Documentation

Alvin Gavel

Contents

1 What are we trying to find out? 2

2 What does the data look like? 3

3 Mathematical model 5

3.1 Basic idea 5
3.2 Linear model of prediction term 5
3.3 Normal model of error term 6

1 What are we trying to find out?

We have given a group of participants five different tests. We have split them into a test group that has been subjected to sleep deprivation and a control group that has not[1]. After each test we asked them to estimate their own performance. We now want to find out how sleep deprivation affects their estimated performance.

[1] At least not systematically, as part of the experiment. Sometimes people just sleep badly.

2 What does the data look like?

There are six csv-files in the directory `Data`[2]:

- `kss_data.csv`
- `arithmetic_data.csv`: Data from arithmetic test
- `working_memory_data.csv`: Data from working memory test
- `episodic_memory_data.csv`: Data from Episodic memory test
- `Stroop_data.csv`: Data from Stroop test
- `simple_attention_data.csv`: Data from simple reaction time test

At the moment, we only use the file `kss_data.csv`. At time of writing I do not know what all columns are, so our best interpretation is this:

- `id`: Participant ID
- `age`: Participant age
- `woman`: Participant gender
- `sd`: Participant sleep status
- `pair`: Participant pair for social interaction tasks
- `dyad_type`: Combination of sleep status in dyad
- `X1`: ¯_(ツ)_/¯
- `date`: Date of test
- `clock`: Time of test
- `order_t`: Test order for each session
- `test_type`: Which cognitive test the participant just finished. This uses the abbreviations:
  - *M*: Arithmetic
  - *W*: Working memory
  - *ST*: Episodic memory
  - *stroop*: Stroop
  - *reactionTime*: Simple reaction time
- `rating1_type`: The type of rating in the column `rating1`
- `rating1`: Self-rated sleepiness after each test. That is, the participant's response to the question "How sleepy are you right now?"
- `rating2_type`: The type of rating in the column `rating2`.

[2] It is currently not in the repo. This may change in the future.

- `rating2`: Self-rated effort on each test. That is, the participant's response to the question "How much of an effort did you make?"

- `rating3_type`: The type of rating in column `rating1`

- `rating3`: Self-rated performance on each test. That is, the participant's response to the question "How did you perform?"

- `time_adjustment_hours`: ¯_(ツ)_/¯

- `old_time`: ¯_(ツ)_/¯

- `correct`: This is a dummy column that can be ignored.

- `time`: Order of the session in the test battery

For now, the columns of interest to us are `id`, `sd` and `rating3`. We want to know how self-rated effort depends on sleep status.

3 Mathematical model

In the article *Too Tired to Know* we implement the cumulative model described in Bürkner and Vuorre (2019). The reader is primarily recommended to that article for the details, or – once it is written – the section in our article that explains the analysis. This is mostly written for myself to force myself to understand how the model works, since the root of at least half of all evil is the use of statistical recipes without first understanding them.

3.1 Basic idea

The cumulative model starts out from the assumption that the ordinal variable that we actually measure, Y, is an increasing function of some underlying unobservable variable $\tilde{Y}$. Specifically it assumes that there are K thresholds τ_k such that if $\tau_{k-1} < \tilde{Y} < \tau_k$ then we will observe $Y = k$. The underlying $\tilde{Y}$ in turn follows some probability density function f such that the probability $P\left(Y = k\right)$ is given by

$$P\left(Y = k\right) = \int_{\tau_{k-1}}^{\tau_k} f\left(\tilde{Y}\right) d\tilde{Y} \tag{1}$$

If we call the corresponding cumulative density function F, this is

$$= F\left(\tau_k\right) - F\left(\tau_{k-1}\right) \tag{2}$$

3.2 Linear model of prediction term

So far the model is very general, but to get anywhere we need to make some assumptions about what the function f actually looks like. We split $\tilde{Y}$ into a predictor term η and an error term ε, so that

$$\tilde{Y} = \eta + \varepsilon \tag{3}$$

We now assume that η is a linear function, so that

$$\eta = \sum_{n=0}^{K} b_n x_n \tag{4}$$

Inserting (4) into (3) gives us

$$\tilde{Y} = \sum_{n=0}^{K} b_n x_n + \varepsilon \tag{5}$$

Since $d\tilde{Y}/d\varepsilon = 1$ we can rewrite (1) as

$$P\left(Y = k\right) = \int_{\tau_{k-1}}^{\tau_k} f\left(\sum_{n=0}^{K} b_n x_n + \varepsilon\right) d\varepsilon \tag{6}$$

Shifting the integration boundaries by the predictor term gives us

$$= F\left(\tau_k - \sum_{n=0}^{K} b_n x_n\right) - F\left(\tau_{k-1} - \sum_{n=0}^{K} b_n x_n\right) \tag{7}$$

3.3 Normal model of error term

We now need some kind of model of what that error term looks like. To make life simpler[3], we assume that it is normally distributed. If we denote the CDF of the normal distribution with Φ, then (7) takes the form:

$$P\left(Y=k\right)=\Phi\left(\tau_k-\sum_{n=0}^{K}b_n x_n\right)-\Phi\left(\tau_{k-1}-\sum_{n=0}^{K}b_n x_n\right) \tag{8}$$

This means we have a regression problem where we need to estimate the thresholds τ_k and τ_{k-1} and the regression coefficients b_n.

[3] It is common to justify assumptions of normality with reference to the central limit theorem. I would argue that this often over-interprets the central limit theorem as being much broader than it actually is, and that the real reason why normal distributions get used so often is that they are just so very convenient.