Learning R

Session 3: Correlation and Regression with R

DR EDMUND RYAN

Overview of this R Course

1. Introduction to R (session 1)
2. Creating graphical plots with R (session 2)

Task 1

3. Correlation and Regression with R (session 3)
4. Creating a model with R using For loops and if-else statements (session 4)

Task 2

5. Hypothesis testing with R (session 5)
6. Using R packages to solve problems (session 6)

Task 3

Overview of this R Course

1. Introduction to R (session 1)
2. Creating graphical plots with R (session 2)

Task 1

3. Correlation and Regression with R (session 3)
4. Creating a model with R using For loops and if-else statements (session 4)

Task 2

5. Hypothesis testing with R (session 5)
6. Using R packages to solve problems (session 6)

Task 3

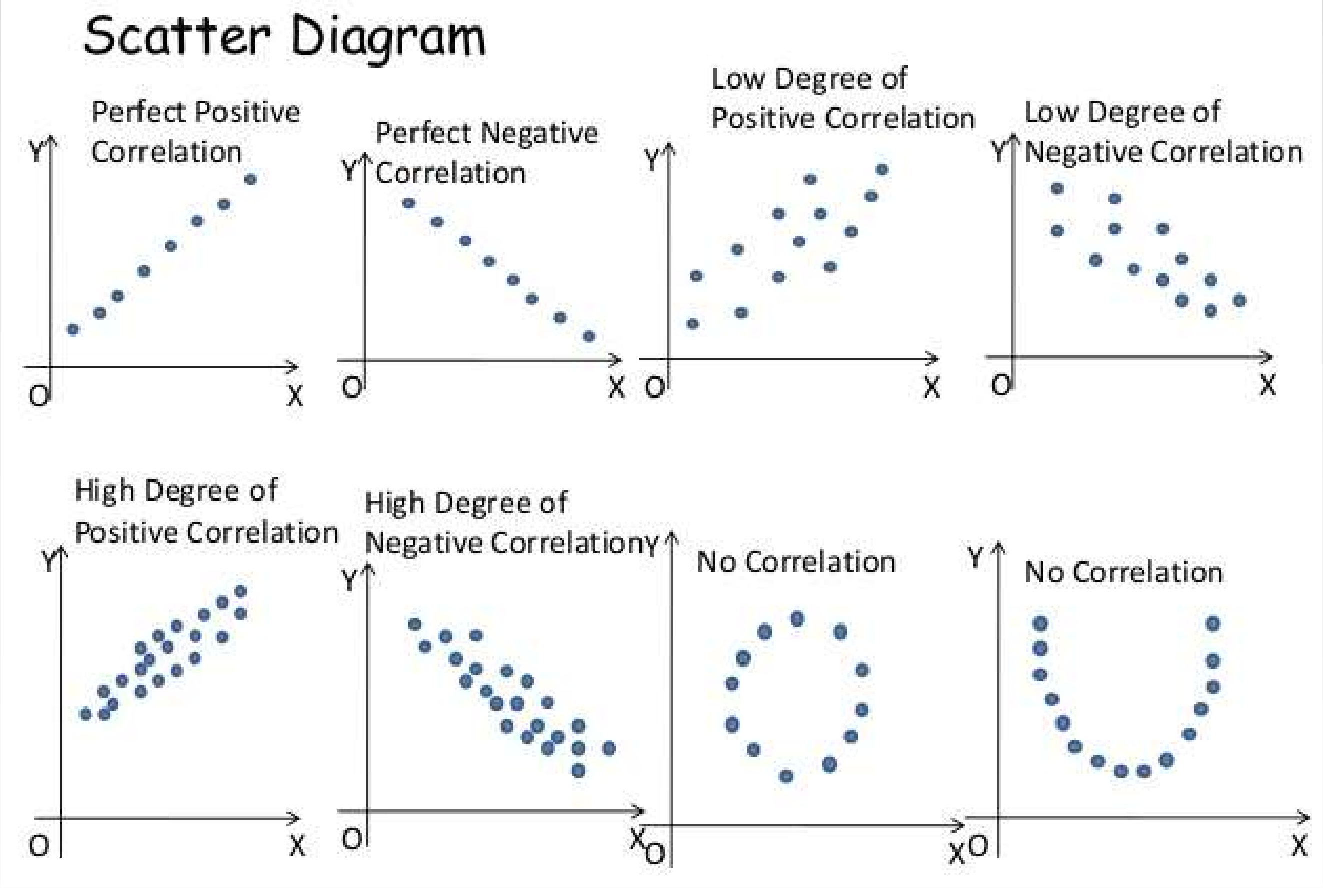
All R code & other files are on Github

https://github.com/edmundryan

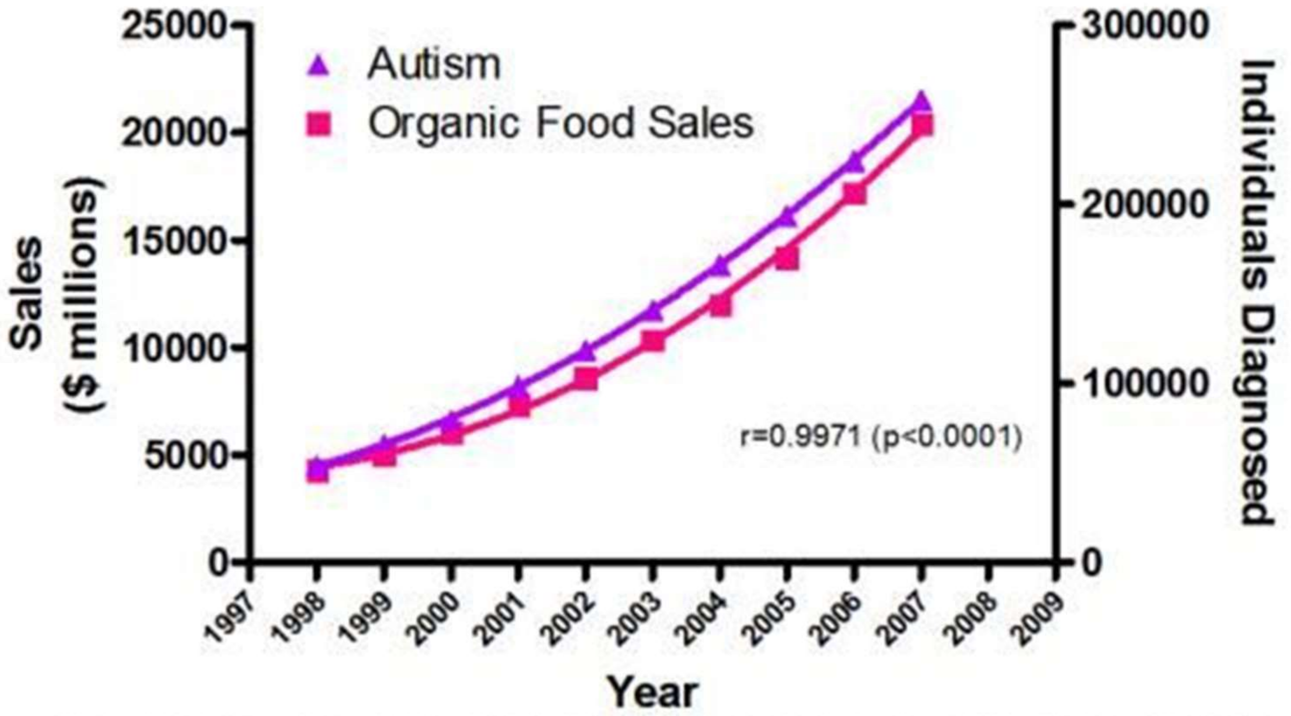
Correlation

Correlation is used to represent the linear relationship between two variables.

Different types of correlation



Spurious Correlations



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals
with Disabilities Education Act

Correlation

Correlation is used to represent the linear relationship between two variables.

Having two variables that correlate does **NOT** mean you have causation. You need demonstrate causality through your experimental design (regression).

