Imperial College London

Week 3 — Multiple linear regression, transformations, and model building

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

24 October 2022

Imperial College London

This week

Time	Session	Topic
Mon 24 Oct 9:30–10:30	Problem Set Review	Problem Set 2: Linear regression
Mon 24 Oct 10:45–12:30	Lecture	Multiple linear regression, transformations, and model building
Mon 24 Oct 13:30–15:30	Applied Statistics Lab	Project 1: performing the analysis
Wed 26 Oct 9:30–11:00	Small group tutorial (Epi)	Problem Set 3
Wed 26 Oct 15:30–17:00	Small group tutorial (HDA)	Problem Set 3

Imperial College London

Announcements

- We plan to start posting email questions + responses also on Blackboard discussion board when they may be of wider interest
- We will aim to post problem set solutions at end of Friday (instead of Sunday)
- Working on creating annotated R script examples for problem set solutions

Imperial College London

Learning objectives

- Define and interpret the correlation coefficient r and r^2, and understand the difference between the correlation coefficient and regression coefficient.
- Use multiple regression to describe, to adjust, and to predict.
- Interpret interactions and decide when to include interaction terms in multiple regression.
- Know why and when to transform outcome and exposure variables in linear regression.
- Develop a model building strategy to identify which and how many variables to include in a model.

Readings

- Kirkwood and Sterne:
  - Chapter 10: Linear regression and correlation
  - Chapter 11: Multiple regression
  - Chapter 12: Regression diagnostics
  - Chapter 13: Transformations
  - Chapter 29: Regression modelling
    - Sections 29.5, 29.7, 29.8
  - Chapter 38: Strategies for analysis
    - Sections 38.5–38.8

Analysis of Variance (ANOVA)

Imperial College London

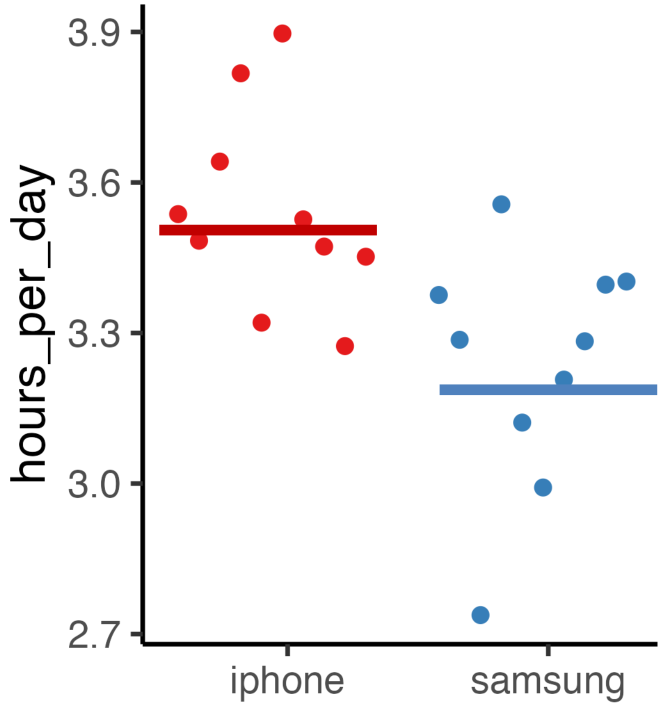
Analysis of variance

How many hours per day do users of different phone manufacturers spend on TikTok?

- Sample n = 10 iPhone users and n = 10 Samsung users

	Mean
iPhone	3.54
Samsung	3.24

- Difference: 0.3
- Pooled SE: 0.097
- t-statistic: 3.149 on 18 degrees of freedom
- p-value: 0.005

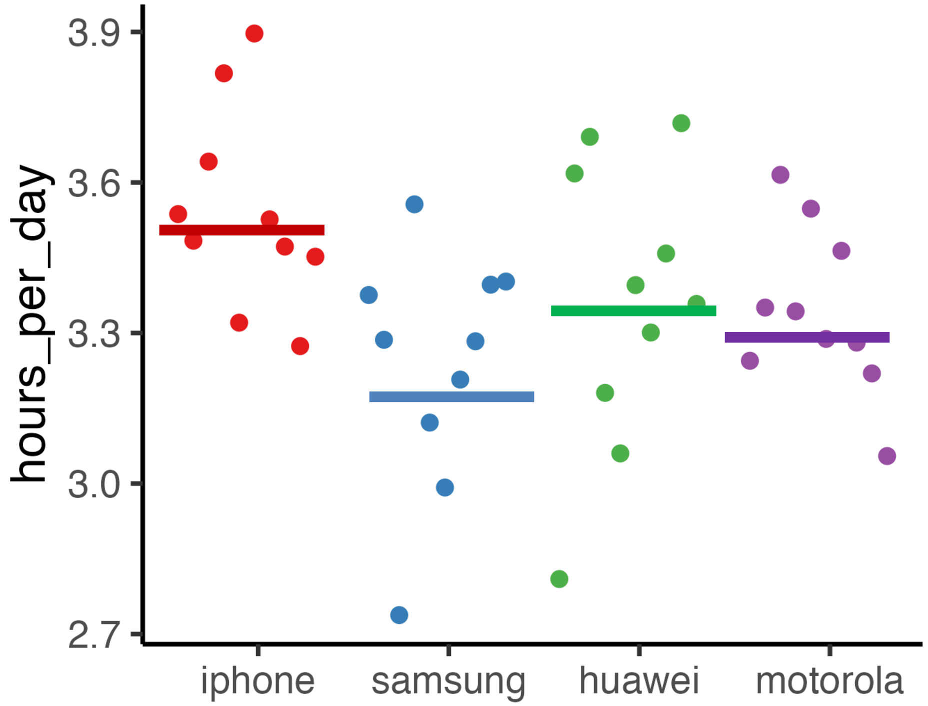


Imperial College London

Analysis of variance

Is the mean number of hours per day on TikTok different by users of different leading phone manufacturers?

- Sample n = 10 of each iPhone, Samsung, Huawei, and Motorola users



Mean hours per day:	3.54 hr	3.24 hr	3.36 hr	3.34 hr

Option: Linear regression w/ categorical

```
> summary(lm(hours_per_day ~ manufacturer, data = phones))

Call:
lm(formula = hours_per_day ~ manufacturer, data = phones)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.54936 -0.10078 -0.00318  0.14492  0.35915 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           3.54223    0.07151  49.538  < 2e-16 ***
manufacturersamsung  -0.30628    0.10112  -3.029  0.00452 ** 
manufacturerhuawei   -0.18321    0.10112  -1.812  0.07837 .  
manufacturermotorola -0.20136    0.10112  -1.991  0.05408 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2261 on 36 degrees of freedom
Multiple R-squared:  0.2089,	Adjusted R-squared:  0.143 
F-statistic: 3.168 on 3 and 36 DF,  p-value: 0.03594
```

- Informs about mean hours per day vs. iPhone users
- Does not answer question is there a difference in the mean across all four groups

Imperial College London

Analysis of variance

Is the mean number of hours per day on TikTok different by users of different leading phone manufacturers?

$$s^2 = \frac{SS}{df} = \frac{\sum(x-\bar{x})^2}{n-1}$$

- Decompose the variance & degrees of freedom:

Group: squared differences between group mean and overall population mean

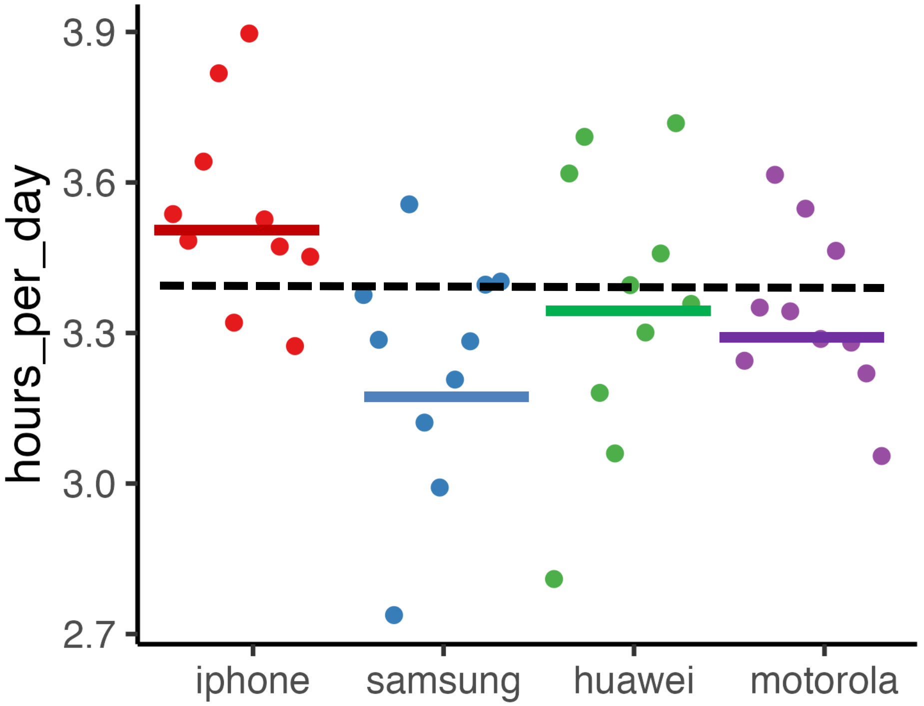
$$s^2 = \frac{SS}{df} = \frac{SS_{\text{group}} + SS_{\text{res}}}{df_{\text{group}} + df_{\text{res}}}$$

