PUBH 7405-Block 4

LAND ACKNOWLEDGEMENT

The School of Public Health at the University of Minnesota Twin Cities is built within the traditional homelands of the Dakota people. Minnesota comes from the Dakota name for this region, Mni Sóta Maḳoce, which loosely translates to the land where the waters reflect the skies.

It is important to acknowledge the peoples on whose land we live, learn, and work as we seek to improve and strengthen our relations with our tribal nations. We also acknowledge that words are not enough. We must ensure that our institution provides support, resources, and programs that increase access to all aspects of higher education for our American Indian students, staff, faculty, and community members.

Galton, Pearson, and the Peas: A Brief History of Linear Regression (and correlation)

Sir Francis Galton

- Worked on inherited characteristics of sweet peas
- Originally conceived modern notions of regression and correlation
- Cousin of Charles Darwin
- Was interested in how strongly characteristics of one generation of living things manifested in the following generation
- Chose sweet peas because this species could self fertilize
- Eliminated genetic contributions from multiple sources

Francis Galton

Francis Galton was an English explorer and anthropologist best known for his research in eugenics and human intelligence. He was the first to study the effects of human selective mating.

UPDATED: MAY 25, 2021

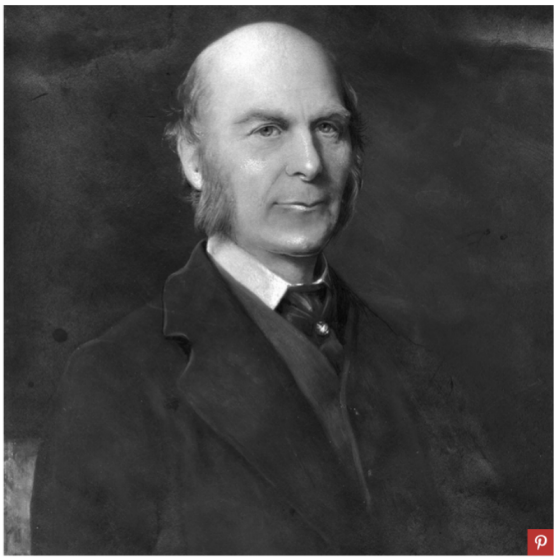
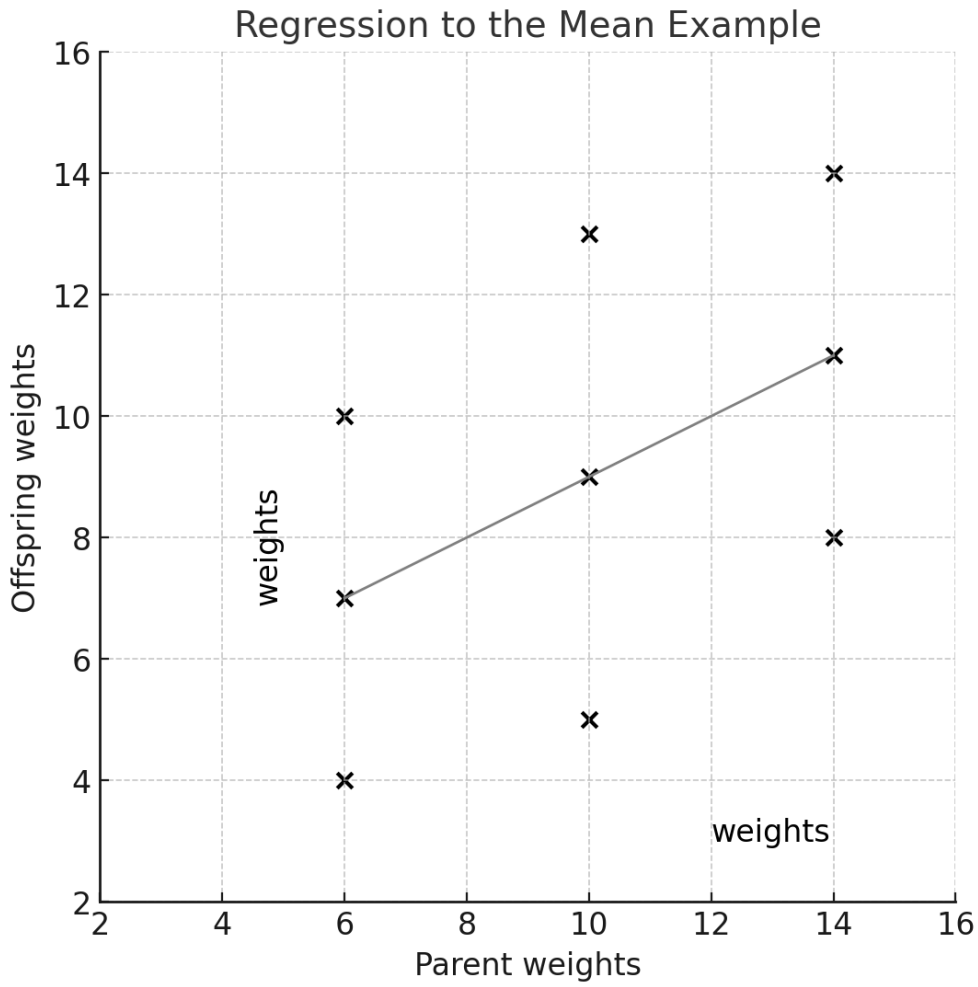


Photo: Hulton Archive/Getty Images

(1822-1911)

Source: biography.com



Explanation of the graph:

- Regression to the mean (Y values closer to $\bar{Y}$ than X values are to $\bar{X}$)
- Approx. equal variability in Y values at a given X value
- Median weights of the daughter seeds from a particular size of mother approx. described straight line with slope < 1.0
- Made up data

$\Rightarrow$ because slope < 1.0, Galton concluded there was a regression to the mean for that generation of peas.

