Lecture 1: Introduction

Applied Statistics and Data Analysis II

Instructor: Jesús Arroyo

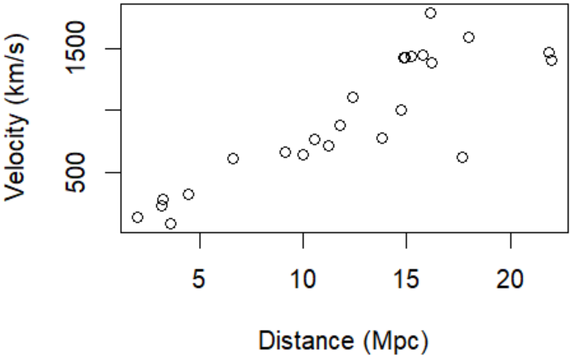
Example: Hubble telescope data

- Consider a sample of measurements $(x_1, y_1), \ldots, (x_n, y_n)$ obtained by the Hubble Space Telescope
  - x_i is the distance between the Earth and a given galaxy (Megaparsecs),
  - y_i is the relative velocity of the galaxy (km/s).
- The Big Bang model states that the universe expands uniformly.
- According to Hubble's law, these quantities follow a linear relation

$$y = \beta x,$$

  where β is Hubble's constant.
- β^{-1} gives the approximate age of the universe.

```
library(gamair)
data("hubble")
plot(hubble$x, hubble$y, xlab = "Distance (Mpc)",
     ylab = "Velocity (km/s)")
```



Simple linear regression

- Measurements are noisy, so in reality the data look like

$$y_i = \beta x_i + \epsilon_i.$$

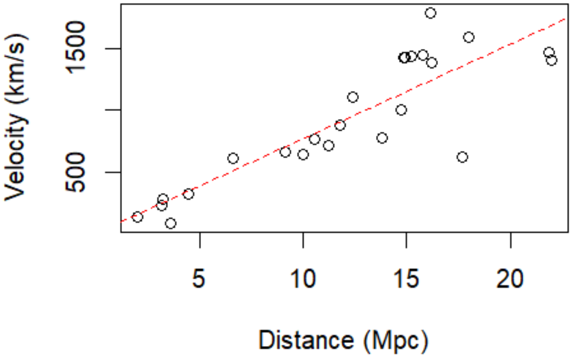
- How to estimate β?
- Find the parameter with the best fit to the data.
- **Simple least squares estimation**: choose β that minimizes

$$L(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2.$$

- The least square estimator for β is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

```
hub.mod <- lm(y ~ x - 1, data = hubble)
plot(hubble$x, hubble$y, xlab = "Distance (Mpc)",
     ylab = "Velocity (km/s)")
abline(hub.mod, lty=2, col="red")
```



Multiple linear regression

- More generally, consider data $(X_1, y_1), \ldots, (X_n, y_n)$, where X_i is a p-dimensional vector.
- Denote $\mathbf{X} = [X_1 \ \cdots X_n]^T$ to the $n \times p$ covariate data matrix
- $Y = (y_1, \ldots, y_n)$ is the response vector.
- Linear model for the response:

$$y_i = X_i^T \boldsymbol{\beta} + \epsilon_i,$$

- $\boldsymbol{\beta}$ is the p-dimensional vector of coefficients.
- $\epsilon_1, \ldots, \epsilon_n$ are unobserved errors.

Least squares estimation

- The fit to the data can be measured with the **ordinary least squares** (OLS) loss function

$$L(\beta) = \sum_{i=1}^{n} (y_i - X_i^T \beta)^2 = \|Y - \mathbf{X}\beta\|_2^2.$$

- The OLS estimator $\hat{\beta}$ can be obtained by differentiating

$$\begin{aligned} \hat{\beta} &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}\, L(\beta) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y, \end{aligned}$$

  provided that $\mathbf{X}^T \mathbf{X}$ is invertible.

Ordinary least squares

What is special about the OLS estimator?

- For each $i = 1, \ldots, n$, assume the following.
  - Linear model for the response: $y_i = X_i \beta + \epsilon_i$.
  - Errors have zero mean: $\mathbb{E}[\epsilon_i | \mathbf{X}_i] = 0$.
- Then, the OLS estimator is **unbiased**, i.e.,

$$\mathbb{E}[\hat{\beta} | \mathbf{X}] = \beta.$$

- The estimated response $\hat{Y} = X\hat{\beta}$ satisfies

$$\mathbb{E}[\hat{Y} | \mathbf{X}] = \mathbf{X}\beta.$$

Ordinary least squares

- Additionally, assume the following on the error vector $\epsilon = (\epsilon_1, \dots, \epsilon_n)$
  - Homoscedasticity on the errors, i.e., constant variance:

$$\mathrm{Var}(\epsilon_i | \mathbf{X}_i) = \sigma^2 < \infty, \qquad i = 1, \dots, n.$$