Residual: difference between each observation and group mean

sum squares/df

	Sum Sq.	d.f.	Mean Sq.
Group	0.486	4 groups - 1 3	0.162
Residual	1.841	36	0.051
Total	2.327	39	

Sample n = 10 of each iPhone, Samsung, Huawei, and Motorola users



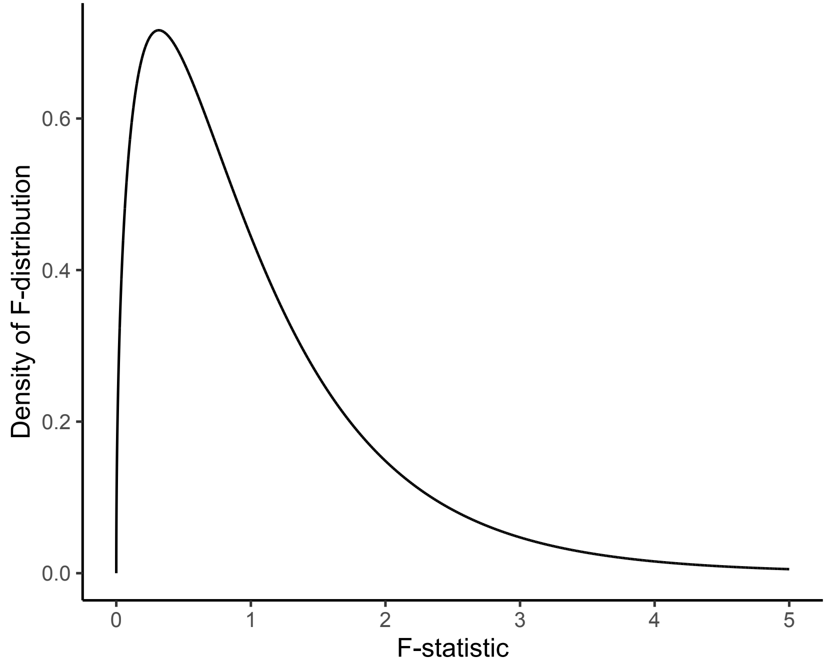
Mean hours per day: **3.54 hr** **3.24 hr** **3.36 hr** **3.34 hr**

F-statistic

$$F = \frac{\text{Between-group mean square}}{\text{Within-group (residual) mean square}}$$

- With degrees of freedom:
  - df1 = between-group d.f.
  - df2 = within-group d.f. residual group freedomo
- *Assuming null hypothesis that all means are equal*, F-statistic follows an F-distribution
- P-value based on tail of the F-disitribution

F-distribution with df_1=3 and df_2=36 degrees of freedom



	Sum Sq.	d.f.	Mean Sq.	F-stat	p-value
Group	0.486	3	0.162	3.168	0.0356
Residual	1.841	36	0.051		
Total	2.327	39			

Imperial College London

F-test in R

```
> fit <- lm(hours_per_day ~ manufacturer, data = phones)

> anova(fit)

Analysis of Variance Table

Response: hours_per_day
             Df Sum Sq Mean Sq F value  Pr(>F)
manufacturer  3 0.4860 0.16200  3.1684 0.03594 *
Residuals    36 1.8407 0.05113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Imperial College London

ANOVA in linear regression output

