Linear Models and Regression Analysis

Chapter 1

Simple Linear Regression

1.

Probabilistic Models

Models

- Representation of some phenomenon
- Mathematical model is a mathematical expression of some phenomenon
- Often describe relationships between variables
- Types
  - Deterministic models (Functional relation)
  - Probabilistic models (Statistical relation)

Deterministic Models

- Hypothesize exact relationships
- Suitable when prediction error is negligible
- Example: force is exactly mass times acceleration

$$F = m \cdot a$$

Probabilistic Models

- Hypothesize two components
  - Deterministic part
  - Random error
- Example: sales volume (y) is 10 times advertising spending (x) + random error
  - $y = 10x + \varepsilon$
  - Random error may be due to factors other than advertising

General Form of Probabilistic Models

y = Deterministic component + Random error

where y is the variable of interest. We always assume that the mean value of the random error equals 0. This is equivalent to assuming that the mean value of y, $E(y)$, equals the deterministic component of the model; that is,

$$E(y) = \text{Deterministic component}$$

A First-Order (Straight Line) Probabilistic Model or Simple Linear Regression

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

Where

$i = 1, \ldots, n$ is the i^{th} observation in the sample

y = **Dependent** *or* **response variable**
(variable to be modeled)

x = **Independent** *or* **predictor variable**
(variable used as a predictor of y)

$E(y) = \beta_0 + \beta_1 x$ = Deterministic component

ε (epsilon) = Random error component

A First-Order (Straight Line) Probabilistic Model

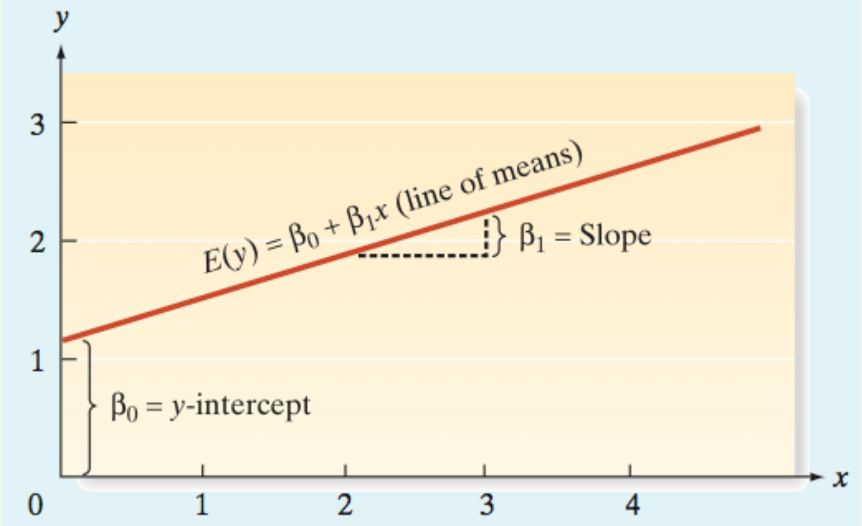
$$y = \beta_0 + \beta_1 x + \varepsilon$$

β_0 (beta zero) = ***y*-intercept of the line**, that is, the point at which the line *intercepts or cuts through the y-axis*

β_1 (beta one) = **slope of the line**, that is, the change (amount of increase or decrease) in the deterministic component of *y* for every 1-unit increase in *x*

A First-Order (Straight Line) Probabilistic Model

Note: A *positive* slope implies that $E(y)$ *increases* by the amount β_1 for each unit increase in x. A *negative* slope implies that $E(y)$ *decreases* by the amount β_1.

