Linear Regression

Dr. Supaporn Erjongmanee

Department of Computer Engineering
Kasetsart University
fengspe@ku.ac.th

1

Outline

- Introduction
- Linear model
- Estimating model parameters
- Linear probabilistic model
- Residuals and Error sum of squares
- Total sum of squares
- Correlation
- Inferences on regression coefficient

2

Introduction

- Regression analysis is to identify relationship between two (or more) variables
- Types of model
  - Linear regression model
  - Logistic regression model
  - Non-linear regression model
- Types of variables
  - Independent variable (x)
  - Dependent variable (e.g., y)
- Example of model:
  - $\boxed{y = f(x) + \varepsilon}$

Note: ε = random deviation such that mean of ε = 0

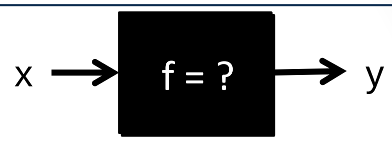
Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

3

Introduction (cont.)

- Process to find model: $y = f(x) + \varepsilon$
  - Collect (x,y)'s data
  - Use the collected data to find function f
    - We pick what type of function f would be
      - Linear
      - Logistic
      - Higher-order function
  - After obtain function f, we can use f to predict value of other x that is not in our collected data
    - Such function f => model



Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

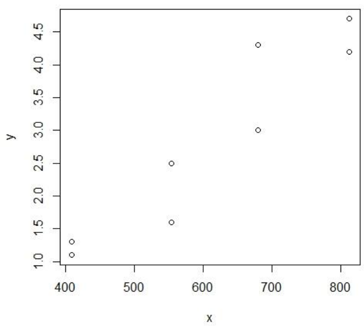
4

Visualization of (x,y) data

- Example
  - To measure global warming, we check effect of CO_2 on tree growth
  - Experiment was performed to measure how tree grew in 11 months
    - X = Atmospheric CO_2 concentration (parts per million (ppm))
    - Y = Mass of tree growth (kilogram)

x	408	408	554	554	680	680	812	812
y	1.1	1.3	1.6	2.5	3.0	4.3	4.2	4.7

- To visualize these data: Use scatter plot



Source: [1]

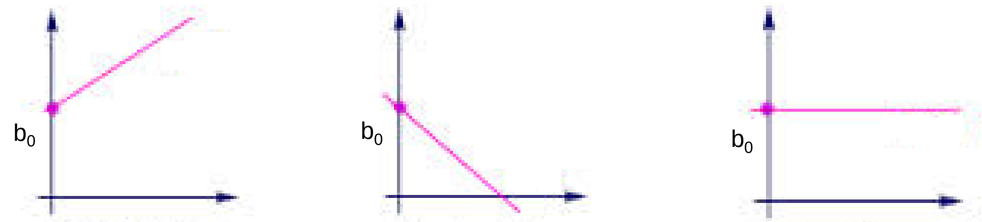
Outline

- Introduction
- Linear model
- Estimating model parameters
- Linear probabilistic model
- Residuals and Error sum of squares
- Total sum of squares
- Correlation
- Inferences on regression coefficient

Linear Model

- Function f :
  - : $Y = b_0 + b_1x$
- Visualization

b_0 = y-intercept
b_1 = slope

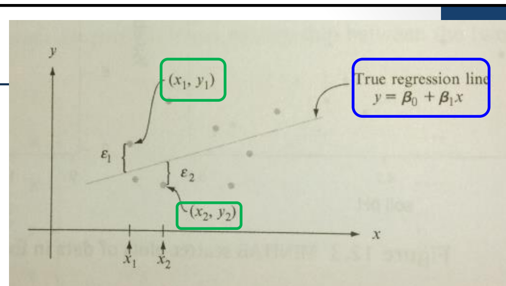


Positive slope: $b_1 > 0$	Negative slope: $b_1 < 0$	Zero slope: $b_1 = 0$

Image source: https://www0.gsb.columbia.edu/premba/analytical/images/s3/9459058040.gif

7

Linear Probabilistic Model



- True regression = Function f (deterministic):
  - $\hat{Y} = b_0 + b_1x$
- Observed data (random):
  - $Y = b_0 + b_1x + \varepsilon$

    where
    - $\varepsilon \sim N(0, \sigma^2)$ = normally distributed with $\mu = 0$, var = σ^2
      - When σ^2 is small, ε is close to zero.
        - Y is closer to true regression line
      - When σ^2 is large, ε is also large.
        - Y is far from true regression line
- Note that b_1 = rate how y increases according to x increases

b_0 = y-intercept
b_1 = slope

8

Outline

- Introduction
- Linear model
- Estimating model parameters
- Linear probabilistic model
- Residuals and Error sum of squares
- Total sum of squares
- Correlation
- Inferences on regression coefficient

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Estimating Model Parameters

- According to Gauss and Legendre, the best fit regression line is the line that has the smallest sum of errors