```
Call:
lm(formula = hours_per_day ~ manufacturer, data = phones)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.54936 -0.10078 -0.00318  0.14492  0.35915 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           3.54223    0.07151  49.538  < 2e-16 ***
manufacturersamsung  -0.30628    0.10112  -3.029  0.00452 ** 
manufacturerhuawei   -0.18321    0.10112  -1.812  0.07837 .  
manufacturermotorola -0.20136    0.10112  -1.991  0.05408 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2261 on 36 degrees of freedom
Multiple R-squared:  0.2089,	Adjusted R-squared:  0.143 
F-statistic: 3.168 on 3 and 36 DF,  p-value: 0.03594
```

same as anova output

Imperial College London

Correlation and coefficient of determination (R^2)

Imperial College London

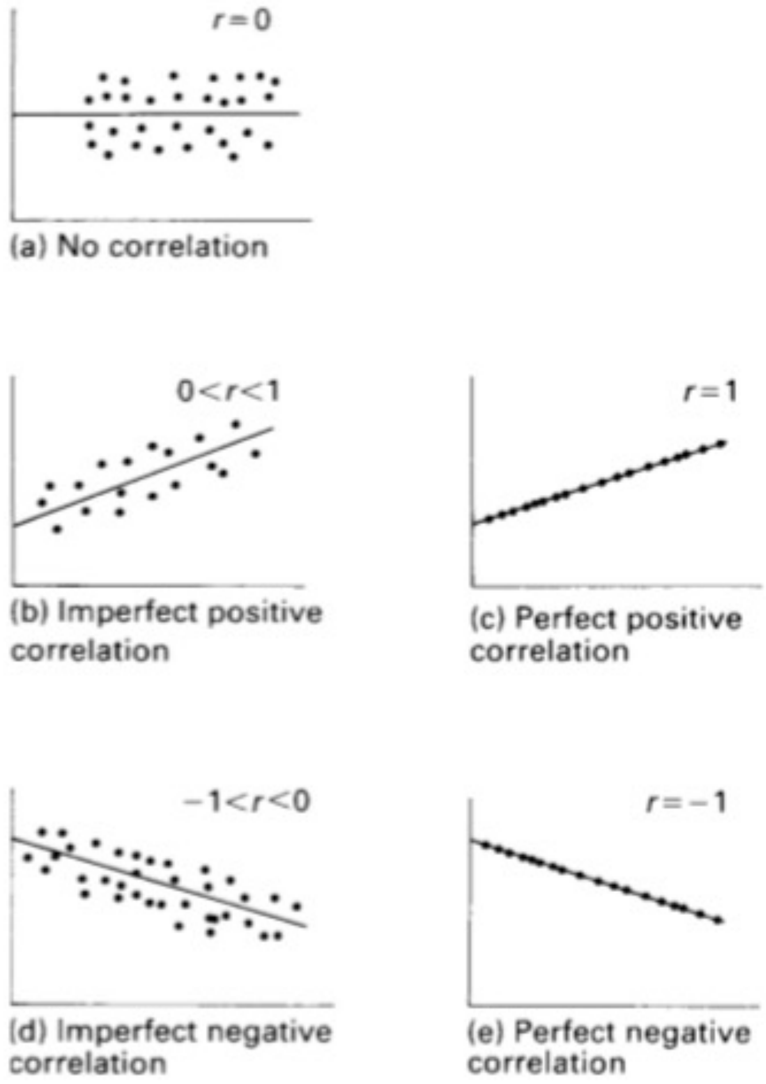
Correlation coefficient

- Regression coefficient (β_1): average change in outcome y for a one unit change in x.
  - Depends on the units of x and y

- Correlation coefficient (r): *strength* of linear association between two variables.
  - Number of standard deviations that y changes for each standard deviation change in x.

- r is between -1 and 1
  - $r = 0$ → no correlation; regression slope also = 0
  - $0 < r < 1$ → positive correlation; $\beta_1 > 0$
  - $-1 < r < 0$ → negative correlation; $\beta_1 < 0$

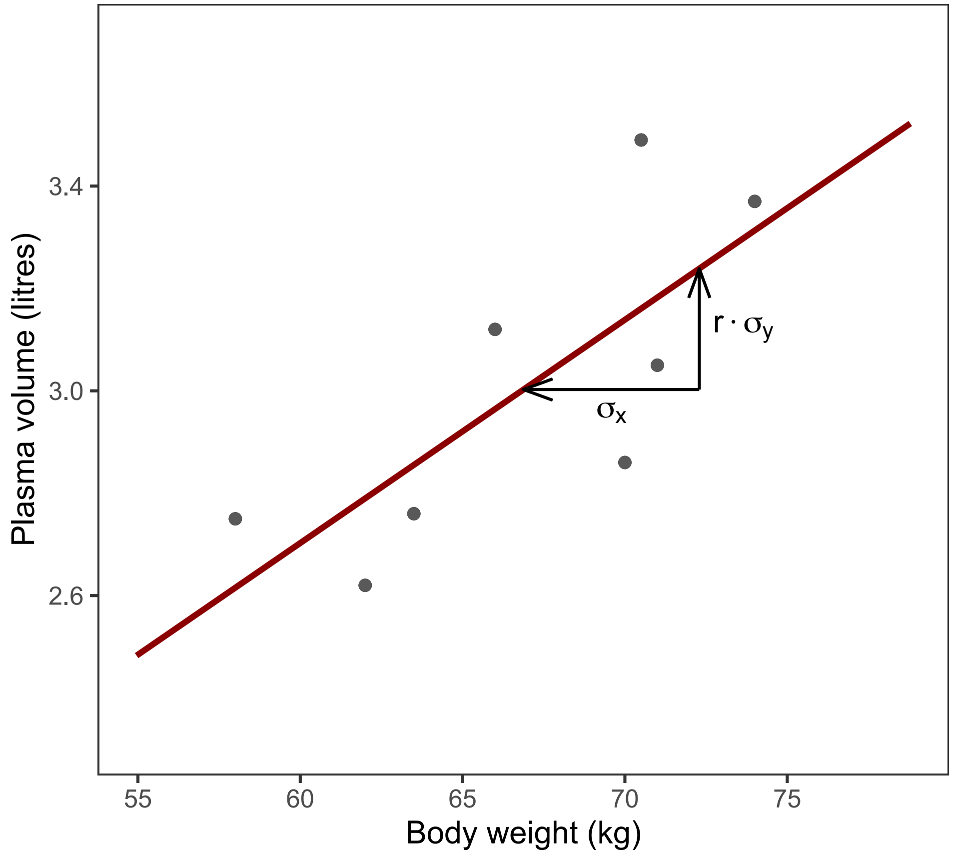
- Two datasets can have the same β_1, but different r.
  - More spread around the regression line → lower absolute correlation.
  - If we standardize x and y, then $\beta_1 = r$
  - But uncommon; usually most interested in interpreting β_1 on natural scale.



Imperial College London

Correlation coefficient

Subject	Weight (kg)	Plasma volume (L)
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



mean(Weight) = 66.88 SD(Weight) = 5.42
mean(Plasma) = 3.003 SD(Plasma) = 0.311
$r = 0.759$

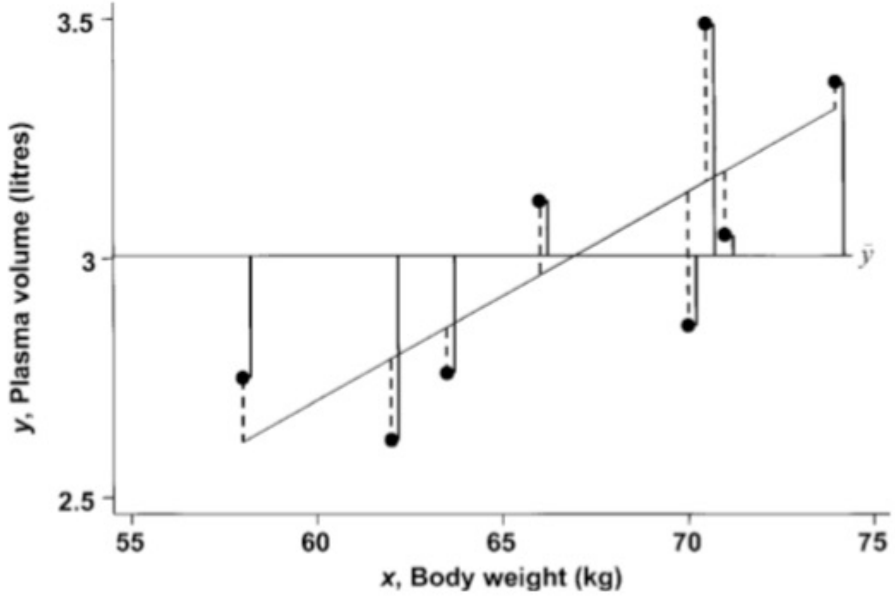
Plasma = 0.0857 + 0.0436 * Weight
SD(Weight) * β_1 = 5.42 * 0.0436 = 0.236
SD(Plasma) * r = 0.311 * 0.759 = 0.236

Coefficient of determination

- **Coefficient of determination—$\boldsymbol{R^2}$:** measures of the proportion of the total variation in the data that has been explained by the regression.
  - Always between 0 and 1.

- Defined by Analysis of Variance (ANOVA):

$$R^2 = \frac{SS_{Total} - SS_{Residual}}{SS_{Total}} = \frac{SS_{Regression}}{SS_{Total}}$$



- For simple linear regression, coefficient of determination is equal to square of correlation (r^2).

- R^2 is distinct from hypothesis testing for regression coefficients.
  - High / low R^2 does not mean a model is 'good'/'bad'; they R^2 and inference for β's answer different questions.

Correlation in R

```
> dat <- 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))

> x <- dat$weight
> y <- dat$plasma
> xbar <- mean(x)
> ybar <- mean(y)

> r <- sum((x - xbar) * (y - ybar)) /
+       sqrt(sum((x-xbar)^2) * sum((y - ybar)^2))
> r
[1] 0.7591266

> cor(dat$weight, dat$plasma)
[1] 0.7591266

## Coefficient of determination
> r^2
[1] 0.5762732
```

```
> summary(lm(plasma ~ weight, data = dat))

Call:
lm(formula = plasma ~ weight, data = dat)

