Applied Statistics: Correlation & Linear Regression

—

4/11/2022

Plan for Today

- What's Upcoming
- Homework
- Graph Discussion: Scatterplot
- Correlation and Regression Team Activity

What's upcoming

- **Checkpoint 3 –** Open through tonight.
- Daily Survey 31 – Due Tuesday.
- Reading Assignment 4 Writeup – due Wednesday, April 13.
- No class on Monday!
- Week 12 Homework – Due next **Tuesday**, April 19.
- Project 2 – Due next Wednesday, April 20.

Homework – Ch. 9 #102

Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

Homework – Ch. 9 #102

Registered nurses earned an average annual salary of \$69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than \$69,110 for California nurses. The sample average was \$71,121 with a sample standard deviation of \$7,489. Conduct a hypothesis test.

Null hypothesis $H_0: \mu = \$69{,}110$

Alternative hypothesis $H_a: \mu > \$69{,}110$

Homework – Ch. 9 #102

Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

Sample mean $\bar{x}$: $71,121

Sample standard deviation s: $7,489

Sample size n: 41

One group test

Homework – Ch. 9 #102

Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

Conclusion:

The p-value is 0.047. Using a significance level of $\alpha = 0.05$, we **reject** the null hypothesis because $0.047 < 0.05$.

We conclude that California RNs have a higher average annual salary than $69,110.

Homework – Ch. 10 #90

Elijah wants to know whether textbook costs are different for different courses of study. He selects a random sample of 33 sociology textbooks offered on a popular online site. The mean price of his sample is $74.64 with a standard deviation of $49.36. He then selects a random sample of 33 math and science textbooks from the same site. The mean price of this sample is $111.56 with a standard deviation of $66.90. Is the mean price of a sociology textbook lower than the mean price of a math or science textbook? Test at a 1% significance level.

Homework – Ch. 10 #90

Elijah wants to know whether textbook costs are different for different courses of study. Is the mean price of a sociology textbook lower than the mean price of a math or science textbook?

Mean price of a sociology textbook: μ_1

Mean price of a math/science textbook: μ_2

Null hypothesis: $\mu_1 - \mu_2 = 0$

Alternative hypothesis: $\mu_1 - \mu_2 < 0$

Homework – Ch. 10 #90

He selects a random sample of 33 sociology textbooks offered on a popular online site. The mean price of his sample is $74.64 with a standard deviation of $49.36. He then selects a random sample of 33 math and science textbooks from the same site. The mean price of this sample is $111.56 with a standard deviation of $66.90.

Sociology sample mean $\bar{x}_1$: $74.64

Sociology sample standard deviation s_1: $49.36

Sociology sample size n_1: 33

Homework – Ch. 10 #90

He selects a random sample of 33 sociology textbooks offered on a popular online site. The mean price of his sample is $74.64 with a standard deviation of $49.36. He then selects a random sample of 33 math and science textbooks from the same site. The mean price of this sample is $111.56 with a standard deviation of $66.90.

