

## 3 Association: Contingency, 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

## 3.1 How Can We Explore the Association between Two Categorical Variables?

Set up a **contingency table**, for comparing two categorical variables.

Within a contingency table, we can determine

**conditional proportions**; i.e., the proportion of the time that a variable takes on a particular value, conditional on some value of the other variable.

**Example:** Consider the following contingency table regarding the gender of an unborn baby (The data are hypothetical but are somewhat consistent with ultrasounds from the 1990's.):

---

	Ultrasound Predicted Female	Ultrasound Predicted Male
Actual gender is female	432	48
Actual gender is male	130	390

---

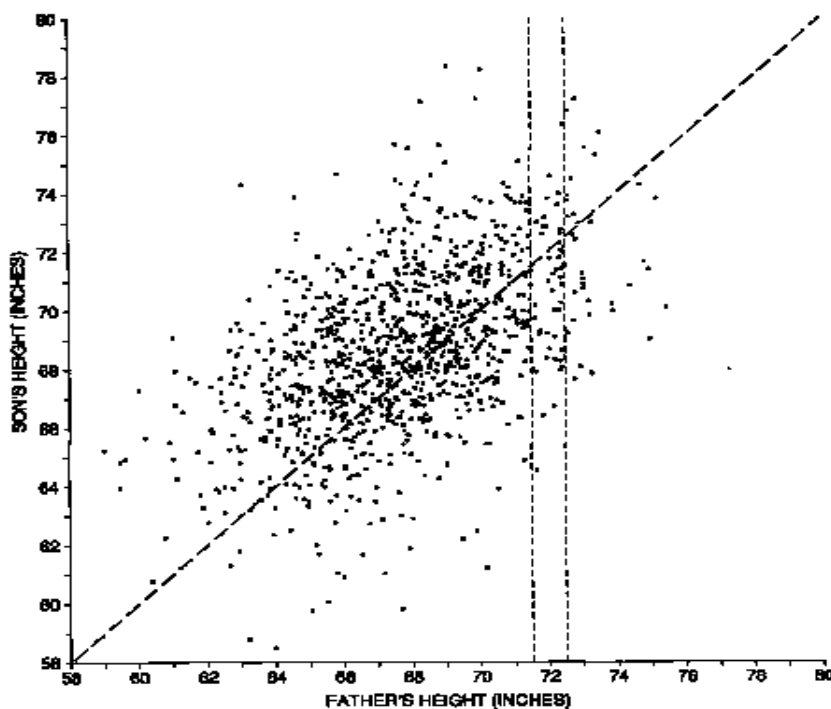
Based on the above contingency table:

- (a) What proportion of babies are *female*, given that the ultrasound predicted that the baby would be *female*?
- (b) What proportion of babies are *male*, given that the ultrasound predicted that the baby would be *male*?
- (c) Is the ultrasound equally reliable for predicting gender for boys and for girls?

