CH 5: Regression

- least-squares regression line
- residuals
- influential observations
- correlation & regression cautions
- association does not imply causation

Introduction

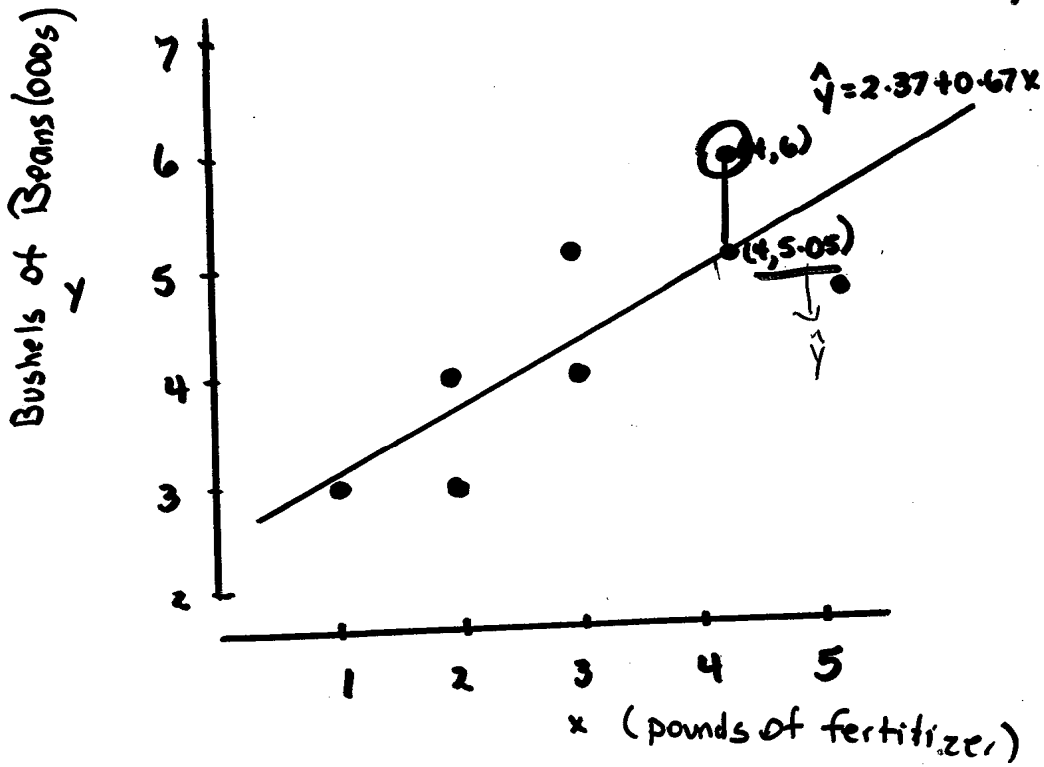
- may **look at regression when two quantitative variables show a linear relationship**

- regression line describes a relationship between an explanatory variable and a response variable(***here you need to know which is which)***

- may use regression line to predict the value of y for a given value of x

- start with revisiting the first Ch. 4 Scatterplot example (fertilizer and crop yield)

Using a Line to Predict
(Fertilizer and Crop Yield Example



Error = observed − predicted

$= y - \hat{y}$ $= 6 - 5.05 = .95$

(Residual)

Least Squares Regression Line

- the *least squares regression line* of y on x is that line that makes the sum of squares of vertical distances of the data points from the line as small as possible

Regression equation: $\hat{y} = a + bx$

$$a = \bar{y} - b\bar{x} \qquad b = r\frac{s_y}{s_x}$$

where

(i) x is the explanatory variable

(ii) y is the response variable

(iii) a is the intercept(value of y when x is 0) and has the same units as y

(iv) b is the slope(slope is the rate of change)

Correlation(r) Example revisited - Is Growth linear?

Child	Age (mths)	Height (cms)	$(x_i - \bar{x})/s_x$	$(y_i - \bar{y})/s_y$	$(x_i - \bar{x})(y_i - \bar{y})/s_x s_y$
1	36	86	**-1.77**	**-1.68**	**2.97**
2	48	90	**-0.35**	**-0.46**	**0.16**
3	51	91	**0.00**	**-0.15**	**0.00**
4	54	93	**0.35**	**0.46**	**0.16**
5	57	94	**0.71**	**0.76**	**0.54**
6	60	95	**1.06**	**1.07**	**1.13**
	$\bar{x}$ **= 51**	$\bar{y}$ **=91.5**			
	s_x**= 8.485**	s_y**= 3.271**			
Total					**4.97**
r					**.9944=r**

1. will show **Least squares equation line is** $\hat{y}$ **= 71.95+.3833x**
2. Predict height for (i)a 32 month child, and (ii) a 2 months child.
3. What about a height prediction for (iii), a 30 year old?

