Exam questions for: Advanced Data Analysis with R (700540)

Master Ecosystem Analysis and Modelling

Johannes Signer & Kai Husmann

Summer term 2025

Contents

1 Vector data

Use the following data set

```r
set.seed(123)

df1 <- data.frame(
  x = runif(100, 0, 100),
  y = runif(100, 0, 100),
  crown_diameter = runif(100, 1, 15),
  sp = sample(letters[1:4], 100, TRUE)
)
```

1. Use `df1` and create a geometry column.
2. Buffer each tree with its canopy radius.
3. Calculate the crown area of each tree and save it in a new column (hint, you my want to use the function `st_area()`).
4. Find the tree with the largest canopy area.
5. Find the tree with the largest canopy area for each species.

2 Wokring with raster data

The aim of this exercise is to get you start working with raster data.

1. Load the Digital Elevation Model (DEM) of Germany saved in `data/raster/dem_3035.tif`.
2. What is the spatial resolution and the CRS of the raster?
3. Cut the DEM to the state of Lower Saxony (use the data set on German states for this; `data/ger/ger_states_3035.shp`).
4. What is the mean elevation of Lower Saxony?
5. Find all pixels in Lower Saxony that have an elevation of 100 m or more. What is the percentage of Lower Saxony with an elevation of 100 m or more?

3 Calculating probabilities

Draw six numbers with replacement from the numbers $1, 2, 3$. Write code for the experiment above and repeat it 10000 times and then calculate the probabilities for

1. each number (1, 2, 3) being drawn exactly twice.
2. there is one number (e.g., 1, 2, 3) that is never drawn.
3. one number being drawn 4 times and one number being drawn twice.
4. one number being drawn three times, one twice and one only once.

4 Random Variables

Consider the very simple random trial of rolling four dices. Use R to answer the following questions:

1. What is the probability that the sum of the four dices is more than 10?
2. What is the probability that the sum of the four dices is less or equal to 5?

5 PMFs

Use R to verify that $f(x) = \frac{2x}{k(k+1)}$ can serve as a PMF for a random variable X, with $x = 1, 2, 3, \dots, k$.

6 Distributions

Many bird species migrate between winter and summer ranges. Assume we have a sample of 20 independent birds and we know that the probability of arriving at the winter range is 0.86.

- What is the probability that at *exactly* 10 birds arrive at their winter range?
- What is the probability that *at least* 10 birds arrive at their winter range?
- What is the probability that more than 5 and less than 16 birds arrive at their winter range?

7 p-value

A test statistics from a two-tailed t-test with 9 degrees of freedom is -1.8452. What is the corresponding p-value?

8 Estimating parameters of a distribution

A researcher measured the distance a bird dispersed (the distance between the nest where it hatched and the position where it established its nest). The measured distances in km are:

```
## c(1.3, 3, 0.5, 1.9, 3.7, 2.2, 0.8, 0.5, 3.1, 2.1, 2.1, 1.9)
```

- Which distribution do you think is suitable to model this data?
- Which parameters does this distribution have?
- Use `optim()` to estimate the parameters of this distribution.

9 Fitting a linear model

Use the `trees` data set (located in `data/trees.rds`) and select only spruce trees for the year 2020. First, investigate graphically, if you think that *bio18* (precipitation in the warmest month) could have a linear effect on `mean_damage`. Then fit a linear model for this relationship and give an interpretation of the estimated intercept, slope and variance parameter.

10 Visualize a fitted model

Use the `trees` data set (located in `data/trees.rds`) and select only spruce trees for the year 2020. Use *bio18* and *elevation* as covariates. Give a brief interpretation of both covariates. Then visualize model by choosing three typical values for elevation (colored lines) and show the values for bio18 on the x-axis and the predicted mean loss on the y-axis. Can you create three variants of the plot? Once without uncertainty, once withe a prediction interval and once with a confidence interval?

11 Linear model 2

Use the iris data set that is already included in R. You can load it with `data(iris)`. Fit the model:

$$\text{sepal.length} = \beta_0 + \beta_1 \times \text{sepal.width} + \beta_2 \times \text{species} + \beta_3 \times \text{sepal.width} \times \text{species}$$

1. Before looking at the model summary, think about how many coefficients need to be estimated.
2. Write the first 5 entries of the model design or model matrix $\mathbf{X}$. Verify with `model.matrix(m1)[1:5, ]` that you did this correct.
3. Give a brief interpretation for each coefficient.
4. Manually predict the sepal.length for *I. setosa* with a sepal width of 3.1 cm. Use the function `predict()` to verify your prediction.
5. Create a plot of the expected sepal lengths for a reasonable range of sepal widths for the different species.

