UCDAVIS
DEPARTMENT OF STATISTICS

STA 108 Applied Statistical Methods: Regression Analysis

Linear Regression

Shizhe Chen, PhD

Spring 2020

You are expected to read the course notes **before** lectures.

Statistics $\neq$ Data analysis

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

R. A. Fisher

Simple linear regression[1] model

$$y_i = x_i\beta_1 + \beta_0 + \epsilon_i, i = 1, \ldots, n$$

- (x_i, y_i) is the ith sample or observation for $i = 1, \ldots, n$ with n being the sample size
- $y_i \in \mathbb{R}$ is the **response** or **dependent variable** of the ith sample
- $x_i \in \mathbb{R}$ is the **covariate**, **explanatory variable**, or **independent variable** of the ith sample
- β_0 is the **intercept term**
- β_1 is the **regression slope**
- $\epsilon_i \in \mathbb{R}$ is the **noise** or **error term** of the ith sample

[1] Regression: a return to a former or less developed state. So why "regression"?

Assumptions on ϵ

$$y_i = x_i\beta_1 + \beta_0 + \epsilon_i, \quad i = 1, \ldots, n$$

- ϵ_i is a (unobserved) random variable
- $\epsilon_1, \ldots, \epsilon_n$ independently and identically distributed (i.i.d.)
- $\mathbb{E}[\epsilon_i] = 0$ for $i = 1, 2, \ldots, n$
- $\mathrm{var}(\epsilon_i) = \sigma^2$ for $i = 1, 2, \ldots, n$

The covariate x_i is considered to be fixed (i.e., not random) in this course

Fitting the simple linear regression

$$y_i = x_i\beta_1 + \beta_0 + \epsilon_i, \quad i = 1, \ldots, n$$

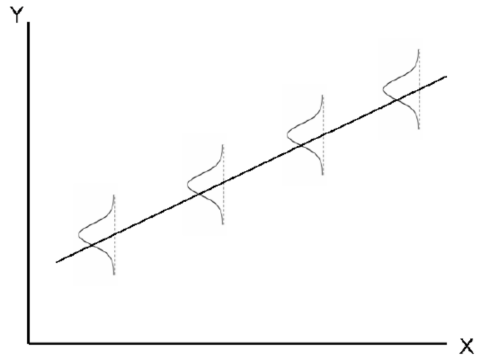
Open `R` to fit a linear regression on[2]

- synthetic data you simulated
- the advertising data
- flu shot data
- Project STAR

[2] The last three are available in Chapter 1 of the course Gitbook

Properties of linear regression

$$y_i = x_i\beta_1 + \beta_0 + \epsilon_i, \quad i = 1, \ldots, n$$

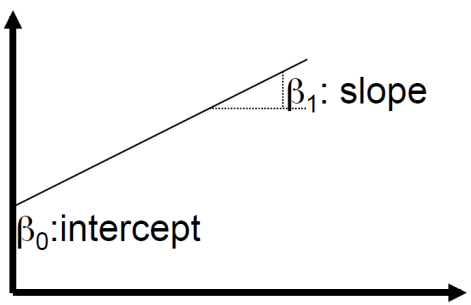


- Expectation of y given $x = a$

$$\mathbb{E}[y \mid x = a] = a\beta_1 + \beta_0$$

- The slope

$$\beta_1 = \mathbb{E}[y \mid x = a + 1] - \mathbb{E}[y \mid x = a]$$



- The intercept

$$\beta_0 = \mathbb{E}[y \mid x = 0]$$

Interpretation of β_1 and β_0

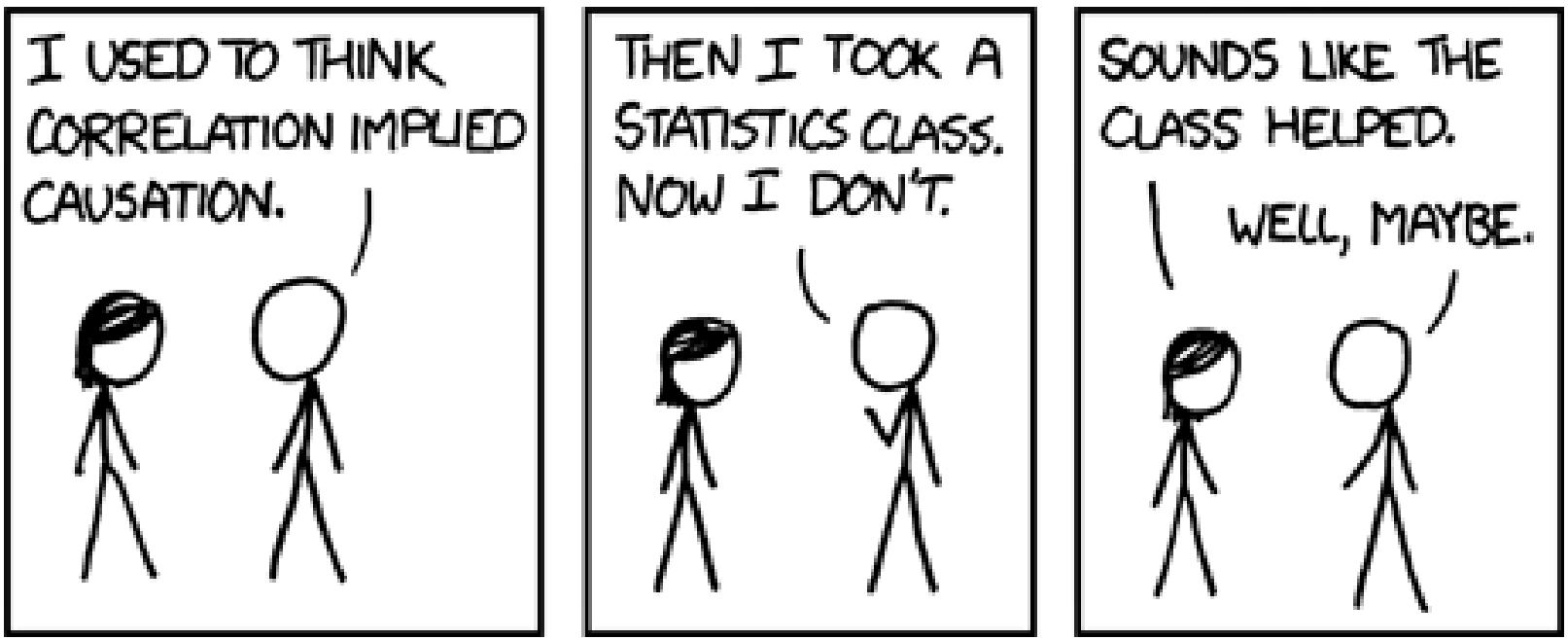
Interpretation of :

β_1 is the difference in the (population) mean response associated with a one unit difference (upwards) in x

(No causality!!!)

Q: How do we interpret β_0?

Correlation/association v.s. Causality[3]

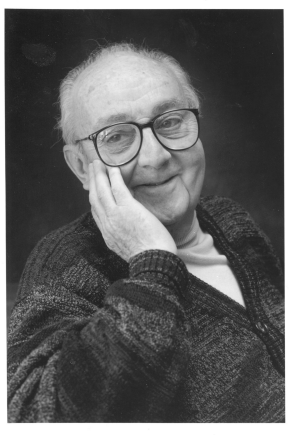


[3] Source

Exercise: interpret the estimated parameters on

- Advertising data
- Flut shot data
- Project STAR

A quote that you will see often



All models are wrong, but some are useful.

George E. P. Box

Estimation

Q: how to find the best line given a batch of samples?

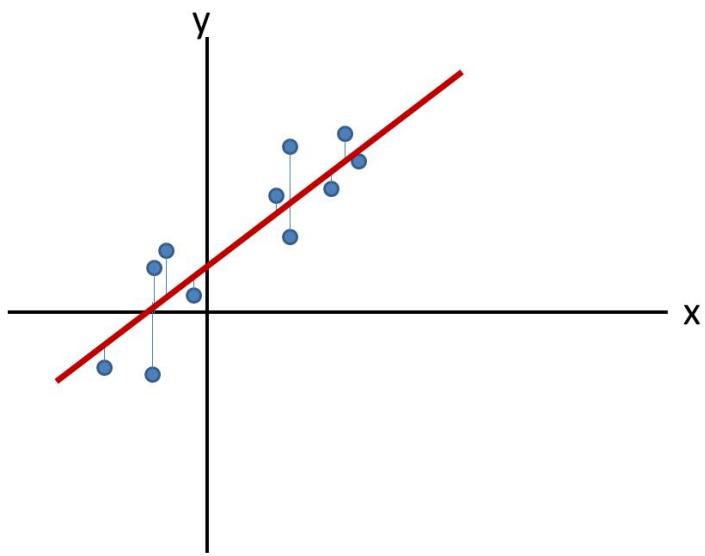
- Define "best"
- Find the best line

Loss function, I

- Residual: $e_i = y_i - x_i\beta_1 - \beta_0$
- **Residual sum of squared** or **squared errors**

$$\mathcal{L}(\beta_1, \beta_0) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - x_i\beta_1 - \beta_0)^2$$

Graphical Interpretation: [Element of Stat Learning: Hastie et al.]



Draw your own version in R!

Least squares estimator

Least squares estimator $(\hat{\beta}_1, \hat{\beta}_0)$: minimizer of $\mathcal{L}(\beta_1, \beta_0)$

Connection with sample mean

Fix $\beta_1 = 0$, and we have

$$\mathcal{L}(0, \beta_0) = \sum_{i=1}^{n} (y_i - \beta_0)^2.$$

Claim: the sample mean $\bar{y}$ is the minimizer of $\mathcal{L}(0, \beta_0)$.

Do one of the following

- Prove this claim via rigorous mathematical derivation
- Verify this claim using simulated data

Connection with sample mean (cont.)

What is the advantage of $\mathcal{L}(\beta_1, \beta_0)$ over $\mathcal{L}(0, \beta_0)$?

$$\mathcal{L}(0, \beta_0) = \sum_{i=1}^{n} (y_i - \beta_0)^2.$$

$$\mathcal{L}(\beta_1, \beta_0) = \sum_{i=1}^{n} (y_i - x_i\beta_1 - \beta_0)^2$$

Finding the least squares estimator

Solve the following optimization problem

$$\underset{\beta_0,\beta_1}{\text{minimize}} \sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2.$$

Do one of the following

- Write a function to solve the optimization problem using `R`
- Derive the estimator using your knowledge in **algebra** or **calculus**

Least squares estimator: solution

$$(\hat{\beta}_1, \hat{\beta}_0)^{\mathrm{T}} = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i \beta_1 - \beta_0)^2,$$

where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Decomposition of sum of squares[4]

- TSS: total sum of squares $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- ESS: explained sum of squares $\mathrm{ESS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- RSS: residual sum of squares $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

We have

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$$

Recall that $\mathcal{L}(0, \beta_0) = \sum_{i=1}^{n}(y_i - \beta_0)^2$. Any connections?

INTERPRETATION: this decomposition allows a characterization of the usefulness of the covariate in predicting the response variable.

[4] The acronyms are merely used for simplicity of notation.

Formalize the characterization: R^2

The sum of squares decomposition allows a characterization of the usefulness of the covariate in predicting the response variable.

We define R^2 (the coefficient of determination) as

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Connection with correlation

Sample correlation between the vectors[5] $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$.

Claim: $R^2 = \widehat{\mathrm{cor}}^2(\boldsymbol{x}, \boldsymbol{y})$.

Do one of the following

- Prove this claim via rigorous mathematical derivation
- Verify this claim using simulated data

[5] The transpose $^{\mathrm{T}}$ ensures that $\boldsymbol{x}$ and $\boldsymbol{y}$ are **column vectors** by convention.

Loss function, II

Assuming the errors $\{\epsilon_i\}_{i=1}^n$[6] are **normally** distributed

- Likelihood: density of normal random variable $z \sim \mathcal{N}(\mu, \sigma^2)$ is

$$f(z) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-(z-\mu)^2/(2\sigma^2)\right\}$$

- **Log-likelihood** for $y_i = x_i\beta_1 + \beta_0 + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$[7]

$$\tilde{\mathcal{L}}(\beta_1, \beta_0) = -\sum_{i=1}^{n}(y_i - x_i\beta_1 - \beta_0)^2$$

Looks familiar?

- Maximum likelihood estimator: estimator that maximizes the likelihood, or minimize the negative log-likelihood

[6] New notation: $\{\epsilon_i\}_{i=1}^n$ is equivalent to $\{\epsilon_i : i = 1, 2, \ldots, n\}$.

[7] Ignoring the variance σ^2 for now...

Are these numbers reliable?

Recall that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Observation: $\hat{\beta}_1$ and $\hat{\beta}_0$ fully depend on the data $\{(x_i, y_i)\}_{i=1}^n$.

Q: If we have a new set of data, will our estimates change?

Point estimates and sampling distributions

- $\hat{\beta}_1$ and $\hat{\beta}_0$ are estimators of β_1 and β_0
- Values of $\hat{\beta}_1$ and $\hat{\beta}_0$ are fixed given a **fixed** set of samples
- Values of $\hat{\beta}_1$ and $\hat{\beta}_0$ are **point estimates** of the true unknown parameters β_1 and β_0 given a **fixed** set of samples
- Samples of $(x_1, y_1), \ldots, (x_n, y_n)$ may vary ($\{\epsilon_i\}$ are random)
- Values of $\hat{\beta}_1$ and $\hat{\beta}_0$ change as the set of samples changes (**sampling distribution**)

Characterizing the sampling distributions

- Expectation (unbiased)

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \quad \mathbb{E}[\hat{\beta}_0] = \beta_0$$

- Variance

$$\text{var}(\hat{\beta}_1) = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\sigma^2 \quad \text{var}(\hat{\beta}_0) = \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}\sigma^2$$

- $\hat{\beta}_1$ and $\hat{\beta}_0$ are both linear in $y_1, \ldots, y_n$.

Do one of the following

- Prove this claim via rigorous mathematical derivation
- Verify this claim using simulated data

Linear unbiased estimators

An estimator $\tilde{\beta}_1$ is a linear unbiased estimator of β_1 if it is

- linear: $\tilde{\beta}_1 = \sum_{i=1}^{n} c_{1,i}\, y_i$
- unbiased: $\mathbb{E}[\tilde{\beta}_1] = \beta_1$

Gauss-Markov theorem: $\mathrm{var}(\hat{\beta}_1) \leq \mathrm{var}(\tilde{\beta}_1)$.

The least squares estimator is the **B**est **L**inear **U**nbiased **E**stimator.

Roses are red, LSE is blue.

Models are wrong, but some are useful.

Do one of the following

- Find and read the proof for this theorem
- Verify this claim using simulated data (How?)

Characterizing the sampling distributions

In these expression, σ is still unknown

- Expectation

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \quad \mathbb{E}[\hat{\beta}_0] = \beta_0$$

- Variance

$$\text{var}(\hat{\beta}_1) = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2 \quad \text{var}(\hat{\beta}_0) = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2$$

Q: how to estimate σ?