Regression analysis using technology

- **various software packages/statistical calculators can be used to determine regression line**
- text shows output from Minitab and Excel (p.131)
- *should be able to pick out value of slope(a) and intercept(b) from various output sources shown*
- will **look at 1 minitab output example in class**

Example: Radioactive Waste and the Columbia River

- the Hanford Atomic Energy Plant in Washington State in the United States had been producing plutonium since World War II

- some of the waste from the plant was stored in open pits and the radioactive waste had leaked into the Columbia River

- eight Oregon counties and the city of Portland have been exposed to radioactive contamination

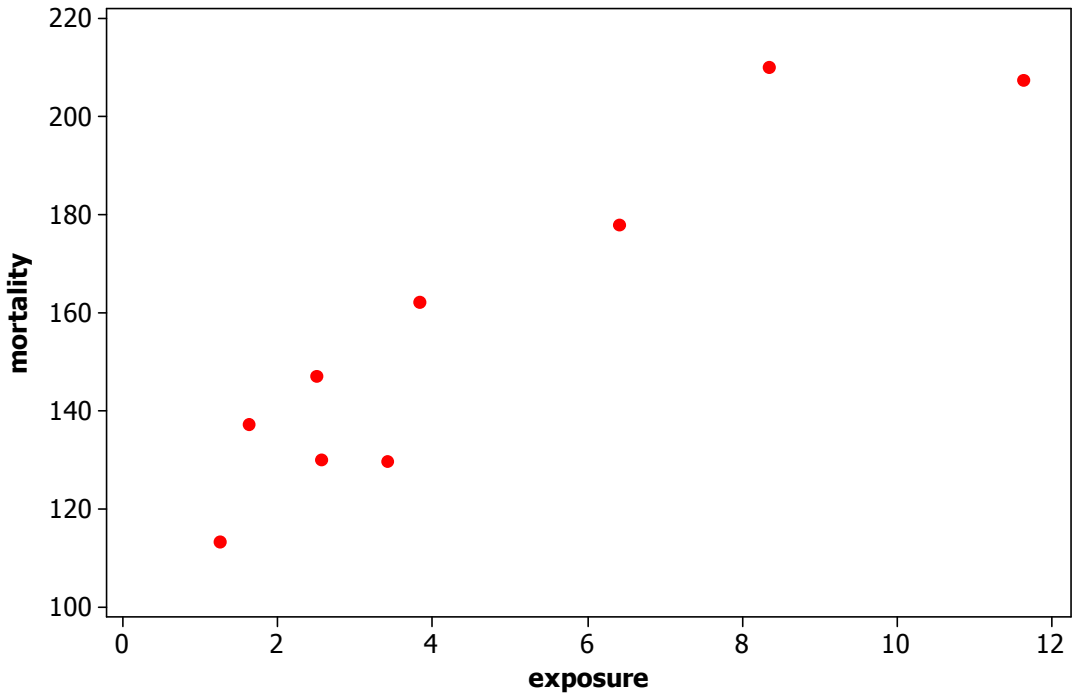
Data

County or city	Exposure Index	Cancer Mortality
Umatilla	2.49	147.1
Morrow	2.57	130.1
Gilliam	3.41	129.9
Sherman	1.25	113.5
Wasco	1.62	137.5
Hood River	3.83	162.3
Portland	11.64	207.5
Columbia	6.41	177.9
Clatsop	8.34	210.3

- original data collected in 1965 based on experience of the previous five years
- problem still exists today

