Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Model Ingredients

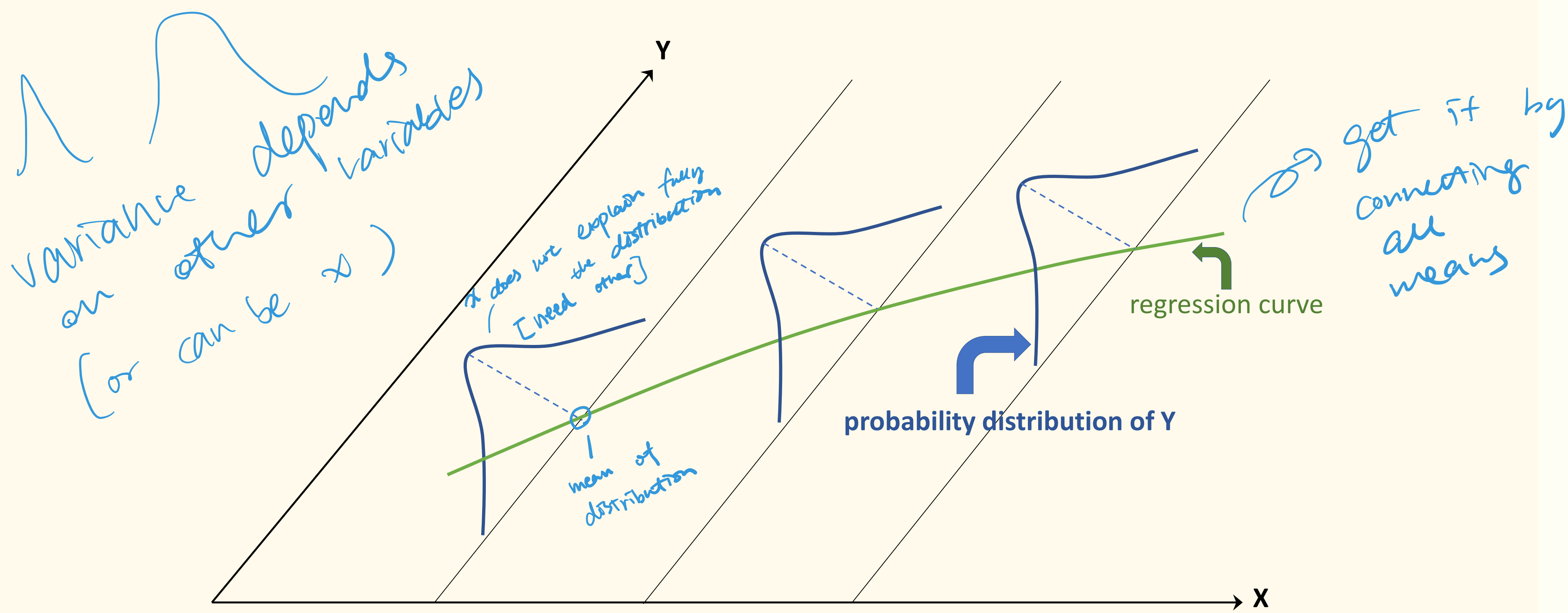
Key Ingredients

(i) Fixed component: How does the mean of the response variable change with the X value(s)?

(ii) Random component: Given the X value(s), what is the distribution of the response variable? assume normal distribution

Notes: We consider ***fixed designs*** *– X variable(s) treated as non-random.* given (for this class)

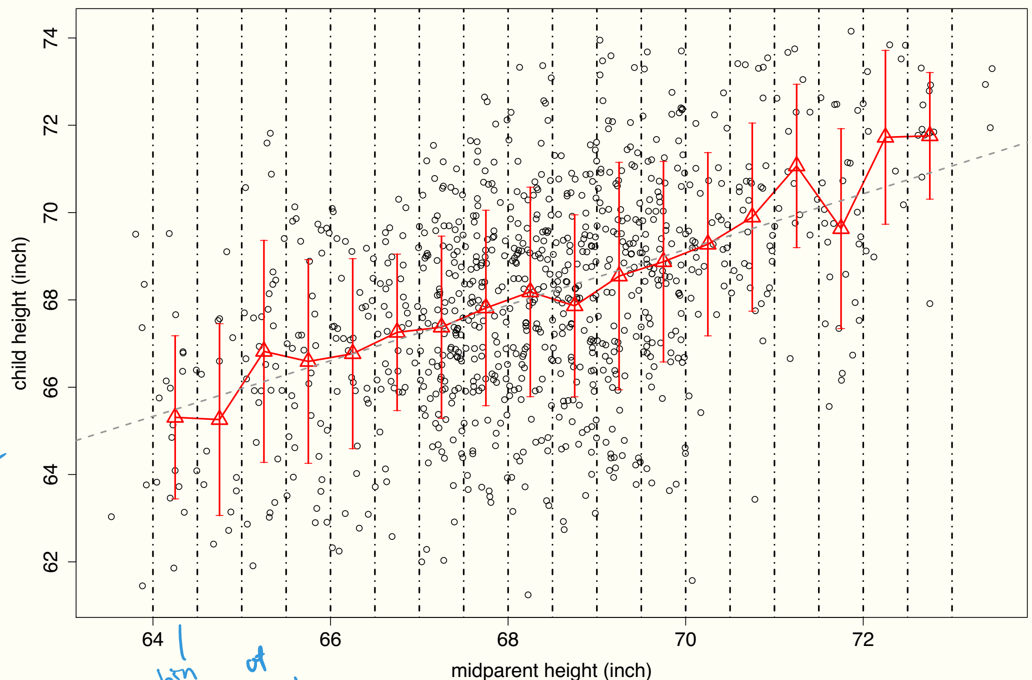
Figure: Illustration of regression model



Heights

Figure: Child's height versus midparent's height

△ = average of all in this bin



bin of 0.5 inch

connect △

↳ almost a straight line

[: width (SD) roughly equal

[degree of variation nearly same across x]

- The average child's height within each vertical strip (**bin**) lies approximately on a straight line.
- The degree of dispersion is roughly the same across bins.

