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.

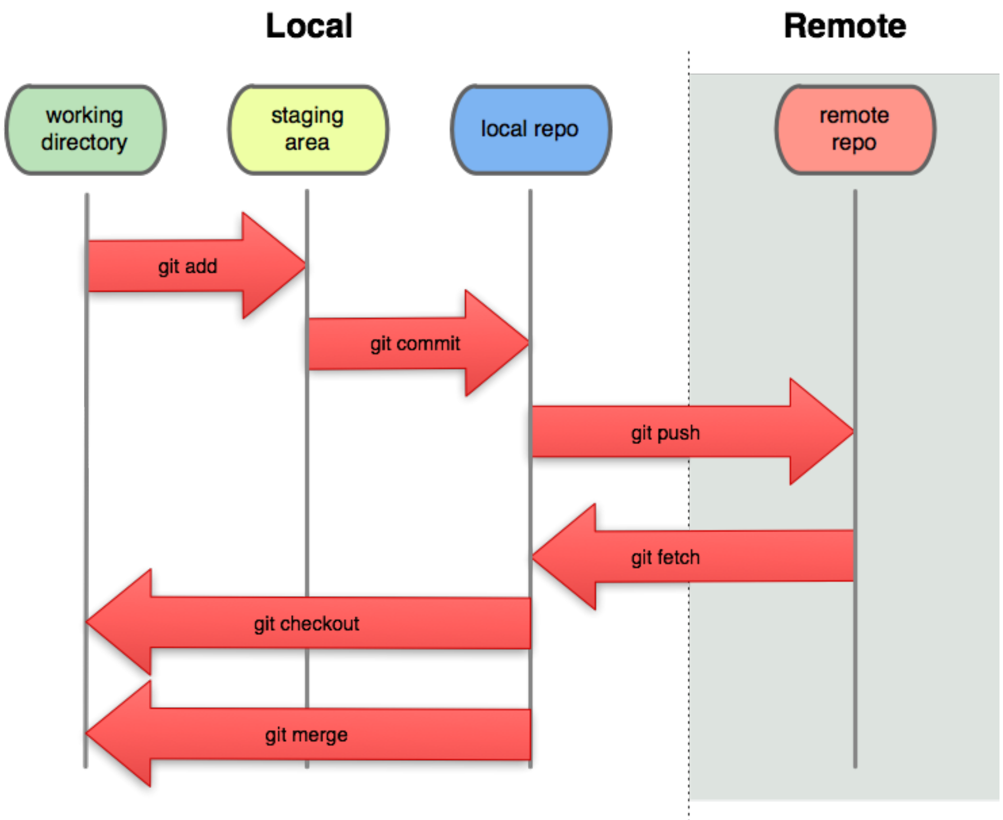

Figure 2.1

- 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,

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

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, $\dim\left(C(\mathbf{A})\right) = \text{rank}(\mathbf{A})$.
- 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.

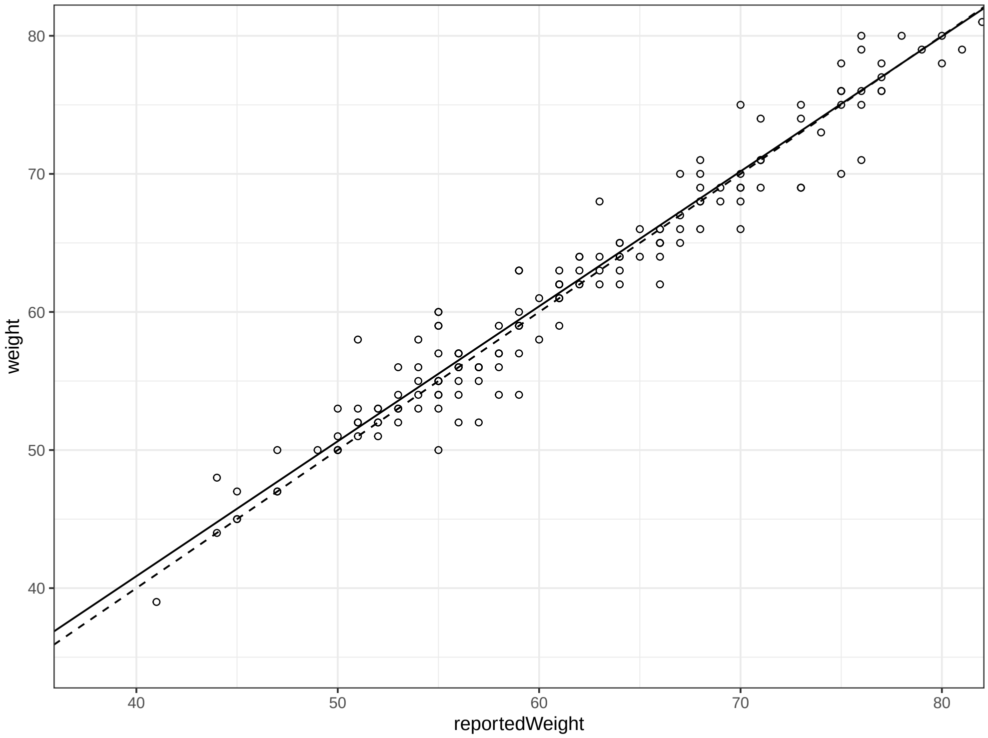

Figure 3.1: Scatterplot of Davis's data on the measured and reported weight of 101 women. The dashed line gives $y = x$.

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

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

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

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

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) $\sum x_i\hat{\epsilon}_i = 0$.
- (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

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

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.

