nabla_w (y_i - x_i \cdot w)^2$$
$$= -2 \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$$

Example: least squares regression

- We obtain iteration: $w^{t+1} \leftarrow w^t - \alpha_t \nabla f(w^t)$.
- Gradient for squared loss:

$$\nabla_w \Big(\sum_{i=1}^N (y_i - x_i \cdot w)^2 \Big) = \sum_{i=1}^N \nabla_w (y_i - x_i \cdot w)^2$$
$$= -2 \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$$

so $w^{t+1} = w^t + 2\alpha_t \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$

Example: least squares regression

- We obtain iteration: $w^{t+1} \leftarrow w^t - \alpha_t \nabla f(w^t)$.
- Gradient for squared loss:

$$\nabla_w \Big(\sum_{i=1}^N (y_i - x_i \cdot w)^2 \Big) = \sum_{i=1}^N \nabla_w (y_i - x_i \cdot w)^2$$
$$= -2 \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$$

so $w^{t+1} = w^t + 2\alpha_t \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$

- w^{t+1} will be a combination of x_i's

Example: least squares regression

- We obtain iteration: $w^{t+1} \leftarrow w^t - \alpha_t \nabla f(w^t)$.
- Gradient for squared loss:

$$\nabla_w \Big(\sum_{i=1}^N (y_i - x_i \cdot w)^2 \Big) = \sum_{i=1}^N \nabla_w (y_i - x_i \cdot w)^2$$
$$= -2 \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$$

so $w^{t+1} = w^t + 2\alpha_t \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$

- w^{t+1} will be a combination of x_i's
- positive residual ($y_i > \hat{y}_i$):
  add fraction of x_i to w^t, increases dot product $x_i \cdot w$

Example: least squares regression

- We obtain iteration: $w^{t+1} \leftarrow w^t - \alpha_t \nabla f(w^t)$.
- Gradient for squared loss:

$$\nabla_w \Big(\sum_{i=1}^N (y_i - x_i \cdot w)^2 \Big) = \sum_{i=1}^N \nabla_w (y_i - x_i \cdot w)^2$$
$$= -2 \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$$

so $w^{t+1} = w^t + 2\alpha_t \sum_{i=1}^N (y_i - x_i \cdot w) x_i^\top$

- w^{t+1} will be a combination of x_i's
- positive residual ($y_i > \hat{y}_i$):
  add fraction of x_i to w^t, increases dot product $x_i \cdot w$
- negative residual: subtract fraction of x_i, decreases dot product

General scheme

start with some w^0
for $t = 0, 1, 2, \ldots$

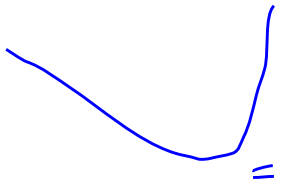
$$w^{t+1} \leftarrow w^t + \underbrace{\alpha_t}_{\text{step size}} \underbrace{d^t}_{\text{direction}}$$

Questions:

1. which direction d^t?
2. step size α_t?
3. how many iterations?

More on step sizes

- Step size determined by quadratic upper bound / increase in slope
- If increase is different in different directions: use smallest step size. slower progress overall.



More on step sizes

- Step size determined by quadratic upper bound / increase in slope
- If increase is different in different directions: use smallest step size. slower progress overall.

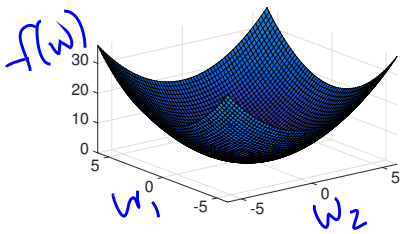
- If we don't know L: Tuning...

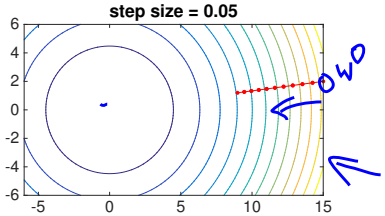
More on step sizes

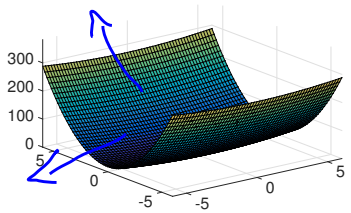
- Step size determined by quadratic upper bound / increase in slope
- If increase is different in different directions: use smallest step size. slower progress overall.
- If we don't know L: Tuning...

