

Correlation & Linear models

2022-03-3

Today we'll cover

- ▶ Correlation
- ▶ Linear Regression

statistical relationship between two variables

- ▶ is there a relationship?
- ▶ how strong is it?
- ▶ what direction is it in?

Covariance

| tumor size | GFP |
|------------|-----|
| 30 | 5 |
| 35 | 8 |
| 40 | 8 |
| 25 | 4 |
| 35 | 5 |

$$\sigma_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1} \quad (1)$$

the sign of σ_{xy} tells you something about the direction of how two variables covary

Covariance

| tumor size | GFP | tumor size | GFP | $x - \bar{x}$ | $y - \bar{y}$ |
|------------|-----|------------|-----|---------------|---------------|
| 30 | 5 | 30 | 5 | -3 | -1 |
| 35 | 8 | 35 | 8 | +2 | +2 |
| 40 | 8 | 40 | 8 | +7 | +2 |
| 25 | 4 | 25 | 4 | -8 | -2 |
| 35 | 5 | 35 | 5 | +2 | -1 |

$$\sigma_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1} \quad (1)$$

the sign of σ_{xy} tells you something about the direction of how two variables covary

Correlation

however covariance is sensitive to the scale of the data (i.e. prove this by multiplying tumor size and GFP by 2?

so we need to scale it

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (2)$$

ρ_{xy} ranges from -1 to +1

What is the correlation of x and y ?

iris

- ▶ what is the correlation of Sepal.Width and Sepal.Length?
- ▶ what is the correlation between Petal.Length and Petal.Width?
- ▶ can you make a correlation matrix?



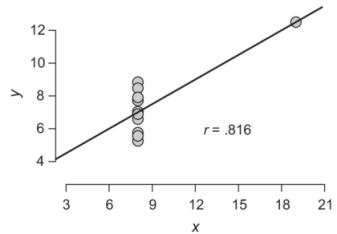
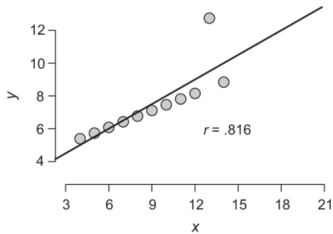
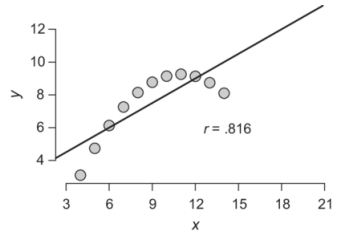
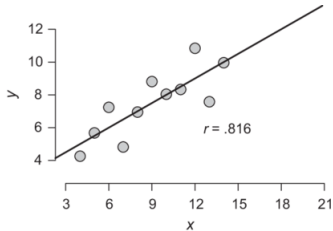
What is the correlation of x and y ?

what is the correlation between x and y

- ▶ $ds1$
- ▶ $ds2$
- ▶ $ds3$

What is the correlation of x and y ?

Anscombe's quartet



?cor.test

iris

- ▶ is there a significant correlation of Sepal.Width and Sepal.Length?
- ▶ is there a significant correlation between Petal.Length and Petal.Width?

what is the null and alternative hypothesis that you're testing?

- ▶ what's the correlation between two random normal variables each with mean 0 and sd 1 ?

- ▶ whats the correlation between two random normal variables each with mean 0 and sd 1 ?

what do you think happens if I was to do this?

```
cor(c(rnorm(100, 0, 1), c(10, 20, 30, 40)),  
    c(rnorm(100, 0, 1), 12, 22, 33, 45))
```

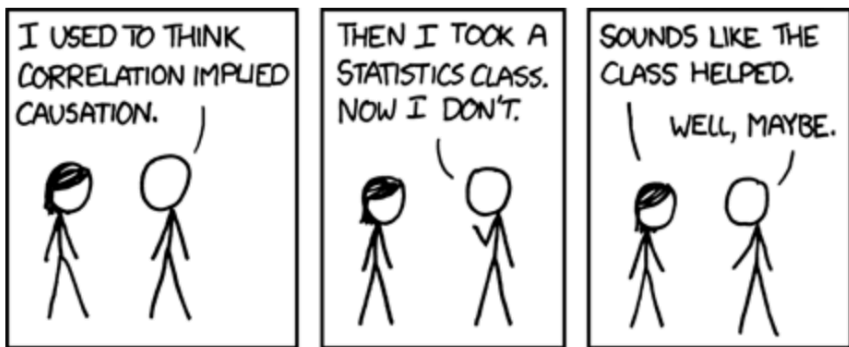
- ▶ whats the correlation between two random normal variables each with mean 0 and sd 1 ?

what do you think happens if I was to do this?

