Regression functions

Two of the most commonly used **R** functions for modelling are:

- `lm()` for **l**inear **m**odels.
- `glm()` for **g**eneralised **l**inear **m**odels.

We have entire stage 3 courses on the use of these commands.

Note for SAS users: `PROC GLM` is **not** the same as `glm` in **R**.

Student’s t-test

Student’s t-test in **R**

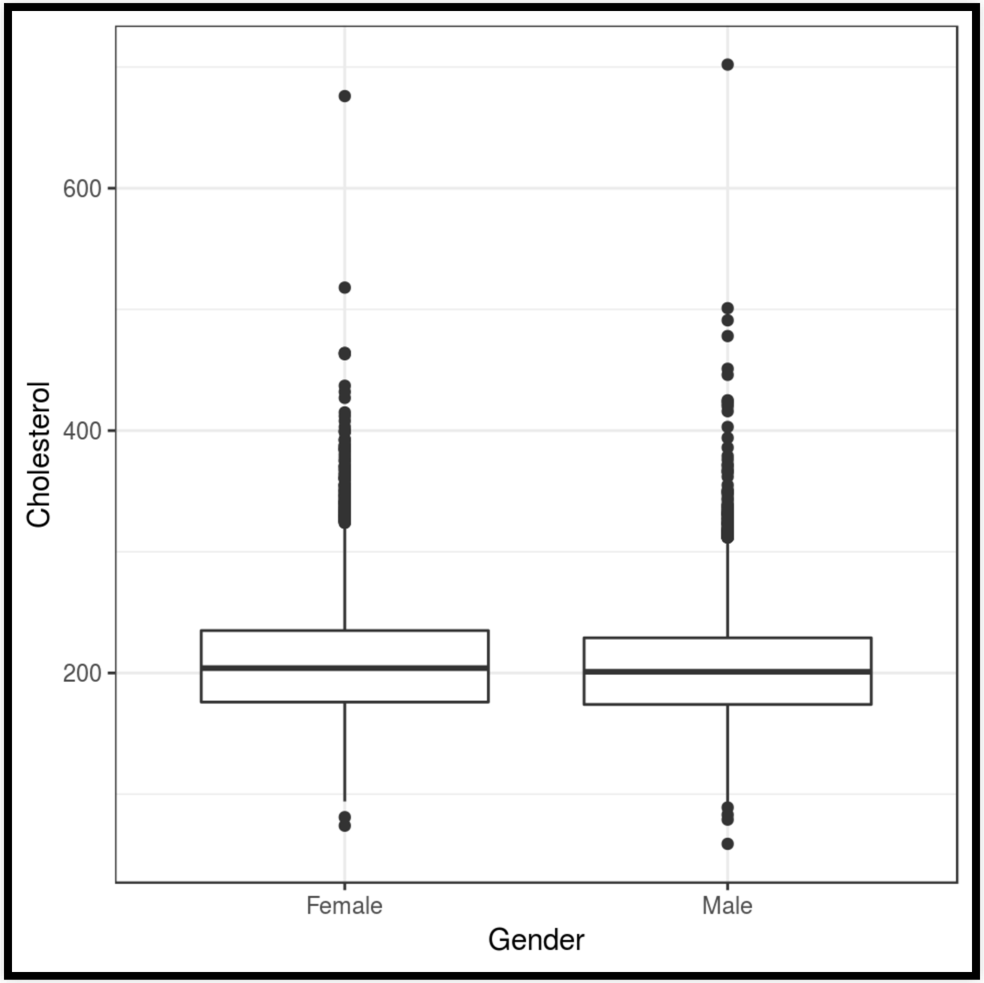
```
t.test(y ~ x, data = dataset)
```

`y`: the continuous response variable. `x`: grouping variable with 2 levels. `data`: name of the dataframe containing the variables.

Suppose we want to test whether males and females have different cholesterol levels. After visualising the data, we can perform the t-test in **R**:

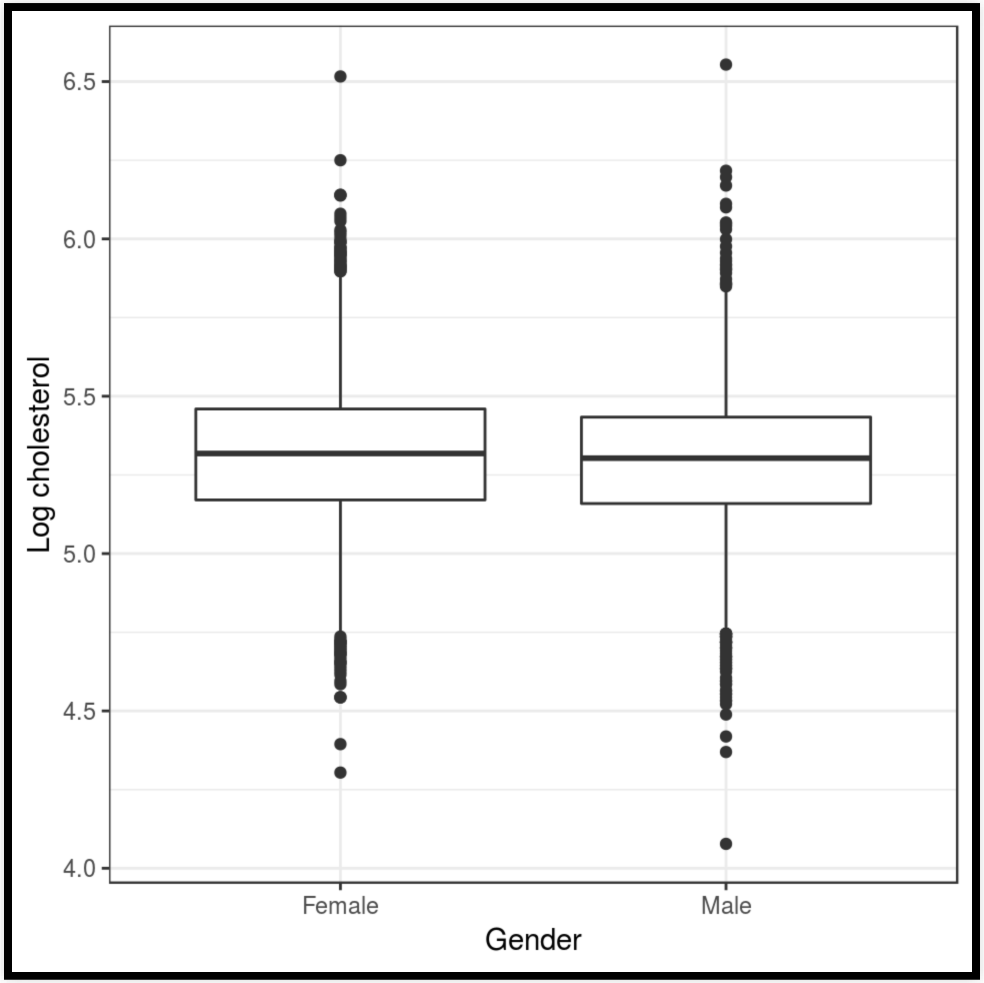
```
t.test(Cholesterol ~ Gender, data = patient.df)
```

Student’s t-test in **R**



Note we could log-transform the cholesterol variable and make inferences on the median.

Student's t-test in **R**



Student’s t-test in **R**

```r
t.test(Cholesterol ~ Gender, data = patient.df)
```

```
#R:
#R:     Welch Two Sample t-test
#R:
#R:  data:  Cholesterol by Gender
#R:  t = 6.4444, df = 16021, p-value = 1.194e-10
#R:  alternative hypothesis: true difference in means is not equal to 0
#R:  95 percent confidence interval:
#R:   3.160295 5.923056
#R:  sample estimates:
#R:  mean in group Female   mean in group Male
#R:              208.1786             203.6370
```

- We have extremely strong evidence to suggest that the average cholesterol level for females is between 3.2 and 5.9 units higher than for males.

Analysis of Variance (ANOVA)

ANOVA in **R**

- Generalises the t-test to more than 2 groups.
- Null hypothesis: all group means are equal.

