Linear Regression

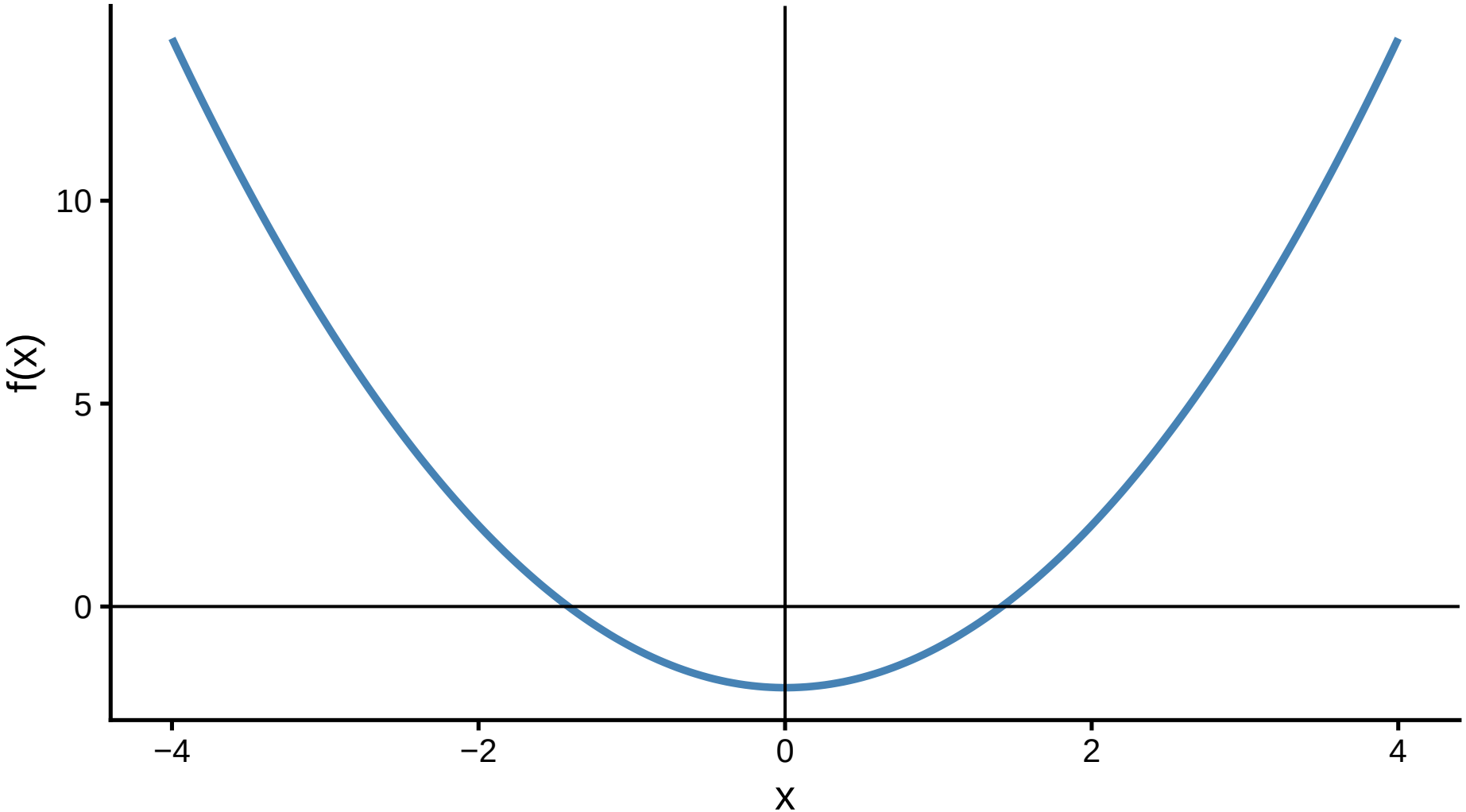
GENOME 560, Spring 2026

Saori Sakaue (sakaue@uw.edu)

Section 1

Recap

Recap on Algebra



Recap on Algebra: Optimization

Derivatives are crucial for finding the minimum or maximum of functions.