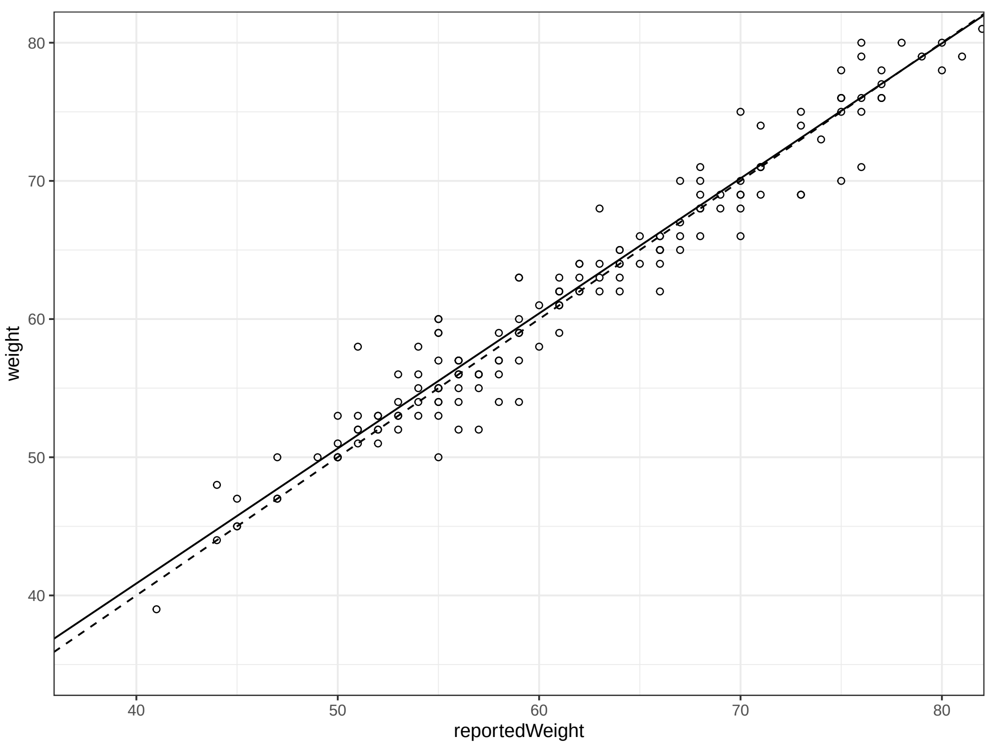

Figure 4.1: Scatterplot of Davis's data on the measured and reported weight of 101 women. The dashed line gives $y = x$. The solid line gives the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

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: $\mathbf{E}(\epsilon) \equiv \mathbf{E}(\epsilon|x_i) = 0$. 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 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.

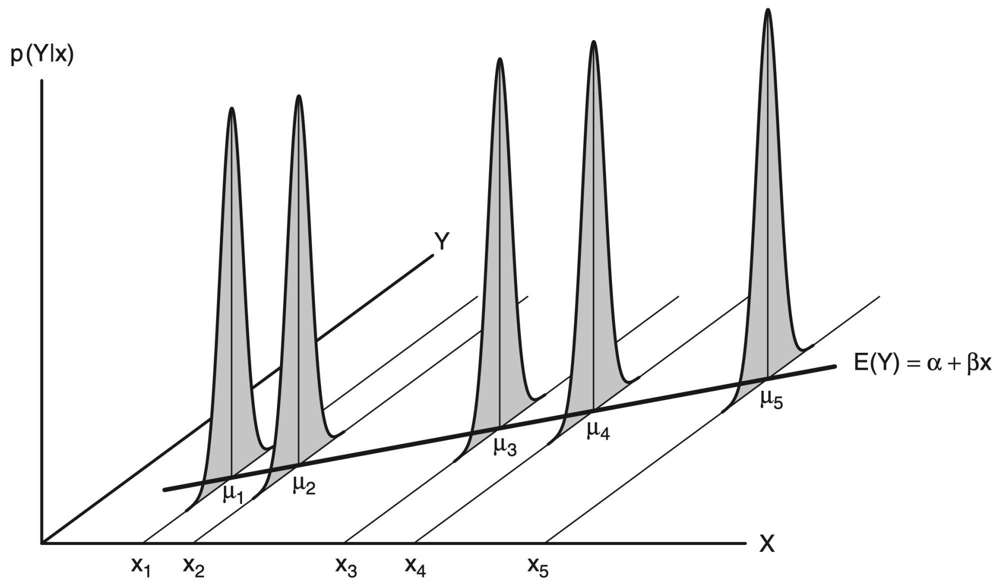

Figure 6.1: The assumptions of linearity, constant variance, and normality in simple regression. The graph shows the conditional population distributions $\Pr(Y|x)$ of Y for several values of the explanatory variable X, labeled as $x_1, x_2, \ldots, x_5$. The conditional means of Y given x are denoted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

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:
$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\sum \mathrm{k_i y_i}) = \sum \mathrm{k_i^2 Var(y_i)} = \sigma_\epsilon^2 \sum \mathrm{k_i^2} = \sigma_\epsilon^2 \frac{\sum_\mathrm{i}(\mathrm{x_i}-\bar{\mathrm{x}})^2}{[\sum_\mathrm{j}(\mathrm{x_j}-\bar{\mathrm{x}})^2]^2} = \frac{\sigma_\epsilon^2}{\sum(\mathrm{x_i}-\bar{\mathrm{x}})^2}$, and
$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{\mathrm{y}} - \hat{\beta}_1\bar{\mathrm{x}}) = \mathrm{Var}(\bar{\mathrm{y}}) + (\bar{\mathrm{x}})^2\mathrm{Var}(\hat{\beta}_1) - 2\bar{\mathrm{x}}\mathrm{Cov}(\bar{\mathrm{Y}}, \hat{\beta}_1)$.
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 $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{(\mathrm{n}-1)\mathrm{S_X^2}}$, 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 $\mathrm{E}[(y_i-\hat{y}_i)^2] = \mathrm{Var(y_i-\hat{y}_i)} + [\mathbf{E}\,(\mathrm{y_i-\hat{y}_i})]^2 = \mathrm{Var(y_i-\hat{y}_i)} = \mathrm{Var(y_i)} + \mathrm{Var(\hat{y}_i)} - 2\mathrm{cov(y_i,\hat{y}_i)}$

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

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

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

And finally, sum over i we get

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

Confidence intervals

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

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

to get the estimated standard errors:

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

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 $\mathbf{E}(Y|X = x_0)$

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):

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

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.