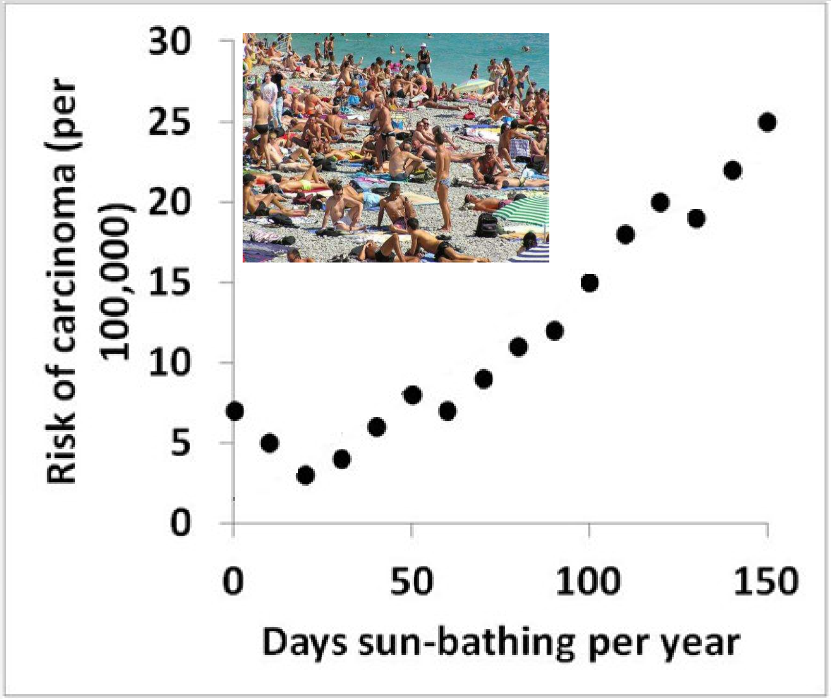
Regression

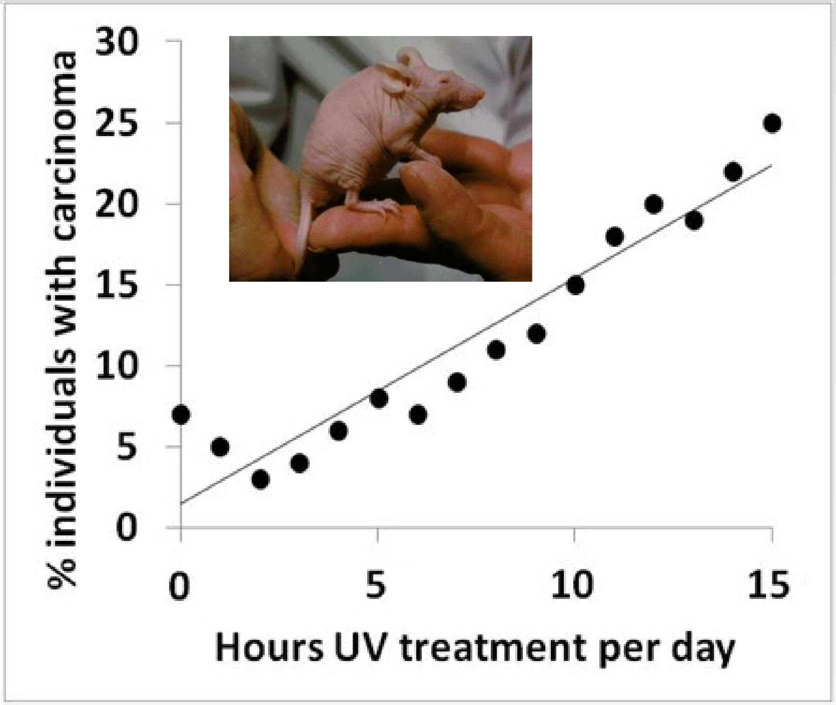
This is similar to correlation, but must have dependent (outcome) and independent (predictor) variables.

Regression describes how the independent variable is numerically related to the dependent variable.

Ideally, we will control the independent variable. Failing that, we need to ensure that we have already done some kind of controlled experiment (see next slide).

