Imperial College London

Week 2 — Hypothesis testing, linear regression

Introduction to Statistical Thinking and Data Analysis
MSc in Epidemiology / Health Data Analytics
Autumn 2022

17 October 2022

Imperial College London

This week

Time	Session	Topic
Mon 17 Oct 9:30–10:30	Problem Set Review	Problem Set 1: Sampling from a population, summarising data, and the normal distribution
Mon 17 Oct 10:45–12:30	Lecture	Hypothesis testing*, linear regression
Mon 17 Oct 13:30–15:30	Applied Statistics Lab	Project 1: Continuous outcome—developing analysis plan
Wed 19 Oct 9:30–11:00	Small group tutorial (Epi)	Problem Set 2
Thu 20 Oct 15:30–17:00	Small group tutorial (HDA)	Problem Set 2

Imperial College London

Learning objectives

- *Describe the principles of hypothesis testing and specify a 'null' hypothesis.*
- *Articulate, interpret, and report statistical hypothesis tests.*
- *Apply the t-test for comparisons of means from two populations*

- ~~Describe one- and two-way analysis of variance (ANOVA) and its assumptions~~
- ~~Describe applications where ANOVA is appropriate and interpret results~~
- Describe simple linear regression and name the assumptions on which it is based
- Interpret linear regression coefficients, their confidence intervals and significance tests
- Fit linear regression models in R and check the assumptions of the regression model

Imperial College London

Readings

- Kirkwood and Sterne:
  - *Chapter 7: Comparison of two means: confidence intervals, hypothesis tests and P-values*
  - *Chapter 8: Using P-values and confidence intervals to interpret the results of statistical analyses*
  - ~~Chapter 9: Comparison of means from several groups: analysis of variance~~
  - Chapter 10: Linear regression and correlation
  - Chapter 12: Goodness of fit and regression diagnostics

Next week:

- Kirkwood and Stern:
  - Chapter 9: Comparison of means from several groups: analysis of variance
  - Chapter 11: Multiple regression
  - Chapter 12: Goodness of fit and regression diagnostics
  - Chapter 13: Transformations

Imperial College London

Recap: sampling, 95% confidence intervals, Student's t-distribution

Imperial College London

Confidence interval for the sample mean

Research question: What is the average height μ of men in the UK?

- Data: measure height among a random sample of 10 adult men:

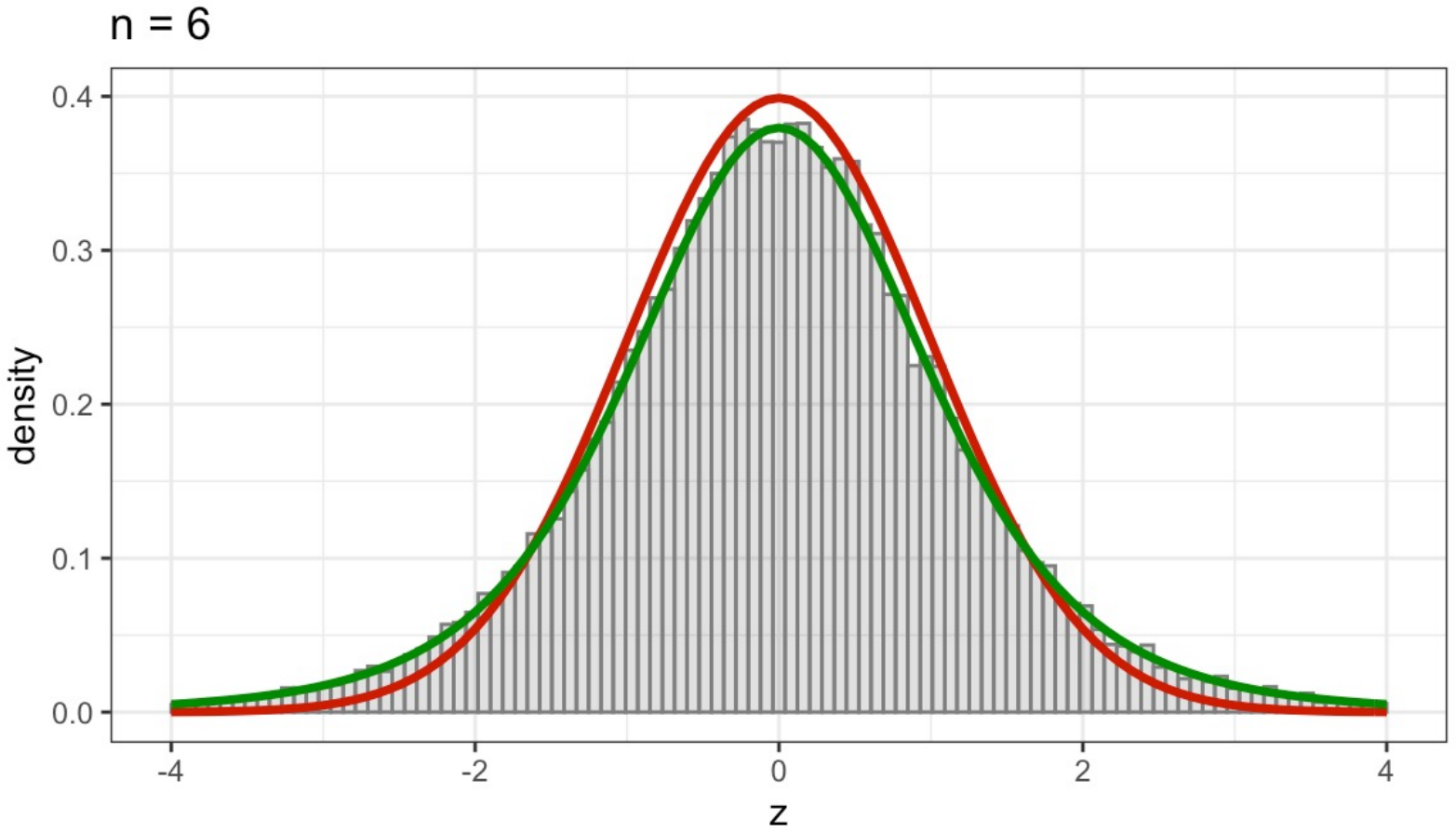
$$x = \{173.1, 168.2, 174.2, 170.6, 167.7, 173.8, 174.6, 169.2, 175.7, 168.8\}$$

- Best estimate of true population mean μ: **Sample mean:** $\bar{x} = \frac{x_1 + \cdots + x_n}{n} = 171.6$
- Estimate of amount of *variation in the population:*
  - Sample variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{RSS}{n-1} = \frac{81.03}{10-1} = 9.00$
  - Sample standard deviation: $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = 3.00$
- *How precise is our estimate $\bar{x}$ of the true population mean value μ?*
  - Expected variation of $\bar{x}$ *around* μ: **Standard error:** $\boldsymbol{s.e.} = \frac{\sigma}{\sqrt{n}} = \frac{3.00}{\sqrt{10}} = 0.95$
- 'Large sample' 95% CI: $\bar{x} \pm 1.96 \times s.e. = 171.6 \pm 1.96 \times 0.95 = (169.73, 173.45)$
- 'Small sample' 95% CI: $\bar{x} \pm t'_{\nu,0.975} \times s.e. = 171.6 \pm 2.26 \times 0.95 = (169.44, 173.74)$ wider

t distribution quartile

Imperial College London

Student's t-disribution



Red: Standard normal distribution

Green: t-distribution with $n - 1$ degrees of freedom

t distribution approximates normal if very large df

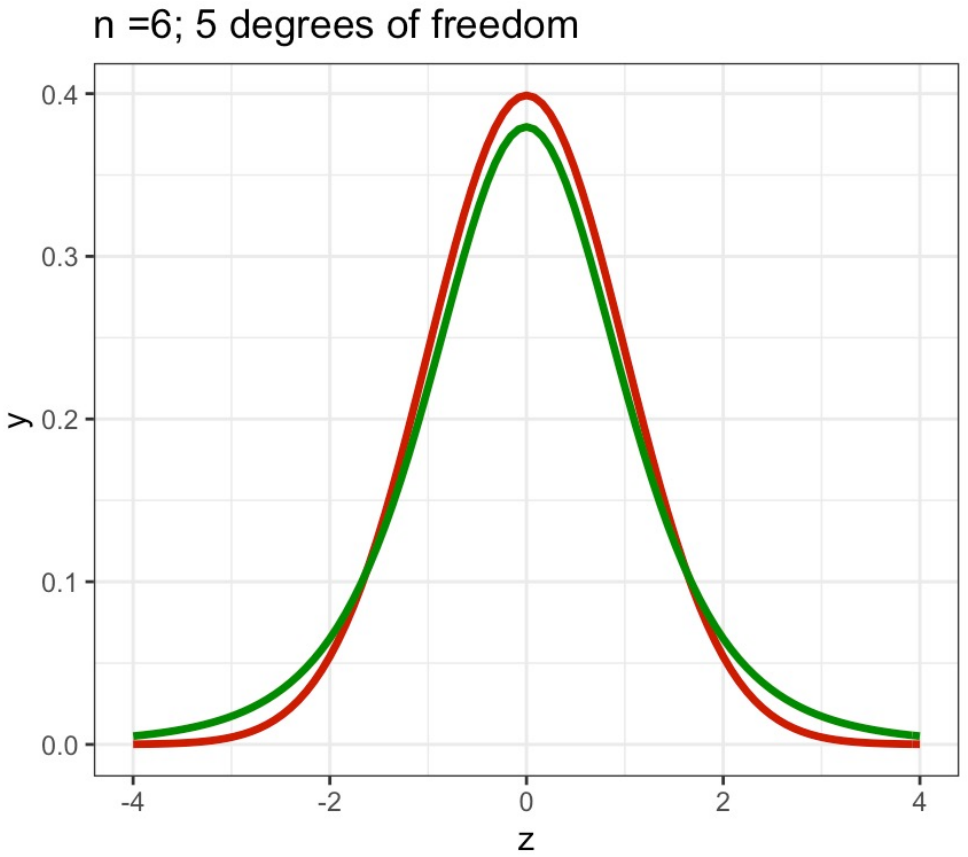
Imperial College London

Small-sample CI: t-distribution

Large-sample 95% CI: $\overline{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$

Small-sample 95% CI: $\overline{x} \pm t'_{\nu,0.975} \cdot \frac{s}{\sqrt{n}}$

t degrees-of-freedom	0.975 quantile
5	2.57
10	2.23
30	2.04
60	2.00
150	1.98
Standard normal	1.96



Imperial College London

Hypothesis testing

Imperial College London

A simple hypothesis

Hypothesis: All adult men in the UK are over 190cm tall.

