11 Correlation and Regression

When comparing two variables, *sometimes* one variable (the **explanatory** variable) can be used to help predict the value of another variable (the **response** variable).

Often we are interested in the association (i.e., a relationship) between two or more variables.

Example:

Example: For the following pairs of variables, which is the explanatory variable, and which is the response variable?

(a) number of years of education and income

(b) blood pressure (systolic) and weight

(c) height of sons and height of fathers

(d) score on midterm exam and score on final exam

(e) score on SAT and final GPA in college

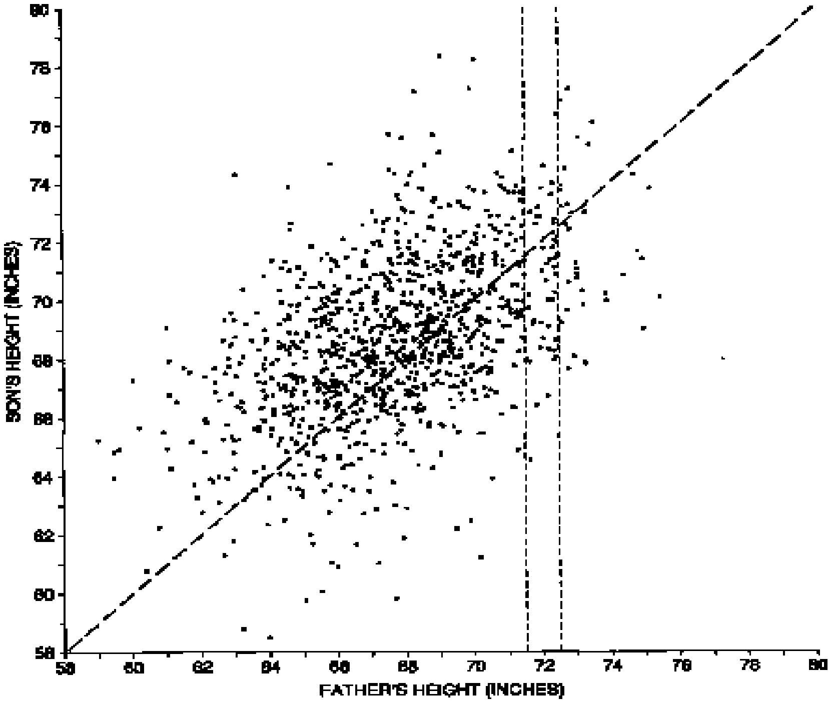
(f) deficit spending and interest rates

(g) temperature and ozone in atmosphere

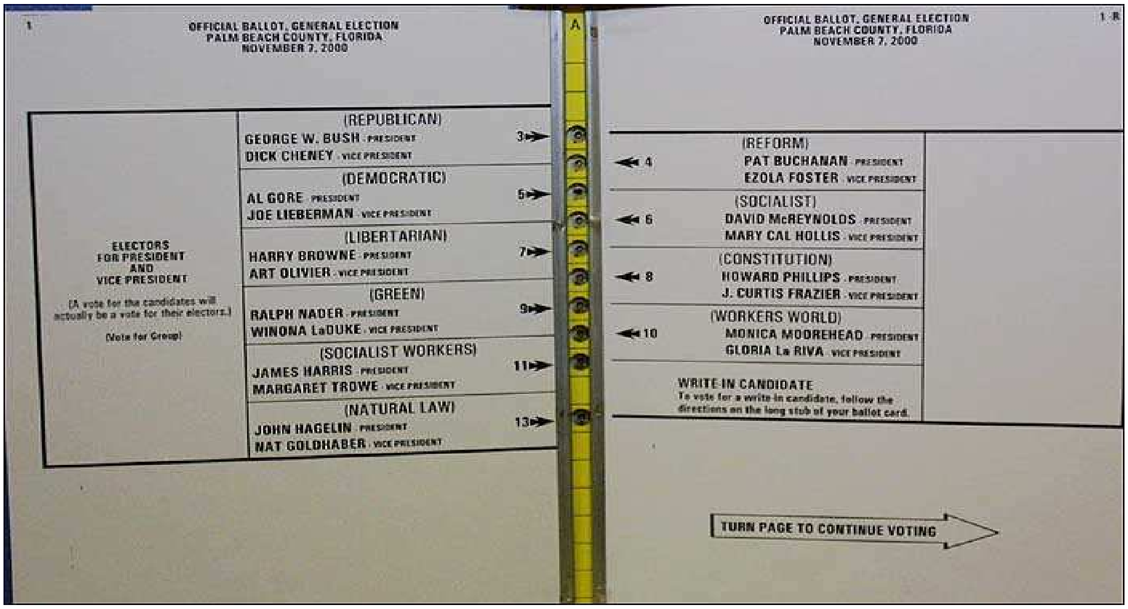
Exploring the Association between Two Quantitative Variables

A **scatterplot** graphically illustrates the relationship between two quantitative variables.