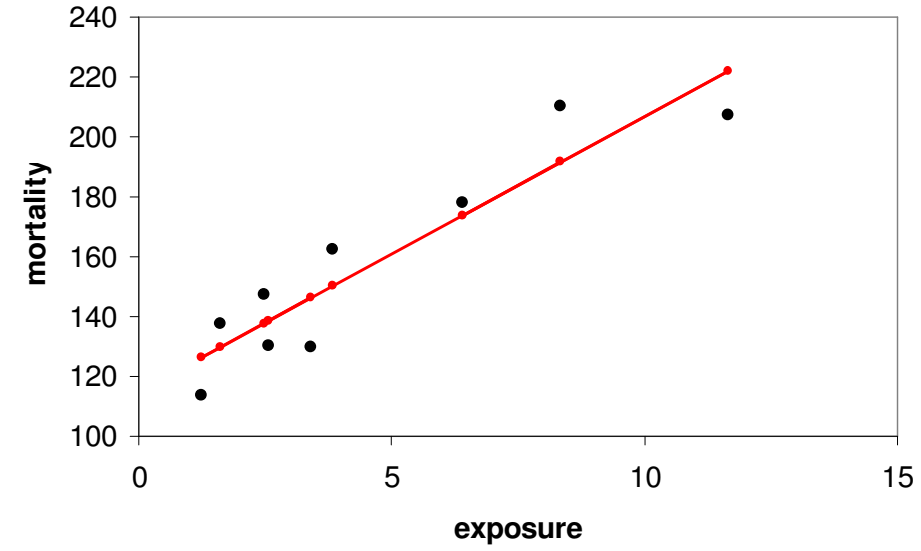
Scatterplot

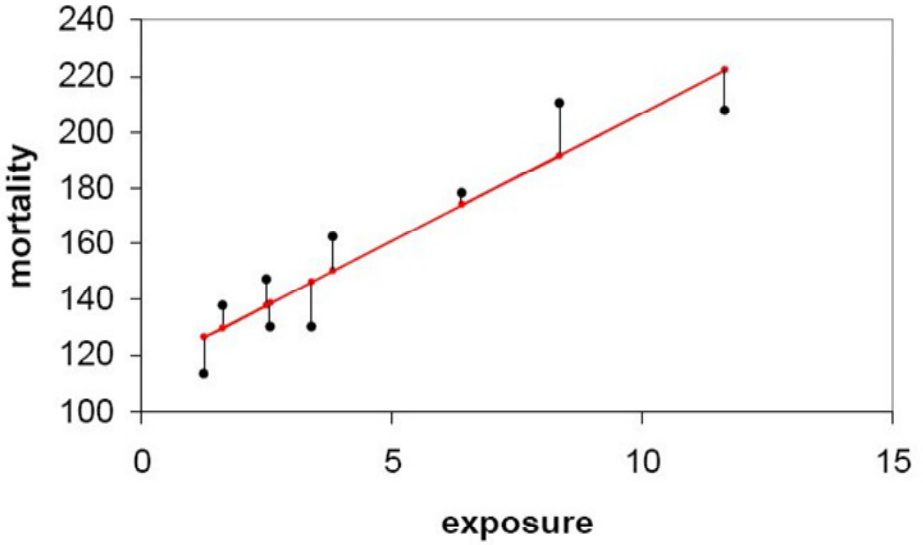
- the exposure index includes factors such as distance of the county from Hanford and the average distance from water frontage
- the cancer mortality rate is deaths per 100,000 residents



Least-squares Regression Line

"the least-squares regression line of *y* on *x* is that line that makes the sum of squares of vertical distances of the data points from the line as small as possible"





EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables, and their correlation r. The least-squares regression line is the line