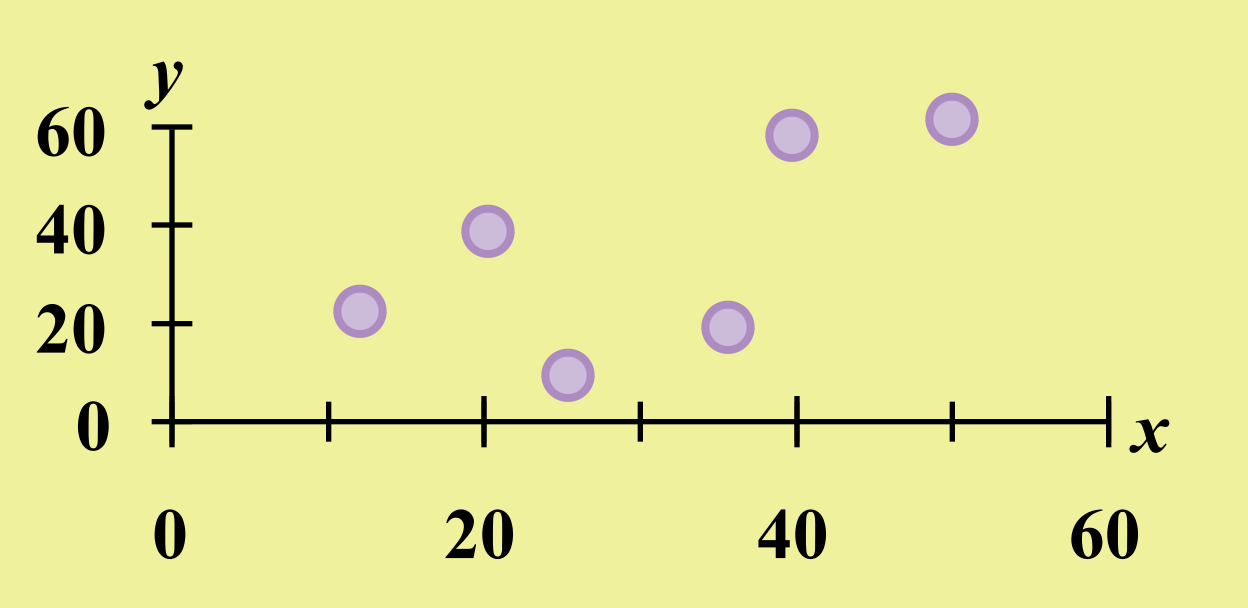


2.

Fitting the Model:
The Least Squares Approach

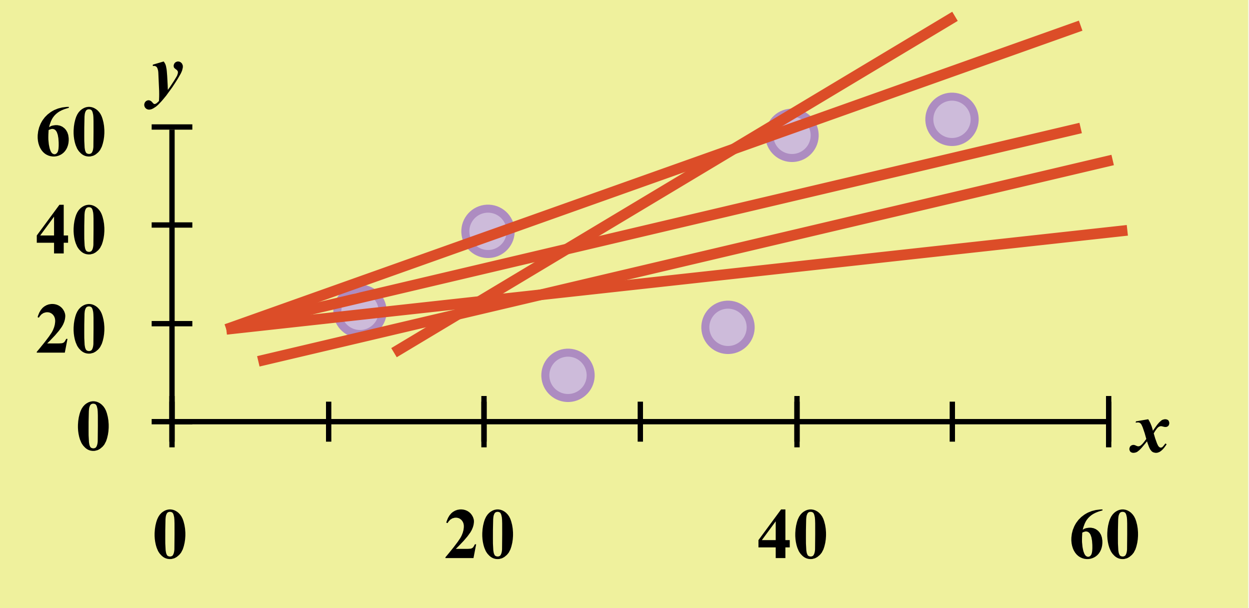
Scatterplot

1. Plot of all (x_i, y_i) pairs
2. Suggests how well model will fit



Thinking Challenge

- How would you draw a line through the points?
- How do you determine which line 'fits best'?