Residuals

- **Definition:**

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \qquad i = 1, \ldots, n.$$

- Intuition: Residuals are the remaining parts of y_i that cannot be explained by x_i in the linear model, or the information that we did not fully capture from our data set
- Properties of e_i:

$$\sum_{i=1}^{n} e_i = 0 \quad \sum_{i=1}^{n} e_i x_i = 0$$

$$\mathrm{Q}: \mathbb{E}[e_i] = ? \quad \mathbb{E}\left[\sum_{i=1}^{n} e_i\right] = ?$$

Residuals and the unknown variance

- It contains the information about the unknown parameter σ^2.

$$e_i = \epsilon_i + (\beta_1 - \hat{\beta}_1)x_i + (\beta_0 - \hat{\beta}_0), \qquad i = 1, \ldots, n.$$

Recall: $\epsilon_i \in \mathbb{R}$ is the error term with $E[\epsilon_i] = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$

- Variance of e_i: [8]

$$\mathrm{var}(e_i) = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right]$$

$$\mathrm{Q} : \mathrm{cov}(e_i, e_j) = ?$$

[8] See a long proof here

Estimate the unknown variance

- Recall the usual variance estimator: $(n-1)^{-1}\sum_{i=1}^{n}(z_i-\bar{z})^2$, for a sample $\{a_1, a_2, \ldots, a_n\}$
- Estimate σ^2 as

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

- Claim: $\hat{\sigma}^2$ is unbiased: $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$

Do one of the following

- Prove this claim using rigorous mathematical derivation
- Verify this claim using simulated data

Sampling distributions: expectation and variance

- Expectation

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \quad \mathbb{E}[\hat{\beta}_0] = \beta_0$$

- Variance estimators

$$\text{var}(\hat{\beta}_1) = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\hat{\sigma}^2 \quad \text{var}(\hat{\beta}_0) = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\hat{\sigma}^2$$

Sampling distributions: standard error

- Standard error: the standard deviation of an estimator's sampling distribution
- The standard error tells us the "average amount that our estimates are different from the actual value".

$$\mathrm{SE}(\hat{\beta}_1) = \left\{\mathbb{E}[\hat{\beta}_1 - \beta_1]^2\right\}^{1/2} = \left\{\mathrm{var}(\hat{\beta}_1)\right\}^{1/2}, \ \mathrm{SE}(\hat{\beta}_0) = \left\{\mathrm{var}(\hat{\beta}_0)\right\}^{1/}$$

- Estimators of the standard errors

$$\widehat{\mathrm{SE}}(\hat{\beta}_0) = \hat{\sigma}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]^{1/2}$$

$$\widehat{\mathrm{SE}}(\hat{\beta}_1) = \hat{\sigma}\left[\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]^{1/2}$$

Sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$

So far, we have derived expectations and variances of $\hat{\beta}_1$ and $\hat{\beta}_0$. However, there are multiple distribution with the same mean and variance

Generate the following random variables in R with expectation 0 and variance 2

- Normal (Gaussian)
- Uniform
- Bernoulli

Q: What are the distributions for $\hat{\beta}_1$ and $\hat{\beta}_0$?

Sampling distribution of $\hat{\beta}_1$

Recall:

- $\hat{\beta}_1 = \sum_{i=1}^n c_{1,i} y_i$ where $c_{1,i} = (x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2$.
- $y_i = x_i \beta_1 + \beta_0 + \epsilon_i$

We can show that

$$\hat{\beta}_1 = \sum_{i=1}^n c_{1,i}(x_i \beta_1 + \beta_0 + \epsilon_i) = \beta_1 + \sum_{i=1}^n c_{1,i} \epsilon_i$$

$$\text{Because } \sum_{i=1}^n c_{1,i} x_i = 1 \text{ and } \sum_{i=1}^n c_{1,i} = 0.$$

Recall: we assume that $\mathbb{E}[\epsilon_i] = 0$ and $\text{var}(\epsilon_i) = \sigma^2$ ...

but the distributions of $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are unknown!

Distribution of the sum of random variables

$\hat{\beta}_1 - \beta_1 = \sum_{i=1}^{n} c_{1,i}\epsilon_i$: $\hat{\beta}_1 - \beta_1$ is the weighted sum of $\{\epsilon_i\}_{i=1}^{n}$ with weights $\{c_{1,i}\}_{i=1}^{n}$.

1. Find the exact distribution of $\hat{\beta}_1$ by assuming some distribution for $\{\epsilon_i\}_{i=1}^{n}$
2. Find the asymptotic distribution of $\hat{\beta}_1$ as the sample size n approaches infinity (i.e., assuming that the sample size is **sufficiently large**)

THEORY $\neq$ REALITY

Recall: Asymptotic distribution[9]

Consider i.i.d. random variables $z_1, \ldots, z_n$ with expectation $\mathbb{E}[z]$ and finite variance σ_z^2.

LARGE OF LARGE NUMBERS:

$\bar{z}$ converges (with probability 1) to $\mathbb{E}[z]$ as n grows, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - \mathbb{E}[z]) \to 0.$$

CENTRAL LIMIT THEOREM:

$n^{1/2}\hat{\sigma}_z^{-1/2}(\bar{z} - \mathbb{E}[z])$ converges (in distribution) to $\mathcal{N}(0,1)$ as n grows, where $\hat{\sigma}_z = \left\{(n-1)^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2\right\}^{1/2}$, i.e.,

$$\frac{\sum_{i=1}^{z}(z_i - \mathbb{E}[z])}{n^{1/2}\hat{\sigma}_z} \to \mathcal{N}(0,1).$$

[9] Don't remember? Verify these claims in R!

Intuition of the asymptotic distribution

Large of large numbers:

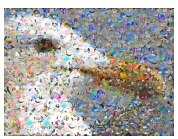
$$\frac{1}{n}\sum_{i=1}^{n}(z_i - \mathbb{E}[z]) \to 0.$$

Central limit theorem:

$$\frac{\sum_{i=1}^{z}(z_i - \mathbb{E}[z])}{n^{1/2}\hat{\sigma}_z} \to \mathcal{N}(0,1).$$

Sum of random variables at "low magnifications"!

Intuition of the asymptotic distribution (cont.)[10]



[10]Source of figure: Wikipedia (link)

Asymptotic distribution of $\hat{\beta}_1$

For $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} c_{1,i}\epsilon_i$, where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. random variables with mean 0 and variance σ^2

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} \to \mathcal{N}(0,1),$$

where

$$\hat{\sigma} = \left\{ \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 \right\} \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Asymptotic distribution of $\hat{\beta}_1$ (cont.)

The Central Limit Theorem we know:

For i.i.d. random variables $z_1, \ldots, z_n$ with expectation $\mathbb{E}[z]$ and finite variance $\mathrm{var}(z)$:

$$\frac{\sum_{i=1}^{z}(z_i - \mathbb{E}[z])}{n^{1/2}\hat{\sigma}_z} \to \mathcal{N}(0,1),$$

where $\hat{\sigma}_z = \left\{(n-1)^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2\right\}^{-1/2}$

Does it apply to $\sum_{i=1}^{n} c_{1,i}\epsilon_i$?

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} = \frac{\sum_{i=1}^{n} c_{1,i}\epsilon_i}{\hat{\sigma}/S_{xx}^{1/2}}$$

No, because $\mathrm{var}(c_{1,i}\epsilon_i) = c_{1,i}^2\sigma^2$ depends on i!

Lindeberg-Feller central limit theorem

Suppose $\{z_1, z_2, \ldots, z_n\}$ is a sequence of independent random variables with $\mathbb{E}[z_i] = \mu_i$ and $\mathrm{var}(z_i) = \sigma_i^2 < \infty$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If the sequence satisfies that, for all $\epsilon > 0$

$$\lim_{n\to\infty} \frac{1}{s_n} \sum_{k=1}^{n} \mathbb{E}\left[(z_i - \mu_i)^2 \mathbf{1}_{\{|z_i-\mu_i|>\epsilon s_n\}}\right] = 0, \qquad (1)$$

then as n approaches infinity, $s_n^{-1} \sum_{i=1}^n (z_i - \mu_i)$ converges in distribution to $\mathcal{N}(0, 1)$.

Lindeberg-Feller central limit theorem (cont.)

The L-F CLT requires

- $z_1, z_2, \ldots, z_n$ are independent with finite means and variances
- Eq. 1 implies that $\lim_{n\to\infty} \sigma_i^2 / \sum_{i=1}^{n} \sigma_i^2 = 0$ for $i = 1, 2, 3, \ldots$, i.e., the information in every variable is incremental compared to the total information from all samples as n grows

The L-F CLT does **not** require

- $z_1, z_2, \ldots, z_n$ are identically distributed
- $z_1, z_2, \ldots, z_n$ have the same mean and variance

Applying Lindeberg-Feller central limit theorem to $\hat{\beta}_1$

Let $z_i = c_{1,i}\epsilon_i$ for $i = 1, \ldots, n$, then

- $\mathbb{E}[z_i] = 0$ and $\text{var}(z_i) = c_{1,i}^2\sigma^2$
- $s_n^2 = \sum_{i=1}^n \text{var}(z_i) = \sum_{i=1}^n c_{1,i}^2\sigma^2 = \sigma^2/S_{xx}$
- $s_n^{-1}\sum_{i=1}^n z_i = (\hat{\beta}_1 - \beta_1)/(\sigma/S_{xx}^{1/2})$

Therefore, $(\hat{\beta}_1 - \beta_1)/(\sigma/S_{xx}^{1/2}) \to \mathcal{N}(0,1)$ by the L-F CLT.

Replacing σ with its estimator $\hat{\sigma} = \{(n-2)^{-1}\sum_{i=1}^n e_i^2\}^{1/2}$ gives

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} \to \mathcal{N}(0,1).$$

Verify this claim in R with errors $\{\epsilon_i\}_{i=1}^p$ that are

- Bernoulli
- Uniform
- Normal

Finding the asymptotic distribution of $\hat{\beta}_1$

(L-F) Central Limit Theorem gives, as sample size approaches infinity,

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} \to \mathcal{N}(0,1).$$

An alternative approach to find the asymptotic distribution is the Bootstrap.

- Also requires large sample sizes
- **Not** useful in simple linear regression model
- Very powerful for complicated models

Bootstrap: Intuition

- The value of $\hat{\beta}_1$ is determined by the samples $\{(x_i, y_i) : i = 1, \ldots, n\}$
- The true sampling distribution of $\hat{\beta}_1$ depends on the true distribution of samples $\{(x_i, y_i) : i = 1, \ldots, n\}$, denoted as F
- We can construct the sampling distribution of $\hat{\beta}_1$ if we can draw samples from F, but F is unknown!
- Bootstrap:
  - Approximate the true distribution F with the empirical distribution F_n
  - Draw samples from F_n
  - Construct the bootstrap distribution of $\hat{\beta}_1$

Bootstrap procedure

Suppose that we want to learn the sampling distribution of a statistics $g(\cdot)$, for instance,

$$g\big(\{(x_i, y_i) : i = 1, \ldots, n\}\big) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\beta}_1.$$

- For $b = 1, \ldots, B$ (e.g., $B = 10,000$)
  1. Draw n samples $\{(x_{i,b}, y_{i,b}) : i = 1, \ldots, n\}$ from F_n (sampling replacement!)
  2. Calculate $\hat{\beta}_{1,b} = g\big(\{(x_{i,b}, y_{i,b}) : i = 1, \ldots, n\}\big)$ on the bootstrap samples
- Construct the bootstrap distribution using $\{\hat{\beta}_{1,b} : b = 1, \ldots, B\}$

Implement the bootstrap in `R`, and compare the distribution with the CLT result

Estimate the sampling distribution via bootstrap

Given $\{\hat{\beta}_{1,b} : b = 1, \ldots, B\}$, we can estimate the sampling distribution of $\hat{\beta}_1$ with

1. the empirical distribution of $\{\hat{\beta}_{1,b} : b = 1, \ldots, B\}$
2. $\mathcal{N}(\tilde{\beta}_B, \tilde{\sigma}^2_{\beta,B})$, where $\tilde{\beta}_B$ and $\tilde{\sigma}^2_{\beta,B}$ are the sample mean and sample variance of $\{\hat{\beta}_{1,b} : b = 1, \ldots, B\}$

- In (2), it is required that the asymptotic distribution of $\hat{\beta}_1$ is normal (e.g., from the central limit theorem)
- Estimation of the **moments** ($\mathbb{E}[z]$, $\mathbb{E}[z^2]$, ...) is more reliable than the empirical distribution

Implement this version of bootstrap in `R`

Finding the asymptotic distribution: Summary

Central Limit Theorem gives, as sample size approaches infinity,

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} \to \mathcal{N}(0,1).$$

Bootstrap yields the bootstrap distribution, based on $\{\hat{\beta}_{1,b} : b = 1, \ldots, B\}$, that approximates the asymptotic distribution of $\hat{\beta}_1$ as sample size approaches infinity

- Both characterize the asymptotic distribution
- Both require large sample size
- Both based on the empirical process theory
- Bootstrap is computationally intense
- CLT is restricted to mathematically tractable case

Q: What is the asymptotic distribution?

What if there are very few samples?

Repeat the previous simulation with

- 100 samples
- 50 samples
- 30 samples
- 10 samples
- 5 samples

Recall: Distribution of sum of random variables

$\hat{\beta}_1 - \beta_1 = \sum_{i=1}^{n} c_{1,i}\epsilon_i$: $\hat{\beta}_1 - \beta_1$ is the weighted sum of $\{\epsilon_i\}_{i=1}^{n}$ with weights $\{c_{1,i}\}_{i=1}^{n}$

1. Find the exact distribution of $\hat{\beta}_1$ by assuming some distribution for $\{\epsilon_i\}_{i=1}^{n}$
2. Find the asymptotic distribution of $\hat{\beta}_1$ as the sample size n approaches infinity (i.e., assuming that the sample size is **sufficiently large**)

THEORY $\neq$ REALITY

1.
2. Asymptotic distributions are unrealistic and does not apply in small samples

Distribution of $\hat{\beta}_1$ in small samples

Distribution of $\hat{\beta}_1 - \beta_1 = \sum_{i=1}^{n} c_{1,i}\epsilon_i$ is known **if** the distributions of $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are known.

The most common assumption: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, n$.

Under the normality assumption

- $\hat{\beta}_1 - \beta_1 = \sum_{i=1}^{n} c_{1,i}\epsilon_i \sim \mathcal{N}(0, S_{xx}\sigma^2)$ (σ^2 unknown!)
- $\left(\hat{\sigma}^2/S_{xx}\right)^{-1/2}(\hat{\beta}_1 - \beta_1) \sim \mathcal{T}(n-2)$ (Student's t distribution)

History of Student's t distribution

WIKIPEDIA:

> In the English-language literature the distribution takes its name from William Sealy Gosset's 1908 paper in Biometrika under the pseudonym "Student". Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples ... where sample sizes might be as few as 3 ... Guinness did not want their competitors to know that they were using the t-test to determine the quality of raw material.