Two approaches to investigate:

- Prove the hypothesis:
  - Find out the height of every man in the UK and confirm that **all** >190cm.

- Disprove the hypothesis:
  - Find **one** adult male in the UK who is not taller than 190cm.

- Science is the process of systematically disproving hypotheses (Karl Popper).
- Statistical methods formalizes this idea.

Imperial College London

The ‘null’ hypothesis

- **Null hypothesis:** Hypothesis that there is no difference between groups or no association between variables.

Examples:

- H_0: Drug A has no effect on cancer survival compared to Drug B.
- H_0: Using fertilizer does not increase crop yield compared to no fertilizer.

Testing the hypothesis: **Are the data are consistent what we would have been expected *if the null hypothesis were true*?**

- **No:** Data are not consistent with no difference → *Reject* (disprove) the null hypothesis; accept the alternative that there is a difference.
- **Yes:** Data are consistent with no difference → *Fail to reject* the null hypothesis; no evidence that there is a difference.

Imperial College London

The ‘null’ hypothesis

key point

- We can only reject *(disprove)* the null hypothesis.
- We **never** accept *(prove)* the null hypothesis, only *fail to reject*.
  - Other outcomes could be consistent with our data (e.g. a small difference)

that we might not have detected; thus cannot prove

Imperial College London

A more subtle hypothesis

Research question: Is the average height of adult men in the UK greater than 170cm?

opposes the research qn

- **Null hypothesis (H0):** Average height of men in the UK ≤170cm.
- **Alternative hypothesis (H1):** Average height of men in the UK >170cm.

matches the research qn

Data: Measure height of 10 adult men in the UK.

$$\mathrm{x} = 170, 182.9, 182.4, 176.6, 172.4, 180.9, 168.9, 177.8, 159.7, 173.1$$

$$\bar{x} = 174.4$$
$$s = SD(x) = 7.2$$

Is observing a sample mean $\bar{x} = 174.4$ *consistent with* $\mu \leq 170$*?*

null hypothesis true mean

What is the *probability* of observing a sample mean *at least as large* ($\bar{x} \geq 174.4$) if the true average height is 170cm ($\mu = 170$)?

Imperial College London

A more subtle hypothesis

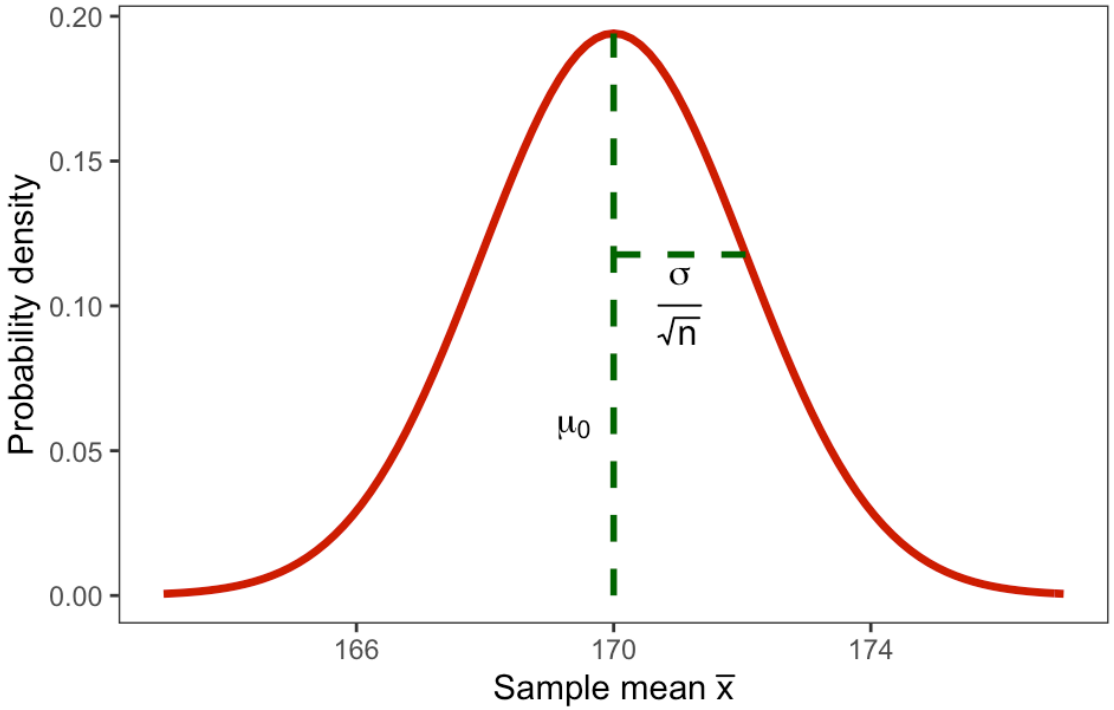
If H_0 is true μ_0 = 170cm:

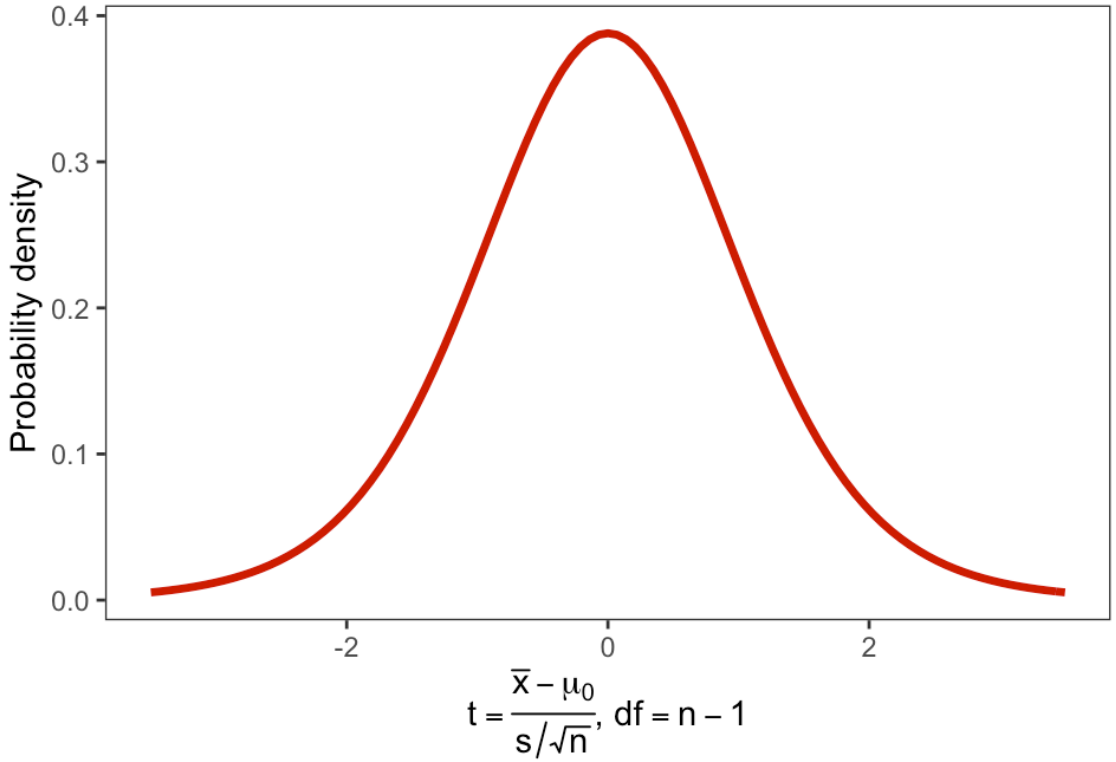
$$\bar{x} \sim \text{Normal}\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$$

z

H_0: $\frac{\mu 0 - 170}{\sigma} = 0$:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim \text{Student}-\text{t}(df = n - 1)$$