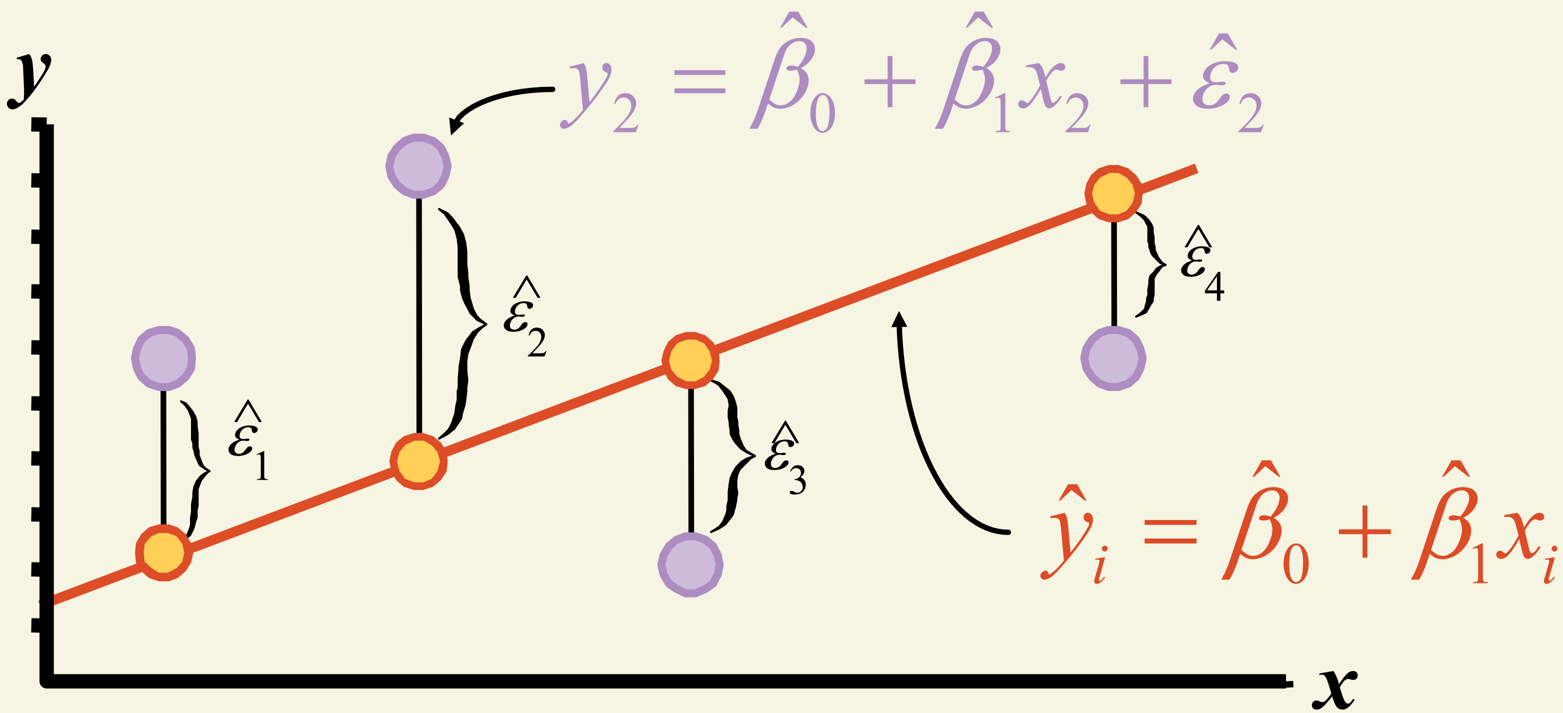
Least Squares Line

The **least squares line** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is one that has the following property:

The sum of squared errors (SSE) is smaller than for any other straight-line model, i.e., the error variance is minimum.

