Lecture 8

Relationships Between Measurement Variables

Thought Question 1:

Judging from the scatterplot, there is a ***positive correlation*** between verbal SAT score and GPA.

For used cars, there is a ***negative correlation*** between the age of the car and the selling price.

Explain what it means for two variables to have a positive correlation or a negative correlation.

Thought Question 2:

Suppose you were to make a scatterplot of (adult) **sons' heights versus fathers' heights** by collecting data on both from several of your male friends.

You would now like to predict how tall your nephew will be when he grows up, based on his father's height.

Could you use your scatterplot to help you make this prediction? Explain.

Thought Question 3:

Do you think each of the following pairs of variables would have a **positive correlation**, a **negative correlation**, or **no correlation**?

a. Calories eaten per day and weight
b. Calories eaten per day and IQ
c. Amount of alcohol consumed and accuracy on a manual dexterity test
d. Number of ministers and number of liquor stores in cities in Pennsylvania
e. Height of husband and height of wife

Thought Question 4:

An article in the *Sacramento Bee* (29 May, 1998, p. A17) noted "Americans are just too fat, researchers say, with 54 percent of all adults heavier than is healthy. **If the trend continues**, experts say that within a few generations **virtually every U.S. adult will be overweight**."

This prediction is based on "**extrapolating**," which assumes the current rate of increase will continue indefinitely. Is that a reasonable assumption? Do you agree with the prediction? Explain.

Scatterplots

- Two-dimensional plot, with one variable's values plotted along the vertical axis and the other along the horizontal axis.
  - X-axis:
  - Y-axis:
- Display the general relationship between 2 quantitative variables graphically.
- 2 variables measured on the same individuals.

Examples

- The weight and height of a person.
- An insurance company reports that heavier cars have less fatal accidents per 10,000 vehicles than lighter cars do.
- A medical study finds that short women are more likely to have heart attacks than women of average height, while tall women have even fewer heart attacks.

Example

- A statistician wanted to purchase a house in a neighborhood. He decided to develop a model to predict the selling price of a house.

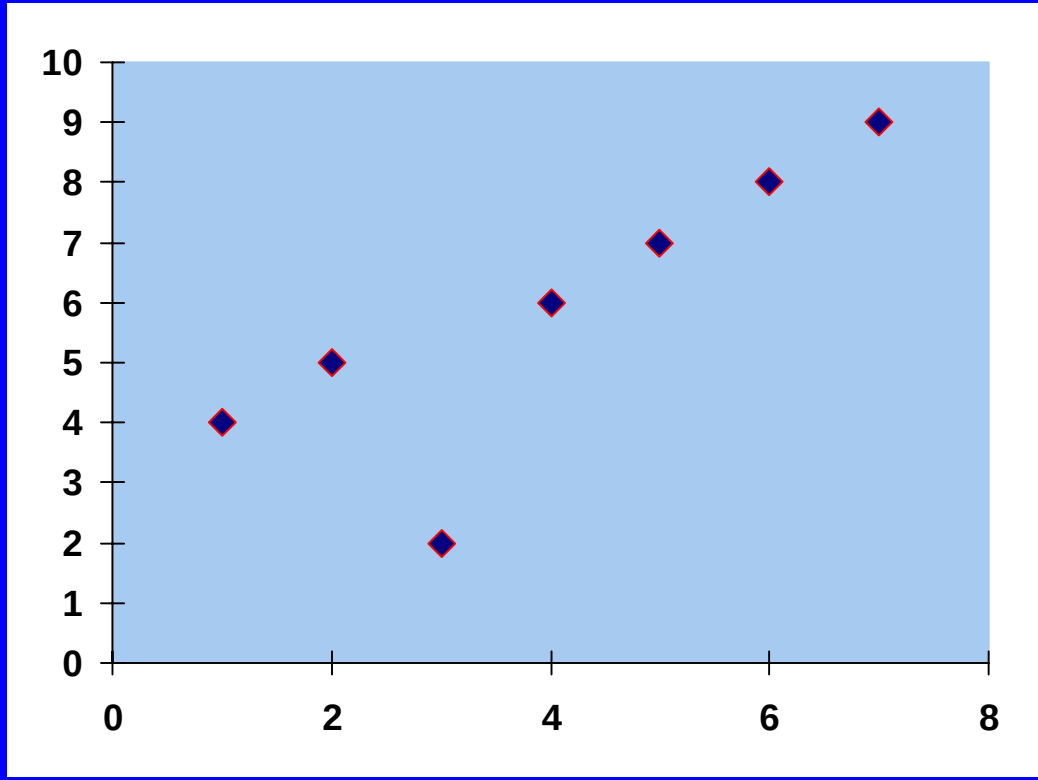
- He took a random sample of 100 houses that recently sold and recorded the selling price, the number of bedrooms, and the size (in square foot).

Example:

Draw a Scatter Plot to represent the following dataset:

x: 1, 3, 2, 4, 7, 6, 5

y: 4, 2, 5, 6, 9, 8, 7


10
9
8
7
6
5
4
3
2
1
0
0
2
4
6
8

Question

Any comments on these two datasets?
Is there anything special about them?

- Looking at a scatter plot can sometimes allow us to determine if a relationship exists between two variables.
- But in general we need to go beyond pictures and develop a numerical measure of how strongly the two variables x and y are **related.**

10.1 Statistical Relationships

Correlation: measures the ***strength*** of a certain type of relationship between two measurement variables.

Regression: gives a numerical method for trying to ***predict*** one measurement variable from another.

Statistical Relationships versus Deterministic Relationships

Deterministic: if we know the value of one variable, we can determine the value of the other *exactly*. e.g. relationship between volume and weight of water.

Statistical: *natural variability* exists in both measurements. Useful for describing what happens to a population or aggregate.

10.2 Strength versus Statistical Significance

A relationship is **statistically significant** if the chances of observing the relationship in the sample when actually nothing is going on in the population are **less than 5%**.

A relationship is statistically significant if that relationship is stronger than 95% of the relationships we would expect to see just by chance.

Two Warnings about Statistical Significance

- **Even a minor relationship will achieve "statistical significance" if the sample is very large.**
- **A very strong relationship won't necessarily achieve "statistical significance" if the sample is very small.**

Example 1: Small but Significant Increase in Risk of Breast Cancer

News Story #12: Working nights may increase breast cancer risk.

"The numbers in our study are small, but they are statistically significant. …The study was based on ***more than 78,000 nurses*** *from 1988 through 1998. It was found that nurses who worked rotating night shifts at least three times a month for one to 29 years were* ***8%*** *more likely to develop breast cancer. For those who worked the shifts more than 30 years, the relative risk went up by* ***36%****."*

The relationship in the sample, while not strong, is "statistically significant".

Example 2: Do Younger Drivers Eat and Drink More while Driving?

News Story #5: Driving while distracted is common, researchers say.

"Stutts' team had to reduce the sample size from 144 people to 70 when they ran into budget and time constraints while minutely cataloging hundreds of hours of video. The reduced sample size does not compromise the researchers' findings, Stutts said, although it does make analyzing population subsets difficult."

"Compared to older drivers, younger drivers appeared more likely to eat or drink while driving … Sample sizes within age groups, however, were small, prohibiting valid statistical testing." (p. 61-62 of report in Original Source 5)

10.3 Measuring Strength Through Correlation

A Linear Relationship

Correlation (or the *Pearson product-moment correlation* or the *correlation coefficient*) represented by the letter ***r*:**

- Indicator of *how closely the values fall to a straight line*.
- Measures *linear relationships* only; that is, it measures how close the individual points in a scatterplot are to a straight line.

- **Properties of r**
- The correct interpretation of r requires an appreciation of some general properties:
- The value of r does not depend on the unit of measurement for either variable, nor does it depend on which variable is labelled x or y.
- The value of r is between -1 and 1.
- A positive value of r indicates a positive linear relationship between the variables. So as x increases so does y.
- A negative value of r corresponds to a negative relationship. As x increases y decreases.

Example

- A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
- A random sample of 100 cars is selected, and the data are summarized as follows.

$$n = 100, \bar{x} = 36{,}009.5, \bar{y} = 5411.4$$

$$s_x = 6597.6, s_y = 254.9, \frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = -1{,}356{,}256.$$

- Find the correlation
- r=-0.806

Other Features of Correlations

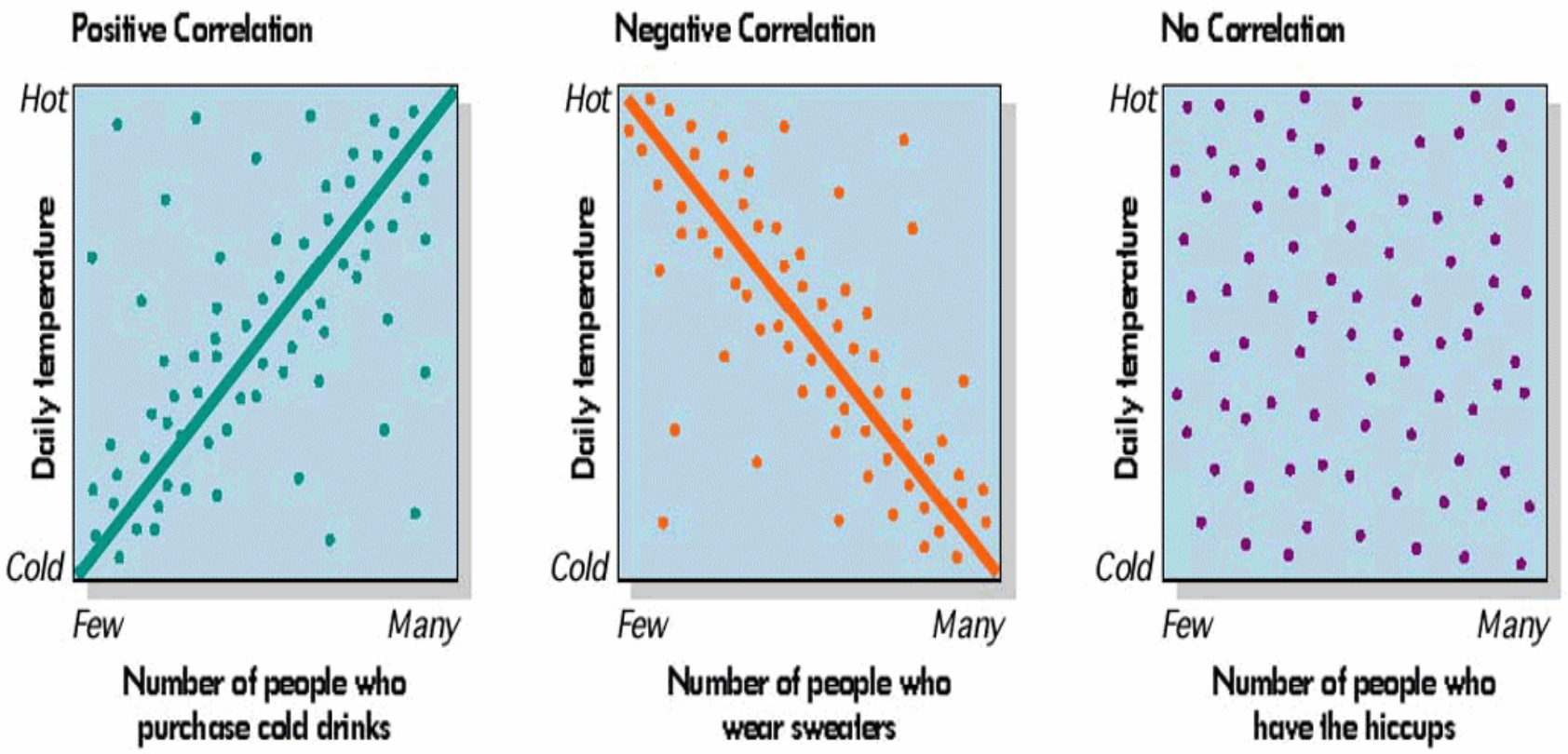
1. Correlation of +1 indicates a perfect linear relationship between the two variables; as one increases, so does the other. All individuals fall on the same straight line (a deterministic linear relationship).
2. Correlation of –1 also indicates a perfect linear relationship between the two variables; however, as one increases, the other *decreases*.
3. Correlation of zero could indicate no linear relationship between the two variables, or that the best straight line through the data on a scatterplot is exactly horizontal.