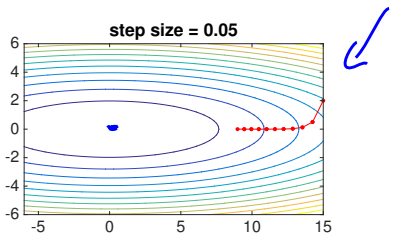
too small: slow progess.
too large: erratic or no convergence.

Step sizes & progress



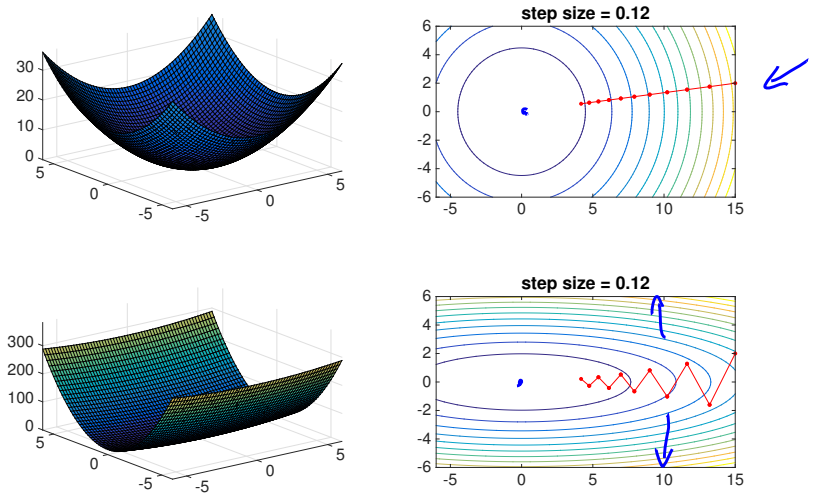




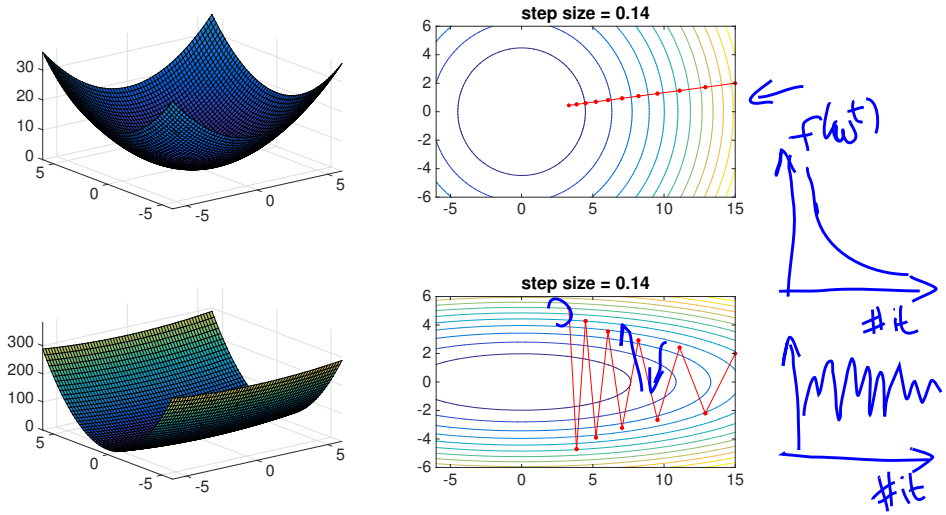


more demo: `http://fa.bianp.net/teaching/2018/eecs227at/gradient_descent.html`

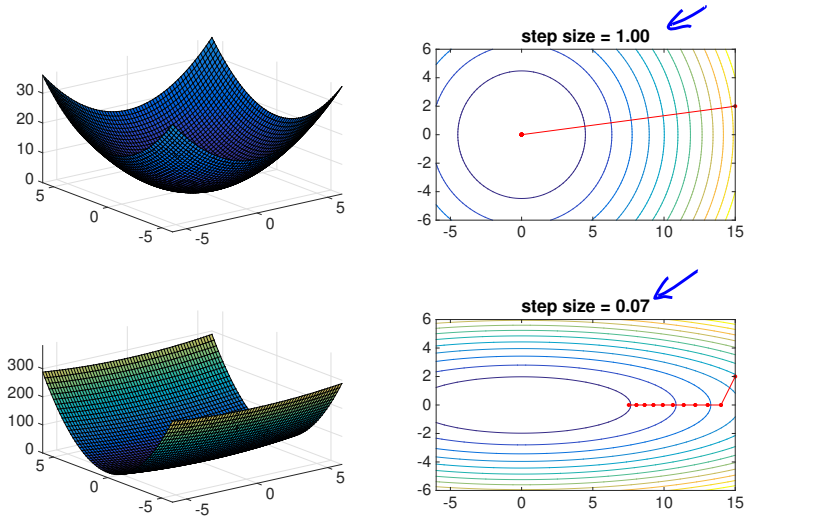
Step sizes & progress



Step sizes & progress



Step sizes: using $\alpha = 1/L$



More on step sizes

- Step size determined by quadratic upper bound / increase in slope
- If increase is different in different directions: use smallest step size. slower progress overall.

