Chapter 11 & 2- Simple Linear Regression Model and Correlation

STAT 251

Lecture 34

Scatterplots, Covariance, Correlation
Simple Linear Regression

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

Correlation Coefficient (r)

- Measures the strength and direction of the linear association between x and y
- Sample correlation coefficient is defined by

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

$$\text{where} \quad s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \; and \; s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$\Rightarrow \quad r = \frac{\text{Cov}(x,y)}{s_x s_y}$$

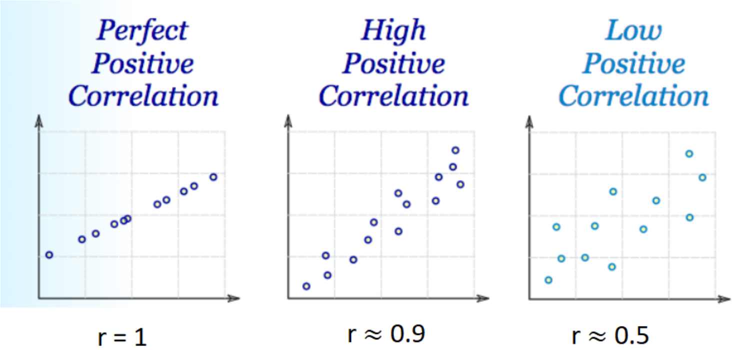
a positive r value indicates a positive association

a negative r value indicates a negative association

r value close to 0 indicates a weak linear association

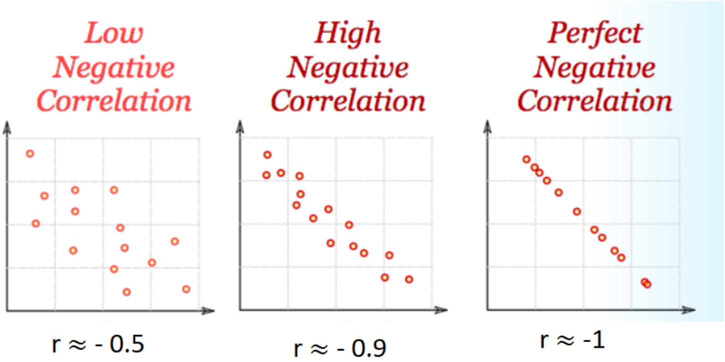
Correlation

Positive Correlation



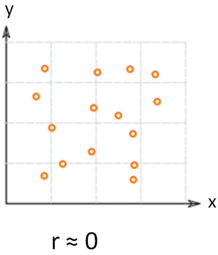
Correlation

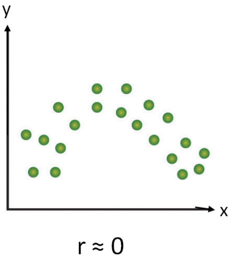
Negative Correlation

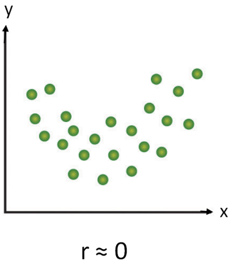


Correlation

No Correlation







Properties of Correlation

- Always falls between -1 and +1, i.e. $-1 \leq r \leq 1$
- Sign of correlation denotes the direction
  (-) indicates negative linear association
  (+) indicates positive linear association.
- Correlation has no units and does not change when we change the units of measurement of x, y or both
- Two variables have the same correlation no matter which is treated as the response variable.

Facts about Correlation

Cautions:

- Correlation requires that both variables be quantitative
- Correlation does not describe curved relationships between variables, no matter how strong the relationship is between them.
- Correlation is not resistant; r is strongly affected by a few outlying observations
- Correlation is not a complete summary of two-variable data.

Does Correlation imply causation?

Data for following variables are available for all fires in greater Vancouver area for last two years.

x is the number of fire fighters at the fire

y is the cost of damage due to the fire

Suppose that the scatter plot shows a positive linear association between two variables. does this mean that having more firefighters at a fire causes damages to be worse?

(A) Yes

(B) No

Does Correlation imply causation?

Data for following variables are available for all fires in greater Vancouver area for last two years.

x is the number of fire fighters at the fire

y is the cost of damage due to the fire

Suppose that the scatter plot shows a positive linear association between two variables. does this mean that having more firefighters at a fire causes damages to be worse?