Least Squares Graphically

$$\text{LS minimizes } \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$$



Normal Equations (1.9)

$$\sum Y_i = nb_0 + b_1 \sum X_i$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

To derive b_0 divide first equation by n:

$$\sum Y_i /n = nb_0/n + b_1 \sum X_i/n$$

$$\Leftrightarrow \bar{Y} = b_0 + b_1 \bar{X}$$

$$\Leftrightarrow b_0 = \bar{Y} - b_1 \bar{X}$$

HW: Derive the expression for b_1

Formula for the Least Squares Estimates

$$Slope: \ \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$y - intercept: \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $SS_{xy} = \sum \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)$

$$SS_{xx} = \sum \left(x_i - \bar{x} \right)^2$$

n = Sample size

Interpreting the Estimates of β_0 and β_1 in Simple Liner Regression

y-intercept: $\hat{\beta}_0$ represents the predicted value of *y* when *x* = 0 (Caution: This value will not be meaningful if the value *x* = 0 is nonsensical or outside the range of the sample data.)

slope: $\hat{\beta}_1$ represents the increase (or decrease) in *y* for every 1-unit increase in *x* (Caution: This interpretation is valid only for *x*-values within the range of the sample data.)

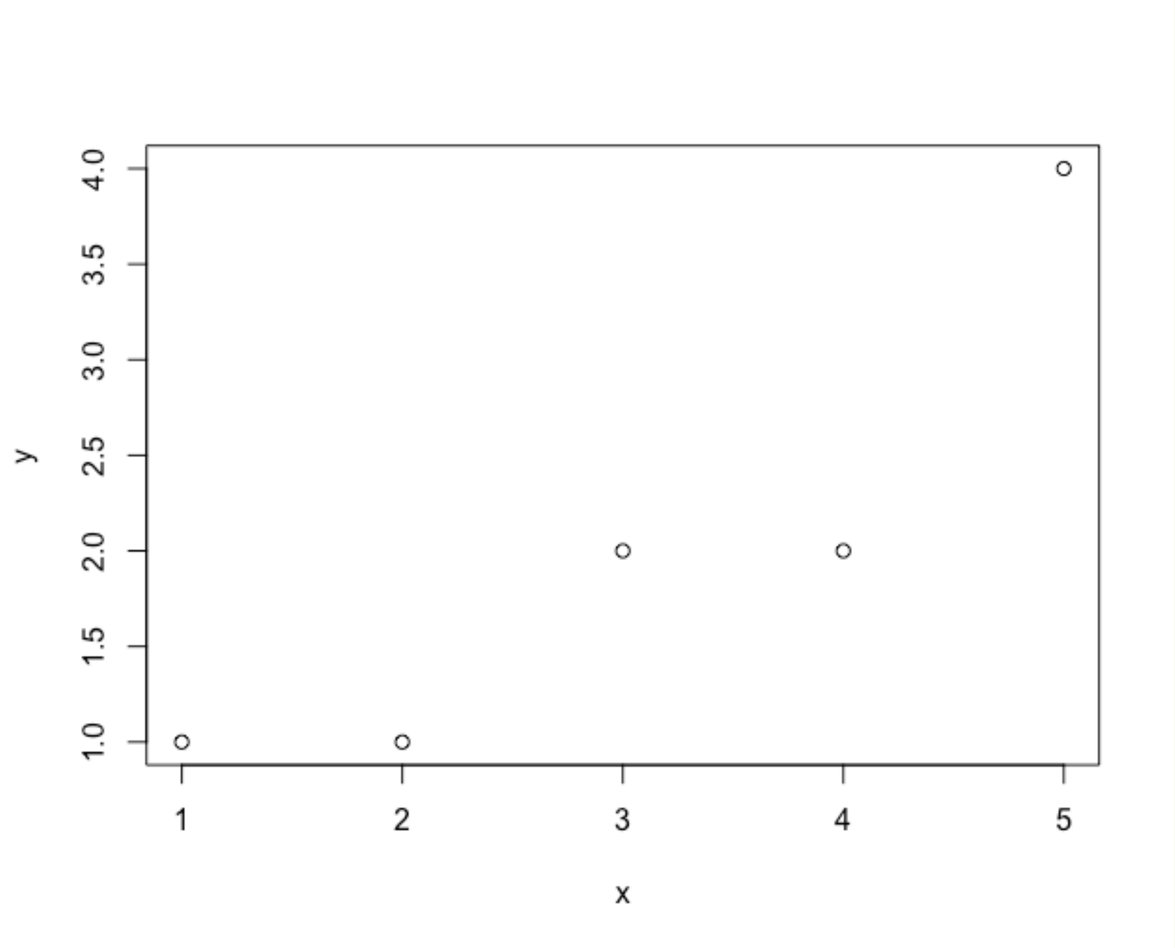
Least Squares Example

You're a marketing analyst for a company. You gather the following data:

Ad Expenditure (100$)	**Sales (1000$)**
1	**1**
2	**1**
3	**2**
4	**2**
5	**4**

Find the **least squares line** relating sales and advertising.

