Math 6040/7260 Linear Models

Mon/Wed/Fri 10:55am - 11:40am

Instructor: Dr. Xiang Ji, xji4@tulane.edu

1 Lecture 1:Jan 20

Today

- Introduction
- Course logistics
- Read JF chapter 1, JM Appendix A

What is this course about?

The term “linear models” describes a wide class of methods for the statistical analysis of multivariate data. The underlying theory is grounded in linear algebra and multivariate statistics, but applications range from biological research to public policy. The objective of this course is to provide a solid introduction to both the theory and practice of linear models, combining mathematical concepts with realistic examples.

A hierarchy of linear models

- The linear mean model:

  $$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}}\underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon}$$

  where $\mathbf{E}(\epsilon) = \mathbf{0}$. Only assumption is that errors have mean 0.

- Gauss-Markov model:

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

  where $\mathbf{E}(\epsilon) = \mathbf{0}$ and $\mathbf{Var}(\epsilon) = \sigma^2\mathbf{I}$. Uncorrelated errors with constant variance.

- Aitken model or general linear model:

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

  where $\mathbf{E}(\epsilon) = \mathbf{0}$ and $\mathbf{Var}(\epsilon) = \sigma^2\mathbf{V}$. $\mathbf{V}$ is fixed and known.

- Variance components models: $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma_1^2\mathbf{V}_1 + \sigma_2^2\mathbf{V}_2 + \cdots + \sigma_r^2\mathbf{V}_r)$ with $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_r$ known.

- General mixed linear Model:

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

  where $\mathbf{E}(\epsilon) = \mathbf{0}$ and $\mathbf{Var}(\epsilon) = \boldsymbol{\Sigma}(\theta)$.

- Generalized linear models (GLMs). Logistic regression, probit regression, log-linear model (Poisson regression), ... Note the difference from the general linear model. GLMs are generalization of the *concept* of linear models. They are covered in Math 7360 - Data Analysis class (https://tulane-math7360.github.io/lectures/).

Syllabus

Check course website frequently for updates and announcements.

https://tulane-math-7260-2021.github.io/

HW submission

Through Github with demo on Friday class.

2 Lecture 2:Jan 22

Last time

- Introduction
- Course logistics

Today

- Introduce yourself (remind remote students to record a short video)
  - basic info (name, department, year, ...)
  - why taking this course
- Git
- Linear algebra: vector and vector space, rank of a matrix

What is git?

Git is currently the most popular system for version control according to Google Trend. Git was initially designed and developed by Linus Torvalds in 2005 for Linux kernel development. Git is the British English slang for unpleasant person.

Why using git?

- GitHub is becoming a de facto central repository for open source development.
- **Advertise** yourself through GitHub (e.g., host a free personal webpage on GitHub).
- a skill that employers look for (according to this AmStat article).

Git workflow

Figure 2.1 shows its basic workflow.

What do I need to use Git?

- A **Git server** enabling multi-person collaboration through a centralized repository.
- A **Git client** on your own machine.
  - Linux: Git client program is shipped with many Linux distributions, e.g., Ubuntu and CentOS. If not, install using a package manager, e.g., `yum install git` on CentOS.
  - Mac: follow instructions at https://www.atlassian.com/git/tutorials/install-git.
  - Windows: Git for Windows at https://gitforwindows.org (GUI) aka Git Bash.

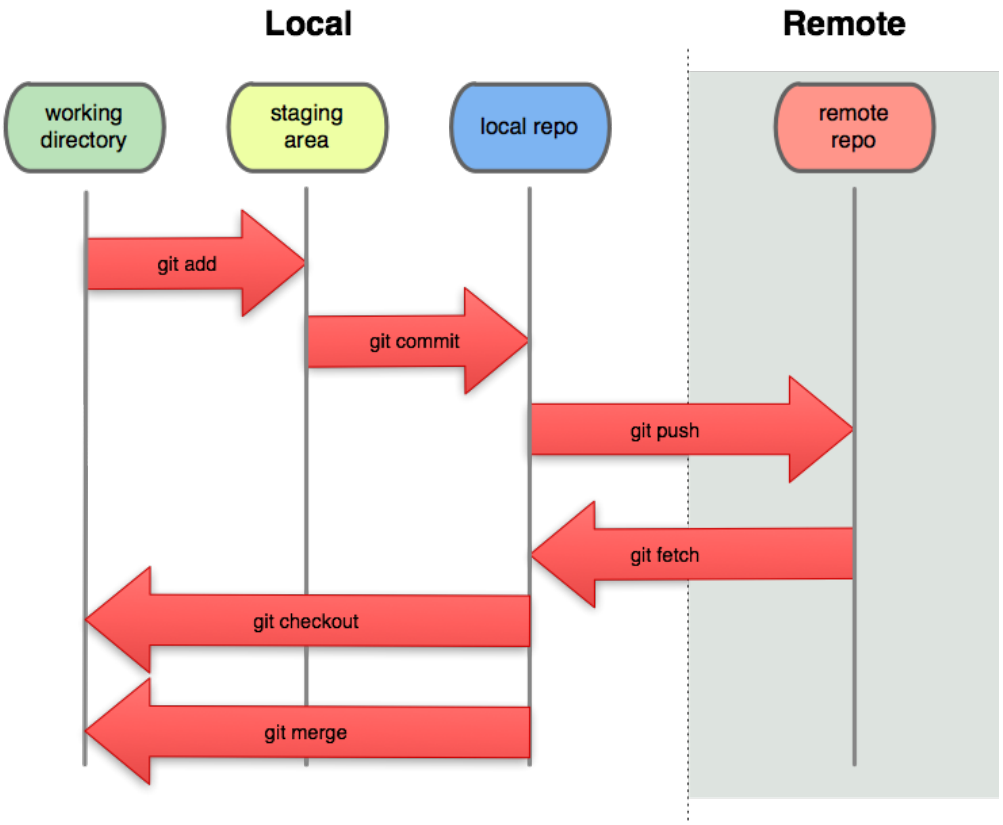

Figure 2.1

- Do **not** totally rely on GUI or IDE. Learn to use Git on command line, which is needed for cluster and cloud computing.

Git survival commands

- `git pull` synchronize local Git directory with remote repository.
- Modify files in local working directory.
- `git add FILES` add snapshots to staging area
- `git commit -m “message”` store snapshots permanently to (**local**) Git repository
- `git push` push commits to remote repository.

Git basic usage

Working with your local copy.

- `git pull`: update local Git repository with remote repository (fetch + merge).
- `git log FILENAME`: display the current status of working directory.

- `git diff` : show differences (by default difference from the most recent commit).
- `git add file1 file2 ...` : add file(s) to the staging area.
- `git commit` : commit changes in staging area to Git directory.
- `git push` : publish commits in local Git repository to remote repository.
- `git reset –soft HEAD 1` : undo the last commit.
- `git checkout FILENAME` : go back to the last commit, discarding all changes made.
- `git rm FILENAME` : remove files from git control.

Vector and vector space

(from JM Appendix A)

- A set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are *linearly dependent* if there exist coefficients c_j for $j = 1, 2, \ldots, n$ such that $\sum_{j=1}^n c_j \mathbf{x}_j = \mathbf{0}$ and $||\mathbf{c}||_2 = \sum_{j=1}^n c_j^2 > 0$. They are *linearly independent* if $\sum_{j=1}^n c_j \mathbf{x}_j = \mathbf{0}$ implies $c_j = 0$ for all j.
- Two vectors are *orthogonal* to each other, written $\mathbf{x} \perp \mathbf{y}$, if their inner product is 0, that is $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} = \sum_j x_j y_j = 0$.
- A set of vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$ are mutually orthogonal iff $\mathbf{x}^{(i)T}\mathbf{x}^{(j)} = 0$ for $\forall i \neq j$.
- The most common set of vectors that are mutually orthogonal are the *elementary* vectors $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}$, which are all zero, except for one element equal to 1, so that $\mathbf{e}_i^{(i)} = 1$ and $\mathbf{e}_j^{(i)} = 0, \forall j \neq i$.
- A *vector space* $\mathcal{S}$ is a set of vectors that are closed under addition and scalar multiplication, that is
  - if $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are in $\mathcal{S}$, then $c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)}$ is in $\mathcal{S}$.
- A vector space $\mathcal{S}$ is *generated* or *spanned* by a set of vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$, written as $\mathcal{S} = \text{span}\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$, if any vector $\mathbf{x}$ in the vector space is a linear combination of $\mathbf{x}_i, i = 1, 2, \ldots, n$.
- A set of linearly independent vectors that generate or span a space $\mathcal{S}$ is called a *basis* of $\mathcal{S}$.

Example A.1

Let

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \text{ and } \mathbf{x}^{(3)} = \begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix}.$$

Then $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are linearly independent, but $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, and $\mathbf{x}^{(3)}$ are linearly dependent since $5\mathbf{x}^{(1)} - 2\mathbf{x}^{(2)} + \mathbf{x}^{(3)} = 0$

Rank

Some matrix concepts arise from viewing columns or rows of the matrix as vectors. Assume $\mathbf{A} \in \mathbb{R}^{m \times n}$.

- $\text{rank}(\mathbf{A})$ is the maximum number of linearly independent rows or columns of a matrix.
- $\text{rank}(\mathbf{A}) \leqslant \min\{m, n\}$.
- A matrix is *full* rank if $\text{rank}(\mathbf{A}) = \min\{m, n\}$. It is *full row rank* if $\text{rank}(\mathbf{A}) = m$. It is *full column rank* if $\text{rank}(\mathbf{A}) = n$.

- a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is *singular* if $\text{rank}(\mathbf{A}) < n$ and *non-singular* if $\text{rank}(\mathbf{A}) = n$.
- $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T) = \text{rank}(\mathbf{A}^T\mathbf{A}) = \text{rank}(\mathbf{A}\mathbf{A}^T)$. (Show this in HW.)
- $\text{rank}(\mathbf{AB}) \leqslant \min\{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\}$. (Hint: Columns of $\mathbf{AB}$ are spanned by columns of A and rows of of $\mathbf{AB}$ are spanned by rows of B.)
- if $\mathbf{Ax} = \mathbf{0}_m$ for some $\mathbf{x} \neq \mathbf{0}_n$, then $\text{rank}(\mathbf{A}) \leqslant n - 1$.

3 Lecture 3:Jan 25

Last time

- Git
- Linear algebra: vector and vector space, rank of a matrix

Today

- Column space and Nullspace (JM Appendix A)
- Simple Linear Regression (JF Chapter 5)

Column space

Definition: The column space of a matrix, denoted by $C(\mathbf{A})$ is the vector space spanned by the columns of the matrix, that is,

$$C(\mathbf{A}) = \{\mathbf{x} : \text{ there exists a vector } \mathbf{c} \text{ such that } \mathbf{x} = \mathbf{A}\mathbf{c}\}.$$

This means that if $\mathbf{x} \in C(\mathbf{A})$, we can find coefficients c_j such that

$$\mathbf{x} = \sum_j c_j \mathbf{a}^{(j)}$$

where $\mathbf{a}^{(j)} = \mathbf{A}_{\cdot j}$ denotes the jth column of matrix $\mathbf{A}$.

- The column space of a matrix consists of all vectors formed by multiplying that matrix by any vector.
- The number of basis vectors for $C(\mathbf{A})$ is then the number of linearly independent columns of the matrix $\mathbf{A}$, and so, $\dim(C(\mathbf{A})) = \text{rank}(\mathbf{A})$.
- The dimension of a space is the number of vectors in its basis.

Example A.2

Let $\mathbf{A} = \begin{bmatrix} 1 & 1 & -3 \\ 1 & 2 & -1 \\ 1 & 3 & 1 \\ 1 & 4 & 3 \end{bmatrix}$ and $\mathbf{c} = \begin{bmatrix} 5 \\ 4 \\ 3 \end{bmatrix}$. Show that $\mathbf{Ac}$ is a linear combination of columns in $\mathbf{A}$.

solution:

$$\mathbf{Ac} = \begin{bmatrix} 1 \times 5 + 1 \times 4 + (-3) \times 3 \\ 1 \times 5 + 2 \times 4 + (-1) \times 3 \\ 1 \times 5 + 3 \times 4 + 1 \times 3 \\ 1 \times 5 + 4 \times 4 + 3 \times 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 10 \\ 20 \\ 30 \end{bmatrix}.$$

You could recognize that

$$\mathbf{Ac} = 5 \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + 4 \times \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} + 3 \times \begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix} = 5\mathbf{a}^{(1)} + 4\mathbf{a}^{(2)} + 3\mathbf{a}^{(3)} = \begin{bmatrix} 0 \\ 10 \\ 20 \\ 30 \end{bmatrix}.$$

Result A.1

$\text{rank}(\mathbf{AB}) \leqslant \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$.

proof: Each column of $\mathbf{AB}$ is a linear combination of columns of $\mathbf{A}$ (i.e. $(\mathbf{AB})_{\cdot j} = \mathbf{Ab}^{(j)}$), so the number of linearly independent columns of $\mathbf{AB}$ cannot be greater than that of A. Similarly, $\text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{B}^T\mathbf{A}^T)$, the same argument gives $\text{rank}(\mathbf{B}^T)$ as an upper bound.

Result A.2

- (a) If $\mathbf{A} = \mathbf{BC}$, then $C(\mathbf{A}) \subseteq C(\mathbf{B})$.
- (b) If $C(\mathbf{A}) \subseteq C(\mathbf{B})$, then there exists a matrix $\mathbf{C}$ such that $\mathbf{A} = \mathbf{BC}$.

proof: For (a), any vector $\mathbf{x} \in C(\mathbf{A})$ can be written as $\mathbf{x} = \mathbf{Ad} = \mathbf{B}(\mathbf{Cd})$.
For (b), $\mathbf{A}_{\cdot j} \in C(B)$, so that there exists a vector $\mathbf{c}^{(j)}$ such that $\mathbf{A}_{\cdot j} = \mathbf{Bc}^{(j)}$. The matrix $\mathbf{C} = (\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(n)})$ satisfies that $\mathbf{A} = \mathbf{BC}$.

Null space

Definition: The null space of a matrix, denoted by $\mathcal{N}(\mathbf{A})$, is $\mathcal{N}(\mathbf{A}) = \{\mathbf{y} : \mathbf{Ay} = \mathbf{0}\}$.

Result A.3

If $\mathbf{A}$ has full-column rank, then $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$.

proof: Matrix $\mathbf{A}$ has full-column rank means its columns are linearly independent, which means that $\mathbf{Ac} = \mathbf{0}$ implies $\mathbf{c} = \mathbf{0}$.

Theorem A.1

Assume $\mathbf{A} \in \mathbb{R}^{m \times n}$, then $\dim(C(\mathbf{A})) = r$ and $\dim(\mathcal{N}(\mathbf{A})) = n - r$, where $r = \text{rank}(\mathbf{A})$.

See JM Appendix Theorem A.1 for the proof.
Interpretation: "dimension of column space + dimension of null space = # columns"
*Mis*Interpretation: Columns space and null space are orthogonal complement to each other. They are of different orders in general! Next result gives the correct statement.

Simple linear regression

Figure 3.1 shows Davis's data on the measured and reported weight in kilograms of 101 women who were engaged in regular exercise.

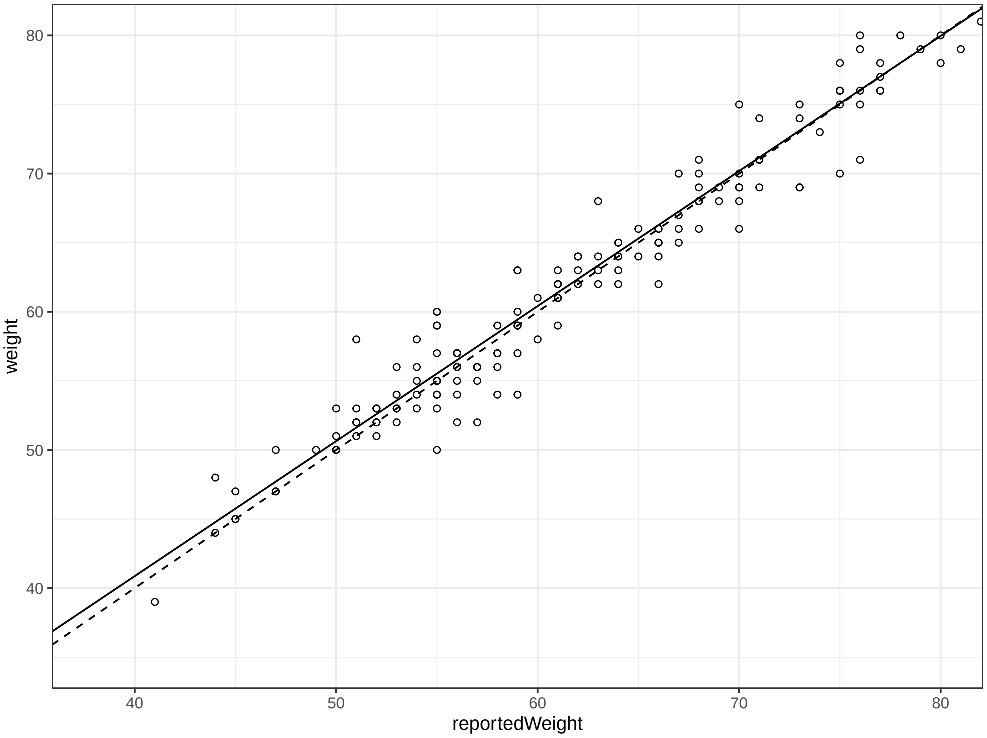

Figure 3.1: Scatterplot of Davis's data on the measured and reported weight of 101 women. The dashed line gives $y = x$.

It's reasonable to assume that the relationship between measured and reported weight appears to be linear. Denote:

- measured weight by y_i: **response variable** or **dependent variable**
- reported weight by x_i: **predictor variable** or **independent variable**
- intercept: β_0
- slope: β_1
- residual/error term ϵ_i.

Then the simple linear regression model writes:

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

For given $(\hat{\beta}_0, \hat{\beta}_1)$ values, the *fitted value* or *predicted value* for observation i is:

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

Therefore, the residual is

$$\epsilon_i = y_i - \hat{y}_i$$

Fitting a linear model

Choose the "best" values for β_0, β_1 such that

$$SS[E] = \sum_1^n \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big)^2 = \sum_1^n (y_i - \hat{y}_i)^2 = \sum_1^n \epsilon_i^2$$

is minimized. These are **least squares** (LS) estimates:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \tfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}. \end{aligned}$$

Definition: The line satisfying the equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

is called the linear regression of y on x which is also called the least squares line.

For Davis's data, we have

$$\begin{aligned} n &= 101 \\ \bar{y} &= \frac{5780}{101} = 57.228 \\ \bar{x} &= \frac{5731}{101} = 56.743 \\ \sum (x_i - \bar{x})(y_i - \bar{y}) &= 4435.9 \\ \sum (x_i - \bar{x})^2 &= 4539.3, \end{aligned}$$

so that

$$\begin{aligned} \hat{\beta}_1 &= \frac{4435.9}{4539.3} = 0.97722 \\ \hat{\beta}_0 &= 57.228 - 0.97722 \times 56.743 = 1.7776 \end{aligned}$$

4 Lecture 4:Jan 27

Last time

- Column space and Nullspace (JM Appendix A)
- Simple Linear Regression (JF Chapter 5)

Today

- HW1 posted, due Feb 12th
- Simple Linear Regression (JF Chapter 5)

Least squares estimates

The simple linear regression (SLR) model writes:

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

The least squares estimates minimizes the sum of squared error (SSE) which is

$$SS[E] = \sum_1^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2 = \sum_1^n (y_i - \hat{y}_i)^2 = \sum_1^n \epsilon_i^2.$$

The **least squares** (LS) estimates (in vector form):

$$\hat{\beta}_{ls} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{pmatrix}.$$

Definition: The line satisfying the equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

is called the linear regression of y on x which is also called the least squares line.

SLR Model in Matrix Form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Jargons

- $\mathbf{X}$ is called the *design matrix*
- β is the vector of parameters
- ϵ is the error vector
- $\mathbf{Y}$ is the response vector.

The Design Matrix

$$\mathbf{X}_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Vector of Parameters

$$\beta_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Vector of Error terms

$$\epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vector of Responses

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Gramian Matrix

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}$$

Therefore, we have

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

Assume the Gramian matrix has full rank (which actually should be the case, why?), we want to show that

$$\hat{\beta}_{ls} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The inverse of the Gramian matrix is

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum_i (x_i - \bar{x})^2}\left[\begin{array}{cc} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{array}\right]$$

Now we have

$$\begin{aligned} \hat{\beta}_{ls} =& (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ =& \frac{1}{n\sum_i (x_i - \bar{x})^2}\left[\begin{array}{cc} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{array}\right]\left[\begin{array}{c} \mathbf{1}_n^T \\ \mathbf{x}^T \end{array}\right]\mathbf{y} \\ =& \frac{1}{n\sum_i (x_i - \bar{x})^2}\left[\begin{array}{cc} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{array}\right]\left[\begin{array}{c} \sum_i y_i \\ \sum_i x_i y_i \end{array}\right] \\ =& \frac{1}{n\sum_i (x_i - \bar{x})^2}\left[\begin{array}{c} (\sum_i x_i^2)(\sum_i y_i) - (\sum_i x_i)(\sum_i x_i y_i) \\ n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i) \end{array}\right] \\ =& \left[\begin{array}{c} \bar{y} - \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\ \bar{x} \\ \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{array}\right] \end{aligned}$$

Some properties:

- (a) $\sum x_i \hat{\epsilon}_i = 0$.
- (b) $\sum \hat{y}_i \hat{\epsilon}_i = 0$ (HW1).

Proof: For (a), we look at

$$\begin{aligned} & \mathbf{X}^T\hat{\epsilon} \\ =& \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ =& \mathbf{X}^T[\mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\ =& \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ =& \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{Y} \\ =& \mathbf{0} \end{aligned}$$

Other quantities in Matrix Form

Fitted values

$$\hat{\mathbf{Y}} = \left[\begin{array}{c} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{array}\right] = \left[\begin{array}{c} \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ \hat{\beta}_0 + \hat{\beta}_1 x_2 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_n \end{array}\right] = \left[\begin{array}{cc} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{array}\right]\left[\begin{array}{c} \hat{\beta}_0 \\ \hat{\beta}_1 \end{array}\right] = X\hat{\beta}$$

Hat matrix

$$\begin{aligned} \hat{\mathbf{Y}} =& \mathbf{X}\hat{\beta} \\ \hat{\mathbf{Y}} =& \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ \hat{\mathbf{Y}} =& \mathbf{H}\mathbf{Y} \end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called "hat matrix" because it turns $\mathbf{Y}$ into $\hat{\mathbf{Y}}$.

Davis's data example

For Davis's data, we have

$$\begin{aligned} n &= 101 \\ \bar{y} &= \frac{5780}{101} = 57.228 \\ \bar{x} &= \frac{5731}{101} = 56.743 \\ \sum(x_i - \bar{x})(y_i - \bar{y}) &= 4435.9 \\ \sum(x_i - \bar{x})^2 &= 4539.3, \end{aligned}$$

so that

$$\begin{aligned} \hat{\beta}_1 &= \frac{4435.9}{4539.3} = 0.97722 \\ \hat{\beta}_0 &= 57.228 - 0.97722 \times 56.743 = 1.7776 \end{aligned}$$

Figure 4.1 shows Davis's data on the measured and reported weight in kilograms of 101 women who were engaged in regular exercise.

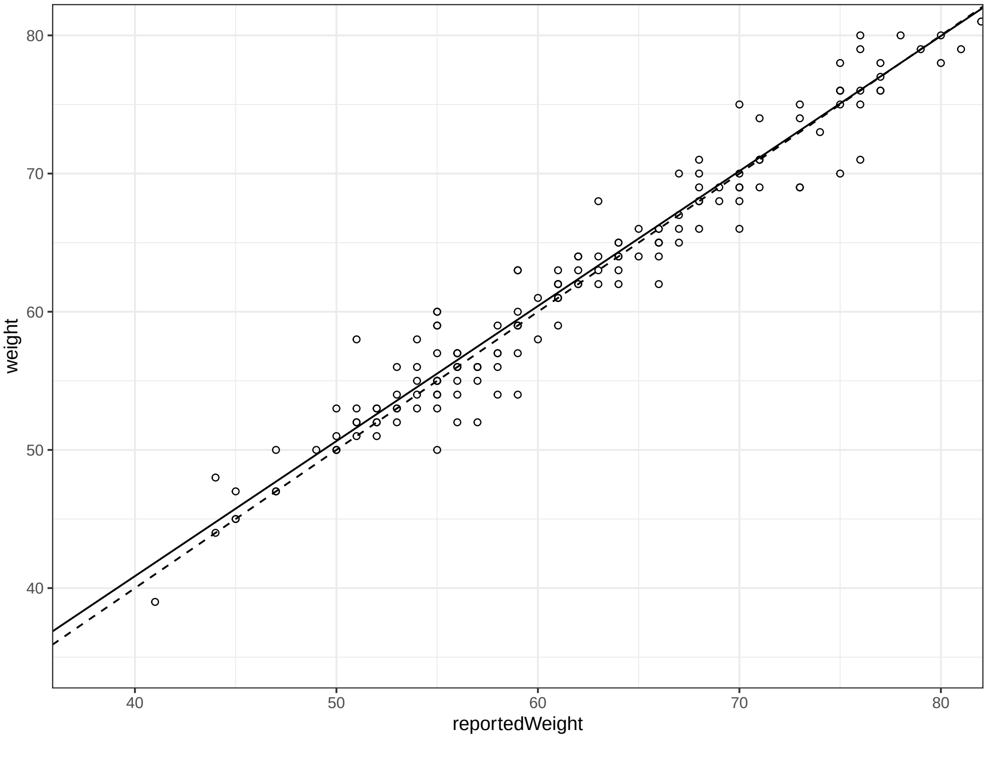

Figure 4.1: Scatterplot of Davis's data on the measured and reported weight of 101 women. The dashed line gives $y = x$. The solid line gives the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

6 Lecture 6:Feb 1

Last time

- SLR in Matrix Form

Today

- Simple correlation
- The statistical model of the SLR (JF chapter 6)

Simple correlation

Having calculated the least squares line, it is of interest to determine how closely the line fits the scatter of points. There are many ways of answering it. The standard deviation of the residuals, S_E, often called the *standard error of the regression* or the *residue standard error*, provides one sort of answer. Because of estimation considerations, the variance of the residuals is defined using *degrees of freedom* $n-2$:

$$S_\epsilon^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2}.$$

The residual standard error is,

$$S_\epsilon = \sqrt{\frac{\sum \hat{\epsilon}_i^2}{n-2}}$$

For the Davis's data, the sum of squared residuals is $\sum \epsilon_i^2 = 418.87$, and thus the standard error of the regression is

$$S_\epsilon = \sqrt{\frac{418.87}{101-2}} = 2.0569\text{kg}.$$

On average, using the least-squares regression line to predict measured weight from reported weight results in an error of about 2 kg.

Sum of squares:

- Total sum of squares (TSS) for Y: TSS $= \sum(y_i - \bar{y})^2$
- Residual sum of squares (RSS): RSS $= \sum(y_i - \hat{y}_i)^2$
- regression sum of squares (RegSS): RegSS = TSS − RSS $= \sum(\hat{y}_i - \bar{y})^2$
- RegSS + RSS = TSS

Sample correlation coefficient

Definition: The sample correlation coefficient r_{xy} of the paired data (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) is defined by

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{\sqrt{\sum (x_i - \bar{x})^2/(n-1) \times \sum (y_i - \bar{y})^2/(n-1)}} = \frac{s_{xy}}{s_x s_y}$$

s_{xy} is called the sample covariance of x and y:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$s_x = \sqrt{\sum (x_i - \bar{x})^2/(n-1)}$ and $s_y = \sqrt{\sum (y_i - \bar{y})^2/(n-1)}$ are, respectively, the sample standard deviations of X and Y.

Some properties of r_{xy}:

- r_{xy} is a measure of the linear association between x and y in a dataset.
- correlation coefficients are always between -1 and 1:
$$-1 \leqslant r_{xy} \leqslant 1$$
- The closer r_{xy} is to 1, the stronger the positive linear association between x and y
- The closer r_{xy} is to -1, the stronger the negative linear association between x and y
- The bigger $|r_{xy}|$, the stronger the linear association
- If $|r_{xy}| = 1$, then x and y are said to be perfectly correlated.
- $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}$

R-square

The ratio of RegSS to TSS is called the *coefficient of determination*, or sometimes, simply “r-square”. it represents the proportion of variation observed in the response variable y which can be “explained” by its linear association with x.

- In simple linear regression, “r-square” is in fact equal to r_{xy}^2. (But this isn't the case in multiple regression.)
- It is also equal to the squared correlation between y_i and $\hat{y}_i$. (This is the case in multiple regression.)

For Davis's regression of measured on reported weight:

$$\begin{aligned} \text{TSS} &= 4753.8 \\ \text{RSS} &= 418.87 \\ \text{RegSS} &= 4334.9 \end{aligned}$$

Thus,

$$r^2 = \frac{4334.9}{4753.8} = 1 - \frac{418.87}{4753.8} = 0.9119$$

The statistical model of Simple Linear Regress

Standard statistical inference in simple regression is based on a *statistical model* that describes the population or process that is sampled:

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

where the coefficients β_0 and β_1 are the *population regression parameters*. The data are randomly sampled from some population of interest.

- y_i is the value of the response variable
- x_i is the explanatory variable
- ϵ_i represents the aggregated omitted causes of y (i.e., the causes of y beyond the explanatory variable), other explanatory variables that could have been included in the regression model, measurement error in y, and whatever component of y is inherently random.

Key assumptions of SLR

The key assumptions of the SLR model concern the behavior of the errors, equivalently, the distribution of y conditional on x:

- *Linearity.* The expectation of the error given the value of x is 0: $\mathbf{E}\left(\epsilon\right) \equiv \mathbf{E}\left(\epsilon|x_i\right) = 0$. And equivalently, the expected value of the response variable is a linear function of the explanatory variable: $\mu_i \equiv \mathbf{E}\left(y_i\right) \equiv \mathbf{E}\left(y_i|x_i\right) = \mathbf{E}\left(\beta_0 + \beta_1 x_i + \epsilon_i|x_i\right) = \beta_0 + \beta_1 x_i$.
- *Constant variance.* The variance of the errors is the same regardless of the value of x: $\mathbf{Var}\left(\epsilon|x_i\right) = \sigma_\epsilon^2$. The constant error variance implies constant conditional variance of y on given x: $\mathbf{Var}\left(y|x_i\right) = \mathbf{E}\left((y_i - \mu_i)^2\right) = \mathbf{E}\left((y_i - \beta_0 - \beta_1 x_i)^2\right) = \mathbf{E}\left(\epsilon_i^2\right) = \sigma_\epsilon^2$. (Question: why the last equal sign?)
- *Normality.* The errors are independent identically distributed with Normal distribution with mean 0 and variance σ_ϵ^2. Write as $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. Equivalently, the conditional distribution of the response variable is normal: $y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2)$.
- *Independence.* The observations are sampled independently.
- *Fixed X, or X measured without error and independent of the error.*
  - For experimental research where X values are under direct control of the researcher (i.e. X's are fixed). If the experiment were replicated, then the values of X would remain the same.
  - For research where X values are sampled, we assume the explanatory variable is measured without error and the explanatory variable and the error are independent in the population from which the sample is drawn.
- *X is not invariant.* X's can not be all the same.

Figure 6.1 shows the assumptions of linearity, constant variance, and normality in SLR model.

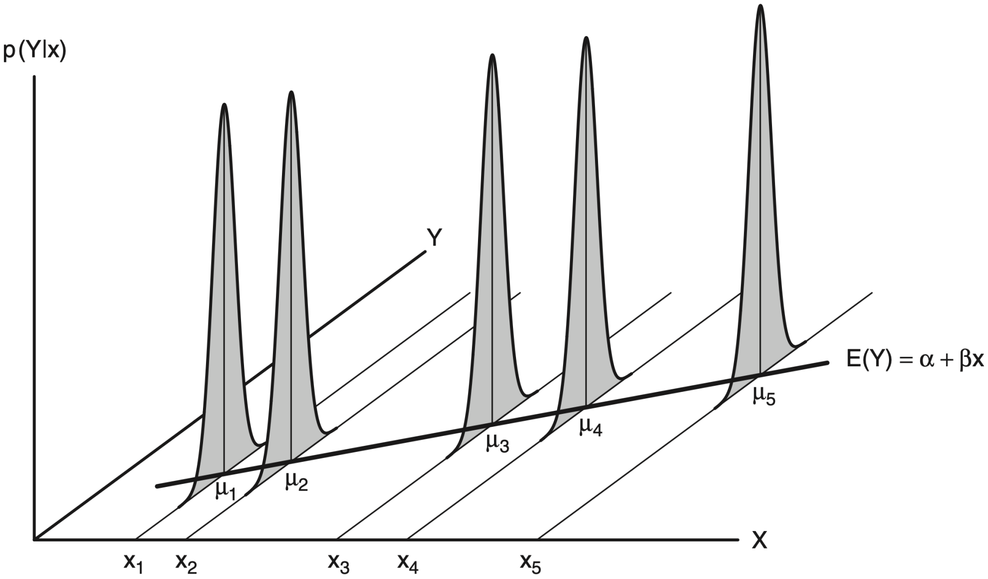

Figure 6.1: The assumptions of linearity, constant variance, and normality in simple regression. The graph shows the conditional population distributions $\Pr(Y|x)$ of Y for several values of the explanatory variable X, labeled as $x_1, x_2, \ldots, x_5$. The conditional means of Y given x are denoted $\mu_1, \ldots, \mu_5$.

7 Lecture 7: Feb 3

Last time

- Statistical model of SLR

Today

- Properties of the LS estimators
- Inference of SLR model

Properties of the Least-Squares estimator

Under the strong assumptions of the simple regression model, the sample least squares coefficients $\hat{\beta}_{ls}$ have several desirable properties as estimators of the population regression coefficients β_0 and β_1:

- The least-squares intercept and slope are *linear estimators*, in the sense that they are linear functions of the observations y_i.
  Proof:
  method (a) $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
  method (b) $\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2} - \frac{\sum(x_i-\bar{x})\bar{y}}{\sum(x_i-\bar{x})^2} = \sum \frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} y_i = \sum k_i y_i$ where $k_i = \frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$
  and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

- The sample least-squares coefficients are *unbiased estimators* of the population regression coefficients:

$$\mathbf{E}\left(\hat{\beta}_0\right) = \beta_0$$
$$\mathbf{E}\left(\hat{\beta}_1\right) = \beta_1$$

  Proof:
  method (a) $\mathbf{E}\left(\hat{\beta}\right) = \mathbf{E}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\right) = \mathbf{E}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta\right) = \beta$. (note: $\mathbf{E}(Y) = \mathbf{E}(\mathbf{X}\beta + \epsilon) = \mathbf{E}(\mathbf{X}\beta) + \mathbf{E}(\epsilon) = \mathbf{X}\beta$)
  method (b) recall that $\hat{\beta}_1 = \sum k_i y_i$ where $k_i = \frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$. First, we want to show

  1. $\sum k_i = 0$
  2. $\sum k_i x_i = 1$

  They are actually quite easy: $\sum k_i = \sum_i \frac{(x_i-\bar{x})}{\sum_j(x_j-\bar{x})^2} = \frac{(\sum_i x_i) - n\bar{x}}{\sum_j(x_j-\bar{x})^2} = 0$, and $\sum k_i x_i = \sum_i \frac{(x_i-\bar{x})x_i}{\sum_j(x_j-\bar{x})^2} = \frac{(\sum_i x_i^2) - \bar{x}(\sum_i x_i)}{\sum_j(x_j-\bar{x})^2} = \frac{(\sum_i x_i^2) - n\bar{x}^2}{\sum_j(x_j-\bar{x})^2} = 1$.
  Now $\mathbf{E}\left(\hat{\beta}_1\right) = \mathbf{E}\left(\sum k_i y_i\right) = \sum[k_i \mathbf{E}(y_i)] = \sum[k_i(\beta_0 + \beta_1 x_i)] = \beta_0 \sum k_i + \beta_1 \sum(k_i x_i) = \beta_1$, and $\mathbf{E}\left(\hat{\beta}_0\right) = \mathbf{E}\left(\bar{y} - \hat{\beta}_1\bar{x}\right) = \mathbf{E}(\bar{y}) - \bar{x}\mathbf{E}\left(\hat{\beta}_1\right) = \mathbf{E}\left(\frac{1}{n}\sum y_i\right) - \bar{x}\beta_1 = \frac{1}{n}[\sum \mathbf{E}(y_i)] - \bar{x}\beta_1 = \frac{1}{n}\sum[\beta_0 + x_i\beta_1] - \bar{x}\beta_1 = \beta_0$

- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ have simple sampling variances:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) =& \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} \\ \mathrm{Var}(\hat{\beta}_1) =& \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}\end{aligned}$$

Proof:
$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\sum \mathrm{k_i y_i}) = \sum \mathrm{k_i^2 Var(y_i)} = \sigma_\epsilon^2 \sum \mathrm{k_i^2} = \sigma_\epsilon^2 \frac{\sum_\mathrm{i} \mathrm{(x_i - \bar{x})^2}}{[\sum_\mathrm{j} \mathrm{(x_j - \bar{x})^2}]^2} = \frac{\sigma_\epsilon^2}{\sum \mathrm{(x_i - \bar{x})^2}}$, and $\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var(\bar{y}} - \hat{\beta}_1 \mathrm{\bar{x}) = Var(\bar{y}) + (\bar{x})^2 Var(}\hat{\beta}_1) - 2\mathrm{\bar{x} Cov(\bar{Y}}, \hat{\beta}_1)$.
Now,

$$\begin{aligned}\mathrm{Var}(\bar{y}) &= \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(y_i) = \frac{\sigma^2}{n}, \\ \mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2},\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) &= \mathrm{Cov}\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i, \frac{\sum_{j=1}^{n} (x_j - \bar{x}) Y_j}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right\} \\ &= \frac{1}{n} \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \mathrm{Cov}\left\{ \sum_{i=1}^{n} Y_i, \sum_{j=1}^{n} (x_j - \bar{x}) Y_j \right\} \\ &= \frac{1}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} (x_j - \bar{x}) \sum_{j=1}^{n} \mathrm{Cov}(Y_i, Y_j) \\ &= \frac{1}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} (x_j - \bar{x}) \sigma^2 \\ &= 0.\end{aligned}$$

Finally,

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) &= \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ &= \frac{\sigma^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \left\{ \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2 \right\} \\ &= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}.\end{aligned}$$

- Rewrite the formula for $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\mathrm{(n-1)S_X^2}}$, we see that the sampling variance of the slope estimate will be small when
  - The error variance σ_ϵ^2 is small
  - The sample size n is large

  - The explanatory-variable values are spread out (i.e. have a large variance, S_X^2)

- (Gauss-Markov theorem) Under the assumptions of linearity, constant variance, and independence, the least-squares estimators are BLUE (Best Linear Unbiased Estimator), that is they have the smallest sampling variance and are unbiased. (show this)
  Proof:
  Let $\widetilde{\beta}_1$ be another linear unbiased estimator such that $\widetilde{\beta}_1 = \sum c_i y_i$. For $\widetilde{\beta}_1$ is still unbiased as above, $\mathbf{E}\left(\widetilde{\beta}_1\right) = \beta_0 \sum c_i + \beta_1 \sum c_i x_i = \beta_1$ for all β_1, we have $\sum c_i = 0$ and $\sum c_i x_i = 1$.
  $\mathbf{Var}\left(\widetilde{\beta}_1\right) = \sigma_\epsilon^2 \sum c_i^2$
  Let $c_i = k_i + d_i$, then

$$\begin{aligned}\mathbf{Var}\left(\widetilde{\beta}_1\right) =&\sigma_\epsilon^2 \sum (k_i + d_i)^2 \\ =&\sigma_\epsilon^2 \left[\sum {k_i}^2 + \sum {d_i}^2 + 2\sum k_i d_i\right] \\ =&\mathbf{Var}\left(\hat{\beta}_1\right) + \sigma_\epsilon^2 \sum d_i^2 + 2\sigma_\epsilon^2 \sum k_i d_i\end{aligned}$$

  Now we show the last term is 0 to finish the proof.

$$\begin{aligned}\sum k_i d_i &= \sum k_i(c_i - k_i) = \sum c_i k_i - \sum k_i^2 \\ &= \sum_i \left[c_i \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}\right] - \frac{1}{\sum_i (x_i - \bar{x})^2} \\ &= 0\end{aligned}$$

- Under the full suite of assumptions, the least-squares coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximum-likelihood estimators of β_0 and β_1. (show this)
  Proof:
  The log likelihood under the full suite of assumptions is $\ell = -\log\left[(2\pi)^{\frac{n}{2}}\sigma_\epsilon^n\right] - \frac{1}{2\sigma_\epsilon^2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$. Maximizing the likelihood is equivalent as minimizing $(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \epsilon^T\epsilon$ which is the SSE.