- If we don't know L: Tuning, or **Backtracking**:

More on step sizes

- Step size determined by quadratic upper bound / increase in slope
- If increase is different in different directions: use smallest step size. slower progress overall.

- If we don't know L: Tuning, or **Backtracking**:

With step size $\alpha_t = 1/L$: $\quad f(w^{t+1}) \leq f(w^t) - \frac{1}{2L}\|f(w^t)\|^2$.

More on step sizes

- Step size determined by quadratic upper bound / increase in slope
- If increase is different in different directions: use smallest step size. slower progress overall.

- If we don't know L: Tuning, or **Backtracking**:

With step size $\alpha_t = 1/L$: $\quad f(w^{t+1}) \leq f(w^t) - \frac{1}{2L}\|f(w^t)\|^2$.

in each step t: select optimistic α_t, check if

$$f(w_t - \alpha_t \nabla f(w_t)) \leq f(w_t) - \frac{\alpha_t}{2}\|\nabla f(w_t)\|^2$$

If yes: use α_t. If no: $\alpha_t = \alpha_t/2$ and check again.

How many iterations?

- **practice:** e.g. until $\|\nabla f(w^t)\|$ is "small enough", or only small change in loss

- **theory:** $f(w^t) - f(w^*) \leq \ldots$ (gap to optimum)

[1] means f is also lower bounded by a quadratic, with constant m (instead of L). See e.g. Boyd & Vandenberghe book.

How many iterations?

- **practice:** e.g. until $\|\nabla f(w)\|$ is "small enough", or only small change in loss

- **theory:** $f(w^t) - f(w^*) \leq \ldots$ (gap to optimum)

[1] means f is also lower bounded by a quadratic, with constant m (instead of L). See e.g. Boyd & Vandenberghe book.

How many iterations?

- **practice:** e.g. until $\|\nabla f(w)\|$ is "small enough", or only small change in loss
- **theory:** $f(w^t) - f(w^*) \leq \ldots$ (gap to optimum)
- *convex functions with bounded L (gradients)*:

$$f(w^t) - f(w^*) \leq \frac{L\|w^0 - w^*\|^2}{2t}$$

(need $O(1/\epsilon)$ iterations for $f(w^t) - f(w^*) \leq \epsilon$)

[1] means f is also lower bounded by a quadratic, with constant m (instead of L). See e.g. Boyd & Vandenberghe book.

How many iterations?

- **practice:** e.g. until $\|\nabla f(w)\|$ is "small enough", or only small change in loss
- **theory:** $f(w^t) - f(w^*) \leq \ldots$ (gap to optimum)
- *convex functions with bounded L (gradients)*:

$$f(w^t) - f(w^*) \leq \frac{L\|w^0 - w^*\|^2}{2t}$$

(need $O(1/\epsilon)$ iterations for $f(w^t) - f(w^*) \leq \epsilon$)

- *m-strongly convex functions*[1]:

$$f(w^t) - f(w^*) \leq \left(1 - \frac{m}{L}\right)^t (f(w^0) - f(w^*))$$

(need $O(\log(1/\epsilon))$ iterations for $f(w^t) - f(w^*) \leq \epsilon$)

[1] means f is also lower bounded by a quadratic, with constant m (instead of L). See e.g. Boyd & Vandenberghe book.

Outline

- Convex functions
- Gradient descent: main scheme
- Direction
- Step size
- Convergence
- Stochastic Gradient Descent

What if N is large?

Stochastic Gradient descent

- Setup: f is a sum: $f(w) = \sum_{i=1}^N f_i(w)$
- Gradient: $\nabla_w f(w) = \sum_{i=1}^N \nabla_w f_i(w)$

What if N is large?

Stochastic Gradient descent

- Setup: f is a sum: $f(w) = \sum_{i=1}^{N} f_i(w)$
- Gradient: $\nabla_w f(w) = \sum_{i=1}^{N} \nabla_w f_i(w)$
- Idea: estimate sum via few (one) term(s)!

What if N is large?

