Lab 03C: Week 10 (Solutions)

Contents

The **specific aims** of this lab are:

- practice performing (two-way) ANOVA and interpreting the output
- conduct post-hoc testing using contrasts
- understand the importance of blocking
- be able to check for interactions (graphically and through an appropriate statistical test)
- identify if treatment effects exist
- introduction to linear regression

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO6 Formulate, evaluate and interpret appropriate linear models to describe the relationships between multiple factors.
- LO8 Create a reproducible report to communicate outcomes using a programming language.

1 Exercises

1.1 Poison and antidotes

In the lectures we considered an experiment with 3 poisons and 4 antidotes (Box and Cox 1964). The

response was **survival time** but a transformed response was used instead, the **reciprocal** of the survival time. The aim of the study was to determine how each antidote affected survival in the presence of each poison.

```
library(tidyverse)
poison_data =
        read_csv("https://raw.githubusercontent.com/DATA2002/data/master/box_cox_survival.

poison_data = poison_data %>%
    mutate(inv_survival = 1/y)  # create the reciprocal survival time variable
glimpse(poison_data)
```

```
Rows: 48
Columns: 4
$ poison        [3m [38;5;246m<chr> [39m [23m "I", "I", "I", "I", "II", "II", "II", "II", "II…
$ antidote      [3m [38;5;246m<chr> [39m [23m "A", "A", "A", "A", "A", "A", "A", "A", "A", "A…
$ y             [3m [38;5;246m<dbl> [39m [23m 0.31, 0.45, 0.46, 0.43, 0.36, 0.29, 0.40, 0.23,…
$ inv_survival  [3m [38;5;246m<dbl> [39m [23m 3.2258065, 2.2222222, 2.1739130, 2.3255814, 2.7…
```

1. Generate summary statistics for each of the treatment combination (including mean, median, standard deviation, interquartile range, sample size). Make sure you don't report too many decimal places in your summary statistics.

```
poison_sum = poison_data %>%
  dplyr::group_by(poison, antidote) %>%
  dplyr::summarise(
    mean = mean(inv_survival),
    median = median(inv_survival),
    sd = sd(inv_survival),
    iqr = IQR(inv_survival),
    n = n()
  )
poison_sum %>% knitr::kable(digits = 2)
```

poison	antidote	mean	median	sd	iqr	n
I	A	2.49	2.27	0.50	0.34	4
I	B	1.16	1.18	0.20	0.18	4
I	C	1.86	1.90	0.49	0.73	4
I	D	1.69	1.56	0.36	0.28	4
II	A	3.27	3.11	0.82	0.96	4
II	B	1.39	1.36	0.55	0.72	4
II	C	2.71	2.68	0.42	0.51	4
II	D	1.70	1.60	0.70	0.70	4
III	A	4.80	4.65	0.53	0.46	4
III	B	3.03	3.02	0.42	0.68	4

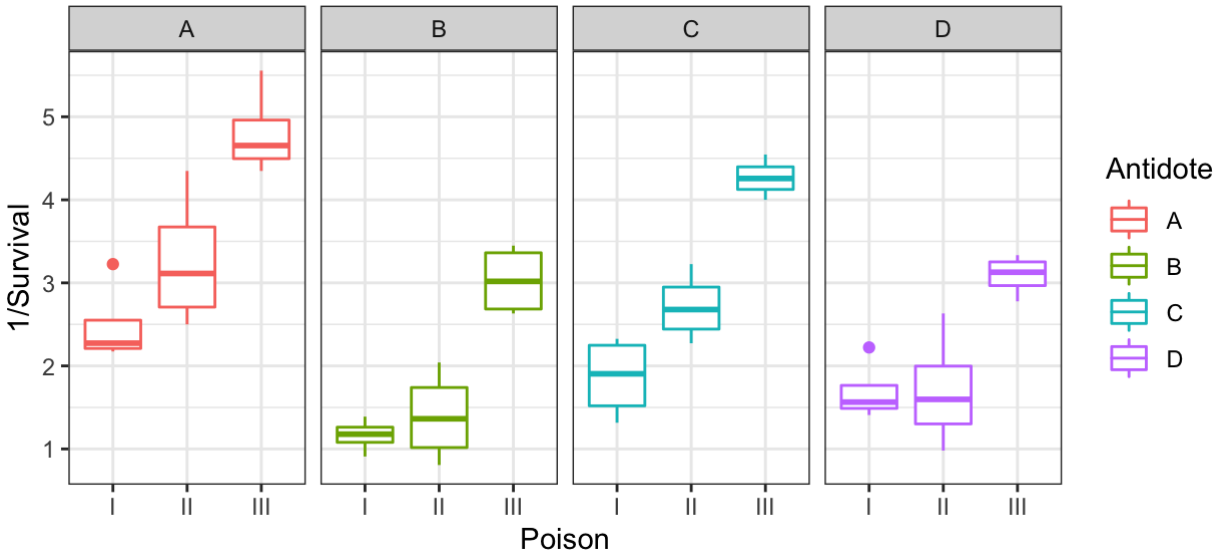
poison	antidote	mean	median	sd	iqr	n
III	C	4.26	4.26	0.23	0.27	4
III	D	3.09	3.13	0.24	0.29	4

2. How many replicates are there in each treatment combination?

There are 4 observations in each treatment combination.

3. Visualise the data using boxplots. Which poison tended to have the lowest survival time (highest reciprocal survival time)?

```
poison_data %>%
  ggplot() +
  aes(y = inv_survival, x = poison,colour = antidote) +
  geom_boxplot() +
  theme_bw() +
  facet_wrap(~ antidote, ncol = 4) +
  labs(y = "1/Survival", x = "Poison", colour = "Antidote")
```



In the boxplots poison III tends to have the highest median reciprocal survival time (lowest median survival time) within each antidote grouping.

4. Write an appropriate model formula for a two-way ANOVA with interactions.

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ijk}$$

where μ is the overall mean, α_i and γ_j are treatment effects (differences between treatment group means and the overall mean), $(\alpha\gamma)_{ij}$ are the interaction effects and $\varepsilon_{ijk} \sim N(0, \sigma^2)$. We also require the following constraints:

- $\sum_i \alpha_i = 0$
- $\sum_j \gamma_j = 0$
- For each j, $\sum_i (\alpha\gamma)_{ij} = 0$
- For each i, $\sum_j (\alpha\gamma)_{ij} = 0$