Imperial College London

A more subtle hypothesis

H0: $\mu \leq 170$

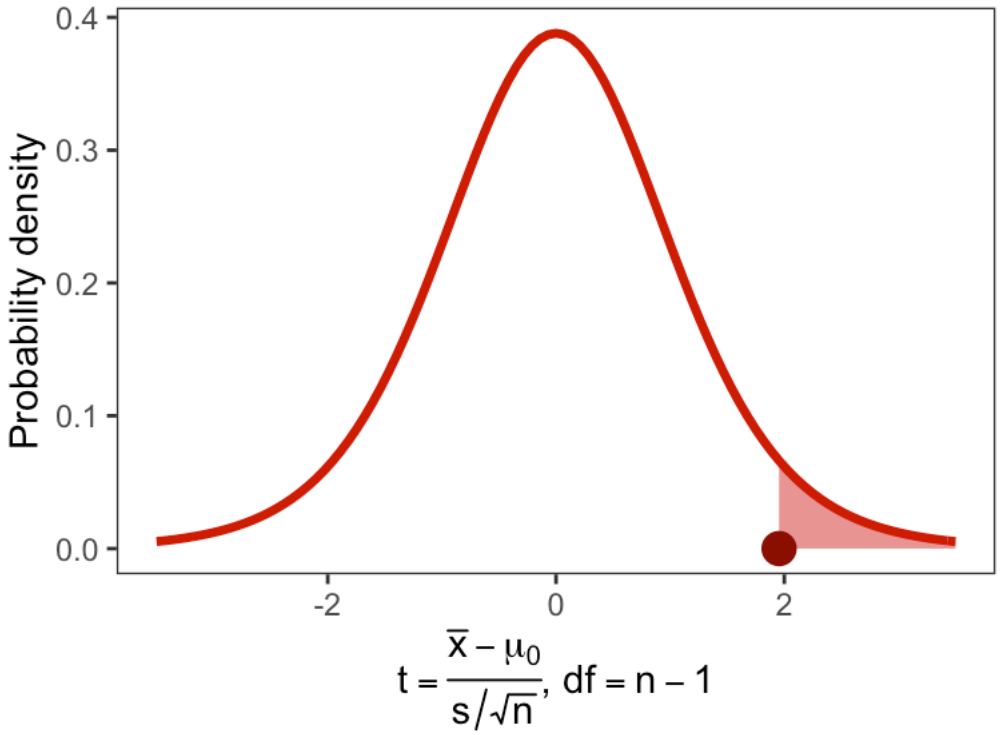
Data: $x = 170, 182.9, 182.4, 176.6, 172.4, 180.9, 168.9, 177.8, 159.7, 173.1$

$\overline{x} = 174.4 \quad s = 7.2$

$$t_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{174.4 - 170}{7.2/\sqrt{10}} = 1.95$$

$P(t \geq t_0 \mid df = 9) = 0.041$

- Probability of observing test statistic at least as large as t_0 if the null hypothesis is true: *p-value.*
- "One-sided" p-value



Imperial College London

Interpreting p-values

- *Test statistic*: Function of sample data measuring how far our observation is from null value of 0.
- *p-value*: Probability observing test statistic as large as t_0 *assuming the null hypothesis is true.*
- Larger test statistic → Smaller *p*-value → stronger evidence *against* the null hypothesis of no association.

Example interpretations:

- p = 0.1: Reasonable chance of observing data if null hypothesis were true → data **do not** provide evidence against the null hypothesis.
- p = 0.001: Very small probability of observing data if null hypothesis were true → data provide **strong evidence** against the null hypothesis.

- Convention: if $p < \alpha$, **reject** the null hypothesis, declare association **statistically significant**.
  - Level α is the pre-determined acceptable level for **rejecting** the null hypothesis when if it is true. (*"Type 1 error rate"*: probability of erroneously detecting an association.)
  - α **must** be determined *a priori* before calculating test statistic. Arbitrary convention: $\alpha = 0.05$
  - Size of *p*-value does not indicate anything about the *magnitude* or medical significance of an association.

Imperial College London

One-sided t-test in R

- **H0:** Average height $\mu \leq 170$.
- **H1:** Average height $\mu > 170$.

```
> x <- c(170.0, 182.9, 182.3, 176.6, 172.4,
         180.9, 168.9, 177.8, 159.6, 173.1)
> xbar <- mean(x);
> s <- sd(x);
> n <- length(x)

> t_stat <- (xbar - 170) / (s / sqrt(n))

> t_stat
[1] 1.952236
```

p value

```
> pt(t_stat, df = n-1, lower.tail = FALSE)
[1] 0.04133462
```

```
> t.test(x, alternative = "greater", mu = 170)

    One Sample t-test

data:  x
t = 1.9522, df = 9, p-value = 0.04133
alternative hypothesis: true mean is greater
than 170
95 percent confidence interval:
 170.2715      Inf
sample estimates:
mean of x
   174.45
```

Imperial College London

Two-sided p-value

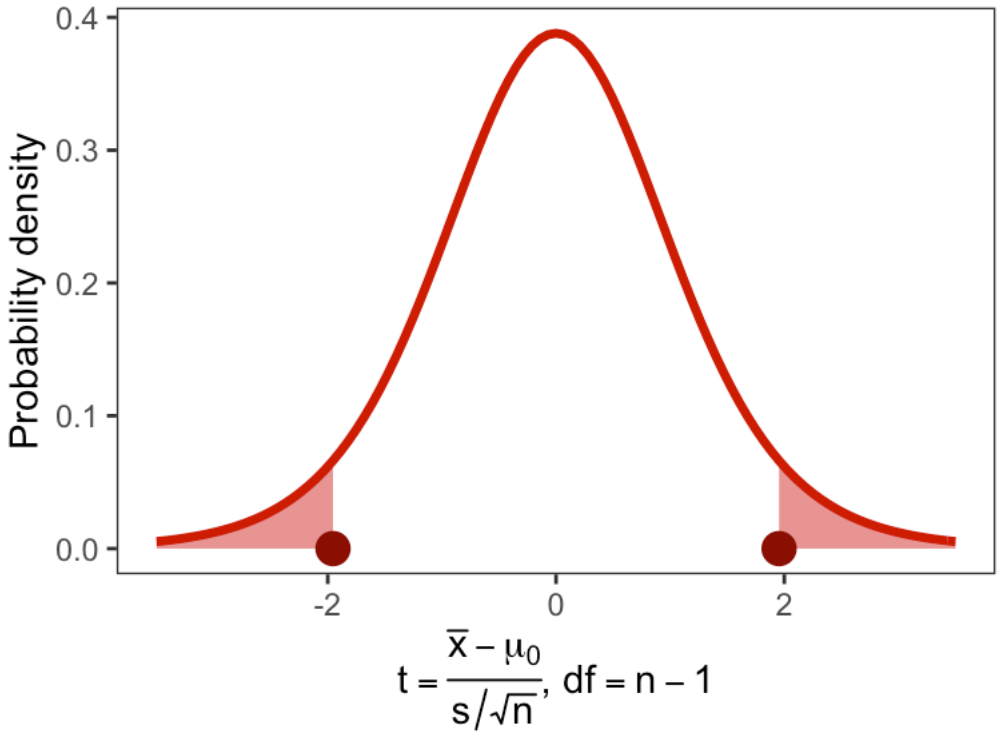
Is the average height of men in the UK exactly equal to 170cm?

- **Null hypothesis:** average height of men in the UK μ = 170cm.
- **Alternative hypothesis:** average height of men in the UK $\mu \neq$ 170cm.

$$t_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{174.4 - 170}{7.2/\sqrt{10}} = 1.95$$

$$\begin{aligned} P(t \geq t0 \; or \; t \leq -t0) &= P(t \geq t0) + P(t \leq -t0) \\ &= 0.041 + 0.041 \\ &= 0.082 \end{aligned}$$

ensure account for both sides!



Imperial College London

Two-sided p-value in R

H0: Average height $\mu = 170$.
H1: Average height $\mu \neq 170$.

```
x <- c(170.0, 182.9, 182.3, 176.6, 172.4,
       180.9, 168.9, 177.8, 159.6, 173.1)
xbar <- mean(x);
s <- sd(x);
n <- length(x)

t_stat <- (xbar - 170) / (s / sqrt(n))

> t_stat
[1] 1.952236

> pt(-t_stat, df = n-1) +
    pt(t_stat, df = n-1, lower.tail = FALSE)
[1] 0.08266924
> 2 * pt(t_stat, df = n-1, lower.tail = FALSE)
[1] 0.08266924
```

```
## t.test(x, alternative = "both", mu = 170)
t.test(x, mu = 170)


> t.test(x, mu = 170)

    One Sample t-test

data:  x
t = 1.9522, df = 9, p-value = 0.08267
alternative hypothesis: true mean is not equal to
170
95 percent confidence interval:
 169.2936 179.6064
sample estimates:
mean of x
   174.45
```

Imperial College London

Comparing means for two groups

- Example: Is there a difference in birthweight between heavy smokers during pregnancy and non-smokers?
- **H0:** There is no difference in the birthweight for children of heavy smokers (μ_1) and children of non-smokers (μ_0).
  - H0: $\mu_1 = \mu_0$
  - H0: $\mu_1 - \mu_0 = 0$

- Data: measured birthweight (kg) among children in 14 heavy smokers and 15 non-smokers.

$$\bar{x}_1 = 3.174 \quad \bar{x}_0 = 3.627$$
$$s_1 = 0.463 \quad s_0 = 0.358$$
$$n_1 = 14 \quad n_2 = 15$$

Table 7.2 Comparison of birth weights (kg) of children born to 14 heavy smokers with those of children born to 15 non-smokers.