Example: Heights of 1078 fathers and sons, England, around year 1900.



Example: In the Presidential Election of 2000, George W. Bush earned 537 votes more than Al Gore in Florida, granting the Presidency to Bush. However, the Palm Beach County, Florida, the "butterfly ballot" possibly caused some individuals to mistakenly vote for Pat Buchanan rather than Gore.



1

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

ELECTORS
FOR PRESIDENT
AND
VICE PRESIDENT

(A vote for the candidates will actually be a vote for their electors.)

(Vote for Group)

(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT
3 →

(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT
5 →

(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT
7 →

(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT
9 →

(SOCIALIST WORKERS)
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT
11 →

(NATURAL LAW)
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT
13 →

A

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

1 R

← 4
(REFORM)
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT

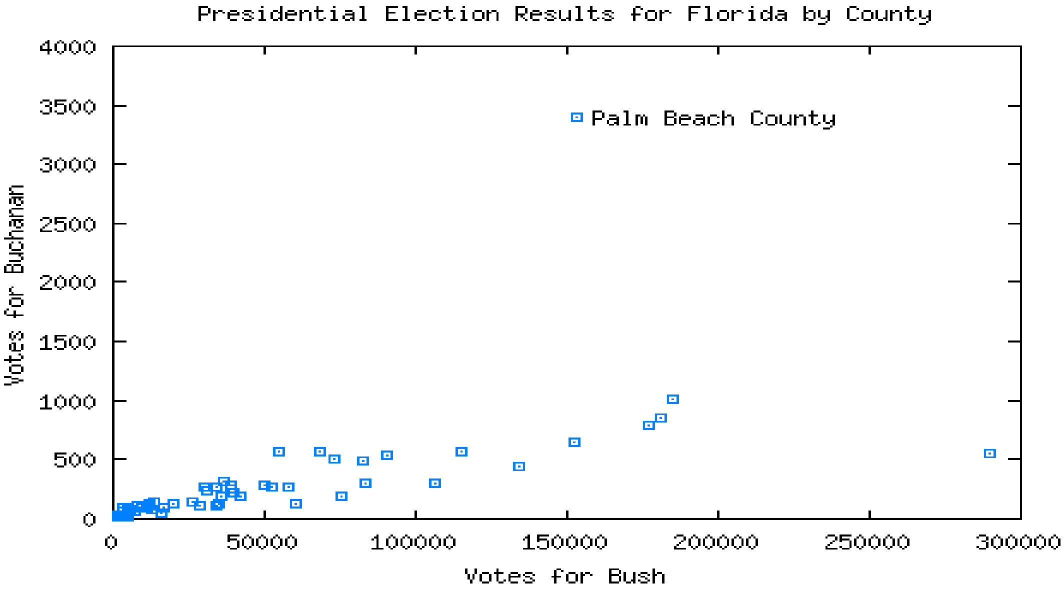
← 6
(SOCIALIST)
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT

← 8
(CONSTITUTION)
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT

← 10
(WORKERS WORLD)
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT

WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.

