Fact: r does not depend on which variable you put on x axis and which on the y axis.

Often: interest centres on whether or not changes in x cause changes in y or on predicting y from x.

In this case: call y the dependent or outcome or response or endogenous variable. **Response** in this course.

Call x explanatory, or exogenous or independent or predictor.

Example: predict son's height from father's height.

Suppose father is 70 inches tall. How to guess height of son?

Simple method: use average height of those sons whose fathers were 70 inches tall.

Pick out cases where father is between 69.5 and 70.5 inches tall. There were 115 such fathers.

Average son's height in this group: 69.8 in

SD son's height in this group: 2.5in



Sons of 70 inch fathers

Now do same for fathers 59 inches tall, then 60 inches tall and so on.

Get a \bar{y} for each x from 59 to 75.



Father's Height

Notice the line: it is called a **regression** or **least-squares** line.

Formula for the least squares line?

$$y = a + bx$$

Where:

$$a = \bar{y} - b\bar{x}$$

and

$$b = r \frac{s_y}{s_x}$$

I prefer to write:

$$\frac{y-\bar{y}}{s_y} = r\frac{x-\bar{x}}{s_x}$$

In words: predict y in standard units to be x in standard units times correlation coefficient.

Jargon:

a is the intercept.

b is the slope also called **regression** coefficient.

"Least squares" because we find formulae for a and b by using calculus to minimize the Error Sum of Squares:

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

Sum of vertical squared deviations between (x_i, y_i) and straight line with slope b and intercept a.

For the height data

Fathers: $\bar{x} = 67.7$, $s_x = 2.74$ (inches)

Sons: $\bar{y} = 68.7$, $s_y = 2.81$ (inches)

Correlation: r = 0.50.

Average weight vs height for STAT 201:



"Fit not as good".

Scatterplot not too oval; mixing sexes in same plot.

Numerical values:

Height: $\bar{H} = 66.8$, $s_H = 3.75$ (inches).

Weight: $\bar{W} = 140$, $s_W = 25.3$ (pounds).

Correlation: r = 0.73.

Regression line:

Slope: $b = 0.73 \times 25.3/3.75 = 4.93$ (pounds per inch)

Intercept: $a = 140 - 4.93 \times 66.8 = -189$ (pounds)

Meaning of intercept: NONE whatever. Not to be understood as weight of person 0 inches tall.

DO NOT USE regression line outside of range of x values!

DO NOT EXTRAPOLATE.

Issues:

1) Regression effect: when r > 0: cases high in one variable predicted to be high in the other BUT closer to mean in standard deviation units. Called *regression to the mean*.

Cases low in one variable predicted to be low in the other but not as low.

For r < 0: cases above mean in one variable predicted to be below mean on other but not as far below.

2) Residual analysis: straight line regression not always appropriate. Watch for non-linearity, outliers, influential observations. Plot residuals

$$y_i - a - bx_i = y_i - (a + bx_i)$$

against x_i to look for problems.

3) Residual variability: for oval shaped scatterplots histogram of y values for a given x value tend to follow normal curve. Mean predicted by regression line; SD is roughly

$$\sqrt{1-r^2}s_y$$

4) Cause and effect. Variables x and y can be highly correlated without changes in one *caus-ing* changes in the other. Watch out for **lurk-ing** or **confounding** variables. Do controlled experiments.

5) Ecological correlations; replacing groups of cases by averages can change correlation dramatically. Illustration of regression effect using height data

Fathers: $\bar{x} = 67.7$, $s_x = 2.74$ (inches)

Sons: $\bar{y} = 68.7$, $s_y = 2.81$ (inches)

Correlation: r = 0.50.

Predict average son's height when father is 72 inches:

My way without remembering formulas:

Convert 72 to standard deviation units:

$$\frac{72 - 67.7}{2.74} = 1.57$$

Predict son's height in Standard units to be 0.50*1.57=0.78

Convert back to original units:

$$68.7 + 2.81 * 0.78 = 70.9$$

OR work out a and b and use regression line:

Slope is $b = rs_y/s_x = 0.50 \times 2.81/2.74 = 0.513$.

Intercept is $a = \bar{y} - b\bar{x} = 68.7 - 0.513 \times 67.7 = 34.0.$

Prediction is

 $\hat{y} = a + 72b = 34.0 + 0.513 * 72 = 70.9$ inches.

Now take sons who are 70.9 inches tall and predict father's height?

NOT just going backwards to 72 inches!

Convert 70.9 to Standard units: get back 0.78.

Multiply by r: predict father's height is 0.39 in standard units.

This is 67.7+0.39*2.71=68.7 inches.

Explanation: picking out 72 inch fathers picks out strip on right side of picture. Picking out 70.9 inch sons picks our strip across top. Different groups of people!



r=0.5

Father's Height (Inches)

Residual plots. plot of $y_i - a - bx_i$ against x_i should be flat, not wider at one end than the other, not curved, no big outliers.



Two inluential observations removed



Distance Driven (km)

Notice that in top plot the main body of dots seems to slope down and to right.

The regression line is not too useful. Difference in two plots: omission of two data points. Gives two different lines:



Plot of SD of son's heights for each different father's height.



Note line across at height

$$\sqrt{1-r^2} imes s_y$$

Idea: SD of y when x is held fixed is smaller than overall SD of y by factor of $\sqrt{1-r^2}$.

Usually expressed in terms of variance: smaller by factor of $1 - r^2$.

Jargon the fraction r^2 for the variance of y is "explained by the variation in x". The rest, the other $1 - r^2$ is "unexplained variation".

This brings up the 4th point.

Our use of regression is for prediction for a new value of x observed in the same way.

NOT to predict what would happen if you changed x.

Example: if you gave the father drugs to make him grow 2 inches the son would not get taller. Manipulating father's height doesn't impact son's height.

But sometimes manipulating x changes y. We say changes in x cause changes in y.

Example: blood pressure (y) regressed on drug dose (x). We hope that changing x will cause a change in y;

If the data are collected in a **study** not an **experiment** we usually cannot tell if changes in x cause changes in y.

Standard example: weekly sales of soft drinks versus weekly cases of polio diagnosed in US during year 1950. Correlation positive.

Why? Both go up in the summer. The **confounding** or **lurking** variable is weather. Good weather brought increases in both. Ecological correlations: correlations computed between averages.

Example: 11 TAs for STAT 2 in 1975.

Relation between average rating of TA in section and average Final exam mark:



r= -0.57

Now make up hypothetical data consistent with known averages:



+

TA Rating

\$

♦

TA Rating

Look at a few TAs for the r=0.8 example. In each section correlation is high. Overall correlation using raw data positive. Correlation using averages negative!