Above figure shows line connecting the means of the columns of data points, which indicates the degree of regression to the mean.

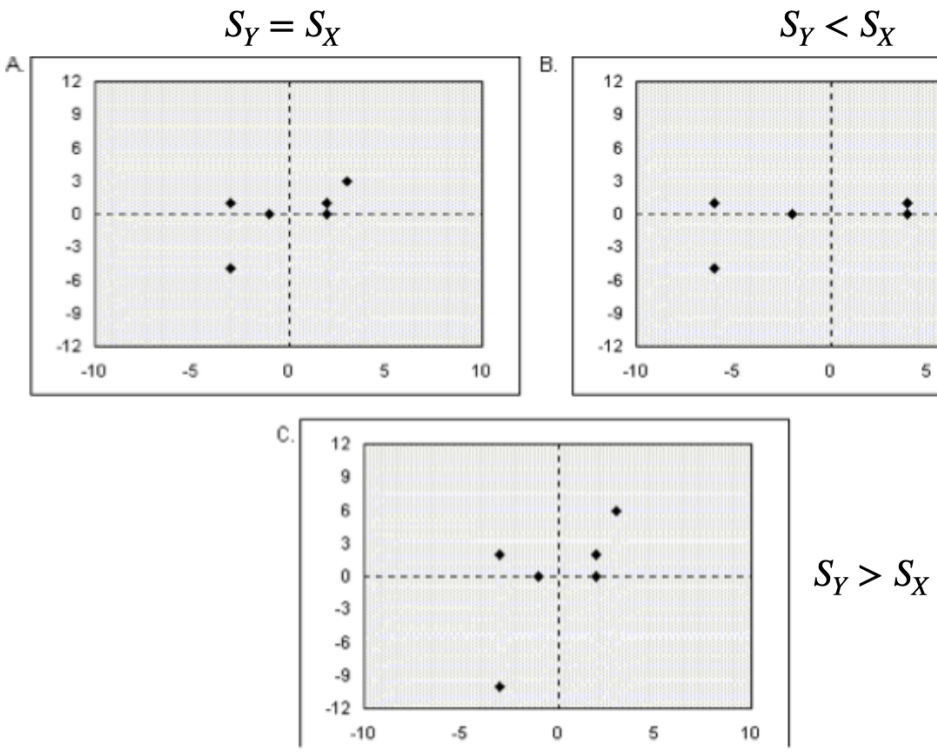
Galton's Recognition of the Generality of Regression Slope

Galton generalized his work to a variety of heredity problems:

- temperament
- artistic ability
- disease incidence

His important breakthrough:

If degree of association between two variables held constant, then slope of the regression line could be described if variability of two variables were known.



($r = 0.64$ for all figures)

Slope of line ↑ as $S_Y > S_X$.

Galton had recognized the rudimentary regression equation:

$$y = r\left(\frac{S_Y}{S_X}\right)x$$

Pearson's Mathematical Development of Correlation and Regression

In 1896, Pearson published the first rigorous treatment of correlation and regression.

Pearson demonstrated that optimum values $\sum_i (y_i - \hat{y}_i)^2$ (smallest) for regression slope and correlation coefficient could be calculated using product-moment:

$$\frac{\sum_i X_i Y_i}{n}$$

Here x_i and y_i are actually deviations from their respective means.

Table 1. Numeric Example Demonstrating that Adding or Subtracting an Offset from the Product-Moment Worsens the Prediction of *Y* from *X*.

	Deviation Data		Product	No Offset (Optimal)		Positive Offset		Negative Offset	
	x	y	x*y	2.0x[b]	Squared Errors[c]	2.1x	Squared Errors	1.9x	Squared Errors
	-1	-3	3	-2	1	-2.1	0.81	-1.9	1.21
	-1	-2	2	-2	0	-2.1	0.01	-1.9	0.01
	-0.5	-1	0.5	-1	0	-1.05	0.025	-0.95	0.025
	-0.5	-1	0.5	-1	0	-1.05	0.025	-0.95	0.025
	0.5	1	0.5	1	0	1.05	0.025	0.95	0.025
	0.5	3	1.5	1	4	1.05	3.8025	0.95	4.2025
	2	3	6	4	1	4.2	1.44	3.8	0.64
Sum	0	0	14		6		6.1375		6.1375
Mean	0	0	2.0[a]						

Sample Covariance

Given n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, *sample covariance* s_{xy} is a measure of the *direction* and *strength* of the linear relationship between X and Y, defined as

