Regression and ANOVA

BIO782P
2017

Recap

Comparing more than one mean

- If 3 means, we would need 3 t-tests

- If 4 means, 6 tests

- If 5 means, 10 tests

- Generally, $\frac{(N-1)(N)}{2}$ pairwise comparisons

Comparing more than one mean

Probability of at least one Type 1 error:

$1-(0.95^{\text{number of tests}})$
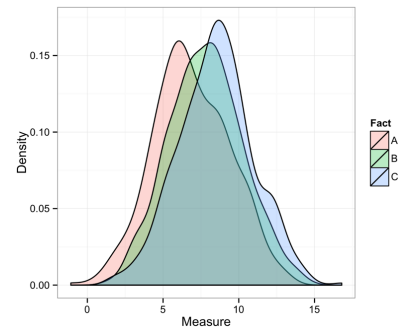
For 3 means p(error) = 0.0975

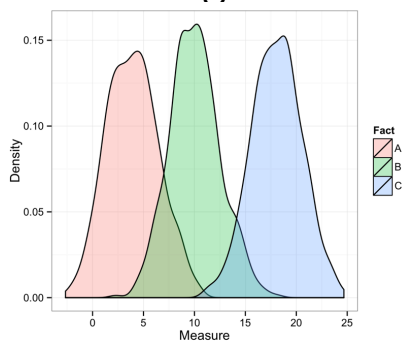For 4 means, 0.265

For 5 means, 0.401

ANOVA

- Analysis of variance

- Relies on partitioning the variance in the data into that explained by the factor(s) and that which is unexplained
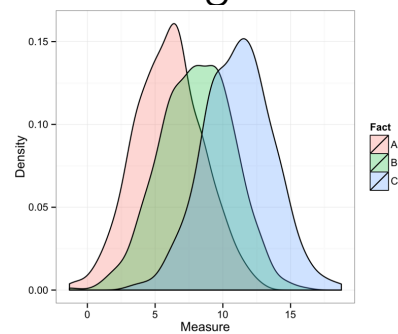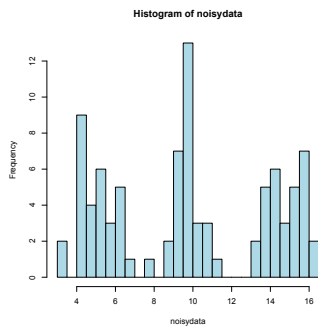
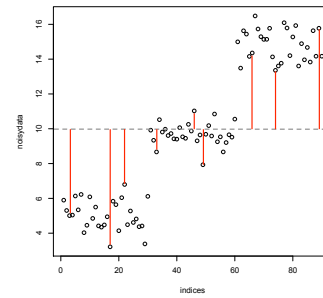# Partitioning variance

Partitioning variance

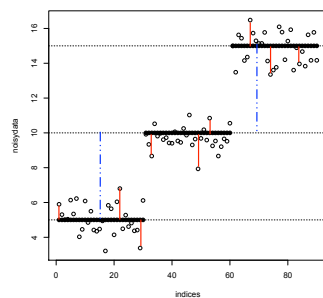# Partitioning variance

Partitioning variance

Histogram of noisydata

Partitioning variance

Partitioning variance

Performing ANOVA

To partition variability use sum of squares (SS) rather than variance (s^2)

$$\sum (x - \overline{x})^2$$

SS = s^2 (variance) x df

- Easier to add and subtract SS than s^2 because don't need to worry about differences in sample size

ANOVA

- We calculate the between group variance, or the factor variance

- This is compared with the within group variance, or error variance, using an F-test.

ANOVA table

Mean squared error: SS/df

Source	df	SS	MS
Factor	2	0.0576	0.0288
Error	15	0.1052	0.007
Total	17	0.1628	

ANOVA table

F ratio:
MSF/MSE

Source	df	SS	MS	F	p
Factor	2	0.0576	0.0288	4.114	0.037
Error	15	0.1052	0.007		
Total	17	0.1628			

ANOVA

- What does a significant ($p < 0.05$) result from an ANOVA mean?

- It tells us that at least one of the group means is different from at least one other

- To find where differences are look at 95%CIs, look at "treatment contrasts" in summary table or use a post-hoc test like the Tukey HSD test.