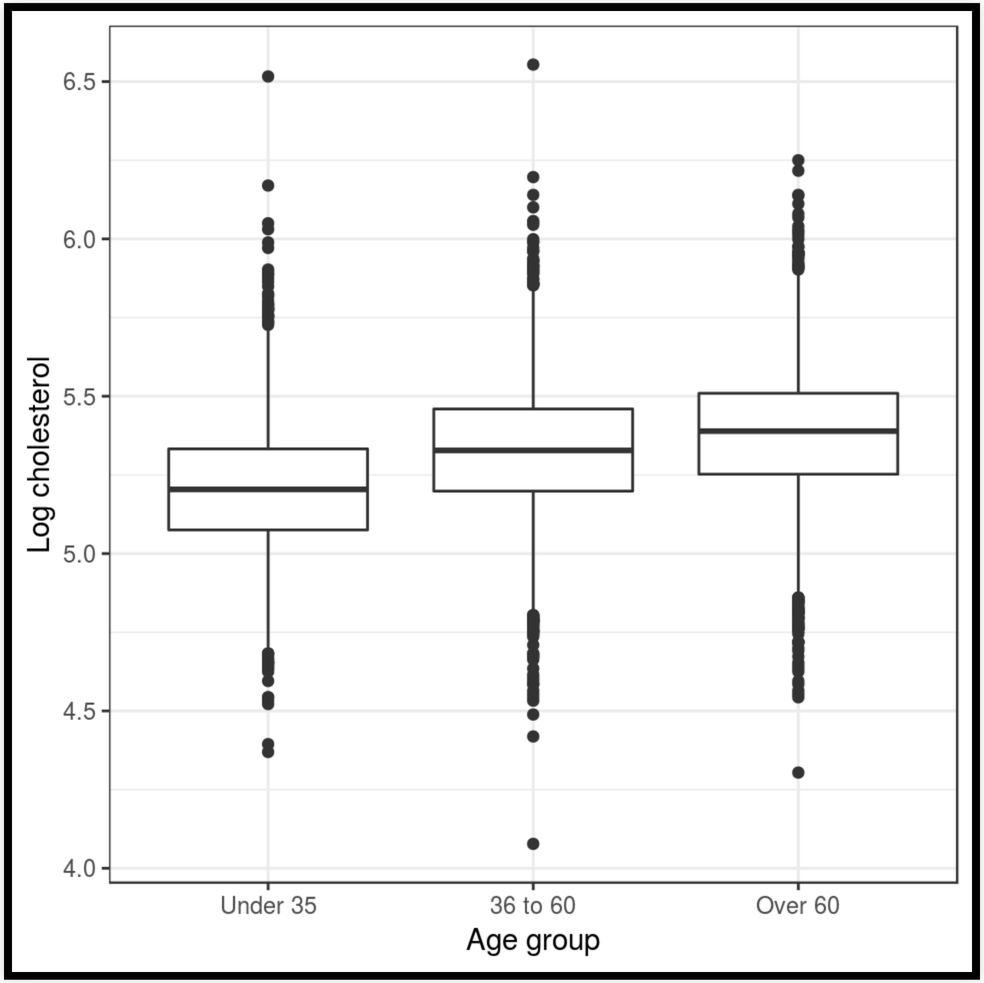
```
aov(y ~ x, data = dataset)
```

Example

Null hypothesis: The mean cholesterol levels are the same for all three age groups.

```
my_aov = aov(Cholesterol ~ age_group, data = patient.df)
```

ANOVA in **R**


6.5
6.0
5.5
5.0
4.5
4.0
Log cholesterol
Under 35
36 to 60
Over 60
Age group

ANOVA in **R**

```r
summary(my_aov)
```

```
#R:                  Df   Sum Sq Mean Sq F value Pr(>F)
#R:  age_group        2  3280912 1640456   908.2 <2e-16 ***
#R:  Residuals   16059 29007528    1806
#R:  ---
#R:  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#R:  968 observations deleted due to missingness
```

- We have extremely strong evidence that the average cholesterol level in at least one age group is different to at least one other age group.

Group means

We can compute a summary table of the results easily with the `model.tables` function:

```
model.tables(my_aov, "means")
```

```
#R:  Tables of means
#R:  Grand mean
#R:
#R:  206.0492
#R:
#R:   age_group
#R:      Under 35 36 to 60 Over 60
#R:         185.9    209.7   221.2
#R:  rep   4949.0   5991.0  5122.0
```

It would be interesting to know which pairs are statistically different from one another.