Finding the sampling distribution of $\hat{\beta}_1$: Summary

1. Small sample size: Find the exact distribution of $\hat{\beta}_1$ by assuming normality for $\{\epsilon_i\}_{i=1}^n$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} \sim \mathcal{T}(n-2)$$

2. Large sample size: Find the asymptotic distribution of $\hat{\beta}_1$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} \to \mathcal{N}(0,1).$$

THEORY $\neq$ REALITY

Lesser of two evils

Normality assumption v.s. Infinite samples

> Only two things are infinite, the universe and human stupidity, and I'm not sure about the former.
>
> Albert Einstein

> If you hear a "prominent" economist using the word "equilibrium", or "normal distribution", do not argue with him; just ignore him, or try to put a rat down his shirt.
>
> Nassim Nicholas Taleb

Objections?

> Evil is Evil. Lesser, greater, middling ... it's all the same. Proportions are negotiated, boundaries blurred. I'm not a pious hermit, I haven't done only good in my life. But if I'm to choose between one evil and another, then I prefer not to choose at all.
>
> Geralt of Rivia[11]

Well, we don't have the luxury of choosing as data hunters...

[11] Actually, Andrzej Sapkowski

Distribution of sum of random variables

$\hat{\beta}_1 - \beta_1 = \sum_{i=1}^{n} c_{1,i}\epsilon_i$: $\hat{\beta}_1 - \beta_1$ is the weighted sum of $\{\epsilon_i\}_{i=1}^{n}$ with weights $\{c_{1,i}\}_{i=1}^{n}$

1. Find the exact distribution of $\hat{\beta}_1$ by assuming some distribution for $\{\epsilon_i\}_{i=1}^{n}$
2. Find the asymptotic distribution of $\hat{\beta}_1$ as the sample size n approaches infinity (i.e., assuming that the sample size is **sufficiently large**)

Theory $\neq$ Reality

1. Normality assumptions are also unrealistic
2. Asymptotic distributions are unrealistic

Sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$

The past 12 slides are all about sampling distributions

- Two sampling distributions (under different assumptions) for $(\hat{\beta}_1 - \beta_1)/(\hat{\sigma}/S_{xx}^{1/2})$: $\mathcal{T}(n-2)$ and $\mathcal{N}(0,1)$
- $\mathcal{T}(n-2) \to \mathcal{N}(0,1)$ as n grows

But why doesn't `R` show the sampling distributions in `summary()`?

> Before I came here I was confused about this subject. Having listened to your lecture I am still confused. But on a higher level.
>
> Enrico Fermi

Sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$

- Two sampling distributions (under different assumptions) for $(\hat{\beta}_1 - \beta_1)/(\hat{\sigma}/S_{xx}^{1/2})$: $\mathcal{T}(n-2)$ and $\mathcal{N}(0,1)$
- $\mathcal{T}(n-2) \to \mathcal{N}(0,1)$ as n grows

These distributions are similar.

We report **summary measures** of the sampling distributions such as, point estimates (learned), standard errors (learned), or confidence intervals (new!).

Confidence interval

For a chosen confidence level $1-\alpha \in (0,1)$, we can defined the $100(1-\alpha)\%$ confidence interval of $\hat{\beta}_1$ as $(\hat{l}, \hat{u})$, where

$$\hat{l} = \hat{\beta}_1 - q(1-\alpha/2; G) \times \hat{\sigma}/S_{xx}^{1/2}, \ \hat{u} = \hat{\beta}_1 - q(\alpha/2; G) \times \hat{\sigma}/S_{xx}^{1/2}.$$

$q(\alpha/2)$ and $q(1-\alpha/2)$ are the $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of the distribution of $(\hat{\beta}_1 - \beta_1)/(\hat{\sigma}/S_{xx}^{1/2})$, denoted as G, e.g.,

- Under the normality assumption: G is $\mathcal{T}(n-2)$
- When the sample size is sufficiently large, G can be
  - $\mathcal{N}(0,1)$ (CLT)
  - the bootstrap distribution (bootstrap)

Quantiles

Definition: The α-quantile of a distribution with cumulative distribution function G (or p.d.f. g) is the value b satisfies that

$$G(b) = \alpha \quad \text{or} \quad \int_{-\infty}^{b} g(u)\mathrm{d}u = \alpha.$$

Denoting the quantile function as $q(\cdot; G)$, we have $q(\alpha; G) = b$.

Confidence interval: Interpretation

INTERPRETATION: $(\hat{l}, \hat{u})$ cover the true parameter β_1 $100(1-\alpha)\%$ of the time.

This is because $\mathbb{P}(\hat{l} < \beta_1 < \hat{u}) = 1 - \alpha$ and $(\hat{l}, \hat{u})$ is random.

Do one of the following

- Prove this statement using rigorous mathematical derivation
- Verify this using simulation in `R`

Confidence interval: Cutoffs

For a chosen confidence level $1-\alpha \in (0,1)$, we can defined the $100(1-\alpha)\%$ confidence interval of $\hat{\beta}_1$ as $(\hat{l}, \hat{u})$, where

$$\hat{l} = \hat{\beta}_1 - q(1-\alpha/2; G) \times \hat{\sigma}/S_{xx}^{1/2}, \ \hat{u} = \hat{\beta}_1 - q(\alpha/2; G) \times \hat{\sigma}/S_{xx}^{1/2}.$$

$q(\alpha/2)$ and $q(1-\alpha/2)$ are the $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of the distribution of $(\hat{\beta}_1 - \beta_1)/(\hat{\sigma}/S_{xx}^{1/2})$, denoted as G.

Q: Why choose $\alpha/2$ and $(1-\alpha/2)$?

Explore other choices of cutoff in `R`

Estimation

Estimating the mean response at x^*: $\mathbb{E}[y \mid x^*]$.

- Point estimate $\hat{y}^* = x^* \hat{\beta}_1 + \hat{\beta}_0$
- Standard error

$$\widehat{\mathrm{SE}}(\hat{y}^*) = \hat{\sigma} \left\{ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}^{1/2}$$

Prediction

Predicting a new response y^* at x^*

- Point estimate $x^*\hat{\beta}_1 + \hat{\beta}_0$, since $y^* = x^*\beta_1 + \beta_0 + \epsilon^*$
- Standard error of $\tilde{y}^*$

$$\mathrm{SE}(\tilde{y}^*) = \left\{\mathbb{E}[\tilde{y}^* - y^*]^2\right\}^{1/2}$$

$$\widehat{\mathrm{SE}}(\tilde{y}^*) = \left\{\hat{\sigma}^2 + \hat{\sigma}^2\frac{1}{n} + \hat{\sigma}^2\frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right\}^{1/2}$$

Prediction v.s. Estimation

Prediction

- Point estimate $\tilde{y}^* = x^* \hat{\beta}_1 + \hat{\beta}_0$
- Standard error of $\tilde{y}^*$

$$\widehat{\mathrm{SE}}(\tilde{y}^*) = \left\{ \hat{\sigma}^2 + \hat{\sigma}^2 \frac{1}{n} + \hat{\sigma}^2 \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}^{1/2}$$

Estimation

- Point estimate $\hat{y}^* = x^* \hat{\beta}_1 + \hat{\beta}_0$
- But the standard error is different

$$\widehat{\mathrm{SE}}(\hat{y}^*) = \hat{\sigma} \left\{ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}^{1/2}$$

Prediction interval

- Point estimate $x^*\hat{\beta}_1 + \hat{\beta}_0$, since $y^* = x^*\beta_1 + \beta_0 + \epsilon^*$
- Standard error of $\tilde{y}^*$

$$\mathrm{SE}(\tilde{y}^*) = \left\{\mathbb{E}[\tilde{y}^* - y^*]^2\right\}^{1/2}$$

$$\widehat{\mathrm{SE}}(\tilde{y}^*) = \left\{\hat{\sigma}^2 + \hat{\sigma}^2\frac{1}{n} + \hat{\sigma}^2\frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right\}^{1/2}$$

- Requires the distribution of ϵ^*!

Q: How about confidence intervals when estimating $\mathbb{E}[y^*]$?

Verify the coverage of the prediction interval and compare it with the confidence interval in `R`

Simultaneous confidence intervals

What if we want to construct the confidence intervals for the mean responses $\{\mathbb{E}[y_j^*] : j = 1, \ldots, m\}$ at $\{x_j^* : j = 1, \ldots, m\}$?

- Bonferroni correction: intersect of **corrected** confidence intervals
- Working-Hotelling procedure: a confidence band for **all** $x^* \in \mathbb{R}$
- Multivariate confidence region: a confidence region on the **joint** distribution of $\{\hat{y}_j^* : j = 1, \ldots, m\}$

Bonferroni correction

For $j = 1, \ldots, m$, let $(\hat{l}_{j,\alpha/m}, \hat{u}_{j,\alpha/m})$ be the $100(1-\alpha/m)\%$ confidence interval for $\hat{y}_j^*$.
The Bonferroni corrected (simultaneous) confidence interval for $\{\hat{y}_j^* : j = 1, \ldots, m\}$ is

$$\prod_{j=1}^{m} (\hat{l}_{j,\alpha/m}, \hat{u}_{j,\alpha/m}) = (\hat{l}_{1,\alpha/m}, \hat{u}_{1,\alpha/m}) \times \cdots \times (\hat{l}_{m,\alpha/m}, \hat{u}_{m,\alpha/m})^{12}$$

Idea: $\mathbb{P}(A \cap B) \geq 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c)$

$$\mathbb{P}\big(\cap_{j=1}^{m} \{\mathbb{E}[y_j^*] \in (\hat{l}_{j,\alpha/m}, \hat{u}_{j,\alpha/m})\}\big)$$

$$\geq 1 - \sum_{j=1}^{m} \mathbb{P}\big(\{\mathbb{E}[y_j^*] \in (\hat{l}_{j,\alpha/m}, \hat{u}_{j,\alpha/m})\}\big) \geq 1 - \alpha$$

The confidence intervals are too wide when m is large!

[12] Here $\times$ is the Cartesian product.

Working-Hotelling(-Scheffe) procedure

For any $x^* \in \mathbb{R}$, a confidence band $(\hat{l}(x^*), \hat{u}(x^*))$for $\mathbb{E}[y^*]$ satisfies that

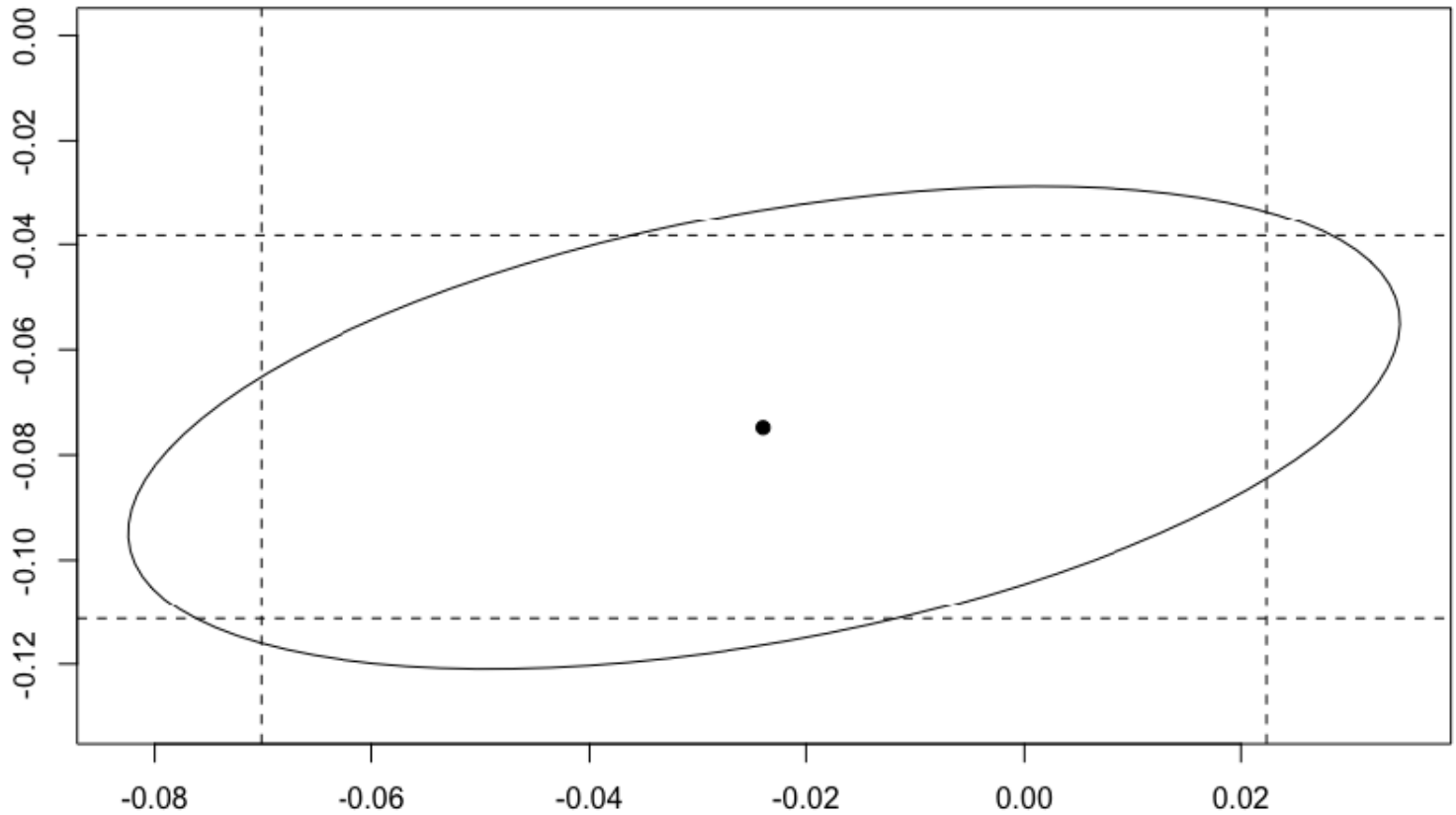
$$\begin{aligned}\hat{l}(x^*) =& x^*\hat{\beta}_1 + \hat{\beta}_0 - \{2q(\alpha; G)\}^{1/2}\widehat{\mathrm{SE}}(\hat{y}^*) \\ \text{and } \hat{u}(x^*) =& x^*\hat{\beta}_1 + \hat{\beta}_0 + \{2q(\alpha; G)\}^{1/2}\widehat{\mathrm{SE}}(\hat{y}^*),\end{aligned}$$

where G is an F-distribution with parameters 2 and $n-2$ (under the normality assumption) and

$$\widehat{\mathrm{SE}}(\hat{y}^*) = \hat{\sigma}\left\{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right\}^{1/2}.$$

Intuition: All $\hat{y}^*$ depend on the same noise $\{\epsilon_i\}_{i=1}^n$.

Multivariate confidence region



The probability that the true values fall into the confidence region is larger than $100(1-\alpha)\%$.