Other Features of Correlations

4. A *positive correlation* indicates that the variables increase together.
5. A *negative correlation* indicates that as one variable increases, the other decreases.
6. Correlations are unaffected if the units of measurement are changed. For example, the correlation between weight and height remains the same regardless of whether height is expressed in inches, feet or millimeters (as long as it isn't rounded off).

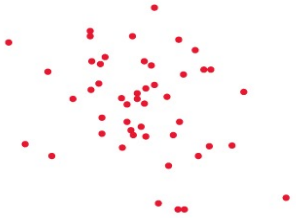
F I G U R E 2 . 1 Correlations: Positive, Negative, and None

Correlations reveal a systematic association between two variables. Positive correlations indicate that variables are in sync: Increases in one variable are associated with increases in the other, decreases with decreases. Negative correlations indicate that variables go in opposite directions: Increases in one variable are associated with decreases in the other. When two variables are not systematically associated, there is no correlation.



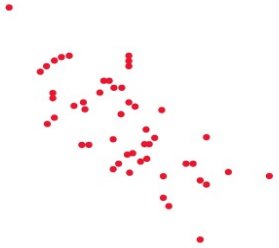
Different correlations

Correlation $r = 0$



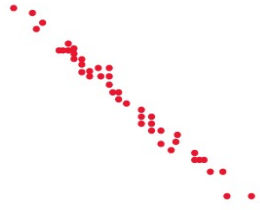
Correlation $r = -0.3$

Correlation $r = 0.5$



Correlation $r = -0.7$

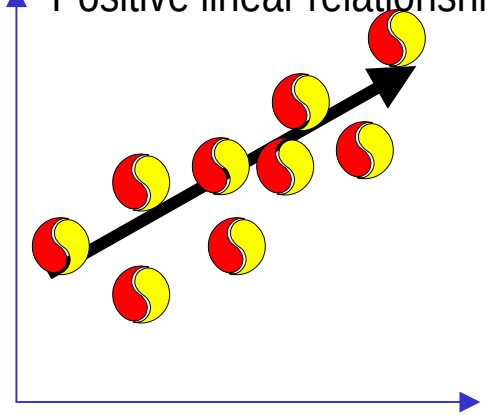
Correlation $r = 0.9$



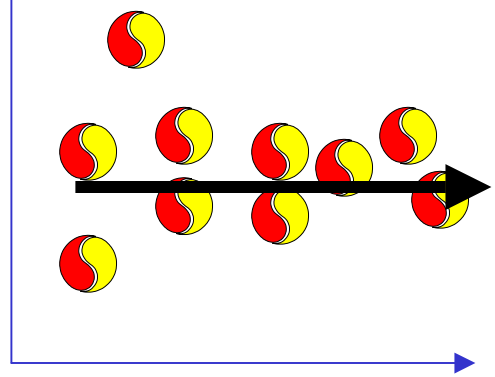
Correlation $r = -0.99$

Typical Patterns of Scatterplots

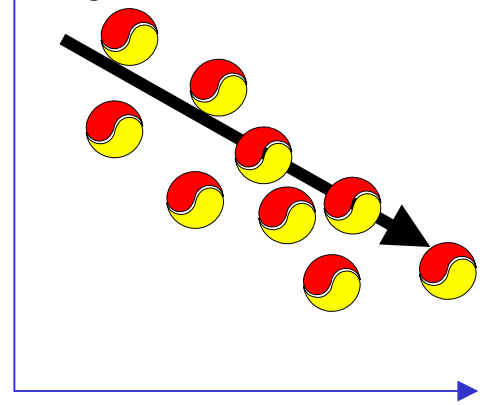
Positive linear relationship



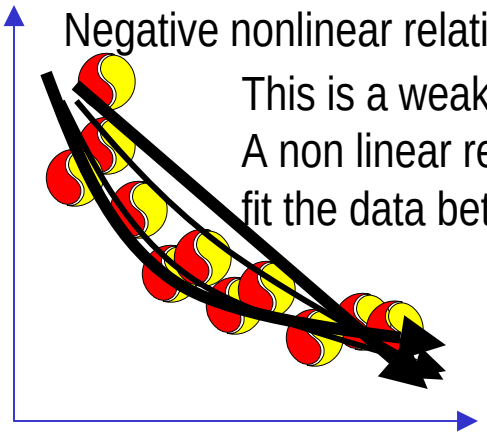
No relationship



Negative linear relationship

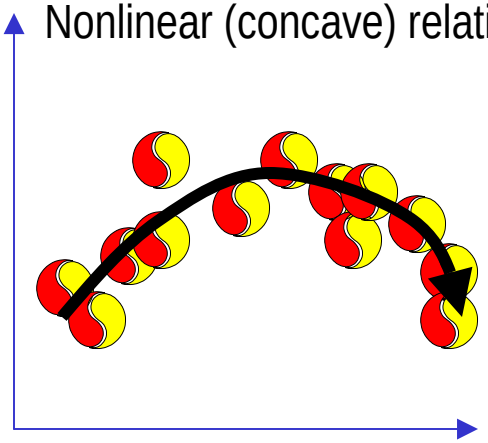


Negative nonlinear relationship

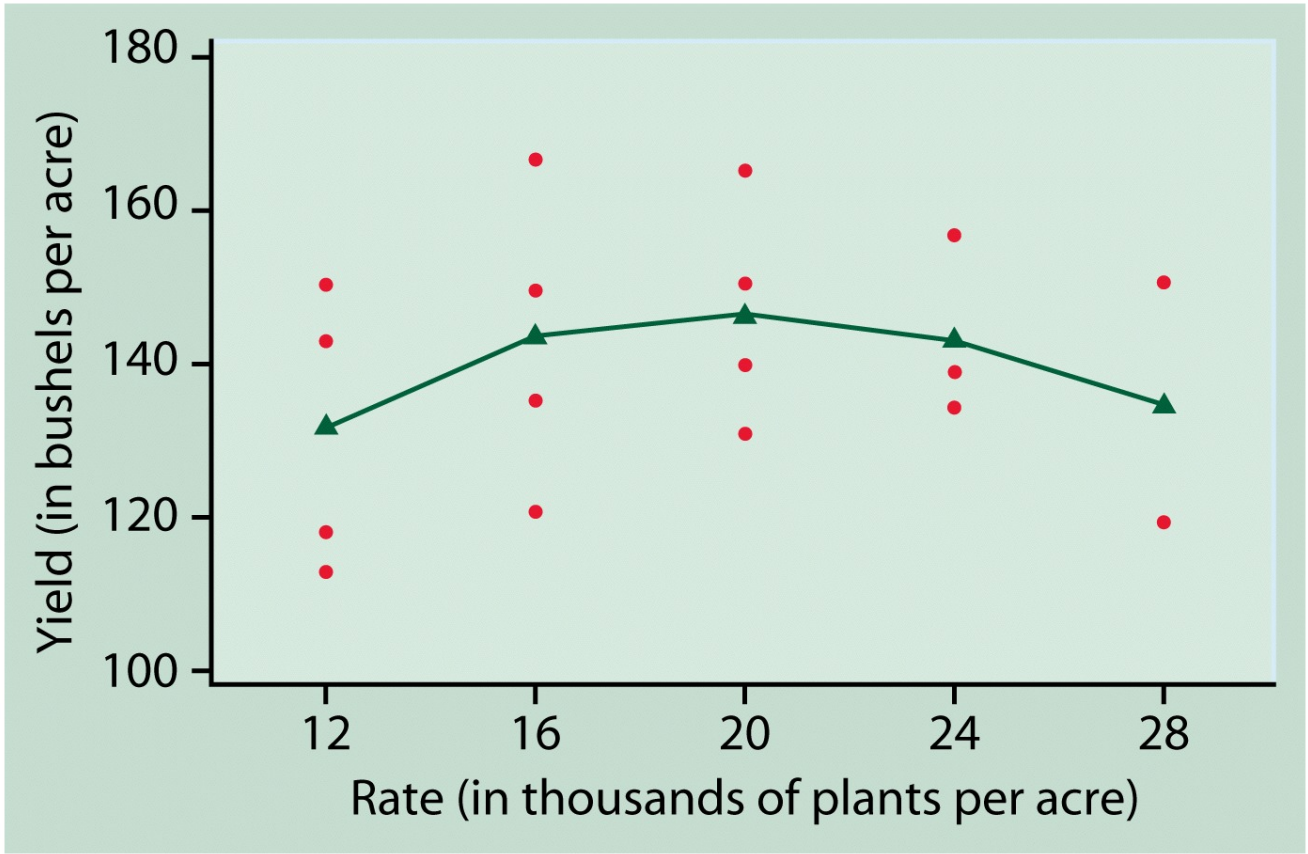


This is a weak linear relationship. A non linear relationship seems to fit the data better.

