

# 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:

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

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

- Gauss-Markov model:

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

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

- Aitken model or general linear model:

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

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

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

- General mixed linear Model:

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

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

- 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  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are *linearly dependent* if there exist coefficients  $c_j$  for  $j = 1, 2, \dots, n$  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  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}$  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  $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \dots, \mathbf{e}^{(n)}$ , 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  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}$ , written as  $\mathcal{S} = \text{span}\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}$ , if any vector  $\mathbf{x}$  in the vector space is a linear combination of  $\mathbf{x}_i, i = 1, 2, \dots, n$ .
- 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  $5\mathbf{x}^{(1)} - 2\mathbf{x}^{(2)} + \mathbf{x}^{(3)} = \mathbf{0}$

## 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.
- $\text{rank}(\mathbf{A}) \leq \min\{m, n\}$ .
- 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.)
- $\text{rank}(\mathbf{AB}) \leq \min\{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\}$ . (Hint: Columns of  $\mathbf{AB}$  are spanned by columns of  $\mathbf{A}$  and rows of  $\mathbf{AB}$  are spanned by rows of  $\mathbf{B}$ .)
- if  $\mathbf{Ax} = \mathbf{0}_m$  for some  $\mathbf{x} \neq \mathbf{0}_n$ , then  $\text{rank}(\mathbf{A}) \leq n - 1$ .

### 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  $j^{\text{th}}$  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  $\mathbf{A}\mathbf{c}$  is a linear combination of columns in  $\mathbf{A}$ .

*solution:*

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



You could recognize that

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

#### Result A.1

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

*proof:* Each column of  $\mathbf{AB}$  is a linear combination of columns of  $\mathbf{A}$  (i.e.  $(\mathbf{AB})_{\cdot j} = \mathbf{A}\mathbf{b}^{(j)}$ ), so the number of linearly independent columns of  $\mathbf{AB}$  cannot be greater than that of  $\mathbf{A}$ . 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  $\mathbf{x} = \mathbf{A}\mathbf{d} = \mathbf{B}(\mathbf{C}\mathbf{d})$ .

For (b),  $\mathbf{A}_{\cdot j} \in C(\mathbf{B})$ , so that there exists a vector  $\mathbf{c}^{(j)}$  such that  $\mathbf{A}_{\cdot j} = \mathbf{B}\mathbf{c}^{(j)}$ . The matrix  $\mathbf{C} = (\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \dots, \mathbf{c}^{(n)})$  satisfies that  $\mathbf{A} = \mathbf{BC}$ .

### Null space

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

#### Result A.3

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

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

#### Theorem A.1

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

*MisInterpretation:* 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.

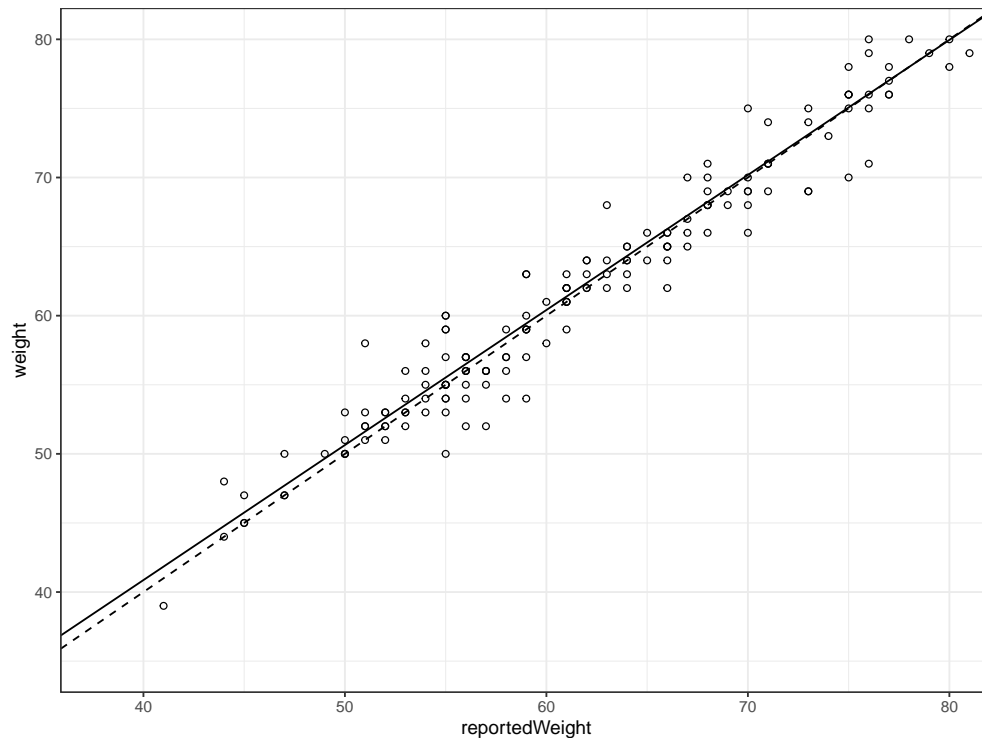


Figure 3.1: Scatterplot of Davis's data on the measured and reported weight of 101 women. The dashed line gives  $y = x$ .