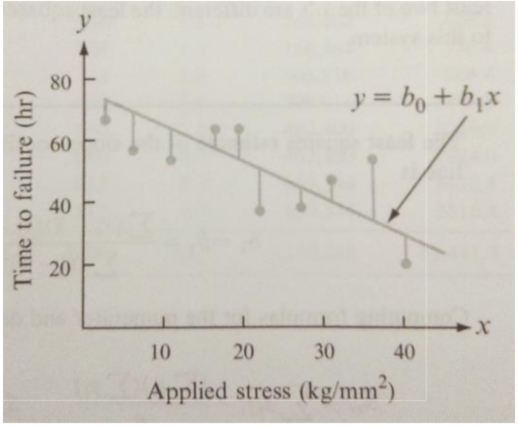


Image source: Figure 12.9 [1]

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Estimating Model Parameters (cont.)

- Sum of errors
  - $f(b_0, b_1) = \sum_{i=1}^{n}(y_i \ - (b_o + b_1 \ x_i))^2$
- Goal is to minimize sum of errors
  - Find b_0 and b_1 that results to minimum $f(b_0, b_1)$
  - Let the resulting b_0 and b_1 be $\hat{b}_0$ and $\hat{b}_1$
    - $f(\hat{b}_0 , \hat{b}_1) \leq f(b_0, b_1)$
- The estimated regression line: $y = \hat{b}_0 + \hat{b}_1 x$

Linear regression model

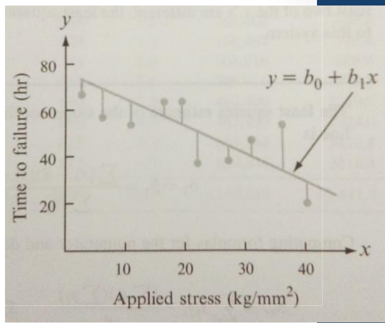


Image source: Figure 12.9 [1]

Estimating Model Parameters (cont.)

$f(b_0, b_1) = \sum_{i=1}^{n}(y_i \ - (b_o + b_1 \ x_i))^2$

- Find $\hat{b}_0$ and $\hat{b}_1$ that results to minimum $f(b_0, b_1)$

$$\boxed{\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum_{i=1}^{n} 2(y_i \ - (b_o + b_1 \ x_i))(-1) \ = 0}$$

$$-\sum_{i=1}^{n} y_i + nb_0 + b_1 \sum_{i=1}^{n} x_i \ = 0$$

$$nb_0 + b_1 \sum_{i=1}^{n} x_i \ = \sum_{i=1}^{n} y_i$$

Equation a

$$\boxed{\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum_{i=1}^{n} 2(y_i \ - (b_o + b_1 \ x_i))(-x_i \ ) \ = 0}$$

$$-\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Equation b

Estimating Model Parameters (cont.)

- Find $\hat{b}_0$ and $\hat{b}_1$ that results to minimum f(b_0, b_1)

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad \text{Equation a}$$

$$b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad \text{Equation b}$$

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

$$\boxed{\hat{b}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{b}_1 \sum_{i=1}^{n} x_i}{n}}$$

Estimating Model Parameters (cont.)

- Find $\hat{b}_0$ and $\hat{b}_1$ that results to minimum f(b_0, b_1) (cont.)

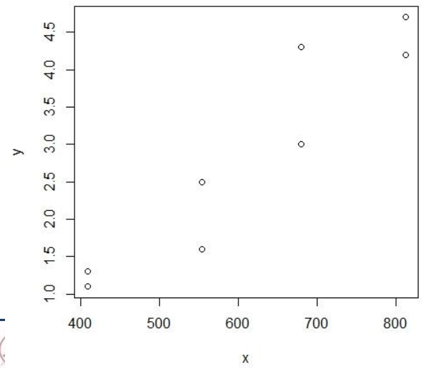
$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{b}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{b}_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{b}_1 \bar{x}$$

Example

- To measure global warming, we check effect of CO_2 on tree growth
- Experiment was performed to measure how tree grew in 11 months
  - X = Atmospheric CO_2 concentration (parts per million (ppm))
  - Y = Mass of tree growth (kilogram)

x	408	408	554	554	680	680	812	812
y	1.1	1.3	1.6	2.5	3.0	4.3	4.2	4.7



Source: [1]

Supaporn Erjongmanee
fengspe@ku.ac.th

Example(cont.)

- Use estimated regression equation to predict y value for other x

x	408	408	554	554	680	680	812	812
y	1.1	1.3	1.6	2.5	3.0	4.3	4.2	4.7

1. What is estimated tree mass ($\hat{y}$) when CO_2 concentration = 365?

2. What is estimated tree mass ($\hat{y}$) when CO_2 concentration = 315?

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Example(cont.)