- At a minimum or maximum, $\frac{df}{dx} = 0$
- Second derivative test: $\frac{d^2f}{dx^2} > 0 \rightarrow$ minimum; $< 0 \rightarrow$ maximum

Partial Derivatives

For a function of multiple variables $f(x_1, x_2, ...)$:

$$\frac{\partial f}{\partial x_1}, \quad \frac{\partial f}{\partial x_2}, \quad ...$$

Each partial derivative treats all other variables as constants.

Section 2

Linear Regression

What is Linear Regression?

- Technique used for the modeling and analysis of numerical data
- Used to predict quantitative output values
- Leverages the **relationship between two or more variables** so that we can gain information about one of them through knowing values of the other

Why Linear Regression?

- Suppose we want to model the outcome variable Y in terms of three predictors, X_1, X_2, X_3

$$Y = f(X_1, X_2, X_3)$$

- Typically we will not have enough data to directly estimate f
- Therefore, we usually have to assume that it has some approximated restricted form, such as **linear**

$$Y = aX_1 + bX_2 + cX_3$$

Regression Terminology

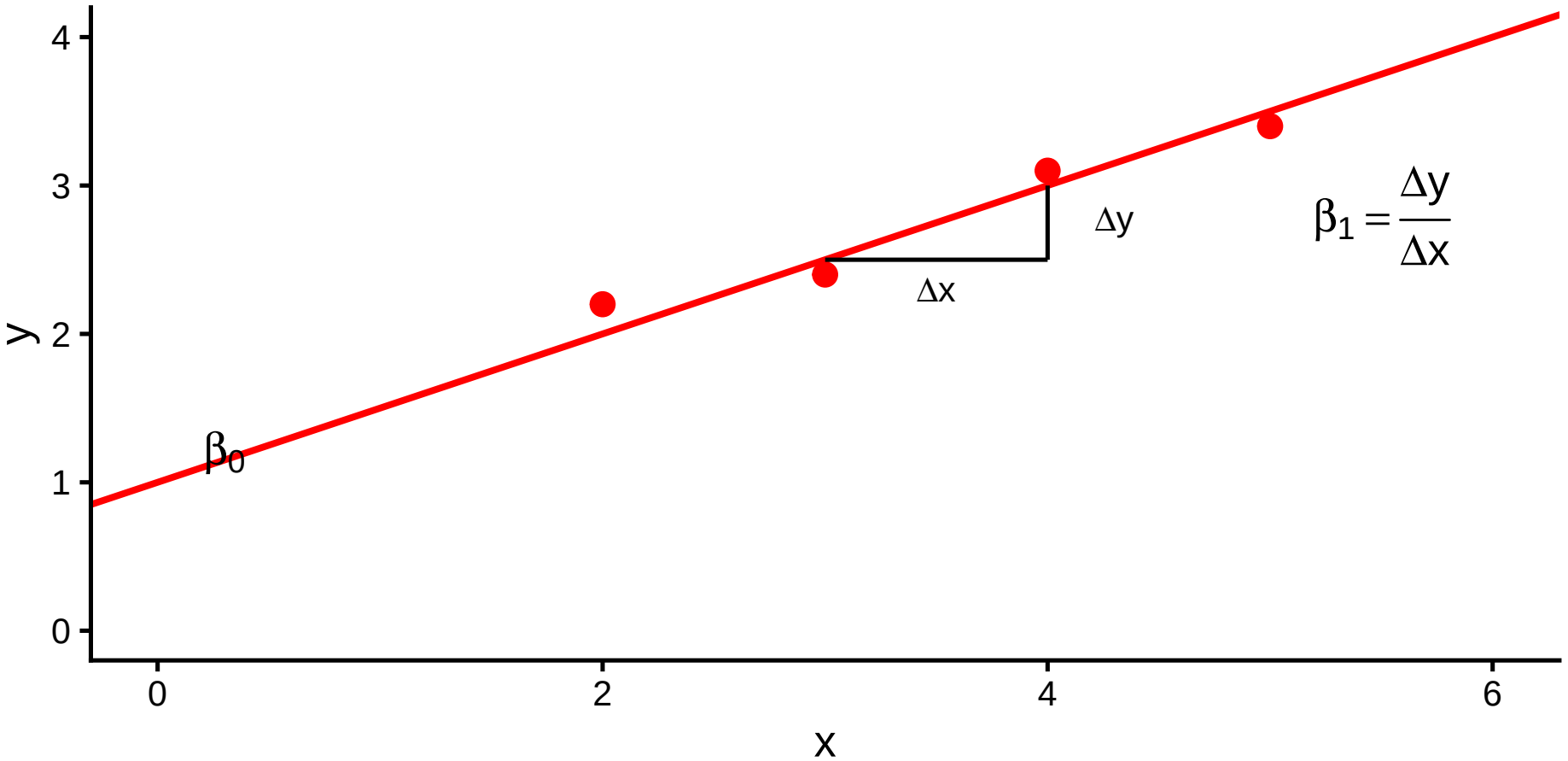
- **Dependent variable** / outcome variable / response variable
- **Independent variable** / predictor variable / explanatory variable