5. Use R to fit the ANOVA model described above and generate an ANOVA table.

```
a1 = aov(inv_survival ~ poison * antidote, data = poison_data)
summary(a1)  # could also use anova(a1)
```

```
               Df Sum Sq Mean Sq F value   Pr(>F)
poison          2  34.88  17.439   72.64 2.31e-13 ***
antidote        3  20.41   6.805   28.34 1.38e-09 ***
poison:antidote 6   1.57   0.262    1.09    0.387
Residuals      36   8.64   0.240
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Can the interaction effect be dropped from the model? Why or why not?

Hypothesis: H_0: $(\alpha\gamma)_{ij} = 0$ for all $i = 1, 2, 3, 4$ and $j = 1, 2, 3$ vs H_1: not all $(\alpha\gamma)_{ij} = 0$

Assumptions: (see below)

Test statistic: $T = \frac{\text{Mean Sq Interaction}}{\text{Mean Sq Residual}}$ which follows an F distribution with $(a-1)(b-1)$ and $ab(n-1)$ degrees of freedom under H_0.

Observed test statistic: $t_0 = 1.09$

p-value: $P(F_{6,36} \geq t_0) = P(F_{6,36} \geq 1.09) = 0.3867$

Conclusion: Since the p-value is greater than 0.05, we do not reject H_0. I.e. the data are consistent with H_0 that the interaction effects are all zero.

Hence, we could drop the interaction term from the model.

7. Test for a poison treatment effect.

Having found that the interaction term is not significant, we can proceed to consider whether or not there are any differences between the means of the *main* effects. I.e. we can look at the p-values associated with `poison` and `antidote`. [Note: we could also refit the model without the interaction term (drop the interaction term from the model), however, if when we were designing the experiment, we hypothesised that there *should* be an interaction, it's safer to leave it in and conduct inferences

using the full model. By "safer", I mean the model won't suffer from potential model misspecification.]

Let α_1, α_2 and α_3 be the treatment effects for the three levels of the poison variable (poisons I, II and III, respectively).

Hypothesis: H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ vs H_1: not all α_j equal to 0

Assumptions: (see below)

Test statistic: $T = \frac{\text{Mean Sq Poison}}{\text{Mean Sq Residual}}$ which follows an F distribution with $a - 1$ and $ab(n - 1)$ degrees of freedom under H_0.

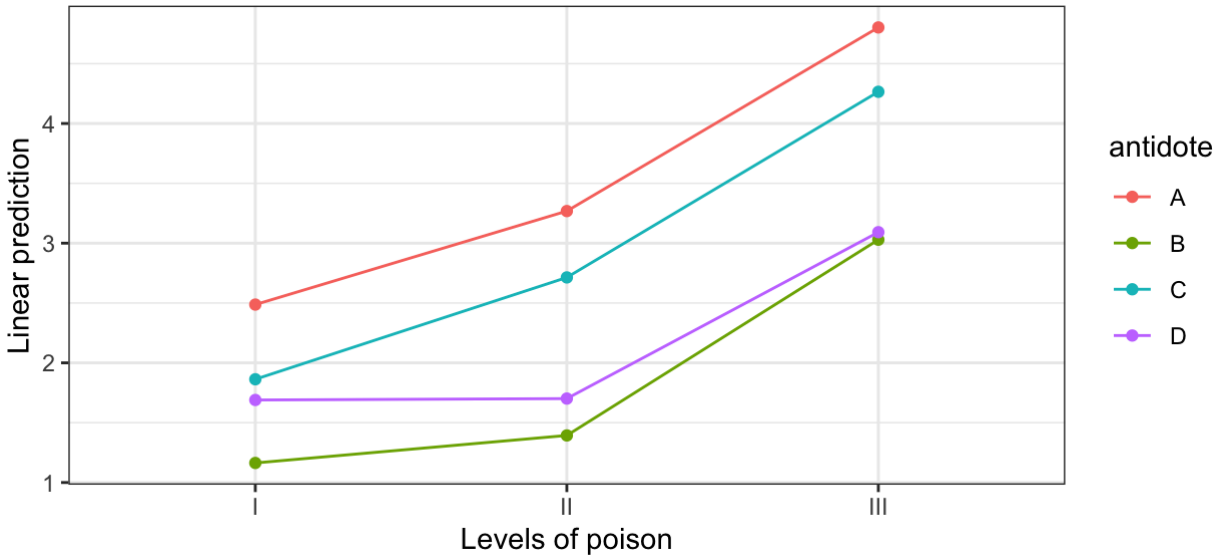
Observed test statistic: $t_0 = 72.64$

p-value: $P(F_{2,36} \geq t_0) = P(F_{2,36} \geq 72.64) < 0.001$

Conclusion: Since the p-value is less than 0.05, we reject H_0. There is strong evidence that the treatment effects are not all the same. I.e. there is a significant difference in the (reciprocal) survival time between the three poisons.

8. Generate an interaction plot and comment on what you see. Do your observations agree with the results from the ANOVA table? Hint: use ggplot to plot the treatment combination means directly or you can the `emmip()` function from the **emmeans** package (Lenth 2018).

```
library(emmeans)
emmip(a1, antidote ~ poison) + theme_bw()
```



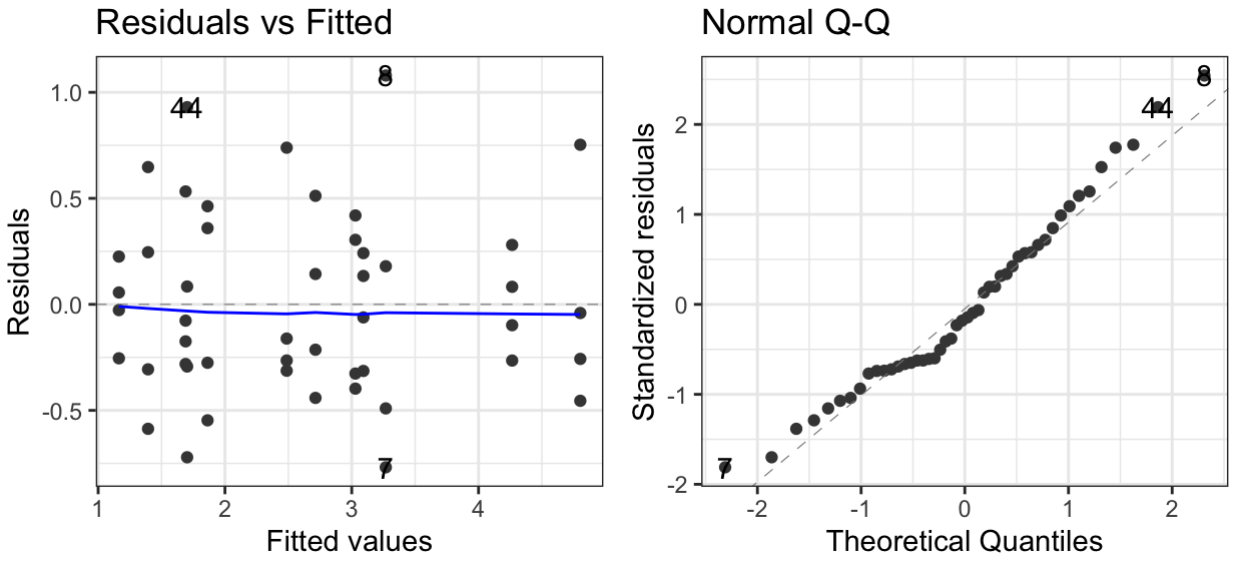
Since there are no intersections between lines this suggests that there's no interaction effect, as found above using the formal test.

9. What are the assumptions required for the ANOVA test to be valid? Generate appropriate diagnostic plots. Comment as to whether or not the assumptions are satisfied with reference to the

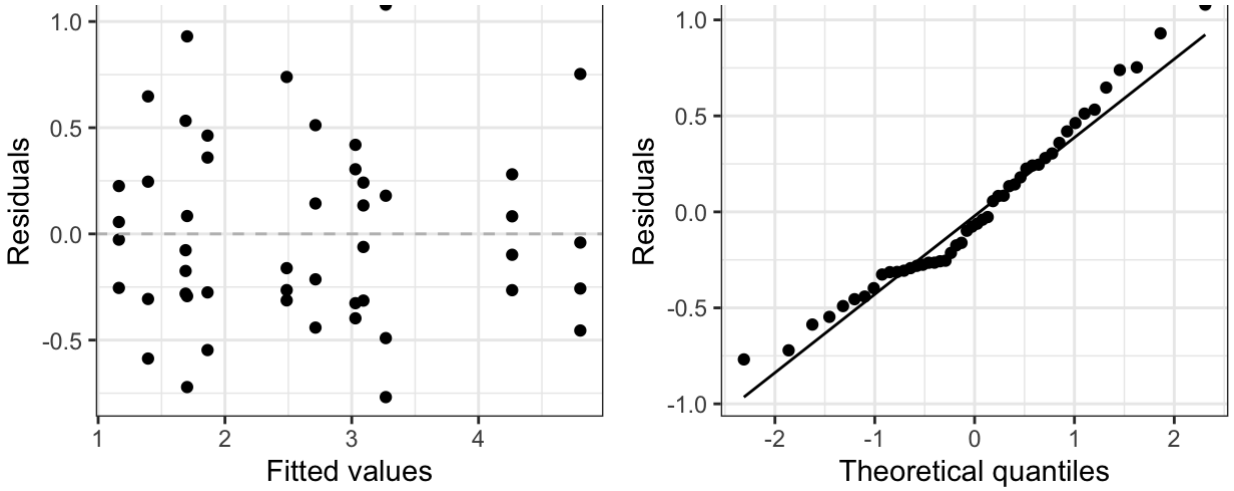
diagnostic plots?

The ANOVA test assumes the residuals to follow a **normal distribution** with **constant variance**. We can check this using a scatter plot of the residuals against the fitted values (looking for **homoskedasticity**: constant error variance over the range of fitted values) and a normal quantile plot (looking to see that the points are close to the diagonal line).