- Use estimated regression equation to predict y value for other x

x	408	408	554	554	680	680	812	812
y	1.1	1.3	1.6	2.5	3.0	4.3	4.2	4.7

3. What is estimated tree mass ($\hat{y}$) when CO_2 concentration = 408?

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Outline

- Introduction
- Linear model
- Estimating model parameters
- Linear probabilistic model
- Residuals and Error sum of squares
- Total sum of squares
- Correlation
- Inferences on regression coefficient

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Example

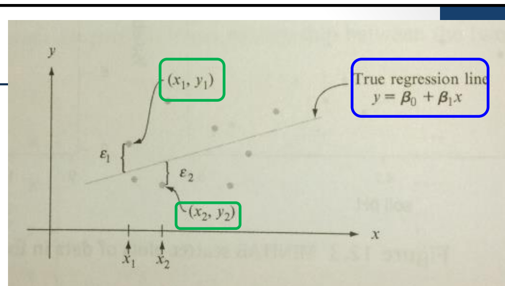
- To measure global warming, we check effect of CO_2 on tree growth
- Experiment was performed to measure how tree grew in 11 months
  - X = Atmospheric CO_2 concentration (parts per million (ppm))
  - Y = Mass of tree growth (kilogram)

x	408	408	554	554	680	680	812	812
y	1.1	1.3	1.6	2.5	3.0	4.3	4.2	4.7

Why are there multiple y's for one value of x?

Source: [1]

Linear Probabilistic Model



- True regression = Function f (deterministic):
  - $\hat{Y} = b_0 + b_1x$
- Observed data (random):
  - $\boxed{Y = b_0 + b_1x + \varepsilon}$

    where
    - $\varepsilon \sim N(0, \sigma^2)$ = normally distributed with $\mu = 0$, var = σ^2
      - When σ^2 is small, ε is close to zero.
        - Y is closer to true regression line
      - When σ^2 is large, ε is also large.
        - Y is far from true regression line

b_0 = y-intercept
b_1 = slope

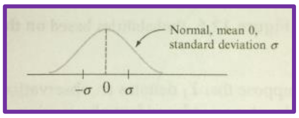
- Note that b_1 = rate how y increases according to x increases

Linear Probabilistic Model (cont.)

- For each fixed x*, corresponding $\hat{Y} = b_0 + b_1 x^* + \varepsilon$ has normal distribution

$\sim N(0, \sigma^2)$

For each x, there are multiple possible values for y.



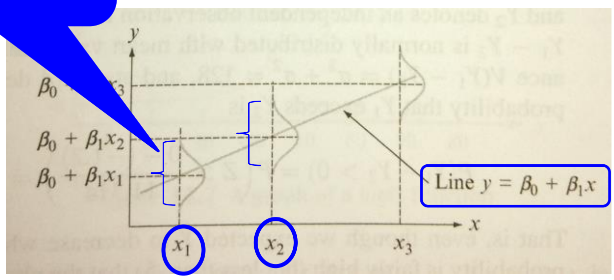


Image source: Figure 12.5 [1]

Supaporn Erjongmanee
fengspe@ku.ac.th

Statistics in Computer Engineering
Slide 21

Department of Computer Engineering
Kasetsart University

21

Linear Probabilistic Model (cont.)

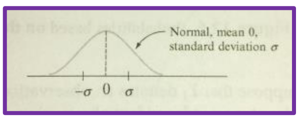
- For each fixed x*, corresponding $\hat{Y} = b_0 + b_1 x^* + \varepsilon$ has normal distribution

$\sim N(0, \sigma^2)$

For each x, there are multiple possible values for y.

Centers of each noise distribution is on the regression line.

What does this mean?



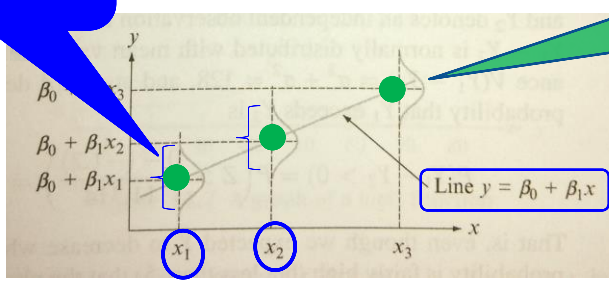


Image source: Figure 12.5 [1]

Supaporn Erjongmanee
fengspe@ku.ac.th

Statistics in Computer Engineering
Slide 22

Department of Computer Engineering
Kasetsart University

22

Linear Probabilistic Model (cont.)

- Let
  - $\mu_{Y_{x^*}}$ = E(Y|x*) = mean of Y when x = x*
  - $\sigma^2_{Y_{x^*}}$ = E(Y|x*) = variance of Y when x = x* = V(Y|x*)

Mean of noise = 0

$$\mu_{Y_{x^*}} = E(b_0 + b_1x^* + \varepsilon) = b_0 + b_1x^* + E(\varepsilon) = b_0 + b_1x^*$$

Mean of Y is on the regression line

No uncertainty

$$\sigma^2_{Y_{x^*}} = V(b_0 + b_1x^* + \varepsilon) = V(b_0 + b_1x^*) + V(\varepsilon) = 0 + \sigma^2 = \sigma^2$$

Variance of Y is same as variance of noise

Supaporn Erjongmanee
fengspe@ku.ac.th

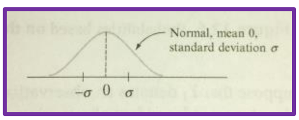
Department of Computer Engineering
Kasetsart University

Linear Probabilistic Model (cont.)