Examples:

- Lung cancer risk $\sim$ Genetic factor + smoking + diet + ...
- Expression level of gene Y $\sim$ Expression levels of gene Y's TFs: A, B, C

Probabilistic vs. Deterministic Models

When we do linear regression, we're interested in understanding the relationship between variables related in a **nondeterministic** fashion.



A Linear Model

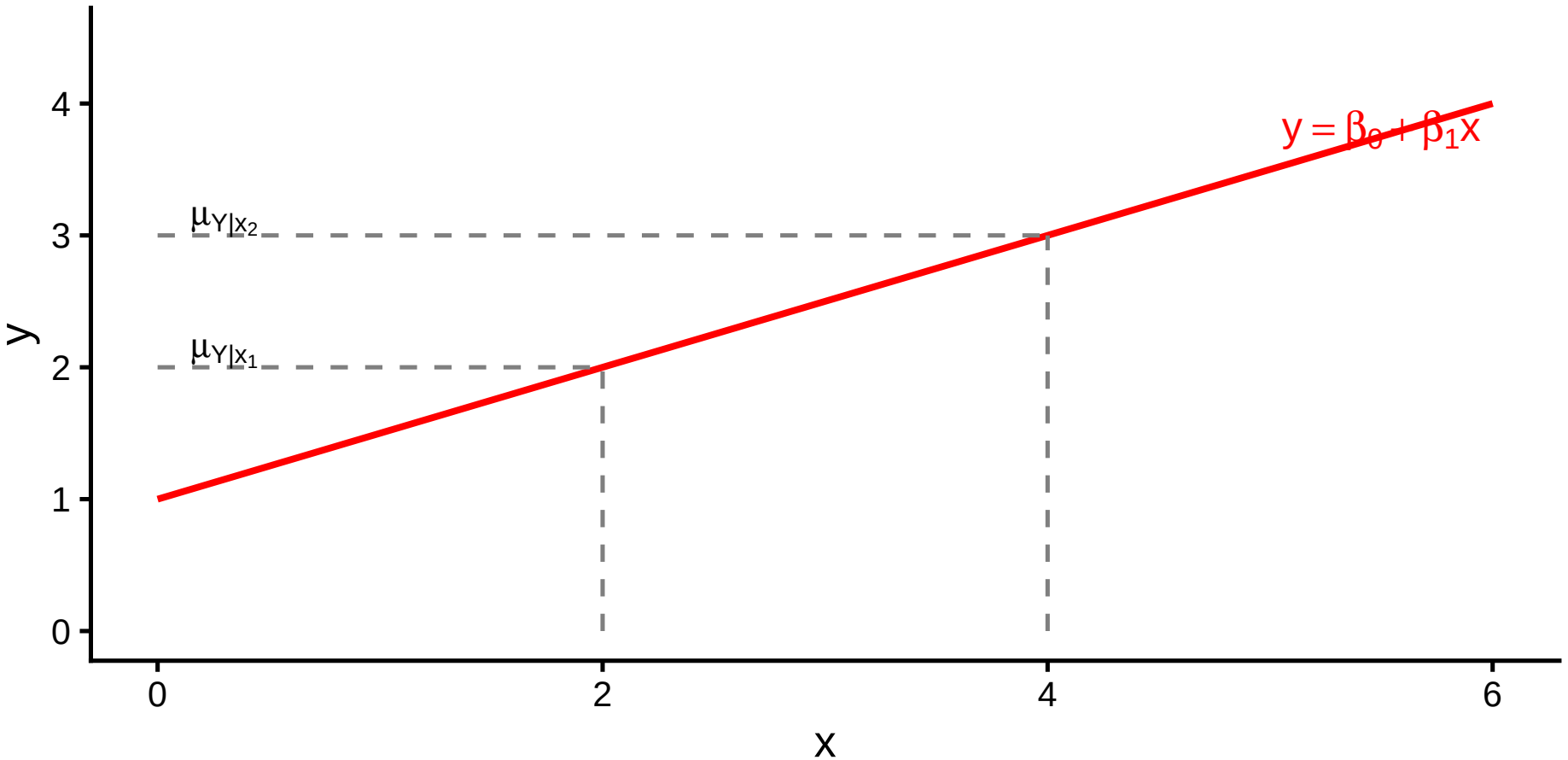
$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- β_0 = y-intercept
- β_1 = slope
- ϵ = random error term (noise)

