

P8131
Biostatistical Methods II
Spring 2019

Instructor: Gen Li

Department of Biostatistics, Columbia

Today's Outline

- ▶ Go over the syllabus
- ▶ Course overview
- ▶ Review linear models

Something about the Course

- ▶ Required textbooks:
 - ▶ Dobson and Barnett (2008) *An Introduction to Generalized Linear Model*. 3rd Ed. Chapman & Hall.
 - ▶ Fitzmaurice, Laird and Ware (2011) *Applied Longitudinal Analysis*. 2nd Ed. Wiley.
 - ▶ Hosmer, Lemeshow and May (2008) *Applied Survival Analysis*. 2nd Ed. Wiley.
- ▶ Recommended textbooks:
 - ▶ McCullagh and Nelder (1989) *Generalized Linear Models*. 2nd Ed. Chapman & Hall.
 - ▶ Faraway (2016) *Extending the Linear Model with R*. 2nd Ed. Chapman & Hall.
 - ▶ Diggle, Heagerty, Liang and Zeger (2013) *Analysis of Longitudinal Data*. 2nd Ed. Oxford.
 - ▶ Klein and Moeschberger (2003) *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd Ed. Springer.

- ▶ Grading policy:
 - ▶ Homework (40%)
 - ▶ 10 times, equal weights.
 - ▶ NO late homework.
 - ▶ Do not copy/paste R outputs; instead, interpret them!
 - ▶ Submit electronically on Canvas.
 - ▶ Midterm exam (30%)
 - ▶ Final exam (30%)
- ▶ Check Canvas frequently for new HW, materials, and grades.
- ▶ My office hours: after class on Tuesdays and Thursdays
- ▶ TA office hours: 10am–11am M.W.F., R627

- ▶ Grading policy:
 - ▶ Homework (40%)
 - ▶ Midterm exam (30%)
 - ▶ March 14, in class, one cheat sheet
 - ▶ Final exam (30%)
- ▶ Check Canvas frequently for new HW, materials, and grades.
- ▶ My office hours: after class on Tuesdays and Thursdays
- ▶ TA office hours: 10am–11am M.W.F., R627

- ▶ Grading policy:
 - ▶ Homework (40%)
 - ▶ Midterm exam (30%)
 - ▶ Final exam (30%)
 - ▶ May 14, 9am-11:50am, two cheat sheet
 - ▶ Notify me of any conflict by Jan 31
- ▶ Check Canvas frequently for new HW, materials, and grades.
- ▶ My office hours: after class on Tuesdays and Thursdays
- ▶ TA office hours: 10am–11am M.W.F., R627

- ▶ Classroom policy:
 - ▶ Classroom participation needed.
 - ▶ When something is unclear, just ASK.
 - ▶ Frequent quizzes, peer review.
 - ▶ Comments and suggestions are always welcome.
 - ▶ Do NOT share any course material online without permission.
 - ▶ Administrative questions should be directed to [Justine Herrera](#) (UNI: jh2477)

- ▶ Classroom policy:
 - ▶ Classroom participation needed.
 - ▶ When something is unclear, just ASK.
 - ▶ Frequent quizzes, peer review.
 - ▶ Comments and suggestions are always welcome.
 - ▶ Do NOT share any course material online without permission.
 - ▶ Administrative questions should be directed to [Justine Herrera](#) (UNI: jh2477)



Something about Me

- ▶ Tenure-track assistant professor in the Department of Biostatistics
- ▶ Got my PhD from UNC-Chapel Hill in 2015
- ▶ Interested in statistical learning (data integration, tensor, high dimension) and cool applications (multi-omics, microbiome, networks, etc)
- ▶ Looking forward to working with talented, self-motivated students

Something about You

- ▶ Your name
- ▶ What is your goal in this course?
- ▶ Anything else you want me to know?

Course Overview

- ▶ This course continues P8130 (Biostatistical Methods I) and generalizes it in several directions.
- ▶ We will cover
 - ▶ Generalized Linear Models
 - ▶ Longitudinal Data Analysis
 - ▶ Survival Analysis
- ▶ The goal is to introduce basic concepts of each topic, and demonstrate how to use them to solve real problems
- ▶ Each topic is a stand-alone course
- ▶ You need to take additional courses to master those subjects
- ▶ R is used in the course. You may use other software (e.g., SAS, Matlab, SPSS, Excel) for data analysis in your HW.

Review of Linear Models

Suppose there are n subjects. For subject i , $i = 1, \dots, n$, we observe response Y_i and covariates X_{i1}, \dots, X_{ip} . Assume $p < n$.

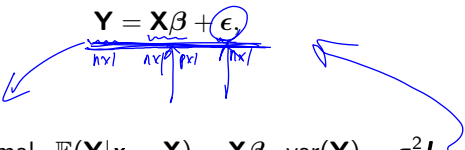
Let $\mathbf{Y} = (Y_1, \dots, Y_n)'$ and $\mathbf{X}_j = (X_{1j}, \dots, X_{nj})'$ for $j = 1, \dots, p$.

Then, the design (or model) matrix is

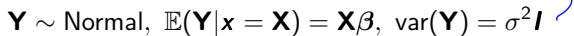
$$\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p).$$

Note that the intercept can be included by setting $\mathbf{X}_1 = \mathbf{1}$.