- For each fixed x*, corresponding $\hat{Y} = b_0 + b_1x^* + \varepsilon$ has normal distribution

$\sim N(0, \sigma^2)$

For each x, there are multiple possible values for y.



Does each of multiple values for y occur with the same probability?

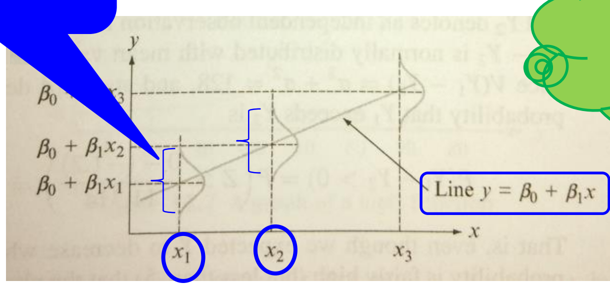


Image source: Figure 12.5 [1]

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Example (cont.)

We can find out probability of obtaining specific values for y, given specific x.

- Let true regression line be y = 65 – 1.2x and $\sigma = 8$

What is probability of obtaining y > 50, when x = 20?

$P(Y > 50 \; when \; x = 20) = ?$

$$P(Y > 50 \; when \; x = 20) = P\left(Z > \frac{50 - \mu}{8}\right)$$

What is μ?

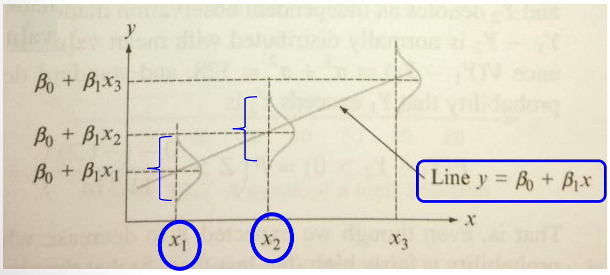


Image source: Figure 12.5 [1]

Example (cont.)

We can find out probability of obtaining specific values for y samples, given specific x.

- Let true regression line be y = 65 – 1.2x and $\sigma = 8$

What is probability of obtaining y > 50, when x = 20?

$P(Y > 50 \; when \; x = 20) = ?$

$$P(Y > 50 \; when \; x = 20) = P\left(Z > \frac{50 - \mu}{8}\right)$$

$$= P\left(Z > \frac{50 - 41}{8}\right)$$

$$= 1 - \Phi(1.13)$$

$$= 0.1292$$

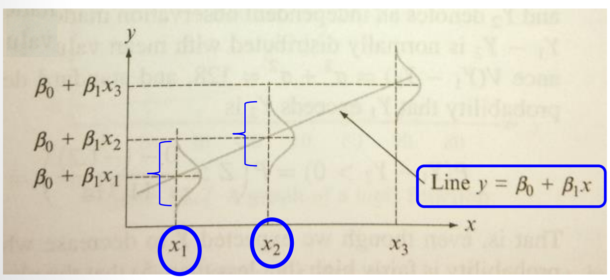


Image source: Figure 12.5 [1]

Example (cont.)

We can find out probability of obtaining specific values for y samples, given specific x.

- Let true regression line be y = 65 – 1.2x and $\sigma = 8$

What is probability of obtaining y > 50, when x = 25?

$P(Y > 50\ when\ x = 25) = ?$

$$P(Y > 50\ when\ x = 20) = P\left(Z > \frac{50 - \mu}{8}\right)$$

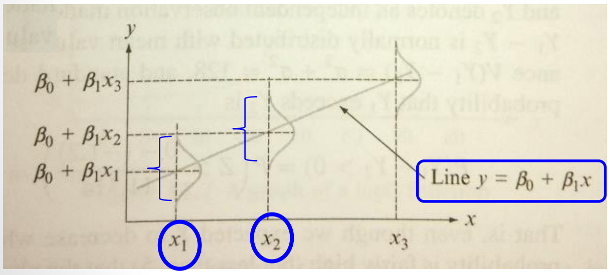


Image source: Figure 12.5 [1]

Example (cont.)