## 3.2 How Can We Explore the Association between Two Quantitative Variables?

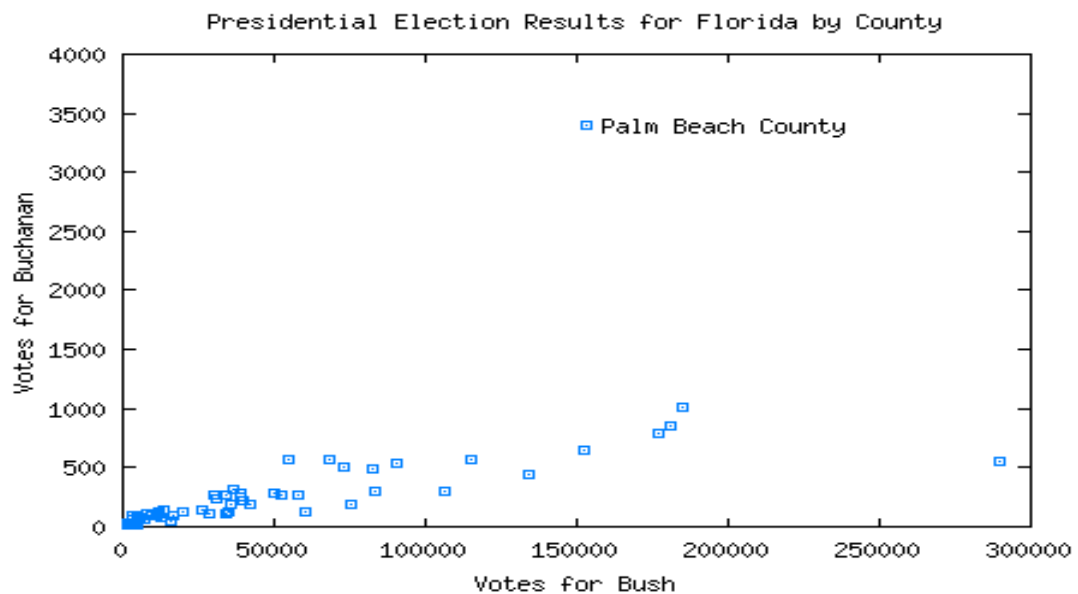
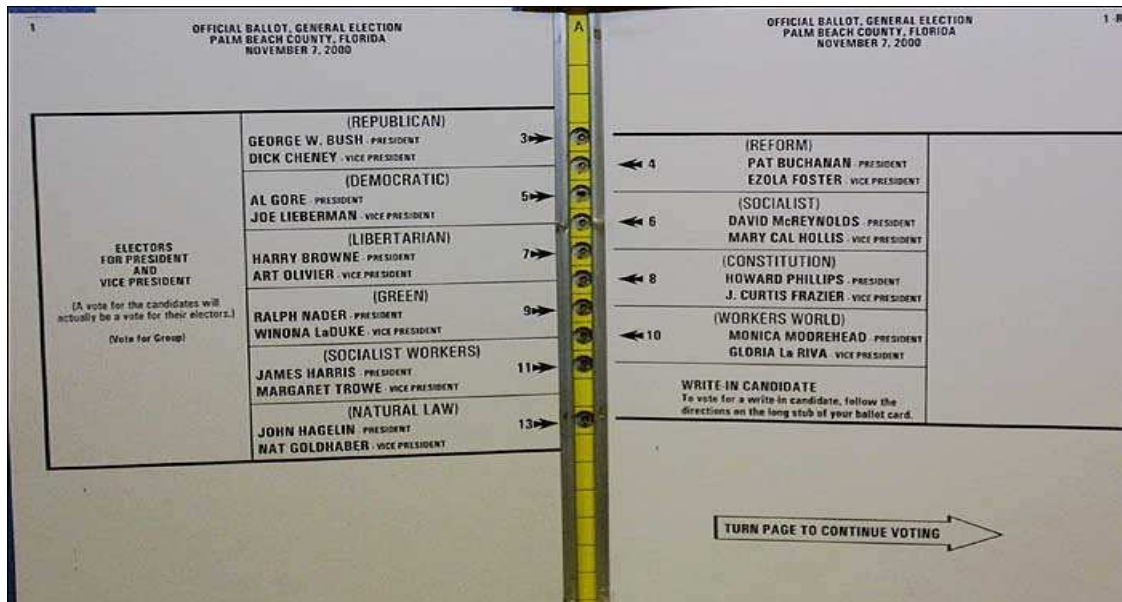
A **scatterplot** graphically illustrates the relationship between two quantitative variables.

