How did we conduct our research project?

1. Data base and research question

I usually download different higher education databases and after reviewing them I think about what I could do with them. Occasionally I think of a topic or research question first and look for databases that might be useful later. In Chile, most of the information on the education system is publicly accessible and is available on the Datos Abiertos (Open Data) website: https://datosabiertos.mineduc.cl/. Specifically, data on the selective higher education admission process is available on the website of the Department of Educational Measurement and Registration (DEMRE) of the Universidad de Chile, one of the most important universities in the country: https://investigador.demre.cl/

Additional information may be requested from the institutions after filling out a form stating the purpose and objective of the use of the data. Each database has its respective user manual to understand its contents (variables, values, and labels).

For this project, I had the databases and had previously participated in similar research. The research question (objective) was the first thing we defined, so that we could organize a concrete database and define the method of analysis.

Personally, I believe that questions should not always be complex or complicated to answer. Sometimes a simple question allows to start the analysis and then it can be modified or complemented with other things. Our initial question was: Is there a significant relationship between socioeconomic factors and the probability of being admitted to selective higher education in Chile?

This allowed us to quickly define the dependent variable (admission), the predictors (academic and non-academic), and the analysis model (binary logistic regression).

2. Data munging

Data wrangling*, sometimes referred to as* ***data munging****, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics* (thanks Wikipedia!).

Except for Lesa's homework, I have never seen a database that can be analyzed after downloading or receiving it. There is always a long, and sometimes tedious, process of reviewing, organizing, cleaning, modifying, combining, and filtering information.

For this project we used RStudio software. First, we loaded the databases into the program. The admissions process has information on 4 stages: enrollment, socio-

economic information, application, and admission. All 4 databases have a variable called 'StudentID', which we used to combine information from different databases.

After having all the information in a single database, we created new variables for further analysis. We identified missing data and assign specific values to them. In addition, we performed descriptive analysis to see if the range and values of the variables were appropriate. For this we reviewed the user's manual of the database.

The last step was to decide on the sample filters. In our case, we were interested in knowing the situation of the average student, that is, the student who had just graduated from regular high school and had taken the admission tests in mathematics and language.

3. Analysis

Once the database had all the information we needed, we started with the descriptive analysis of the data. Since one of the principles of the admissions process is that all students have similar probabilities of entering higher education, we began by analyzing the scores on the academic admissions factors according to different socioeconomic characteristics of the students.

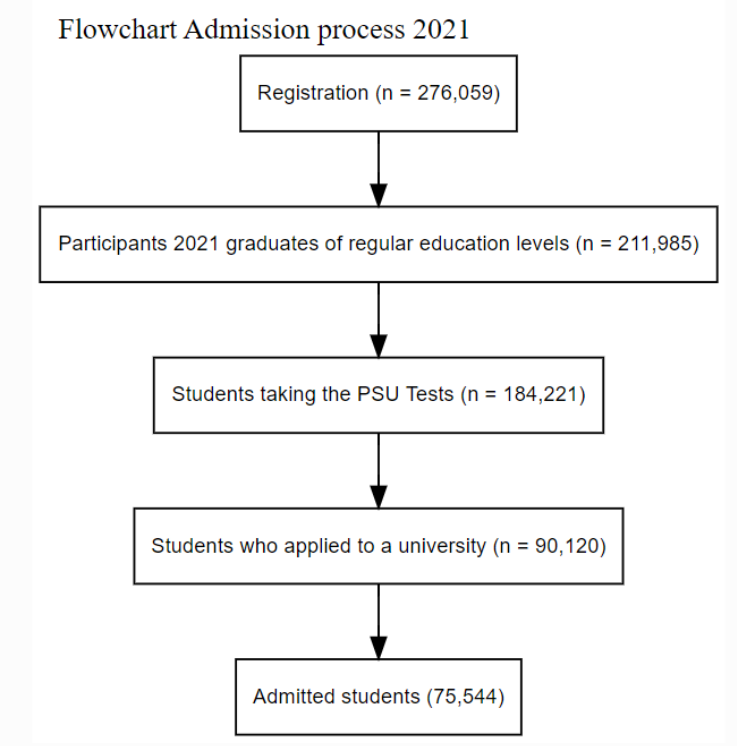
If the average scores on these factors were similar between different groups, then the principle may be defensible. To represent the information we chose box plots, which show the concentration of the data around the median and you can assign colors to the groups using ggplot2.

