Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Model Ingredients

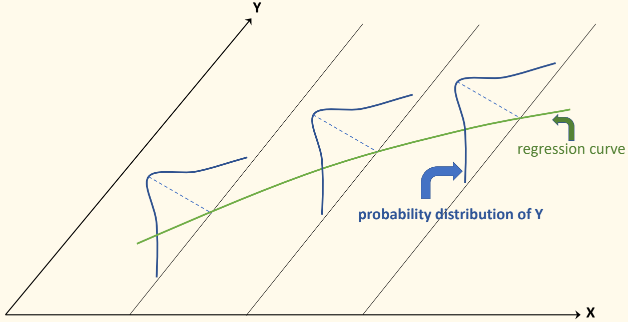
Key Ingredients

(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?

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

Figure: Illustration of regression model



Heights

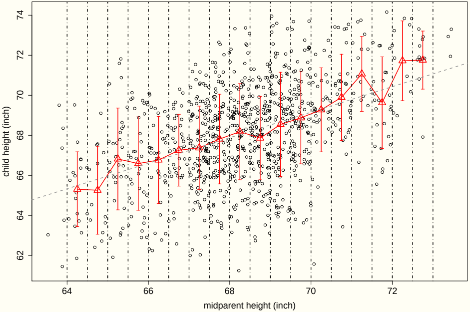


Figure: Child's height versus midparent's height

- 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.

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:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

- Y_i – value of the response variable in the *ith* case; X_i – value of the X variable in the *ith* case.
- **random errors**: ϵ_i – uncorrelated, zero-mean, equal-variance random variables
- **Unknown parameters**: β_0 – **regression intercept**; β_1 – **regression slope**; σ^2 – **error variance**

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:

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

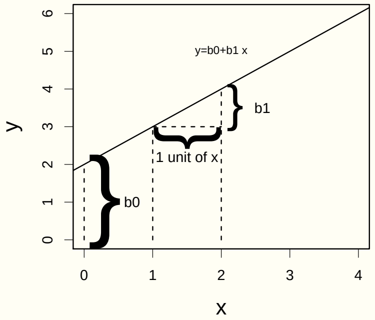
The simple linear regression model says:

- 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

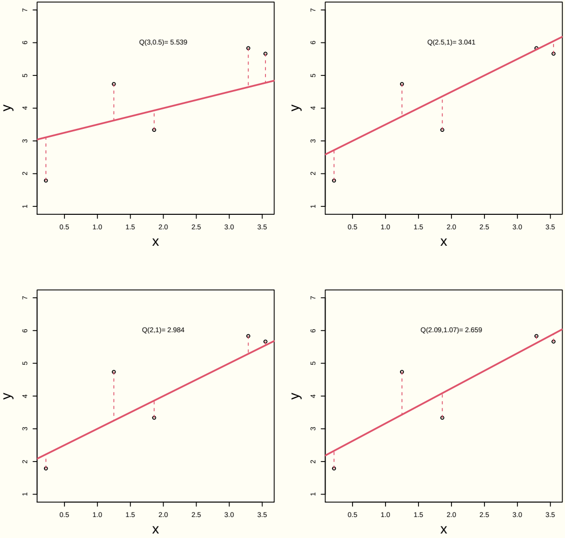
$$y = \beta_0 + \beta_1 x$$

- 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

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

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

- $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i, \ \ \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, 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

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \overline{Y} + r_{XY}\frac{s_Y}{s_X}(x - \overline{X}).$$

- 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 $\hat{\beta}_1 = r_{XY}$.
- **Regression effect**: One standard deviation change in X leads to r_{XY} standard deviation change in Y. (Recall $|r_{XY}| \leq 1$)

Derive the LS Estimator

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

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

This leads to the **normal equations**:

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

The solution is the LS estimator.

Fitted Values and Residuals

Fitted Values and Residuals

- **Fitted values** are predictions by the LS line :

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

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

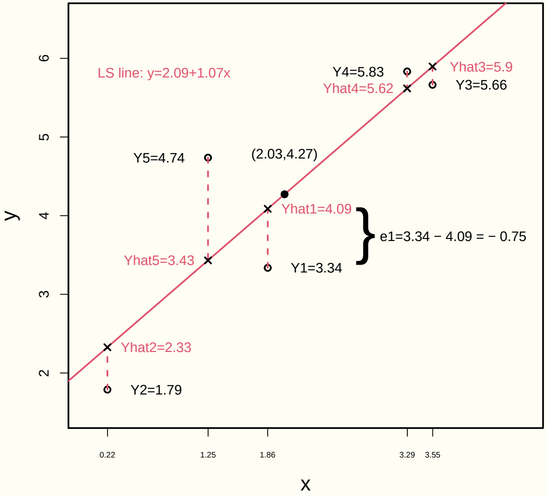
$$\begin{aligned} e_i &= Y_i - \widehat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i), \quad i = 1, \cdots n. \\ &= (Y_i - \overline{Y}) - \hat{\beta}_1 (X_i - \overline{X}). \end{aligned}$$

