Evaluation of a survey on factors influencing case reporting in Switzerland – how to handle missing data in covariates

Birgit Schrödle

Zurich, July 1st, 2010

Overview

1. Survey details
2. Example: Stockholders and disease awareness
3. Dealing with missing covariates

The survey

- Registry data of the Federal Veterinary Office (BVET) for passive surveillance of about 80 notifiable diseases
- But: Case reporting can be biased due to, e.g., limited case detection or low reporting motivation (underreporting)
- Complex reporting chain:
  stockholder → veterinary → laboratory → cantonal veterinary → BVET
- Which factors influence the quality of case reporting?

→ Large study among all members of the reporting chain, conducted in October 2009

- Participants:

Stockholders: 1177/∼ 8000	Veterinaries: 132/∼ 500
Laboratories: 19/29	Cantonal veterinaries: 17/22

Connect registry data and results from the survey

- Registry data are analyzed using a spatio-temporal disease mapping model
- Idea: Include data from the survey as covariates
  → Ecological regression
- Problem 1: Reported cases are based on 185 Swiss regions
  → Survey data have to be assigned to Swiss regions (post code)
  → Problematic for laboratories,...
- Problem 2: Several items of the questionnaire are not disease-specific
- Problem 3: Vast amount of data
  → List of hypotheses was generated

Example: Stockholders and disease awareness

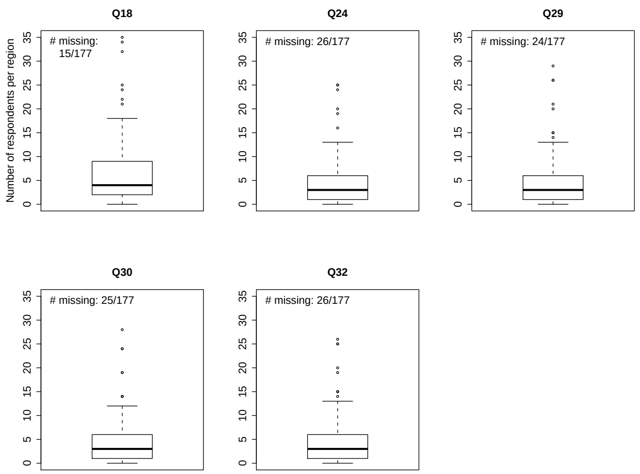
- Hypothesis: The disease awareness of a stockholder influences the reporting behavior of a stockholder and, hence, the number of reported cases in a region
- Disease-specific (e.g. Salmonellosis in cattle) items on "disease awareness" in the questionnaire:
  Q18: Do you know this disease? (no/a bit/yes)
  Q24: I am able to recognize this disease from the symptoms.
  (no/rather no/rather yes/yes)
  Q29: I have a good knowledge about this disease.
  (no/rather no/rather yes/yes)
  Q30: I am aware of symptoms of this disease among my cattle.
  all the time (no/rather no/rather yes/yes)
  Q32: I think that I am jointly responsible to fight against this disease. (no/rather no/rather yes/yes)

What to do next? - Some drawbacks

- Form an average covariate for each region from all respondents within one region
- Number of respondents is different for each region
- No respondent for several regions → covariate is missing
- No Italian translation of the questionnaire:
  → no data for all regions in the canton Tessin
  → exclude the canton Tessin from the analysis

Available covariate data

- Number of regions where covariate information is missing
- Boxplot of the number of respondents per region



Missing pattern for all regions in the canton Graubünden

0≙ covariate available
x≙ covariate missing

RegionID	Q18	Q24	Q29	Q30	Q32
1801	0	0	0	0	0
1802	×	×	×	×	×
1803	0	0	0	0	0
1804	0	0	0	0	0
1805	0	0	0	0	0
1806	×	×	×	×	×
1807	0	×	0	0	×
1808	×	×	×	×	×
1809	×	×	×	×	×
1810	×	×	×	×	×
1811	0	0	0	0	0
1812	0	×	×	×	×
1813	0	0	0	0	0
1814	0	0	0	0	0

→ non-monotone missing pattern

Missing data mechanisms

- When subjects with missing covariate values differ systematically from those with complete data with respect to the outcome of interest, results from a traditional data analysis omitting the missing cases may no longer be valid. (Ibrahim et al., 2005, JASA)
- Consider data $\mathbf{X}$ which can be decomposed in $\mathbf{X}_{\mathbf{obs}}$ and $\mathbf{X}_{\mathbf{mis}}$
- Missing completely at random (MCAR): The missingness of $\mathbf{X}_{\mathbf{mis}}$ is independent of $\mathbf{X}_{\mathbf{obs}}$ and $\mathbf{X}_{\mathbf{mis}}$
- Missing at random (MAR): The missingness of $\mathbf{X}_{\mathbf{mis}}$ is independent of $\mathbf{X}_{\mathbf{mis}}$, but dependent on $\mathbf{X}_{\mathbf{obs}}$
  $\longrightarrow$ Missingness is random after controlling for missingness due to observed quantities
- Missing not at random (MNAR): The missingness of $\mathbf{X}_{\mathbf{mis}}$ is dependent on $\mathbf{X}_{\mathbf{obs}}$ and $\mathbf{X}_{\mathbf{mis}}$

Methods

- Imputation of the mean value, last value carried forward,... are not statistically valid
- Statistically valid methods (MAR):
  - EM algorithm (likelihood framework)
  - Multiple imputation
  - Fully Bayesian approach
  - ...
- Stockholder survey: Hierarchical Bayesian model
  $\longrightarrow$ use multiple imputation or fully Bayesian model
- Multiple imputation, 3 steps:
  1. Impute plausible values for missing covariates using an imputation model; repeat this process about 5-10 times.
  2. Analyze each obtained dataset using complete-data methods.
  3. Combine the results appropriately, taking into account the uncertainty regarding the imputation.

Some remarks on the imputation model

- If the missing pattern is non-monotone, a truly multivariate imputation model has to be used:
  Joint modelling of all covariates or full conditional specification of each covariate
- The complete-data analysis should not use a richer model than that used for imputation.
- Any knowledge about the process that generated the missing data should be included.
- A sensitivity analysis with regard to the imputation model should be conducted.
- The process of specifying the imputation model is a scientific modelling activity on its own.
- Useful guidelines for "reporting any analysis potentially affected by missing data" in a medical publication were recently published in Sterne et al. (2009, BMJ).