11 Regression

- The Correlation Coefficient
- The Least-Squares Regression Line

The Correlation Coefficient

Introduction

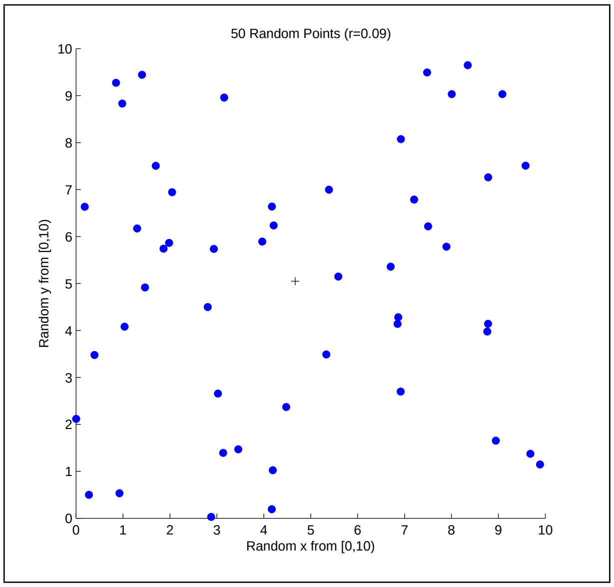
A *bivariate* data set consists of n ____________________________, $(x_1, y_1), \cdots, (x_n, y_n)$.

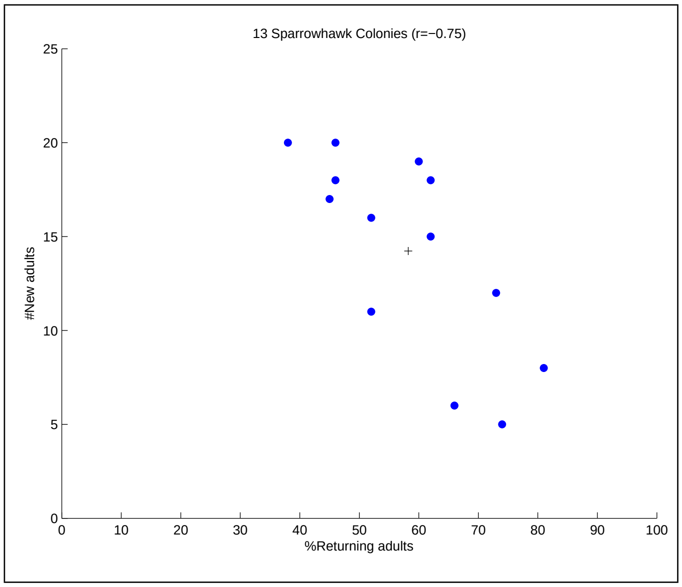
A *scatterplot* is a ______________________ of a bivariate data set.

e.g. Here are data for 13 sparrowhawk colonies relating the % of adult sparrowhawks in a colony that return from the previous year and the number of new adults that join the colony:

%Returning adults	74	66	81	52	73	62	52	45	62	46	60	46	38
#New adults	5	6	8	11	12	15	16	17	18	18	19	20	20

The right-hand scatterplot, below, is from these data. It shows $\cdots$





The Correlation Coefficient

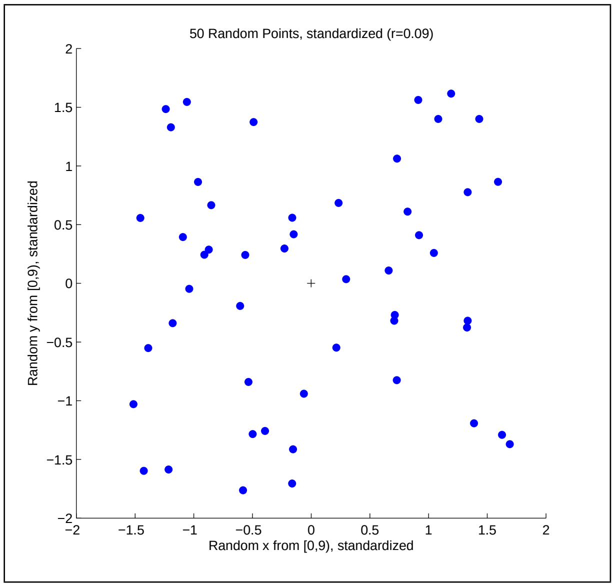
The *correlation coefficient*, r, measures the _____________ and _____________ of the linear relationship (if any) between x and y:

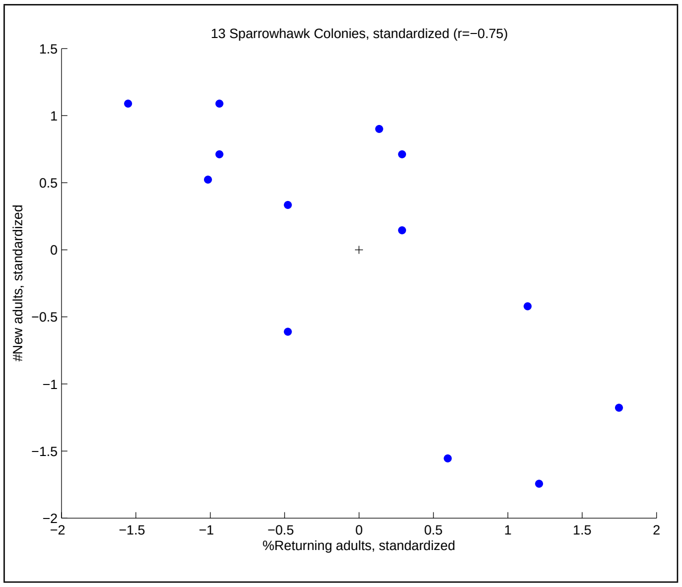
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

An Informal Explanation of r

- Start with a scatterplot.
- Shift reference point to _____________ by subtracting $\bar{x}$ from each x_i and $\bar{y}$ from each y_i.
- Rescale the x-axis by dividing each x coordinate by _______, and rescale the y-axis by dividing each y coordinate by s_y.

  Now x coordinates, $\frac{x_i-\bar{x}}{s_x}$, have mean _______ and standard deviation _______. y coordinates, $\frac{y_i-\bar{y}}{s_y}$, have the same mean and standard deviation.
- Analyze the sign of the i^{th} term in the sum above, $\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$, by quadrant:





e.g. For the sparrowhawk data, $r =$ _____________. For the random data, $r =$ _____________.

Properties of r

- $-1 \leq r \leq 1$, and

  $r = \pm 1 \implies$ data are _____________; $r \approx \pm 1 \implies$ data are __________________________

  $r \not\approx 0 \implies$ some linear relationship: x and y are *correlated*

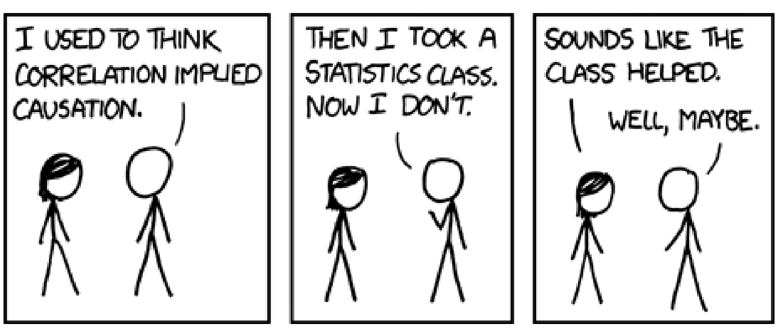
  $r > 0 \implies$ slope of line is _____________

  $r < 0 \implies$ slope of line is _____________

  $r \approx 0 \implies$ no linear relationship: x and y are ____________________

- r doesn't distinguish between _______ and _______

- r doesn't depend on _____________ or _____________



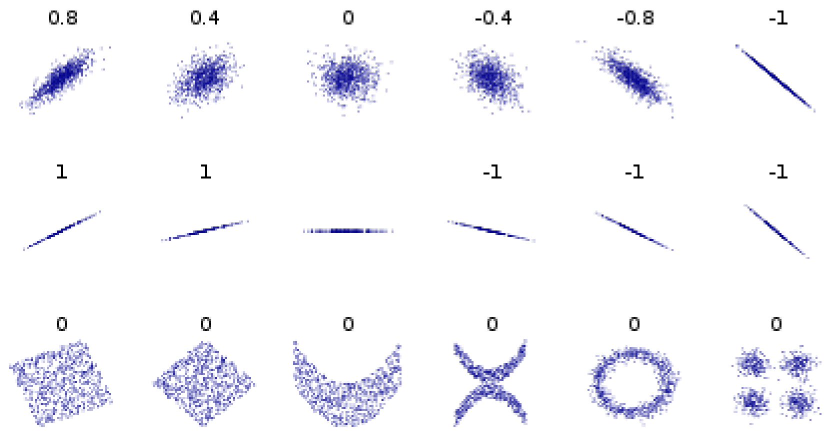
http://imgs.xkcd.com/comics/correlation.png

Cautions

- r measures strength of a *linear* relationship; check scatterplot to avoid using r for a ____________

  e.g. The data { (-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4) } fit ____________, but $r = 0$ because the data have no ____________ relationship (draw).

  e.g. (from `http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`)



- r is not resistant to the influence of ____________: don't use it for a data set with ____________

  e.g. Adding $(0, 0)$ to the sparrowhawk data changes r to ____________.