Math/science sample mean $\bar{x}_1$: $111.56

Math/science sample standard deviation s_1: $66.90

Math/science sample size n_1: 33

Two group test

Homework – Ch. 10 #90

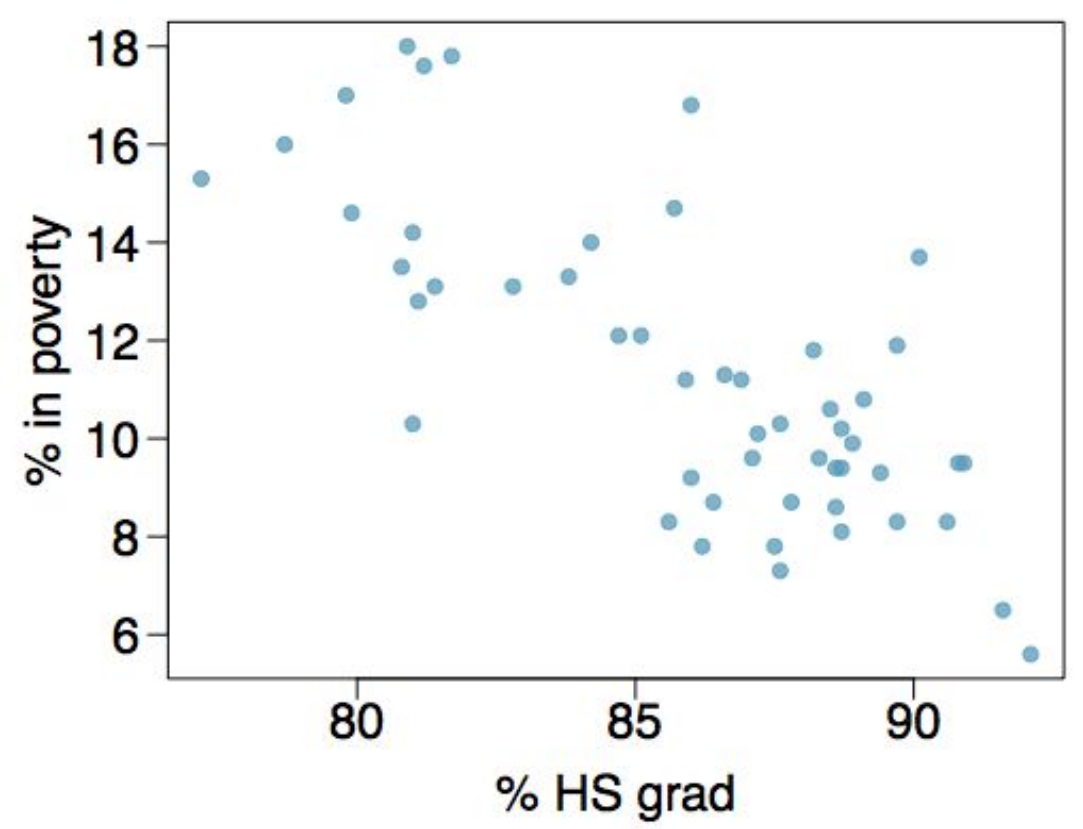
Is the mean price of a sociology textbook lower than the mean price of a math or science textbook? Test at a 1% significance level.

Conclusion:

The p-value is 0.008. The probability of seeing this difference in sample mean prices if the population means are equal is only 0.008. We **reject** the null hypothesis because this is lower than the significance level of 0.01.

We conclude that sociology textbooks are, on average, lower in price than math/science textbooks.

State Poverty Rate vs. State HS Graduation Rate



Team Activity, Models 1 & 2

What does it mean for two variables to have a strong linear correlation?

What does it mean for two variables to be positively related?

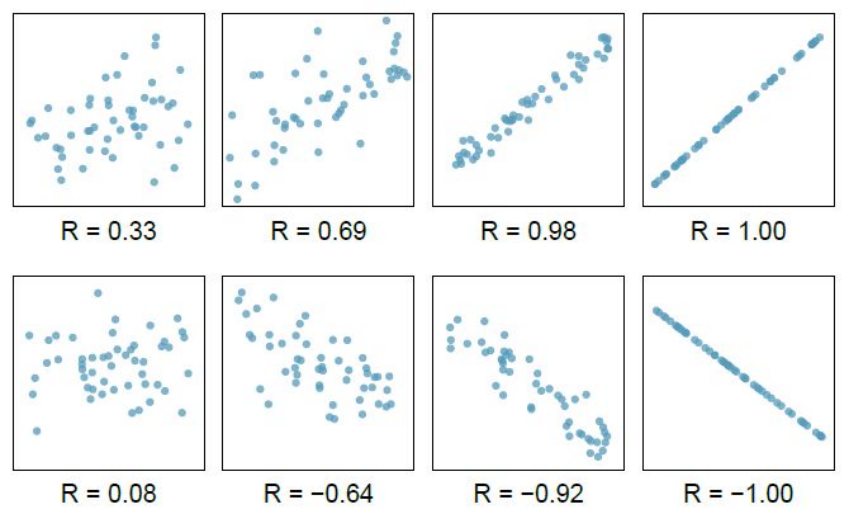
What is the difference in correlation and causation?

