Lesson 13. Coefficient of Determination for Simple Linear Regression – Part 1

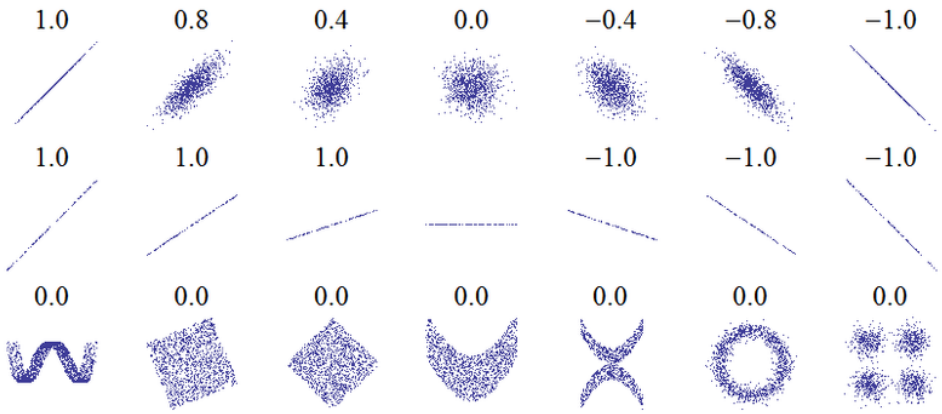
Note. In Part 2 of this lesson, you can run the R code that generates the outputs here in Part 1.

1 Overview

- **Correlation** quantifies the strength of the linear relationship between X and Y

Population correlation	Sample correlation

- Some examples that illustrate different correlation values:



https://commons.wikimedia.org/wiki/File:Correlation_examples.png

2 Properties

- Possible values are from ____ to ____
- A larger magnitude means a ____ linear relationship
- $\rho > 0$ means larger values of Y are associated with ____ values of X
- $\rho < 0$ means larger values of Y are associated with ____ values of X
- Relation to slope:

⇒ In simple linear regression, testing whether $\beta_1 = 0$ (versus $\beta_1 \neq 0$) is equivalent to testing whether $\rho = 0$ (versus $\rho \neq 0$)

3 t-test for correlation

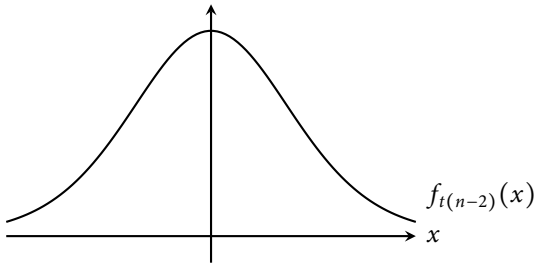
- Question: **Are the variables X and Y correlated?**
- Formal steps:

1. State the hypotheses:

2. Calculate the test statistic:

3. Calculate the p-value:
   - If the conditions for simple linear regression hold, then the test statistic t follows

   $\Rightarrow$ p-value =



4. State your conclusion, based on the given significance level α:

   If we reject H_0 (p-value $\leq \alpha$):

   > We reject H_0 because the p-value is less than the significance level α. We see significant evidence that X and Y are correlated.

   If we fail to reject H_0 (p-value $> \alpha$):

   > We fail to reject H_0 because the p-value is greater than the significance level α. We do not see significant evidence that X and Y are correlated.

The underlined parts above should be rephrased to correspond to the context of the problem

Example 1. Consider once again our regression model with the `AccordPrice` data. Recall that there are 30 cars in the data.

```
library(Stat2Data)
data(AccordPrice)
```

We can find the sample correlation between *Price* and *Mileage* in R using the `cor()` function:

```
cor(AccordPrice$Price, AccordPrice$Mileage)
```

We find that the sample correlation between *Price* and *Mileage* is -0.8489.

In fact, we can find the sample correlation between all the variables in a data frame using the `cor()` function, as long as all the variables are quantitative:

```
cor(AccordPrice)
```

Here is the output:

A matrix: 3 × 3 of type dbl

	Age	Price	Mileage
Age	1.0000000	-0.8956156	0.7895010
Price	-0.8956156	1.0000000	-0.8489441
Mileage	0.7895010	-0.8489441	1.0000000

Perform a t-test to determine whether there is a significant ($\alpha = 0.05$) correlation between *Price* and *Mileage*.

- Note that we can perform the t-test for correlation without performing simple linear regression
    - Sometimes it isn't clear which variable is explanatory vs. response, so it is handy to be able to test correlation without doing regression

Example 2. Continuing with the `AccordPrice` example...

We can also perform the t-test for correlation in R using the following code:

```
cor.test(AccordPrice$Price, AccordPrice$Mileage,
         alternative = "two.sided", conf.level = 0.95)
```

The output is as follows:

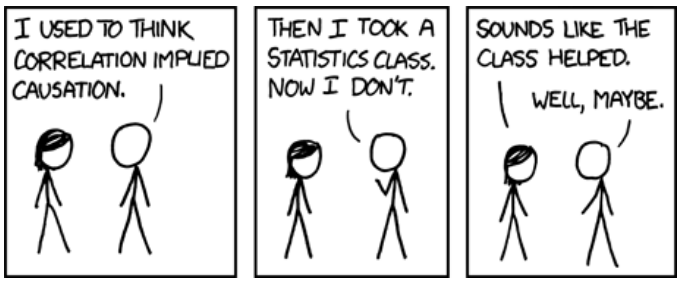
```
  Pearson's product-moment correlation

data:  AccordPrice$Price and AccordPrice$Mileage
t = -8.5002, df = 28, p-value = 3.055e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9259982 -0.7039888
sample estimates:
       cor
-0.8489441
```

4 Correlation does not imply causation

- For example, suppose

  X = number of firefighters Y = damage in dollars

  - X and Y probably have a strong correlation
  - Do more firefighters present cause more damage?
  - Size of fire is responsible for both
- A significant correlation only means the variables are associated, not that one causes the other



https://xkcd.com/552/

5 Coefficient of determination (r^2)

- The **coefficient of determination** r^2 tells us how much of the variability in the response variable is explained by the regression model
  - For simple linear regression, the coefficient of determination is directly related to the sample correlation:

$$\text{coefficient of determination} = (\text{sample correlation})^2$$

Example 3. Continuing with the `AccordPrice` example...

Look at the R output in Lessons 11 and 12.

a. Using the ANOVA table output by R, calculate the coefficient of determination (r^2). Interpret it.
b. Look at the summary output by R, where do you see the value you calculated in part a?

6 Summary

- We have three distinct ways to test for a significant linear relationship between two quantitative variables:
    1. t-test for simple linear regression slope
    2. ANOVA F-test for simple linear regression
    3. t-test for correlation
- These three procedures are exactly equivalent for simple linear regression
- When we have more than one predictor, these tests will have different purposes
- A significant linear relationship does not mean that a line is the best way to describe the relationship
- Correlation does not imply causation!