- Under the assumption of normality, the least-squares coefficients are themselves normally distributed. Summing up,

$$\begin{aligned}\hat{\beta}_0 \sim& N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2}) \\ \hat{\beta}_1 \sim& N(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2})\end{aligned}$$

8 Lecture 8: Feb 5

Last time

- Properties of the LS estimators

Today

- Inference of SLR model
- Lab 1

Statistical inference of the SLR model

Now we have the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\begin{aligned}\hat{\beta}_0 \sim& N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n\sum(x_i-\bar{x})^2})\\ \hat{\beta}_1 \sim& N(\beta_1, \frac{\sigma_\epsilon^2}{\sum(x_i-\bar{x})^2}).\end{aligned}$$

However, σ_ϵ is never known in practice. Instead, an *unbiased* estimator of σ_ϵ^2 is given by

$$\hat{\sigma_\epsilon}^2 = MS[E] = \frac{SS[E]}{n-2}.$$

Proof:

$$MS[E] = \frac{\sum(y_i-\hat{y}_i)^2}{n-2},$$

we want to show $\mathbf{E}\left(\sum(y_i-\hat{y}_i)^2\right) = \sigma_\epsilon^2(n-2)$.
LHS: $\mathbf{E}\left(\sum(y_i-\hat{y}_i)^2\right) = \sum_i\left[\mathbf{E}\left(y_i-\hat{y}_i\right)^2\right]$

and $\mathrm{E}[(y_i-\hat{y}_i)^2] = \mathrm{Var}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i}) + [\mathbf{E}\,(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})]^2 = \mathrm{Var}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i}) = \mathrm{Var}(\mathrm{y_i}) + \mathrm{Var}(\hat{\mathrm{y}}_\mathrm{i}) - 2\mathrm{cov}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i})$

$$\begin{aligned}
\mathrm{Var}(\mathrm{y_i}) &= \sigma_\epsilon^2 \\
\mathrm{Var}(\hat{\mathrm{y}}_\mathrm{i}) &= \mathrm{Var}(\bar{\mathrm{y}} + \hat{\beta}_1(\mathrm{x_i}-\bar{\mathrm{x}})) \\
&= \mathrm{Var}(\bar{\mathrm{y}}) + (\mathrm{x_i}-\bar{\mathrm{x}})^2\mathrm{Var}(\hat{\beta}_1) + 2(\mathrm{x_i}-\bar{\mathrm{x}})\mathrm{Cov}(\bar{\mathrm{y}}, \hat{\beta}_1) \\
\mathrm{Cov}(\bar{y}, \hat{\beta}_1) &= \mathrm{Cov}(\bar{y}, \sum k_i y_i) \\
&= \sum_i \mathrm{Cov}(\bar{y}, k_i y_i) \\
&= \sum_i \frac{k_i}{n}\mathrm{Var}(\mathrm{y_i}) \\
&= \frac{1}{n}\sum k_i \\
&= 0 \\
\therefore\ \mathrm{Var}(\hat{\mathrm{y}}_\mathrm{i}) &= \mathrm{Var}(\bar{\mathrm{y}}) + (\mathrm{x_i}-\bar{\mathrm{x}})^2\mathrm{Var}(\hat{\beta}_1) \\
&= \frac{1}{n}\sigma_\epsilon^2 + \frac{\sigma_\epsilon^2(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2} \\
&= \sigma_\epsilon^2\left[\frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]
\end{aligned}$$

Now, we derive the last term $\mathrm{cov}(y_i, \hat{y}_i)$:

$$\begin{aligned}
\mathrm{cov}(y_i, \hat{y}_i) &= \mathrm{cov}(y_i, \bar{y} + \hat{\beta}_1(x_i-\bar{x})) \\
&= \mathrm{cov}(y_i, \frac{1}{n}\sum_j y_j + (x_i-\bar{x})\sum_j k_j y_j) \\
&= \mathrm{cov}(y_i, \sum_j\left[\frac{1}{n} + (x_i-\bar{x})k_j\right]y_j) \\
&= \sigma_\epsilon^2\left[\frac{1}{n} + (x_i-\bar{x})k_i\right] \\
&= \sigma_\epsilon^2\left[\frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]
\end{aligned}$$

Therefore, we have for ith residue

$$\begin{aligned}
\mathrm{Var}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i}) &= \mathrm{Var}(\mathrm{y_i}) + \mathrm{Var}(\hat{\mathrm{y}}_\mathrm{i}) - 2\mathrm{cov}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}) \\
&= \sigma_\epsilon^2 + \sigma_\epsilon^2\left[\frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right] - 2\sigma_\epsilon^2\left[\frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right] \\
&= \sigma_\epsilon^2\left[1 - \frac{1}{n} - \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right].
\end{aligned}$$

And finally, sum over i we get

$$\sum_i \mathrm{Var}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i}) = \sigma_\epsilon^2\sum_\mathrm{i}\left[1 - \frac{1}{\mathrm{n}} - \frac{(\mathrm{x_i}-\bar{\mathrm{x}})^2}{\sum(\mathrm{x_i}-\bar{\mathrm{x}})^2}\right] = (\mathrm{n}-2)\sigma_\epsilon^2$$

Confidence intervals

Now we substitute $\hat{\sigma}_\epsilon^2$ into the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\begin{aligned}\hat{\beta}_1 &\sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum(x_i - \bar{x})^2}) \\ \hat{\beta}_0 &\sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum(x_i - \bar{x})^2})\end{aligned}$$

to get the estimated standard errors:

$$\begin{aligned}\widehat{SE}(\hat{\beta}_1) &= \sqrt{\frac{MS[E]}{\sum(x_i - \bar{x})^2}} \\ \widehat{SE}(\hat{\beta}_0) &= \sqrt{MS[E]\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)}\end{aligned}$$

And the $100(1-\alpha)\%$ confidence intervals for β_1 and β_0 are given by

$$\hat{\beta}_1 \pm t(n-2, \alpha/2)\sqrt{\frac{MS[E]}{S_{xx}}}$$

$$\hat{\beta}_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

where $S_{xx} = \sum(x_i - \bar{x})^2$

Confidence interval for $\mathbf{E}\left(Y|X = x_0\right)$

The conditional mean $\mathbf{E}\left(Y|X = x_0\right)$ can be estimated by evaluating the regression function $\mu(x_0)$ at the estimates $\hat{\beta}_0$, $\hat{\beta}_1$. The conditional variance of the expression isn't too difficult (already shown):

$$\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_0 | \mathrm{X} = \mathrm{x}_0) = \sigma^2(\frac{1}{\mathrm{n}} + \frac{(\mathrm{x}_0 - \bar{\mathrm{x}})^2}{\mathrm{S}_{\mathrm{xx}}})$$

This leads to a confidence interval of the form

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Prediction interval

Often, prediction of the response variable Y for a given value, say x_0, of the independent variable of interest. In order to make statements about future values of Y, we need to take into account

- the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

- the randomness of a future value Y.

We have seen the predicted value of Y based on the linear regression is given by $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

The 95% prediction interval has the form

$$\hat{Y}_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Hypothesis test

To test the hypothesis $\boxed{H_0 : \beta_1 = \beta_{slope_0}}$ that the population slope is equal to a specific value β_{slope_0} (most commonly, the null hypothesis has $\beta_{slope_0} = 0$), we calculate the test statistic (T-statistics) with $df = n - 2$

$$t_0 = \frac{\hat{\beta}_1 - \beta_{slope_0}}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

9 Lecture 9: Feb 8

Last time

- Inference of SLR model
- Lab 1

Today

- SLR questions
- Multiple Linear Regression

Some questions to answer using regression analysis:

1. What is the meaning, in words, of β_1?
   Answer: β_1 is the population slope parameter of the SLR model that represents the amount of increase in the mean of the response variable with a unit increase of the explanatory variable.
2. True/False: (a) β_1 is a statistic (b) β_1 is a parameter (c) β_1 is unknown.
   Answer: (a) False (b) True (C) True. In reality, the true population parameters are almost never known. However, in simulation studies, we do know them.
3. True/False: (a) $\hat{\beta}_1$ is a statistic (b) $\hat{\beta}_1$ is a parameter (c)$\hat{\beta}_1$ is unknown
   Answer: (a) True (b) False (C) False. $\hat{\beta}_1$ is an estimate of the population parameter β_1.
4. Is $\hat{\beta}_1 = \beta_1$?
   Answer: No. However, $\mathbf{E}\left(\hat{\beta}_1\right) = \beta_1$

Multiple linear regression

JF 5.2+6.2

Multiple linear regression - an example

An example on the prestige, education, and income levels of 45 U.S. occupations (Duncan's data):

	income	education	prestige
accountant	62	86	82
pilot	72	76	83
architect	75	92	90
author	55	90	76
chemist	64	86	90
minister	21	84	87
professor	64	93	93
dentist	80	100	90
reporter	67	87	52
engineer	72	86	88
lawyer	76	98	89
teacher	48	91	73

"prestige" represents the percentage of respondents in a survey who rated an occupation as "good" or "excellent" in prestige, "education" represents the percentage of incumbents in the occupation in the 1950 U.S. Census who were high school graduates, and "income" represents the percentage of occupational incumbents who earned incomes in excess of $3,500.

Using the `pairs` command in R, we can look at the pairwise scatter plot between the three variables as in Figure 9.1.

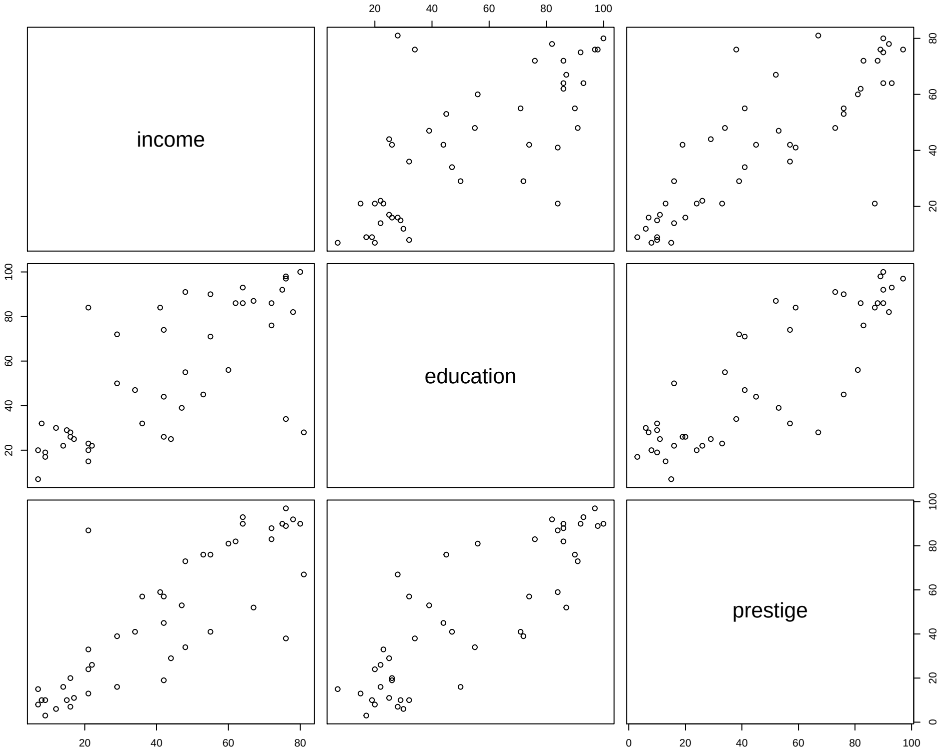

Figure 9.1: Scatterplot matrix for occupational prestige, level of education, and level of income of 45 U.S. occupations in 1950.

Consider a regression model for the "prestige" of occupation i, Y_i, in which the mean of Y_i is a linear function of two predictor variables $X_{i1} = \mathit{income}, X_{i2} = \mathit{education}$ for occupations $i = 1, 2, \ldots, 45$:

$$Y = \beta_0 + \beta_1 \mathit{income} + \beta_2 \mathit{education} + \mathit{error}$$

or

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

or

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \epsilon_2 \\ \vdots &= \vdots \\ Y_{45} &= \beta_0 + \beta_1 X_{45,1} + \beta_2 X_{45,2} + \epsilon_{45} \end{aligned}$$

A multiple linear regression (MLR) model w/ p independent variables

Let p independent variables be denoted by $x_1, \ldots, x_p$.

- Observed values of p independent variables for i^{th} subject from sample denoted by $x_{i1}, \ldots, x_{ip}$
- response variable for i^{th} subject denoted by Y_i
- For $i = 1, \ldots, n$, MLR model for Y_i:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- As in SLR, $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$

Least squares estimates of regression parameters minimize $SS[E]$:

$$SS[E] = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

$$\boxed{\hat{\sigma}^2 = \frac{SS[E]}{n-p-1}}$$

Interpretations of regression parameters:

- σ^2 is unknown <u>error variance</u> parameter
- $\beta_0, \beta_1, \ldots, \beta_p$ are $p+1$ unknown regression parameters:
  - β_0: average response when $x_1 = x_2 = \cdots = x_p = 0$
  - β_i is called a <u>partial slope</u> for x_i. Represents mean change in y per unit increase in x_i *with all other independent variables held fixed.*

Matrix formulation of MLR

Let a $(1 \times (p+1))$ vector for p observed independent variables for individual i be defined by

$$x_{i\cdot} = (1, x_{i1}, x_{i2}, \ldots, x_{ip}).$$

The MLR model for $Y_1, \ldots, Y_n$ is given by

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \epsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \epsilon_2 \\
\vdots &= \vdots \\
Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \epsilon_n
\end{aligned}$$

This system of n equations can be expressed using matrices:

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

where

- $\mathbf{Y}$ denotes a response vector of size $n \times 1$
- $\mathbf{X}$ denotes a design matrix of size $n \times (p+1)$
- β denotes a vector of regression parameters of size $(p+1) \times 1$
- ϵ denotes an error vector of size $n \times 1$

Here, the error vector ϵ is assumed to follow a multivariate normal distribution with variance-covariance matrix $\sigma^2 \mathbf{I}_n$. For individual i,

$$Y_i = x_{i\cdot}\beta + \epsilon_i.$$

Some simplified expressions: ($\mathbf{a}$ is a known $p \times 1$ vector)

$$\boxed{\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
\mathbf{Var}\left(\hat{\beta}\right) &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \mathbf{\Sigma} \\
\widehat{\text{Var}}(\hat{\beta}) &= MS[E](\mathbf{X}^T\mathbf{X})^{-1} \\
&= \widehat{\mathbf{\Sigma}} \\
\widehat{\text{Var}}(\mathbf{a}^T\hat{\beta}) &= \mathbf{a}^T\widehat{\mathbf{\Sigma}}\mathbf{a}
\end{aligned}}$$

Question: what are the dimensions of each of these quantities?

- $(\mathbf{X}^T\mathbf{X})^{-1}$ may be verbalized as " x transposed x inverse"
- $\widehat{\mathbf{\Sigma}}$ is the estimated variance-covariance matrix for the estimate of the regression parameter vector $\hat{\beta}$

- $\mathbf{X}$ is assumed to be of full *rank*.

Some more simplified expressions:

$$\boxed{\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\beta}\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\\ &= \mathbf{H}\mathbf{Y}\\ \hat{\epsilon} &= \mathbf{Y}-\hat{\mathbf{Y}}\\ &= \mathbf{Y}-\mathbf{X}\hat{\beta}\\ &= (\mathbf{I}-\mathbf{H})\mathbf{Y}\end{aligned}}$$

- $\hat{\mathbf{Y}}$ is called the vector of <u>fitted</u> or <u>predicted values</u>
- $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the <u>hat matrix</u>
- $\hat{\epsilon}$ is the vector of <u>residuals</u>

For the Duncan's data example on income, education and prestige, with $p = 2$ independent variables and $n = 45$ observations,

$$\mathbf{X} = \begin{bmatrix} 1 & 62 & 86\\ 1 & 72 & 76\\ \vdots & \vdots & \vdots\\ 1 & 8 & 32 \end{bmatrix}$$

and

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 45 & 1884 & 2365\\ 1884 & 105148 & 122197\\ 2365 & 122197 & 163265 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.10211 & -0.00085 & -0.00084\\ -0.00085 & 0.00008 & -0.00005\\ -0.00084 & -0.00005 & 0.00005 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} -6.0646629\\ 0.5987328\\ 0.5458339 \end{bmatrix} =?$$

$$SS[E] = \epsilon^T\epsilon = (\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}) = 7506.7$$

$$MS[E] = \frac{SS[E]}{df} = \frac{7506.7}{45-2-1} = 178.73$$

$$\widehat{\mathbf{\Sigma}} = MS[E](\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 18.249481 & -0.151845008 & -0.150706025\\ -0.151845 & 0.014320275 & -0.008518551\\ -0.150706 & -0.008518551 & 0.009653582 \end{bmatrix}$$

10 Lecture 10: Feb 10

Last time

- SLR questions
- Multiple Linear Regression

Today

- Multiple correlation
- Confidence intervals and hypothesis tests
- R practice with questions

Multiple correlation, JF 5.2.3

The sums of squares in multiple regression are defined in the same manner as in SLR:

$$\begin{aligned}
TSS &= \sum (Y_i - \bar{Y})^2 \\
RegSS &= \sum (\hat{Y}_i - \bar{Y})^2 \\
RSS &= \sum (Y_i - \hat{Y}_i)^2 = \sum \epsilon_i^2
\end{aligned}$$

Not surprisingly, we have a similar analysis of variance for the regression:

$$TSS = RegSS + RSS$$

The squared multiple correlation R^2, representing the proportion of variation in the response variable captured by the regression, is defined in terms of the sums of squares:

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Because there are several slope coefficients, potentially with different signs, the *multiple correlation coefficient* is, by convention, the positive square root of R^2. The multiple correlation is also interpretable as the simple correlation between the fitted and observed Y values, i.e. $r_{\hat{Y}Y}$.

Adjusted-R^2

Because the multiple correlation can only rise, never decline, when explanatory variables are added to the regression equation (HW1), investigators sometimes penalize the value of R^2 by a "correction" for degrees of freedom. The corrected (or "adjusted") R^2 is defined as:

$$\begin{aligned}
R^2_{adj} =& 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}} \\
=& 1 - \left[\frac{(1 - R^2)(n-1)}{n-p-1} \right]
\end{aligned}$$

Confidence intervals

Confidence intervals and hypothesis tests for individual coefficients closely follow the pattern of simple-regression analysis:

1. substitute an estimate of the error variance (MSE) for the unknown σ^2 into the variance term of $\hat{\beta}_i$
2. find the estimated standard error of a slope coefficient $\widehat{SE}(\hat{\beta}_i)$
3. $t = \frac{\hat{\beta}_i - \beta_i}{\widehat{SE}(\hat{\beta}_i)}$ follows a t-distribution with degrees of freedom as associated with SSE.

Therefore, we can construct the $100(1-\alpha)\%$ confidence interval for a single slope parameter by (why?):

$$\hat{\beta}_i \pm t(n-p-1, \alpha/2)\widehat{SE}(\hat{\beta}_i)$$

Hand-waving proof:
we know that $t = \frac{\hat{\beta}_i - \beta_i}{\widehat{SE}(\hat{\beta}_i)} \sim t_{n-p-1}$, such that

$$\begin{aligned}1 - \alpha &= \Pr\left(-t_c < t < t_c\right) \\ &= \Pr\left(t_c < \frac{\hat{\beta}_i - \beta_i}{\widehat{SE}(\hat{\beta}_i)} < t_c\right) \\ &= \Pr\left(\hat{\beta}_i - t_c \cdot \widehat{SE}(\hat{\beta}_i) < \beta_i < \hat{\beta}_i + t_c \cdot \widehat{SE}(\hat{\beta}_i)\right)\end{aligned}$$

where $t_c = t(n-p-1, \alpha/2)$ is the critical value.

Hypothesis tests

We first test the null hypothesis that all population regression slopes are 0:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

The test statistics,

$$F = \frac{RegSS/p}{RSS/(n-p-1)}$$

follows an F-distribution with p and $n-p-1$ degrees of freedom.

We can also test a null hypothesis about a *subset* of the regression slopes, e.g.,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0.$$

Or more generally, test the null hypothesis

$$H_0 : \beta_{q_1} = \beta_{q_2} = \cdots = \beta_{q_k} = 0$$

where $0 \leqslant q_1 < q_2 < \cdots < q_k \leqslant p$ is a subset of k indices. To get the F-statistic for this case, we generally perform the following steps:

1. Fit the *full* ("unconstrained") model, in other words, model that provides context for H_0. Record SSR_{full} and the associated df_{full}

2. Fit the *reduced* ("constrained") model, in other words, full model constrained by H_0. Record SSR_{red} and the associated df_{red}

3. Calculate the F-statistic by

$$F = \frac{[SSR_{red} - SSR_{full}]/(df_{red} - df_{full})}{SSR_{full}/df_{full}}$$

4. Find p-value (the probability of observing an F-statistic that is at least as high as the value that we obtained) by consulting an F-distribution with numerator $df(ndf) = df_{red} - df_{full}$ and denominator $df(ddf) = df_{full}$. Notation: $F_{ndf,ddf}$, see Figure 10.1.

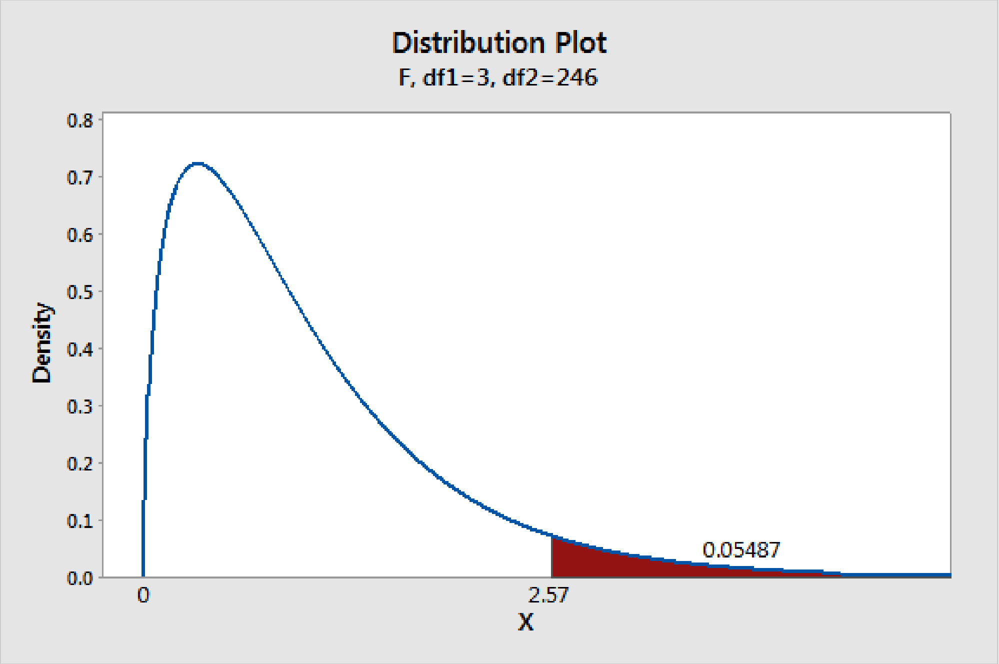

Figure 10.1: An example for p-value for F-statistic value 2.57 with an $F_{3,246}$ distribution

Now, open the `Lecture10_to_fill.Rmd` file and start working on the following questions:

1. What is the estimate of β_1? Interpretation?
   Answer: $\hat{\beta}_1 = 0.60$ (second element of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, "prestige" increase per unit income for occupations with the same level of education)

2. What is the standard error of $\hat{\beta}_1$?
   Answer: $\sqrt{0.014320275} = 0.12$ (square root of middle element of $\widehat{\mathbf{\Sigma}}$)

3. Is $\beta_1 = 0$ plausible, while controlling for possible linear associations between Prestige and Education? ($t(0.025, 42) = 2.02$)
   Answer: $\boxed{H_0 : \beta_1 = 0}$, T-statistic: $t = (\hat{\beta}_1 - 0)/SE(\hat{\beta}_1) = 0.60/0.12 = 5.0 > 2.02$, ("$\hat{\beta}_1$ differs significantly from 0.")

4. Estimate the mean prestige among the population of ALL occupations with $income = 42$ and $education = 84$.
   Answer: Unknown population mean: $\theta = \beta_0 + \beta_1(42) + \beta_2(84)$
   Estimate: $\hat{\theta} = (1, 42, 84)\hat{\beta} = 64.9$

5. Report a standard error
   Answer: $SE(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})} = \sqrt{\text{Var}(\mathbf{a}^{\text{T}}\hat{\beta})} = \sqrt{\mathbf{a}^T\widehat{\mathbf{\Sigma}}\mathbf{a}} = 3.67$

6. Report a 95% confidence interval
   Answer: $\hat{\theta} \pm t(0.025, 42)SE(\hat{\theta})$ or $64.9 \pm 2.02(3.67)$ or $(57.49, 72.31)$

7. Test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$
   Answer: we follow the more general formula for calculating the F-statistic:

   (a) The full model $Y = \beta_0 + \beta_1 income + \beta_2 education + error$ has $SSR_{full} = 7507$ with $df_{full} = 42$.

   (b) The reduced model $Y = \beta_0 + error$ has $SSR_{red} = 43688$ with $df_{red} = 40$.

   (c) F-statistic: $F = \frac{[SSR_{red} - SSR_{full}]/(df_{red} - df_{full})}{SSR_{full}/df_{full}} = 101.22$

   (d) use the R software to find the p-value: ≈ 0

12 Lecture 12: Feb 15

Last time

- R practice with questions

Today

- Probability review
- HW2 posted
- HW1 review on Wednesday

Reference:

- Statistical Inference, 2nd Edition, by George Casella & Roger L. Berger
- Review of Probability Theory by Arian Maleki and Tom Do

Probability theory review

A few basic elements to define a probability on a set:

- **Sample space** S is the set that contains all possible outcomes of a particular experiment.
- An **event** is any collection of possible outcomes of an experiment, that is , any subset of S (including S itself).
- Event operations
  1. Union: The union of A and B, written $A \cup B$, is the set of elements that belong to either A or B or both:
     $$A \cup B = \{x : x \in A \text{ or } x \in B\}$$
  2. Intersection: The intersection of A and B, written $A \cap B$, is the set of elements that belong to both A and B:
     $$A \cap B = \{x : x \in A \text{ and } x \in B\}$$
  3. Complementation: The complement of A, written as A^c, is the set of all elements that are not in A:
     $$A^c = \{x : x \notin A\}.$$
- **Sigma algebra (or Borel field)**: A collection of subsets of S is called a sigma algebra (or Borel field), denoted by $\mathcal{B}$, if it satisfies the following three properties:
  1. $\emptyset \in \mathcal{B}$ (the empty set is an element of $\mathcal{B}$)

2. If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$ ($\mathcal{B}$ is closed under complementation).

3. If $A_1, A_2, \cdots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

- **Axioms of probability:** Given a sample space S and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function $\Pr()$ with domain $\mathcal{B}$ that satisfies

  1. $\Pr(A) \geqslant 0$ for all $A \in \mathcal{B}$
  2. $\Pr(S) = 1$.
  3. If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $\Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \Pr(A_i)$.

Properties:

If $\Pr()$ is a *probability function* and A and B are any sets in $\mathcal{B}$, then

- $\Pr(\varnothing) = 0$, where $\varnothing$ is the empty set
  Proof: $1 = \Pr(S) = \Pr(S \cup \varnothing)$
- $\Pr(A) \leqslant 1$
  Proof: see below and remember $\Pr(A^c) \geqslant 0$
- $\Pr(A^c) = 1 - \Pr(A)$
  Proof: $1 = \Pr(S) = \Pr(A \cup A^c) = \Pr(A) + \Pr(A^c)$
- $\Pr(B \cap A^c) = \Pr(B) - \Pr(A \cap B)$
  Proof: $B = \{B \cap A\} \cup \{B \cap A^c\}$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
  Proof: $A \cup B = A \cup \{B \cap A^c\}$ and use the above property.
- $Pr(A \cup B) = \Pr(A) + \Pr(B \cap A^c) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- If $A \subset B$, then $\Pr(A) \leqslant \Pr(B)$.
  Proof: If $A \subset B$, then $A \cap B = A$ and use $\Pr(B \cap A^c) = \Pr(B) - \Pr(A \cap B)$.

Conditional probability

Definition: If A and B are events in S, and $\Pr(B) > 0$, then the conditional probability of A given B, written $\Pr(A|B)$, is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Note that what happens in the conditional probability calculation is that B becomes the sample space: $\Pr(B|B) = 1$, in other words, $\Pr(A|B)$ is the probability measure of the event A after observing the occurrence of event B.

Definition: Two events A and B are statistically independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$. When A and B are independent events, then $\Pr(A|B) = \Pr(A)$ and the following pairs are also independent

- A and B^c
  proof:

$$\begin{aligned}\Pr(A \cap B^c) &= \Pr(A) - \Pr(A \cap B)\\ &= \Pr(A) - \Pr(A)\Pr(B)\\ &= \Pr(A)(1 - \Pr(B))\\ &= \Pr(A)\Pr(B^c)\end{aligned}$$

- A^c and B
- A^c and B^c

Random variables

Definition: A <u>random variable</u> is a function from a sample space S into the real numbers.

Experiment	Random variable
Toss two dice	X = sum of the numbers
Toss a coin 25 times	X = number of heads in 25 tosses
Apply different amounts of fertilizer to corn plants	$X = \textit{yield/acre}$

Suppose we have a sample space

$$S = \{s_1, \ldots, s_n\}$$

with a probability function Pr and we define a random variable X with range $\mathcal{X} = \{x_1, \ldots, x_m\}$. We can define a probability function $\Pr_X$ on $\mathcal{X}$ in the following way. We will observe $X = x_i$ if and only if the outcome of the random experiment is an $s_j \in S$ such that $X(s_j) = x_i$. Thus,

$$\Pr_X(X = x_i) = \Pr(\{s_j \in S : X(s_j) = x_j\}).$$

We will simply write $\Pr(X = x_i)$ rather than $\Pr_X(X = x_i)$.
A note on notation: Randon variables are often denoted with uppercase letters and the realized values of the variables (or its range) are denoted by corresponding lowercase letters.

Distribution functions

Definition: The <u>cumulative distribution function</u> or <u>*cdf*</u> of a random variable (r.v.) X, denoted by $F_X(x)$ is defined by

$$F_X(x) = \Pr(X \leqslant x), \text{ for all } x.$$

The function $F(x)$ is a cdf if and only if the following three conditions hold:

1. $\lim_{x\to\infty} F(x) = 1$.
2. $F(x)$ is a nondecreasing function of x.
3. $F(x)$ is right-continuous; that is, for every number x_0, $\lim_{x\downarrow x_0} = F(x_0)$.

Definition: A random variable X is continuous if $F(x)$ is a continuous function of x. A random variable X is discrete if $F(x)$ is a step function of x.

The following two statements are equivalent:

1. The random variables X and Y are identically distributed.
2. $F_X(x) = F_Y(x)$ for every x.

Density and mass functions

Definition: The probability mass function (pmf) of a discrete random variable X is given by

$$f_X(x) = \Pr(X = x) \text{ for all } x.$$

Example (Geometric probabilities) For the geometric distribution, we have the pmf

$$f_X(x) = \Pr(X = x) = \begin{cases} p(1-p)^{x-1} & \text{for } \text{ x} = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Definition: The probability density function or *pdf*, $f_X(x)$, of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad \text{for all } x.$$

A note on notation: The expression "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$", where we read the symbol "$\sim$" as " is distributed as".

Example (Logistic distribution) For the logistic distribution, we have

$$F_X(x) = \frac{1}{1+e^{-x}}$$

and, hence,

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$

A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if

1. $f_X(x) \geqslant 0$ for all x
2. $\sum_x f_X(x) = 1$ *(pmf) or* $\int_{-\infty}^{\infty} f_X(x)dx = 1$ *(pdf)*.

Expectations

The expected value, or expectation, of a random variable is merely its average value, where we speak of "average" value as one that is weighted according to the probability distribution.

Definition: The expected value or mean of a random variable $g(X)$, denoted by $\mathbf{E}(g(X))$, is

$$\mathbf{E}(g(X)) = \begin{cases} \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{if } X \text{ is continuous} \\ \sum_{x\in\mathcal{X}} g(x)f_X(x) = \sum_{x\in\mathcal{X}} g(x)\Pr(X = x) & \text{if } X \text{ is discrete,} \end{cases}$$

Exponential mean

Suppose $X \sim Exp(\lambda)$ distribution, that is, it has pdf given by

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leqslant x < \infty, \quad \lambda > 0$$

Then $\mathbf{E}\left(X\right)$ is:

$$\begin{aligned}
\mathbf{E}\left(X\right) &= \int_0^\infty \frac{1}{\lambda} x e^{-x/\lambda} dx \\
&= -xe^{-x/\lambda}\Big|_0^\infty + \int_0^\infty e^{-x/\lambda} dx \\
&= \int_0^\infty e^{-x/\lambda} dx = \lambda
\end{aligned}$$

13 Lecture 13: Feb 17

Last time

- Probability review

Today

- HW1 review
- Probability review, cont

Reference:

- Statistical Inference, 2nd Edition, by George Casella & Roger L. Berger
- Review of Probability Theory by Arian Maleki and Tom Do

Binomial mean

IF X has binomial distribution, i.e. $X \sim \mathit{binomial}(n,p)$, its pmf is given by

$$\Pr(X=x) = \left(\begin{array}{c} \text{n} \\ \text{x} \end{array}\right) p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n,$$

where n is a positive integer, $0 \leqslant p \leqslant 1$, and for every fixed pair n and p the pmf sums to 1. The expected value of a binomial random variable is then given by

$$\mathbf{E}(X) = \sum_{x=0}^{n} x \left(\begin{array}{c} \text{n} \\ \text{x} \end{array}\right) p^x(1-p)^{n-x}$$

Now, use the identity $x \left(\begin{array}{c} \text{n} \\ \text{x} \end{array}\right) = n \left(\begin{array}{c} \text{n - 1} \\ \text{x - 1} \end{array}\right)$ to derive the Expected value.

$$\begin{aligned}
\mathbf{E}(X) &= \sum_{x=1}^{n} x \left(\begin{array}{c} \text{n} \\ \text{x} \end{array}\right) p^x(1-p)^{n-x} \\
&= \sum_{x=1}^{n} n \left(\begin{array}{c} \text{n - 1} \\ \text{x - 1} \end{array}\right) p^x(1-p)^{n-x} \\
&= \sum_{y=0}^{n-1} n \left(\begin{array}{c} \text{n - 1} \\ \text{y} \end{array}\right) p^{y+1}(1-p)^{n-(y+1)} \\
&= np \sum_{y=0}^{n-1} \left(\begin{array}{c} \text{n - 1} \\ \text{y} \end{array}\right) p^{y}(1-p)^{n-1-y} \\
&= np,
\end{aligned}$$

since the last summation must be 1, being the sum over all possible values of a $\mathit{binomial}(n-1,p)$ pmf.

properties:

Let X be a random variable and let a, b and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

1. $\mathbf{E}\left(a \cdot g_1(X) + b \cdot g_2(X) + c\right) = a\mathbf{E}\left(g_1(X)\right) + b\mathbf{E}\left(g_2(X)\right) + c.$
2. If $g_1(x) \geqslant 0$ for all x, then $\mathbf{E}\left(g_1(X)\right) \geqslant 0$.
3. If $g_1(x) \geqslant g_2(x)$ for all x, then $\mathbf{E}\left(g_1(X)\right) \geqslant \mathbf{E}\left(g_2(X)\right)$.
4. If $a \leqslant g_1(x) \leqslant b$ for all x, then $a \leqslant \mathbf{E}\left(g_1(X)\right) \leqslant b$.

Moments

The various moments of a distribution are an important class of expectations.

Definition: For each integer n, the n^{th} moment of X (or $F_X(x)$), μ'_n, is

$$\mu'_n = \mathbf{E}\left(X^n\right).$$

The n^{th} central moment of X, μ_n, is

$$\mu_n = \mathbf{E}\left((X-\mu)^n\right),$$

where $\mu = \mu'_1 = \mathbf{E}\left(X\right)$.

Variance

Definition: The variance of a random variable X is its second central moment, $\mathbf{Var}\left(X\right) = \mathbf{E}\left((X - EX)^2\right)$. The positive square root of $\mathbf{Var}\left(X\right)$ is the standard deviation of X.