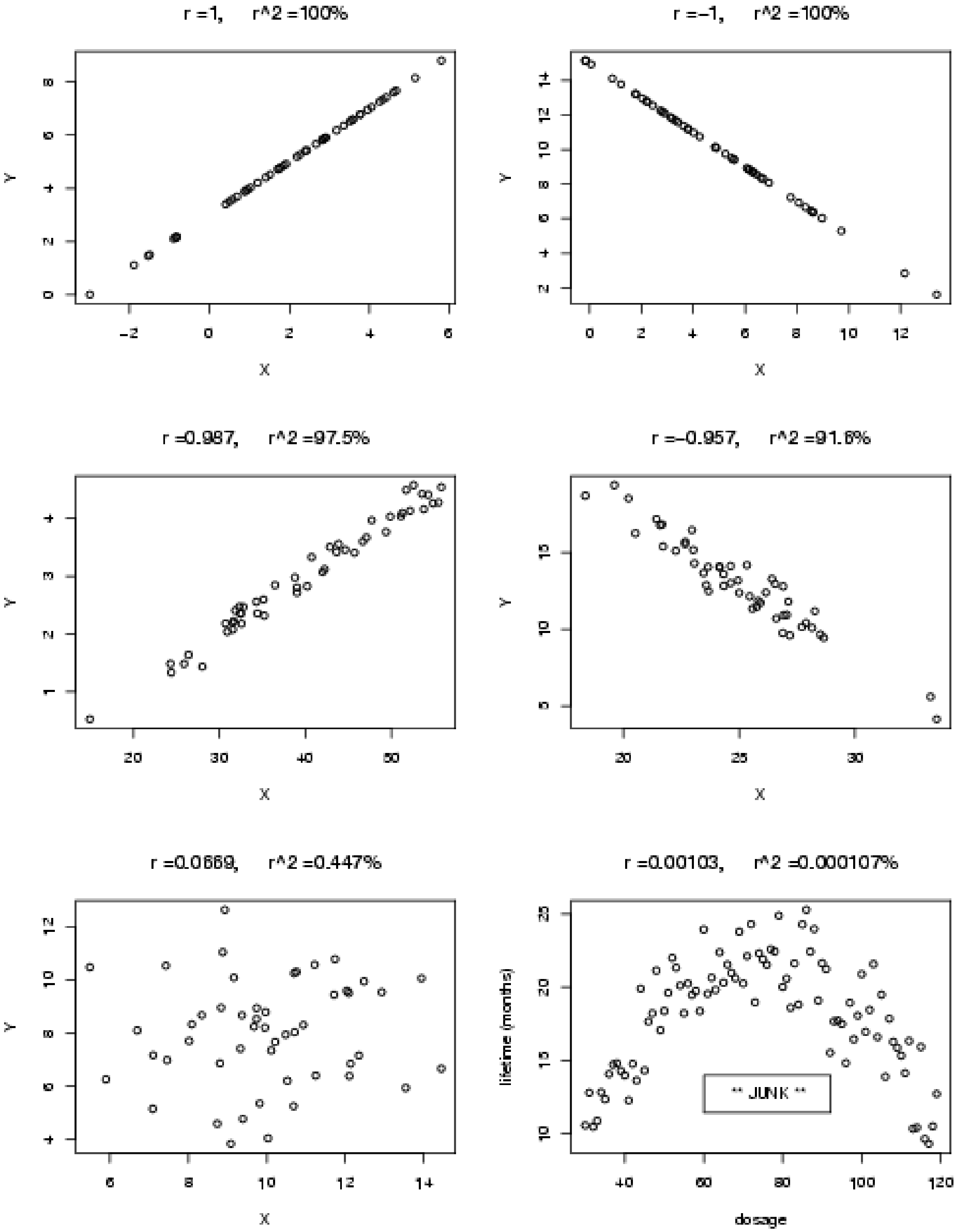
TURN PAGE TO CONTINUE VOTING

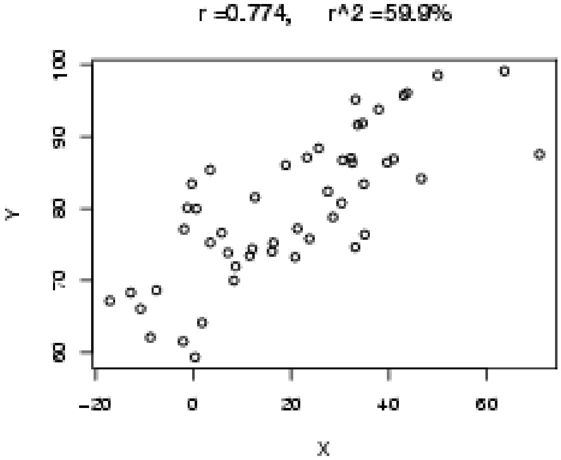


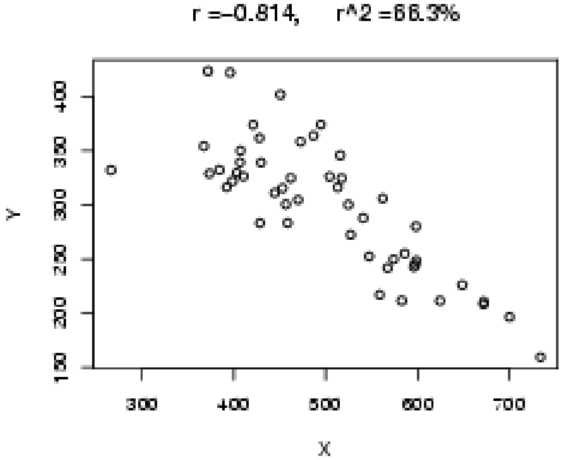
11.1 Correlation

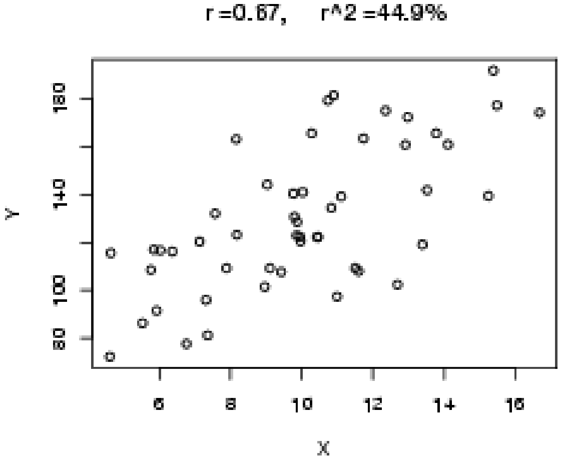
How Can Summarize Strength of Association?