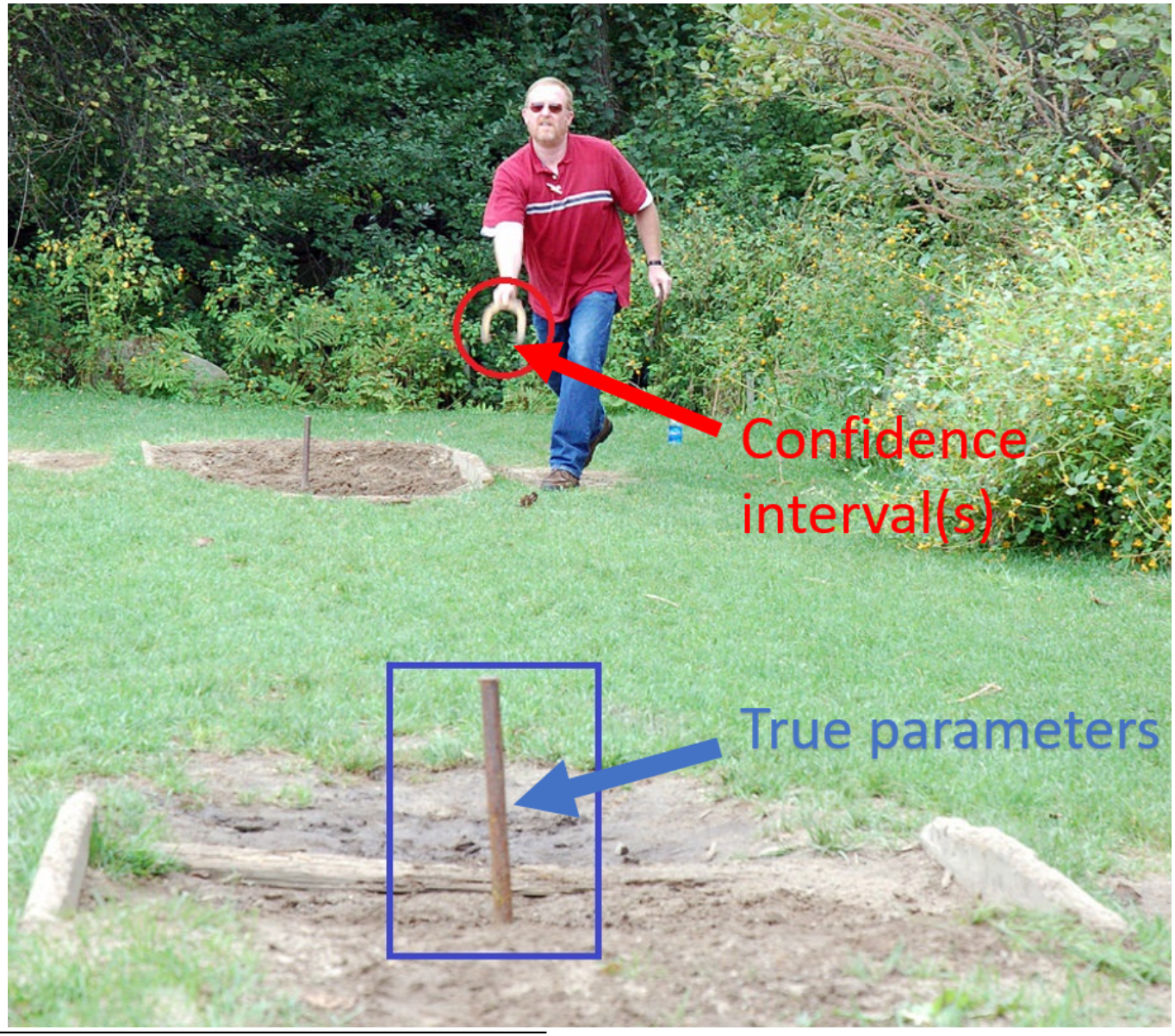
This statement is wrong, but what is the correct statement?

Simultaneous confidence intervals: Summary

- Bonferroni correction: intersect of **corrected** confidence intervals
- Working-Hotelling procedure: a confidence band for **all** $x^* \in \mathbb{R}$
- Multivariate confidence region: a confidence region on the **joint** distribution of $\{\hat{y}_j^* : j = 1, \ldots, m\}$

Construct and compare simultaneous confidence intervals in R

Confidence intervals: Summary[13]



[13] Source

Hypothesis testing

Null hypothesis:

$$H_0 : \text{There is no association between } x \text{ and } y$$

Alternative hypothesis:

$$H_1 : \text{There is some association between } x \text{ and } y.$$

The hypotheses are equivalent to (why?)

$$H_0 : \beta_1 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0$$

Statistical reasoning

We want to test the hypothesis given data

$$H_0 : \beta_1 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0$$

- In real life, we reject a hypothesis if the null is given our observations.
- The null hypothesis is if the observed data is very extreme under the null hypothesis.

Q: How to quantify "extreme": the probability of observing an even more "extreme" value

Test statistics

Under the null hypothesis

$$H_0 : \beta_1 = 0$$

1. Small sample size: t-test (normality assumption):

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/S_{xx}^{1/2}} \sim \mathcal{T}(n-2)$$

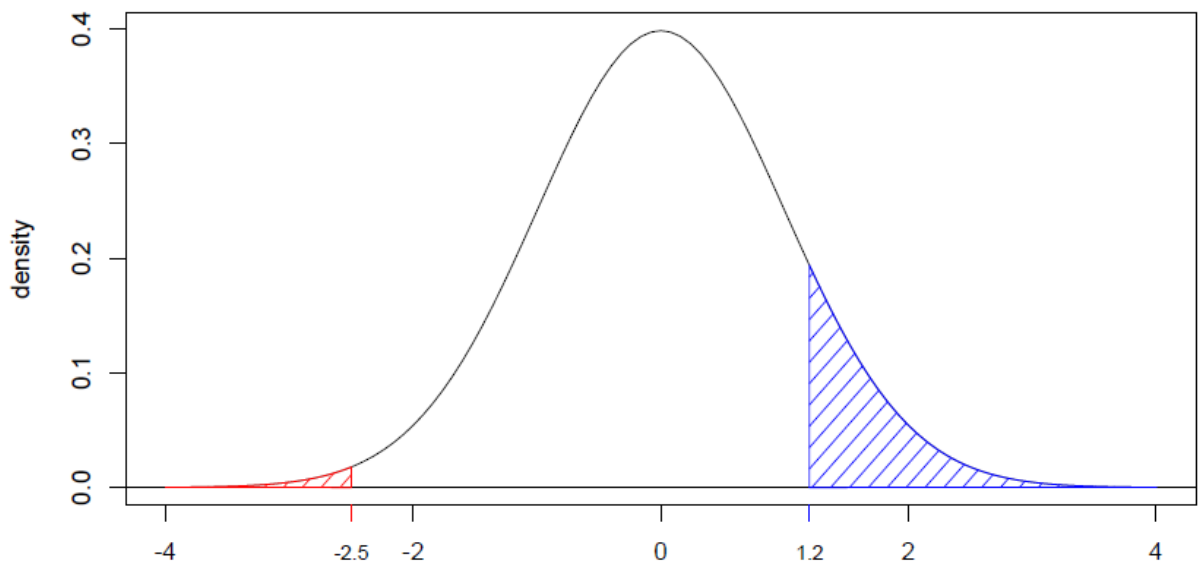
2. Large sample size: z-test (asymptotic distribution)

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/S_{xx}^{1/2}} \to \mathcal{N}(0,1).$$

Bootstrap can also approximate the null distribution, but is a bit complicated...

P-value

- Definition: the probability of observing a **more extreme** test statistics T than the test statistics t calculated on the current sample, i.e.,
  p-value $= 2\min(\mathbb{P}(T < t), \mathbb{P}(T > t))$.
- A summary measure of the plausibility of the null hypothesis.



Significance level

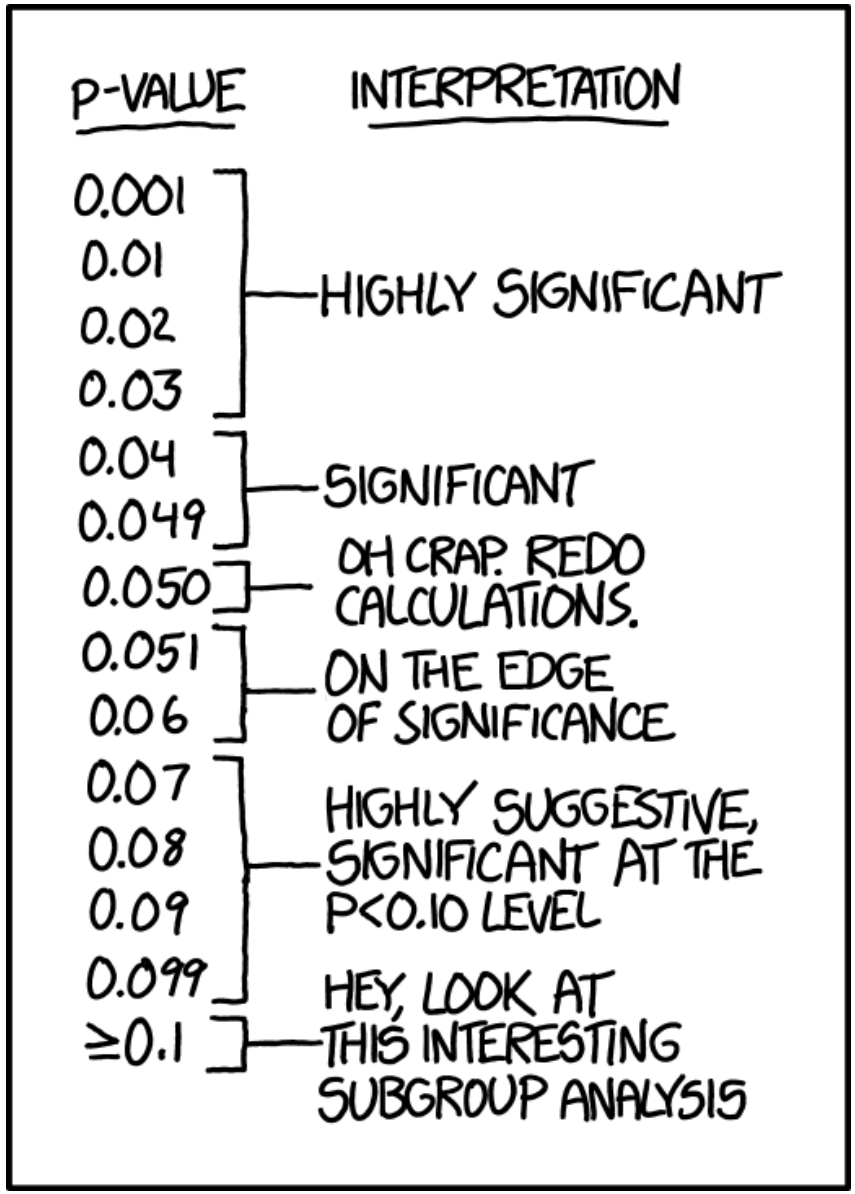
$$H_0 : \beta_1 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0$$

We reject the null hypothesis $H_0 : \beta_1 = 0$ if the p-value is less than α, a chosen significance level.
Typical choices of α include 0.05, 0.025, 0.01, etc.

> We reject the null hypothesis at the significance level of α.

INTERPRETATION: At the significance level of α, the data is unlikely to be from a population distribution where the null hypothesis holds.

P-values[14]



14

Alternative rejection criteria

We reject the null hypothesis $H_0 : \beta_1 = 0$ if the p-value is less than α, or

- t falls into the rejection region: $(-\infty, q(\alpha/2; G)] \cup [q(1-\alpha/2; G), \infty)$.
- $|t|$ is larger than the critical value $|q(1-\alpha/2; G)|$ (if G is symmetric)

Errors and power

$$H_0 : \beta_1 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0$$

	Fail to reject H_0	Reject H_0
H_0 true	correct	Type I error, (significance level)[15]
H_1 true	Type II error	correct, (power)

In R,

- verify the equivalence between the significance level and type I error **rate**
- write simulation to learn the power against $H_1 : \beta_1 = a$ for $a = 0.1, 0.2, 0.3$

[15] See xkcd

Hypothesis testing: permutation test

Consider the hypotheses

$$H_0 : \beta_1 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0$$

- Under the null, there is no associations between y_i and x_i
- PERMUTATION test: permuting the index of $\{x_i : i = 1, \ldots, n\}$
  - Each permutation gives one test statistics under the null
  - Multiple permutations generate a distribution of the test statistics under the null
  - p-value: the frequency of permuted test statistics being more extreme than the test statistics on the original samples
- Permutation: sampling without replacement; Bootstrap: sampling with replacement

Implement the permutation test in R

Other hypothesis

$$H_0 : \beta_1 = a \text{ v.s. } H_1 : \beta_1 \neq a$$

1. Small sample size:

$$t = \frac{\hat{\beta}_1 - a}{\hat{\sigma}/S_{xx}^{1/2}} \sim \mathcal{T}(n-2)$$

2. Large sample size:

$$t = \frac{\hat{\beta}_1 - a}{\hat{\sigma}/S_{xx}^{1/2}} \to \mathcal{N}(0,1).$$

Other Hypothesis: One-sided Test

$$H_0 : \beta_1 \leq a \text{ v.s. } H_1 : \beta_1 > a$$

1. Small sample size:

$$t = \frac{\hat{\beta}_1 - a}{\hat{\sigma}/S_{xx}^{1/2}} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} \sim \mathcal{T}(n-2)$$

2. Large sample size:

$$t = \frac{\hat{\beta}_1 - a}{\hat{\sigma}/S_{xx}^{1/2}} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/S_{xx}^{1/2}} \to \mathcal{N}(0,1).$$

The p-value is $\mathbb{P}(T \geq t)$ instead of $\mathbb{P}(|T| \geq |t|)$.

Testing simultaneous hypothesis

$$H_0 : \beta_1 = \beta_0 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0 \text{ or } \beta_0 \neq 0$$

- The F-statistics:

$$t = \frac{\left[\sum_{i=1}^n y_i^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2\right]/2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2)}.$$

- Under the normality assumption, $t \sim F(2, n-2)$[16] under H_0

$$t = \frac{[\mathcal{L}(0,0) - \mathcal{L}(\hat{\beta}_0, \hat{\beta}_1)]/2}{\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1)/(n-2)}.$$

Verify the claim above and

simulate the case without the normality assumption?

[16] See F-distribution, χ^2-distribution

F-statistics

Consider the null hypothesis

$$H_0 : \beta_1 = 0 \text{ v.s. } H_1 : \beta_1 \neq 0$$

We can also define the F-statistics

$$t_F = \frac{[\mathcal{L}(\bar{y}, 0) - \mathcal{L}(\hat{\beta}_0, \hat{\beta}_1)]/1}{\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1)/(n-2)}.$$

Then $t_F \sim F(1, n-2)$ under H_0.
Recall that

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/S_{xx}^{1/2}}$$

Then $t_t^2 = t_F$,
because $\mathcal{L}(\bar{y}, 0) - \mathcal{L}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \hat{\beta}_1^2$

Multiple testings

Suppose that we want to test two separate hypotheses:

$$H_{0,1} : \beta_1 = 0 \text{ v.s. } H_{1,1} : \beta_1 \neq 0$$

$$H_{0,0} : \beta_0 = 0 \text{ v.s. } H_{1,0} : \beta_0 \neq 0$$

Recall: The type I error is probability of falsely rejecting the null hypothesis under the null.