Exponential variance

Let X have the exponential(λ) distribution, $X \sim Exp(\lambda)$. Then the variance of X is

$$\begin{aligned}
\mathbf{Var}\left(X\right) &= \mathbf{E}\left((X-EX)^2\right) = \mathbf{E}\left((X-\lambda)^2\right) \\
&= \int_0^\infty (x-\lambda)^2 \frac{1}{\lambda} e^{-x/\lambda} dx \\
&= \int_0^\infty (x^2 - 2x\lambda + \lambda^2) \frac{1}{\lambda} e^{-x/\lambda} dx \\
&= \lambda^2.
\end{aligned}$$

properties

1. $\mathbf{Var}\left(aX+b\right) = a^2\mathbf{Var}\left(X\right)$.
   proof:

$$\begin{aligned}\mathbf{Var}\,(aX+b) &= \mathbf{E}\left(((aX+b)-\mathbf{E}\,(aX+b))^2\right)\\ &= \mathbf{E}\left((aX-aEX)^2\right)\\ &= a^2\mathbf{E}\left((X-EX)^2\right)\\ &= a^2\mathbf{Var}\,(X)\end{aligned}$$

2. $\mathbf{Var}\,(X) = \mathbf{E}\,(X^2) - (\mathbf{E}\,(X))^2$.
   proof:

$$\begin{aligned}\mathbf{Var}\,(X) &= \mathbf{E}\,(X-EX)^2\\ &= \mathbf{E}\left(X^2 - 2X\mathbf{E}\,(X) + (\mathbf{E}\,(X))^2\right)\\ &= \mathbf{E}\left(X^2\right) - 2\mathbf{E}\,(X)\mathbf{E}\,(X) + (\mathbf{E}\,(X))^2\\ &= \mathbf{E}\left(X^2\right) - (\mathbf{E}\,(X))^2\end{aligned}$$

Moment generating function

Definition: Let X be a random variable with cdf F_X. The moment generating function or mgf of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = \mathbf{E}\left(e^{tX}\right),$$

provided that the expectation exists for t in some neighborhood of 0. That is, there exists an $h > 0$ such that for all t in $-h < t < h$, $\mathbf{E}\left(e^{tX}\right)$ exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

Property: If X has mgf $M_X(t)$, then

$$\mathbf{E}\,(X^n) = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \left.\frac{d^n}{dt^n}M_X(t)\right|_{t=0}.$$

Some common random variables

Discrete random variables

- $X \sim \mathit{Bernoulli}(p)$ (where $0 \leqslant p \leqslant 1$):

$$\Pr(x) = \begin{cases} p & \text{if } x = 1\\ 1-p & \text{if } x = 0\end{cases}$$

- $X \sim \mathit{Binomial}(n,p)$ (where $0 \leqslant p \leqslant 1$):

$$\Pr(x) = \begin{pmatrix} n\\ x\end{pmatrix} p^x(1-p)^{n-x}$$

- $X \sim Geometric(p)$ (where $0 \leqslant p \leqslant 1$):

$$\Pr(x) = p(1-p)^{x-1}$$

- $X \sim Poisson(\lambda)$ (where $\lambda > 0$):

$$\Pr(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Continuous random variables

- $X \sim Uniform(a, b)$ (where $a < b$):

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } \ a \leqslant x \leqslant b \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Exponential(\lambda)$ (where $\lambda > 0$):

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } \ x \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim Normal(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

The following table provides a summary of some of the properties of these distributions.

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p & \text{if } \ x = 1 \\ 1-p & \text{if } \ x = 0 \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x}$, for $0 \leqslant k \leqslant n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{x-1}$, for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\frac{\lambda^x}{x!}$, for $k = 1, 2, \ldots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a} I(a \leqslant x \leqslant b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x} I(x \geqslant 0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

14 Lecture 14: Feb 19

Last time

- HW1 review
- Probability review, cont

Today

- Probability review
- Lab session

Reference:

- Statistical Inference, 2nd Edition, by George Casella & Roger L. Berger
- Review of Probability Theory by Arian Maleki and Tom Do

Chi-square, t-, and F-Distributions

Let $Z_1, Z_2, \dots, Z_k \overset{iid}{\sim} N(0,1)$, then $X^2 \equiv Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi_k^2$ (with k degrees of freedom). If $X \sim \chi_k^2$

$$\begin{aligned} \mathbf{E}(X) &= k \\ \mathbf{Var}(X) &= 2k. \end{aligned}$$

Student's t versus χ^2

If $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

When σ is unknown,

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}, \quad \text{where } \hat{\sigma} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}.$$

Note that

$$\begin{aligned} \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} &= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{1}{\frac{\hat{\sigma}}{\sigma}} \\ &= Z \cdot \frac{1}{\sqrt{\frac{\sum (X_i - \bar{X})^2}{(n-1)\sigma^2}}} \\ &= \frac{Z}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}} \end{aligned}$$

F **versus** χ^2

$$F_{ndf,ddf} \equiv \frac{\chi^2_{ndf}/ndf}{\chi^2_{ddf}/ddf}$$

t **versus** F

$$\begin{aligned} t_k &= \frac{Z}{\sqrt{\chi^2_k/k}} \\ &= \frac{\sqrt{\chi^2_1/1}}{\sqrt{\chi^2_k/k}} \\ &= \sqrt{F_{1,k}} \end{aligned}$$

or, in other words, $t_k^2 = F_{1,k}$

Random vectors and matrices

The cdf for random vector

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \text{ is } F_{\mathbf{Y}}(\mathbf{y}) = \Pr(Y_1 \leqslant y_1, Y_2 \leqslant y_2, \ldots, Y_n \leqslant y_n)$$

If a joint pdf exists, then $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}}(y_1, \ldots, y_n)$ and

$$F_{\mathbf{Y}}(\mathbf{y}) = \int\limits_{-\infty}^{y_1} \int\limits_{-\infty}^{y_2} \ldots \int\limits_{-\infty}^{y_n} f_{\mathbf{Y}}(\mathbf{t}) d\mathbf{t}$$

Moments

$$\begin{aligned} \mathbf{E}(\mathbf{Y}) = \mu_{\mathbf{Y}} &= \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \\ \mathbf{Var}(\mathbf{Y}) &= \mathbf{E}\left((\mathbf{Y} - \mu_{\mathbf{Y}})(\mathbf{Y} - \mu_{\mathbf{Y}})^T\right) \\ &= \mathbf{E}\left(\begin{bmatrix} (Y_1-\mu_1)^2 & (Y_1-\mu_1)(Y_2-\mu_2) & \cdots \\ (Y_2-\mu_2)(Y_1-\mu_1) & (Y_2-\mu_2)^2 & \cdots \\ \cdots & & \end{bmatrix}\right) \\ &= \mathbf{E}\left([(Y_i-\mu_i)(Y_j-\mu_j), i=1,2,\ldots,n, j=1,2,\ldots,n]\right) \\ &= (\sigma_{ij})_{i=1,2,\ldots,n;j=1,2,\ldots,n} \end{aligned}$$

where $\sigma_{ij} = Cov(Y_i, Y_j)$

Linear functions

Let $\mathbf{X} \in \mathbb{R}^{k\times 1}$, $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ and $\mathbf{A} \in \mathbb{R}^{k\times 1}$, $\mathbf{B} \in \mathbb{R}^{k\times n}$ be non-random, then

$$\begin{aligned} \underset{k\times 1}{\mathbf{X}} &= \underset{k\times 1}{\mathbf{A}} + \underset{k\times n}{\mathbf{B}}\underset{n\times 1}{\mathbf{Y}} \\ \mathbf{E}\left(\mathbf{X}\right) &= \mathbf{A} + \mathbf{B}\mathbf{E}\left(\mathbf{Y}\right) \\ \mathbf{Var}\left(\mathbf{X}\right) &= \mathbf{B}\mathbf{Var}\left(\mathbf{Y}\right)\mathbf{B}^T \end{aligned}$$

Sums of random vectors

$$\begin{aligned} \underset{n\times 1}{\mathbf{X}} &= \underset{n\times 1}{\mathbf{Y}} + \underset{n\times 1}{\mathbf{Z}} \\ \mathbf{E}\left(\mathbf{X}\right) &= \mathbf{E}\left(\mathbf{Y}\right) + \mathbf{E}\left(\mathbf{Z}\right) = \mathbf{E}\left(\mathbf{Y}+\mathbf{Z}\right) \end{aligned}$$

Note that there is no independence assumed above.

$$\mathbf{Var}\left(\mathbf{X}\right) = \mathbf{Var}\left(\mathbf{Y}+\mathbf{Z}\right) = \mathbf{Var}\left(\mathbf{Y}\right) + \mathbf{Var}\left(\mathbf{Z}\right) + Cov(\mathbf{Y},\mathbf{Z}) + Cov(\mathbf{Z},\mathbf{Y})$$

If $\mathbf{Y}, \mathbf{Z}$ are uncorrelated, then $\mathbf{Var}\left(\mathbf{X}\right) = \mathbf{Var}\left(\mathbf{Y}\right) + \mathbf{Var}\left(\mathbf{Z}\right)$

15 Lecture 15: Feb 22

Last time

- Probability review
- Lab session

Today

- Dummy-Variable regression
- Interactions

Dummy-variable regression

For categorical data (factor), we use dummy variable regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i$$

where D, called a dummy variable regressor or an indicator variable, is coded 1 for one level and 0 for all others,

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}.$$

Therefore, for women, the model becomes

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and for men

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 + \epsilon_i = (\beta_0 + \beta_2) + \beta_1 X_i + \epsilon_i$$

For example, Figure 15.1 (a) and (b) represents two small (idealized) populations. In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income: Holding education constant, the "effect" of gender is the vertical distance between the two regression lines, which, for parallel lines, is everywhere the same.

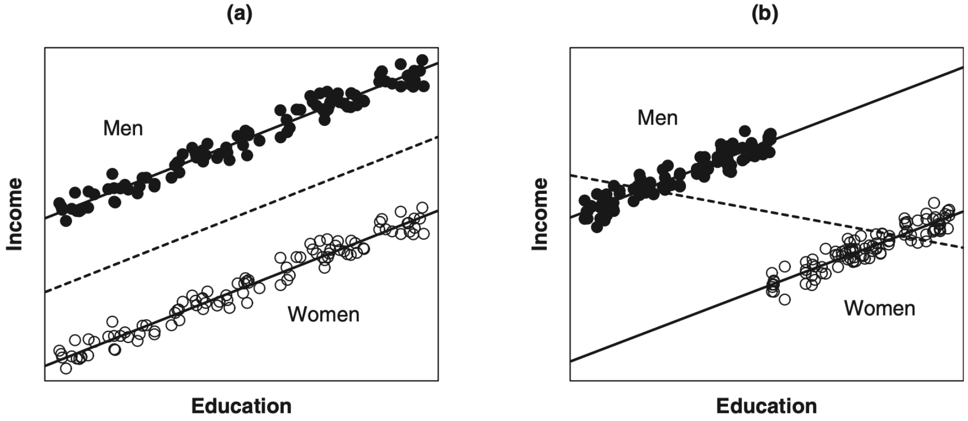

Figure 15.1: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are parallel. In each graph, the overall (i.e. marginal) regression of income on education (ignoring gender) is given by the broken line. JF Figure 7.1.

Multi-level factor

We can model the effects of classification factors with m categories (levels) by using $m - 1$ indicator variables.

For example, the three-category occupational-type factor can be represented in the regression equation by introducing two dummy regressors:

Category	D_1	D_2
Professional and managerial	1	0
White collar	0	1
Blue collar	0	0

A model for the regression of prestige on income, education, and type of occupation is then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

where X_1 is income and X_2 is education. This model describes three parallel regression planes, which can differ in their intercepts:

$$\begin{array}{ll} \text{Professional:} & Y_i = (\beta_0 + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\ \text{White collar:} & Y_i = (\beta_0 + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\ \text{Blue collar:} & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \end{array}$$

Therefore, the coeficient β_0 gives the intercept for blue-collar occupations; γ_1 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income); and γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations (again, fixing education and income).

In the above prestige example, we chose “blue collar” as the baseline category. Sometimes, it is natural to pick a particular category as the baseline category, for example, the “control group” in an experiment. However, in most applications, the choice of a baseline category is entirely arbitrary.

Matrix representation

For the above prestige model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

we have the design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & D_{11} & D_{12} \\ 1 & X_{21} & X_{22} & D_{21} & D_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & D_{n1} & D_{n2} \end{bmatrix}$$

and the vector of coefficients β is

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}$$

such that we have (again) the linear model in matrix form:

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

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, in other words, $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Interactions

Two explanatory variables are said to interact in determining a response variable when the partial effect of one depends on the value of the other. Consider the hypothetical data shown in Figure 15.2.

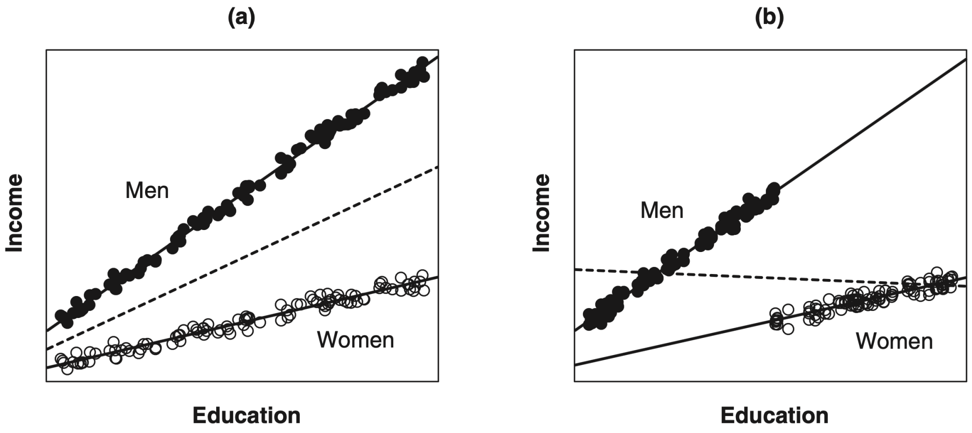

Figure 15.2: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are not parallel. The slope for men is greater than the slope for women, and consequently education and gender interact in affecting income. In each graph, the overall regression of income on education (ignoring gender) is given by the broken line. JF Figure 7.7.

It is apparent in both Figure 15.2 (a) and (b) the within-gender regressions of income on education are not parallel: In both cases, the slope for men is larger than the slope for women.

Modeling interactions

We accommodate the interaction of education and gender by:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i D_i) + \epsilon_i$$

where we introduce the interaction regressor XD into the regression equation. For women, the model becomes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 \cdot 0 + \beta_3 (X_i \cdot 0) + \epsilon_i \\ &= \beta_0 + \beta_1 X_i + \epsilon_i \end{aligned}$$

and for men

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 \cdot 1 + \beta_3 (X_i \cdot 1) + \epsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + \epsilon_i \end{aligned}$$

The parameters β_0 and β_1 are, respectively, the intercept and slope for the regression of income on education among women (the baseline category for gender); β_2 gives the difference in intercepts between the male and female groups; and β_3 gives the difference in slopes between the two groups.

Usual guidance: Models that include an interaction between two predictors should also include the individual predictors by themselves regardless of the statistical significance of the associated β's.

Test for the interaction

We can simply test the hypothesis $H_0 : \beta_3 = 0$ and construct the test statistic $t = \frac{\hat{\beta}_i - 0}{\widehat{SE}(\hat{\beta}_i)} \sim t_{n-4}$ $(p = 3)$.

Interactions with multi-level factor

We can easily extend the method for modeling interactions by forming product regressors to multi-level factors, to several factors, and to several quantitative explanatory variables. Using the occupational prestige example, the occupational type could possibly interact both with income (X_1) and with education (X_2):

$$\begin{aligned} Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \epsilon_i \end{aligned}$$

The model therefore permits different intercepts and slopes for the three types of occupations:

$$\begin{array}{llrrrl} \text{Professional:} & Y_i = & (\beta_0 + \gamma_1)+ & (\beta_1 + \delta_{11})X_{i1}+ & (\beta_2 + \delta_{21})X_{i2}+ & \epsilon_i \\ \text{White collar:} & Y_i = & (\beta_0 + \gamma_2)+ & (\beta_1 + \delta_{12})X_{i1}+ & (\beta_2 + \delta_{22})X_{i2}+ & \epsilon_i \\ \text{Blue collar:} & Y_i = & \beta_0+ & \beta_1 X_{i1}+ & \beta_2 X_{i2}+ & \epsilon_i \end{array}$$

16 Lecture 16: Feb 24

Last time

- Dummy-Variable regression (JF chapter 7)
- Interactions

Today

- Unusual and influential data (JF chapter 11)

Unusual and influential data

Linear models make strong assumptions about the structure of data, assumptions that often do not hold in applications. The method of least squares can be very sensitive to the structure of the data and may be markedly influenced by one or a few unusual observations.

Outliers

In simple regression analysis, an outlier is an observation whose response-variable value is *conditionally* unusual *given* the value of the explanatory variable: see Figure 16.1.

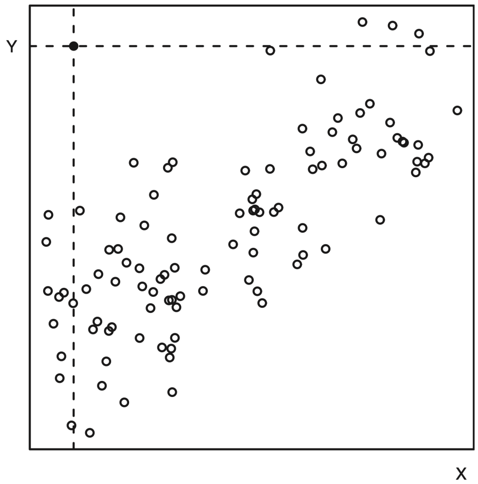

Figure 16.1: The black point is a regression outlier because it combines a relatively large value of Y with a relatively small value of X, even though neither its X-value nor its Y-value is unusual individually. Because of the positive relationship between Y and X, points with small X-values also tend to have small Y-values, and thus the black point is far from other points with similar X-values. JF Figure 11.1.

Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis. Their presence may be a signal that the model fails to capture important characteristics of the data.

Figure 16.2 illustrates some distinctions for the simple-regression model $Y = \beta_0 + \beta_1 X + \epsilon$.

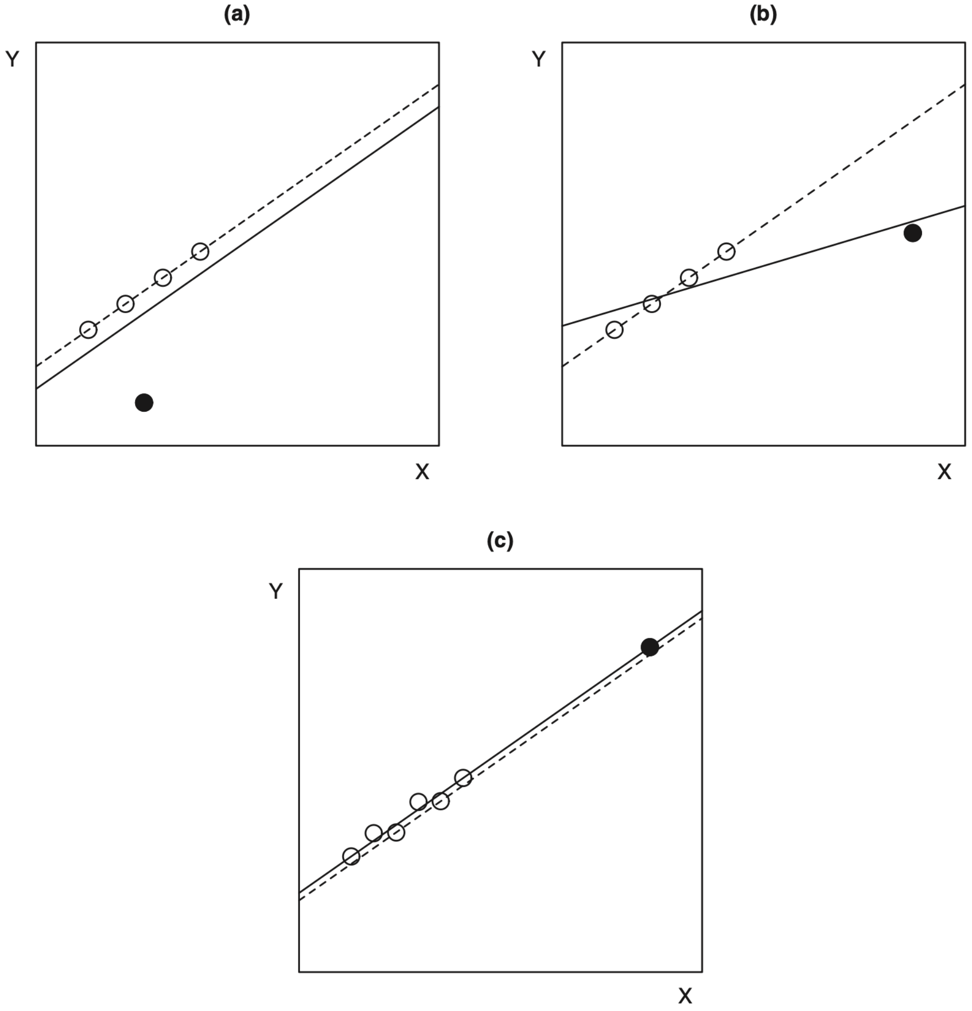

Figure 16.2: Leverage and influence in simple regression. In each graph, the solid line gives the least-squares regression for all the data, while the broken line gives the least-squares regression with the unusual data point (the black circle) omitted. (a) An outlier near the mean of X has low leverage and little influence on the regression coefficients. (b) An outlier far from the mean of X has high leverage and substantial influence on the regression coefficients. (c) A high-leverage observation in line with the rest of the data does not influence the regression coefficients. In panel (c), the two regression lines are separated slightly for visual effect but are, in fact, coincident JF Figure 11.2.

Some qualitative distinctions between outliers and high leverage observations:

- An outlier is a data point whose response Y does not follow the general trend of the rest of the data.
- A data point has high leverage if it has "extreme" predictor X values:

  - With a single predictor, an extreme X value is simply one that is particularly high or low.
  - With multiple predictors, extreme X values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values .

And the influence of a data point is the combination of leverage and discrepancy ("outlyingness") though the following heuristic formula:

$$\text{Influence on coefficients} = \text{Leverage} \times \text{Discrepancy}.$$

Assessing leverage: hat-values

The hat-value h_i is a common measure of leverage in regression. They are named because it is possible to express the fitted values $\hat{Y}_j$ ("Y-hat") in terms of the observed values Y_i:

$$\hat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jj}Y_j + \cdots + h_{nj}Y_n = \sum_{i=1}^{n} h_{ij}Y_i.$$

The weight h_{ij} captures the contribution of observation Y_i to the fitted value $\hat{Y}_j$: If h_{ij} is large, then the ith observation can have a considerable impact on the jth fitted value. With the least square solutions, for the fitted values:

$$\hat{\mathbf{Y}} = \mathbf{X}\beta = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

we (already) get the hat matrix:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Properties:

- (idempotent) $\mathbf{H} = \mathbf{H}\mathbf{H}$
- $h_i \equiv h_{ii} = \sum_{j=1}^{n} h_{ij}^2$
- $\frac{1}{n} \leqslant h_i \leqslant 1$ (a proof by Mohammad Mohammadi)
- $\bar{h} = (p+1)/n$

In the case of SLR, the hat-values are:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

Detecting outliers: studentized residuals

The variance of the residuals ($\hat{\epsilon}_i = Y_i - \hat{Y}_i$) do not have equal variances (even if the errors ϵ_i have equal variances):

$$\text{Var}(\hat{\epsilon}) = \text{Var}(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \text{Var}[(\mathbf{I} - \mathbf{H})\mathbf{Y}] = (\mathbf{I} - \mathbf{H})\text{Var}(\mathbf{Y})(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

so that for $\hat{\epsilon}_i$,

$$\mathrm{Var}(\hat{\epsilon}_\mathrm{i}) = \sigma^2(1 - \mathrm{h}_\mathrm{i}).$$

High-leverage observations tend to have small residuals (in other words, these observations can pull the regression surface toward them).

The standardized residual (sometimes called internally studentized residual)

$$\hat{\epsilon}_i^{'} \equiv \frac{\hat{\epsilon}}{\hat{\sigma}\sqrt{1 - h_i}}$$

, however, does not follow a t-distribution, because the numerator and denominator are not independent.

Suppose, we refit the model deleting the ith observation, obtaining an estimate $\hat{\sigma}_{(-i)}$ of σ that is based on the remaining $n-1$ observations. Then the studentized residual (sometimes called externally studentized residual)

$$\hat{\epsilon}_i^{*} \equiv \frac{\hat{\epsilon}}{\hat{\sigma}_{(-i)}\sqrt{1 - h_i}}$$

has an independent numerator and denominator and follows a t-distribution with $n - p - 2$ degrees of freedom.

The studentized and the standardized residuals have the following relationship (Beckman and Trussell, 1974):

$$\hat{\epsilon}_i^{*} = \hat{\epsilon}_i' \sqrt{\frac{n - p - 2}{n - p - 1 - \hat{\epsilon}_i'^2}}$$

For large n,

$$\hat{\epsilon}_i^{*} \approx \hat{\epsilon}_i' \approx \frac{\hat{\epsilon}}{\hat{\sigma}}$$

Test for outlier

It is of our interest to pick the studentized residual $\hat{\epsilon}_{max}^{*}$ with the largest absolute value among $\hat{\epsilon}_1^{*}, \hat{\epsilon}_2^{*}, \ldots, \hat{\epsilon}_n^{*}$ to test for outlier. However, by doing so, we are effectively picking the biggest of n test statistics such that it is not legitimate simply to use t_{n-p-2} to find a p-value. We need a correction on the p-value because of multiple-comparisons.

Suppose that we have $p' = \Pr(t_{n-p-2} > |\hat{\epsilon}_{max}^{*}|)$, the p-value before correction. Then the Bonferroni adjusted p-value is $p = np'$.

Measuring influence

Influence on the regression coefficients combines leverage and discrepancy. The most direct measure of influence simply expresses the impact on each coefficient of deleting each observation in turn:

$$D_{ij} = \hat{\beta}_j - \tilde{\beta}_{j(-i)} \quad \text{for } i = 1, \ldots, n \text{ and } j = 0, 1, \ldots, p$$

where $\hat{\beta}_j$ are the least-squares coefficients calculated for all the data, and the $\tilde{\beta}_{j(-i)}$ are the least-squares coefficients calculated with the ith observation omitted. To assist in interpretation, it is useful to scale the D_{ij} by (deleted) coefficient standard errors:

$$D^*_{ij} = \frac{D_{ij}}{\widehat{SE}_{(-i)}(\tilde{\beta}_{j(-i)})}$$

Following Belsley, Kuh, and Welsh (1980), the D_{ij} are often termed DFBETA$_{ij}$, and D^*_{ij} are called DFBETAS$_{ij}$. One problem associated with using D_{ij} or D^*_{ij} is their large number $n(p+1)$ of each.

Cook's distance calculated as

$$D_i = \frac{\sum_{j=1}^{n}(\tilde{y}_{j(-i)} - \hat{y}_j)^2}{(p+1)\hat{\sigma}^2} = \frac{\hat{\epsilon}_i'^2}{p+1} \times \frac{h_i}{1-h_i}$$

In effect, the first term in the formula for Cook's D is a measure of discrepancy, and the second is a measure of leverage. We look for values of D_i that stand out from the rest.

A similar measure suggested by Belsley et al. (1980)

$$\text{DFFITS}_i = \hat{\epsilon}^*_i \frac{h_i}{1-h_i}$$

Except for unusual data configurations, Cook's $D_i \approx \text{DFFITS}_i^2/(p+1)$.

Numerical cutoffs (suggested)

Diagnostic statistic	Cutoff value
h_i	$2\bar{h} = \frac{2(p+1)}{n}$, ($3\bar{h}$ for small sample)
D^*_{ij}	$\lvert D^*_{ij}\rvert > 1$ or 2 ($2/\sqrt{n}$ for large samples)
Cook's D_i	$D_i > \frac{4}{n-p-1}$
DFFITS	$\lvert\text{DFFITS}_i\rvert > 2\sqrt{\frac{p+1}{n-p-1}}$

17 Lecture 17: Feb 26

Last time

- Unusual and influential data (JF chapter 11)

Today

- Added-variable plots
- Should unusual data be discarded

Added-variable plots

Unlike the case of SLR, the scatterplot with the response variable and one predictor gives only the marginal effect in MLR. Instead, the added-variable plot (also called a partial-regression plot or a partial-regression leverage plot) gives a graphical inspection over each dimension.

Let $\hat{\tilde{Y}}_i^{(1)}$ represent the residuals from the least-squares regression of Y on all the Xs except X_1, in other words, the residuals from the following fitted regression equation:

$$Y_i = \tilde{\beta}_0^{(1)} + \tilde{\beta}_2^{(1)} X_{i2} + \cdots + \tilde{\beta}_p^{(1)} X_{ip} + \tilde{Y}_i^{(1)}$$

where the parenthetical superscript (1) indicates the omission of X_1 from the right-hand side of the regression equation. Likewise, $X_i^{(1)}$ is the residual from the least-squares regression of X_1 on all the other Xs:

$$X_{i1} = \check{\beta}_0^{(1)} + \check{\beta}_2^{(1)} X_{i2} + \cdots + \check{\beta}_p^{(1)} X_{ip} + \check{X}_i^{(1)}$$

Then, the residuals $\tilde{Y}_i^{(1)}$ and $\check{X}_i^{(1)}$ have the following interesting properties:

1. The slope from the least-squares regression of $\tilde{Y}_i^{(1)}$ on $\check{X}_i^{(1)}$ is simply the least-squares slope $\hat{\beta}_1$ from the *full* multiple regression.
2. The residuals from the simple regression of $\tilde{Y}_i^{(1)}$ on $\check{X}_i^{(1)}$ are the same as those from the full regression, that is
$$\tilde{Y}_i^{(1)} = \hat{\beta}_1 \check{X}_i^{(1)} + \hat{\epsilon}_i$$
3. The variation of $\check{X}_i^{(1)}$ is the *conditional variation* of X_1 holding the other Xs constant.

Figure 17.1 shows that the conditional variation is smaller than its marginal variation – much smaller when X_1 is strongly collinear with other Xs,

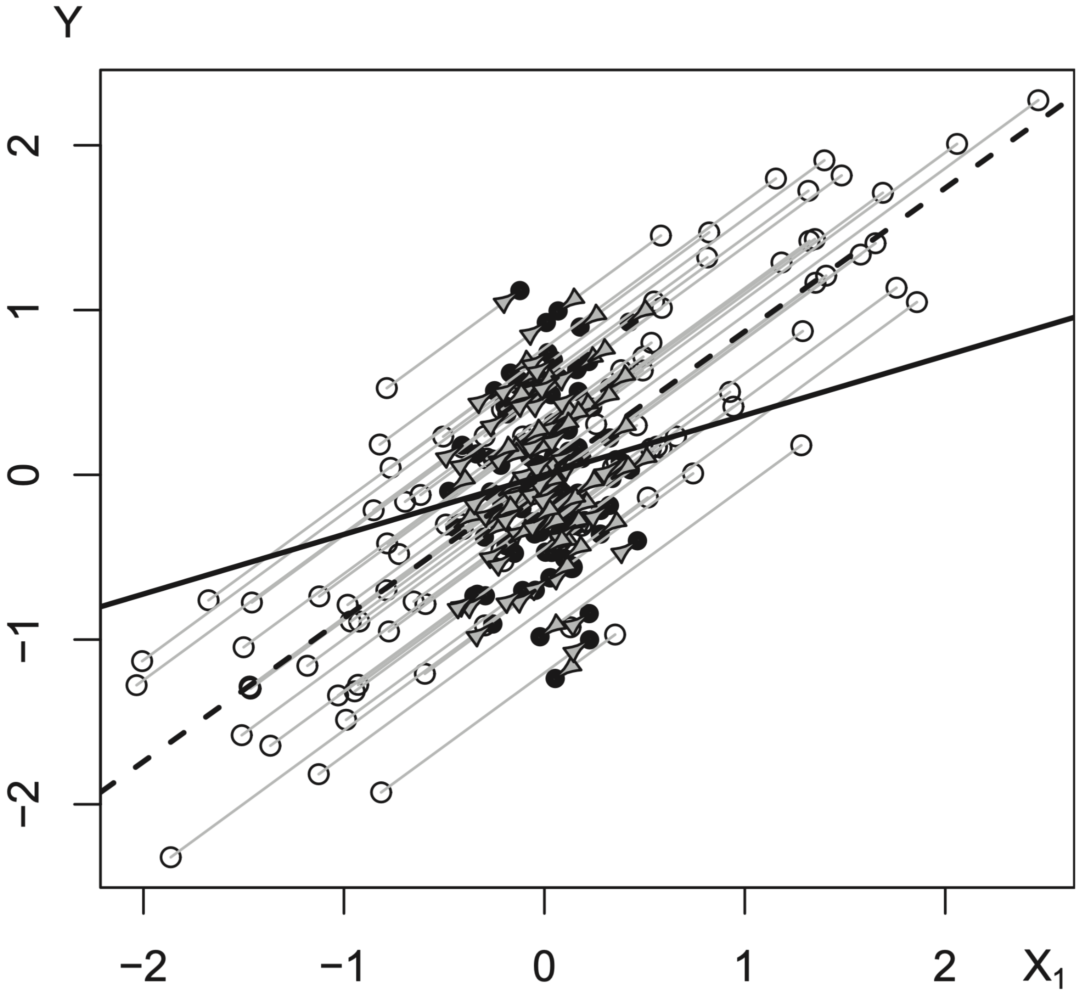

Figure 17.1: The marginal scatterplot (open circles) for Y and X_1 superimposed on the added-variable plot (filled circles) for X_1 in the regression of Y on X_1 and X_2. The variables Y and X_1 are centered at their means to facilitate the comparison of the two sets of points. The arrows show how the points in the marginal scatterplot map into those in the AV plot. In this contrived data set, X_1 and X_2 are highly correlated ($r_{12} = 0.98$), and so the conditional variation in X_1 (represented by the horizontal spread of the filled points) is much less than its marginal variation (represented by the horizontal spread of the open points). The broken line gives the slope of the marginal regression of Y on X_1 alone, while the solid line gives the slope $\hat{\beta}_1$ of X_1 in the MLR of Y on both Xs. JF Figure 11.9.

Figure 17.2 illustrates the added-variable plots using the Duncan's data.

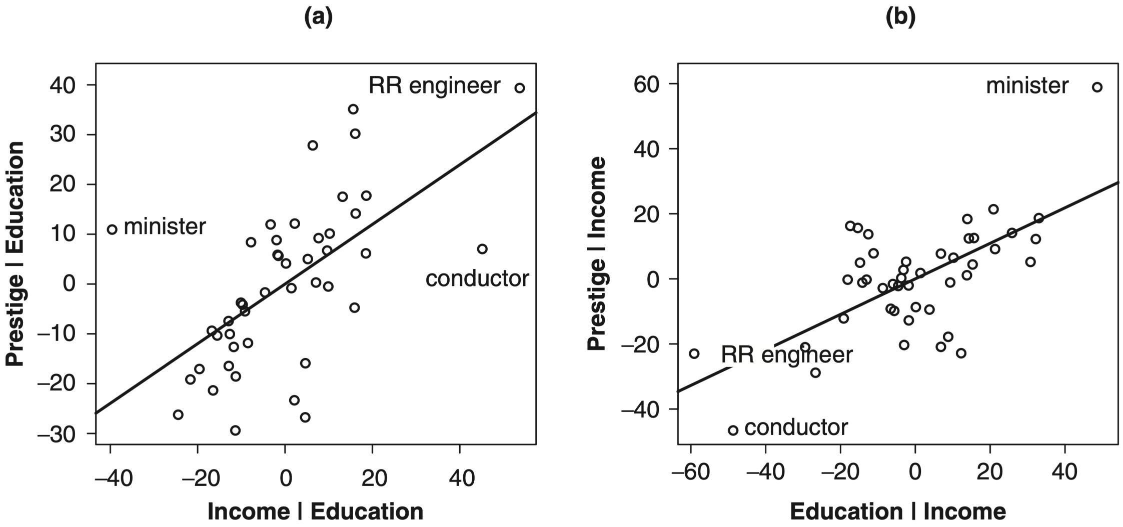

Figure 17.2: Added-variable plots for Duncan's regression of occupational prestige on the (a) income and (b) education levels of 45 US occupations in 1950. Three unusal observations, *miniters*, *conductors*, and *railroadengineers*, are identified on the plots. The added-variable plot for the intercept $\hat{\beta}_0$ is not shown. JF Figure 11.10.

Should unusual data be discarded?

In practice, although problematic data should not be ignored, they also should not be deleted automatically and without reflection:

- It is important to investigate *why* an observation is unusual. Truly "bad" data (e.g., an error in data entry) can often be corrected or, if correction is not possible, thrown away. When a discrepant data point is correct, we may be able to understand why the observation is unusual. For Duncan's data, for example, it makes sense that ministers enjoy prestige not accounted for by the income and educational levels of the occupation and for a reason not shared by other occupations. In a case like this, where an outlying observation has characteristics that render it unique, we may choose to set it aside from the rest of the data.
- Alternatively, outliers, high-leverage points, or influential data may motivate model respecification, and the pattern of unusual data may suggest the introduction of additional explanatory variables. We noticed, for example, that both conductors and railroad engineers had high leverage in Duncan's regression because these occupations combined relatively high income with relatively low education. Perhaps this combination of characteristics is due to a high level of unionization of these occupations in 1950, when the data were collected. If so, and if we can ascertain the levels of unionization of all of the occupations, we could enter this as an explanatory variable, perhaps shedding further light on the process determining occupational prestige.
- Except in clear-cut cases, we are justifiably reluctant to delete observations or to respecify the model to accommodate unusual data. Some researchers reasonably adopt

alternative estimation strategies, such as robust regression, which continuously down-weights outlying data rather than simply discarding them. Because these methods assign zero or very small weight to highly discrepant data, however, the result is generally not very different from careful application of least squares, and , indeed, robust-regression weights can be used to identify outliers.

- Finally, in large samples, unusual data substantially alter the results only in extreme instances. Identifying unusual observations in a large sample, therefore, should be regarded more as an opportunity to learn something about the data not captured by the model that we have fit, rather than as an occasion to reestimate the model with the unusual observations removed.

18 Lecture 18: March 1

Last time

- Unusual and influential data (JF chapter 11)

Today

- HW2 deadline extends to end of this week.
- Diagnosing non-normality, non-constant error variance, and nonlinearity (JF chapter 12)
- Data transformation (JF chapter 4)

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of iid random variables whose mgfs exist in a neighborhood of 0 (that is, $M_{X_i}(t)$ exists for $|t| < h$, for some positive h). Let $\mathrm{E}X_i = \mu$ and $\mathrm{VarX_i} = \sigma^2 > 0$. (Both μ and σ^2 are finite since the mgf exists.). Define $\bar{X}_n = (1/n) \sum_{i=1}^{n} X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any x, $-\infty < x < \infty$,

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution. (Refer to Casella & Berger p.237 - p.238 for a proof.)

Delta Method

Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose $g'(\theta)$ exists and is not 0. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to N(0, \sigma^2[g'(\theta)]^2) \text{ in distribution.}$$

(Refer to Casella & Berger p.243 for a proof using Taylor expansion.)

Second-order Delta Method

Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2} \chi_1^2 \text{ in distribution.}$$

Non-normally distributed errors

The assumption of normally distributed errors is almost always arbitrary. Nevertheless, the central limit theorem ensures that, under very broad conditions, inference based on the least-squares estimator is approximately valid in all but small samples. Why concern about non-normal errors?