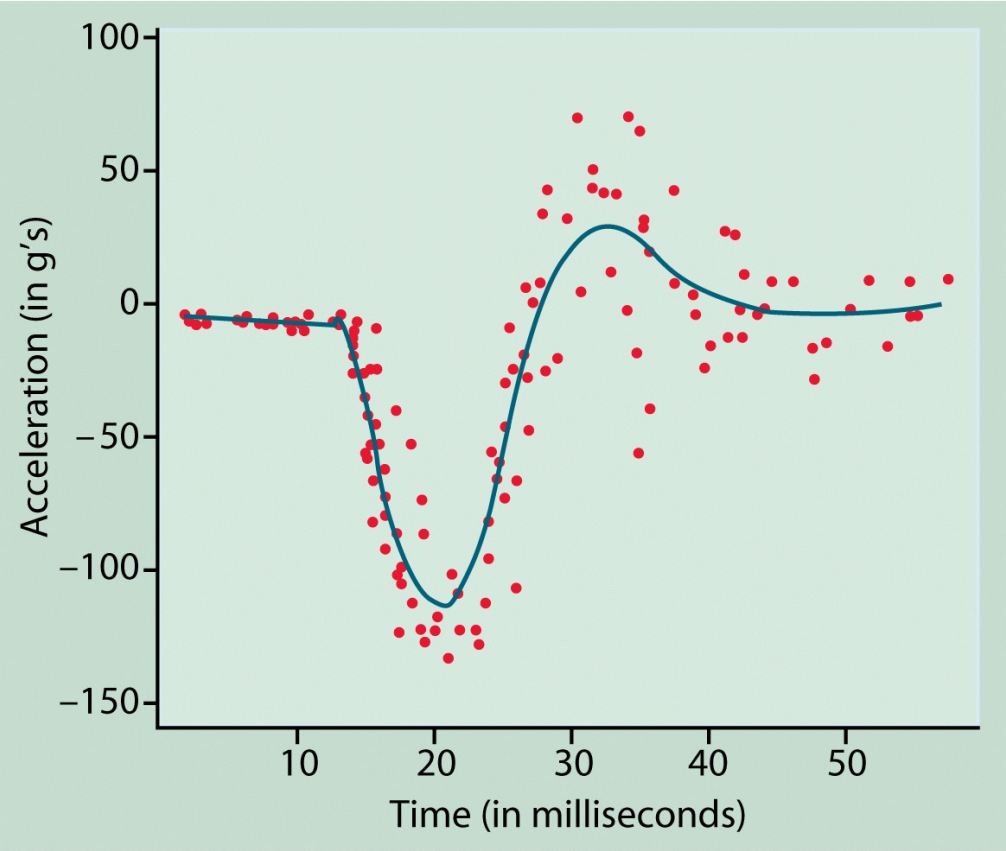
Nonlinear (concave) relationship



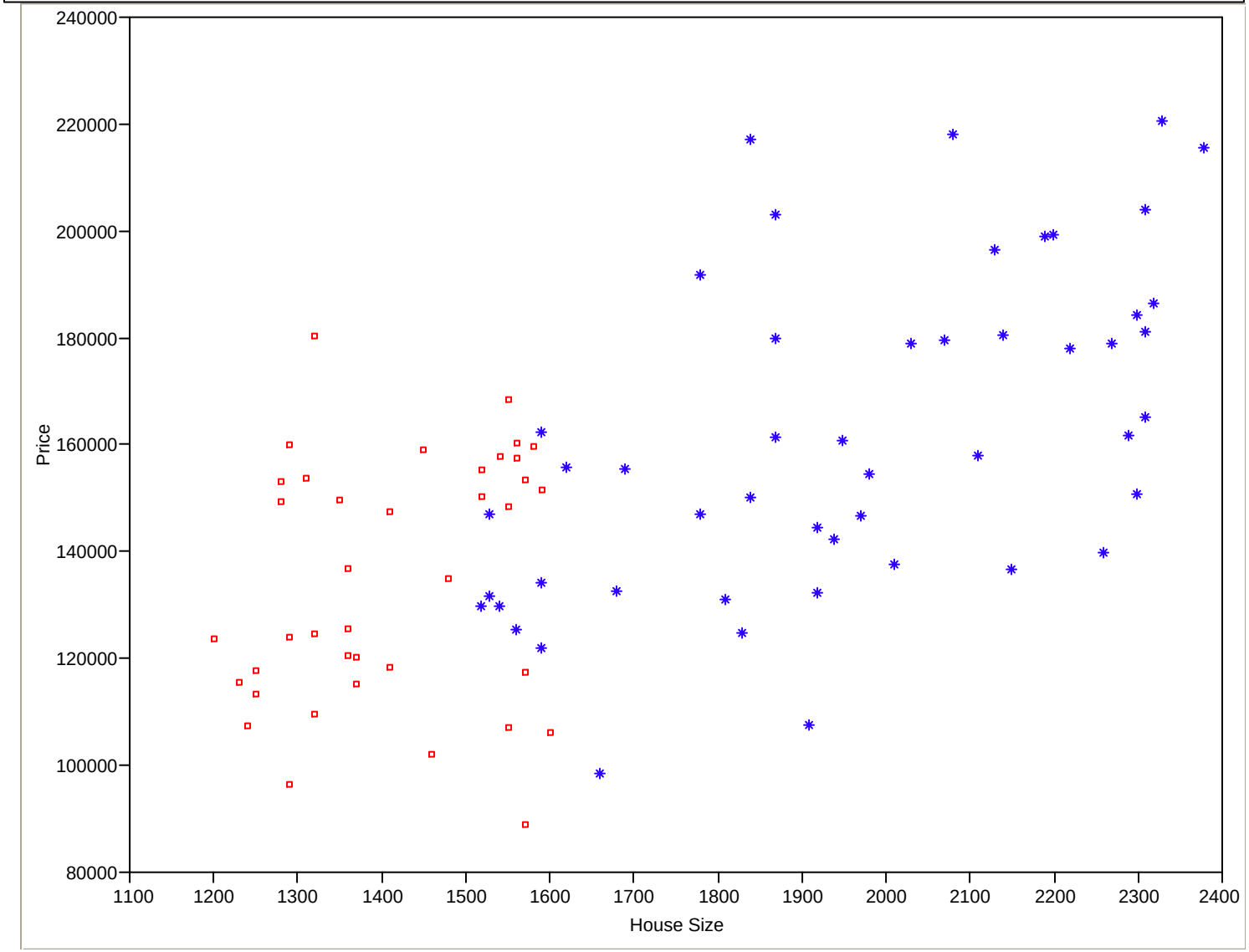
Example: Yield of corn



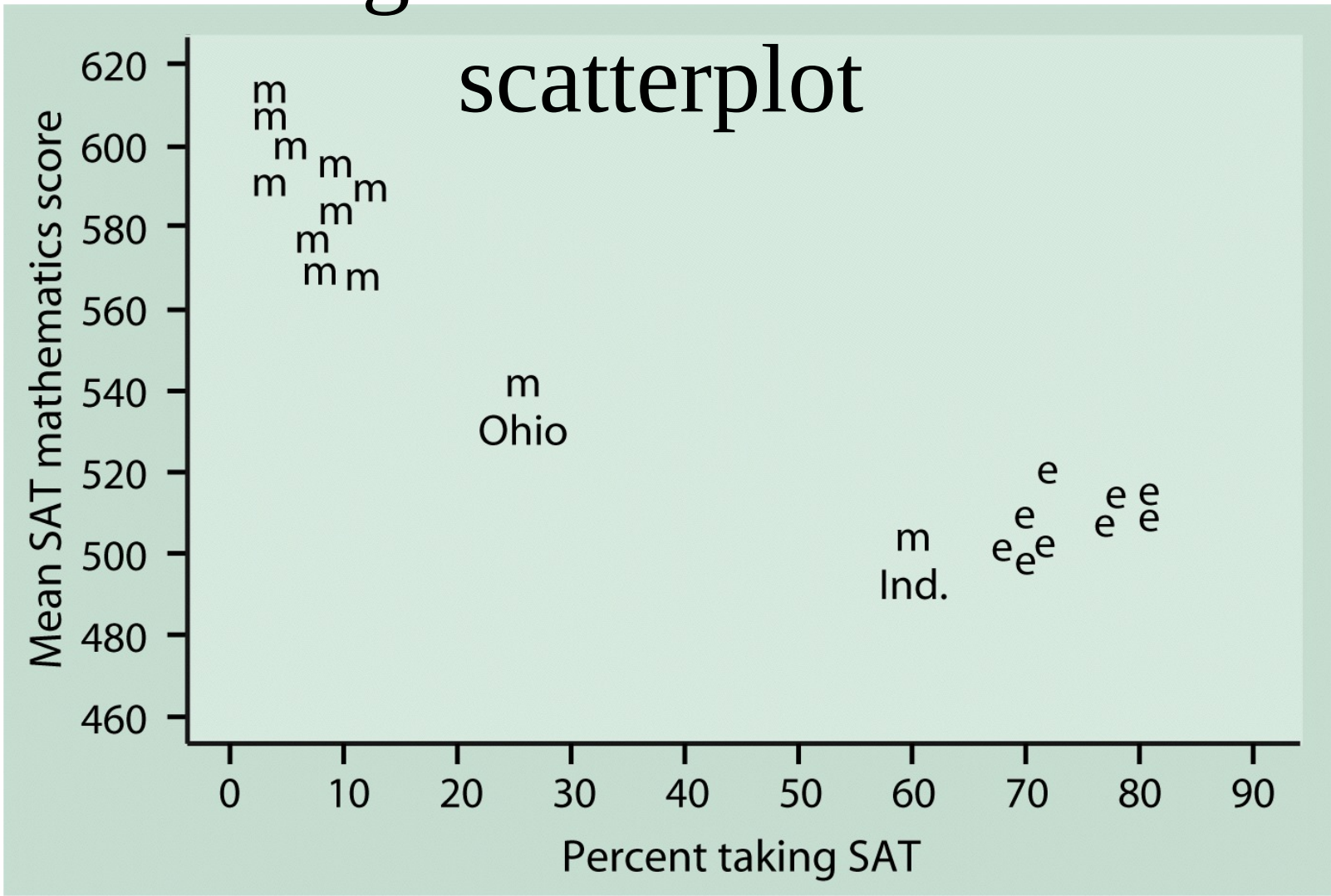
Example: Crash test



Bivariate Fit of Price By H Size

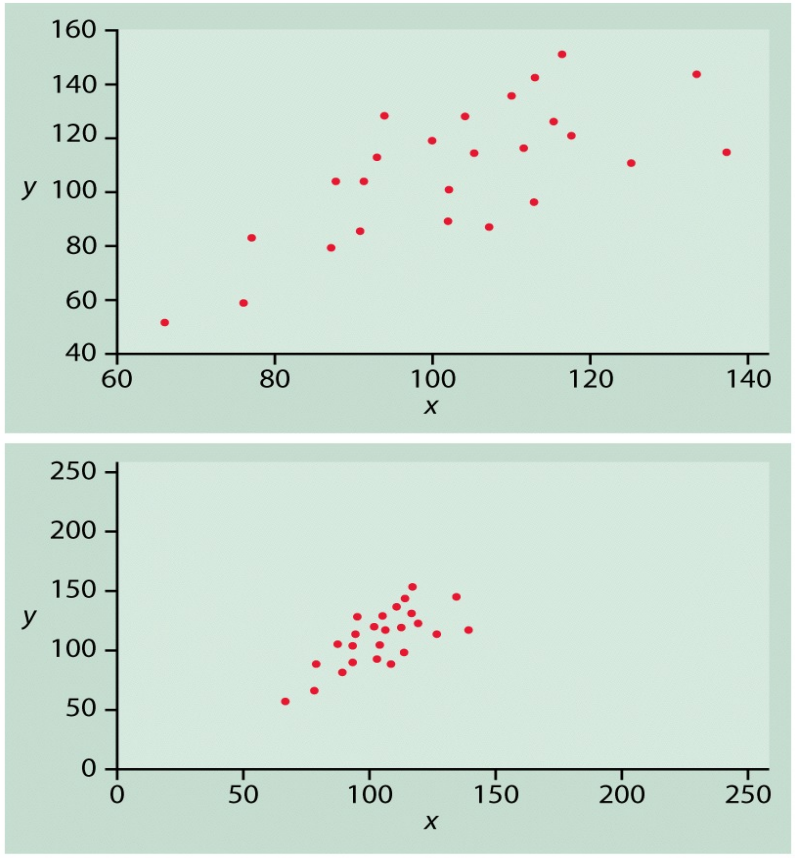


Categorical variable in a scatterplot



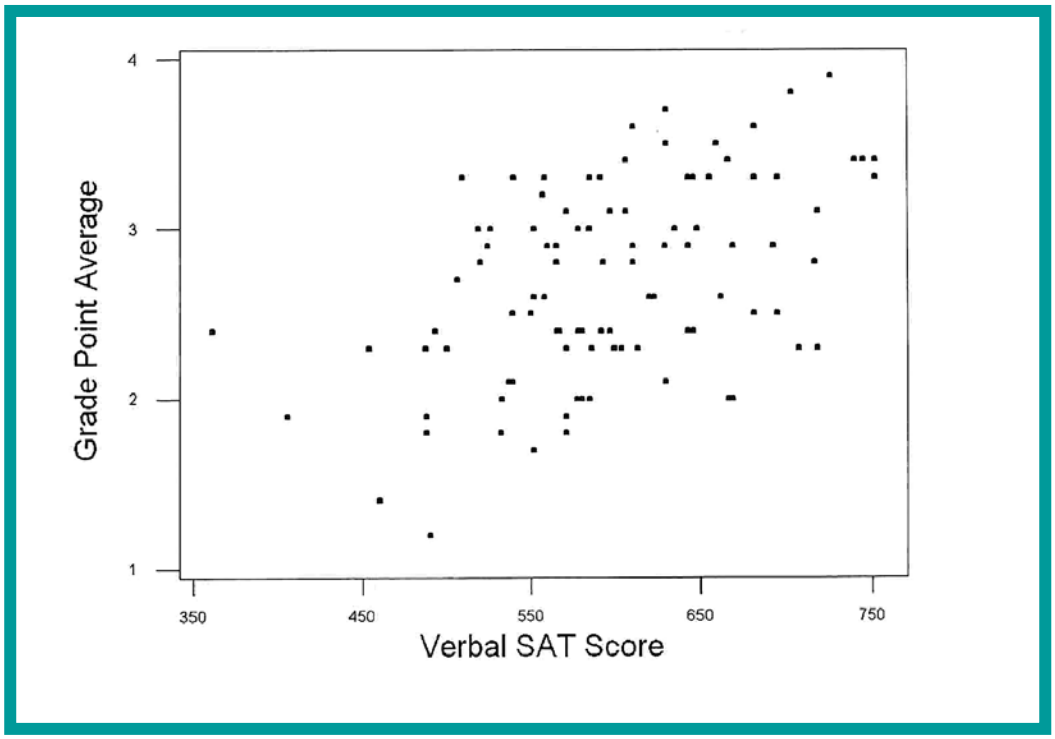
- **Examples** For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation or no correlation.
- Minimum daily temperature and heating costs
- Interest rate and number of loan applications
- Incomes of husbands and wives when both have full-time jobs
- Ages of boyfriends and girlfriends
- Height and IQ
- Height and shoe size
- Your Maths score in the Leaving Cert and your Irish score in the Leaving Cert

Which one shows a stronger relationship?



Example 3: Verbal SAT and GPA

Scatterplot of GPA and verbal SAT score.

