Correlation and Regression

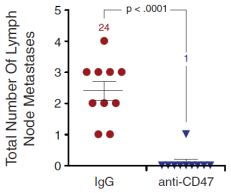
PSLS chapters 3 and 4

Part II: issues and examples (flipped lesson)

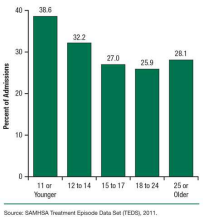
Copyright Brigitte Baldi 2019 ©

Studying patterns of association

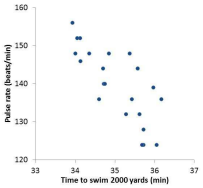
Between a quantitative variable and a categorical variable



Between two categorical variables



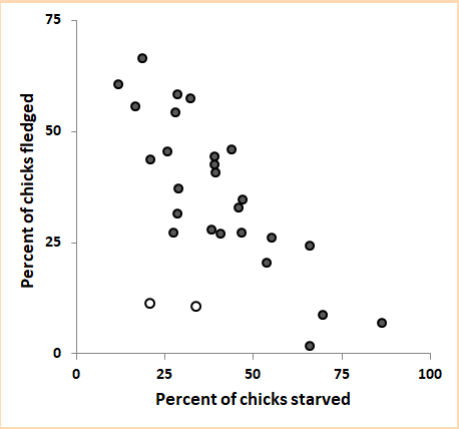
Between two quantitative variables



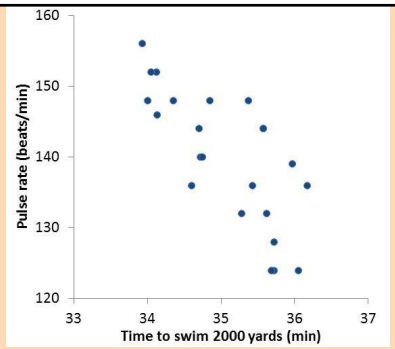
→ If the pattern is linear, we can analyze the data with correlation and regression

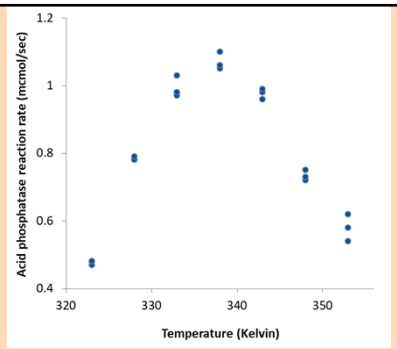
Long-term study of Magellanic penguins at Punta Tombo, Argentina (1983–2010)

Climate Change Increases Reproductive Failure in Magellanic Penguins (2014)
doi:10.1371/journal.pone.0085602.g003



Punta Tombo is arid with low annual precipitation. The 2 open circles represent 1991 and 1999, when rain killed over 40% of chicks each year, and were not included in the regression.





If we computed the **correlation *r*** for these 2 graphs, which value would be closest to zero?

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

A) The one on the left.
B) The one on the right.
C) They would be fairly similar.
D) We can't tell from the graphs.

The **linear correlation coefficient** is a meaningful measure of the direction and strength of an association *only* when the association is **linear**.

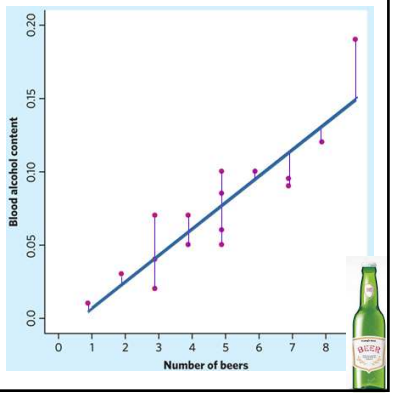
The least-squares regression line

The **least-squares regression line** is the unique line such that the sum of the squared vertical distances between the data points and the line is the smallest possible. We use this line to "**model**" the behavior of y as a linear function of x.

sample data = model + residuals

residual = actual – predicted
$= y - \hat{y}$

model:
$y = \text{constant} + \text{slope}.x$

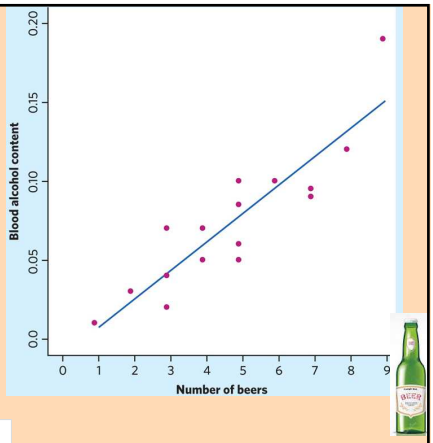


The **constant (y intercept)** and the **slope** are the two **coefficients** of the regression line.

