Chapter 10: Correlation and Regression

Scatterplot--a two-dimensional graph of data values

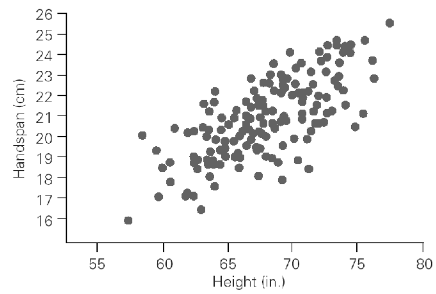
- Use a ***scatterplot*** to look at the relationship between two quantitative variables
- Plot has one variable's values along the vertical axis and the other variable's values along the horizontal axis

Simple Linear Regression—a method to generate linear equation that describes the average relationship between a response and explanatory variable

Correlation, a statistic that measures the ***strength*** and ***direction*** of a linear relationship

Relationships Between Quantitative Variables

heights (in inches) and fully stretched handspans (in centimeters) of 167 college students



Relationships Between Quantitative Variables

Questions that might be asked

1. What is the *average* pattern? Does it look like a straight line or is it curved?
2. What is the direction of the pattern?
3. How much do individual points vary from the average pattern?
4. Are there any unusual data points?

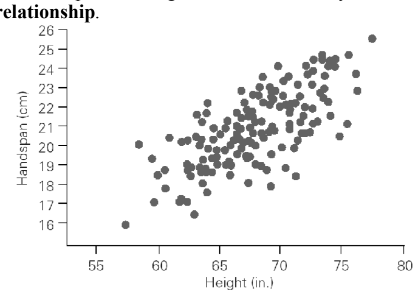
Relationships Between Quantitative Variables

Positive/Negative Association

- **Positive association--**the values of one variable tend to increase as the values of the other variable increase.
- **Negative association--t**he values of one variable tend to decrease as the values of the other variable increase.

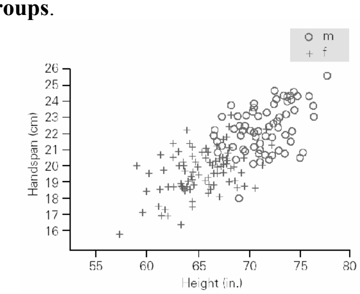
Positive Association

- Taller people tend to have greater handspan measurements than shorter people do
- The handspan and height measurements may have a **linear relationship**.



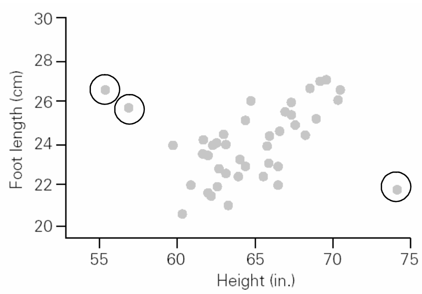
Looking at Subgroups

Are there differences between genders?--use different plotting symbols or colors to represent **different subgroups**.



Looking for Outliers

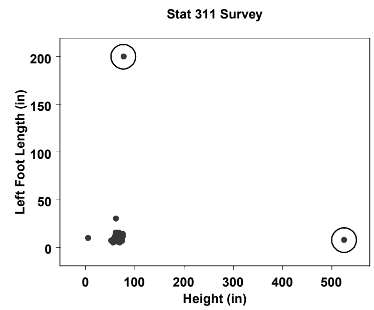
Identify points that have an usual combination of data values.



7

Relationships Between Quantitative Variables

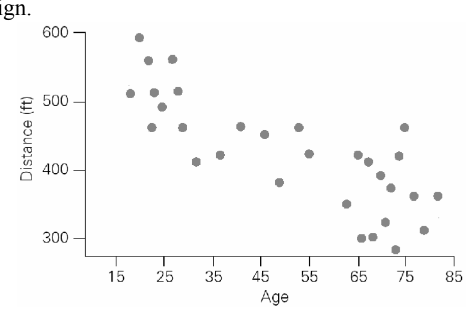
heights (in inches) and fully left foot length (in inches) of Stat 311 students



8

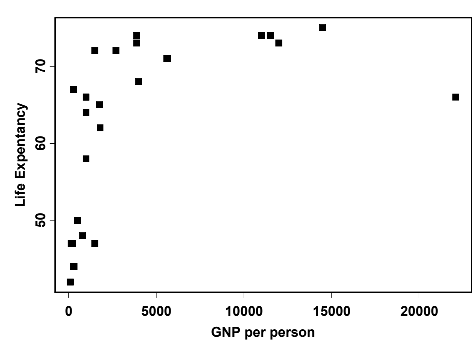
Negative Association

A research firm determined the **maximum distance** at which each of 30 drivers could read a newly designed sign.