```
# using autoplot() from the ggfortify package
library(ggfortify)
autoplot(a1, which = 1:2) + theme_bw()
```




```
# manually extracting the fitted values and residuals
poison_data = poison_data %>%
  mutate(
    fitted = a1$fitted.values,
    resid = a1$residuals
  )
d1 = ggplot(poison_data, aes(x = fitted, y = resid)) +
  geom_point() + geom_hline(yintercept = 0, colour = "gray", lty = 2) +
  theme_bw() + labs(title = "Residuals vs fitted", x = "Fitted values", y = "Residuals")
d2 = ggplot(poison_data, aes(sample = resid)) +
  geom_qq() + geom_qq_line() + theme_bw() +
  labs(title = "Normal QQ of the residuals", x = "Theoretical quantiles", y = "Residuals")
gridExtra::grid.arrange(d1, d2, ncol = 2)
```



- **Residual plot:** It shows that the spread of residuals is roughly even above and below the central line and across the range of fitted values. Hence the equality of variance assumption is approximately satisfied.
- **QQ plot:** The points are all reasonably close to the diagonal line. Hence the normality assumption for residuals is approximately satisfied.

We could also go on and do post hoc pairwise tests, see the lecture for details.

1.2 Manufacturing

The data below gives the number of units produced in a day by 4 different machines, A, B, C and D, on each of 5 different days. The days may be regarded as a nuisance factor. We wish to compare the production levels of the machines and consider the days as blocks.

```
library(tidyverse)
manufacturing =
        read_csv("https://raw.githubusercontent.com/DATA2002/data/master/manufacturing.csv

knitr::kable(manufacturing)
```

Day	A	B	C	D
Mon	293	308	323	333
Tue	298	353	343	363
Wed	280	323	350	368
Thu	288	358	365	345
Fri	260	343	340	330

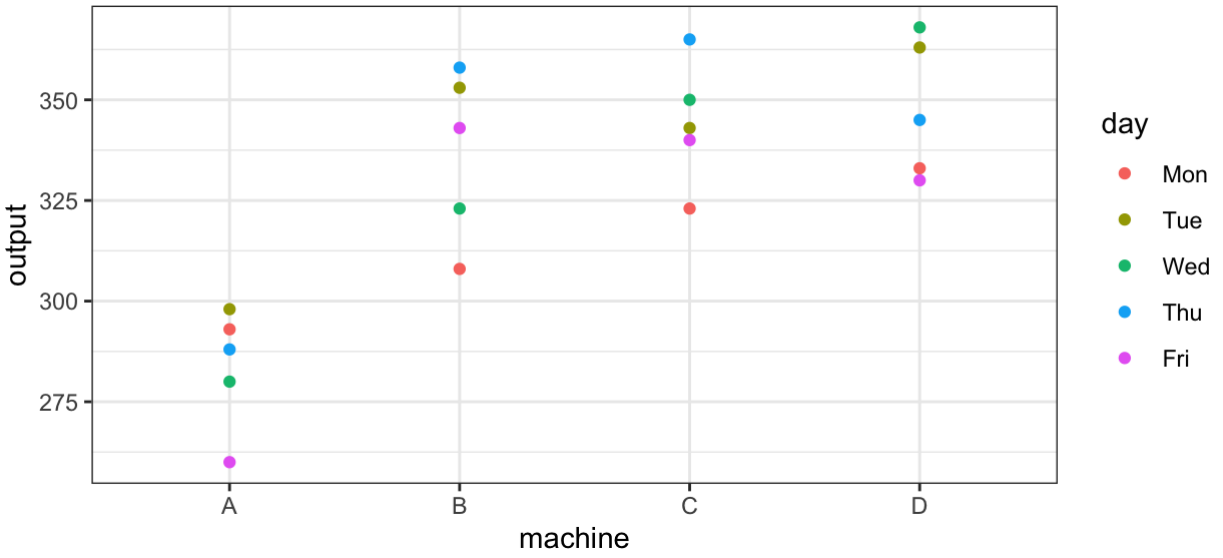
1. Convert the data from its current “wide” format to “long” format.

```
manuf = gather(manufacturing, key = "machine", value = "output", A:D) %>%
    mutate(day = factor(Day, levels = c("Mon", "Tue", "Wed", "Thu", "Fri")))
glimpse(manuf)
```

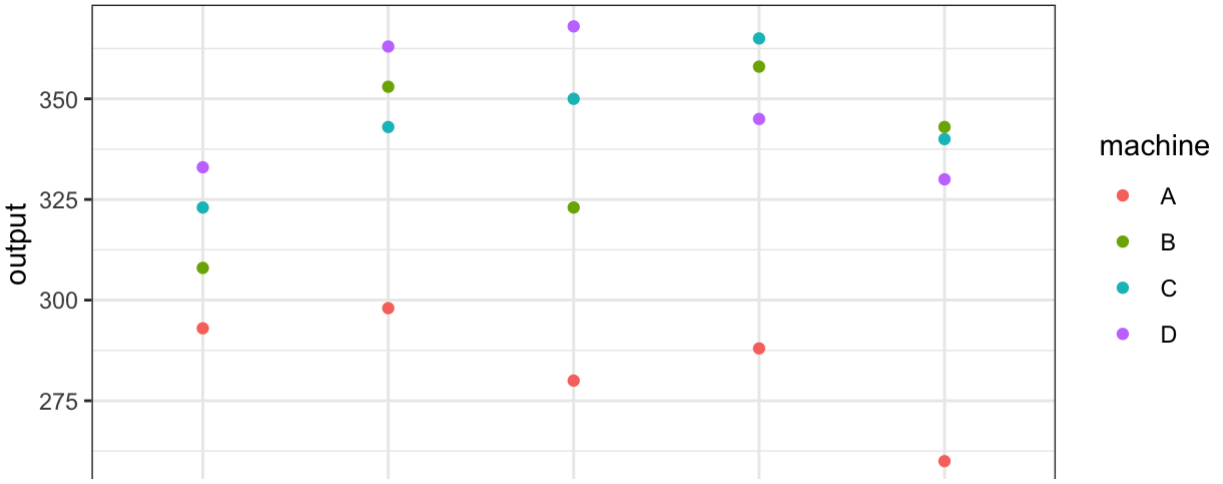
```
Rows: 20
Columns: 4
$ Day      [3m [38;5;246m<chr> [39m [23m "Mon", "Tue", "Wed", "Thu", "Fri", "Mon", "Tue", "We…
$ machine  [3m [38;5;246m<chr> [39m [23m "A", "A", "A", "A", "A", "B", "B", "B", "B", "B", "C…
$ output   [3m [38;5;246m<dbl> [39m [23m 293, 298, 280, 288, 260, 308, 353, 323, 358, 343, 32…
$ day      [3m [38;5;246m<fct> [39m [23m Mon, Tue, Wed, Thu, Fri, Mon, Tue, Wed, Thu, Fri, Mo…
```

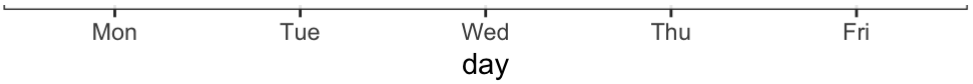
2. Summarise and visualise the data. What do you notice?

```
manuf %>%
    ggplot(aes(x = machine, y = output, colour = day)) + geom_point() +
    theme_bw()
```




```
manuf %>%
    ggplot(aes(x = day, y = output, colour = machine)) + geom_point() +
    theme_bw()
```






```
manuf %>%
    group_by(day) %>%
    dplyr::summarise(mean = mean(output), median = median(output), sd = sd(output),
        n = n()) %>%
    knitr::kable(digits = 2)
```

day	mean	median	sd	n
Mon	314.25	315.5	17.50	4
Tue	339.25	348.0	28.69	4
Wed	330.25	336.5	38.27	4
Thu	339.00	351.5	35.00	4
Fri	318.25	335.0	39.23	4

```
manuf %>%
    group_by(machine) %>%
    dplyr::summarise(mean = mean(output), median = median(output), sd = sd(output),
        n = n()) %>%
    knitr::kable(digits = 2)
```

machine	mean	median	sd	n
A	283.8	288	14.87	5
B	337.0	343	21.04	5
C	344.2	343	15.29	5
D	347.8	345	17.20	5

- Machine A seems to be outputting less than the other three machines.
- Monday and Friday seem to be lower output days.

3. How many observations do we have in each treatment group?

The common sample (block) size is $n = 5$ (5 days per machine).

4. Write an appropriate model formula for a two-way ANOVA with blocks.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where μ is the overall mean effect, α_i is the treatment effect for machine i ($i = 1, 2, 3, 4$) and β_j is the block effect for day j ($j = 1, 2, 3, 4, 5$) and the ε_{ij}'s are iid $N(0, \sigma^2)$. We require the following constraints:

- $\sum_i \alpha_i = 0$
- $\sum_j \beta_j = 0$

5. Test if there is a machine effect.