Scatterplot
Sales vs. Advertising



Parameter Estimation Solution

$$\bar{x} = \frac{\sum x}{5} = \frac{15}{5} = 3 \qquad \bar{y} = \frac{\sum y}{5} = \frac{10}{5} = 2$$

$$\begin{aligned} SS_{xy} &= \sum\left(x - \bar{x}\right)\left(y - \bar{y}\right) \\ &= \sum\left(x - 3\right)\left(y - 2\right) = 7 \end{aligned}$$

$$\begin{aligned} SS_{xx} &= \sum\left(x - \bar{x}\right)^2 \\ &= \sum\left(x - 3\right)^2 = 10 \end{aligned}$$

Parameter Estimation Solution

The slope of the least squares line is:

$$\hat{B}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{7}{10} = .7$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 2 - \left(.70\right)\left(3\right) = -.10$$

$$\hat{y} = -.1 + .7x$$

Parameter Estimation Computer Output

Parameter Estimates

$\hat{\beta}_0$

Variable	DF	Parameter Estimate	Standard Error	T for H0: Param=0	Prob>|T|
INTERCEP	1	-0.1000	0.6350	-0.157	0.8849
ADVERT	1	0.7000	0.1914	3.656	0.0354

$\hat{\beta}_1$

$$\hat{y} = -.1 + .7x$$

Coefficient Interpretation Solution

1. Slope ($\hat{\beta}_1$)
   - Sales Volume (*y*) is expected to increase by \$700 for each \$100 increase in advertising (*x*)

2. *y*-Intercept ($\hat{\beta}_0$)
   - Since 0 is outside of the range of the sampled values of *x*, the *y*-intercept has no meaningful interpretation

Properties of Fitted Regression Line

- The i^{th} residual is $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$
- The sum of the residuals is 0:

$$\sum_{i=1}^{n} e_i = 0$$

- The sum of the squared residuals is a minimum.
- The sum of the observed Ys equals the sum of fitted Ys:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i \qquad (1.18)$$

Proof of (1.18)

$\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n}(b_0 + b_1 X_i)$ (by def. of fitted value)

$= \sum_{i=1}^{n} b_0 + \sum_{i=1}^{n} b_1 X_i$ (by linearity of summation)

$= n b_0 + b_1 \sum_{i=1}^{n} X_i$ (by linearity of summation)

$= n(\bar{Y} - b_1 \bar{X}) + b_1 n\, \bar{X}$ (by formula 1.10b)

$= n\bar{Y} - b_1 n\, \bar{X} + b_1 n\, \bar{X}$ (by distributive law)

$= n\bar{Y} = \sum_{i=1}^{n} Y_i$ (by definition of mean)

3.

Estimation of Error Terms Variance σ^2

Point Estimator of σ^2

Recall the SLR model (1.1)

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

where ε_i is a random error term such that

- $E(\varepsilon_i) = 0$
- $Var(\varepsilon_i) = \sigma^2$
- $cov(\varepsilon_i, \varepsilon_j) = 0$

The variance σ^2 needs to be estimated!

Estimation of σ^2 for a (First-Order) Straight-Line Model

$$s^2 = \frac{\text{SSE}}{\text{Degrees of freedom for error}} = \frac{\text{SSE}}{n-2}$$

$$\text{where SSE} = \sum \left(y_i - \hat{y}_i \right)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

$$SS_{yy} = \sum \left(y_i - \bar{y} \right)^2$$

To estimate the standard deviation σ of ε, we calculate

$$s = \sqrt{s^2} = \sqrt{\frac{\text{SSE}}{n-2}}$$

We will refer to *s* as the **estimated standard error of the regression model.**

Calculating SSE, s^2, s Example

You're a marketing analyst for Hasbro Toys. You gather the following data:

Ad Expenditure (100$)	Sales (Units)
1	1
2	1
3	2
4	2
5	4

Find **SSE**, s^2, and s.

Calculating s^2 and s Solution

$$s^2 = \frac{SSE}{n-2} = \frac{1.1}{5-2} = .36667$$

$$s = \sqrt{.36667} = .6055$$