Heavy smokers (group 1)	Non-smokers (group 0)
3.18	3.99
2.74	3.89
2.90	3.60
3.27	3.73
3.65	3.31
3.42	3.70
3.23	4.08
2.86	3.61
3.60	3.83
3.65	3.41
3.69	4.13
3.53	3.36
2.38	3.54
2.34	3.51
	2.71

Imperial College London

Comparing means for two groups

- Estimate for magnitude of effect—difference in sample means: estimate for how large difference is

$$\bar{x}_1 - \bar{x}_0 = 3.174 - 3.627 = -0.452 \text{ kg}$$

one option:

- Assume two groups have equal population standard deviation.
  - Calculate combined SD: of variation in birth rate

weighted variance

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}} = \sqrt{\frac{13 \cdot 0.463^2 + 14 \cdot 0.358^2}{14 + 15 - 2}} = 0.412 \text{ kg}$$

dont need to memorize formula

  - Standard error for difference in sample means $\bar{x}_1 - \bar{x}_0$:

$$\text{s. e.} = s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_0}} = 0.412 \cdot \sqrt{\frac{1}{14} + \frac{1}{15}} = 0.153 \text{ kg}$$

- $t = \frac{\bar{x}_1 - \bar{x}_0}{s.e.}$ has t-distribution with $df = n1 + n0 - 2$.
  - 95% CI: $(\bar{x}_1 - \bar{x}_0) \pm t'_{df,0.975} \cdot s.e. = -0.452 \pm 2.052 \cdot 0.153 = (-0.767, -0.138)$
  - $t_0 = -0.452/0.153 = -2.954$
  - *P*-value: $2 \cdot P(t > |t_0|) = 0.006$

Imperial College London

Two sample t-test in R

```
## Kirkwood and Sterne, Example 7.2
smokers <- c(3.18, 2.74, 2.90, 3.27, 3.65, 3.42, 3.23, 2.86, 3.60, 3.65, 3.69, 3.53, 2.38, 2.34)
nonsmokers <- c(3.99, 3.89, 3.60, 3.73, 3.31, 3.70, 4.08, 3.61, 3.83, 3.41, 4.13, 3.36, 3.54, 3.51, 2.71)

xbar1 <- mean(smokers); xbar0 <- mean(nonsmokers)
s1 <- sd(smokers); s0 <- sd(nonsmokers)
n1 <- length(smokers); n0 <- length(nonsmokers)

xbar_diff <- xbar1 - xbar0
s_pooled <- sqrt(((n1-1)*s1^2 + (n0-1)*s0^2) /
                   (n1+n0-2))
se_diff <- s_pooled * sqrt(1/n1 + 1/n0)

t_stat <- (xbar1 - xbar0) / (se_diff)
```

```
> xbar_diff
[1] -0.452381

> ## 95% confidence interval
> xbar_diff + c(-1, 1) * qt(0.975, n1+n0-2) * se_pooled
[1] -0.7666542 -0.1381077
> t_stat
[1] -2.953509

> ## two-sided p-value
> 2 * pt(t_stat, n1+n0-2)
[1] 0.006437305
```

- Estimate that birthweight for babies born to heavy smokers are about 0.45kg lower than birthweight of babies of non-smokers (95% CI 0.14–0.77kg).
- *P*-value $p = 0.006$ indicates we would be very unlikely to observe such a large difference if there was no true difference → strong evidence to reject the null hypothesis of no difference in birthweight.

Imperial College London

Two-sample t-test in R

```
xbar1 <- mean(smokers); xbar0 <- mean(nonsmokers)
s1 <- sd(smokers); s0 <- sd(nonsmokers)
n1 <- length(smokers); n0 <- length(nonsmokers)

xbar_diff <- xbar1 - xbar0
s_pooled <- sqrt(((n1-1)*s1^2 + (n0-1)*s0^2) /
                   (n1+n0-2))
se_diff <- s_pooled * sqrt(1/n1 + 1/n0)

t_stat <- (xbar1 - xbar0) / (se_diff)
> xbar_diff
[1] -0.452381

> ## 95% confidence interval
> xbar_diff +
    c(-1, 1) * qt(0.975, n1+n0-2) * se_pooled
[1] -0.7666542 -0.1381077

> t_stat
[1] -2.953509

> ## two-sided p-value
> 2 * pt(t_stat, n1+n0-2)
[1] 0.006437305
```

```
> t.test(smokers, nonsmokers, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  smokers and nonsmokers
## t = -2.9535, df = 27, p-value = 0.006437
## alternative hypothesis: true difference in means
is not equal to 0
## 95 percent confidence interval:
##  -0.7666542 -0.1381077
## sample estimates:
## mean of x mean of y
##  3.174286  3.626667
```

Imperial College London

Comparing means with unequal population standard deviation

- If believe the **standard deviation** is different between two groups, or large difference in sample standard deviation (e.g. >2x greater), assumption of same standard deviation between groups may be inappropriate.

- Welch's t-test / unequal variance t-test:
  - Denominator *not* based on pooled variance estimate:

$$t = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}}$$

  - Approximation for degrees of freedom:

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_0^4}{n_0^2(n_0 - 1)}}$$

Imperial College London

Comparing means with unequal population standard deviation

- Unequal variance assumption is the default in R when calling `t.test(x, y)`.
  - Argument `t.test(..., var.equal = FALSE)`. default assumes unequal var dist
  - Slightly wider CIs; slightly larger p-value.
  - Performs relatively well even when variance is the same.

Equal variance assumption:

```
> t.test(smokers, nonsmokers, var.equal = TRUE)

   Two Sample t-test

data:  smokers and nonsmokers
t = -2.9535, df = 27, p-value = 0.006437
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
-0.7666542 -0.1381077
sample estimates:
mean of x mean of y
3.174286  3.626667
```

Unequal variance assumption:

```
> t.test(smokers, nonsmokers, var.equal = FALSE)

    Welch Two Sample t-test

data:  smokers and nonsmokers
t = -2.9271, df = 24.489, p-value = 0.007281
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -0.7710160 -0.1337459
sample estimates:
mean of x mean of y
 3.174286  3.62666
```

Imperial College London

Paired t-test

- Data often consist of *paired* observations.
  - Two observations of different exposures on same individual (e.g. treatment A & treatment B).
  - Observations in different groups matched on key characteristics (e.g. age, sex).

- Calculate *difference* within each pair.
  - Differences become observations.
  - Sample size: number of pairs.
  - Paired t-test = one-sample t-test on differences

- Typically, pairing **reduces** standard error for difference → reduce *p*-value compared to two-sample test.
  - If pairs are 'well matched'.
  - Each pair acts as it's own 'control': removes other sources of variation → Isolates effect treatment difference.

Imperial College London

Paired t-test example

- Example: Does new sleeping drug improve number of hours of sleep per night?
- **H0:** No difference in the average number of hours sleep per night for patients who receive the drug compared to placebo.

- **Data:** Ten patients observed one night with drug, one night with placebo. Outcome is number of hours slept.

Table 7.3 Results of a placebo-controlled clinical trial to test the effectiveness of a sleeping drug.

	Hours of sleep		
Patient	Drug	Placebo	Difference
1	6.1	5.2	0.9
2	6.0	7.9	−1.9
3	8.2	3.9	4.3
4	7.6	4.7	2.9
5	6.5	5.3	1.2
6	5.4	7.4	−2.0
7	6.9	4.2	2.7
8	6.7	6.1	0.6
9	7.4	3.8	3.6
10	5.8	7.3	−1.5

- Estimated magnitude of effect:

$$\overline{\text{drug}} = 6.66 \qquad \overline{\text{placebo}} = 5.58$$

$$\overline{\text{drug}} - \overline{\text{placebo}} = 1.08$$

- Standard error:

$$s = SD(diff) = 2.31$$

$$n = 10 \quad df = n - 1 = 9$$

$$\text{s. e.} = \frac{2.31}{\sqrt{10}} = 0.73$$

s/sqrt(n)

t val

- 95% CI: $1.08 \pm 2.26 \cdot 0.73 = (-0.57, 2.73)$
- Test statistic and p-value:

$$t_0 = \frac{1.08}{0.73} = 1.48$$

p-value = 0.173

Imperial College London

Paired t-test in R

```
> sleep <- data.frame(patient = 1:10,          ## Kirkwood and Sterne, Example 7.3
                      drug = c(6.1, 6.0, 8.2, 7.6, 6.5, 5.4, 6.9, 6.7, 7.4, 5.8),
                      placebo = c(5.2, 7.9, 3.9, 4.7, 5.3, 7.4, 4.2, 6.1, 3.8, 7.3))


> diff <- sleep$drug - sleep$placebo
> xbar_diff <- mean(diff)  # equivalent: xbar1 - xbar0
> se_diff <- sd(diff) / sqrt(nrow(sleep))

> t_stat
[1] 1.479529

> 2 * pt(abs(t_stat), df=nrow(sleep)-1, lower.tail = FALSE)
[1] 0.1731279

## Built-in R function
> t.test(sleep$drug, sleep$placebo, paired = TRUE)
...
```

Imperial College London

Paired t-test in R

Equivalent results: paired t-test and one-sample t-test on differences

```
> t.test(sleep$drug, sleep$placebo, paired = TRUE)

    Paired t-test

data:  sleep$drug and sleep$placebo
t = 1.4795, df = 9, p-value = 0.1731
alt. hypothesis: true difference != 0
95 percent confidence interval:
 -0.5712886  2.7312886
sample estimates:
mean of the differences
                   1.08
```

```
> t.test(sleep$drug - sleep$placebo)

    One Sample t-test

data:  sleep$drug - sleep$placebo
t = 1.4795, df = 9, p-value = 0.1731
alt. hypothesis: true difference != 0
95 percent confidence interval:
 -0.5712886  2.7312886
sample estimates:
mean of x
     1.08
```