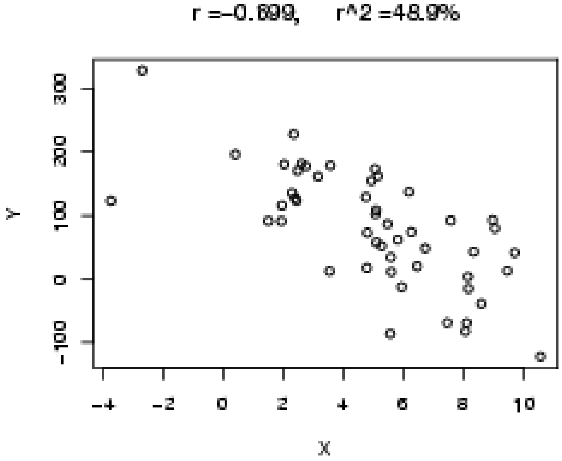
Correlation is a numerical measure of the **linear** association between two variables.

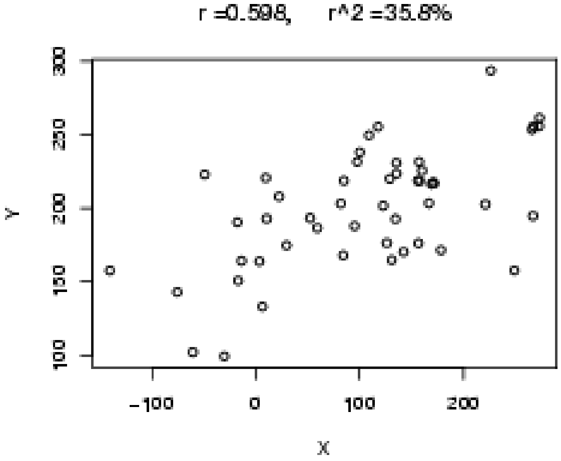

r =1,   r^2 =100%
Y
0
2
4
6
8
X
-2
0
2
4
6
r =-1,   r^2 =100%
Y
2
4
6
8
10
12
14
X
0
2
4
6
8
10
12
r =0.987,   r^2 =97.5%
Y
1
2
3
4
X
20
30
40
50
r =-0.957,   r^2 =91.6%
Y
5
10
15
X
20
25
30
r =0.0669,   r^2 =0.447%
Y
4
6
8
10
12
X
6
8
10
12
14
r =0.00103,   r^2 =0.000107%
lifetime (months)
10
15
20
25
dosage
40
60
80
100
120
** JUNK **