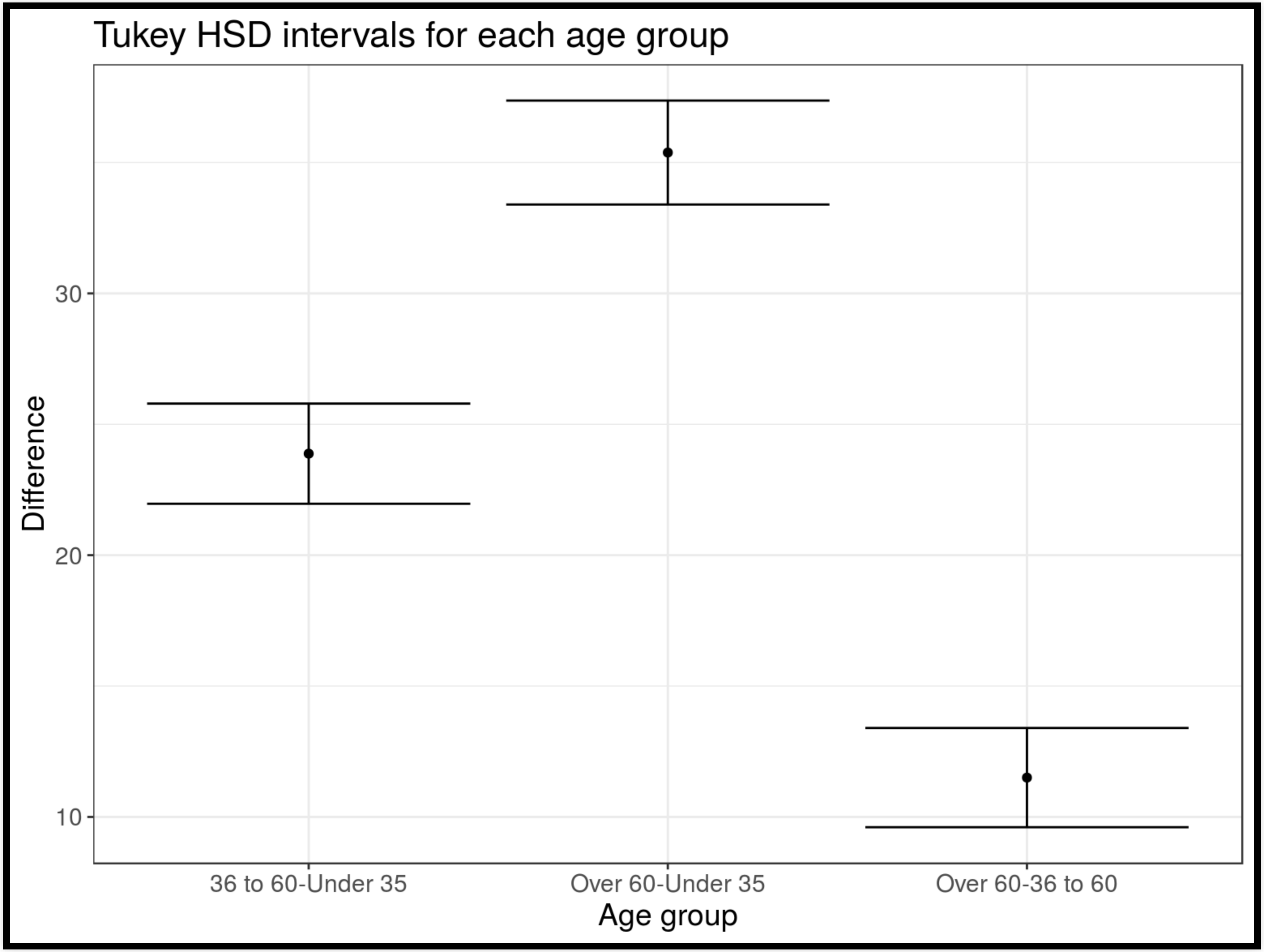
Post-hoc multiple comparisons

We can calculate Tukey’s Honest Significant Difference intervals for our post-hoc tests:

```
TukeyHSD(my_aov)
```

```
#R:    Tukey multiple comparisons of means
#R:      95% family-wise confidence level
#R:
#R:  Fit: aov(formula = Cholesterol ~ age_group, data = patient.df)
#R:
#R:  $age_group
#R:                        diff       lwr      upr p adj
#R:  36 to 60-Under 35 23.87793 21.964387 25.79148     0
#R:  Over 60-Under 35  35.38133 33.395712 37.36695     0
#R:  Over 60-36 to 60  11.50340  9.607634 13.39917     0
```

Post-hoc multiple comparisons



Two-way ANOVA

Two-way ANOVA in **R**

- The last ANOVA model was fitted using on categorical variable (`age_group`), hence a *one*-way ANOVA.
- If we fit a linear model using two categorical, explanatory variables, we have a *two*-way ANOVA.

