Correlation & Regression, II

9.07

4/6/2004

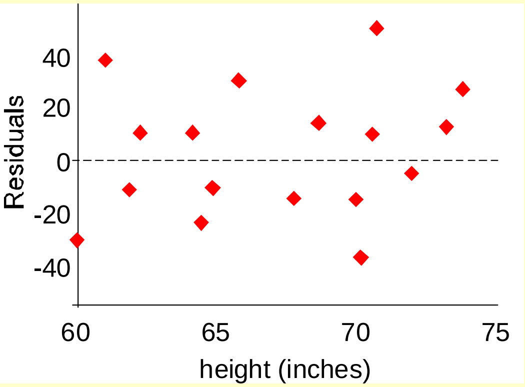
Steps in regression analysis (so far)

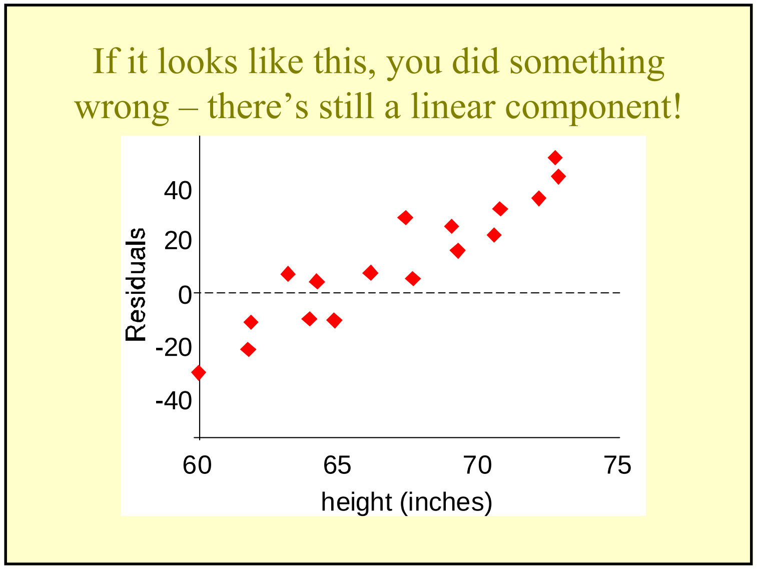
- Plot a scatter plot
- Find the parameters of the best fit regression line, y'=a+bx
- Plot the regression line on the scatter plot
- Plot the residuals, x_i vs. $(y_i – y_i')$, as a scatter plot, for diagnostic purposes

Residual Plots

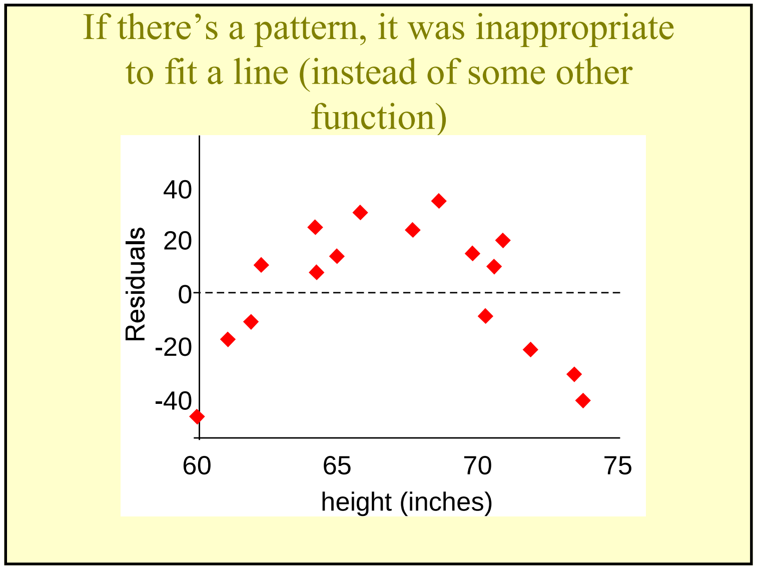
- Plotting the residuals $(y_i – y_i')$ against x_i can reveal how well the linear equation explains the data
- Can suggest that the relationship is significantly non-linear, or other oddities
- The best structure to see is no structure at all

What we like to see: no pattern

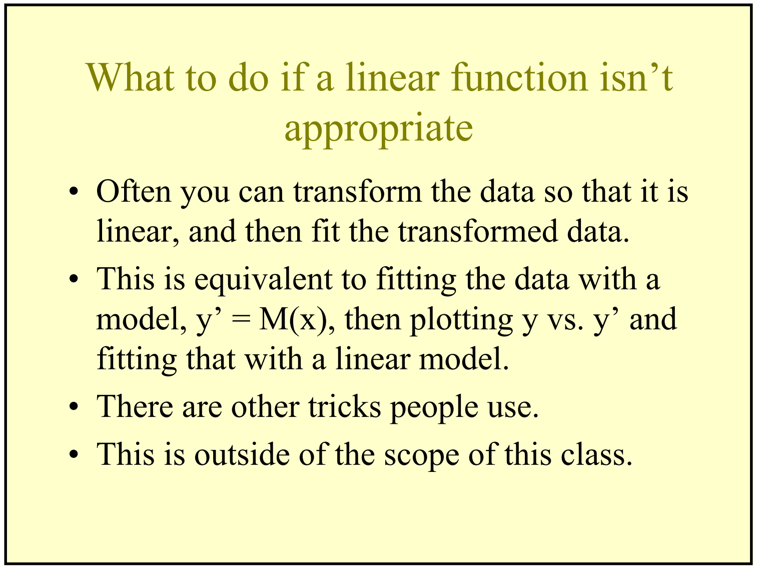




If it looks like this, you did something wrong – there's still a linear component!

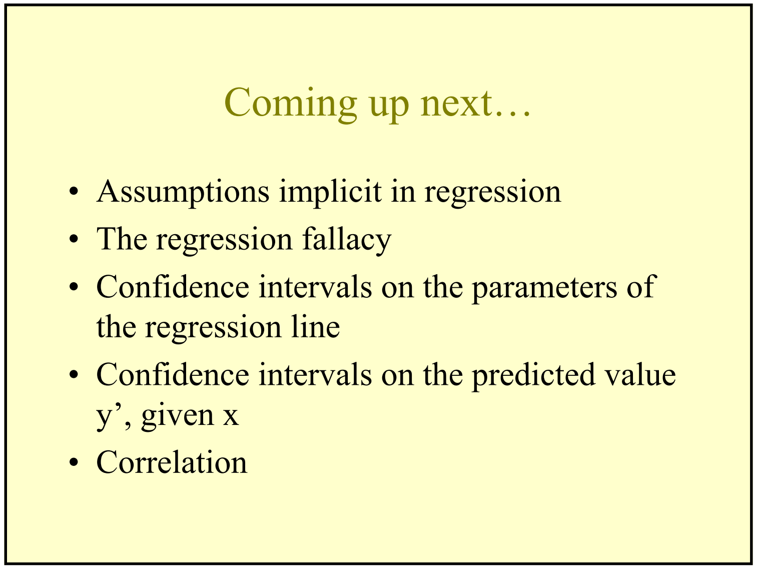


If there's a pattern, it was inappropriate to fit a line (instead of some other function)



What to do if a linear function isn't appropriate

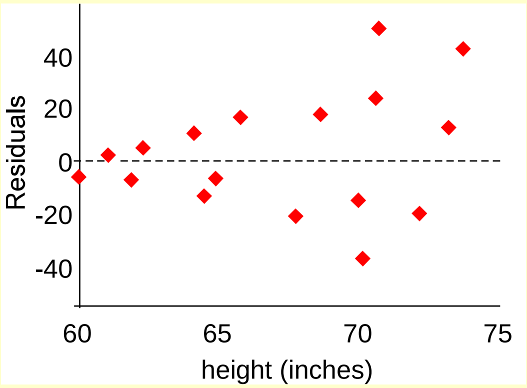
- Often you can transform the data so that it is linear, and then fit the transformed data.
- This is equivalent to fitting the data with a model, y' = M(x), then plotting y vs. y' and fitting that with a linear model.
- There are other tricks people use.
- This is outside of the scope of this class.