```
manuf_aov = aov(output ~ day + machine, data = manuf)
summary(manuf_aov)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
day          4   2146     537   2.452    0.103
machine      3  13445    4482  20.478 5.18e-05 ***
Residuals   12   2626     219
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let α_1, α_2, α_3 and α_4 be the treatment effects for the four machines (A, B, C and D).

Hypothesis: H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ vs H_1: not all $\alpha_j = 0$.

Assumptions: residuals follow a normal distribution with common variance.

Test statistic: $T = \frac{\text{Mean Sq Machine}}{\text{Mean Sq Residual}}$ which follows an F distribution with $a - 1$ and $(a-1)(b-1)$ degrees of freedom under H_0.

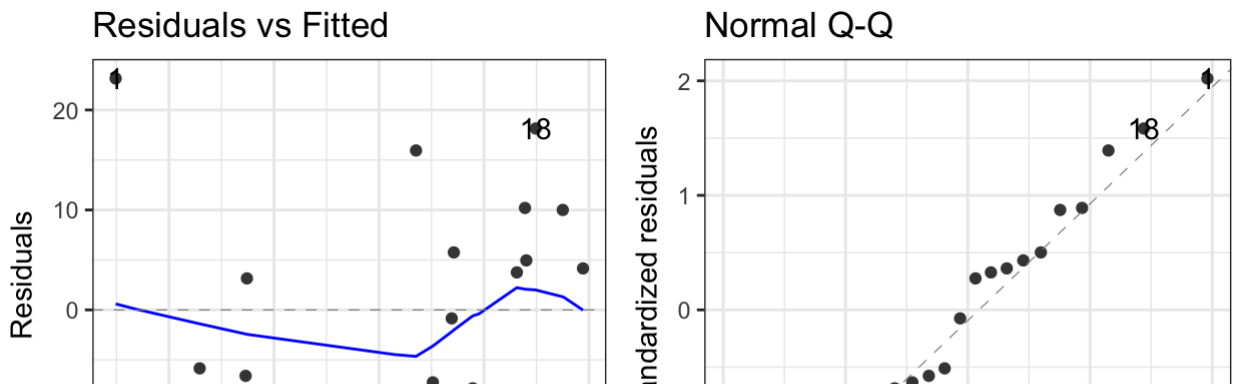
Observed test statistic: $t_0 = 20.478$

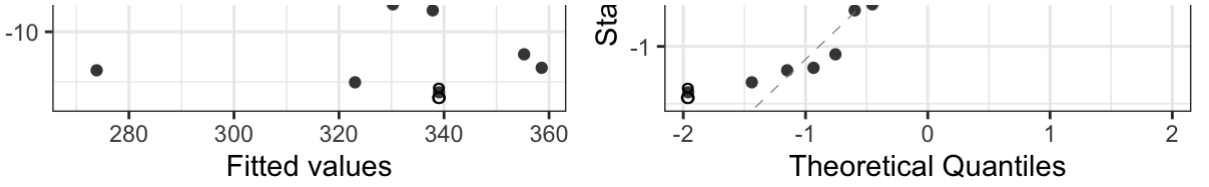
p-value: $P(F_{3,12} \geq t_0) = P(F_{3,12} \geq 20.478) < 0.001$

Conclusion: Since the p-value is less than 0.05, we reject H_0. There is strong evidence that the treatment effects are not all the same. I.e. there is a significant difference between the mean outputs of the four difference machines.

6. Check and comment on the ANOVA assumptions.

```
autoplot(manuf_aov, which = 1:2) + theme_bw()
```





There is no apparent pattern in the residual vs fitted values plot, hence the common variance assumption is OK. Similarly, the points in the normal QQ plot are all reasonably close to the diagonal line, which suggests that the normality assumption is at least approximately satisfied.

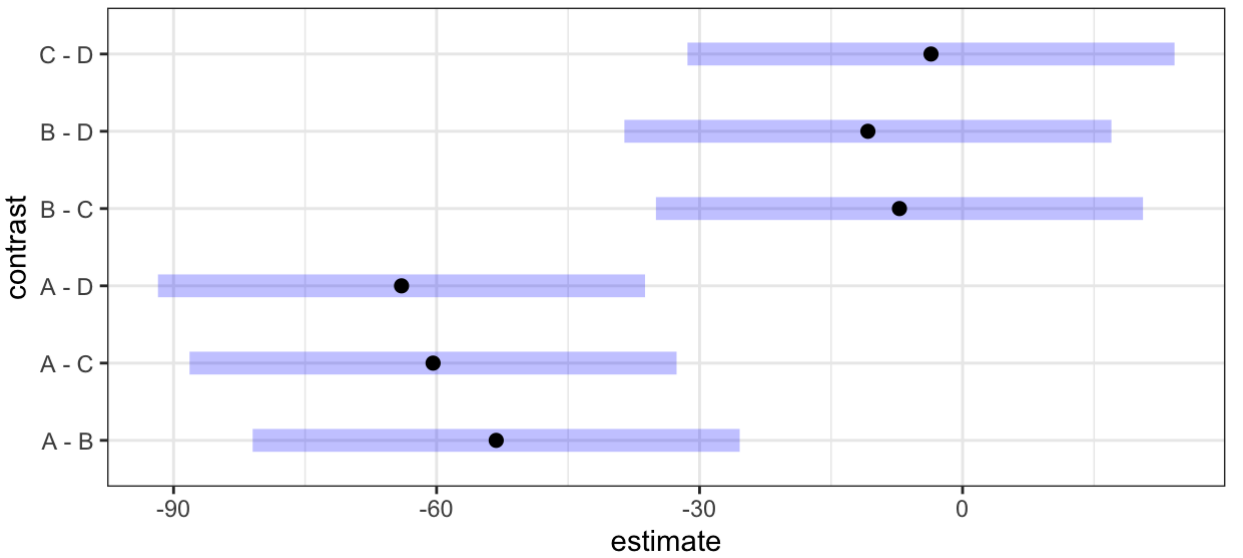
7. Perform post hoc tests to see which pairs of machines have significantly different means.

```
library(emmeans)
em_machine = emmeans(manuf_aov, ~machine)
contrast(em_machine, method = "pairwise", adjust = "tukey")
```

```
contrast estimate   SE df t.ratio p.value
A - B       -53.2 9.36 12  -5.686  0.0005
A - C       -60.4 9.36 12  -6.456  0.0002
A - D       -64.0 9.36 12  -6.840  0.0001
B - C        -7.2 9.36 12  -0.770  0.8666
B - D       -10.8 9.36 12  -1.154  0.6649
C - D        -3.6 9.36 12  -0.385  0.9797

Results are averaged over the levels of: day
P value adjustment: tukey method for comparing a family of 4 estimates
```

```
contrast(em_machine, method = "pairwise", adjust = "tukey") %>%
    plot() + theme_bw()
```



We see that machine A is significantly different to the other machines (which in turn are not significantly different to each other).

Comment: this is a block design, so we're not really interested in considering if *day* is significant - looking at the p-value for day, it isn't significant, but it has still played an important role in reducing the residual mean square and hence improved the sensitivity of the tests for differences among machines.

1.3 Hubble

Hubble (1929) investigated the relationship between distance of a galaxy from the earth and the velocity with which it appears to be receding. This information can then be used to estimate the time since "Big Bang".

Hubble's law is as follows:

$$\text{Recession velocity} = H_0 \times \text{Distance},$$

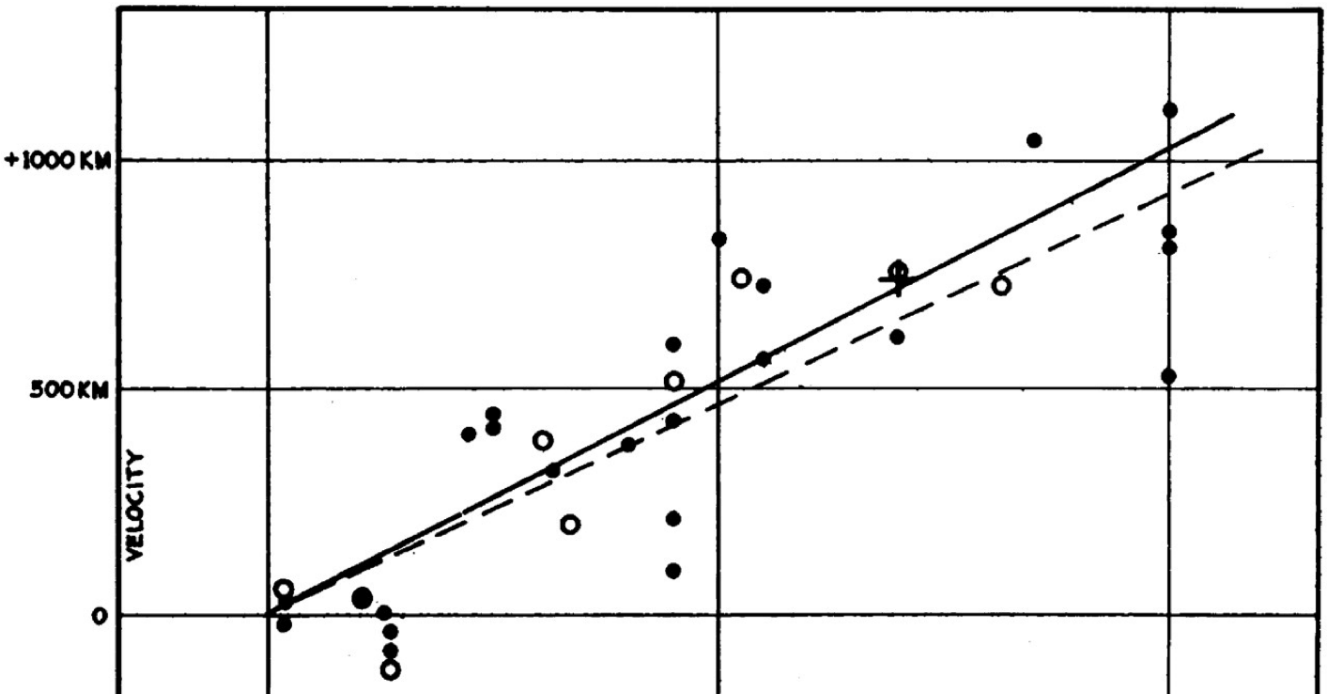
where H_0 is Hubble's constant thought to be about 75 km/sec/Megaparsec.

The data can be imported as follows:

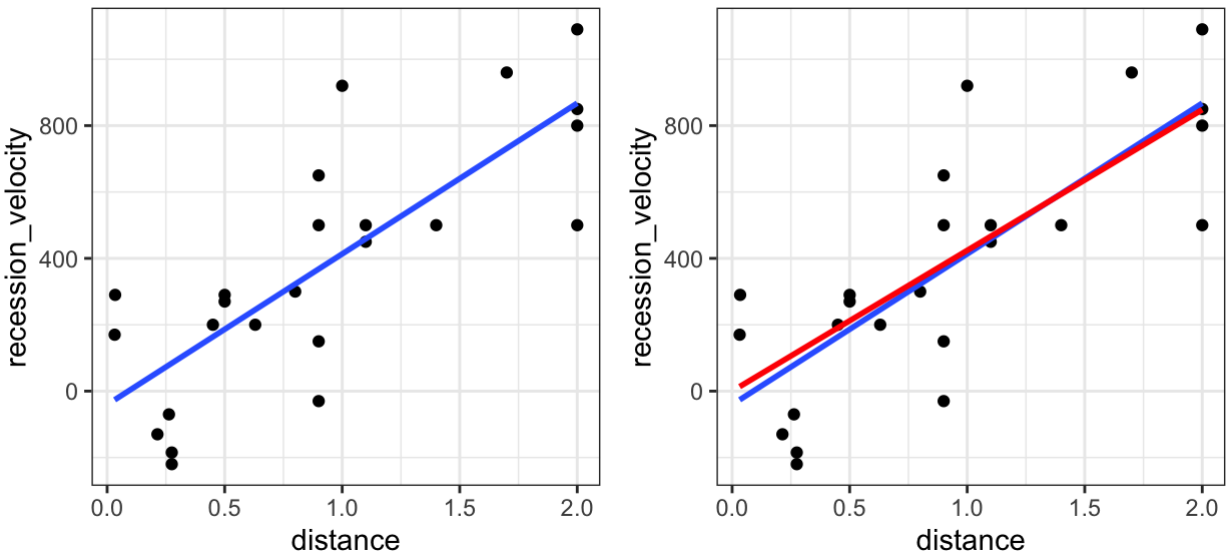
```
library(tidyverse)
hubble = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/Hubble.txt")
glimpse(hubble)
```

```
Rows: 24
Columns: 2
$ distance           [3m [38;5;246m<dbl> [39m [23m 0.032, 0.034, 0.214, 0.263, 0.275, 0.275,…
$ recession_velocity [3m [38;5;246m<dbl> [39m [23m 170, 290, -130, -70, -185, -220, 200, 290…
```

1. What will be the most effective visualisation to look at this data. Add a line of best fit to your plot. Compare your results to Figure 1 from the PNAS paper (below).




```
hubble_scatter = hubble %>%
  ggplot() +
  aes(x = distance, y = recession_velocity) +
  geom_point() +
  theme_bw()
## Adding a regression line
hubble_lm = hubble_scatter +
  geom_smooth(method="lm", se = FALSE)
## Adding a different line with intercept being zero
hubble_lm2 = hubble_lm +
  geom_smooth(method='lm', formula = y ~ -1 + x,
              col="red", se = FALSE)
gridExtra::grid.arrange(hubble_lm, hubble_lm2, ncol=2)
```



2. Does the regression make sense with the constant term = 0? (if the distance from the earth is zero, is the velocity from the earth 0?) Fit the model allowing for an intercept and test the null hypothesis that the intercept is equal to zero. Fit another regression that does not allow an intercept and write down your estimate for Hubble's constant. You can force the regression line to have a zero intercept by putting a `-1` in the model formula, e.g. `lm(y ~ x - 1)`

It's quite unusual to force your regression model to not have an intercept, in almost all future settings we will allow an intercept in the model (even if it is "insignificant" i.e. has a large p-value). We're only forcing it out of the model here because it makes sense from a physics perspective to have the estimated line pass through the origin.

```
hfit1 = lm(recession_velocity ~ distance, data = hubble)
summary(hfit1)
```

```
Call:
```

```
Call:
lm(formula = recession_velocity ~ distance, data = hubble)

Residuals:
    Min      1Q  Median      3Q     Max
-397.96 -158.10  -13.16  148.09  506.63

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -40.78      83.44  -0.489     0.63
distance      454.16      75.24   6.036 4.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 232.9 on 22 degrees of freedom
Multiple R-squared:  0.6235,    Adjusted R-squared:  0.6064
F-statistic: 36.44 on 1 and 22 DF,  p-value: 4.477e-06
```

Let the population model be:

$$\text{Recession velocity} = \beta_0 + \beta_1 \text{Distance} + \varepsilon,$$

we want to test if $\beta_0 = 0$.

Hypothesis: $H_0\colon\ \beta_0 = 0$ vs $H_1\colon\ \beta_0 \neq 0$

Assumptions: The residuals ε_i are iid $N(0, \sigma^2)$ and there is a linear relationship between y and x. [Checked below.]

Test statistic: $T = \dfrac{\hat{\beta}_0}{\text{SE}(\hat{\beta}_0)} \sim t_{n-2}$ under H_0.

Observed test statistic: $t_0 = -0.489$ (from R output)

p-value: $2P(t_{n-2} \geq |t_0|) = 2P(t_{n-2} \geq 0.489) = 0.63$

Conclusion: We do not reject H_0 as the p-value is quite large. I.e. the intercept is not significantly different to zero.

We can fit the model forcing the intercept to be exactly 0 (i.e. don't allow for an intercept in the model). *We're only doing this because it is dictated by the underlying physics that the model is trying to describe - in general you wouldn't be checking for the significance of the intercept, you'd just leave it in the model regardless.*

```
hfit2 = lm(recession_velocity ~ -1 + distance, data = hubble)
summary(hfit2)
```

```
Call:
```

```
lm(formula = recession_velocity ~ -1 + distance, data = hubble)

Residuals:
   Min     1Q Median     3Q    Max
-411.5 -191.3   -7.1  128.0  496.1

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
distance   423.94      42.15   10.06 6.87e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 229 on 23 degrees of freedom
Multiple R-squared:  0.8147,    Adjusted R-squared:  0.8067
F-statistic: 101.1 on 1 and 23 DF,  p-value: 6.869e-10
```

We can compare the two models nicely using the **stargazer** package or the **sjPlot** package.