(does not depend on x) for this distribution

Heights

- Model the mean of child's height (Y) as a linear function of the midparent's height (X):

$$E(Y) = \beta_0 + \beta_1 X$$

- Model the distribution of child's height as having a constant variance:

$$Var(Y) = \sigma^2$$

Simple Regression Model

Simple Linear Regression Model

Only one X variable:

n observations

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{fixed component}} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

X contribute to variance

- Y_i – value of the response variable in the *ith* case; X_i – value of the X variable in the *ith* case.

assumptions (need to be checked)

- **random errors**: ϵ_i – uncorrelated, zero-mean, equal-variance random variables

  ϵ : covariance = $\emptyset$

- **Unknown parameters**: β_0 – **regression intercept**; β_1 – **regression slope**; σ^2 – **error variance**

  ↳ not depend on x

Given X_i, the response Y_i is the sum of two terms:

- Non-random term:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

- Random error:

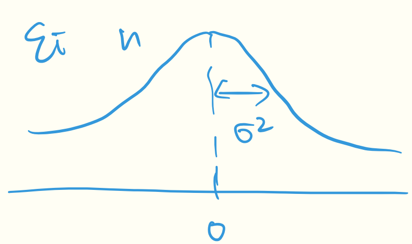
Assumption

$$\epsilon_i \sim \text{zero mean, common variance, uncorrelated}$$

$E(\varepsilon_i) = 0$

$\mathrm{var}(\varepsilon_i) = \sigma^2$

$\mathrm{Cor}(\varepsilon_i, \varepsilon_j) = 0$ $i \neq j$



does not have to be true

The simple linear regression model says:

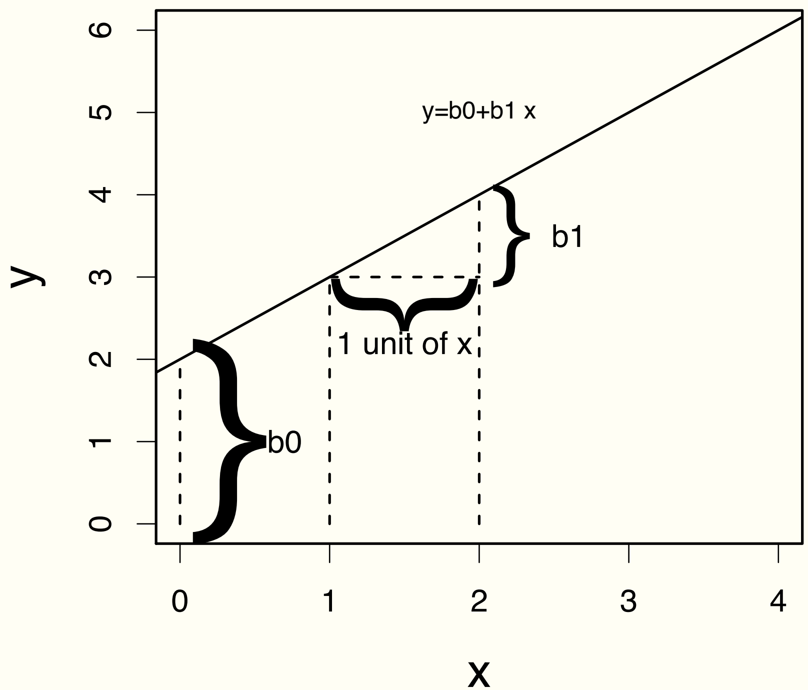
- The response variable Y_i is a random variable.
- Its mean is linearly related to X_i.
- Its variance is a constant (i.e., not depending on X_i).
- Two responses Y_i and Y_j $(i \neq j)$ are uncorrelated.