We'll spend 20-25 minutes on these two sections.

Model 1 Debrief

When there is a strong linear relationship between an increase in one variable and an increase in another variable, R is _____.

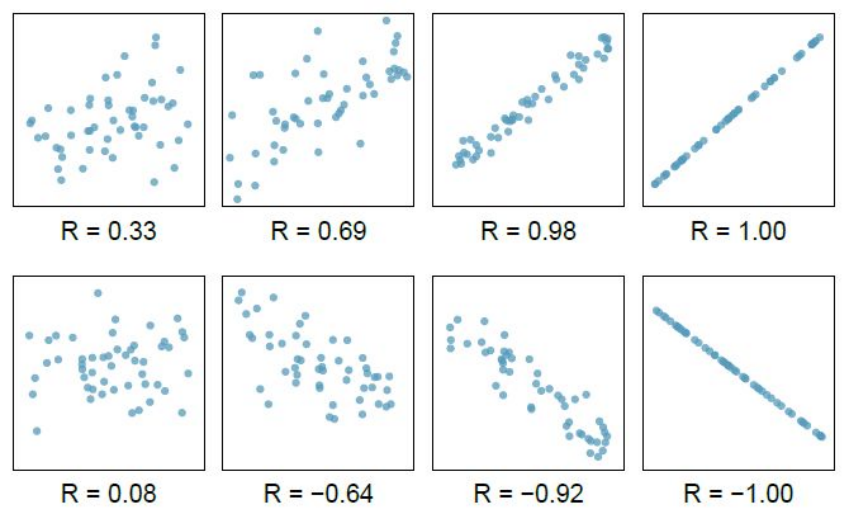
When there is a strong linear relationship between an increase in one variable and a decrease in another variable, R is _____.



Model 1 Debrief

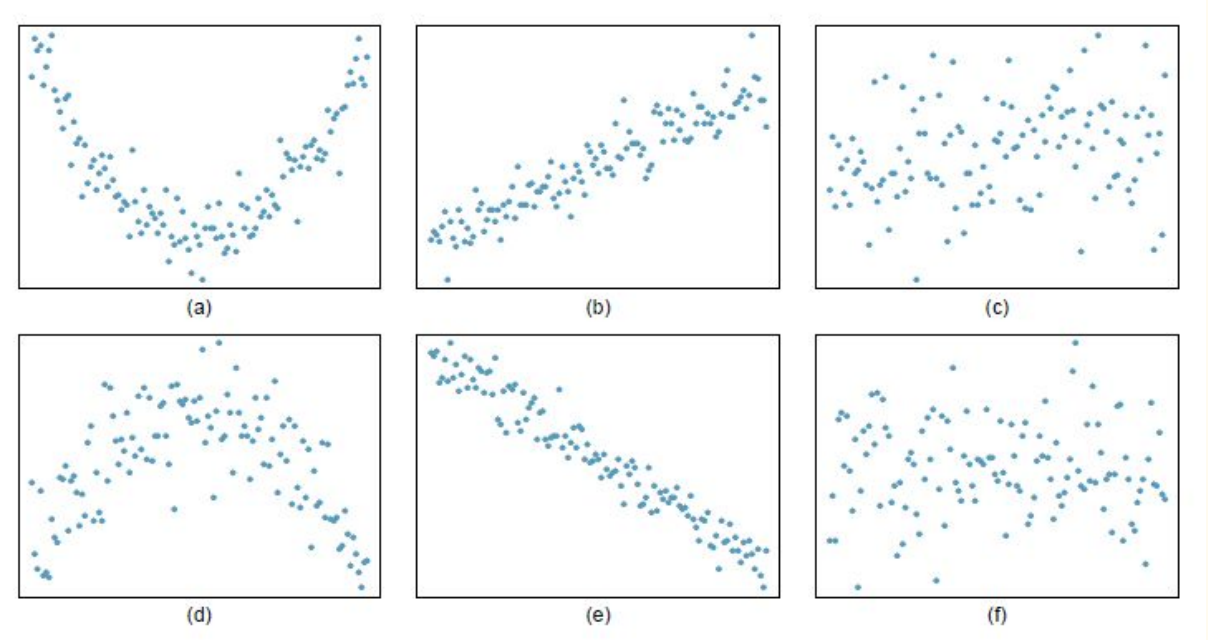
When there is a strong linear relationship between an increase in one variable and an increase in another variable, R is **positive, close to 1.**