Coming up next…

- Assumptions implicit in regression
- The regression fallacy
- Confidence intervals on the parameters of the regression line
- Confidence intervals on the predicted value y', given x
- Correlation

Assumption #1: your residual plot should not look like this:



Heteroscedastic data

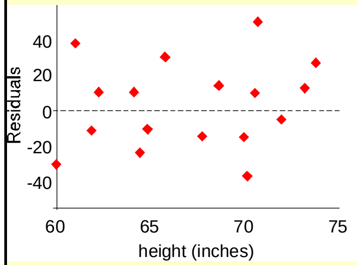
- Data for which the amount of scatter depends upon the x-value (vs. "homoscedastic", where it doesn't depend on x)
- Leads to residual plots like that on the previous slide
- Happens a lot in behavioral research because of Weber's law.
  - Ask people how much of an increment in sound volume they can just distinguish from a standard volume
  - How big a difference is required (and how much variability there is in the response) depends upon the standard volume
- Can often deal with this problem by transforming the data, or doing a modified, "weighted" regression
- (Again, outside of the scope of this class.)

Why we care about heteroscedasticity vs. homoscedasticity

- Along with the residual plots, we often want to look at the rms (root-mean-square) error for the regression:

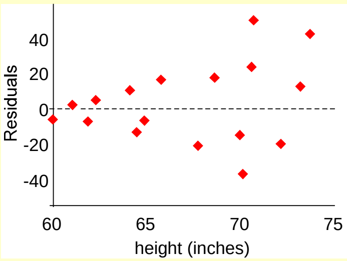
  rms = sqrt($\Sigma(y_i - y_i')^2/N$) = $s_{y'}$
- This gives us a measure of the spread around the regression line
- For this measure to be meaningful and useful, we want the data to be *homoscedastic*, i.e. we want the data to be spread out to the same degree for every value of x_i.

Homoscedastic



Here, rms error is a good measure of the amount of spread of the data about y_i', for any value of x_i.

Heteroscedastic



Here, rms error is not such a good measure of the spread -- for some x_i it will be an overestimate of spread, for some an underestimate.

Another assumption for regression analysis

- Assume the y scores at each x form an approximately normal distribution
- Because the rms error, $s_{y'}$ is like a standard deviation, if the above assumptions hold, then we can say things like, “approximately 68% of all y scores will be between ±1 $s_{y'}$ from the regression line”

Assumptions

- Homoscedastic data
- y scores approximately normally distributed about the regression line

The regression effect and the regression fallacy

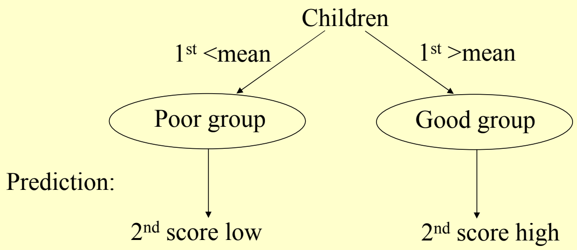
- Your book’s example:
  - A preschool program is aimed at increasing children’s IQ’s. Children are given a pre-test prior to entering the program, and a post-test at the end of the program.
  - On both tests, the mean score is about 100, and the SD is about 15. The program seems to have no effect.
  - However, closer inspection shows that students who started out with lower IQ’s had an average gain of 5 IQ pts. Students who started out with a higher IQ showed a drop in IQ of 5 pts, on average.

Is something interesting going on, here?

- Does the program act to equalize intelligence? Does it make smart kids less smart, and less bright kids more bright?
- No, nothing much is going on. This is just the *regression effect*.
  - In virtually all test-retest situations, the group that performs poorly on the first test will show some improvement (on average) on the second, and the group that performs better than average will do worse on the second test.
- The *regression fallacy* is assuming that this effect is due to something important (like the program equalizing intelligence). As we’ll see, it’s not.

Simple intuition about the regression effect

- Divide children into two groups, based upon their performance on the first test
  - "Poor" group vs. "Good" group


Children

1st <mean

1st >mean

Poor group

Good group

Prediction:

2nd score low

2nd score high

Simple intuition about the regression effect

- However, by chance, some of the children who belong in the "poor" group got a high enough score on test 1 to be categorized in the "good" group.
- And vice versa.
- If there's a real difference between the groups, on test 2 we'd expect some of these mis-labeled children to score more like the group they really belong to.

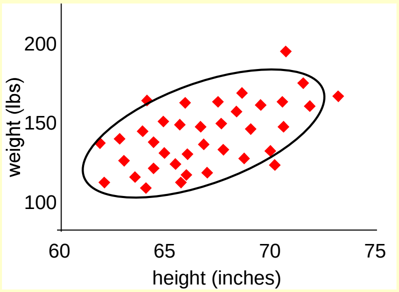
-> "poor" group scores, on average:
better on test 2 than test 1,
"good" group scores, on average:
worse on test 2 than on test 1

The regression effect: more involved explanation

- To explain the regression effect the way your book does, it will help first to talk some more about what kind of distribution we expect to see in our scatter plots, and more about some intuitions for what the least squares linear regression is telling us.

"Football-shaped" scatter plots

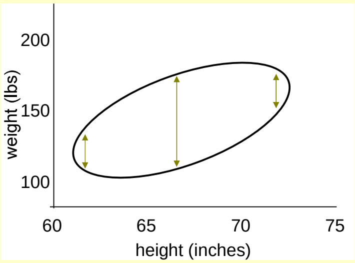
- For many observational studies, the scatter plot will tend to be what your book calls "football-shaped"



Why "football-shaped"?

- Just as weights and heights tend to have a (1-dimensional) normal distribution, a plot of bivariate (height, weight) data will often tend to have a 2-dimensional normal distribution
- The two-dimensional form of the normal "bell-shaped" distribution is a cloud of points, most of them falling within an elliptical "football-shaped" region about the mean(height, weight).
  - This is equivalent to most of the points in a normal distribution falling within +/- one standard deviation from the mean.

Wait, are football-shaped scatter plots homoscedastic?

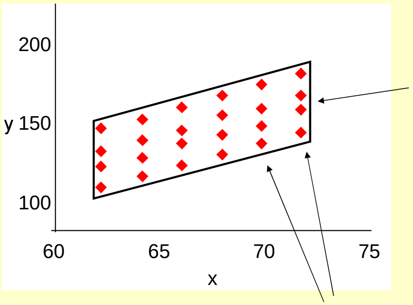


- If it's really football-shaped, isn't the spread of the data in the y-direction slightly smaller at the ends (for small and large x)?