ANOVA in R

- lm() or aov()

- Both can carry out ANOVA

- with lm() use anova() on a model object to get an ANOVA table

Reporting an ANOVA

There were no significant differences in mean response between any factor levels (ANOVA, $F_{x,y}$ = Z, p=0.YYY)

There were no significant differences in mean response between any factor levels (table 1)

Two-factor ANOVA

- We can use ANOVA to analyse the results of experiments where more than on factor has been used

- Example: trial measuring how inflammation is affected by drug treatment, patients given a placebo, a low dose or a high dose and also classified by sex

Two-factor ANOVA

- Two factor ANOVA allows us to test for MAIN EFFECTS and also for INTERACTIONS

- A main effect is the effect of one factor in isolation

- An interaction is the effect of one factor when the level of the other factor is taken into account

Main effects and interactions
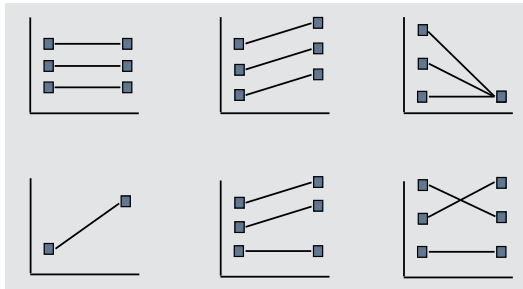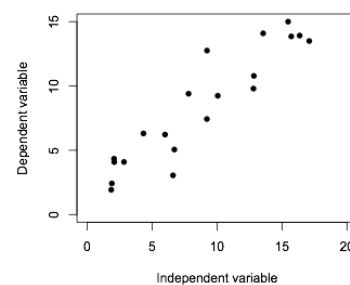

ANOVA assumptions

- Normally distributed errors
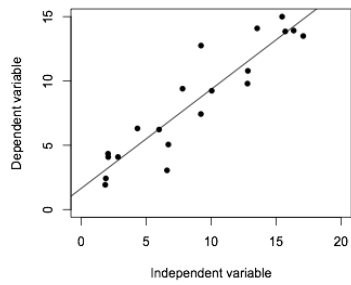- Homoscedasticity
- Observations are independent

Regression

Sometimes we want to make predictions from one variable of another

Use regression analysis to fit a line

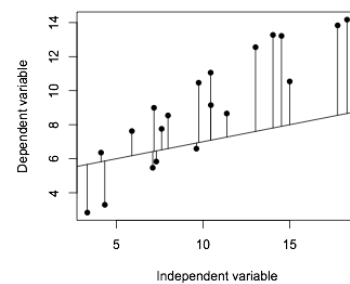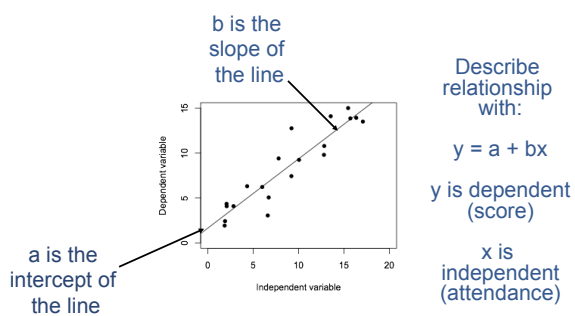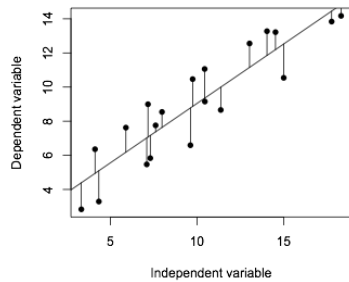One variable is independent and one is dependent

Examples

Height (dependent variable) against time (independent)

Patient response (dependent) against drug dose (independent)

MCQ test score (dependent) against attendance (independent)

b is the slope of the line

a is the intercept of the line

Describe relationship with:

y = a + bx

y is dependent (score)

x is independent (attendance)

y = a + bx

If we have values for a and b, we can predict the value of y for a given value of x

a is the score when x = 0

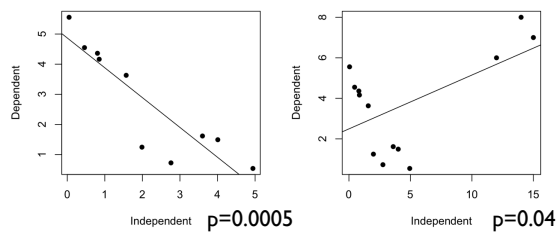b is the increase in y per unit increase in x

In R

Use lm() function

The variables used for the regression must be specified as a formula
dependent~independent

Checking your regression

- Several things you need to check including:
  - Structure in the data
  - Error distribution
  - Variance structure
  - Linearity

Structure in the data

p=0.0005   p=0.04

Error distribution

- Linear regression assumes normal errors

- Check by looking at histogram of residuals or qq-plot

Homogeneity of variance

- Linear regression assumes variance is constant for all values of the independent variable

