MTH 23.5 LECTURE NOTES (Ojakian)

Topic 20: Correlation and Scatter Diagrams

OUTLINE

References (**Algebra Book**: None; **Statistics Book**: 12.2, 12.3)

1. Correlation
2. Best-Fit Lines

---

1. How are two variables related?

   (a) Example: Guilded Exercise 1 (ch. 4, p. 122, from 5th edition): Look at just table of numbers.

   (b) Two variables are correlated if: The value of one variable can be used to predict the value of the other variable.

   (c) Goal: Determine how correlated two variables are.

   PROBLEM 1. *In the example, guess the work hours lost for various choices of training hours.*

2. Scatter Diagram

   PROBLEM 2. *Verify the scatter plot of data for guilded exercise.*

   (a) Terminology

      i. Horizontal axis: Explanatory variable

      ii. Vertical axis: Response variable

      iii. Correlation ...

   (b)

   PROBLEM 3. *Make a scatter plot for the following data:*

   $$X : 4, \quad 7, \quad 8, \quad 12, \quad 17$$

   $$Y : 2, \quad 5, \quad 10, \quad 11, \quad 20$$

   Does the data look "correlated"? What is its rough shape?

3. <u>Correlation Coefficient</u>

   (a) How good is the Best-Fil line? ...
   Correlation Coefficient = `Correl`([column 1], [column 2])

   (b) Measures how close to a line the scatter plot looks. Denoted r.

      i. It is between -1 and 1, inclusive.
      ii. If r close to 0: Little or no linear correlation.
      iii. If r close to +1: Positive correlation
      iv. If r close to -1: Negative correlation

   (c)

      PROBLEM 4.

      i. *Make up a table of two columns of data, with at least 10 individuals and find the correlation coefficient. Try to choose the data so that r is close to 0.9.*
      ii. *Make up a table of two columns of data, with at least 10 individuals and find the correlation coefficient. Try to choose the data so that r is close to* -0.9.
      iii. *Make up a table of two columns of data, with at least 10 individuals and find the correlation coefficient. Try to choose the data so that r is close to* 0.

   PROBLEM 5. *Pick two variables from class data that you think might be correlated and check.*

4. <u>Correlation versus Causation</u>

   “Correlation does not imply causation!”

   (a) **Lurking variable (or hidden variable)**: A third variable (not X or Y) that is simultaneously responsible for the changes in X and Y.

   (b)

   PROBLEM 6. *From section 4.1 (5th edition) do problems: 8, 9.*

   (c) See webpage: http://www.tylervigen.com/spurious-correlations

5. <u>Best-Fit Line</u>

   (a) Rough Definition: It is the line that is simultaneously as close as possible to all the data.

   (b) Precise Definition: The line that minimizes the sum of the squares of the vertical distances between the data and the line.

   (c) Finding using Excel.

      i. First make scatter plot
      ii. Select scatter plot
      iii. `Layout` → `Trendline` → `Linear Trendline`
      iv. Find for examples above, along with the correlation coefficient.

6. Calculate r by hand

Follow the handout. Summary of steps:

(a) Find the mean for each list of data

(b) Find Standard deviation for each list of data

(c) Find z-scores: (value - mean) / (sample standard deviation)

(d) Find products of z-scores

(e) Find the sum of these products

(f) Divide by $n-1$ where n is the number of individuals

(g)

PROBLEM 7. *Compute the sample correlation coefficient for the followed paired data:*

$$X = 7, 5, 3 \quad \textit{and} \quad Y = 30, 20, 10.$$

PROBLEM 8. *Compute the sample correlation coefficient for the followed paired data:*

$$X = 1, 3, 5 \quad \textit{and} \quad Y = 10, 5, 0$$

PROBLEM 9. *Suppose we have paired data where the z-values of the first data are:*

$$-0.6,\ -0.3,\ -1.2,\ 0.6,\ 1.3,$$

and the z-values for the second list of data are:

$$-0.9,\ -0.4,\ -0.7,\ 0.9,\ 1.2.$$

i. *What is the sample correlation coefficient?*

ii. *What is r if all the first data are negated and the second remain the same? (do this without further calculation)*

iii. *In the first data list, which z-value corresponds to the data item furtherest from the mean and which corresponds to the one closest to the mean?*

iv. *In the first data list, which items are above the mean and which are below its mean?*

v. *In the second data list, which items are above the mean and which are below its mean?*

vi. *Try to make r smaller by just changing the the signs of some of the z-values.*

vii. *In general, why is r a reasonable measure of how correlated two variables are?*