Lab 03C: Week 10

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.
2. How many replicates are there in each treatment combination?
3. Visualise the data using boxplots. Which poison tended to have the lowest survival time (highest reciprocal survival time)?
4. Write an appropriate model formula for a two-way ANOVA with interactions.
5. Use R to fit the ANOVA model described above and generate an ANOVA table.
6. Can the interaction effect be dropped from the model? Why or why not?
7. Test for a poison treatment effect.
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).
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?

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

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.
2. Summarise and visualise the data. What do you notice?
3. How many observations do we have in each treatment group?
4. Write an appropriate model formula for a two-way ANOVA with blocks.
5. Test if there is a machine effect.
6. Check and comment on the ANOVA assumptions.
7. Perform post hoc tests to see which pairs of machines have significantly different means.

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.

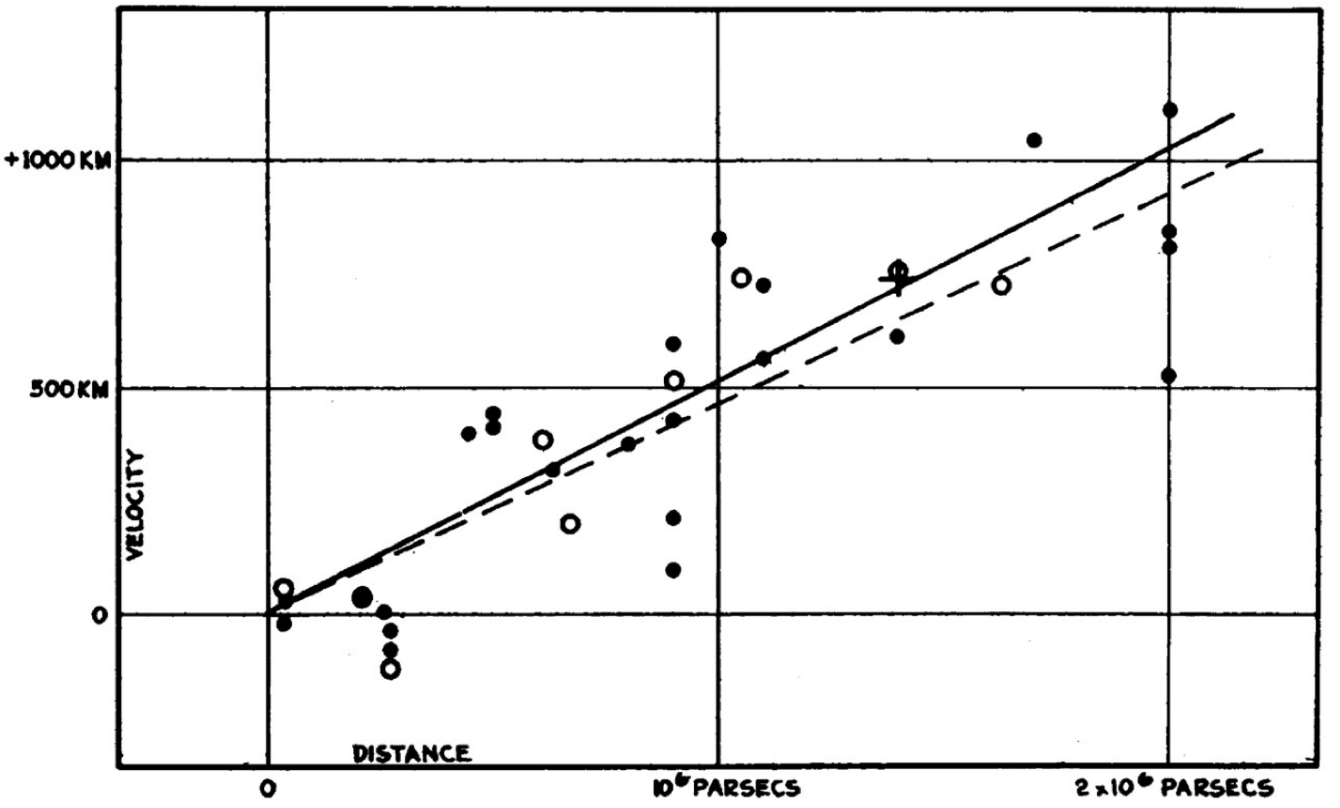
The data can be imported as follows:

```
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).



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.

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

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.

2 For practice after the computer lab

2.1 Tooth growth

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)
```

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).
2. Visualise the data using side-by-side boxplots.
3. Generate summary statistics for each of the treatment combinations. How many observations are there in each treatment combination?
4. Generate interaction plots. Does it look like there's an interaction between supplement and dose?
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.
6. Perform post hoc tests to identify which treatment combinations are significantly different.

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.

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.