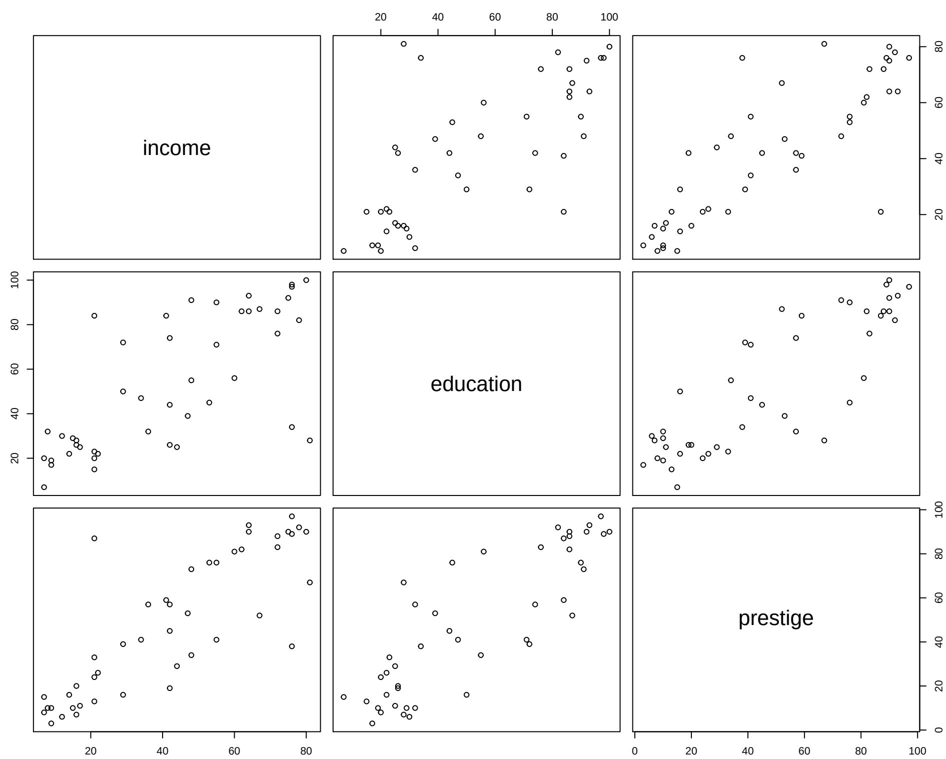

Figure 9.1: Scatterplot matrix for occupational prestige, level of education, and level of income of 45 U.S. occupations in 1950.

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 $X_{i1} = income, X_{i2} = education$ for occupations $i = 1, 2, \ldots, 45$:

$$Y = \beta_0 + \beta_1 income + \beta_2 education + error$$

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 error variance 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 partial slope 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 <u>response vector</u> of size $n \times 1$
- $\mathbf{X}$ denotes a <u>design matrix</u> of size $n \times (p+1)$
- β denotes a vector of <u>regression parameters</u> of size $(p+1) \times 1$
- ϵ denotes an <u>error vector</u> 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

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

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

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

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

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

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

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.

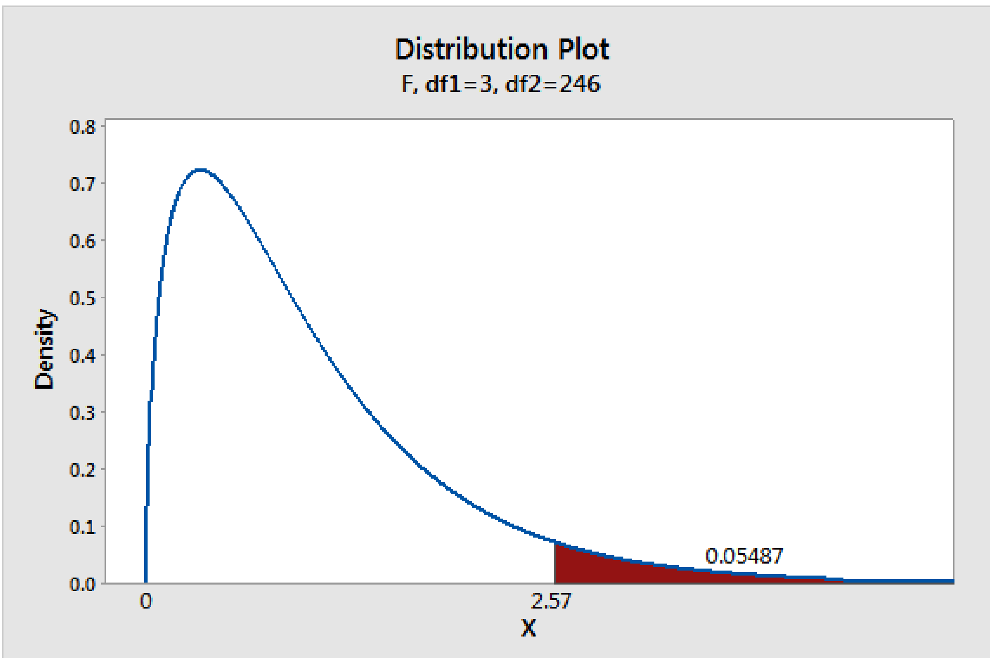

Figure 10.1: An example for p-value for F-statistic value 2.57 with an $F_{3,246}$ distribution

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 $\widehat{\boldsymbol{\Sigma}}$)

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: $SE(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})} = \sqrt{\mathrm{Var}(\mathbf{a}^{\mathrm{T}}\hat{\beta})} = \sqrt{\mathbf{a}^T\widehat{\boldsymbol{\Sigma}}\mathbf{a}} = 3.67$

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. $\varnothing \in \mathcal{B}$ (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 $A_1, A_2, \dots \in \mathcal{B}$, 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 $A_1, A_2, \dots \in \mathcal{B}$ 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 random variable 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	$X = yield/acre$

Suppose we have a sample space
$$S = \{s_1, \dots, s_n\}$$
with a probability function Pr and we define a random variable X with range $\mathcal{X} = \{x_1, \dots, x_m\}$. 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 cumulative distribution function or *cdf* 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

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

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

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

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

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

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

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

Variance

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

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

properties

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

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

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

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

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

- $X \sim Geometric(p)$ (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	$p(1-p)$
$Binomial(n, p)$	$\begin{pmatrix} n \\ x \end{pmatrix} p^x(1-p)^{n-x}$, 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 $Z_1, Z_2, \ldots, Z_k \overset{iid}{\sim} N(0,1)$, 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 $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, 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

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

Moments

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

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

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

Sums of random vectors

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

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.

