

Class 3: Two Variables Sections 2.1, 2.2, 2.3

Math 263, Section 5, Deborah Hughes Hallett

Example: Spam. What determines how much spam you get? (Based on “Do Zebras get more Spam than Aardvarks?” Richard Clayton, Computer Laboratory, University of Cambridge, presented at *Fifth Conference on Email and Anti-Spam*, August 2008.) <http://www.lightbluetouchpaper.org/2008/08/25/zebras-and-aardvarks/>

1. QUANTITATIVE AND CATEGORICAL VARIABLES

Quantitative variables:

Categorical variables:

2. SCATTERPLOTS

We look at the association between two quantitative variables. In this case, what is the relationship between the numbers of emails and the proportion of spam? To compare them, we can use a clustered column graph (below) or a scatterplot (next page).

If you are interested in the relationship between the first letter/number in your email address and the proportion of spam, which graph is easier?

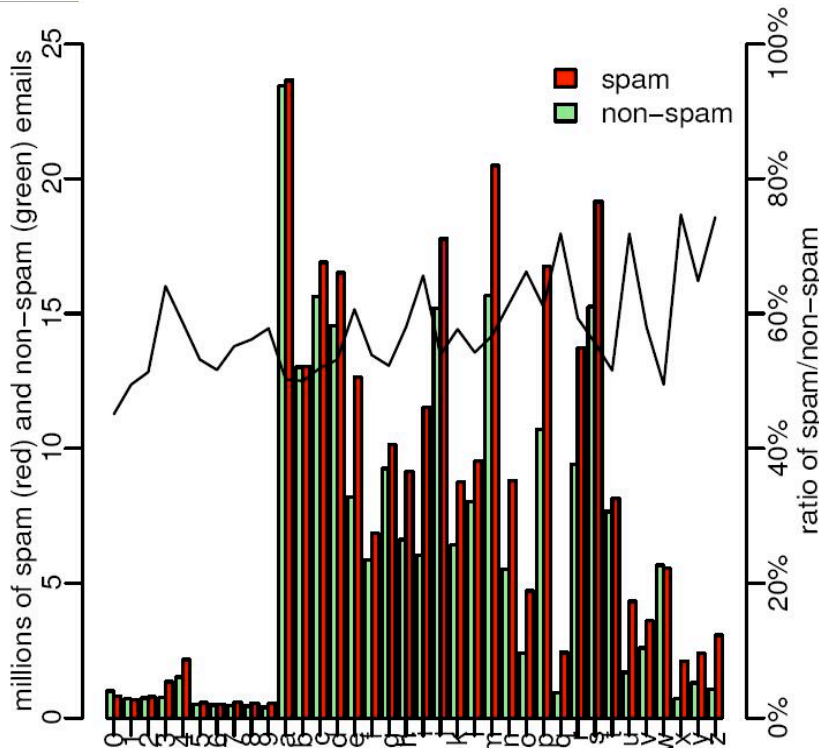


Figure 1: Spam (red) and non-spam (green) email for 8 week period, where local parts begin with particular letters. Line shows percentage of email that is spam.

Notice:

- Less email for addresses beginning with a number.
- Addresses starting with an “a” get more email and more spam than those starting with a “z”.
- The “a”s get a slightly smaller proportion of spam than the “z”s.