- Let true regression line be y = 65 – 1.2x and $\sigma = 8$

$$P(Y > 50\ when\ x = 20) = P\left(Z > \frac{50 - 41}{8}\right) = 1 - \Phi(1.13) = 0.1292$$

$$P(Y > 50\ when\ x = 25) = P\left(Z > \frac{50 - 35}{\sigma_{Y_{x^*}}}\right) = 1 - \Phi(1.88) = 0.0301$$

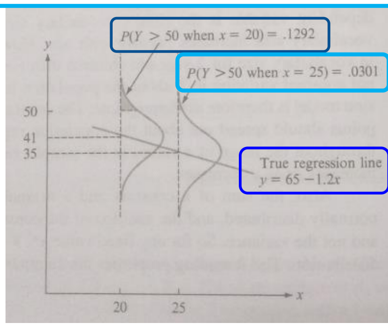


Image source: Figure 12.6 [1]

Outline

- Introduction
- Linear model
- Estimating model parameters
- Linear probabilistic model
- Residuals and Error sum of squares
- Total sum of squares
- Correlation
- Inferences on regression coefficient

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Residuals

- Deviation between and the observed data and the fitted (predicted value) = $\boxed{y_i - \hat{y}_i}$

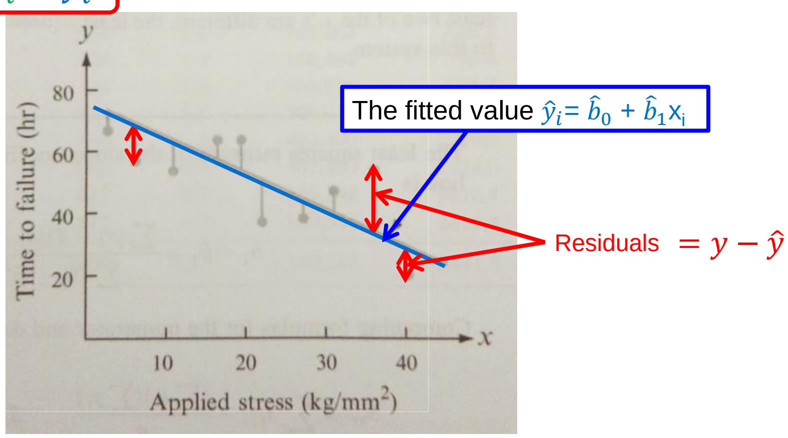


Image source: Figure 12.9 [1]

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Error Sum of Squares and Estimated Variance

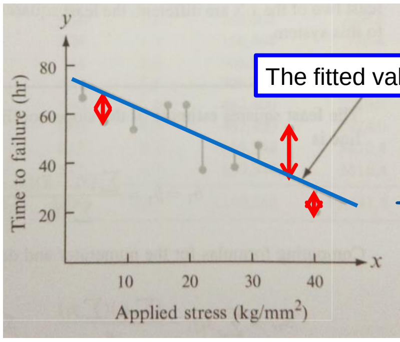
- Let SSE = Error sum of squares (or residual sum of squares)

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

Formula 1

SSE can be described as variation that cannot be explained by linear model

- Estimated variance = $\hat{\sigma}^2 = \frac{SSE}{n-2}$ ⇒ $\hat{\sigma} = s = \sqrt{\frac{SSE}{n-2}}$



The fitted value $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$

Variance of noise and SSE are related.

Image source: Figure 12.9 [1]

31

Example 2 (cont.)

- Find Residuals and SSE

x	y	$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$	Residual
125.3	77.9	78.100	-0.200
98.2	76.8	76.988	-0.188
201.4	81.5	81.223	0.277
147.3	79.8	79.003	0.797
145.9	78.2	78.945	-0.745
124.7	78.3	78.075	0.225
112.2	77.5	77.563	-0.063
120.2	77.0	77.891	-0.891
161.2	80.1	79.573	0.527
178.9	80.2	80.299	-0.099
…	…	…	…
110.7	78.6	77.501	1.099

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

SSE = $(-0.200)^2 + (-0.188)^2 + \ldots + (1.099)^2$
= 7.968

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{7.968}{20-2} = 0.4427$$

$$\hat{\sigma} = \sqrt{0.4427} = 0.665$$

On average, distance from observed and fitted values = 0.665

32

Error Sum of Squares (cont.)

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

$$= \sum_{i=1}^{n} {y_i}^2 - (\hat{b}_0 \sum_{i=1}^{n} y_i + \hat{b}_1 \sum_{i=1}^{n} x_i y_i)$$ Formula 2

- Note that formula 2 is very sensitive to decimal numbers
  - Use as many decimal points as you can

Example 2 (cont.)

SSE = $(-0.200)^2 + (-0.188)^2 + \ldots + (1.099)^2$
= 7.968

$$SSE = \sum_{i=1}^{n} {y_i}^2 - (\hat{b}_0 \sum_{i=1}^{n} y_i + \hat{b}_1 \sum_{i=1}^{n} x_i y_i)$$ Formula 2

- Note that formula 2 is very sensitive to decimal numbers
  - Use as many decimal points as you can

$$SSE = 124{,}039.6 - (\underline{72.95855} \times 1{,}574.8 + \underline{0.04103377} \times 222{,}657.9)$$

$$= 124{,}039.6 - 124{,}031.6$$

$$= 7.96799$$

$$SSE = \cancel{124{,}039.6} - (\underline{72.958} \times 1{,}574.8 - \underline{0.041033} \times 222{,}657.9)$$
$$= 9.019989$$