Stochastic Gradient descent

- Setup: f is a sum: $f(w) = \sum_{i=1}^{N} f_i(w)$
- Gradient: $\nabla_w f(w) = \sum_{i=1}^{N} \nabla_w f_i(w)$
- Idea: estimate sum via few (one) term(s)!

start with some w^0
for $t = 0, 1, 2, \ldots$
draw (data point) i uniformly at random, $1 \leq i \leq N$

$$w^{t+1} \leftarrow w^t - \alpha_t \nabla f_i(w^t)$$

What if N is large?

Stochastic Gradient descent

- Setup: f is a sum: $f(w) = \sum_{i=1}^{N} f_i(w)$
- Gradient: $\nabla_w f(w) = \sum_{i=1}^{N} \nabla_w f_i(w)$
- Idea: estimate sum via few (one) term(s)!

start with some w^0
for $t = 0, 1, 2, \ldots$
draw (data point) i uniformly at random, $1 \leq i \leq N$

$$w^{t+1} \leftarrow w^t - \alpha_t \nabla f_i(w^t)$$

- with "right" step size, also converges to minimum. Step size must shrink ($\approx \frac{1}{t+1}$)

What if N is large?

Stochastic Gradient descent

- Setup: f is a sum: $f(w) = \sum_{i=1}^{N} f_i(w)$
- Gradient: $\nabla_w f(w) = \sum_{i=1}^{N} \nabla_w f_i(w)$
- Idea: estimate sum via few (one) term(s)!

start with some w^0
for $t = 0, 1, 2, \ldots$
draw (data point) i uniformly at random, $1 \leq i \leq N$

$$w^{t+1} \leftarrow w^t - \alpha_t \nabla f_i(w^t)$$

- with "right" step size, also converges to minimum. Step size must shrink ($\approx \frac{1}{t+1}$)
- more erratic, but standard for large data

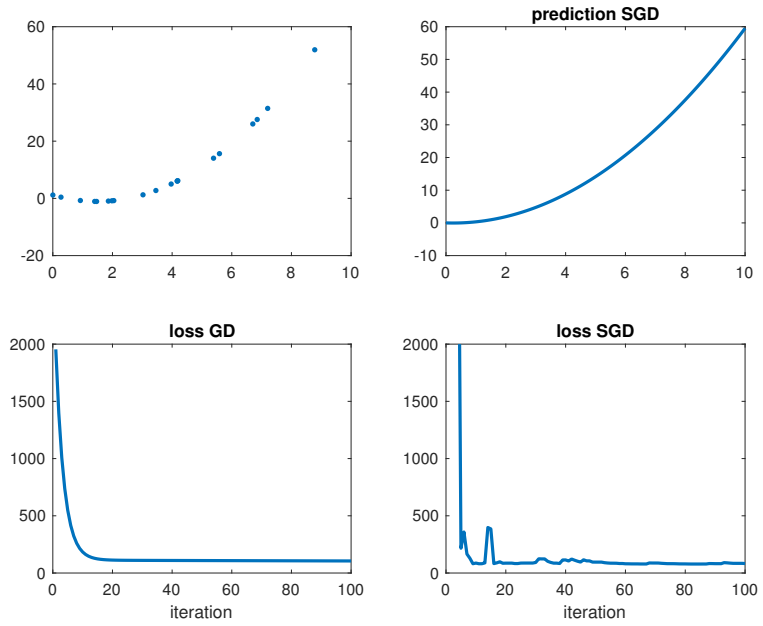
Demo

- Fit a function to observations from $y = \beta_1 x + \beta_2 x^2$ (upper left: observations) via least squares regression with gradient descent and stochastic gradient descent.

Demo

- Fit a function to observations from $y = \beta_1 x + \beta_2 x^2$ (upper left: observations) via least squares regression with gradient descent and stochastic gradient descent.
- Bottom row: Loss function (sum of squared residuals) values for gradient descent (GD, left) and stochastic gradient descent (SGD,right). Note that SGD only uses *one* data point per iteration, whereas GD uses all data points in each iteration.

Demo: plots



Summary: Gradient descent

- (Stochastic) Gradient Descent is a pillar of modern machine learning
- iterative method to find a point with zero gradient
- small steps in direction of negative gradient
- for convex functions, finds the global minimum

References

Convex Optimization

- S. Boyd & L. Vandenberghe. *Convex Optimization*. Available online: `http://stanford.edu/~boyd/cvxbook/`
  Parts of Chapters 3.1, (3.2 for additional optional reading); parts of Chapter 9.1 and 9.3

For some background in linear algebra

- very short: Appendix in Boyd & Vandenberghe. Or: Many statistics books have a chapter on it, e.g., D. Freedman. *Statistical Models – Theory and Practice*.
- a bit longer:
  T.A. Garrity. *All the Mathematics you missed: But need to know for Graduate School*. Cambridge University Press.