- Family-wise error rate (FWER): the overall rate of type I error all tests $\mathbb{P}(\text{making at least one false rejection})$.
  - Use Bonferroni correction (or methods)
- False discovery rate (FDR):
  $\{\text{number of false rejections}\}/\{\text{number of total rejections}\}$
  - Use Benjamini–Hochberg procedure.

False positive & false negatives

Recall the table in lecture notes

	Fail to reject H_{null}	Reject H_{null}
H_{null} true	correct	Type I error
H_{alt} true	Type II error	correct

We will give them new names

	Fail to reject H_{null}	Reject H_{null}
H_{null} true	true negative	false positive/discovery
H_{alt} true	false negative	true positive

False discovery rate

Suppose that we have a lot of (null) hypotheses $H_{1,\text{null}}, \ldots, H_{m,\text{null}} \cdots$

- Family-wise error rate

$$\text{FWER} \equiv \mathbb{P}\left(\text{at least one hypothesis is falsely rejected}\right)$$

- False discovery rate

$$\text{FDR} \equiv \mathbb{E}\left[\frac{\#\text{false discoveries}}{\#\text{false discoveries} + \#\text{true discoveries}}\right]$$

- You may define the false negative rate in a similar manner

Example: Coronavirus detection test

False positive: a healthy subject that tests positive

False negative: an infected subject that tests negative

- Usually a trade-off between false positive rate and false negative rate
- Cost and benefit
  - each test costs $\$200$ (UW medicine[17])
  - each false negative case would cost ...?

[17] link

Controlling the FDR

The Benjamini–Hochberg procedure (Benjamini and Hochberg 2000)

Suppose that there are a set of null hypotheses $H_1, \ldots, H_m$

1. Choose a level α at which to control the FDR
2. Calculate the p-values for the hypotheses $H_1, \ldots, H_m$, denoted as $p_1, \ldots, p_m$
3. Order the p-values as $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$
4. Find the largest k such that $p_{(k)} \leq k\alpha/m$
5. Reject the null hypotheses that correspond to $p_{(1)}, \ldots, p_{(k)}$

In R

Use the function `p.adjust()`

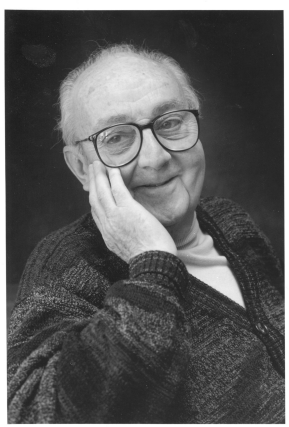
Then implement your own version of B-H procedure

Recall: Simple linear regression model

The **simple linear regression** model takes the form

$$y_i = x_i\beta_1 + \beta_0 + \epsilon_i, \quad i = 1, 2, \ldots, n$$

where $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$.



ALL MODELS ARE WRONG, BUT SOME ARE USEFUL.

George E. P. Box

Assumptions on simple linear regression model[19]

The conventional assumptions

Linearity	$\mathbb{E}[y \mid x]$ is linearly related to x
Independence	$\epsilon_1, \ldots, \epsilon_n$ are (mutually) independent[18]
Normality	Distribution of $y_i \mid x$ is normal
Equal variance	$\mathrm{var}(y_i \mid x) = \mathrm{var}(\epsilon_i)$ does not depend on x_i

LINE $\rightarrow$ LIE

Q1: Which assumptions are necessary?

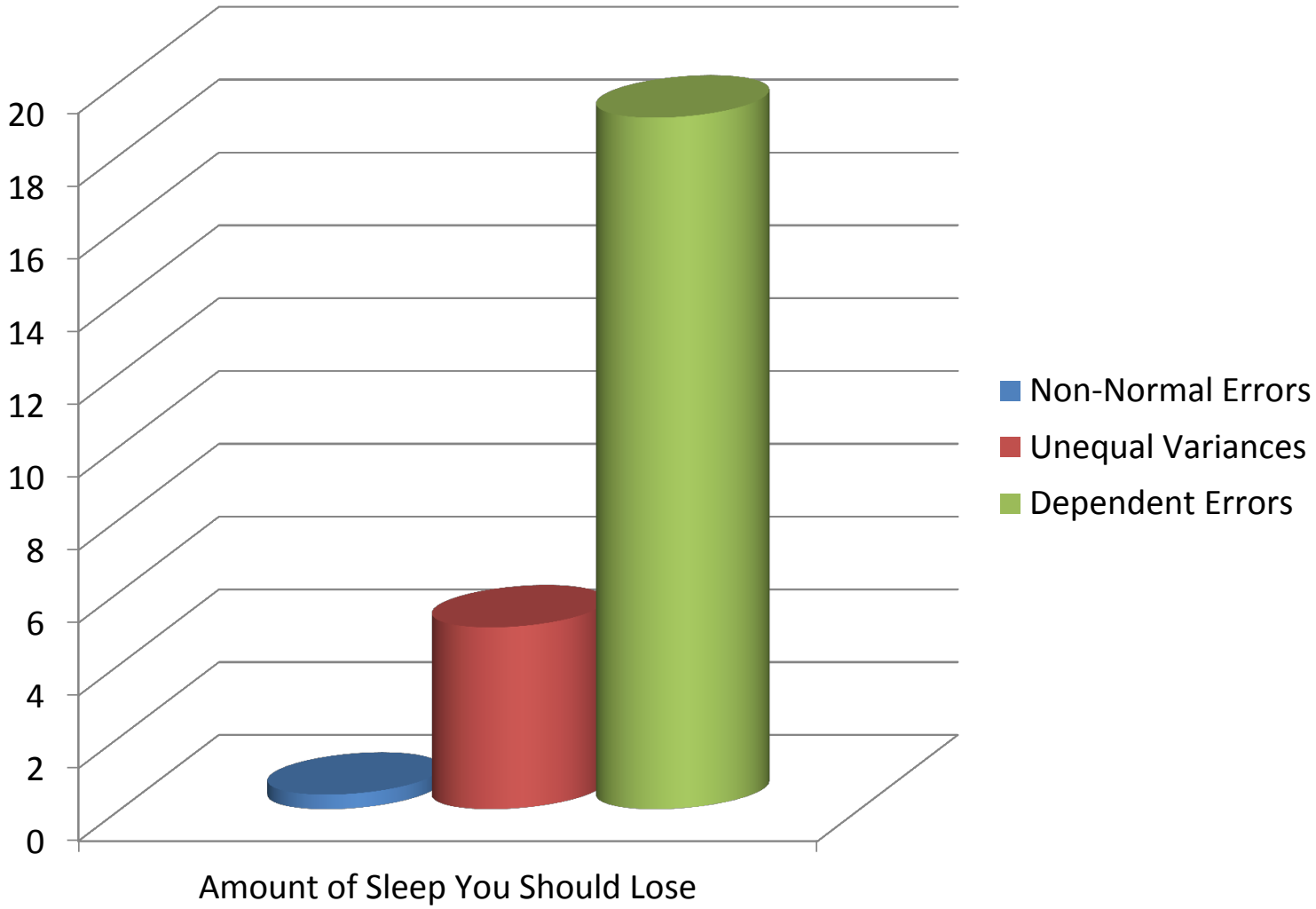
Q2: How to verify these assumptions?

Q3: What are the remedial measures?

[18] ... and are independent of $x_1, \ldots, x_n$, although we treat $x_1, \ldots, x_n$ as fixed in this class.

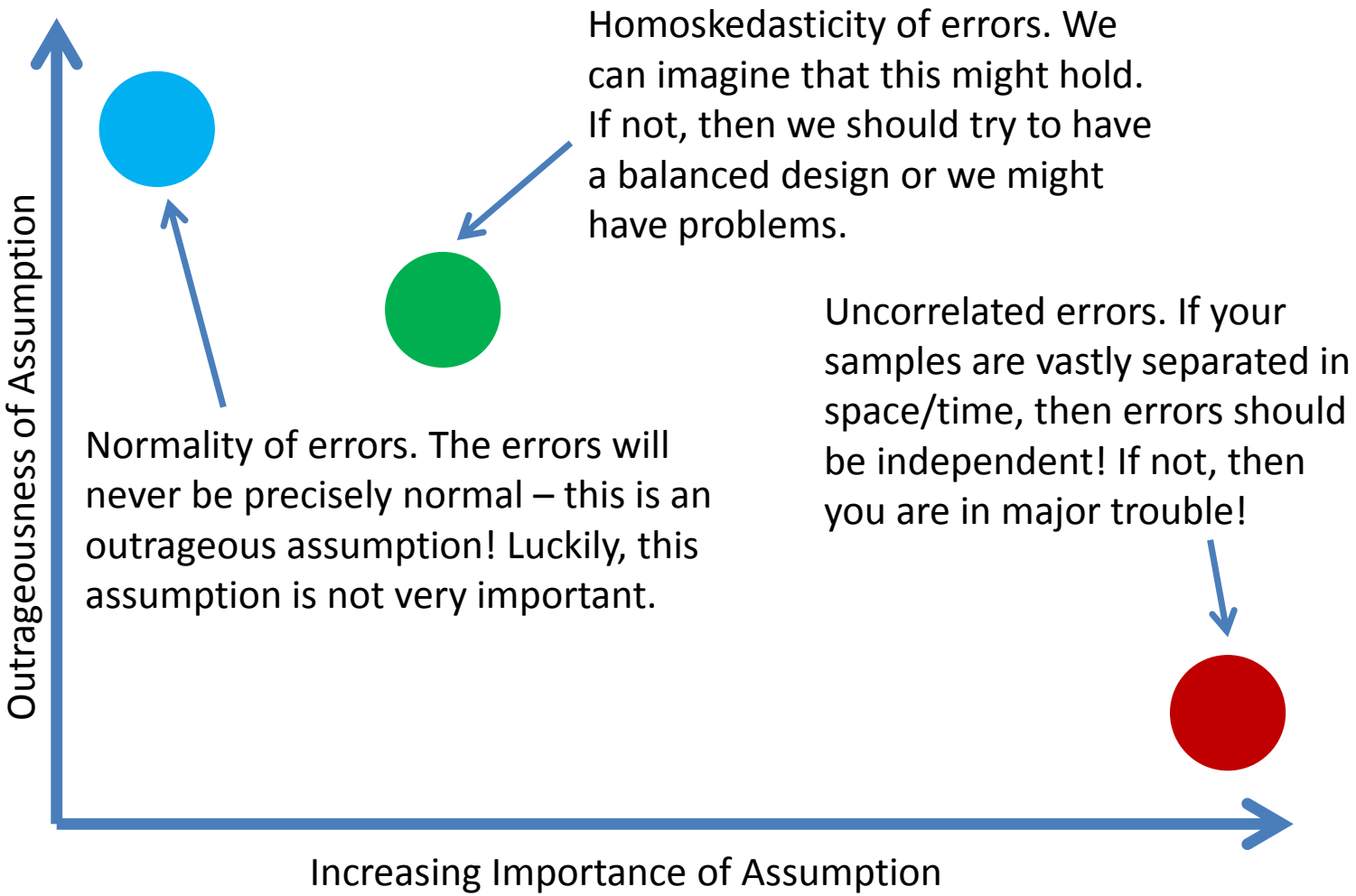
[19] The term "linear" refers to the fact that the mean is a linear function of the **unknown parameters** β_0, β_1.

Overview[20]



[20] Figure courtesy of Prof. Witten at University of Washington.

Overview (cont.)[21]



[21] Figure courtesy of Prof. Witten at University of Washington.

Evaluate the plausibility of assumptions

- Common sense
- Scientific knowledge
- Analysis of **residuals**

  Recall: residuals are defined as

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \qquad i = 1, \ldots, n.$$

Numerical examples

> All happy families are alike; each unhappy family is unhappy in its own way.
>
> Leo Tolstoy

Simulate data that violate the aforementioned assumptions Run model diagnostics on the synthetic data (you can read ahead for diagnostic tools[22])

[22] There are a lot of other diagnostic tools, including formal hypothesis testings, which are not covered in this class

Model diagnostics with residuals: Summary

- **L**inearity
  - Necessities: ???
  - Diagnostics: scatter plots of e_i v.s. x_i, ...
  - Remedy: transformation of x_i
    (e.g, $y_i \sim \beta_0 + \beta_1 x_i + \beta_2 \exp(x_i)$)
    Q: Is the model still linear?

- **I**ndependence
  - Necessities: ???
  - Diagnostics: scatter plots of e_i v.s. x_i, ...
  - Remedy: generalized least squares (if dependence structure is known)

Model diagnostics with residuals: Summary (cont.)

- **N**ormality
  - Necessities: ???
  - Diagnostics: Q-Q plots of e_i, ...
  - Remedy: CLT or bootstrap

- **E**qual variance
  - Necessities: ???
  - Diagnostics: scatter plots of e_i v.s. x_i, ...
  - Remedy: generalized least squares, transforming y

Model diagnostics with residuals: Summary (cont.)

- Influential observations
  - **Not** a violation of assumptions
  - Diagnostics: scatter plots of e_i v.s. x_i
  - Remedy: investigate these samples

- Outliers
  - **Maybe** a violation of assumptions
    Imply anomalies in these samples
  - Diagnostics: scatter plots of e_i v.s. x_i
  - Remedy: investigate these samples

You will start to see a lot of repetitive materials from this point

...because multiple linear regression is not that different from simple linear regression!

Multiple linear regression model

The **multiple linear regression** model takes the form

$$y = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p + \epsilon,$$

where

- $y \in \mathbb{R}$ is the real-valued response
- $x_j \in \mathbb{R}$ is the jth covariate
- β_0 is the intercept term
- β_j is the regression slope for the jth covariate
- $\epsilon \in \mathbb{R}$ is the error term with $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$

Note: The term "linear" refers to the fact that the mean is a linear function of the **unknown parameters** $\beta_0, \ldots, \beta_p$.

With n observations

With n observations of y and $x_1, \ldots, x_p$, the complete model becomes

$$\begin{aligned} y_1 &= \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p + \epsilon_1 \\ y_2 &= \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p + \epsilon_n, \end{aligned}$$

where the error terms are assumed to have the following properties

- $\mathbb{E}[\epsilon_i] = 0$
- $\mathrm{var}(\epsilon_i) = \sigma^2$ (constant for all i)
- $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ for $j \neq i$

Note: Sometimes stronger assumptions are imposed, such as ϵ's are i.i.d. with mean 0 and variance σ^2.

Interpretation of multiple linear regression

In a multiple linear regression model $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p + \epsilon$, β_1 is the expected mean difference in y per unit difference in x_1 if $x_2, \ldots, x_p$ is held constant (or adjusted/controlling for $x_2, \ldots, x_p$)

Example: Suppose that y is the systolic blood pressure of newborns, x_1 is days of age, and x_2 is the weight at birth in ounces. We say that

- β_1: We estimate that two groups of newborns with the same age and who differ by one ounce at birth will have systolic blood pressure that differs on average by 0.13 mm Hg (95% CI: 0.05, 0.20).
- β_2: We estimate that two groups of newborns with the same weight at birth and who differ by one day of age will have systolic blood pressure that differs on average by 5.89 mm Hg (95% CI: 4.42, 7.36).