Outline

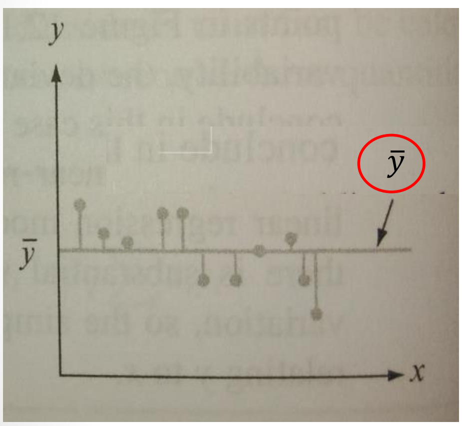
- Introduction
- Linear model
- Estimating model parameters
- Linear probabilistic model
- Residuals and Error sum of squares
- Total sum of squares
- Correlation
- Inferences on regression coefficient

Total Sum of Squares

Sum of squared difference between <u>each y</u> and <u>its average</u> (average of y's)

- Let *SST* = total sum of squares

$$= \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$



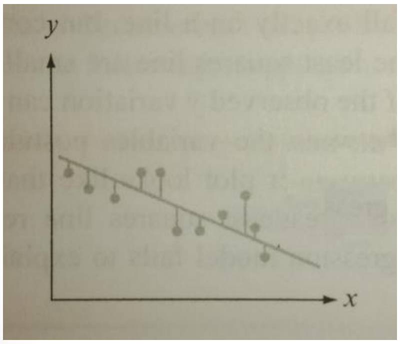
Use $\bar{y}$ as value of $\hat{y}$ for every y

Image source: Figure 12.13 [1]

SSE vs. SST

Which value is larger ? SSE or SST?

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



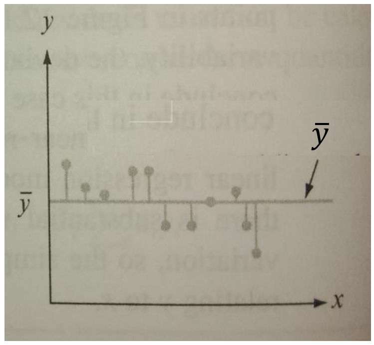


Image source: Figure 12.13 [1]

Coefficient of Determination

SSE is variation that cannot be explained by linear model

- Let r^2 = coefficient of determination

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

- Value of r^2 is between 0 and 1
- Note that SST = SSR + SSE

$$\boxed{\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

- $\frac{SSE}{SST}$ =*proportion* of total variation that *cannot be described by linear regression model*
- $r^2 = 1 - \frac{SSE}{SST}$ = *proportion* of total variation that *can be described by linear regression model*

Coefficient of Determination (cont.)

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

- $r^2 = 1 - \frac{SSE}{SST}$ = *proportion* of total variation that *can be described by linear regression model*
  - The higher r^2, the better that linear regression model can explain variation of data
  - When r^2 is small, then linear regression model may not be appropriate.
- r^2 = proportion that SSE is reduced by linear regression line ($\hat{y}$) compared to average y ($\bar{y}$)

Example: SSE = 2, SST = 20, $r^2 = 1 - (2/20) = 0.90$
Hence, regression reduces SSE by 90%

Example : Tree Growth vs. CO_2 (cont.)

Sample#	x	y	x^2	xy	y^2
Sum	4908	22.7	3,190,248	15,441.4	78.93

$\hat{b}_1 = 0.00845443$ $\quad\hat{b}_0 = -2.349293$ $\quad n = 8$

$$\boxed{SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} {y_i}^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}$$

$$SST = 78.93 - \frac{22.7^2}{8} = 14.519$$

$$\boxed{SSE = \sum_{i=1}^{n} {y_i}^2 - (\hat{b}_0 \sum_{i=1}^{n} y_i + \hat{b}_1 \sum_{i=1}^{n} x_i y_i) \quad (2)}$$

$$SSE = 78.93 - [(-2.349293)(22.7) + (0.00845443)(15{,}441.4)] = 1.711$$

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{14.519}{1.711} = 0.882$$

88.2% of observed variation can be explained by linear regression model

Outline

- Introduction
- Linear model
- Estimating model parameters
- Linear probabilistic model
- Residuals and Error sum of squares
- Total sum of squares
- Correlation
- Inferences on regression coefficient

Correlation

Note that correlation does not imply causation

- Paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- How strongly x's and y's are related to each other
- Sample correlation coefficient (r)

What is magnitude of change per one unit change of X and Y?

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

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Characteristics of Correlation

- Value of *r* does not depend which data is labeled x or y

Different from regression analysis

x is independent variable. Y is not.