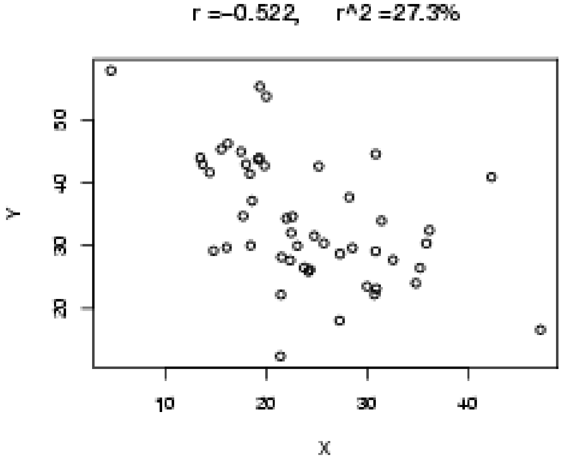

r =0.774, r^2 =59.9%
Y
60
70
80
90
100
-20
0
20
40
60
X


r =−0.814, r^2 =66.3%
Y
150
200
250
300
350
400
300
400
500
600
700
X

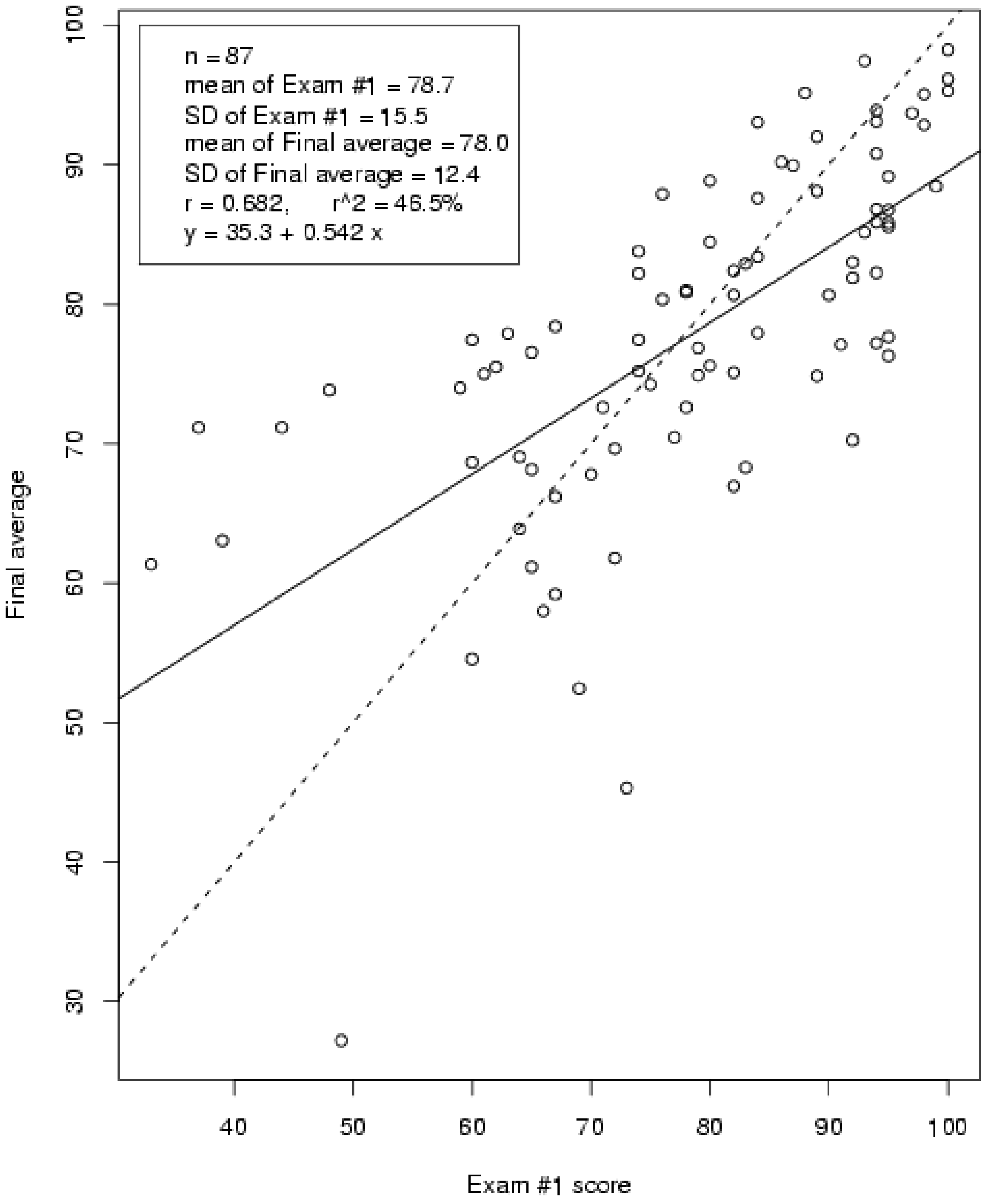

r =0.67, r^2 =44.9%
Y
80
100
140
180
6
8
10
12
14
16
X

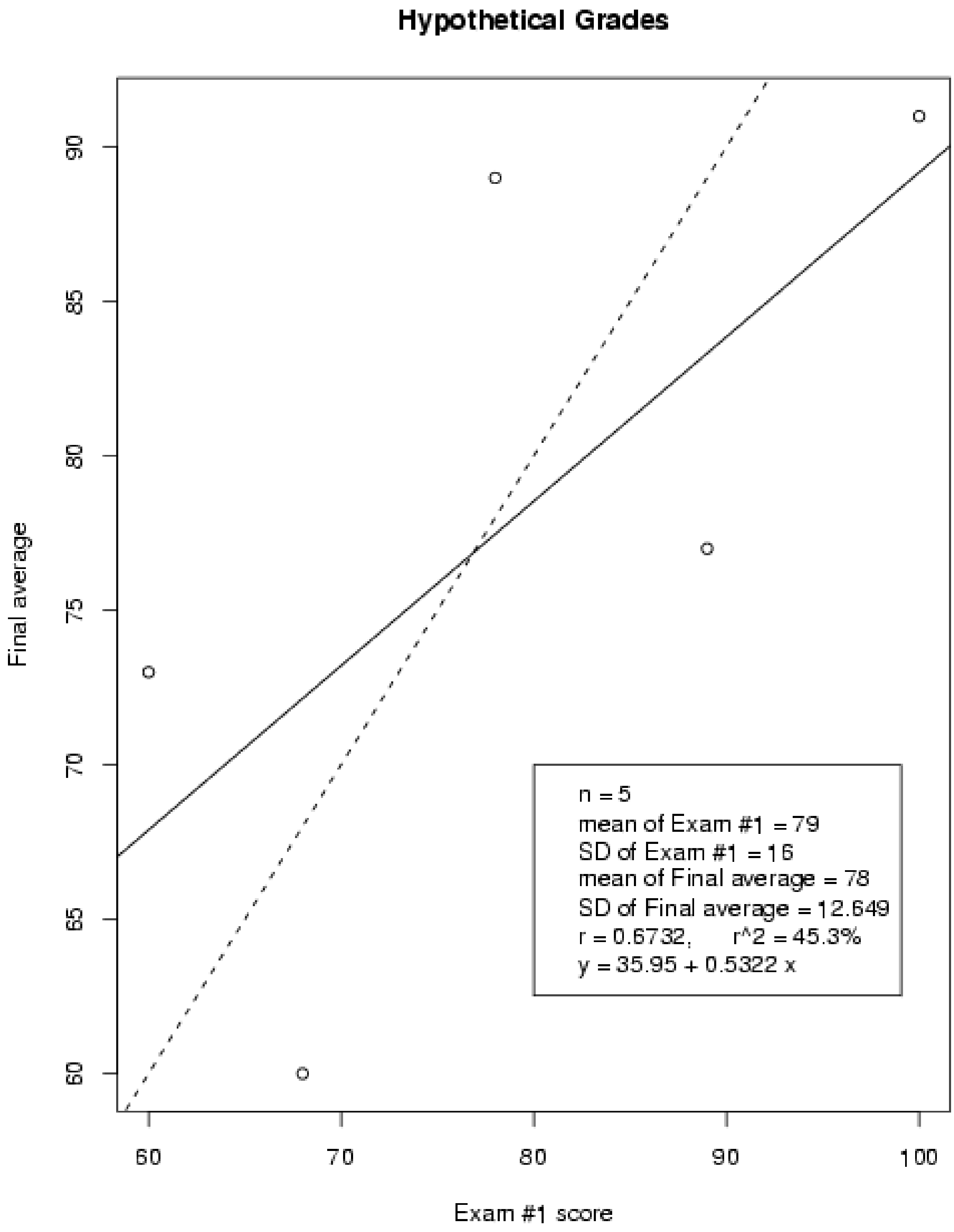

r =−0.699, r^2 =48.9%
Y
−100
0
100
200
300
−4
−2
0
2
4
6
8
10
X


r =0.598, r^2 =35.8%
Y
100
150
200
250
300
−100
0
100
200
X


r =−0.522, r^2 =27.3%
Y
20
30
40
50
10
20
30
40
X

Real Grades from Math 220




Hypothetical Grades
Final average
60
65
70
75
80
85
90
Exam #1 score
60
70
80
90
100
n = 5
mean of Exam #1 = 79
SD of Exam #1 = 16
mean of Final average = 78
SD of Final average = 12.649
r = 0.6732, r^2 = 45.3%
y = 35.95 + 0.5322 x

Calculation of **(Pearson's) correlation**, $\boldsymbol{r}$, for n pairs of data (x, y).

$$r = \frac{\frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y})}{s_x\ s_y}.$$

The textbook gives the formula

$$r = \frac{\sum z_x\ z_y}{n-1},$$

where $z_x = (x-\bar{x})/s_x$ and $z_y = (y-\bar{y})/s_y$.

Example: Determine the correlation for the following data:

Exam #1 score	Final score
x	y
68	60
100	91
89	77
78	89
60	73

	X	Y	$X-\bar{X}$	$(X-\bar{X})^2$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
	68	60	-11	121	-18	324	198
	100	91	21	441	13	169	273
	89	77	10	100	-1	1	-10
	78	89	-1	1	11	121	-11
	60	73	-19	361	-5	25	95
sum	395	390	0	1024	0	640	545

Remarks:

(a) Is r random or fixed?

(b) What are the units on r?

(c) What are the possible values of r?

(d) $r = 1$ implies what type of correlation?

(e) $r = -1$ implies what type of correlation?

(f) Is selection of x and y relevant when calculating r?

(g) r makes sense for linear associations only.

(h) A linear transformation on the data does not affect $|r|$.

(i) As the number of (x, y) data pairs becomes huge, r "converges" to the **population** correlation.

Assessing the Fit of a Line: r^2, the Coefficient of Determination

Definition: $\boldsymbol{r^2}$ is the fraction of the variation in the values of y that is explained by the least-squares regression on x.

Example: (FIVE PAIRS OF GRADES) Return to the data for the grades of the five hypothetical students of (exam #1 score, final score): (68, 60), (100, 91), (89, 77), (78, 89), and (60, 73). Determine the fraction of the variation in the values of y that is explained by the least-squares regression on x.