Using the **stargazer** package:

```
library(stargazer)
stargazer(hfit1, hfit2, type = "html")
```

	Dependent variable:	
	recession_velocity	
	(1)	(2)
distance	454.158***	423.937***
	(75.237)	(42.154)
Constant	-40.784	
	(83.439)	
Observations	24	24
R^2	0.624	0.815
Adjusted R^2	0.606	0.807
Residual Std. Error	232.911 (df = 22)	229.025 (df = 23)
F Statistic	36.438*** (df = 1; 22)	101.140*** (df = 1; 23)
Note:		*p<0.1; **p<0.05;*** p<0.01

Using the **sjPlot** package:

```
library(sjPlot)
tab_model(hfit1, hfit2, show.ci = FALSE)
```

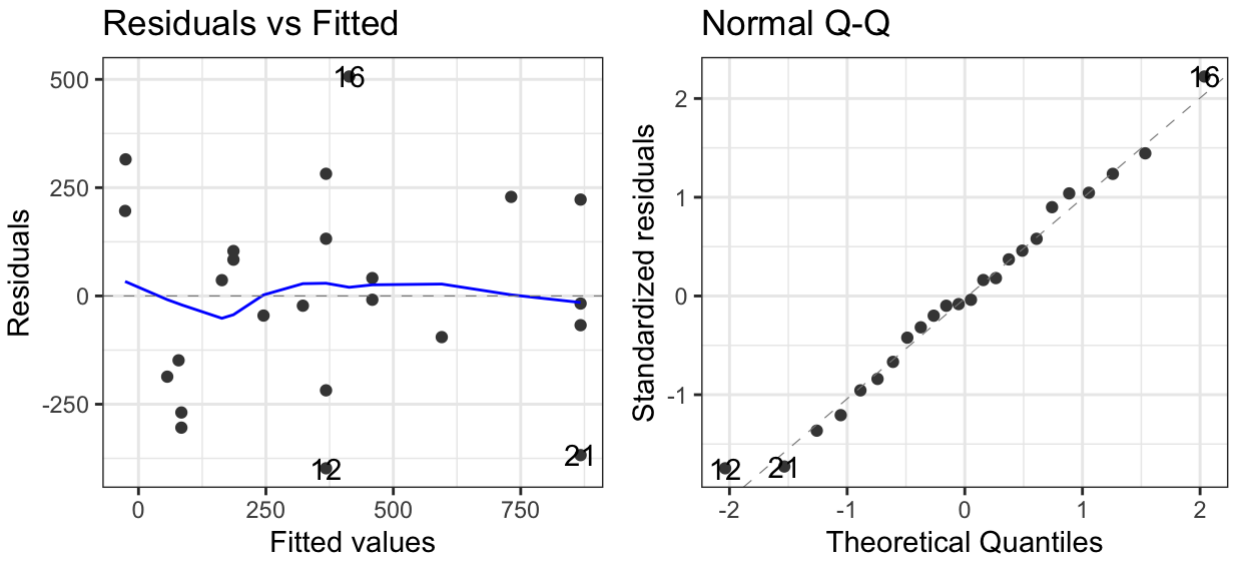
	recession velocity		**recession velocity**	
Predictors	*Estimates*	*p*	*Estimates*	*p*

(Intercept)	-40.78	0.630		
distance	454.16	**<0.001**	423.94	**<0.001**
Observations	24		24	
R^2 / R^2 adjusted	0.624 / 0.606		0.815 / 0.807	

Note that it looks like the R^2 is higher for the model without an intercept, but the reported R^2 value is calculated differently for models where an intercept is not allowed, and it cannot be compared to models which do allow an intercept. See here for some discussion around this. In general forcing your estimated regression model to pass through the origin is not a good idea.

3. Generate plots to check for equal variance and the normality of the residuals.

```
library(ggfortify)
autoplot(hfit1, which = 1:2) + theme_bw()
```



In the residual vs fitted values plot, there is no obvious pattern in the spread of the residuals across the range of fitted values. It looks like the homoskedasticity assumption is satisfied as the points are roughly equally spread over the range of fitted values.

In the normal QQ plot, the points are all quite close to the diagonal line, suggesting that the normality assumption is comfortably satisfied.

4. (Optional) Why isn't our estimated coefficient 75? For fun, have a look on the web to see the various Hubble constant estimate over the years.

Hubble started off being way off (around 500), and successive experiments over the years brought the estimate down as they got better at measuring things. https://en.wikipedia.org/wiki/Hubble%27s_law

2 For practice after the computer lab

2.1 Tooth growth

The data set `ToothGrowth.txt` has measurements of tooth growth (len) of guinea pigs for different dosages of Vitamin C (dose) and two different delivery methods (supp). The response is the length of odontoblasts (len) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods: orange juice or ascorbic acid (McNeil 1977).

> *Perform a two-way analysis of variance of tooth growth modelled by dosage and delivery method. The following questions help you with this analysis.*

```
library(tidyverse)
tooth = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/toothgrowth.txt")
glimpse(tooth)
```

```
Rows: 10
Columns: 6
$ VC05  [3m [38;5;246m<dbl> [39m [23m 4.2, 11.5, 7.3, 5.8, 6.4, 10.0, 11.2, 11.2, 5.2, 7.0
$ VC1   [3m [38;5;246m<dbl> [39m [23m 16.5, 16.5, 15.2, 17.3, 22.5, 17.3, 13.6, 14.5, 18.8, 1…
$ VC2   [3m [38;5;246m<dbl> [39m [23m 23.6, 18.5, 33.9, 25.5, 26.4, 32.5, 26.7, 21.5, 23.3, 2…
$ OJ05  [3m [38;5;246m<dbl> [39m [23m 15.2, 21.5, 17.6, 9.7, 14.5, 10.0, 8.2, 9.4, 16.5, 9.7
$ OJ1   [3m [38;5;246m<dbl> [39m [23m 19.7, 23.3, 23.6, 26.4, 20.0, 25.2, 25.8, 21.2, 14.5, 2…
$ OJ2   [3m [38;5;246m<dbl> [39m [23m 25.5, 26.4, 22.4, 24.5, 24.8, 30.9, 26.4, 27.3, 29.4, 2…
```

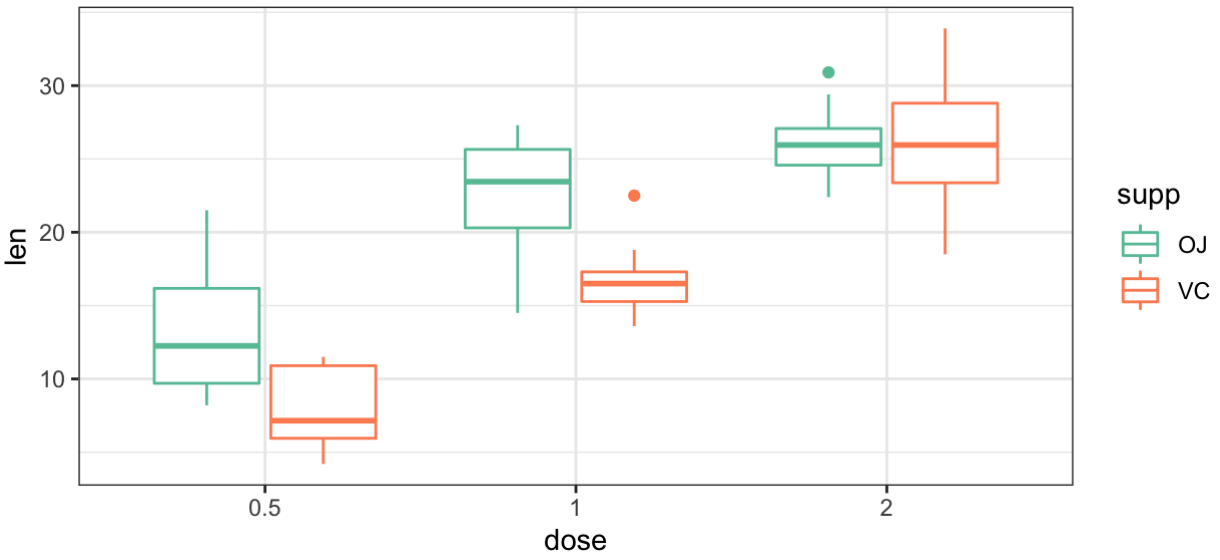
1. Reshape the data so that it is in long format (as opposed to wide format). Use the `gather()` function to do this. Call the `key` variable `group` and the `value` variable `len`. Separate the newly created `group` variable into the two variables that make it up (`supp` and `dose`). Hint: the `stringr::str_extract()` function might be useful here, it can extract characters and/or numbers from a column. In the dose column convert `"05"` to `"0.5"` (`ifelse()` might work well for you here, `dplyr::case_when()` is another alternative, but it's overkill in this situation).

```
tooth_df = tooth %>%
  gather(key = "group", value = "len") %>%
  mutate(supp = stringr::str_extract(group, "[aA-zZ]+"), # extract the letters
         dose = stringr::str_extract(group, "[0-9]+"), # extract the numbers
         dose = ifelse(dose == "05", "0.5", dose))
```

Note that dose is still a character variable (not a numeric variable) but that's OK because we want to treat the different doses as different levels of a factor variable.