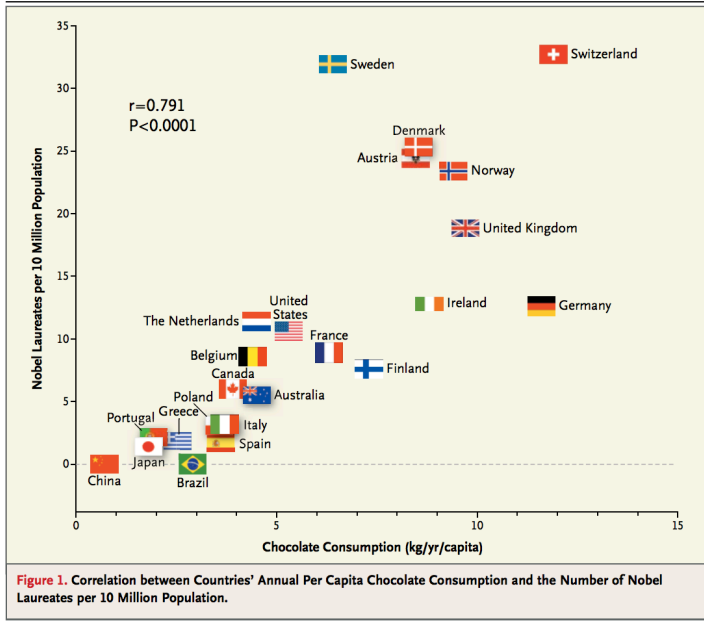
```
cor(c(rnorm(100, 0, 1), c(10, 20, 30, 40)),  
    c(rnorm(100, 0, 1), 12, 22, 33, 45))
```

beware of outliers

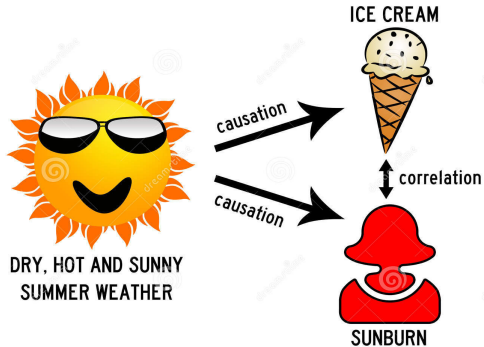
Correlation and Causation



Correlation and Causation



Correlation is not Causation



Download from
Dreamstime.com

The content does not constitute an offer or a recommendation for any financial product or service.



37981889

Alan Lacroix | Dreamstime.com

What is Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation (a straight line) to the observed data. One variable is considered to be an independent variable or predictor and the other is considered to be a dependent variable, or response.

The dependent variable in linear regression should be continuous and normally distributed. The independent variable can be continuous or categorical.

What is Linear Regression

Linear Regression

Simple linear regression model

The diagram illustrates the simple linear regression model equation, $y = \beta_0 + \beta_1 x + \epsilon$, enclosed in a yellow box. Labels with arrows point to specific parts of the equation: 'Dependent Variable' points to y ; 'Population y intercept' points to β_0 ; 'Population Slope Coefficient' points to β_1 ; 'Independent Variable' points to x ; and 'Random Error term, or residual' points to ϵ . Below the box, two purple curly braces group the terms: the first brace under $\beta_0 + \beta_1 x$ is labeled 'Linear component', and the second brace under ϵ is labeled 'Random Error component'.

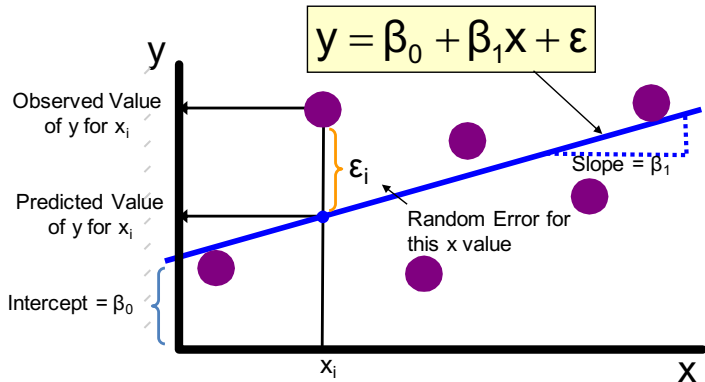
$$y = \beta_0 + \beta_1 x + \epsilon$$

Labels and components:

- Dependent Variable: y
- Population y intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: x
- Random Error term, or residual: ϵ
- Linear component: $\beta_0 + \beta_1 x$
- Random Error component: ϵ

Linear Regression

The Line of Best Fit



Clinical Example

Ovarian Cancer

This dataset from a cohort of ovarian cancer patients. It contains clinical information, including age, treatment group, presence of residual disease, quality of life, blood pressure* and cholesterol levels*. The patients were followed up for ~ 3.5 years after treatment.

