Missing data

Silje Synnøve Lyder Hermansen

2023-05-11

Recap: our course

Recap: our course

We are entering the last part of this course

1. R-skills (week 1-3)
2. Limited and categorical outcome variables (GLMs) (week 4-10)
3. **Data structures** (week 11-14)

The purpose of this course

The purpose of this course

$\Rightarrow$ *The purpose of this course is to find solutions when the assumptions of the linear model are not satisfied*

Two assumptions in ordinary regression

Two assumptions in ordinary regression

Linear models (OLS) rely on two assumptions that are often violated

1. **Assumption 1:** outcomes are continuous and unbounded (week 4-10)
2. **Assumption 2:** observations are independent and identically distributed (iid) (week 11-14)
   - independent: probabiliy of observing one unit is not dependent on observing another
   - identically distributed: observations come from the same data generating process/probability distribution

$\Rightarrow$ *strategies for when these are not satisfied*

Solutions to violations of those assumptions

1. **Assumption 1:** Limited and categorical outcome variables (GLMs): - recode the dependent variable and describe the data generating process w/probability distribution - choice of model depends on the data generating process - e.g. logit, multinomial, ordinal, poisson, neg.bin, zero-inflated, coxph…
2. **Assumption 2:** Observations are not iid: - hierarchical/nested data - missing data

$\Rightarrow$ *what do we do when observations are not iid?*

Today (week 13 and 14)

Sources of missing data

Sources of missing data

Most data contain missing observations

- missing data (`NA`) is the result of a "lurking" variable that :
  - assigns `NA` to some of the other variables
  - ... possibly affecting both x and y
- the "lurking" means that the assignment mechanism is not observed
  - think about the data generating process of the `NA`
  - we have to theorize/make assume

$\Rightarrow$ *addressing/reducing the problem is often easier than what we think*

Classifications of missing data:

Take 1

The original classification by Rubin (1979)

- **MCAR** (Missing Completely at Random)
  - probability of NA is the same for all cases
- **MAR** (Missing at Random):
  - probability of NA depends on *observable* data (known sources)
- **MNAR** (Missing Not at Random)
  - probability of NA depends on *unobservable* data (unknown sources)

$\Rightarrow$ *these are assumptions that we can never test*

Why is it a problem

- **statistical power (MCAR):** only a problem if it reduces the N too much $\rightarrow$ *a representative sample*
- **information bias (MAR, MNAR):** we only record parts of a phenomenon (recall bias, missclassification, observer bias. . .)
  - independent variables:
    - we might not get the full “elasticity” of the variable
  - dependent variable: do we underestimate our phenomenon?
- **selection bias (MAR, MNAR):**
  - our estimate is biased because the unobserved assignment of NA affects both x and y

Take 2

We can subdivide the last category

- **MCAR** (Missing Completely at Random)
  - NA are *not* dependent on any predictors (observed or not): not conditional → *you can ignore the problem, unless you have too little statistical power*
- **MAR** (Missing at Random):
  - NA depends on the value of other observed predictors: it is conditional → *ignorability; you can condition on the other predictors*
- **MNAR** (Missing Not at Random)
  - NA depends on unobserved data
  - NA depends on the value of *the predictor itself* (e.g. censoring)

→ *NA must be modeled, or you will have to accept a biased estimate*

Strategies

Simple strategies

Discard data

Ignore the problem

- complete case analysis:
  - the usual “listwise.exclusion”
- available data analysis:
  - analyze subsets of data separately
  - exclude variables with missing observations
- weighing of `NA` according to predictors
  - common in surveys → *some cases may be underrepresented in the data, because of NA*

Replace each NA by a single value

We can also infer the missing values in fairly simple ways

- **mean imputation:**
  - replace the missing data by the variable mean
- **conditional mean imputation:** use information from other variables
  - group mean, regression predictions

$\Rightarrow$ *still possible to insert bias, and doesn't take into account the uncertainty from our estimate*

Multipe imputation

Multipe imputation

Multiple imputation generates several predictions for each missing value to account for the uncertainty

- step 1: make predictions for the missing values by adding som random noise for each model

$\rightarrow$ *we end up with several data sets (5-20 frames)*

- step 2: estimate the main model on all the different data sets

$\rightarrow$ *pool over the regression parameters*

EM algorithm

EM algorithm

The EM is the base-line approach, and only has one data frame in the end

- we have several variables
- E-step:
  - give your NA some initial values
  - predict your x_{miss} using the observed values and initial values of x (and all other predictors)
- M-step:
  - re-do until you your predictions of x_{miss} don't change any more (set a value at which you stop)

$\Rightarrow$ *classic maximum likelihood with a twist*

Multiple Imputation via Chained Equations (MICE)

Multiple Imputation via Chained Equations (MICE)

We assume a set of variables are correlated, and use them to predict for each other in turn (a cycle)

Figure 1: Mice thrive in holes...

Imagine x, y and z:

- cycle 1:
  - $x \ \alpha + \beta_1 y + \beta_2 z$:give y and z some starting value; regress x on all other models
  - $y \ \alpha + \beta_1 \hat{x} + \beta_2 z$: replace missing values of x by predicted $\hat{x}$
  - $z \ \alpha + \beta_1 \hat{x} + \beta_2 \hat{y}$: same
- cycle 2-...: rinse and repeat until nothing changes (convergence)

. . . and add som random noise

This is usually done together with a bit of (random) noise at each step

- for each iteration, create a new data set with imputed variables
- run regular (g)lm on each data set:
  - regression parameters (β) are averaged over
  - the standard error is a fusion of both:
    - within-model variation: standard errors from the regression
    - between-model variation: the deviation between the regression parameters

..then check

. . . were my imputations appropriate?

- problem of **overfitting**:
  - you may have perfect in-sample predictions, that are useless for out-of-sample imputation
- **overimputation**: randomly leave observations out, check if you predict correctly

Literature