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:  $\beta_0$
- slope:  $\beta_1$
- residual/error term  $\epsilon_i$ .

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  $\beta_0, \beta_1$  such that

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

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

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

*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*
- $\beta$  is the vector of parameters
- $\epsilon$  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

$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{n \sum_i (x_i - \bar{x})^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}$$

Now we have

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

Some properties:

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

*Proof:* For (a), we look at

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

## Other quantities in Matrix Form

Fitted values

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ \hat{\beta}_0 + \hat{\beta}_1 x_2 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \mathbf{X} \hat{\beta}$$

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*  $n - 2$ :

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

The residual standard error is,

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

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



## Sample correlation coefficient

*Definition:* The sample correlation coefficient  $r_{xy}$  of the paired data  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  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  $1$ :

$$-1 \leq r_{xy} \leq 1$$

- 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:

$$\text{TSS} = 4753.8$$

$$\text{RSS} = 418.87$$

$$\text{RegSS} = 4334.9$$

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  $\beta_0$  and  $\beta_1$  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
- $\epsilon_i$  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:  $\mathbf{E}(\epsilon) \equiv \mathbf{E}(\epsilon|x_i) = 0$ . And equivalently, the expected value of the response variable is a linear function of the explanatory variable:  $\mu_i \equiv \mathbf{E}(y_i) \equiv \mathbf{E}(y_i|x_i) = \mathbf{E}(\beta_0 + \beta_1 x_i + \epsilon_i|x_i) = \beta_0 + \beta_1 x_i$ .
- *Constant variance*. The variance of the errors is the same regardless of the value of  $x$ :  $\mathbf{Var}(\epsilon|x_i) = \sigma_\epsilon^2$ . The constant error variance implies constant conditional variance of  $y$  on given  $x$ :  $\mathbf{Var}(y|x_i) = \mathbf{E}((y_i - \mu_i)^2) = \mathbf{E}((y_i - \beta_0 - \beta_1 x_i)^2) = \mathbf{E}(\epsilon_i^2) = \sigma_\epsilon^2$ . (Question: why the last equal sign?)
- *Normality*. The errors are independent identically distributed with Normal distribution with mean 0 and variance  $\sigma_\epsilon^2$ . Write as  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$ . Equivalently, the conditional distribution of the response variable is normal:  $y_i \stackrel{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2)$ .
- *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.

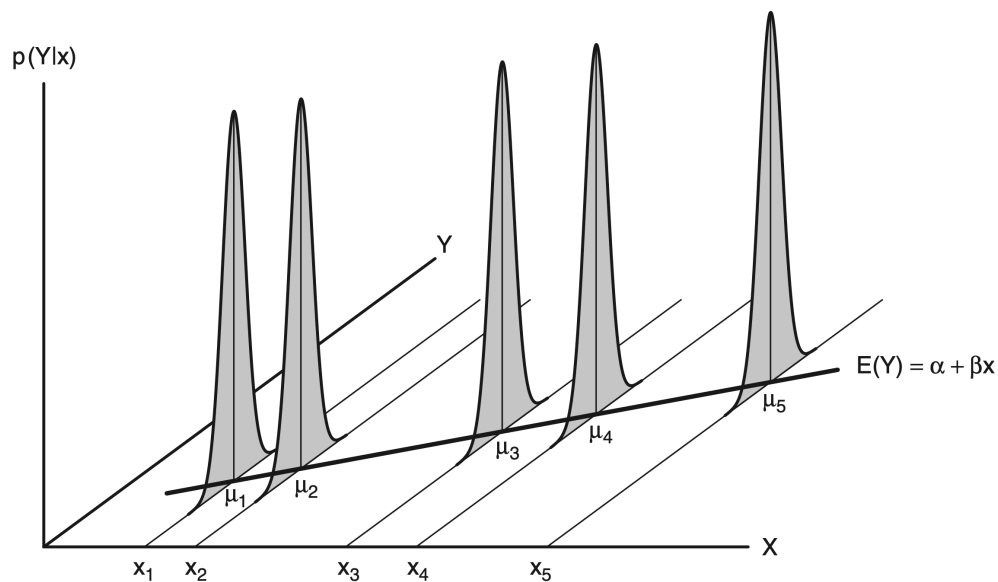


Figure 6.1: The assumptions of linearity, constant variance, and normality in simple regression. The graph shows the conditional population distributions  $\Pr(Y|x)$  of  $Y$  for several values of the explanatory variable  $X$ , labeled as  $x_1, x_2, \dots, x_5$ . The conditional means of  $Y$  given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