2 different software outputs:

Predictor	Coef	SE Coef	T	P
Constant	-0.01270	0.01264	-1.00	0.332
Beers	0.017964	0.002402	7.48	0.000

S = 0.0204410 R-Sq = 80.0% R-Sq(adj) = 78.6%



SUMMARY OUTPUT		
Regression Statistics		
Multiple R	0.894338148	← *r*
R Square	0.799840723	
Adjusted R Square	0.785543632	
Standard Error	0.020440951	
Observations	16	

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.012700604	0.012637502	-1.004993204	0.331955132
Number of Beers	0.017963762	0.002401703	7.479592058	2.96948E-06

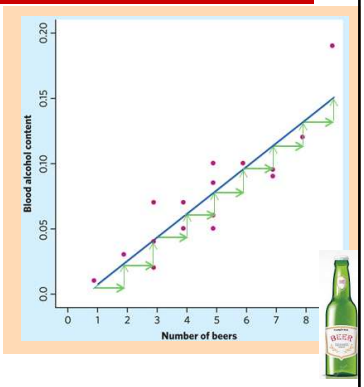
The value of the **slope** is:

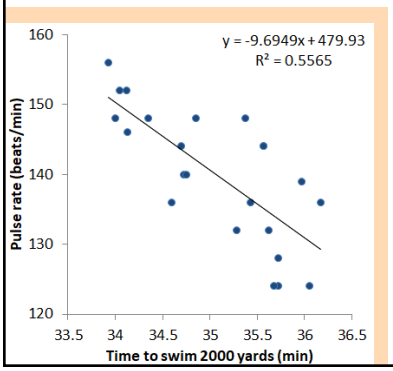
A) 0.01796
B) –0.0127
C) 0.8943
D) 0.0204

What does it mean in context?

Slope and intercept of the regression line

The **slope** of the regression line describes how much we expect *y* to change, **on average**, for every unit change in *x*.





The **intercept** is a necessary mathematical descriptor of the regression line. It does not describe a specific property of the data.

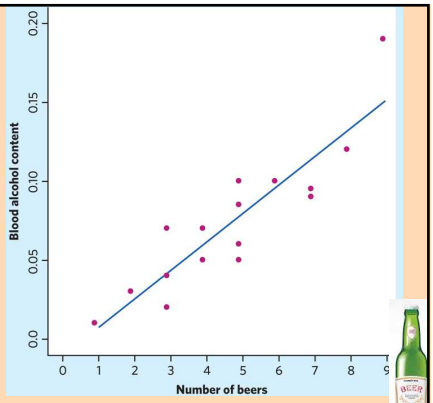
y or $\hat{y} = 0.01796x - 0.013$

$\text{BAC} = 0.01796\text{Beers} - 0.013$

Results - Simple Linear Regression

Export

Fitted Equation:	BAC = -0.01270 + 0.01796 * Beers			
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.01270	0.01264	-1.005	0.3320
Beers	0.01796	0.002402	7.480	<0.0001
estimated sigma:	0.02044			



What does the slope mean in context?

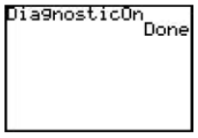
TI calculator : linear regression / correlation

First, you need to set up the regression function in your calculator.

This must be done ONCE only. No need to repeat next time.

In order to compute the correlation coefficient r between paired data of quantitative variables, we first must make sure that the calculator's diagnostics are turned on. To turn on the setting, press [CATALOG] (i.e., [2nd] 0) and scroll down to the **DiagnosticOn** command. Press [ENTER] to bring the command to the Home screen, then press [ENTER] again.

[CATALOG] ([2nd] 0).

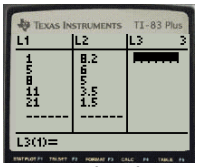


Press [ENTER].

Now if paired data is entered into lists, then we can find the correlation with the **LinReg(ax+b)** or **LinReg(a+bx)** commands from the [STAT] CALC screen.