12 Interactions

Use the `trees` data set (located in `data/trees.rds`) and select only the year 2020. Select the following species: spruce, beeches and oaks. Fit two models: 1) a linear model where you try to explain variation in the mean leaf loss as a function of *bio18* and *species*, and 2) second where you allow an interaction between *bio18* and *species*. Give a brief interpretation of the results. Visualize model both model, what do you notice?

13 Fitting a GLM

Linden and Rohloff (2015) studied white-headed woodpeckers in California (USA) using point counts. The file `woodp_Occ.csv` contains data from three visits at 66 sites. For each site and visit we know the if a woodpecker was present or not (columns `y.1`, `y.2`, `y.3`), the day of the year (columns `date.1`, `date.2`, `date.3`) and the snag density (the number of tree snags per ha; column `snags`). Tree snags are standing but dead or dying trees. Only use the first session of data collection for this exercise (i.e., `y.1`).

1. What would be a suitable model for this kind of data (i.e., think of the distribution of y)?
2. Fit the model that you decided on in 1).
3. Give an interpretation of the model results.
4. Create a suitable plot to communicate the results.
5. Validate your model.

14 Splines

The code, below, reads in data containing estimates of the number of Moose in Minnesota between 2005 and 2020:

```
mnmoose <- data.frame(
  year=2005:2020,
  estimate = c(8160, 8840, 6860, 7890, 7840, 5700,
               4900, 4230, 2760, 4350, 3450, 4020, 3710,
               3030, 4180, 3150))
```

Fit different linear regression (polynomials up to order 5 and natural splines with 3 and 5 knots) model to the data and evaluate whether the assumptions are reasonable.

Plot the data and the model.

15 Linear Mixed Models

The file `pines/Data1.txt` contains estimates of DBH and age for each tree at different ages. Each tree has an individual id (`core.code`). Perform the following tasks:

1. Plot a growth curve for each tree (in one plot). Where you plot the age on the x-axis and the DBH on the y-axis.
2. Fit the following models:
   a. A model $DBH = \beta_0 + \beta_1 age$ for each individual tree.
   b. A global model $DBH = \beta_0 + \beta_1 age$.
   c. A global model $DBH = \beta_0 + \beta_1 age + \beta_2 id + \beta_3 age \cdot id$.
   d. The same model as in b), but with a random intercept.
   e. The same model as in b), but with a random intercept and random slope.
3. Create a plot with the model type (2b, 2d, 2e) on the x-axis and the estimate (with a confidence interval) on the y-axis. Distinguish between the two terms (intercept and slope) by using different panels. Hint, you may find the function(s) `broom::tidy()` with the argument `conf.int = TRUE`, and `bind_rows` useful.

16 Time series 1

Run the following code to generate an annual time series from year 1 to 70.

```
set.seed(1)
c(arima.sim(n = 30, list(ar = c(.9, -.5, .3))) + .05 * c(1 : 30) + runif(30),
  arima.sim(n = 10, list(ar = c(.9))) + .05 * c(31 : 40) + runif(10),
  arima.sim(n = 30, list(ar = c(.9, -.5, .3))) + .05 * c(41 : 70) + runif(30)) |> ts()
```

1. Create an appealing plot.
2. Visualise all relevant information that are of interest for further analyses of the time series.
3. Create an autocorrelation function.

17 Time series 2

- Load the number of births per month in New York city, from January 1946 to December 1959 as `births <- scan("http://robjhyndman.com/tsdldata/data/nybirths.dat")`.
- Transform and save it as a time series object.

1. Plot the time Series
2. Estimate parameters for an intercept, a linear trend and seasonality.
3. Which of these 3 components are significant?

18 Time series 3

- Load the number of births per month in New York city, from January 1946 to December 1959 as `births <- scan("http://robjhyndman.com/tsdldata/data/nybirths.dat")`.
- Transform and save it as a time series object.

1. Describe the series using descriptive statistics.
2. Which model(s) do you use?
3. How do you interpret the results of that model(s)?

19 Time series 4

- Load the number of births per month in New York city, from January 1946 to December 1959 as `births <- scan("http://robjhyndman.com/tsdldata/data/nybirths.dat")`.
- Transform and save it as a time series object.

1. Create an appropriate ar model.
2. Do a one year forecast.