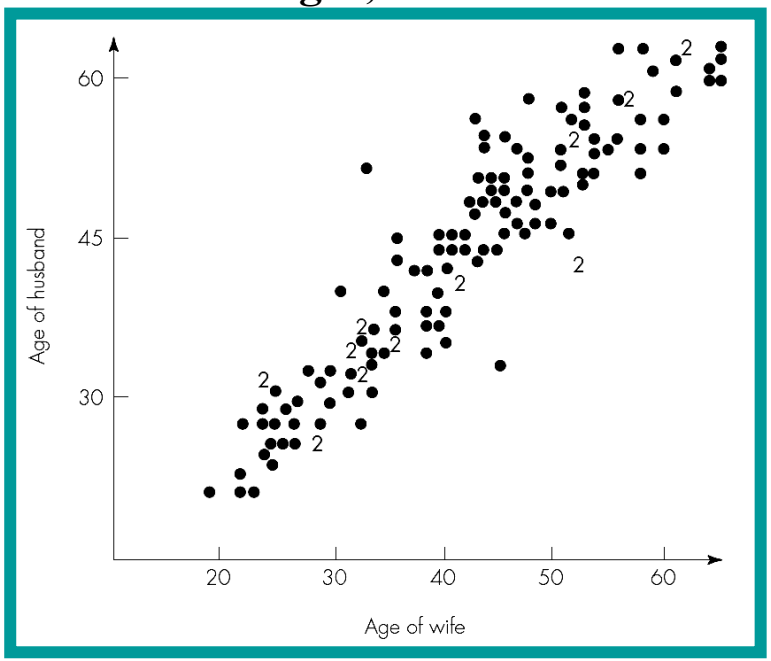


The correlation is **.485**, indicating a moderate positive relationship.

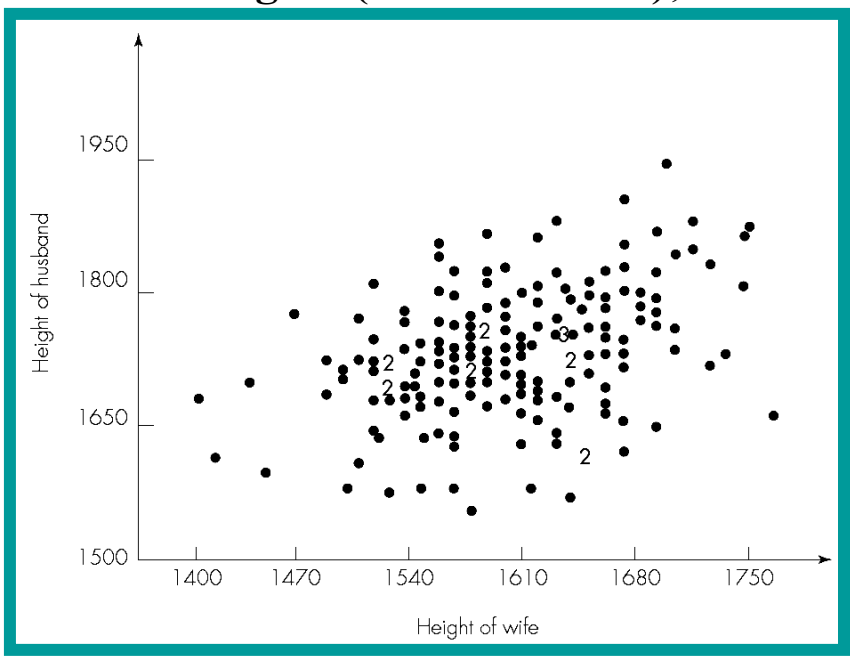
Higher verbal SAT scores tend to indicate higher GPAs as well, but the relationship is nowhere close to being exact.

Example 4: Husbands' and Wifes' Ages and Heights

Scatterplot of British husbands' and wives' ages; $r = .94$



Scatterplot of British husbands' and wives' heights (in millimeters); $r = .36$



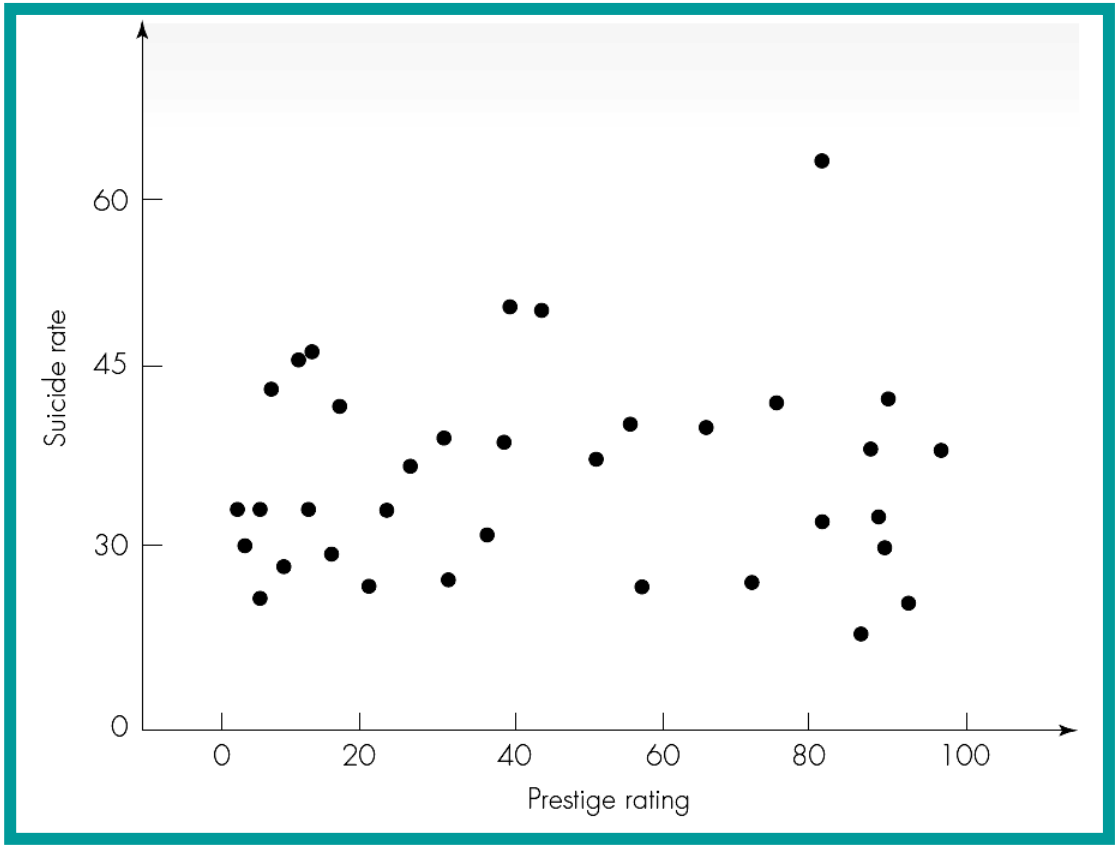
Husbands' and wives' ages are likely to be closely related, whereas their heights are less likely to be so.

Source: Marsh (1988, p. 315) and Hand et al. (1994, pp. 179-183)

Example 5: Occupational Prestige and Suicide Rates

Plot of suicide rate versus occupational prestige for 36 occupations.

Correlation of **.109** – these is not much of a relationship. If outlier removed *r* drops to **.018**.