This is the **true regression line** (population level).

Implications



Graphical Interpretation

Example: $X =$ height and $Y =$ weight.

Then $\mu_{Y|x=60}$ is the average weight for all individuals 60 inches tall in the population.

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

What's the Best Solution?

What's the best solution to find the best fit linear model?

Minimize the distance between the observed data and the model predictions!

Section 3

Least Squares Regression

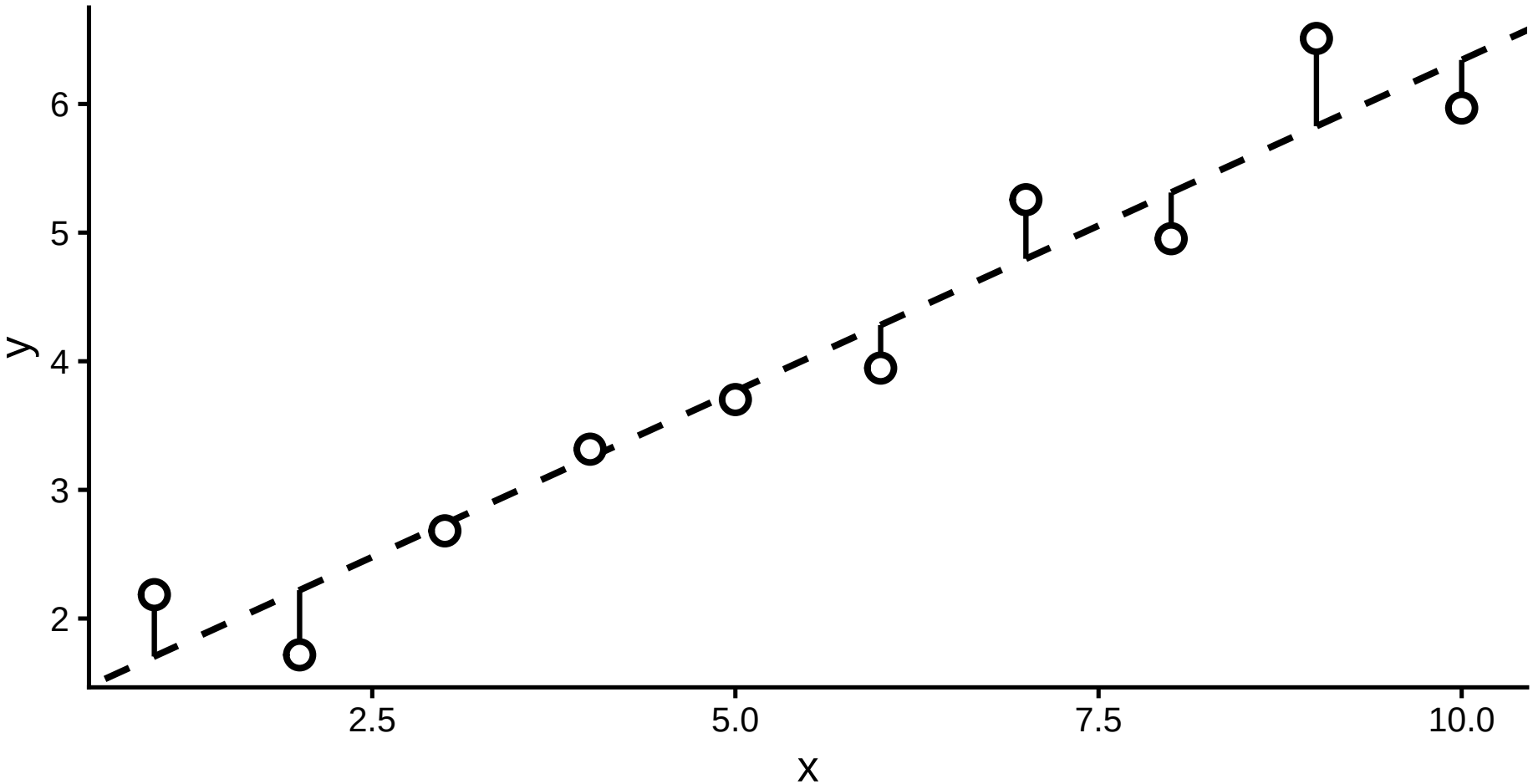
Some Regression Terminology

For a sample of n observations $(x_1, y_1), \ldots, (x_n, y_n)$:

- **Fitted values:** $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$
- **Residuals:** $e_i = y_i - \widehat{y}_i$

Residuals Are Useful!

Residuals = vertical distances from points to line



Least Squares Regression

- The **sum of squared errors** (SSE) tells us how well the line fits the data:

$$\text{SSE} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **Least squares regression:** Find β_0 and β_1 that **minimize** SSE

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

- Denote the solutions by $\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{\beta_0, \beta_1} f(\beta_0, \beta_1)$

Let's Derive Least Square Solutions

Take partial derivatives and set to zero:

$$\frac{\partial f}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial f}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Solutions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Three Key Assumptions in Linear Regression

1. **Linearity:** Observations are appropriately modeled by a linear equation
2. **Homoscedasticity:** The variance of residuals is independent of the predicted value (constant variance)
3. **Independence:** The residuals are independent of each other

$$\epsilon_i \sim N(0, \sigma^2) \quad \text{i.i.d.}$$

Put in Other Terminology...

- **Linearity:** $E[Y|X] = \beta_0 + \beta_1 X$
- **Homoscedasticity:** $\text{Var}(\epsilon_i) = \sigma^2$ for all i (the variance of residuals is the same for any value of X)
- **Independence:** ϵ_i and ϵ_j are independent for $i \neq j$
- **Normality:** $\epsilon_i \sim N(0, \sigma^2)$

Variance of $\hat{\beta}$ and Hypothesis Testing

Under the assumptions:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Hypothesis test: $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