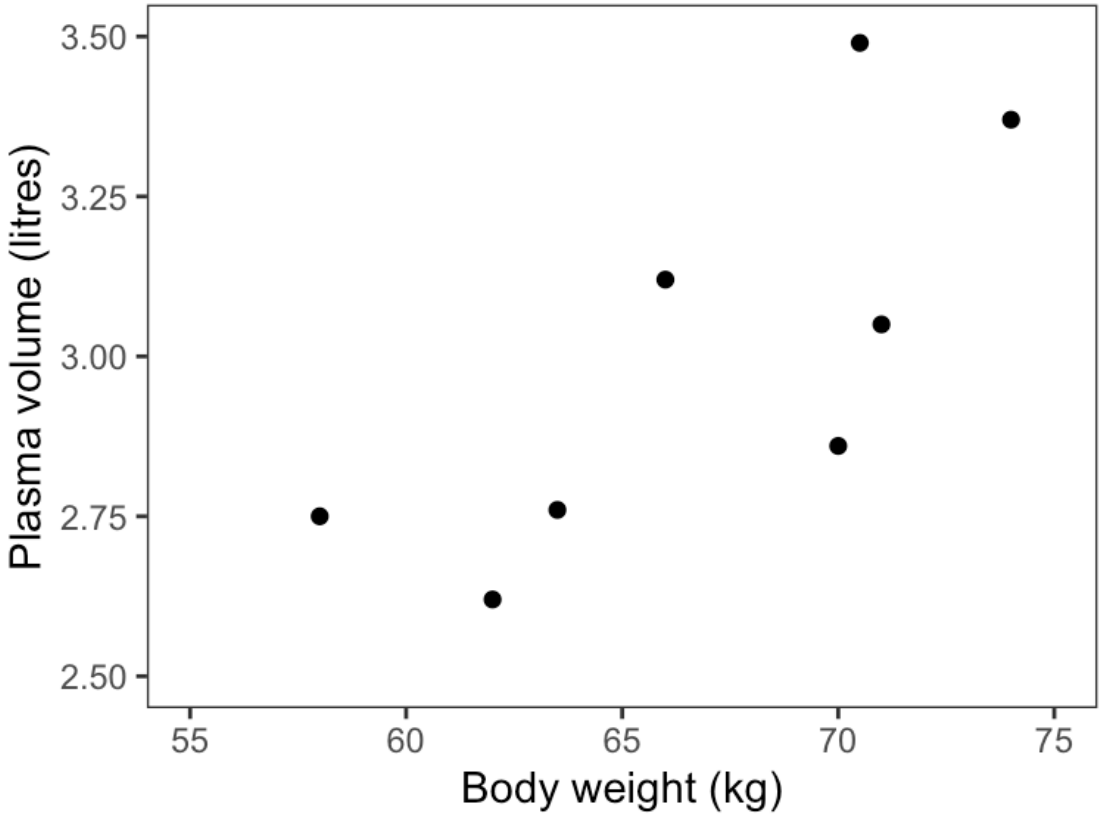
Linear regression

Imperial College London

Linear regression

Subject	Body weight (kg)	Plasma volume (litres)
1	58.0	2.75
2	70.0	2.86
3	74.0	3.37
4	63.5	2.76
5	62.0	2.62
6	70.5	3.49
7	71.0	3.05
8	66.0	3.12

Measurements of body weight and blood plasma volume in eight healthy men.

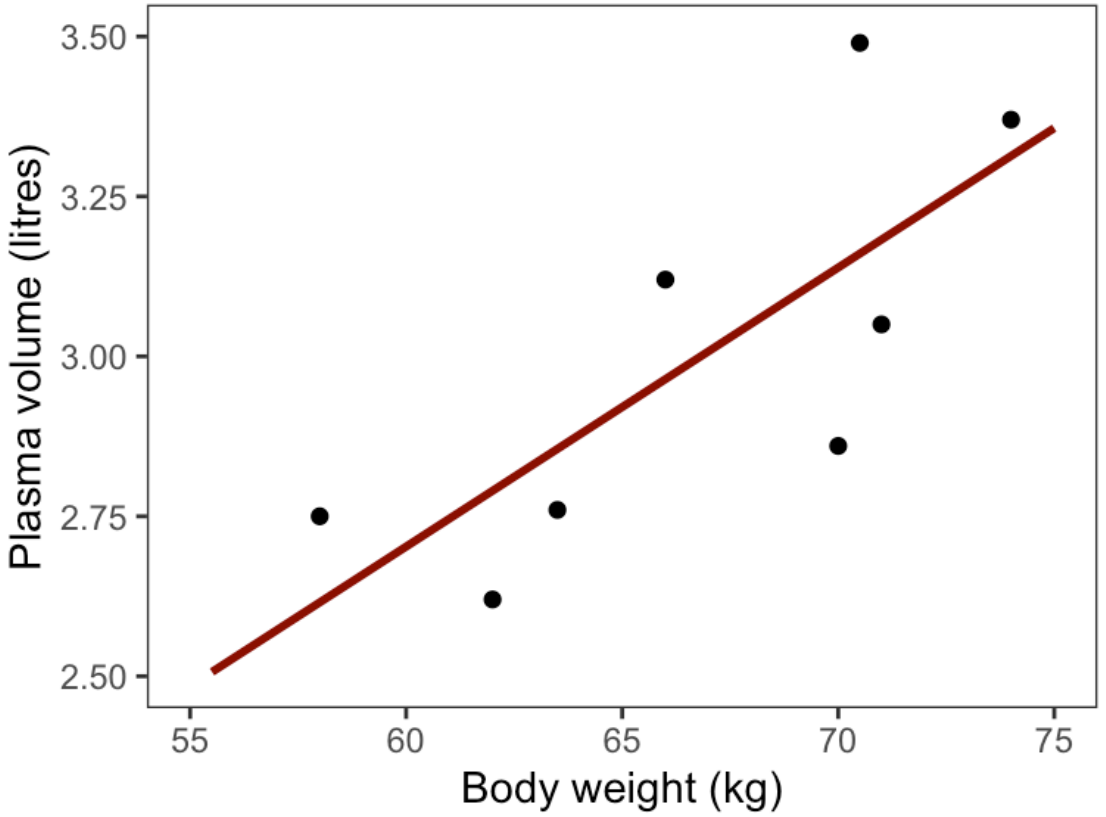


Imperial College London

Linear regression

Subject	Body weight (kg)	Plasma volume (litres)
1	58.0	2.75
2	70.0	2.86
3	74.0	3.37
4	63.5	2.76
5	62.0	2.62
6	70.5	3.49
7	71.0	3.05
8	66.0	3.12

Measurements of body weight and blood plasma volume in eight healthy men.



Imperial College London

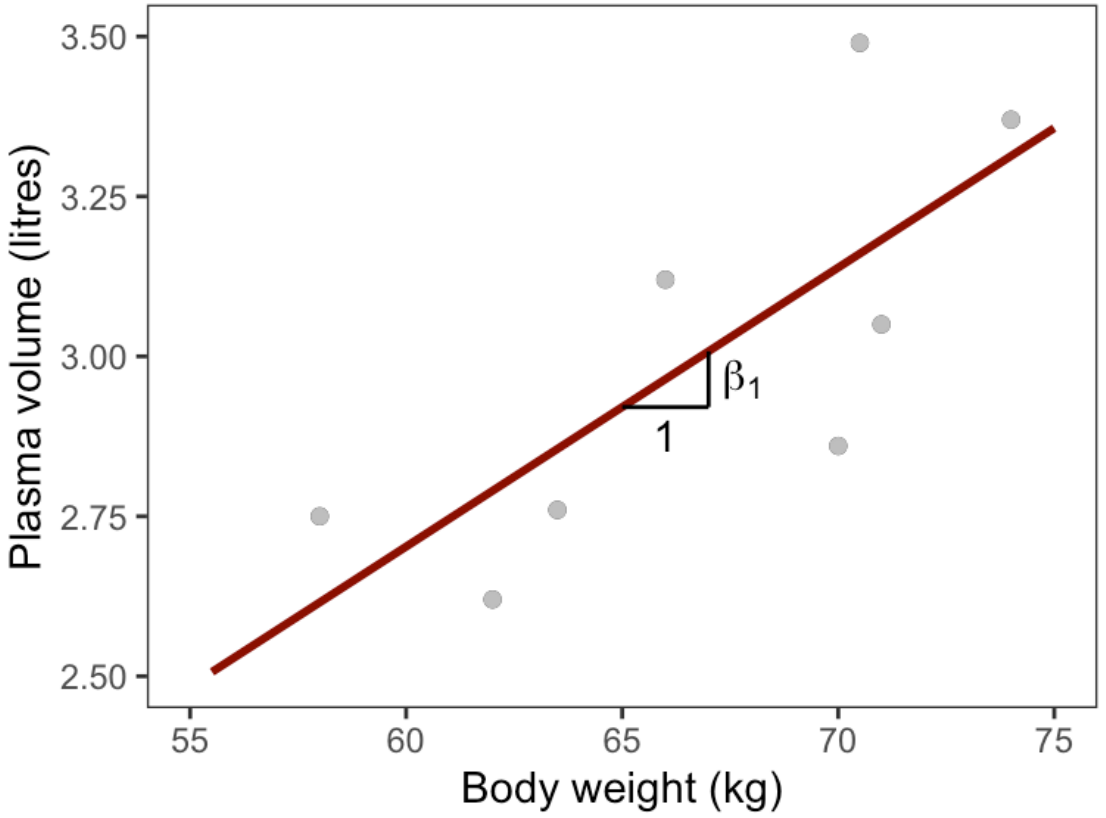
Linear regression

two parameters

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

- β_0 and β_1 are called *regression coefficients*
  - β_0 = intercept: value of y when $x = 0$
  - β_1 = slope: change in y per 1 unit change in x

Measurements of body weight and blood plasma volume in eight healthy men.