9

Curvilinear Relationship



10

Correlation Coefficient

Correlation is a measure of the linear association between two variables, x and y.

Pearson product moment coefficient of correlation, $\boldsymbol{r}$, is a measure of the strength of the linear relationship between two variables x and y. It is computed for a ***sample*** of n measurements on x and y as:

$$r = \frac{1}{n-1}\sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

11

Correlation Coefficient

Can also be written as $r = \dfrac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$

where SS stands for sum of squares

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$SS_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

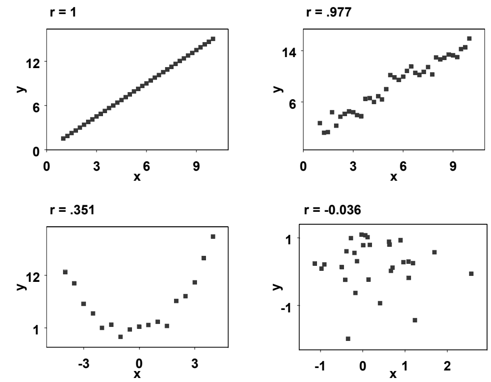
$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

12

Correlation Coefficient

1. r is positive when the slope is positive and likewise negative when the slope is negative

2. If $SS_{xy} = 0$, then $r = 0$ ➔ that there is no association between the magnitudes of the two variables; or, a change in the magnitude of one variable does not imply a change in the magnitude of the other variable
3. The correlation coefficient, r, is unitless and assumes a value between –1 and +1, regardless of the units of x and y
4. The Cor(X,Y) = Cor(Y,X)



<u>Example: From Zar, 1974, page 238</u>

Data (subset) for wing and tail lengths, in cm, among birds of a particular species

Wing (x)	Tail (y)	x^2	y^2	xy
10.4	7.4	108.16	54.76	76.96
10.8	7.6	116.64	57.76	82.08
11.1	7.9	123.21	62.41	87.69
10.2	7.2	104.04	51.84	73.44
10.3	7.4	106.09	54.76	76.22
10.2	7.1	104.04	50.41	72.42
10.7	7.4	114.49	54.76	79.18
10.5	7.2	110.25	51.84	75.6
84.2	**59.2**	**886.92**	**438.54**	**623.59**

<u>Example: From Zar, 1974, page 238 (continued)</u>

Wing (x)	Tail (y)	x^2	y^2	xy
84.2	**59.2**	**886.92**	**438.54**	**623.59**

← Column sums

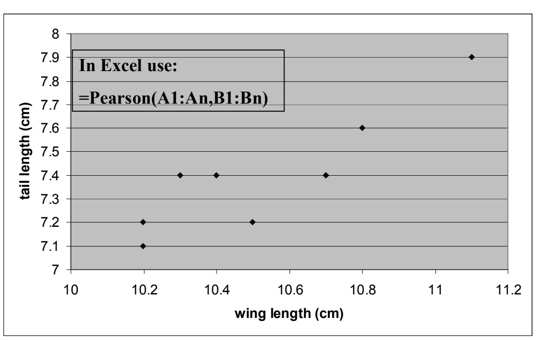
$$SS_{xx} = 886.92 - \frac{(84.2)^2}{8} = 0.715 \text{ cm}^2$$

$$SS_{yy} = 438.54 - \frac{(59.2)^2}{8} = 0.46 \text{ cm}^2$$

$$SS_{xy} = 623.59 - \frac{(84.2)(59.2)}{8} = 0.51 \text{ cm}^2$$

$$r = \frac{0.51}{\sqrt{0.715 \times 0.46}} = \frac{0.51}{0.573} = 0.89$$

<u>Example: From Zar, 1974, page 238 (continued)</u>



Correlation Coefficient--Interpretations

For the wing/tail length data, r is strongly positive, indicating that tail length tends to increase as wing length increases---***for this sample of 8 birds***

<u>Interpretations of an observed association</u>

1. Correlation does not prove causation, although it is possible that there is causation.

2. There may be causation, but confounding factors contribute as well, making this causation difficult to prove.

Correlation Coefficient--Interpretations

Interpretations of an observed association (continued)

3. There is no causation. The association is explained by how the explanatory and response variables are both affected by other variables.
4. The response variable is causing a change in the explanatory variable.