2. Visualise the data using side-by-side boxplots.

```
ggplot(tooth_df, aes(x = dose, y = len, color = supp)) + geom_boxplot() +
    scale_color_brewer(palette = "Set2") + theme_bw()
```



3. Generate summary statistics for each of the treatment combinations. How many observations are there in each treatment combination?

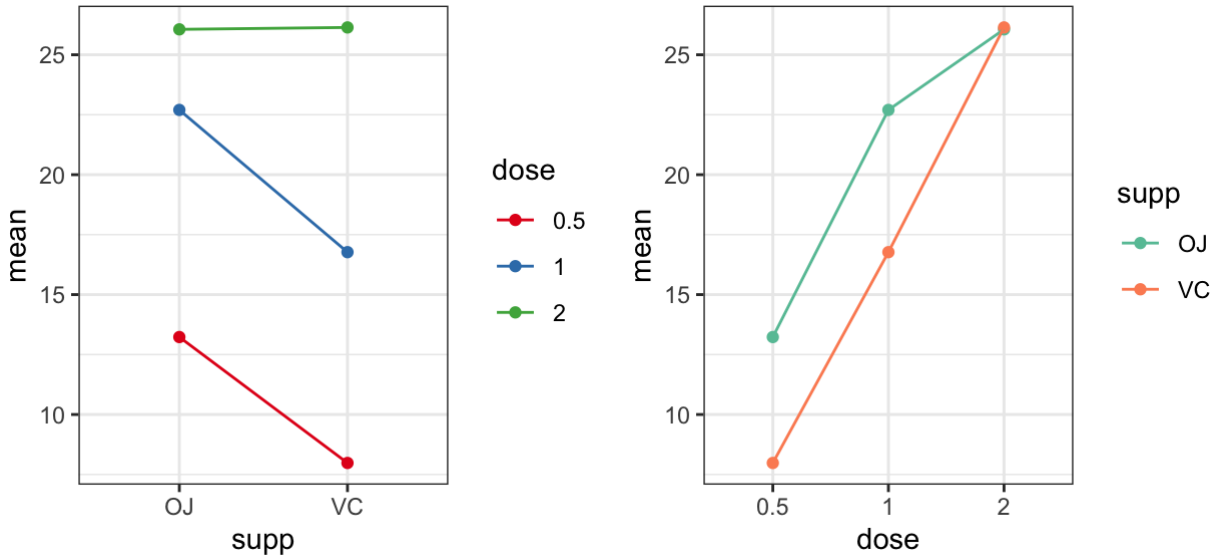
```
tooth_sum = tooth_df %>%
    group_by(supp, dose) %>%
    dplyr::summarise(mean = mean(len), median = median(len), sd = sd(len),
        iqr = IQR(len), n = n())
tooth_sum %>%
    knitr::kable(digits = 2)
```

supp	dose	mean	median	sd	iqr	n
OJ	0.5	13.23	12.25	4.46	6.48	10
OJ	1	22.70	23.45	3.91	5.35	10
OJ	2	26.06	25.95	2.66	2.50	10
VC	0.5	7.98	7.15	2.75	4.95	10
VC	1	16.77	16.50	2.52	2.03	10
VC	2	26.14	25.95	4.80	5.43	10

4. Generate interaction plots. Does it look like there's an interaction between supplement and dose?

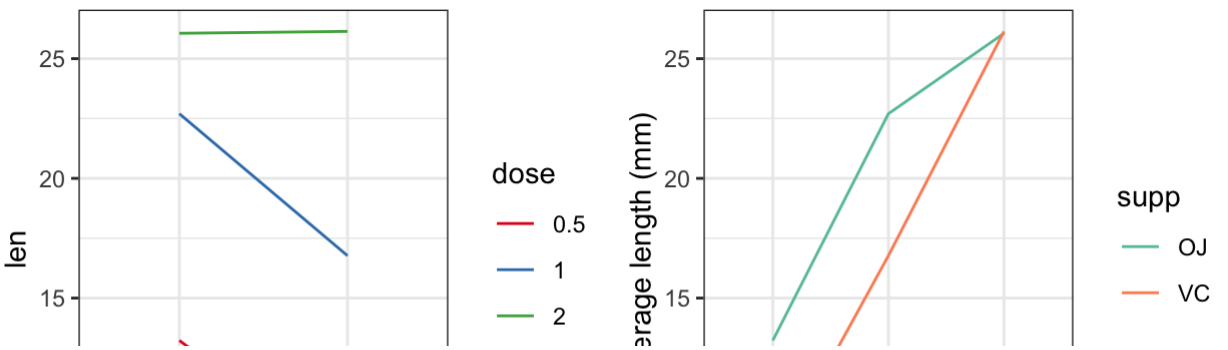
```
p1 = ggplot(tooth_sum, aes(x = supp, y = mean, colour = dose, group = dose)) +
    geom_point() + geom_line() + scale_color_brewer(palette = "Set1") +
    theme_bw()
```

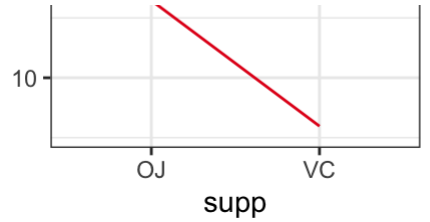
```
p2 = ggplot(tooth_sum, aes(x = dose, y = mean, colour = supp, group = supp)) +
    geom_point() + geom_line() + scale_color_brewer(palette = "Set2") +
    theme_bw()
gridExtra::grid.arrange(p1, p2, ncol = 2)
```



This could also be done with the original data (i.e. without having to pre-compute the group means) using `stat_summary()`:

```
p1 = tooth_df %>%
    ggplot() + aes(x = supp, y = len, color = dose, group = dose) + stat_summary(fun =
        mean,
    geom = "line") + scale_color_brewer(palette = "Set1") + theme_bw()
p2 = tooth_df %>%
    ggplot() + aes(x = dose, y = len, color = supp, group = supp) + stat_summary(fun =
        mean,
    geom = "line") + scale_color_brewer(palette = "Set2") + theme_bw() +
    labs(y = "Average length (mm)")
gridExtra::grid.arrange(p1, p2, ncol = 2)
```





The lines aren't totally parallel, a dose of 2mg in particular looks like it interacts differently with supplement than the 0.5mg and 1mg doses.

