UNIVERSITY OF OSLO

Survey analysis and data bases

Analysis of survey data and key decisions

Martin Moland

ARENA, Centre for European Studies

September 11 2024

Overview

1. Introduction
2. Variable coding and recoding
3. Latent variables: Creating indices
4. Estimation choices
5. Unusual and little used values
6. Missing data
7. Dealing with outliers

Our running example

“What is the relationship between social conservatism and wanting or opposing EU labor migration?”

Different types of variables

- **Binary:** Only two response categories
- **Ordinal:** Ordered response categories, going from instance to lowest to highest.
- **Continuous:** (Theoretically) infinite range of responses, with equal distance between all of them.

Our dependent variable

Our DV asks people about the following statement: “People must be Free to Move/Work across Borders”, with the following response categories:

1. Strongly Disagree
2. Somewhat Disagree
3. Neither
4. Somewhat Agree
5. Strongly Agree

What kind of variable is this?

Recoding ordinal variables

- **First choice: What do you recode it to?**
  - Fewer categories (from 5 to 3, for instance)
  - Binary

- **What impact does variable recording have?**
  - Determines how much info you have (coarser variables = less info)
  - Determines the kind of analysis you can do.

Practical example (I)

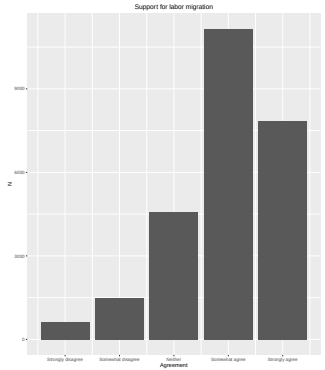

Figure: Ordinal variable (Support for labor migration)

Practical example (II)

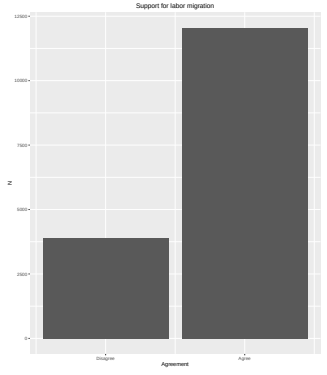

Figure: Binary variable (Support for labor migration)

What changed between the two?

- We changed “Neither” to “Disagree”. Increased negative responses substantially.

- Hides the fact that only a very small number of people really hate EU labor migration.

A middle ground

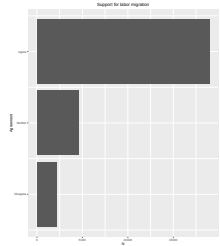

Figure: The middle ground (Support for labor migration)

- **Shows a middle ground:**
  - We still get the many who somewhat or strongly agree.
  - We do not hide the fact that quite a few people have no opinion.

Recoding: Summary

- The more categories you merge, the coarser picture you'll be able to paint.

- It will also impact the kind of analysis you can run.

What are indices and when would you use them?

- Typically used to measure **latent variables** (i.e. factors you can't easily observe)

- Create **an average of various measures** that correlate to an underlying factor.

A practical example: Social conservatism

- We want to measure the impact of social conservatism on support for labor migration

- **Use three components:**
  - Environmental concerns
  - Views on multiculturalism
  - Views on international cooperation

Before you begin

- **Make sure the variables are on the same scale.**
- **Ensure they go in the same direction**:
  - Sometimes one or more of the response categories will go in the opposite direction.
- **Lastly, run a separate "sanity check" for each of the included variables:**
  1. Do the variables (on the face of it) seem to measure what you actually want them to measure (the distinct component of a broader measure)?
  2. Are the questions "double"- or "single"-barrelled (asking about only one thing)?
  3. Ideally: Has someone else used this question to measure the same thing before?