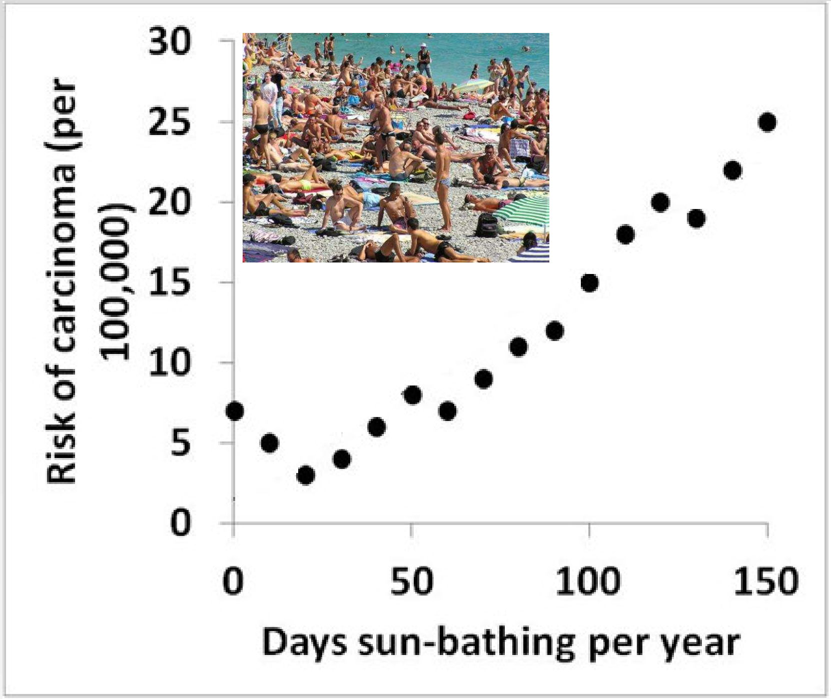
Causation

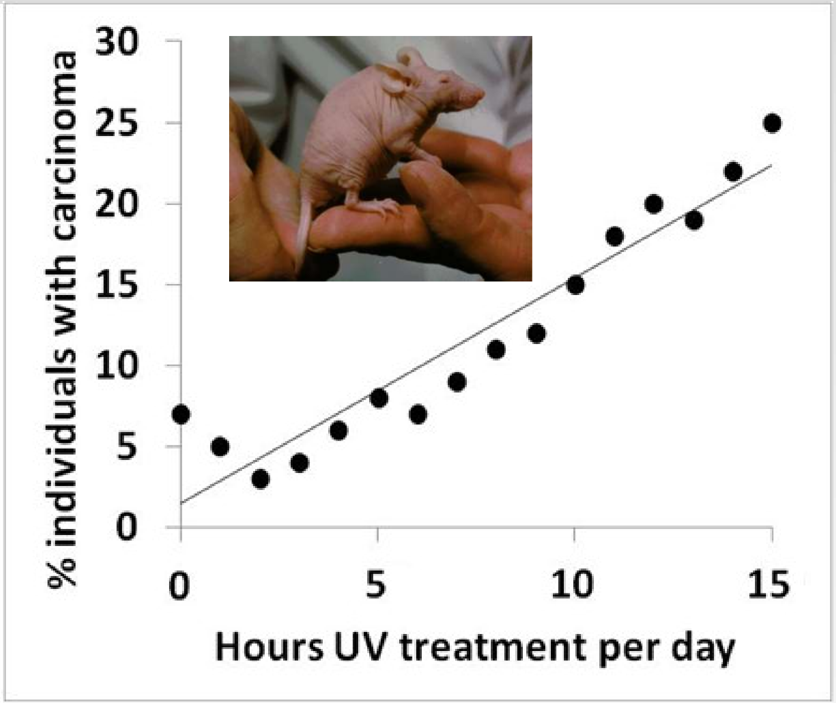




What’s the difference in how the measurements are made?