  - Uncorrelated errors: defining I to be the $n \times n$ identity matrix,

$$\mathrm{Cov}(\epsilon) = \sigma^2 I.$$

- Then, the covariance matrix of $\hat{\beta}$ satisfies

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

Optimality

- Is there a better estimator? It depends how "better" is defined.
- Consider all linear estimators of β, that is, all $\tilde{\beta}$ of the form

$$\tilde{\beta} = \mathbf{M}Y$$

  for some matrix $\mathbf{M} \in \mathbb{R}^{p \times n}$.
- In particular, $\mathbf{M} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ is the OLS estimator $\hat{\beta}$.
- There are many matrices M for which $\tilde{\beta}$ is unbiased

Optimality

- **Gauss-Markov theorem:** The OLS estimator is the best linear unbiased estimator, in the sense that its covariance matrix is smallest
$$\mathrm{Cov}(\tilde{\beta}|\mathbf{X}) \succeq \mathrm{Cov}(\hat{\beta}|\mathbf{X}).$$
Here, $A \succeq B$ indicates that $A - B$ is positive semidefinite.
- Are there better linear *biased* estimators? The answer will come later in the course.

Hubble telescope data

- Returning to the example, we can use OLS to estimate Hubble's constant

```
> hub.mod <- lm(y ~ x - 1, data = hubble)
> hubble.constant <- coef(hub.mod)
> # convert Mega-parsecs to km
> hubble.constant.km <- hubble.constant/3.09e19
> age <- 1/hubble.constant.km
> # convert seconds to years
> age / (60*60*24*365)
12794692825
```

- Based on our plug-in estimate, the age of the universe is approximately 12.8 billion years.

Parameter inference

In addition to estimate β, there are other questions we might want to answer:

- Calculating confidence intervals
- Testing hypothesis about β
- Model selection

Normally distributed errors

- Additional assumption: the errors are normally distributed

$$\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2), \qquad i = 1, \ldots, n$$

$$f(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right).$$

- In a normal distribution, uncorrelated variables are also independent, so $\epsilon_1, \ldots, \epsilon_n$ are i.i.d.
- Under the Gaussian distribution assumption, the OLS estimator has a multivariate Gaussian distribution

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{n}(X^T X)^{-1}\right).$$

Estimating the variance of the errors

- The variance of the errors is unknown. An unbiased estimator for it is given by
$$\hat{\sigma}^2 = \frac{\|Y - \mathbf{X}\hat{\beta}\|_2^2}{n-p}.$$
- Under the normal distribution assumption, its distribution can be shown to follow
$$\frac{(n-p)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p}.$$
- Another useful property: $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

Hypothesis testing

- Suppose that want to test if β_i, the i-th entry of β, is equal to some value $b_0 \in \mathbb{R}$.
- Let $\hat{\beta}_{\text{obs}}$. be the observed estimate of β for a given dataset.
- Define $\mathbf{V} = (\mathbf{X}^T\mathbf{X})^{-1}$.
- A pivotal quantity for β_i:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 V_{ii}}} \sim t_{n-p}.$$

Hypothesis testing

- Under the null hypothesis $\mathscr{H}_0 : \beta_i = b_0$, the p-value of the test is

$$\mathbb{P}\left(|\hat{\beta}_i - \beta_i| > |\hat{\beta}_{\text{obs},i} - \beta_i| \,\middle|\, \beta_i = b_0 \right) =$$

$$\mathbb{P}\left(\frac{|\hat{\beta}_i - \beta_i|}{\sqrt{\hat{\sigma}^2 \mathbf{V}_{ii}}} > \frac{|\hat{\beta}_{\text{obs},i} - \beta_i|}{\sqrt{\hat{\sigma}^2 \mathbf{V}_{ii}}} \,\middle|\, \beta_i = b_0 \right)$$

- The above p-value can be calculated explicitly using the t_{n-p} distribution.

Hubble telescope data

- Current estimates of the Hubble's constant β are around 67.31. The current estimate of the age of the universe is around 13.8 billions of years (source: Wikipedia).
- Our estimate for β is 76.58. Is this estimate compatible with the current estimates?

```
> hubble.ct.wiki <- 67.31
> n <- length(hubble$y)
> p <- 1
> t.stat <- abs(hubble.constant - hubble.ct.wiki) /
+   vcov(hub.mod)[1,1]^0.5
> # tail probability of a t distribution
> (1 - pt(t.stat, df=n-p))*2
0.02842646
```