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

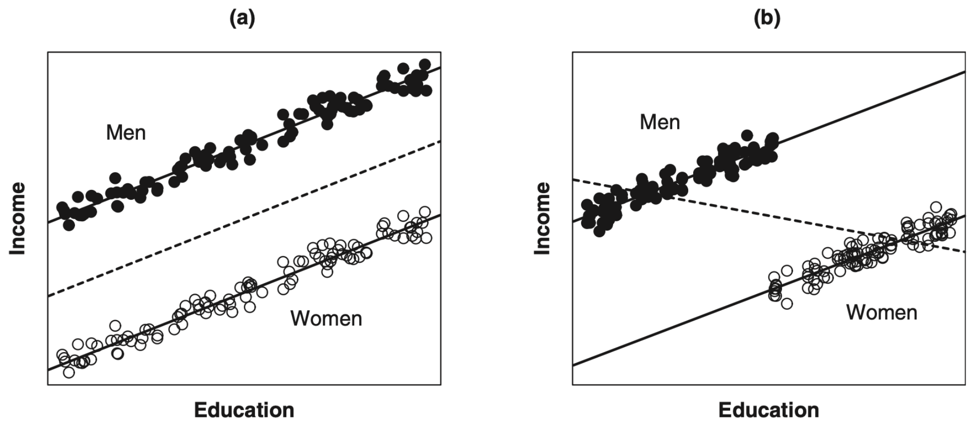

Figure 15.1: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), 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.

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. It is apparent in both Figure 15.2 (a) and (b) the within-gender regressions

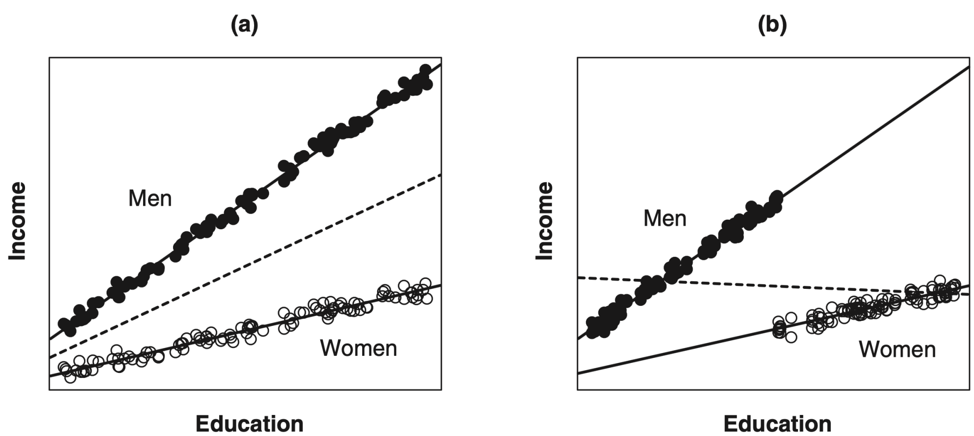

Figure 15.2: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), 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.

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:

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

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.

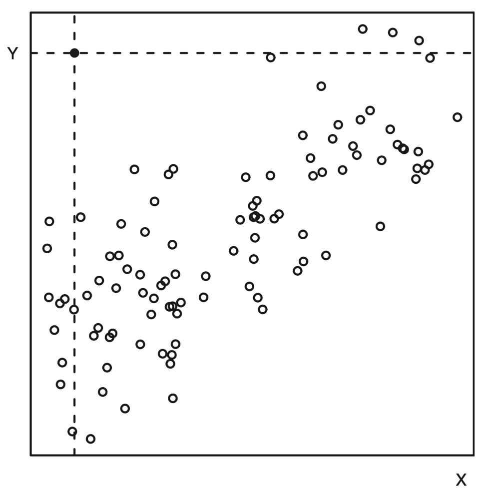

Figure 16.1: The black point is a regression outlier because it combines a relatively large value of Y with a relatively small value of X, even though neither its X-value nor its Y-value is unusual individually. Because of the positive relationship between Y and X, points with small X-values also tend 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 $Y = \beta_0 + \beta_1 X + \epsilon$.