TI calculator : linear regression / correlation

see flipped video

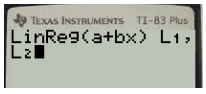


Enter the data into 2 lists

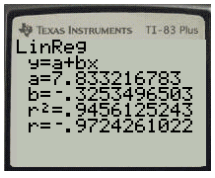
STAT / **CALC** then

Use **LinReg(a+bx)**

then enter: **L1, L2**



Select the list for x values first, then the list for y values



→ here, a is the intercept and b is the slope

→ r is the linear correlation coefficient

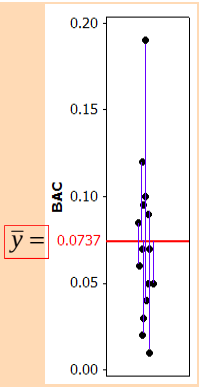
$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{SSTO}{n-1}$$

residuals = $y_i - \hat{y}$

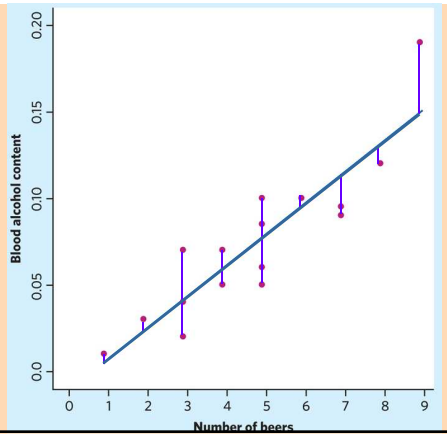
$$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

$$r^2 = \frac{SSTO - SSE}{SSTO}$$

The least-squares regression line is by definition the line with the smallest sum of squared residuals.



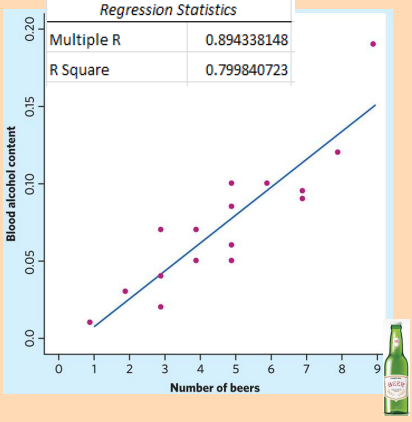
Variable	N	Mean	Variance
BAC (y)	16	0.0737	0.0019

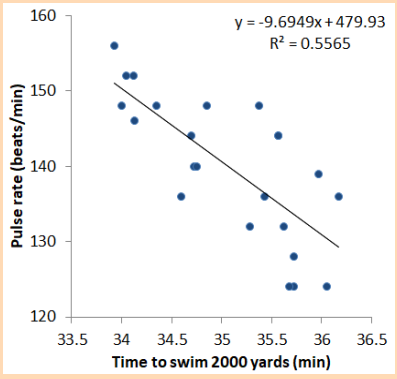


r represents the direction and strength a of a linear relationship

r^2 indicates what fraction of the variation in *y* can be explained by the linear regression model

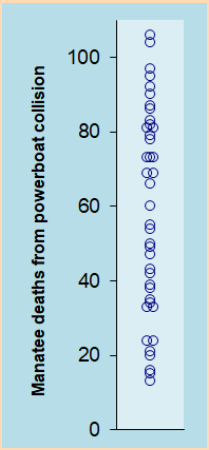
SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.894338148
R Square	0.799840723

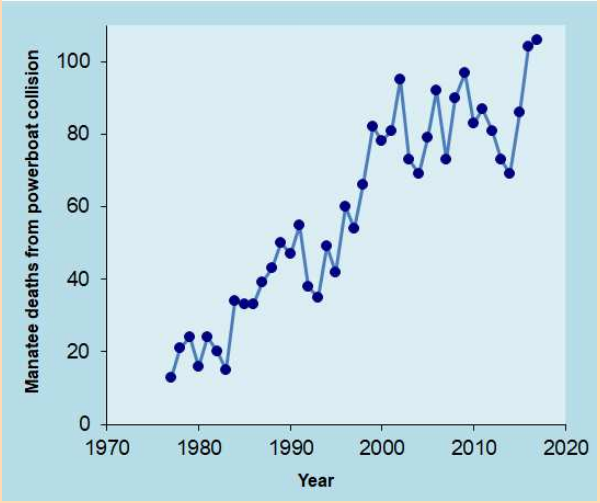




Manatee population data for Florida, 1977–2017

http://myfwc.com/research/manatee/rescue-mortality-response/mortality-statistics/

