École polytechnique fédérale de Lausanne

Regression Methods Project Report

Coin-Data Regression Study

Authors:
Tara Fjellman
Rayan Harfouche

Professor:
Anthony Davison

Contents

Abstract

context of the original paper (with their claims). The paper that won the 2024 IgNobel Prize in Probability took a Bayesian approach to studying the statistics behind the coin-flipping process. Its main goal was to confirm a prediction made by a physical model of human coin tossing developed by Diaconis, Holmes, and Montgomery (DHM; 2007); i.e. that when people flip an ordinary coin, the probability of it landing on the same side it started is about 51%. It also revealed considerable between-people variation in the degree of this same-side bias, as well as its decrease as more coins were flipped.

The goal of this report is two fold. On the one side, we aim to investigate similar questions with a regression approach : is there evidence for between-person, between-coin, or even person-coin pair differences ? To what extent does flipping experience affect the observed same-side bias ? In addition, we also investigate the differences between GLM and WLS approaches; as well as muscle memory effects (through outcomes of recent flips).

1 Introduction

Before diving into the analysis, we provide a brief overview of the datasets main features and the models we consider. Our exploratory analysis is available as a Jupyter notebook, and provides more detail.

The dataset is composed of throws from 48 people using 44 coins in total. Given the study did not impose strict guidelines for coins to be used, the design is heavily unbalanced. Eighteen coins have only been thrown by a single person, while some of them have been thrown by more than 20 people. Also, more than half the people have only flipped 5 or less different coins, while someone threw 11 different ones. As for the the person-coin pairs, most have fewer or equal to 1000 throws, while some have around 10000 ones. This severe unbalance must be kept in mind during the study, given it can pose challenges during model interpretation.

After plotting the same-side rates across people, coin and person-coin combinations, we deemed it relevant to investigate models considering both person and coin as covariates, as well as individual person-coin pairs. Following the advice given on the project statement, we also branched our analysis in Binomial-response GLM and WLS approaches.

Motivated by impacts of muscle-memory on the flipping, we additionally investigated aspects such as time-varying same-side rates and memory between successive throws.

2 Analysis

2.1 Model Comparison

In this section, we introduce and compare different models for the same-side rate.
Should explain no a priori response transformation ...
For each, the considered formulas in terms of the covariates are:

- `1`, corresponding to a constant model.
- `1+C(person)`, corresponding to a model with the person as a covariate.
- `1+C(person)+C(coin)`, corresponding to a model with the person and the coin as covariates.
- `1+C(person)+C(coin)+C(person):C(coin)`, corresponding to a model with the person, the coin, and the interaction between the person and the coin as covariates.

* model 4 could seem redundant due to nesting-main effect, but we .... *
Should explain why eliminated some covariates ...
In the next section, we start by introducing the theory on which we base our model selection.

2.1.1 Tools For Selection

Following the suggestion in the project statement, we compare candidate models both with help of Akaike's Information Criterion (AIC) and Likelihood Ratio Tests (LRTs).

We recall that AIC is defined as

$$\text{AIC} := 2p - 2\ln(\hat{L}), \tag{1}$$

with p the number of degreas of freedom of the model and $\hat{L}$ the value of the maximised likelihood function for the model. By minimising AIC over candidate models, one is therefore rewarding model fit, while penalising by a $2p$ term to counter overfitting. Its goal is therefore to yield parsimonious predictive models. It does not make any assumption about models being of similar interpretation or nested.

In contrast, LRT is specifically designed for comparing a model A with a nested (or restricted) one B. It is based on the fact that

$$\lambda_{LR} = -2\ln\left[\frac{\sup_{\theta\in\Theta_B}\mathcal{L}(\theta)}{\sup_{\theta\in\Theta_A}\mathcal{L}(\theta)}\right] \overset{H_0}{\sim} \chi^2_{p_A-p_B}, \tag{2}$$

where $\Theta_A \supset \Theta_B$ respectively are the parameters spaces (of dimension $p_a > p_b$) associated to models A and B, and H_0 is the null hypothesis : $\theta^* \in \Theta_B$. Given this definition, LRTs are used to shed light on whether extra degrees of freedom improve the model significantly better than chance. This has applications in explanatory models, but might be of limited use in the context of model selection due to being prone to overfitting. This is particularly true when testing many models, as tests become correlated and give rise to spurious results [cite davison slides ?].

2.1.2 Diagnostic Plots

In this report, the diagnostics we decided to include where

- A QQ plot to check normality of normalised residuals. Pearson studentized residuals were used for WLS models while the presented r^* were used for GLM models. Normality corresponds to a roughly straight line, while outliers, skewness and heavy tails are easily spotted as deviations from it.

