Descriptive Analysis

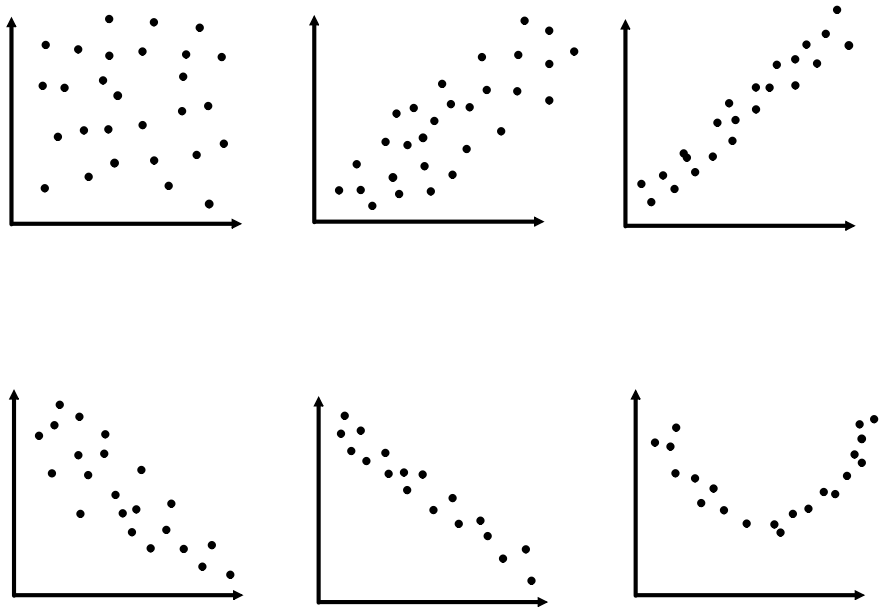
The values of two different variables that are obtained from the same population element are called **bivariate data**.

The data we study will be the result of two quantitative variables, we express the data as ordered pairs (x,y).

To form a **scatter diagram** we plot the x variable on the horizontal axis and the y variable on the vertical axis.

The purpose of **linear correlation analysis** is to measure the strength of a linear relationship between two variables.

Illustration of scatter diagrams and correlation



Sums of Squares

The following three sums will be required for most of our calculations

(Recall total variation)

$$SS(x) = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = (n-1)S_x^2$$

Homework: show this is true.

Where S_x^2 is the variance with respect to the x variable.

Similarly

$$SS(y) = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = (n-1)S_y^2$$

With data presented as (x,y)

- x is considered the independent variable
- y is considered the dependent variable

More emphasis is placed on the dependent variable y.

We refer to the total variation in y as the total variation (SST)

$$SS(y) = \text{SST}$$

$SS(x)$ and $SS(y)$ add together the squares of the deviations from the mean, $x_i - \bar{x}$ and $y_i - \bar{y}$ respectively

Variation due to pairing x_i and y_i

$$SS(xy) = \sum (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$= \sum x_i \cdot y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

.

Regression Analysis

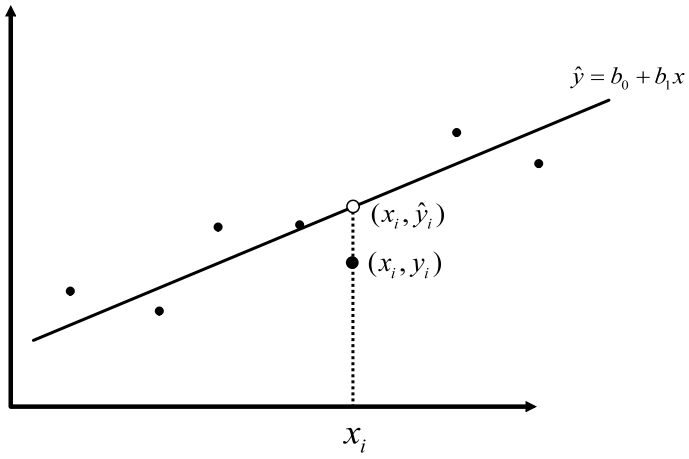
If a straight line model seems appropriate, **Regression Analysis** finds the equation of the line that best describes the relationship between two variables.

To do this we use the **method of least squares**.

We want to find the equation of a line so it will have the equation

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

$\hat{y}$ represents the **predicted value of y** that corresponds to a value of x. This line is often called the line of best fit.