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



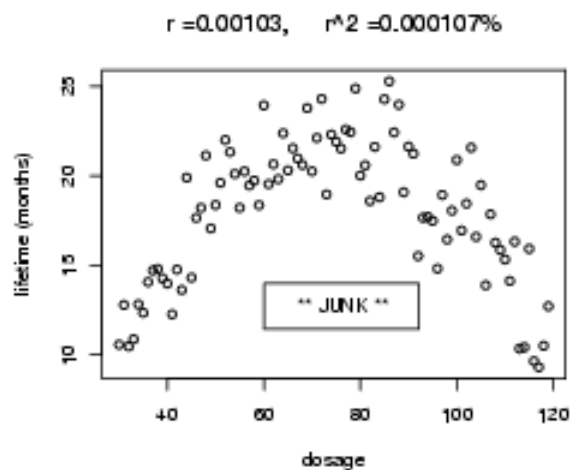
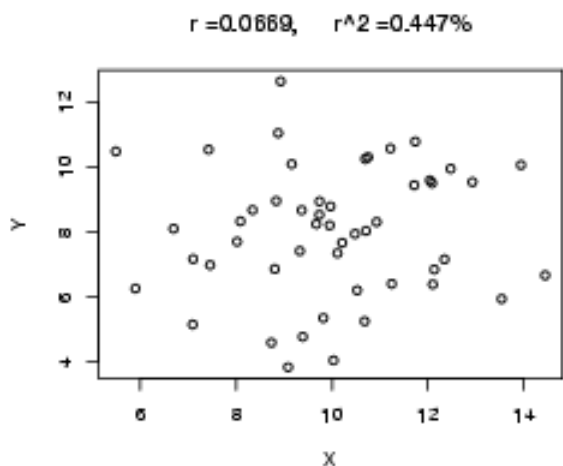
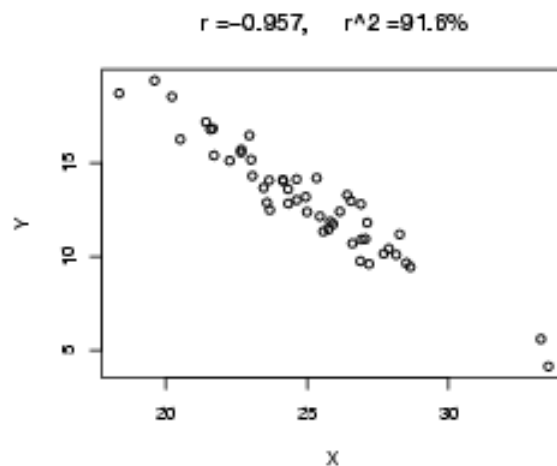
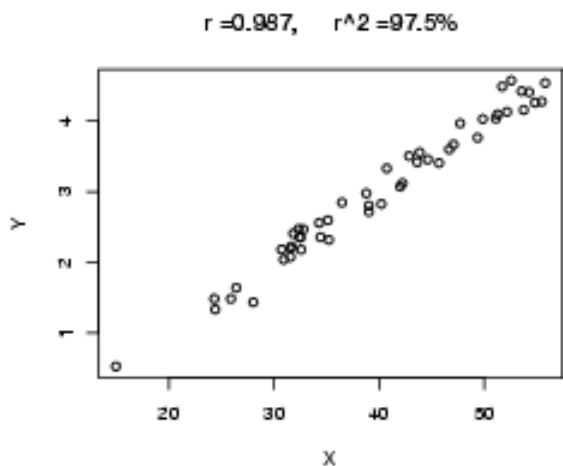
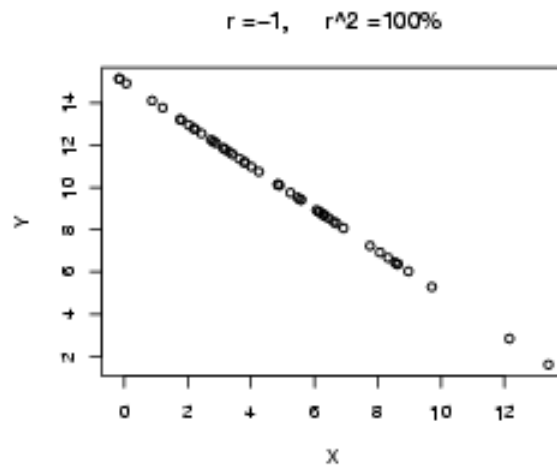
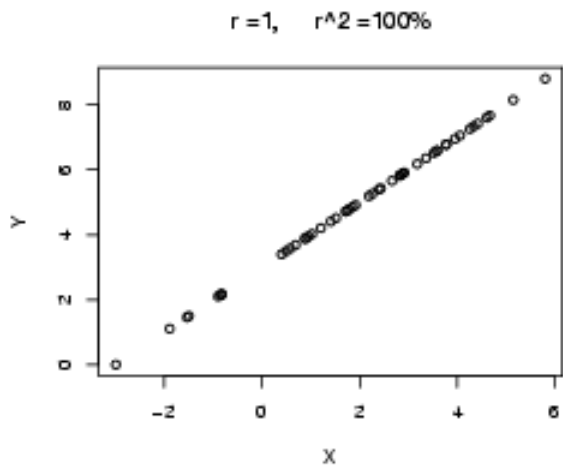
**Example:** In the Presidential Election of 2000, George W. Bush earned 537 votes more than Al Gore