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
```

- Body weight accounts for 57.6% of the total variation in plasma volume.

interpretation of r2

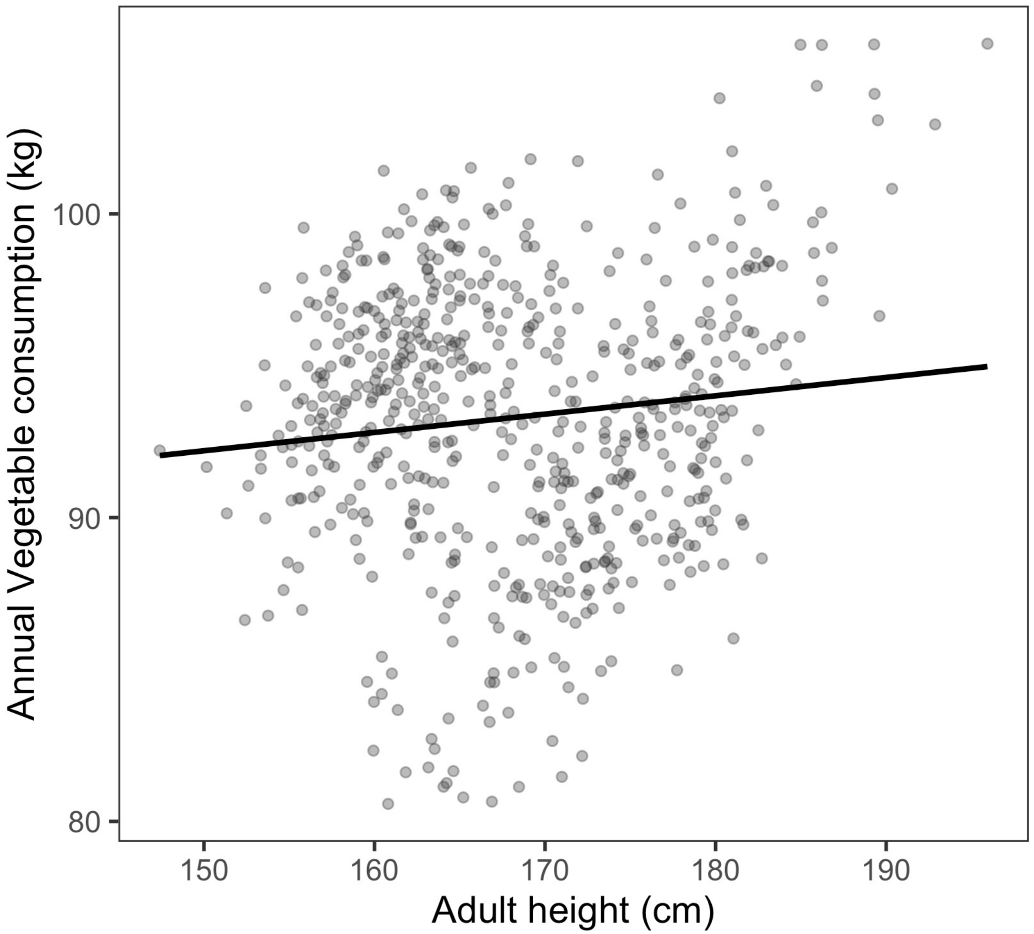
Imperial College London

Multiple linear regression

Example: height and vegetable consumption

<u>Data:</u> random sample of 600 adults in the UK

$$\text{Veg cons.} = \beta_0 + \beta_{height} \times \text{height} + \epsilon$$



Param.	Estim.	Std. err.	t-value	p-value	95% CI
β_0	83.1503	3.6202	22.97	<0.001	(76.04–99.26)
β_{height}	0.0603	0.0214	2.81	0.005	(0.018–0.103)
R^2	0.013				

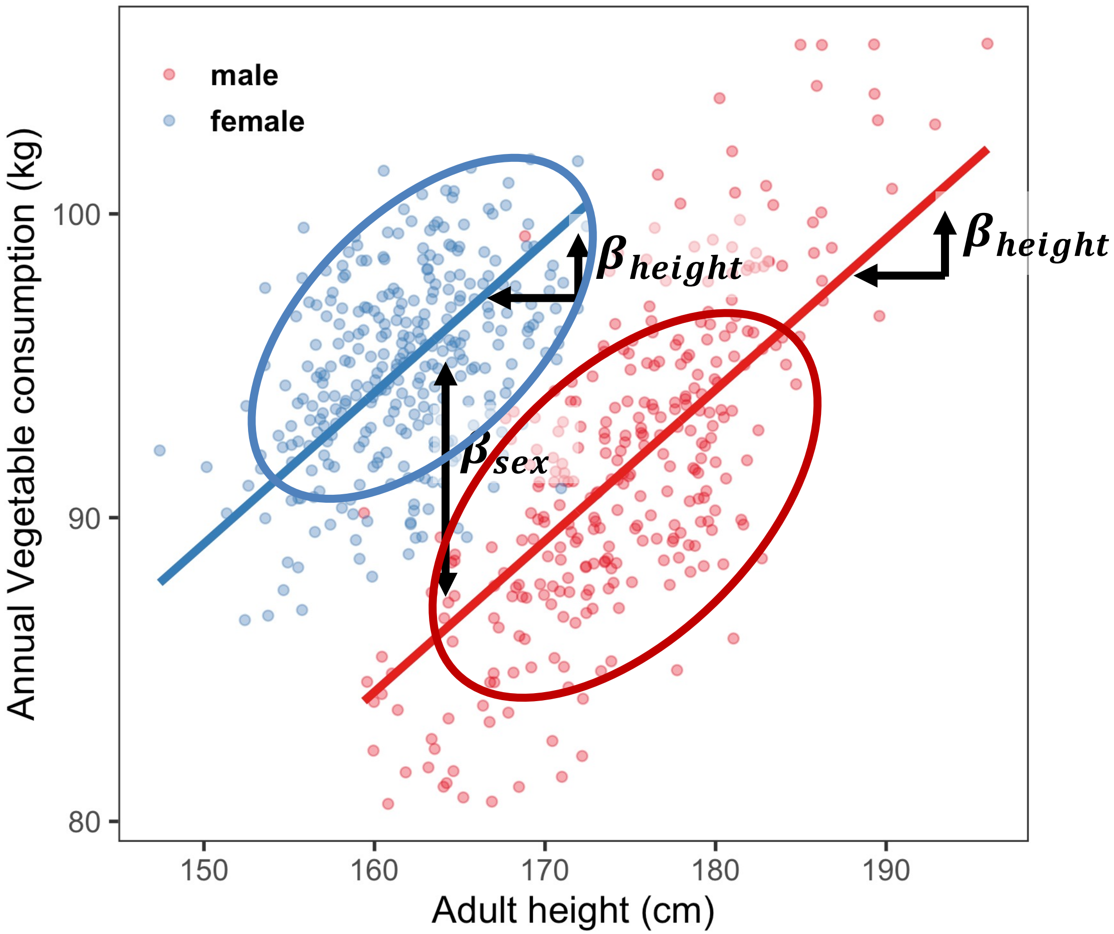
Among population of adults in the UK:

- On average, consume 60g (0.06 kg) more vegetables each year per 1 cm taller height; 95% CI 18g to 103g.
- Strong evidence to **reject** null hypothesis of no association between vegetable consumption and height.
- Height explains about 1.3% of variation in vegetable consumption.

statistically significant relationship

Example: height and vegetable consumption

Data: random sample of 600 adults in the UK
- 300 men, 300 women



$$\text{Veg cons.} = \beta_0 + \beta_{sex}\times[\text{sex} = \text{F}] + \beta_{height}\times\text{height} + \epsilon$$

Param.	Estim.	Std. err.	t-value	p-value	95% CI
β_0	4.669	4.136	1.129	0.259	(-3.45–12.79)
β_{sex}	9.854	0.407	24.237	<0.001	(9.06–10.65)
β_{height}	0.497	0.024	21.052	<0.001	(0.451–0.544)
R^2	0.503				

Among population of adults in the UK:
- Women eat around 9.9kg more vegetables per year, *adjusted for height* (95% CI 9.1–10.7kg)
- Additional 1 cm height associated with 498g (95% CI 451–544g) more veg consumption, adjusted for sex.
- Sex and height explain about 50% of the variation in annual vegetable consumption.

Multiple linear regression

- Examine dependency of outcome on **<u>several</u>** exposure variables, not just one.

- Two reasons for including additional exposure variables:
  1. Estimate an exposure effect after accounting for the effect of other variables—adjusting for ***confounding*** factors.
     - Example: *Women were on average shorter and ate more vegetables than men of similar height→ sex **confounded** the relationship between height and vegetable consumption*
  2. Systematically explain additional variation in the data → reduce residual variation → decrease standard error of regression coefficient
     - Increases accuracy of coefficient estimate; likelihood that hypothesis will detect any real effect that exists
     - Only applies to linear regression (not, e.g., logistic or Poisson regression)

- *What is the association between vegetable consumption and height, adjusted for effect of sex?*
- *What is the association between vegetable consumption and sex, adjusted for effect of height?*
  - ‘adjusted for’ = having taken into account
  - Also often referred to as ‘controlled for’; ‘adjusted for’ generally preferred language now.

Multiple regression in R

Simple linear regression

```
> fit1 <- lm(veg ~ height, data = dat)

> summary(fit1)

Call:
lm(formula = veg ~ height, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5693  -2.9901   0.3621   3.2112  11.2556

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 83.15026    3.62024  22.968  < 2e-16 ***
height       0.06034    0.02149   2.808  0.00514 **

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.528 on 598 degrees of freedom
Multiple R-squared:  0.01302,    Adjusted R-squared:  0.01137
F-statistic: 7.887 on 1 and 598 DF,  p-value: 0.005143
```

Multiple linear regression

```
> fit2 <- lm(veg ~ sex + height, data = dat)

> summary(fit2)

Call:
lm(formula = veg ~ sex + height, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7129 -2.2744  0.1207  2.1465 10.6118

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.66867    4.13555   1.129    0.259
sexfemale    9.85417    0.40658  24.237   <2e-16 ***
height       0.49747    0.02363  21.052   <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.217 on 597 degrees of freedom
Multiple R-squared:  0.5025,     Adjusted R-squared:  0.5008
F-statistic: 301.5 on 2 and 597 DF,  p-value: < 2.2e-16
```

Imperial College London

Comparing multiple regression models

- Use ANOVA to test the null hypothesis that no association of veg consumption with sex, having already accounted for height.

- Reverse order: null hypothesis no association of veg consumption and height, having already accounted for sex.

Notes:

- Models must be ‘nested’
  - All terms in model 1 must also be included in model 2
- For single additional variable, equivalent to t-test for the added term in multiple regression table.
  - t-statistic is square-root of the F-statistic
- More useful for categorical variables (multiple levels)
  - H0: all of the additional categories are = 0
  - H1: at least one of additional categories != 0