When there is a strong linear relationship between an increase in one variable and a decrease in another variable, R is **negative, close to -1**.



Model 1 Debrief

Weak/strong? Linear/nonlinear? Positive/negative?



Model 2 Debrief

How are temperature, sunburns, and ice cream sales related?

Team Activity, Model 3

How do we use a line of best fit (**a least squares regression line**) for prediction?

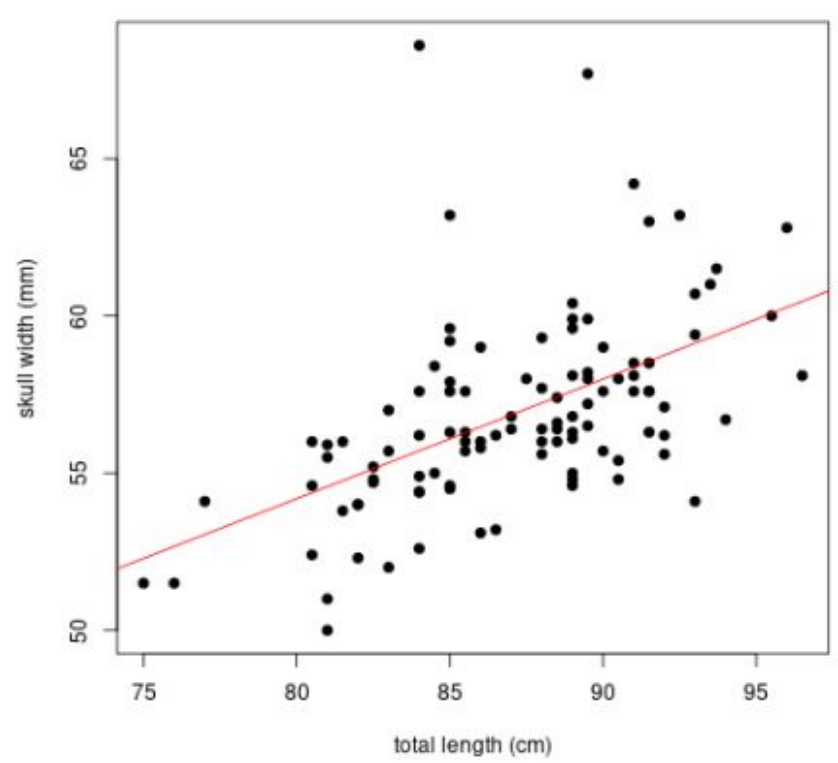
When do we use it? – When there is a linear relationship between two variables!

Model 3 Debrief

Explanatory variable: The one we use to predict the response, plotted on the horizontal axis. Also called the **independent** variable.

Response variable: What we're trying to predict, plotted on the vertical axis. Also called the **dependent** variable.