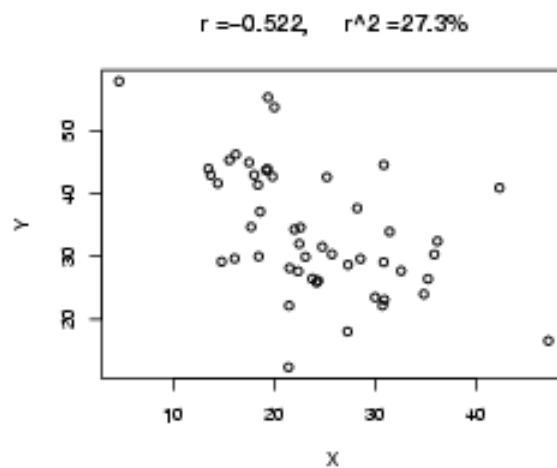
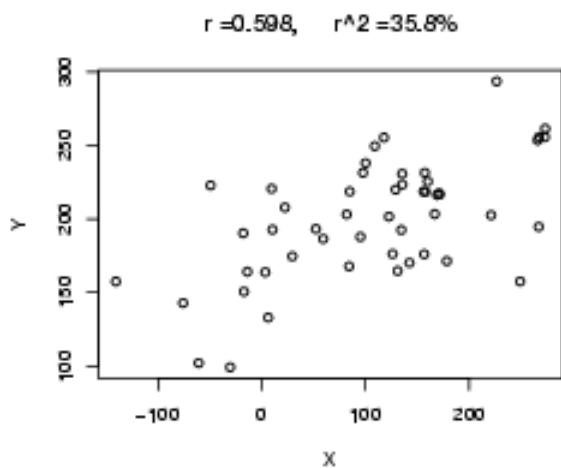
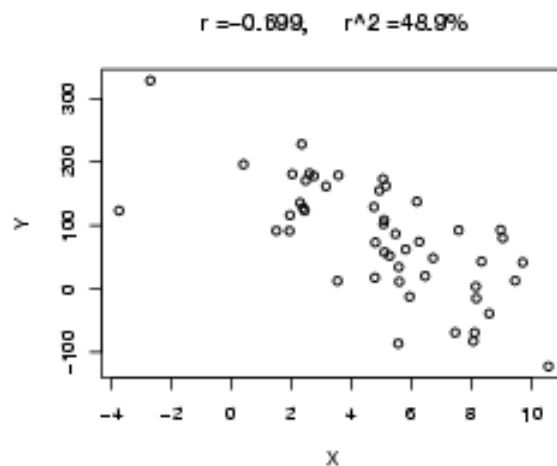
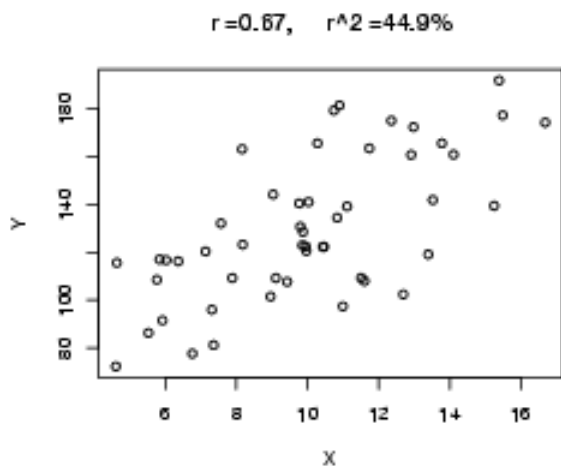
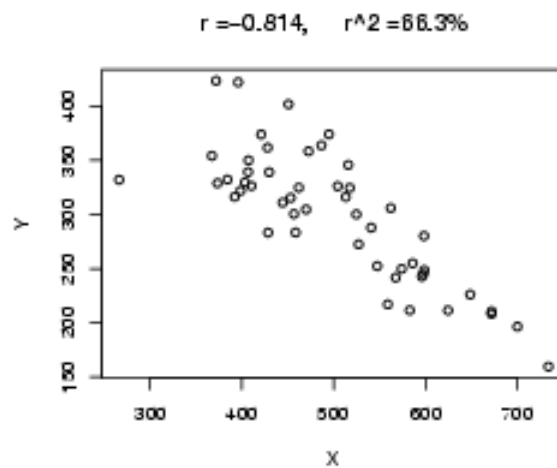
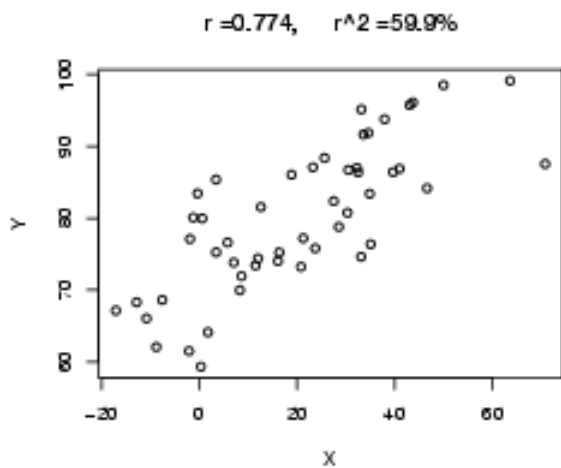
in Florida, granting the Presidency to Bush. However, the Palm Beach County, Florida, the “butterfly ballot” possibly caused some individuals to mistakenly vote for Pat Buchanan rather than Gore.