```
> fit1a <- lm(veg ~ height, data = dat)
> fit2 <- lm(veg ~ height + sex, data = dat)

> anova(fit1a, fit2)
Analysis of Variance Table

Model 1: veg ~ height
Model 2: veg ~ height + sex
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    598 12260.0
2    597  6179.7  1    6080.4 587.41 < 2.2e-16 ***
---
```

```
> fit1b <- lm(veg ~ sex, data = dat)
> fit2 <- lm(veg ~ sex + height, data = dat)
> anova(fit1b, fit2)
```

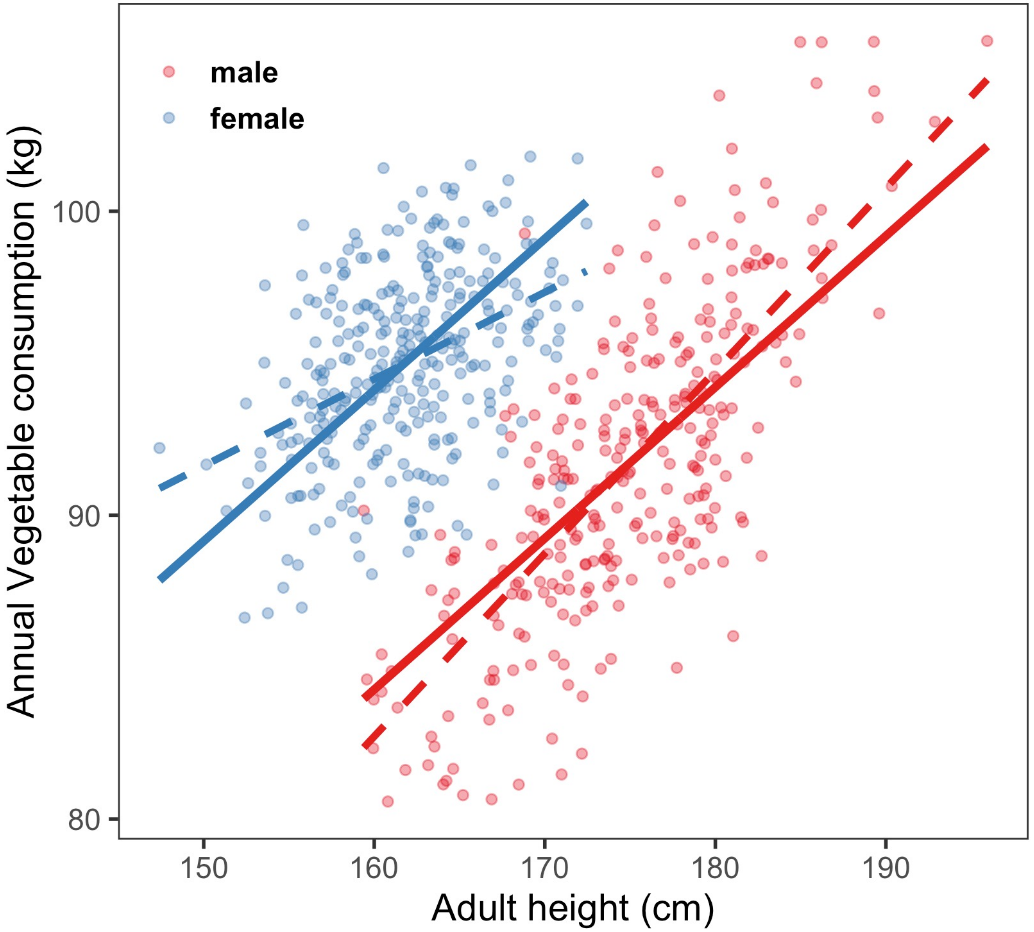
compares model 1 and 2

```
Analysis of Variance Table

Model 1: veg ~ sex
Model 2: veg ~ sex + height
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    598 10767.2
2    597  6179.7  1    4587.6 443.19 < 2.2e-16 ***
---
```

Interactions

Is the relationship between height and vegetable consumption different for women and men?



$$\text{Veg cons.} = \beta_0 + \beta_{sex} \times [\text{sex} = \text{F}] + \beta_{height} \times \text{height} + \beta_{sexF,height} \times [\text{sex} = \text{F}] \times \text{height} + \epsilon$$

Param.	Estim.	Std. err.	t-value	p-value	95% CI
β_0	-13.503	4.882	-2.766	0.006	(-23.09– -3.91)
β_{sex}	62.272	8.085	7.703	<0.001	(46.39–78.15)
β_{height}	0.601	0.028	21.551	<0.001	(0.547–0.656)
$\beta_{sexF,height}$	-0.316	0.049	-6.491	<0.001	(-0.411– -0.220)
R^2	0.535				

- $\beta_{sexF,height}$: is the effect of the *interaction* between sex and height
  - Also called *effect modification*: sex modifies the effect of height on vegetable consumption.

Interactions

$$\text{Veg cons.} = \beta_0 + \beta_{sex} \times [\text{sex} = \text{F}] + \beta_{height} \times \text{ height} + \beta_{sexF,height} \times [\text{sex} = \text{F}] \times \text{height} + \epsilon$$

Param.	Estim.	Std. err.	t-value	p-value	95% CI
β_0	-13.503	4.882	-2.766	0.006	(-23.09– -3.91)
β_{sex}	62.272	8.085	7.703	<0.001	(46.39–78.15)
β_{height}	0.601	0.028	21.551	<0.001	(0.547–0.656)
$\beta_{sexF,height}$	-0.316	0.049	-6.491	<0.001	(-0.411– -0.220)
R^2	0.535				

- β_0: The predicted vegetable consumption for *men* with *height = 0cm* is -13.5kg per year.
- β_{sex}: The predicted difference between female veg consumption and male veg consumption is 62.3kg per year <u>when height = 0</u> (for both sexes)
- β_{height}: <u>For men,</u> vegetable consumption increases by 0.6kg for each cm height.
- $\beta_{sexF,height}$: <u>For women,</u> each cm increase in height is associated with <u>**0.32kg less increase**</u> in veg consumption than for men.
  - <u>For women,</u> vegetable consumption increases by 0.60 - 0.32 = 0.28 kg per cm increase in height

check updated slide deck

Interactions in R

Multiple regression with interaction

```
> fit3 <- lm(veg ~ sex + height + sex:height, data = dat)

> summary(fit1)

Call:
lm(formula = veg ~ height, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.5693  -2.9901   0.3621   3.2112  11.2556 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 83.15026    3.62024  22.968  < 2e-16 ***
height       0.06034    0.02149   2.808  0.00514 **

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.528 on 598 degrees of freedom
Multiple R-squared:  0.01302,    Adjusted R-squared:  0.01137
F-statistic: 7.887 on 1 and 598 DF,  p-value: 0.005143
```

Multiple linear regression

```
> fit2 <- lm(veg ~ sex + height, data = dat)

> summary(fit2)

Call:
lm(formula = veg ~ sex + height, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.7129 -2.2744  0.1207  2.1465 10.6118 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.66867    4.13555   1.129    0.259
sexfemale    9.85417    0.40658  24.237   <2e-16 ***
height       0.49747    0.02363  21.052   <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.217 on 597 degrees of freedom
Multiple R-squared:  0.5025,    Adjusted R-squared:  0.5008
F-statistic: 301.5 on 2 and 597 DF,  p-value: < 2.2e-16
```

NB: `lm(veg ~ sex + height + sex:height, ...)` and `lm(veg ~ sex*height, ...)` are equivalent.

Centering covariates

only include if there's a clear motivation behind why the confounder could affect

- Often 'centering' covariates at their mean makes results more interpretable.
  - E.g.: height = 0 -> height = mean(height)

```
> mean(dat$height)
[1] 168.2679
> dat$height_c <- dat$height - mean(dat$height)

> fit3_c <- lm(veg ~ sex + height_c + sex:height_c, data=dat)