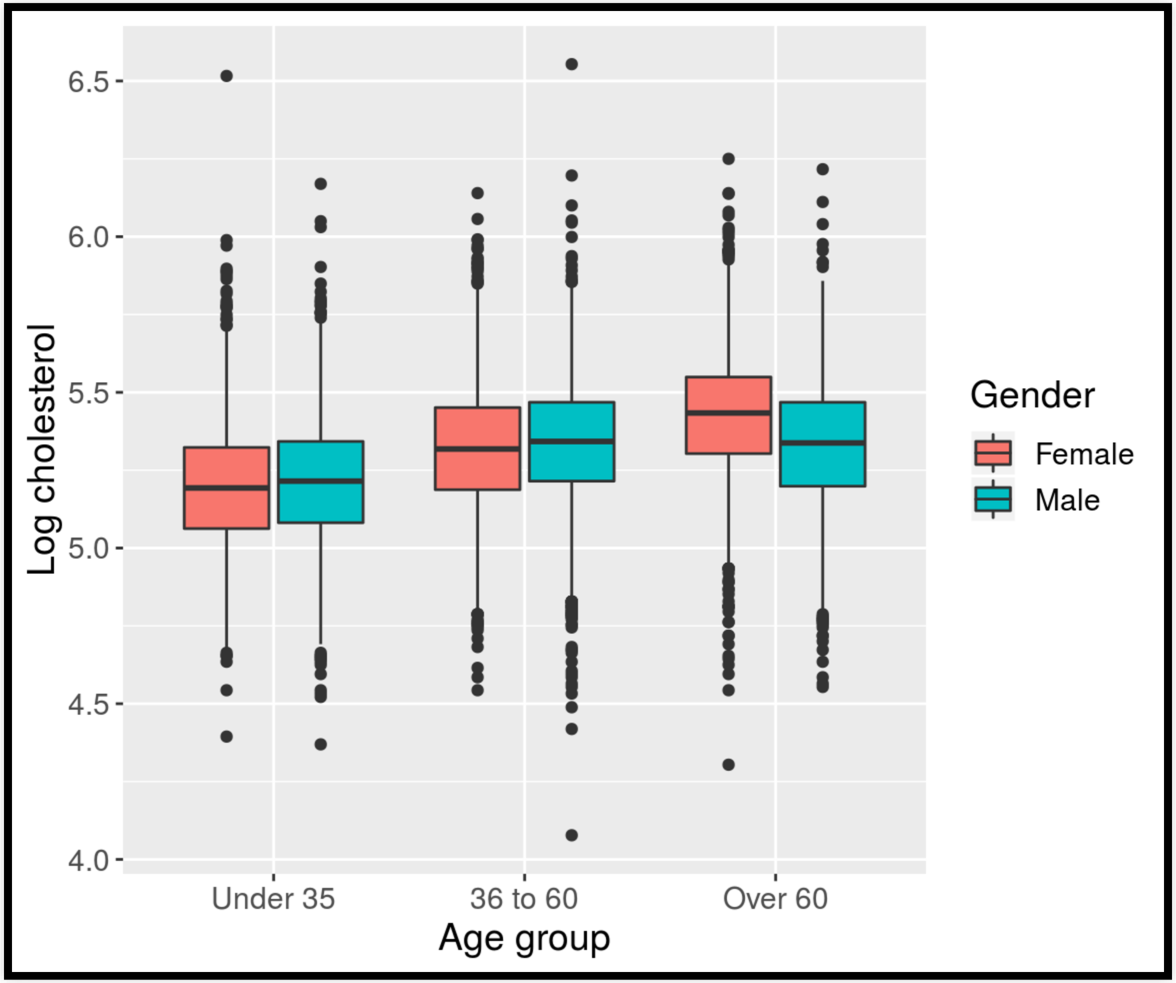
Example research question

Do the differences in cholesterol levels differ between the various age groups *and* genders?

```
my_2way_aov = aov(Cholesterol ~ Gender * age_group,
                  data = patient.df)
```

Note that we are fitting an interaction using `*`.

Visualise the data



Two-way ANOVA in R

```
summary(my_2way_aov)
```

```
#R:                        Df   Sum Sq Mean Sq F value   Pr(>F)    
#R:  Gender                 1    82506   82506   46.63 8.87e-12 ***
#R:  age_group              2  3292611 1646305  930.48  < 2e-16 ***
#R:  Gender:age_group       2   505233  252617  142.78  < 2e-16 ***
#R:  Residuals          16056 28408089    1769                     
#R:  ---
#R:  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#R:  968 observations deleted due to missingness
```

There is a significant 2-way interaction between gender and age group, i.e. the magnitude of the difference in the mean cholesterol levels between males and females is not constant across the age groups.

Group means

```r
model.tables(my_2way_aov, "means")
```

```
#R:  Tables of means
#R:  Grand mean
#R:
#R:  206.0492
#R:
#R:   Gender
#R:      Female   Male
#R:       208.2  203.6
#R:  rep 8531.0 7531.0
#R:
#R:   age_group
#R:      Under 35 36 to 60 Over 60
#R:         185.8    209.7   221.3
#R:  rep   4949.0   5991.0  5122.0
#R:
#R:   Gender:age_group
#R:          age_group
#R:  Gender   Under 35 36 to 60 Over 60
```

Post-hoc multiple comparisons

```r
TukeyHSD(my_2way_aov)
```

```
#R:    Tukey multiple comparisons of means
#R:      95% family-wise confidence level
#R:
#R:  Fit: aov(formula = Cholesterol ~ Gender * age_group, data = patient
#R:
#R:  $Gender
#R:                   diff       lwr       upr p adj
#R:  Male-Female -4.541675 -5.845313 -3.238038     0
#R:
#R:  $age_group
#R:                        diff      lwr      upr p adj
#R:  36 to 60-Under 35 23.86896 21.97511 25.76281     0
#R:  Over 60-Under 35  35.45641 33.49123 37.42159     0
#R:  Over 60-36 to 60  11.58745  9.71120 13.46370     0
#R:
#R:  $`Gender:age_group`
#R:                                        diff        lwr        upr
#R:  Male:Under 35-Female:Under 35     3.110853 -0.3060589   6.527764
```

Tests of independence

Table of counts

Example

Does smoking depend on age group?

- Two Categorical variables
- Test for independence between rows and columns

```
smoke_tab = with(patient.df, table(Smoke, age_group))
smoke_tab
```

```
#R:        age_group
#R:   Smoke Under 35 36 to 60 Over 60
#R:     No       580     1611    2064
#R:     Yes     1629     1943     799
```

Pearson’s Chi-squared test

We can use the `chisq.test` function in **R** to perform a Pearson’s Chi-square test for independence:

```r
chisq.test(smoke_tab)
```

```
#R: 
#R:     Pearson's Chi-squared test
#R: 
#R:  data:  smoke_tab
#R:  X-squared = 1086.7, df = 2, p-value < 2.2e-16
```

- We have extremely strong evidence to suggest that smoking and age group are *not* independent of one another.
- Whether a patient smokes or not likely depends on their age group.