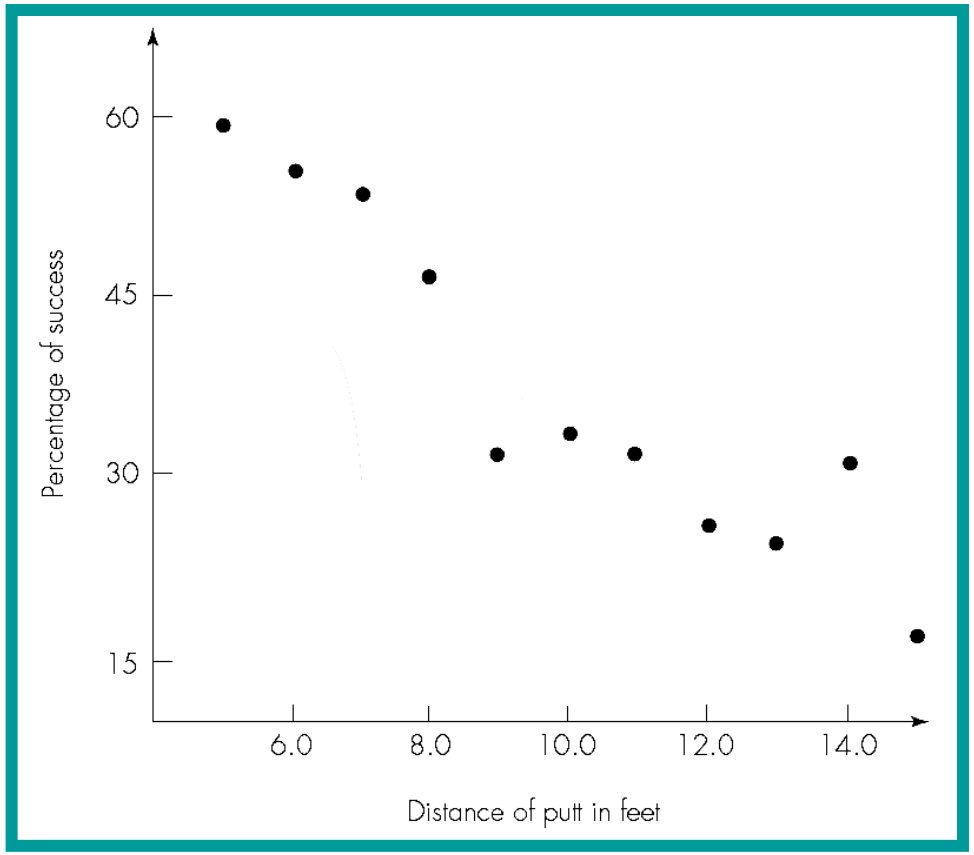


Source: Labovitz (1970, Table 1) and Hand et al. (1994, pp. 395-396)

Example 6: Professional Golfers' Putting Success

Scatterplot of distance of putt and putting success rates.

Correlation $r = -.94$. Negative sign indicates that as distance goes up, success rate goes down.



Source: Iman (1994, p. 507)

10.4 Specifying Linear Relationships with Regression

Goal: Find a straight line that comes as close as possible to the points in a scatterplot.

- Procedure to find the line is called **regression.**
- Resulting line is called the **regression line.**
- Formula that describes the line is called the **regression equation.**
- Most common procedure used gives the **least squares regression line.**

Linear Regression

- Correlation measures the direction and strength of the linear relationship between two quantitative variables
- A regression line
  - summarizes the relationship between two variables if the form of the relationship is linear.
  - describes how a response variable y changes as an explanatory variable x changes.
  - is often used as a mathematical model to predict the value of a response variable y based on a value of an explanatory variable x.

The Equation of the Line

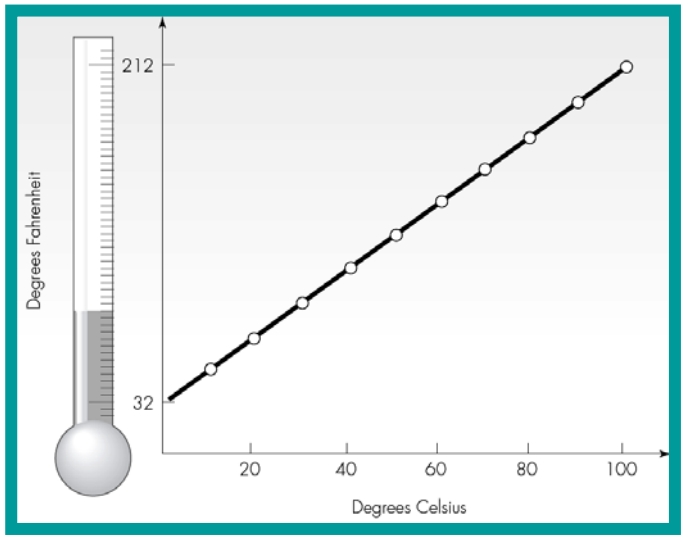
$$y = a + bx$$

- ***a* = intercept** – where the line crosses the vertical axis when $x = 0$.
- ***b* = slope** – how much of an increase there is in y when x increases by one unit.

***y* = temperature in Fahrenheit**
***x* = temperature in Celsius**

$$y = 32 + 1.8x$$

Intercept of 32 = temperature in F when C temperature is zero. Slope of 1.8 = amount by which F temperature increases when C temperature increases by one unit.



Least Squares

Introduction

- We have just mentioned that one should not always conclude that because two variables are correlated that one variable is causing the other to behave a certain way. However, sometimes this is the case, eg: interest rate and number of loan applications.
- In this section we will deal with datasets which are correlated and in which one variable, x, is classed as an **independent variable** and the other variable, y, is called a **dependent variable** as the value of y depends on x.

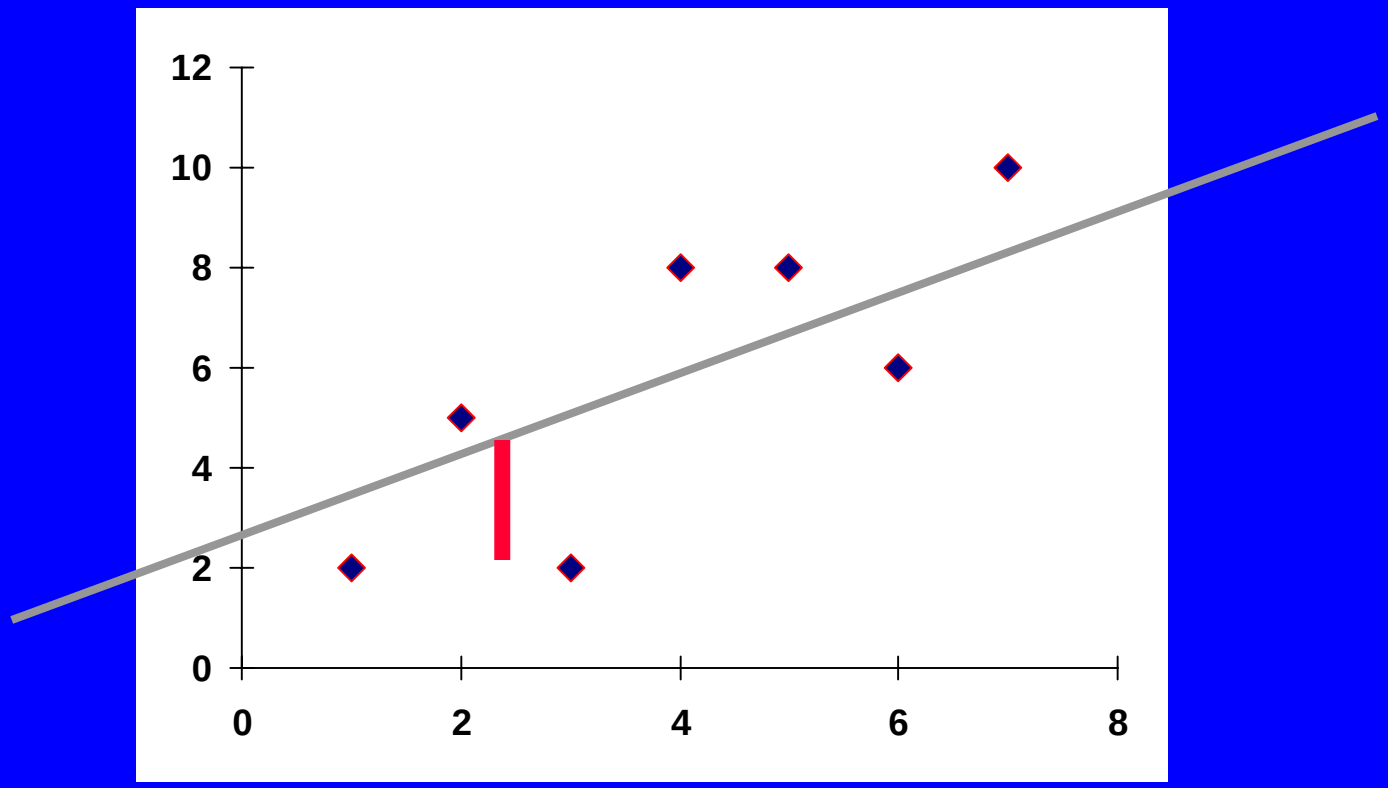
Least Squares

- We saw that correlation implies a linear relationship. Well a line is described by the equation

  - **$y = a + bx$**

- where **b** is the **slope** of the line and **a** is the **intercept** i.e. where the line cuts the **y** axis.
- The intercept a is just the value that y takes when x is zero.
- The slope b is how much y increases by when x increases by one unit.

- Suppose we have a dataset which is strongly correlated and so exhibits a linear relationship, how would we draw a line through this data so that it fits all points best?
- We use the principle of least squares, we draw a line through the dataset so that the sum of the squares of the deviations of all the points from the line is minimised.

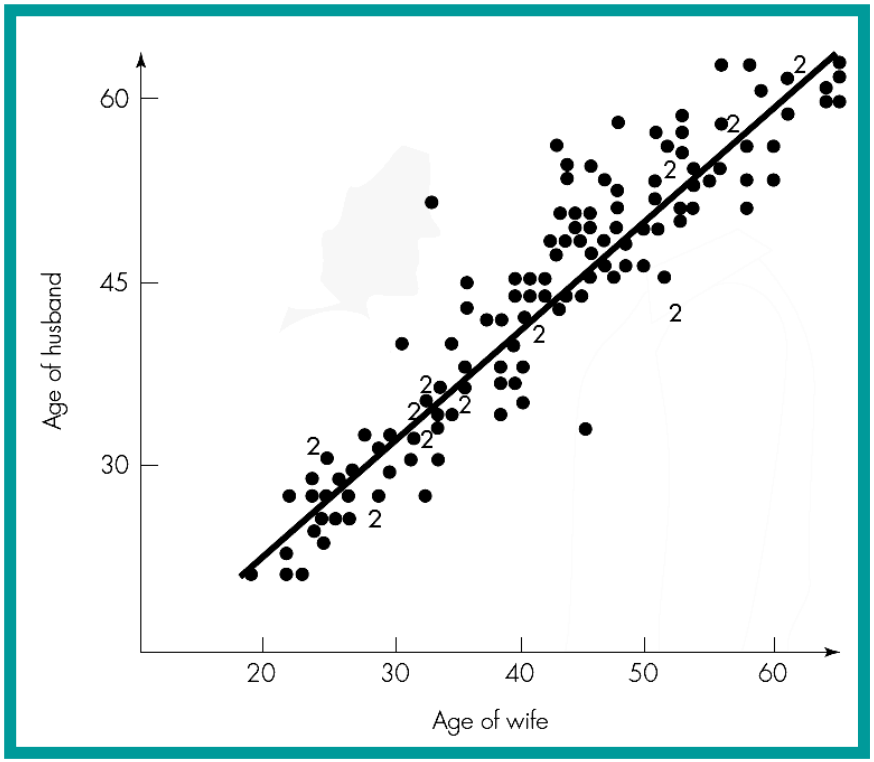

12
10
8
6
4
2
0
0
2
4
6
8

Example 7: Husbands' and Wifes' Ages, Revisited

Scatterplot of British husbands' and wives' ages with regression equation: $y = 3.6 + 0.97x$

Intercept: has no meaning.
Slope: for every year of difference in two wives ages, there is a difference of about 0.97 years in their husbands ages.

