Chapter 11 & 2- Simple Linear Regression Model and Correlation

STAT 251

Lecture 35

Correlation, Simple Linear Regression, coefficient of determination
Least Squares Method
Examples

Dr. Lasantha Premarathna

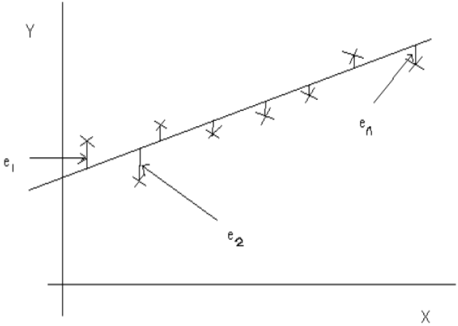
Chapter 11 & 2 - Learning Outcomes

- Scatter plot
- Covariance & Correlation
- Simple linear regression
- Least squares estimates in simple linear regression
- Interpret the parameters in a fitted linear model
- Inference for the slope parameter - Confidence interval & hypothesis testing

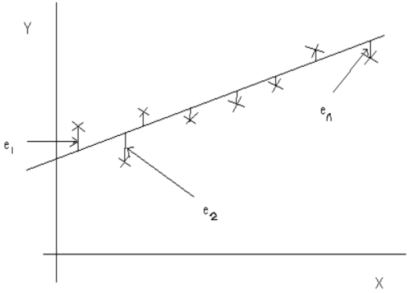
Residuals

- For each point (x_i, y_i)
  - e_i: **vertical** distance from the point to the line fitted
    - point above the line → e_i is positive
    - point below the line → e_i is negative
    - point on the line → e_i is zero

- **residual** $= e_i = y_i - \hat{y}_i$



Regression line



- Regression line minimizes the sum of the **squares** of the errors
  - i.e., $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = e_1^2 + e_2^2 + \cdots + e_n^2$
- **Least squares Regression Line**

  The least squares regression line is the line that minimizes the residual sum of squares.

Applet- Guess the Least Squares Regression Line

Guess the Least Squares Regression Line

⇒ https://www.geogebra.org/m/ZWSy5SxE

Least Squares Method

- Consider residual sum of squares

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

- The least squares regression line is the line that minimizes the residual sum of squares.

$$\Rightarrow \quad \text{minimize} \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

consider $\hat{y} = b_0 + b_1 x$

then,

$$\Rightarrow \quad f(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x)]^2$$

Least Squares Method

$$\Rightarrow \quad f(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x)\right]^2$$

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1 and the equating them to zero.

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

$$\Rightarrow \quad n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad ----(1)$$

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

$$\Rightarrow \quad b_0 \sum_{1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad ----(2)$$

- solve these two equations to get b_0 and b_1

Least Squares Estimates

- The least squares estimate of the slope coefficient β_1 of the regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2} = \frac{r s_y}{s_x}$$

  - r: sample correlation coefficient,
  - s_y and s_x: sample standard deviations,
  - $\bar{x}$ and $\bar{y}$: sample means

- The least squares estimate of the intercept β_0 of the regression line is

$$b_0 = \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Regression line always passes through the point $(\bar{x}, \bar{y})$

Interpreting the intercept & slope

- Intercept
  - There predicted value for y when $x = 0$
  - helps in plotting the line
  - may not have any interpretative value if no observations had x value near 0

- Slope
  - slope measures the change in the predicted variable(y) for a 1 unit increase in the explanatory variable (x).

Regression Line

- At a given vale of x, the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (i.e. $\hat{y} = b_0 + b_1 x$)
  - predicts a single value of the response variable
  - But, we should not expect all subjects at that value of x to have the same value of y
    - variability occurs in the y values

- Regression line connects the estimated means of y at the various x values.

- That is, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (i.e. $\hat{y} = b_0 + b_1 x$) describes the relationship between x and the estimated means of y at the various values of x.

Coefficient of Determination (r^2)

Coefficient of determination (r^2) - squared correlation

- r^2 is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model.

- The higher the value of r^2, the more successful is the simple linear model in explaining y variation.

Ex: consider correlation between x and y is 0.9. Then,

$$r^2 = 0.9^2 = 0.81$$

$$\Rightarrow \quad r^2 = 81\%$$

$\Rightarrow$ 81% of the variation in y values can be explained by the linear relationship between y and x

Estimating (σ^2) and Sum of Squares

- **Error Sum of Squares (residual sum of squares)** denoted by SSE, is

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x) \right]^2$$

- The estimate of σ^2 is

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

- SSE can be interpreted as a measure of how much variation in y is left unexplained by the model.

Sum of Squares in Simple Linear Regression

- **Total Sum of Squares (SST)**

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

  SST is the sum of squared deviation about the sample mean of the observed y values.

- **Regression Sum of Squares (SSR)**

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

  SSR is interpreted as the amount of total variation that is explained by the model.

Sum of Squares and Coefficient of Determination

- We have the following relation

$$SST = SSR + SSE$$

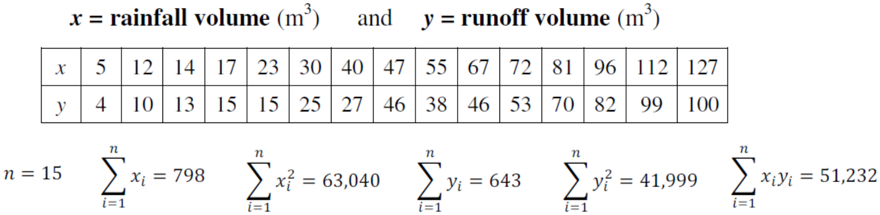
- Coefficient of Determination can be given using SST, SSR, and SSE.