Assumptions

- Pearson’s Chi-squared tests have certain assumptions.
- These assumptions are primarily to do with sample size.
- **R** will give you a warning if these assumptions are not met:

```
#R:  Warning in chisq.test(my_table): Chi-squared approximation may be i
```

```
#R: 
#R:     Pearson's Chi-squared test
#R: 
#R:  data:  my_table
#R:  X-squared = 8.0496, df = 3, p-value = 0.045
```

If the sample size is too small for a Pearson’s Chi-square test, one alternative is to use a Fisher’s exact test.

Fisher’s exact test

If we assume that our sample size was much smaller, and our assumptions for a Chi-square test were not met, we could perform a Fisher’s exact test using `fisher.test`:

```
no_na.df = subset(patient.df, !is.na(Smoke) & !is.na(age_group))
set.seed(3)
smoke_tab = table(no_na.df[sample(seq_along(no_na.df$Age), 30), c("Smoke
fisher.test(smoke_tab)
```

```
#R:
#R:     Fisher's Exact Test for Count Data
#R:
#R:  data:  smoke_tab
#R:  p-value = 0.005496
#R:  alternative hypothesis: two.sided
```

Linear regression

Simple linear regression

We can perform a simple linear regression in **R** using the `lm` function, for example:

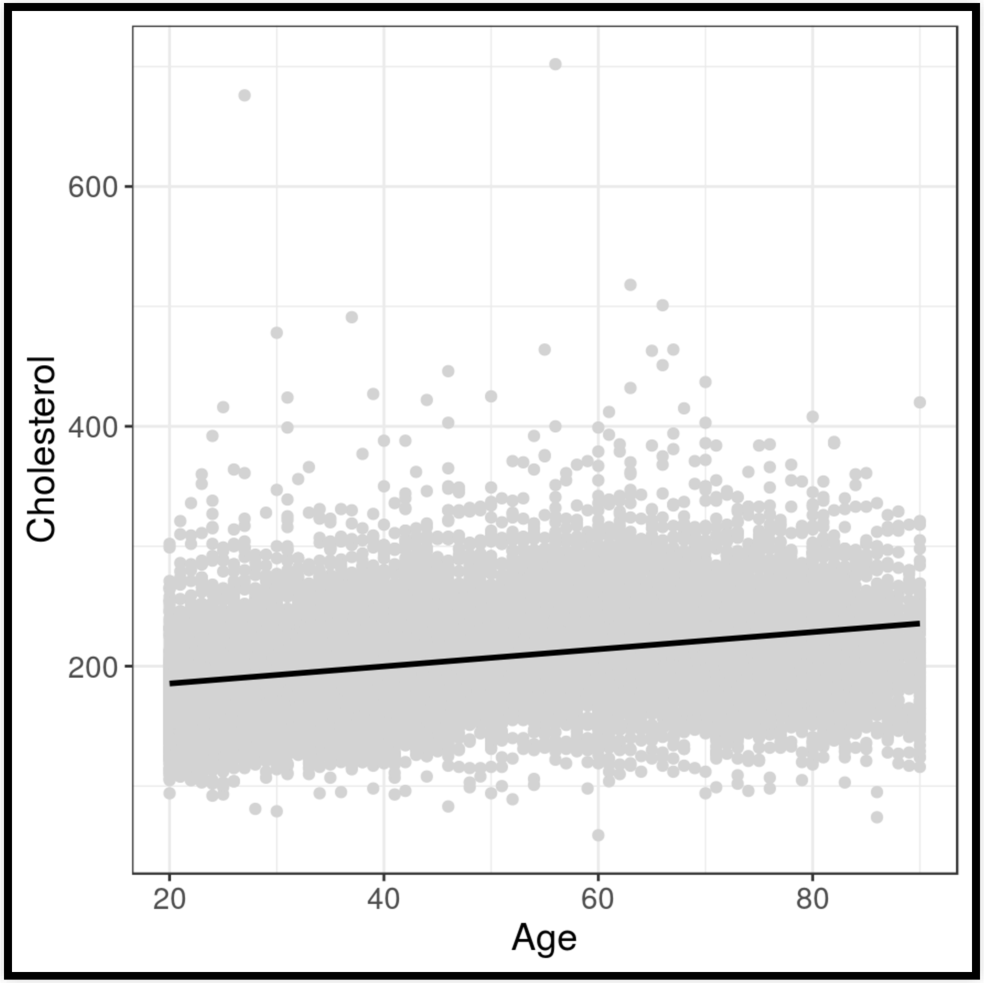
```
lm(y ~ x, data = dataset)
```

`y`: the continuous response variable. `x`: the continuous explanatory variable. `data`: name of the dataframe containing the variables.