Imperial College London

Estimating regression parameters

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}(0, \sigma)$$

Estimating parameters via least squares:

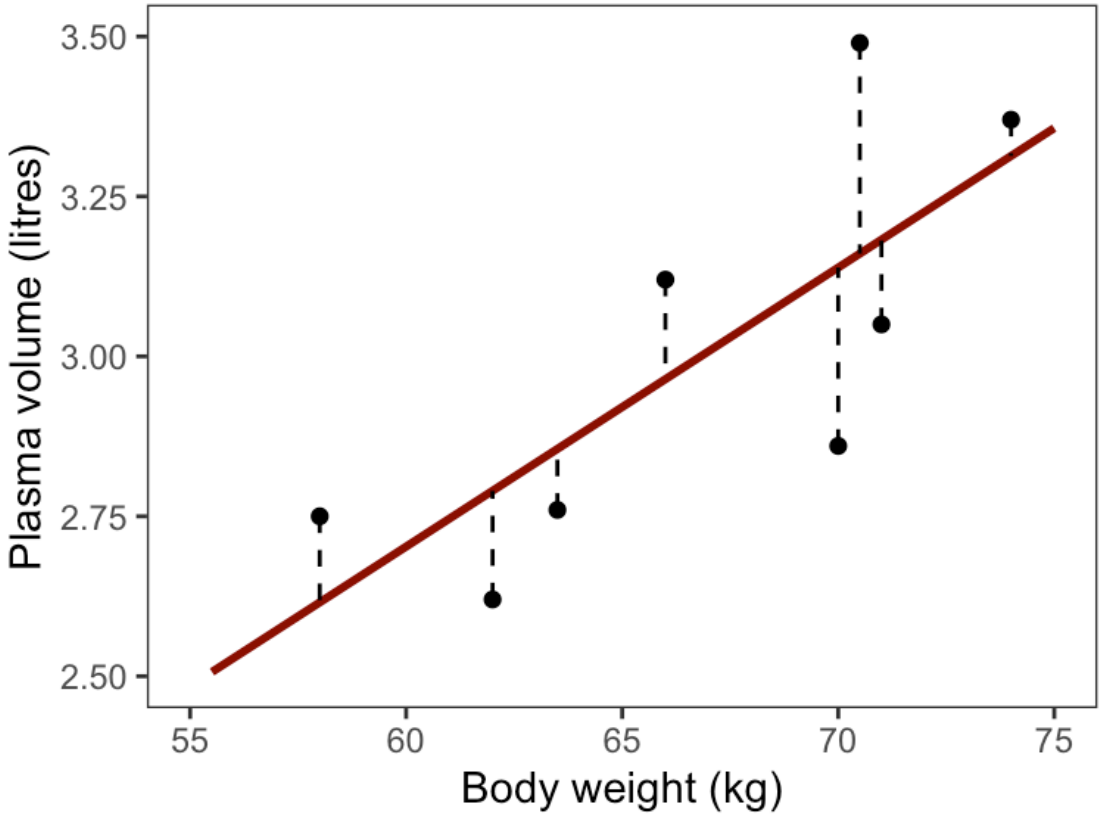
- Find values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the *residual sum of squares*

$$RSS = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

- Satisfied by

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Measurements of body weight and blood plasma volume in eight healthy men.



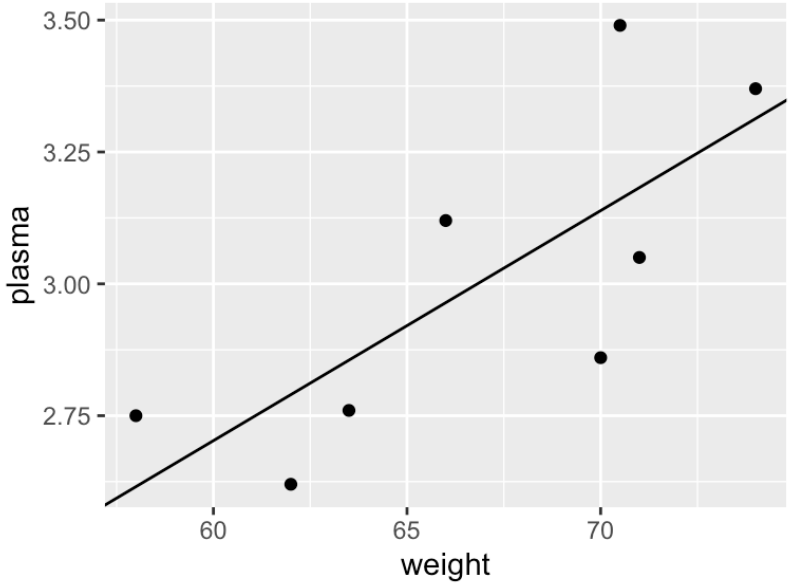
Imperial College London

Linear regression in R

```
pv <- data.frame(subject = 1:8,
                 weight = c(58, 70, 74, 63.5, 62, 70.5, 71, 66),
                 plasma = c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12))

xbar <- mean(pv$weight)
ybar <- mean(pv$plasma)

beta1_hat <- sum((pv$weight - xbar) * (pv$plasma - ybar)) / sum((pv$weight - xbar)^2)
beta0_hat <- ybar - beta1_hat * xbar
```

```
> beta0_hat
## [1] 0.08572428

> beta1_hat
## [1] 0.04361534
```

```
ggplot(pv, aes(x = weight, y = plasma)) +
geom_point() +
geom_abline(slope = beta1_hat,
            intercept = beta0_hat)
```



Imperial College London

Linear regression in R

```
pv <- data.frame(subject = 1:8,
                 weight = c(58, 70, 74, 63.5, 62, 70.5, 71, 66),
                 plasma = c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12))

xbar <- mean(pv$weight)
ybar <- mean(pv$plasma)

beta1_hat <- sum((pv$weight - xbar) * (pv$plasma - ybar)) / sum((pv$weight - xbar)^2)
beta0_hat <- ybar - beta1_hat * xbar
```

- Best fitting regression line: **Plasma volume = 0.086 + 0.044 x weight**
- Interpretation: *Among our sample of eight men,* for every 1 kilogram greater weight, plasma volume increases by an average of 0.044 litres.

- What about the association between weight and plasma volume *among all men in the population?*

Imperial College London

Standard error of regression coefficients

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}(0, \sigma)$$

- Estimate of σ depends on the *residual sum of squares* (RSS):

$$\hat{\sigma} = s = \sqrt{\frac{RSS}{df}} = \sqrt{\frac{\sum\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2}{n-2}}$$

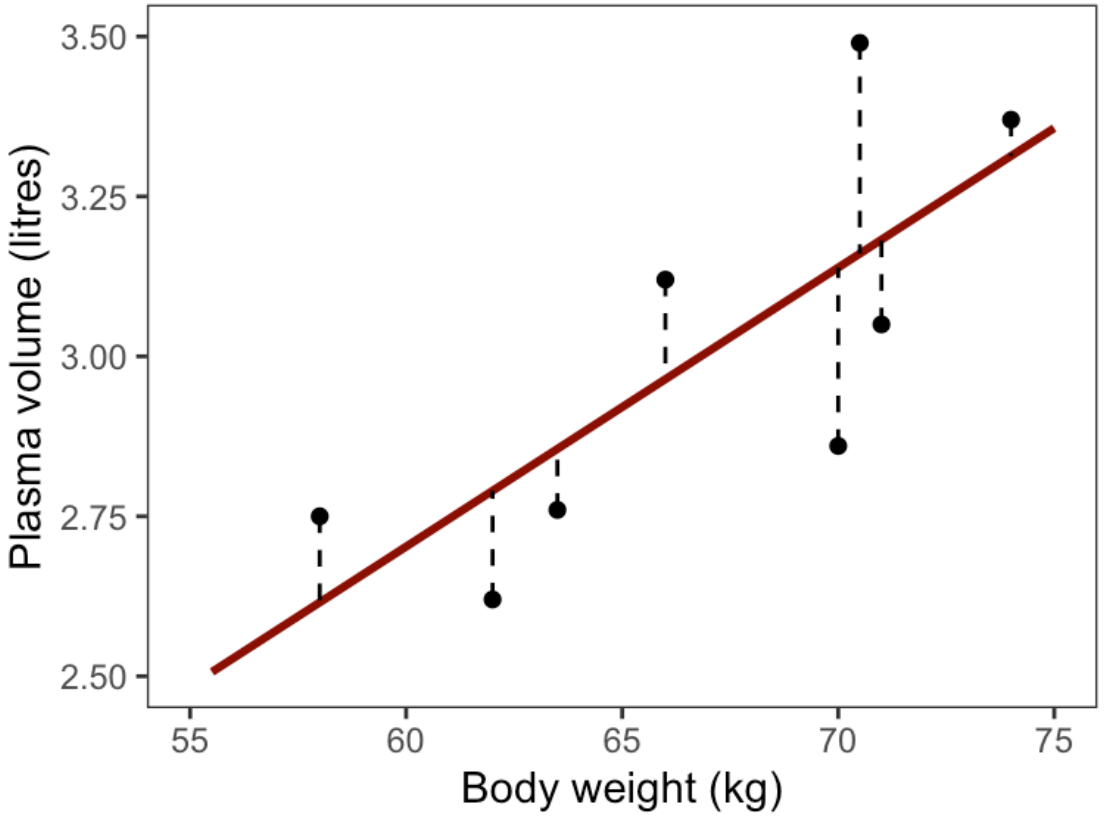
- $df = n - k$: Observations - parameters

- Standard error of regression coefficients:

$$\text{s.e.}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \overline{x})^2}}$$

$$\text{s.e.}(\hat{\beta}_0) = \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}}$$

Measurements of body weight and blood plasma volume in eight healthy men.