Section 4

Regression, Correlation, and R^2

Relationship Among Regression, Correlation, and R^2

Decomposing Variance

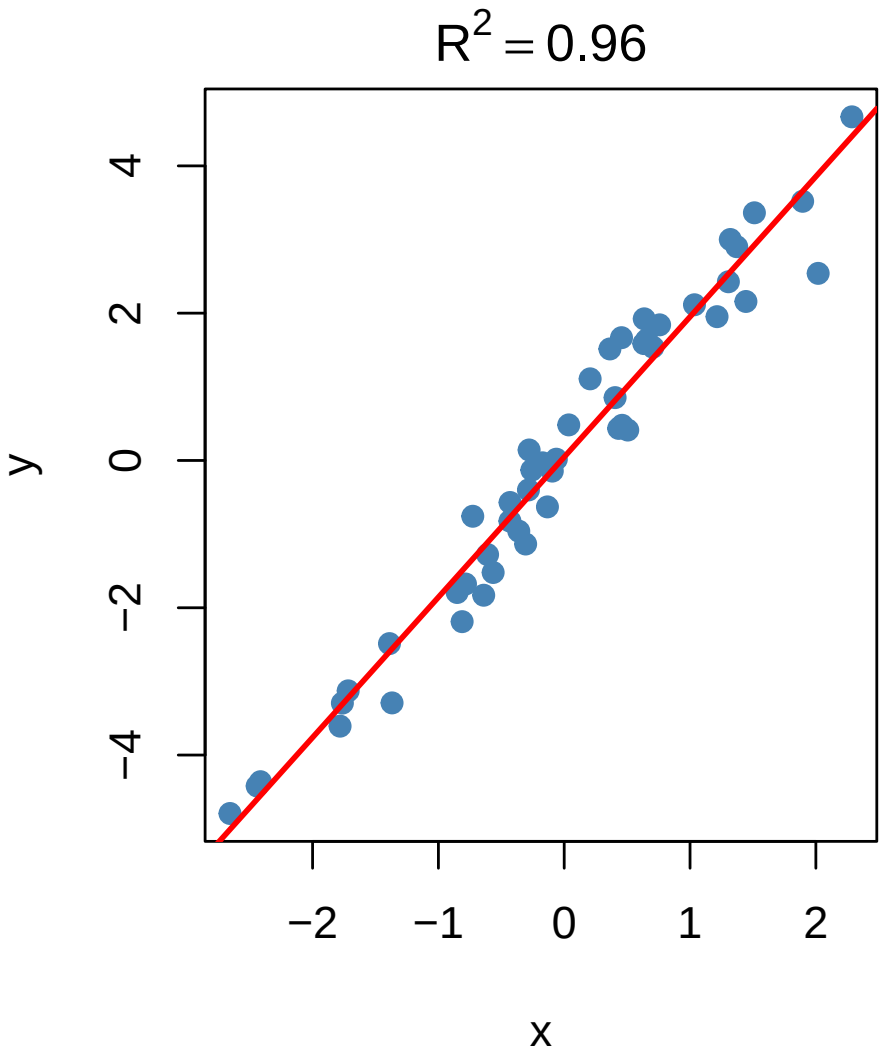
$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST (total)}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR (explained by } x)} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE (not explained)}}$$