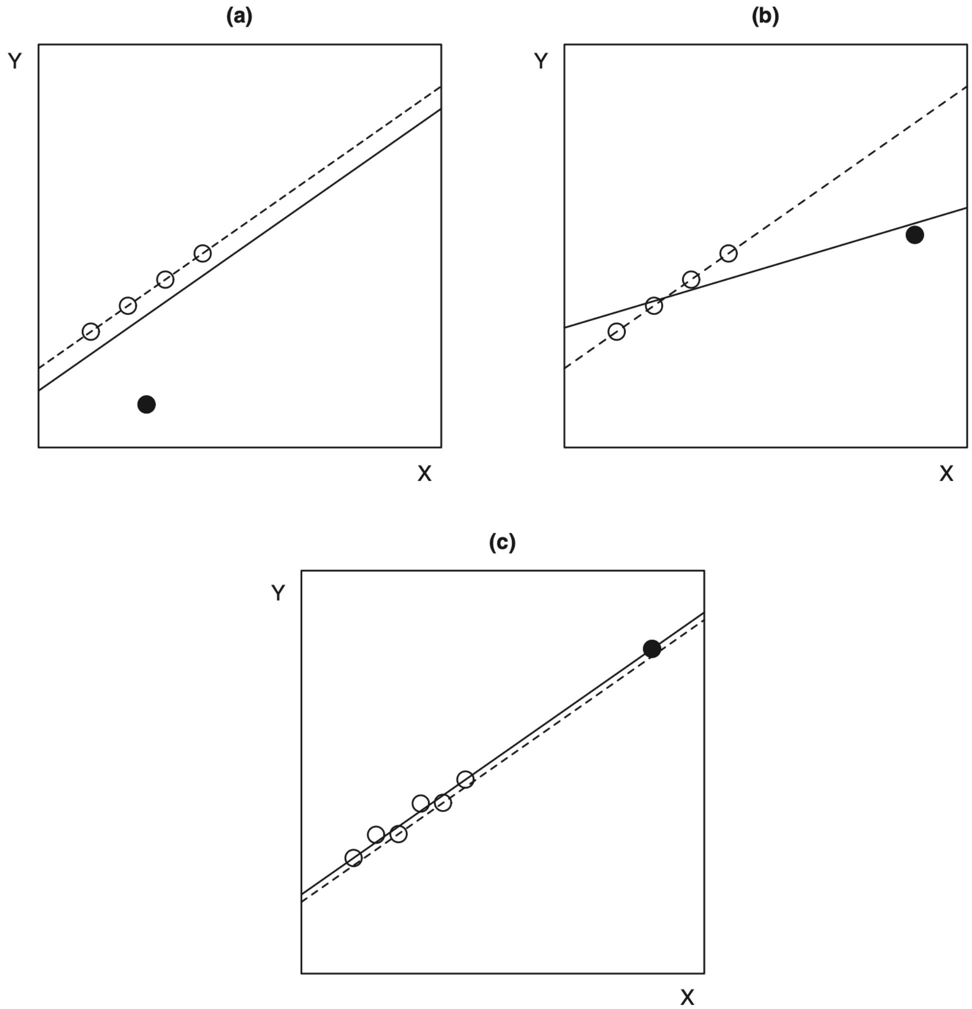

Figure 16.2: Leverage and influence in simple regression. In each graph, the solid line gives the least-squares regression for all the data, while the broken line gives the least-squares regression with the unusual data point (the black circle) omitted. (a) An outlier near the mean of X has low leverage 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 <u>outlier</u> is a data point whose response Y does not follow the general trend of the rest of the data.
- A data point has high <u>leverage</u> 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 ("outlying-ness") 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 $\hat{\epsilon}_i$,

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

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):

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

For large n,

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

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 $\hat{Y}_i^{(1)}$ 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 17.2 illustrates the added-variable plots using the Duncan's data.

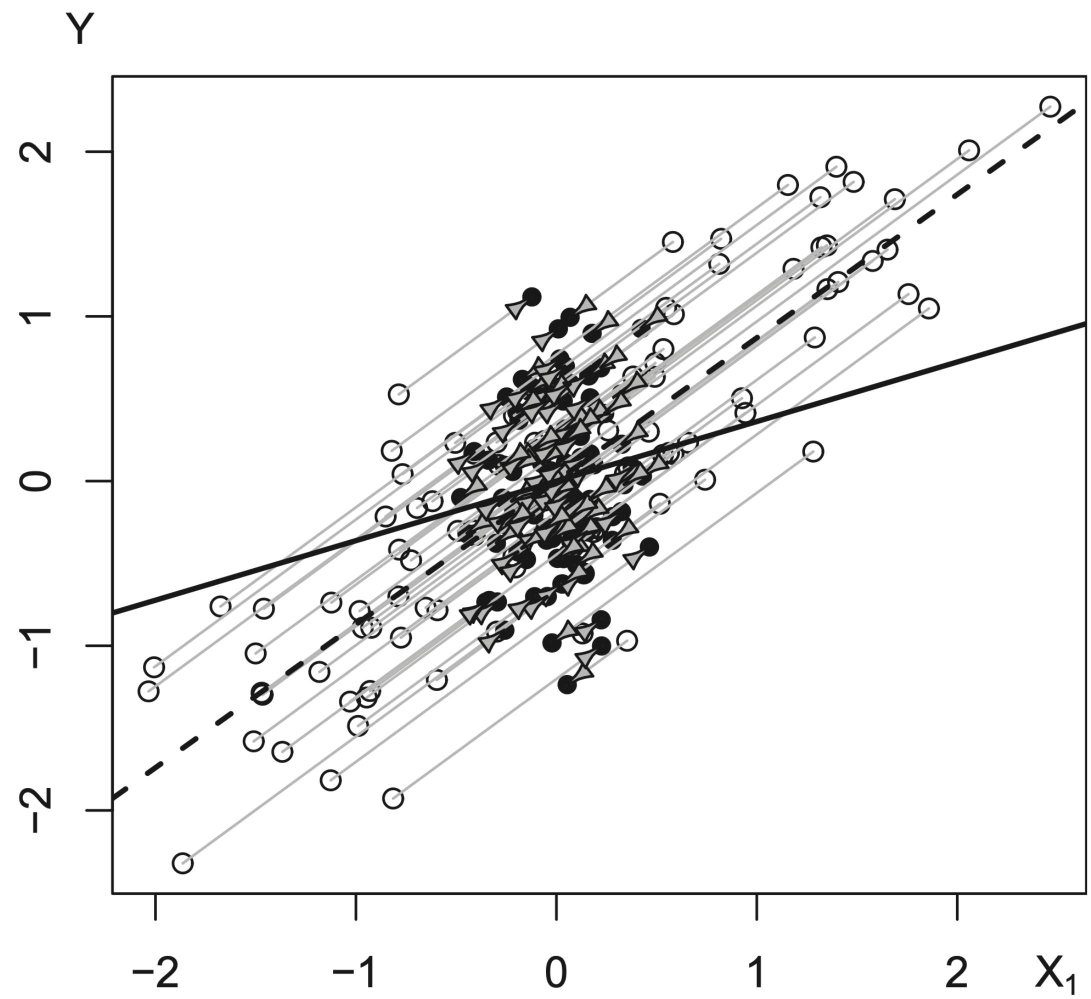

Figure 17.1: The marginal scatterplot (open circles) for Y and X_1 superimposed on the added-variable plot (filled circles) for X_1 in the regression of Y on X_1 and X_2. The variables Y and X_1 are centered at their means to facilitate the comparison of the two sets of points. The arrows show how the 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.