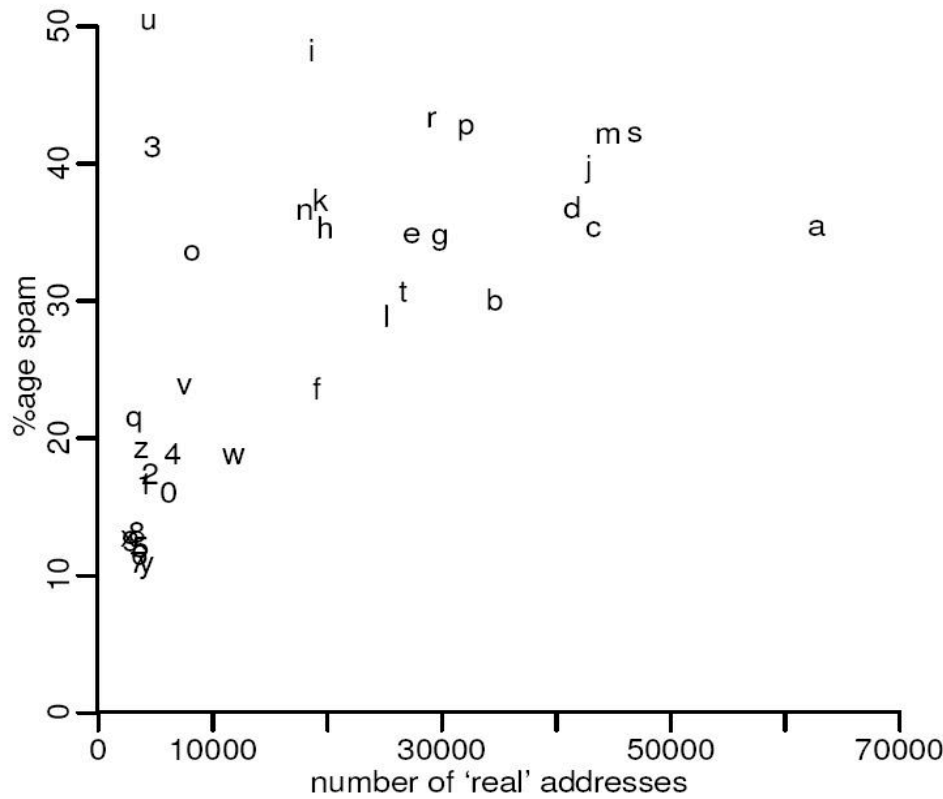


Figure 3: Relationship between number of email addresses receiving 28 or more non-spam emails over 8 weeks, and the proportion of email these addresses receive that is spam.

What is on each axis in the scatterplot?

Explanatory Variable: Horizontal axis:

Response Variable: Vertical axis:

What does the scatterplot tell us? What does the scatterplot *not* tell us?

A strong association between two quantitative variables does *not* necessarily mean there is causation.

3. NUMERICAL DESCRIPTORS: Strength of an association given by the correlation coefficient, r .

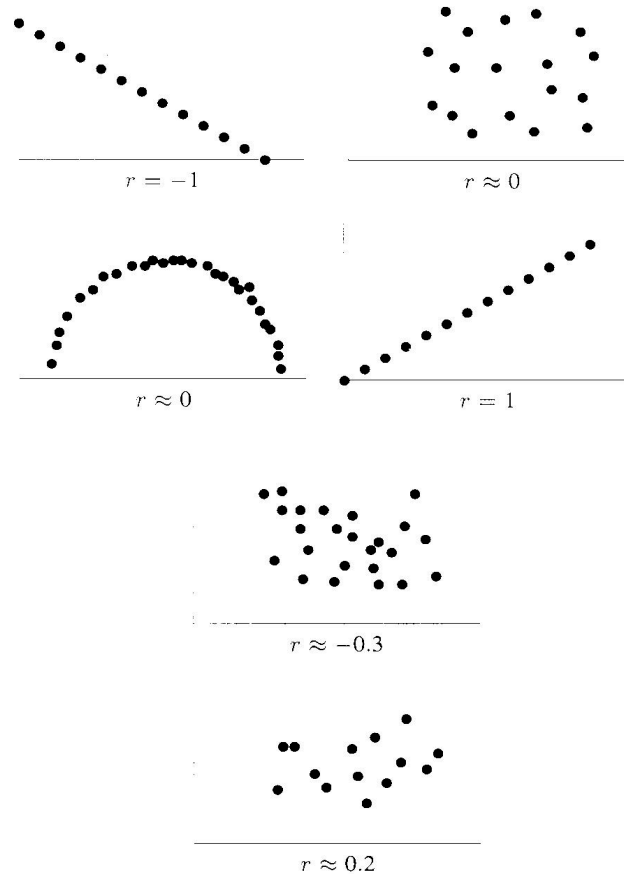
What is formula for r ?

Suppose there are n data points, $(x_1, y_1), (x_2, y_2)$, etc. Let a typical x -value be x_i and a typical y -value be y_i and the corresponding means be \bar{x} and \bar{y} . Writing s_x for the standard deviation of the x -values and s_y for the standard deviation of the y -values, the formula for the correlation coefficient is