- For some types of error distributions, particularly those with heavy tails, the efficiency of least-squares estimation decreases markedly.
- Highly skewed error distributions, aside from their propensity to generate outliers in the direction of the skew, compromise the interpretation of the least-squares fit. This fit is a conditional mean (of Y given the Xs), and the mean is not a good measure of the center of a highly skewed distribution.
- A multimodal error distribution suggests that omission of one or more discrete explanatory variables that divide the data naturally into groups. An examination of the distribution of the residuals may motivate respecification of the model.

Note: The skewness α_3 is defined as $\alpha_3 \equiv \frac{\mu_3}{(\mu_2)^{3/2}}$ where μ_n denotes the nth central moment of a random variable X. The skewness measures the lack of symmetry in the pdf.

Quantile-comparison plot, JF 3.1.3

Quantile-comparison plots are useful for comparing an empirical sample distribution with a theoretical distribution, such as the normal distribution.

Let $P(x)$ represent the theoretical cumulative distribution function (cdf) with which we want to compare the data, that is $P(x) = \Pr(X \leqslant x)$. The quantile-comparison plot is constructed by:

1. Order the data values from smallest to largest, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The $X_{(i)}$ are called the order statistics of the sample.
2. By convention, the cumulative proportion of the data "below" $X_{(i)}$ is given by
$$P_i = \frac{i - \frac{1}{2}}{n}$$
3. Use the inverse of the cdf to find the value z_i corresponding to the cumulative probability P_i, that is
$$z_i = P^{-1}(\frac{i - \frac{1}{2}}{n})$$
4. Plot the z_i as horizontal coordinates against the $X_{(i)}$ as vertical coordinates. If X is sampled from the distribution P, then $X_{(i)} \approx z_i$.
   - if the distributions are identical except for location, then the plot is approximately linear with nonzero intercept, $X_{(i)} \approx \mu + z_i$

- if the distributions are identical except for scale, then the plot is approximately linear with a slope different from 1, $X_{(i)} \approx \sigma z_i$
- if the distributions differ both in location and scale but have the same shape, then $X_{(i)} \approx \mu + \sigma z_i$

5. It is often helpful to place a comparison line on the plot to facilitate the perception of departures from linearity. For a normal quantile-comparison plot (comparing the distribution of the data with the standard normal distribution), we can alternatively use the median as a robust estimator of μ and the interquartile range/1.39 as a robust estimator of σ.

6. We expect some departure from linearity because of sampling variation. It therefore assists interpretation to display the expected degree of sampling error in the plot. The standard error of the order statistic $X_{(i)}$ is

$$\mathrm{SE}(X_{(i)}) = \frac{\hat{\sigma}}{p(z_i)}\sqrt{\frac{P_i(1-P_i)}{n}}$$

where $p(z_i)$ is the probability density function, pdf, corresponding to the CDF $P(z)$. The values along the fitted line are given by $\hat{X}_{(i)} = \hat{\mu} + \hat{\sigma} z_i$. An approximate 95% confidence "envelope" around the fitted line is, therefore,

$$\hat{X}_{(i)} \pm 2 \times \mathrm{SE}(X_{(i)})$$

- Figure 18.1 plots a sample of $n = 100$ observations from a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 10$.

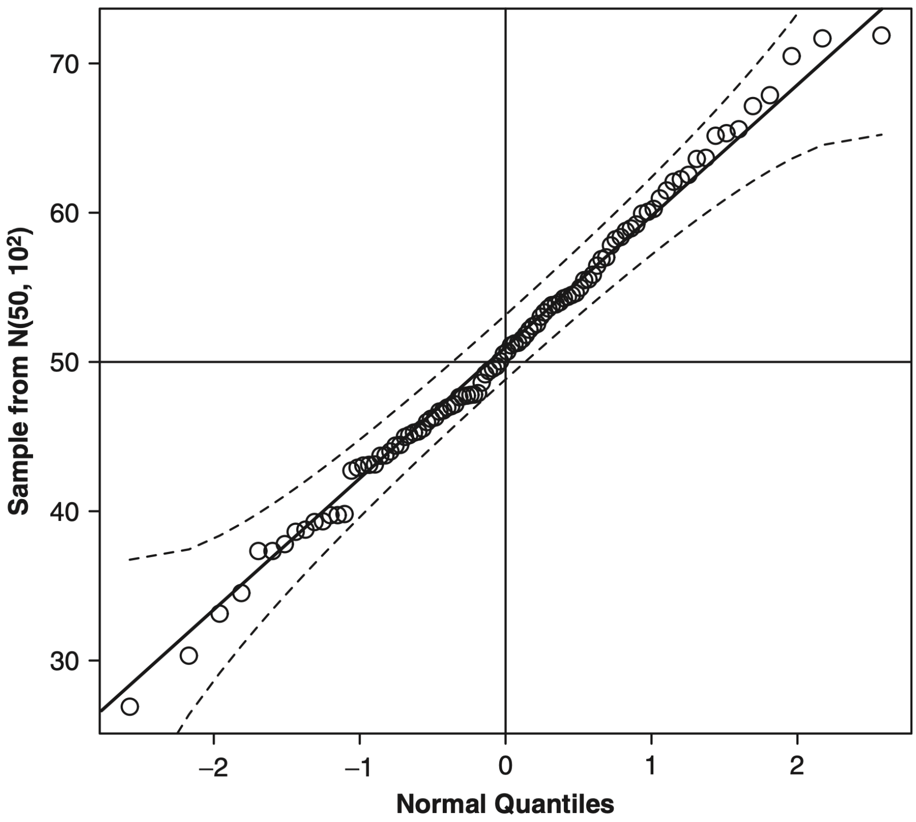

Figure 18.1: Normal quantile-comparison plot for a sample of 100 observations drawn from a normal distribution with mean 50 and standard deviation 10. The fitted line is through the quartiles of the distribution, the broken lines give a pointwise 95% confidence interval around the fit. JF Figure 3.8.

The plotted points are reasonably linear and stay within the rough 95% confidence envelope.

- Figure 18.2 plots a sample of $n = 100$ observations from the positively skewed chi-square distribution with 2 degrees of freedom. The positive skew of the data is reflected in points that lie *above* the comparison line in both tails of the distribution. (In contrast, the tails of negatively skewed data would lie *below* the comparison line.)

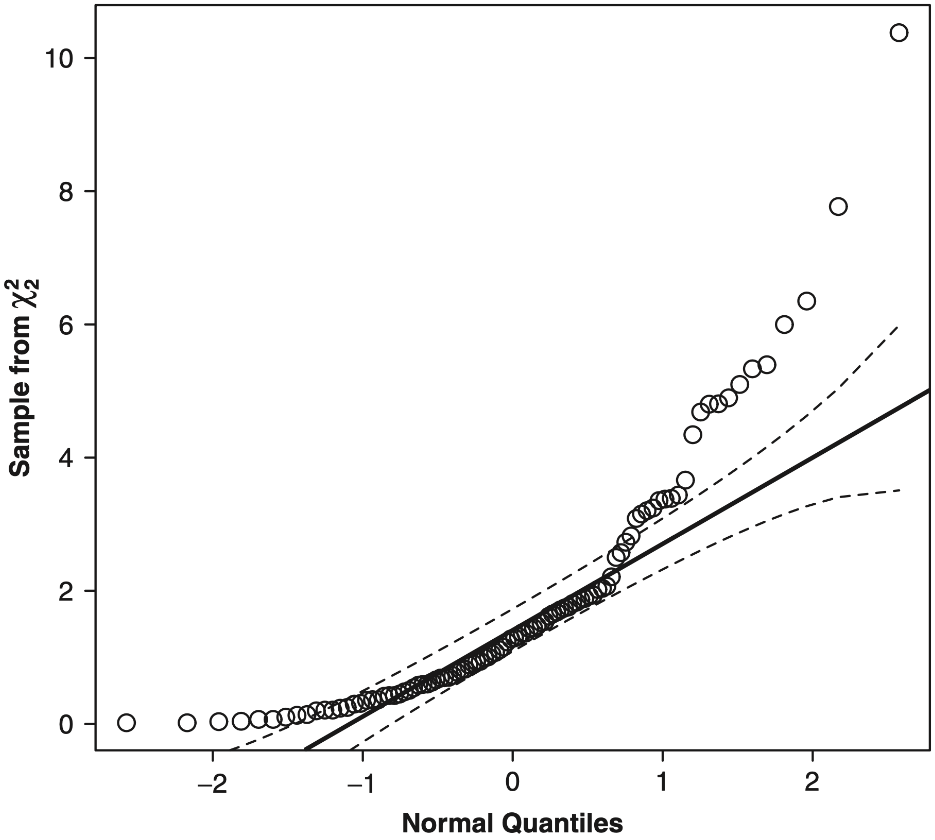

Figure 18.2: Normal quantile-comparison plot for a sample of 100 observations drawn from the positively skewed chi-square distribution with 2 degrees of freedom. JF Figure 3.9.

- Figure 18.3 plots a sample of $n = 100$ observations from the heavy-tailed t distribution with 2 degrees of freedom. In this case, values in the upper tail lie above the corresponding normal quantiles, the values in the lower tail below the corresponding normal quantiles.

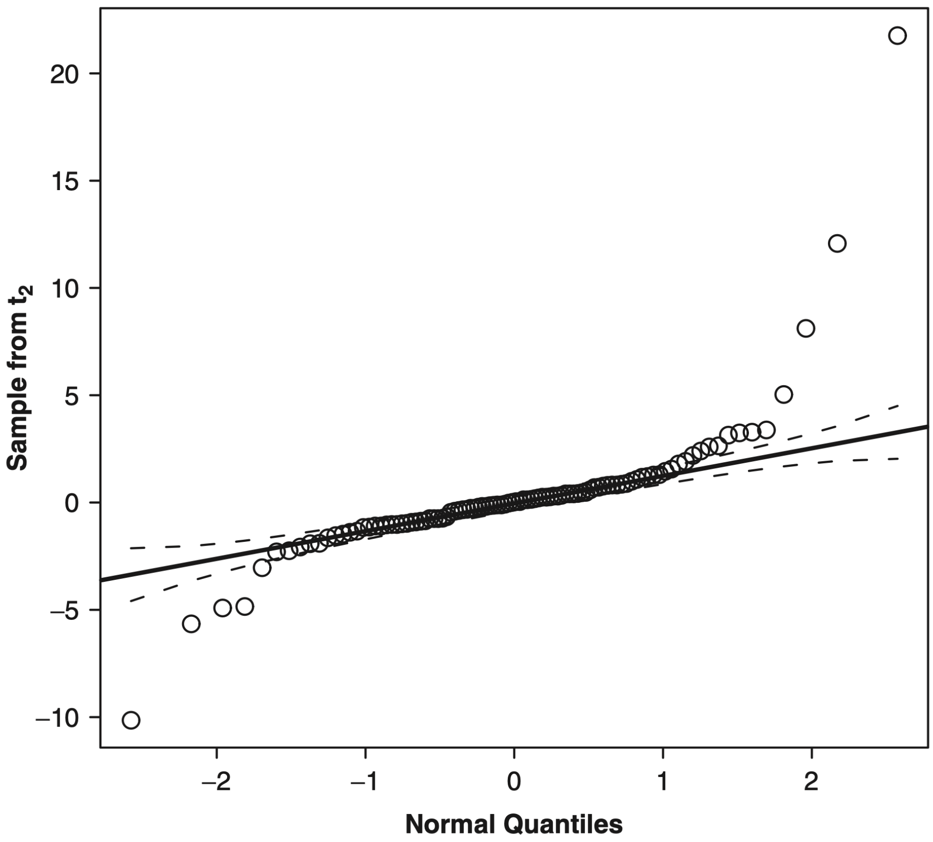

Figure 18.3: Normal quantile-comparison plot for a sample of 100 observations drawn from heavy-tailed t-distribution with 2 degrees of freedom. JF Figure 3.10.

- Figure 18.4 shows the normal quantile-comparison plot for the distribution of infant mortality. The positive skew of the distribution is readily apparent.

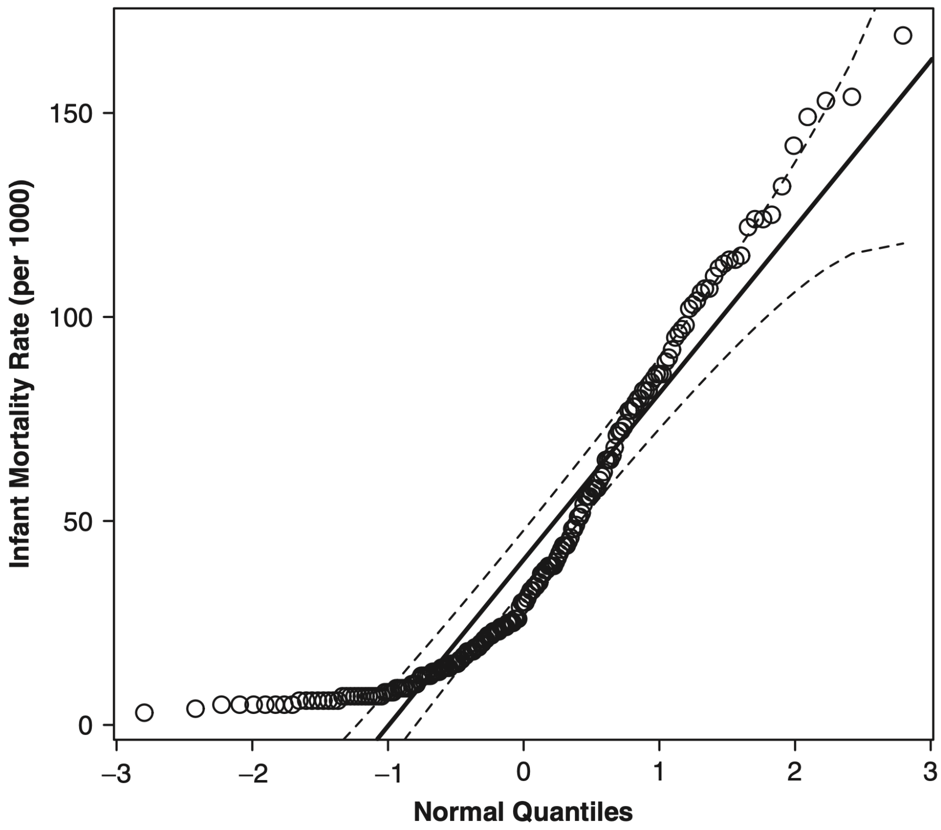

Figure 18.4: Normal quantile-comparison plot for the distribution of infant mortality. Note the positive skew. JF Figure 3.11.

Nonconstant error variance

One of the assumptions of the regression model is that the variation of the response variable around the regression surface (the error variance) is everywhere the same:

$$\mathrm{Var}(\epsilon) = \mathrm{Var}(\mathrm{Y}|\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}) = \sigma_\epsilon^2$$

Constant error variance is often termed homoscedasticity, and similarly, nonconstant error variance is termed heteroscedasticity. We detect nonconstant error variances through graphical methods.

Residual plots

Because the least square residuals have unequal variance even when the constant variance assumption is correct:

$$\mathrm{Var}(\hat{\epsilon}_\mathrm{i}) = \sigma^2(1 - \mathrm{h}_\mathrm{i}).$$

It is preferable to plot studentized residuals against fitted values. A pattern of changing spread is often more easily discerned in a plot of absolute studentized residuals, $|\hat{\epsilon}_i^*|$, or squared studentized residuals, $\hat{\epsilon}_i^{*2}$, against $\hat{Y}$. If the values of $\hat{Y}$ are all positive, then we can plot $\log|\hat{\epsilon}_i^*|$ against $\log\hat{Y}$. Figure 18.5 shows a plot of studentized residuals against fitted values and spread-level plot of studentized residuals, several points with negative fitted values were omitted.

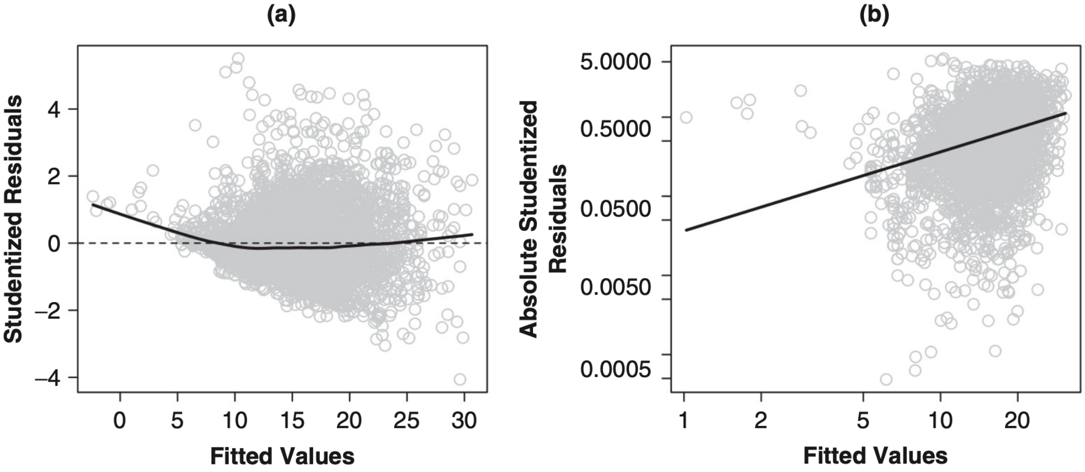

Figure 18.5: (a) Plot of studentized residuals versus fitted values and (b) spread-level plot for studentized residuals. JF Figure 12.3.

It is apparent from both graphs that the residual spread tends to increase with the level of the response, suggesting a violation of constant error variance assumption.

Weighted-least-squares estimation

Weighted-least-squares (WLS) regression provides an alternative approach to estimation in the presence of nonconstant error variance. Suppose that the errors from the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ are independent and normally distributed, with zero means but *different* variances: $\epsilon_i \sim N(0, \sigma_i^2)$. Suppose further that the variances of the errors are known up to a constant of proportionality σ_ϵ^2, so that $\sigma_i^2 = \sigma_\epsilon^2/w_i^2$. Then the likelihood for the model is

$$L(\beta, \sigma_\epsilon^2) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\beta)\right]$$

where $\mathbf{\Sigma}$ is the covariance matrix of the errors,

$$\mathbf{\Sigma} = \sigma_\epsilon^2 \times \text{diag}\{1/w_1^2, \ldots, 1/w_n^2\} \equiv \sigma_\epsilon^2\mathbf{W}^{-1}$$

The maximum-likelihood estimators of β and σ_ϵ^2 are then

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$
$$\hat{\sigma}_\epsilon^2 = \frac{\sum(w_i\hat{\epsilon}_i)^2}{n}$$

Correcting OLS standard errors for nonconstant variance

The covariance matrix of the ordinary-least-squares (OLS) estimator is

$$\begin{aligned}\mathbf{Var}\left(\hat{\beta}\right) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Var}\left(\mathbf{Y}\right)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma_\epsilon^2(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

under the standard assumptions, including the assumption of constant error variance, $\mathbf{Var}\left(\mathbf{Y}\right) = \sigma_\epsilon^2\mathbf{I}_n$. If, however, the errors are heteroscedastic but independent then $\mathbf{\Sigma} \equiv \mathbf{Var}\left(\mathbf{Y}\right) = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$, and

$$\mathbf{Var}\left(\hat{\beta}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

White (1980) shows that the following is a consistent estimator of $\mathbf{Var}\left(\hat{\beta}\right)$

$$\tilde{\text{Var}}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{\Sigma}}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

with $\hat{\mathbf{\Sigma}} = \text{diag}\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2\}$, where $\hat{\sigma}_i^2$ is the OLS residual for observation i.

Subsequent work suggested small modifications to White's coefficient-variance estimator, and in particular simulation studies by Long and Ervin (2000) support the use of

$$\tilde{\text{Var}}^*(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{\Sigma}}^*\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

where $\hat{\mathbf{\Sigma}}^* = \text{diag}\{\hat{\sigma}_i^2/(1-h_i)^2\}$ and h_i is the hat-value associated with observation i. In large samples, where h_i is small, the distinction between $\tilde{\text{Var}}(\hat{\beta})$ and $\tilde{\text{Var}}^*(\hat{\beta})$ essentially disappears.

A rough *rule* is that nonconstant error variance seriouly degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more (or, more conservatively, about 4 or more).

19 Lecture 19: March 3

Last time

- Diagnosing non-normality, non-constant error variance (JF chapter 12)

Today

- Diagnosing nonlinearity (JF chapter 12)
- Data transformation (JF chapter 4)

Nonlinearity

If $\mathbf{E}\left(\mathbf{Y}|\mathbf{X}\right)$ is not linear in $\mathbf{X}$ (in other words, $\mathbf{E}\left(\epsilon|\mathbf{X}\right) \neq 0$ for some x), $\hat{\beta}$ may be biased and inconsistent. Usually we employ "linearity by default" but we should try to make sure this is appropriate: **detect** non-linearities and **model** them accurately.

Lowess smoother, JF 2.3

We can employ local averaging plots to help with diagnostics. Lowess method is in many respects similar to local-averaging smoothers, except that instead of computing an average Y-value within the neighborhood of a focal x, the lowess smoother computes a *fitted* value based on a locally weighted least-squares line, giving more weight to observations in the neighborhood that are close to the focal x than to those relatively far away. The name "lowess" is an acronym for *lo*cally *we*ighted *s*catterplot *s*moother and is sometimes rendered as *loess*, for *lo*cal regr*ess*ion. (If time permitted, we will revisit local regression in Nonparametric Regression.)

Component-plus-residual plots

Component-plut-residual plots are constructed by

1. Compute residuals from full regression:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

2. Compute "linear component" of the partial relationship:

$$C_i = \hat{\beta}_j X_{ij}$$

3. Add linear component to residual to get partial residual for the jth explanatory variable

$$\hat{\epsilon}_i^{(j)} = \hat{\epsilon}_i + C_i = \hat{\epsilon}_i + \hat{\beta}_j X_{ij}$$

4. Plot $\hat{\epsilon}_{\cdot}^{(j)}$ against $X_{\cdot j}$

Figure 19.1 shows the component-plus-residual plots for the regression of log wages on variables (age, education and sex) of the 1994 wave of Statistics Canada's Survey of Labour and Income Dynamics (SLID) data. The SLID data set includes 3997 employed individuals who were between 16 and 65 years of age and who resided in Ontario.

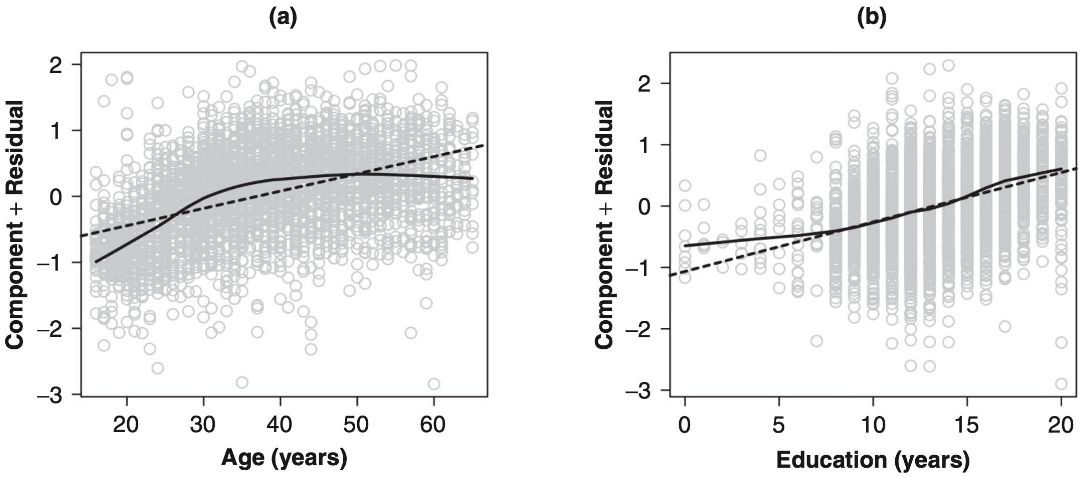

Figure 19.1: Component-plus-residual plots for age and education in SLID regression of log wages on these variables and sex. The solid lines are for lowess smooths with spans of 0.4, and the broken lines are for linear least-squares fits. JF Figure 12.6.

Data transformation

The family of powers and Roots, JF 4.1

A particularly useful group of transformations is the "family" of powers and roots:

$$X \rightarrow X^p$$

wehre the arrow indicates that we intend to replace X with the transformed variable X^p. If p is negative, then the transformation is an inverse power. For example, $X^{-1} = 1/X$. If p is a fraction, then the transformation represents a root. For example, $X^{1/3} = \sqrt[3]{X}$.

It is more convenient to define the family of power transformations in a slightly more complex manner, called the Box-Cox family of transformations (introduced in a seminal paper on transformations by Box & Cox, 1964):

$$X \rightarrow X^{(p)} = \frac{X^p - 1}{p}$$

Because $X^{(p)}$ is a linear function of X^p, the two transformations have the same essential effect on the data, but, as is apparent in Figure 19.2

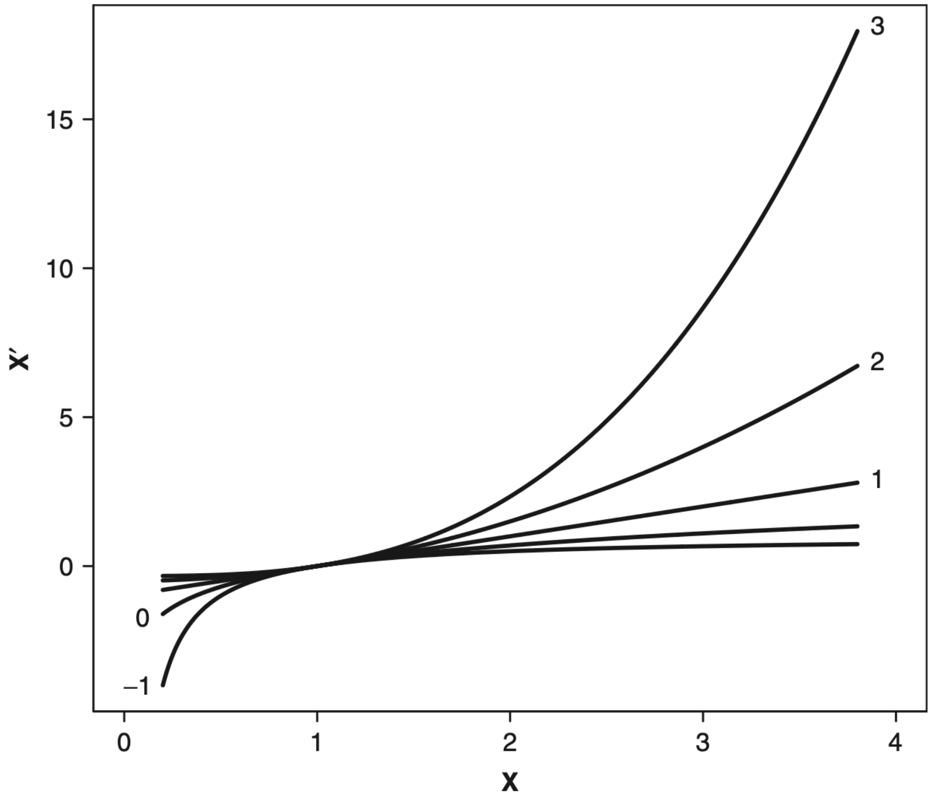

Figure 19.2: The Box-Cox family of power transformations X' of X. The curve labeled p is the transformation $X^{(p)}$, that is $(X^p - 1)/p$; $X^{(0)}$ is $\log_e(X)$. JF Figure 4.1.

- Dividing by p preserves the direction of X, which otherwise would be reversed when p is negative.
- The transformations $X^{(p)}$ are "matched" above $X = 1$ both in level and in slope:
  1. $1^{(p)} = 0$, for all values of p
  2. each transformation has a slope of 1 at $X = 1$.
- Descending the "ladder" of powers and roots towards $X^{(-1)}$ compresses the large values of X and spreads out the small ones. Ascending the ladder of powers and roots towards $X^{(2)}$ has the opposite effect. As p moves further from $p = 1$ (i.e. no transformation) in either direction, the transformation grows more powerful, increasingly "bending" the data.
- The power transformation X^0 is useless because it changes all values to 1, but we can think of the log transformation as a kind of "zeroth" power:

  $$\lim_{p \to 0} \frac{X^p - 1}{p} = \log_e X$$

  and by convention, $X^{(0)} \equiv \log_e X$.

Box-Cox transformation of Y

Box and Cox (1964) suggested a power transformation of Y with the object of normalizing the error distribution, stabilizing the error variance, and straightening the relationship of Y to the Xs. The general Box-Cox model is

$$Y_i^{(\lambda)} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, and

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log_e Y_i & \text{for } \lambda = 0 \end{cases}$$

Note: in statistics, $\log_e$ is often written as log.

For a particular choice of λ, the conditional maximized log-likelihood (see JF 12.5.1 p.324 footnote 55) is

$$\begin{aligned} \log_e L(\beta_0, \beta_1, \ldots, \beta_p, \sigma_\epsilon^2 | \lambda) = & -\frac{n}{2}(1 + \log_e 2\pi) \\ & -\frac{n}{2} \log_e \hat{\sigma}_\epsilon^2(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log_e Y_i \end{aligned}$$

where $\hat{\sigma}_\epsilon^2(\lambda) = \sum \hat{\epsilon}_i^2(\lambda)/n$ and where $\hat{\epsilon}_i(\lambda)$ are the residuals from the least-squares regression of $Y^{(\lambda)}$ on Xs. The least-squares coefficients from this regression are the maximum-likelihood estimates of βs conditional on the values of λ.

A simple procedure for finding the maximum-likelihood estimator $\hat{\lambda}$ is to evaluate the maximized $\log_e L$ (called the profile log-likelihood) for a range of values of λ. To test:$H_0 : \lambda = 1$, calculated the likelihood-ratio statistic

$$G_0^2 = -2[\log_e L(\lambda = 1) - \log_e L(\lambda = \hat{\lambda})]$$

which is asymptotically distributed as χ_1^2 with one degree of freedom under H_0. A 95% confidence interval for λ includes those values for which

$$\log_e L(\lambda) > \log_e L(\lambda = \hat{\lambda}) - 1.92$$

The number 1.92 comes from $\frac{1}{2}\chi_{1,0.05}^2 = 0.5 \times 1.96^2$.

Figure 19.3 shows a plot of the profile log-likelihood against λ for the original SLID regression of composite hourly wages on sex, age, and education. The maximum-likelihood estimate of λ is $\hat{\lambda} = 0.09$, and a 95% confidence interval runs from 0.04 to 0.13. Although 0 is outside of the CI (confidence interval), it is essentially the same transformation of wages as $\lambda = 0.09$ (the correlation between log wages and wages$^{0.09}$ is 0.9996).

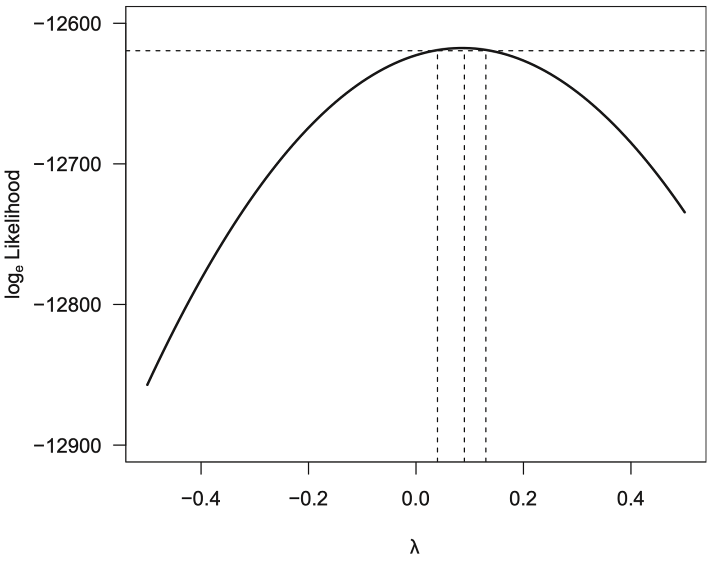

Figure 19.3: Box-Cox transformations for the SLID regression of wages on sex, age, and education. The maximized (profile) log-likelihood is plotted against the transformation parameter λ. The intersection of the line near the top of the graph with the profile log-likelihood curve marks off a 95% confidence interval for λ. The maximum of the log-likelihood corresponds to the MLE of λ. JF Figure 12.14.

Box-Tidwell transformation of Xs

Now, consider the model

$$Y_i = \beta_0 + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_p X_{ip}^{\gamma_p} + \epsilon_i$$

where the errors are independently distributed as $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ and all the X_{ij} are positive.

The parameters of this model ($\beta_0, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_p$, and σ_ϵ^2) could be estimated by general nonlinear least squares. Box and Tidwell (1962) suggested the following computationally more efficient procedure (also yields a constructed-variable diagnostic):

1. Regress Y on $X_1, \ldots, X_p$, obtaining $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. (“Regress A on Bs” is the same as “fitting the linear regression model with A as the response variable and Bs as the explanatory variables”.)

2. Regress Y on $X_1, \ldots, X_p$ and the constructed variables $X_1 \log_e X_1, \ldots, X_p \log_e X_p$ (again, by fitting the model of $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \delta_1 X_1 \log_e X_1 + \cdots + \delta_p X_p \log_e X_p + \epsilon_i$) to obtain $\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_p, \tilde{\delta}_1, \ldots, \tilde{\delta}_p$. In general $\hat{\beta}_i \neq \tilde{\beta}_i$. (The constructed variables result from the first-order Taylor-series approximation to $X_j^{\gamma_j}$ evaluated at $\gamma_j = 1$: $X_j^{\gamma_j} \approx X_1 + (\gamma_1 - 1) X_1 \log_e X_1$.)

3. The constructed variable $X_j \log_e X_j$ can be used to assess the need for a transformation of X_j by testing the null hypothesis $H_0 : \delta_j = 0$. Added-variable plots for the constructed variables are useful for assessing leverage and influence on the decision to transform the Xs.

4. A preliminary estimate of the transformation parameter γ_j (not the MLE) is

$$\tilde{\gamma_j} = 1 + \frac{\tilde{\delta}_j}{\hat{\beta}_j}$$

where $\tilde{\delta}_j$ is from step 2 and $\hat{\beta}_j$ is from step 1.

Polynomial regression

A machinery of multiple regression to fit non-linear relationships between predictor(s) and response.

- Linear: $y = \beta_0 + \beta_1 x + \epsilon$
- Quadratic: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- Cubic: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$
- k^{th} order polynomial: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$

Question:
Does quadratic model provide a significantly better fit than linear model?
Solution: Test $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$.
Alternatively, compare the corresponding adjusted-R^2 values.

21 Lecture 21: March 15

Last time

- Diagnosing nonlinearity (JF chapter 12)
- Data transformation (JF chapter 4)

Today

- Collinearity (JF chapter 13, RD 8.3.2)
- Principal component analysis (JF 13.1.1, RD 8.3.4)

Additional reference

"A First Course in Linear Model Theory" by Nalini Ravishanker and Kipak K. Dey.

Collinearity

In linear model

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\beta + \epsilon\\ \epsilon &\sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)\end{aligned}$$

Collinearity (or multicollinearity) exists when there is "near-dependency" between the columns of the design matrix $\mathbf{X}$.

- Two or more columns.
- In other words, high correlation between explanatory variables.
- the data/model pair is ill-conditioned when $\mathbf{X}^T\mathbf{X}$ is nearly singular.

Perfect collinearity leads to rank-deficiency in $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X}$ is singular. In the case of perfect collinearity, two or more columns are linear-dependent.

An example of perfect collinearity

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

Consider the case, where

- Y_i represents the amount of sales.
- $X_{i1}, X_{i2}, ..., X_{i4}$ are categorical that represent the quarter in which the sample is collected: $X_{ij} = \mathbf{1}(\text{sample } i \text{ collected in quarter } j)$.
- X_{i5} represents expense spent in advertising.

The dummy variable trap $X_{i4} = 1 - X_{i1} - X_{i2} - X_{i3}$. Recall that we need $m - 1$ dummy variables for m categories.

An example of high correlation between predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Consider the case, where

- Y_i represents the salary of individual i.
- X_{i1} represents the age of individual i.
- X_{i2} represents the experience of individual i.

How to interpret β_1?

We expect high correlation between age and experience.

Problems caused by multicollinearity

1. large standard errors of the regression coefficients
   - small associated t-statistics
   - conclusion that truly useful explanatory variables are insignificant in explaining the regression
2. the sign of regression coefficients may be the opposite of what a mechanistic understanding of the problem would suggest
3. deleting a column of the predictor matrix will cause large changes in the coefficient estimates for other variables

However, multicollinearity does **not** greatly affect the **predicted values**.

Signs and detections of multicollinearity

Some signs for multicollinearity:

1. Simple correlation between a pair of predictors exceeds 0.9 or R^2.
2. High value of the multiple correlation coefficient with some high partial correlations between the explanatory variables.
3. Large F-statistics with some small t-statistics for individual regression coefficients

Some approaches for detecting multicollinearity:

1. Pairwise correlations among the explanatory variables
2. Variance inflation factor
3. Condition number

Variance inflation factor

For a multiple linear regression with k explanatory variables. We can regress X_j on the $(k-1)$ other explanatory variables and denote R_j as the coefficient of determination.

Then the <u>variance inflation factor</u> (VIF) is defined as

$$\text{VIF}_j = \frac{1}{1-R_j^2}$$

- $\text{VIF}_j \in [1, +\infty)$
- A suggested threshold is 10
- May use the averaged $\overline{\text{VIF}} = \sum_{j=1}^{k} \text{VIF}_j \Big/ k$.

Condition index and condition number

We first scale the design matrix $\mathbf{X}$ into column-equilibrated predictor matrix $\mathbf{X}_E$ such that $\{X_E\}_{ij} = X_{ij}/\sqrt{\mathbf{X}_j^T\mathbf{X}_j}$.

Let $\mathbf{X}_E = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the singular-value decomposition (SVD) of the $n \times p$ matrix $\mathbf{X}_E$ where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D} = diag(d_1, d_2, ..., d_p)$ is a diagonal matrix with $d_j \geqslant 0$.

The j^{th} <u>condition index</u> is defined as

$$\eta(\mathbf{X}_E) = d_{\max}/d_j, \; j = 1, 2, ..., p$$

The <u>condition number</u> is defined as

$$C = d_{\max}/d_{\min}$$

$C \geqslant 1$, $d_{\max} = \max_{1 \leqslant j \leqslant p} d_j$ and $d_{\min} = \min_{1 \leqslant j \leqslant p} d_j$

Some properties of the condition number

- Large condition number indicates evidence of multicollinearity
- Typical cutoff values, 10, 15 to 30.

Some problems with the condition number

- practitioners have different opinions of whether $\mathbf{X}$ should be centered around their means for SVD.
  - centering may remove nonessential ill conditioning, e.g. $Cor(X, X^2)$
  - centering may mask the role of the constant term in any underlying near-dependencies

- the degree of multicollinearity with dummy variables may be influenced by the choice of reference category
- condition number is affected by the scale of the $\mathbf{X}$ measurements
    - By scaling down any column of $\mathbf{X}$, the condition number can be made arbitrarily large
    - Known as *artificial ill-conditioning*
    - The condition number of the scaled matrix $\mathbf{X}_E$ is also referred to as the *scaled condition number*

Recall that $\mathbf{X}_E = \mathbf{UDV}^T$ is the singular-value decomposition (SVD) of $\mathbf{X}_E$, where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D} = diag(d_1, d_2, ..., d_p)$ is a diagonal matrix with $d_j \geqslant 0$.

