통계분석
Statistical Analysis

Regression

Regression Problems

Variable X Variable Y

$$\{x_1, x_2, \cdots, x_n\} \longleftrightarrow \{y_1, y_2, \cdots, y_n\}$$

Relationship between X and Y?

$$y_i = f(x_i)?$$

X and Y are related to each other in a <u>nondeterministic</u> way.

[EX1] X = age of a child, Y = size of that child's vocabulary

[EX2] X = size of engine, Y = fuel efficiency of that engine

- Obviously X can affect Y, but X is not related to Y in a deterministic way.
- Individual by individual, X can have a slightly different value of Y.

The Linear Regression Model

Variable x | Random Variable Y

$$\{x_1, x_2, \cdots, x_n\} \longleftrightarrow \{y_1, y_2, \cdots, y_n\}$$

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

- x: the variable fixed by the experimenter, which is called the <u>independent</u>, <u>predictor</u>, or <u>explanatory variable</u>.

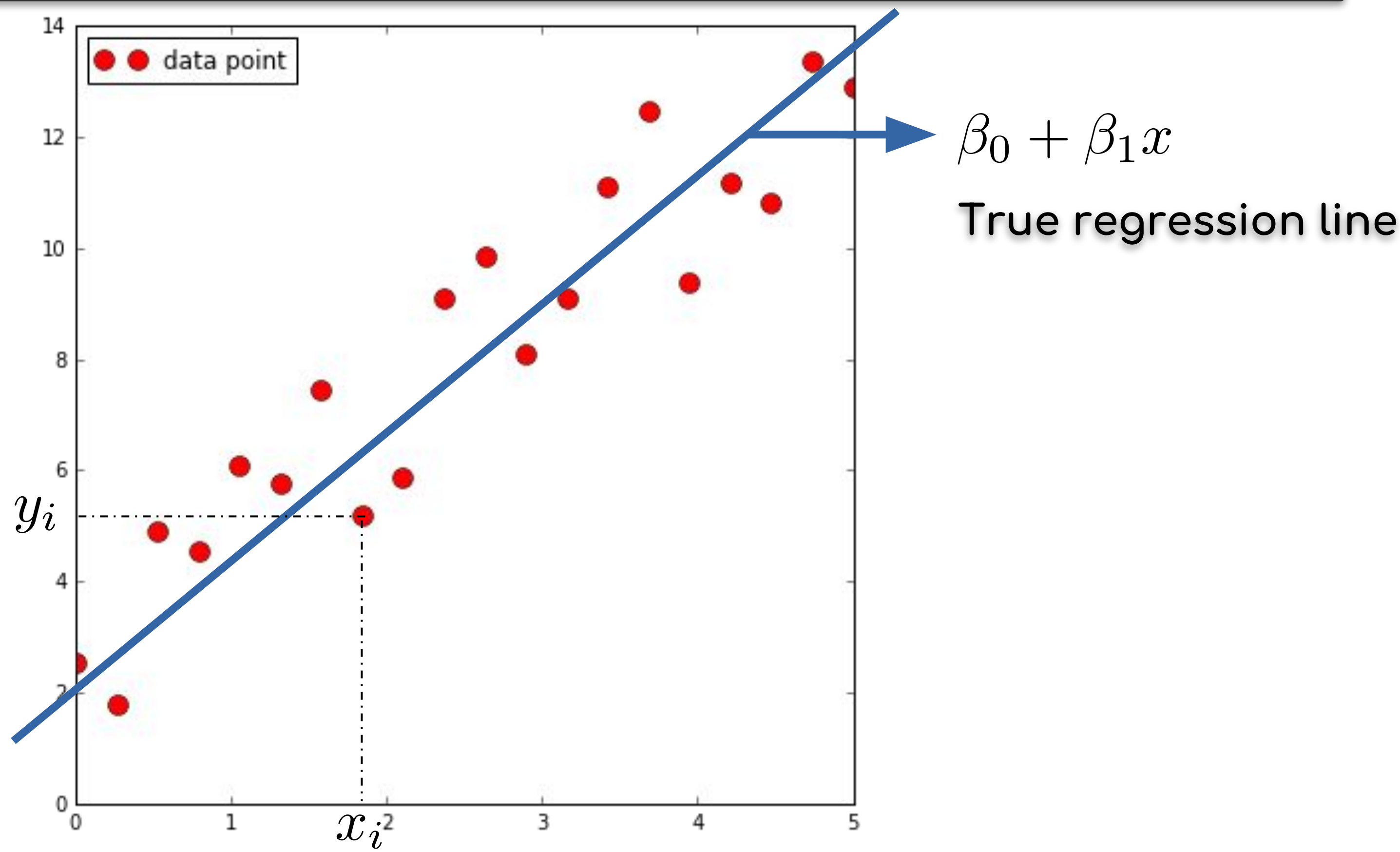
- Y: the random variable <u>affected by randomness</u> for a fixed value of x, dependent or response variable.

The Linear Regression Model

Variable x Random Variable Y

$$\{x_1, x_2, \cdots, x_n\} \longleftrightarrow \{y_1, y_2, \cdots, y_n\}$$

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



The Linear Regression Model

Variable x | Random Variable Y

$$\{x_1, x_2, \cdots, x_n\} \longleftrightarrow \{y_1, y_2, \cdots, y_n\}$$

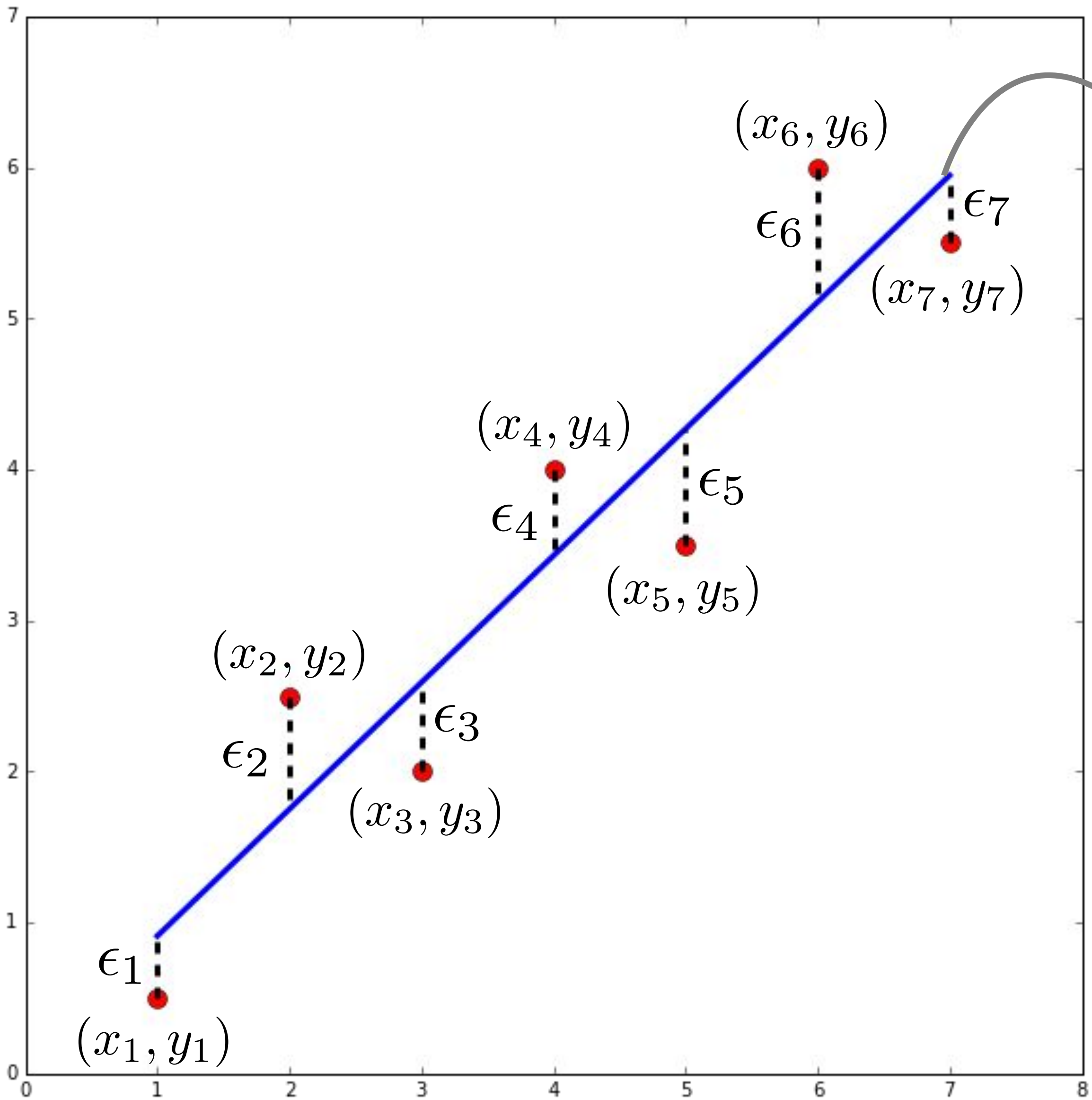
$$Y = \underbrace{\beta_0 + \beta_1 x}_{\text{Deterministic part}} + \epsilon$$

Random deviation or random error

ϵ is a rando variable, normally distributed.

$$\epsilon \sim N(0, \sigma^2) \begin{cases} E(\epsilon) = 0 \\ \mathrm{Var}(\epsilon) = \sigma^2 \end{cases}$$

The Linear Regression Model



linear regression equation:

$$f(x) = \beta_0 + \beta_1 x$$

data points (x_i, y_i) are NOT necessarily on the fit equation.