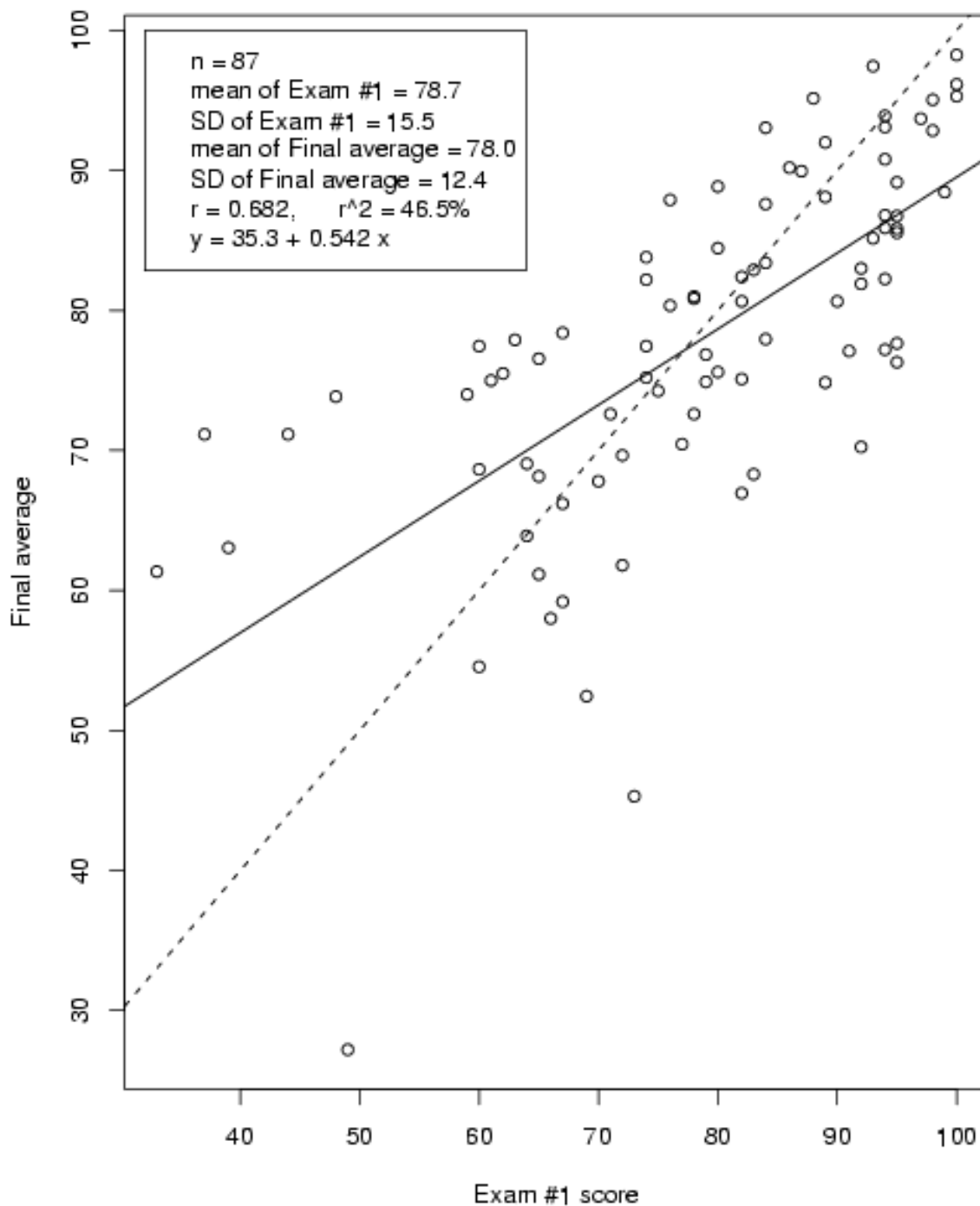
## How Can Summarize Strength of Association?