$$\hat{y} = a + bx$$

with **slope**

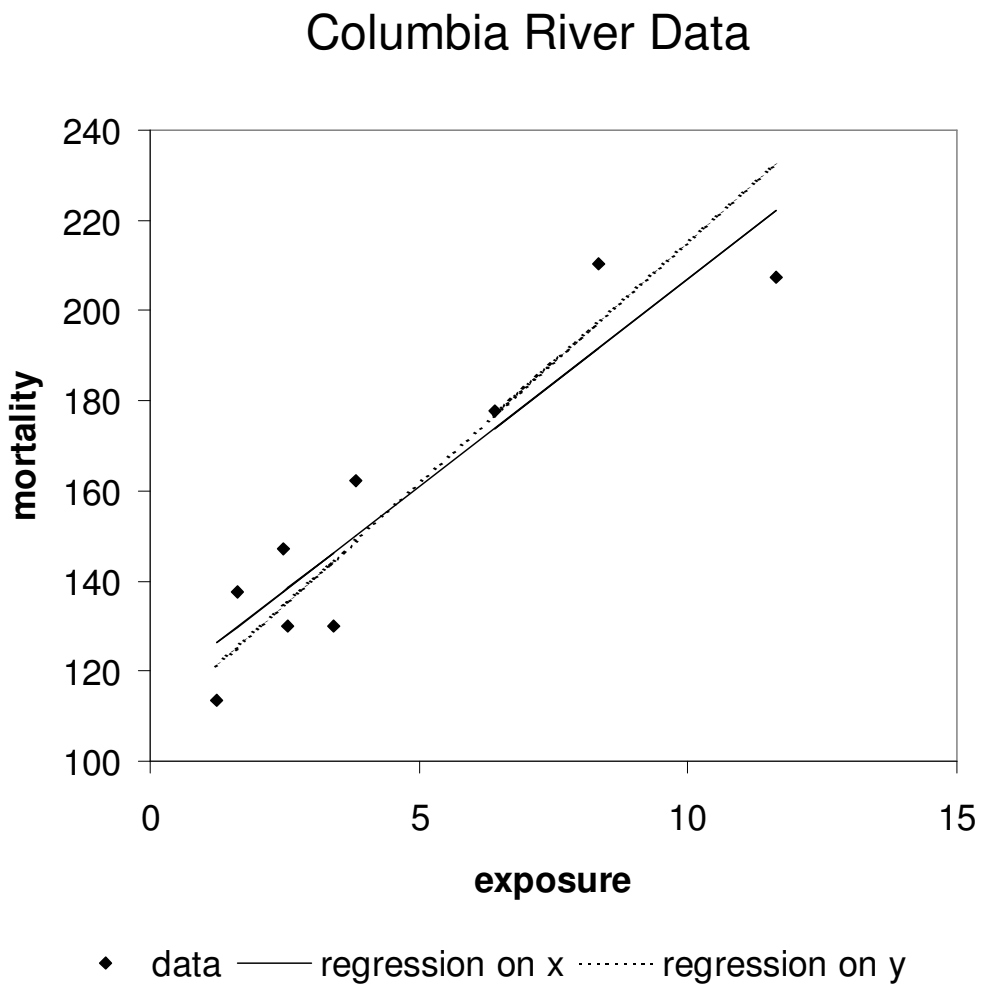
$$b = r\frac{s_y}{s_x}$$

and **intercept**

$$a = \bar{y} - b\bar{x}$$

Regression Fact 1

- the distinction between explanatory and response variables is essential
- regression of *y* on *x* is different from regression of *x* on *y*



Regression Fact 2

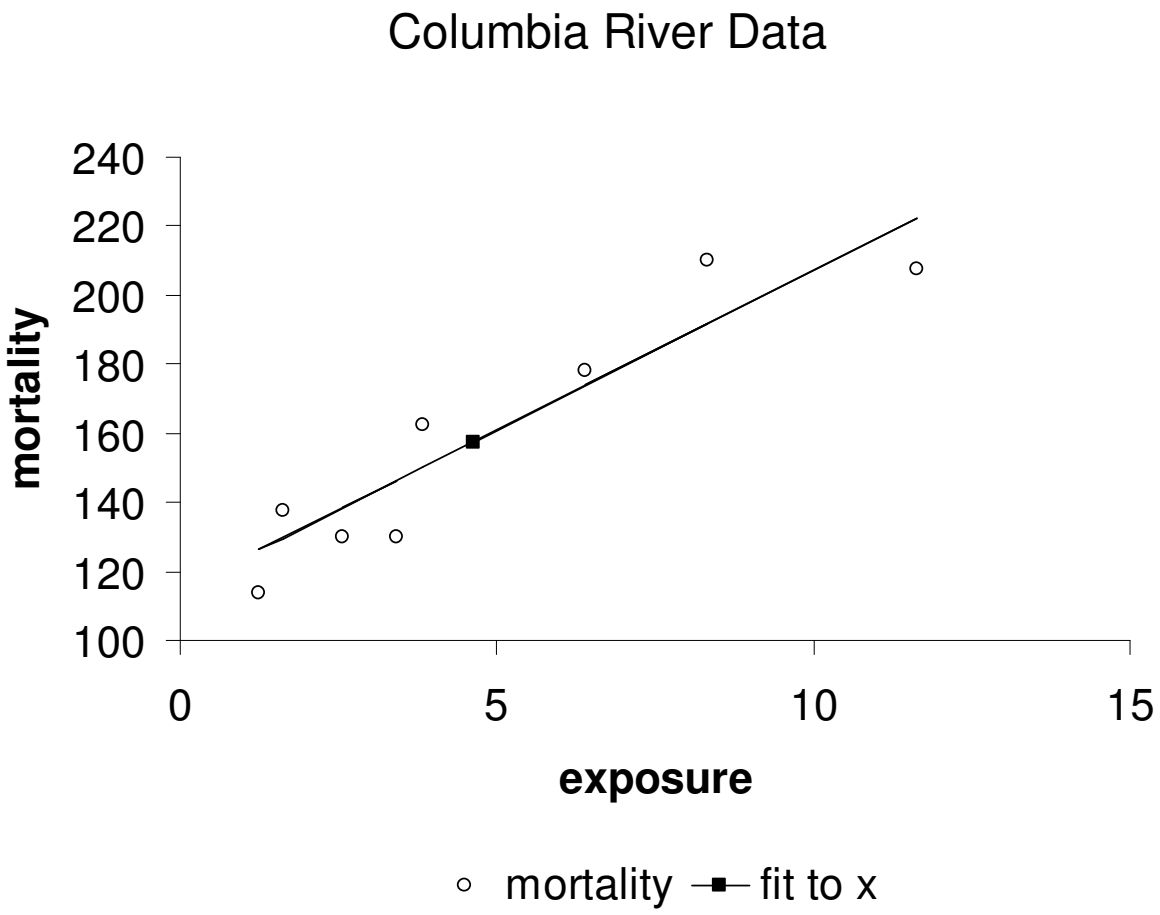
$$\hat{y} = a + bx \qquad b = r\frac{s_y}{s_x}$$