- Value of *r* does not depend on unit of x or y
- Value of *r* is between [-1, 1]
- *r* = 1 if and only if <u>all</u> (x_i, y_i) are on the line with positive slope
- *r* = -1 if and only if <u>all</u> (x_i, y_i) are on the line with negative slope
- r^2 is coefficient of determination

43

Correlation (cont.)

$$\boxed{r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \sqrt{\frac{SSR}{SST}}}$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(\hat{b}_1 x_i + \hat{b}_0 - \hat{b}_1\bar{x} + \hat{b}_0)^2 = \sum_{i=1}^{n}(\hat{b}_1 x_i - \hat{b}_1\bar{x})^2$$

$$= \hat{b}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

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

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

$$= \boxed{r^2} SST$$

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

Coefficient of determination

44

Strong vs. Weak Correlation

- What is value of r to identify weak or strong correlation?
  - $0.8 \leq |r| \leq 1$: Strong
  - $0 \leq |r| \leq 0.5$: Weak

Why r = 0.5 is weak?

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Outline

- Introduction
- Linear model
- Estimating model parameters
- Linear probabilistic model
- Residuals and Error sum of squares
- Total sum of squares
- Correlation
- Inferences on regression coefficient

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Slope of Linear Regression

- $\hat{b}_1$ indicates linear relationship between x and y
- How much do we know about $\hat{b}_1$?
  - Distribution of $\hat{b}_1$
  - Variance of $\hat{b}_1$
  - Hypothesis test on $\hat{b}_1$

Distribution of $\hat{b}_1$

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

$\bar{y}\sum_{i=1}^{n}(x_i-\bar{x}) = \bar{y}[(\sum_{i=1}^{n}x_i) - n\bar{x}] = 0$

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

Constants for each y_i

$$\hat{b}_1 = \sum_{i=1}^{n} c_i Y_i \quad \text{where } c_i = \frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{x_i-\bar{x}}{Sxx}$$

$\hat{b}_1$ is a linear of Y_i

Each Y_i's are normally distributed.

Hence, $\hat{b}_1$ is also normal distributed

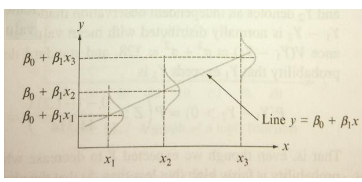
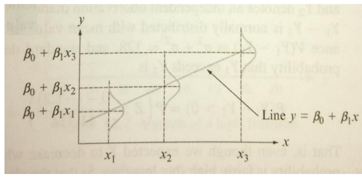


Image source: Figure 12.5 [1]

Variance of $\hat{b}_1$

$$\hat{b}_1 = \sum_{i=1}^{n} c_i Y_i \quad \text{where} \quad c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{Sxx}$$

$$Var(\hat{b}_1) = \sigma_{\hat{b}_1}^2 = \frac{Var(Y)}{S_{xx}} = \frac{\sigma^2}{S_{xx}}$$



$$\sigma_{\hat{b}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

$$s_{\hat{b}_1} = \frac{s}{\sqrt{S_{xx}}}$$

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Hypothesis Testing of $\hat{b}_1$

σ, s: standard deviation of noise $\varepsilon \sim N(0, \sigma^2)$
$\sigma_{\hat{b}_1}, s_{\hat{b}_1}$: standard deviation of $\hat{b}_1$

- Null hypothesis (H_0): $b_1 = b$
- Test statistic = $t = \frac{\hat{b}_1 - b}{s_{\hat{b}_1}}$

Note: $s_{\hat{b}_1} = \frac{s}{\sqrt{Sxx}} = \frac{\sqrt{SSE/(n-2)}}{\sqrt{Sxx}}$

Alternative Hypothesis	Rejection Region at α level
$H_a : b_1 > b$	$t \geq t_{\alpha, n-2}$
$H_a : b_1 < b$	$t \leq -t_{\alpha, n-2}$
$H_a : b_1 \neq b$	Either $t \leq -t_{\alpha/2, n-2}$ or $t \geq t_{\alpha/2, n-2}$

Model utility test:
- When $b = 0$, we test H_0: $b_1 = 0$.
- If H_0 is true, this means that linear regression model: $y = b_0$ only
  - Equivalently, model does not depend on x.
- If H_0 is rejected, r^2 will be large. Linear model is appropriate for use.

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Example : Tree Growth vs. CO_2 (cont.)

Sample#	x	y	x^2	xy	y^2
Sum	4908	22.7	3,190,248	15,441.4	78.93

Test statistic $= t = \frac{\hat{b}_1 - b}{s_{\hat{b}_1}}$

$$s_{\hat{b}_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{\sqrt{SSE/(n-2)}}{\sqrt{S_{xx}}}$$

$\hat{b}_1 = 0.00845443$ $\quad \hat{b}_0 = -2.349293$ $\quad n = 8$

$SST = 14.519$ $\quad SSE = 1.710705$

$$s = \sqrt{\frac{SSE}{n-2}} = 0.533964$$

$$Sxx = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 179,190$$

$$s_{\hat{b}_1} = \frac{0.533964}{\sqrt{179,910}} = 0.001261407$$

Example : Tree Growth vs. CO_2 (cont.)

