Exercises for Day 6

Applied Statistics & Statistical methods for the Biosciences

Anton Rask Lundborg

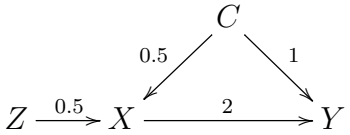
January 2024

Datasets and R scripts can be downloaded in a ZIP archive from the Absalon page (Applied Statistics) or from

`https://www.arlundborg.com/assets/SmS/data/day6.zip`

Exercise 6.1 Causal inference using an instrumental variable

The purpose of this exercise is to try the *two-step procedure* using an instrumental variable. In order to know the true value of the parameters we simulate a dataset from the following model:



Here the regression coefficients are written above the arrows. Furthermore, normal errors with unit standard deviation are added to all variables. Simulating $N = 100$ samples from this model may be done via the R code:

```
N <- 100
C <- rnorm(N)
Z <- rnorm(N)
X <- 0.5 * Z + 0.5 * C + rnorm(N)
Y <- 2 * X + C + rnorm(N)
```

We wish to recover the causal effect of X on Y (i.e. the coefficient 2) from the dataset, where only Z, X and Y are available. Thus, the *confounder* C is not available, but we have the *instrumental variable* Z at our disposal. Answer the following questions:

- Simulate the dataset in R using the above code.
- Create a simple linear regression of Y on X. Do you recover the coefficient 2 from this regression? Why not? Why is the estimate biased to return too large a value?
- Implement the *two-step procedure* described on lecture slide 17 (Day 6). Hint: To find the predicted values for X given the instrumental variable Z you may use the code

```
hatX <- predict(lm(X ~ Z))
```

- Do you recover the coefficient 2 from the *two-step procedure*? Is this still true if you simulate a new dataset several times?
- Try to see if you can recover the coefficient 2 when the sample size equals $N = 10000$. What does this say about the power of the *two-step procedure*?
- Now assume that the confounder C is available. Can you recover the true causal effect of X on Y from $N = 100$ observations?

Exercise 6.2 Latin square

The effect of insulin on the blood concentration of glucose was studied on rabbits. Three rabbits received insulin doses A, B and C (corresponding to respectively 0, 1 and 2 units) on different days. The experiment is given below (dataset available in file `rabbit.txt`) with the glucose measurements (mg pr. 100 ml blood) taken 50 minutes after injection.

	Rabbit					
Day	1		2		3	
1	A	50	C	39	B	36
2	C	37	B	51	A	53
3	B	51	A	60	C	37

Make the *Table of Variables*, set up the associated statistical model, and analyze the data. Remember to obtain estimates of effects of interest.

Why is it not possible to investigate whether there is an interaction between rabbit and dose based on these data?

(Data are from Young & Romans (1948): Assay of insulin with one blood sample per rabbit per day. *Biometrics*, **4**, 122–131.)

Exercise 6.3 Cover crops for apples

In East Malling the total harvest of apples (in pounds) in a four-year experimental period was investigated in a randomized block design with six treatments (cover crops A, ..., F) and four blocks. The design was implemented with 6 plots per block randomized over treatments. Beside the total harvest y in the experimental period, the total harvest x in a 4 years period prior to treatment was also recorded. The data are as follows (dataset available in file `apples.txt`):

	Block							
	1		2		3		4	
Cover crop	x	y	x	y	x	y	x	y
A	8.2	287	9.4	290	7.7	254	8.5	307
B	8.2	271	6.0	209	9.1	243	10.1	348
C	8.2	234	7.0	210	9.7	286	9.9	371
D	5.7	189	5.5	205	10.2	312	10.3	375
E	6.1	210	7.0	276	8.7	279	8.1	344
F	7.6	222	10.1	301	9.0	238	10.5	257

Analyze the data! Is there a significant effect of cover crop on the total harvest? Does the total harvest in the experimental period depend on the total harvest in the preceding period?

Assume that the previous harvest x was not measured. Is it then possible to find significant differences between the effect of cover crop?

Remark: The variables y and x appear to be recorded on different scales, i.e. the values of y approximately 30 times as big as the values of x. You can ignore this difference in scale, and simply use x as a covariate in the analysis of y.