Isn't the spread in the data smaller for small and large x?

- No, this is just because, in an observational study, there are fewer data points in the ends of the football – there just aren't that many people who are really short or really tall, so we have the illusion that there's less spread for those heights.
- In fact, for a controlled experiment, where an experimenter makes sure there are the same number of data points for each value of x_i, the scatter plot will tend to look more like this:

Typical scatter plot for a controlled experiment

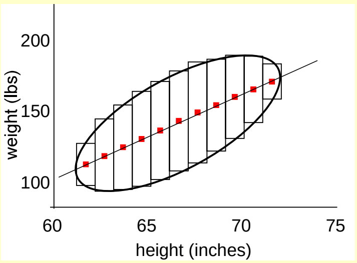


same # of data points for each x

collected data only for specific values of x

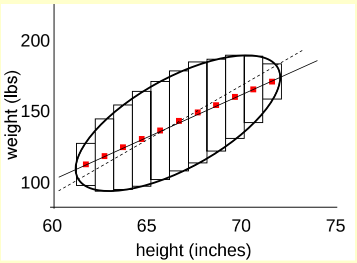
What happens in regression – an alternate view

- Divide the scatter plot into little vertical strips.
- Take the mean in each vertical strip.
- Least squares regression attempts to fit a line through these points.



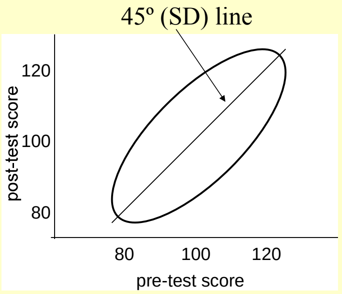
Note that the regression line is not the line through the axis of the ellipse

- The axis line, shown dashed, is what your book calls the "SD line"
- It starts at the mean, and as it moves right one s_x, it moves up one s_y



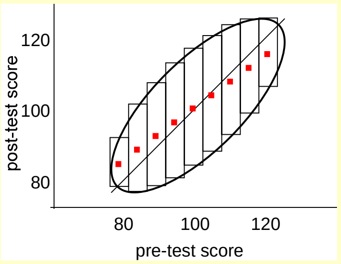
Consider what our ellipse looks like, in the preschool program situation

- If $\text{score}_{\text{test1}} = \text{score}_{\text{test2}}$ for each child, the data would fall exactly on a 45º line.
- However, this doesn't often happen. Instead we see spread about this line.
  - Spread is in both pre- and post-test scores.
  - So it looks like we've centered our "football" about the 45º line = the SD line



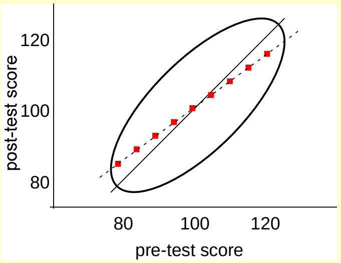
For each score range on test 1, look at the mean score expected on test 2

- To do this, we again look at a vertical strip through our football, and find the mean in that strip.
- This will be the mean score on the 2nd test for the students whose 1st test score lies within the given range.



Predicted scores on test 2 are closer to the mean than corresponding scores on test 1

- Solid line corresponds to $\text{score}_{\text{test1}} = \text{score}_{\text{test2}}$
- For $\text{score}_{\text{test1}} <$ mean predict $\text{score}_{\text{test2}} > \text{score}_{\text{test1}}$
- For $\text{score}_{\text{test1}} >$ mean predict $\text{score}_{\text{test2}} < \text{score}_{\text{test1}}$



The regression effect

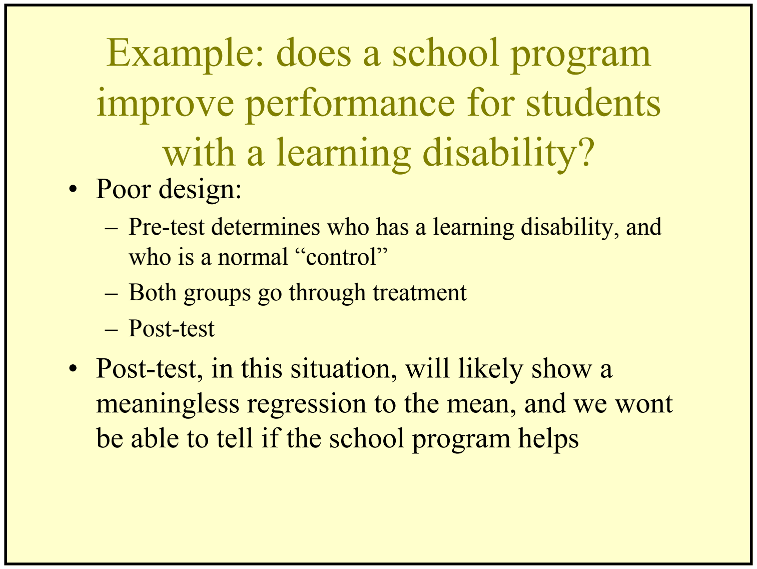
- We don’t expect scores on the second test to be exactly the same as the scores on the first test (the SD line)
- The scatter about the SD line leads to the familiar football-shaped scatter plot
- The spread around the line makes the mean second score come up for the bottom group, and go down for the top group.
- There’s nothing else to the regression effect – the preschool program does not equalize IQ’s.

“Regression to the mean”

- This effect was first noticed by aristocrat Galton, in his study of family resemblances.
  - Tall fathers tended to have sons with height closer to the average (i.e., shorter)
  - Short fathers also tended to have sons with height closer to the average (i.e., taller)
- Galton referred to this effect as “regression to mediocrity”
- Regression = “movement backward to a previous and especially worse or more primitive state or condition”
- This is where the term “regression analysis” comes from, although now it means something quite different

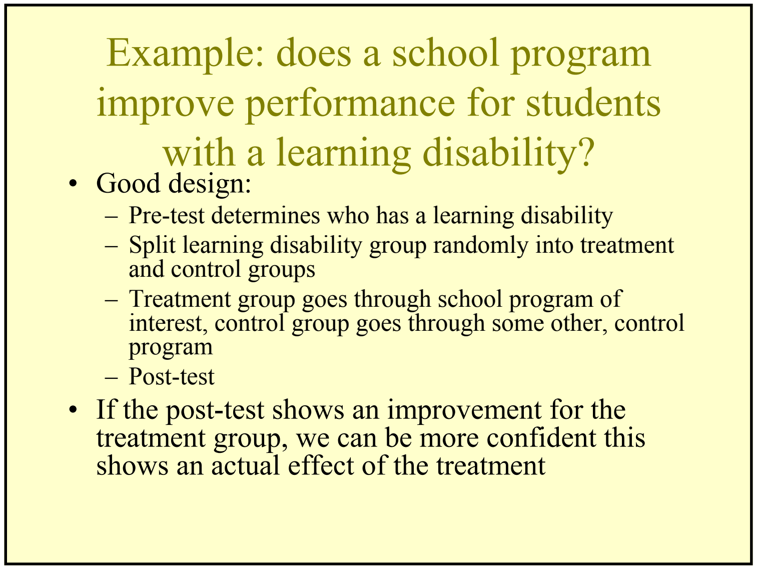
Important experimental design lesson!