Regression Line

lower case (y) for non-random variables

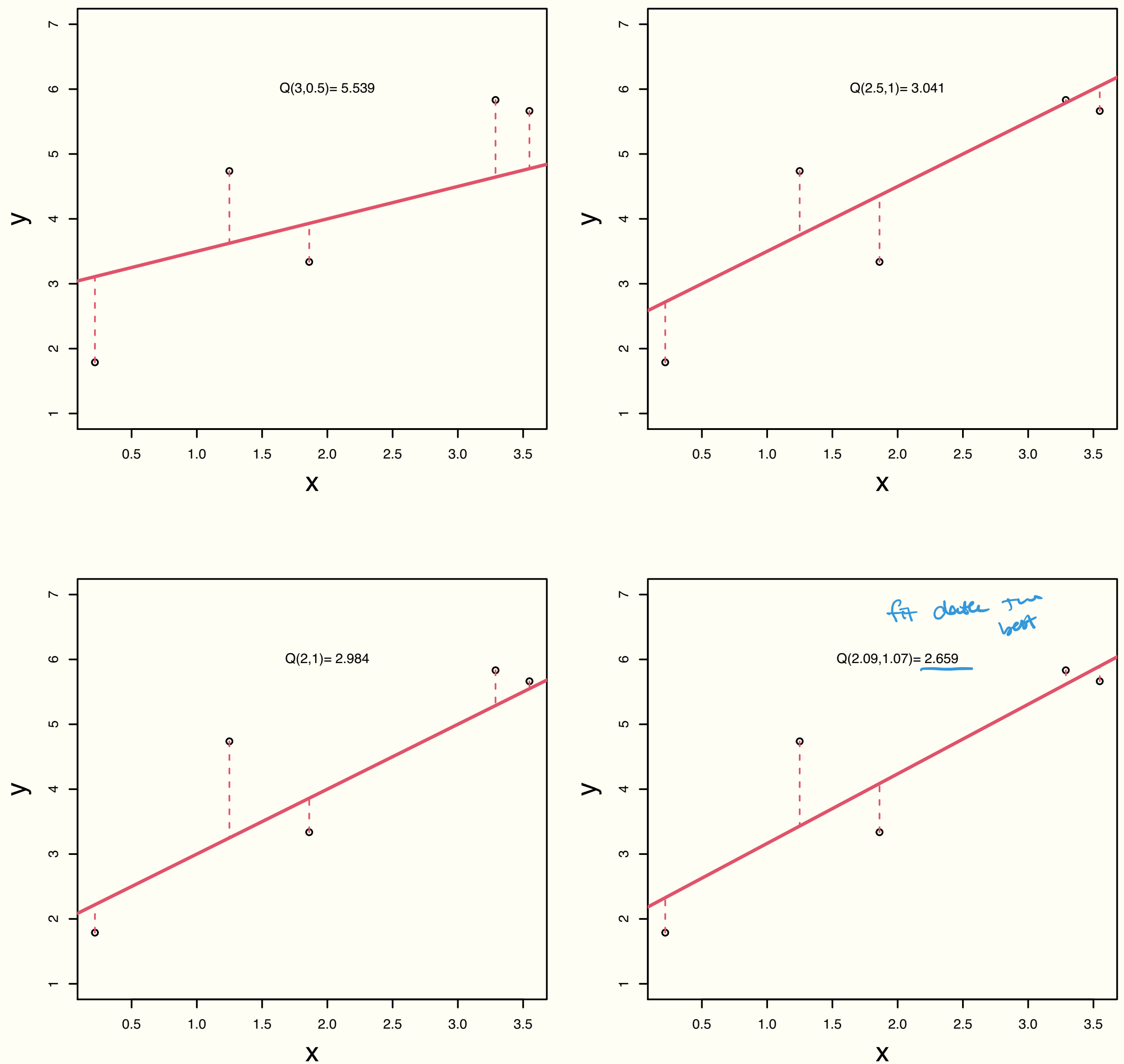
$$y = \beta_0 + \beta_1 x$$

- β_1 – regression slope: the change in mean of Y per unit change of X.
- β_0 – regression intercept: the value of $E(Y)$ when $X = 0$.



Least-Squares Estimator

Which Line is the “Best” Fit?



Least-Squares Principle

$(\hat{\beta}_0, \hat{\beta}_1) \xrightarrow{\text{LS estimate}} (\beta_0, \beta_1)$

$(\hat{\beta}_0, \hat{\beta}_1)$ = argument that min $Q(b_0, b_1)$
criterion function

The *sum of squared vertical deviations* of the observations

$\{(X_i, Y_i)\}_{i=1}^n$ from line $y = b_0 + b_1 x$:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2 .$$

- The **least squares (LS) principle** is to fit the observed data by a line that minimizes the sum of squared vertical deviations.

Least-Squares Estimator

$Q(\hat{\beta}_0, \hat{\beta}_1)$

$= \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \sum (y_i - \hat{y}_i)^2$

smallest

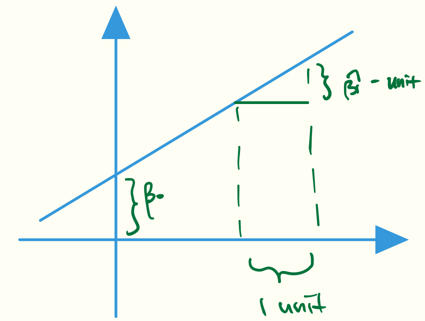
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = r_{XY}\frac{s_Y}{s_X}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, are the **sample means.**
- $s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2}$, $s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2}$, are the **sample standard deviations.**
- r_{XY} is the **sample correlation** between X and Y.
- If X_is are all equal, then LS estimator is not defined.

Least-Squares Line

change in y ≤ change in x

$|r_{xy}| \le 1$



$$y = \hat\beta_0 + \hat\beta_1 x = \overline{Y} + r_{XY}\frac{s_Y}{s_X}(x - \overline{X}).$$

- The LS line passes through the **center of the data** – $(\overline{X}, \overline{Y})$.
- If the data are **centered** (i.e., $\overline{X} = 0, \overline{Y} = 0$), then $\hat\beta_0 = 0$ and the LS line passes the origin $(0, 0)$.
- If the data are **standardized**, then $\hat\beta_0 = 0$ and $\hat\beta_1 = r_{XY}$.

  → $\bar{x} = 0$, $\bar{y} = 0$, $s_x = s_y = 1$
- **Regression effect**: One standard deviation change in X leads to r_{XY} standard deviation change in Y. (Recall $|r_{XY}| \le 1$)

  only

  in terms of SD (not absolute value)

Derive the LS Estimator

The pair (b_0, b_1) that minimizes the function $Q(\cdot, \cdot)$ satisfies:

partial derivative

$$\frac{\partial Q(b_0, b_1)}{\partial b_0} = 0, \quad \frac{\partial Q(b_0, b_1)}{\partial b_1} = 0.$$

This leads to the **normal equations**: HW1

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i \\ b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i \end{aligned}$$

The solution is the LS estimator.