- A scatter plot of residuals as a function of fitted values, to assess linearity and homoscedasticity. [TO FILL] were used for WLS models, while deviance residuals where used for GLM models. Linearity corresponds to residuals that stay centred around 0 across the fitted values, while homoscedasticity corresponds to errors whose spread is constant across the fitted values.

- Cook's distance as a function of data index to reveal highly influential data points. Points over the $8/(n-2p)$ threshold are given a closer look (in order of importance).

- Scatter plots (resp. box plots for categorical covariates) of residuals as a function of the covariates of interest. These are used to check for independence between residuals and the covariates. If there is independence, the residuals look uncorrelated to the covariates.

This decision was motivated by our understanding of section 1.4 about diagnostics found in [CITE DAVISON SLIDES].

2.1.3 WLS Approach

In this section, we consider the normal approximation for the binomial variable R with denominator m. Having that the success probability is fairly close to 0.5 we get $p(1-p)$ approximately equal to $1/4$. Hence we end up with: $R/m \sim N(p, 1/(4m))$.

We then make a linear fit using weighted least squares, where weight associated to each entry is proportional to the inverse of the variance, that only depending on m.

We consider different models, starting with a constant model, then adding in the following order the

Tab 1: *Model comparison based on AIC values.*

Model Formula (RHS)	AIC
1	159.76
1 + C(person)	12.89
1 + C(person) + agg	0.00
1 + C(person) + agg + C(coin)	29.21
1 + C(person) + agg + C(person):C(coin)	123.98

2.1.4 GLM Approach

In this section, instead of relying on an approximation to fit a Least Squares models, we take into account the nature of the data by using a GLM. Indeed, given the binomial nature of aggregated data, using a binomial-response GLM seems natural. More specifically, we consider a Logit

Tab 2: *Model comparison for different models.*

Model	Deviance	AIC	Model DF
`1`	3942.13	187.84	0
`1+person`	3676.20	13.91	46
`1+person+agg`	3660.29	0.00	47
`1+person+agg+coin`	3602.26	25.97	89

link as it leads to easily interpretable results, meaning we model $\mathbb{E}[y \mid x] = \exp\left(\sum_i \beta_i x_i\right)$ with binomial errors. We expect this to lead to more accurate results than the WLS approximation, especially for entries having $R/M \not\approx 1/2$. As for the considered covariates, we follow what is done for the WLS approach. The only difference being in the interpretation of the coefficients and the in the fact that we omit the model with coins nested within people. Indeed, even the `0+person:coin` model did not give any signs of convergence after 100 IRLS iterations. We tried fitting it with BFGS, but did not succeed either as the hessian resulted to be non-full rank. The analysis of deviance and LRT tables for the models that converged are Tab.2 and Tab.3. The diagnostic plots for the `1+person+agg+coin` model are in Fig.1 and Fig.2.

Tab 3: *Likelihood ratio tests between models.*

Tested model	Restricted model	p-value
`1+person`	`1`	0.00e+00
`1+person+agg`	`1+person`	6.63e-05
`1+person+agg+coin`	`1+person+agg`	5.09e-02

2.2 Unusual Observations

2.3 Zoom on Learning Effects

- bias comes from start
- amount of bias (considerable)
- wobble interpretation (consistent with physical model, citing the paper)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2.4 Memory Effects

In this section we shift our focus to memory effects. We do so motivated by the fact that we deem it probable a priori that successive throws are more similar to each other than randomly selected ones. Indeed, one could imagine that after two same-side throws, the next could end being a same-side one two with a probability higher than the base rate.

To test this we start by considering the data consisting of individual throw outcomes. To this, we add columns corresponding to

- same-side indicator variables,
- same-side indicator variables for the penultimate throw,
- same-side indicator variables for the antepenultimate throw.

To deal with the boundary effects between sequences of flips, we removed the two first entries of each sequence.

We then define the models

- `1`, same-side indicator variables,
- `1+hop1_mem`, same-side indicator variables for the penultimate throw,
- `1+hop1_mem+hop2_mem`, same-side indicator variables for the antepenultimate throw.

and carry out an analysis of deviance. The results associated are found in Tab.4. Given the uni-directionality of these results, no further analysis was made.