After reviewing the bivariate relationships between predictors, we fit the binary logistic regression model with the glm function. For this analysis, we first centered the academic predictors on a constant = 550, which is a score close to the average. For the categorical predictors, we used dummy coding, setting as reference the group that, theoretically, was less likely to be admitted (vulnerable student profile). Obviously, the decision on the reference group does not change the role of the variables in the model, only the interpretation of some fixed effects. You can choose any coding scheme and reference.

Our intercept would represent the probability, in logits, that a student with average performance in the academic factors, female, from a public school, with parents with basic education, from a low-income family and from a humanistic school, would be admitted to selective higher education in the year 2021.

Here we found a problem. Of all the students who register to take the admission tests, some did not attend the administration of the tests. In addition, some of them graduated from schools with a different curriculum from the one used to design the admission tests (general curriculum for secondary education), so they have a clear disadvantage. These

students were removed in the first phase of the study. However, we had not considered that some students, after taking the admission exams, decided not to apply. These students formed more than half of the sample!



In order not to overestimate the probability of not being admitted, we discarded the cases that did not apply to a selective program.

The first model we fitted had some predictor variables that were not significant, so we fitted other versions and then compared them with the LRT test. These kinds of decisions can be controversial. Usually, one looks for significant effects, since one wants to show that the predictors contribute statistically to the estimate of the outcome. However, there are variables that are theoretically important and that could be maintained in the model despite their lack of statistical significance. In this case, we decided to remove the non-significant predictor variables because we had several predictors, and we were not missing important variables.

It is important to mention that the models we fitted only included main effects. We did not explore interactions between predictors, mainly due to lack of time. This project was part of a course I took with Mubarak, so time was an important factor. In order not to spend too much time fitting the models, we left this for later. Also, we did not consider the nested structure of the data. Each student applying to a selective program attended a school with other students, who are more likely to be related to each other than to the rest of the sample. When we did this project, we did not know how to perform multilevel analysis. However, we now know how to do it!

In addition to checking the individual significance of each predictor (p-values), we checked the significance of all the predictors with a multivariate Wald test, calculated the *SurrogateR2* of the model (a type of pseudo R2), transformed the slopes into ORs to compare the size of the effect between the predictors, and calculated probabilities according to different attributes of the students (with linear combinations and the inverse function).

The *SurrogateR2* is a relatively new approach that is explained in Liu et al. (2022) article. In their words: *The general idea is to simulate a continuous variable and use it as a surrogate for the original discrete response* (p.195). As we saw in class, and as the authors mention, there is no consensus among researchers as to which goodness-of-fit measure is best for models that do not use OLS. The authors propose to simulate a continuous variable *S* as a surrogate for the original discrete dependent variable, apply a linear regression model with the predictors and this surrogate as the dependent variable, and treat the proportion of variance explained in this model as the R2 of the original model. If you are interested in more details, you can read the article here. Something important to mention is that this GOF measure was originally intended for probit models, although as we know, the results of logit and probit models are similar if we apply a linear transformation: logit coefficient * 1.7 = probit coefficient.

For the ORs we created a graph comparing between academic and non-academic factors, and then we created boxplots to show the distribution of admission probabilities among students.

4. Discussion and conclusion

Once we finished reviewing the descriptive results and model estimation, we focused on interpreting them according to the Chilean educational context, the implications they have for the admission process, the socioeconomic differences among students, and how they are evidence of an apparent disconnect between the assumptions of the admission process and how this works.

5. Take away

It is crucial to have a clear and precise **research objective**. This helps in the exploration, analysis, and interpretation of the data.