- Do NOT choose the groups for your experiment based upon performance above or below a threshold on a pre-test
- The group that did worse on the pre-test will likely do better after the treatment, but this could just be the regression effect (and thus meaningless)



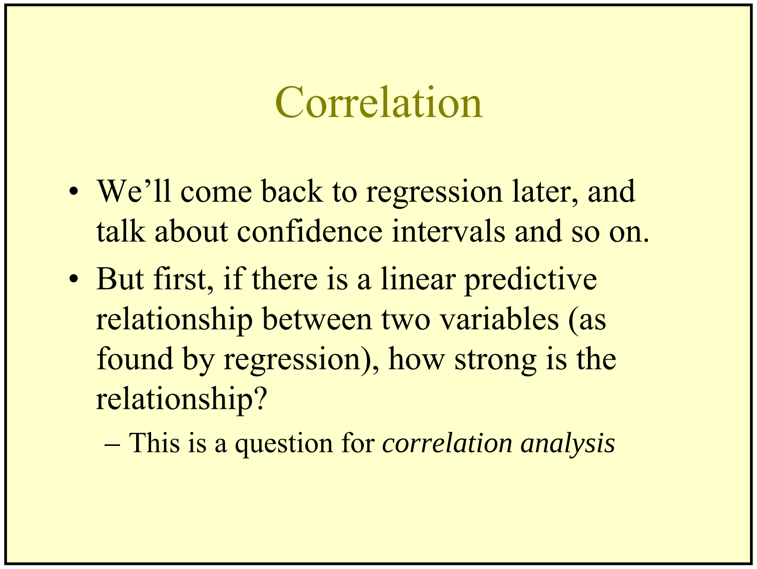
Example: does a school program improve performance for students with a learning disability?

- Poor design:
  - Pre-test determines who has a learning disability, and who is a normal “control”
  - Both groups go through treatment
  - Post-test
- Post-test, in this situation, will likely show a meaningless regression to the mean, and we wont be able to tell if the school program helps



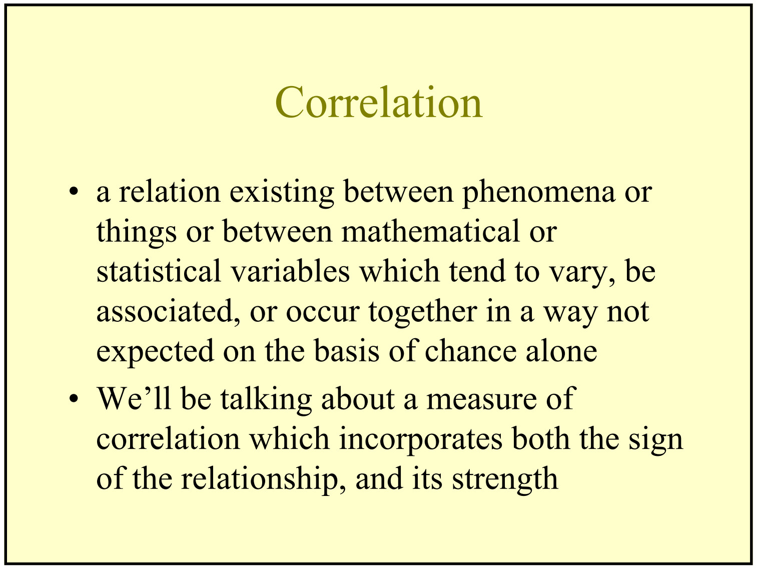
Example: does a school program improve performance for students with a learning disability?

- Good design:
  - Pre-test determines who has a learning disability
  - Split learning disability group randomly into treatment and control groups
  - Treatment group goes through school program of interest, control group goes through some other, control program
  - Post-test
- If the post-test shows an improvement for the treatment group, we can be more confident this shows an actual effect of the treatment



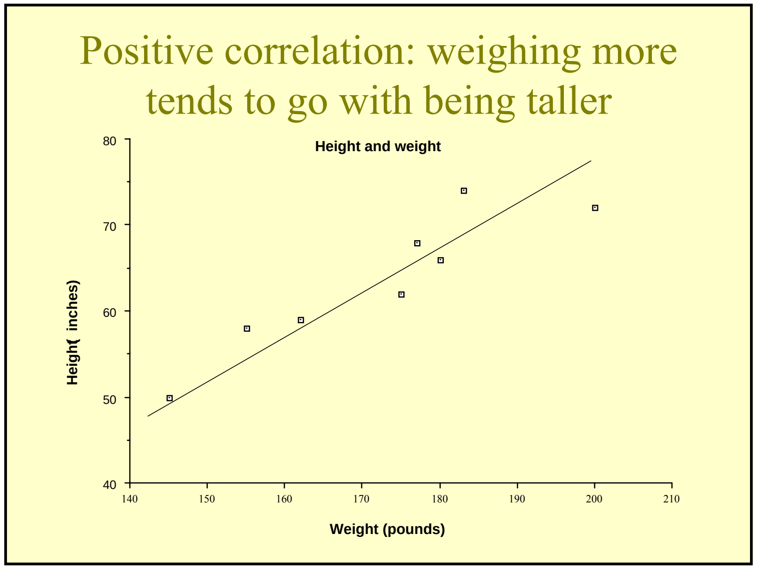
Correlation

- We’ll come back to regression later, and talk about confidence intervals and so on.
- But first, if there is a linear predictive relationship between two variables (as found by regression), how strong is the relationship?
  - This is a question for *correlation analysis*