Model 3 Debrief



$\hat{y} = 0.38x + 23.8$

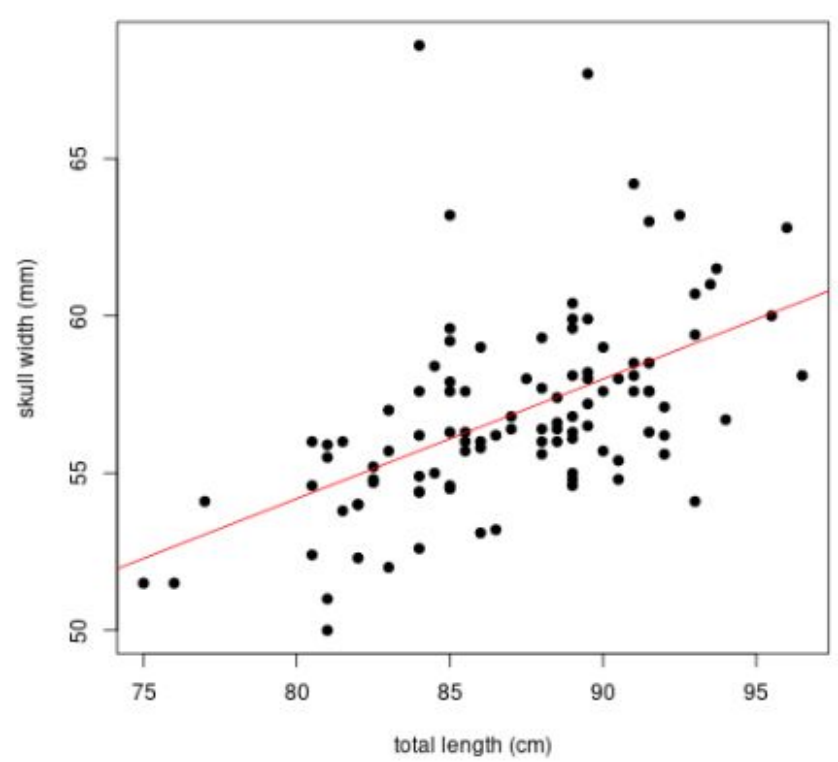
Slope: 0.38 mm/cm

For every cm longer the possum is, we expect the skull to be 0.38 mm wider

Intercept: 23.8 mm

A hypothetical possum with 0 length (?!) would have a skull width of 23.8 mm.

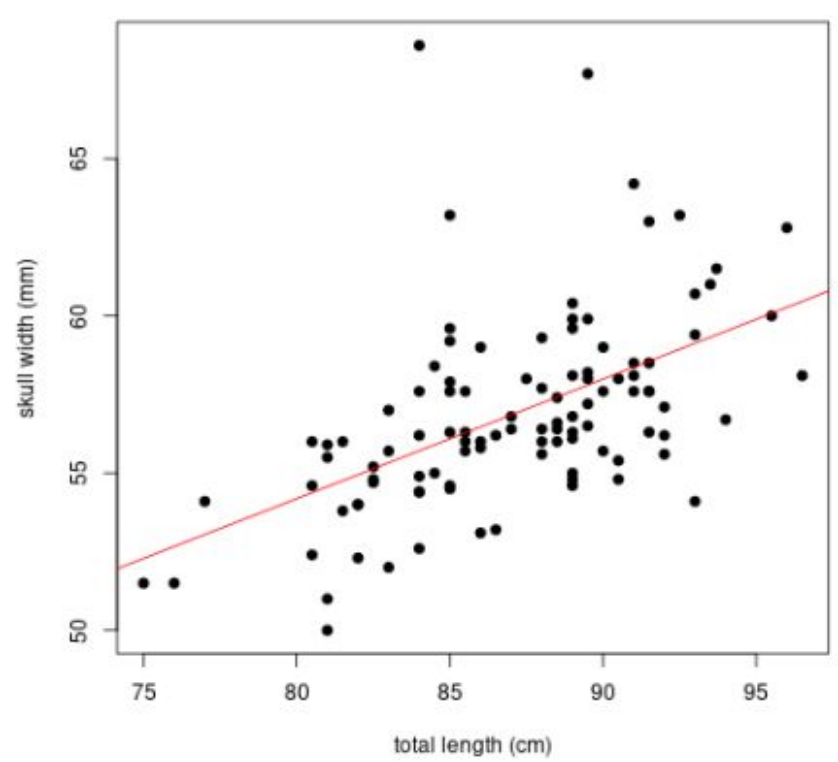
Model 3 Debrief



$$\hat{y} = 0.38x + 23.8$$

What is the skull width of a possum with total length 88cm?