Vertical deviations

$$\begin{aligned} \epsilon_i &= y_i - f(x_i) \\ &= y_i - (\beta_0 + \beta_1 x_i) \end{aligned}$$

$\epsilon_i =$ error,residual, or deviation

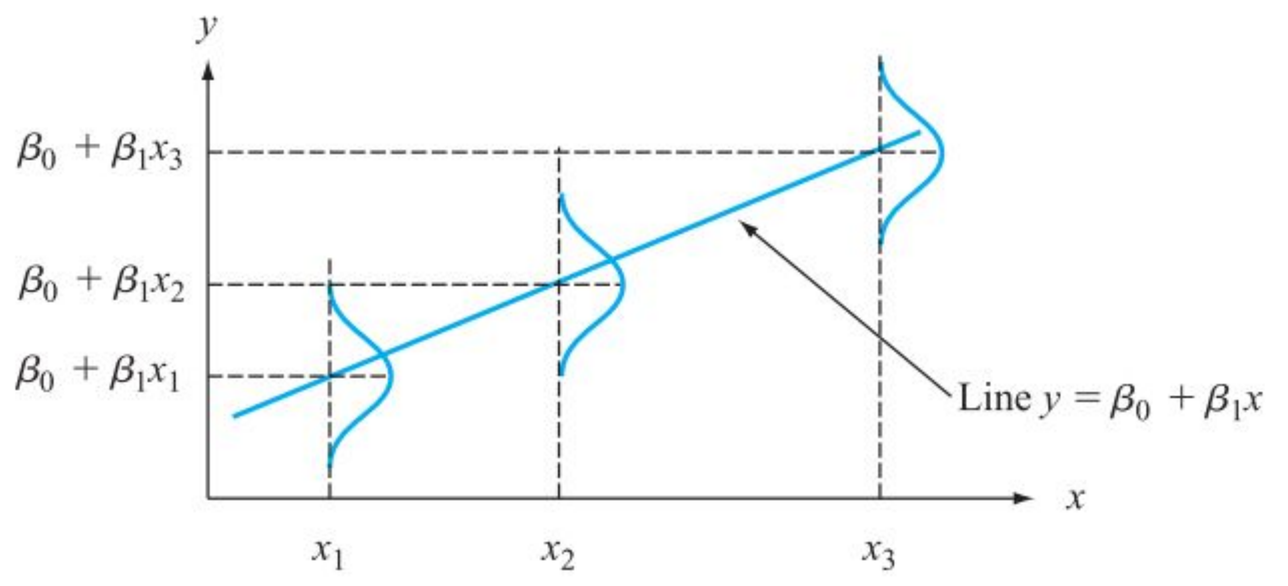
The Linear Regression Model

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

- $E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x$

  When the independent variable x is fixed, the mean value of the random variable Y

- $\mathrm{Var}(Y|x) = \mathrm{Var}(\epsilon) = \sigma^2$

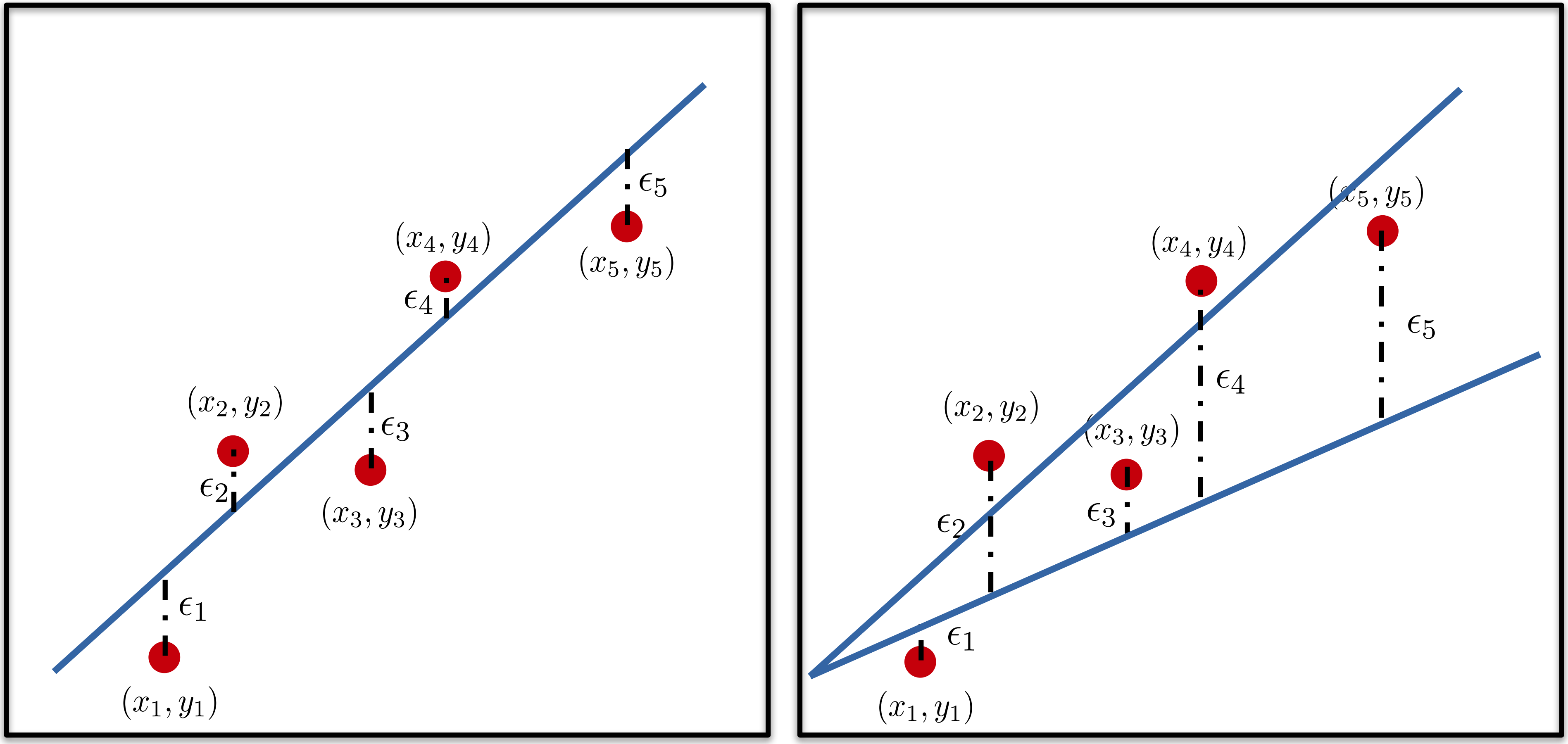


Probability and Statistics for Engineering and the Sciences, J. L. Devore

The Linear Regression Model

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

How can we determine these parameters?



→ Determine parameters by minimizing errors.

Least-Square Fit

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

→ Determine parameters by minimizing errors.