Call:
lm(formula = veg ~ sex + height_c + sex:height_c, data = dat)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       87.69415    0.25663 341.720  < 2e-16 ***
sexfemale          9.14602    0.40811  22.411  < 2e-16 ***
height_c           0.60141    0.02791  21.551  < 2e-16 ***
sexfemale:height_c -0.31572   0.04864  -6.491 1.79e-10 ***
---

Residual standard error: 3.112 on 596 degrees of freedom
Multiple R-squared:  0.5354,    Adjusted R-squared:  0.533
F-statistic: 228.9 on 3 and 596 DF,  p-value: < 2.2e-16
```

- Estimate for `height` slope, `sexfemale:height` interaction, R^2, and statistical inferences are unchanged.

- Intercept and sex coefficients are more interpretable:
  - β_0: The predicted vegetable consumption for *men* with *height = 168cm* is 87.7kg per year.
  - β_{sex}: For men and women both with height = 168cm, women consume 9.1kg more vegetables per year than men.

Imperial College London

Interaction of continuous covariates

Peru lung function data set (636 children aged 7 to 10 years)

- Estimate interaction between effect of height and age on lung function (FEV1)

```
> peru_fit <- lm(fev1 ~ age*height, peru)

> summary(peru_fit)

Call:
lm(formula = fev1 ~ age * height, data = peru)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.793279   1.956125   1.939  0.05292 .
age         -0.587443   0.216724  -2.711  0.00690 **
height      -0.024679   0.015950  -1.547  0.12230
age:height   0.005497   0.001755   3.133  0.00181 **
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2275 on 632 degrees of freedom
Multiple R-squared:  0.4442,    Adjusted R-squared:  0.4416
F-statistic: 168.4 on 3 and 632 DF,  p-value: < 2.2e-16
```

- *For every 1 year older age*, the effect of an additional cm of height on FEV1 is increased by 0.005 L/sec

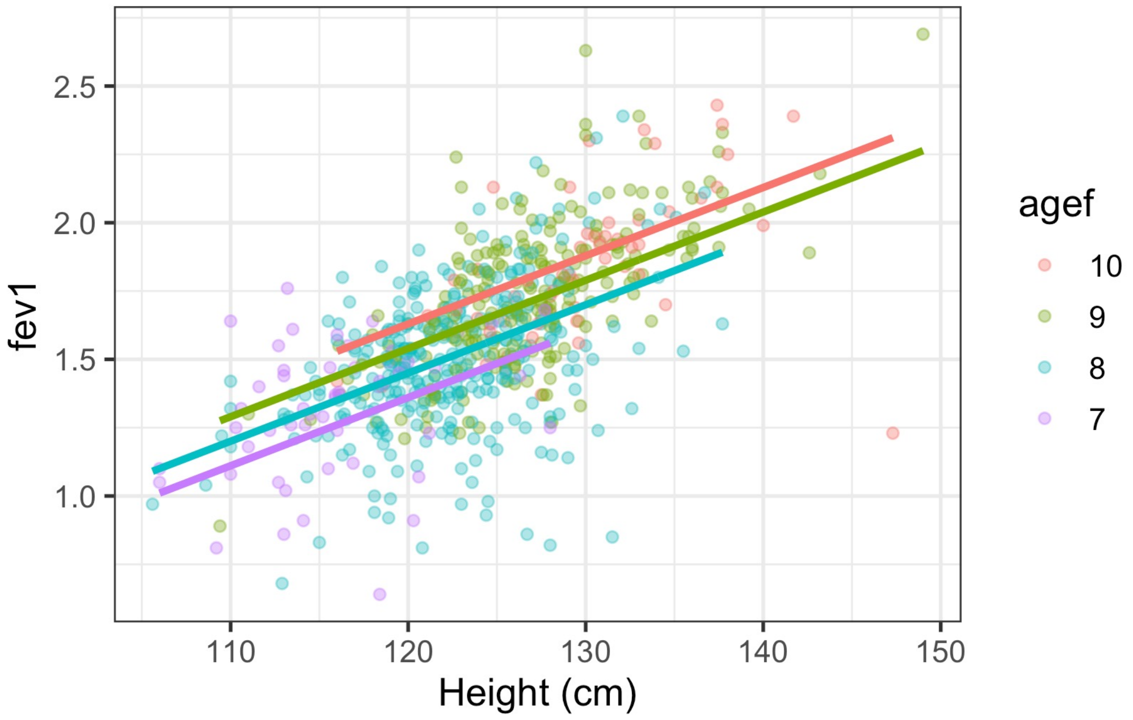
- *For each cm additional height*, the effect of an additional year of age on FEV1 is increased by 0.005 L/sec

- Often useful / clearer to communicate slopes for indicative values:
  - At age 7, each ⬆ cm height **0.016 L/sec** ⬆ FEV1
  - At age 8, each ⬆ cm height -> **0.022 L/sec** ⬆ FEV1
  - At age 9, each ⬆ cm height -> **0.028 L/sec** ⬆ FEV1
  - At age 10, each ⬆ cm height -> **0.033 L/sec** ⬆ FEV1

Interaction of continuous covariates

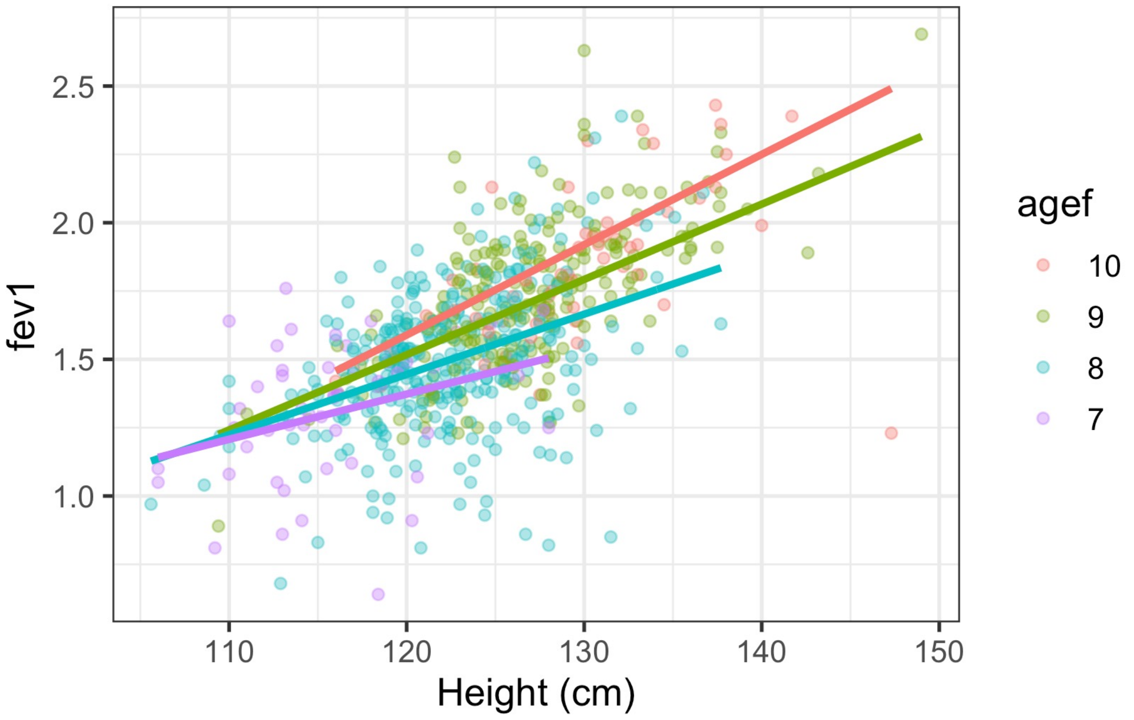
Multiple regression (no interaction)

```
> lm(fev1 ~ age + height, data = peru)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.308601   0.181175 -12.742  < 2e-16 ***
age          0.089714   0.015717   5.708 1.76e-08 ***
height       0.024967   0.001813  13.774  < 2e-16 ***
```



With age x height interaction