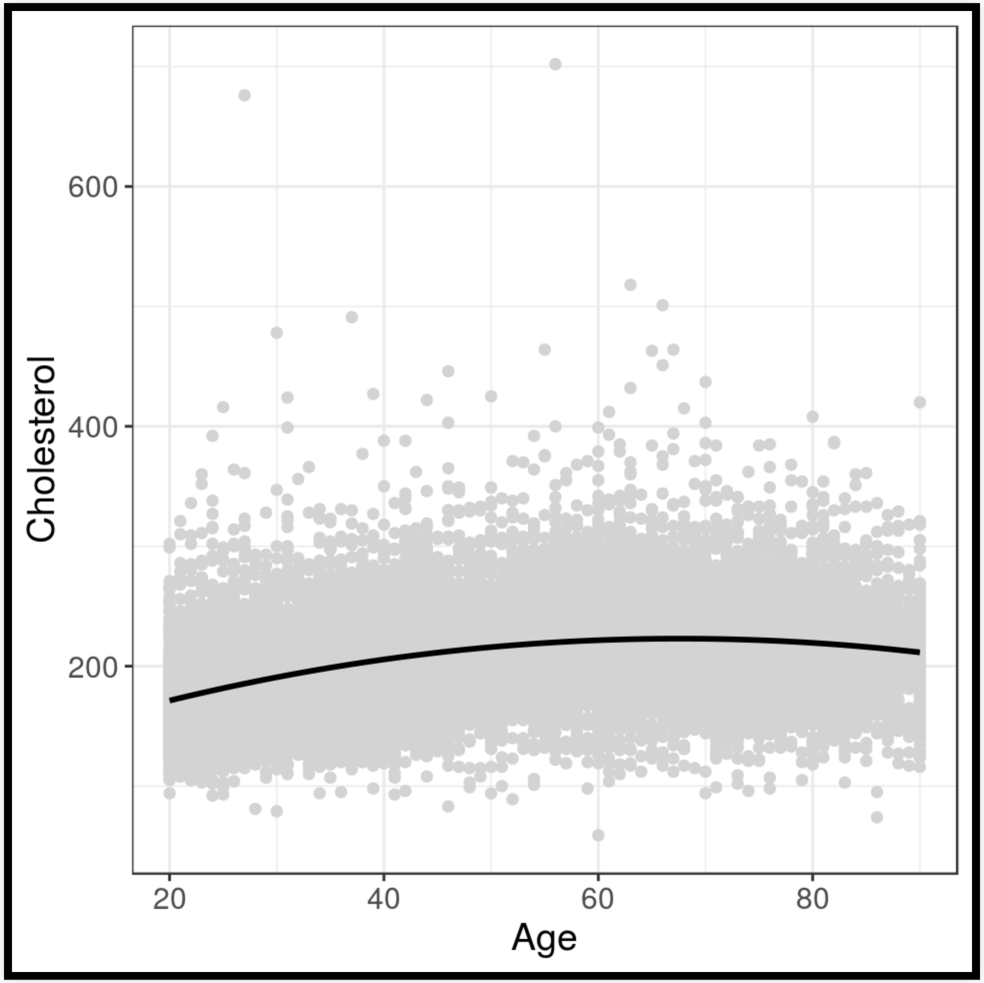
- There can be more than one explanatory variable for a multiple linear regression.
- Since there is only one explanatory variable here, we refer to this a simple linear regression.

Visualise the relationship

`geom_smooth(method = lm)` gives you the fitted line of the simple linear regression!



Visualise the relationship with a quadratic term



Fit the regression model

- We saw that we will need a quadratic term in the model.
- We can fit a quadratic term in **R** using `I(x^2)`:

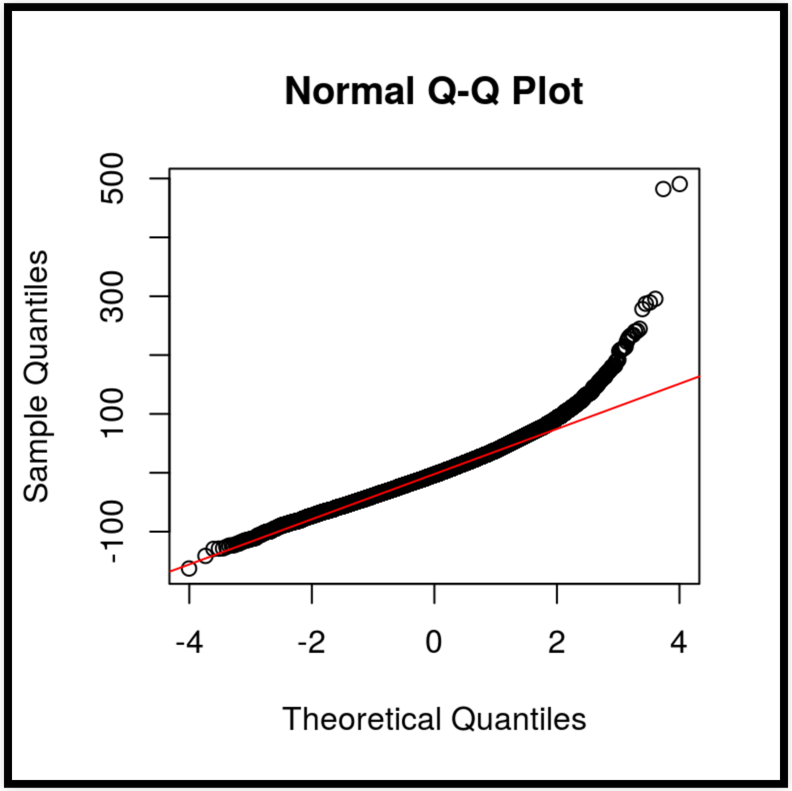
```r
my_lm = lm(Cholesterol ~ Age + I(Age^2), data = patient.df)
```