- Sum of squared vertical deviations (errors)

$$S(b_0, b_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [y_i - b_0 - b_1 x_i]^2$$

- Least-square fit: Minimizing the sum of squared errors

$$\beta_0 = b_0,\ \ \beta_1 = b_1 \ \text{satisfying}$$

$$\underline{\frac{\partial}{\partial b_0} S(b_0, b_1) = 0 \quad \frac{\partial}{\partial b_1} S(b_0, b_1) = 0}$$

Minimization conditions

Least-Square Fit

$$\frac{\partial}{\partial b_0} S(b_0, b_1) = (-1) \sum_{i=1}^{n} [y_i - b_0 - b_1 x_i] = 0$$

$$\frac{\partial}{\partial b_1} S(b_0, b_1) = \sum_{i=1}^{n} [y_i - b_0 - b_1 x_i] (-x_i) = 0$$

Normal Equations

$$n b_0 + \boxed{\left(\sum_{i}^{n} x_i \right)} b_1 = \boxed{\sum_{i=1}^{n} y_i}$$

$$b_0 \boxed{\left(\sum_{i=1}^{n} x_i \right)} + \boxed{\left(\sum_{i=1}^{n} x_i^2 \right)} b_1 = \boxed{\sum_{i=1}^{n} x_i y_i}$$

Least-Square Fit

$$nb_0 + \left(\sum_{i}^{n} x_i\right) b_1 = \sum_{i=1}^{n} y_i$$

$$b_0 \left(\sum_{i=1}^{n} x_i\right) + \left(\sum_{i=1}^{n} x_i^2\right) b_1 = \sum_{i=1}^{n} x_i y_i$$

$$\longrightarrow \quad \hat{\beta}_1 = b_1 = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = b_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

We need to calculate $\sum x_i,\ \sum x_i^2,\ \sum y_i,\ \sum x_i y_i$

True Regression Line vs Estimated Regression Line

Variable x $\longleftrightarrow$ Random Variable Y

$$Y = \underbrace{\beta_0 + \beta_1 x}_{\text{True regression line}} + \epsilon$$

random error

To estimate true values of β_1 and β_2,

$\hat{\beta}_1$ and $\hat{\beta}_2$ are calculated from sample data by using the least-square fit.

Sample data

$\{x_1, x_2, \cdots, x_n\}$

$\{y_1, y_2, \cdots, y_n\}$

$\xrightarrow{\textbf{Least-square fit}}$

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

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estimated regression line

Estimators of True Regression Line

Variable x $\longleftrightarrow$ Random Variable Y

$$Y = \underbrace{\beta_0 + \beta_1 x}_{\text{True regression line}} + \epsilon$$

random error

Sample data

$\{x_1, x_2, \cdots, x_n\}$

$\{y_1, y_2, \cdots, y_n\}$

$\xrightarrow{\textbf{Least-square fit}}$

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

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estimated regression line

$\hat{\beta}_{0,1}$ = Estimators of the true regression line, $\beta_{0,1}$

Least-Square Fit: Example

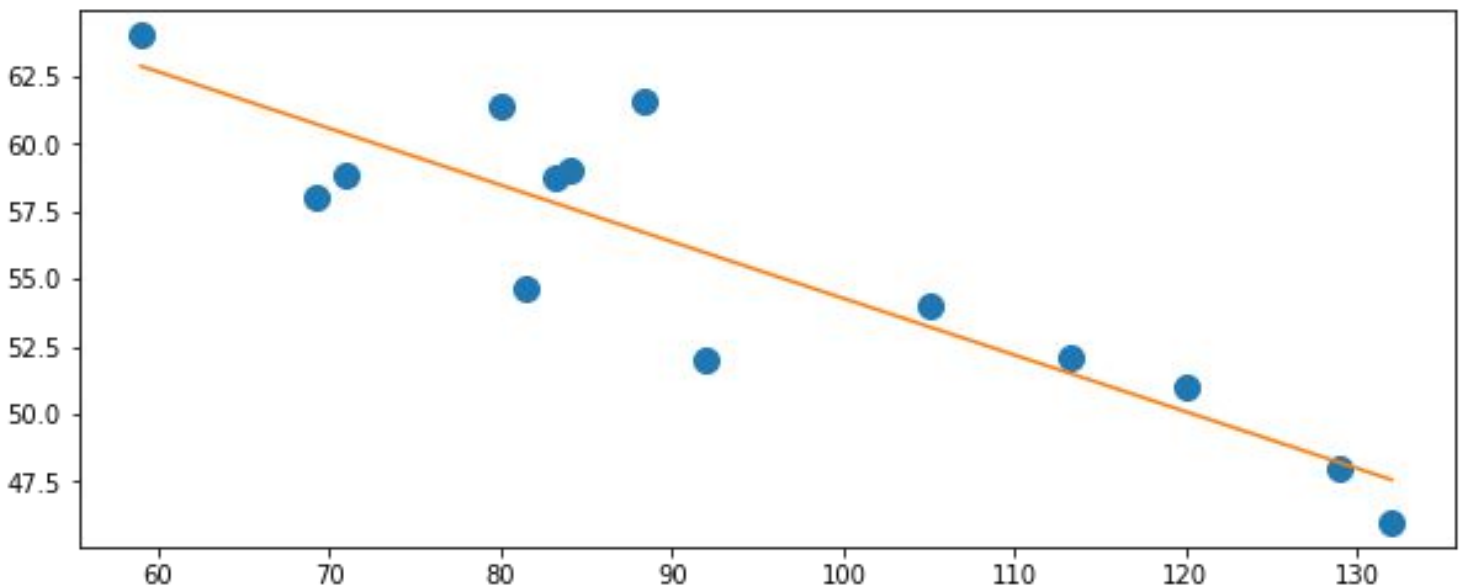
x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

$$\sum x_i = 1307.5 \qquad \sum y_i = 779.2$$

$$\sum x_i^2 = 128913.93 \qquad \sum x_i y_i = 71347.30$$

$$\longrightarrow \hat{\beta}_1 = -0.20938742, \ \hat{\beta}_1 = 75.212432$$

Least-Square Fit: Example



$$\sum x_i = 1307.5 \qquad \sum y_i = 779.2$$

$$\sum x_i^2 = 128913.93 \qquad \sum x_i y_i = 71347.30$$

$$\longrightarrow \hat{\beta}_1 = -0.20938742,\ \hat{\beta}_1 = 75.212432$$

Estimation of Residual Variance

$\{y_1, y_2, \cdots, y_n\}$: observed values

$\{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n\}$: predicted values by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- **Sum of squared errors (SSE)**

$$\begin{aligned}\mathrm{SSE} = \sum(y_i - \hat{y}_i)^2 &= \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i\end{aligned}$$

- **Unbiased Estimator of Residual Variance**

$$\hat{\sigma}^2 = s^2 = \frac{\mathrm{SSE}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{\underline{n-2}}$$

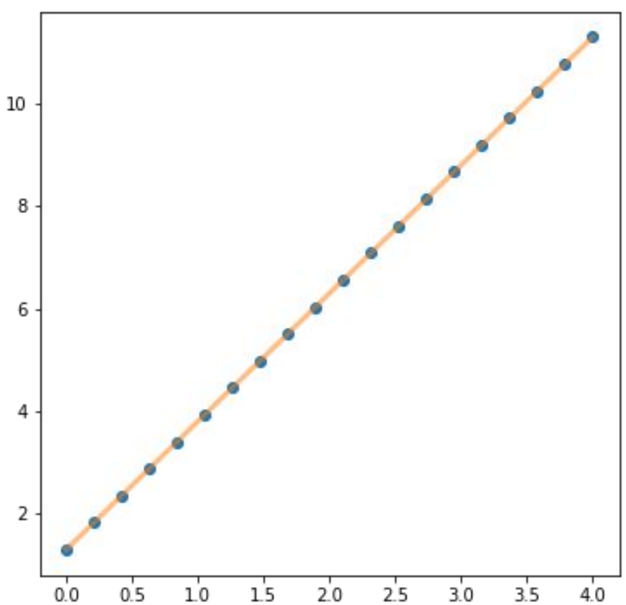
Why is degree of freedom reduced by 2?