## 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  $\beta_0$  and  $\beta_1$ :

- 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:

$$\mathbf{E}(\hat{\beta}_0) = \beta_0$$

$$\mathbf{E}(\hat{\beta}_1) = \beta_1$$

*Proof:*

method (a)  $\mathbf{E}(\hat{\beta}) = \mathbf{E}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}) = \mathbf{E}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta) = \beta$ . (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  $\mathbf{E}(\hat{\beta}_1) = \mathbf{E}(\sum k_i y_i) = \sum [k_i \mathbf{E}(y_i)] = \sum [k_i (\beta_0 + \beta_1 x_i)] = \beta_0 \sum k_i + \beta_1 \sum (k_i x_i) = \beta_1$ , and  $\mathbf{E}(\hat{\beta}_0) = \mathbf{E}(\bar{y} - \hat{\beta}_1 \bar{x}) = \mathbf{E}(\bar{y}) - \bar{x} \mathbf{E}(\hat{\beta}_1) = \mathbf{E}(\frac{1}{n} \sum y_i) - \bar{x} \beta_1 = \frac{1}{n} [\sum \mathbf{E}(y_i)] - \bar{x} \beta_1 = \frac{1}{n} \sum [\beta_0 + x_i \beta_1] - \bar{x} \beta_1 = \beta_0$

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

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}$$

*Proof:*

$$\text{Var}(\hat{\beta}_1) = \text{Var}(\sum k_i y_i) = \sum k_i^2 \text{Var}(y_i) = \sigma_\epsilon^2 \sum k_i^2 = \sigma_\epsilon^2 \frac{\sum_i (x_i - \bar{x})^2}{[\sum_j (x_j - \bar{x})^2]^2} = \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}, \text{ and}$$

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) = \text{Var}(\bar{y}) + (\bar{x})^2 \text{Var}(\hat{\beta}_1) - 2\bar{x} \text{Cov}(\bar{Y}, \hat{\beta}_1).$$

Now,

$$\text{Var}(\bar{y}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(y_i) = \frac{\sigma^2}{n},$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2},$$

and

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

Finally,

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

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

- The error variance  $\sigma_\epsilon^2$  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  $\tilde{\beta}_1$  be another linear unbiased estimator such that  $\tilde{\beta}_1 = \sum c_i y_i$ . For  $\tilde{\beta}_1$  is still unbiased as above,  $\mathbf{E}(\tilde{\beta}_1) = \beta_0 \sum c_i + \beta_1 \sum c_i x_i = \beta_1$  for all  $\beta_1$ , we have  $\sum c_i = 0$  and  $\sum c_i x_i = 1$ .

$$\mathbf{Var}(\tilde{\beta}_1) = \sigma_\epsilon^2 \sum c_i^2$$

Let  $c_i = k_i + d_i$ , then

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

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  $\beta_0$  and  $\beta_1$ . (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,

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

## 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  $\hat{\beta}_1$

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

However,  $\sigma_\epsilon$  is never known in practice. Instead, an *unbiased* estimator of  $\sigma_\epsilon^2$  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  $\mathbf{E}(\sum (y_i - \hat{y}_i)^2) = \sigma_\epsilon^2(n-2)$ .

LHS:  $\mathbf{E}(\sum (y_i - \hat{y}_i)^2) = \sum_i [\mathbf{E}(y_i - \hat{y}_i)^2]$

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

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

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

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

Therefore, we have for  $i$ th residue

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

And finally, sum over  $i$  we get

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



## 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  $\beta_1$  and  $\beta_0$  are given by

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

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

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

The conditional mean  $\mathbf{E}(Y|X = x_0)$  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):

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 | X = x_0) = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

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  $\beta_{slope_0}$  (most commonly, the null hypothesis has  $\beta_{slope_0} = 0$ ), we calculate the test statistic ( $T$ -statistics) with  $df = n - 2$

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