A linear regression model assumes that

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


or equivalently

$$\mathbf{Y} \sim \text{Normal}, \mathbb{E}(\mathbf{Y}|\mathbf{x} = \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}, \text{var}(\mathbf{Y}) = \sigma^2 \mathbf{I}$$


where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is the regression coefficient vector and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ are the regression errors, with $\mathbb{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$.

Ordinary Least Squares

The OLS estimate of β is defined as

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

$\hat{\beta}$ satisfies the normal equation

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\beta}.$$

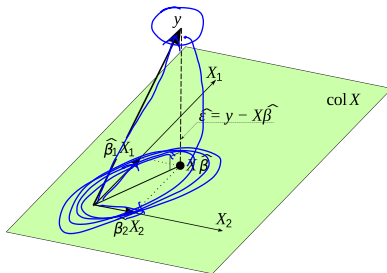
If $\operatorname{rank}(\mathbf{X}'\mathbf{X}) = p$, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

$\hat{\beta}$ has the following properties:

- ▶ $\mathbb{E}(\hat{\beta}) = \beta$;
- ▶ $\operatorname{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$;
- ▶ $\hat{\beta}$ is BLUE.

Furthermore, let $\operatorname{RSS} = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2$, then $\hat{\sigma}^2 = \operatorname{RSS}/(n - p)$ is unbiased for σ^2 .

Geometric representation of OLS:



Maximum Likelihood Estimation

$$Y = X\beta + \varepsilon$$

Estimate β and σ by maximizing the likelihood function

$$l(\beta, \sigma) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\|Y - X\beta\|^2}{2\sigma^2} \right\}.$$

- ▶ $\hat{\beta}_{ML} = \hat{\beta}_{OLS} = X^T X^{-1} X^T Y$ with normal indep. errors;
- ▶ $\hat{\sigma}_{ML}^2 = \frac{RSS}{n} \neq \frac{RSS}{n-p} = \hat{\sigma}_{OLS}^2$; biased, yet asymptotically unbiased and efficient.

Residuals and Model Diagnostics

Residual $\hat{\epsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ has variance $\sigma^2(\mathbf{I} - \mathbf{H})$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat (projection) matrix.

Define the standardized residual as

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}(1 - h_{ii})^{1/2}},$$

where $\hat{\sigma}$ is the OLS estimate of σ , and h_{ii} (i.e., leverage of the i th observation) is the i th diagonal value of \mathbf{H} . The standardized residuals can be used to check the normality assumption, goodness-of-fit, and homoscedasticity.

Hypothesis Testing

$$I_{p \times p} \cdot \beta = 0_{p \times 1}$$

$$I \cdot \beta = \beta \neq 0$$

$$H_0: \mathbf{C}\beta = \mathbf{h} \quad \text{vs.} \quad H_1: \mathbf{C}\beta \neq \mathbf{h},$$

where \mathbf{C} is a $q \times p$ matrix and $\text{rank}(\mathbf{C}) = q$, \mathbf{h} is a $q \times 1$ vector.

For example,

$$\blacktriangleright \mathbf{C} = \mathbf{e}_1^t = (1, 0, \dots, 0), \mathbf{h} = 0: H_0: \beta_1 = 0;$$

$$\blacktriangleright \mathbf{C} = \mathbf{I}_p, \mathbf{h} = \mathbf{0}: H_0: \beta = \mathbf{0}.$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

$$H_0: \beta_1 = \beta_2 \quad H_1: \beta_1 \neq \beta_2$$

$$\mathbf{C} = (1, -1, 0) \quad \mathbf{h} = 0$$

$$\hat{\beta}_0 = \underset{\beta: \mathbf{C}\beta = \mathbf{h}}{\text{argmin}} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \quad H_0: (1, -1, 0) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \beta_1 - \beta_2 = 0$$

$$= \hat{\beta} - (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{C}^t [\mathbf{C} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{C}^t]^{-1} (\mathbf{C} \hat{\beta} - \mathbf{h}),$$

$$\hat{\sigma}_0^2 = \frac{RSS_0 = \|\mathbf{r} - \mathbf{X}\hat{\beta}_0\|^2}{n},$$

which can be obtained from the Lagrange method.

F-test:

Under H_0 , $\text{RSS}_0 = \|\mathbf{Y} - \mathbf{X}\hat{\beta}_0\|^2$;

Under the full model, $\text{RSS} = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2$.

Theorem

Suppose $\epsilon \sim N(0, \sigma^2 I_n)$, then $\text{RSS}/\sigma^2 \sim \chi_{n-p}^2$. Under H_0 , $(\text{RSS}_0 - \text{RSS})/\sigma^2 \sim \chi_q^2$ and is independent of RSS ; Furthermore,

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p)} \sim F_{q, n-p}.$$

The F -test covers a general class of linear hypothesis testing problems in the form of $\mathbf{C}\beta = \mathbf{h}$.

Likelihood ratio test:

$$\lambda = 2\{\log l(\hat{\beta}, \hat{\sigma}) - \log l(\hat{\beta}_0, \hat{\sigma}_0)\} = n \log(\text{RSS}_0/\text{RSS}).$$

It can be shown that under $H_0 : \mathbf{C}\beta = \mathbf{h}$, λ tends to a χ_q^2 distribution as $n \rightarrow \infty$, which can be used to define the cutoff value for large samples.

In fact,

$$F = \frac{n-p}{q}(e^{\lambda/n} - 1).$$