Then

$$\begin{aligned}\mathbf{X}_E^T\mathbf{X}_E &= \mathbf{VDU}^T\mathbf{UDV}^T \\ &= \mathbf{VD}^2\mathbf{V}^T\end{aligned}$$

is the spectral decomposition of the Gramian matrix $\mathbf{X}_E^T\mathbf{X}_E$ with $\{d_j^2\}$ being the eigenvalues and $\mathbf{V}$ being the corresponding eigen vector matrix. This relationship links the condition numbers to the eigen values of the Gramian matrix.

Variance decomposition method

The variance-covariance matrix of the coefficient

$$\begin{aligned}Cov(\hat{\beta}) &= \sigma^2(\mathbf{X}_E^T\mathbf{X}_E)^{-1} \\ &= \sigma^2\mathbf{VD}^{-2}\mathbf{V}^T\end{aligned}$$

Its j^{th} diagonal element is the estimated variance of the j^{th} coefficient, $\hat{\beta}_j$. Then

$$Var(\hat{\beta}_j) = \sigma^2 \sum_{h=1}^{p} \frac{v_{jh}^2}{d_h^2}$$

- Let $q_{jh} = \frac{v_{jh}^2}{d_h^2}$ and $q_j = \sum\limits_{h=1}^{p} q_{jh}$.
- The variance decomposition proportion is $\pi_{jh} = q_{jh}/q_j$.
- π_{jh} denotes the proportion of the variance of the j^{th} regression coefficient associated with the h^{th} component of its decomposition.
- The variance decomposition proportion matrix is $\mathbf{\Pi} = \{\pi_{jh}\}$.

Condition	Proportions of variance			
Index	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	...	$Var(\hat{\beta}_3)$
η_1	π_{11}	π_{12}	$\cdots$	π_{1p}
η_2	π_{21}	π_{22}	$\cdots$	π_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
η_p	π_{p1}	π_{p2}	$\cdots$	π_{pp}

Table 1: Table of condition index and proportions of variance

In practice, it is suggested to combine condition index and proportions of variance for multicollinearity diagnostic. Identify multicollinearity if

- Two or more elements in the j^{th} row of matrix $\mathbf{\Pi}$ are relatively large
- And its associated condition index η_j is large too

Principal Components

The method of principal components, introduced by Karl Pearson (1901) and Harold Hotelling (1933), provides a useful representation of the correlational structure of a set of variables. Some advantages of the principal component analysis include

- more unified
- linear transformation of the original predictors into a new set of orthogonal predictors
- the new orthogonal predictors are called principal components

Principal components regression is an approach that inspects the sample data $(\mathbf{Y}, \mathbf{X})$ for directions of variability and uses this information to reduce the dimensionality of the estimation problem. The procedure is based on the observation that every linear regression model can be restated in terms of a set of orthogonal predictor variables, which are constructed as linear combinations of the original variables. The new orthogonal variables are called the principal components of the original variables.

Let $\mathbf{X}^T\mathbf{X} = \mathbf{Q}\mathbf{\Delta}\mathbf{Q}^T$ denote the spectral decomposition of $\mathbf{X}^T\mathbf{X}$, where $\mathbf{\Delta} = diag\{\lambda_1, \ldots, \lambda_p\}$ is a diagonal matrix consisting of the (real) eigenvalues of $\mathbf{X}^T\mathbf{X}$, with $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ and $\mathbf{Q} = (\mathbf{q_1}, \ldots, \mathbf{q_p})$ denotes the matrix whose columns are the orthogonal eigenvectors of $\mathbf{X}^T\mathbf{X}$ corresponding to the ordered eigenvalues. Consider the transformation

$$\mathbf{Y} = \mathbf{X}\mathbf{Q}\mathbf{Q}^T\beta + \epsilon = \mathbf{Z}\theta + \epsilon,$$

where $\mathbf{Z} = \mathbf{X}\mathbf{Q}$, and $\theta = \mathbf{Q}^T\beta$.
The elements of θ are known as the regression parameters of the principal components. The matrix $\mathbf{Z} = \{\mathbf{z_1}, \ldots, \mathbf{z_p}\}$ is called the matrix of principal components of $\mathbf{X}^T\mathbf{X}$. $\mathbf{z}_j = \mathbf{X}\mathbf{q}_j$ is the jth principal component of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{z}_j^T\mathbf{z}_j = \lambda_j$, the jth largest eigenvalue of $\mathbf{X}^T\mathbf{X}$.

Principal components regression consists of deleting one or more of the variables $\mathbf{z}_j$ (which correspond to small values of λ_j), and using OLS estimation on the resulting reduced regression model.

Derivation under standardized predictors, JF 13.1.1

Consider the vectors of standardized predictors, $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_p^*$ (obtained by subtracting the mean and divided by standard deviation of the original predictor vectors). Because the principal components are linear combinations of the original predictors, we write the first principal component as

$$\begin{aligned}\mathbf{w}_1 &= A_{11}\mathbf{x}_1^* + A_{21}\mathbf{x}_2^* + \cdots + A_{p1}\mathbf{x}_p^* \\ &= \mathbf{X}^*\mathbf{a}_1\end{aligned}$$

The variance of the first component becomes

$$\begin{aligned}S_{w_1}^2 &= \frac{1}{n-1}\mathbf{w}_1^T\mathbf{w}_1 \\ &= \frac{1}{n-1}\mathbf{a}_1^T\mathbf{X}^{*T}\mathbf{X}^*\mathbf{a}_1 \\ &= \mathbf{a}_1^T\mathbf{R}_{XX}\mathbf{a}_1\end{aligned}$$

where $\mathbf{R}_{XX} = \frac{1}{n-1}\mathbf{X}^{*T}\mathbf{X}^*$. We want to maximize $S_{w_1}^2$ under the normalizing constraint $\mathbf{a}_1^T\mathbf{a}_1 = 1$ (otherwise $S_{w_1}^2$ can be arbitrarily large by inflating $\mathbf{a}_1$). Consider

$$F_1 \equiv \mathbf{a}^T\mathbf{R}_{XX}\mathbf{a}_1 - L_1(\mathbf{a}_1^T\mathbf{a}_1 - 1)$$

where L_1 is a Lagrange multiplier. By differentiating this equation with respect to $\mathbf{a}_1$ and L_1,

$$\begin{aligned}\frac{\partial F_1}{\partial \mathbf{a}_1} &= 2\mathbf{R}_{XX}\mathbf{a}_1 - 2L_1\mathbf{a}_1 \\ \frac{\partial F_1}{\partial L_1} &= -(\mathbf{a}_1^T\mathbf{a}_1 - 1)\end{aligned}$$

Setting the partial derivatives to 0 produces

$$\begin{aligned}(\mathbf{R}_{XX} - L_1\mathbf{I}_p)\mathbf{a}_1 &= \mathbf{0} \\ \mathbf{a}_1^T\mathbf{a}_1 &= 1\end{aligned}$$

From the first equation, we see that L_1 is an eigenvalue of $\mathbf{R}_{XX}$ such that $\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1$ such that

$$S_{w_1}^2 = \mathbf{a}_1^T\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1^T\mathbf{a}_1 = L_1$$

To maximize $S_{w_1}^2$, we only need to pick the largest eigenvalue of $\mathbf{R}_{XX}$.

22 Lecture 22: March 17

Last time

- Collinearity (JF chapter 13, RD 8.3.2)
- Principal component analysis (JF 13.1.1, RD 8.3.4)

Today

- Midterm exam review
- Biased estimation (JF 13.2.3, CG's notes)
  - Ridge regression
  - Lasso regression

Additional reference

Lecture notes by Cedric Ginestet

Ridge Regression

Ridge regression and the Lasso regression are two forms of regularized regression. These methods can be used to alleviate the consequences of multicollinearity.

1. When variables are highly correlated, a large coefficient in one variable may be alleviated by a large coefficient in another variable, which is negatively correlated to the former.
2. Regularization imposes an upper threshold on the values taken by the coefficients, thereby producing a more parsimonious solution, and a set of coefficients with smaller variance.

Constrained optimization

Ridge regression is motivated by a constrained minimization problem, which can be formulated as

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2$$

$$\text{subject to } ||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2 \leqslant t$$

for $t \geqslant 0$.

Use a Lagrange multiplier, we can rewrite the formula as

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \}$$

for $\lambda \geqslant 0$ and where there is a one-to-one correspondence between t and λ. λ is an arbitrary constant usually referred to as the "ridge constant".

Analytical solutions

The ridge-regression estimator has analytical solution

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

This is obtained by differentiating the objective function with respect to β and set it to 0:

$$\begin{aligned}&\frac{\partial}{\partial\beta}\{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta\}\\ =&2(\mathbf{X}^T\mathbf{X})\beta - 2\mathbf{X}^T\mathbf{Y} + 2\lambda\beta\\ =&0\end{aligned}$$

Therefore,

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^T\mathbf{Y}$$

Since we are adding a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$, we are , in general, producing an invertible matrix, $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ even if $\mathbf{X}^T\mathbf{X}$ is singular. Historically, this particular aspect of ridge regression was the main motivation behind the adoption of this particular extension of OLS theory.

The ridge regression estimator is related to the classical OLS estimator, $\hat{\beta}^{OLS}$, in the following manner

$$\hat{\beta}^{ridge} = \left[\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right]^{-1}\hat{\beta}^{OLS},$$

assuming $\mathbf{X}^T\mathbf{X}$ is non-singular. This relationship can be verified by applying the definition of $\hat{\beta}^{OLS}$,

$$\begin{aligned}\hat{\beta}^{ridge} &= \left[\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}\end{aligned}$$

using the fact $\mathbf{B}^{-1}\mathbf{A}^{-1} = (\mathbf{AB})^{-1}$.

Moreover, when $\mathbf{X}$ is composed of orthonormal variables, such that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, it then follows that

$$\hat{\beta}^{ridge} = \frac{1}{1+\lambda}\hat{\beta}^{OLS}$$

Bias and variance of ridge estimator

Ridge estimation produces a biased estimator of the true parameter β. With the definition of $\hat{\beta}^{ridge}$ and the model assumption $\mathbf{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$, we obtain,

$$\begin{aligned}\mathbf{E}\left(\hat{\beta}^{ridge}|\mathbf{X}\right) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\beta\\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \lambda\mathbf{I})\beta\\ &= \beta - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\beta\end{aligned}$$

where the bias of the ridge estimator is proportional to λ. The variance of the ridge estimator is

$$\mathbf{Var}\left(\hat{\beta}^{ridge}|\mathbf{X}\right) = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}.$$

When λ increases, the inverted term $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$ is increasingly dominated by $\lambda\mathbf{I}$. The variance of the ridge estimator, therefore, is a decreasing function of λ. This result is intuitively reasonable because the estimator itself is driven toward $\mathbf{0}$.

Variance-bias tradeoff

The mean-squared error of an estimator can be decomposed into the sum of its squared bias and sampling variance.

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbf{E}\left((\hat{\theta} - \theta)^2\right) = \mathbf{E}(\hat{\theta}^2) + \theta^2 - 2\theta\mathbf{E}(\hat{\theta}) \\ Bias^2(\hat{\theta}) &= \left[\mathbf{E}(\hat{\theta}) - \theta\right]^2 = \mathbf{E}^2(\hat{\theta}) + \theta^2 - 2\theta\mathbf{E}(\hat{\theta}) \\ \mathrm{Var}(\hat{\theta}) &= \mathbf{E}(\hat{\theta}^2) - \mathbf{E}^2(\hat{\theta}) \end{aligned}$$

Therefore

$$MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})$$

The essential idea here is to trade a small amount of bias in the coefficient estimates for a large reduction in coefficient sampling variance. Hoerl and Kennard (1970) prove that it is always possible to choose a positive value of the ridge constant λ so that the mean-squared error of the ridge estimator is less than the mean-squared error of the least-squares estimator. These ideas are illustrated heuristically in Figure 22.1

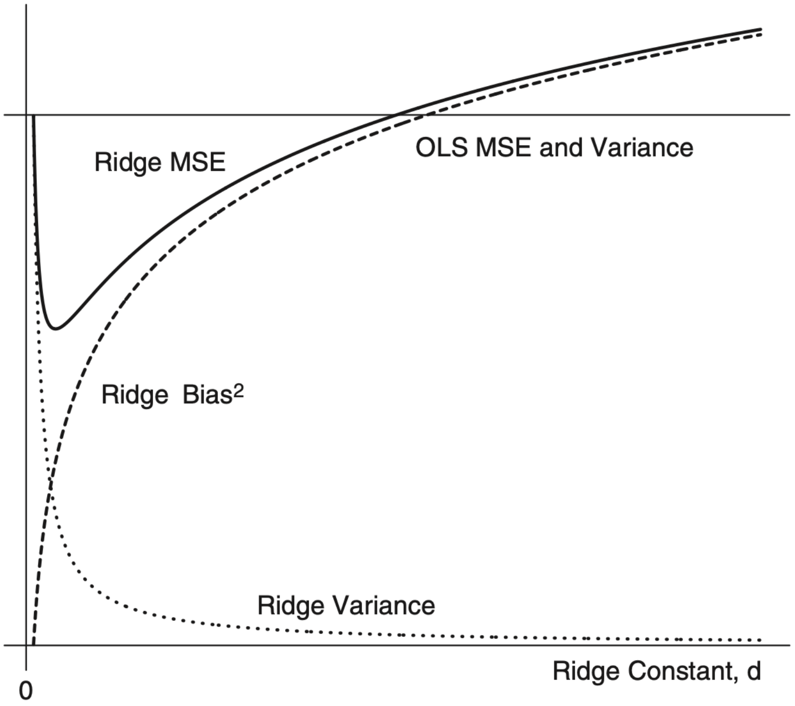

Figure 22.1: Trade-off of bias and against variance for the ridge-regression estimator. The horizontal line gives the variance of the least-squares (OLS) estimator; because the OLS estimator is unbiased, its variance and mean-squared error are the same. The broken line shows the squared bias of the ridge estimator as an increasing function of the ridge constant d (i.e. λ in our notes). The dotted line shows the variance of the ridge estimator. The mean-squared error (MSE) of the ridge estimator, given by the heavier solid line, is the sum of its variance and squared bias. For some values of d, the MSE error of the ridge estimator is below the variance of the OLS estimator. JF Figure 13.9.

Lasso regression

We have seen that ridge regression essentially re-scales the OLS estimates. The lasso, by contrast, tries to produce a *sparse* solution, in the sense that several of the slope parameters will be set to zero.

Constrained optimization

Different from the L_2 penalty for ridge regression, the Lasso regression employs L_1-penalty.

$$\hat{\beta}^{lasso} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2$$

$$\text{subject to } ||\beta||_1 = \sum_{j=1}^{p} |\beta_j| \leqslant t$$

for $t \geqslant 0$; which can again be re-formulated using the Lagrangian for the L_1-penalty,

$$\hat{\beta}^{lasso} = \arg\min_{\beta \in \mathbb{R}^p} \{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \}$$

where $\lambda > 0$ and, as before, there exists a one-to-one correspondence between t and λ.

Parameter estimation

Contrary to ridge regression, the Lasso does not have a closed-form solution. The L_1-penalty makes the solution non-linear in y_i's. The above constrained minimization is a quadratic programming problem, for which many solvers exist.

Choice of Hyperparameters

Regularization parameter

The choice of λ in both ridge and lasso regressions is more of an art than a science. This parameter can be constructed as a complexity parameter, since as λ increases, less and less effective parameters are likely to be included in both ridge and lasso regressions. Therefore, one can adopt a model selection perspective and compare different choices of λ using cross-validation or an information criterion. That is, the value of λ should be chosen adaptively, in order to minimize an estimate of the expected prediction error (as in cross-validation), for instance, which is well approximated by AIC. We will discuss model selection in more detail later.

Bayesian perspective

The penalty terms in ridge and lasso regression can also be justified, using a Bayesian framework, whereby these terms arise as aresult of the specification of a particular prior distribution on the vector of slope parameters.

1. The use of an L_2-penalty in multiple regression is analogous to the choice of a Normal prior on the β_j's, in Bayesian statistics.

$$\begin{aligned} y_i &\overset{iid}{\sim} \mathcal{N}(\beta_0 + \mathbf{x}_i^T \beta, \sigma^2), \quad i = 1, \ldots, n \\ \beta_j &\overset{iid}{\sim} \mathcal{N}(0, \tau^2), \quad j = 1, \ldots, p \end{aligned}$$

2. Similarly, the use of an L_1-penalty in multiple regression is analogous to the choice of a Laplace prior on the β_j's, such that

$$\beta_j \overset{iid}{\sim} Laplace(0, \tau^2), \quad j = 1, \ldots, p$$

In both cases, the value of the hyperparameter, τ^2, will be inversely proportional to the choice of the particular value for λ. For ridge regression, λ is exactly equal to the shrinkage parameter of the hierarchical model, $\lambda = \sigma^2/\tau^2$.

23 Lecture 23: March 24

Last time

- Midterm exam review
- Biased estimation (JF 13.2.3, CG's notes)
  - Ridge regression

Today

- Midterm evaluation suggestions
- Poll on alternative grade path
- Lasso regression (CG's notes)
- Model selection (JF 22)

Additional reference

Lecture notes by Cedric Ginestet

Midterm evaluation suggestions

1. What has been most helpful for your learning in this class so far?
   - lecture notes
   - lab session
2. What has caused you the most difficulty in terms of learning in this class so far?
   - Prior knowledge
   - R
   - More explanation in writing besides talking
   - prewritten slides
   - disconnection between lecture notes and homework (XJ: not really sure)
3. What suggestion(s) do you have that would enhance your learning experience in this class?
   - more examples
   - more R practice
   - partially filled slides and questions
   - more comments on homework grading (XJ: probably redundant with in-class reviews)

4. Please share any additional comments about the content, format or instructor that you have about the course. Your comments are appreciated!

   - appreciate TA and instructor's effort
   - It's cool that we get to see student presentations throughout the semester instead of all at once during finals week like other classes.
   - I know that the class is meant to move at a fast pace, but I feel I am struggling to stay on top of everything.

Alternative grade path proposal

The alternative grade path would be an addition to the original grading scheme. Let $S_{original} = w_{homework}s_{homework} + w_{mid-term}s_{mid-term} + w_{final}s_{final} + w_{presentation}s_{presentation}$ represent the final grade from the original grading scheme. Then the proposed alternative grading scheme has $S_{alternative} = w_{homework}s_{homework} + (w_{mid-term} + w_{final})s_{final} + w_{presentation}s_{presentation}$. And the total score will be $S_{total} = \max\{S_{original}, S_{alternative}\}$.

Lasso regression

We have seen that ridge regression essentially re-scales the OLS estimates. The lasso, by contrast, tries to produce a *sparse* solution, in the sense that several of the slope parameters will be set to zero.

Constrained optimization

Different from the L_2 penalty for ridge regression, the Lasso regression employs L_1-penalty.

$$\begin{aligned}\hat{\beta}^{lasso} &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2 \\ \text{subject to } ||\beta||_1 &= \sum_{j=1}^{p}|\beta_j| \leqslant t\end{aligned}$$

for $t \geqslant 0$; which can again be re-formulated using the Lagrangian for the L_1-penalty,

$$\hat{\beta}^{lasso} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\}$$

where $\lambda > 0$ and, as before, there exists a one-to-one correspondence between t and λ.

Parameter estimation

Contrary to ridge regression, the Lasso does not have a closed-form solution. The L_1-penalty makes the solution non-linear in y_i's. The above constrained minimization is a quadratic programming problem, for which many solvers exist.

Choice of Hyperparameters

Regularization parameter

The choice of λ in both ridge and lasso regressions is more of an art than a science. This parameter can be constructed as a complexity parameter, since as λ increases, less and less effective parameters are likely to be included in both ridge and lasso regressions. Therefore, one can adopt a model selection perspective and compare different choices of λ using cross-validation or an information criterion. That is, the value of λ should be chosen adaptively, in order to minimize an estimate of the expected prediction error (as in cross-validation), for instance, which is well approximated by AIC. We will discuss model selection in more detail later.

Bayesian perspective

The penalty terms in ridge and lasso regression can also be justified, using a Bayesian framework, whereby these terms arise as aresult of the specification of a particular prior distribution on the vector of slope parameters.

1. The use of an L_2-penalty in multiple regression is analogous to the choice of a Normal prior on the β_j's, in Bayesian statistics.

$$\begin{aligned} y_i &\overset{iid}{\sim} \mathcal{N}(\beta_0 + \mathbf{x}_i^T\beta, \sigma^2), \quad i = 1, \dots, n \\ \beta_j &\overset{iid}{\sim} \mathcal{N}(0, \tau^2), \quad j = 1, \dots, p \end{aligned}$$

2. Similarly, the use of an L_1-penalty in multiple regression is analogous to the choice of a Laplace prior on the β_j's, such that

$$\beta_j \overset{iid}{\sim} Laplace(0, \tau^2), \quad j = 1, \dots, p$$

In both cases, the value of the hyperparameter, τ^2, will be inversely proportional to the choice of the particular value for λ. For ridge regression, λ is exactly equal to the shrinkage parameter of the hierarchical model, $\lambda = \sigma^2/\tau^2$.

Model selection

Model selection is conceptually simplest when our goal is *prediction* – that is, the development of a regression model that will predict new data as accurately as possible. However, prediction is not often the only desirable characteristic in a statistical model that model interpretation, data summary and explanations are also desired. We discuss several criteria for selecting among m competing statistical models $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$ for n observations of a response variable Y and associated predictors Xs.

Adjsted-R^2

The squared multiple correlation "corrected" (or "adjusted") for degrees of freedom is intuitively reasonable criterion for comparing linear-regression models with different numbers of parameters. Suppose model M_j is one of the models under consideration. If M_j has s_j regression coefficients (including the regression constant) and is fit to a data set with n observations, then the adjusted-R^2 for the model is

$$R^2_{adj,j} = 1 - \frac{n-1}{n-s_j} \times \frac{RSS_j}{TSS}$$

Models with relatively large numbers of parameters are penalized for their lack of parsimony. The model with the highest adjusted-R^2 value is selected as the best model. Beyond this intuitive rationale, however, there is no deep justification for using R^2_{adj} as a model selection criterion.

Cross-validation and generalized cross-validation

The key idea in cross-validation (more accurately, leave-one-out cross-validation) is to omit the ith observation to obtain an estimate of $E(Y|x_i)$ based on the other observations as $\hat{Y}^{(j)}_{-i}$ for model M_j. Omitting the ith observation makes the fitted value $\hat{Y}^{(j)}_{-i}$ independent of the observed value Y_i. The cross-validation criterion for model M_j is

$$CV_j \equiv \frac{\sum_{i=1}^{n} \left[\hat{Y}^{(j)}_{-i} - Y_i\right]^2}{n}$$

We prefer the model with the smallest value of CV_j.

In linear least-squares regression, there are efficient procedures for computing the leava-one-out fitted values $\hat{Y}^{(j)}_{-i}$ that do not require literally refitting the model (recall the discussions of standardized residuals). However, in other applications, leave-one-out cross-validation can be computationally expensive (that requires literally refitting the model n times).

An alternative is to divide the data into a relatively small number of subsets of roughly equal size and to fit the model omitting one subset at a time, obtaining fitted values for all observations in the omitted subset. This method is termed as K-fold cross-validation where K is the number of subsets. The cross-validation criterion is defined the same way as before.

An alternative criterion is to approximate CV by the generalized cross-validation criterion

$$GCV_j \equiv \frac{n \times RSS_j}{df^2_{res_j}}$$

which however is less popular given the increasing computational power we have in the modern era.

AIC and BIC

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also popular model selection criteria. Both are members of a more general family of *penalized* model-fit statistics (in the form of "*IC"), applicable to regression models fit by maximum likelihood, that take the form

$$*IC_j = -2\log_e L(\hat{\theta}_j) + cs_j$$

where $L(\hat{\theta}_j)$ is the maximized likelihood under model M_j; $\hat{\theta}_j$ is the vector of parameters of the model (including, for example, regression coefficients and an error variance); s_j is the number of parameters in $\hat{\theta}_j$; and c is a constant that differs from one model selection criterion to another. The first term, $-2\log_e L(\hat{\theta}_j)$, is the residual deviance under the model; for a linear model with normal errors, it is simply the residual sum of squares.

The model with the smallest *IC is the one that receives most support from the data (the selected model). The AIC and BIC are defined as follows:

$$\begin{aligned} AIC_j &\equiv -2\log_e L(\hat{\theta}_j) + 2s_j \\ BIC_j &\equiv -2\log_e L(\hat{\theta}_j) + s_j \log_e(n) \end{aligned}$$

The lack-of-parsimony penalty for the BIC grows with the sample size, while that for the AIC does not. When $n \geqslant 8$ the penalty for the BIC is larger than that for the AIC resulting in BIC tends to nominate models with fewer parameters. Both AIC and BIC are based on deeper statistical considerations, please refer to JF 22.1 sections **A closer look at the AIC** and **A closer look at the BIC** for more details.

Sequential procedures

Besides the ranking systems above, there is another class loosely defined as sequential procedures for model selection.

1. Forward selection
2. Backwards elimination
3. Stepwise selection

Forward selection :

1. Choose a threshold significance level for adding predictors, "SLENTRY" (SL stands for significance level). For example, $SLENTRY = 0.10$.
2. Initialize with $y = \beta_0 + \epsilon$.
3. Form a set of candidate models that differ from the working model by addition of one new predictor
4. Do any of the added predictors have $p - value \leqslant SLENTRY$?
   - Yes: add predictor with smallest p-value to working model + repeat steps 3 to 4.
   - No: stop. Final model = working model.

Backwards elimination

1. Choose threshold level for removing predictors. For example, $SLSTAY = 0.05$.
2. Initialize with most general model (biggest possible): $y = \beta_0 + \beta_1 x_1 + \cdots + \epsilon$.
3. Form a set of candidate models that differ from working model by deletion of one term
4. Do any $p - value > SLSTAY$ (from fitting the current working model)?
   - Yes: remove the term with largest p-value and repeat steps 3 and 4.
   - No: stop. Final model = working model.

Stepwise Alternate forwards + backwards steps. Initialize with $y = \beta_0 + \epsilon$. Stop when consecutive forward + backward steps do not change working model. $(SLENTRY \leqslant SLSTAY)$

Some examples

- Model selection by AIC
- Model selection by AIC and Lasso

25 Lecture 25: March 29

Last time

- Lab session

Today

- Announcement: alternative grading path didn't pass (5:5 from last poll + a fail on the first poll)
- Analysis of Variance (JF chapter 8)
  - one-way anova
  - two-way anova

Additional reference

Course notes by Dr. Jason Osborne.

Analysis of Variance

The term analysis of variance is used to describe the partition of the response-variable sum of squares into "explained" and "unexplained" components, noting that this decomposition applies generally to linear models. For historical reasons, analysis of variance (abbreviated ANOVA) also refers to procedures for fitting and testing linear models in which the explanatory variables are categorical.

One-way ANOVA

Suppose that there are *no* quantitative explanatory variables, but only a single factor (categorical data). For example, for a three-category classification, we have the model

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i \tag{3}$$

employing the following coding for the dummy regressors:

Group	D_1	D_2
1	1	0
2	0	1
3	0	0

The expectation of the response variable in each group (i.e. in each category or level of the factor) is the population group mean, denoted by μ_j for the jth group. Equation 3 produces the following relationship between group means and model parameters:

$$\begin{aligned}
&\text{Group 1: } E(Y_i|D_{i1}=1, D_{i2}=0) = \alpha + \gamma_1 \times 1 + \gamma_2 \times 0 = \alpha + \gamma_1 \\
&\text{Group 2: } E(Y_i|D_{i1}=0, D_{i2}=1) = \alpha + \gamma_1 \times 0 + \gamma_2 \times 1 = \alpha + \gamma_2 \\
&\text{Group 3: } E(Y_i|D_{i1}=0, D_{i2}=0) = \alpha + \gamma_1 \times 0 + \gamma_2 \times 0 = \alpha
\end{aligned}$$

There are three parameters (α, γ_1 and γ_2) and three group means, so we can solve uniquely for the parameters in terms of the group means:

$$\begin{aligned}
\hat{\alpha} &= \mu_3 \\
\hat{\gamma}_1 &= \mu_1 - \mu_3 \\
\hat{\gamma}_2 &= \mu_2 - \mu_3
\end{aligned}$$

Not surprisingly, α represents the mean of the baseline category (Group 3) and that γ_1 and γ_2 captures differences between the other group means and the mean of the baseline category.

notations

Because observations are partitioned according to groups, it is convenient to let Y_{jk} denote the kth observation within the jth of m groups. The number of observations in the jth group is n_j, and the total number of observations is $n = \sum_{j=1}^{m} n_j$. Let $\mu_j \equiv E(Y_{jk})$ be the population mean in group j.

The one-way ANOVA model is

$$Y_{jk} = \mu + \alpha_j + \epsilon_{jk}$$

where μ represents the general level of response variable in the population; α_j represents the effect on the response variable of membership in the jth group; ϵ_{jk} is an error variable that follows the usual linear-model assumptions: $\epsilon_{jk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

By taking expectations, we have

$$\mu_j = \mu + \alpha_j$$

The parameters of the model are, therefore, underdetermined, for there are $m+1$ parameters (including μ) but only m population group means (recall the dummy variable trap introduced in collinearity). To produce easily interpretable parameters and that estimates and generalizes usefully to more complex models, we impose the sum-to-zero constraint

$$\sum_{j=1}^{m} \alpha_j = 0$$

With the sum-to-zero constraint, we solve for the parameters

$$\begin{aligned}
\hat{\mu} &= \frac{\sum \mu_j}{m} \\
\hat{\alpha}_j &= \mu_j - \mu
\end{aligned}$$

The fitted Y values are the group means for the one-way ANOVA model:

$$\hat{Y}_{jk} = \hat{\mu} + \hat{\alpha}_j$$

and the regression and residual sums of squares therefore take particularly simple forms in one-way ANOVA:

$$RegSS = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(\hat{Y}_{jk} - \bar{Y})^2 = \sum_{j=1}^{m} n_j(\bar{Y}_j - \bar{Y})^2$$
$$RSS = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(Y_{jk} - \hat{Y}_{jk})^2 = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(Y_{jk} - \bar{Y}_j)^2$$

and can be presented in an ANOVA table.

Table 2: General one-way ANOVA table

Source	Sum of Squares	df	Mean Square	F	H_0
Groups	$\sum n_j(\bar{Y}_j - \bar{Y})^2$	$m-1$	$\frac{RegSS}{m-1}$	$\frac{RegMS}{RMS}$	$\alpha_1 = \cdots = \alpha_m = 0$
Residuals	$\sum\sum(Y_{jk} - \bar{Y}_j)^2$	$n-m$	$\frac{RSS}{n-m}$		
Total	$\sum\sum(Y_{jk} - \bar{Y})^2$	$n-1$			

Sometimes, the column of Source can also be denoted with Treatments (for Groups) and Error (for Residuals). And a balanced one-way ANOVA model has the same number of observations in one group (or treatment), in other words, $n_1 = \cdots = n_m = \frac{n}{m}$.

one-way ANOVA example

The following data come from study investigating binding fraction for several antibiotics using $n = 20$ bovine serum samples:

Antibiotic	Binding Percentage	Sample mean
Penicillin G	29.6 24.3 28.5 32.0	28.6
Tetracyclin	27.3 32.6 30.8 34.8	31.4
Streptomycin	5.8 6.2 11.0 8.3	7.8
Erythromycin	21.6 17.4 18.3 19	19.1
Chloramphenicol	29.2 32.8 25.0 24.2	27.8

Question: Are the population means for these 5 treatments plausibly equal?
Answer:
One model parameterizes antibiotic effects as differences from mean

$$Y_{jk} = \mu + \alpha_j + \epsilon_{jk}$$

for $j = 1, \ldots, 5$ and $k = 1, \ldots, 4$, where $\epsilon_{jk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ errors.

Unknown parameters

1. μ - overall population mean (average of 5 treatment population means)
2. α_j - difference between (population) mean for treatment j and μ
3. σ^2 - (population) variance of binding fraction for a given antibiotic

To test $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_5 = 0$, we just carry out one-way ANOVA:

Source	Sum of Squares	df	Mean Square	F
Groups	1481	4	370	41
Residuals	136	15	9	
Total	1617	19		

Compared to $F(0.05, 4, 15) = 3.06$, we have $F = 41 > 3.06$.
Conclusion: we reject the null hypothesis of all population means for 5 treatment being equal at 0.05 significance level.

What do we obtain standard errors of parameter estimates? (HW)

26 Lecture 26: March 31

Last time

- One-way ANOVA

Today

- Announcement: alternative grading path didn't pass (5:5 from last poll + a fail on the first poll)
- Analysis of Variance (JF chapter 8)
    - two-way ANOVA

Additional reference

Course notes by Dr. Jason Osborne.

Two-Way ANOVA

The inclusion of a second factor permits us to model and test partial relationships, as well as to introduce interactions. Let's take a look at the patterns of relationship that can occur when a quantitative response variable is classified by two factors.

Patterns of Means in the two-way classification

Consider the following table:

	C_1	C_2	$\ldots$	C_c	
R_1	μ_{11}	μ_{12}	$\ldots$	μ_{1c}	$\mu_{1\cdot}$
R_2	μ_{21}	μ_{22}	$\ldots$	μ_{2c}	$\mu_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
R_r	μ_{r1}	μ_{r2}	$\ldots$	μ_{rc}	$\mu_{r\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\ldots$	$\mu_{\cdot c}$	$\mu_{\cdot\cdot}$

The factors, R and C (for "rows" and "columns" of the table of means), have r and c categories, respectively. The factor categories are denoted R_j and C_k. Within each cell of the design - that is, for each combination of categories $\{R_j, C_k\}$ of the two factors - there is a population cell mean μ_{jk} for the response variable. Extending the dot notation, we have

$$\mu_{j\cdot} \equiv \frac{\sum_{k=1}^{c} \mu_{jk}}{c}$$

is the marginal mean of the response variable in row j.

$$\mu_{\cdot k} \equiv \frac{\sum_{j=1}^{r} \mu_{jk}}{r}$$

is the marginal mean in column k. And

$$\mu_{\cdot\cdot} \equiv \frac{\sum_j \sum_k \mu_{jk}}{r \times c}$$

is the grand mean.

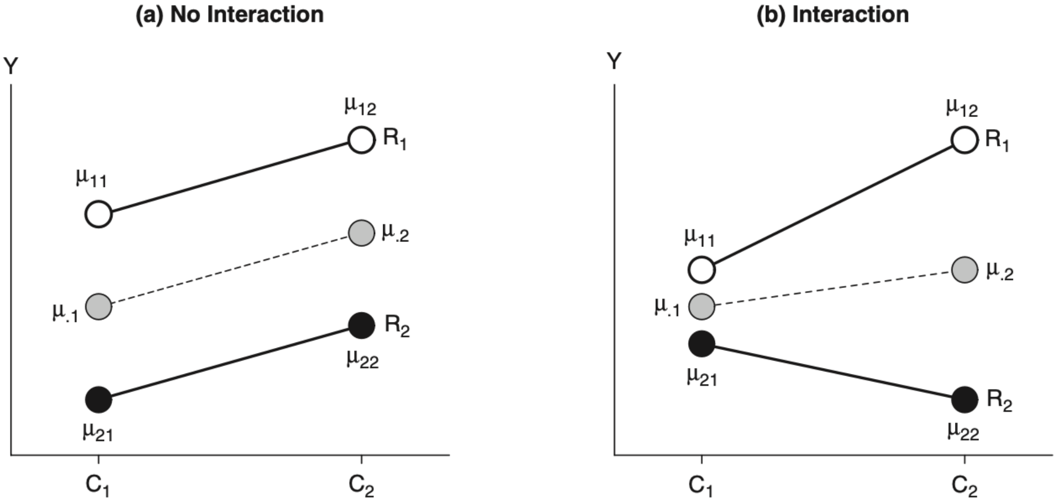

Figure 26.1: Interaction in the two-way classification. In (a), the parallel profiles of means (given by the white and black circles connected by solid lines) indicate that R and C do not interact in affecting Y. The R-effect – that is, the difference between the two profiles – is the same at both C_1 and C_2. Likewise, the C-effect – that is , the rise in the line from C_1 to C_2 – is the same for both profiles. In (b), the R-effect differs at the two categories of C, and the C-effect differs at the two categories of R: R and C interact in affecting Y. In both graphs, the column marginal means $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$ are shown as averages of the cell means in each column (represented by the gray circles connected by broken lines). JF Figure 8.2.

Two-way ANOVA model

The two-way ANOVA model, suitably defined, provides a convenient means for testing the hypotheses concerning interactions and main effects. The model is

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk}$$

where Y_{ijk} is the ith observation in row j, column k of the RC table; μ is the general mean of Y; α_j and β_k are the main-effect parameters; γ_{jk} are interaction effect parameters; and

ϵ_{ijk} are errors satisfying the usual linear-model assumptions (i.e. $\epsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$). By taking expectations, we have

$$\mu_{jk} \equiv E(Y_{ijk}) = \mu + \alpha_j + \beta_k + \gamma_{jk}$$

We have $r \times c$ population cell means with $1 + r + c + r \times c$ model parameters. Similar to one-way ANOVA model, we add in additional constraints to make the model identifiable.

$$\sum_{j=1}^{r} \alpha_j = 0$$
$$\sum_{k=1}^{c} \beta_k = 0$$
$$\sum_{j=1}^{r} \gamma_{jk} = 0 \quad \text{for all } k = 1, \ldots, c$$
$$\sum_{k=1}^{c} \gamma_{jk} = 0 \quad \text{for all } j = 1, \ldots, r$$

The constraints produce the following solution for model parameters in terms of population cell and marginal means (and we add a hat for their estimates using the sample means):

$$\begin{aligned}
\mu &= \mu_{\cdot\cdot} \\
\alpha_j &= \mu_{j\cdot} - \mu_{\cdot\cdot} \\
\beta_k &= \mu_{\cdot k} - \mu_{\cdot\cdot} \\
\gamma_{jk} &= \mu_{jk} - \mu - \alpha_j - \beta_k \\
&= \mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu_{\cdot\cdot}
\end{aligned}$$

Hypotheses with two-way ANOVA

Some interesting hypotheses:

1. Are the cell means all equal? (Equivalent to one-factor ANOVA's "overall F-test")
   $H_0 : \mu_{11} = \mu_{12} = \cdots = \mu_{rc}$ vs. H_a : At least two μ_{ij} differ

2. Are the marginal means for row main effect equal?
   $H_0 : \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{r\cdot}$ vs H_a : At least two $\mu_{j\cdot}$ differ
   which is equivalent as testing for no row main effects H_0 : all $\alpha_j = 0$ (why?)
   Answer: This is because $\alpha_j = \mu_{j\cdot} - \mu_{\cdot\cdot}$ such that all $\alpha_j = 0$ is the equivalent as all marginal means are equal $\mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{r\cdot}$.