5. Fit the full model including interactions and obtain the corresponding analysis of variance table. Next, use the F-test to compare the full model with the additive model (no interaction model) and comment on the results.

```
tooth_aov = aov(len ~ supp * dose, tooth_df)
summary(tooth_aov)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
supp         1  205.4   205.4  15.572 0.000231 ***
dose         2 2426.4  1213.2  92.000  < 2e-16 ***
supp:dose    2  108.3    54.2   4.107 0.021860 *
Residuals   54  712.1    13.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we specify the full model as,

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ijk}$$

where μ is the overall mean, α_i and γ_j are treatment effects (differences between treatment group means and the overall mean), $(\alpha\gamma)_{ij}$ are the interaction effects and $\varepsilon_{ijk} \sim N(0, \sigma^2)$. We require the following constraints:

- $\sum_i \alpha_i = 0$
- $\sum_j \gamma_j = 0$
- For each j, $\sum_i (\alpha\gamma)_{ij} = 0$
- For each i, $\sum_j (\alpha\gamma)_{ij} = 0$

When testing for an interaction, we specify the null and alternative hypotheses as follows,

H_0: $(\alpha\gamma)_{ij} = 0$ for $i = 1, 2$ and $j = 1, 2, 3$ (i.e. there is no interaction effect between supp and dose)

H_1: at least one $(\alpha\gamma)_{ij} = 0$ (i.e. There is an interaction effect between supp and dose).

Looking at the ANOVA table, the F-statistic is 4.11 with the corresponding p-value is $P(F_{2,54} \geq 4.11) = 0.022$. At the 5% level of significance, the small p-value provides sufficient

evidence to reject the null hypothesis and conclude that the data favors the alternative hypothesis that is an interaction effect between the different dosage of Vitamin C (`dose`) and the delivery methods (`supp`).

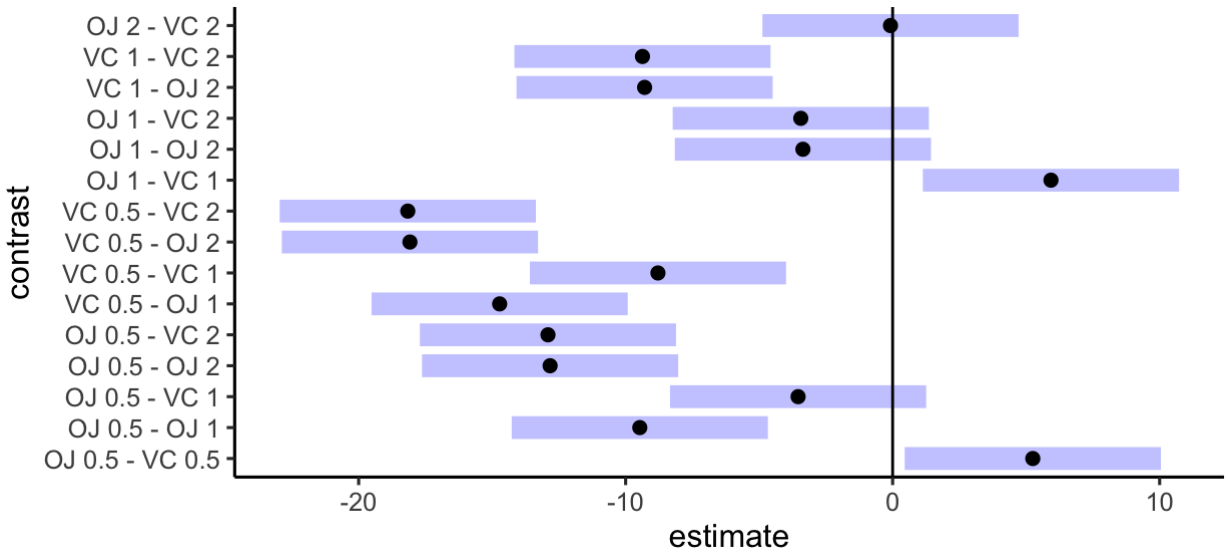
6. Perform post hoc tests to identify which treatment combinations are significantly different.

```
library(emmeans)
em_tooth = emmeans(tooth_aov, ~supp + dose)
em_pair = contrast(em_tooth, method = "pairwise")
em_pair
```

```
contrast        estimate   SE df t.ratio p.value
OJ 0.5 - VC 0.5     5.25 1.62 54   3.233  0.0243
OJ 0.5 - OJ 1      -9.47 1.62 54  -5.831  <.0001
OJ 0.5 - VC 1      -3.54 1.62 54  -2.180  0.2640
OJ 0.5 - OJ 2     -12.83 1.62 54  -7.900  <.0001
OJ 0.5 - VC 2     -12.91 1.62 54  -7.949  <.0001
VC 0.5 - OJ 1     -14.72 1.62 54  -9.064  <.0001
VC 0.5 - VC 1      -8.79 1.62 54  -5.413  <.0001
VC 0.5 - OJ 2     -18.08 1.62 54 -11.133  <.0001
VC 0.5 - VC 2     -18.16 1.62 54 -11.182  <.0001
OJ 1 - VC 1         5.93 1.62 54   3.651  0.0074
OJ 1 - OJ 2        -3.36 1.62 54  -2.069  0.3187
OJ 1 - VC 2        -3.44 1.62 54  -2.118  0.2936
VC 1 - OJ 2        -9.29 1.62 54  -5.720  <.0001
VC 1 - VC 2        -9.37 1.62 54  -5.770  <.0001
OJ 2 - VC 2        -0.08 1.62 54  -0.049  1.0000

P value adjustment: tukey method for comparing a family of 6 estimates
```

```
plot(em_pair) + geom_vline(xintercept = 0)
```

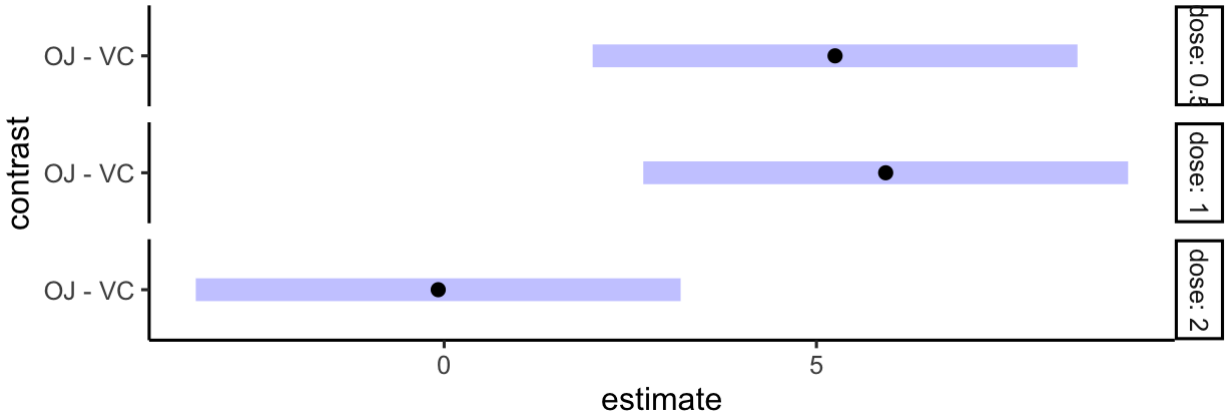


There is no significant difference between OJ and VC when the dose is 2mg. There's also no

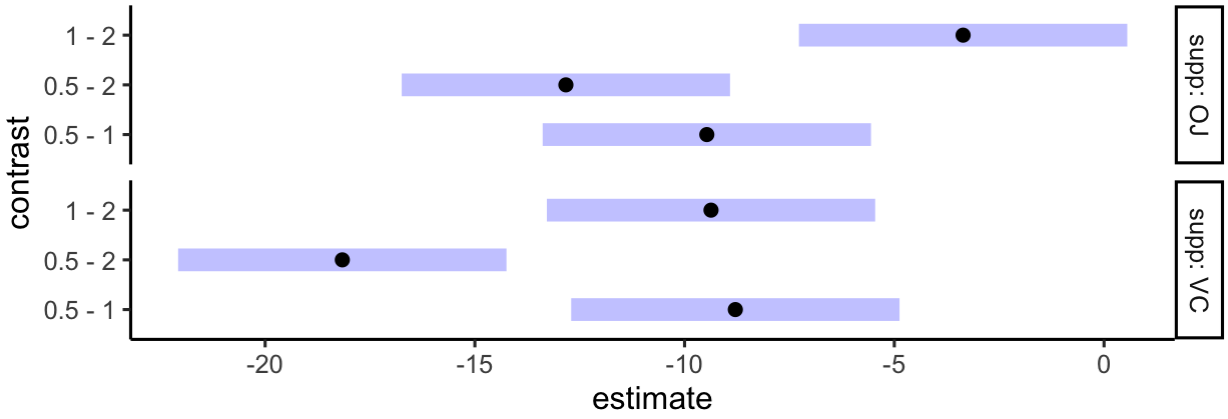
difference between 1mg OJ and 2mg VC or between 0.5mg OJ and 1mg VC. There's no significant difference between 1mg OJ and 2mm OJ. All other pairwise comparisons were significant.

We could be more targeted with our contrasts comparisons as follows:

```
emmeans(tooth_aov, ~supp | dose) %>%
    contrast(method = "pairwise") %>%
    plot()
```




```
emmeans(tooth_aov, ~dose | supp) %>%
    contrast(method = "pairwise") %>%
    plot()
```



References

Box, G. E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)* 26 (2): 211–52. http://www.jstor.org/stable/2984418.

Hubble, Edwin. 1929. "A Relation Between Distance and Radial Velocity Among Extra-Galactic Nebulae." *Proceedings of the National Academy of Sciences* 15 (3): 168–73. https://doi.org/10.1073/pnas.15.3.168.

National Academy of Sciences 15 (3): 168–73. https://doi.org/10.1073/pnas.15.3.168.

Lenth, Russell. 2018. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans.

McNeil, D. R. 1977. *Interactive Data Analysis*. New York: Wiley.