We have two relations of $\{x_i\}, \{y_i\}$, which are $\hat{\beta}_0, \hat{\beta}_1$

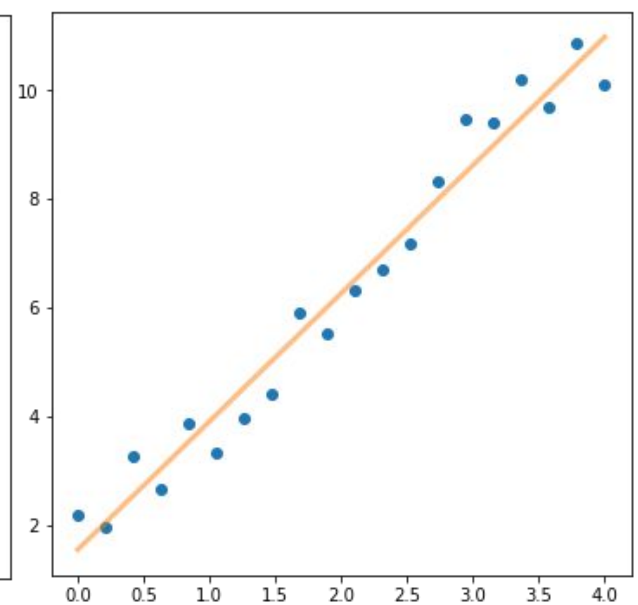
Coefficient of Determination

$$y = \underbrace{\overbrace{\beta_0 + \beta_1 x}^{\text{deterministic}}}_{\text{1. linearity of x and y}} + \underbrace{\overbrace{\epsilon}^{\text{non-deterministic}}}_{\text{2. random error}}$$

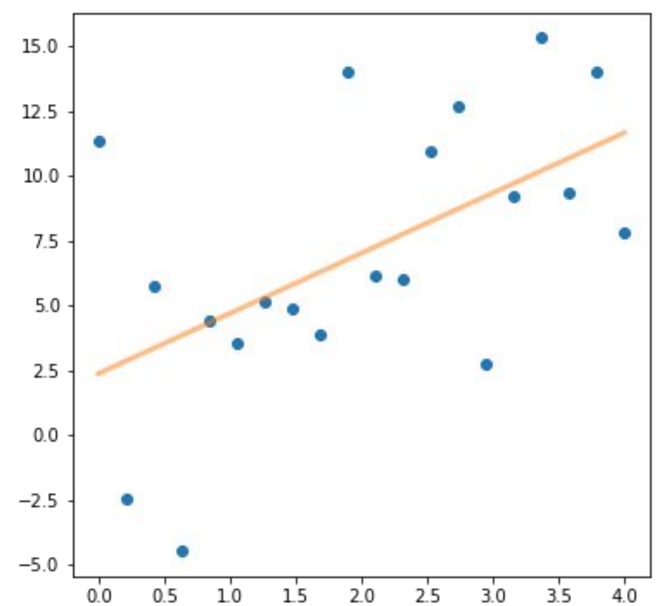
What determines the sample variation in y?



No random error
Definitely, linear between x and y.



Very small random error
Linear model well explains relationship between x and y.



A large variation in y implies that a simple linear model fails to explain the relationship between x and y.

Coefficient of Determination

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

What determines the sample variation in y?

1. linearity of x and y

2. random error

- **Coefficient of Determination (결정계수)**

Numeric measure to show the contribution of linearity between x and y to the sample variation of y data, especially, in comparison with random error contributions

We need to define

1) The sample variation of y data
2) The contribution of the linear relationship between x and y
3) The contribution of random errors

Sum of Squares

Total Sum of Squares (SST, 총제곱합)

$$\mathrm{SST} = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

Sample mean of y $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- The total sum of squares (SST) represents the sample variation (variance) of y data

Sum of Squared Errors (SSE, 오차제곱합)

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

- SSE measures squared deviation between y data and the (estimated) regression line.

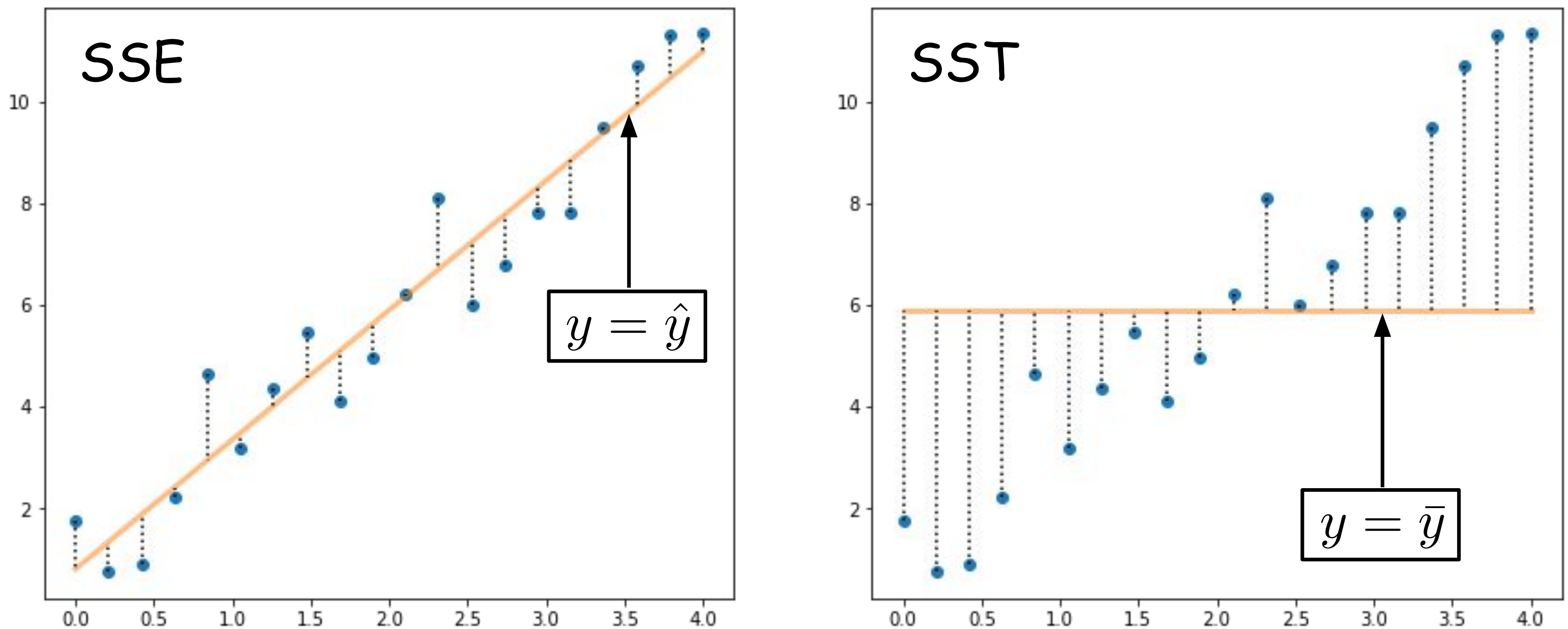
Regression Sum of Squares (SSR, 회귀제곱합)

$$\mathrm{SSR} = \sum (\bar{y} - \hat{y}_i)^2 = \sum (\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- SSR measures squared deviations between the mean of y data and the regression line.