Imperial College London

Linear regression in R

```
pv <- data.frame(subject = 1:8,
                 weight = c(58, 70, 74, 63.5, 62, 70.5, 71, 66),
                 plasma = c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12))
xbar <- mean(pv$weight)
ybar <- mean(pv$plasma)

beta1_hat <- sum((pv$weight - xbar) * (pv$plasma - ybar)) / sum((pv$weight - xbar)^2)
beta0_hat <- ybar - beta1_hat * xbar

y_pred <- beta0_hat + beta1_hat * pv$weight
sd sigma_hat <- sqrt( sum((pv$plasma - y_pred)^2) / (nrow(pv) - 2))

se_beta1 <- sigma_hat / sqrt(sum((pv$weight - xbar)^2))
se_beta0 <- sigma_hat * sqrt(1/nrow(pv) + xbar^2 / sum((pv$weight - xbar)^2))
```

	Estimate	S.E.
$\hat{\beta}_0$	0.0857	1.0240
$\hat{\beta}_1$	0.0436	0.0153
$\hat{\sigma}$	0.2188	

Imperial College London

Hypothesis testing for regression coefficients

- Is plasma volume associated with weight *(among adult men in the population)*?
  - H0: No association between plasma volume and weight ⇔ $\beta_1 = 0$
  - H1: Plasma volume is associated with weight ⇔ $\beta_1 \neq 0$
- Familiar looking ingredients…:

	Estimate	S.E.
$\hat{\beta}_0$	0.0857	1.0240
$\hat{\beta}_1$	0.0436	0.0153
$\hat{\sigma}$	0.2188	

df = 6

- 95% CI for β_1: $\hat{\beta}_1 \pm t'_{df,0.975} \times \text{s.e.}(\hat{\beta}_1)$ = 0.044 ± 2.447 x 0.015 = (0.006, 0.081)
- t-statistic: $t_0 = \frac{\hat{\beta}_1}{\text{s.e.}(\hat{\beta}_1)} = \frac{0.0436}{0.0153}$ = 2.86 on 6 degrees of freedom
- Two-sided p-value: 0.0289 (t-distribution, 6-degrees of freedom)

Imperial College London

Linear regression in R

```
fit <- lm(plasma ~ weight, data = pv)
summary(fit)

## Call:
## lm(formula = plasma ~ weight, data = pv)
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.27880 -0.14178 -0.01928 0.13986 0.32939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0 08572    1.02400   0.084   0.9360
## weight       0.04362    0.01527   2.857   0.0289 *
## ---
## Signif. Codes   0 '*' 0.001 '' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2188 on 6 degrees of freedom
## Multiple R-squared: 0.5763, Adjusted R-squared: 0.5057
## F-statistic: 8.16 on 1 and 6 DF, p-value: 0.02893
```

two sided

Imperial College London

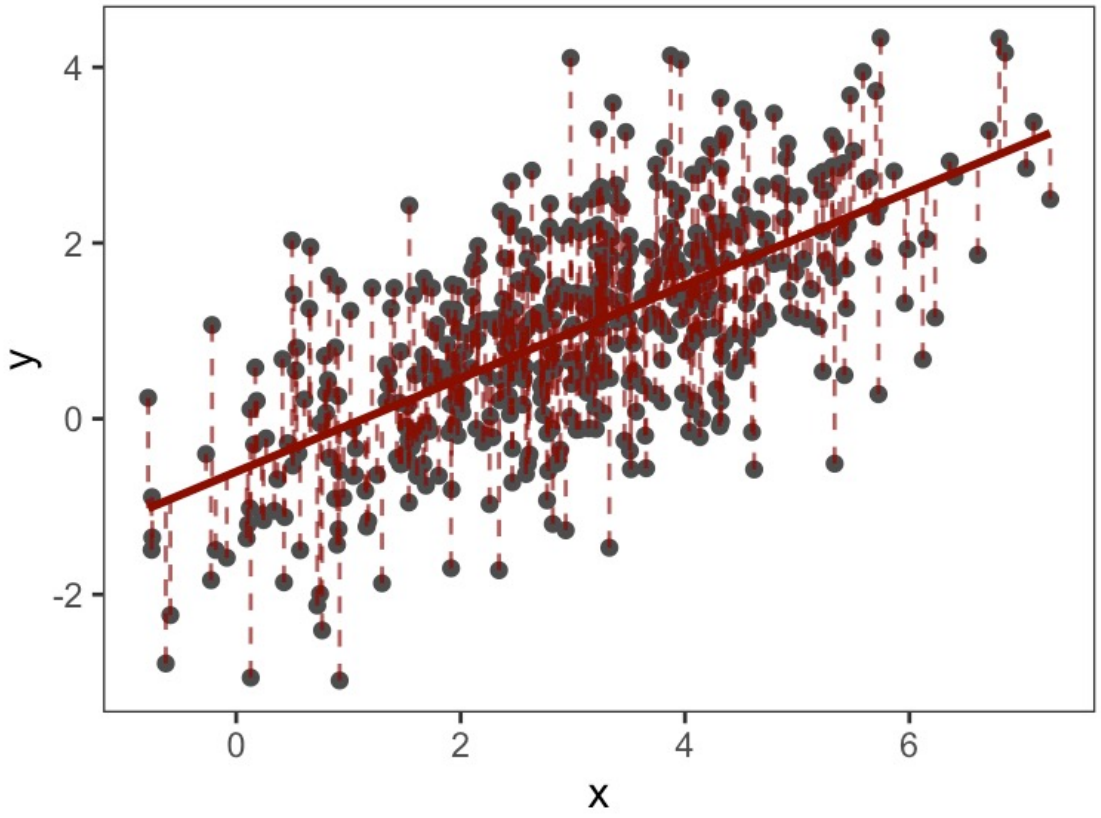
Linear regression assumptions

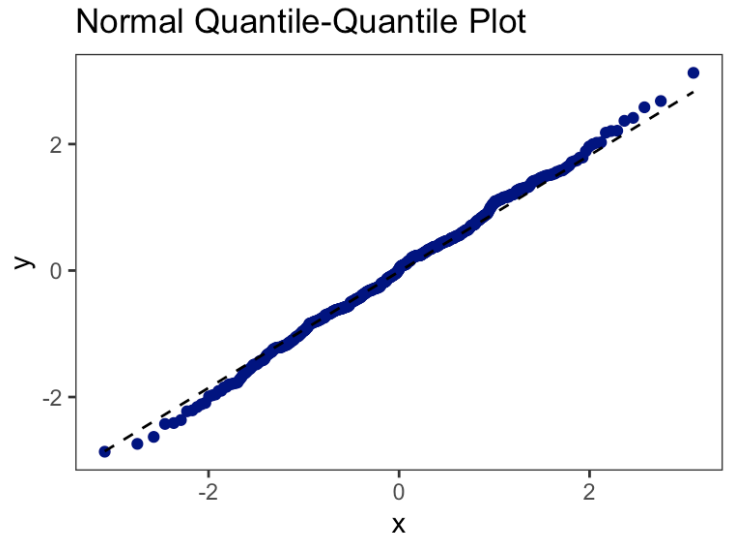
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}(0, \sigma)$$

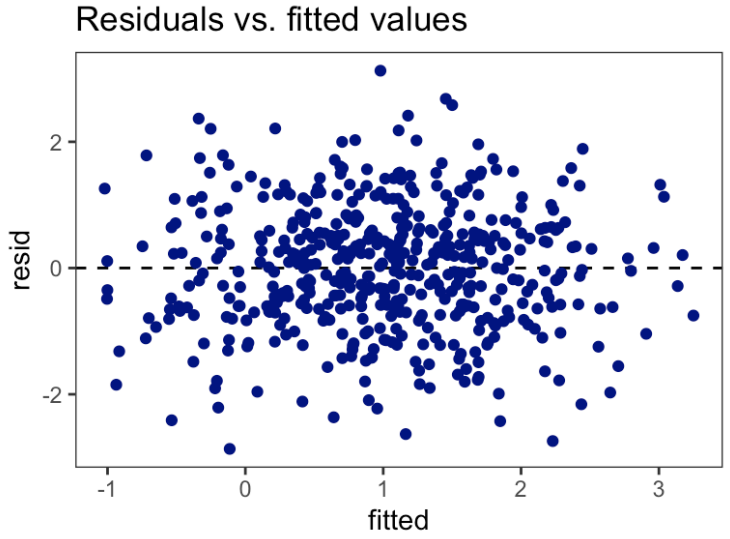
1. **Linearity:** Relationship between y and x is linear.
2. **Independence:** Residuals (ϵ_i) are uncorrelated.
3. **Normality:** *Residuals* ϵ_i follow a normal distribution.
4. **Constant variance:** Residuals ϵ_i have constant variance for all values of y.