3. Are the marginal means for column main effect equal?
   $H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot c}$ vs H_a : At least two $\mu_{\cdot k}$ differ

4. Do the factors interact? In other words, does effect of one factor depend on the other factor? $H_0 : \mu_{ij} = \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot})$ vs H_a : At least one $\mu_{ij} \neq \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot})$
   The null hypothesis is also equivalent as H_0 : all $\gamma_{jk} = 0$.

Testing hypotheses in two-way ANOVA

We follow the notations of JF for incremental sums of squares in ANOVA:

$$\begin{aligned}
\mathbf{SS}(\gamma|\alpha,\beta) &= \mathbf{SS}(\alpha,\beta,\gamma) - \mathbf{SS}(\alpha,\beta)\\
\mathbf{SS}(\alpha|\beta,\gamma) &= \mathbf{SS}(\alpha,\beta,\gamma) - \mathbf{SS}(\beta,\gamma)\\
\mathbf{SS}(\beta|\alpha,\gamma) &= \mathbf{SS}(\alpha,\beta,\gamma) - \mathbf{SS}(\alpha,\gamma)\\
\mathbf{SS}(\alpha|\beta) &= \mathbf{SS}(\alpha,\beta) - \mathbf{SS}(\beta)\\
\mathbf{SS}(\beta|\alpha) &= \mathbf{SS}(\alpha,\beta) - \mathbf{SS}(\alpha)
\end{aligned}$$

where $\mathbf{SS}(\alpha,\beta,\gamma)$ denotes the regression sum of squares for the full model which includes both sets of main effects and the interaction. $\mathbf{SS}(\alpha,\beta)$ denotes the regression sum of squares for the no-interaction model and $\mathbf{SS}(\alpha,\gamma)$ denotes the regression for the model that omits the column main-effect regressors. Note that the last model violates the principle of marginality because it includes the interaction regressors but omits the column main effects. However, it is useful for constructing the incremental sum of squares for testing the column main effects.

Additional readings: Notes on 3 types of Sum of Squares by Dr. Nancy Reid.

We now have the two-way ANOVA table

Table 3: Two-way ANOVA table

Source	Sum of Squares	df	H_0
R	$\mathbf{SS}(\alpha\|\beta,\gamma)$	$r-1$	all $\alpha_j = 0$
	$\mathbf{SS}(\alpha\|\beta)$	$r-1$	all $\alpha_j = 0 \mid$ all $\gamma_{jk} = 0$
C	$\mathbf{SS}(\beta\|\alpha,\gamma)$	$c-1$	all $\beta_k = 0$
	$\mathbf{SS}(\beta\|\alpha)$	$c-1$	all $\beta_k = 0 \mid$ all $\gamma_{jk} = 0$
RC	$\mathbf{SS}(\gamma\|\alpha,\beta)$	(r -1)(c - 1)	all $\beta_k = 0$
Residuals	$\mathbf{TSS} - \mathbf{SS}(\alpha,\beta,\gamma)$	n - rc	
Total	**TSS**	n -1	

where the residual sum of squares

$$RSS = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{jk})^2$$

When test for the hypothesis, use the corresponding SS and df together with the residual SS and df to construct the F-statistic.

$$F = \frac{SS/df}{RSS/df_{residual}}$$

There are two reasonable procedures for testing main-effect hypotheses in two-way ANOVA:

1. Tests based on $\mathbf{SS}(\alpha|\beta,\gamma)$ and $\mathbf{SS}(\beta|\alpha,\gamma)$ ("type III" tests) employ models that violate the principle of marginality, but the tests are valid whether or not interactions are present.
2. Tests based on $\mathbf{SS}(\alpha|\beta)$ and $\mathbf{SS}(\beta|\alpha)$ ("type II" tests) conform to the principle of marginality but are valid only if interactions are absent, in which case they are maximally powerful.

Some more jargon:

- Experimental unit (EU): entity to which experimental treatment is assigned.
  For example, Assign fertilizer treatment to fields. Fields = EU.
- Measurement unit (MU): entity that is measured.
  For example, Measure yields at several subplots within each field. MU: subplot
- Treatment structure: describes how different experimental factors are combined to generate treatments.
  For example, Fertilizers: A, B, C; Irrigation: High, Low.
- Randomization structure: how treatments are assigned to EUs.
- Simplest treatment structure: single experimental factor with multiple levels. Ex. Fertilizers A vs B vs C.
- Simplest randomization structure: Completely randomized design – Experimental treatments assigned to EUs entirely at random.

Example: Honeybee data

Entomologist records energy expended (y) by $N = 27$ honeybees at $a = 3$ temperature (A) levels $(20, 30, 40°C)$ consuming liquids with $b = 3$ levels of sucrose concentration (B) $(20\%, 40\%, 60\%)$ in a balanced, completely randomized crossed 3×3 design.

Temp	Suc	Sample		
20	20	3.1	3.7	4.7
20	40	5.5	6.7	7.3
20	60	7.9	9.2	9.3
30	20	6	6.9	7.5
30	40	11.5	12.9	13.4
30	60	17.5	15.8	14.7
40	20	7.7	8.3	9.5
40	40	15.7	14.3	15.9
40	60	19.1	18.0	19.9

1. What is the experimental unit?
   EU = honeybee.
2. What is the treatment structure?
   Three levels of temperature (A) are combined with each of the three sucrose concentrations (B).
3. Finish the table below

Source	df
A	
B	
$A \times B$	
Residual	
Total	

Answer:

Source	df
A	2
B	2
$A \times B$	4
Residual	18
Total	26

4. Consider the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where $i = 1, 2, \dots, a$, $j = 1, 2, \dots, b$ and $k = 1, 2, \dots, n$ for a balanced design. Deviation:

- total: $y_{ijk} - \bar{y}_{+++}$
- due to level i of factor A: $\bar{y}_{i++} - \bar{y}_{+++}$
- due to level j of factor B: $\bar{y}_{+j+} - \bar{y}_{+++}$
- due to levels i of factor A and j of factor B after subtracting main effects:

$$\bar{y}_{ij+} - \bar{y}_{+++} - (\bar{y}_{i++} - \bar{y}_{+++}) - (\bar{y}_{+j+} - \bar{y}_{+++}) = \bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++}$$

Use the following equations to calculate the Sum of Squares and fill out the ANOVA table.

$$\begin{aligned}
SS[Tot] &= \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{+++})^2 \\
SS[A] &= \sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2 \\
SS[B] &= \sum_i \sum_j \sum_k (\bar{y}_{+j+} - \bar{y}_{+++})^2 \\
SS[AB] &= \sum_i \sum_j \sum_k (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2 \\
SS[E] &= \sum_i \sum_j \sum_k (\bar{y}_{ijk} - \bar{y}_{ij+})^2
\end{aligned}$$

where

$$
\begin{aligned}
\bar{y}_{ij+} &= \frac{1}{n}\sum_k y_{ijk} \\
\bar{y}_{i++} &= \frac{1}{b}\sum_j \bar{y}_{ij+} = \frac{1}{bn}\sum_j\sum_k y_{ijk} \\
\bar{y}_{+j+} &= \frac{1}{a}\sum_i \bar{y}_{ij+} = \frac{1}{an}\sum_i\sum_k y_{ijk} \\
\bar{y}_{+++} &= \frac{1}{a}\sum_i \bar{y}_{i++} = \frac{1}{b}\sum_j \bar{y}_{+j+} \\
&= \frac{1}{abn}\sum_i\sum_j\sum_k y_{ijk}
\end{aligned}
$$

Source	df	Sum of Squares	Mean Square	F
A				
B				
$A \times B$				
Residual				
Total				

Answer:

Source	df	Sum of Squares	Mean Square	F
Temp	2	293.16	146.58	162.00
Suc	2	309.96	154.98	171.28
Temp × Suc	4	27.13	6.78	7.50
Residual	18	16.29	0.90	
Total	26	646.53		

28 Lecture 28: April 5

Last time

- two-way ANOVA

Today

- Announcement: per requested by three students, we will do a third poll on Wednesday for the alternative grading path (the last time).
- Lab session review
- ANCOVA
- Linear contrasts of means

Additional reference

Course notes by Dr. Jason Osborne.

A three-factor example

In a balanced, complete, crossed design, $N = 36$ shrimp were randomized to $abc = 12$ treatment combinations from the factors below:

- A1: Temperature at $25°C$
- A2: Temperature at $35°C$
- B1: Density of shrimp population at 80 shrimp/$40l$
- B2: Density of shrimp population at 160 shrimp/$40l$
- C1: Salinity at 10 units
- C2: Salinity at 25 units
- C3: Salinity at 40 units

The response variable of interest is weight gain Y_{ijkl} after four weeks.

Three-way ANOVA model

$$\begin{aligned} Y_{ijkl} = &\mu + \alpha_i + \beta_j + \gamma_k \\ &+ (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \\ &+ (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl} \end{aligned}$$

$$\begin{aligned} i &= 1, 2 \\ j &= 1, 2 \\ k &= 1, 2, 3 \\ l &= 1, 2, 3 \\ \epsilon_{ijkl} &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned}$$

Many constraints such as (over one dimension):

$$\sum_i \alpha_i = 0$$

$$\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0 \quad \text{for all } i, j$$

$$\sum_i (\alpha\beta\gamma)_{ijk} = \sum_j (\alpha\beta\gamma)_{ijk} = \sum_k (\alpha\beta\gamma)_{ijk} = 0 \quad \text{for all } i, j, k$$

Now, please finish the table below

Source	df
A	
B	
C	
$A \times B$	
$A \times C$	
$B \times C$	
$A \times B \times C$	
Residual	
Total	

Answer:

Source	df
A	1
B	1
C	2
$A \times B$	1
$A \times C$	2
$B \times C$	2
$A \times B \times C$	2
Residual	24
Total	35

The three-way ANOVA model includes parameters for

- Main effects: α_i, β_j and γ_k.
- Two-way interactions between each pair of factors: $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ and $(\beta\gamma)_{jk}$.
- Three-way interaction among all three factors: $(\alpha\beta\gamma)_{ijk}$.

Readings:

1. JF 8.3.1 on parameter estimates and hypothesis testing for three-way ANOVA model.
2. JF 8.3.2 on Higher-order classifications.

Analysis of Covariance

Analysis of covariance (ANCOVA) is a term used to describe linear models that contain both qualitative and quantitative explanatory variables. The method is, therefore, equivalent to dummy-variable regression, discussed in the previous lectures, although the ANCOVA model is parametrized differently from the dummy-regression model.

Covariate is a variable known to affect the response that

1. differs among EUs
2. reflects differences that exist independently of experimental treatment.

A nutrition example

A nutrition scientist conducted an experiment to evaluate the effects of four vitamin supplements on the weight gain of laboratory animals. The experiment was conducted in a completely randomized design with $N = 20$ animals randomized to $a = 4$ supplement groups, each with sample size $n \equiv 5$. The response variable of interest is weight gain, but calorie intake z was measured simultaneously.

Diet	$y(g)$	Diet	y	Diet	y	Diet	y
1	48	2	65	3	79	4	59
1	67	2	49	3	52	4	50
1	78	2	37	3	63	4	59
1	69	2	75	3	65	4	42
1	53	2	63	3	67	4	34
1	$\bar{y}_{1+} = 63$	2	$\bar{y}_{2+} = 57.8$	3	$\bar{y}_{3+} = 65.2$	4	$\bar{y}_{4+} = 48.8$
1	$s_1 = 12.3$	2	$s_2 = 14.9$	3	$s_3 = 9.7$	4	$s_4 = 10.9$

Question: Is there evidence of a vitamin supplement effect?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	3	797.8	265.9	1.823	0.184
Residuals	16	2334.4	145.9		

Conclusion: at $\alpha = 0.05$ level, there is no significant difference between vitamin supplement levels on weight gain.

But calorie intake z was measured simultaneously:

Diet	$y(g)$	z	Diet	y	z	Diet	y	z	Diet	y	z
1	48	350	2	65	400	3	79	510	4	59	530
1	67	440	2	49	450	3	52	410	4	50	520
1	78	440	2	37	370	3	63	470	4	59	520
1	69	510	2	75	530	3	65	470	4	42	510
1	53	470	2	63	420	3	67	480	4	34	430

Question: How and why could these new data be incorporated into analysis?
Answer: ANCOVA can be used to reduce unexplained variation.

ANCOVA model,

$$y_{ij} = \mu + \alpha_i + \beta z_{ij} + \epsilon_{ij}$$

where μ is the reference level, α_i is the main effect of treatment, β is the partial regression coefficient, and $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The model is equivalent as the dummy-variable regression model,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_z z_i + \epsilon_i \quad \text{for } i = 1, \ldots, 20$$

Finish the table below

Source	df
Diet	
Covariate	1
Residual	
Total	

Answer:

Source	df
Diet	3
Covariate	1
Residual	15
Total	19

To test for difference among treatments. The null hypothesis in terms of α_i is
$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_4 = 0$ v.s. $H_a :$ at least one $\alpha_i \neq 0$
And the null hypothesis in terms of β_i is
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ v.s. $H_a :$ at least one $\beta_i \neq 0$

Question: which two models do we compare when testing the above null hypothesis? *Answer:*

- In terms of ANOVA and ANCOVA models, we compare the one-way "ANOVA" model (actually the simple linear regression model) with only the covariate term to the ANCOVA model that has both the covariate and the treatment.
  `aov(y ~ z, data = vitamin.supplement)` vs `aov(y ~ Diet + z, data = vitamin.supplement)`
- In terms of the dummy-variable regression model, we compare the simple linear regression model of regression y on z to the model that includes the dummy-variable for Diet (treatments).
  `lm(y ~ z, data = vitamin.supplement)` vs `lm(y ~ Diet + z, data = vitamin.supplement)`

Linear contrasts of means

With ANOVA (or ANCOVA) models, we do not generally test hypotheses about individual coefficients (but we can do so if we wish). For dummy-coded regressors in one-way ANOVA, a t-test or F-test of $H_0 : \alpha_1 = 0$, for example, is equivalent to testing for the difference in means between the first group and the baseline group, $H_0 : \mu_1 = \mu_m$.

Consider the one-way ANOVA model:

$$Y_{ij} = \mu_i + \epsilon_{ij}, i = 1, 2, \dots, t, \text{ and } j = 1, 2, \dots, n_i$$

with $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

A linear function of the group means of the form

$$\theta = c_1\mu_1 + c_2\mu_2 + \cdots + c_t\mu_t$$

is called a <u>linear combination</u> of the treatment means. And the c_i's are the <u>coefficients</u> of the linear combination. If

$$c_1 + c_2 + \cdots + c_t = \sum_{j=1}^{t} c_j = 0,$$

the linear combination is called a <u>contrast</u>. Contrasts with more than two non-zero coefficients are called <u>complex contrasts</u>.

Let two contrasts θ_1 and θ_2 be given by

$$\theta_1 = c_1\mu_1 + \cdots + c_t\mu_t = \sum_{j=1}^{t} c_j\mu_j$$
$$\theta_2 = d_1\mu_1 + \cdots + d_t\mu_t = \sum_{j=1}^{t} d_j\mu_j,$$

then the two contrasts θ_1 and θ_2 are <u>mutually orthogonal</u> if the products of their coefficients sum to zero:

$$c_1d_1 + \cdots + c_td_t = \sum_{j=1}^{t} c_jd_j = 0$$

θ_i and θ_j are orthogonal $\implies$ $\hat{\theta}_i$ and $\hat{\theta}_j$ are statistically independent.

Types of effects

Consider the following two-way ANOVA model:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \\ & i = 1, 2 = a \text{ and } j = 1, 2 = b \text{ and } k = 1, 2, \dots, 7 = n. \end{aligned}$$

$\epsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Parameter constraints: $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i (\alpha\beta)_{ij} = 0$ for each j and $\sum_j (\alpha\beta)_{ij} = 0$ for each i.

- Factor A: AGE has $a = 2$ levels - A_1 : younger and A_2 : older
- Factor B: GENDER has $b = 2$ levels - B_1 : female and B_2 : male

Three kinds of effects in this 2×2 design:

1. Simple effects are simple contrasts.
   - $\mu(A_1B) = \mu_{12} - \mu_{11}$ - simple effect of gender for young folks.
   - $\mu(AB_1) = \mu_{21} - \mu_{11}$ - simple effect of age for women.
2. Interaction effects are differences of simple effects: $\mu(AB) = \mu(AB_2) - \mu(AB_1) = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})$
   - difference between simple age effects for men and women
   - difference between simple gender effects for old and young folks
   - interaction effect of AGE and GENDER.
3. Main effects are averages or sums of simple effects

$$\mu(A) = \frac{1}{2}(\mu(AB_1) + \mu(AB_2))$$
$$\mu(B) = \frac{1}{2}(\mu(A_1B) + \mu(A_2B))$$

29 Lecture 29: April 7

Last time

- ANCOVA
- Linear contrasts of means

Today

- One last poll on alternative grading path
  - Consider the votes on the three polls (including today's), if we model them as binomial distributed random variables with probability for 'yes' as p_1, p_2 and p_3
  - What is the likelihood function?
  - How do we test the null hypothesis of $H_0 : p_1 = p_2 = p_3$?
  - What do you expect? Why?
- Final exam will be posted on April 30th (per requested by Grace), Due May 11th 11:59pm
- Sampling distribution of linear contrasts
- Multiple comparisons
- Sample size computations for one-way ANOVA
- Lack of fit test

Additional reference

Course notes by Dr. Jason Osborne.

Sampling distribution of linear contrast estimates

For a linear contrast

$$\theta = c_1\mu_1 + \cdots + c_t\mu_t$$

The *best* estimator for a contrast of interest can be obtained by substituting treatment group sample means $\bar{y}_{i+}$ for treatment population means μ_i in the contrast θ:

$$\hat{\theta} = c_1\bar{Y}_{1+} + c_2\bar{Y}_{2+} + \cdots + c_t\bar{Y}_{t+}$$

Example

Recall the binding fraction data that investigate binding fraction for several antibiotics using $n = 20$ bovine serum samples:

Antibiotic	Binding Percentage	Sample mean
Penicillin G	29.6 24.3 28.5 32.0	28.6
Tetracyclin	27.3 32.6 30.8 34.8	31.4
Streptomycin	5.8 6.2 11.0 8.3	7.8
Erythromycin	21.6 17.4 18.3 19	19.1
Chloramphenicol	29.2 32.8 25.0 24.2	27.8

Consider the pairwise contrast comparing penicillin (population) mean to Tetracyclin mean:

$$\theta = \mu_1 - \mu_2 = (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 + (0)\mu_5$$

Obtain a point estimator of θ.
Answers:

$$\begin{aligned}\hat{\theta} &= \hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_{1+} - \bar{Y}_{2+} \\ &= 28.6 - 31.4 = -2.8\end{aligned}$$

Question: How good is this estimate? In other words, how much uncertainty associated with the estimate?

We want to characterize the sampling distribution of $\hat{\theta}$. According to our model setup, Y_{ij} follow normal distributions. $\hat{\theta}$ is a linear function of Y_{ij}, so that $\hat{\theta}$ follows a normal distribution. We want to derive the mean and variance (the two sufficient statistics) to characterize the normal distribution that $\hat{\theta}$ follows:

$$\hat{\theta} \sim \mathcal{N}(\theta, Var(\hat{\theta}))$$

Derive expressions for the mean and the variance:
For the mean, we have

$$\begin{aligned}E(\hat{\theta}) &= E(\hat{\mu}_1) - E(\hat{\mu}_2) = E(\bar{Y}_{1+}) - E(\bar{Y}_{2+}) \\ &= \mu_1 - \mu_2 = \theta\end{aligned}$$

The variance follows

$$\begin{aligned}Var(\hat{\theta}) &= Var(\sum_j c_j \bar{Y}_{j+}) \\ &= \sum_j Var(c_j \bar{Y}_{j+}) \\ &= \sum_j c_j^2 Var(\bar{Y}_{j+}) \\ &= \sum_j \frac{c_j^2}{n_j} \sigma^2\end{aligned}$$

Therefore, the standard error:

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})} = \sqrt{\sigma^2 \sum_{j=1}^{t} \frac{c_j^2}{n_j}}$$

which is estimated by

$$\widehat{SE}(\hat{\theta}) = \sqrt{MS[E] \sum_{j=1}^{t} \frac{c_j^2}{n_j}}$$

To test $H_0 : \theta = \theta_0$ (often 0) versus $H_1 : \theta \neq \theta_0$, use t-test:

$$t = \frac{\hat{\theta} - \theta_0}{\widehat{SE}(\hat{\theta})} \overset{H_0}{\sim} t_{N-t}$$

At level α, the critical value for this test is $t(N-t, \alpha/2)$ and $100(1-\alpha)\%$ confidence interval for a contrast $\theta = \sum c_j \mu_j$ is given by

$$\sum c_j \bar{Y}_{j+} \pm t(N-t, \alpha/2) \sqrt{MS[E] \sum \frac{c_j^2}{n_j}}$$

Multiple Comparisons

Let's first review type I and type II errors.

	H_0 is True	H_0 is False
Don't reject H_0	Probability $1-\alpha$	Probability β
Reject H_0	Probability α	Probability $1-\beta$

- Type I error: rejection of a true null hypothesis (false positive).
- Type II error: failure to reject a false null hypothesis (false negative).
- Type I error rate or significance level (α): the probability of rejecting the null hypothesis given the null hypothesis is true.
- Type II error rate (β): the probability of failure to reject the null hypothesis given the null hypothesis is false. $1-\beta$ gives the power of a test.

Now, let's consider all simple (pairwise) contrasts for the binding fraction data with $t = 5$ antibiotic treatments of the form $\theta = \mu_i - \mu_j$.

- We have $\left(\begin{array}{c} 5 \\ 2 \end{array} \right) = 10$ tests for significance each at level $\alpha = 0.05$

- what is the probability of committing at least one type I error?

$$1 - (1-\alpha)^{10}$$

We need to consider the familywise error rate (fwe) when testing k contrasts:

$$fwe = \Pr(\text{at least one type I error})$$

Methods for simultaneous inference for multiple contrasts include

- Bonferroni
- Scheffé
- Tukey

When the number of comparisons is in the hundreds or thousands (e.g. genome-wide association studies), and FWE control is hopeless, more manageable type I error rate is the False Discovery Rate (FDR):

$$FDR = E(\frac{\text{Falsely rejected null hypotheses}}{\text{Number of rejected null hypotheses}})$$

Bonferroni correction

Suppose interest lies in exactly k contrasts. The Bonferroni adjustment to α controls fwe is

$$\alpha_{bonferroni} = \frac{\alpha}{k}$$

and simultaneous 95% confidence intervals for the k contrasts are given by

$$a_1\bar{Y}_{1+} + \cdots + a_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{a_j^2}{n_j}}$$

$$b_1\bar{Y}_{1+} + \cdots + b_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{b_j^2}{n_j}}$$

$$\cdots$$

$$k_1\bar{Y}_{1+} + \cdots + k_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{k_j^2}{n_j}}$$

where ν denotes *df* for error.

Example: for the binding fraction example, consider only pairwise comparisons with Penicillin:

$$\theta_1 = \mu_1 - \mu_2, \theta_2 = \mu_1 - \mu_3, \theta_3 = \mu_1 - \mu_4, \theta_4 = \mu_1 - \mu_5$$

We have $k = 4, \alpha_{bonferroni} = 0.05/k = 0.0125$ and $t(\frac{\alpha_{bonferroni}}{2}, 15) = 2.84$. Substitution leads to

$$t(\frac{\alpha_{bonferroni}}{2}, 15)\sqrt{MS[E]\left(\frac{1^2}{4} + \frac{(-1)^2}{4} + \frac{0^2}{4} + \cdots + \frac{0^2}{4}\right)}$$
$$=2.84\sqrt{(9.05)\frac{2}{4}} = 6.0$$

so that **simultaneous** 95% confidence intervals for θ_1, θ_2, θ_3 and θ_4 take the form

$$\bar{y}_{1+} - \bar{y}_{i+} \pm 6.0$$

Scheffé

Another method to construct **simultaneous** 95% confidence intervals for **ALL** contrasts, use

$$\sum_{j=1}^{t} c_j \bar{y}_{j+} \pm \sqrt{(t-1)(F^*)MS[E]\sum_{j=1}^{t}\frac{c_j^2}{n_j}}$$

where $F^* = F(\alpha, t-1, N-t)$. For a pairwise comparisons of means, μ_j and μ_k, this yields

$$\bar{y}_{j+} - \bar{y}_{k+} \pm \sqrt{(t-1)(F^*)MS[E](1/n_j + 1/n_k)}$$

Using $\alpha = 0.05$, need to specify

- t (from the design)
- F^* (same critical value as for $H_0 : \alpha_i \equiv 0$).
- $MS[E]$ (from the data)
- $\bar{y}_{j+}$, $\bar{y}_{k+}$
- n_j, n_k (from the data)

For binding fraction data,

$$\sqrt{(t-1)(F^*)MS[E](\frac{1}{n_j} + \frac{1}{n_k})} = \sqrt{(5-1)(3.06)9.05(\frac{1}{4} + \frac{1}{4})} = 7.44$$

If any two sample means differ by more than 7.44, they differ significantly.

Tukey

Tukey's method is better than Scheffé's method when making **all pairwise** comparisons in balanced designs ($n = n_1 = n_2 = \cdots = n_t$). It is conservative, controlling the experimentwise error rate, and has a lower type II error rate in these cases than Scheffé. (It is more powerful.)

For simple contrasts of the form

$$\theta = \mu_j - \mu_k$$

to test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

reject H_0 at level α if

$$|\hat{\theta}| > q(t, N-t, \alpha)\sqrt{\frac{MS[E]}{n}}$$

where $q(t, N-t, \alpha)$ denotes α level studentized range for t means and $N-t$ degrees of freedom, the quantity $q(t, N-t, \alpha)\sqrt{\frac{MS[E]}{n}}$ is referred to as Tukey's honestly significant difference (HSD). The studentized ranges can be calculated using R function qtukey($1-\alpha$, t, $N-t$).

31 Lecture 31: April 12

Last time

- One last poll on alternative grading path
- Sampling distribution of linear contrasts
- Multiple comparisons

Today

- **Last** poll on alternative grading path result: *passed* with 4 + 4 yes, 4 + 1 no (0 by emails...?)
- Sample size computations for one-way ANOVA
- Lack of fit test
- One-way random effect model (JF Chapter 23 + Dr. Osborne's notes)

Additional reference

Course notes by Dr. Jason Osborne.

Sample size computations for one-way ANOVA

Now consider the null hypothesis in a balanced experiment using one-way ANOVA to compare t treatment means and $\alpha = 0.05$:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t = \mu$$

versus the alternative

$$H_a : \mu_i \neq \mu_j \text{ for some } i \neq j$$

Suppose that we intend to use a balanced design. How big does our sample size $n_1 = n_2 = \cdots = n_t = n$ need to be?

The answer depends on lots of things, namely, σ^2 and how many treatment groups t and how much of a difference among the means we hope to be able to detect, and with how big a probability.

Given α, $\mu_1, \dots, \mu_t$ and σ^2, we can choose n to ensure a power of at least β (i.e. type II error rate) using the noncentral F distribution.

Recall that the critical region for the statistic $F = MS[Trt]/MS[E]$ is everything bigger than $F(\alpha, t-1, N-t) = F^*$.

The power of the F-test conducted using $\alpha = 0.05$ to reject H_0 under this alternative is given by

$$1 - \beta = \Pr(MS[Trt]/MS[E] > F^*; H_1 \text{ is true}). \tag{4}$$

Let $\tau_i = \mu_i - \mu$ for each treatment i so that

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

When some H_1 is true and the sample size n is used in each group, it can be shown that the F ratio has the noncentral F distribution with noncentrality parameter

$$\gamma = \sum_{j=1}^{t} n_j \left(\frac{\tau_j}{\sigma}\right)^2 = n \sum_{j=1}^{t} \left(\frac{\tau_j}{\sigma}\right)^2$$

This is the parameterization for the F distribution used in both SAS and R.

One way to obtain an adequate sample size is trial and error. Software packages can be used to get probabilities of the form 4 for various values of n.

Example

Suppose we want to test equal mean binding fractions among antibiotics against the alternative

$$H_1 : \mu_P = \mu + 3, \mu_T = \mu + 3, \mu_S = \mu - 6, \mu_E = \mu, \mu_C = \mu$$

so that

$$\tau_1 = \tau_2 = 3, \tau_3 = -6, \tau_4 = \tau_5 = 0.$$

Assume $\sigma = 3$ (is it arbitrary? any idea of how to guess?) and we need to use $\alpha = 0.05$. The noncentrality parameter is given by

$$\gamma = n[(\frac{3}{3})^2 + (\frac{3}{3})^2 + (\frac{-6}{3})^2]$$

The $\alpha = 0.05$ critical value for H_0 is given by

$$F^* = F(5 - 1, 5n - 5, 0.05).$$

We need the area to the right of F^* for the noncentral F distribution with degrees of freedom 4 and $5(n - 1)$ and noncentrality parameter $\gamma = 6n$ to be greater or equal to the desired power level of $1 - \beta = 0.8$.

We will revisit this example in the lab session on Friday.

Lack-of-fit test

Hiking example: completely randomized experiment involving alpine meadows in the White Mountains of New Hampshire. $N = 20$ lanes of dimension $0.5m \times 1.5m$ randomized to 5 trampling treatments:

i: trt group	x: Number of passes	y_{ij}: Height (cm)			
1	0	20.7	15.9	17.8	17.6
2	25	12.9	13.4	12.7	9.0
3	75	11.8	12.6	11.4	12.1
4	200	7.6	9.5	9.9	9.0
5	500	7.8	9.0	8.5	6.7

Two models for mean plant height:

$$\begin{aligned} \text{SLR model: } & \mu(x) = \beta_0 + \beta_1 x \\ \text{one-factor ANOVA model: } & \mu_{ij} = \mu + \alpha_i \end{aligned}$$

When the t treatments have an interval scale, the SLR model, and all polynomials of degree $p \leqslant t-2$ (why?), are nested in one-factor ANOVA model with t treatment means.

Answer:
For t levels, there are $t-1$ degrees of freedom. A polynomial model of degree p has $p+1$ number of parameters (in other words, takes $p+1$ degrees of freedom).

F-ratio for lack-of-fit test

To test for lack-of-fit of a polynomial (reduced) model of degree p, use extra sum-of-squares F-ratio on $t-1-p$ and $N-t$ df:

$$F = \frac{SS[\text{lack of fit}]/(t-1-p)}{MS[\text{pure error}]}$$

where

$$MS[\text{pure error}] = MS[E]_{full}$$

and

$$\begin{aligned} SS[\text{lack-of-fit}] &= SS[Trt] - SS[Reg]_{poly} \\ &= SS[E]_{poly} - SS[E]_{full} \end{aligned}$$

What is the SS[lack of fit] for the meadows data? In a simple linear ($p=1$) model for the meadows data,

$$SS[\text{lack-of-fit}] = 243.163 - 141.295 = 101.867 \quad \text{on } t-1-p = 3 \text{ df}$$

and the sum of squares for the full model is $SS[E]_{full} = 30.93$ that leads to

$$F = \frac{101.867/3}{30.93/15} \approx \frac{34}{2.1} = 16.5$$

(highly significant since $F(0.01, 3, 15) = 5.42$.) $\Rightarrow$ model misspecified: SLR model suffers from lack of fit.

Next step: either go with the one-factor ANOVA model or specify some other model, such as quadratic.

One-way random effects model

Let's first consider an example.

- Genetics study w/ beef animals. Measure birthweight Y (lbs).
- $t = 5$ sires, each mated to a separate group of $n = 8$ dams.
- $N = 40$, completely randomized.

Sire #	Level	Sample Birthweights								$\bar{y}_{i+}$	s_i
177	1	61	100	56	113	99	103	75	62	83.6	22.6
200	2	75	102	95	103	98	115	98	94	97.5	11.2
201	3	58	60	60	57	57	59	54	100	63.1	15.0
202	4	57	56	67	59	58	121	101	101	77.5	25.9
203	5	59	46	120	115	115	93	105	75	91.0	28.0

Question: Statistical model for these data?
Answer: One-way fixed effects model?

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where τ_i denotes the difference between the mean birthweight of population of offspring from sire i and μ, mean of whole population.

The one-way random effects model

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{\epsilon_{ij}}_{\text{random}} \qquad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n$$

with

- $T_1, T_2, \ldots, T_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2)$
- $\epsilon_{11}, \ldots, \epsilon_{tn} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
- $T_1, T_2, \ldots, T_t$ independent of $\epsilon_{11}, , \ldots, \epsilon_{tn}$

Features

- $T_1, T_2, \ldots$ denote random effects, drawn from some population of interest. That is $T_1, T_2, \ldots$ is a random sample.
- σ_T^2 and σ^2 are called variance components
- conceptually different from one-way fixed effects model

For beef animal genetic study, with $t = 5$ and $n = 8$, the random effects $T_1, T_2, \ldots, T_5$ reflect sire-to-sire variability.

No particular interest in $\tau_1, \tau_2, \ldots, \tau_5$ from the fixed effects model:

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{\tau_i}_{\text{fixed}} + \underbrace{\epsilon_{ij}}_{\text{random}} \qquad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n$$

with

- $\tau_1, \tau_2, \ldots, \tau_t$ unknown model parameters
- $\epsilon_{11}, \ldots, \epsilon_{tn} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Exercise

Using the random effects model, specify

$$E(Y_{ij}) \text{ and } Var(Y_{ij})$$

Recall:

- Two *components* to variability in data: σ^2, σ_T^2
- $T_1, T_2, \ldots, T_5$ a random sample of sire effects
- Sire effects is a population in its own right.
- Model parameters: σ^2, σ_T^2, μ.

Answer:

$$E(Y_{ij}) = E(\mu) + E(T_i) + E(\epsilon_{ij}) = \mu$$
$$Var(Y_{ij}) = Var(\mu + T_i + \epsilon_{ij}) = Var(T_i) + Var(\epsilon_{ij}) = \sigma_T^2 + \sigma^2$$

Sums of squares and mean squares are the same as in one-way fixed effects ANOVA:

$$SS[T] = \sum\sum(\bar{y}_{i+} - \bar{y}_{++})^2$$
$$SS[E] = \sum\sum(y_{ij} - \bar{y}_{i+})^2$$
$$SS[Tot] = \sum\sum(y_{ij} - \bar{y}_{++})^2$$

ANOVA table

The ANOVA table is almost the same, it just has a different expected mean squares column:

Source	SS	df	MS	Expected MS
Treatment	$SS[T]$	$t-1$	$MS[T]$	$\sigma^2 + n\sigma_T^2$
Error	$SS[E]$	$N-t$	$MS[E]$	σ^2
Total	$SS[Tot]$	$N-1$		

Estimating parameters of one-way random effects model

1. **Method of moment (M.o.M.) estimation**: Equate EMS with observed MS and solve. Problem: M.o.M estimation can give $\hat{\sigma}^2 < 0$ (estimates of variances < 0)
2. **Maximum likelihood (ML) / Restricted maximum likelihood (REML)**: Numerical procedures that avoid negative variance estimates.
   - ML: full maximum-likelihood estimation maximizes the likelihood with respect to all of the parameters of the model simultaneously (i.e., both the fixed-effects parameters and the variance components).
   - REML: restricted (or residual) maximum-likelihood estimation integrates the fixed effects out of the likelihood and estimates the variance components; given the resulting estimates of the variance components, estimates of the fixed effects are recovered. REML estimates are the same as M.o.M. estimates with balanced data.

For one-way random-effects model:

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{++} \\
\hat{\sigma}^2 &= MS[E] \\
\hat{\sigma}_T^2 &= \frac{MS[T] - MS[E]}{n}
\end{aligned}$$

For sires data, $\bar{y}_{++} = 82.6$ and

Source	SS	df	MS	Expected MS
Sire	5591	4	1398	$\sigma^2 + 8\sigma_T^2$
Error	16233	35	464	σ^2
Total	21824	39		

Obtain the parameter estimates. *Answers:*

$$\begin{aligned}\hat{\mu} &= 82.6\\ \hat{\sigma}^2 &= 464\\ \hat{\sigma}_T^2 &= \frac{1398 - 464}{8}\\ &= 117\end{aligned}$$

32 Lecture 32: April 14

Last time

- Sample size computations for one-way ANOVA
- Lack of fit test
- One-way random effect model (JF Chapter 23 + Dr. Osborne's notes)

Today

- hypothesis test and confidence intervals for one-way random-effects model
- review of one-way random effects ANOVA model
- nested design

Additional reference

Course notes by Dr. Jason Osborne.
Lecture notes from Lukas Meier on ANOVA using R

Other parameters of interest in random effects models

Coefficient of variation (CV):

$$CV(Y_{ij}) = \frac{\sqrt{Var(Y_{ij})}}{|E(Y_{ij})|} = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{|\mu|}$$

Intraclass correlation coefficient:

$$\rho_I = \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{Var(Y_{ij})\, Var(Y_{ik})}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}$$

- Interpretation: the correlation between two responses receiving the same level of the random factor.
- Bigger values of ρ_I correspond to (bigger/smaller?) random treatment effects.

For sires,

$$\widehat{CV} = \frac{\sqrt{117 + 464}}{82.6} = 0.29$$

$$\hat{\rho}_I = \frac{117}{117 + 464} = 0.20$$

Interpretations:

- The estimated standard deviation of a birthweight, 24.1 is 29% of the estimated mean birthweight, 82.6.
- The estimated correlation between any two calves with the same sire for a male parent, or the estimated *intrasire* correlation coefficient, is 0.20.

Testing a variance component - $H_0 : \sigma_T^2 = 0$

Recall that $\sigma_T^2 = Var(T_i)$, the variance among the population of treatment effects.

$$F = \frac{MS[T]}{MS[E]}$$

reject H_0 at level α if $F > F(\alpha, t-1, N-t)$.

For the sires data,

$$F = \frac{1398}{464} = 3.01 > 2.64 = F(0.05, 4, 35)$$

so H_0 is rejected at $\alpha = 0.05$. (The p-value is 0.0309)

Interval estimation of some model parameters

A 95% confidence interval for μ derived by considering $SE(\bar{Y}_{++})$:

$$\begin{aligned}
\bar{Y}_{++} &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij} \\
&= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n} (\mu + T_i + \epsilon_{ij}) \\
&= \mu + \bar{T}_{+} + \bar{\epsilon}_{++}
\end{aligned}$$

where $\bar{T}_{+} = (T_1 + \cdots + T_t)/t$ and $\bar{\epsilon}_{++} = (\sum\sum \epsilon_{ij})/N$, so that

$$\begin{aligned}
Var(\bar{Y}_{++}) &= Var(\bar{T} + \bar{E}_{++}) \\
&= \frac{\sigma_T^2}{t} + \frac{\sigma^2}{nt} \\
&= \frac{1}{nt}(n\sigma_T^2 + \sigma^2) \\
&= \frac{1}{nt}E(MS[T]).
\end{aligned}$$

If the data are normally distributed, then

$$\frac{\bar{Y}_{++} - \mu}{\sqrt{\frac{MS[T]}{nt}}} \sim t_{t-1}$$

and a 95% confidence interval for μ is given by

$$\bar{Y}_{++} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}$$

For the sires data: $\bar{y}_{++} = 82.6$, $MS[T] = 1398$, $nt = 40$. Critical value $t(0.025, 4) = 2.78$ yields the interval

$$82.6 \pm 2.78(5.91) or (66.1, 99.0).$$

Confidence interval for ρ_I:
A 95% confidence interval for ρ_I can be obtained from the expression

$$\frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}} < \rho_I < \frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}}$$

where $F_{\alpha/2} = F(\alpha/2, t-1, N-t)$ and F_{obs} is the observed F-ratio for treatment effect from the ANOVA table.