Relationships among Sum of Squares

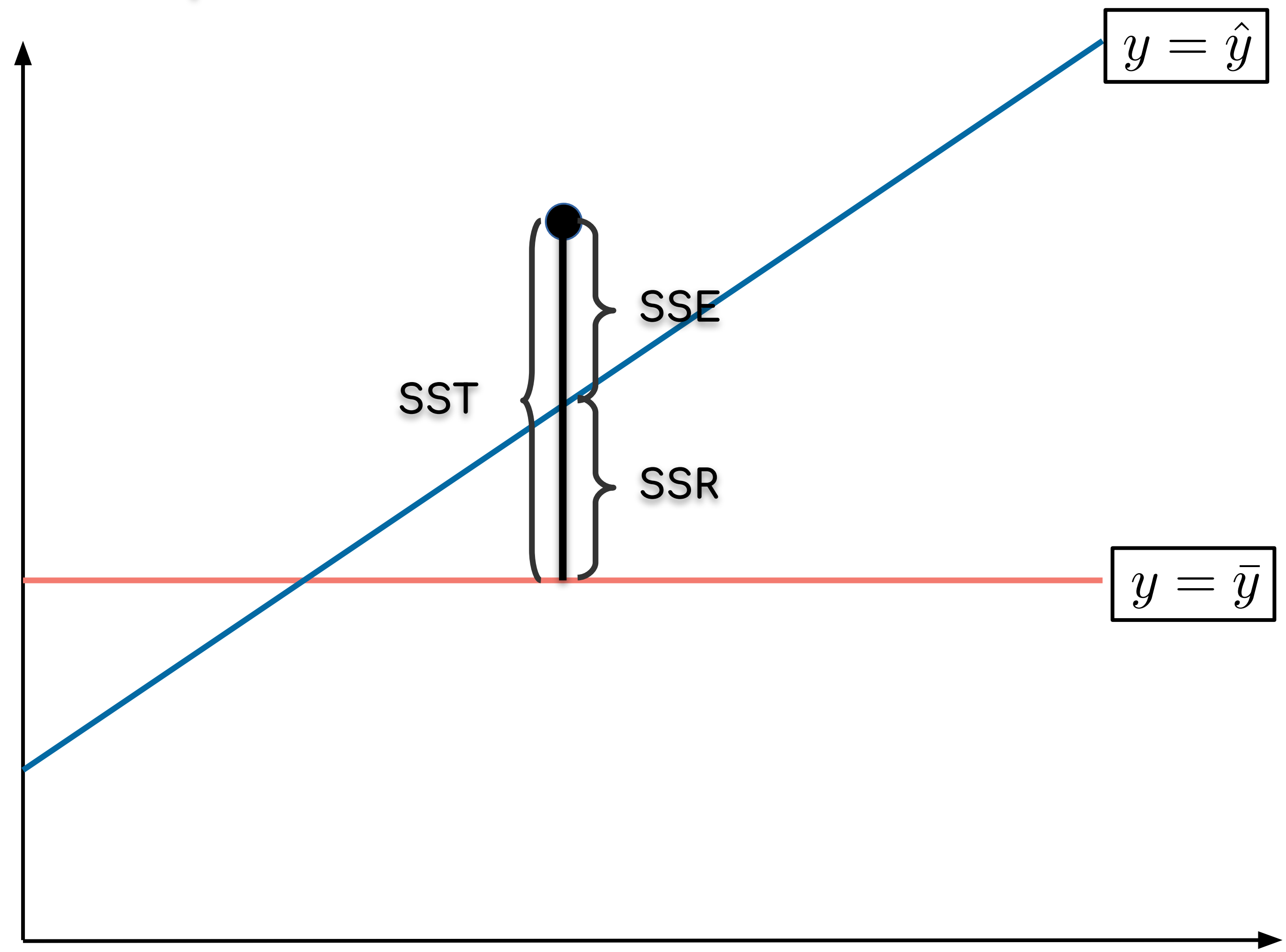
Graphical Interpretation



$$\mathrm{SSE} < \mathrm{SST}$$

Relationships among Sum of Squares

Graphical Interpretation



Relationships among Sum of Squares

Mathematical Relation

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$$

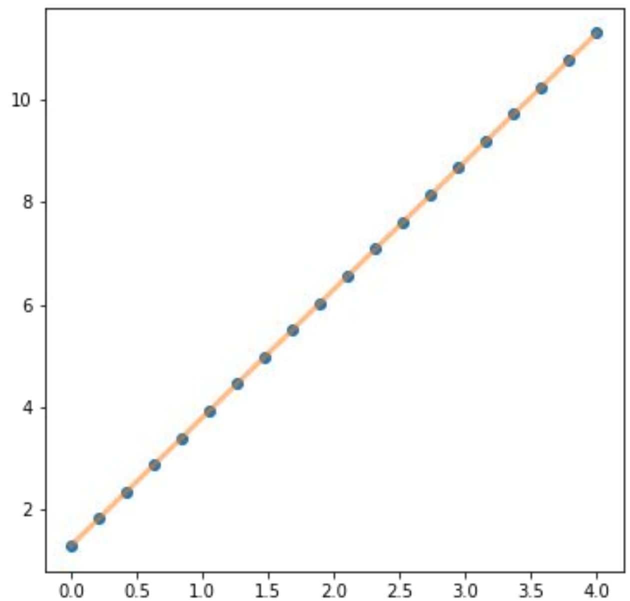
$$\underbrace{\boxed{\sum_{i=1}^{n} (y_i - \bar{y})^2}}_{\text{SST}} = \underbrace{\boxed{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}_{\text{SSE}} + \underbrace{\boxed{\sum_{i=1}^{n} (\bar{y} - \hat{y}_i)^2}}_{\text{SSR}}$$

The Coefficient of Determination

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

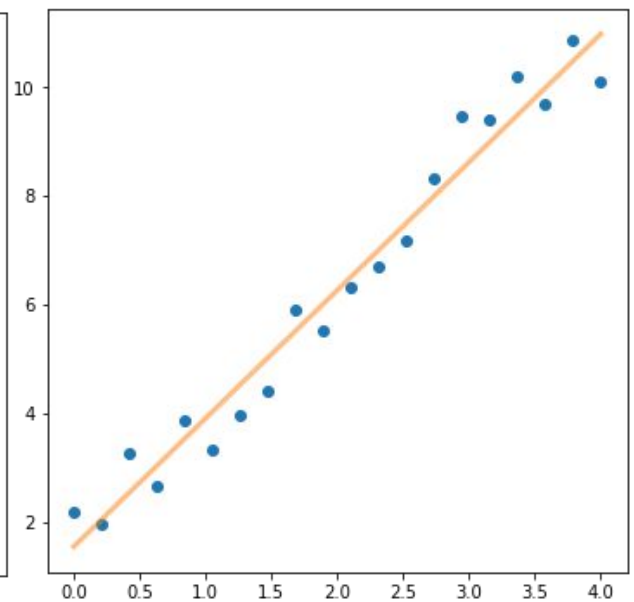
The proportion of observed y variation that can be explained by the simple linear regression model.

→ The higher coefficient of determination, the better the regression model explains the variation of y.