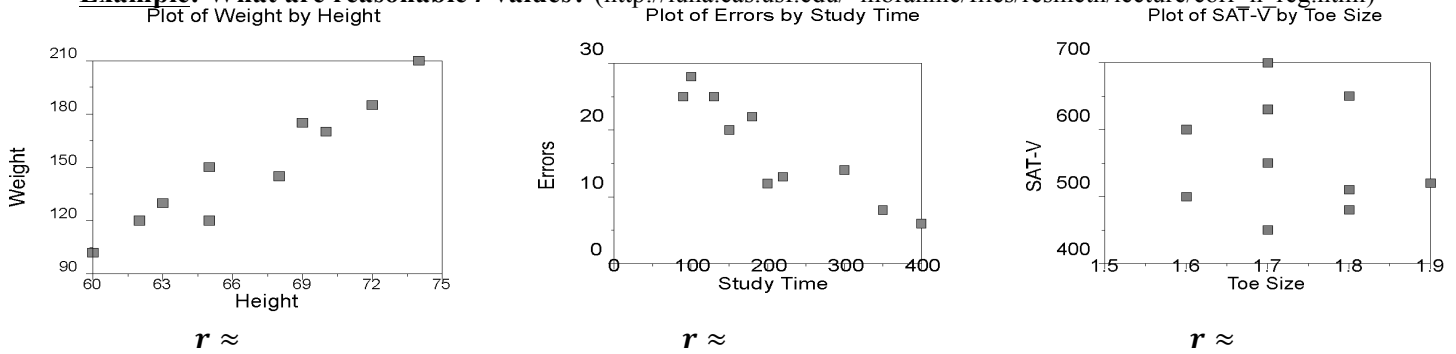
$$r = \frac{1}{n-1} \sum_{i=1}^{i=n} \frac{(x_i - \bar{x})}{s_x} \cdot \frac{(y_i - \bar{y})}{s_y}.$$

What can you tell about r by looking at the formula?

What values can r take? What does the value tell us? See the examples pictured below.



Example: What are reasonable r values? (http://luna.cas.usf.edu/~mbrannic/files/resmeth/lecture/corr_n_reg.html)



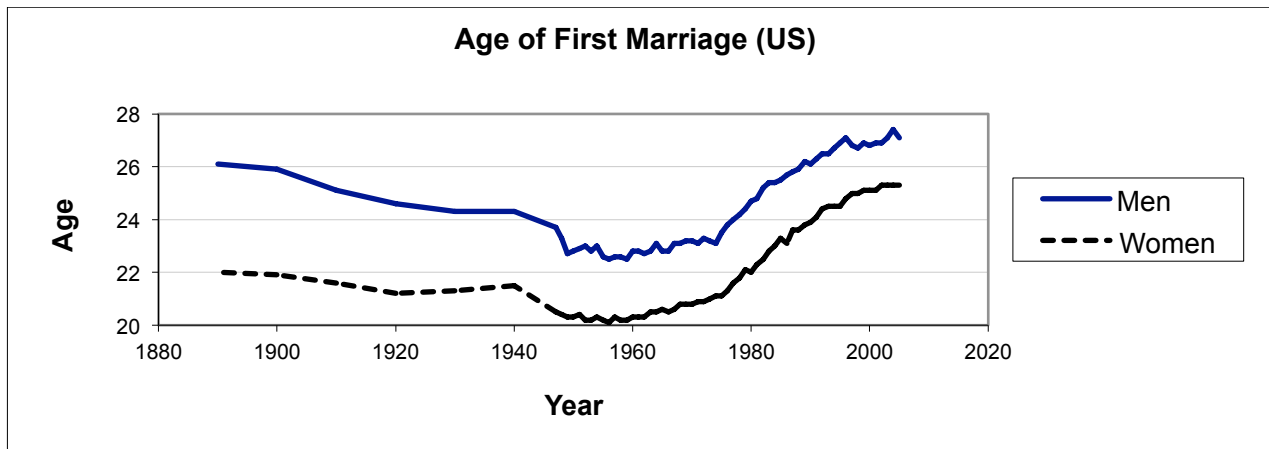
Example: What does the following abstract tell us?

“The present study¹ is based on the association of hand grip strength (both left and right) with height, weight and BMI on randomly selected 600 normal healthy individuals(300 boys and 300 girls) aged 6-25years of Amritsar, Punjab. The findings of present study indicate a strong association of right and left hand grip strength with height ($r = 0.925$ and $r = 0.927$ respectively in boys and $r = 0.800$ and $r = 0.786$ respectively in girls), weight ($r = 0.882$ and $r = 0.878$ respectively in boys and $r = 0.698$ and $r = 0.690$ respectively in girls) and with BMI ($r = 0.636$ and $r = 0.632$ respectively in boys and $r = 0.477$ and $r = 0.472$ respectively in girls).”

¹ From “An Association of Hand Grip Strength with Height, Weight and BMI in Boys and Girls aged 6-25 years of Amritsar, Punjab, India” Koley, Gandhi, Singh *The Internet Journal of Biological Anthropology* . 2008. Volume 2 Number 1. <http://www.ispub.com/ostia/index.php?xmlFilePath=journals/ijba/vol2n1/grip.xml>