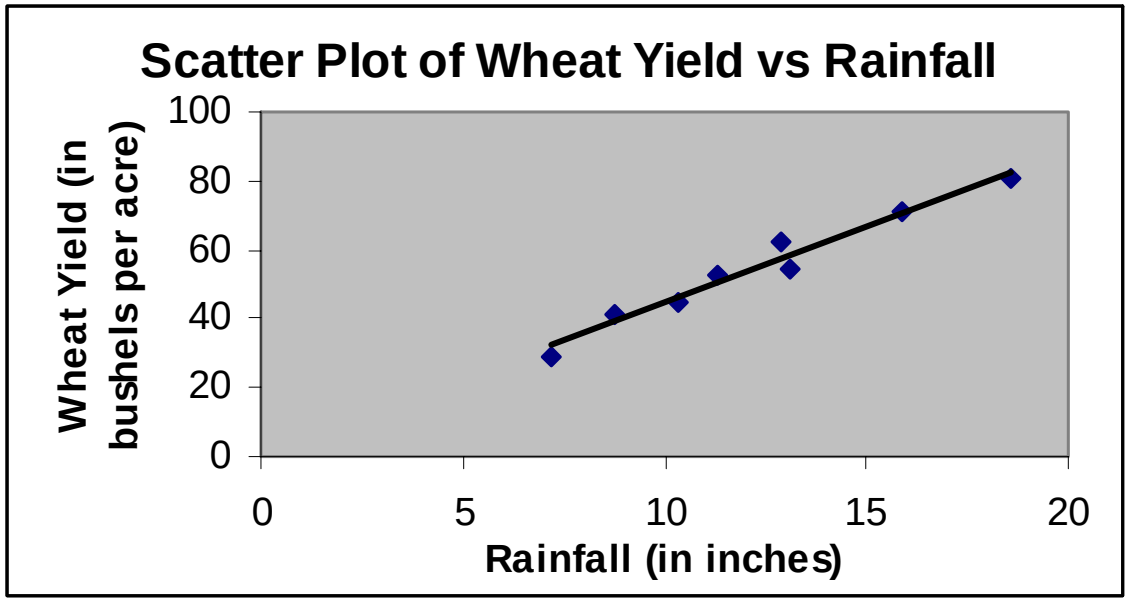
Wife's Age	Predicted Age of Husband
20 years	3.6 + (.97)(20) = 23.0 years
25 years	3.6 + (.97)(25) = 27.9 years
40 years	3.6 + (.97)(40) = 42.4 years
55 years	3.6 + (.97)(55) = 57.0 years



husband's age = 3.6 + (.97)(wife's age)

Example: Predicting the Wheat Yield based on Rainfall

Rainfall (X) (in inches)	Wheat Yield (Y) (bushels per acre)
12.9	62.5
7.2	28.7
11.3	52.2
18.6	80.6
8.8	41.6
10.3	44.5
15.9	71.3
13.1	54.4



The regression equation is: $y = 0.2292 + 4.4237x$

Correlation = $r = 0.9823$

Regression equation: $y = 0.2292 + 4.4237x$

Use this equation to predict the wheat yield when there are 15 inches of rain.

We use the symbol $\hat{y}$ to indicate a predicted value.

$\hat{y} = 0.2292 + 4.4237*15 = 66.5847$

So when there is 15 inches of rain we predict a 66.6 bushel yield.

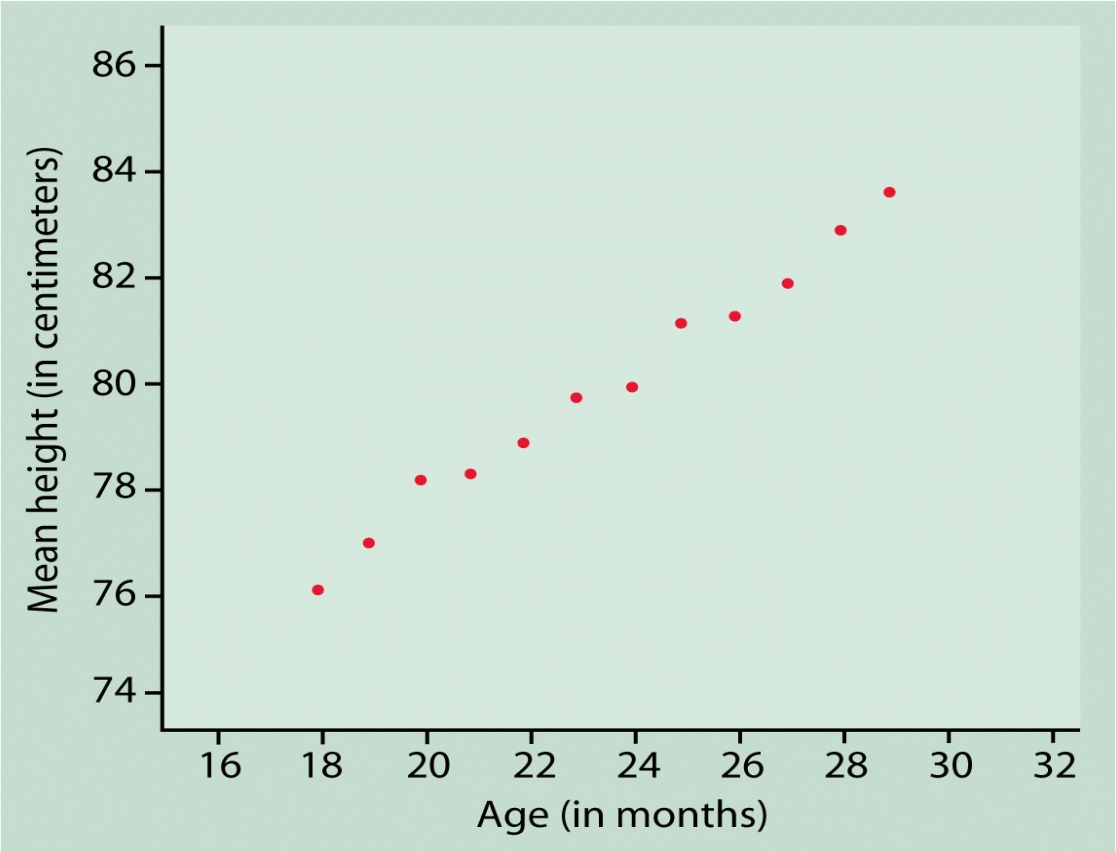
What is the intercept? Explain what it tells us.

The intercept is 0.2292. It represents the predicted yield when there is no rain.

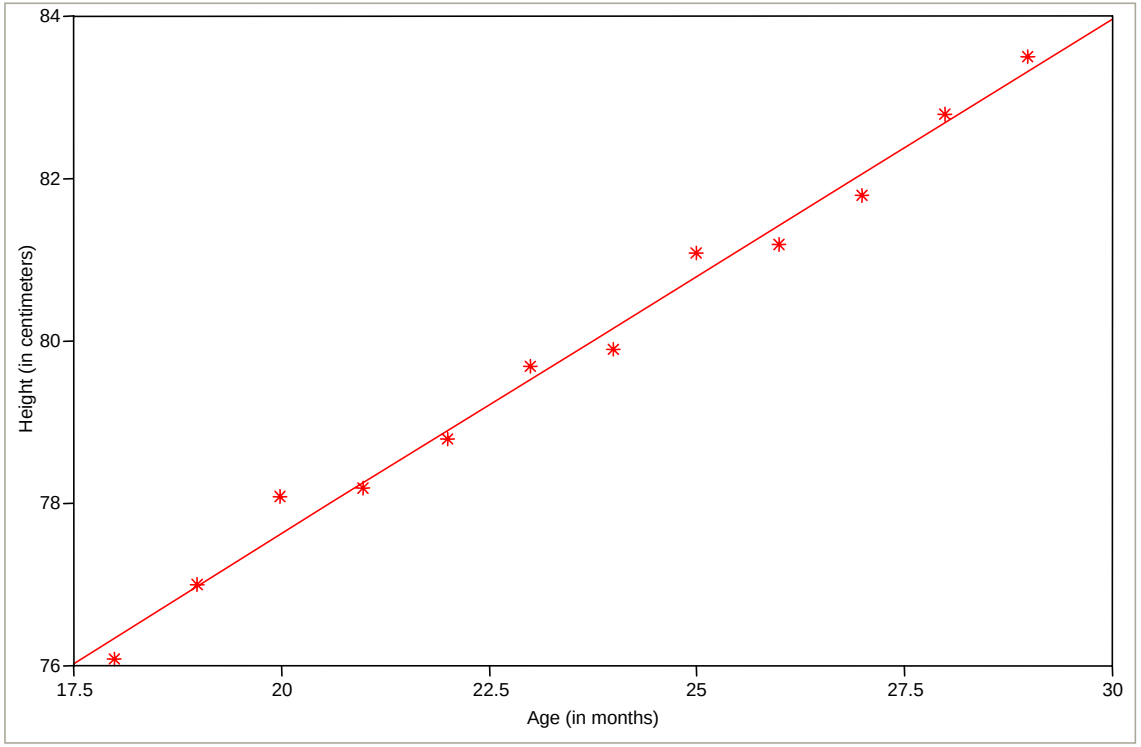
What is the slope? Explain what it tells us.

The slope is 4.4237. For each increase of 1 inch of rain we would expect to see an increase of 4.4237 bushels in yield.