Example

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				

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

Figure: LS line, fitted values and residuals



Properties of Residuals

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

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

- Error variance $\sigma^2 = \mathrm{Var}(\epsilon_i)$. (Recall $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$)
- Idea: Estimate σ^2 by the "variance" of residuals. (Recall residual $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$)
- **Error sum of squares (SSE)**:

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

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

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

Degrees of Freedom

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

Example (Cont'd)

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

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

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

LS Estimator: Properties

Mean and Variance

- **LS estimators are unbiased**:

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

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

$$\begin{aligned} \sigma^2\{\hat{\beta}_0\} &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right] \\ \sigma^2\{\hat{\beta}_1\} &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}. \end{aligned}$$

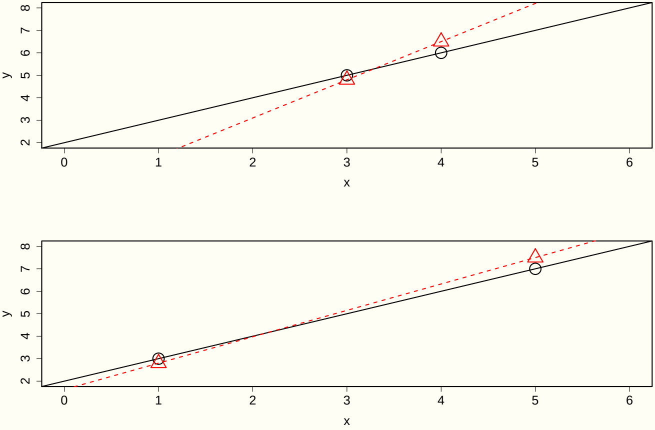
Standard Error (SE)

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

$$
\begin{aligned}
s\{\hat{\beta}_0\} &= \sqrt{MSE\left[\frac{1}{n}+\frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]} \\
s\{\hat{\beta}_1\} &= \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i-\overline{X})^2}}
\end{aligned}
$$

- 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.

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



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:

$$Y_i = 2 + X_i + \epsilon_i, \quad i = 1, \cdots, 5.$$

- Repeat 100 times $\rightarrow$ 100 data sets.

data set 1	case	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

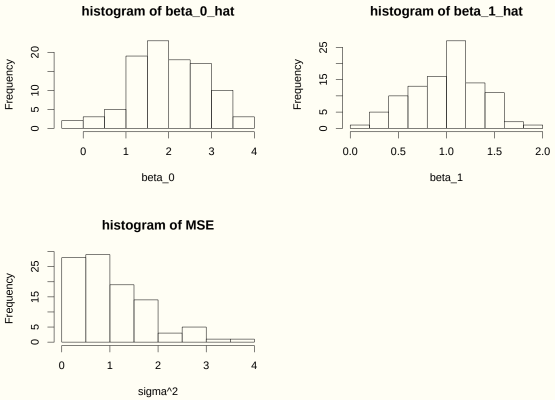
$\hat{\beta}_0 = 1.34,\ \hat{\beta}_1 = 0.94,\ MSE = 0.79.$

$\cdots, \cdots$

data set 100	case	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

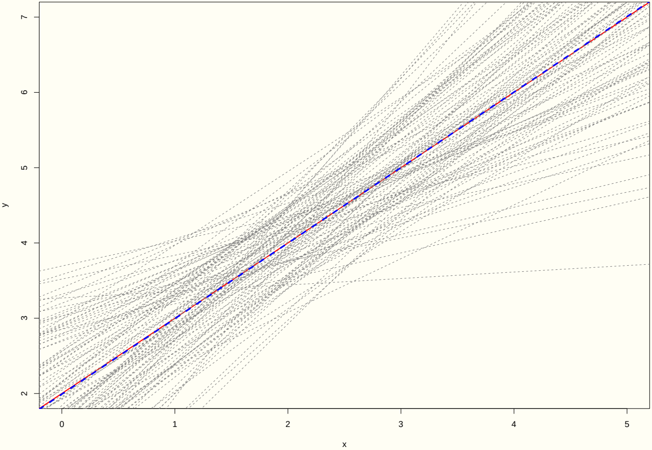
$\hat{\beta}_0 = 1.75,\ \hat{\beta}_1 = 1.09,\ MSE = 0.24.$

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



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: $1.99, 0.847$.

- $\hat{\beta}_1$: 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: $1.002, 0.36$.