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

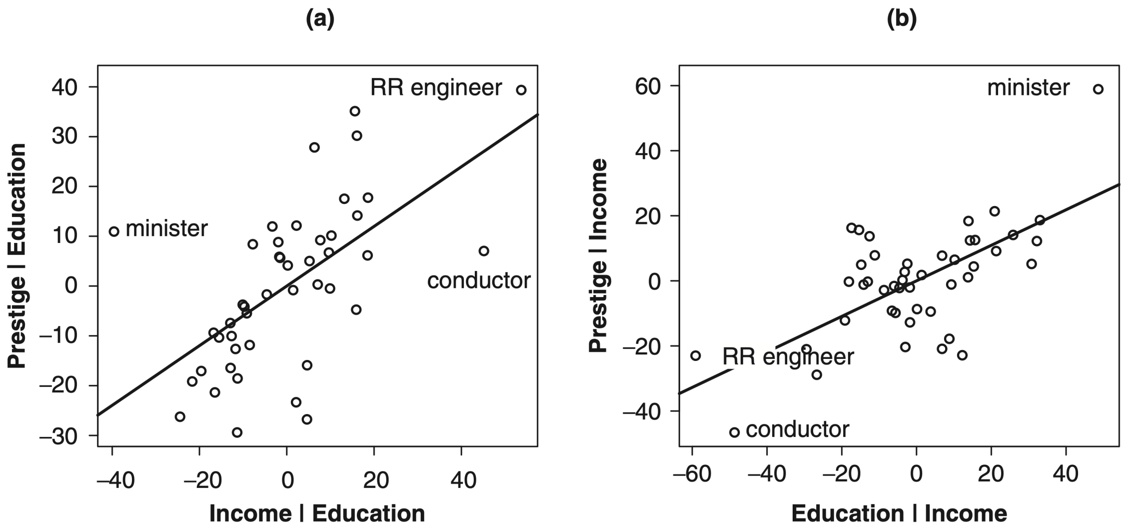

Figure 17.2: Added-variable plots for Duncan's regression of occupational prestige on the (a) income and (b) education levels of 45 US occupations in 1950. Three unusual observations, *miniters*, *conductors*, and *railroadengineers*, are identified on the plots. The added-variable plot for the intercept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

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 downweights 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 $\sigma = 10$. 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 18.3 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 18.4 shows the normal quantile-comparison plot for the distribution of infant mortality. The positive skew of the distribution is readily apparent.

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:

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

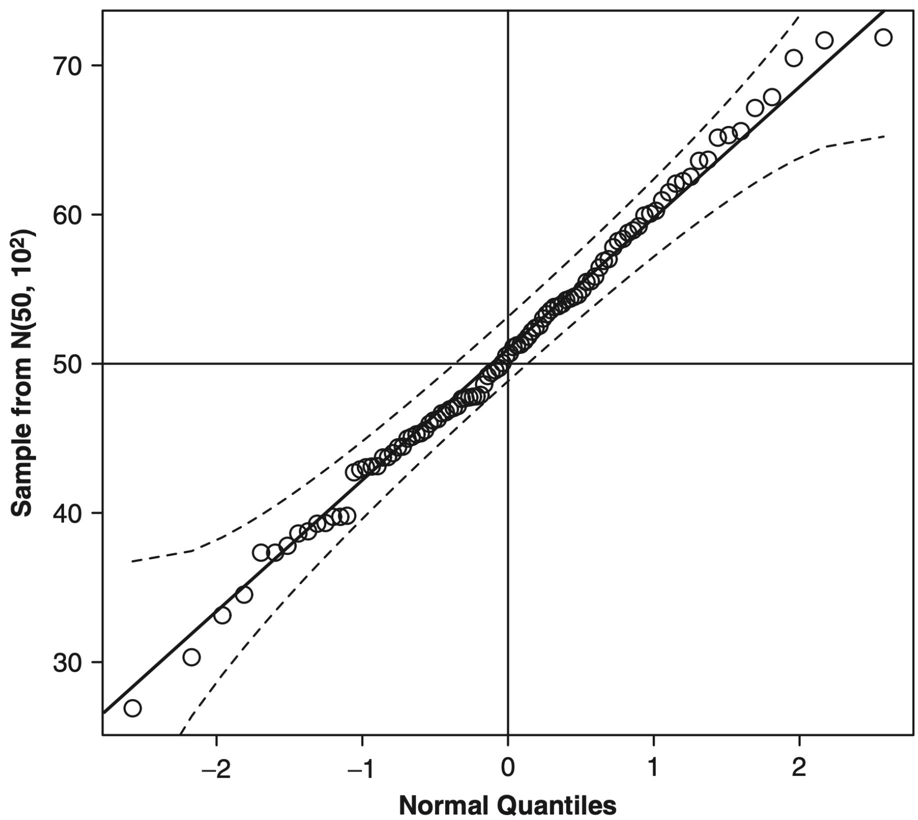

Figure 18.1: Normal quantile-comparison plot for a sample of 100 observations drawn from a normal distribution with mean 50 and standard deviation 10. The fitted line is through the quartiles of the distribution, the broken lines give a pointwise 95% confidence interval around the fit. JF Figure 3.8.

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. 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.