Imperial College London

Checking assumptions: residual analysis



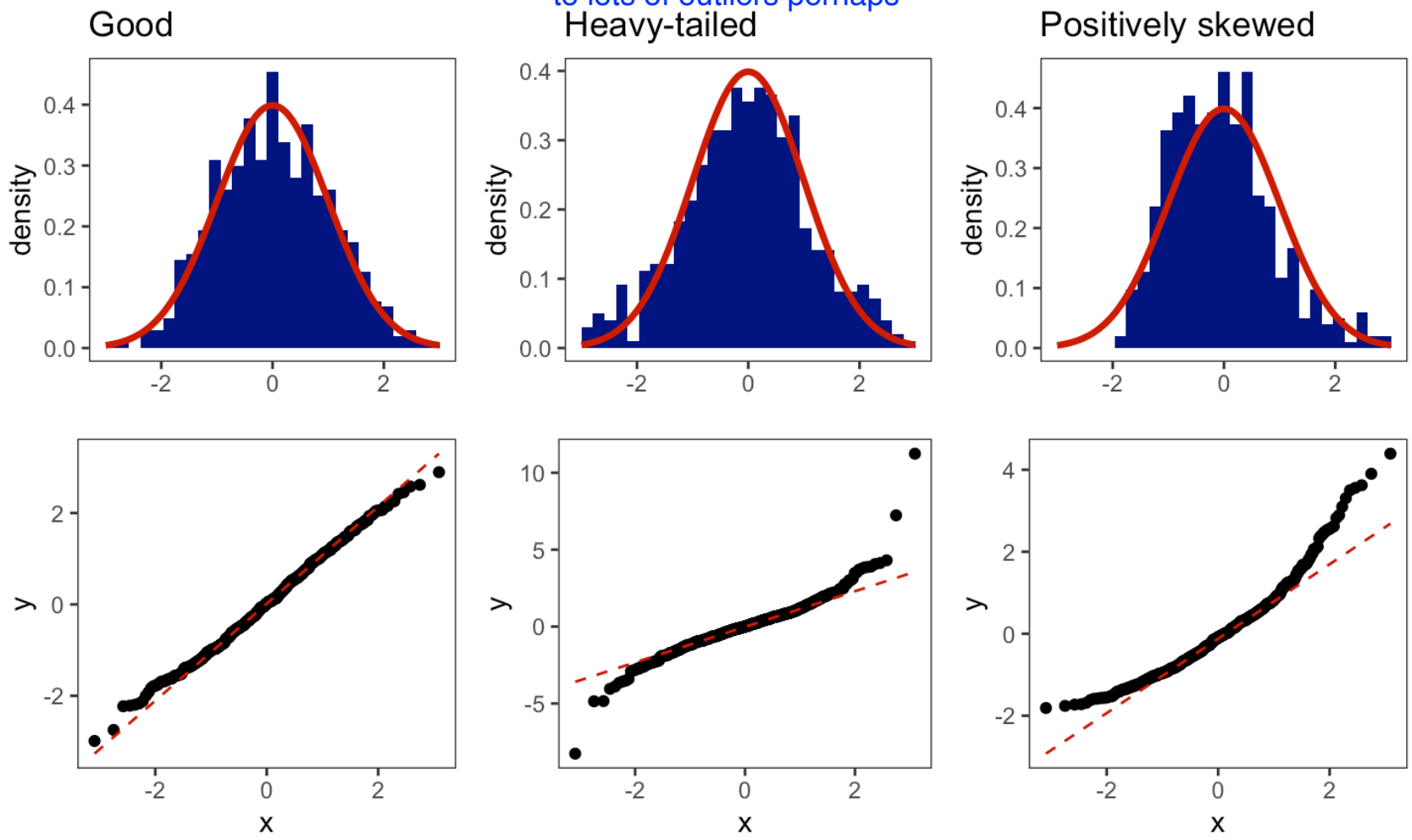




Imperial College London

Normal Q-Q plot: checking normality

more extreme values in residuals than expected due to lots of outliers perhaps

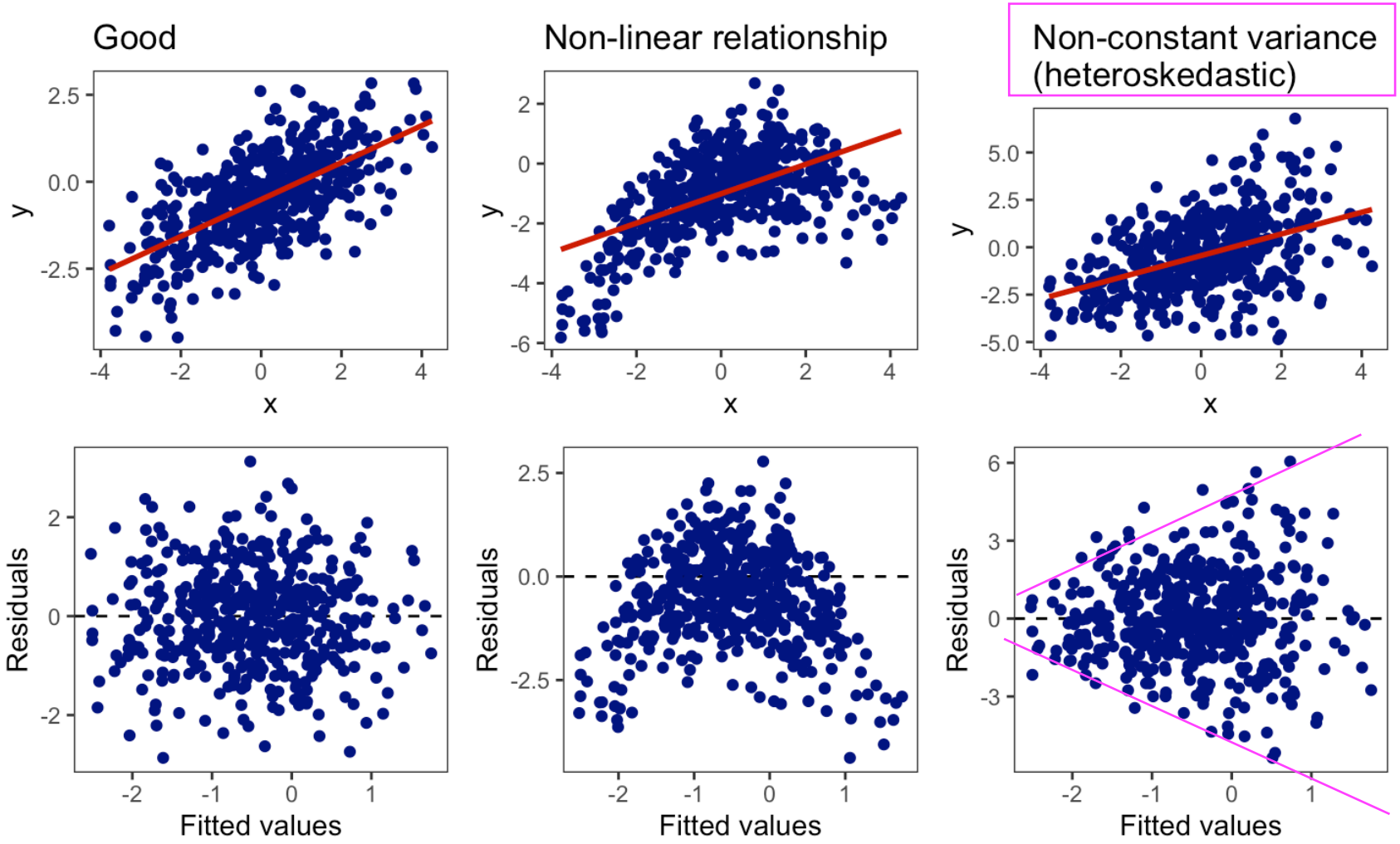


u shaped curve

Imperial College London

predicted y for observed x

Residuals vs. fitted values

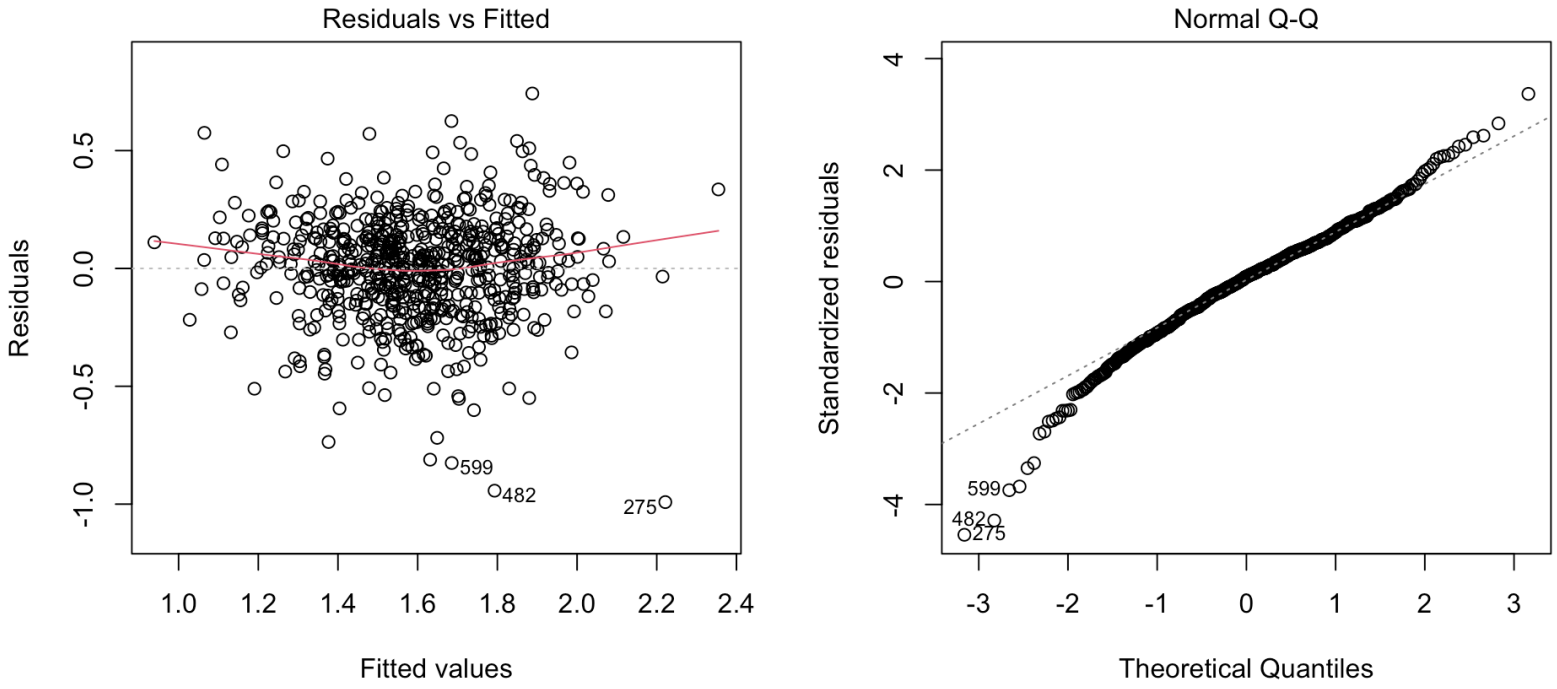


funnel like plot

Imperial College London

Residuals analysis in R

```
perulung <- read.csv("perulung_ems.csv")
fit <- lm(fev1 ~ height + age + sex, data = perulung)
par(mfrow = c(1, 2))
plot(fit, 1:2)
```



Any questions?