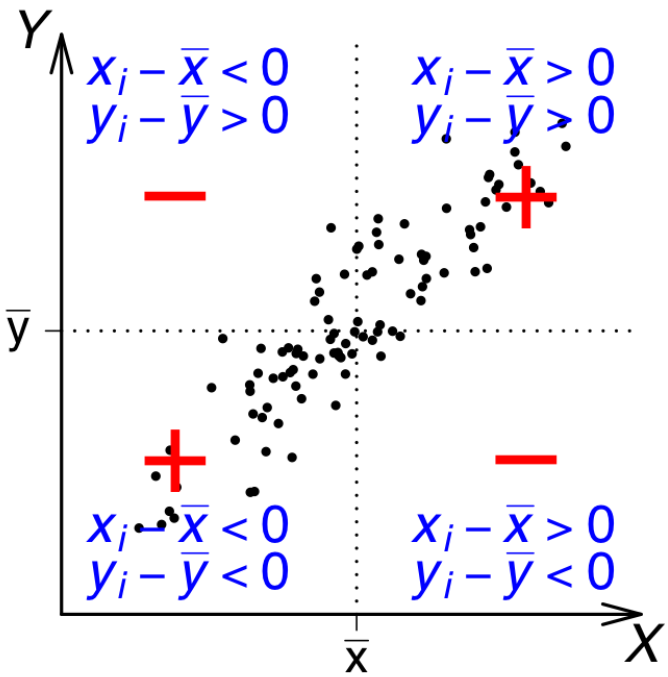
$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- $s_{xy} > 0$: Positive linear relation;
- $s_{xy} < 0$: Negative linear relation
- The *magnitude* of covariance reflects the *strength* of the relation
- The covariance of a variable X with itself is its *sample variance*

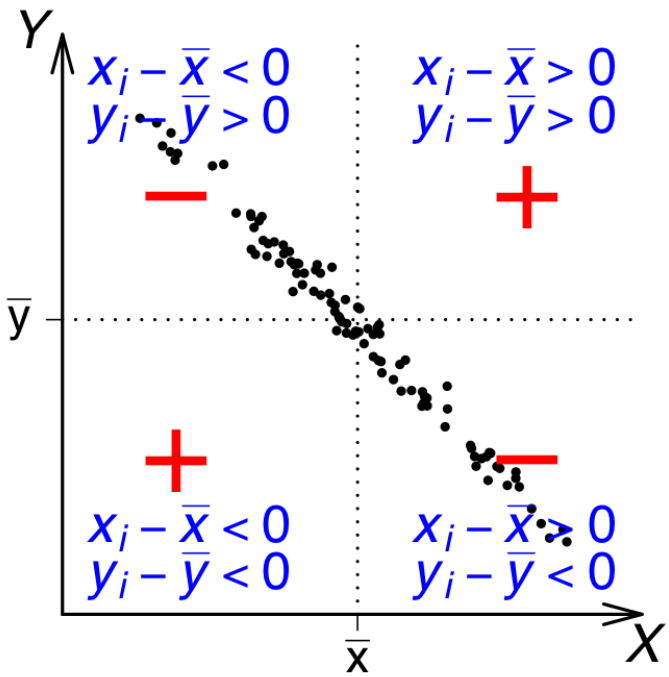
$$s_{xx} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s_x^2$$

Sample Covariance Reflects the Direction of a Linear Relation

What is the sign of $(x_i - \bar{x})(y_i - \bar{y})$?