- a change of one standard deviation in *x* corresponds to a change of *r* standard deviations in *y*

Regression Fact 3

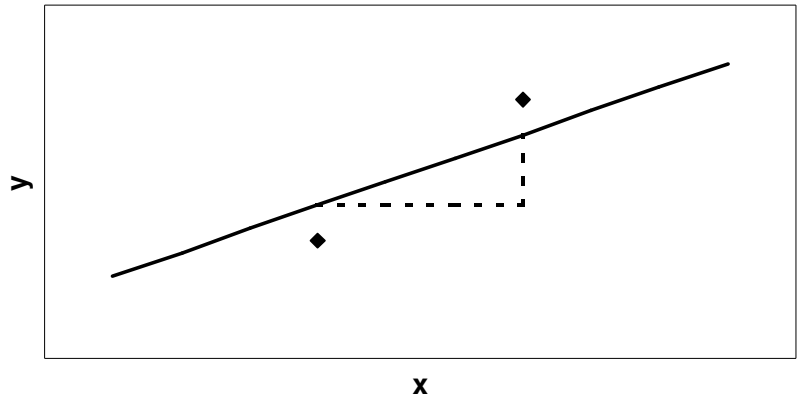
- the regression line passes through the point: (4.62,157.3)

$$(\overline{x}, \overline{y})$$

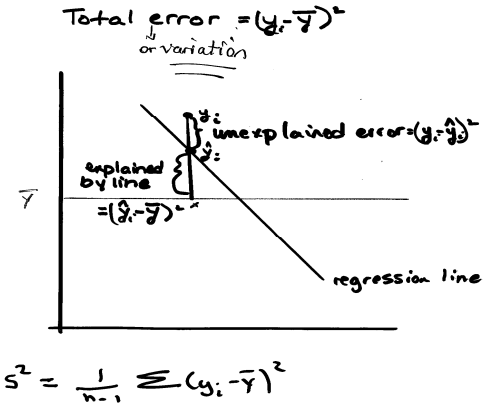


Regression Fact 4

- the square of the correlation, i.e. r^2, is the fraction of the variation in the values of y that is explained by the least-squares regression of *y* on *x*
- two types of variation: (1) due to the line and (2) about the line.



More on Regression Fact 4



- Least squares regression makes the sum of squared residuals as small as possible
- **r^2 tells how well the regression fits the data**
- r^2 is the fraction of the variation in y that is explained by the least squares regression
  - ❖ **r^2= explained variation/total variation**

Example: Constructing a Least Squares Regression Line

Given:

$r = 0.8$

$\bar{x} = 10$

$\bar{y} = 60$

$s_x = 0.4$

$s_y = 2.0$

Construct the least squares regression line.