Interpretation of multiple linear regression

In a multiple linear regression model $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p + \epsilon$, β_1 is the expected mean difference in y per unit difference in x_1 if $x_2, \ldots, x_p$ is held constant (or adjusted/controlling for $x_2, \ldots, x_p$)

Note: Interpretation of parameters might not make sense in the multiple linear regression!

Interpretation of multiple linear regression

In $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p + \epsilon$, what if

- $x_3 = x_1 \times x_2$?
  Effect of x_1 on y will differ depends on the value of x_2.
  Example: When comparing two groups of newborns that differ by one day of age and with the same birthweight, the difference in systolic blood pressure depends on the babies' birthweight, with the difference in mean systolic blood pressure decreasing by 0.13 mm/Hg for each ounce difference in birth.

- $x_2 = x_1^3$?
  x_1 is constant if x_2 is held constant.
  Interpret all the terms that depend on x_1!

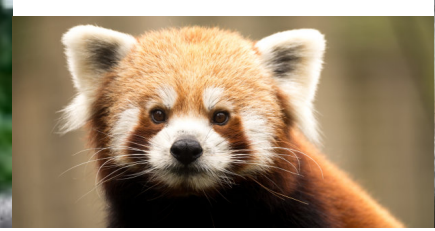
- $x_1, \ldots, x_p$ are dummy variables for a categorical variable z that has $p + 1$ categories?
  $x_1 = 1$ means $x_2 = \cdots = x_p = 0$.
  How will you interpret β_1?

Multiple linear regression model: Categorical covariates

Consider the question in Homework #1, we code x_i as 0 or 1 to distinguish between ducks and pandas

- What if we found out that there are actually red pandas and raccoons in the data set?
- $x_i = 0, 1, 2, 3$ for ducks, pandas, red pandas, or raccoons?
- $x_{i1} = 1$ for a panda, $x_{i2} = 1$ for a red panda, $x_{i3} = 1$ for a raccoon, and zero otherwise!






Multiple linear regression model: Categorical covariates (cont.)

Create $K-1$ dummy variables for a categorical variable with K categories

Also known as the ANalysis Of VAriance, a.k.a., ANOVA

Multiple linear regression model: Polynomial regression[23]

Consider a true (and unknown!) model where y is non-linear in x

$$y = (x-3)^4 + \epsilon \iff y = x^4 - 12x^3 + 54x^2 - 108x + 81 + \epsilon$$

Suppose that we have n observations of x and y, can we learn the above model using linear regression?

[23] Check out Wolfram Alpha

Real data

Fit multiple linear regression on

- advertising data
- flut shot data
- Project STAR

and interpret the estimated coefficients

Building a linear model

Suppose that you are interested in studying the relationship between y and x_1, and you have the resources to collect data (via experiments or surveys). How will you build a linear model?

$$y = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p + \epsilon$$

Consider the following scenarios

- y is the body weight and x_1 is the length of sleep per day
- y is the lung function (measured by forced expiratory volume, or FEV) and x_1 is a dummy variable for smoking
- y is the occurrence of an heart attack and x_1 is a dummy variable for depression

What other covariates do you want to include?

Classification of Variables

- Variable of interest (or exposure, treatment, etc.)
- Response variable (or outcome)
- Confounder
- Effect modifier
- Precision variable
- Instrument

Confounding

- Confounding is an effect of some uncontrolled variable on the response variable that hinders interpretation of the relationship between the response and the predictor variable of interest
- Confounding describes real or imagined effects that distort the relationship one wishes to observe between the predictor variable and response
- More of a problem for observational studies
  - in contrast to a designed experiment

Examples: effects of smoking on lung function (measured by forced expiratory volume, FEV) may be confounded by age

Controlling for confounding

- Implicitly with appropriate study designs
- Explicitly by measuring it and including it in the model

Controlling for confounding: Design

- Match the observations that are similar in terms of confounding variables (confounders), e.g., comparing the FEV between smokers and non-smokers of the same age
  - Relatively easy to implement
  - Infeasible when there are too many confounders
- Conduct a randomized experiment, e.g., randomly assign participants to the smoking group or the non-smoking group
  - Destroy all confounding possibilities (grant causality)
  - Infeasible in many cases

Controlling for confounding: Model-based

Using knowledge in this class, we can

- Fit the unadjusted model

$$y = \beta_0 + x_1\beta_1 + \epsilon$$

- Fit the adjusted model

$$y = \beta_0' + x_1\beta_1' + x_2\beta_2' + \epsilon$$

- Compare the fitted values of β_1' and β_1
  - Eyeballing
  - Hypothesis testing

You can also use other advanced statistical methods to control for unmeasured confounders, e.g., propensity score, instrumental variable

Effect modifier

- Variable that modifies the effect (or association) of the variable of interest on the response
- Modeling approaches
  - Stratify analysis (for categorical variables)
  - Multiple linear regression with an interaction term

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3 + \epsilon$$

Precision variable

- Variable that only affects the response variable
- Improve the precision of the model fits if included

Linear model in matrix notation

$$
\begin{aligned}
y_1 &= \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p + \epsilon_1 \\
y_2 &= \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p + \epsilon_n,
\end{aligned}
$$

where the error terms are assumed to have the following properties $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$, and $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ for $j \neq i$.

$$
\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0} \text{ and } \mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}
$$

Linear model in matrix notation, I

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

We write

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

- $\boldsymbol{y}$ is the $n \times 1$ response column vector
- $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix
- $\boldsymbol{\beta}$ is the $n \times (p+1)$ design matrix
- $\boldsymbol{\epsilon}$ is the random error $n \times 1$ column vector

Linear model in matrix notation, II

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Recall: The error terms are assumed to have the following properties $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$, and $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ for $j \neq i$. Alternatively, you can write[24]

- $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$
- $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$

[24] Again, we consider $\boldsymbol{X}$ to be fixed in this course

Some notes on linear models

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

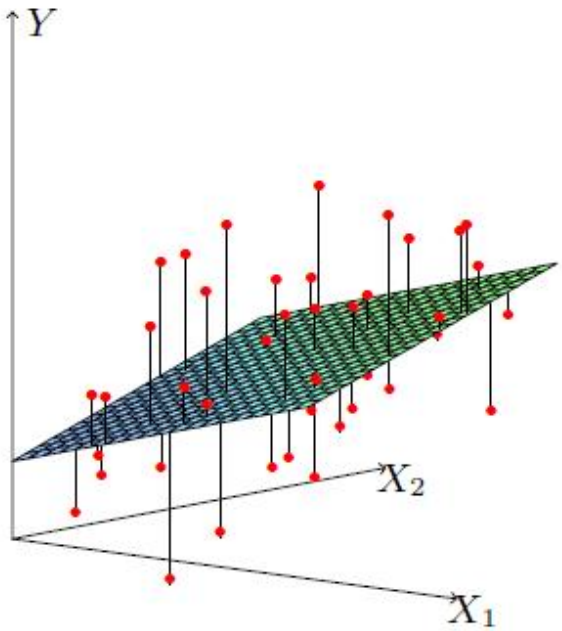
1. Usually the first column of $\boldsymbol{X}$ is the vector of ones, corresponding to the intercept β_0
2. The jth column of $\boldsymbol{X}$, denoted as $\boldsymbol{X}_j$, is the jth predictor variable for the n observations
3. $\boldsymbol{\epsilon}$ is the random part of the model ($\boldsymbol{y}$ is random because $\boldsymbol{\epsilon}$ is random)
4. $\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{\beta}] + \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{X}\boldsymbol{\beta}$, i.e., $\mathbb{E}[\boldsymbol{y}]$ is a linear combination of $\{\boldsymbol{x}_j : j = 1, \ldots, p+1\}$

Least squares estimator for multiple linear regression

Estimates $\beta_0, \beta_1, \ldots, \beta_p$ are **least squares estimators** of $\beta_0, \beta_1, \ldots, \beta_p$ if they minimize

$$\mathcal{L}(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

Graphical Interpretation: [Element of Stat Learning: Hastie et al.]



Least squares estimator for multiple linear regression

Solve the least squares problem in matrix notation:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

where $\|\cdot\|_2$ is the ℓ_2-norm $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

Claim: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbf{y}$.

Do one of the following

- Prove this claim with rigorous derivation
- Solve the optimization problem, and verify this claim in `R`

Simple linear regression as a special case

Recall: In simple linear regression, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1.$$

Properties of the LSE

Projection

Residuals

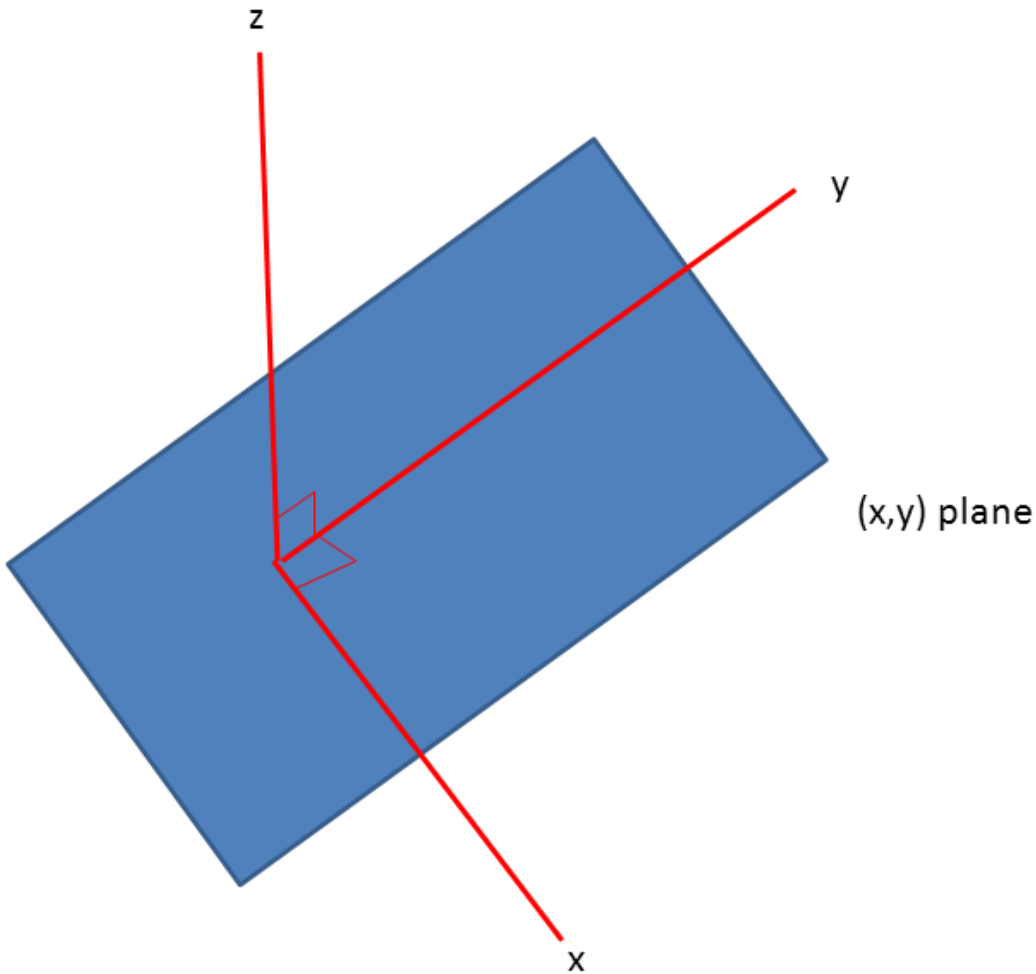
Underfitting and Overfitting

Multicollinearity

Sampling Distribution

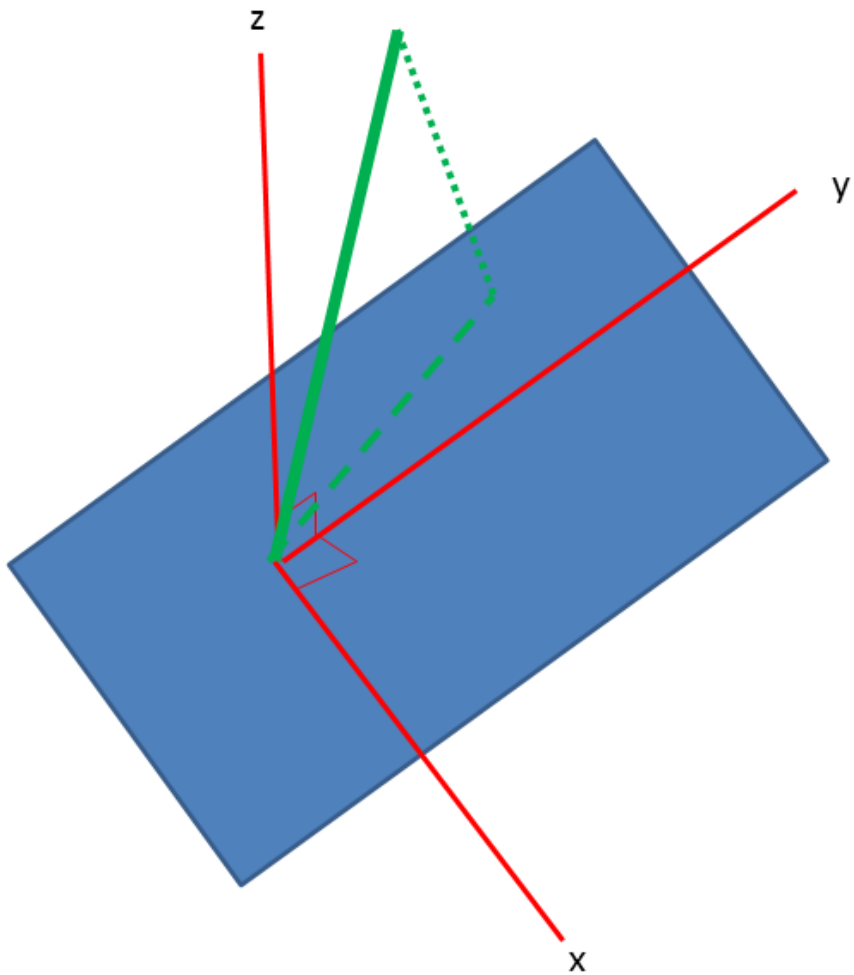
Understanding projection

We have a three dimensional vector.



