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, $i = 1, \ldots, n$, we observe response Y_i and covariates $X_{i1}, \ldots, X_{ip}$. Assume $p < n$.
Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ and $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})'$ for $j = 1, \ldots, p$.
Then, the design (or model) matrix is

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

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

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

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

Ordinary Least Squares

The OLS estimate of $\boldsymbol{\beta}$ is defined as

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

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

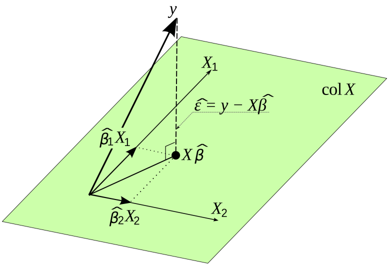
$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.$$

If $\mathrm{rank}(\mathbf{X}'\mathbf{X}) = p$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
$\hat{\boldsymbol{\beta}}$ has the following properties:

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

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

Geometric representation of OLS:



Maximum Likelihood Estimation

Estimate $\boldsymbol{\beta}$ and σ by maximizing the likelihood function

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

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

Residuals and Model Diagnostics

Residual $\hat{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$ has variance $\sigma^2(\boldsymbol{I} - \boldsymbol{H})$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 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 ith observation) is the ith diagonal value of $\boldsymbol{H}$. The standardized residuals can be used to check the normality assumption, goodness-of-fit, and homoscedasticity.

Hypothesis Testing

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

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

- $\boldsymbol{C} = \boldsymbol{e}_1^t = (1, 0, \ldots, 0)$, $\boldsymbol{h} = 0$: $H_0: \ \beta_1 = 0$;
- $\boldsymbol{C} = \boldsymbol{I}_p$, $\boldsymbol{h} = \mathbf{0}$: $H_0: \ \boldsymbol{\beta} = \mathbf{0}$.

Under H_0,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_0 &= \mathrm{argmin}_{\boldsymbol{\beta}:\boldsymbol{C}\boldsymbol{\beta}=\mathbf{h}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \\
&= \hat{\boldsymbol{\beta}} - (\mathbf{X}^t\mathbf{X})^{-1}\boldsymbol{C}^t[\boldsymbol{C}(\mathbf{X}^t\mathbf{X})^{-1}\boldsymbol{C}^t]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \mathbf{h}), \\
\hat{\sigma}_0^2 &= \frac{RSS_0}{n},
\end{aligned}$$

which can be obtained from the Lagrange method.

F-test:

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

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

Theorem

Suppose $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$, *then* $\mathrm{RSS}/\sigma^2 \sim \chi^2_{n-p}$. *Under* H_0, $(\mathrm{RSS}_0 - \mathrm{RSS})/\sigma^2 \sim \chi^2_q$ *and is independent of* RSS; *Furthermore,*

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

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

Likelihood ratio test:

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

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

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