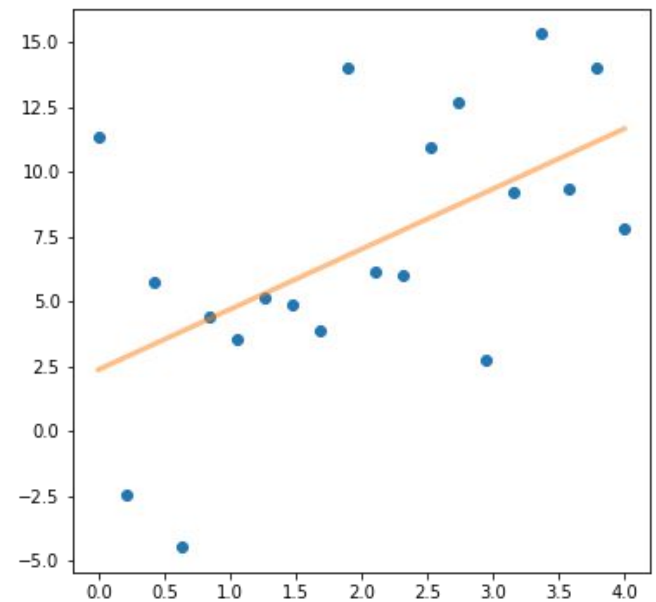
$$\text{SSE} = 0$$

r^2 = 1



$$\text{SSE} << 1$$

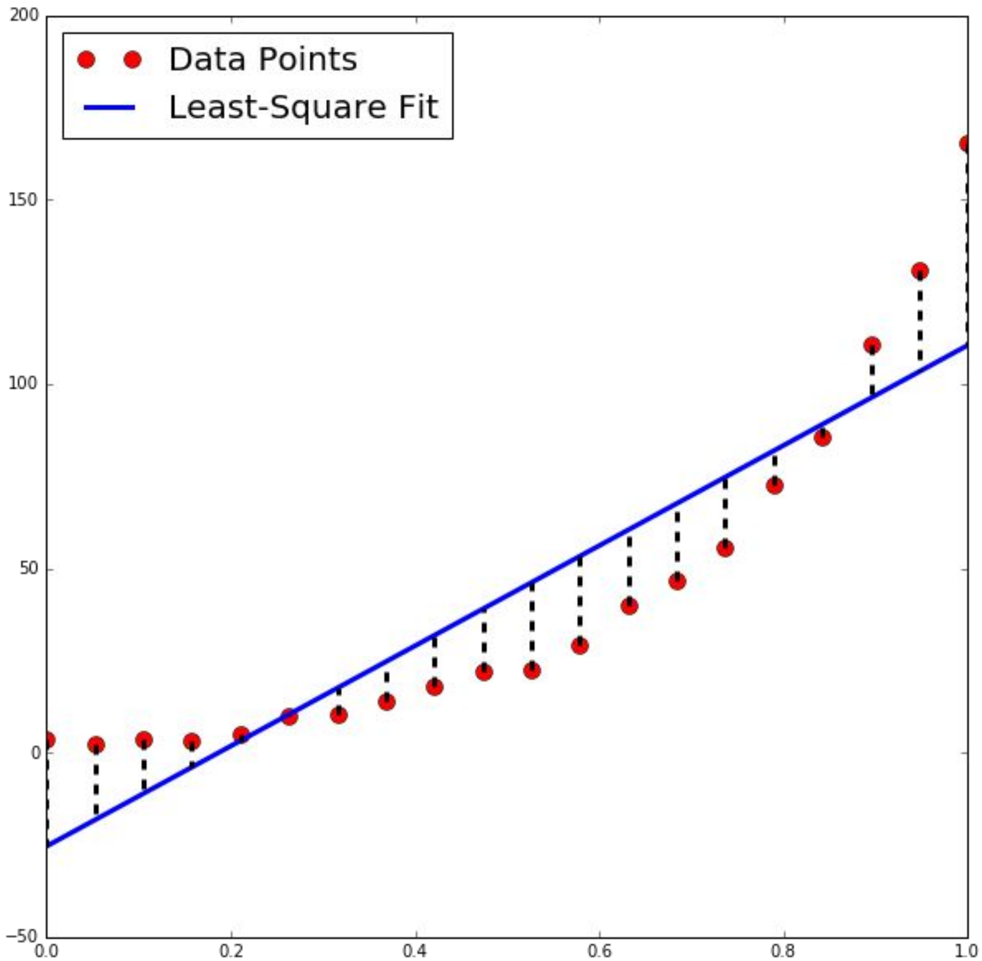
r^2 < 1, very close to 1



$$\text{SSE} >> 1$$

r^2 >0, very close to 0

Linearization with Variable Transformation

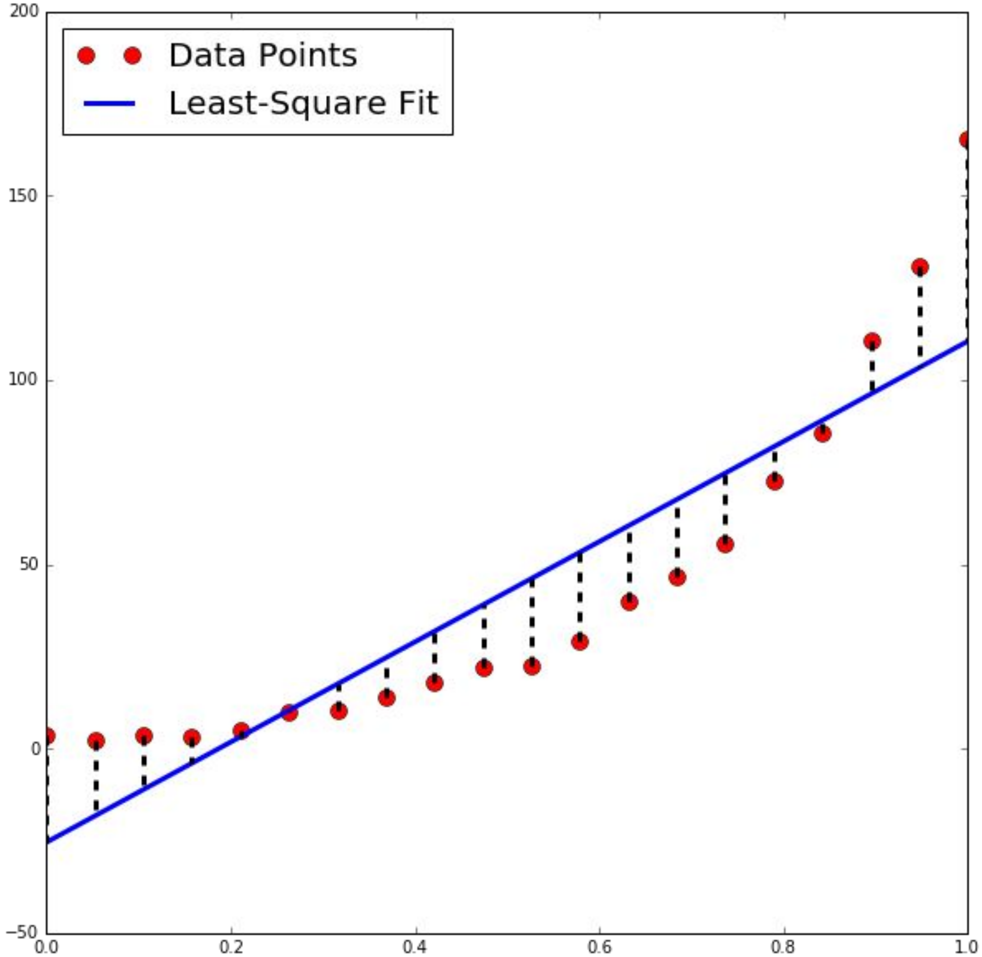


The sum of squared residuals

$$\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \sim 10^4$$

It seems that data points do not show a linear behavior.
Least-square linear fit might not be a good model for them.