Age vs. Mean Height



Example: Age vs. Height



$\bar{x} = 23.5, \bar{y} = 79.85$

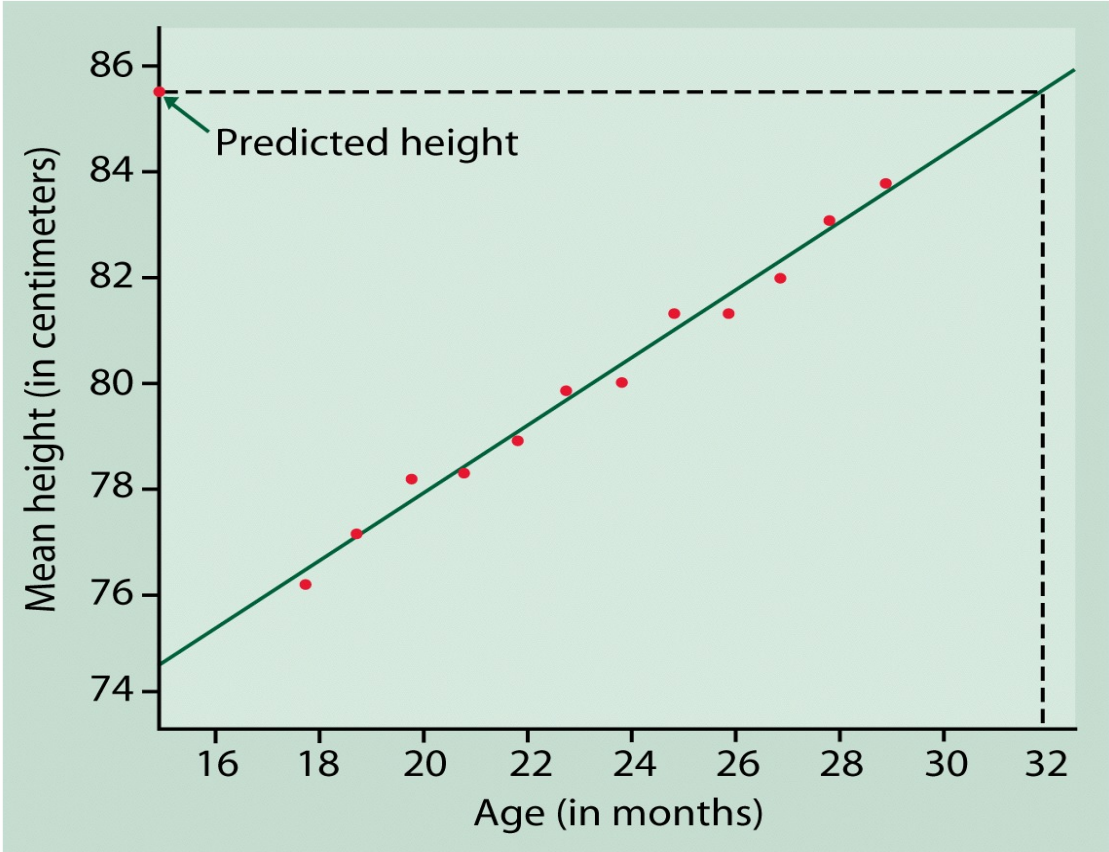
$s_x = 3.606, s_y = 2.302$

$r = 0.9944$

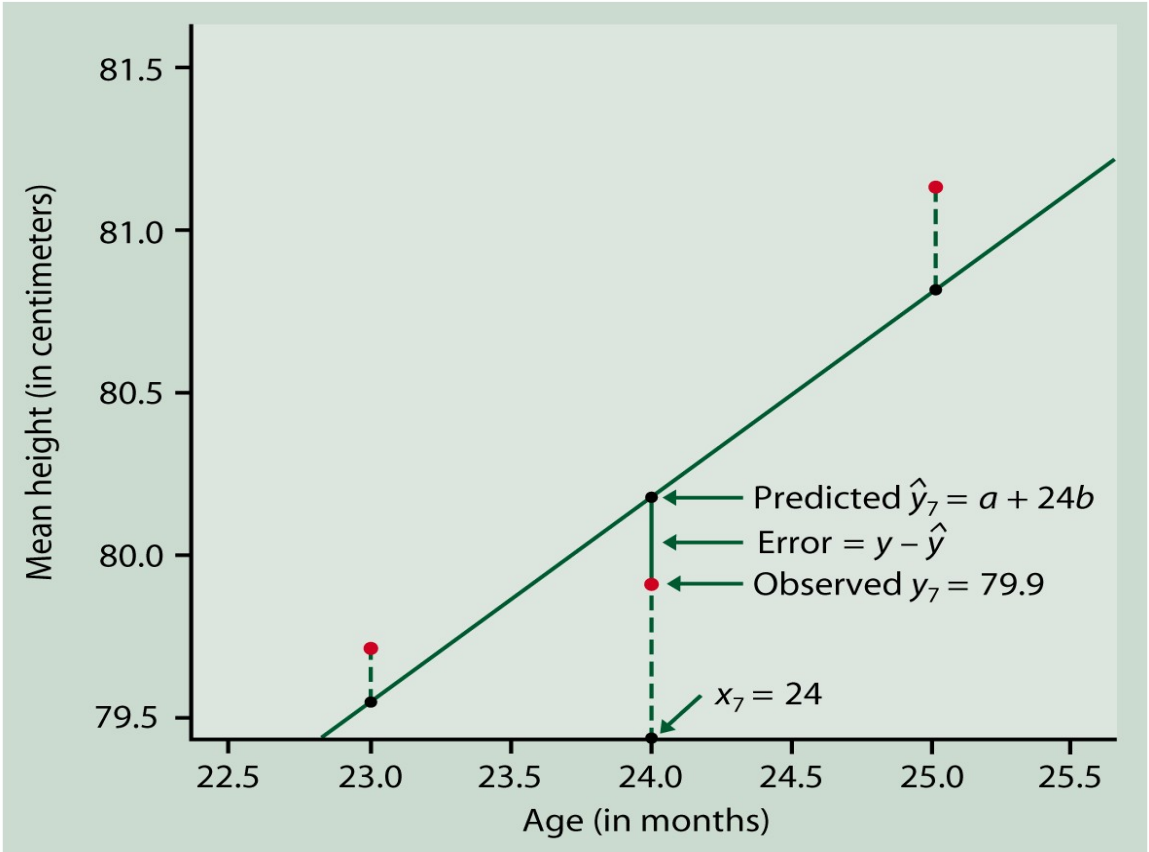
$$b = r \frac{s_y}{s_x}$$

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

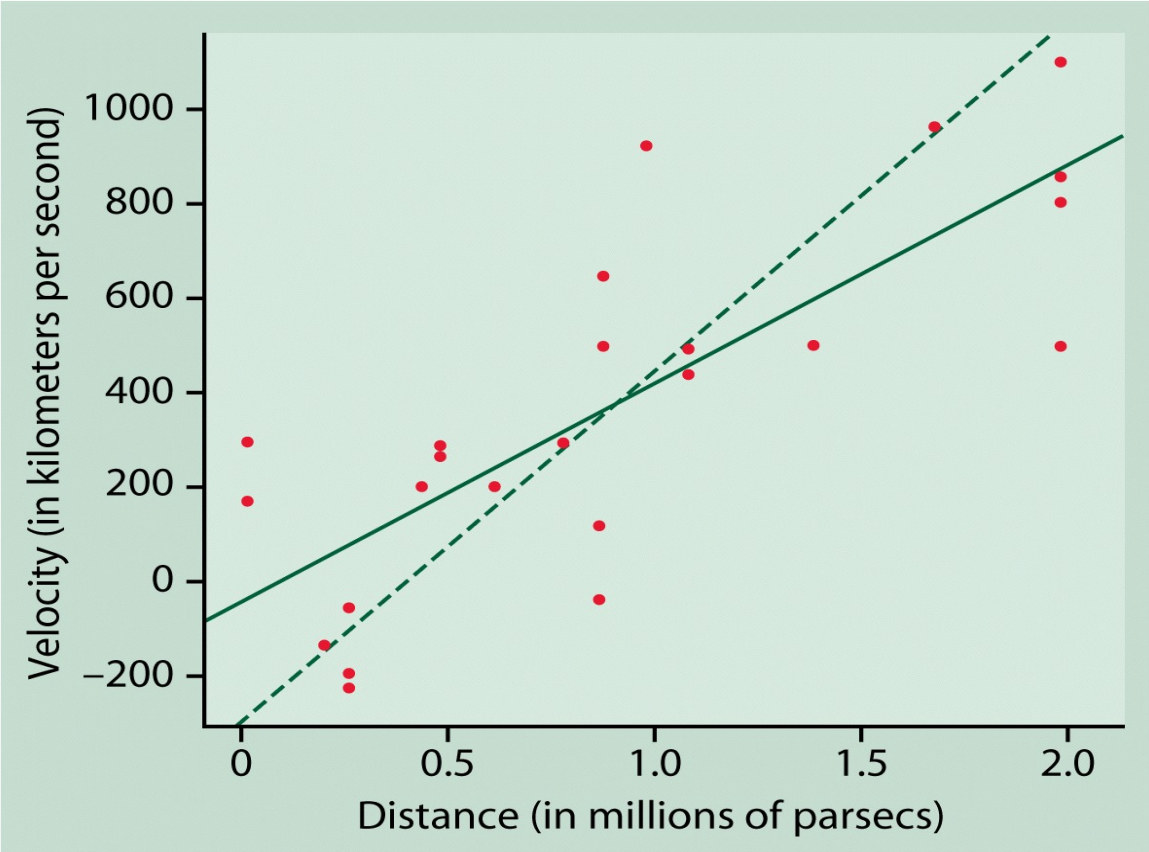
To predict mean height at age 32 months?



Error


81.5
81.0
80.5
80.0
79.5
Mean height (in centimeters)
22.5
23.0
23.5
24.0
24.5
25.0
25.5
Age (in months)
Predicted $\hat{y}_7 = a + 24b$
Error $= y - \hat{y}$
Observed $y_7 = 79.9$
$x_7 = 24$

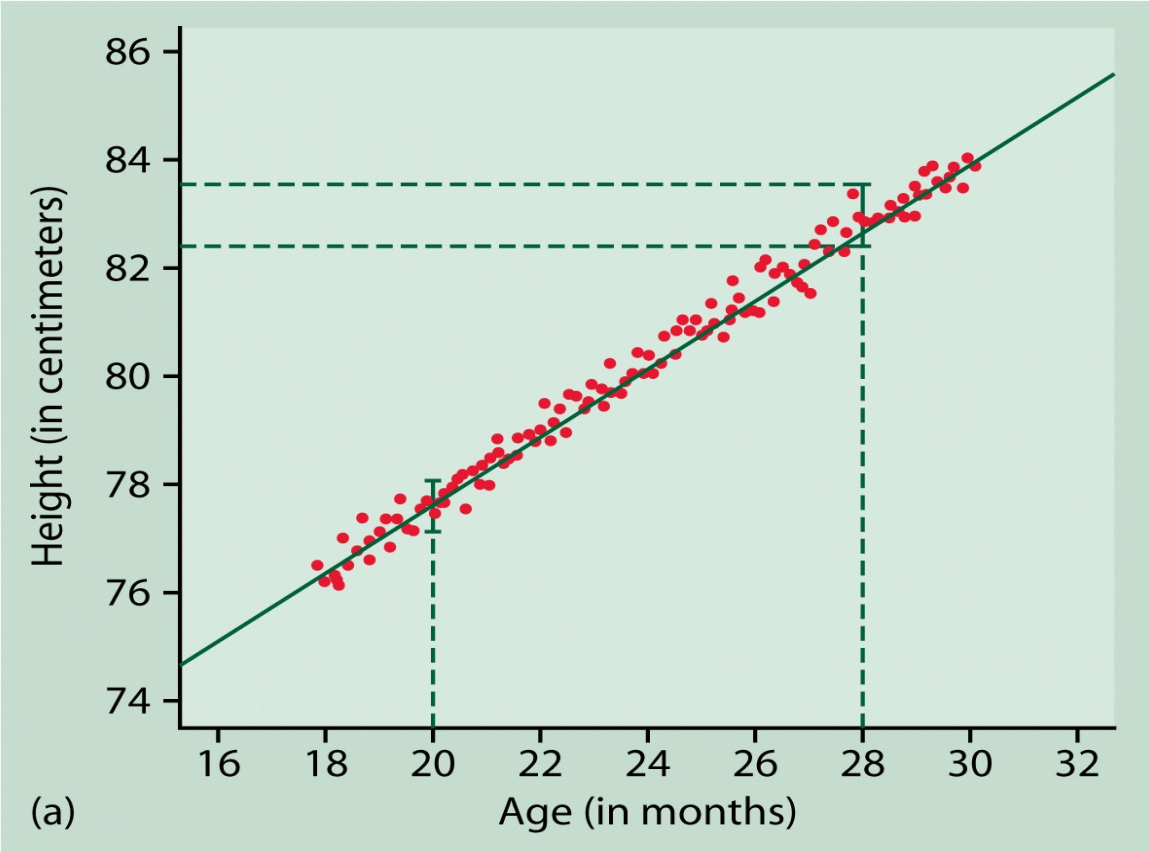
Regression lines depend on (x,y) or (y,x).