```
> lm(fev1 ~ age * height, data = peru)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.793279   1.956125   1.939  0.05292 .
age         -0.587443   0.216724  -2.711  0.00690 **
height      -0.024679   0.015950  -1.547  0.12230
age:height   0.005497   0.001755   3.133  0.00181 **
```



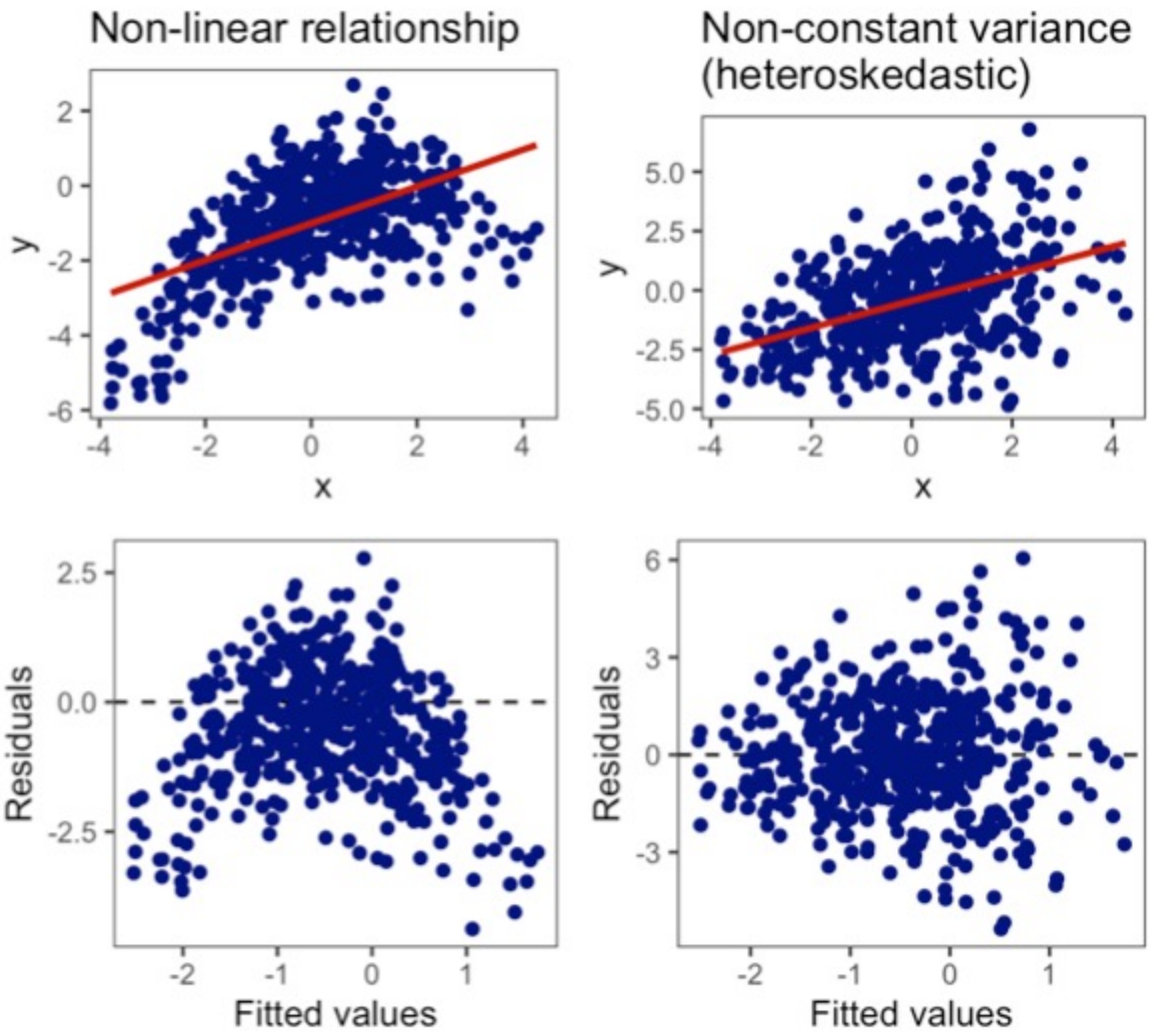
Imperial College London

Transformations

Imperial College London

Linear regression assumptions

1. Linearity
2. Normality
3. Independence
4. Equal variance



Strategies for addressing assumption violations

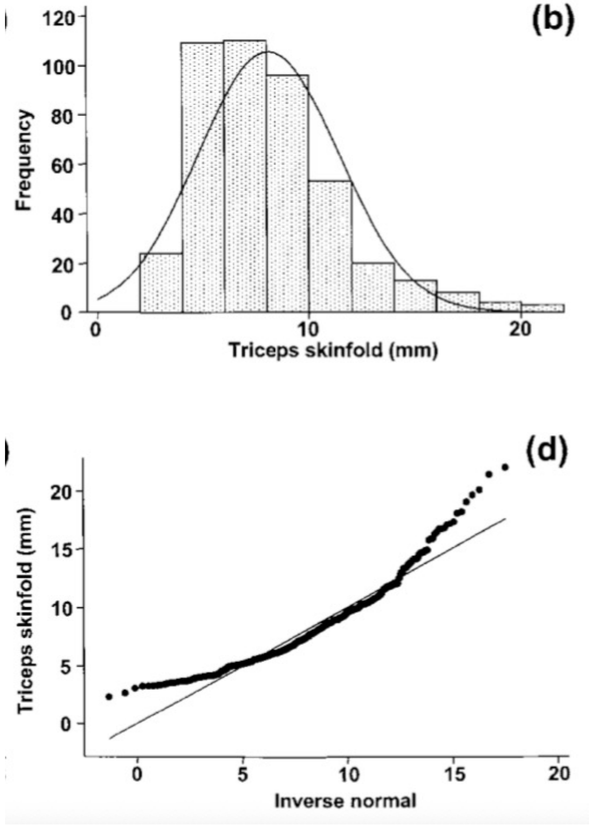
- Check for errors in the data (coding, entry) resulting in outliers.
- Explore non-linear relationships between outcome and exposure variables.
  - Convert continuous variables to categorical.
  - Add quadratic term. allows us to capture curved r/s
- Conduct **sensitivity analysis** to check whether Text excluding outliers / alternative specifications change the key conclusions of the analysis.
- **Use transformations** of outcome or exposure variables.

rnorm() ?

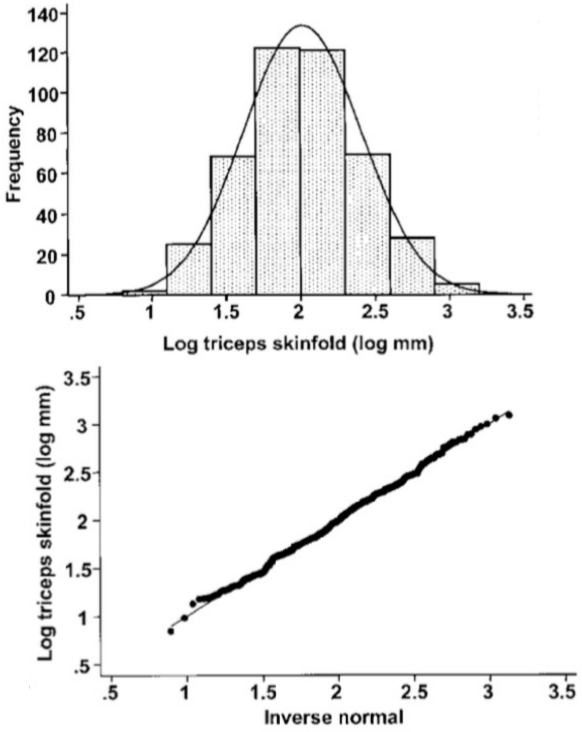
- Use robust standard errors or bootstrapping to derive CIs without requiring distributional assumptions (Chapter 30; Problem Set 3).

Log transformation

- Positively (right) skewed data: Triceps skinfold measurements of 440 men

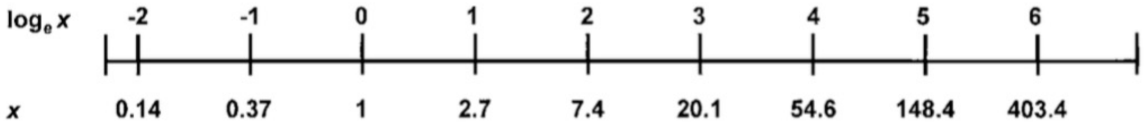


- Log transformation triceps skinfold measurement



Convenient properties:

- Log(y) ~ Normal() → y ~ Log-normal()
- $\log(y) = \beta_0 + \beta_1 x$ → $y = e^{\beta_0} e^{\beta_1 x}$
- "1 unit increase in x → e^{β_1} **<u>times</u>** increase in y"



Imperial College London

Non-linear relationship

log functions as useful

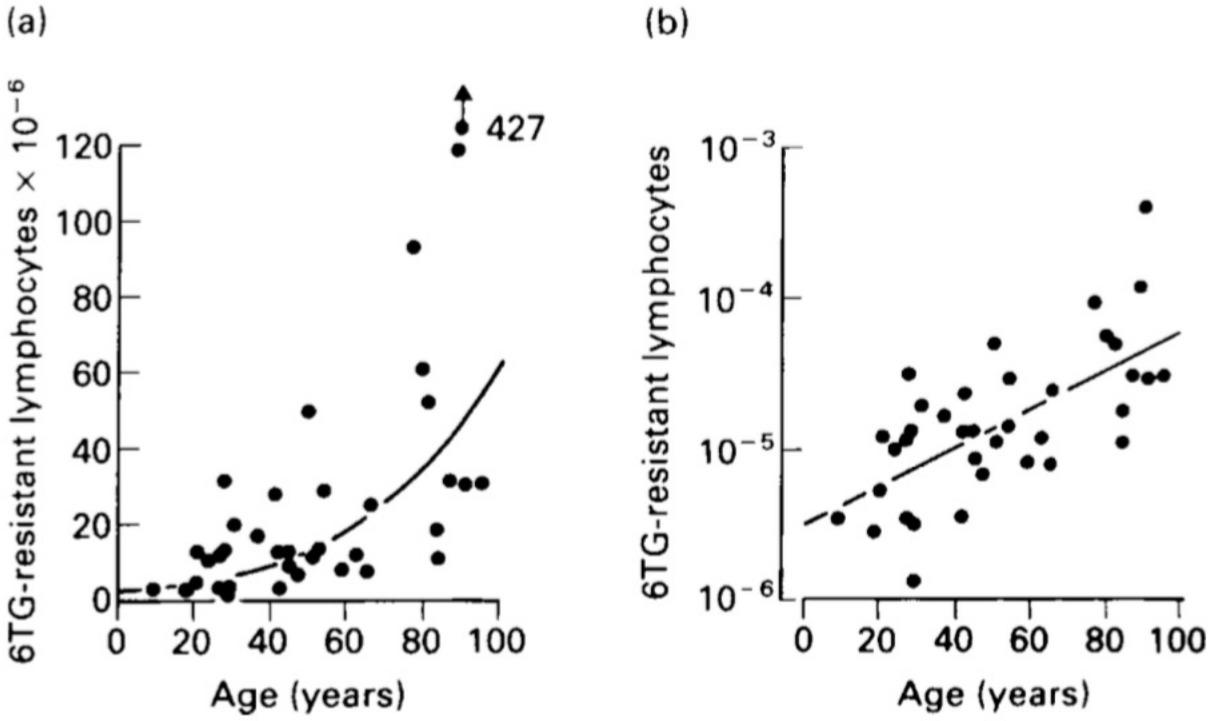


Fig. 13.4 Relationship between frequency of 6TG-resistant lymphocytes and age for 37 individuals drawn using (a) a linear scale, and (b) a logarithmic scale for frequency. Reprinted from Morley *et al*. *Mechanisms of Ageing and Development* **19**: 21–6, copyright (1982), with permission from Elsevier Science.

Imperial College London

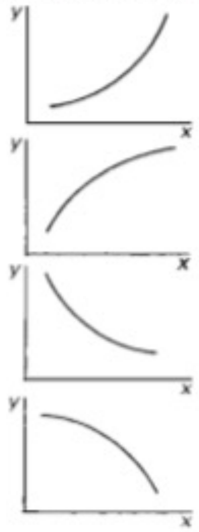
Transformations: summary

- **Unequal variance:** variance increasing with mean is very common situation
  - Diagnose with residuals vs. fitted values—funnel shape.
  - Try logarithmic transformation of outcome variable (S.D. increases linearly with mean).
  - Reciprocal or square-root transformation for more/less severe.

- **Non-linear relationship:**
  - Diagnose with residual vs. fitted values—'U-shaped' pattern.
  - May be resolved by transforming either outcome or exposure variable (see Table 13.3)
  - Often becomes difficult to interpret: "a 1 unit increase in reciprocal of soda intake is associated with a 0.2 unit reduction in square-root of BMI"

- Goodness-of-fit tests (Shapiro-Wilk, chi-squared) tend to be not that useful for data sets of moderate size.
  - Very sensitive to detect some difference from Normal distribution; unclear practical importance for inference.

Table 13.3 Summary of different choices of transformations. Those removing positive skewness are called group A transformations, and those removing negative skewness group B.

Situation	Transformation
Positively skewed distribution (group A)	
Lognormal	Logarithmic ($u = \log x$)
More skewed than lognormal	Reciprocal ($u = 1/x$)
Less skewed than lognormal	Square root ($u = \sqrt{x}$)
Negatively skewed distribution (group B)	
Moderately skewed	Square ($u = x^2$)
More skewed	Cubic ($u = x^3$)
Unequal variation	
s.d. proportional to mean	Logarithmic ($u = \log x$)
s.d. proportional to mean^2	Reciprocal ($u = 1/x$)
s.d. proportional to $\sqrt{\text{mean}}$	Square root ($u = \sqrt{x}$)