- Check by looking at a plot of residuals vs fitted values

Linearity

- Linear regression assumes a straight line relationship between variables

- Check by looking at scatterplots and plots of residuals vs fitted values

In R

- To get a histogram of residuals use hist(model$residuals)
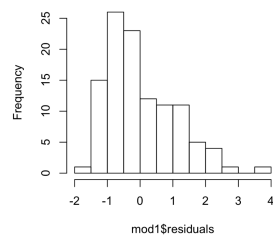
In R
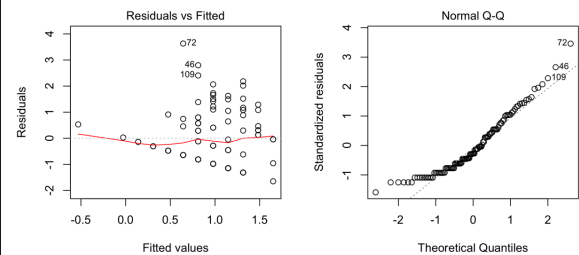
Confidence intervals

Can estimate 95% CI for fitted line

Indicates region we are 95% certain the best fit of the line lies

Prediction intervals

Can calculate 95% PI for estimates of the dependent variable

Indicates region we are 95% certain predictions of the dependent lie

95% PI always exceed CI

Extrapolation

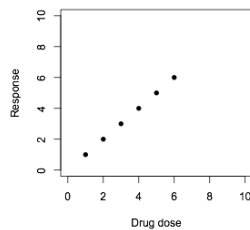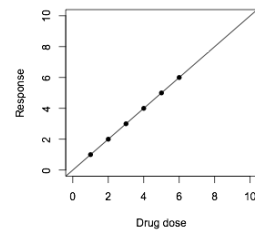Avoid EXTRAPOLATION: estimating dependent variables from a regression equation outside the range of your data
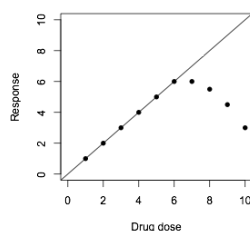
Response

Drug dose

Extrapolation

Avoid EXTRAPOLATION: estimating dependent variables from a regression equation outside the range of your data

Response

Drug dose

Extrapolation

Avoid EXTRAPOLATION: estimating dependent variables from a regression equation outside the range of your data

Response

Drug dose

BBC NEWS UK EDITION

WATCH BBC NEWS IN VIDEO

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend   Printable version

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.

An Oxford University study found that women are running faster than they have ever done over 100m.

Women are set to become the dominant sprinters

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times are also going to improve."

A team led by Dr Tatem, from the Department of Zoology at Oxford University, calculated that by 2156, a woman sprinter could cover the 100m in 8.079 seconds.

That would put women ahead of their male colleagues, who

News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Education
Science/Nature
Technology
Entertainment
Have Your Say
Magazine
In Pictures
Week at a Glance
Country Profiles
In Depth
Programmes

BBC SPORT
BBC WEATHER
CBBC news
BBC ON THIS DAY

LANGUAGES
NEWYDDION

SEE ALSO:
▸ Top sprinters may have key gene
27 Aug 03 | Health
▸ How to eat like an Olympian
20 Aug 04 | Health

RELATED BBC LINKS:
▸ Athletics

TOP UK STORIES NOW
▸ Clarke retreats over house arrest
▸ Terror suspect admits plane plot
▸ Blair bids to woo parental vote
▸ Britain braced for latest freeze

100m sprint time plot

100m sprint time

13
12
11
10
9
8

1900 1950 2000 2050 2100 2150
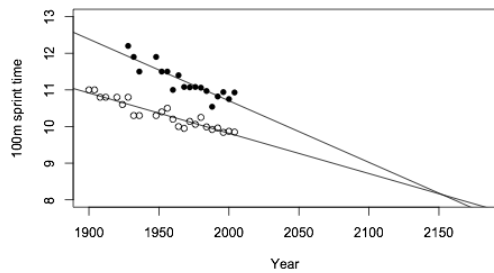
Year

Women's sprint times

- Fitted model is y=44.34-0.001682*year

- Light takes 3.3×10^{-7} seconds to travel 100m

- 3.3×10^{-7}=44.34-0.001682*year

- $(3.3 \times 10^{-7}$-44.34)/-0.001682=2636

- Women will be sprinting at the speed of light in the 2636 olympics.

Comparing regression lines

Several ways

Easiest to use ANCOVA

Use the independent variable as covariate

E.g. compare regressions of male & female scores against attendance

Multiple regression

Can use more than 1 independent variable to predict a dependent

y = a + b1x1 + b2x2

E.g. plant growth is dependent on light and rainfall

Summary

- Correlation and regression are not the same
- Correlation is used to measure the strength and significance of a relationship
- Regression fits a line to your data to allow estimates of the dependent variable to be made from the independent
- For both correlation and regression
    - Data must have normal errors
    - Variances must be similar across the relationship