Factor analysis: Social conservatism

```
Factor Analysis using method =  minres
Call: psych::fa(r = fa_data_galtan, nfactors = 1, rotate = "oblimin")
Standardized loadings (pattern matrix) based upon correlation matrix
                              MR1   h2   u2 com
Transnational solidarity     0.58 0.34 0.66   1
Immigration enriches culture 0.64 0.41 0.59   1
Environmentalism             0.62 0.38 0.62   1
```

Figure: Output from factor analysis

- **Short takeaway:**
  - I did the factor analysis with *psych* in R.
  - All variables load relatively strongly on a “social conservatism” variable ($\beta \geq 0.5$)

The DV-estimator connection

- **The choices we made earlier impact what kind of model we use:**
  - If we made a binary DV: Logistic regression.
  - If it is ordinal (between three or ten categories): Ordinal logistic regression.
  - If it is continuous: OLS.

What to do with unusual values?

- **Some values are typically less used (and often less informative):**
  - Usually middle categories like “Neither” or “Either-or”.

- **One big question:**
  - Is the middle category something you actually care about?

Returning to our example

- Our middle category: Those who “neither” agree nor disagree on whether they support labor migration.
- We can either **take it out**…
  - Generally the right choice if we don’t want to know anything about those that do not have any opinion.
- …Or we can **leave them in**:
  - The right choice if we are interested in for instance what determines *whether* someone has an opinion, rather than what that opinion is.

Second question: The Don’t knows

- **Surprisingly tricky in opinion research:**
  - Is “Don’t know” an absence of opinion or just indifference?
- The most common choice is to just **remove the Don’t knows** from the data.
- **I would also think about some likely reasons for why you get the Don’t know answers:**
  - If it is an attitude question: They might just not have thought all that much about it.
  - If it’s about a behavior: They might not remember. You are likely to see more of these if asking about specific behaviors, like signing petitions and so on.

Our example: Support for labor migration

```
GLM estimation, family = binomial, Dep. Var.: people_free_dico
Observations: 12,855
Weights: subset_data$sample_weight
Fixed-effects: country_code: 8
Standard-errors: Clustered (country_code)
```

Figure: Output from *fixest*

- **I've made a couple of choices here:**
  - Made a binary dependent variable (**dico = dummy variable**)
  - Used a logit model (**family = binomial**)
  - **Important point:** Data originally has more than 16.000 units. Observations suggest a lot of missing values. What to do about those?

Types of missing data

- **MCAR:** No pattern to the missing data
- **MAR:** Missing data is related to factors you already have in your dataset.
- **MNAR:** Missing data is related to factors you can control for *or* other factors (this can get really tricky).

In our case we have 342 people with no valid value on the question. Given our discussion about why people may not enter valid responses, what is the most likely reason for these missing values?

Three solutions to missing data

- **Listwise deletion:** When analyzing the data, remove everyone with a missing value on one of the variables in the model.
- **Single imputation:** Use one number to stand in for the missing value (such as the mean value).
- **Multiple imputation:** Uses correlated information to impute (several) possible values.

Which should you use?

- **With very little missing data:** You are generally safe with just removing every unit with a missing value.

- **Moderate to high levels of missing data:** Multiple imputation is generally considered the gold standard.

How do you do this in practice?

- **For single imputation:** Calculate a mean value for a given variable. Specify that every case of “NA” should be replaced by the mean value.

- **Multiple imputation:** Easiest to use R software that does this procedure for you.
  - *Amelia* is the easiest one to use (in my view), but *mice* is also well-regarded.
  - Most regular R functions will work with output from *mice*.

What are outliers?

- Outliers are values that lie significantly outside of the distribution of the rest of your data.
- **But beware:** Outliers can either be “real” outliers or just weird coding choices (coding “Don’t know” as 99 is a common one).
- **To test for** outliers, you can use box plots to visualize the data.
  - Quick way to visualize whether you actually have outliers and what they look like.

How do you handle them?

- **One choice:** Delete them
  - However, that mainly works if you know you *really* don't need them.

- **A better choice:** Log-transform variables (typically done for incomes that are really strong outliers)

Tying it all together

- **So how do you approach working with a new data set?**
  1. Figure out your dependent and independent variables
  2. Spend some time cleaning data
  3. Think about how to handle unusual values like outliers and “Don’t know”
  4. Think through what your dependent variable is and how this impacts the estimator.