Correlation Coefficient

Population Correlation Coefficient

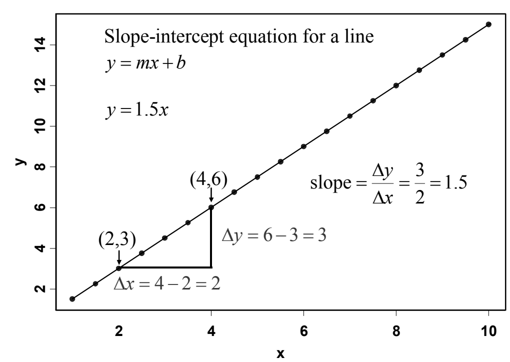
Denoted by ρ (rho)

ρ is estimated by the sample statistic, r

Hypothesis test for ρ given on page 527—we will not specifically go over this

Simple Linear Regression

Regression equation—an equation that describes the average relationship between a response (dependent) and an explanatory (independent) variable.

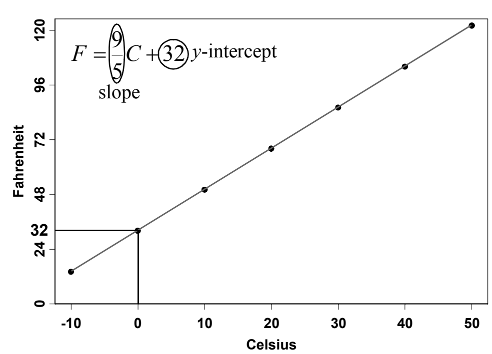


Deterministic Model

A model that defines an exact relationship between variables.

Example: $y = 1.5x$

There is no allowance for error in the prediction of y for a given x.



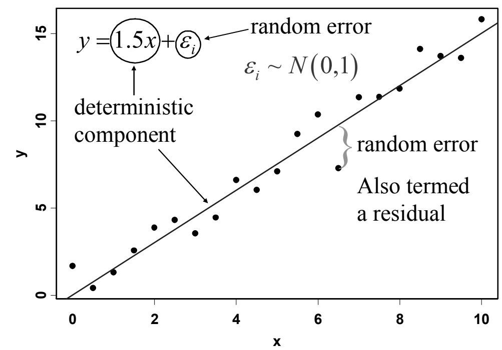
Probabilistic Model

A model that accounts for ***random error***.

Includes both a deterministic component and a random error component.

$$y = 1.5x + \text{ random error}$$

This model hypothesizes a probabilistic relationship between *y* and *x*.



Probabilistic Model—General Form

y = Deterministic component + Random component

where *y* is the "variable of interest".

Assume that the mean value of the random error is zero ➔ the mean value of *y*, *E*(*y*), equals the deterministic component of the model

First-Order (Straight Line) Probabilistic Model

$y = \boxed{\beta_0 + \beta_1 x} + \varepsilon$ where *y* = ***Dependent variable***
x = ***Independent variable***

β_0 = ***population y-intercept of the line***—the point at which the line intersects or cuts through the *y*-axis

β_1 = ***population slope of the line***—the amount of increase (or decrease) in the deterministic component of *y* for every 1-unit increase (or decrease) in *x*.

ε = random error component

First-Order (Straight Line) Probabilistic Model

β_0 and β_1 are population parameters. They will only be known if the population of all (*x*, *y*) measurements are available.

β_0 and β_1, along with a specific value of the independent variable *x* determine the ***mean value*** of the dependent variable *y*.

Model Development

β_0 and β_1 will generally be unknown.

