**Data Preprocessing & Checklist\* for Experimental Design (Litson, PSY 6302)**

By Leandro Ledesma

Sep 10, 2024

**Section1:**

1. **Hypothesis/Research Question: What is the research question and hypothesis?**

* **Q1**: Do adults with poor spelling ability show different patterns of EEG activity compared to normal spelling adults?
* **Q2**: If so, does the EEG activity of poor spelling adults look more similar to that of younger individuals with normal spelling ability than age matched peers?
* **H1**: Adults with poor spelling ability will exhibit distinct EEG patterns, such as differences in lower frequency bands (delta and theta) and higher frequency bands (alpha and beta) absolute power compared to normal spelling adults.
* **H2**: Poor spelling adults will have brain activity that looks similar to that of someone younger with normal spelling ability, particularly by having higher absolute power for lower frequency bands (delta and theta) and lower absolute power for higher frequency bands (alpha and beta).

1. **What is the research design?**

* Cross-sectional study where subjects of ages 18-30 years engaged in a 3-minute resting-state EEG, which measured brain activity during a rested state, and a spelling task that measured spelling performance. Additionally, data from an IQ measure known as the CFIT (non-verbal IQ test) and participant’s sex information was collected.

1. **Are variables relevant for answering the research question?**

* Yes, the main predictor of interest is a measure of spelling performance from a spelling task and the outcome is the measure of EEG brain activity during a rested state. Other variables included will function as covariates.

1. **Do you need to control for any extraneous effects, if so, what variable can you use to do so?**

* Yes, a few studies have reported EEG differences based on IQ measures. Thus, we will introduce CFIT scores as an index of IQ to control for its effects on brain activity. Additionally, age influences EEG activity as well, so that will also be included as a covariate.

1. **Based on the hypothesis/research question, what models will you be evaluating?**

* I will be using two regression models. The first will include each of the following variables as predictors: spelling errors, age, IQ scores, and sex with the outcome being EEG absolute power for delta, theta, alpha and beta frequency bands. The second regression will include the predictors: spelling errors, age, spelling errors and age interaction, IQ scores and sex with the same outcome variables.

1. **Are the models being used to test the relevant component(s) of your research question? If not, revisit the research question and/or model development.**

* Yes, the models are testing my research question. The first is looking for a main effect of spelling errors on EEG brain activity while controlling for the effects of age, IQ and sex. The second model introduces an interaction between spelling errors and age, since it could be likely that the relationship between spelling errors and brain activity differs across the individual’s age. Thus, both models get at the bottom of answering the research question, with the latter model taking a more complex approach.

1. **What are the variable types being used in the analysis?**

* Spelling errors, age, and IQ will be a numeric variable. Sex will be a categorical variable with two distinct values, male and female.

1. **What are the distributions of each variable used in the analysis?**

* Total spelling errors, age, and the outcome variables (delta, theta, alpha, and beta) are all heavily skewed to the right. IQ on the other hand is skewed to the left. Thus, there is a good chance that running a regression using these variables will lead to non-normally distributed errors, which violates the assumption of normality. If that happens to be the case, then then variables will be log-transformed to see if that results in normally distributed errors.

1. **Is** **there a true 0 in the variable distribution? If so, is it coded as 0?**

* There are true zeroes in spelling error, and it is coded as the numeric 0. Observations with the value 0 for this variable indicate no spelling errors. IQ and age all contain values greater than zero. The variable sex has the characters male and female as values- there are no numbers.

1. **Do binary variables have codes 0/1 or 1/2 or otherwise (note: ideally code binary variables as 0/1)?**

* In the data processing stage, the values for sex were changed to female = 0 and male = 1. This will mean that the beta coefficient for sex in the regression will represent the effect of being male (on average) on EEG outcomes compared to being female.

1. **Are data standardized? Should they be standardized?**

* Right now, they are not. I don’t think I will standardize them because I can use a function in R to calculate the standardized beta coefficients without needing to transform my numeric variables into z-scores. Thus, I can have a model with both unstandardized and standardized beta coefficients, which is much more informative.

1. **Reliability and Validity: Do variables measure what you think they measure?**

* I want to say yes but I am not 100% confident. So, these data have already been collected in Russia, I am now just doing this analysis to answer a research question separate from the grant that funded the study. They used a spelling measure, which is a subtest from the ARFA, a behavioral assessment that is the first of its kind. It has not gone through any psychometric tests to see if it is actually measuring what it is supposed to be measuring. However, on face value, it is measuring spelling errors, I just cannot know for certain if the current scoring system for this test is accurate. For IQ, I think it is measuring non-verbal IQ. The CFIT has been used by other studies and has gone through psychometric tests. Additionally, studies have correlated scores from the CFIT to other well known non-verbal IQ tests like Raven’s Progressive Matrices and there was a moderate to high correlation. Age and Sex are measuring what I think they measure.