Summary: linear models

The following assumptions are convenient in regression problems

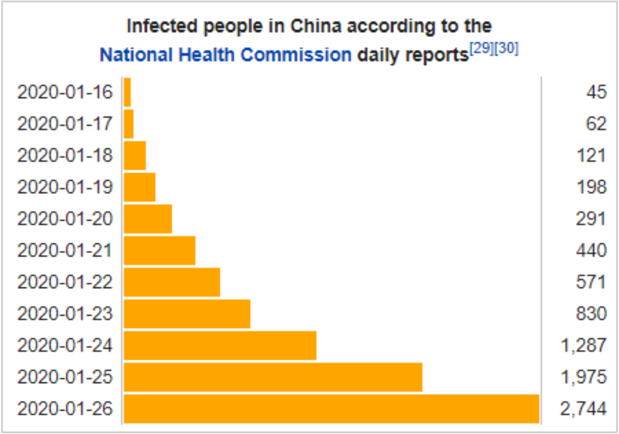
- Linear relationship between the covariates and the response
- Mean-zero, equal variance and uncorrelated errors
- Gaussian distribution of the errors

How can we get over those assumptions?

Case 1: non-linear responses

- Some possible violations to the model
  - The response y might not be linear in X
  - The distribution of the error might not be normal
- **Example:** epidemiological models suggest that the rate in which new infections occur at early stages of an outbreak is exponential
  - $y_i =$ number of new cases on day x_i.
  - $\mathbb{E}[y_i] = \gamma \exp(\eta x_i)$

Example: 2020 Coronavirus outbreak



Source: Wikipedia
(see also `https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6`)

Generalized linear models

How to model those data?

- **Generalized linear models:** suppose that there exist a link function g such that

$$g(\mathbb{E}[y|X]) = X\beta,$$

$$y \sim \text{some distribution.}$$

- **Example (continued):** The epidemic data can be modeled using $g(\cdot) = \log(\cdot)$.
- Since y_i represents counts, we can use a Poisson distribution for Y.

Case 2:

- The errors and the covariates can be related
  - $\mathbb{E}[\epsilon_i|X_i] = b_i$.
  - $\mathrm{Var}(\epsilon_i|X_i) = \sigma_i^2$.
- **Example:** Consider height and weight data (x_i, y_i) from the same individual.
- Suppose that in addition, we know that individuals are clustered by family.

Linear mixed models:

- The linear mixed models extend the linear model

$$Y = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon.$$

- $\mathbf{b}$ is a random d-vector that contains random effects.
- $\mathbf{Z}$ is a $n \times d$ model matrix for the random effects.
- In the example, $\mathbf{Z}$ can be an indicator function of each family

Case 3: collinearities in the data

- Recall that the covariance of the OLS estimator is given by

$$\mathrm{Cov}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

.

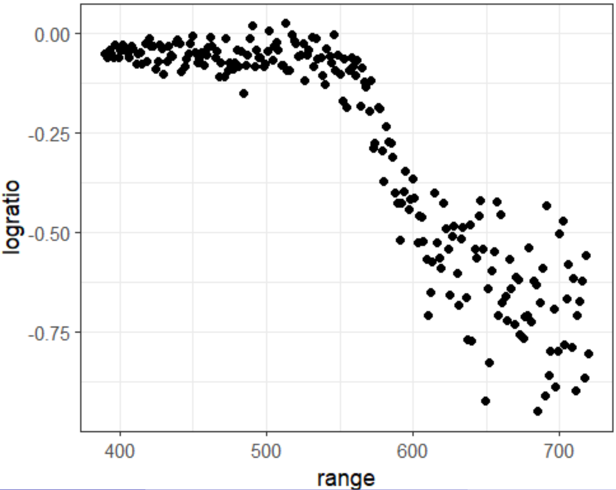
- When $\det(\mathbf{X}^T\mathbf{X}) \to 0$, the variance of the estimator for β grows to infinity.
- **Example:** High-dimensional data.
- When $p > n$, $\det(\mathbf{X}^T\mathbf{X}) = 0$.
- Consider gene expresion measurements $X_1, \ldots, X_n$ from a sample of individuals from cancer patients. The response $y_1, \ldots, y_n$ indicates cancer type
- Usually, more than 5,000 genes are measured, while the sample size is in the order of hundreds.

Penalized least squares

- Consider a biased estimator for β, obtained by penalizing the least squares loss
$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) + \lambda \text{pen}(\beta).$$
- $\text{pen}(\cdot)$ is a non-negative function that penalizes certain values of β.
- $\lambda > 0$ is a penalty parameter.
- Penalized estimators can reduce the variance in estimating β to produce a more accurate estimator.

Case 4: unknown non-linear model

- In some cases, the relation between the response and the covariates is unknown, and non-linear
- Spectroscopy data: light detection and ranging experiment



Nonparametric methods

One solution: construct a piecewise linear function