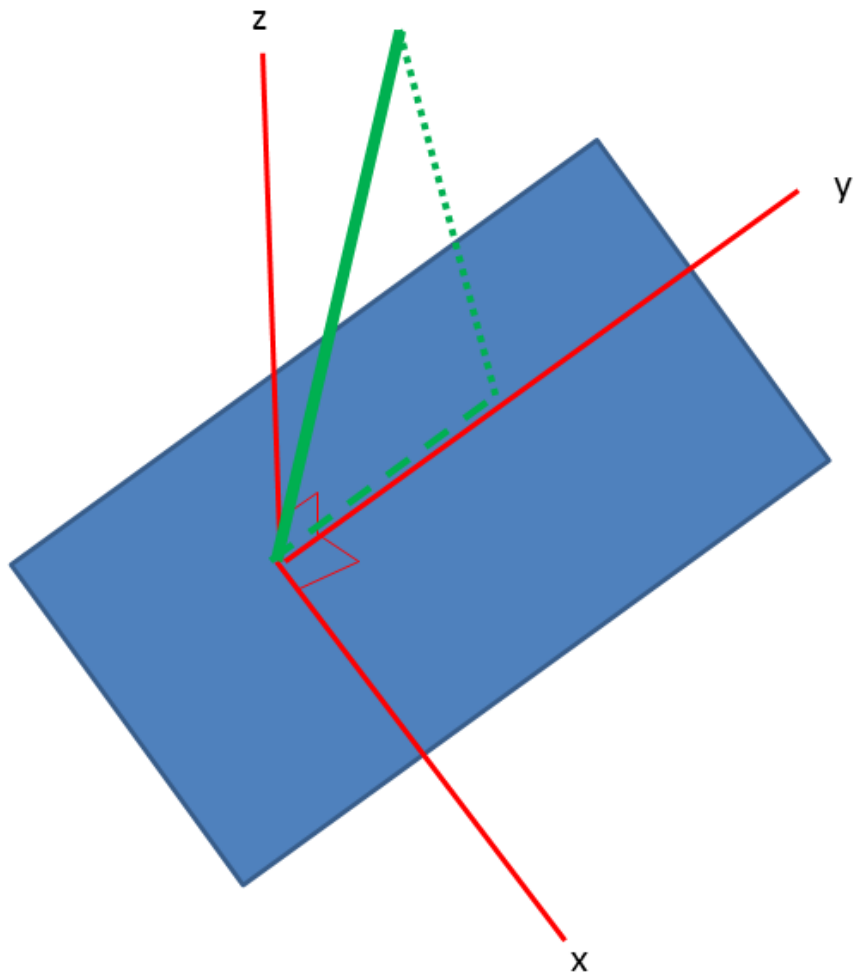
Understanding projection

We are projecting the three dimensional vector onto the (x,y) plane.



Understanding projection

We are projecting the three dimensional vector onto the y axis



Understanding regression via projection

Let $\boldsymbol{X}_0$ be a vector of ones. Then, we have

$$\begin{aligned}\boldsymbol{X}\boldsymbol{\beta} &= (\boldsymbol{x}_0 \quad \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \dots \quad \boldsymbol{x}_p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \\ &= \boldsymbol{x}_0\beta_0 + \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \cdots + \boldsymbol{x}_p\beta_p \\ &\in \mathcal{R}(\boldsymbol{X})\end{aligned}$$

Here $\mathcal{R}(\boldsymbol{X})$ is the subspace spanned by the columns of $\boldsymbol{X}$.

Orthogonal projection onto columns of $\boldsymbol{X}$

Theorem: The observation vector $\mathbf{y}$ can be uniquely decomposed as $\mathbf{y} = \hat{\mathbf{y}} + \boldsymbol{e}$, where

$$\hat{\mathbf{y}} \in \mathcal{R}(\boldsymbol{X}), \boldsymbol{e} \in \mathcal{R}(\boldsymbol{X})^{\perp},$$

with

$$\mathcal{R}(\boldsymbol{X})^{\perp} = \text{orthogonal complement of } \mathcal{R}(\boldsymbol{X})$$

In other words, if two vectors $\boldsymbol{a} \in \mathcal{R}(\boldsymbol{X})$ and $\boldsymbol{b} \in \mathcal{R}(\boldsymbol{X})^{\perp}$, we have

$$\boldsymbol{a}^T \boldsymbol{b} = 0.$$

Properties of the LSE

Projection ✓

Residuals

Underfitting and Overfitting

Multicollinearity

Sampling Distribution

Residuals

Again, residuals can be interpreted as what's left over after projecting $\boldsymbol{y}$ onto $\boldsymbol{X}$.

Definition: The residual vector is

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

Definition: The residual sum of squares is defined as

$$\begin{aligned}\boldsymbol{e}^T\boldsymbol{e} &= \sum_{i=1}^{n} e_i^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\end{aligned}$$

Hat matrix and fitted values

Definition: Let $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{y}$ denote the fitted values of $\boldsymbol{y}$, where

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

Then $\boldsymbol{P}$ is called the hat matrix. The residuals can be rewritten as

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.$$

Interpretation: $\boldsymbol{P}$ is a projection matrix that projects $\boldsymbol{y}$ onto $\mathcal{R}(\boldsymbol{X})$.

Properties of projection matrix

The projection matrix $\boldsymbol{P}$ satisfies the following properties:

- $\boldsymbol{P}$ is the projection matrix onto $\mathcal{R}(\boldsymbol{X})$.
- $\boldsymbol{I} - \boldsymbol{P}$ is the projection matrix onto $\mathcal{R}(\boldsymbol{X})^{\perp}$.
- $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$
- $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$
- Projection matrices are idempotent, i.e., $(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P}) = (\boldsymbol{I} - \boldsymbol{P})$ and $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$.
- $\boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{0}$.

Sum of squares

Recall: The three sum of squares are

- **Total Sum of Squares:** $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- **Explained Sum of Squares:** $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- **Residual Sum of Squares:** $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Sum of squares in matrix notation

Let $\boldsymbol{J} = \mathbf{1}\mathbf{1}^T$ be an $n \times n$ matrix of ones. Then we have

$$\begin{aligned}
\textbf{Total Sum of Squares:} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \boldsymbol{y}^T \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} \right) \boldsymbol{y} \\
\textbf{Explained Sum of Squares:} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
&= \boldsymbol{y}^T \left(\boldsymbol{P} - \frac{1}{n}\boldsymbol{J} \right) \boldsymbol{y} \\
\textbf{Residual Sum of Squares:} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \boldsymbol{y}^T (\boldsymbol{I} - \boldsymbol{P}) \boldsymbol{y}
\end{aligned}$$

Sum of squares (cont.)

Therefore, we have

$$\begin{aligned}
\textbf{Total Sum of Squares:} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \boldsymbol{y}^T \left(\boldsymbol{I} - \frac{1}{n} \boldsymbol{J} \right) \boldsymbol{y} \\
&= \boldsymbol{y}^T \left(\boldsymbol{I} - \frac{1}{n} \boldsymbol{J} + \boldsymbol{P} - \boldsymbol{P} \right) \boldsymbol{y} \\
&= \boldsymbol{y}^T \left(\boldsymbol{P} - \frac{1}{n} \boldsymbol{J} \right) \boldsymbol{y} + \boldsymbol{y}^T (\boldsymbol{I} - \boldsymbol{P}) \boldsymbol{y} \\
&= \textbf{Residual Sum of Squares} + \\
&\quad \textbf{Explained Sum of Squares}
\end{aligned}$$

Sum of Squares: Degrees of Freedom

The corresponding **degrees of freedom** are

- **Total Sum of Squares:** $df_T = n - 1$
- **Explained Sum of Squares:** $df_E = p$
- **Residual Sum of Squares:** $df_R = n - p - 1$

Coefficient of multiple determination

Recall: The coefficient of multiple determination is defined as

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Interpretation: Gives the amount of variation in y that is explained by the linear relationships with the covariates $\boldsymbol{X}$.

Note: When interpreting the R^2 values, note that

- $0 \leq R^2 \leq 1$
- Large R^2 values do not necessarily imply a good model
- The more covariates we include in our model, the higher R^2 is (Why? Test this in R!)

Adjusted coefficient of multiple determination

Problem of R^2: Including more and more predictors can artificially inflate R^2

- Capitalizing on spurious effects present in noisy data
- Phenomenon of over-fitting the data

The adjusted R^2 is a relative measure of fit:

$$R^2_\alpha = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / df_R}{\sum_{i=1}^n (y_i - \bar{y})^2 / df_T} = 1 - \frac{\hat{\sigma}^2}{s_y^2}.$$

where $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ is the sample estimate of the variance of $\boldsymbol{y}$.

Properties of the LSE

Projection ✓

Residuals ✓

Underfitting and Overfitting

Multicollinearity

Sampling Distribution

What happens when you underfit the model?

Suppose that the true underlying model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon},$$

but we instead fit the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In fact, in real data application, we often does this because it is impossible for us to know which variables are in the true underlying model.

For simplicity, assume that the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are linearly independent.

Bias due to underfitting

Naive Argument: I am only interested in the parameters $\boldsymbol{\beta}$, so why bother estimating $\boldsymbol{\eta}$?

Claim: if we fit the smaller model, $\mathbb{E}[\hat{\boldsymbol{\beta}}] \neq \boldsymbol{\beta}$. The estimates we get is biased! Even the fitted values are biased!

- $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{\eta}$
- $\mathbb{E}[\hat{\boldsymbol{y}}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{\eta}$

Example I: Underfitting

Suppose that the true model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

but instead we fit the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

What is the bias of β_1?

Variance if we underfit the model?

Suppose that the true underlying model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon},$$

but we instead fit the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Claim: $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}$. But

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{\boldsymbol{e}^{\text{T}}\boldsymbol{e}}{n-p-1}\right] = \sigma^2 + \frac{\boldsymbol{\eta}^{\text{T}}\boldsymbol{Z}^{\text{T}}(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{Z}\boldsymbol{\eta}}{n-p-1} > \sigma^2.$$

Implication: We overestimate the variance!

What happens when you overfit the model?

Suppose that the true underlying model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon},$$

but we instead fit the model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta},$$

where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

What will happen to our estimates $\hat{\boldsymbol{\beta}}_1$?

Bias due to overfitting

Claim: if we fit the larger model, $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. The estimates we get is **unbiased**! Even the fitted values and $\hat{\sigma}^2$ are **unbiased**!

Proof:

Why don't we keep overfitting then?

Claim: The variance of $\hat{\boldsymbol{\beta}}$ will be larger!!! Too complicated and we will skip the results.

Scenario True $\boldsymbol{\beta}_1 = (1, 1)^T$

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$$

Overfit the model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

Summary of effects of underfitting and overfitting

	Underfitting	Overfitting
$\hat{\boldsymbol{\beta}}$	biased	unbiased
$\hat{\boldsymbol{y}}$	biased	unbiased
$\hat{\sigma^2}$	biased upward	unbiased
$\mathrm{cov}(\hat{\boldsymbol{\beta}})$	still $\sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$	increased

Verify the claims in the tables in R

Model selection: How many variables are sufficient, and what variables should we include?

Properties of the LSE

Projection ✓

Residuals ✓

Underfitting and Overfitting ✓

Multicollinearity

Sampling Distribution

Understanding multicollinearity using projection

Multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

1. This means that there are strong linear dependencies among the columns of $\boldsymbol{X}$.
2. We refer to such $\boldsymbol{X}$ as almost singular matrix.
3. Does multicollinearity affect $\mathbb{E}[\hat{\boldsymbol{\beta}}]$?
4. Does multicollinearity affect $\mathrm{var}(\hat{\boldsymbol{\beta}})$?
5. What are the implications of multicollinearity?

Solutions

- Remove excessive variables
- Shrinkage estimator

Properties of the LSE

Projection ✓

Residuals ✓

Underfitting and Overfitting ✓

Multicollinearity ✓

Sampling Distribution

Sampling distribution: prerequisite

In multiple linear regression, the sampling distribution is the joint distribution of $\left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right)^{\mathrm{T}}$.
We need to introduce some new concepts to describe the distribution of $\hat{\boldsymbol{\beta}}$.

- Covariance between two random vectors $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^n$
- Multivariate normal distribution

Covariance matrix

- Covariance between two random vectors $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^n$

$$\operatorname{cov}(\boldsymbol{a}, \boldsymbol{b}) = \begin{pmatrix} \operatorname{cov}(a_1, b_1) & \operatorname{cov}(a_1, b_2) & \cdots & \operatorname{cov}(a_1, b_n) \\ \operatorname{cov}(a_2, b_1) & \operatorname{cov}(a_2, b_2) & \cdots & \operatorname{cov}(a_2, b_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(a_m, b_1) & \operatorname{cov}(a_m, b_2) & \cdots & \operatorname{cov}(a_m, b_n) \end{pmatrix}$$

- Covariance between $\boldsymbol{A}\boldsymbol{a}$ and $\boldsymbol{B}\boldsymbol{b}$

$$\operatorname{cov}(\boldsymbol{A}\boldsymbol{a}, \boldsymbol{B}\boldsymbol{b}) = \boldsymbol{A}\operatorname{cov}(\boldsymbol{a}, \boldsymbol{b})\boldsymbol{B}^{\mathrm{T}}$$

- Partitioned covariance

$$\boldsymbol{a} = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \quad \operatorname{cov}(\boldsymbol{a}, \boldsymbol{b}) = \begin{pmatrix} \operatorname{cov}(\boldsymbol{a}_1, \boldsymbol{b}_1) & \operatorname{cov}(\boldsymbol{a}_1, \boldsymbol{b}_2) \\ \operatorname{cov}(\boldsymbol{a}_2, \boldsymbol{b}_1) & \operatorname{cov}(\boldsymbol{a}_2, \boldsymbol{b}_2) \end{pmatrix}$$

- $\operatorname{var}(\boldsymbol{a}) = \operatorname{cov}(\boldsymbol{a}, \boldsymbol{a})$

Multivariate normal distribution

Let $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ and let $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The **multivariate normal density** function is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where

- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ is the p-dimensional **mean vector**
- $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$ is the **covariance matrix**

Properties of multivariate normal distribution

Properties on the mean and covariance parameters:

- $\mu_j \in \mathbb{R}$ for all j
- $\sigma_{jj} > 0$ for all j
- $\sigma_{ij} = \rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}$, where ρ_{ij} is the correlation between X_i an X_j
- $\sigma_{ij}^2 \leq \sigma_{ii}\sigma_{jj}$ for all $i, j \in \{1, \ldots, p\}$

The **marginals** of a multivariate normal is a **univariate normal**

$$X_j \sim N(\mu_j, \sigma_{jj}) \qquad \text{for all } j \in \{1, \ldots, p\}$$

Affine transformation of multivariate normal

Let $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Let $\mathbf{A} = \{a_{ij}\}_{n \times p}$ be a **non-random matrix** and consider a **non-random vector** $\mathbf{b} = (b_1, \ldots, b_n)^T$.

Define $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ with $\mathbf{A} \neq \mathbf{0}_{n \times p}$. Then

$$\mathbf{y} \sim N_n(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathbf{T}})$$

Linear combinations of normal variables are normally distributed!!!

Conditional distribution of multivariate random variables

If $\boldsymbol{\Sigma}$ is positive definite and

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_y \end{pmatrix} \right).$$

Then, the **conditional distribution of $\mathbf{x}$ given $\mathbf{y}$** is

$$\mathbf{x} \mid \mathbf{y} \sim N(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx})$$

Important Property: $\mathbf{x}$ and $\mathbf{y}$ are independent if and only if $\boldsymbol{\Sigma}_{xy} = \mathbf{0}$.