- Identify a third variable (lurking variable) that could be considered a common cause of x and y?

  a) Distance from the fire station

  b) Intensity of the fire

Lurking Variable

- A lurking variable is a variable, usually unobserved, that influences the association between the variables of primary interest.

Difference between correlation and causation

- **Correlation** means there is a relationship or pattern between the values of two quantitative variables.
- **Causation** means that one event causes another event to occur

Question: Causation can be determined from

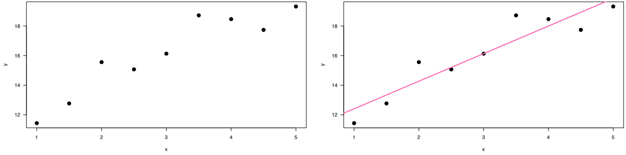
(A) an observational study

(B) an appropriately designed experiment.

(C) Both A and B

Chapter 11 - Simple Linear Regression Model

The regression line is a line that best describe the relationship between X and Y. Linear regression consists of finding the best-fitted straight line through the points.

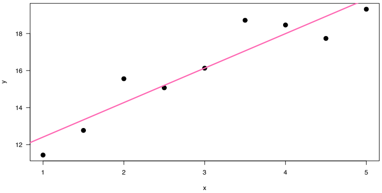


- We can suggest that a **linear model** exists between X and Y?
- A linear model implies the following relationship:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  - where:
    - ε: error term and $\varepsilon \sim \mathrm{N}(0, \sigma^2)$
    - Y and ε are random
    - β_0, β_1, and σ^2 are parameters

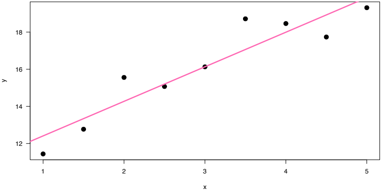
Chapter 11 - Simple Linear Regression Model



- We have bivariate data, (x_i, y_i)
  - (x): explanatory variable
  - (y): response variable

- Linear regression
  - **Linear model:** $Y = \beta_0 + \beta_1 X + \epsilon$
  - **Regression line (Y on X):** $\hat{y} = b_0 + b_1 x$
    - ★ b_0 is an estimate of the population intercept, β_0
    - ★ b_1 is an estimate of the population slope, β_1
  - Line $E(Y) = \beta_0 + \beta_1 X$ is called the true (or population) regression line.

Simple Linear Regression Model

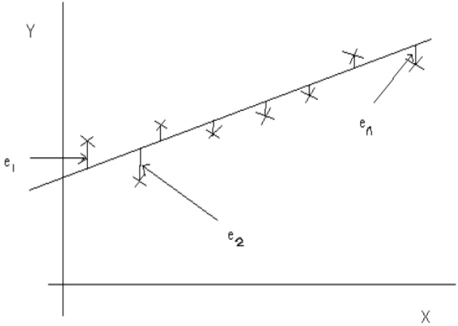


$$Y = \beta_0 + \beta_1 X + \varepsilon$$

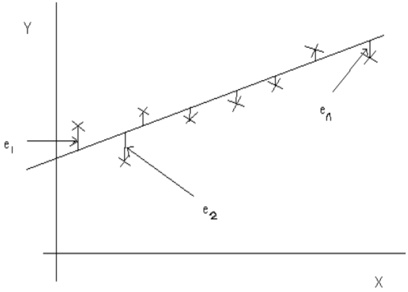
- The regression line: $\hat{y} = b_0 + b_1 x$
  - this is the best fitted-line constructed from data using the **least squares method**.
- Why do we need an error term?
  - Unlikely that the line will fit **exactly** to real data
- We assume that ε has a Normal distribution with mean zero
  - $\varepsilon \sim \mathrm{N}(0, \sigma^2)$

Residuals

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

$\Rightarrow$ 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

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

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.

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

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

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

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

- 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 successfull is the simple linear model in explaining y variation.

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

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

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

$\Rightarrow \quad$ 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.

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

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

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.

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

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

Before the next class ...

Visit the course website at canvas.ubc.ca

- Review Lecture 34 and related sections in the text book
- Topic of next class: **Simple Linear Regression, Examples**