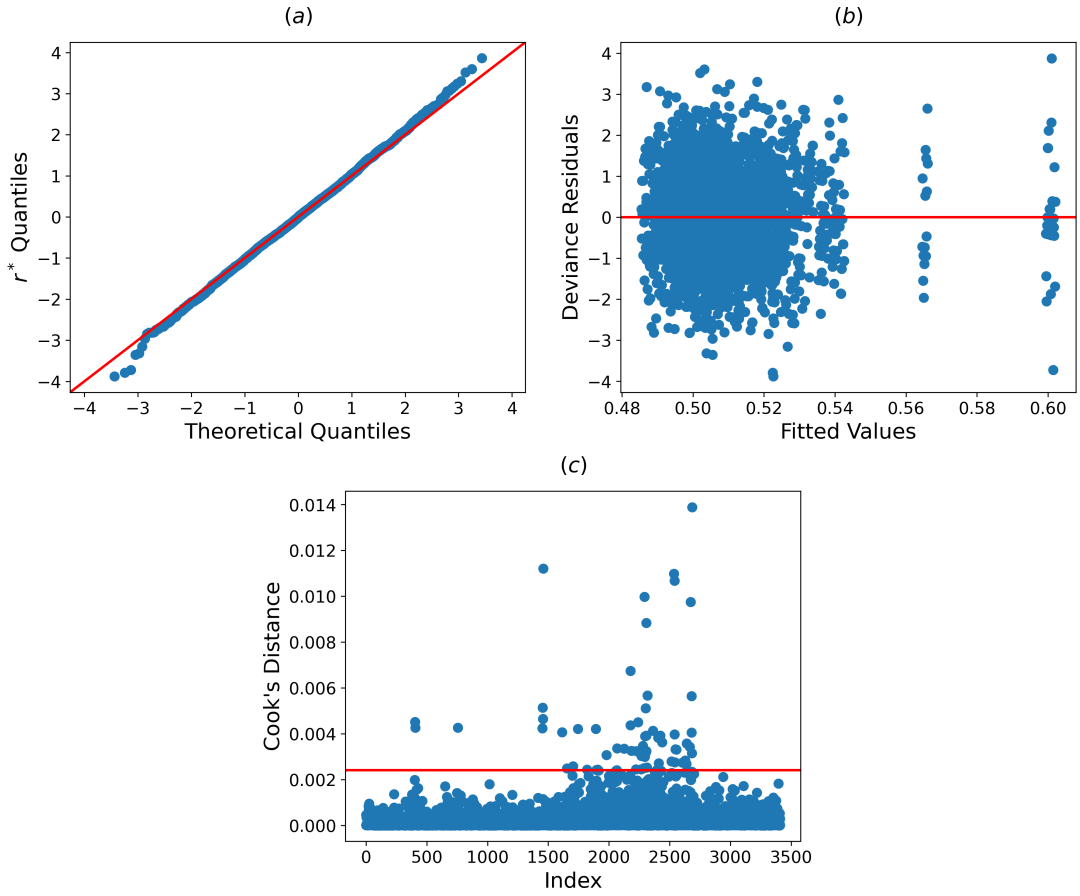

Fig 1: *Diagnostics for the selected GLM model. (a).*

3 Discussion

3.1 Model Comparison

3.1.1 WLS Approach

3.1.2 GLM Approach

We first consider the model comparison tables Tab.2 and Tab.3. Upon inspection of these, we see that the evidence for between-person variations is extremely strong. The deviance decreases by more than 5.5 points per degree of freedom on average, and the AIC drops by more than 170. Similarly, there is overwhelming evidence for time-dependence in the same-side bias. By itself, the term for example yields a deviance decrease of 15. A closer look at this contribution is given in the Sec.3.3. When it comes to between-coin differences, the evidence is nowhere near

Tab 4: *Model comparison for models including : no memory, 1-hop memory and 2-hop memory.*

Model	Deviance	AIC	Model DF
`1`	474381.54	0.00	0
`1+hop1_mem`	474380.73	1.19	1
`1+hop1_mem+hop2_mem`	474380.53	2.98	2