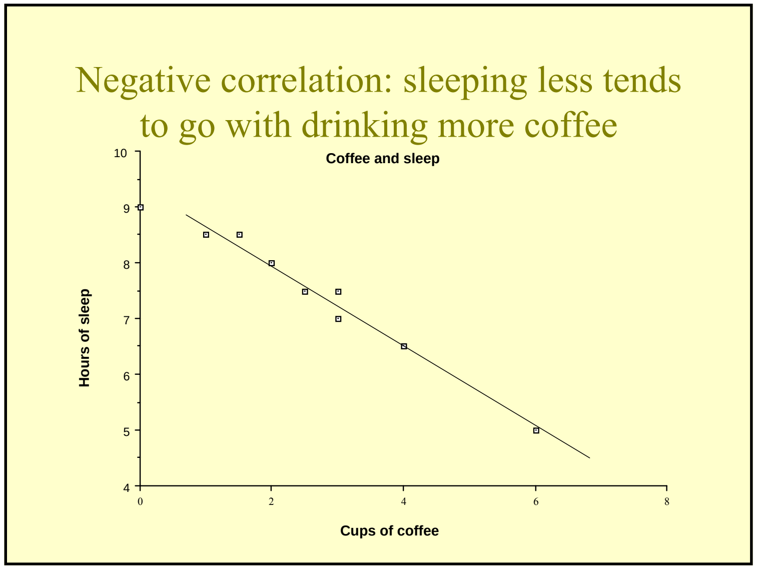


Correlation

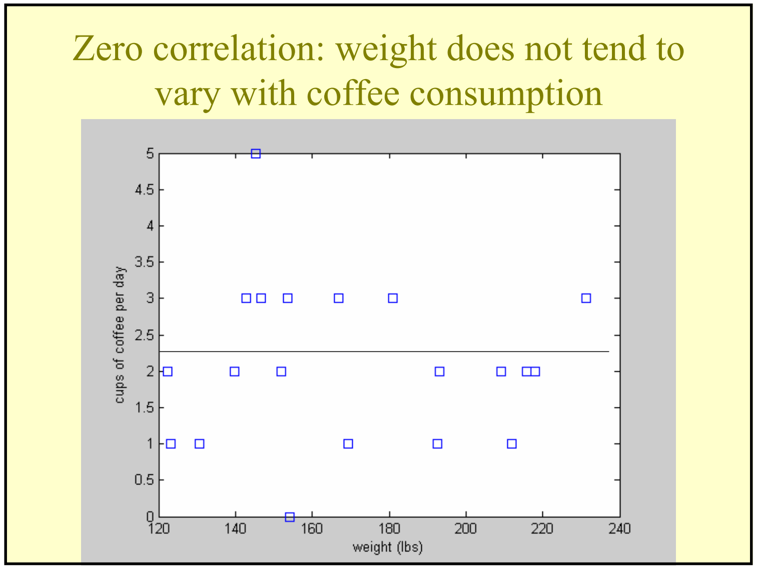
- a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone
- We’ll be talking about a measure of correlation which incorporates both the sign of the relationship, and its strength



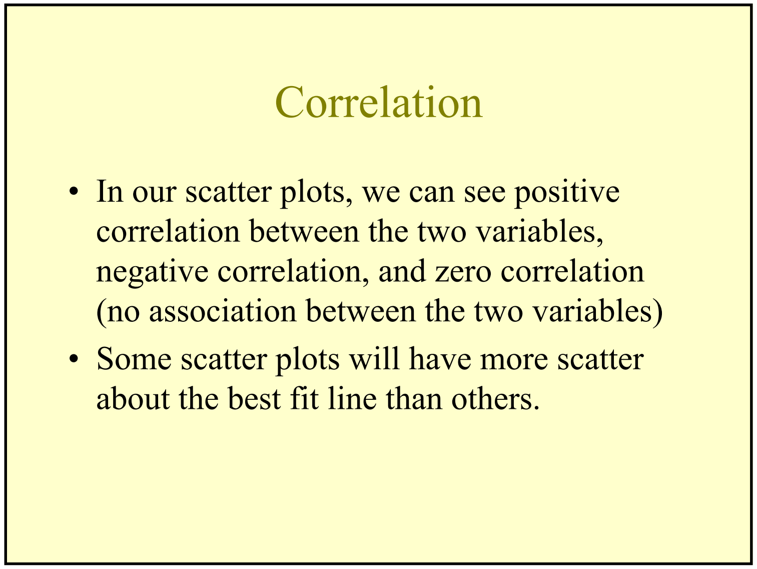
Positive correlation: weighing more tends to go with being taller



Negative correlation: sleeping less tends to go with drinking more coffee

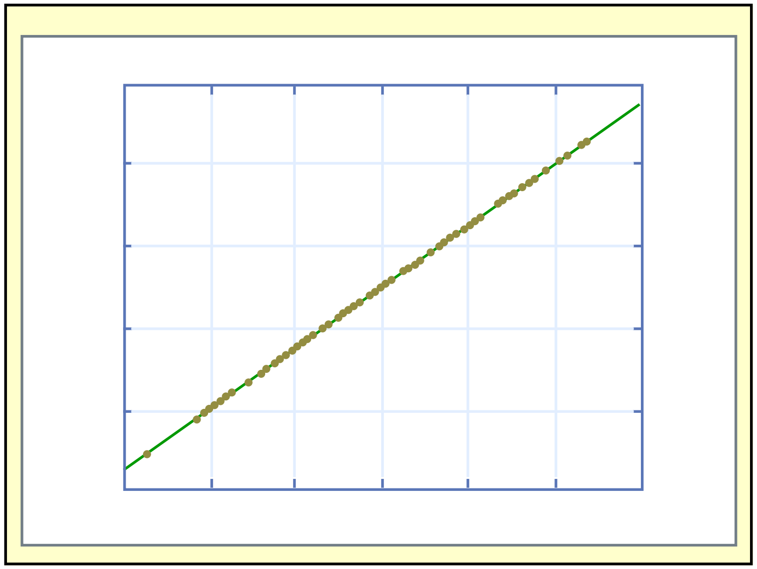


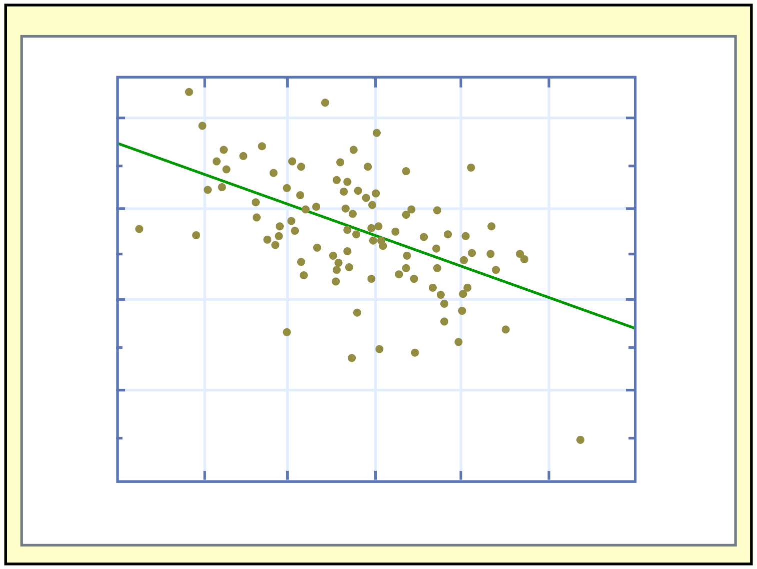
Zero correlation: weight does not tend to vary with coffee consumption



Correlation

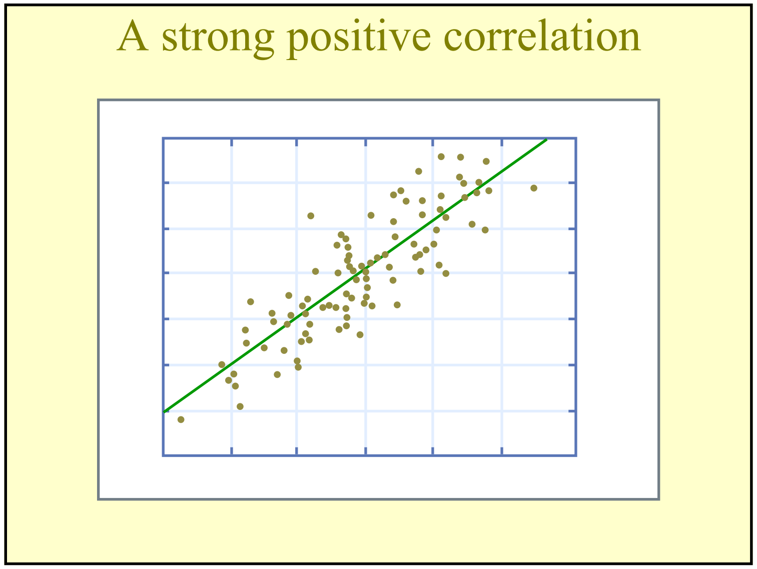
- In our scatter plots, we can see positive correlation between the two variables, negative correlation, and zero correlation (no association between the two variables)
- Some scatter plots will have more scatter about the best fit line than others.