Fitted Values and Residuals

Fitted Values and Residuals

- **Fitted values** are predictions by the LS line : $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1 (X_i - \overline{X}), \quad i = 1, \cdots n.$$

fitted value
for ith case

- **Residuals** are differences between the observed values and their respective fitted values:

$$\begin{aligned} e_i &= Y_i - \widehat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i), \quad i = 1, \cdots n. \\ &= (Y_i - \overline{Y}) - \hat{\beta}_1 (X_i - \overline{X}). \end{aligned}$$

residual for
ith case

centered version

Example

$\sum_i (x_i - \bar{x}) = 0$

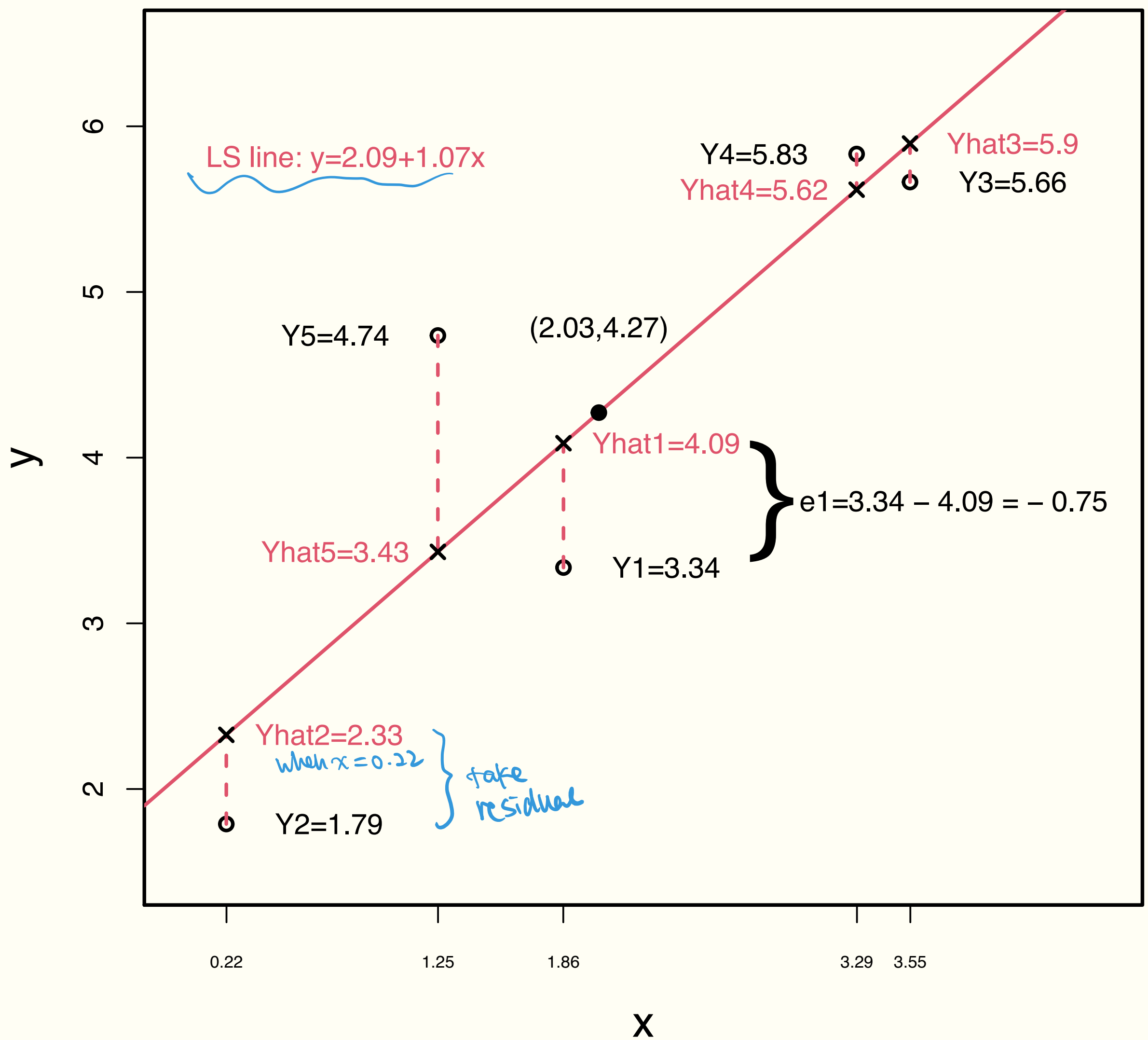
centered

Case	X_i	Y_i	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
1	1.86	3.34	-0.17	-0.94	0.03	0.16
2	0.22	1.79	-1.81	-2.48	3.29	4.50
3	3.55	5.66	1.52	1.39	2.30	2.11
4	3.29	5.83	1.26	1.56	1.58	1.96
5	1.25	4.74	-0.78	0.47	0.61	-0.36
Col. Sum	10.17	21.36	0.00	0.00	7.81	8.37
Col. Mean	2.03	4.27				
	$\bar{x}$	$\bar{y}$			$\sum (x_i - \bar{x})^2$	$\sum (x_i - \bar{x})(y_i - \bar{y})$

$$\hat{\beta}_1 = 8.37/7.81 = 1.07, \quad \hat{\beta}_0 = 4.27 - 1.07 \times 2.03 = 2.09$$

$\bar{y}$ $\hat{\beta}_1$ $\bar{x}$

Figure: LS line, fitted values and residuals



Properties of Residuals

(i) + (ii)

HW1

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

(i) $\sum_{i=1}^n e_i = 0$; (ii) $\sum_{i=1}^n X_i e_i = 0$; (iii) $\sum_{i=1}^n \widehat{Y}_i e_i = 0$