For the sires data, $F_{obs} = 3.01$ and $F_{0.025} = 3.179$, $F_{0.975} = 0.119$. The formula gives $(-0.01, -0.75)$.

These formulas arrived at via some distributional results:

- $(t-1)\frac{MS[T]}{\sigma^2 + n\sigma_T^2} \sim \chi^2_{t-1}$
- $(N-t)\frac{MS[E]}{\sigma^2} \sim \chi^2_{N-t}$
- $MS[T]$ and $MS[E]$ are independent
- Ratio of independent χ^2 random variables divided by *df* has an F distribution
- $\left(\frac{MS[T]}{\sigma^2 + n\sigma_T^2}\right) / \left(\frac{MS[E]}{\sigma^2}\right) \sim F_{t-1,N-t}$
  (which explains the F test for $H_0 : \sigma_T^2 = 0$)
- Rearranging the probability statement below

$$1 - \alpha = \Pr\left(F(1-\frac{\alpha}{2}, t-1, N-t) < \frac{\frac{MS[T]}{\sigma^2 + n\sigma_T^2}}{\frac{MS[E]}{\sigma^2}} < F(\frac{\alpha}{2}, t-1, N-t)\right)$$

Confidence interval for variance components:

The estimated residual variance component for the sire data was $\hat{\sigma}^2 = MS[E] = 464$ lbs^2.

A 95% confidence interval for this variance component is given by

$$\left(\frac{(40-5)464}{53.2} < \sigma^2 < \frac{(40-5)464}{20.6}\right)$$

or $(305.2, 789.5)$ lbs^2

This can be derived using the distributional result

$$(N-t)\frac{MS[E]}{\sigma^2} \sim \chi^2_{N-t}$$

setting up the probability statement

$$1 - \alpha = \Pr\left(\chi^2(1-\frac{\alpha}{2}, N-t) < (N-t)\frac{MS[E]}{\sigma^2} < \chi^2(\frac{\alpha}{2}, N-t)\right)$$

Rearranging to get σ^2 in the middle yields the $100(1-\alpha)\%$ confidence interval for σ^2:

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2}}, \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2}}\right).$$

Question: what are the mean and variance of χ^2_{35} distribution? *Answer:* $E(\chi^2_k) = k$ and $Var(\chi^2_k) = 2k$.

Confidence interval for σ^2_T:

The estimated variance component for the random sire effect was $\hat{\sigma}^2_T = 117$.

Q: How can we get a 95% confidence interval for σ^2_T?
A: In a similar fashion, but the confidence level based on Satterthwaite's approximation to the degrees of freedom of the linear combination of MS terms:

$$\left(\frac{\widehat{df}\,\hat{\sigma}^2_T}{\chi^2_{\alpha/2,\widehat{df}}}, \frac{\widehat{df}\,\hat{\sigma}^2_T}{\chi^2_{1-\alpha/2,\widehat{df}}}\right)$$

where

$$\widehat{df} = \frac{(n\hat{\sigma}^2_T)^2}{\frac{MS[T]^2}{t-1} + \frac{MS[E]^2}{N-t}}$$

For the sire data,

$$\widehat{df} = \frac{(8 \times 117)^2}{\frac{1398^2}{4} + \frac{464^2}{35}} = 1.76$$

and

$$\chi^2_{0.975,1.79} = 0.029,\ \chi^2_{0.025,1.76} = 6.87$$

yielding the 95% confidence interval

$$\left(\frac{1.76(117)}{6.87}, \frac{1.76(117)}{0.29}\right)$$

or

$$(30, 7051)$$

Review of one-way random effects ANOVA

The one-way random effects model

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{\epsilon_{ij}}_{\text{random}} \qquad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n$$

with

- $T_1, T_2, \ldots, T_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2)$
- $\epsilon_{11}, \ldots, \epsilon_{tn} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
- $T_1, T_2, \ldots, T_t$ independent of $\epsilon_{11}, , \ldots, \epsilon_{tn}$

Remarks:

- $T_1, T_2, \ldots$ randomly drawn from population of treatment effects.
- Only three parameters: μ, σ^2, and σ_T^2
- Several functions of these parameters of interest
    - Coefficient of variation: $CV(Y) = \frac{\sqrt{\sigma^2+\sigma_T^2}}{\mu}$
    - Intraclass correlation coefficient: $\rho_I = Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_T^2}{\sigma^2+\sigma_T^2}$
- Two observations from same treatment group are **not** independent

Exercise: match up the formulas for confidence intervals below with their targets, ρ_I, σ^2, σ_T^2, μ:

$$\bar{Y}_{++} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}$$

$$(\frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}}, \frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}})$$

$$(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2}}, \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2}})$$

$$(\frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{\alpha/2,\widehat{df}}}, \frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{1-\alpha/2,\widehat{df}}})$$

Modelling factorial effects: fixed, or random?

	Random	Fixed
Levels		
- selected from conceptually ∞ population of collection of levels	X	
- finite number of possible levels		X
Another experiment		
- would use same levels		X
- would involve new levels sampled from same population	X	
Goal		
- estimate variance components	X	
- estimate longrun means		X
Inference		
- for these levels used in this experiment		X
- for the population of levels	X	

Nested design

Factor B is nested in factor A if there is a new set of levels of factor B for every different level of factor A.

To illustrate the concept of nested design, we consider the "Pastes" data set in "lme4" package in R. The strength of a chemical paste product was measured for a total of 60 samples coming from 10 randomly selected delivery batches each containing 3 randomly selected casks. Hence, two samples were taken from each cask. We want to check what part of the variability of strength is due to batch and cask.

Let Y_{ijk} be the strength of the kth sample of cask j in batch i. We can use the model

$$Y_{ijk} = \mu + A_i + B_{j(i)} + \epsilon_{ijk}$$

where A_i is the random effect of batch and $B_{j(i)}$ is the random effect of cask **within** batch. Note the special notation $B_{j(i)}$ emphasizes that cask is nested in batch.

34 Lecture 34: April 19

Last time

- hypothesis test and confidence intervals for one-way random-effects model
- review of one-way random effects ANOVA model

Today

- HW3 deadline extended to Tuesday 04/18 midnight.
- nested design
- Two-factor designs

Additional reference

Course notes by Dr. Jason Osborne.
Lecture notes from Lukas Meier on ANOVA using R

Nested design

Factor B is nested in factor A if there is a new set of levels of factor B for every different level of factor A.

To illustrate the concept of nested design, we consider the "Pastes" data set in "lme4" package in R. The strength of a chemical paste product was measured for a total of 60 samples coming from 10 randomly selected delivery batches each containing 3 randomly selected casks. Hence, two samples were taken from each cask. We want to check what part of the variability of strength is due to batch and cask.

Let Y_{ijk} be the strength of the kth sample of cask j in batch i. We can use the model

$$Y_{ijk} = \mu + A_i + B_{j(i)} + \epsilon_{ijk}$$

where A_i is the random effect of batch and $B_{j(i)}$ is the random effect of cask **within** batch. Note the special notation $B_{j(i)}$ emphasizes that cask is nested in batch.

Two-factor designs with factors that are fixed/random and nested/crossed

There are in total six types of two-factor models with fixed/random effects factors that are either crossed or nested.

1. $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}$	crossed/random
2. $Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$	nested/fixed
3. $Y_{ijk} = \mu + A_i + B_{j(i)} + \epsilon_{ijk}$	nested/random
4. $Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + \epsilon_{ijk}$	crossed/mixed
5. $Y_{ijk} = \mu + \alpha_i + B_{j(i)} + \epsilon_{ijk}$	nested/mixed
6. $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$	crossed/fixed

In the models above,

- GREEK symbols parameterized FIXED, unknown treatment means (except for the error term)
- CAPITAL letters represent RANDOM effects (including the error term)
- for Model 1, $A_i, B_j, (AB)_{ij}$ are all independent
- for Model 2, $\sum \alpha_i = \sum_j \beta_{j(i)} \equiv 0$
- for Model 3, $A_i, B_{j(i)}$ are all independent
- for Model 4, $\sum \alpha_i = 0$ and $B_j, (\alpha B)_{ij}$ are all independent
- for Model 5, $\sum \alpha_i = 0$
- for Model 6, $\sum \alpha_i = \sum \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} \equiv 0$

Now let's consider the following 6 examples from Dr. Osborne's notes.

1. Entomologist records energy expended (y) by $N = 27$ honeybees
   - at three TEMPERATURES (20, 30, 40°C)
   - consuming three levels of SUCROSE (20%, 40%, 60%)

Temp	Suc	Sample		
20	20	3.1	3.7	4.7
20	40	5.5	6.7	7.3
20	60	7.9	9.2	9.3
30	20	6	6.9	7.5
30	40	11.5	12.9	13.4
30	60	17.5	15.8	14.7
40	20	7.7	8.3	9.5
40	40	15.7	14.3	15.9
40	60	19.1	18.0	19.9

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $\qquad\qquad +\epsilon_{ijk}$

Answer:

- First factor: Temperature
- Second factor: Sucrose
- Fixed or random? Fixed (Temp and Sucrose)
- Crossed or nested? Crossed
- Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$

2. Experiment to study effect of drug and method of administration on fasting blood sugar in a random sample of $N = 18$ diabetic patients.

- First factor is drug: brand I tablet, brand II tablet, insulin injection
- Second factor is type of administration (see table)

Drug (i)	Type of Administration (j)	Mean $\bar{y}_{j(i)}$	Variance $s^2_{j(i)}$
($i = 1$) Brand I tablet	($j = 1$) $30mg \times 1$	15.7	6.3
	($j = 2$) $15mg \times 1$	19.7	9.3
($i = 2$) Brand II tablet	($j = 1$) $20mg \times 1$	20	1
	($j = 2$) $10mg \times 1$	17.3	6.3
($i = 3$) Brand I tablet	($j = 1$) before breakfast	28	4
	($j = 2$) before supper	33	9

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $\qquad\qquad +\epsilon_{ijk}$

Answer:

- First factor: Drug
- Second factor: Administration
- Fixed or random? Drug (fixed), Administration (fixed)
- Crossed or nested? Nested
- Model: $Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$

3. An experiment is conducted to determine variability among laboratories (interlaboratory differences) in their assessment of bacterial concentration in milk after pasteurization. Milk w/ various degrees of contamination was tested by randomly drawing four samples of milk from a collection of cartons at various stages of spoilage. Y is colony-forming units/μl. Labs think they are receiving 8 independent samples.

	Sample			
Lab	1	2	3	4
1	2200	3000	210	270
	2200	2900	200	260
2	2600	3600	290	360
	2500	3500	240	380
3	1900	2500	160	230
	2100	2200	200	230
4	2600	2800	330	350
	4300	1800	340	290
5	4000	4800	370	500
	3900	4800	340	480

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $\qquad\qquad$ $+\epsilon_{ijk}$

Answer:

- First factor: Lab
- Second factor: Sample
- Fixed or random? Lab(random), Carton (random)
- Crossed or nested? Crossed
- Model: $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}$

4. An experiment measures Campylobacter counts in $N = 120$ chickens in a processing plant, at four locations, over three days. Means (std) for $n = 10$ chickens sampled at each location tabulated below:

	Sample			
	Before	After	After	After
Day	Washer	Washer	mic. rinse	chill tank
1	70070.00	48310.00	12020.00	11790.00
	(79034.49)	(34166.80)	(3807.24)	(7832.05)
2	75890.00	52020.00	8090.00	8690.00
	(74551.32)	(17686.27)	(4848.01)	(5526.19)
3	95260.00	33170.00	6200.00	8370.00
	(03176.00)	(22259.08)	(5028.81)	(5720.15)

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $+\epsilon_{ijk}$

Answer:

- First factor: Day
- Second factor: Location
- Fixed or random? Day(random), Location (fixed)
- Crossed or nested? Crossed
- Model: $Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + \epsilon_{ijk}$

5. An experiment to assess the variability of a particular acid among plants and among leaves of plants:

Plant i	1			2			3			4		
Leaf j	1	2	3	1	2	3	1	2	3	1	2	3
$k = 1$	11.2	16.5	18.3	14.1	19.0	11.9	15.3	19.5	16.5	7.3	8.9	11.3
$k = 2$	11.6	16.8	18.7	13.8	18.5	12.4	15.9	20.1	17.2	7.8	9.4	10.9
$k = 3$	12.0	16.1	19.0	14.2	18.2	12.0	16.0	19.3	16.9	7.0	9.3	10.5

- First factor:

- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $+\epsilon_{ijk}$

Answer:

- First factor: Plant
- Second factor: Leaf
- Fixed or random? Plant(random), Leaf(random)
- Crossed or nested? Nested
- Model: $Y_{ijk} = \mu + A_i + B_{j(i)} + \epsilon_{ijk}$

6. Plant heights from 20 pots randomized to 10 treatment combinations

Treatment	Dark	Source	Intensity	Pot	Seedling 1	Seedling 2
DD	1	D	D	1	32.94	35.98
DD	1	D	D	2	34.76	32.40
AL	0	A	L	1	30.55	32.64
AL	0	A	L	2	32.37	32.04
AH	0	A	H	1	31.23	31.09
AH	0	A	H	2	30.62	30.42
BL	0	B	L	1	34.41	34.88
BL	0	B	L	2	34.07	33.87
BH	0	B	H	1	35.61	35.00
BH	0	B	H	2	35.65	32.91

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $+\epsilon_{ijk}$

Answer:

- First factor: Treatment
- Second factor: Pot
- Fixed or random? Treatment(fixed), Pot(random)
- Crossed or nested? Nested
- Model: $Y_{ijk} = \mu + \alpha_i + B_{j(i)} + \epsilon_{ijk}$

Tables of expected mean squares (EMS)

When factors A and B are CROSSED, and no sum-to-zero assumptions are made on random effects, expected means associated with sums of squares are given in the table below:

Source	*df*	A, B fixed	A, B random	A fixed B random
A	$a-1$	$\sigma^2 + nb\psi_A^2$	$\sigma^2 + nb\sigma_A^2 + n\sigma_{AB}^2$	$\sigma^2 + nb\psi_A^2 + n\sigma_{\alpha B}^2$
B	$b-1$	$\sigma^2 + na\psi_B^2$	$\sigma^2 + na\sigma_B^2 + n\sigma_{AB}^2$	$\sigma^2 + na\sigma_B^2 + n\sigma_{\alpha B}^2$
AB	$(a-1)(b-1)$	$\sigma^2 + n\psi_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$	$\sigma^2 + n\sigma_{\alpha B}^2$
Error	$ab(n-1)$	σ^2	σ^2	σ^2

When factor B is NESTED in factor A, expected means associated with sums of squares are given in the table below:

Source	*df*	A, B fixed	A, B random	A fixed B random
A	$a-1$	$\sigma^2 + nb\psi_A^2$	$\sigma^2 + nb\sigma_A^2 + n\sigma_{B(A)}^2$	$\sigma^2 + nb\psi_A^2 + n\sigma_{B(A)}^2$
$B(A)$	$a(b-1)$	$\sigma^2 + n\psi_{B(A)}^2$	$\sigma^2 + n\sigma_{B(A)}^2$	$\sigma^2 + n\psi_{B(A)}^2$
Error	$ab(n-1)$	σ^2	σ^2	σ^2

where ψ^2 and σ^2 values are defined below

$$\psi_A^2 = \frac{1}{a-1}\sum_1^a \alpha_i^2 \quad \text{effect size of factor } A$$

$$\psi_B^2 = \frac{1}{b-1}\sum_1^b \beta_i^2 \quad \text{effect size of factor } B$$

$$\psi_{AB}^2 = \frac{1}{(a-1)(b-1)}\sum_{i=1}^a\sum_{j=1}^b (\alpha\beta)_{ij}^2 \quad \text{effect size of interaction}$$

$$\psi_{B(A)}^2 = \frac{1}{a(b-1)}\sum_{i=1}^a\sum_{j=1}^b \beta_{j(i)}^2 \quad \text{effect size of factor } B$$

$$\begin{aligned}
\sigma_A^2 &= \mathrm{Var}(A_i) \quad \text{variance component for factor } A \\
\sigma_B^2 &= \mathrm{Var}(B_i) \quad \text{variance component for factor } B \\
\sigma_{AB}^2 &= \mathrm{Var}((AB)_{ij}) \quad \text{variance component for interaction} \\
\sigma_{B(A)}^2 &= \mathrm{Var}(B_{j(i)}) \quad \text{variance component for factor } B \\
\sigma^2 &= \mathrm{Var}(E_{ijk}) \quad \text{error variance}
\end{aligned}$$

The term *effect size* is often used in power considerations and sometimes involves division by σ^2.

Using expected mean squares to analyze data in mixed-effects models

F-tests and estimating variance components.

1. To test for interaction effect, use $F_{AB} = \frac{MS[AB]}{MS[E]}$
2. To test for main effect of A, use $F_A = \frac{MS[A]}{MS[AB]}$
3. To test for main effect of B, use $F_B = \frac{MS[B]}{MS[AB]}$

Note the departure from fixed-effects analysis, where $MS[E]$ is always used in the denominator.

The estimated variance components satisfy the system of equations by equate (observed) mean squares to their expected values.

For example, for a 2 factor crossed, random-effects model

$$\begin{aligned}
MS[E] &= \hat{\sigma}^2 \\
MS[AB] &= \hat{\sigma}^2 + n\hat{\sigma}_{AB}^2 \\
MS[A] &= \hat{\sigma}^2 + nb\hat{\sigma}_A^2 + n\hat{\sigma}_{AB}^2 \\
MS[B] &= \hat{\sigma}^2 + na\hat{\sigma}_B^2 + n\hat{\sigma}_{AB}^2
\end{aligned}$$

Analysis of variance in nested designs

Consider a two-factor design in which factor B is nested in factor A. Let Y_{ijk} denote the k^{th} response at level j of factor B within level i of factor A. A model:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

for $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b_i$, $k = 1, 2, \ldots, n$ $SS[Tot]$ can be broken down into components reflecting variability due to A, $B(A)$ and variability not due to either factor ($SS[E]$):

$SS[Tot] = SS[A] + SS[B(A)] + SS[E]$

$$\begin{aligned}
SS[Tot] &= \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{+++})^2 \\
SS[A] &= \sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2 \\
SS[B(A)] &= \sum_i \sum_j \sum_k (\bar{y}_{ij+} - \bar{y}_{i++})^2 \\
SS[E] &= \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij+})^2
\end{aligned}$$

The ANOVA table looks like

Source	*df*	Sum of Squares	Mean Square	F
A	$a-1$	$SS[A]$	$MS[A] = \frac{SS[A]}{a-1}$	$F_A = \frac{MS[A]}{MS[E]}$
$B(A)$	$\sum_i (b_i - 1)$	$SS[B(A)]$	$MS[B(A)] = \frac{SS[B(A)]}{\sum_i (b_i - 1)}$	$F_{B(A)} = \frac{MS[B(A)]}{MS[E]}$
Error	$N - \sum b_i$	$SS[E]$	$MS[E] = \frac{SS[E]}{N - \sum b_i}$	
Total	$N-1$	$SS[Tot]$		

And with random-effect, the test statistic becomes

Test for	$A, B(A)$ fixed	A fixed, $B(A)$ random	$A, B(A)$ random
Factor A	$MS[A]/MS[E]$	$MS[A]/MS[B(A)]$	$MS[A]/MS[B(A)]$
Factor $B(A)$	$MS[B(A)]/MS[E]$	$MS[B(A)]/MS[E]$	$MS[B(A)]/MS[E]$

35 Lecture 35: April 21

Last time

- nested design
- Two-factor designs

Today

- HW3 deadline extended to Friday 04/23 midnight.
- Theoretical background of linear models

Additional reference

Course notes by Dr. Hua Zhou
"A Primer on Linear Models" by Dr. John F. Monahan

Linear Models in the matrix form

Recall the matrix form of the linear model

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} \underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon}$$

Simple linear regression model

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Multiple linear regression model

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1,p-1} \\ 1 & x_{21} & \dots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

One-way ANOVA model

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1,n_1} \\ y_{21} \\ \vdots \\ y_{2,n_2} \\ \vdots \\ y_{a,1} \\ \vdots \\ y_{a,n_a} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & & & \\ \mathbf{1}_{n_2} & & \mathbf{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \mathbf{1}_{n_a} & & & & \mathbf{1}_{n_a} \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{a,1} \\ \vdots \\ \epsilon_{a,n_a} \end{bmatrix}$$

Two-way ANOVA model without interaction Model $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$, $i = 1, \ldots, a$ (a levels in factor 1), $j = 1, \ldots, b$ (b levels in factor 2), and $k = 1, \ldots, n_{ij}$ (n_{ij} observations in the (i, j)-th cell). In total we have $n = \sum_{i,j} n_{ij}$ observations and $p = a + b + 1$ parameters. For simplicity, we consider the case without replicates, i.e., $n_{ij} = 1$ and only write out $\mathbf{X}\beta$. Note adding more replicates to each cell does *not* change the rank of $\mathbf{X}$.

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\beta = \begin{bmatrix} \mathbf{1}_b & \mathbf{1}_b & & & \mathbf{I}_b \\ \mathbf{1}_b & & \mathbf{1}_b & & \mathbf{I}_b \\ \vdots & & & \ddots & \vdots \\ \mathbf{1}_b & & & \mathbf{1}_b & \mathbf{I}_b \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \\ \beta_1 \\ \vdots \\ \beta_b \end{bmatrix}$$

Two-way ANOVA with interaction Model $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, $i = 1, \ldots, a$ (a levels in factor 1), $j = 1, \ldots, b$ (b levels in factor 2), and $k = 1, \ldots, n_{ij}$ (n_{ij} observations in the (i, j)-th cell). In total we have $n = \sum_{i,j} n_{ij}$ observations and $p = 1 + a + b + ab$ parameters. For simplicity, we consider the case without replicates, i.e., $n_{ij} = 1$ and only write out $\mathbf{X}\beta$.

Note adding more replicates to each cell does *not* change the rank of $\mathbf{X}$.

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\beta = \begin{bmatrix} \mathbf{1}_b & \mathbf{1}_b & & & & \mathbf{I}_b & \mathbf{I}_b & & & \\ \mathbf{1}_b & & \mathbf{1}_b & & & \mathbf{I}_b & & \mathbf{I}_b & & \\ \vdots & & & \ddots & & \vdots & & & \ddots & \\ \mathbf{1}_b & & & & \mathbf{1}_b & \mathbf{I}_b & & & & \mathbf{I}_b \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \\ \beta_1 \\ \vdots \\ \beta_b \\ \gamma_{11} \\ \vdots \\ \vdots \\ \gamma_{ab} \end{bmatrix}$$

For all the above models, we have 5the most general assumption over the error term, i.e. $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

Mixed effects models For mixed effects models, we generally have

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a design matrix for fixed-effects $\mathbf{b} \in \mathbb{R}^p$
- $\mathbf{Z} \in \mathbb{R}^{n\times q}$ is a design matrix for random-effects $\mathbf{u} \in \mathbb{R}^q$
- The most general assumption is $\mathbf{e} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{R})$, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}_q, \mathbf{G})$, and $\mathbf{e}$ is independent of $\mathbf{u}$.

In many applications, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ and

$$\mathbf{Z}\mathbf{u} = (\mathbf{Z}_1, \dots, \mathbf{Z}_m) \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m \end{pmatrix} = \mathbf{Z}_1\mathbf{u}_1 + \cdots + \mathbf{Z}_m\mathbf{u}_m,$$

where $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}_{q_i}, \sigma_i^2\mathbf{I}_{q_i})$, $\sum_{i=1}^m q_i = q$. $\mathbf{e}$ and $\mathbf{u}_i$, $i = 1, \dots, m$, are jointly independent. Then the covariance of responses $\mathbf{y}$

$$\mathbf{V}(\sigma^2, \sigma_1^2, \dots, \sigma_m^2) = \sigma^2\mathbf{I} + \sum_{i=1}^m \sigma_i^2\mathbf{Z}_i\mathbf{Z}_i^T$$

Linear equations and generalized inverse

For the linear model

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\;\underset{p\times 1}{\mathbf{b}} + \underset{n\times 1}{\mathbf{e}},$$

we obtain the least square estimator by minimize the objective function $Q(\mathbf{b}) = \sum\limits_{i=1}^{n} e_i^2 = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T(\mathbf{Y} - \mathbf{X}\mathbf{b})$. By taking derivative with respect to $\mathbf{b}$ and setting it to zero, we get

$$\left(\frac{\partial Q}{\partial \mathbf{b}}\right)^T = \left(\frac{\partial Q}{\partial b_1}, \frac{\partial Q}{\partial b_2}, \ldots, \frac{\partial Q}{\partial b_p}\right)^T = \left[\frac{\partial\left(\mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\mathbf{b} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right)}{\partial \mathbf{b}}\right]^T = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b}$$

where we used the fact that for constant vector $\mathbf{a} \in \mathbb{R}^{p\times 1}$, constant matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{x} \in \mathbb{R}^{p\times 1}$, we have the two derivatives:

1. $\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^T$
2. $\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$

By setting $\left(\frac{\partial Q}{\partial \mathbf{b}}\right)^T = \mathbf{0}_{p\times 1}$, we get the Normal equations

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$$

Consistency

Assume $\mathbf{A} \in \mathbb{R}^{m\times n}$
Definition: The linear system $\mathbf{A}\mathbf{x} = c$ is consistent if there exists an $\mathbf{x}^*$ such that $\mathbf{A}\mathbf{x}^* = \mathbf{c}$.

- If $\mathbf{A}$ is square and $\mathbf{A}^{-1}$ exists, then $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$.
- Proposition (g1): If $\mathbf{A}\mathbf{x} = \mathbf{c}$ is consistent, and if $\mathbf{G}$ is any matrix such that $\underset{m\times n}{\mathbf{A}}\,\underset{n\times m}{\mathbf{G}}\,\underset{m\times n}{\mathbf{A}} = \underset{m\times n}{\mathbf{A}}$, then $\mathbf{x}^\psi = \mathbf{G}\mathbf{c}$ is a solution to $\mathbf{A}\mathbf{x} = \mathbf{c}$.
  Proof: Let $\mathbf{x}^*$ satisfy $\mathbf{A}\mathbf{x}^* = \mathbf{c}$. Now consider $\mathbf{A}\mathbf{x}^\psi = \mathbf{A}\mathbf{G}\mathbf{c} = \mathbf{A}\mathbf{G}\mathbf{A}\mathbf{x}^* = \mathbf{A}\mathbf{x}^* = \mathbf{c}$
- A matrix $\mathbf{G}$ satisfying $\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}$ is a generalized inverse of $\mathbf{A}$ with notation $\mathbf{A}^-$.
- If $\mathbf{A}$ is square and $\mathbf{A}^{-1}$ exists, then $\mathbf{A}^- = \mathbf{A}^{-1}$ is unique.

The set of all solutions to $\mathbf{A}\mathbf{x} = \mathbf{c}$

Suppose that $\mathbf{A}\mathbf{x} = \mathbf{c}$ is consistent. Then $\mathbf{x}^*$ is a solution to $\mathbf{A}\mathbf{x} = \mathbf{c}$ if and only if $\mathbf{x}^* = \mathbf{A}^-\mathbf{c} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{z}$ for some $\mathbf{z}$ and $\mathbf{A}^-$.
Proof:

1. "If part": By proposition (g1) $\mathbf{x}^+ = \mathbf{A}^-\mathbf{c}$ is a solution. So if $\mathbf{x}^* = \mathbf{A}^-\mathbf{c} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{z}$, then $\mathbf{A}\mathbf{x}^* = \mathbf{A}\mathbf{x}^+ + (\mathbf{A} - \mathbf{A}\mathbf{A}^-\mathbf{A})\mathbf{z} = \mathbf{c}$.
2. "Only if part:" If $\mathbf{A}\mathbf{x}^* = \mathbf{c}$, then $\mathbf{x}^* = \mathbf{A}^-\mathbf{c} + \mathbf{x}^* - \mathbf{A}^-\mathbf{c} = \mathbf{A}^-\mathbf{c} + \mathbf{x}^* - \mathbf{A}^-\mathbf{A}\mathbf{x}^* = \mathbf{A}^-\mathbf{c} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{x}^*$

Moore-Penrose inverse

Assume $\mathbf{A} \in \mathbb{R}^{m\times n}$

- The Moore-Penrose inverse of $\mathbf{A}$ is a matrix $\mathbf{A}^+ \in \mathbb{R}^{n\times m}$ with the following properties

1. $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$ (Generalized inverse, g_1 inverse, or inner pseudo-inverse)
2. $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$. (outer pseudo-inverse. Any g_1 inverse that satisfies this condition is called a g_2 inverse, or reflexive generalized inverse)
3. $\mathbf{A}^+\mathbf{A}$ is symmetric
4. $\mathbf{A}\mathbf{A}^+$ is symmetric

- $\mathbf{A}^+$ exists and is unique for any matrix $\mathbf{A}$.
- In practice, the Moore-Penrose inverse $\mathbf{A}^+$ is easily computed from the singular value decomposition of $\mathbf{A}$.
- $(\mathbf{A}^-)^T$ is a generalized inverse of $\mathbf{A}^T$

General form of the least squares solution

Now we have derived the general form of the least squares solution with generalized inverse.

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} + [\mathbf{I}_p - (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}]\mathbf{q}$$

where $\mathbf{q} \in \mathbb{R}^p$ is arbitrary.

Positive (semi)definite matrix

Assume $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric (i.e. $\mathbf{A} = \mathbf{A}^T$)

- A real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is positive semi-definite (or nonnegative definite, or p.s.d.) if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geqslant 0$ for all $\mathbf{x}$. Notation $\mathbf{A} \succeq_{p.s.d.} \mathbf{0}$
- E.g., the Gramian matrix $\mathbf{X}^T\mathbf{X}$ is p.s.d.
- We write $\mathbf{A} \succeq_{p.s.d.} \mathbf{B}$ means $\mathbf{A} - \mathbf{B} \succeq_{p.s.d.} \mathbf{0}$
- Cholesky decomposition. Each positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ can be factorized as $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ for some lower triangular matrix $\mathbf{L} \in \mathbb{R}^{n\times n}$ with nonnegative diagonal entries.
- $\mathbf{A} \in \mathbb{R}^{n\times n}$ is positive semidefinite if and only if $\mathbf{A}$ is a covariance matrix of a random vector.
  Proof:

  1. "If part": Let $\mathbf{A} = \mathrm{Cov}(\mathbf{x})$ for some random vector $\mathbf{x}$. Then for any constant $\mathbf{c}$ of same length as $\mathbf{x}$, $\mathbf{c}^T\mathbf{A}\mathbf{c} = \mathbf{c}^T\mathrm{Cov}(\mathbf{x})\mathbf{c} = \mathrm{Var}(\mathbf{c}^T\mathbf{x}) \geqslant 0$.
  2. "Only if part": Let $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ be the Cholesky decomposition and $\mathbf{x}$ a vector of iid standard normal. Then $\mathbf{L}\mathbf{x}$ has covariance matrix $\mathbf{L}\mathrm{Cov}(\mathbf{x})\mathbf{L}^T = \mathbf{L}\mathbf{I}_n\mathbf{L}^T = \mathbf{A}$.

37 Lecture 37: April 26

Last time

- Theoretical background of linear model

Today

- Course evaluation (5/17)
- HW3 review
- Theoretical background of linear models cont.
  - Estimability
  - Idempotent matrix
  - Projections
  - Geometry of least squares solution

Additional reference

Course notes by Dr. Hua Zhou
"A Primer on Linear Models" by Dr. John F. Monahan

Estimable function

Assume the linear mean model: $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, $\mathrm{E}(\mathbf{e}) = \mathbf{0}$. One main interest is estimation of the underlying parameter $\mathbf{b}$. Can $\mathbf{b}$ be estimated or what functions of $\mathbf{b}$ can be estimated?

- A parametric function $\boldsymbol{\Lambda}\mathbf{b}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{m\times p}$ is said to be (linearly) estimable if there exists an affinely unbiased estimator of $\boldsymbol{\Lambda}\mathbf{b}$ for all $\mathbf{b} \in \mathbb{R}^p$. That is there exist constants $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{c} \in \mathbb{R}^m$ such that $\mathrm{E}(\mathbf{A}\mathbf{y} + \mathbf{c}) = \boldsymbol{\Lambda}\mathbf{b}$ for all $\mathbf{b}$.
- Theorem: Assuming the linear mean model, the parametric function $\boldsymbol{\Lambda}\mathbf{b}$ is (linearly) estimable if and only if $\mathcal{C}(\boldsymbol{\Lambda}) \subset \mathcal{C}(\mathbf{X}^T)$, or equivalently $\mathcal{N}(\mathbf{X}) \subset \mathcal{N}(\boldsymbol{\Lambda})$.
  "$\boldsymbol{\Lambda}\mathbf{b}$ is estimable $\iff$ the row space of $\boldsymbol{\Lambda}$ is contained in the row space of $\mathbf{X}$ $\iff$ the null space of $\mathbf{X}$ is contained in the null space of $\boldsymbol{\Lambda}$."
  Proof: Let $\mathbf{A}\mathbf{y} + \mathbf{c}$ be an affine estimator of $\boldsymbol{\Lambda}\mathbf{b}$. Unbiasedness requires

  $$\mathrm{E}(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{A}\mathrm{E}(\mathbf{y}) + \mathbf{c} = \mathbf{A}\mathbf{X}\mathbf{b} + \mathbf{c} = \boldsymbol{\Lambda}\mathbf{b}$$

  for all $\mathbf{b} \in \mathbb{R}^p$. Taking the special value $\mathbf{b} = \mathbf{0}$ shows that $\mathbf{c} = \mathbf{0}$. Thus $(\mathbf{A}\mathbf{X} - \boldsymbol{\Lambda})\mathbf{b} = \mathbf{0}$ for all $\mathbf{b}$. Taking special values $\mathbf{b} = \mathbf{e}_i$ shows that columns of the matrix $\mathbf{A}\mathbf{X} - \boldsymbol{\Lambda}$ are all zeros. This means $\mathbf{A}\mathbf{X} = \boldsymbol{\Lambda}$. Therefore, the matrix $\mathbf{A}$ exists if and only if rows of $\boldsymbol{\Lambda}$ are linear combinations of the rows of $\mathbf{X}$, that is, if and only if $\mathcal{C}(\boldsymbol{\Lambda}) \subset \mathcal{C}(\mathbf{X}^T)$
- $\lambda^T\mathbf{b}$ is linearly estimable if and only if $\lambda^T\mathbf{b}$ is a linear combination of the components in $\mu_Y = \mathrm{E}(\mathbf{Y})$

- the 'if' part: $\lambda^T\mathbf{b} = \mathbf{a}^T\mu_Y$ for some $\mathbf{a} \in \mathbb{R}^{n\times 1}$, then by definition, $\lambda^T\mathbf{b}$ is estimable.
- the 'only if' part: if $\lambda^T\mathbf{b}$ is estimable, then $\lambda^T = \mathbf{a}^T\mathbf{X}$ for some $\mathbf{a} \in \mathbb{R}^{n\times 1}$. Then $\lambda^T\mathbf{b} = \mathbf{a}^T\mathbf{X}\mathbf{b} = \mathbf{a}^T\mathrm{E}(\mathbf{Y})$

- Corollary: $\mathbf{Xb}$ is estimable.
  "Expected value of any observation $\mathrm{E}(y_i)$ and their linear combinations are estimable."
- Corollary: If $\mathbf{X}$ has full column rank, then any linear combinations of $\mathbf{b}$ are estimable.
- If $\mathbf{\Lambda b}$ is (linearly) estimable, then its *least squares estimator* $\mathbf{\Lambda}\hat{\mathbf{b}}$ is invariant to the choice of the least squares solution $\hat{\mathbf{b}}$.
  Proof: Let $\hat{\mathbf{b}}_1$, $\hat{\mathbf{b}}_2$ be two least squares solutions. Then $\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2 \in \mathcal{N}(\mathbf{X}^T\mathbf{X}) = \mathcal{N}(\mathbf{X}) \subset \mathcal{N}(\mathbf{\Lambda})$. Hence, $\mathbf{\Lambda}(\hat{\mathbf{b}}_1 - \hat{\mathbf{b}}_2) = \mathbf{0}$, that is $\mathbf{\Lambda}\hat{\mathbf{b}}_1 = \mathbf{\Lambda}\hat{\mathbf{b}}_2$
- The least squares estimator $\mathbf{\Lambda}\hat{\mathbf{b}}$ is a linearly unbiased estimator of $\mathbf{\Lambda b}$. *Proof:* The least squares solution takes the general form

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} + [\mathbf{I}_p - (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}]\mathbf{q}$$

where $\mathbf{q} \in \mathbb{R}^p$ is arbitrary. Thus the least squares estimator

$$\begin{aligned}\mathbf{\Lambda}\hat{\mathbf{b}} &= \mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} + \mathbf{\Lambda}[\mathbf{I}_p - (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}]\mathbf{q} \\ &= \mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}\end{aligned}$$

is a linear function of $\mathbf{y}$. Now

$$\begin{aligned}\mathrm{E}(\mathbf{\Lambda}\hat{\mathbf{b}}) &= \mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathrm{E}(\mathbf{y}) \\ &= \mathbf{\Lambda}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X}\mathbf{b} \\ &= \mathbf{\Lambda b}\end{aligned}$$

since $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-$ is a projection onto $\mathcal{C}(\mathbf{X}^T\mathbf{X}) = \mathcal{C}(\mathbf{X}^T)$ and $\mathcal{C}(\mathbf{\Lambda}^T) \subset \mathcal{C}(\mathbf{X}^T)$. Therefore the least squares estimator is unbiased.

Estimability example: One-way ANOVA model

Consider the following example with one-way ANOVA model.

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad i = 1,2,3,\ j = 1,2$$

In matrix form:

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{12} \\ Y_{22} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \end{bmatrix}$$

Note: replication doesn't help with estimability. What functions of $\lambda^T \mathbf{b}$ are estimable?

Solutions: $\mu_Y = \begin{bmatrix} \mu+\alpha_1 \\ \mu+\alpha_2 \\ \mu+\alpha_3 \\ \mu+\alpha_1 \\ \mu+\alpha_2 \\ \mu+\alpha_3 \end{bmatrix} = \begin{bmatrix} \mu_{Y_1} \\ \mu_{Y_2} \end{bmatrix}.$

$\lambda^T \mathbf{b} = \mathbf{a}^T \mathrm{E}(\mathbf{Y})$ for some $\mathbf{a} \in \mathbb{R}^{6\times 1}$ if an only if $\lambda^T \mathbf{b} = \mathbf{a}^T \mu_{Y_1}$ for some $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$.