- Correlation does not imply causation:

  A __________________ (or *lurking*) *variable* is one ____________ under consideration that correlates with both the independent and dependent variables of interest.

  e.g.

  - Increasing ice cream sales are correlated with increasing ____________ rates. Does ice cream cause ____________? ______

    The confounding variable is _________________________.

  - Sleeping with shoes on is correlated with ______________________________________.
    Does sleeping with shoes on cause ____________? ______

    The confounding variable is _________________________.

  If either the independent variable under study, or a __________ confounding variable, affects the dependent variable, then both will seem to by the (___________) criterion of correlation.

  ____ cartoon

The Least-Squares Regression Line

A ______________ line is one that describes how a dependent variable, y, changes as an independent variable, x, changes in a data set $(x_1, y_1), \cdots, (x_n, y_n)$. We use it to predict y for a given x.

The *least-squares regression line* is the line that ___________________ the data (according to a reasonable criterion).

Notation includes:

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$: an unknown true (model) regression line,
  where β_0 is the y-intercept, β_1 is the slope,
  and ϵ_i is the ith random error

- $y = \hat{\beta}_0 + \hat{\beta}_1 x$: estimated regression line, where
  - x: _______________ variable
  - y: dependent variable
  - $\hat{\beta}_0$: estimated y-intercept
  - $\hat{\beta}_1$: estimated _____________

- (x_i, y_i): i^{th} data point

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: ___________________ value of y given $x = x_i$:

- $e_i = y_i - \hat{y}_i$: *residual*, the difference between observed y_i and predicted $\hat{y}_i$; estimates ϵ_i

We predict y from x, so minimize vertical error in the "least squares" sense by minimizing a "sum of squared errors"

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

(Alas, really it should be called a "sum of squared _____________.")   Ten lines of calculus gives:

For the data set $(x_1, y_1), \cdots, (x_n, y_n)$, the least-squares line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$, where

$$\hat{\beta}_1 = \frac{s_y}{s_x} r \text{ (slope)}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ (}y\text{-intercept)}$$

e.g. Here again are data for 13 sparrowhawk colonies relating the % of adults in a colony that return from the previous year and the number of new adults that join the colony:

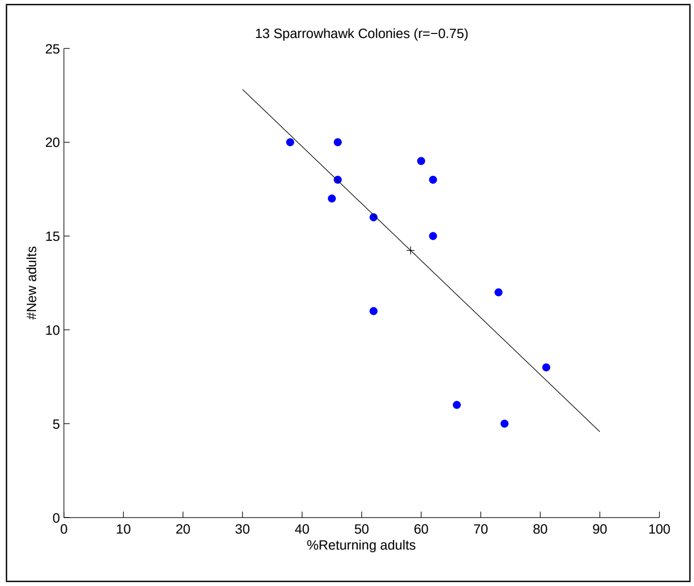
x = %Returning adults	74	66	81	52	73	62	52	45	62	46	60	46	38
y = #New adults	5	6	8	11	12	15	16	17	18	18	19	20	20

Use a calculator to find the least-squares line (recall slope $\hat{\beta}_1 = \frac{s_y}{s_x} r$, y-intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$):

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

$\Longrightarrow$

- $\hat{\beta}_1 =$
- $\hat{\beta}_0 =$



So our model is $y =$

Or we can do it more directly. (Figure out your _____________ labels.)

e.g. Predict the number of new adults in a colony to which 60% of last year's adults return.

$\hat{y} =$_____________

R code for correlation and regression

```
returning = c(74,66,81,52,73,62,52,45,62,46,60,46,38)
new = c( 5,6,8,11,12,15,16,17,18,18,19,20,20)
cor(x=returning, y=new)              # cor() gives correlation
model = lm(new ~ returning)          # lm(y ~ x) gives linear model
model
plot(x=returning, y=new, xlim=c(0, 85), ylim=c(0, 35)) # scatterplot
abline(model)                        # abline() adds line
summary(model)                       # test H_0: beta_i = 0
confint(model)                       # CIs for beta_i
```