Chapter 4: Describing the Relation Between Two Variables

Univariate Data v.s. Bivariate Data

Section 4.1: Scatter Diagrams and Correlation

- **Scatter Diagrams: For two quantitative variables.**
- **Response Variable v.s. Explanatory Variable**
  (Dependent Variable) (Predictor Variable)
- **Correlation vs. Causation**
  In observation studies, it is usually not appropriate to interpret observed correlation as causation. There may be lurking variables that influence both variables and therefore can better explain the seen correlation.
- **Linearly Related**
  - **Positively Associated**
  - **Negatively Associated**
- **Linear Correlation Coefficient (Pearson Product Moment Correlation Coefficient)**
  - **Population Correlation Coefficient ρ ("rho")**

$$\rho = \frac{\sum\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)}{N} = \frac{\sum(x-\mu_x)(y-\mu_y)}{\sqrt{\sum(x-\mu_x)^2}\sqrt{\sum(y-\mu_y)^2}}$$

$$= \frac{N\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{N\left(\sum x^2\right)-\left(\sum x\right)^2}\sqrt{N\left(\sum y^2\right)-\left(\sum y\right)^2}}$$

  - **Sample Correlation Coefficient r**

$$r = \frac{\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}{n-1} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

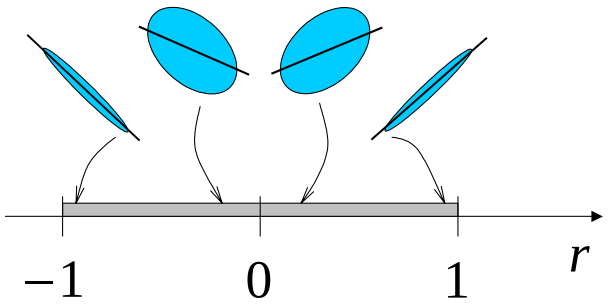
$$= \frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right)-\left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right)-\left(\sum y\right)^2}}$$

- **Linear correlation coefficients are unitless. They don't change when the unit of measurement for either x or y is changed.**
- **The Linear correlation coefficient doesn't change when the roles of x and y are exchanged.**

- **Interpretation of the Linear Correlation Coefficient. (Applet "Correlation by Eye")**

$$-1 \le r \le 1$$

  - **When r is close to 0 → week linear correlation.**
  - **When r is close to 1 → strongly and positively associated.**
  - **When r is close to -1 → strongly and negatively associated.**



**** Try Applet "Correlation by Eye"**
**** Pay attention to the point made by Figure 3 (p. 179) in comparison with Figure 1.**

- **Determine Whether There Is a Linear Correlation: Table VIII (page A-14)**

- **Linear correlation coefficients should be used in conjunction with scatter diagrams in order to provide reliable analysis. (See Problem 35 in Section 4.1, page 191.)**

- Accuracy Issues in Computing the Linear Correlation Coefficient. Please keep as many decimal places as possible for all numbers at intermediate steps. Don't do rounding until the very end!!!

- **The effect of outliers.**

Section 4.2: Least-Squares Regression

- **Residuals**
- **Least-Squares Regression Criterion: Minimize the "sum of squares of errors", i.e. "sum of squares of residuals".**
- **Formula**

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

Important Property:

This regression line passing through $(\bar{x}, \bar{y})$ **!!!!!**

(1) Calculate b_1 using

$$b_1 = r \cdot \frac{s_y}{s_x}$$

(2) Determine b_0 by plugging in $(\bar{x}, \bar{y})$ and solve for b_0

$$b_0 = \bar{y} - b_1 \bar{x}$$

- **Important: Don't extrapolate. (Don't use the regression equation to make predictions outside the scope of the model!)**

Section 4.3: The Coefficient of Determination

Total Deviation = (Unexplained Deviation) + (Explained Deviation)

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$$

Total Variation = (Unexplained Variation) + (Explained Variation)

$$\sum (y - \bar{y})^2 = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$$

$$1 = \frac{\text{Unexplained Variation}}{\text{Total Variation}} + \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$1 = \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} + \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

R^2 means $\dfrac{\text{Explained Variation}}{\text{Total Variation}}$

Fact: $R^2 = r^2$ $\qquad 0 \le R^2 \le 1$