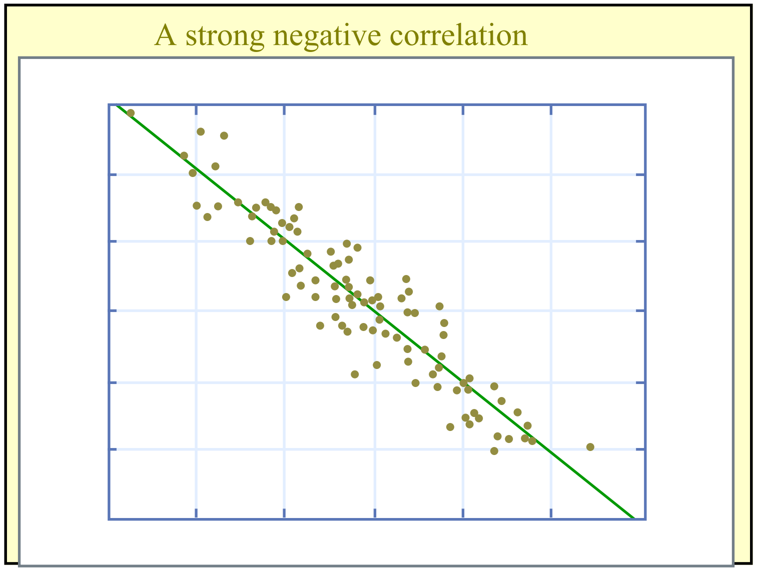
Strength of association

- When the points cluster closely about the best fit line, the association between the two variables is *strong*.
  - Knowing information about one variable (e.g. a person’s height is 65”) tells you a lot about the person’s weight (it’s probably about 177 lbs)
- When the spread of points increases, the association weakens
  - Knowing information about one variable doesn’t help you much in pinning down the other variable

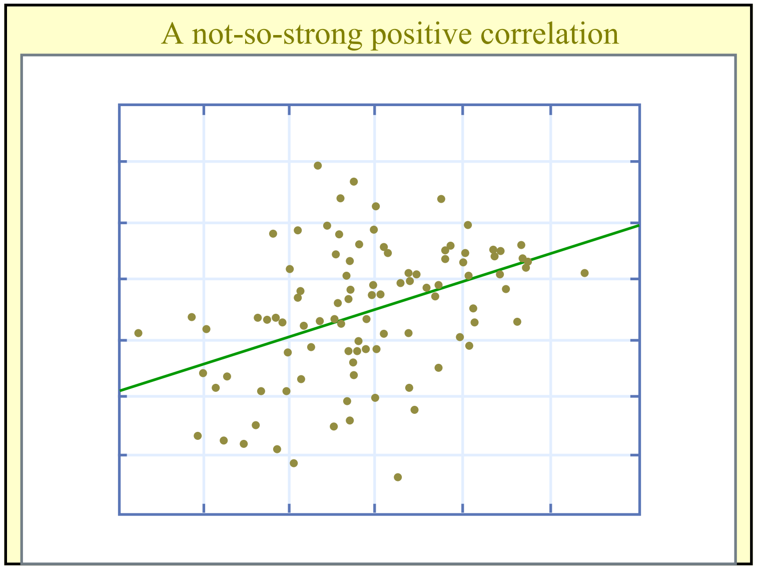


A strong positive correlation

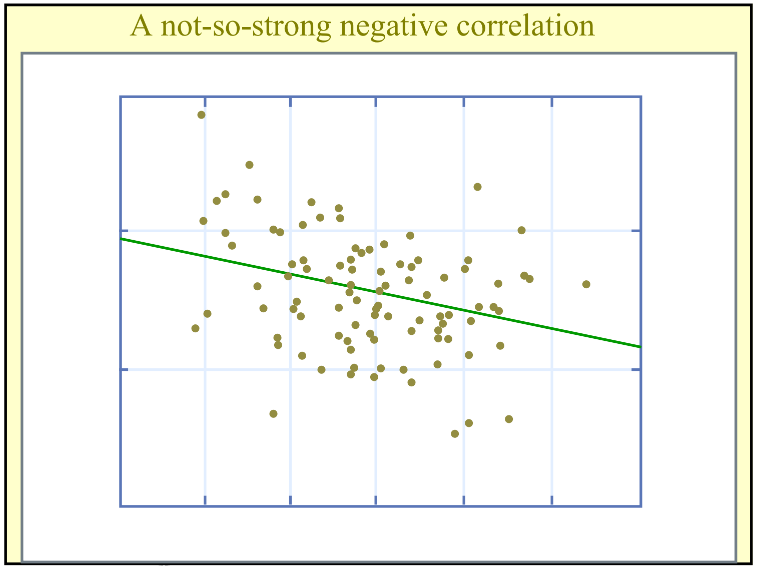
Figures by MIT OCW.



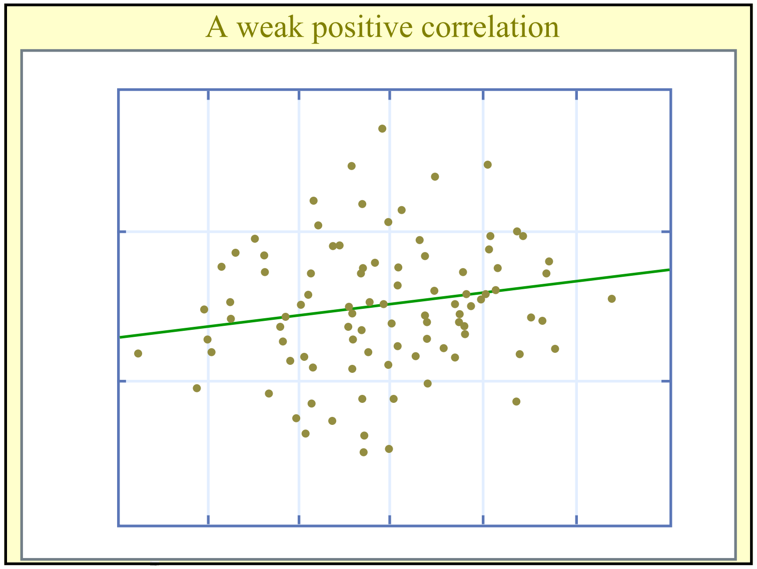
A strong negative correlation



A not-so-strong positive correlation



A not-so-strong negative correlation



A weak positive correlation

Figures by MIT OCW.