Causation

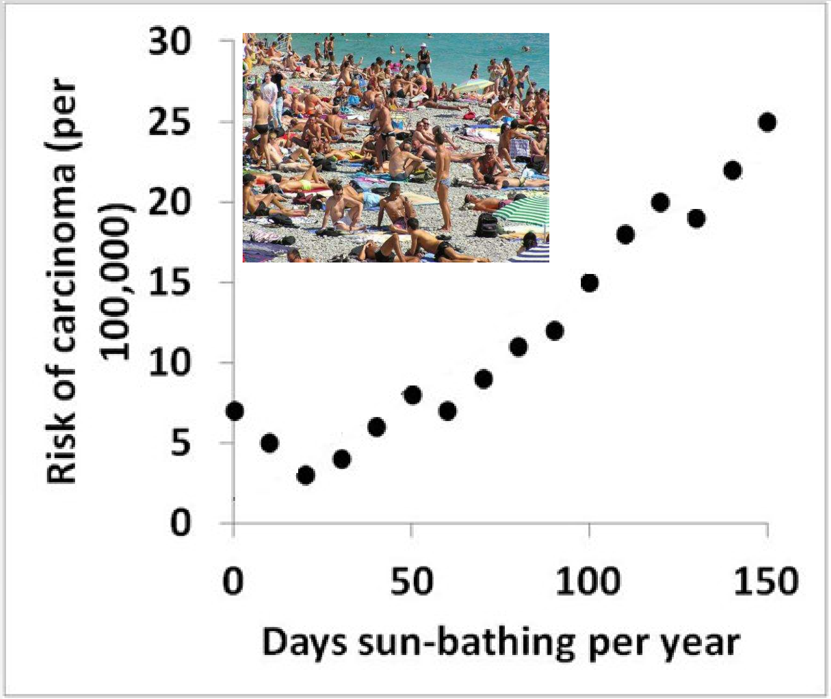


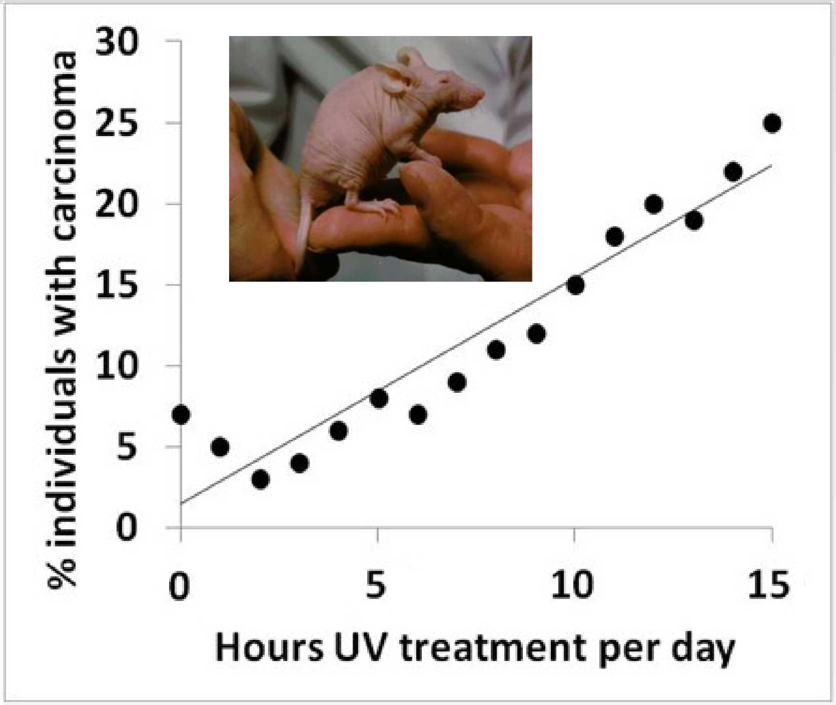


Measurements in x and y = Correlation

Manipulation in x and measurement in y = Regression

Causation





We could do regression on data from left plot but only once we've established the link between exposure and cancer using a controlled experiment (right plot).

Coefficient of determination (r^2)

- It quantifies the proportion of the variation in the dependent variable (y) that is explained by the independent variables (x). We refer to it as r^2 or R^2.
- For simple linear regression, r^2 is the square of Pearson's product moment correlation coefficient (r).

The assumptions of regression

- Linearity of the data – relationship between the X variable(s) (independent variables) and Y variable (dependent variable) is assumed to be linear.
- Normality of the residuals– the residual errors are assumed to be normally distributed.
- Homogeneity of the residuals' variance – the residuals are assumed to have a constant variance (also known as homoscedasticity).
- Independence of the residual error terms

Potential problems

- Non-linearity of the outcome.
- Non constant variance of the residuals (heteroscedasticity).
- Non-normality of the residuals.
- Presence of influential data points that can be:
  - Outliers: extreme values in the dependent variable
  - High-leverage points: extreme values in the independent variables.