The process of developing a model, estimating model parameters, and using the model can be summarized in these 5-steps:

1. Hypothesize the deterministic component of the model that relates the mean, *E*(*y*) to the independent variable *x*

$$E(y) = \beta_0 + \beta_1 x$$

2. Use sample data to estimate unknown model parameters

$$\text{find estimates: } \hat{\beta}_0 \text{ or } b_0, \hat{\beta}_1 \text{ or } b_1$$

Model Development (continued)

3. Specify the probability distribution of the random error term and estimate the SD of this distribution

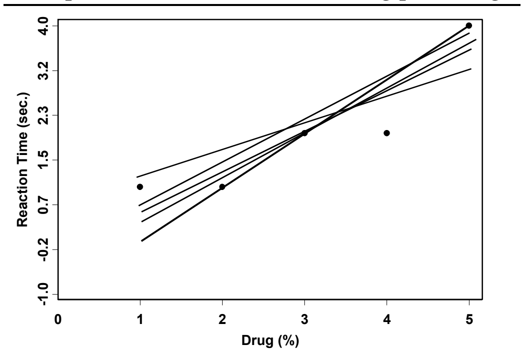
$$\varepsilon_i \sim N(0, \sigma) \text{ --will revisit this later}$$

4. Statistically evaluate the usefulness of the model
5. Use model for prediction, estimation or other purposes

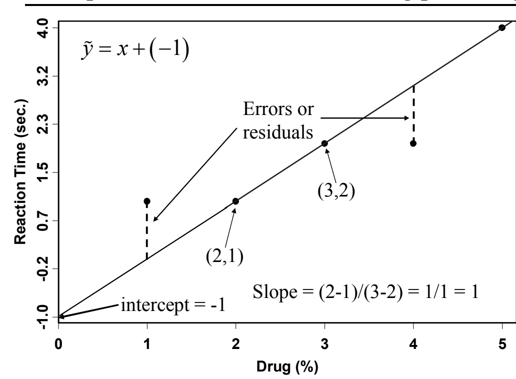
Example: Reaction time versus drug percentage

Subject	Amount of Drug (%) x	Reaction Time (seconds) y
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Example: Reaction time versus drug percentage



Example: Reaction time versus drug percentage



Example: Reaction time versus drug percentage

Errors of prediction---vertical differences between the observed and the predicted values of y

x	y	$\tilde{y} = -1 + x$	$(y - \tilde{y})$	$(y - \tilde{y})^2$
1	1	0	(1-0) = 1	1
2	1	1	(1-1) = 0	0
3	2	2	(2-2) = 0	0
4	2	3	(2-3) = -1	1
5	4	4	(4-4) = 0	0
			Sum of errors = 0	Sum of squared errors (SSE) = 2

Least Squares Line

Also called ***regression line***, or the ***least squares prediction equation***

Method to find this line is called the ***method of least squares***

For our example, we have a sample of n = 5 pairs of (x, y) values. The fitted line that we will calculate is written as $\hat{y} = b_0 + b_1 x$

$\hat{y}$ is an estimator of the mean value of y, $E(y)$;

b_0 and b_1 are estimators of β_0 and β_1

Least Squares Line (continued)

Define the sum of squares of the deviations of the *y* values about their predicted values for all *n* data points as:

$$\text{SSE} = \sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2 = \sum_{i=1}^{n}\left[y_i - \left(b_0 + b_1 x_i\right)\right]^2$$

We want to find b_0 and b_1 to make the SSE a minimum---termed ***least squares estimates***

$\hat{y} = b_0 + b_1 x$ is called the least squares line

37

Formulas for the Least Squares Estimates

Slope: $b_1 = \dfrac{\text{SS}_{xy}}{\text{SS}_{xx}}$ or $b_1 = r\dfrac{\text{SD}_y}{\text{SD}_x}$

$$\text{SS}_{xy} = \sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$\text{SS}_{xx} = \sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

y-intercept: $b_0 = \boxed{\bar{y}} - b_1\boxed{\bar{x}} = \dfrac{\sum y_i}{n} - b_1\dfrac{\sum x_i}{n}$

n = sample size

38

LS Calculations for Drug/Reaction Example

x_i	y_i	x_i^2	$x_i y_i$
1	1	1	1
2	1	4	2
3	2	9	6
4	2	16	8
5	4	25	20
$\sum x_i = 15$	$\sum y_i = 10$	$\sum x_i^2 = 55$	$\sum x_i y_i = 37$