$\sum \hat{Y}_i e_i = \sum \hat{\beta}_0 e_i + \sum \hat{\beta}_1 x_i e_i$

(i) 0 (ii) 0

constant

$\downarrow - \downarrow = \downarrow$

Case	X_i	Y_i	$\widehat{Y}_i$	e_i
1	1.86	3.34	4.09	-0.75
2	0.22	1.79	2.33	-0.54
3	3.55	5.66	5.90	-0.23
4	3.29	5.83	5.62	0.22
5	1.25	4.74	3.43	1.31

Mean Squared Error

Estimation of Error Variance

random error

- Error variance $\sigma^2 = \mathrm{Var}(\epsilon_i)$. (Recall $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$)
  not observed!
- Idea: Estimate σ^2 by the "variance" of residuals. (Recall residual $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$)
  calculated

  $(e_1, \cdots e_n)$

  $\sum e_i = 0 \Rightarrow \bar{e} = 0$
- **Error sum of squares (SSE)**:

obtained smalles value for LS estimator @ $\hat{\beta}_0, \hat{\beta}_1$

$$SSE := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

$\sum_{i=1}^{n} (e_i - \bar{e})^2 = \sum e_i^2$

$Q(\hat{\beta}_0, \hat{\beta}_1)$ = smallest valle of LS estimator function $Q(b_0, b_1)$

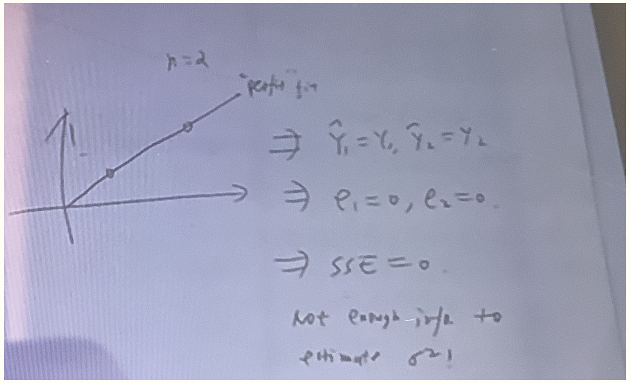
- **Mean squared error (MSE)**:

$$MSE = \frac{SSE}{n-2}$$

only n-2 residuals are free

(iii) not independent
↳ does not count

Degrees of Freedom



properties (1) & (2) ⇒ 2 constraints

[although n res] only n-2 are free ↳ know these 2, know rest (not free anymore)

- The **degrees of freedom** of a random vector is the number of its components that are free to vary.

- Recall $\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} X_i e_i = 0 \rightarrow$ degrees of freedom of $(e_1, \cdots, e_n)$ is $n-2$.

$MSE = \frac{SSE}{n-2}, \quad E(MSE) = \frac{E(SSE)}{n-2} = \sigma^2$

- $d.f.(SSE) = n - 2$.

- $E(SSE) = (n-2)\sigma^2$ and thus $E(MSE) = \sigma^2 \rightarrow$ MSE is an **unbiased estimator** of σ^2.

no systematic error → centered around estimator it tries to estimate

$E(\text{param}) = \text{param}$
↳ unbiased estimator

Example (Cont'd)

^2

Case	X_i	Y_i	$\widehat{Y}_i$	e_i
1	1.86	3.34	4.09	-0.75
2	0.22	1.79	2.33	-0.54
3	3.55	5.66	5.90	-0.23
4	3.29	5.83	5.62	0.22
5	1.25	4.74	3.43	1.31

$$SSE = (-0.75)^2 + (-0.54)^2 + (-0.23)^2 + 0.22^2 + 1.31^2 = 2.6715$$

$$MSE = \frac{2.6715}{5-2} = 0.8905.$$

→ estimate of σ^2

LS Estimator: Properties

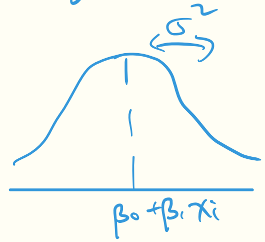
Mean and Variance

$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

r.v. r.v. r.v.

- **LS estimators are unbiased:**

$y_i \sim$ random var.



$\hat{\beta}_0 \sim$

$\hat{\beta}_1 \sim$

$\mathrm{cov}(y_i, y_j) = 0$

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$$

- Variance of $\hat{\beta}_0, \hat{\beta}_1$:

$$\mathrm{var}(\hat{\beta}_0) = \sigma^2\{\hat{\beta}_0\} = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right]$$

$$\mathrm{var}(\hat{\beta}_1) = \sigma^2\{\hat{\beta}_1\} = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

estimated sd. , sd is $\sqrt{\text{var}}$

Standard Error (SE)

① ↑ sample var. of x
x spread out, SE ↓

Replace σ^2 by *MSE* and take square-root:

$$SE(\hat{\beta}_0) = s\{\hat{\beta}_0\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\underbrace{\sum_{i=1}^n (X_i - \overline{X})^2}_{= (n-1)\cdot S_x^2}}\right]}$$

$$SE(\hat{\beta}_1) = s\{\hat{\beta}_1\} = \sqrt{\frac{MSE}{\underbrace{\sum_{i=1}^n (X_i - \overline{X})^2}_{= (n-1)\cdot S_x^2}}}$$