Non-linear relationship	Transform: y variable and/or x variable
	Group A (y) Group B (x) Group B (y) Group A (x) Group A (y) Group A (x) Group B (y) Group B (x)

Imperial College London

Model building

Imperial College London

What to include in my regression?

1. Any variables that are *confounders* of the relationship between the outcome and primary exposure of interest.

2. For linear regression (only): other variables that are clearly associated with the outcome → improves precision
   - But keep the number relatively small: interpretation becomes harder with large number of variables.
   - Rule of thumb, at least 10 observations per additional covariate. (Often much more for large or noisy data sets.)

like p val

- For hypothesis testing (e.g. RCTs): full analysis plan must be specified before you start.
- For observational studies (exploratory analysis): some model building required, but write down clear hypotheses and analysis plan before you start.
  - Usually involves **sensitivity analysis**: testing interactions, non-linear terms

- Avoid model selection procedures that involve choosing covariates based on looking at the data:
  - to avoid: Including all covariates with univariate associations; e.g. $p < 0.05$, $p < 0.2$
  - to avoid: 'Stepwise selection': recursively adding / remove variables until no further improvement in model fit

Why should we avoid stepwise models

- Over-optimistic results:
  - P-values for the selected variables tend to be small, CIs narrow and R2 high
  - Higher chance of spurious associations with increasing number of initial variables

- Regression coefficients tend to be larger and as a result perform poorly in test data

- Slight changes in data can lead to large in the variables selected in the final model

- Stepwise methods often used instead of expert knowledge and correct problem definition

Imperial College London

Dangers of collinearity

- Collinearity: high correlation between two exposure variables.

reduces ability to identify main exposure of interest as it is being mocked up by the other covariate

- Not possible to identify linearly independent association with each variable.
  - Only a linear combination of the covariates → biased coefficient estimates
- Can cause very large standard errors; spurious null association with the outcome

fake

Table 29.11 Demonstration of the effect of collinearity, using data from the study of lung disease in children in Lima, Peru. Variable *newage* is variable *age* plus a random error whose standard deviation is given in the first column in the table.

s.d. of random error	Correlation between *age* and *newage*	Regression of height on *newage*	Regression of height on *age* and *newage*		
		Coefficient (s.e.) for *newage*	Coefficient (s.e.) for *age*	Coefficient (s.e.) for *newage*	Sum of coefficients
1	0.57	1.61 (0.20)	5.31 (0.33)	− 0.17 (0.20)	5.16
0.1	0.9904	5.06 (0.28)	6.81 (2.00)	− 1.66 (1.99)	5.15
0.01	0.9999	5.16 (0.28)	21.76 (19.94)	−16.62 (19.94)	5.14

Common pitfalls

Common pitfalls in (Kirkwood & Sterne, section 38.8)

1. **Multiple comparisons:**
   - Even if there is no association between exposure and outcome, 1 in 20 comparisons to be statistically significant at 5% level.
   - Try to minimize the number of hypothesis tests; if testing many, adjust for it.

2. **Subgroup analyses:**
   - Be very cautious about interpreting apparent associations in subgroups of data, particularly when no evidence of overall association.
   - Tempting to emphasise an ‘interesting’ finding in an otherwise null study!

3. **Data-driven comparisons:**
   - Avoid choosing definitions for specifying exposure or outcome variable after running several analyses

Imperial College London

Any questions?