Lecture 13/Chapter 10
Relationships between Measurement (Quantitative) Variables

- Scatterplot; Roles of Variables
- 3 Features of Relationship
- Correlation
- Regression

Definition

- **Scatterplot** displays relationship between 2 quantitative variables:
  - Explanatory variable (x) on horizontal axis
  - Response variable (y) on vertical axis

Example: *Explanatory/Response Roles*

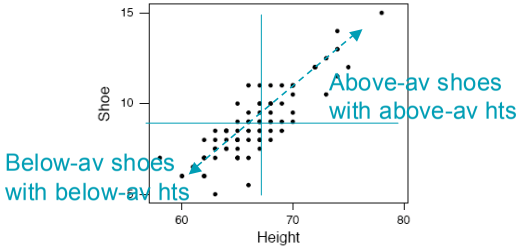
- **Background**: We're interested in the relationship between students' shoe sizes and heights; also, relationship between HW1 score and Exam 1 score.
- **Question:** Which variable should be graphed along the horizontal axis of each scatterplot?
- **Response:**
  - Shoe sizes and heights:
  - HW1 and Exam 1:

Definitions

- **Form:** relationship is linear if scatterplot points cluster around some straight line
- **Direction:** relationship is
  - positive if points slope upward left to right
  - negative if points slope downward left to right
- **Strength** (assuming linear):
  - strong: points tightly clustered around a line (explanatory var. tells us a lot about response)
  - weak: points loosely scattered around a line (explanatory var. tells us little about response)

Example: *Form and Direction*

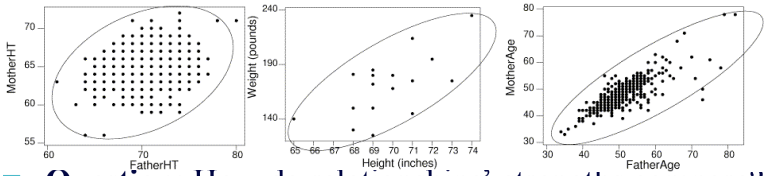
- **Background**: Scatterplot displays relationship between students' heights and shoesizes.



- **Question:** What are the form and direction of the relationship?
- **Response:**

Example: *Relative Strengths*

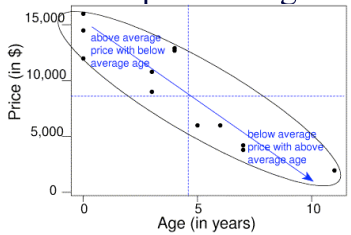
- **Background**: Scatterplots display:
  - mothers' ht. vs. fathers' ht. (left)
  - males' wt. vs. ht. (middle)
  - mothers' age vs. fathers' age (right):



- **Question:** How do relationships' strengths compare? (Which is strongest, which is weakest?)
- **Response:** _____ strongest, _____ weakest

Example: *Negative Relationship*

- **Background**: Plot of price vs. age for 14 used Grand Am's.



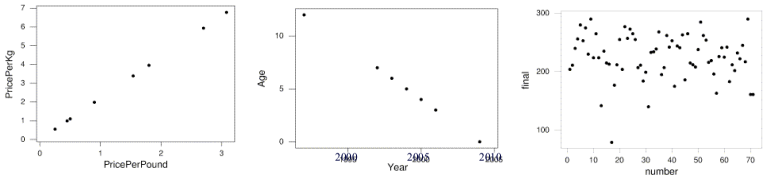
- **Questions:**
  - Why should we expect the relationship to be negative?
  - Does it appear linear? Is it weak or strong?
- **Responses:**
  - _____
  - _____

Definition

- **Correlation *r*:** tells direction and strength of linear relation between 2 quantitative variables
  - **Direction:** *r* is
    - positive for positive relationship
    - negative for negative relationship
    - zero for no relationship
  - **Strength:** *r* is between -1 and +1; it is
    - close to 1 in absolute value for strong relationship
    - close to 0 in absolute value for weak relationship
    - close to 0.5 in absolute value for moderate relationship

Example: *Extreme Values of Correlation*

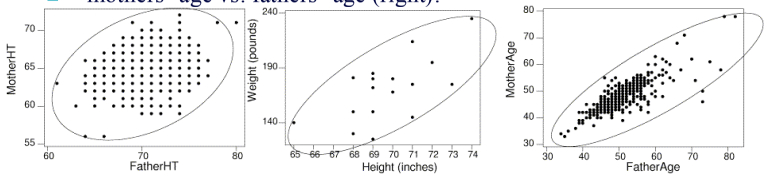
- **Background**: Scatterplots show relationships…
  - Price per kilogram vs. price per pound for groceries
  - Used cars' age vs. year made
  - Students' final exam score vs. (number) order handed in



- **Question:** Which has $r = -1$? $r = 0$? $r = +1$?
- **Response:** left has $r =$___, middle has $r =$___, right has $r =$__