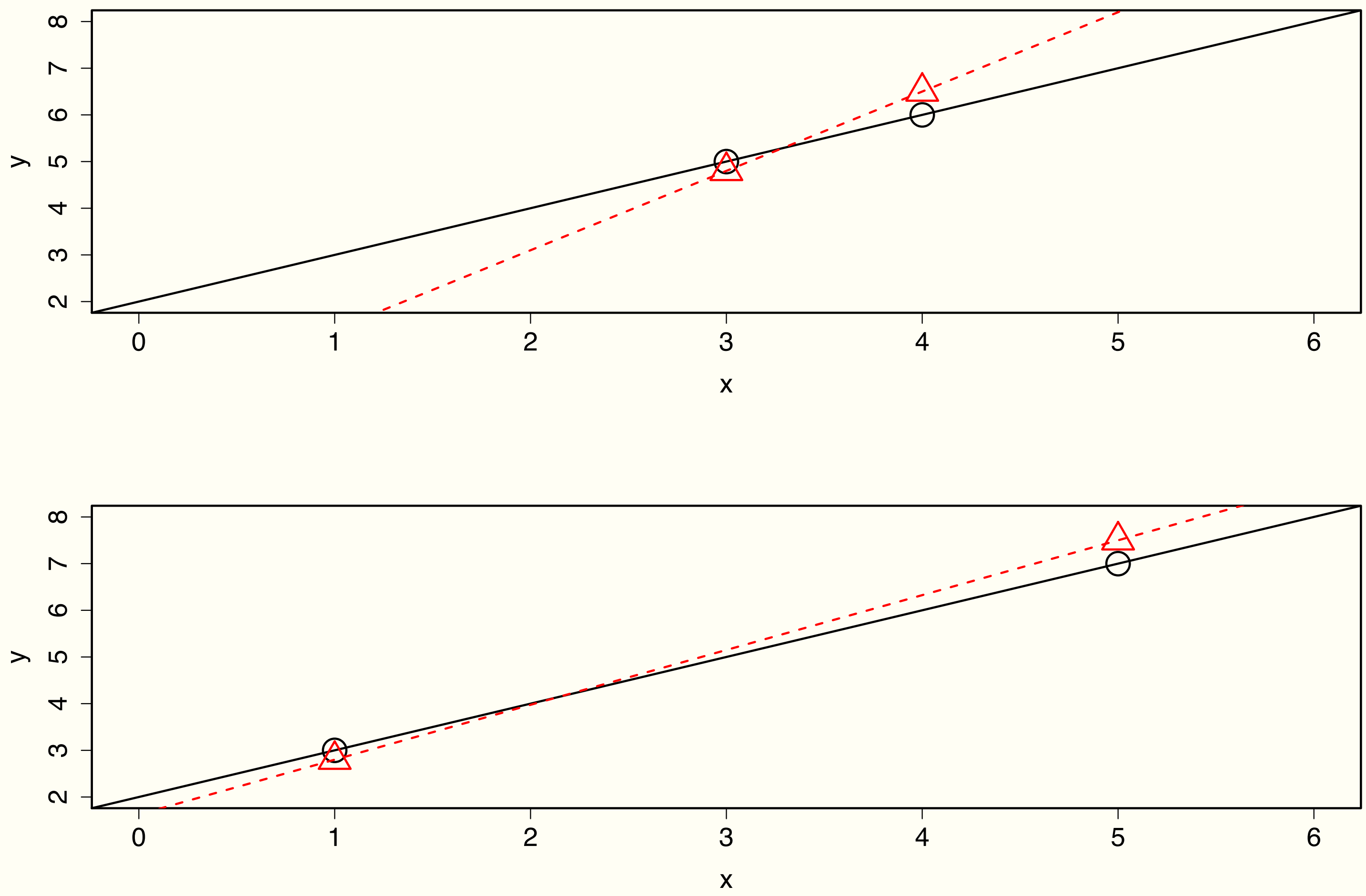
if n↑, ∵ ① @ denominator
② ↑ spread out,
standard error ↓↓

- SE decreases with the increase of the sample size n or the sample variance s_X^2. (Recall $\sum_{i=1}^n (X_i - \overline{X})^2 = (n-1)s_X^2$)
- SE tends to increase with the increase of the error variance σ^2.

amount fluctuation in data, correspond ↑ MSE
↑↑↑ fluc, ↓ reliable

Figure: Effects of the dispersion of X on the sampling variability of the LS line

smaller spread
↳ larger variability



same fluctuation in y.

but effect larger in ①

∵ x closer

Simulation Experiment

Simulation

- $n = 5$ cases with the X values

$$X_1 = 1.86, \quad X_2 = 0.22, \quad X_3 = 3.55, \quad X_4 = 3.29, \quad X_5 = 1.25,$$

  fixed throughout.
- The responses:
  - First generate $\epsilon_1, \cdots, \epsilon_5$ i.i.d. from $N(0, 1)$.
  - Then set the response variable as:

fixed

$\beta_0 = 2$ $\beta_1 = 1$

$\sigma^2 = 1$

$$Y_i = 2 + X_i + \underbrace{\epsilon_i}_{\text{random}}, \quad i = 1, \cdots, 5.$$

- Repeat 100 times → 100 data sets.

data set 1

~~case X Y~~

	x	y
1	1.86	3.08
2	0.22	2.27
3	3.55	4.38
4	3.29	5.12
5	1.25	1.38

$\hat{\beta}_0 = 1.34,\ \hat{\beta}_1 = 0.94,\ MSE = 0.79.$

fixed x (design)