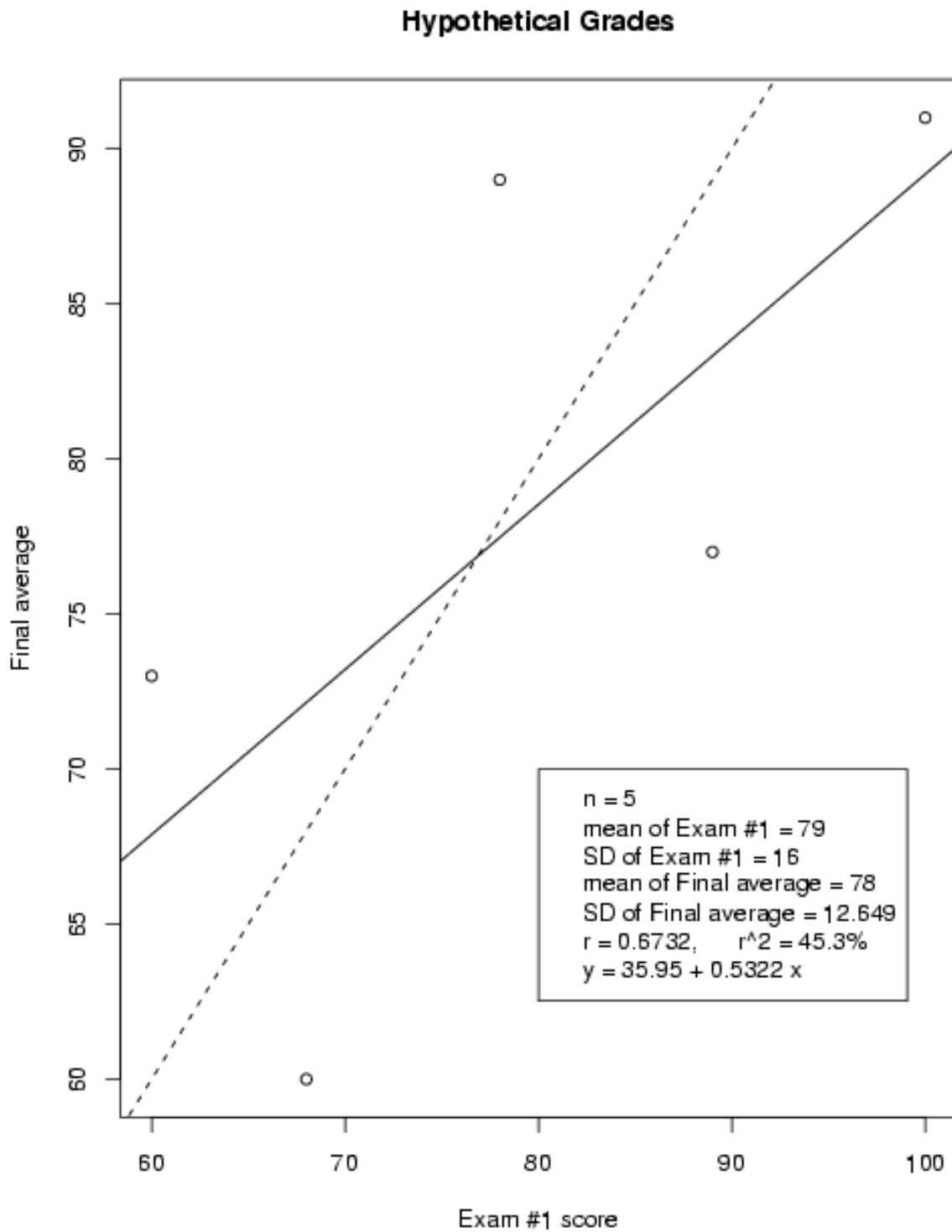
**Correlation** is a numerical measure of the  
**linear** association between two variables.





