Tutorial 11:

Introduction to Data Science

Causality

Ernst C. Wit

December 2, 2024

Exercises

1. **ANOVA: causal regression**. Sir Ronald Aylmer Fisher (1890 – 1962) was a British polymath who was active as a statistician, biologist and geneticist. He is the father of modern statistics and in genetics, Fisher was the one to most comprehensively combine the ideas of Gregor Mendel and Charles Darwin. From 1919, he worked at the Rothamsted Experimental Station for 14 years. This was an agricultural station, where the aim was to improve farming practices. There, he analyzed its immense body of data from crop experiments since the 1840s, and developed the analysis of variance (ANOVA).

   At the end of his time at Rothamsted, Fisher encountered a study studying the effect of sulphate of ammonia, poultry manure, soot and rape dust on the yield of Brussels sprouts. Each of the four fertilizers were applied to random plots in low and high concentrations and compared to the yield without any treatment. In total, 48 plots of land were available for the randomized experiment, whereby 1 of the 9 treatments got randomly assigned to any of the plots. The data are available in the R-package `agridat` as the object `Rothamsted.brussels`. To load the data, type the following:

   ```
   > install.packages("agridat")
   > library(agridat)
   > ? rothamsted.brussels
   ```

   (a) **Causality.** Why could the researchers not simply ask farmers to report their yields while applying the various fertilizers? What kind of confounding may occur? Sketch a causal diagram involving *treatment*, *confounders* and *yield*.

   (b) **Randomization.** How does randomization of the treatments alter the causal diagram?

   (c) **Research Question.** What is the causal question of interest?

   (d) **Sample and population.** What are the 48 statistical units in this trial? What is the population?

   (e) **Exploratory analysis.** Plot the yield data in a way to show the effect of the various fertilizers.

   (f) **Formal analysis (I).** Perform an analysis of variance to see if the fertilizer treatment has an effect. If it does, then use regression to detect which method works best. Order them from top to bottom.

(g) **Plots of land.** The plots of land where the brussels sprouts have been planted constitute a large field with 8 rows and 6 columns. Plot the yield visually in each plot:

```
libs(desplot)
desplot(rothamsted.brussels, yield~col*row,
        num=trt, out1=block, cex=1,
        main="rothamsted.brussels")
```

Visually what do you suspect about the fertility of the land where the experiment has been performed?

(h) **Formal analysis (II).** To account for the varying fertility of the various plots, the experimenters have identified 4 more or less equally fertile plots, called `blocks`. Include the `blocks` in the formal analysis and explain which conclusions change and which stay the same.

(i) **Conclusions.** What are your conclusions with respect to the original research question. Are there any aspects in the study that needs further attention?