Linearization with Variable Transformation



The exponential model

$$\hat{y} = \alpha \exp[\beta x]$$

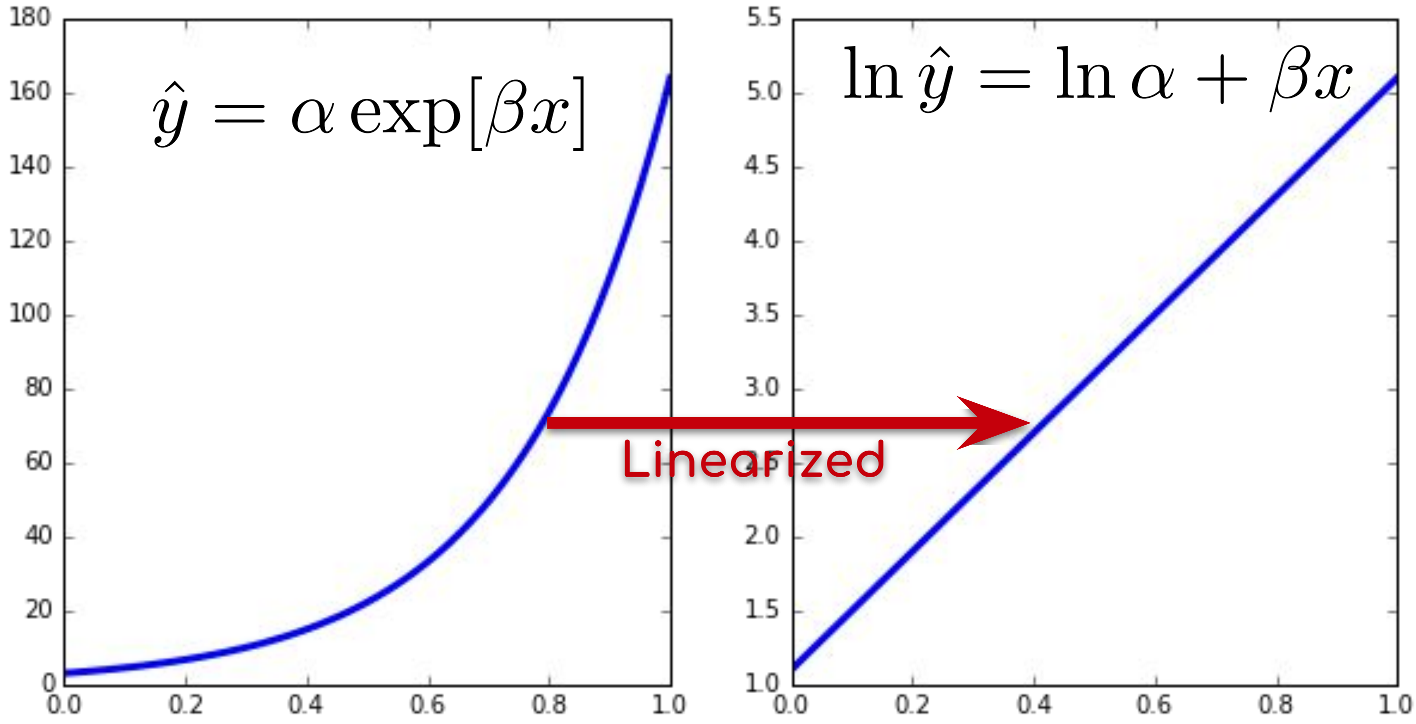
$$\downarrow$$

$$\ln \hat{y} = \ln \alpha + \beta x$$

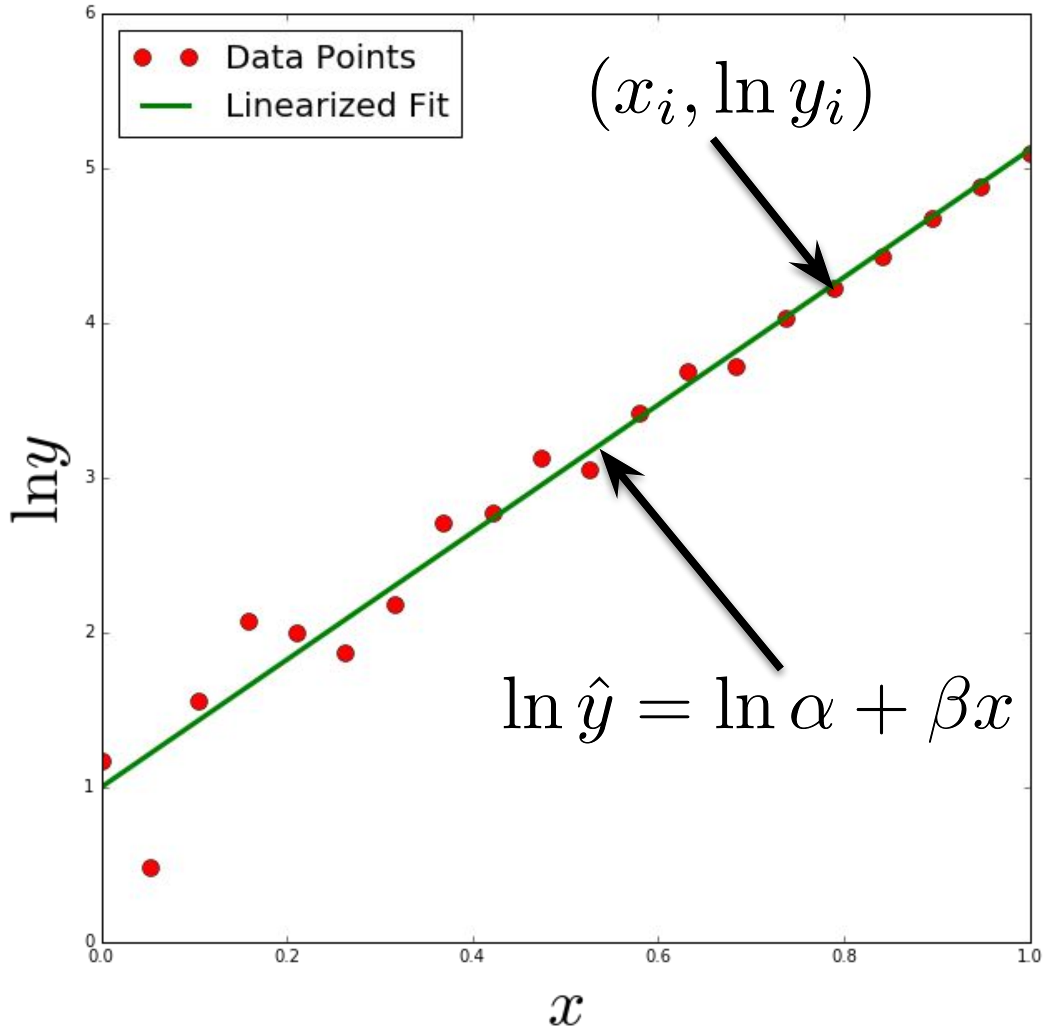
Data points

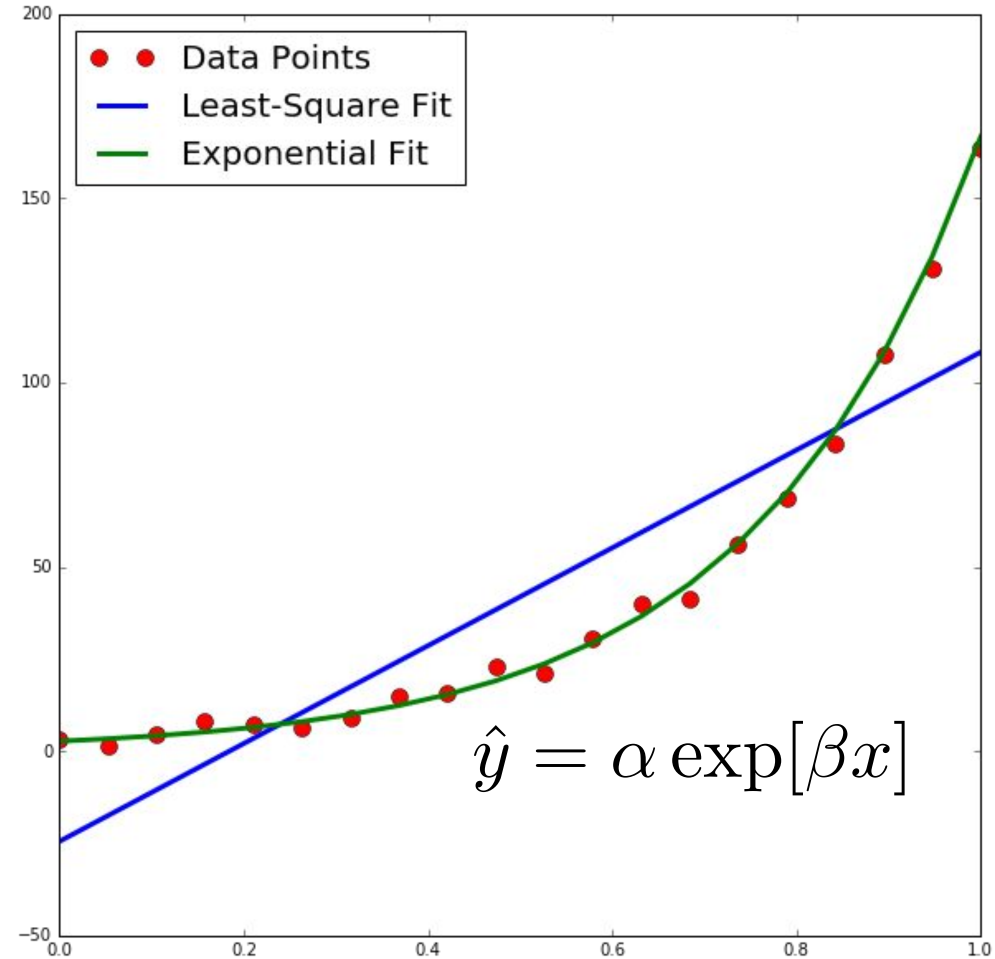
$$(x_i, y_i) \to (x_i, \ln y_i)$$

Linearization with Variable Transformation



Linearization with Variable Transformation





Linearization with Variable Transformation

- <u>Exponential Equation</u>

$$\hat{y} = \alpha \exp[\beta x] \xrightarrow[\text{Linearized}]{} \ln \hat{y} = \ln \alpha + \beta x$$

- <u>Power Equation</u>

$$\hat{y} = \alpha x^{\beta} \xrightarrow[\text{Linearized}]{} \log_{10} \hat{y} = \log_{10} \alpha + \beta \log_{10} x$$

- <u>Saturation-Growth-Rate Equation</u>

$$\hat{y} = \alpha \frac{x}{\beta + x} \xrightarrow[\text{Linearized}]{} \frac{1}{\hat{y}} = \frac{1}{\alpha} + \frac{\beta}{\alpha} \frac{1}{x}$$

Polynomial Regression

Nonlinear Curve-fitting

- Linearization of nonlinear equations
- Extension of linear fitting to polynomials

- Observed data points

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

- Polynomial regression equation

$$f_m(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$$

- Error (Residual): Vertical deviation from data points

$$e_i = y_i - f(x_i) = y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x_i^m\right)$$

Best fit to Data Points

- The sum of squared vertical deviations

$$S(a_0, a_1, \cdots, a_m) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (a_0 + a_1 x_i + \cdots + a_m x_i^m)\right]^2$$

- The best fit? ⟶ Minimizing the sum of squared errors

$$\frac{\partial}{\partial a_k} S(a_0, a_1, \cdots, a_m) = 0$$

$$\to \sum_{i=1}^{n} 2\left[y_i - (a_0 + a_1 x_i + \cdots + a_m x_i^m)\right](-x_i^k) = 0$$

Least-Square Fit

$$\frac{\partial}{\partial a_k} S(a_0, a_1, \cdots, a_m) = 0$$

$$\rightarrow \sum_{i=1}^{n} \left[y - (a_0 + a_1 x_i + \cdots + a_m x_i^m) \right] (-x_i^k) = 0$$

$$\rightarrow \left(\sum_{i=1}^{n} x_i^k \right) a_0 + \left(\sum_{i=1}^{n} x_i^{k+1} \right) a_1 + \cdots + \left(\sum_{i=1}^{n} x_i^{k+m} \right) a_m = \sum_{i=1}^{n} x_i^k y_i$$

Normal Equations

Least-Square Fit

- <u>Normal Equations</u>

$$na_0 + \left(\sum_{i=1}^{n} x_i\right) a_1 + \cdots + \left(\sum_{i=1}^{n} x_i^m\right) a_m = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i\right) a_0 + \left(\sum_{i=1}^{n} x_i^2\right) a_1 + \cdots + \left(\sum_{i=1}^{n} x_i^{m+1}\right) a_m = \sum_{i=1}^{n} x_i y_i$$

$$\vdots$$

$$\left(\sum_{i=1}^{n} x_i^m\right) a_0 + \left(\sum_{i=1}^{n} x_i^{m+1}\right) a_1 + \cdots + \left(\sum_{i=1}^{n} x_i^{2m}\right) a_m = \sum_{i=1}^{n} x_i^m y_i$$

$$\left.\begin{array}{l} \boxed{\text{(m+1) linear eqns}} \\ \text{(m+1) unknowns} : a_0, a_1, \cdots, a_m \end{array}\right\} \text{Uniquely determined}$$