4. REGRESSION LINE: ORDINARY LEAST SQUARES (OLS): Modeling a Trend

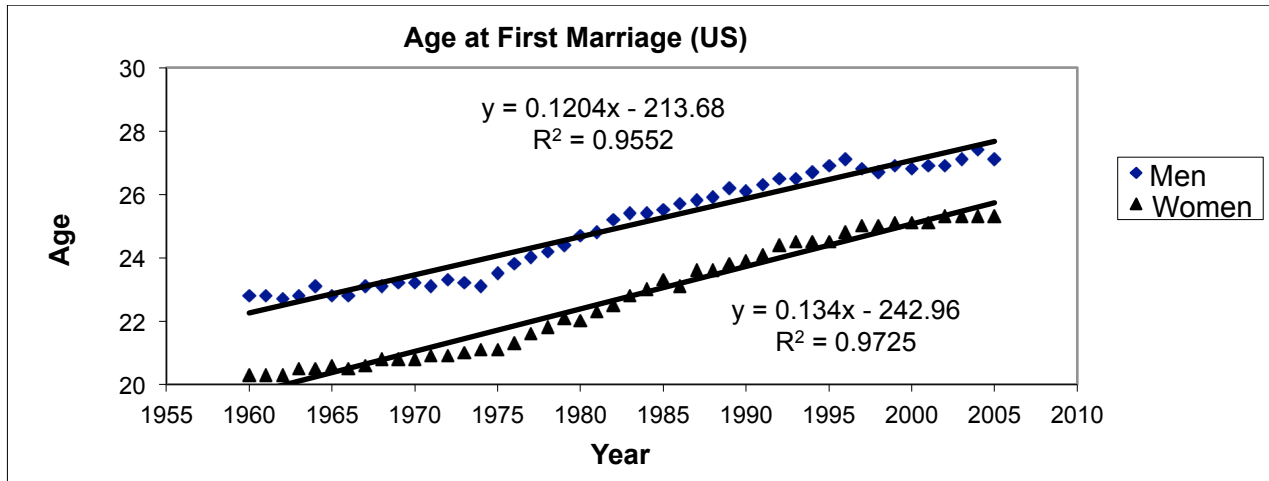
Example: What is the trend in the age of marriage in the US?²



² Source: U.S. Census Bureau: "Estimated Median Age at First Marriage, by Sex: 1890 to the Present" Sept 21, 2006

³ Decennial Census and American Community Survey, 2010.

Example: Using the data from 1960-2005, fit a line and use it to predict the age of marriage for men and women in 2010. In what year is the marriage age for men predicted to reach 30? Comment on the reliability of these answers.

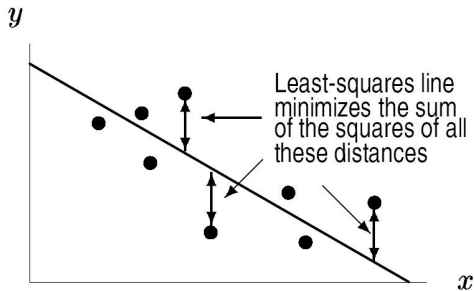


Example: Interpret the coefficient of x in the marriage age regression lines. What are the units?

Example: If the current trends continue, will men's and women's marriage ages ever be equal? If so, when?

5. HOW IS THE REGRESSION LINE CALCULATED? WHAT DOES “LEAST SQUARES” MEAN?

We call the x -variable the **explanatory** variable and the y -variable the **response** variable.



Given a set of data, the least squares regression line is chosen to minimize the sum of the squared vertical distances from the line. This is called the **ordinary least squares** (OLS) line.

- The distances are squared so that the positive and negative values don't cancel
- The distances are measured vertically because we are interested in predicting y for a given (fixed) value of x .

What is the relationship between the means, \bar{x} and \bar{y} , and the regression line?

It can be shown that the line goes through the point with coordinates (\bar{x}, \bar{y}) .

What is the relation between r and the regression line?

A line is usually written $y = b + mx$. If s_x is the standard deviation of the x s, and s_y is the standard deviation of the y s, and r is the correlation coefficient, then slope of line is given by

$$m = \text{Slope} = r \frac{s_y}{s_x}$$

Thus the intercept is

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

Interpretation of parameters:

In the linear relationship $y = b + mx$, with response variable y and explanatory variable x :

—The **constant b** is the initial value of y , that is the value of y when $x = 0$. Its units are the units of y .

—The **coefficient m** is the change in y if x increases by one unit. The units of m are the units of y divided by the units of x .

What is the interpretation of the coefficient of determination, written r^2 or R^2 ?

Let the original data points be (x_i, y_i) for $i = 1, 2, 3, \dots, n$. Let (x_i, \hat{y}_i) be the corresponding point on the line.

The value of R^2 gives the fraction of the vertical variation from the mean explained by the line.

In other words,

$$R^2 = \frac{\text{Variation of points on line from } \bar{y}}{\text{Variation of original data points from } \bar{y}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Ex: For the data on marriage age, we see that $R^2 = 0.9725$. Interpret in the context of age of marriage.