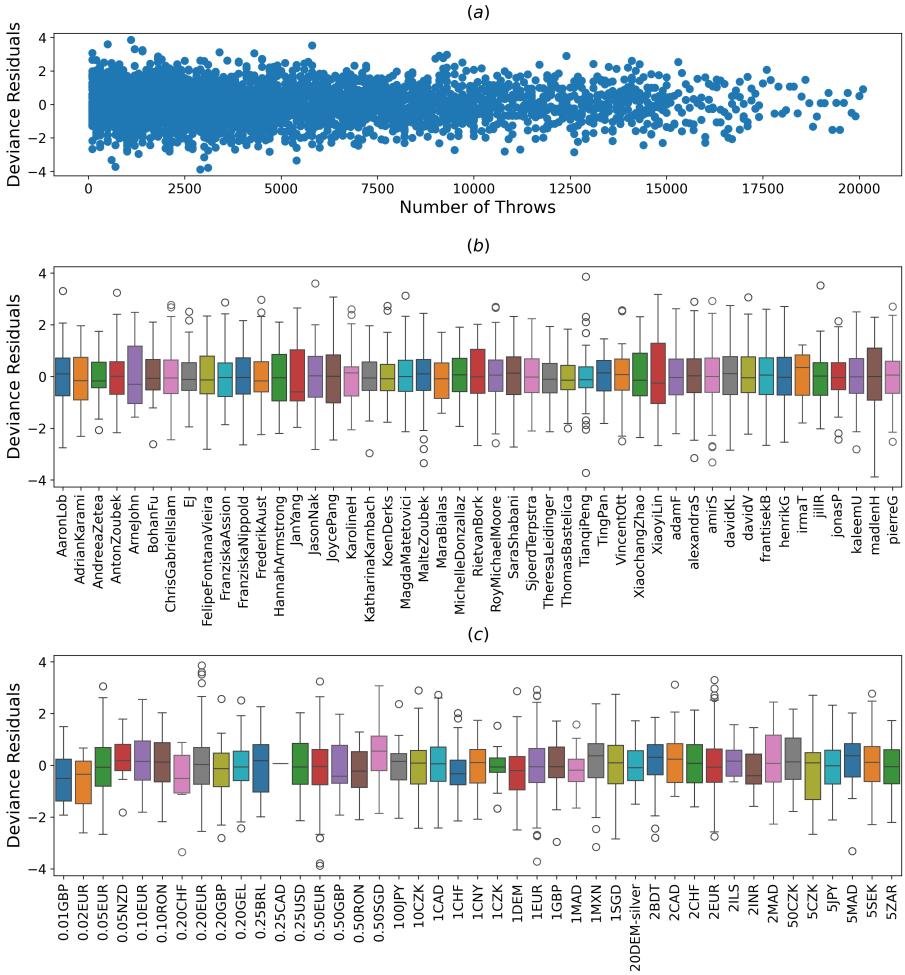

Fig 2: *Dev-resid as a function of (a) person and (b) coin.*

as strong. Indeed, AIC increases by around 25 when compared to the `1+person+agg` model, reflecting clear overfitting. The associated LRT is around .05, meaning coin-effects might still help in explaining part of the deviance. A better evaluation of this could probably be obtained by considering a smaller scale, balanced-design study. Indeed, in the considered dataset, coin and time effects where aliased[1]. This was a result of there being a strong time effect on same-side bias, and some coins being flipped many more times than others. Given these elements, one would definitely prefer the `1+person+agg` model for prediction purposes.

We now move to analysing the diagnostics of the selected model. Looking at (a) and (b) of Fig.1, we see that the residuals are for the most part appropriately normal and homoschedastic. Slight anomalies are observed in the QQ plot for the largest and smallest residuals, as well as for the residuals associated to the largest fitted same-side rate. Looking up the entries associated to the largest Cooks distances, we notice these are associated to sequences of throws where [TO FILL]

- best model (AIC vs LRT) + interpretation of significance (wobble)
- discussion of diagnostics

[1]Adding `coin` to the `1+person` model explained around 10 units of deviance more than when it was added in the `1+person+agg` model.

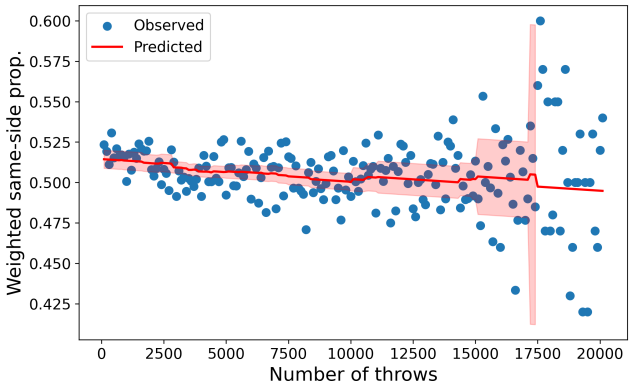

Fig 3: *Learning effects.*

- (rational of including agg, presentation of used residuals with assumptions)
-
- comparison with WLS (coefs ?? accuracy ?)

3.2 Unusual Observations

3.3 Zoom on Learning Effects

3.4 Memory Effects

Looking at Tab.4, we see there is no support in favour of memory effects relative to the constant model. The deviance decreases by around .8 due to the introduction of the penultimate throw memory and only a further .2 when including memory of the antepenultimate throw. Another factor that shows this is the increase by >1 and ≈ 3 respectively in AIC compared to the constant model. The fact that memory about antepenultimate outcome seems to matter even less than memory about the penultimate outcome does make nonetheless intuitive sense.

This (non-) finding can in itself be regarded as reassuring in a way. Specifically, it might contribute to rule out concerns of the authors of the original paper regarding the potential same-side bias induced by participants knowing about the goal of the study. Indeed it seems far-fetched that someone could bias their throws without relying on muscle memory.

4 Conclusion

Acknowledgements

References