Sampling distribution: Model assumptions

Assume that

1. The errors have mean zero: $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.
2. The errors are uncorrelated with common variance $\text{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$.

These imply that

1. $\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] = \boldsymbol{X}\boldsymbol{\beta}$
2. $\text{var}(\boldsymbol{y}) = \text{var}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \text{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$

Mean and variance

1. The least squares estimate is unbiased: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
2. The covariance matrix of the least squares estimate is $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}$.

Do one of the following

- Prove these claims with rigorous derivations
- Verify these claims in `R`

How about $\boldsymbol{e}$?

Recall that $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$:

1. $\mathbb{E}[\boldsymbol{e}] = \boldsymbol{0}$
2. $\mathrm{var}(\boldsymbol{e}) = \sigma^2[\boldsymbol{I} - \boldsymbol{P}]$
3. $\mathbb{E}[\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}] = (n - p - 1)\sigma^2$.
4. Implication: An unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}}{n - p - 1}$$

Sampling distributions

- Asymptotic Distributions
  - Central Limit Theorem
  - Bootstrap
- Exact Distributions
  - Normality assumption: multivariate t-distribution
  - Other distribution assumptions...

Construct confidence intervals/regions given the sampling distribution as in simple linear regression

Inferences about multiple $\hat{\beta}_j$'s (cont)

Test Statistic:

$$
\begin{aligned}
t_F &= \frac{\mathcal{L}(\hat{\boldsymbol{\beta}}_{\text{reduced}}) - \mathcal{L}(\hat{\boldsymbol{\beta}}_{\text{full}})}{df_{\text{reduced}} - df_{\text{full}}} \Big/ \frac{\mathcal{L}(\hat{\boldsymbol{\beta}}_{\text{full}})}{df_{\text{full}}} \\
&= \frac{\mathcal{L}(\hat{\boldsymbol{\beta}}_{\text{reduced}}) - \mathcal{L}(\hat{\boldsymbol{\beta}}_{\text{full}})}{(n-q-1) - (n-p-1)} \Big/ \frac{\mathcal{L}(\hat{\boldsymbol{\beta}}_{\text{full}})}{n-p-1} \\
&\sim F_{(p-q,n-p-1)},
\end{aligned}
$$

where

- $\mathcal{L}(\hat{\boldsymbol{\beta}}_{\text{reduced}})$ is the residual sum of squares for the reduced model
- $\mathcal{L}(\hat{\boldsymbol{\beta}}_{\text{full}})$ is the residual sum of squares for the full model
- df_{reduced} is the error degrees of freedom for the reduced model
- df_{full} is the error degrees of freedom for the full model

Testing linear combinations of parameters

What if we want to test

$$H_0 : 2\beta_1 + \beta_3 = 0 \text{ v.s. } H_a : 2\beta_1 + \beta_3 \neq 0$$

- Transform the data to fit a new linear regression
- Use the Wald test
  - Estimators follow a multivariate normal distribution (asymptotically)
  - Need to know the covariance

Constructing statistical models

- Use domain knowledge
- Use statistical methods for model selection

Selecting an appropriate model

Model selection has many meanings

- It can mean selecting covariates and how they are to be included in the model

  known as feature selection, and sometimes feature generation
- It can mean choosing between a regression tree (random forest), neural network (deep learning), or logistic regression
- It can mean selecting a λ-value in the lasso

  known as tuning parameter selection

We will focus on the feature/variable selection

Trade-off on model complexity

There are two keys for a method to do well on your data:

- It doesn't completely miss important structure
- It admits a proper amount of complexity for the number of observations in your data.

The "proper" amount of complexity

How can we tell if we are choosing a model that is...

sufficiently complex (to capture the signal)...

but not overly complex! (such that we overfit our training data)?

The proper degree of complexity is **extremely** data-dependent

- Information criteria
- Loss: residual sum of squares (or likelihood) evaluated using cross-validation

Information criteria

Information criterion measures the trade-off between the goodness of fit and model complexity. In general, the information criterion admits the following form:

$$\text{information criterion} = \text{loss in model fitting} + \text{penalty on model complexity}$$

- Select a model that minimizes the chosen information criterion.
- A good information criterion need to carefully balance the trade-off between model fitting and model complexity
- Two famous information criteria: AIC and BIC

Akaike Information Criterion (AIC)

Formulated by Japanese statistician Hirotugu Akaike in 1973.

$$\mathrm{AIC} = 2k + n\log(\mathrm{RSS}/n) \text{ or } \mathrm{AIC} = 2k - 2\log\mathcal{L}_k$$

- Loss of model fitting: log of residual sum of squares (for linear regression), or two times negative log-likelihood (in general)
- Penalty of model complexity: two times the number of variables k.

When the sample size n is large, AIC tends to be conservative on "punishing" model complexity.

Bayesian Information Criterion (BIC)

BIC was developed by statistician Gideon E. Schwarz in 1978. Therefore, BIC is also named as Schwarz information criterion (SIC).

$$\mathrm{BIC} = \log(n)k + n\log(\mathrm{RSS}/n) \quad \text{or } \mathrm{BIC} = \log(n)k - 2\log\mathcal{L}_k$$

- Loss of model fitting: log of residual sum of squares (for linear regression), or two times negative log-likelihood (in general)
- Penalty of model complexity: $\log(n)$ times the number of variables

Model selection

Choosing a proper complexity for your model

- Information criteria
- Loss function: residual sum of squares (or likelihood) evaluated using cross-validation

Split-sample validation

Q: What is overfitting?

Ideally, we wish to have two datasets, where we can

- explore all possible models on the training data
- evaluate/confirm our findings on the test data

In practice, often times we only have one dataset....
so we split the dataset into a training set and a test set.

How to split?

What proportion should be used for training vs testing/evaluating?

Often something like $2/3$ training - $1/3$ testing is good.

This still feels like inefficient data use.

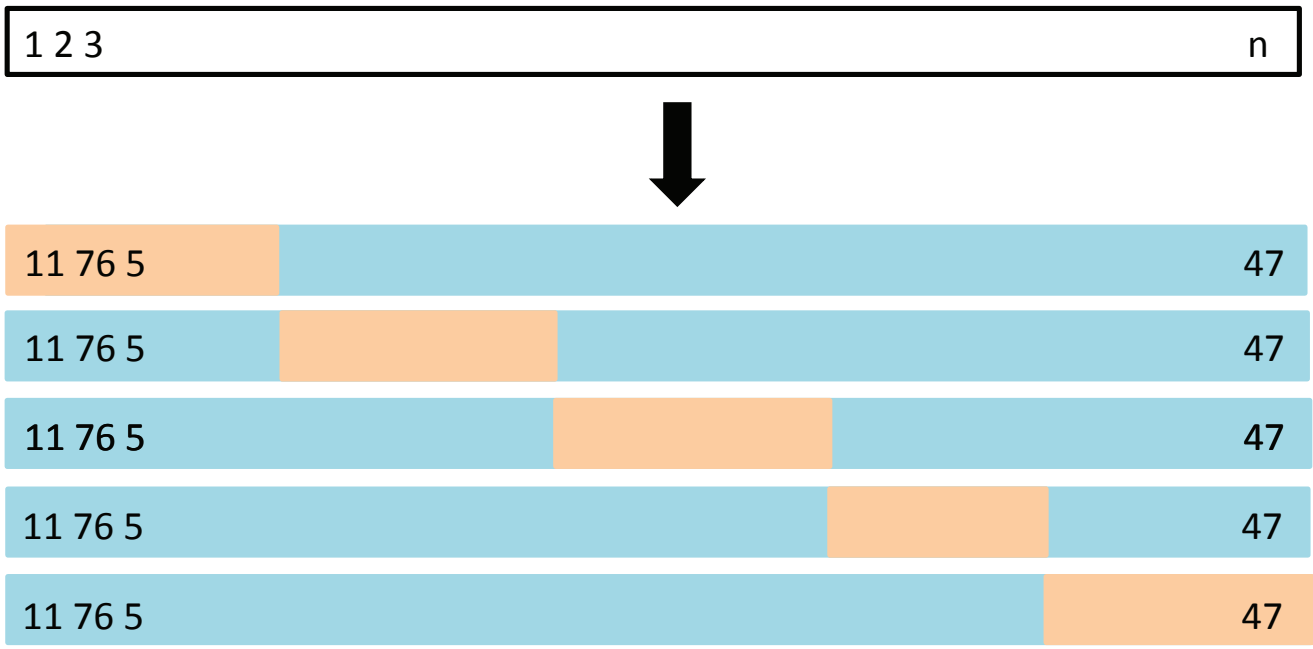
Is there some way to use the majority of the data for both training and testing?

Cross-validation - I

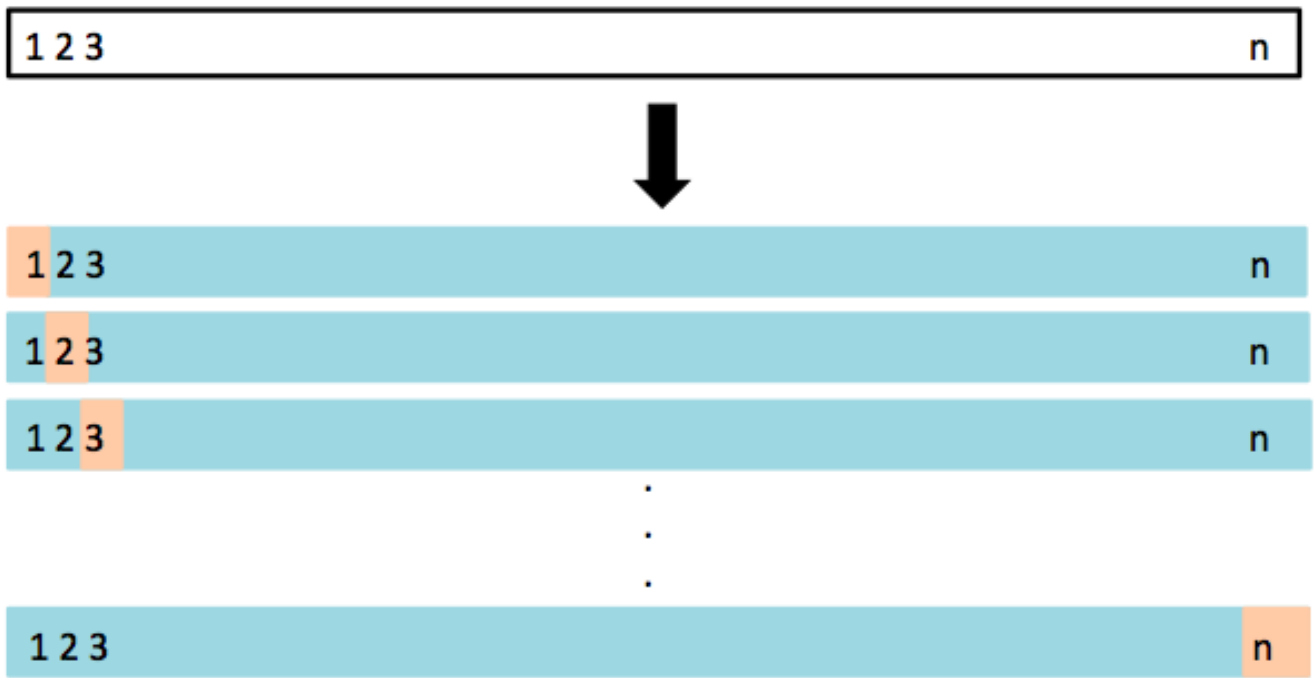
Let's use cross-validation:

- Partition our data into multiple folds...
- Each time use 1 fold as test, and all other folds as training

K-fold cross-validation



Leave-one-out cross-validation



Cross-validation II

Procedure for a k-fold validation

1. Randomly split the dataset into k non-overlapping subsets (folds)
2. For $i = 1, \ldots, k$,
   - 2.1 Fit the chosen model on data from the data excluding the ith fold
   - 2.2 Evaluate the loss of the fitted model on the ith fold, denoted as $\mathcal{L}_i$
3. The final loss is $\sum_{i=1}^{k} \mathcal{L}_i / k$.

Model selection

Choosing a proper complexity for your model

- Information criteria
- Loss function: residual sum of squares (or likelihood) evaluated using cross-validation

Next, how to select the "best" model?

Selection procedure

- Best subset selection
- Stepwise selection
- Penalization

Best subset selection

An intuitive procedure:

1. List all possibilities
2. The model with the best criteria (AIC, BIC, CV loss) wins

Algorithm: Best Subset Selection

1. Choose a selection criterion
2. For $k = 1, \ldots, p$,
   - 2.1 Fit all possible models that contain exactly k covariates
   - 2.2 Pick the best among these models (the one with the best criterion)
   - 2.3 Denote this model as M_k
3. Select a single best model among $M_1, \ldots, M_k$

Best subset selection

An intuitive procedure:

1. List all possibilities
2. The model with the best criteria (AIC, BIC, CV loss) wins

There are a lot of models to fit (2^p)!!!

Stepwise selection

Stepwise selection:

- Computationally efficient alternative
- Explore a restricted set of models

  No guarantee that it can find the best possible model!

Two directions

- Forward
- Backward

Forward stepwise selection

Algorithm: Forward Adding

1. Choose a selection criterion
2. Let M_0 denote the null model, which contains no predictors
3. For $k = 0, \ldots, p-1$
   - 3.1 Fit all $p-k$ models by adding one additional variable to M_k
   - 3.2 Pick the best among these $p-k$ models, denoted this model as M_{k+1}
4. Select a single best model among $M_0, \ldots, M_p$ using the chosen criterion.

You may stop early if you use p-values as the criterion...

Backward stepwise selection

Algorithm: Backward Deleting

1. Choose a selection criterion
2. Let M_p denote the full model, which contains all predictors
3. For $k = p - 1, \ldots, 1$
   - 3.1 Fit all k models by deleting one variable from M_k
   - 3.2 Pick the best among these k models, denoted this model as M_{k-1}
4. Select a single best model among $M_0, \ldots, M_p$ using the chosen criterion.

You may stop early if you use p-values as the criterion...

Penalization

We can approximate the best subset selection using sparsity-inducing penalties.

Choose $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$, to minimize

$$\mathcal{L}(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})$$

Lasso

One notable example is the **lasso** where $P(\boldsymbol{\beta}) = |\beta_1| + \ldots + |\beta_p|$, a function that penalizes complexity in β

Minimizing β-values, will "magically" give sparse model
(meaning: many β-values exactly equal to 0)

To modulate the degree of sparsity, we change λ.

Q: How would you choose λ?

Linear regression assumptions

Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

1. Constant variance assumption $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.
2. Uncorrelated error.

What if these assumptions are violated? How does it affect our solution if we fit the ordinary multiple linear regression?

Non-constant variance and correlated error

Suppose that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

1. Non-constant variance $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$ for some positive definite matrix $\mathbf{V}$.
2. Correlated error.

How should we estimate β?