$\cdots, \cdots$

data set 100

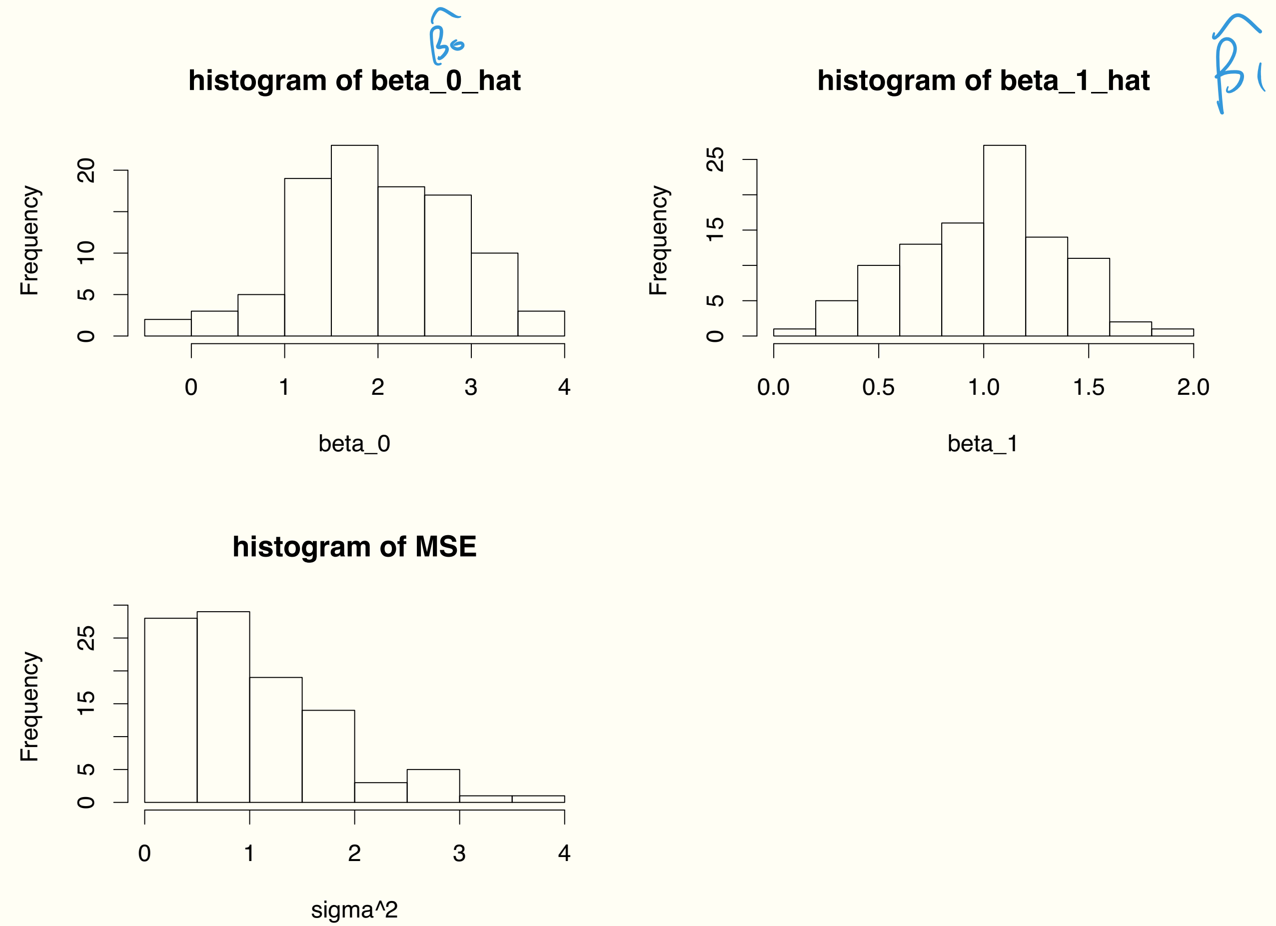
~~case X Y~~

	x	y
1	1.86	3.36
2	0.22	2.50
3	3.55	5.93
4	3.29	5.36
5	1.25	2.67

Sampling variability

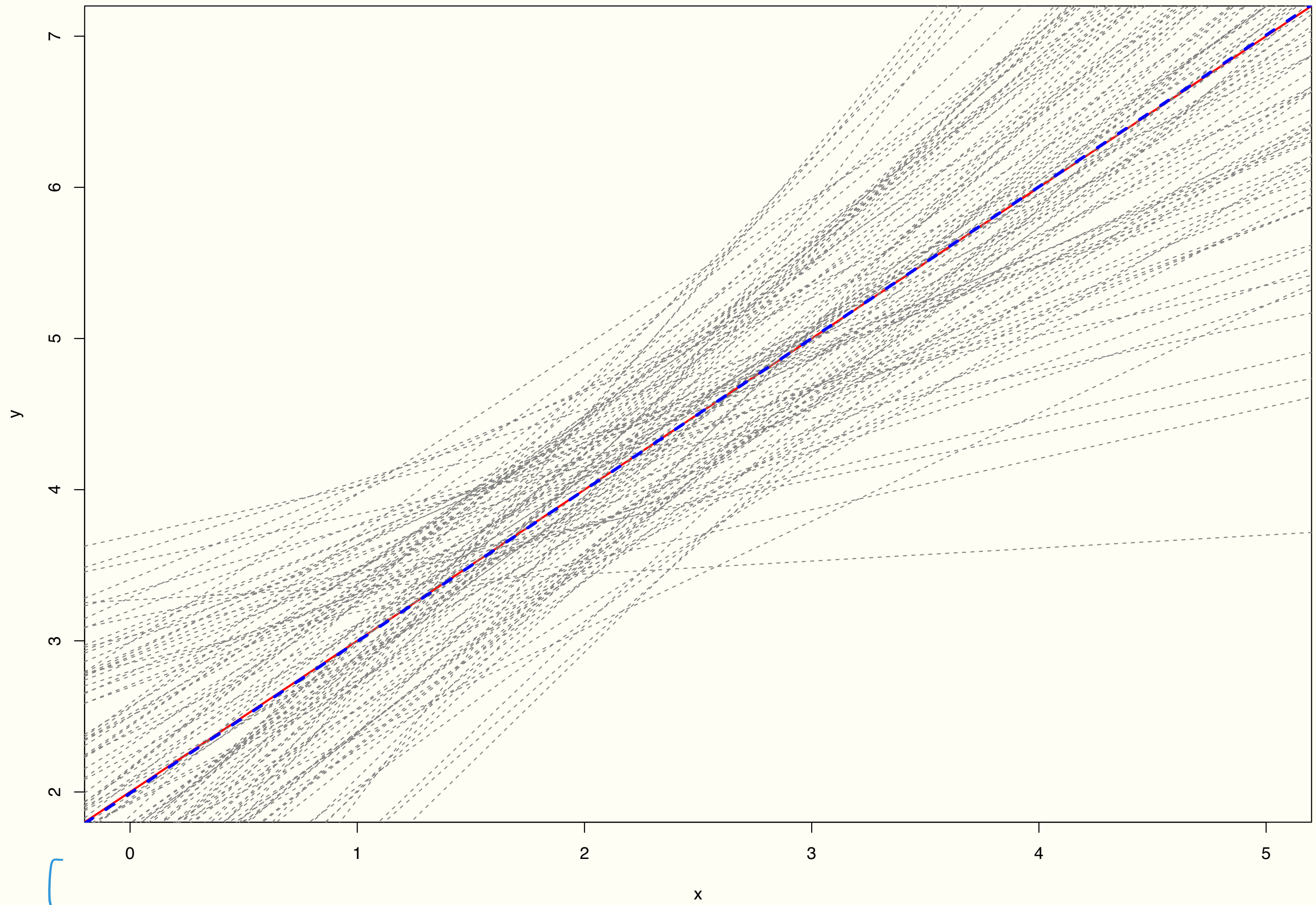
$\hat{\beta}_0 = 1.75,\ \hat{\beta}_1 = 1.09,\ MSE = 0.24.$