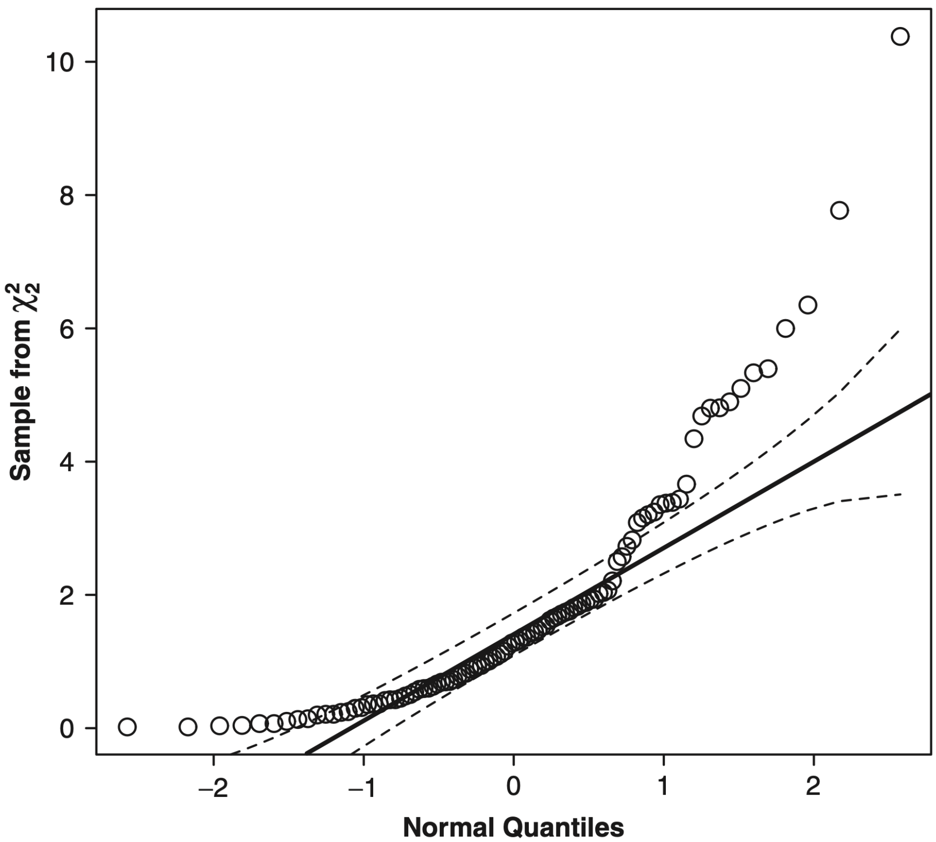

Figure 18.2: Normal quantile-comparison plot for a sample of 100 observations drawn from the positively skewed chi-square distribution with 2 degrees of freedom. JF Figure 3.9.

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

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

$$\hat{\sigma}_\epsilon^2 = \frac{\sum (w_i \hat{\epsilon}_i)^2}{n}$$

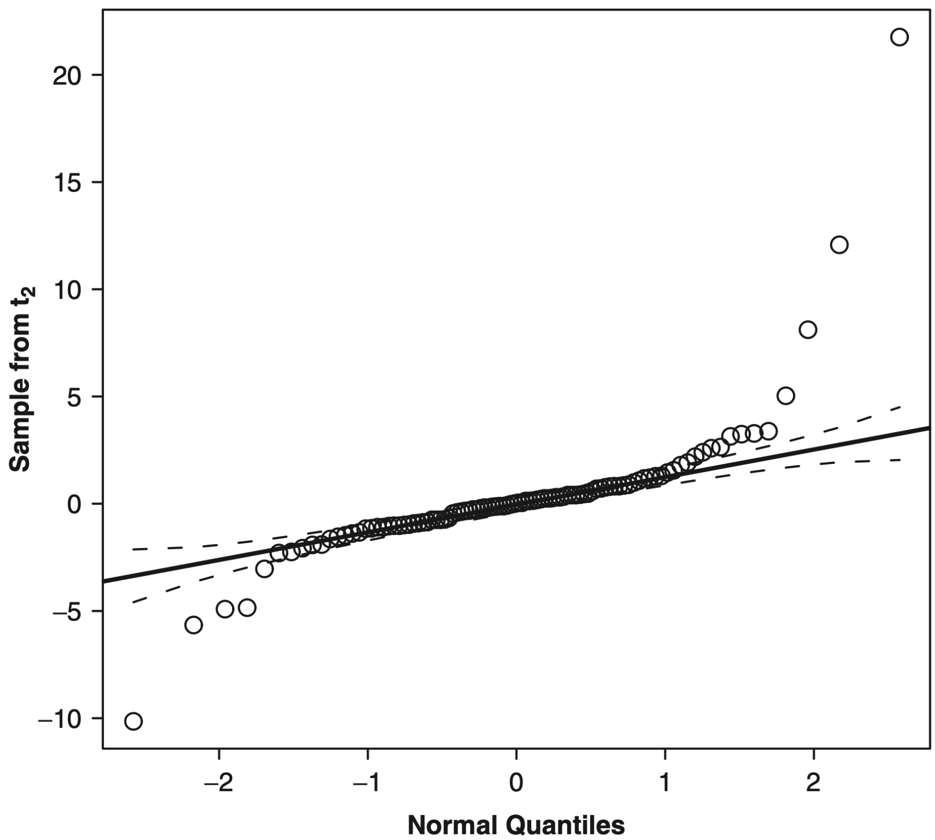

Figure 18.3: Normal quantile-comparison plot for a sample of 100 observations drawn from heavy-tailed t-distribution with 2 degrees of freedom. JF Figure 3.10.

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 $\boldsymbol{\Sigma} \equiv \mathbf{Var}\left(\mathbf{Y}\right) = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$, and