Example: *Other Values of r*

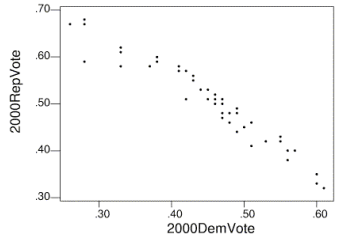
- **Background**: Scatterplots display:
  - mothers' ht. vs. fathers' ht. (left)
  - males' wt. vs. ht. (middle)
  - mothers' age vs. fathers' age (right):



- **Question:** Which graphs go with which correlation: $r = 0.78$, $r = 0.65$, $r = 0.23$?
- **Response:**
  left has r=____, middle has r=____, right has r=____

Example: *Imperfect Relationships*

- **Background**: For 50 states, % voting Republican vs. % Democrat in 2000 presidential election had r =-0.96.



- **Questions:** Why is the relationship negative? Why imperfect?
- **Responses:**
  - _____: more voting Democratic→_____ Republican
  - Imperfect: _____

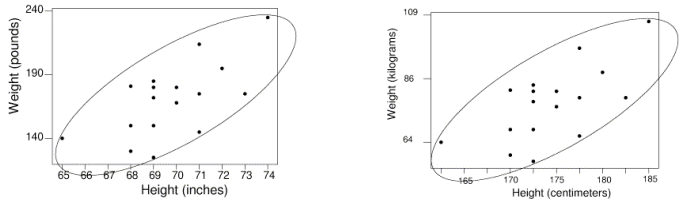
More about Correlation *r*

- Correlation is a standardized measure of the direction and strength of the linear relation between 2 quantitative variables
  - A strong curved relationship may have r close to 0
  - r is unaffected by change of units
  - r based on averages overstates strength (next time)

Example: *Correlation when Units are Changed*

- **Background**: For 17 male students plotted…

Left: wt (lbs) vs. ht (in) or **Right:** wt (kg) vs. ht (cm)



- **Question:** How do directions, strengths, and correlations compare, left vs. right?
- **Response:**

Least Squares Regression Line

If form appears linear, then we picture points clustered around a straight line.

- **Questions** (Rhetorical): Is there only one "best" line? If so, how can we find it? If found, how can we use it?
- **Responses:**

If found, we'd use the line to make predictions. Use calculus to find the line that makes the best predictions. There is a unique best line.

Least Squares Regression Line

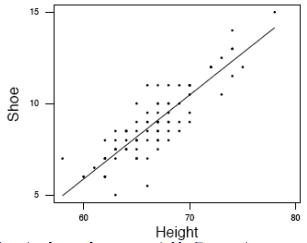
Summarize linear relationship between explanatory (x) and response (y) values with line $y = a+bx$ that minimizes sum of squared prediction errors (called *residuals*).

- **Slope *b*:** predicted change in response y for every unit increase in explanatory value x
- **Intercept *a*:** where best-fitting line crosses y-axis (predicted response for x=0?)

Note: In Algebra, we use $y=mx+b$ as equation of a line.

Example: *Least Squares Regression Line*

- **Background**: We regress shoe size on height and get y=-22+0.5x [shoe=-22+0.5height]



- **Question:**What do slope=+0.5, intercept=-22 tell us?
- **Response:**
  - For each additional inch in ht, predict shoe_____
  - The "best" line crosses the y axis at y=_____

Definition

- **Extrapolation:** using the regression line to predict responses for explanatory values outside the range of those used to construct the line.

Example: *Extrapolation*

- **Background**: A regression of 17 male students' weights (lbs.) on heights (inches) yields the equation $y = -438+8.7x$
- **Question:** What weight does the line predict for a 20-inch-long infant?
- **Response:**

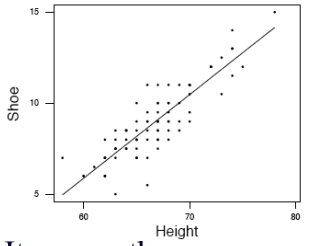
More about intercept and slope

Consider slope and intercept of the least squares regression line $y=a+bx$

- **Slope:** $b = r \frac{\text{standard deviation in y}}{\text{standard deviation in x}}$ so if x increases by a standard deviation, predict y to increase by r standard deviations
  - **$|r|$ close to 1:** y responds closely to x
  - **$|r|$ close to 0:** y hardly responds to x
- **Intercept:** a=average y-b(average x), so the line passes through the point of averages.

Example: *Summaries, Intercept, Slope*

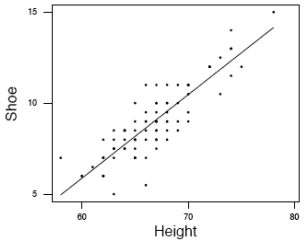
- **Background**: means and sds are 67in and 4in for hts, 9 and 2 for shoe sizes; $r = +0.9$.
- **Question:**How does the regression line relate to these?



- **Response:** It passes thru _____. If ht is 4 in. more, predict shoesize up by _____.