Test statistic $= t = \frac{\hat{b}_1 - b}{s_{\hat{b}_1}}$

$$s_{\hat{b}_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{\sqrt{SSE/(n-2)}}{\sqrt{S_{xx}}}$$

- H_0: $b_1 = 0$
- H_a: $b_1 \neq 0$

$s_{\hat{b}_1} = \frac{0.533964}{\sqrt{179,910}} = 0.001261407$ $\quad \hat{b}_1 = 0.00845443$

$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.00845443}{0.001261407} = 6.702386$ ➡ p-value = 0.0005355

Given $\alpha = 0.05, n = 6,$ then
Rejection region: t ≥ 2.447 or t ≤ - 2.447

Given $\alpha = 0.05, p < \alpha$

Reject H_0

b1 is not zero. ⇨ X linearly relates with Y. ⇨ X affects Y.

Example : Tree Growth vs. CO_2 (cont.)

- $H_0: b_1 = 0$
- $H_a: b_1 \neq 0$

$$s_{\hat{b}_1} = \frac{0.533964}{\sqrt{179{,}910}} = 0.001261407 \qquad \hat{b}_1 = 0.00845443$$

$$t = \frac{\hat{b}_1 - b}{s_{\hat{b}_1}} = \frac{0.00845443}{0.001261407} = 6.702386$$

p-value = 0.0005355

```
> model <- lm(y ~ x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73446 -0.33671  0.08271  0.18819  0.90028

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.349295   0.796567  -2.949 0.025637 *
x            0.008454   0.001261   6.702 0.000536 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.534 on 6 degrees of freedom
Multiple R-squared:  0.8822,    Adjusted R-squared:  0.8625
F-statistic: 44.92 on 1 and 6 DF,  p-value: 0.0005355
```

Given $\alpha = 0.05, n = 6,$ then
Rejection region: t ≥ 2.447 or t ≤ - 2.447

Reject H_0

b1 is not zero.

Example : Tree Growth vs. CO_2 (cont.)

Sample#	x	y	x²	xy	y²
Sum	4908	22.7	3,190,248	15,441.4	78.93

Test statistic $= t = \frac{\hat{b}_1 - b}{s_{\hat{b}_1}}$

$$s_{\hat{b}_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{\sqrt{SSE/(n-2)}}{\sqrt{S_{xx}}}$$

$\hat{b}_1 = 0.00845443 \quad \hat{b}_0 = -2.349293 \quad n = 8$

$SST = 14.519 \quad SSE = 1.710705$

$$s = \sqrt{\frac{SSE}{n-2}} = 0.533964$$

$$Sxx = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 179{,}190$$

$$s_{\hat{b}_1} = \frac{0.533964}{\sqrt{179{,}910}} = 0.001261407$$

```
> model <- lm(y ~ x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73446 -0.33671  0.08271  0.18819  0.90028

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.349295   0.796567  -2.949 0.025637 *
x            0.008454   0.001261   6.702 0.000536 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.534 on 6 degrees of freedom
Multiple R-squared:  0.8822,    Adjusted R-squared:  0.8625
F-statistic: 44.92 on 1 and 6 DF,  p-value: 0.0005355
```

ANOVA Relationship with $\hat{b}_1$

- H_0: $b_1 = 0$
- H_a: $b_1 \neq 0$

	df	Sum of Squares (SS)	Mean Square (MS)	f
Regression	1	SSR	SSR	$\frac{SSR}{SSE/(n-2)}$
Error	n-2	SSE	S^2 = SSE / (n-2)	
Total	n-1	SST		

- Reject H_0 if f ≥ $F_{\alpha, 1, n-2}$

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Example : Tree Growth vs. CO_2 (cont.)

- H_0: $b_1 = 0$
- H_a: $b_1 \neq 0$

$SST = 14.519$ $SSE = 1.710705$

	df	Sum of Squares (SS)	Mean Square (MS)	f
Regression	1	**14.51875-1.710705 = 12.80804**	**12.80804**	**44.92197**
Error	**6**	**1.710705**	**0.2851176**	
Total	**7**	**14.51875**		

- Given α = 0.05, $F_{0.05, 1, 6}$ = 5. 987378 ➡ p-value = 0.000535
- Reject H_0
- b1 is not zero. ⇨ X linearly relates with Y. ⇨ X affects Y.

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Example : Tree Growth vs. CO_2 (cont.)

```
> model <- lm(y ~ x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73446 -0.33671  0.08271  0.18819  0.90028

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.349295   0.796567  -2.949 0.025637 *
x            0.008454   0.001261   6.702 0.000536 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.534 on 6 degrees of freedom
Multiple R-squared:  0.8822,    Adjusted R-squared:  0.8625
F-statistic: 44.92 on 1 and 6 DF,  p-value: 0.0005355
```

ı Square	f
804	44.92197
1176	

- Given $\alpha = 0.05$, $F_{0.05, 1, 6} = 5.987378$ ➡ p-value = 0.000535
- Reject H_0
- b1 is not zero

References

1. J.L. Devore and K.N.Berk, Modern Mathematical Statistics with Applications, Springer, 2012.