Model 3 Debrief



$\hat{y} = 0.38x + 23.8$

What is the skull width of a possum with total length 88cm?

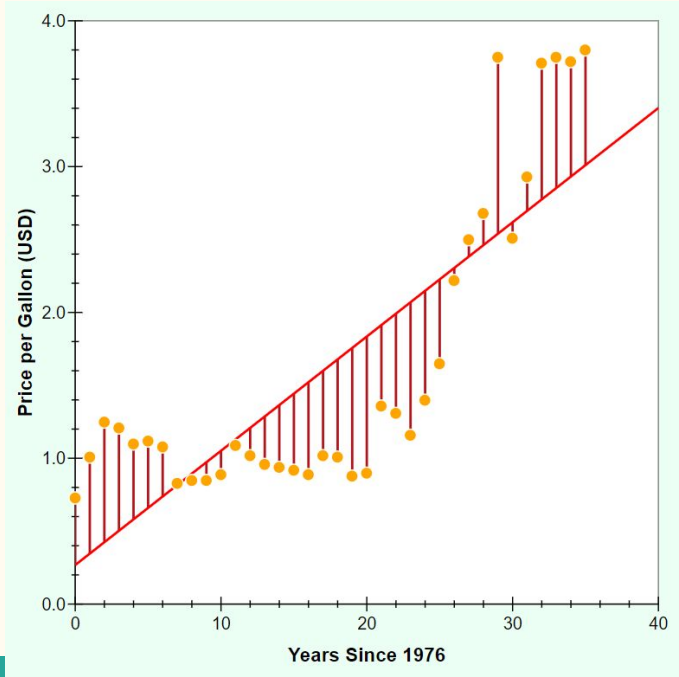
We have $x = 88$.

So $\hat{y} = 0.38(88) + 23.8$

$= 57.24$ mm

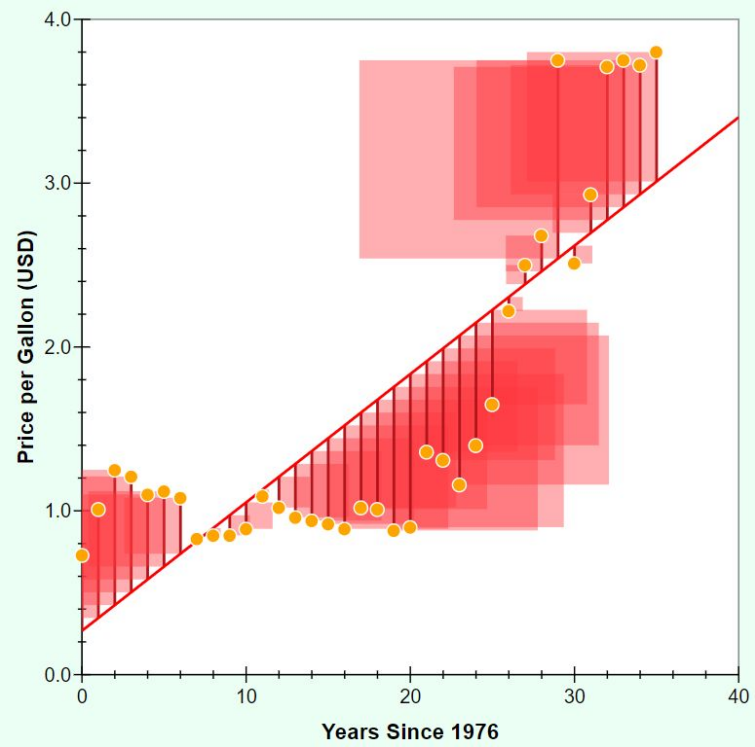
Where does the best fit line come from?

It is called the “least squares” line because it minimizes the sum of squares of the **residuals.** Residuals are the difference between the prediction made by the line and the actual values:



Where does the best fit line come from?

We find the line that minimizes the **square** of the residuals (so that they're all positive).



Wednesday

- Making scatterplots
- Finding best fit lines
- Using them to understand relationships between two variables and to make predictions about the future