Grades for Math 220, Fall 2003





Calculation of (Pearson's) correlation,  $r$ , for  $n$

pairs of data  $(x, y)$ .

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

The textbook gives the formula

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

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

**Example:** Determine the correlation for the following data:

---

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

---

**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,  $r$  “gets close” to the **population** correlation.

### 3.3 How Can We Predict the Outcome of a Variable?

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

$$\hat{y} = a + b x.$$

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

be

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

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.

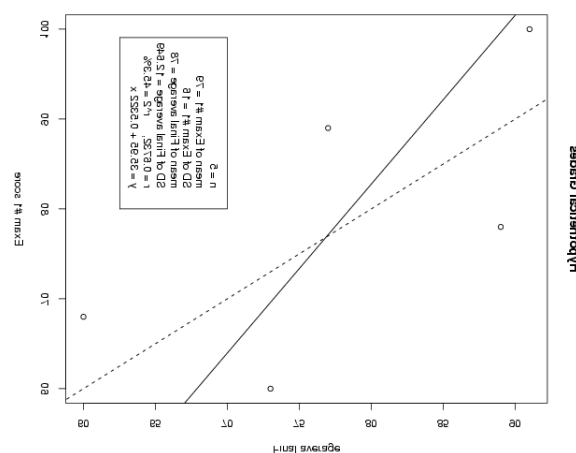
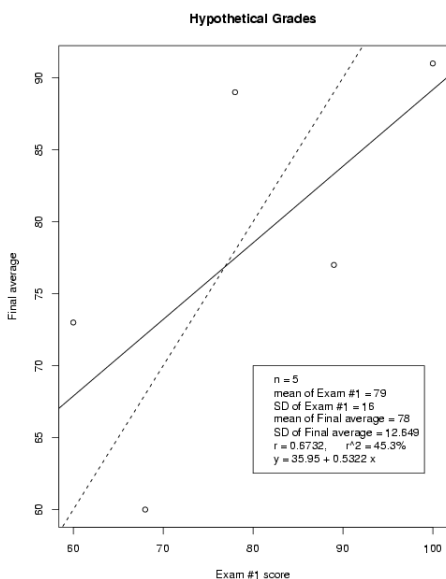
Predict a new value of  $y$  if  $x$  is 85.

Estimate the mean of  $y$  if  $x$  is 85.

Predict a new value of  $y$  if  $x$  is 40.



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



## Assessing the Fit of a Line: r-Squared

**Definition:** The **proportional reduction in error**, denoted by  $r^2$ , gives the proportion of variation in  $y$  which can be explained by  $x$ , when the data are linear.

**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 *proportional reduction in error*.

**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?

□

What are the possible values of  $r^2$ ?

### 3.4 What are Some Cautions in Analyzing Association?

“Extrapolation is dangerous.”

“Correlation 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 **lurking variable** is a third variable which confuses the relationship between the two variables of interest.

In general, “association does not imply causation.”

**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 *lurking variable*?

□