Measuring strength of relationship

- The Pearson Product-Moment Correlation Coefficient, r
- Provides a numerical measure of the strength of the relationship between two variables
- $-1 \leq r \leq 1$
- Sign indicates direction of relationship
- Magnitude indicates strength

The formula for r

- Remember, r has to do with ‘correlation’ not with ‘regression’

$$r = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - m_x)}{s_x}\frac{(y_i - m_y)}{s_y} = \frac{1}{N}\sum_{i=1}^{N} z_x z_y$$

(Use the 1/N form of s_x and s_y)

Unpacking the formula for *r*

- r is based on z-scores -- it is the average product of the standard scores of the two variables
- r will be positive if both variables tend to be on the same side of their means at the same time, negative if on opposite sides
- r will be 0 if there’s no systematic relationship between the two variables

r = 1.00

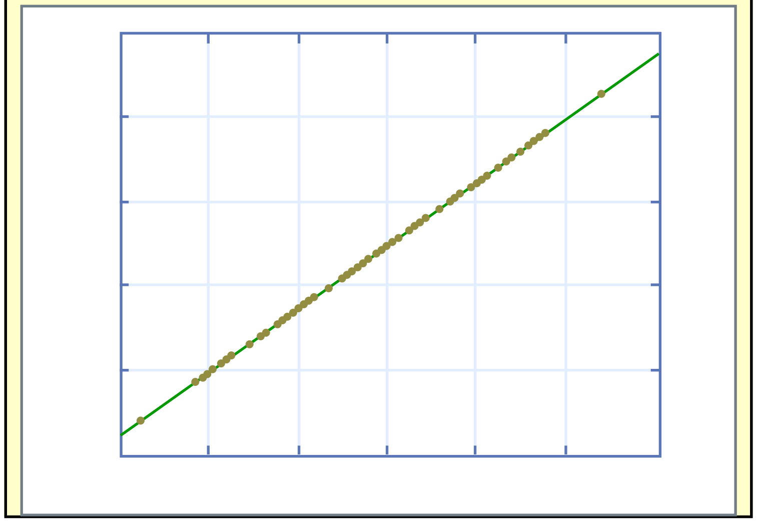
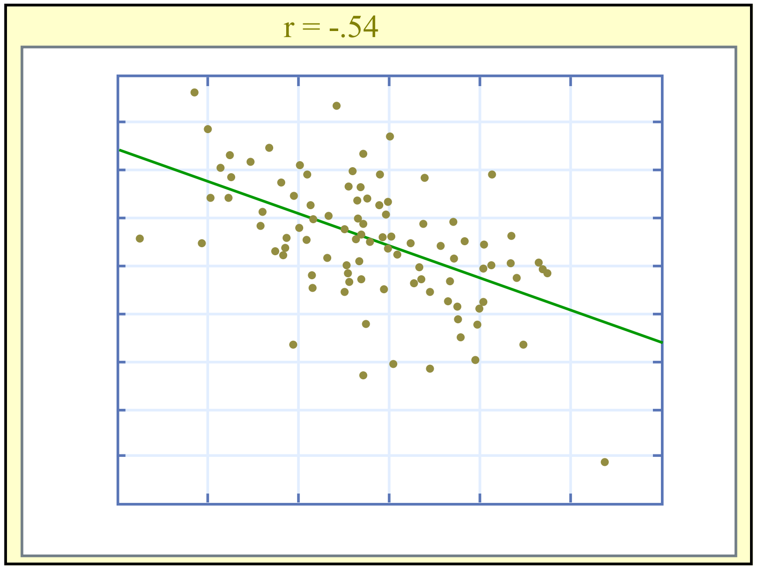
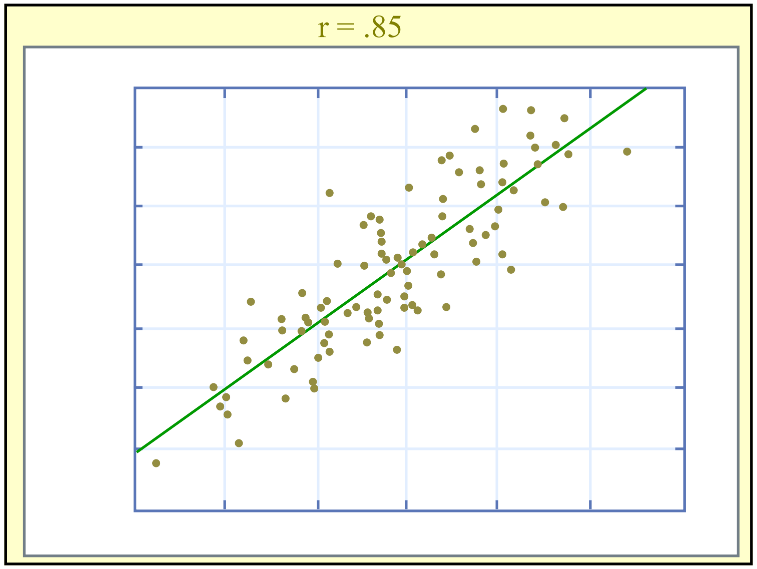


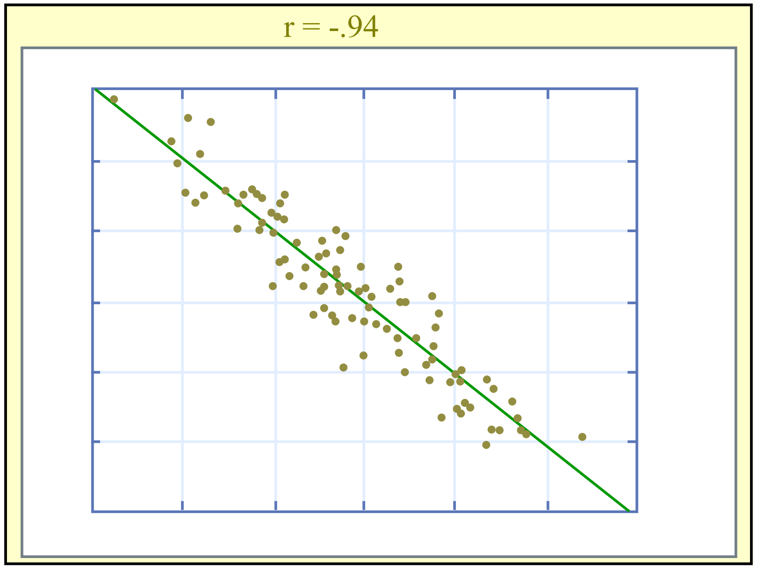
Figure by MIT OCW.