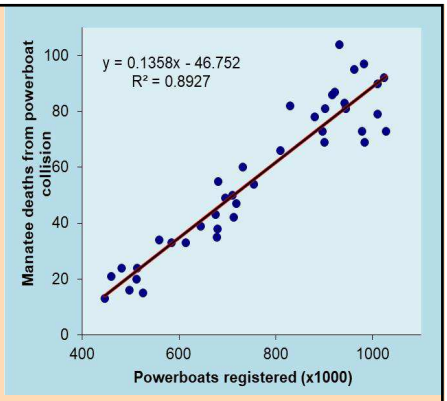




Florida, 1977–2016

Positive linear relationship

$\hat{y} = 0.1358x - 46.8$



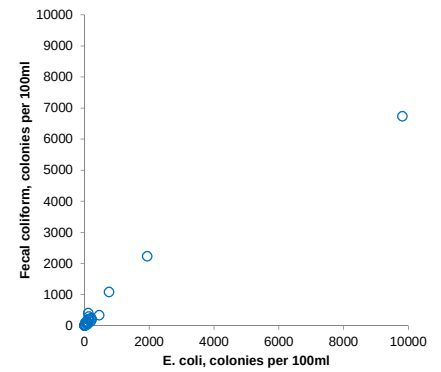
If Florida were to limit the number of powerboats to 500,000, what could we expect the number of manatee deaths to be in that year?

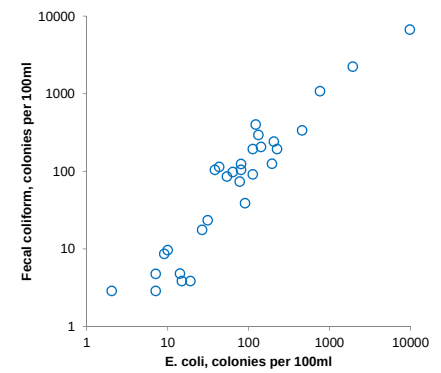
A) ~21 **B) ~ 65** **C) ~109** **D) ~65,006**

What if Florida were to limit the number of powerboats to 200,000?

Illinois River between Hennepin and Peoria, Illinois: 2007–08

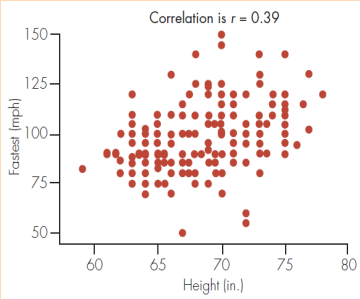
pubs.usgs.gov/of/2012/1075/pdf/OFR-Dupre-050212.pdf



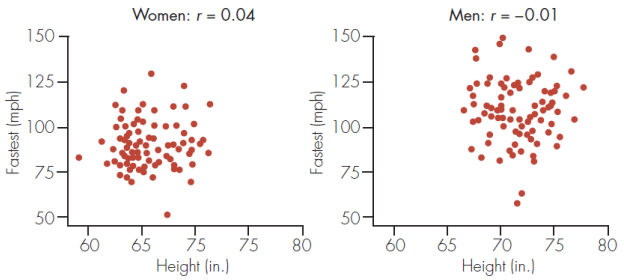


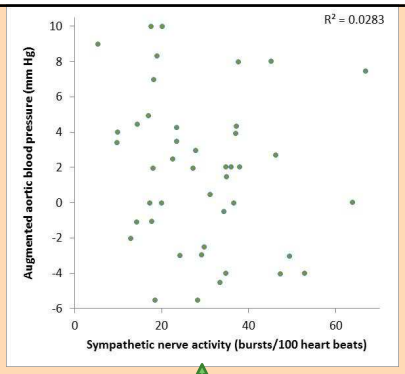
Log transformations are common in biology

College student heights and responses to the question “What is the fastest you have ever driven a car?” Looks like a mild positive linear relationship.

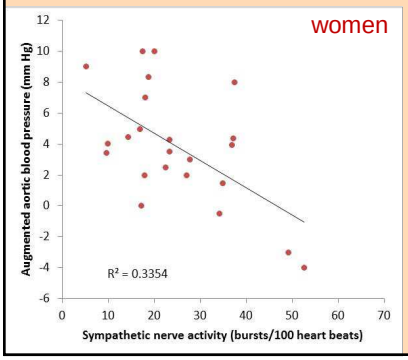


Same data broken down by sex.

