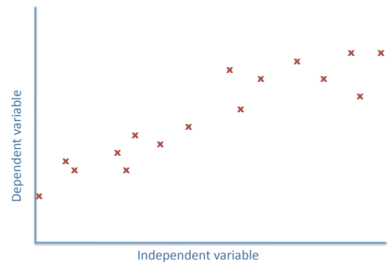
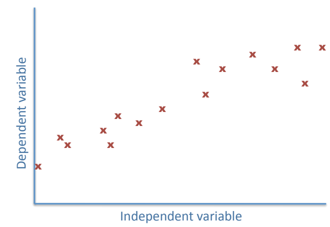
## Evaluating Correlation or Association
### Regression Analysis




- Test: comparison of two variables (continuous data / one sample )

## Example:
### continuous data / one sample / two variables




- Test: comparison of two variables:
  - Correlation coefficient / linear regression

## Pearson Sample Correlation Coefficient

Does the Dependent variable correlate with the Independent variable?

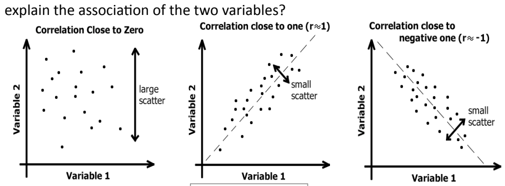
$$r = \frac{n\,\Sigma xy - [(\Sigma x)(\Sigma y)]}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\ \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$-1 \leq r \leq +1$

## Pearson Sample Correlation Coefficient

Does the Dependent variable correlate with the independent variable?

- How well does a linear regression (straight line) explain the association of the two variables?




$-1 \leq r \leq +1$

- **r** is related to scatter (standard deviation) and slope
- If **r = 0**, y is not correlated with x (I.e., Δy is not associated with Δx)
- If **r > 0**, y (variable 2) is positively correlated with x (variable 1)
- If **r < 0**, y is negatively correlated with x

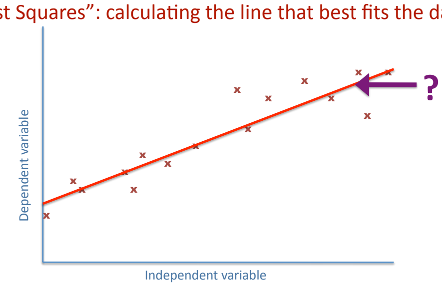
## Is the Correlation Coefficient significantly different from 0?

$$r = \frac{n\,\Sigma xy - [(\Sigma x)(\Sigma y)]}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\ \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

- **r** is related to scatter (standard deviation) and slope
- If **r = 0**, y is not correlated with x (I.e., Δy is not associated with Δx)
- If **r > 0**, y (variable 2) is positively correlated with x (variable 1)
- If **r < 0**, y is negatively correlated with x

Compare |**r**| with Critical Value for **n**.

## Linear Regression
### "Least Squares": calculating the line that best fits the data




- Linear regression line
- Linear regression equation: **y = a + bx**
  - **b** = slope of he line (Δy / Δx)
  - **a** = y-intercept (value of y if x=0)

## Least-Squares Line

- Exact equation for "line of best fit"

$$\hat{y} = a + bx$$

- Slope, b = $\frac{n\sum xy-(\sum x)(\sum y)}{n\sum x^2-(\sum x)^2}$
- Intercept, a = $\bar{y}$ - b$\bar{x}$
- Using standard deviation…b = r($\frac{s_y}{s_x}$)
- ($\bar{x}$, $\bar{y}$) is on the least-squares line

## Regression

### Definition

**Regression Equation**

**Given a collection of paired data, the regression equation**

$$\hat{y} = b_0 + b_1 x$$

**algebraically describes the relationship between the two variables**

**Regression Line**

**(line of best fit or least-squares line) is the graph of the regression equation**

18

**Least-Squares Regression Line**

We can use technology to find the equation of the least-squares regression line. We can also write it in terms of the means and standard deviations of the two variables and their correlation.

Definition: **Equation of the least-squares regression line**

We have data on an explanatory variable *x* and a response variable *y* for *n* individuals. From the data, calculate the means and standard deviations of the two variables and their correlation. The least squares regression line is the line $\hat{y} = a + bx$ with

**slope** $b = r\frac{s_y}{s_x}$

and ***y* intercept** $a = \bar{y} - b\bar{x}$

Least-Squares Regression

Stats Chapter 5 - Least Squares Regression

***Definition of a regression line***:

A regression line is a straight line that describes how a response variable (y) changes as an explanatory variable (x) changes…

- Used to predict a y value given an x value.
- Requires an explanatory and a response variable.
- Given as an equation of a line in slope intercept form:

$$\hat{y} = a + bx$$

Read as: "y-hat" | a = y-intercept | b = slope

## summary

- Regression equation: describes how a dependent variable (**y**) changes in association with an independent variable (**x**).
  - **y = a + bx**
- **a** = y-intercept: the value of y when x=0.
- **b** = slope: the rate at which y varies in association with x.
- **r** = **correlation coefficient**: What is the probability that the change in y is related to the change in x?
  - p-value: What is the probability that the change in y is **not** related to the change in x ($H_0$)?
- **$r^2$** = **determination coefficient**: *How much* (i.e., what fraction) of the variation in y is related to the variation in x?

Remember: correlation does not prove causation!

## Scatterplots and Line-fitting in Excel

- https://www.youtube.com/watch?v=Ohp1PpzrRhE