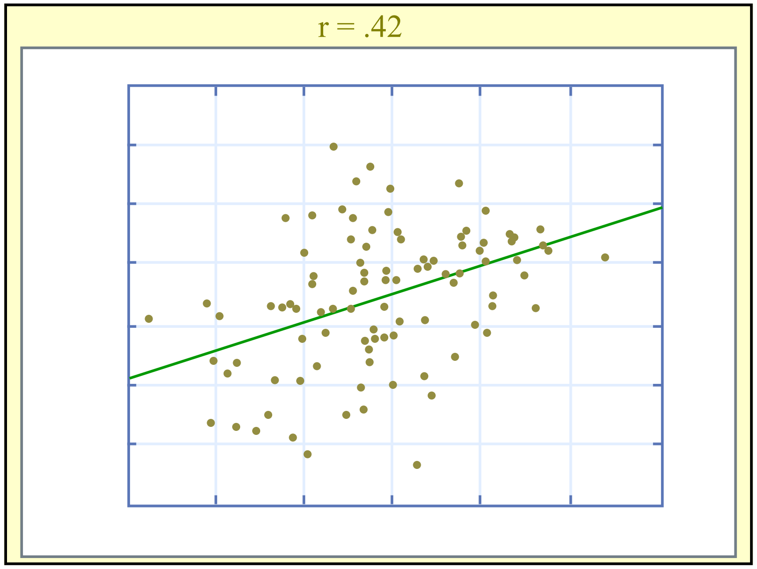
r = -.54



r = .85



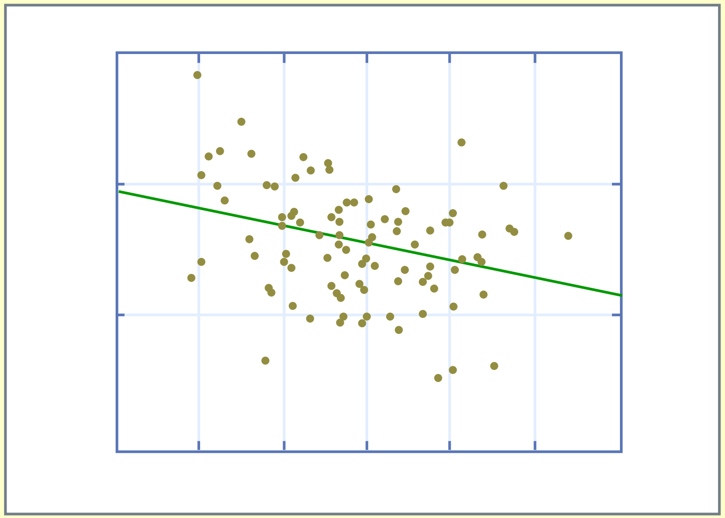
r = -.94



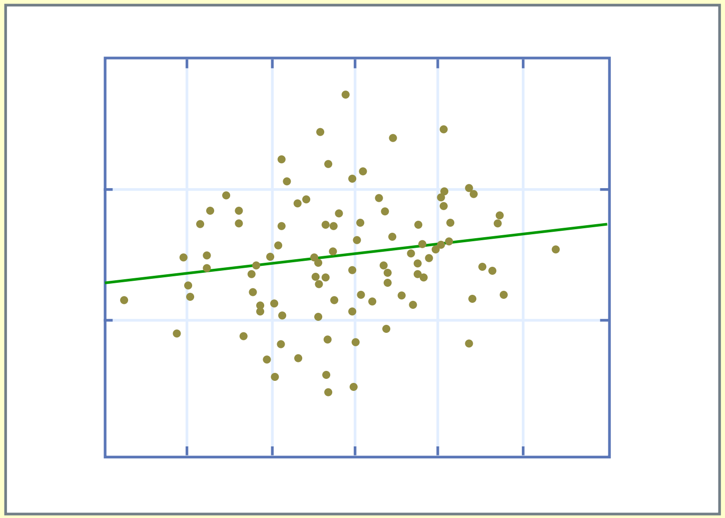
r = .42

Figures by MIT OCW.

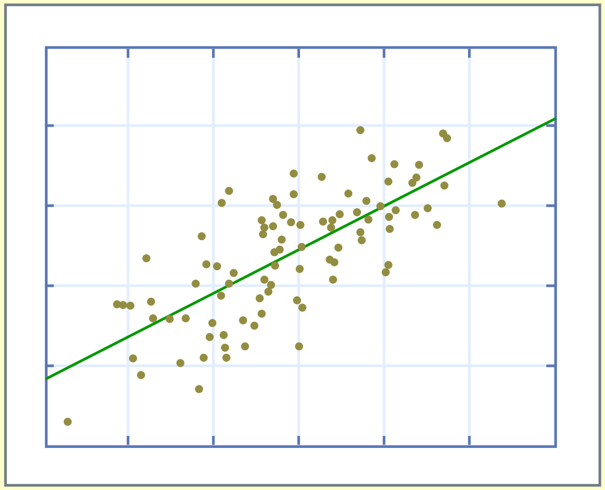
r = -.33



r = .17



r = .39

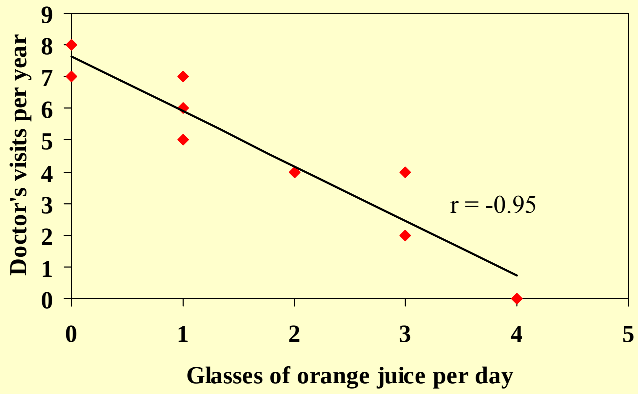


Example

- Looking at the correlation between # of glasses of juice consumed per day (x), and doctor visits per year (y)
- x = [0 0 1 1 1 2 2 3 3 4];
- y = [8 7 7 6 5 4 4 4 2 0];
- zx = (x-mean(x))/std(x, 1);
- zy = (y-mean(y))/std(y, 1);
- r = mean(zx.*zy) = -0.95

Figures by MIT OCW.

Scatter plot



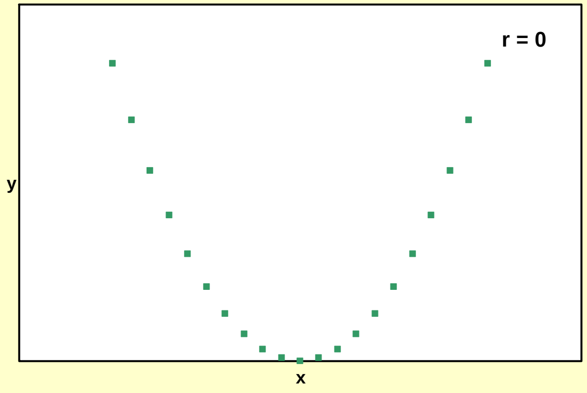
Behavioral research

- (Only rules of thumb:)
- Typical r within ±0.30 or ±0.50
- ±0.20 typically considered weak
- Between ±0.60 and ±0.80 quite strong, impressive
- Greater than ±0.80 is extremely strong and unlikely in this sort of research
- Greater than ±1.0? Something's wrong.

Limitations of r

- r only tells you whether two variables tend to vary together -- nothing about the nature of the relationship
  - E.G: *correlation is not causation!!*
- r only measures the strength of the ***linear*** relationship between two variables
  - Other kinds of relationships can exist, so look at the data!

No correlation?



An alternative computational formula for r

$$r = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - m_x)}{s_x}\frac{(y_i - m_y)}{s_y} = \frac{1}{N}\sum_{i=1}^{N} z_x z_y$$

- Note we can write this as:

  $r = E[(x-m_x)/s_x \cdot (y-m_y)/s_y] = \mathrm{cov}(x, y)/(s_x\, s_y)$
- Plugging in equations for cov(x, y), s_x, and s_y, you’ll also see this version:

$$r = \frac{N(\sum xy) - (\sum x)(\sum y)}{\sqrt{[N(\sum x^2) - (\sum x)^2][N(\sum y^2) - (\sum y)^2]}}$$

Notes on computing r

- That last equation is intended for computing by hand/with a calculator. If you have to compute by hand, it’s probably the most efficient version.
- These days, in MATLAB the easiest thing to do is just to compute cov(x, y), s_x, and s_y.

Juice & doctor visits example, again

- x = [0 0 1 1 1 2 2 3 3 4];
- y = [8 7 7 6 5 4 4 4 2 0];
- tmpmx = cov(x, y, 1)/(std(x, 1)*(std(y, 1))
- % Recall: cov returns the matrix

  % [var(x) cov(x,y); cov(x,y) var(y)]
- r = tmpmx(1, 2) = -0.95