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

$$\Rightarrow \quad r^2 = \frac{SSR}{SST}$$

Example:

The article ”Characterization of Highway Runoff ” for a particular location in BC gave following data and summaries



x = rainfall volume (m^3) and **y = runoff volume** (m^3)

x	5	12	14	17	23	30	40	47	55	67	72	81	96	112	127
y	4	10	13	15	15	25	27	46	38	46	53	70	82	99	100

$$n = 15 \qquad \sum_{i=1}^{n} x_i = 798 \qquad \sum_{i=1}^{n} x_i^2 = 63{,}040 \qquad \sum_{i=1}^{n} y_i = 643 \qquad \sum_{i=1}^{n} y_i^2 = 41{,}999 \qquad \sum_{i=1}^{n} x_i y_i = 51{,}232$$

(a) Does a scatter plot of the data support the use of the simple linear regression model? Predict the value of correlation coefficient r using the scatter plot.

(b) Calculate the correlation coefficient r.

(c) Calculate point estimate of the slope and intercept of the population regression line.

Example:

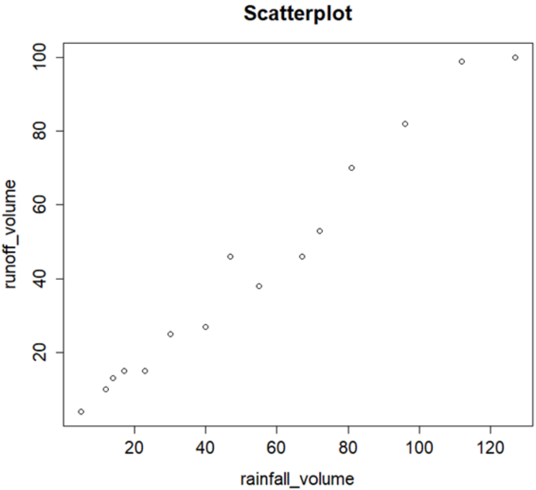
(d) Interpret the slope of the least squares line.

(e) Calculate point estimate of the true average runoff volume when rainfall volume is 50.

(f) Calculate the residuals corresponding to the last two observations.

(g) What proportion of observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall?

Example: Scatterplot



Example: Solutions

(a) Does a scatter plot of the data support the use of the simple linear regression model? Predict the value of correlation coefficient r using the scatter plot.

Yes, the scatter plot shows a strong positive linear relationship between rainfall volume and runoff volume. Therefore it supports the use of simple linear regression model. Since the relationship is strong positive linear relationship, r can be a value greater than 0.9

Example: Solutions

(b) Calculate the correlation coefficient r.

$$r = \frac{\mathrm{cov}(x,y)}{s_x s_y} = ?$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}} = \sqrt{\frac{63040 - 15\left(\frac{798}{15}\right)^2}{15-1}} = 38.35$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\left(\sum_{i=1}^{n} y_i^2\right) - n\bar{y}^2}{n-1}} = \sqrt{\frac{41999 - 15\left(\frac{643}{15}\right)^2}{15-1}} = 32.11$$

$$cov(x,y) == \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{n-1} = \frac{51232 - 15\left(\frac{798}{15}\right)\left(\frac{643}{15}\right)}{15-1} = 1216.029$$

$$\Rightarrow \quad r = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{1216.029}{(38.35)(32.11)} = 0.987$$

Example: Solutions

(c) Calculate point estimate of the slope and intercept of the population regression line.

$$b_1 = \hat{\beta}_1 = \frac{r\ s_y}{s_x} = \frac{0.987\ (32.11)}{38.35} = 0.826$$

$$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{643}{15} - 0.826 \left(\frac{798}{15} \right) = -1.077$$

estimated regression line $\Rightarrow \quad \hat{y} = -1.077 + 0.826\ x$

(d) Interpret the slope of the least squares line.

When the rainfall volume increase by 1 unit ($1\ m^3$), the runoff volume is increased by $0.826\ m^3$

Example: Solutions

(e) Calculate point estimate of the true average runoff volume when rainfall volume is 50.

$$\text{when } x = 50 \quad \Rightarrow \quad \hat{y} = -1.077 + 0.826(50) = 40.22\ m^3$$

(f) Calculate the residuals corresponding to the last two observations.

$$\text{residual} = y_i - \hat{y}_i$$

$$\text{when } x = 112 \quad \Rightarrow \quad \hat{y} = -1.077 + 0.826(112) = 91.435\ m^3$$

$$\Rightarrow \text{residual} = 99 - 91.435 = 7.565$$

$$\text{when } x = 127 \quad \Rightarrow \quad \hat{y} = -1.077 + 0.826(127) = 103.825\ m^3$$

$$\Rightarrow \text{residual} = 100 - 103.825 = -3.825$$

Example: Solutions

(g) What proportion of observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall?

From part (b) we have $r = 0.987$

$$\Rightarrow r^2 = (0.987)^2 = 0.974 \;\Rightarrow\; 97.4\%$$

Therefore, 97.4% of the observed variation in runoff volume can be explained by the simple linear regression relationship between runoff volume and rainfall volume.

Before the next class ...

Visit the course website at canvas.ubc.ca

- Review Lecture 35 and related sections in the text book
- Topic of next class: **Chapter 11: Inference Simple Linear Regression**