1. **Check items, validity of scales/measured used, etc.**

* I do not yet have the stats knowledge to run psychometric tests on the spelling task to see if the items are valid or not. Maybe I will be able to do this next semester after taking more advanced statistic classes.

1. **Are composite variables reliable?**

* Yes, so the CFIT is a composite score of four subtests that measure non-verbal reasoning performance. The subtests are series, matrices, classification, and conditions. The author of the CFIT showed in the manual how these subtests are reliable with each other to create a final raw score of IQ.

1. **Preprocessing: Have variables already been preprocessed? If so, do you know:**

* Yes, I have preprocessed the EEG data and scored the spelling errors task and the CFIT.

1. **Whether any cases have been removed? If yes, why?**

* Yes, I removed all observations that were missing any of the data from above. I removed them because I do not know how to interpolate data. I also have 449 observations remaining out of ~500, so I should have enough power to answer my research question regardless. However, if I learn how to interpolate missing data then I will use that approach instead.

1. **Whether missing data is coded as a value (e.g., -99, 999, 0, NA)?**

* All missing data showed as *NA* in the dataset.

1. **Is data consistently missing for a specific item/person, within treatment, or for a specific outcome?**

* Most of the missing data is actually for the EEG. However, the data itself is not necessarily missing, but instead it was not included into the final sample size for analysis because it was too noisy. There are metrics we can use to obtain the data quality of the EEG recordings, those that did not meet this criterion were removed- so in a sense a good number of subjects have all data but are missing EEG data due to poor data quality. Aside from that, other missing data from the predictors or covariates were missing at random (I think).

1. **Is missing data has been imputed? What mechanisms of missingness was assumed (missing completely at random, missing at random, missing not at random)?**

* Missing data was not imputed.

1. **If measures are across time/situations: Are the measures administered using the same items and design at each wave of assessment?**

* This is a cross-sectional analysis, so there is no repeated measures design.

1. **Are the data in a format relevant for the analysis you want to use (wide vs long)?**

* The data initially was processed in wide format. It has been changed to long format for the analysis.

**Section 2**: **Update your problem statement appropriately. After completing the data preprocessing checklist, are there any concerns you have about being able to answer your research question.**

* I do not think there are serious concerns with how I am going about analyzing the data to answer my research question. The main shortcomings I see is that a) I do not know how valid one of my measures is because I do not know how to do psychometric tests and b) I do not know how to properly interpolate missing data. Overall, I think these are not huge issues since at face value it seems that spelling errors are measuring spelling errors, just not super accurately and also I have a larger sample size remaining.

**Section 3: Write at least two models that include the different components/variables that you will use to answer your research question. You need not write these in mathematical form, instead, you can write something like:** *Reading recall at age 5* = *treatment (conditions: teaching as usual vs novel teaching approach) + classroom (100 classrooms total) + parent education (measured ordinally) + (treatment X parent education)*

* I will use the same models mentioned in Question 1 part D. These models will both help me understand the relationship between spelling errors and EEG brain activity. The difference between the models is that the second will include an interaction between the main predictor spelling errors and the covariate age.
* **Model1**: *delta absolute power = spelling errors (numeric) + age (numeric) + IQ (numeric) + sex (binary)*
* **Model2**: *delta absolute power = spelling errors (numeric) + age (numeric) + IQ (numeric) + sex (binary) + spelling errors X age*

**Section 4: Describe how these two models help you answer your research question.**

* My main research question is how spelling errors are related to EEG activity. Specifically, I want to investigate if poor spellers have brain activity that matches that of someone who is younger in age but show normal spelling ability. To do this, I will be using two models that include spelling errors as a predictor, EEG absolute power for different frequency bands (delta, theta, alpha, and beta) as outcomes, and age, IQ, and sex as covariates. However, one model will consist of only main effects while the other will include an interaction. There is a chance that the levels of age (numeric not categorical) will produce different regression coefficient for spelling error predicting EEG absolute power. Thus, this implies that maybe for younger people, the relationship between spelling performance and EEG absolute power may be different than in older people. Therefore, having a second model with the interaction between spelling errors and age included as a predictor could be uses to see if it explains more of the variance in the outcome variable than the model without this interaction included.