Age - age at treatment

Resid Disease - was there residual disease after treatment

Rx - which drug were they put on, A or B

ECOG - quality of life

BP - blood pressure at start of treatment

Chol - cholesterol levels at start of treatment

Clinical Example

Ovarian Cancer

| | Age | Resid Disease | Rx | ECOG | BP | Chol |
|----|-------|---------------|----|------|--------|-------|
| 1 | 72.33 | yes | A | good | 117.83 | 13.58 |
| 2 | 74.49 | yes | A | good | 114.00 | 7.78 |
| 3 | 66.47 | yes | A | bad | 117.55 | 10.95 |
| 4 | 74.50 | yes | A | bad | 113.50 | 22.50 |
| 5 | 43.14 | yes | A | good | 139.19 | 22.11 |
| 6 | 63.22 | no | B | bad | 124.80 | 8.46 |
| 7 | 64.42 | yes | B | good | 118.09 | 23.19 |
| 8 | 58.31 | no | B | good | 130.09 | 26.51 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 25 | 44.21 | yes | B | good | 138.34 | 26.25 |
| 26 | 59.59 | no | B | bad | 129.18 | 22.93 |

Linear Regression - Ovarian Cancer

Ovarian Cancer

| | Age | Resid Disease | Rx | ECOG | BP | Chol |
|----|-------|---------------|----|------|--------|-------|
| 1 | 72.33 | yes | A | good | 117.83 | 13.58 |
| 2 | 74.49 | yes | A | good | 114.00 | 7.78 |
| 3 | 66.47 | yes | A | bad | 117.55 | 10.95 |
| 4 | 74.50 | yes | A | bad | 113.50 | 22.50 |
| 5 | 43.14 | yes | A | good | 139.19 | 22.11 |
| 6 | 63.22 | no | B | bad | 124.80 | 8.46 |
| 7 | 64.42 | yes | B | good | 118.09 | 23.19 |
| 8 | 58.31 | no | B | good | 130.09 | 26.51 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 25 | 44.21 | yes | B | good | 138.34 | 26.25 |
| 26 | 59.59 | no | B | bad | 129.18 | 22.93 |

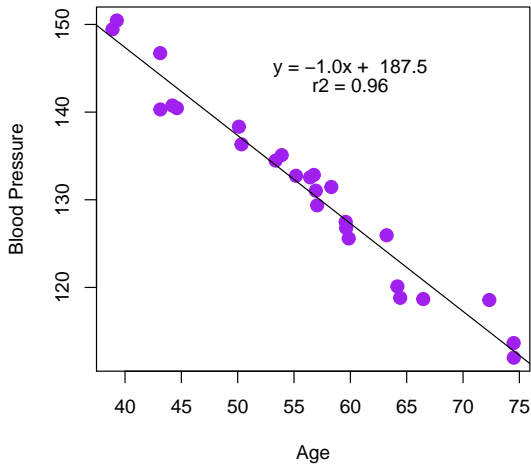
Linear Regression - Ovarian Cancer

In this cohort of ovarian cancer patients,

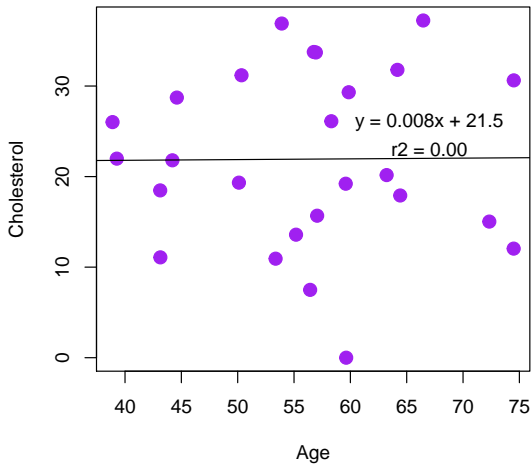
- ▶ is there a linear relationship between age and blood pressure?
- ▶ is there a linear relationship between age and cholesterol levels?

In these instances, age is the independent variable. Age can influence blood pressure and cholesterol, but blood pressure and cholesterol are not going to change a subjects age.