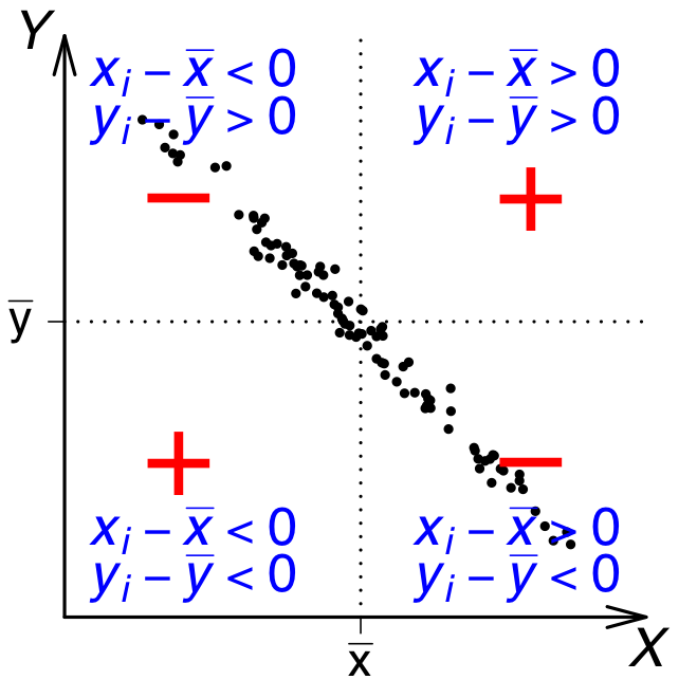


Cov > 0 as most points have $(x_i - \bar{x})(y_i - \bar{y}) > 0$

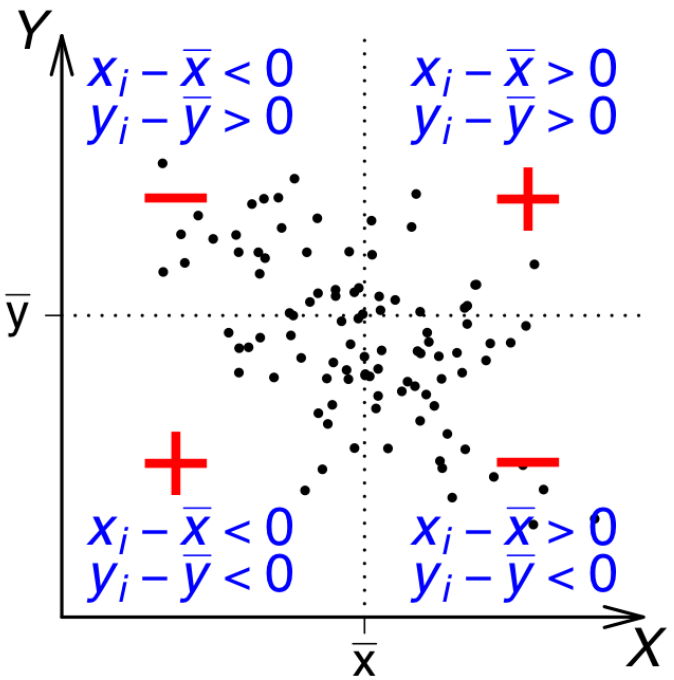


Cov < 0 as most points have $(x_i - \bar{x})(y_i - \bar{y}) < 0$

Sample Covariance Reflects the Strength of a Linear Relation



Cov Has a Larger Magnitude



Cov Has a Smaller Magnitude

Covariance is of a smaller magnitude in the right plot than in the left because the $(x_i - \bar{x})(y_i - \bar{y})$ of most points in the left plot are of the different signs and get cancelled out when adding up.

How Large the Covariance is Large Enough?

It can be shown in the next slide that

$$|s_{xy}| \leq s_x s_y = (\text{SD of } X) \times (\text{SD of } Y)$$

Moreover, the sample covariance reaches its maximum possible magnitude if and only if all the points (x_i, y_i) fall on a straight line.

Thus, one can determine whether a linear relation is strong by comparing the Cov with the product of the SDs of the two variables.

Proof of $|s_{xy}| \leq s_x s_y$

For any two sequences $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$, the Cauchy Schwartz Inequality below is always true

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

Moreover, the inequality becomes an equality if and only if

$$\alpha a_i + \beta b_i = 0 \quad \text{for all } i \text{ for some non-zero constants } \alpha \text{ and } \beta.$$

Applying Cauchy Schwartz Inequality with $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$, we get

$$\underbrace{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right)^2}_{[(n-1)s_{xy}]^2} \leq \underbrace{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)}_{(n-1)s_x^2}\underbrace{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)}_{(n-1)s_y^2}.$$

Dividing both sides by $(n-1)^2$, and taking square-root, we get

$$|s_{xy}| \leq s_x s_y.$$

Proof of $|s_{xy}| \leq s_x s_y$ (Cont'd)

Moreover, recall the the inequality becomes an equality if and only if

$$\alpha a_i + \beta b_i = 0 \quad \text{for all } i \text{ for some nonzero constants } \alpha \text{ and } \beta.$$

Now with $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$, we get that $|s_{xy}|$ reach its max $s_x s_y$ if and only if

$\alpha(x_i - \bar{x}) + \beta(y_i - \bar{y}) = 0$ for all i for some nonzero constants α and β,

or equivalently all the points (x_i, y_i) fall on the straight line

$$\alpha x_i + \beta y_i = \alpha \bar{x} + \beta \bar{y}$$

Shortcut Formula for the Sample Covariance

There are various formula for computing the sample covariance:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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

The last one is the *shortcut formula* for calculating the *sample covariance*, similar to the shortcut formula for the sample variance

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}$$

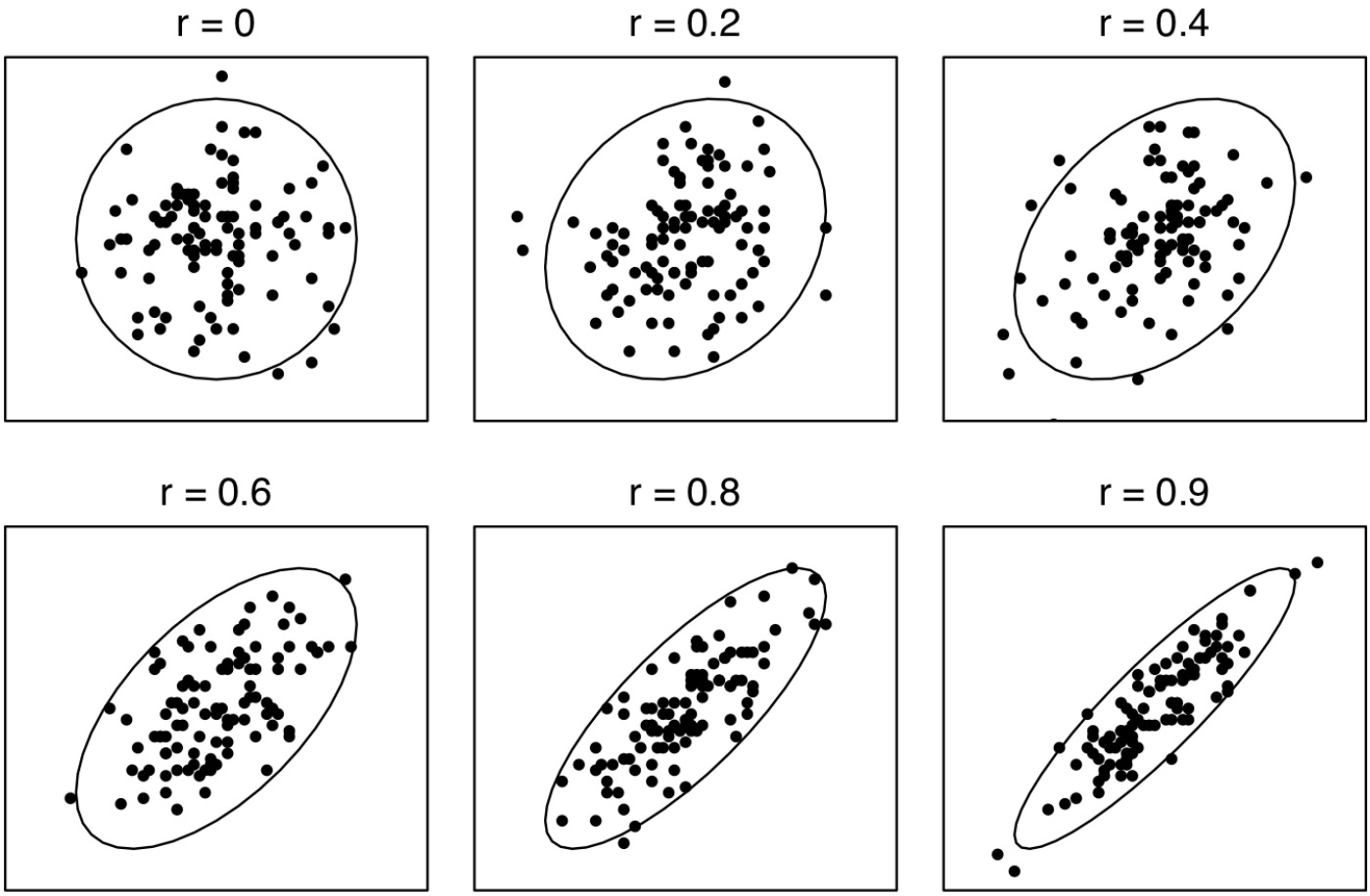
Sample Correlation = Correlation Coefficient r