We now need to check our assumption that the residuals are normally distributed.

Check normality of the residuals

- We can extract the residuals of the model with `resid`.
- We can plot a Quantile-Quantile (QQ) plot with `qqnorm

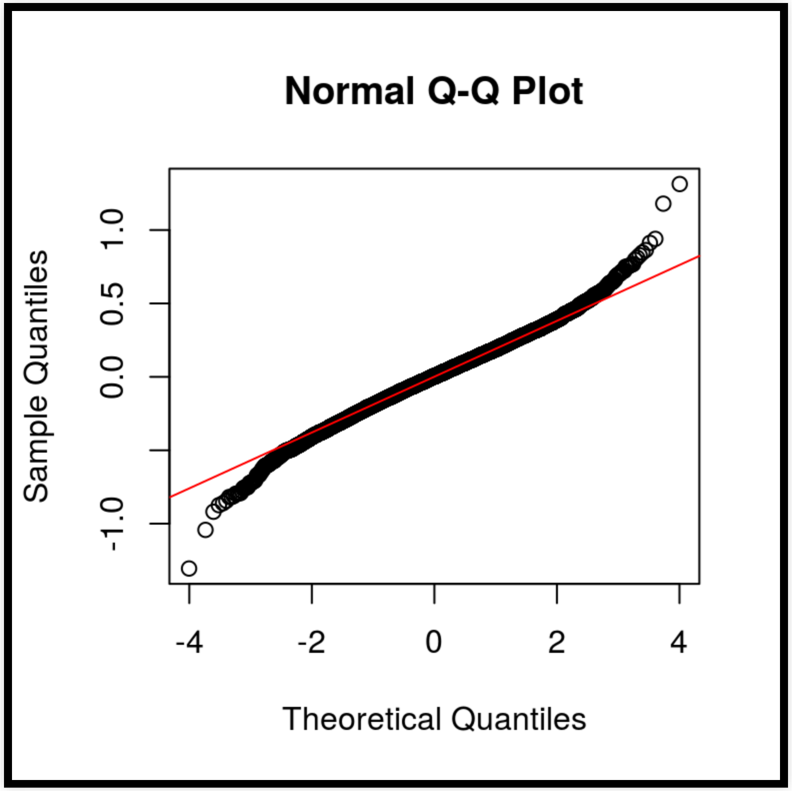
```r
qqnorm(resid(my_lm))
qqline(resid(my_lm), col = "red")
```



Check normality of the residuals

- We can refit the model using log cholesterol as our response variable instead:

```r
my_lm = lm(log(Cholesterol) ~ Age + I(Age^2), data = patient.df)
```



Final regression model

$$\text{log cholesterol} = 4.87 + 0.02 \times \text{Age} + 0.0001 \times \text{Age}^2$$

```r
summary(my_lm)
```

```
#R: 
#R: Call:
#R: lm(formula = log(Cholesterol) ~ Age + I(Age^2), data = patient.df)
#R: 
#R: Residuals:
#R:      Min       1Q   Median       3Q      Max 
#R: -1.30527 -0.12697  0.00406  0.12949  1.31316 
#R: 
#R: Coefficients:
#R:               Estimate Std. Error t value Pr(>|t|)    
#R: (Intercept)  4.868e+00  1.142e-02  426.44   <2e-16 ***
#R: Age          1.555e-02  4.855e-04   32.02   <2e-16 ***
#R: I(Age^2)    -1.161e-04  4.613e-06  -25.17   <2e-16 ***
#R: ---
#R: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#R: 
#R: Residual standard error: 0.1998 on 16059 degrees of freedom
#R:   (968 observations deleted due to missingness)
```

Summary

Model	Function
Student’s t-test	`t.test`
One-way ANOVA	`aov`
Two-way ANOVA	`aov`
Pearson’s Chi-square test	`chisq.test`
Fisher’s exact test	`fisher.test`
Linear regression	`lm`