A. C. Davison **Regression Methods Project 2024** MATH EPFL

Problem

The paper that won the 2024 IgNobel Prize in Probability summarises an experiment resulting in 350,757 coin flips by 48 people using 211 different coins. It suggests that a coin showing heads when it is flipped into the air will land heads up with probability around 0.51, significantly higher than 0.5, and likewise for tails. A summary of the data in `data-agg.csv` gives for each participant and coin the number of heads when the coin started heads up (1) and tails up (0); we can label these events $1 \mapsto 1$ and $0 \mapsto 1$, and likewise define $0 \mapsto 1$ and $0 \mapsto 0$. The original paper takes a Bayesian approach with a variety of prior densities on the probabilities $p_{11} = \mathrm{P}(1 \mapsto 1)$ and $p_{00} = \mathrm{P}(0 \mapsto 0)$, but the goal of this project is to use regression methods to analyse the data. The simplest possible model would treat all the flips as independent Bernoulli trials, but clearly the outcomes might be influenced by the different groups of participants, the individual participant, the coin, by time (maybe there are learning effects?) and perhaps even the outcomes of recent flips. Some coins appear several times (e.g., 1 euro coins for `frantisekB` and for `pierreG`), but these are presumably not the same, so coins are nested within participants.

To read the data into `R` and make a data frame for a simple analysis, use the code

```
df <- read.csv("data-agg.csv",header=T)
n <- nrow(df)
df1 <- df[,-c(2,4)] # heads to heads
df2 <- df[,-c(1,3)] # tails to heads
df2[,1] <- df2[,2]-df2[,1] # tails to tails
names(df1) <- names(df2) <- c("y","m","person","coin")
start <- rep(c("heads","tails"),c(n,n))
df <- rbind(df1,df2)
df$person <- factor(df$person); df$coin <- factor(df$coin); df$start <- factor(start)
summary(glm(cbind(y,m-y)~1,family=binomial,data=df))
```

but you will need to manipulate the data (maybe also data from other files) for more complex analysis.

The data and code for the analyses in the original paper can be found `here`.

Report

Submission and deadlines

A short report of at most **12 pages** (excluding cover page, table of contents and bibliography) is to be **submitted on Moodle** as a PDF file (nothing else will be accepted) by midnight on **11 January 2024**. Later submission will not be accepted. Reports must be written **in pairs**; you can discuss with your classmates, but the code and writing should be your own. You should upload your code in a separate file and ensure the output is reproducible (not proper to your laptop or computer).

Your file should be named `NameA-NameB-RMProject-2023.pdf` (in alphabetical order of surname, e.g., `Hardy-Laurel-RMProject-2024.pdf`).

This will count as **40%** of the final grade for the course.

Structure of the report

The report should be typed in English. Some notes on report-writing can be found here and an example report is posted on Moodle.

Introduction: Briefly state the purpose of the analysis, discuss the main features of the data (e.g., via exploratory data analysis), and outline what will follow.

Analysis: describe the model(s) fitted, using your own words. Give the key elements only: you can refer to the lecture notes and to books, but should give careful references (to pages and equations etc.). It is not enough simply to give a list of sources at the end of the work: references should be mentioned in the text, and only those mentioned in the text should be listed at the end. Use BibTeX or similar to ensure that the references appear properly; check a book or journal article to see what details should appear in the bibliography.

Discussion of the results in more detail. Include crucial graphs and tables only, make sure that their contents are understandable without reference to the text, and that their axis labels and captions are clear and informative; each graph and/or table should tell the reader a coherent story. Give appropriate numbers of digits for tables. The text should give detailed interpretations of the plots and tables, with more details, if they are needed, and should show where the graph/table fits into the overall picture.

Conclusions: the take-away message from your analysis. Convince the reader that you know what you did and are aware of its strengths and limitations. Sketch what more you might do, if you had more time.

Discussion points

Exploratory data analysis: are there obvious outlying persons, coins, or person/coin combinations?

Modelling: the raw data are binary, so it would be natural to fit a suitable generalized linear model. On the other hand the success probability is fairly close to 0.5 and $p(1-p) \approx 1/4$ when $p \approx 1/2$, so the approximation $R/m \dot{\sim} \mathcal{N}\{p, 1/(4m)\}$ for binomial variables R with denominator m should be reasonable and might be fitted using weighted least squares. Do such analyses agree? Is there evidence of differences between the persons and coins (within persons)? Are there components of variance between coins within persons? Are there obvious bad residuals from your fits? Are certain person/coin combinations unusual? Are the conclusions from the original paper well-supported, or are matters more nuanced?

Discussion: discuss the results, showing (e.g.,) analysis of variance/deviance and/or AIC for model comparisons, give estimates and their standard errors, explain any disagreement between models. Ensure that you carefully interpret your preferred models (if any) in terms of the original problem.

Suggestions and caveats

- We recommend (but do not insist) that you use LaTeX and `R`.
- Your report and code should allow a reader to reproduce your results, with code adequately commented.
- Figures and tables should be numbered and have captions briefly explaining their contents. Reference should be made to each figure/table from within the text.

- Read your report carefully before handing it in and use a spell checker to find any typos.
- Mention any references you have used and provide a detailed bibliography. References should be made to scientific articles or books. Detailed (chapter, section, page, equation) references to books are usually needed, so that the reader does not have to figure out which page(s) of a book you are referring to.
- Pasting plain computer output is unacceptable.
- Due to space limitations, you should provide only relevant output. Your code can however contain exploratory data analysis and other model fits and diagnostics that are not reported in the text.
- When writing a report, do not answer questions directly. Instead, make sure your report covers the material discussed in each point, but structure it as a scientific paper.
- Common problems to avoid: many students don't give enough (or sometimes any!) interpretation of fitted models; often tables have too many digits; often figures are too small to be read properly or don't use the page layout well; often captions to tables or figures are uninformative; often the discussion section is insufficiently detailed; often the bibliography has missing details; often references to publications are inadequate; often equations are not (or are incorrectly) punctuated; often the English has persistent spelling errors.

Marking scheme

Correctness	Accurate, appropriate use of statistical tools 10 9 8 7 6 5	Incorrect, many errors 4 3 2 1 0	Score ______
Discussion	Thoughtful, detailed, apposite 10 9 8 7 6 5	Banal, obvious, thin 4 3 2 1 0	Score ______
Graphics and tables	Clearly labelled, well-chosen, good captions and discussion, appropriate numbers of digits 10 9 8 7 6 5	Poorly labelled, no discussion, unmotivated, unedited output 4 3 2 1 0	Score ______
Originality and scope	Wide range of tools/ideas 5 4 3	Limited range of tools/ideas 2 1 0	Score ______
Quality of writing	Good grammar and punctuation, including mathematics 5 4 3	Poor grammar and punctuation 2 1 0	Score ______
Referencing	Full, accurate, and detailed references given 5 4 3	Inadequate citation of sources 2 1 0	Score ______
		Grand total (max 45)	______