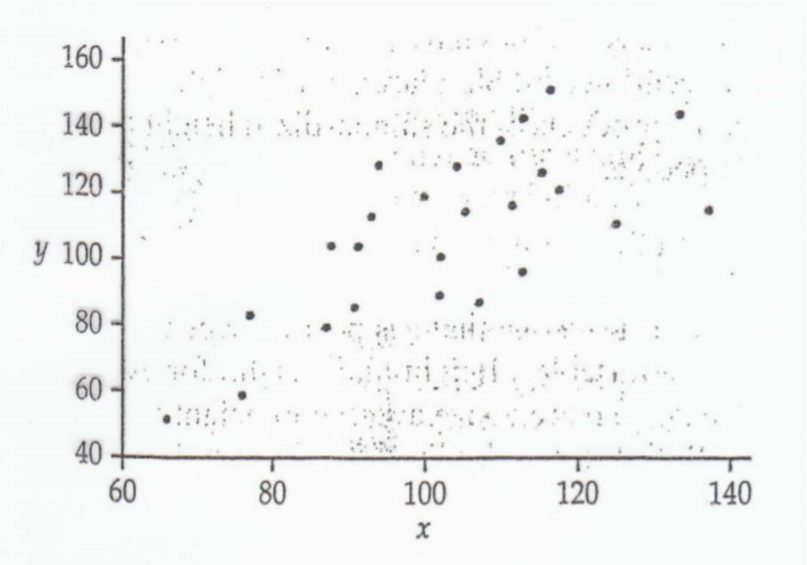
STAB22 TUT21

Chong Chen
University of Toronto, Scarborough
Department of Computer and Mathematical Sciences

February 15, 2019

1 Scatterplots

Previously, we looked at one categorical variable, two categorical variables, and one quantitative variable. Today we shall look at two quantitative variables. First tool to do so: a scatterplot!



Describe the overall pattern of a scatterplot by the form, direction, and strength of the relationship.

1.1

Definition: A dependent variable measures an outcome of a study. An independent variable explains or causes changes in the dependent variable.

1.2 Form

Form: straight (linear relationship), curved (curvilinear relationship), exotic, no pattern

1.3 Direction

Direction: positive, negative, neither

1.4 Strength

Strength: strong, moderate, weak, no relationship

1.5 Unusual features

Unusual features: outliers, subgroups

2 Linear regression

A regression line is a straight line that describes how a dependent variable y changes as an independent variable x changes.

2.1 Correlation

We say a linear relationship is:

STRONG if the points lie close to a straight line, and
WEAK if they are widely scattered about a line

2.1.1 Definition

Definition: The correlation measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r.

2.1.2 Properties

1. The correlation r is always a number between -1 and 1. Values of r near 0 indicate a very weak linear relationship. The strength of the relationship increases as r moves away from 0 toward either -1 or 1. The extreme values $r = -1$ and $r = 1$ occur only when the points in a scatterplot lie exactly along a straight line.
2. Positive r indicates positive association between the variables, and negative r indicates negative association.

2.2 Equations

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

where slope

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

and intercept

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

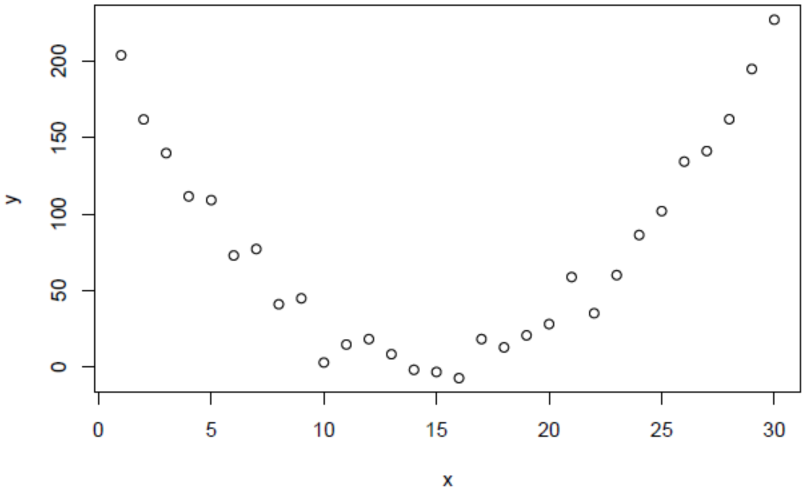
3 Coefficient of determination

The square of the correlation (r^2) is the fraction of the variation in the values of y that is explained by the linear regression of y on x.

4 Examples

4.1 Scatterplot

The figure below shows the scatterplot of two variables, x and y.



Which of the following statements is/are correct?
(I) Variables x and y have a strong relationship
(II) Variables x and y have a negative relationship
(III) Variables x and y are strongly correlated

(a) Only statement (I) is correct
(b) Only statement (II) is correct
(c) Only statements (I) and (II) are correct

(d) Only statements (I) and (III) are correct
(e) All three statements are correct

4.2 Linear regression

The breathing rate of goldfish (opercular breathing rate, in counts per minute) were measured in a biology laboratory. Counts were made at various temperatures ranging from 9 degrees Celsius to 27 degrees Celsius in order to study the relationship between temperature and breathing rate of goldfish. The researchers found that the regression of breathing rate on temperature yields a regression line with slope 4.54 and intercept -1.57
Use this information for this question and the next three questions.

Which of the following statements is/are known for sure?
(I) The correlation between breathing rate of goldfish and temperature is positive
(II) The correlation between breathing rate of goldfish and temperature is negative
(III) The coefficient of determination R^2 is negative

(a) Only statement (I) is known for sure
(b) Only statement (II) is known for sure
(c) Only statements (II) and (III) are known for sure
(d) None of the three statements are known for sure

The predicted breathing rate of goldfish at 20 degrees Celsius is closest to:
(a) 5 counts per minute
(b) 16 counts per minute
(c) 27 counts per minute
(d) 89 counts per minute

4.3 Correlation

In a regression calculation, a researcher finds that the explanatory variable x has mean 100 and standard deviation 10, and the response variable y has mean 250 and standard deviation 40. The regression equation is found to be:

$$\hat{y} = 450 - 2x$$

What is the correlation between x and y?
(a) -0.8
(b) -0.5
(c) 0.1
(d) 0.4
(e) cannot tell from the information available