Residuals and Residual Plots

- **Residual = difference between observed & predicted value**

  Residual = $(y - \hat{y})$

  positive value ➔observed value falls above regression line

  negative value ➔observed value falls below regression line

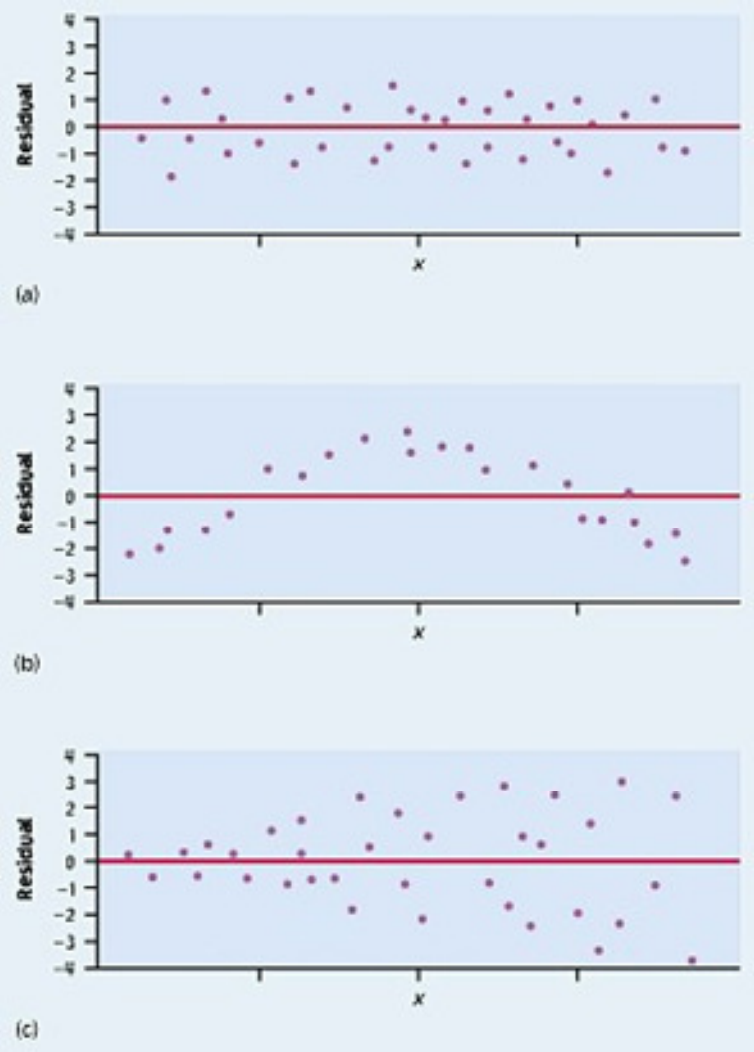
- **Sum of all residuals is 0**

- **Examining plots of residuals against explanatory variable help assess fit of the regression line**

  - Ideally want no pattern in the residuals(if there's an unstructured horizontal band of points centered a 0, a linear model is appropriate)

  - a definite pattern of residuals(e.g. curved) could indicate a linear model is not appropriate

Residual Plots-Examples



Residuals are randomly scattered—good!

A curved pattern—means the relationship you are looking at is not linear.

A change in variability across plot is a warning sign. You need to find out why it is and remember that predictions made in areas of larger variability will not be as good.

Outliers&Influential Observations(see text exs.)

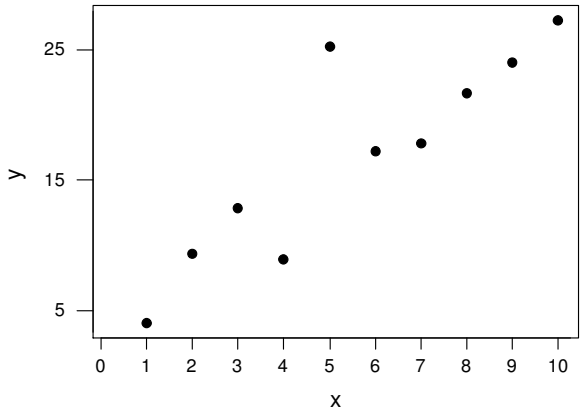
- An **outlier i**s an observation that lies outside the overall pattern of the observation(but with regression not all outliers are influential)