Given n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the (sample) **corelation** is defined to be

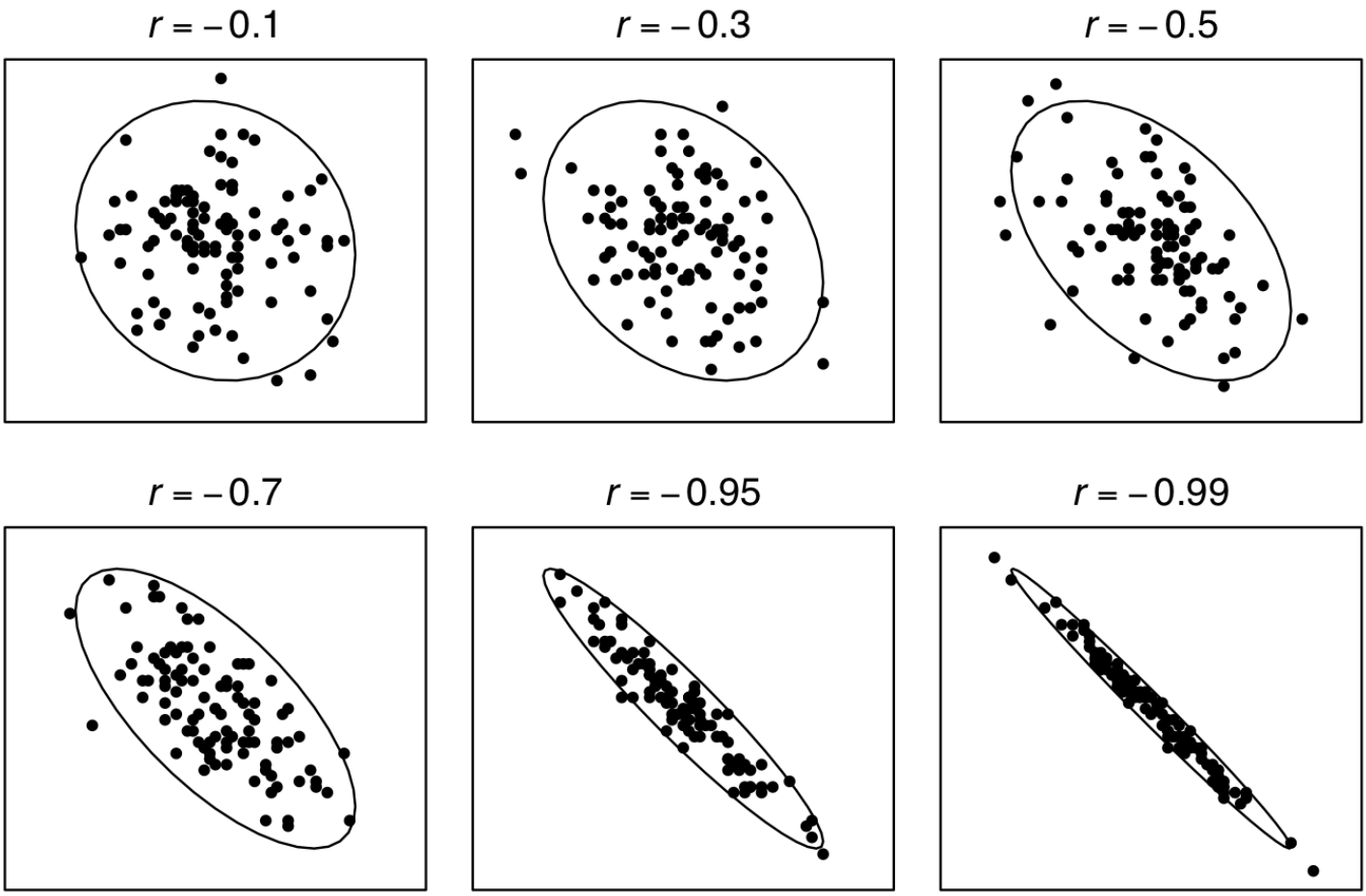
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $-1 \leq r \leq 1$ since $|s_{xy}| \leq s_x s_y$
- The closer r is to 1 or -1, the stronger the linear relation
- $r = 1$ or -1 if and only if all the points (x_i, y_i) fall on a straight line

Positive Correlations



Negative Correlations



Sample Correlation r v.s. Population Correlation ρ

Recall in Lecture 11 we introduced the *correlation* between two random variables X, Y,

$$\rho = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\mathrm{E}[(X - \mu_X)^2]\,\mathrm{E}[(Y - \mu_Y)^2]}}.$$

The sample correlation r

$$r_{xy} = r = \widehat{\rho} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} = \frac{s_{xy}}{s_x s_y},$$

is an estimate for the population correlation ρ if $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are i.i.d. pairs of observations from the joint distribution of (X, Y).