Example: Under the *lofty* assumption that the final score is based upon 5 equally weighted INDEPENDENT exams with a common variance, then r^2 (at least for the entire data set of 87 students) should be about what number (where r is the correlation between the first exam score and the final score)?

□

What are the possible values of r^2?

Cautions in Analyzing Association

Avoid extrapolation.

Association does not imply causation.

Example: Consider the two variables "weight of **older** brother at age 5" and "weight of **younger** brother at age 5."

A **confounding variable** or **confounder** is a third variable which confuses the relationship between the two variables of interest.

Example: Suppose in a large survey on alcohol consumption and lung cancer, it is determined that people who consume *a lot of alcohol* have a significantly *higher* rate of *lung cancer* than people who consume *little or no alcohol*.

Is it reasonable to conclude that heavy alcohol consumption **causes** lung cancer?

What might be a *confounding variable*?

□

11.2 The Least-Squares Regression Line

Examining the relationship between variables is called **regression analysis**.

Examining the **linear** relationship between **two** variables is called **simple linear regression**.

Two purposes of regression analysis:

1. explain

2. predict

Typically,

x is the **explanatory** variable.

y is the **response** variable.

Goal is to fit a reasonable line through the scatter plot.

The unique line which minimizes the sum of squares of the vertical distances is called the **least squares line** or **fitted regression line**.

The equation of the **least squares** line can be written

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

The **slope** of the least squares line can be shown to be

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

The **intercept** of the least squares line may be computed by noting that the least squares line goes through the point $(\bar{x}, \bar{y})$.

Example: (FIVE PAIRS OF GRADES) Return to the data for the grades of the five hypothetical students of (exam #1 score, final score): (68, 60), (100, 91), (89, 77), (78, 89), and (60, 73). Fit the regression line.

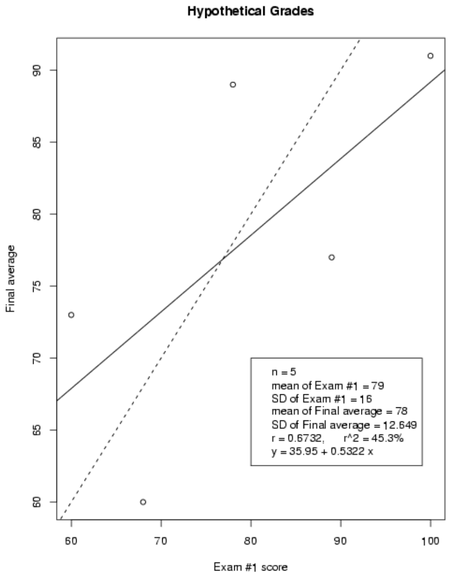
Predict a new value of y if x is 85.

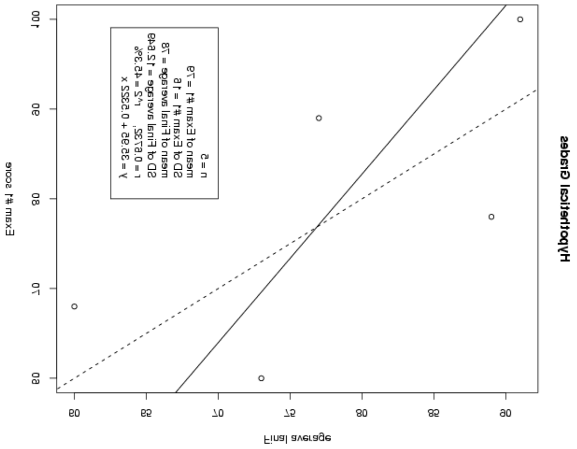
Estimate the mean of y if x is 85.

Predict a new value of y if x is 40.

$\square$

Remark: The least squares line of y on x differs from the least squares line of x on y.





11.3 Inference on the Slope of the Regression Line

The Linear Model

For the entire population of (x, y) values, we define the (population) linear model as:

$$\mu_{y|x} = \beta_0 + \beta_1\ x,$$

where the population intercept β_0 and the population slope β_1 are unknown.

The sample slope b_0 and the sample intercept b_1 estimate β_0 and β_1.

What is an interesting hypothesis test?

Recall: $b_1 = r\ s_y/x_x$

Hypothesis test can be performed and confidence intervals can be constructed on β_1 (or ρ) when all of the following hold.

(1) Observations (x, y) are sampled independently.

(2) The data follow a linear pattern.

(3) The y-values within a vertical strip are approximately normally distributed with the same standard deviation for each x, or the sample size is large $(n > 30)$.