Let $\lambda = \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}$. We have

$$\begin{aligned} \lambda_0\mu + \lambda_1\alpha_1 + \lambda_2\alpha_2 + \lambda_3\alpha_3 &= a_1(\mu+\alpha_1) + a_2(\mu+\alpha_2) + a_3(\mu+\alpha_3) \\ &= (a_1+a_2+a_3)\mu + a_1\alpha_1 + a_2\alpha_2 + a_3\alpha_3 \end{aligned}$$

$$\begin{cases} \lambda_0 = a_1+a_2+a_3 \\ \lambda_1 = a_1 \\ \lambda_2 = a_2 \\ \lambda_3 = a_3 \end{cases} \Longrightarrow \begin{cases} a_1 = \lambda_1 \\ a_2 = \lambda_2 \\ a_3 = \lambda_3 \\ \lambda_0 = \lambda_1+\lambda_2+\lambda_3 \end{cases}$$

In other words, $\lambda^T \mathbf{b}$ is linearly estimable if and only if $\lambda_0 = \lambda_1 + \lambda_2 + \lambda_3$.

Idempotent matrix

Assume $\mathbf{A} \in \mathbb{R}^{n\times n}$.

- A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is idempotent if and only if $\mathbf{A}^2 (= \mathbf{AA}) = \mathbf{A}$.
- Any idempotent matrix $\mathbf{A}$ is a generalized inverse of itself.
- The only idempotent matrix of full rank is $\mathbf{I}$.
  Proof. Since $\mathbf{A}$ has full rank, the inverse $\mathbf{A}^{-1}$ exists. Then $\mathbf{A} = \mathbf{A}^{-1}\mathbf{AA} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.Interpretation: all idempotent matrices are singular except for the identity matrix.
- $\mathbf{A}$ is idempotent if and only if $\mathbf{A}^T$ is idempotent if and only if $\mathbf{I}_n - \mathbf{A}$ is idempotent.
- For a general matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the matrices $\mathbf{A}^-\mathbf{A}$ and $\mathbf{AA}^-$ are idempotent and

$$\begin{aligned} \mathrm{rank}(\mathbf{A}) &= \mathrm{rank}(\mathbf{A}^-\mathbf{A}) = \mathrm{rank}(\mathbf{AA}^-) \\ \mathrm{rank}(\mathbf{I}_n - \mathbf{A}^-\mathbf{A}) &= n - \mathrm{rank}(\mathbf{A}) \\ \mathrm{rank}(\mathbf{I}_m - \mathbf{AA}^-) &= m - \mathrm{rank}(\mathbf{A}). \end{aligned}$$

38 Lecture 38: April 28

Last time

- Theoretical background of linear model

Today

- Course evaluation (7/17)
- Typo in HW3_keys
- Theoretical background of linear models cont.
  - Projections
  - Geometry of least squares solution
  - Multivariate normal distribution
  - Independence and Cochran's theorem

Additional reference

Course notes by Dr. Hua Zhou
"A Primer on Linear Models" by Dr. John F. Monahan

Projection

- A matrix $\mathbf{P} \in \mathbb{R}^{m \times n}$ is a projection onto a vector space $\mathcal{V}$ if and only if
  1. $\mathbf{P}$ is idempotent
  2. $\mathbf{P}\mathbf{x} \in \mathcal{V}$ for any $\mathbf{x} \in \mathbb{R}^n$
  3. $\mathbf{P}\mathbf{z} = \mathbf{z}$ for any $\mathbf{z} \in \mathcal{V}$.
- Any idempotent matrix $\mathbf{P}$ is a projection onto its own column space $\mathcal{C}(\mathbf{P})$.
  Proof: Property (1) is free. Property (2) is trivial since $\mathbf{P}\mathbf{x} \in \mathcal{C}(\mathbf{P})$ for any $\mathbf{x}$. For property (3), note $\mathbf{P}\mathbf{P} = \mathbf{P}$ says $\mathbf{P}\mathbf{p}_i = \mathbf{p}_i$ for each column $\mathbf{p}_i$ of $\mathbf{P}$. Therefore, $\mathbf{P}\mathbf{z} = \mathbf{z}$ for any $\mathbf{z} \in \mathcal{C}(\mathbf{P})$.
- $\mathbf{A}\mathbf{A}^-$ is a projection onto the column space $\mathcal{C}(\mathbf{A})$.
  Proof:
  1. Idempotent: $\mathbf{A}\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}\mathbf{A}^-$ by definition of generalized inverse.
  2. $\mathbf{A}\mathbf{A}^-\mathbf{v} = \mathbf{A}(\mathbf{A}^-\mathbf{v}) \in \mathcal{C}(\mathbf{A})$
  3. Let $\mathbf{z} \in \mathcal{C}(\mathbf{A})$, then $\mathbf{z} = \mathbf{A}\mathbf{c}$ for some $\mathbf{c}$. Therefore, $\mathbf{A}\mathbf{A}^-\mathbf{z} = \mathbf{A}\mathbf{A}^-\mathbf{A}\mathbf{c} = \mathbf{A}\mathbf{c} = \mathbf{z}$

- Start with $\mathbf{P_X X} = \mathbf{X}$, we have $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X} = \mathbf{X}$. Therefore, $(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is a generalized inverse of $\mathbf{X}$ which is sometimes called the least-squares inverse. And $\mathbf{P_X}$ is a projection onto $\mathcal{C}(\mathbf{X})$.

- The projection matrix
$$\underset{n\times n}{\mathbf{P_X}} = \underset{n\times p}{\mathbf{X}}\underset{p\times p}{(\mathbf{X}^T\mathbf{X})^-}\underset{p\times n}{\mathbf{X}^T}$$
is unique.
Proof: Let $\mathbf{G}_1$ and $\mathbf{G}_2$ be two generalized inverse of $\mathbf{X}^T\mathbf{X}$. Define $\mathbf{P}_{\mathbf{X},1} = \mathbf{X}\mathbf{G}_1\mathbf{X}^T$ and $\mathbf{P}_{\mathbf{X},2} = \mathbf{X}\mathbf{G}_2\mathbf{X}^T$.
$$\mathbf{X}^T\mathbf{X}\mathbf{G}_1\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}\mathbf{G}_2\mathbf{X}^T\mathbf{X}$$
By the proposition below, we have
$$\mathbf{X}\mathbf{G}_1\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{G}_2\mathbf{X}^T\mathbf{X}$$
and taking transpose on both sides, we get $\mathbf{X}^T\mathbf{X}(\mathbf{X}\mathbf{G}_1)^T = \mathbf{X}^T\mathbf{X}(\mathbf{X}\mathbf{G}_2)^T$. Applying the proposition below again, we get
$$\mathbf{X}(\mathbf{X}\mathbf{G}_1)^T = \mathbf{X}(\mathbf{X}\mathbf{G}_2)^T$$
which is
$$\mathbf{X}\mathbf{G}_1\mathbf{X}^T = \mathbf{X}\mathbf{G}_2\mathbf{X}^T.$$
We showed that $\mathbf{P}_{\mathbf{X},1} = \mathbf{P}_{\mathbf{X},2}$ is unique.

- Proposition: Let $\mathbf{X}, \mathbf{A}, \mathbf{B}$ be matrices, then $\mathbf{X}^T\mathbf{X}\mathbf{A} = \mathbf{X}^T\mathbf{X}\mathbf{B}$ if and only if $\mathbf{X}\mathbf{A} = \mathbf{X}\mathbf{B}$.
Proof:
  - 'if' part: $\mathbf{XA} = \mathbf{XB} \implies \mathbf{X}^T\mathbf{XA} = \mathbf{X}^T\mathbf{XB}$
  - 'only if' part: if $\mathbf{X}^T\mathbf{XA} = \mathbf{X}^T\mathbf{XB}$, then $\mathbf{X}^T\mathbf{XA} - \mathbf{X}^T\mathbf{XB} = \mathbf{0} \implies \mathbf{X}^T\mathbf{X}(\mathbf{A}-\mathbf{B}) = \mathbf{0} \implies (\mathbf{A}-\mathbf{B})^T\mathbf{X}^T\mathbf{X}(\mathbf{A}-\mathbf{B}) = \mathbf{0} \implies (\mathbf{XA}-\mathbf{XB})^T(\mathbf{XA}-\mathbf{XB}) = \mathbf{0} \implies \mathbf{XA}-\mathbf{XB} = \mathbf{0}$

- $\underset{n\times n}{\mathbf{P_X}}\ \underset{n\times p}{\mathbf{X}} = \underset{n\times p}{\mathbf{X}}$
Proof:
$$\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}\mathbf{I}$$
with the above proposition, we have
$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X} = \mathbf{X}$$
which is $\mathbf{P_X X} = \mathbf{X}$.

- Predicted values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}_{ls}$ are invariant to choice of solution to the normal equation, where
$$\hat{\mathbf{b}}_{ls} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y}$$
is not necessarily unique.
Proof:

1. $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}_{ls} = \mathbf{P_X}\mathbf{Y}$ and apply the uniqueness of $\mathbf{P_X}$
2. Given $\hat{\mathbf{b}}_1$ and $\hat{\mathbf{b}}_2$ are two solutions to the Nomral Equations, then

$$\hat{\mathbf{b}}_2 = \hat{\mathbf{b}}_1 + \left\{\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{X}\right\}\mathbf{z}$$

for some vector $\mathbf{z}$. Then $\mathbf{X}\hat{\mathbf{b}}_2 = \mathbf{X}\hat{\mathbf{b}}_1 + \mathbf{X}\left\{\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{X}\right\}\mathbf{z} = \mathbf{X}\hat{\mathbf{b}}_1$

Recall, multicollinearity doesn't affect prediction.

Geometry of least squares

- $\mathbf{P}^2_{\mathbf{X}} = \mathbf{P_X}$ and $\hat{\mathbf{Y}} = \mathbf{P_X}\mathbf{Y}$ is unique.
- Recall the column space of $\mathbf{X}$ is $\mathcal{C}(\mathbf{X}) = \left\{ \underset{n\times 1}{\mathbf{y}} : \mathbf{y} = \mathbf{X} \underset{p\times 1}{\mathbf{b}} \text{ for some } \mathbf{b} \right\}$
- The vector in $\mathcal{C}(\mathbf{X})$ that is closest in terms of squared norm (L_2 norm: $||\mathbf{a} - \mathbf{b}||_2 = \sqrt{(\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b})}$) to $\mathbf{Y}$ is given by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}_{ls} = \mathbf{P_X}\mathbf{Y}$.
  Proof: $\hat{\mathbf{b}}_{ls}$ minimizes $||\mathbf{Y} - \mathbf{X}\mathbf{b}||^2$ over all $\mathbf{b} \in \mathbb{R}^p$.
- $\hat{\mathbf{Y}} \in \mathcal{C}(\mathbf{X})$
- $\underset{n\times 1}{\hat{\mathbf{e}}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P_X})\mathbf{Y} \in \mathcal{N}(\mathbf{X}^T)$ where $\mathcal{N}(\mathbf{X}^T) = \left\{ \underset{n\times 1}{\mathbf{v}} : \mathbf{X}^T\mathbf{v} = \mathbf{0} \right\}$ is the null space of $\mathbf{X}^T$.
  Proof: For any $\mathbf{y} \in \mathcal{C}(\mathbf{X})$, such that $\mathbf{y} = \mathbf{X}\mathbf{b}$ for some $\mathbf{b}$.

$$\begin{aligned}\hat{\mathbf{e}}^T\mathbf{y} &= \mathbf{Y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{X}\mathbf{b} \\ &= \mathbf{Y}^T(\mathbf{X} - \mathbf{P_X}\mathbf{X})\mathbf{b} \\ &= 0\end{aligned}$$

  Therefore $\hat{\mathbf{e}}$ is orthogonal to every vector in $\mathcal{C}(\mathbf{X})$.

Normal distribution in scaler case

- A random variable Z has a standard normal distribution, denoted $Z \sim \mathcal{N}(0,1)$, if

$$F_Z(t) = \Pr(Z \leqslant t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz,$$

  or equivalently Z has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

  or equivalently, Z has moment generating function (mgf)

$$m_Z(t) = \mathrm{E}(e^{tZ}) = e^{t^2/2}, \quad -\infty < z < \infty$$

- Non-standard normal random variable
  - Definition 1: A random variable X has normal distribution with mean μ and variance σ^2, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if

    $$X = \mu + \sigma Z$$

    where $Z \sim \mathcal{N}(0, 1)$
  - Definition 2: $X \sim \mathcal{N}(\mu, \sigma^2)$ if

    $$m_X(t) = \mathrm{E}(e^{tX}) = e^{t\mu + \sigma^2 t^2/2}, \quad -\infty < t < \infty$$

  - In both definitions, $\sigma^2 = 0$ is allowed. If $\sigma^2 > 0$, it has a density

    $$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

Multivariate normal distribution

- The standard multivariate normal is a vector of independent standard normals, denoted $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. The joint density is

  $$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}} e^{-\sum_{i=1}^p z_i^2/2}.$$

  The mgf is

  $$m_{\mathbf{Z}}(\mathbf{t}) = \prod_{i=1}^p m_{Z_i}(t_1) = \prod_{i=1}^p e^{t_i^2/2} = e^{\mathbf{t}^T\mathbf{t}/2}.$$

- Consider the affine transformation $\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. $\mathbf{X}$ has mean and variance

  $$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}, \quad \mathrm{Var}(\mathbf{X}) = \mathbf{A}\mathbf{A}^T$$

  and the moment generating function is

  $$m_{\mathbf{X}}(\mathbf{t}) = \mathrm{E}(e^{\mathbf{t}^T(\boldsymbol{\mu}+\mathbf{A}\mathbf{Z})}) = e^{\mathbf{t}^T\boldsymbol{\mu}}\mathrm{E}(e^{\mathbf{t}^T\mathbf{A}\mathbf{Z}}) = e^{\mathbf{t}^T\boldsymbol{\mu}+\mathbf{t}^T\mathbf{A}\mathbf{A}^T\mathbf{t}/2}.$$

- $\mathbf{X} \in \mathbb{R}^p$ has a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\mathbf{V} \in \mathbb{R}^{p\times p}, \mathbf{V} \succeq_{p.s.d.} \mathbf{0}$, denoted $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, if its mgf takes the form

  $$m_{\mathbf{X}}(\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu}+\mathbf{t}^T\mathbf{V}^T\mathbf{t}/2}, \quad \mathbf{t} \in \mathbb{R}^p$$

- if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{V}$ is non-singular, then
  - $\mathbf{V} = \mathbf{A}\mathbf{A}^T$ for some non-singular $\mathbf{A}$
  - $\mathbf{A}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$

∗ The density of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{V}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}.$$

- (Any affine transform of normal is normal) If $\mathbf{X} \in \mathbb{R}^p, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$, where $\mathbf{a} \in \mathbb{R}^q$ and $\mathbf{B} \in \mathbb{R}^{q \times p}$, then $\mathbf{Y} \sim \mathcal{N}(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{V}\mathbf{B}^T)$.
- (Marginal of normal is normal) If $\mathbf{X} \in \mathbb{R}^p, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, then any subvector of $\mathbf{X}$ is normal too.
- A convenient fact about normal random variables/vectors is that zero correlation/covariance implies independence.
  If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and is partitioned as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_m \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \dots & \mathbf{V}_{1m} \\ \vdots & & \vdots \\ \mathbf{V}_{m1} & \dots & \mathbf{V}_{mm} \end{bmatrix}$$

  then $\mathbf{X}_1, \dots, \mathbf{X}_m$ are jointly independent if and only if $\mathbf{V}_{ij} = \mathbf{0}$ for all $i \neq j$.
  Proof: If $\mathbf{X}_1, \dots, \mathbf{X}_m$ are jointly independent, then $\mathbf{V}_{ij} = \mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_j) = \mathrm{E}(\mathbf{X}_i - \boldsymbol{\mu}_i)(\mathbf{X}_j - \boldsymbol{\mu}_j)^T = \mathrm{E}(\mathbf{X}_i - \boldsymbol{\mu}_i)\mathrm{E}(\mathbf{X}_j - \boldsymbol{\mu}_j)^T = \mathbf{0}_{p_i}\mathbf{0}_{p_j}^T = \mathbf{0}_{p_i \times p_j}$.
  Conversely, if $\mathbf{V}_{ij} = \mathbf{0}$ for all $i \neq j$, then the mgf of $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_m)^T$ is

$$\begin{aligned} m_{\mathbf{X}}(\mathbf{t}) &= e^{\mathbf{t}^T\boldsymbol{\mu} + \mathbf{t}^T\mathbf{V}\mathbf{t}/2} \\ &= e^{\sum_{i=1}^m \mathbf{t_i}^T\boldsymbol{\mu_i} + \sum_{i=1}^m \mathbf{t_i}^T\mathbf{V_{ii}}\mathbf{t_i}/2} \\ &= m_{\mathbf{X}_1}(\mathbf{t}_1) \dots m_{\mathbf{X}_m}(\mathbf{t}_m) \end{aligned}$$

  Therefore $\mathbf{X}_1, \dots, \mathbf{X}_m$ are jointly independent.

Independence and Cochran's theorem

- (Independence between two linear forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{Y}_1 = \mathbf{a}_1 + \mathbf{B}_1\mathbf{X}$ and $\mathbf{Y}_2 = \mathbf{a}_2 + \mathbf{B}_2\mathbf{X}$. Then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if $\mathbf{B}_1\mathbf{V}\mathbf{B}_2^T = 0$.
  Proof: Note $\mathrm{Cov}(\mathbf{Y}_1, \mathbf{Y}_2) = \mathbf{B}_1\mathrm{Cov}(\mathbf{X})\mathbf{B}_2^T = \mathbf{B}_1\mathbf{V}\mathbf{B}_2^T$.
- Consider the normal linear model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}_n)$
  - Using $\mathbf{A} = (1/\sigma^2)(\mathbf{I} - \mathbf{P_X})$, we have

$$SSE/\sigma^2 = ||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2 = \mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2_{n-r},$$

    where $r = \mathrm{rank}(\mathbf{X})$. Note the noncentrality parameter is

$$\phi = \frac{1}{2}(\mathbf{X}\mathbf{b})^T(1/\sigma^2)(\mathbf{I} - \mathbf{P_X})(\mathbf{X}\mathbf{b}) = 0 \text{ for all } \mathbf{b}.$$

  - Using $\mathbf{A} = (1/\sigma^2)\mathbf{P}_{\mathbf{X}}$, we have

    $$SSR/\sigma^2 = ||\hat{\mathbf{y}}||_2^2/\sigma^2 = \mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi_r^2(\phi),$$

    with the noncentrality parameter

    $$\phi = \frac{1}{2}(\mathbf{X}\mathbf{b})^T(1/\sigma^2)\mathbf{P}_{\mathbf{X}}(\mathbf{X}\mathbf{b}) = \frac{1}{2\sigma^2}||\mathbf{X}\mathbf{b}||_2^2.$$

  - The joint distribution of $\hat{\mathbf{y}}$ and $\hat{\boldsymbol{\epsilon}}$ is

    $$\left[\begin{array}{c} \hat{\mathbf{y}} \\ \hat{\boldsymbol{\epsilon}} \end{array}\right] = \left[\begin{array}{c} \mathbf{P}_{\mathbf{X}} \\ \mathbf{I}_n - \mathbf{P}_{\mathbf{X}} \end{array}\right]\mathbf{y} \sim \mathcal{N}\left(\left[\begin{array}{c} \mathbf{X}\mathbf{b} \\ \mathbf{0}_n \end{array}\right], \left[\begin{array}{cc} \sigma^2\mathbf{P}_{\mathbf{X}} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) \end{array}\right]\right).$$

    So $\hat{\mathbf{y}}$ is independent of $\boldsymbol{\epsilon}$. Thus $||\hat{\mathbf{y}}||_2^2/\sigma^2$ is independent of $||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2$ and

    $$F = \frac{||\hat{\mathbf{y}}||_2^2/\sigma^2/r}{||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2/(n-r)} \sim F_{r,n-r}(\frac{1}{2\sigma^2}||\mathbf{X}\mathbf{b}||_{\mathbf{2}}^{\mathbf{2}}).$$

- (Independence between linear and quadratic forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$. Then $\mathbf{A}$ is symmetric with rank s. If $\mathbf{B}\mathbf{V}\mathbf{A} = \mathbf{0}$, then $\mathbf{B}\mathbf{X}$ and $\mathbf{X}^T\mathbf{A}\mathbf{X}$ are independent.
  Proof: By eigen-decomposition, $\mathbf{A} = \mathbf{Q}_1\boldsymbol{\Lambda}_1\mathbf{Q}_1^T$, where $\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I}_s$ and $\boldsymbol{\Lambda}_1 \in \mathbb{R}^{s\times s}$ is non-singular. Consider the joint distribution

  $$\left[\begin{array}{c} \mathbf{B}\mathbf{X} \\ \mathbf{Q}_1^T\mathbf{X} \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \mathbf{B}\boldsymbol{\mu} \\ \mathbf{Q}_1^T\boldsymbol{\mu} \end{array}\right], \left[\begin{array}{cc} \mathbf{B}\mathbf{V}\mathbf{B}^T & \mathbf{B}\mathbf{V}\mathbf{Q}_1 \\ \mathbf{Q}_1^T\mathbf{V}\mathbf{B}^T & \mathbf{Q}_1^T\mathbf{V}\mathbf{Q}_1 \end{array}\right]\right)$$

  By hypothesis

  $$\mathbf{B}\mathbf{V}\mathbf{A} = \mathbf{B}\mathbf{V}\mathbf{Q}_1\boldsymbol{\Lambda}_1\mathbf{Q}_1^T = \mathbf{0}$$

  Post-multiplying both sides by $\mathbf{Q}_1\boldsymbol{\Lambda}_1^{-1}$ gives $\mathbf{B}\mathbf{V}\mathbf{Q}_1 = \mathbf{0}$, which implies $\mathbf{B}\mathbf{X}$ is independent of both $\mathbf{Q}_1^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Q}_1\boldsymbol{\Lambda}_1\mathbf{Q}_1\mathbf{X} = \mathbf{X}^T\mathbf{A}\mathbf{X}$.

- (Independence between two quadratic forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{A}$ be symmetric with rank r, and $\mathbf{B}$ be symmetric with rank s. If $\mathbf{B}\mathbf{V}\mathbf{A} = \mathbf{0}$, then $\mathbf{X}^T\mathbf{A}\mathbf{X}$ and $\mathbf{X}^T\mathbf{B}\mathbf{X}$ are independent.
  Proof: Again, by eigen-decomposition.

  $$\begin{array}{ll} \mathbf{A} = \mathbf{Q}_1\boldsymbol{\Lambda}_1^{-1}\mathbf{Q}_1, & \text{where } \mathbf{Q}_1 \in \mathbb{R}^{p\times r}, \boldsymbol{\Lambda}_1 \in \mathbb{R}^{r\times r} nonsingular \\ \mathbf{B} = \mathbf{Q}_2\boldsymbol{\Lambda}_2^{-1}\mathbf{Q}_2, & \text{where } \mathbf{Q}_2 \in \mathbb{R}^{p\times s}, \boldsymbol{\Lambda}_2 \in \mathbb{R}^{s\times s} nonsingular \end{array}$$

  Now consider the joint distribution

  $$\left[\begin{array}{c} \mathbf{Q}_1^T\mathbf{X} \\ \mathbf{Q}_2^T\mathbf{X} \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \mathbf{Q}_1^T\boldsymbol{\mu} \\ \mathbf{Q}_2^T\boldsymbol{\mu} \end{array}\right], \left[\begin{array}{cc} \mathbf{Q}_1^T\mathbf{V}\mathbf{Q}_2 & \mathbf{Q}_1^T\mathbf{V}\mathbf{Q}_2 \\ \mathbf{Q}_2^T\mathbf{V}\mathbf{Q}_1 & \mathbf{Q}_2^T\mathbf{V}\mathbf{Q}_2 \end{array}\right]\right)$$

  By hypothesis

  $$\mathbf{B}\mathbf{V}\mathbf{A} = \mathbf{Q}_2\boldsymbol{\Lambda}_2\mathbf{Q}_2^T\mathbf{V}\mathbf{Q}_1\boldsymbol{\Lambda}_1\mathbf{Q}_1^T = \mathbf{0}$$

Pre-multiplying both sides by $\mathbf{\Lambda}_2^{-1}\mathbf{Q}_2^T$ and then post-multiplying both sides by $\mathbf{Q}_1\mathbf{\Lambda}_1^{-1}$ gives

$$\mathbf{Q}_2^T\mathbf{V}\mathbf{Q}_1 = \mathbf{0}.$$

Therefore $\mathbf{Q}_1^T\mathbf{X}$ is independent of $\mathbf{Q}_2^T\mathbf{X}$, which implies $\mathbf{X}^T\mathbf{A}\mathbf{X} = \mathbf{X}^T\mathbf{Q}_1\mathbf{\Lambda}_1^{-1}\mathbf{Q}_1$ is independent of $\mathbf{X}^T\mathbf{B}\mathbf{X} = \mathbf{X}^T\mathbf{Q}_2\mathbf{\Lambda}_2^{-1}\mathbf{Q}_2$.

- (Cochran's theorem) Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ and $\mathbf{A}_i$, $i = 1, \ldots, k$ be symmetric idempotent matrix with rank s_i. If $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n$, then $(1/\sigma^2)\mathbf{y}^T\mathbf{A}_i\mathbf{y}$ are independent $\chi^2_{s_i}(\phi_i)$, with $\phi_i = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}$ and $\sum_{i=1}^k s_i = n$.
Proof: Since $\mathbf{A}_i$ is symmetric and idempotent with rank s, $\mathbf{A}_i = \mathbf{Q}_i\mathbf{Q}_i^T$ with $\mathbf{Q}_i \in \mathbb{R}^{n\times s}$ and $\mathbf{Q}_i^T\mathbf{Q}_i = \mathbf{I}_{s_i}$. Define $\mathbf{Q} = (\mathbf{Q}_1, \ldots, \mathbf{Q}_k) \in \mathbb{R}^{n\times\sum_{i=1}^k s_i}$. Note

$$\begin{aligned}\mathbf{Q}^T\mathbf{Q} &= \mathbf{I}_{\sum_{i=1}^k s_i}\\ \mathbf{Q}\mathbf{Q}^T &= \sum_{i=1}^k \mathbf{Q}_i\mathbf{Q}_i^T = \sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n\end{aligned}$$

Now

$$\mathbf{Q}^T\mathbf{y} = \begin{bmatrix}\mathbf{Q}_1^T\mathbf{y}\\ \vdots\\ \mathbf{Q}_k^T\mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{Q}_1^T\boldsymbol{\mu}\\ \vdots\\ \mathbf{Q}_k^T\boldsymbol{\mu}\end{bmatrix}, \sigma^2\mathbf{I}_n\right)$$

implying that $\mathbf{Q}_i^T\mathbf{y} \sim \mathcal{N}(\mathbf{Q}_i^T\boldsymbol{\mu}, \sigma^2\mathbf{I}_{s_i})$ are jointly independent. Therefore $(1/\sigma^2)\mathbf{y}^T\mathbf{A}_i\mathbf{y} = (1/\sigma^2)||\mathbf{Q}_i^T\mathbf{y}||_2^2 \sim \chi^2_{s_i}(\frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu})$ are jointly independent.

- Application to the one-way ANOVA: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$. We have the classical ANOVA table

Source	df	Projection	SS	Noncentrality
Mean	1	$\mathbf{P_1}$	$SSM = n\bar{y}^2$	$\frac{1}{2\sigma^2}n(\mu+\bar{\alpha})^2$
Group	$a-1$	$\mathbf{P_X} - \mathbf{P_1}$	$SSA = \sum_{i=1}^a n_i\bar{y}_i^2 - n\bar{y}^2$	$\frac{1}{2\sigma^2}\sum_{i=1}^a n_i(\alpha_i - \bar{\alpha})^2$
Error	$n-a$	$\mathbf{I} - \mathbf{P_X}$	$SSE = \sum_{i=1}^a\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	0
Total	n	$\mathbf{I}$	$SST = \sum_i\sum_j y_{ij}^2$	$\frac{1}{\sigma^2}\sum_{i=1}^a n_i(\mu+\alpha_i)^2$

39 Lecture 39: May 3

Last time

- Theoretical background of linear model

Today

- Course evaluation (12/17)
- Bootstrap (JF Chapter 21)
- Logistic Regression (JF Chapter 14)

Additional reference

"Essential Statistical Inference Theory and Methods" by Dr. Dennis D. Boos and Dr. L. A. Stefanski.
Dr. Hua Zhou's Computational Statistics notes.

Bootstrap

We follow JF Chapter 21 to discuss the version of nonparametric bootstrap here. The term *bootstrapping*, coined by Efron (1979), refers to using the sample to learn about the sampling distribution of a statistic without reference to external assumptions – as in "pulling oneself up by one's bootstraps."

Bootstrapping offers a number of advantages:

- The bootstrap is quite general, although there are some cases in which it fails.
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small.
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically.
- It is relatively simple to apply the bootstrap to complex data collection plans.

Bootstrap standard errors

For simplicity, we start with an iid sample $Y_1, \ldots, Y_n$ with each Y_i having distribution function F, and a real parameter θ is estimated by $\hat{\theta}$. When necessary, we think of $\hat{\theta}$ as a function of the sample, $\hat{\theta}(Y_1, \ldots, Y_n)$. The variance of $\hat{\theta}$ is then

$$\mathrm{Var}_F(\hat{\theta}) = \int \left\{ \hat{\theta}(y_1, \ldots, y_n) - \mathrm{E}_F(\hat{\theta}) \right\}^2 dF(y_1) \ldots dF(y_n),$$

where

$$\mathrm{E}_F(\hat{\theta}) = \int \hat{\theta}(y_1, \ldots, y_n) dF(y_1) \ldots dF(y_n).$$

The nonparametric bootstrap estimate of $\mathrm{Var}(\hat{\theta})$ is just to replace F by the empirical distribution function $F_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \leqslant y)$:

$$\mathrm{Var}_{F_n}(\hat{\theta}) = \int \left\{ \hat{\theta}(y_1, \ldots, y_n) - \mathrm{E}_{F_n}(\hat{\theta}) \right\}^2 dF_n(y_1) \ldots dF_n(y_n),$$

Please refer to Chapter 11 of Boos and Stefanski for a complete discussion.

A practical bootstrapping procedure follows:

1. Create r number of bootstrap replications or pseudo-replicates – that is, for each bootstrap sample (replicate) $b = 1, \ldots, r$, we randomly draw n observations $\{Y_{b_1}^*, Y_{b_2}^*, \ldots, Y_{b_n}^*\}$ with replacement from the original sample $\{Y_1, Y_2, \ldots, Y_n\}$.
2. Obtain an estimate $\hat{\theta}_b^*$ of each bootstrap sample.
3. Use the distribution of $\hat{\theta}_b^*$ to estimate properties of the sampling distribution of $\hat{\theta}$.For example, the sample standard deviation of $\hat{\theta}_b^*$ gives the bootstrap standard error estimates of $\widehat{SE}^*(\hat{\theta})$.

Bootstrap example

We use the example in JF 21.1 for illustration. Imagine that we sample (fake) ten working, married couples, determining in each case the husband's and wife's income, as recorded in the table (JF table 21.3) below.

Observation	husband's Income	Wife's Income	Difference Y_i
1	34	28	6
2	24	27	-3
3	50	45	5
4	54	51	3
5	34	28	6
6	29	19	10
7	31	20	11
8	32	40	-8
9	40	33	7
10	34	25	9

A point estimate of this population mean difference μ is the sample mean,

$$\bar{Y} = \frac{\sum Y_i}{n} = 4.6$$

Elementary statistical theory tells us that the standard deviation of the sampling distribution of sample means is $\text{SD}(\bar{Y}) = \sigma/\sqrt{n}$, where σ is the population standard deviation of Y. Because we do not know σ in most real applications, the usual estimator of σ is the sample standard deviation

$$\hat{S} = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}$$

and we obtain the 95% confidence interval by

$$\bar{Y} \pm t_{n-1,0.025}\frac{\hat{S}}{\sqrt{n}}$$

In the present case, $\hat{S} = 5.948$, $\widehat{SE}(\bar{Y}) = 5.948/\sqrt{10} = 1.881$, and $t_{9,0.025} = 2.262$. The 95% confidence interval for the population mean μ is therefore

$$4.6 \pm 2.262 \times 1.881 = 4.6 \pm 4.255$$

or equivalently,

$$0.345 < \mu < 8.855$$

To illustrate the bootstrap procedure,

1. We can draw $r = 2000$ bootstrap samples (using a computer), each of size $n = 10$, from the original data given in table 21.3.
2. We then calculate the mean $\bar{Y}_b^*$, with $b = 1, \ldots, r$ for each bootstrap sample.
3. The bootstrap estimate of the standard error is then given by $\widehat{SE}^*(\bar{Y}^*) = \sqrt{\frac{\sum_{b=1}^{r}\left(\bar{Y}_b^* - \bar{\bar{Y}}*\right)^2}{r-1}}$

From the 2000 replicates that Dr. Fox drew, he obtained $\bar{\bar{Y}}^* = 4.693$ and $\widehat{SE}(\bar{Y}^*) = 1.750$. Both are quite close to the theoretical values (read JF 21.1 for a discussion over $\sqrt{n/n-1}$ for the differences in calculating the standard errors, which is often negligible, especially when n is large).

Now, we can get a bootstrap estimate for the $100(1-\alpha)\%$ confidence interval by using the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the bootstrap sampling distribution of $\hat{\theta}_b^*$ which means

1. We order $\hat{\theta}_b^*$ such that $\hat{\theta}_{(1)}^* \leqslant \hat{\theta}_{(2)}^* \leqslant \cdots \leqslant \hat{\theta}_{(r)}^*$.
2. Find the two quantiles $\hat{\theta}_{(lower)}^* = \hat{\theta}_{(\alpha/2\times r)}^*$ and $\hat{\theta}_{(upper)}^* = \hat{\theta}_{((1-\alpha/2)\times r)}^*$
3. Construct the confidence interval by $(\hat{\theta}_{(lower)}^*, \hat{\theta}_{(upper)}^*)$.

In this case,

$$\begin{aligned}
\text{lower} &= 2000(0.05/2) = 50 \\
\text{upper} &= 2000(1 - 0.05/2) = 1950 \\
\bar{Y}^*_{(50)} &= 0.7 \\
\bar{Y}^*_{(1950)} &= 7.8 \\
0.7 < \mu &< 7.8
\end{aligned}$$

Bias-corrected bootstrap intervals

We introduce the bias-corrected version of the above bootstrap intervals through two "correction factors" Z and A defined below:.

1. Calculate

   $$Z \equiv \Phi^{-1}\left[\frac{\sum_{b=1}^{r} I(\hat{\theta}^*_b < \hat{\theta})}{r}\right]$$

   where $\Phi^{-1}(\cdot)$ is the inverse of the standard-normal distribution and $\sum_{b=1}^{r} I(\hat{\theta}^*_b < \hat{\theta})/r$ is the proportion of bootstrap replicates below the estimate $\hat{\theta}$. If the bootstrap sampling distribution is symmetric and if $\hat{\theta}$ is unbiased, then this proportion will be close to 0.5, and the "correction factor" Z will be close to 0.

2. Let $\hat{\theta}_{(-i)}$ represent the value of $\hat{\theta}$ produced when ith observation is deleted from the sample (known as the jackknife values of $\hat{\theta}$). There are n of these quantities. Let $\bar{\theta} = \sum \hat{\theta}_{(-i)}/n$. Then calculate

   $$A \equiv \frac{\sum_{i=1}^{n}(\bar{\theta} - \hat{\theta}_{(-i)})^3}{6\left[\sum_{i=1}^{n}(\bar{\theta} - \hat{\theta}_{(-i)})^2\right]^{3/2}}$$

   With the correction factors Z and A, compute

   $$\begin{aligned}
   A_1 &\equiv \Phi\left[Z + \frac{Z - z_{\alpha/2}}{1 - A(Z - z_{\alpha/2})}\right] \\
   A_2 &\equiv \Phi\left[Z + \frac{Z + z_{\alpha/2}}{1 - A(Z + z_{\alpha/2})}\right]
   \end{aligned}$$

   And the corrected interval is

   $$\hat{\theta}^*_{(lower)} < \theta < \hat{\theta}^*_{(upper)}$$

   where lower* $= rA_1$ and upper* $= rA_2$ (rounding or interpolating as required).

When the correction factors Z and A are both 0, $A_1 = \Phi(-z_{\alpha/2}) = \alpha/2$ and $A_2 = \Phi(z_{\alpha/2}) = 1 - \alpha/2$.

For the 2000 bootstrap samples that Dr. Fox drew, there are 926 bootstrapped means below $\bar{Y} = 4.6$, and so $Z = \Phi^{-1}(926/2000) = -0.09288$. The $\bar{Y}_{(-i)}$ are $4.444, 5.444, \ldots, 4.111$. And

$A = -0.05630$. Using the correction factors Z and A,

$$\begin{aligned}
A_1 &= \Phi\left[-0.09288 + \frac{-0.09288 - 1.96}{1 - [-0.05630(-0.09288 - 1.96)]}\right] \\
&= \Phi(-2.414) = 0.007889 \\
A_2 &= \Phi\left[-0.09288 + \frac{-0.09288 + 1.96}{1 - [-0.05630(-0.09288 + 1.96)]}\right] \\
&= \Phi(1.597) = 0.9449
\end{aligned}$$

Multiplying by r, we have $2000 \times 0.007889 \approx 16$ and $2000 \times 0.9449 \approx 1890$, from which

$$\begin{aligned}
&\bar{Y}^*_{(16)} < \mu < \bar{Y}^*_{(1890)} \\
&-0.4 = \mu < 7.3
\end{aligned}$$

Logistic regression

So far, we only considered cases where the response variable is continuous. Logistic regression belongs in the family of Generalized Linear Model that can be used for analyzing binary responses.

Motivation Let p be the probability of a specific outcome. We are interested in how this probability is affected by the explanatory variables. A naive approach could be:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Problem p must be between 0 and 1.

Solution Model log odds of p (i.e. logit of p) which are defined as

$$\begin{aligned}
\text{odds} &= \frac{p}{1-p} \in [0, \infty] \\
\text{logit} &= \log(\frac{p}{1-p}) \in (-\infty, \infty)
\end{aligned}$$

This forms the logistic regression

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Note that

1. Increase in log odds $\iff$ increase in p.
   Decrease in log odds $\iff$ decrease in p.
2. No ϵ in logistic regression because we observe a binary outcome y_i, not p itself.

The density

$$\begin{aligned} f(y_i|p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= e^{y_i \log(p_i) + (1-y_i)\log(1-p_i)} \\ &= e^{y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i)} \end{aligned}$$

where

$$\begin{aligned} \mathrm{E}(y_i) = p_i &= \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \quad \text{(mean function, inverse link function)} \\ \mathbf{x}_i^T \boldsymbol{\beta} &= \log(\frac{p_i}{1-p_i}) \quad \text{(logit link function)} \end{aligned}$$

We obtain parameter estimates by maximum likelihood. Read page 131 - page 133 of Dr. Hua Zhou's Computational Statistics notes (link) for algorithms to find these MLE (maximum likelihood estimates).