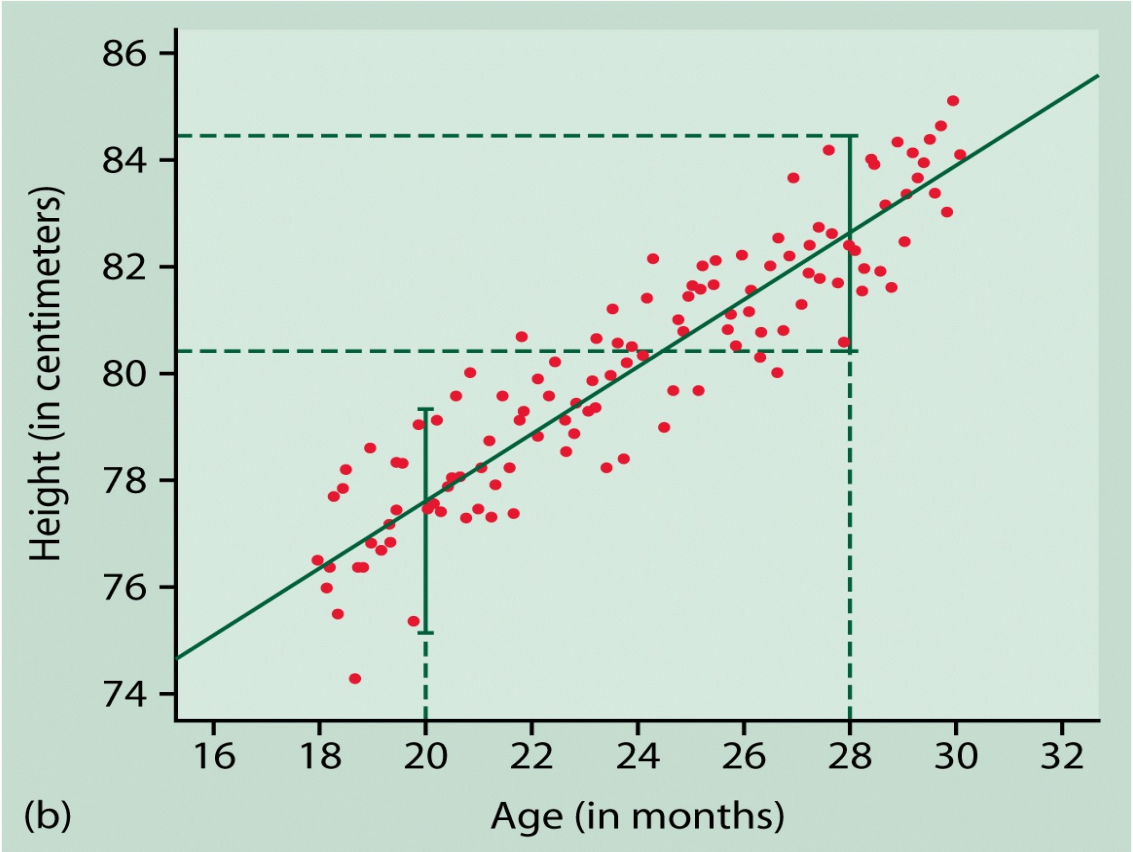
Coefficient of Determination r^2

- The **square of the correlation** is the proportion of variation in the values of y that is explained by the regression model with x.
- $0 \leq r^2 \leq 1$.
- The larger r^2 , the stronger the *linear* relationship.
- The closer r^2 is to 1, the more confident we are in our prediction.

Age vs. Height: r^2=0.9888.



Age vs. Height: r^2=0.849.



Extrapolation

Not a good idea to ***use a regression equation to predict values far outside the range where the original data fell.***

No guarantee that the relationship will continue beyond the range for which we have data.

Use the equation only for a *minor extrapolation* beyond the range of the original data.

Final Cautionary Note:
Easy to be misled by inappropriate interpretations and uses of correlation and regression.
Chapter 11: how that can happen, and how you can avoid it.

Case Study 10.1: Are Attitudes about Love and Romance hereditary?

Study Details:

- 342 pairs of **monozygotic (MZ) twins** (share 100% of genes);
  100 pairs of **dizygotic (DZ) twins** (share about 50% of genes);
  172 **spouse pairs** (a twin and his or her spouse)
- Each filled out "**Love Attitudes Scale**" (LAS) questionnaire;
  42 statements (7 questions on each of **6 love styles**);
  Respondents ranked 1 (strongly agree) to 5 (strongly disagree).
- **Six scores** (1 for each love type) determined for each person.
  Correlations were computed for each of three types of pairs.

Source: Waller and Shaver, September 1994.

Case Study 10.1: Are Attitudes about Love and Romance hereditary?

Key: If love styles are genetic then matches between MZ twins should be much higher than those between DZ twins.

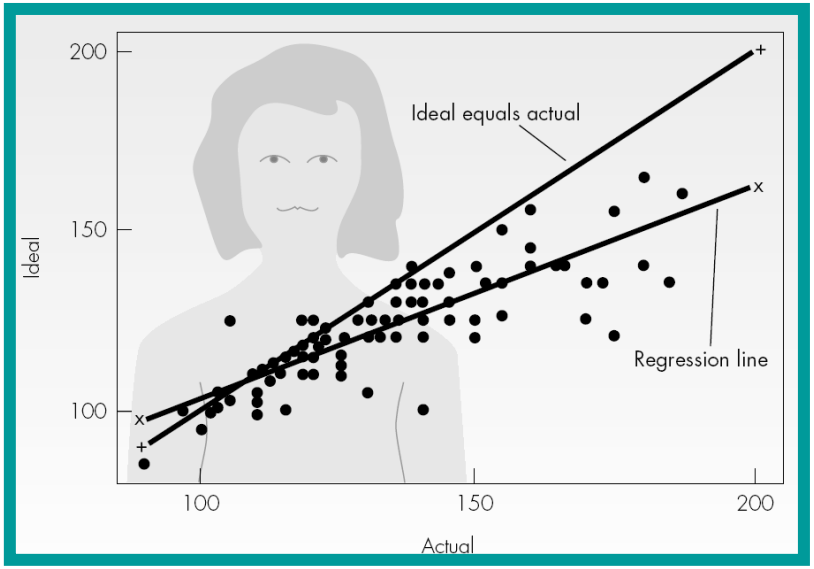
Results: Correlations are ***not*** higher for the MZ twins than they are for the DZ twins.

	Correlation		
	Monozygotic Twins	Dizygotic Twins	Spouses
Love Style			
Eros	.16	.14	.36
Ludus	.18	.30	.08
Storge	.18	.12	.22
Pragma	.40	.32	.29
Mania	.35	.27	−.01
Agape	.30	.37	.28
Personality Trait			
Well-being	.38	.13	.04
Achievement	.43	.16	.08
Social closeness	.38	.01	−.04

This surprising, and very unusual, finding suggests that genes are not important determinants of attitudes toward romantic love. Rather, the common environment appears to play the cardinal role in shaping familial resemblance on these dimensions. (p. 271)

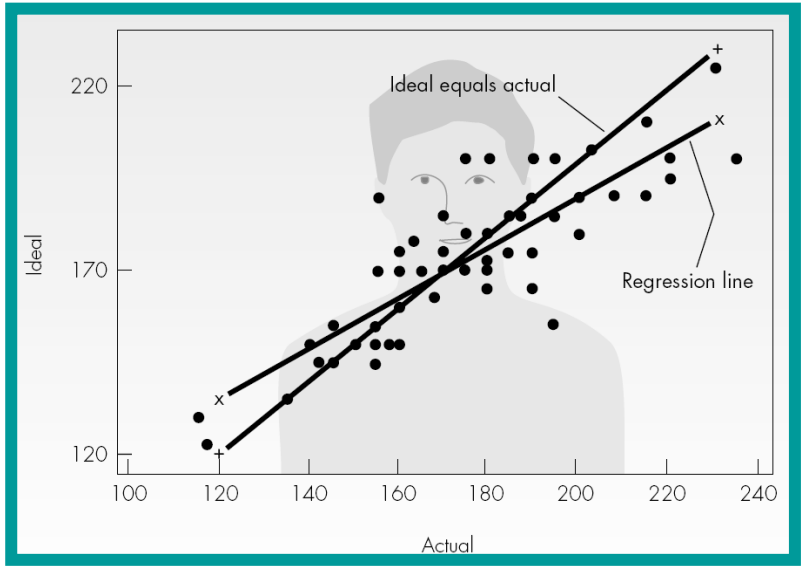
Case Study 10.2: A Weighty Issue: Women Want Less, Men Want More

Ideal versus actual weight for females



Equation: ideal = 43.9 + 0.6(actual)

Ideal versus actual weight for males



Equation: ideal = 52.5 + 0.7(actual)

- If everyone at their ideal weight, all points fall on line *Ideal = Actual*.
- Most women fall below that line.
- Men under 175 pounds would prefer to weight same or more, while men over 175 pounds would prefer to weight same or less.

For Those Who Like Formulas

The Data

n pairs of observations, (x_i, y_i), $i = 1,2, \ldots, n,$ where x_i is plotted on the horizontal axis and y_i on the vertical axis.

Summaries of the Data, Useful for Correlation and Regression

$$\text{SSX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$\text{SSY} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

$$\text{SXY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

Correlation for a Sample of n Pairs

$$r = \frac{\text{SXY}}{\sqrt{\text{SSX}}\sqrt{\text{SSY}}}$$

The Regression Slope and Intercept

$$\text{slope} = b = \frac{\text{SXY}}{\text{SSX}}$$

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