Exercise 6.4 Constructing a latin square design

The objective of this exercise is to use the AlgDesign-package in R to generate the latin square design from Exercise 6.2. The full factorial design with $3^3 = 27$ observations may be generated by the following code:

```
library(AlgDesign)
full.factorial <- gen.factorial(levels=3, nVars=3,
                                varNames=c("Rabbit", "Day", "Dose"))
full.factorial
```

In this design the levels of the 3 variables are the numbers -1, 0, 1. To make it easier for us to look at the variables we may recode the levels, which may be done using the following code:

```
full.factorial <- with(full.factorial,data.frame(
  Rabbit=factor(c("Rabbit 1", "Rabbit 2", "Rabbit 3")[2+Rabbit]),
  Day=factor(c("Day 1", "Day 2", "Day 3")[2+Day]),
  Dose=factor(c("Dose A", "Dose B", "Dose C")[2+Dose])))
```

Have a look at the full factorial design with the recoded variable levels. Now use the `optFederov()` function to make a design with 9 observations where you have interest on the main effects of the 3 variables. Do you find the design from Exercise 6.2?

Remarks:

1. The `optFederov()` looks for a D-optimal design. And in this particular situation the *latin square design* is the D-optimal design, so you would expect that `optFederov()` will find the latin square. However, `optFederov()` does a random search, so you may be unlucky that is does not find the true optimum. To improve on this you may increase the `nRepeats`-option, e.g. using `nRepeats=1000`.

2. You may need to rename the levels of *Dose* and *Rabbit* to have an exact match with the table in Exercise 6.2. Using different names for the levels does, of course, not change the basic properties of the design.

Exercise 6.5 Non-linear regression

The following experiment was carried out in a greenhouse: 15 pots were sown with barley seeds: 3, 7, 15, 34, 77 barley seeds per pot, respectively, with three pots for each number of barley seeds. After harvest, the total fresh weight yields (in grams) were measured for each pot. The results are listed in the table below:

No. of seeds	Yield		
3	7.5	9.8	9.0
7	18.8	27.7	27.1
15	64.7	30.2	37.0
34	84.3	110.0	71.2
77	125.8	85.7	91.9

We want to use the following non-linear model for the relationship between number of barley seeds, $x =$ `seeds`, and the logarithmic yield, $y = \log($`yield`$)$:

$$y \approx a - b \cdot \mathrm{e}^{-cx} \tag{1}$$

Please answer the following questions:

1. Plot the logarithmic yield (variable y) against `seeds`.

2. What is the interpretation of the parameters a and b? Make a qualified guess on the values for a and b.

   Hint: What happens for $x = 0$ and x very large?

3. Although more difficult it is also possible to make a qualified guess on the c parameter, namely:

   $$c \approx \frac{\log(2)}{15} \approx 0.045$$

   You are welcome to explain the reasoning behind this guess (if you can), but otherwise you may simply take it for granted.

4. Assuming that that data frame with the observations is called `barley`, the following R code makes an interactive plot in RStudio (remember to install the `manipulate`-package if you have not done it before):

```
library(manipulate)
manipulate(
  {plot(log(yield) ~ seeds, data=barley)
   x <- 0:80
   y <- a - b * exp(-c * x)
   lines(x, y)},
  a=slider(2, 8, initial=5, step=0.1),
  b=slider(0, 4, initial=2, step=0.1),
  c=slider(0, 0.1, initial=0.045, step=0.005)
)
```

Try this and click on the *gear sprocket* (in danish: *tandhjul*) icon in the graphical window to manipulate the a, b and c parameters interactively. See if you can change the parameters such that the data points are fitted closely by the non-linear curve.

Remark: In this way the `manipulate()` function may be used to derive initial guesses for the parameters in a non-linear regression. However, if you already have an adequate guess, then this is not necessary.

5. Use `nls()` with the initial guesses given in the `start`-option to fit the parameters a, b and c by a non-linear regression.

6. Give estimates and confidence intervals for the parameters a, b and c.

7. Is the non-linear model valid?

   Hint: You may create a *residual plot* and *normal quantile plot* by "hand" using the code on slide 9 (Day 6).

8. For some of the "classical" and often used non-linear functions there exist so-called *self-starting* functions in R. The non-linear function used in this exercise is one of these functions. The associated R function is called `SSasymp()`. Try the following R code and relate it to the results you found above:

   ```
   m2 <- nls(log(yield) ~ SSasymp(seeds, a, a.minus.b, log.c), data=barley)
   cbind(estimate=coef(m2), confint(m2))
   ```

   Remark: `SSasymp()` uses a different parametrization than the one used in Eq. (1). Can you describe how you pass between these parametrizations?

(Reference: Based on exercise 8.6 from Anders Tolver & Helle Sørensen: *Lecture notes for Applied Statistics*.)