Example: *Predicting from Regression Line*

- **Background**: The regression equation is shoe=-21.7+0.46 height



- **Question/Response:**What are the following?
  - Predicted shoe for ht=65?
  - Predicted shoe for ht=70?
  - Predicted shoe for ht=67?
  - Predicted shoe for ht=78?

Example: *Predicting from Regression Line*

- **Background**: The regression equation is Exam1=100 +1.24 HW1 ["standard error"=11]
- **Question:** Predict your own Exam score. Is it close?
- **Response:**

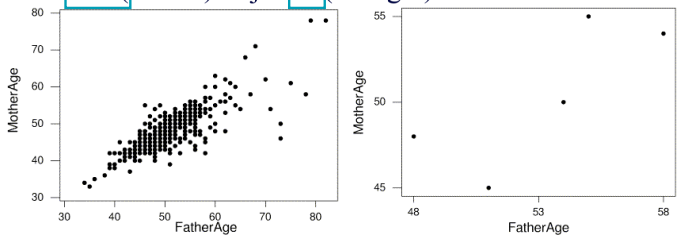
Definition

- **Statistically significant relationship:** one that cannot easily be attributed to chance. (If there were actually no relationship in the population, the chance of seeing such a relationship in a random sample would be less than 5%.)

(We'll learn to assess statistical significance in Chapters 13, 22, 23.)

Example: *Sample Size, Statistical Significance*

- **Background**: Relationship between ages of students' mothers and fathers both have r=+0.78, but sample size is over 400 (on left) or just 5 (on right):



- **Question:** Which plot shows a relationship that appears to be statistically significant?
- **Response:** The one on the_____. (Relationship on _____ could be due to chance.)

CHILDREN BEWARE: WATCHING TV CAN MAKE YOUR FAT

While the debate over TV's effects on children focuses on what they watch, a new study of some 4,000 children underscores the importance of how much they watch, showing that the more time children spend in front of the tube, the fatter they tend to be. Moreover, the study firmly documents for the first time that black and Latino youths watch more TV than do whites, putting them at greater risk of obesity. Spending more than four hours a day in front of the TV were 43% of black children, 30% of Mexican Americans, and 20% of non-Latino whites. One reason for the ethnic and racial differences in viewing trends, researchers speculate, is that parents in urban neighborhoods may discourage their children from playing outside because of crime. Thus the fear of crime appears to contribute to the "epidemic of obesity," researchers say. Though it may seem obvious that watching TV and shirking exercise is behind the childhood obesity epidemic, researchers have had surprising difficulty nailing down these factors…

(cont'd) The study's results, made public in the Journal of the American Medical Assn, "are consistent, make sense, and indicate a serious problem in the U.S.," said Steven Gortmaker, a sociologist at Harvard who has studied TV viewing and obesity.
In the most comprehensive study of its kind, the researchers analyzed data from lifestyle interviews with 4,063 children between 1988 and 1994. Consistent with previous surveys, the study found high rates of TV viewing overall: 67% watched at least 2 hours a day, and 26% racked up 4 or more hours.
The central finding was that children who watched a lot of TV were measurably fatter than those who watched relatively little. For instance, children who watched at least 4 hrs daily had about 20% more body fat than children who watched fewer than 2 hrs..
Some researchers say the new study cannot definitively claim that watching TV caused the children's weight problems. It may be that overweight children just watch more TV than other children, as Dr. Thomas Robinson of Stanford pointed out in an editorial of the AMA journal.

BEST DIET FOR CHILDREN MAY MEAN NO TV Diet and exercise programs usually fail to prevent obesity in kids, but a simpler approach may yield better results: turning off the television, even if for an hour or less a day.The idea already has the support of Surgeon General David Satcher, who last month urged parents nationwide to limit their children's TV viewing as part of an effort to improve public health.In a study presented in San Francisco on Tuesday, researchers at Stanford University helped persuade nearly 100 3rd and 4th graders in San Jose to cut their TV viewing by one-quarter to one-third. At the end of the school year, the youngsters had gained about two pounds less on average than a matched set of their peers who kept up their normal habits.
The results, presented to an audience at the Pediatric Academic Societies' annual meeting, dovetail with other studies showing that adults and children who watch more TV tend to be heavier than those who watch less.

(continued) Obesity nationwide is on the rise, with 33 percent of U.S adults considered overweight. Twenty years ago, 10 percent of children were considered overweight; now the figure is 20 percent, according to Len Epstein of the State University of New York at Buffalo, who studies pediatric obesity.
"It's really kind of scary the rate at which we're seeing increasing obesity in kids and adults, in this country and around the world," said Thomas N. Robinson, assistant professor of pediatrics and medicine at Stanford University. Robinson conducted the study.
At the same time, TV-watching in the United States remains high, with the average child younger than 11 watching nearly 20 hours per week.
Obesity is one of the hardest conditions to prevent and treat. Programs aimed at keeping people from smoking, drinking and taking drugs have far higher success rates than those aimed at keeping people fit and slim, Robinson said.