Let (x_i, y_i) be a data point in a data set that we are looking at and let $\hat{y}_i$ be the predicted value of y for x_i. We see that $y_i - \hat{y}_i$ is the difference between the observed value and the predicted value for y. This difference is called the **residual**.

Notice that if $y_i - \hat{y}_i$ is positive then (x_i, y_i) is above the line.

if $y_i - \hat{y}_i$ is negative then (x_i, y_i) is below the line.

Each residual tells us how far off our prediction is from the observed value.

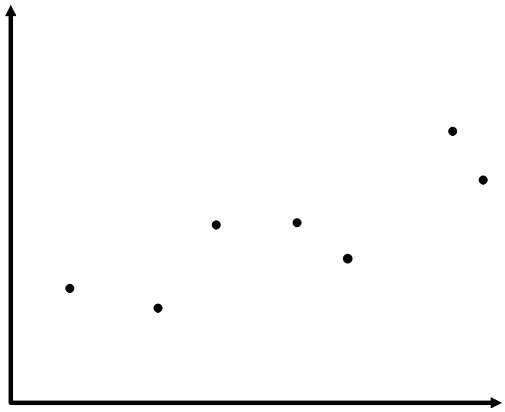
If we want to get a sense of how far off our entire line is from predicting all of the observed data we want to sum up the residuals somehow. But since some residuals are positive and some are negative taking the sum of the residuals will not be a good measure for how far off our line is.

If we instead square each residual $(y_i - \hat{y}_i)^2$ they will all be positive. And so we will look at the sum of the squares of the residuals.

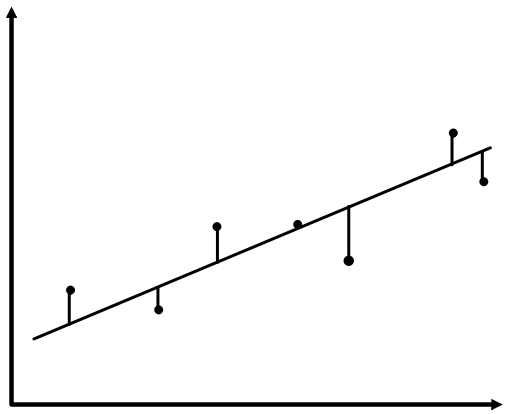
$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

We want to find a line that minimizes SSE.

The **least squares criterion** requires that we find the constants b_0 and b_1 such that $\sum(y-\hat{y})^2$ is as small as possible.



Scatter diagram



Line of best fit with residuals

We need formula to find the slope b_1 and the y-intercept b_0 of the line of best fit $\hat{y} = b_0 + b_1 x$

Coefficients for the Line of Best Fit

Slope $$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

$$= \frac{SS(xy)}{SS(x)}$$

y-intercept $$b_0 = \frac{\sum y - (b_1 \cdot \sum x)}{n}$$

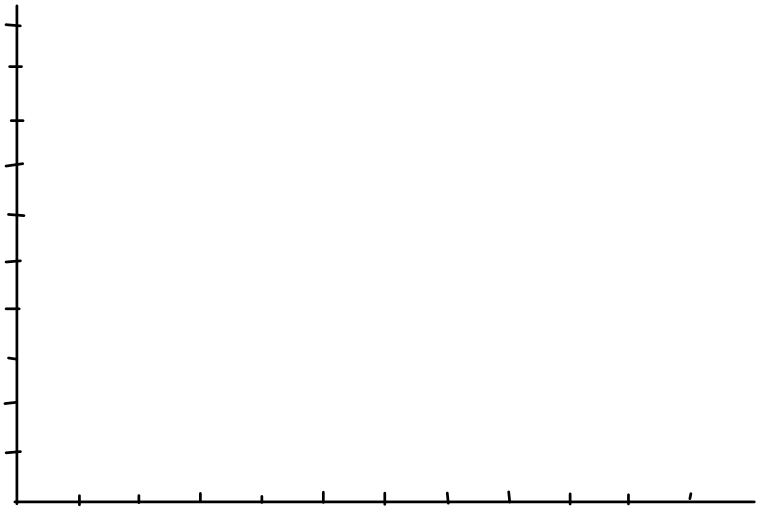
$$= \bar{y} - (b_1 \cdot \bar{x})$$

Proof:

Consider the data of distance from a golf hole (in yards) and the corresponding success rate of a shot.

(x) distance (yards)	(y) success rate (%)
213	44
188	53
163	61
138	68
113	72
88	78
63	85

a) Construct a scatter diagram



b) Calculate the line of best fit

c) Predict the success rate of a shot from 100 yards away.

•