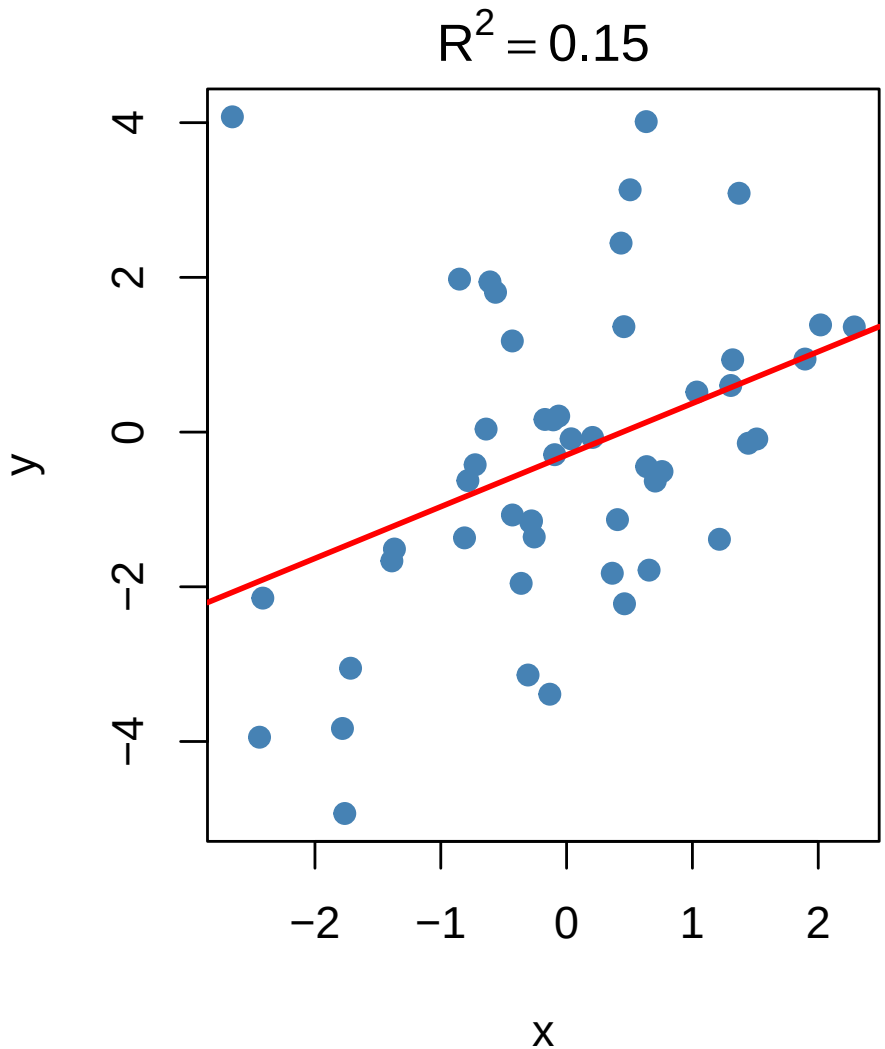
Coefficient of Determination R^2

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

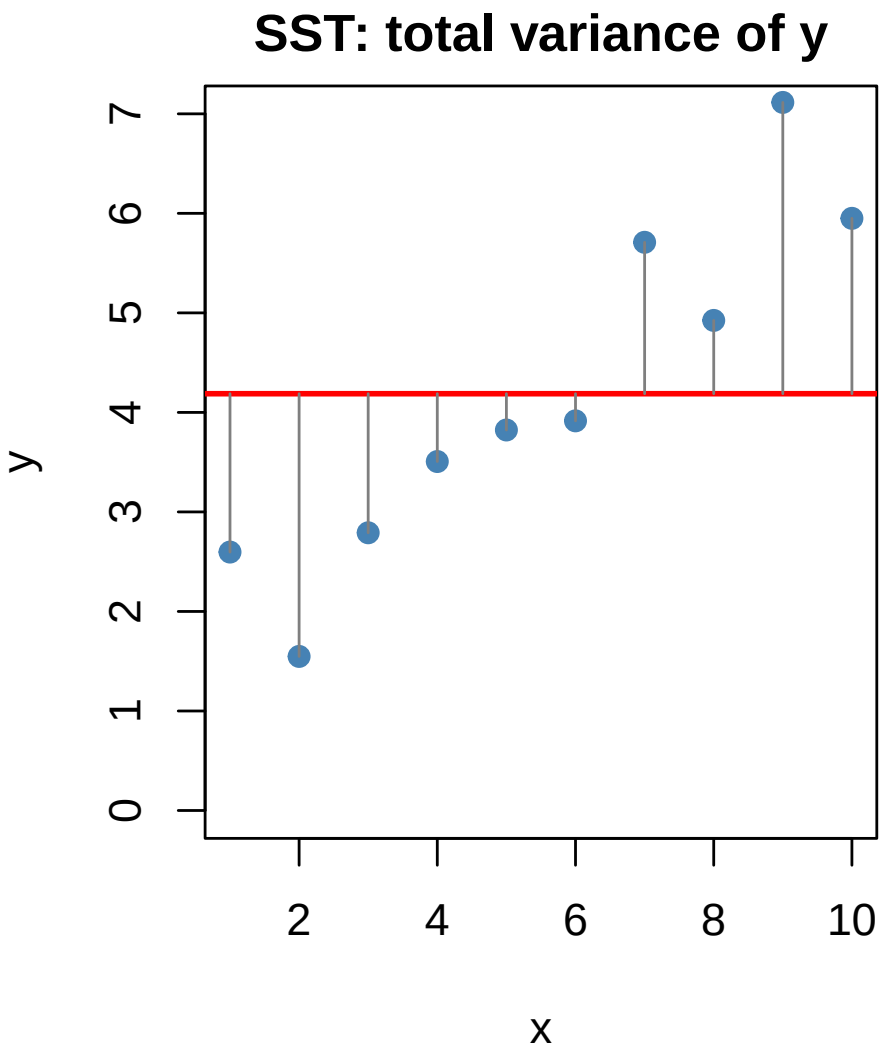
- R^2 measures the **proportion of variance in Y explained by X**
- $0 \leq R^2 \leq 1$
- $R^2 = 1$: perfect fit
- $R^2 = 0$: X explains none of the variance in Y

