**Regression without regrets: Workflow of initial data analyses**

The focus of this document/website is to provide guidance on conducting initial data analysis in a reproducible manner in the context of intended regression analyses.

1. IDA Framework
2. Scope of regression models for the examples
3. IDA tasks for data screening
4. Dataset 1
   1. Introduction to the dataset
      1. Overview
      2. Meta data/links to data dictionary
      3. References where the data set was used in research studies. Caveat of what we did differently and assumptions we made. Source.
   2. Statistical analysis plan
   3. Univariate summary
      1. Numerical summary tools of some sort
      2. More detailed histograms as they relate to variables in the SAP and to be identified in step i.
      3. Grids of barplots for categorical variables
      4. Missing data
      5. Session info
   4. Multivariate summary
      1. Numerical summaries stratified by categorical variables
      2. Grids of scatterplots (continuous variables)
      3. Facet grid (categorical or continuous variables by categorical variables)
      4. Session info
   5. Consequences of IDA for the analysis plan
   6. Reporting of IDA results in a research paper
5. Dataset 2
6. Dataset 3
7. Reference to our IDA for regression paper and slides from talks
8. IDA framework

The IDA framework consists of six steps [Huebner et al 2018, Figure 1], here we assume that metadata (step I) exist in sufficient detail, and that data cleaning (step II) was already performed. Metadata summarize background information about the data to properly conduct IDA steps, and a data cleaning process identifies and corrects technical errors.  The data screening (step III) examines data properties to inform decisions about the intended analysis. Initial data reporting (step IV) document insight of the previous steps and can be referred to when interpreting results from the regression modeling. Consequences of these analyses can be that the analysis plan needs to be refined or updated (step V). Finally, reporting of IDA results in research papers (step VI) are necessary to ensure transparency regarding key findings that influence the analysis or interpretation of results. Further details about the elements of IDA are discussed  in [TG3 papers].

A screenshot of a cell phone

Description automatically generated

References

Huebner M, le Cessie S, Schmidt CO, Vach W . A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. [Link](https://www.google.com/url?q=https%3A%2F%2Fobsstudies.org%2Fcontemporary-conceptual-framework-initial-data-analysis%2F&sa=D&sntz=1&usg=AFQjCNEiIZaQtDqVKwCEjDKJRnnstrbAYw)

Huebner M, Vach W, le Cessie S, Schmidt C, Lusa L. Hidden Analyses: a eview of reproting practice and recommnedations for more transparent reporting of initial data analyses. BMC Med Res Meth 2020; 20:61. [Link](https://www.google.com/url?q=https%3A%2F%2Fbmcmedresmethodol.biomedcentral.com%2Ftrack%2Fpdf%2F10.1186%2Fs12874-020-00942-y&sa=D&sntz=1&usg=AFQjCNFwlg30WgMnyKO_BigNZELyDjN_7w)

1. **Scope of the regression analyses for the examples**

Regression models can be used for a wide range of purposes, for the purpose of these examples the assumptions on the regression analysis set-up in this paper are listed in Table 1. Thus, IDA tasks will be explained in a well-defined, practically relevant setting. Since a key principle is that IDA does not touch the research question no associations between dependent (outcome)  and independent (non-outcome) variables are considered

**Table 1:** The scope of the regression analyses considered for IDA tasks

|  |  |  |
| --- | --- | --- |
| **Aspects of the research plan** | **Assumptions in this paper** | **Reason for the assumption** |
| Dependent (outcome) variable | One dependent variable that can be continuous or binary;  exclude time-to-event or longitudinal outcomes | Explain IDA tasks in a well-defined, practically relevant  setting |
| Regression models | Models with linear predictors |
| Purpose of regression model | Adjust effect of one variable of interest for confounders; quantify the effects of explanatory variables on the outcome |
| Independent variables | “explanatory” or “confounder” depending on purpose of model;  small to moderate number of mixed types; Not high dimensional; no repeated measurements | To demonstrate IDA approaches for a mix of variables likely to be encountered in practice |
| Statistical analysis plan | Exists,  defines the outcome variable, the type of regression model to be used, and a set of independent variables | IDA does not touch the research question,  but may lead to an update or refinement of the analysis plan |

References:

Vach W. Regression Models as a Tool in Medical Research. Chapman/Hall CRC 2012

Harrell FE. Regression Modeling Strategies. Springer (2nd ed) 2015

Royston P and Sauerbrei W. Multivariable Model Building. Wiley (2008)

[…]

1. Data screening and possible actions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **What to look at** | **Possible actions: Interpretation** | **Possible actions: SAP** | **Possible actions: Presentation** |
| **Univariate distributions** | | | | |
|  |  | Help in interpreting results | Update SAP | Update intended presentation of results |
| Continuous variables | General skewness | Wide CI for coefficients | Use variable as log-transformed |  |
|  | Outliers | Disproportional impact on results | Winsorize or transform | Model involves winsorization |
|  | Spike at 0 | Narrow CI at 0 | Use appropriate representation of variable in model | Use 2 (or more) coefficents to distinguish 0 from non-0 continuous part |
| Categorical variables | Frequencies | Comparisons to default reference probably irrelevant | Change reference category | Contrasts compare to (new) reference category |
|  | Rare categories | Wide CI for coefficients | Collapse/exclude | Fewer categories to present |
|  | One very frequent category | Comparisons irrelevant? | Exclude variable | Variable omitted |
| **Bivariate distributions** | | | | |
| Continuous by continuous | Outliers (from the cloud) | Disproportional impact on results | Winsorize or transform | Model involves winsorization |
|  | Correlations | Wide CI for coefficients |  |  |
|  |  |  |  |  |
| Continuous by categorical | Outliers (only visible in bivariate plot) |  |  |  |
|  |  |  |  |  |
| Categorical by categorical | Frequent/rare combinations | Comparison to default reference irrelevant | Change reference category | Contrasts compare to (new) reference category |
|  |  | interactions relevant? | Remove interaction from model | Fewer interactions to present |
| **Missing values** | | | | |
| Per variable | Number and proportion | Wide CI for coefficients | Remove variable if many missing values |  |
| Pattern | Variables missing independently or together |  | Omit variables together | Changes model |
| Pattern | Variables missing dependent on levels of other variables | Systematic missingness? Model still based on representative? | IPW needed? | Weighted analysis |
| Complete cases | Number and proportion | Few cases left for main CCO analysis | Multiple imputation (or other way of dealing with missing values)? | Result from MI analysis? Or applicability restricted to a subpopulation? |

References:

Huebner M, le Cessie S, Schmidt CO, Vach W . A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. [Link](https://www.google.com/url?q=https%3A%2F%2Fobsstudies.org%2Fcontemporary-conceptual-framework-initial-data-analysis%2F&sa=D&sntz=1&usg=AFQjCNEiIZaQtDqVKwCEjDKJRnnstrbAYw)

Harrell

[…]