Ovarian Cancer - Age \sim Blood Pressure



Ovarian Cancer - Age \sim Cholesterol



$$r^2$$

coefficient of determination - proportion of variance in the dependent variable that can be explained by the independent variable.

H_0 : horizontal line is correct

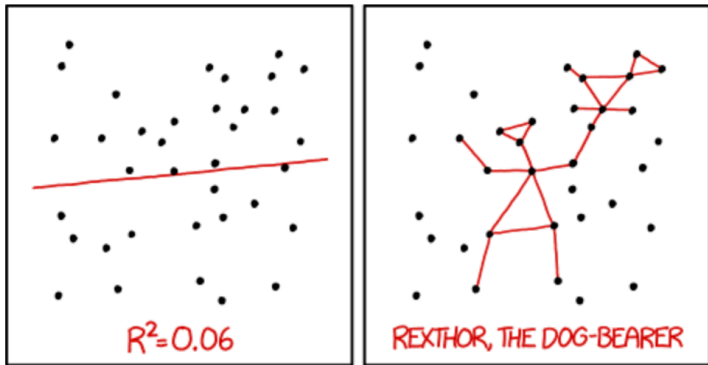
H_1 : the slope of the line differs from 0

Model fitting

we're actually fitting two models and comparing them against each other -
does a horizontal line fit the data better than the regression line?

lots of ways of comparing which model is the best?
 R^2

Linear Regression - Look at the data!



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Assumptions of Linear Regression

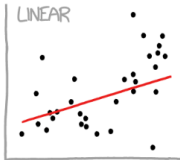
- ▶ The sample must be representative of the population to which the inference will be made
- ▶ The data (residuals) should be normally distributed
- ▶ There should be no multicollinearity
- ▶ The error terms are independently and identically normally distributed

Summary - Linear Regression

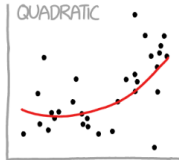
- ▶ Linear regression attempts to model the relationship between two variables by fitting a linear equation $y = \beta_0 + \beta_1 x + \epsilon$, to the observed data.
- ▶ The intercept, β_0 and slope, β_1 , from the observed data can be used to predict the relationship between variables.

Linear Regression and Other Models

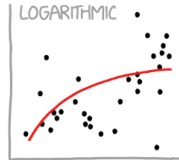
CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



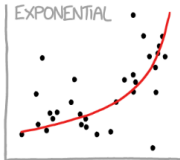
"HEY, I DID A
REGRESSION."



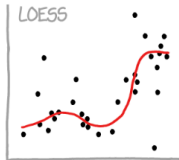
"I WANTED A CURVED
LINE, SO I MADE ONE
WITH MATH."



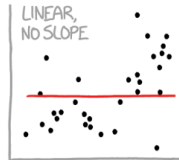
"LOOK, IT'S
TAPERING OFF!"



"LOOK, IT'S GROWING
UNCONTROLLABLY!"

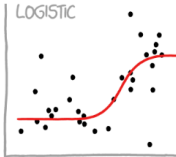


"I'M SOPHISTICATED, NOT
LIKE THOSE BUMBLING
POLYNOMIAL PEOPLE."



"I'M MAKING A
SCATTER PLOT BUT
I DON'T WANT TO."

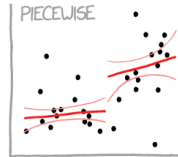
Linear Regression and Other Models



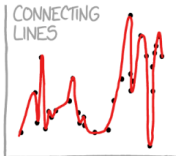
"I NEED TO CONNECT THESE
TWO LINES, BUT MY FIRST IDEA
DIDN'T HAVE ENOUGH MATH."



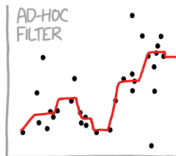
"LISTEN, SCIENCE IS HARD.
BUT I'M A SERIOUS
PERSON DOING MY BEST."



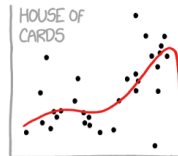
"I HAVE A THEORY,
AND THIS IS THE ONLY
DATA I COULD FIND."



"I CLICKED 'SMOOTH
LINES' IN EXCEL."



"I HAD AN IDEA FOR HOW
TO CLEAN UP THE DATA.
WHAT DO YOU THINK?"



"AS YOU CAN SEE, THIS
MODEL SMOOTHLY FITS
THE— WAIT NO NO DON'T
EXTEND IT AAAAAA!!!"

Questions?

- ▶ Next lecture: Multiple regression