Appendix: some linear algebra concepts

- **rank** of matrix: max. number of linearly independent columns (rows)

- $\mathbf{X}^\top\mathbf{X}$ is invertible if it has full rank (no zero eigenvalues), i.e., if $N \geq p$ and all columns of $\mathbf{X}$ are linearly independent.

Appendix: some linear algebra concepts

- **rank** of matrix: max. number of linearly independent columns (rows)
- **eigenvalue**: an eigenvector $\mathbf{u}$ of a matrix $\mathbf{A}$ and the corresponding eigenvalue λ satisfy $\mathbf{Au} = \lambda\mathbf{u}$.

$$\lambda_{\max} = \max_{\mathbf{u}\in\mathbb{R}^p, \|\mathbf{u}\|=1} \mathbf{u}^\top\mathbf{Au} \qquad \lambda_{\min} = \min_{\mathbf{u}\in\mathbb{R}^p, \|\mathbf{u}\|=1} \mathbf{u}^\top\mathbf{Au}$$

- $\mathbf{X}^\top\mathbf{X}$ is invertible if it has full rank (no zero eigenvalues), i.e., if $N \geq p$ and all columns of $\mathbf{X}$ are linearly independent.

Appendix: some linear algebra concepts

- **rank** of matrix: max. number of linearly independent columns (rows)
- **eigenvalue**: an eigenvector $\mathbf{u}$ of a matrix $\mathbf{A}$ and the corresponding eigenvalue λ satisfy $\mathbf{Au} = \lambda\mathbf{u}$.

$$\lambda_{\max} = \max_{\mathbf{u}\in\mathbb{R}^p, \|\mathbf{u}\|=1} \mathbf{u}^\top\mathbf{Au} \qquad \lambda_{\min} = \min_{\mathbf{u}\in\mathbb{R}^p, \|\mathbf{u}\|=1} \mathbf{u}^\top\mathbf{Au}$$

- symmetric $\mathbf{A}$ is *positive semidefinite* (psd) if

$$\mathbf{v}^\top\mathbf{Av} \geq 0 \text{ for all } \mathbf{v} \in \mathbb{R}^p$$

  Equivalent: $\mathbf{A}$ is symmetric and all eigenvalues are nonnegative.

- $\mathbf{X}^\top\mathbf{X}$ is invertible if it has full rank (no zero eigenvalues), i.e., if $N \geq p$ and all columns of $\mathbf{X}$ are linearly independent.

Appendix: Additional note: What is L?

- We want (scalar version):

$$f(u) \leq f(w) + f'(w)(u-w) + \frac{L}{2}(u-w)^2$$

Taylor expansion:

$$f(u) = f(w) + f'(w)(u-w) + \tfrac{1}{2}f''(z)(u-w)^2$$

so $L = \max_z f''(z)$.

Appendix: Additional note: What is L?

- We want (scalar version):

$$f(u) \leq f(w) + f'(w)(u-w) + \frac{L}{2}(u-w)^2$$

Taylor expansion:

$$f(u) = f(w) + f'(w)(u-w) + \tfrac{1}{2}f''(z)(u-w)^2$$

so $L = \max_z f''(z)$.

- for vectors, this condition becomes:

$$v^\top [\nabla^2 f(z)] v \;\leq\; L v^\top v \qquad \text{for } v = (u-w)$$

i.e., L must be larger than largest eigenvalue of $\nabla^2 f(z)$ for all z.

Appendix: Additional note: What is L?

- We want (scalar version):

$$f(u) \leq f(w) + f'(w)(u-w) + \frac{L}{2}(u-w)^2$$

Taylor expansion:

$$f(u) = f(w) + f'(w)(u-w) + \tfrac{1}{2}f''(z)(u-w)^2$$

so $L = \max_z f''(z)$.

- for vectors, this condition becomes:

$$v^\top [\nabla^2 f(z)] v \;\leq\; L v^\top v \qquad \text{for } v = (u-w)$$

i.e., L must be larger than largest eigenvalue of $\nabla^2 f(z)$ for all z.

- if $f(w) = w^\top \mathbf{A} w + bw$, then $\nabla^2 f(z) = \mathbf{A}$ and L is maximum eigenvalue of $\mathbf{A}$: $L = \lambda_{\max}$