Figure: Sampling distributions of $\hat{\beta}_0, \hat{\beta}_1$ and MSE



Sample means are $1.99, 1.02, 1.04$, respectively. True parameters are $2, 1, 1$, respectively.

Figure: True: red solid; LS lines: grey broken; mean LS line: blue broken



almost same → unbiased

Compare sample mean and sample standard deviation of these 100 realizations of $\hat{\beta}_0, \hat{\beta}_1$ to the respective theoretical values.

- $\hat{\beta}_0$: Theoretical mean and standard deviation:

$$E(\hat{\beta}_0) = \beta_0 = 2, \quad \sigma\{\hat{\beta}_0\} = \sqrt{\sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]} = 0.854$$

Sample mean and sample standard deviation: $1.99, 0.847$.

very close

- $\hat{\beta}_1$: Theoretical mean and standard deviation:

$$E(\hat{\beta}_1) = \beta_1 = 1, \quad \sigma\{\hat{\beta}_0\} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}} = 0.358$$

Sample mean and sample standard deviation: $1.002, 0.36$.

$$E(\hat{\beta}_1) = E\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left(\sum_i (x_i-\bar{x})Y_i\right)$$

constant ∵ fixed design

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} E(\underbrace{(x_i-\bar{x})}_{\text{non random (fixed) ∴ pull out}} Y_i) = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i-\bar{x}) \cdot E(Y_i)$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i-\bar{x}) \cdot [\beta_0 + \beta_1 x_i]$$

omit $\bar{Y}$ (does not depend on i)

$$= \underbrace{\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i-\bar{x})\beta_0}_{= 0} + \underbrace{\frac{1}{S_{xx}} \sum (x_i-\bar{x}) x_i \cdot \beta_1}$$

(b.c. $\sum(x_i-\bar{x}) = 0$)

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

$$\beta_1 \boxed{\frac{1}{S_{xx}} \sum(x_i-\bar{x})x_i}$$

$$\|$$

$$\beta_1 \cdot \frac{\sum(x_i-\bar{x})(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

S_{xx}

S_{xx}

check

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}$$

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $E(Y_i) = \beta_0 + \beta_1 x_i$, $i = 1 \dots n$

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

- Simple linear regression model:

obs. → $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \dots, n$

random error → $\varepsilon_i \overset{\text{uncorrelated}}{\sim} (0, \sigma^2)$ ← assumptions on ε_i's

↑ common variance

- L-S estimator:

(Hw1) ← $\begin{cases} \hat{\beta}_1 = \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \dfrac{\sum (X_i - \bar{X}) Y_i}{S_{xx}} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$, $\boxed{S_{xx} = \sum (X_i - \bar{X})^2}$

- $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_0) = \beta_0$ ⟶ these are unbiased estimators

↑ Hw 1

non-random

- $\boxed{Var(\hat{\beta}_1)} = Var\left(\dfrac{\sum (X_i - \bar{X}) Y_i}{S_{xx}} \right) = \sum_{i=1}^{n} Var\left(\dfrac{X_i - \bar{X}}{S_{xx}} Y_i \right) = \sum_{i=1}^{n} \dfrac{(X_i - \bar{X})^2}{S_{xx}^2} \underbrace{Var(Y_i)}_{\sigma^2}$

↑ because Y_i's are uncorrelated

$= \sum_{i=1}^{n} \dfrac{(X_i - \bar{X})^2}{S_{xx}^2} \cdot \sigma^2 = \dfrac{\sigma^2}{S_{xx}^2} \underbrace{\sum (X_i - \bar{X})^2}_{S_{xx}} = \dfrac{\sigma^2}{S_{xx}^2} \cdot S_{xx} = \dfrac{\sigma^2}{S_{xx}} = \boxed{\dfrac{\sigma^2}{\sum (X_i - \bar{X})^2}}$

↓

$\sigma^2 \{ \hat{\beta}_1 \}$

$\hat{\beta}_1 \sim$