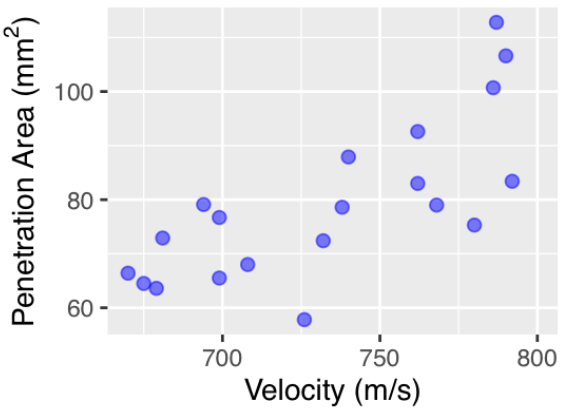
Covariance & Correlation Do Not Distinguish Between X & Y

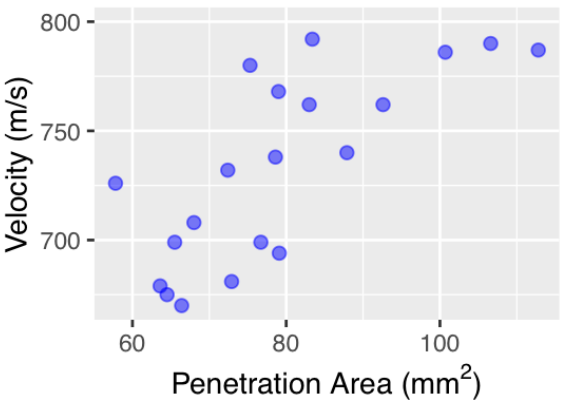
When one uses X to predict Y, X is called the *explanatory variable*, and Y the *response*. Covariance and correlation do not distinguish between X & Y. They treat X and Y symmetrically.

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = s_{yx};$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{yx}}{s_x s_y} = r_{yx}$$

Swapping the x-, y-axes doesn't change r (both $r \approx 0.74$.)





Scaling Property of Sample Covariance

$$
\begin{array}{cc}
(X,\ Y) & \longrightarrow \quad (aX+b,\ cY+d) \\
\hline
(x_1,\ y_1) & \qquad\quad (ax_1+b,\ cy_1+d) \\
(x_2,\ y_2) & \qquad\quad (ax_2+b,\ cy_2+d) \\
(x_3,\ y_3) & \Rightarrow \quad (ax_3+b,\ cy_3+d) \\
\vdots & \qquad\quad \vdots \\
(x_n,\ y_n) & \qquad\quad (ax_n+b,\ cy_n+d)
\end{array}
$$

The sample covariance has the scaling property:

$$
\begin{aligned}
S_{aX+b,cY+d} &= \frac{1}{n-1}\sum_{i=1}^{n}[ax_i+b-(a\bar{x}+b)][cy_i+d-(c\bar{y}+d)] \\
&= \frac{1}{n-1}\sum_{i=1}^{n} ac(x_i-\bar{x})(y_i-\bar{y}) \\
&= ac\,S_{XY}.
\end{aligned}
$$

Correlation is Scale Invariant

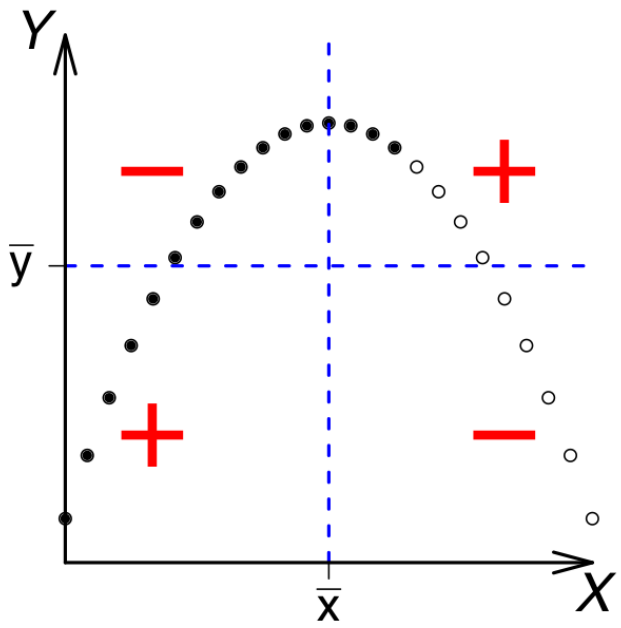
The sample correlation is *scaling invariant* and *has no units*!

$$r_{aX+b,cY+d} = \frac{S_{aX+b,cY+d}}{S_{aX+b}S_{cY+d}} = \frac{ac\,S_{XY}}{|a|S_X\,|c|S_Y} = (\text{sign of } ac) \times \frac{s_{XY}}{s_X s_Y}$$