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

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

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

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

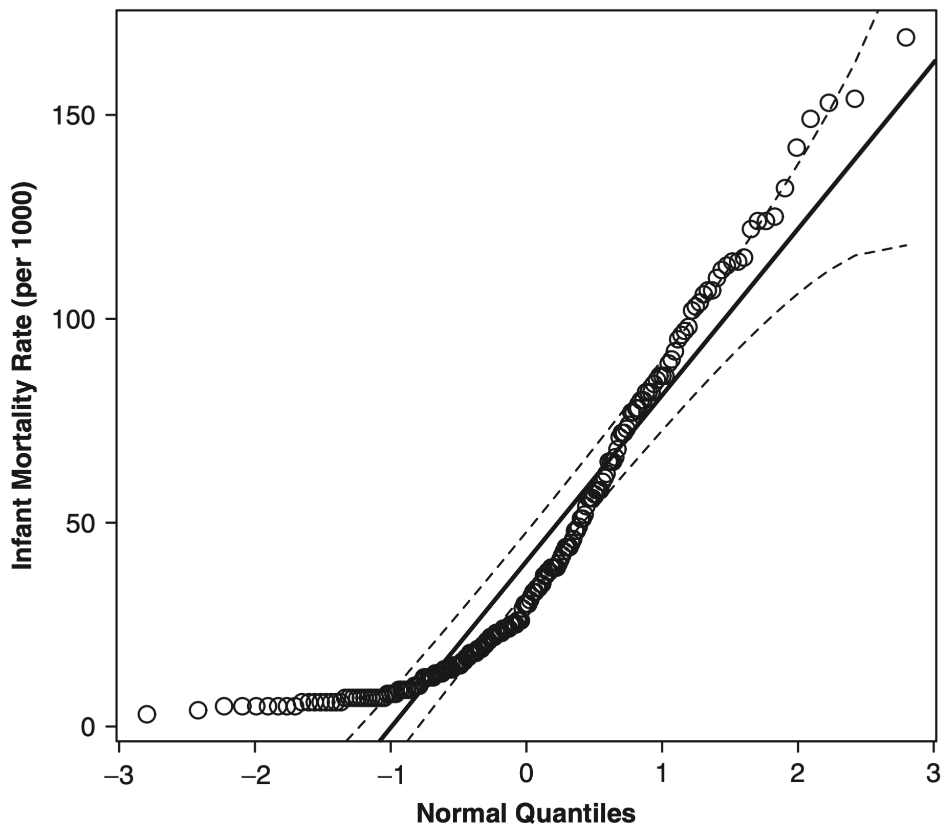

Figure 18.4: Normal quantile-comparison plot for the distribution of infant mortality. Note the positive skew. JF Figure 3.11.

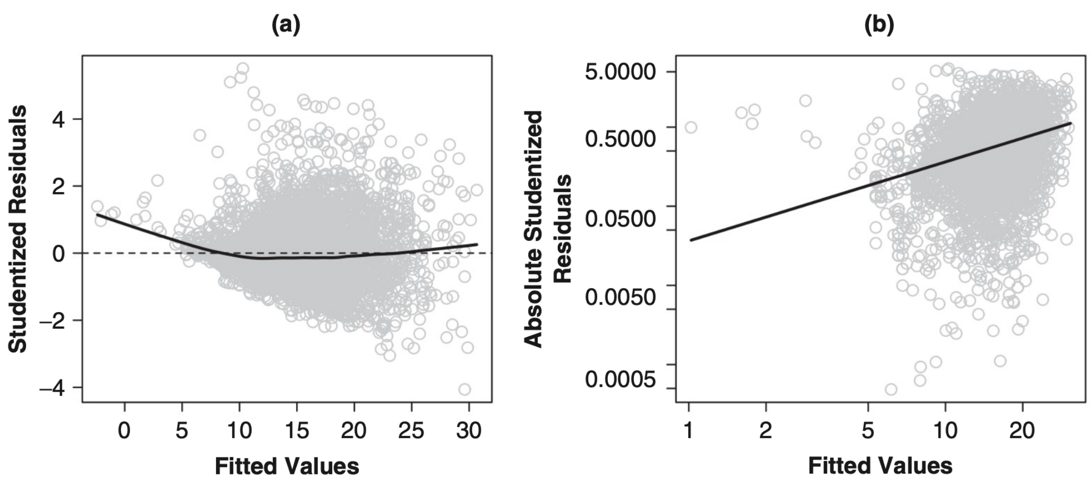

Figure 18.5: (a) Plot of studentized residuals versus fitted values and (b) spread-level plot for studentized residuals. JF Figure 12.3.

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

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

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

Data transformation

The family of powers and Roots

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

$$X \to X^p$$

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):

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

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 18.6

- 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.

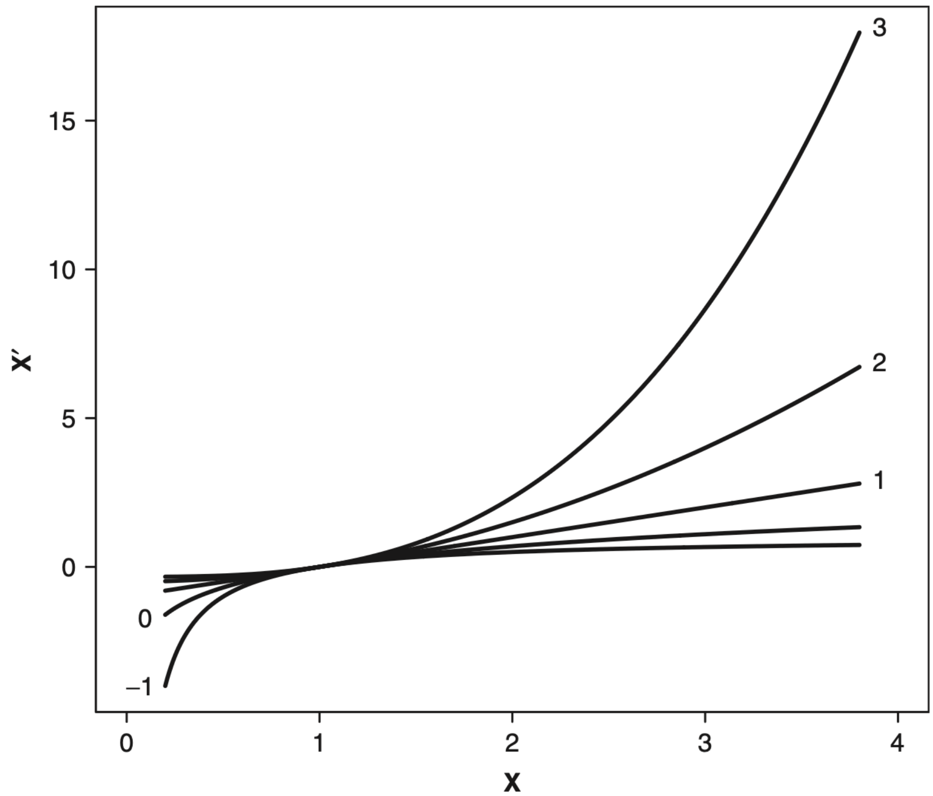

Figure 18.6: The Box-Cox family of power transformations X' of X. The curve labeled p is the transformation $X^{(p)}$, that is $(X^p - 1)/p$; $X^{(0)}$ is $\log_e(X)$. JF Figure 4.1.

- 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, $X^{(0)} \equiv \log_e X$.