Assumptions #2 and #3 may be checked using a **residual plot** (see p. 506), which plots the points $(x,\ y - \hat{y})$, where the values of $y - \hat{y}$ are called the **residuals.**

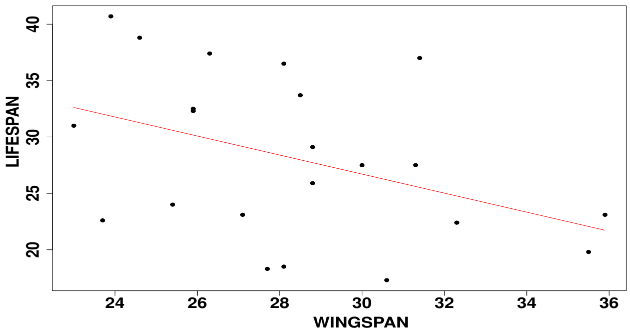
If the **residual plot** shows no obvious pattern, contains no outliers, and has roughly the same vertical spread for each x value, then assumptions #2 and #3 are valid.

Compute the test statistic $T = (b_1 - \beta_1)/s_b$, which is t-distributed with $(n - 2)$ degrees of freedom, when performing hypothesis tests on β_1 (or ρ).

The confidence interval on β_1 is $b_1 \pm t_{n-2}\ s_b$.

The formula for s_{b_1} is given in the textbook but is ugly, so we will use statistical software to calculate s_{b_1}.

Example: Is there a relationship between the size and the lifespan of butterflies? Test at level $\alpha = 0.05$. The wingspan (in millimeters) and the lifespan in the adult state (in days) were measured for 22 species of butterfly. The summary statistics are $\bar{x} = 28.309$ mm, $s_x = 3.526$ mm, $\bar{y} = 28.136$ days, $s_y = 7.225$ days, $r = -0.412$, and $s_{b_1} = 0.4175$ days / mm.



(a) State the null and alternative hypotheses.

(b) Discuss the assumptions needed to perform this hypothesis test.

(c) Determine the value of the estimated slope.

(d) Determine the value of the standardized test statistic.

(e) Determine the number of degrees of freedom.

(f) Determine the P-value.

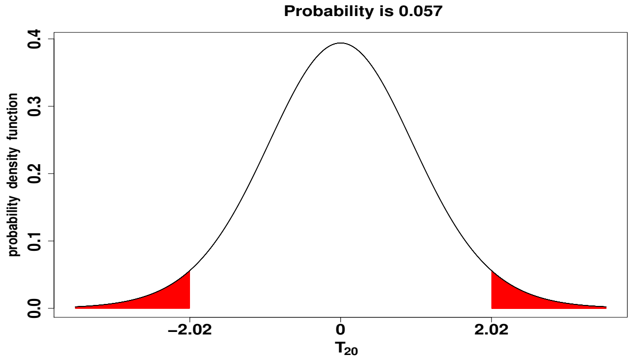
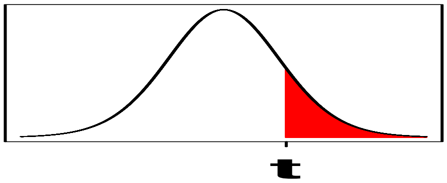


Table A.3 Critical Values for the Student's t Distribution, p. A-8



Degrees of Freedom	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	Area in Right Tail									
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(g) State the conclusion in statistical terms.

(h) State the conclusion in regular English.

We fail to conclude that the slope parameter (change in lifetime per change in wingspan) is nonzero.

We fail to conclude that the correlation between wingspan and lifetime is statistically significant.

(i) Construct a 95% confidence interval on β_1.

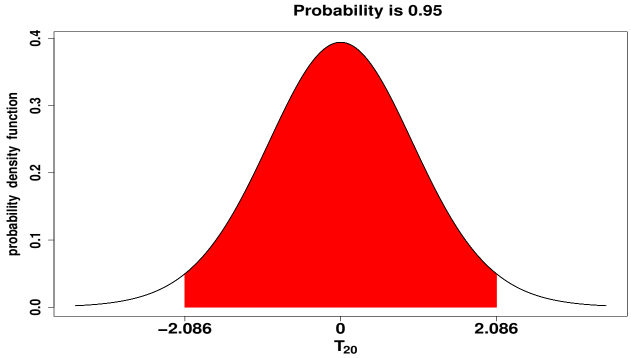
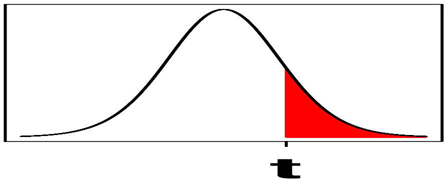


Table A.3 Critical Values for the Student's t Distribution, p. A-8



Degrees of	Area in Right Tail									
Freedom	**0.40**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	0.254	0.677	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
200	0.254	0.676	1.286	1.653	1.972	2.345	2.601	2.839	3.131	3.340
z	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291
	20%	50%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
	Confidence Level									

(j) Discuss the relevance of the relationship between your confidence interval and the number zero.

□