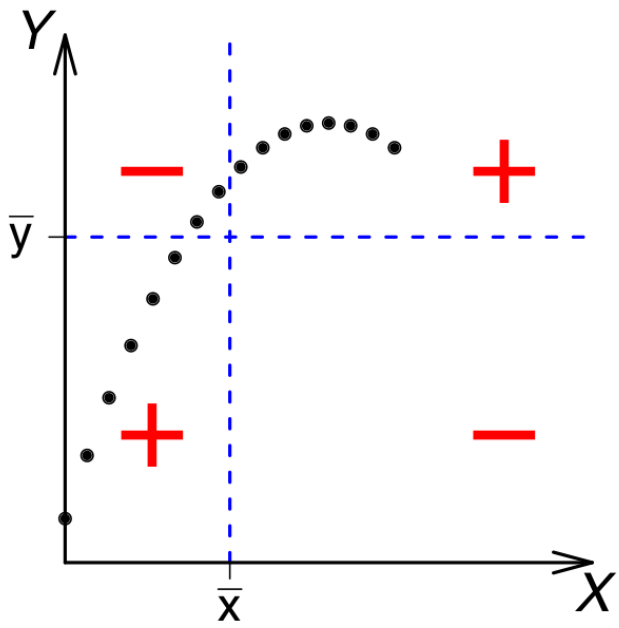
$$= (\text{sign of } ac) \times r_{XY}.$$

Correlation Doesn't Reflect Strength of Nonlinear Relations

Both scatter plots below show perfect nonlinear relations. All points fall on the quadratic curve $y = 2 - x^2/2$.



r = 0 (why?)
(black + white dots)



r = 0.91
(black dots only)

Testing Population ρ

Suppose interested in testing the following hypotheses:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Fact:

$$t(X,Y) = \frac{r(XY)\sqrt{n-2}}{\sqrt{1-r(XY)^2}} \quad \overset{H_0}{\sim} t_{n-2}$$

Reject H_0 at level α if:

$$|t_{\text{obs}}| > t_{1-\alpha/2,n-2}$$

(t_{obs}: observed value from sample)

A one-tailed test for a positive correlation between X and Y tests:

$$H_0 : \text{when } X \uparrow \text{ does } Y \uparrow \text{ in the population?}$$

Confidence Intervals for ρ

We use Fisher's Z transform:

$$Z = \frac{1}{2}\log_e\left(\frac{1+r(XY)}{1-r(XY)}\right)$$

For large n, this is distributed under H_0 as:

$$N(0, \left(\frac{1}{\sqrt{n-3}}\right)^2)$$

$\therefore$ First, find $100 \times (1-\alpha)\%$ CI based on Z and then transform back to the $r(XY)$ scale.

How?

$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

$$2 \cdot Z = \log_e \left(\frac{1+r}{1-r} \right)$$

$$\exp(2Z) = \frac{1+r}{1-r}$$

$$\exp(2Z)(1-r) = 1+r$$

$$\exp(2Z) - r\exp(2Z) = 1+r$$

$$\exp(2Z) - 1 = r(\exp(2Z)+1)$$

$$\Rightarrow r = \frac{\exp(2Z)-1}{\exp(2Z)+1}$$

Example

Pearson's sample r for a study investigating the association of basal metabolic rate with total energy expenditure was estimated as 0.7283 in $n = 13$ women. Derive a 95% CI for ρ.

$$Z = \frac{1}{2}\log_e \left(\frac{1+0.7283}{1-0.7283} \right) = 0.9251$$

Lower limit of 95% CI:

$$0.9251 - 1.96 \left(\frac{1}{\sqrt{13-3}} \right) = 0.3053$$

Upper limit of 95% CI:

$$0.9251 + 1.96 \left(\frac{1}{\sqrt{13-3}} \right) = 1.545$$

Now transform back to the r scale:

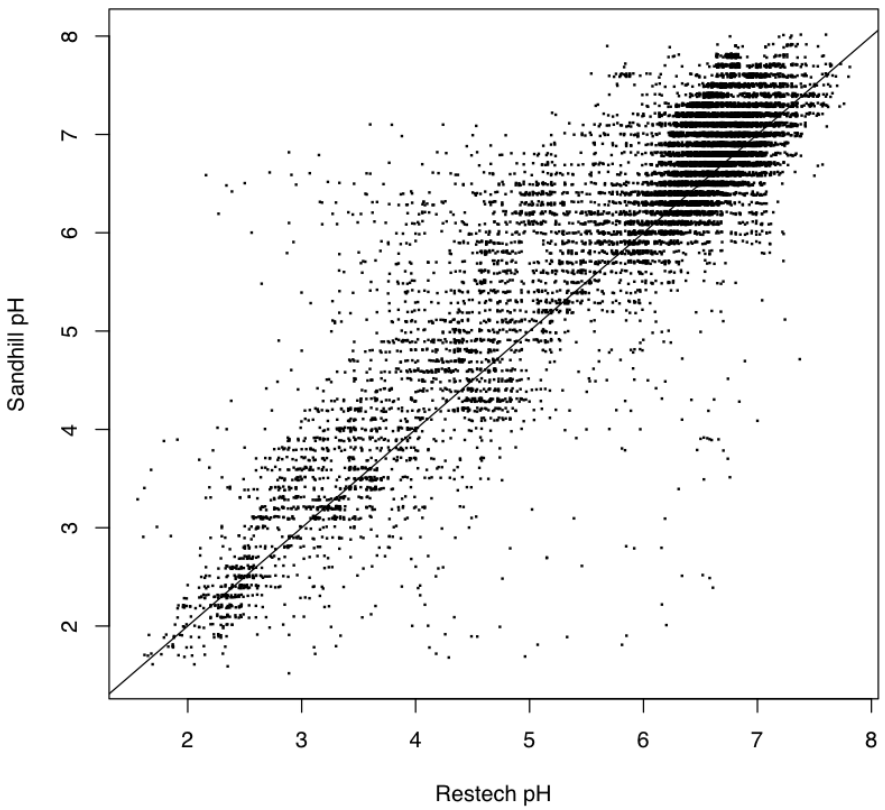
$\therefore$ the 95% confidence interval for ρ is:

$$\left(\frac{\exp(2 \times 0.3053) - 1}{\exp(2 \times 0.3053) + 1}, \frac{\exp(2 \times 1.545) - 1}{\exp(2 \times 1.545) + 1} \right)$$

$$= (0.30, 0.91)$$

Correlation and Agreement

- Compare two methods of measuring the same underlying value
  - Lung function measured using a spirometer (expensive, accurate) or peak flow meter (cheap, less accurate)
  - Two devices (Restech and Sandhill) used to mesured acidity (pH) in the esophagus
- Typical (incorrect) approach begins with scatterplot of Restech versus Sandhill with a 1:1 line indicating perfect agreement



- Problems with the correlation approach

1. r measures the degree of linear association between two variables, not the agreement. If,for example, the Sandhill consistently gave pH values that were 0.5 unit higher than the Restech, we could still have high correlation, but poor agreement between the two devices. We can have high correlation if the two devices lie closely to any line, not just a 1:1 line that indicates perfect agreement.
2. A change in scale does not affect correlation, but does influence agreement. For example, if the Sandhill always registered 2 times larger than the Restech, we would have perfect correlation but the agreement would get progressively worse for larger values of pH.
3. Correlation depends on the range of the data so that larger ranges lead to larger correlations. This can lead to vary strange interpretations

	r	ρ
all data	0.90	0.73
avg pH ≤ 4	0.51	0.58
avg pH > 4	0.74	0.65

4. Tests of significance (testing if $\rho = 0$) are irrelevant to the question at hand, but often reported to demonstrate a significant association. The two devices are measuring the same quantity, so it would be shocking if we did not observe a highly significant p-value. A $p < .0001$ is not impressive. A regression analysis with a highly significant slope would be similarly unimpressive.
5. Data can have high correlation, but poor agreement. There are many examples in the literature, but even in our analysis with $r = 0.90$, the correlation is high, but we will show that the agreement is not as good as the high correlation implies.

Using r to Compute Sample Size

- Without knowledge of population variances, etc., r can be useful for planning studies

- Choose n so that margin for error (half-width of C.L.) for r is acceptable

- Precision of r in estimating ρ is generally worst when $\rho = 0$

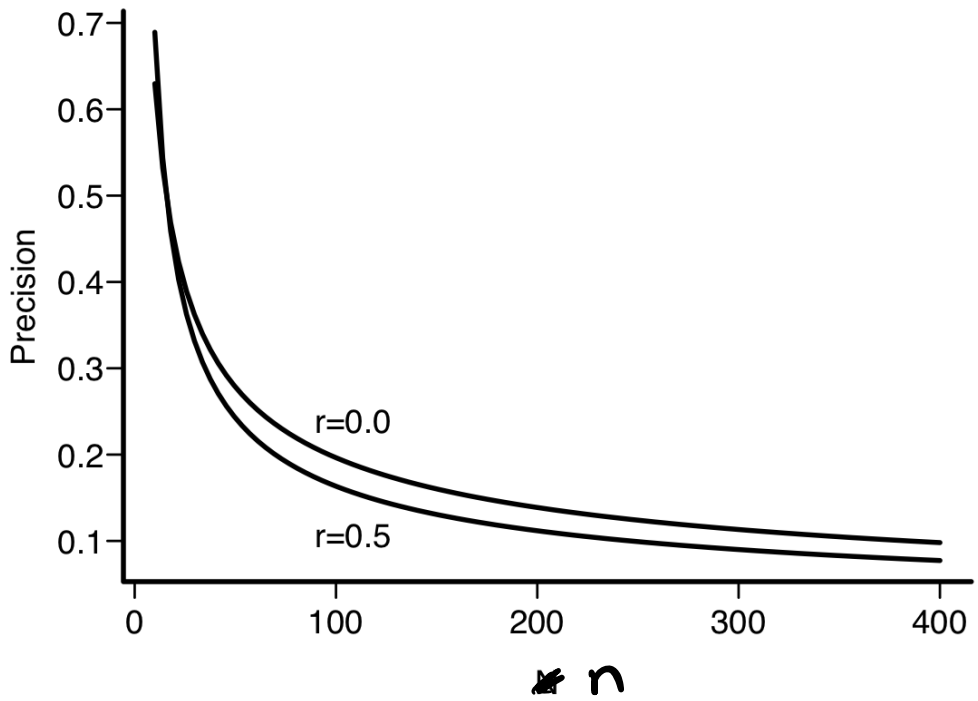


Figure 6.6: *Margin for error (length of longer side of asymmetric 0.95 confidence interval) for r ir estimating ρ, when $\rho = 0$ and $\rho = 0.5$. Calculations are based on Fisher's z transformation of r.*

Comparing Two r's

- Rarely appropriate
- Two r's can be the same even though slopes may differ
- Usually better to compare effects on a real scale (slopes)