Least-Square Fit

<u>Normal Equations: Solving linear equations</u>

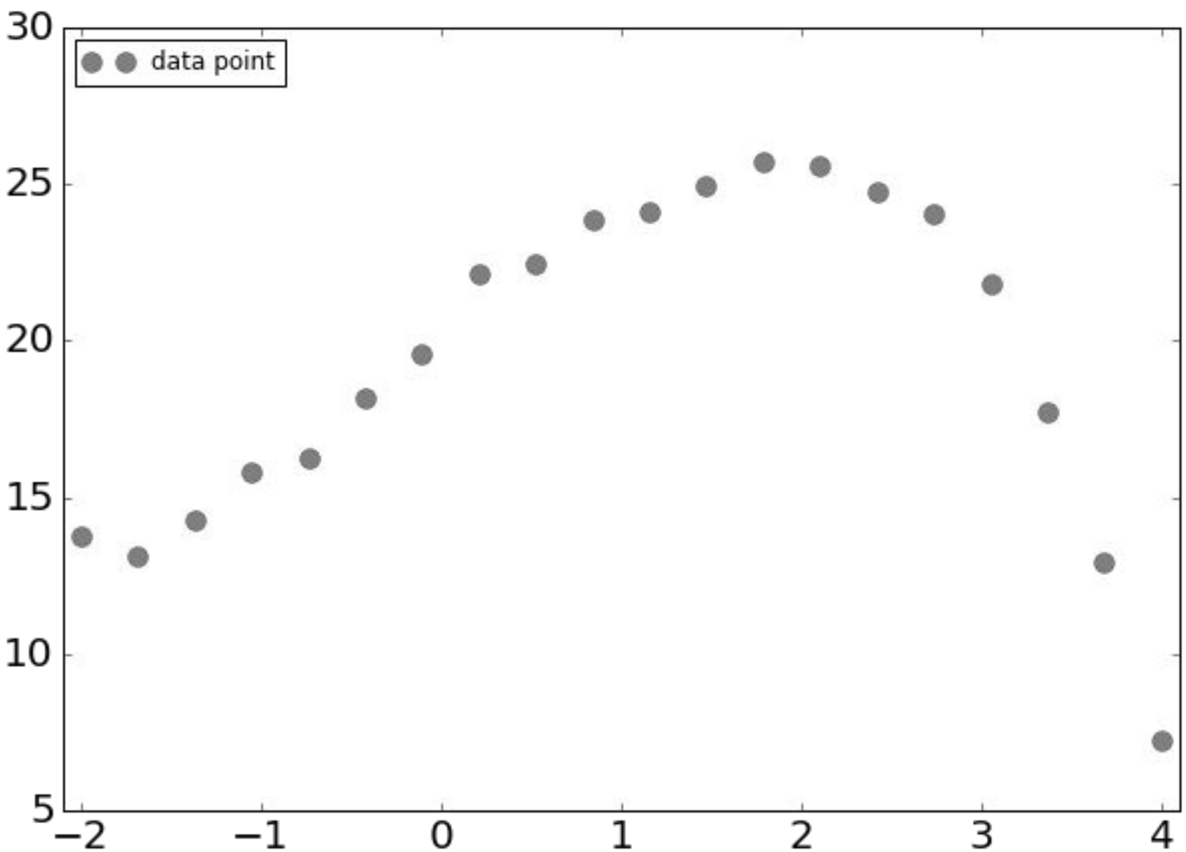
$$\begin{pmatrix} n & \sum x_i & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \cdots & \sum x_i^{2m} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^m y_i \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} n & \sum x_i & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \cdots & \sum x_i^{2m} \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^m y_i \end{pmatrix}$$

$$\text{Here, } \sum \equiv \sum_{i=1}^{n}$$

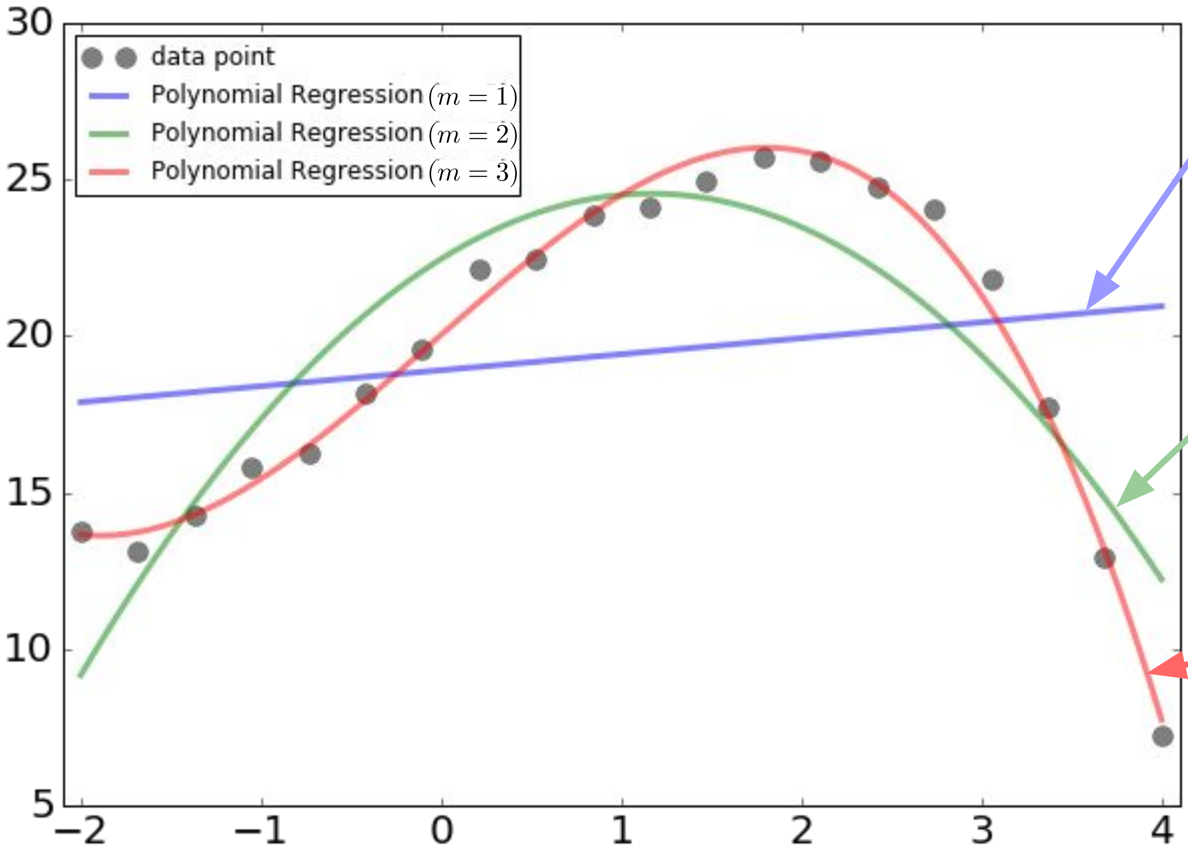
Polynomial Least-Square Fit

data points: $y = 20 + 5x - 0.5x^3 + e \begin{cases} e = \text{random number} \\ |e| \leq 0.5 \end{cases}$



Polynomial Least-Square Fit

$$\text{data points: } y = 20 + 5x - 0.5x^3 + e \begin{cases} e = \text{random number} \\ |e| \leq 0.5 \end{cases}$$



$$f_1(x) = 18.91 + 0.511x$$

$$S = \sum_{i=1}^{n} [y_i - f_1(x_i)]^2 \cong 523.19$$

$$f_2(x) = 22.47 + 3.584x - 1.536x^2$$

$$S = \sum_{i=1}^{n} [y_i - f_2(x_i)]^2 \cong 111.07$$

$$f_3(x) = 20.03 + 5.032x$$
$$-0.063x^2 - 0.491x^3$$

$$S = \sum_{i=1}^{n} [y_i - f_3(x_i)]^2 \cong 5.448$$

Multiple Regression Analysis

Variables $x_1, x_2, \cdots, x_k$ $\longleftrightarrow$ Random Variable Y

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k}_{\text{True regression line}} + \underbrace{\epsilon}_{\text{random error}}$$

$\updownarrow$

$$\text{Sample data : } \{(x_{1j}, x_{2j}, \cdots, x_{kj}),\ y_j\} \text{ for } j = 1, \cdots, n$$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_k x_k$$

Estimated regression line from the least-square fit