$$b_1 = \frac{7}{10} = 0.7$$

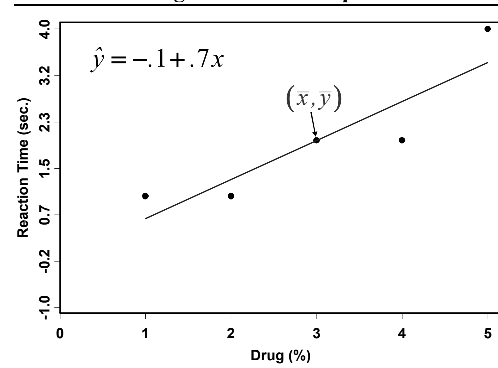
$$b_0 = \frac{10}{5} - (.7)\frac{15}{5} = 2 - (.7)(3) = 2 - 2.1 = -.1$$

$$\text{SS}_{xy} = 37 - \frac{(15)(10)}{5} = 37 - 30 = 7 \qquad \text{SS}_{xx} = 55 - \frac{(15)^2}{5} = 55 - 45 = 10$$

$$\boxed{\hat{y} = -.1 + .7x}$$

39

LS Line for Drug/Reaction Example



40

LS Calculations for Drug/Reaction Example

x	y	$\hat{y} = -.1 + .7x$	$(y - \hat{y})$	$(y - \hat{y})^2$
1	1	.6	(1-.6) = .4	.16
2	1	1.3	(1-1.3) = -.3	.09
3	2	2.0	(2-2.0) = 0	.00
4	2	2.7	(2-2.7) = -.7	.49
5	4	3.4	(4-3.4) = .6	.36
			Sum of errors = 0	Sum of squared errors (SSE) = 1.10

The LS line has a sum of errors = 0, but
SSE = 1.1 < 2.0 for visual model

41

Least Squares Line—Interpretation of $\hat{y} = -.1 + .7x$

Estimated intercept is negative
➔that the estimated ***mean reaction time*** is equal to -0.1 seconds when the amount of drug is 0%.

What does this mean since negative reaction times are not possible?

Model parameters should be interpreted only within the sampled range of the independent variable.

42

Least Squares Line—Interpretation of $\hat{y} = -.1 + .7x$

The slope of 0.7 implies that for every unit increase of x, the ***mean value*** of y is estimated to increase by 0.7 units.

In the context of the problem:

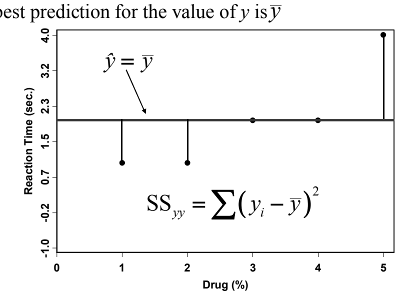
For every 1% increase in the amount of drug in the bloodstream, the mean reaction time is estimated to increase by 0.7 seconds over the sampled range of drug amounts from 1% to 5%.

43

Coefficient of Determination

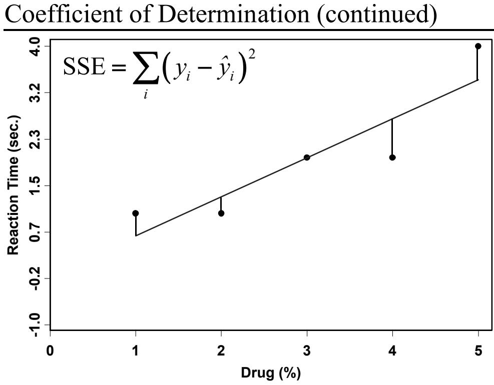
A measure of the contribution of x in predicting y

Assuming that x provides no information for the prediction of y, the best prediction for the value of y is $\bar{y}$



44

Coefficient of Determination (continued)



45

Coefficient of Determination (continued)

$SS_{yy} = \sum (y_i - \bar{y})^2$ --total sample variation around mean

$SSE = \sum (y_i - \hat{y}_i)^2$ --unexplained sample variability after fitting

$SS_{yy} - SSE$ --explained sample variability attributable to linear relationship

$\frac{SS_{yy} - SSE}{SS_{yy}} = \frac{\text{explained}}{\text{total}} =$ proportion of total sample variability explained by the linear relationship

46

Coefficient of Determination (continued)

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \underbrace{\frac{SSE}{SS_{yy}}}_{\text{Unexplained variability}}$$

In simple linear regression r^2 is computed as the square of the correlation coefficient, r

$0 \le r^2 \le 1$

<u>Interpretation</u>— $r^2 = .75$ means that the sum of squared deviations of the y values about their predicted values has been reduced by 75% by the use $\hat{y}$, instead of $\bar{y}$, to predict y of the least squares equation.

47