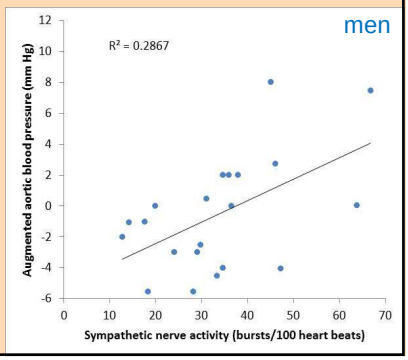




Relationship between muscle sympathetic nerve activity and a measure of arterial stiffness in young adults.



Same data broken down by sex.



Researchers examined the relationship between breastfeeding and infant mortality in the USA, based on state data for 2010.

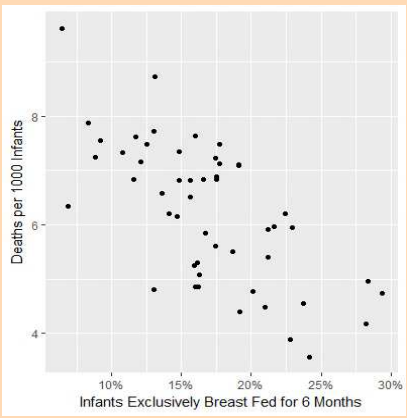
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.08191	0.46358	19.591	< 2e-16
breastfeeding	-0.16884	0.02633	-6.413	5.38e-08

R-squared: 0.4563

Is it appropriate to conclude that, in the U.S., breast-feeding infants until they are 6-months old reduces infant mortality?

A. Yes, because of the negative association between death rate and breastfeeding rate.
B. No, because the association is not strong (r^2 is only 46%).
C. No, because the association was observed and we cannot rule out the role of confounding variables.



Association does not imply causation

Association, however strong, does NOT necessarily imply causation. An observed association could have an external cause (a confounding variable) or be a coincidence.

Establishing causation from an observed association can be done if, in a variety of statistical studies:

1) The association is strong.
2) The association is consistent.
3) Higher doses are associated with stronger responses.
4) The alleged cause precedes the effect.
5) The alleged cause is plausible.

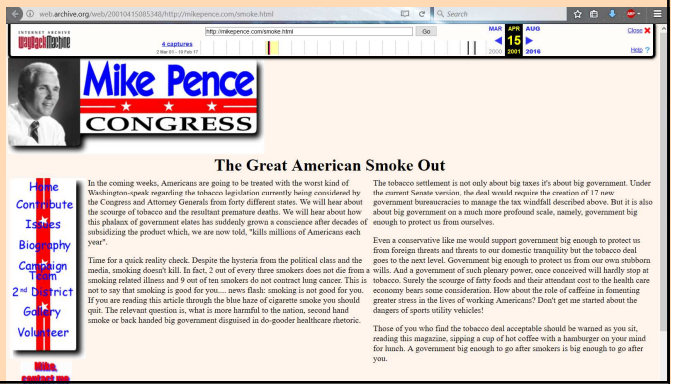
http://web.archive.org/web/20010415085348/http://mikepence.com/smoke.html

Mike Pence wrote an editorial in 2001 titled “The Great American Smoke Out”:

…

“Time for a quick reality check. Despite the hysteria from the political class and the media, smoking doesn't kill. In fact, 2 out of every three smokers does not die from a smoking related illness and 9 out of ten smokers do not contract lung cancer.”

…



Observed associations with an established conclusion of causality

- Smoking cause of lung cancer, heart disease, etc.
- Second-hand smoking cause of lung cancer, heart disease, etc.
- Man-made activity source of increased lead pollution and cause of neurodevelopmental damage
- Zika virus infection during pregnancy and microcephaly in newborn (WHO declaration, 2016)
  who.int/emergencies/zika-virus/situation-report/31-march-2016/en/

Observed associations with a causal component still hotly argued

- Consumption of added sugar and obesity / metabolic syndrome
- Man-made activity and global climate change
- Concussions and depression / CTE (chronic traumatic encephalopathy)
  www.nytimes.com/2016/03/25/sports/football/nfl-concussion-research-tobacco.html

Completely debunked causal association

- Vaccines do NOT cause autism – fraudulent study www.bmj.com/content/342/bmj.c5347.full
- Spicy food and stress do NOT cause gastric ulcers – it's mostly *H. pilori*