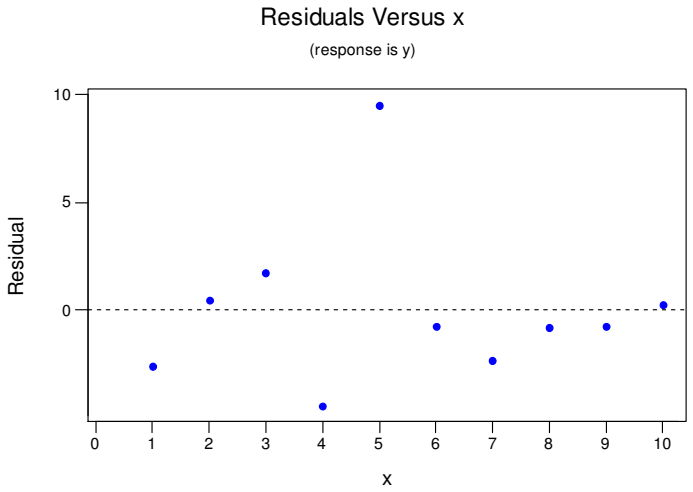
- an observation is **influential** for a statistical calculation if removing it markedly changes the result of the calculation

- **outliers in the x direction are often influential observations** on the regression line (but may not always have a large residual)

- **regression is therefore not a resistant measure!**

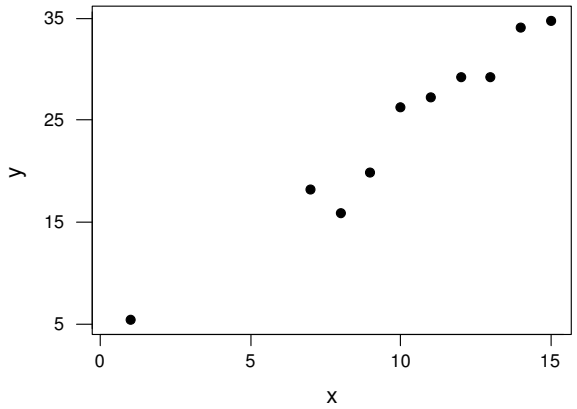
Points with Large Residuals

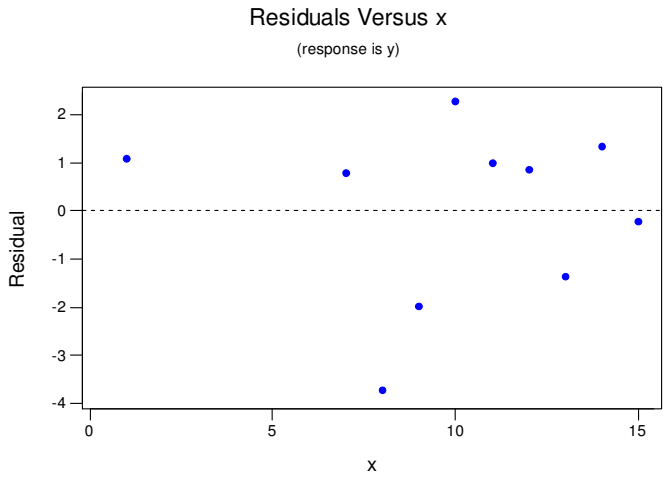




- points are outliers in the vertical direction because they lie far from the line that describes the overall pattern

Extreme Points in the *X* Direction





- the residual associated with the smallest *x* here is not a large residual but it could have a considerable influence on the fit of the line

CAUTIONS

- **Cautions discussed for both correlation& regression and (also review text examples for (i) to (iii) below)**

(i) Extrapolation;

- Extrapolation is use of regression line to predict far outside range of values used to obtain the line

- **infants age & height example** using regression line based on data we had to predict age of a 2 month baby or a 30 year old would both be examples of extrapolation

- few relationships are linear for *all* values of x

CAUTIONS (continued)

(ii) Lurking Variables

- **A lurking variable is** a variable that isn't among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables

- A lurking variable can suggest a strong relationship between between two other variable(x and y) or hide a strong relationship between two other variables

Example: Do Left Handers Die early?

A relatively recent analysis of California 1,000 deaths showed that average age at death for lefthanders is age 66, and for right handed people the average age was 75.Does this mean lefthanders die early? **What is the lurking variable?**

CAUTIONS (continued)

(iii) Association does not imply causation

- association is sometimes explained by a lurking variable
- event when direct causation is present, there could still be a lurking variable
- best evidence comes from a controlled experiment
- **Review Text Examples**
  - Does having More cars make you live longer?
  - Overweight mothers, overweight daughter
  - Does smoking cause lung cancer?