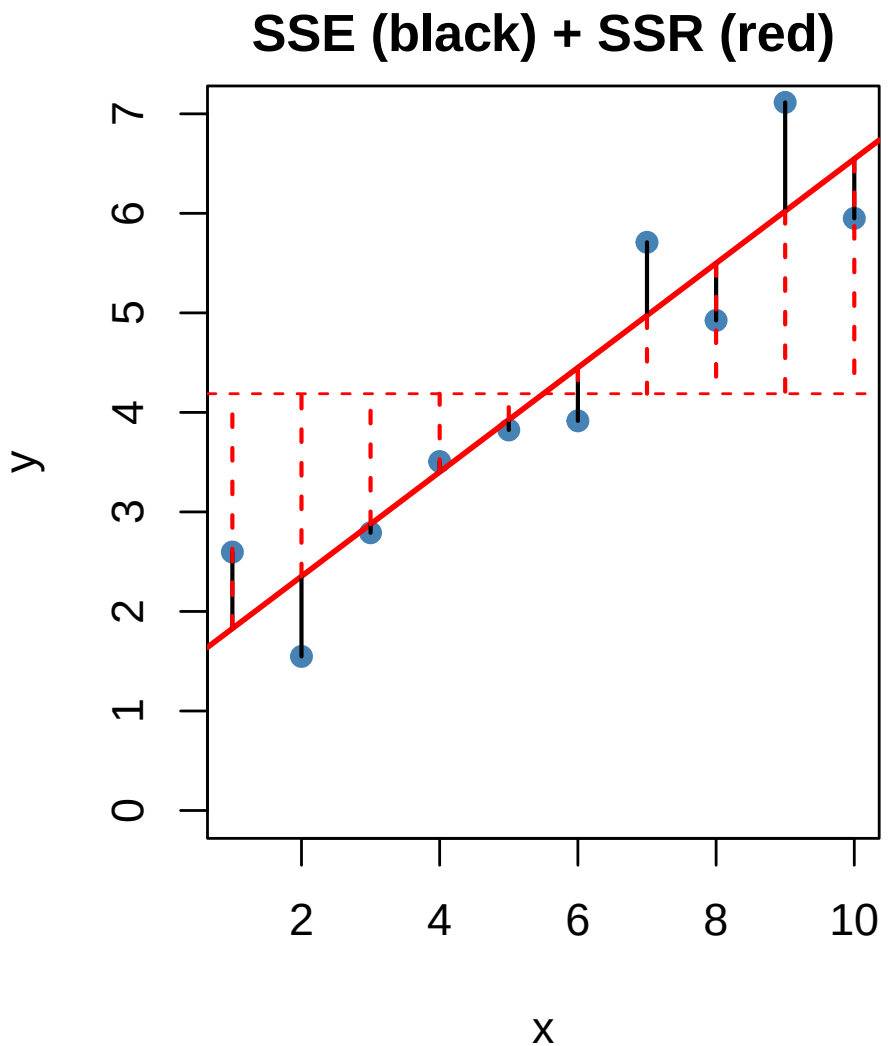
Visualizing R^2





Decomposing Variance — Visually





Correlation Coefficient r

Used to describe the degree of **sample correlation**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $-1 \leq r \leq 1$
- $r > 0$: positive correlation
- $r < 0$: negative correlation
- $r = 0$: no linear correlation
- Note: $R^2 = r^2$ for simple linear regression