Modeling accounting for data collection (ch 7)

- *Ignorability*
- Complete, observed and missing data
- Stability and stable treatment
- Super population, finite population

Slide 1

- Sample surveys, designed experiments, observational studies
- Censoring and truncation

Modeling accounting for data collection – example

- Outcomes of ten rolls of a die and all are 6's
- Does following information affect your inference
  - these were the only rolls performed
  - the die was rolled 60 times, but only 6's were reported
  - the die was rolled until 10 6's was obtained

Slide 2

Modeling accounting for data collection – example

- Course feedback is asked from
  - all students
  - random students
  - students attending lectures
  - students who received the highest grade

Slide 3

Modeling accounting for data collection – example

- How the data was collected is important information
- *Designs*, which are *ignorable*, are useful
  - inference less sensitive to assumptions made in the model
  - eg. *randomization*
- More explanatory variables means more valid inferential conclusions conditionally, but possibly more sensitive to the model specifications relating to the explanatory variables

Slide 4

Modeling accounting for data collection

- *Observed* vs. *Missing* data
  - there is a complete data, which we have observed partially
  - eg: course feedback
    - complete data: all students answer
    - observed data: only some of the students answer
    - missing data: some of the students did not answer

Slide 5

- Inference is conditional on observed data and on the pattern of observed and missing observations
- Missing data can be
  - unintentional
    - eg: nonresponse and censored measurements
  - intentional
    - eg: data from people not in the survey or treatments not applied

Modeling accounting for data collection – examples

- Sampling
  - observed: values for n observations
  - complete: values for every N unit in the population
- Medical experiment
  - observed: results of treatments for patients
  - complete: results of every treatment for every patient

Slide 6

- Rounding
  - observed: rounded values
  - complete: unrounded values
- Unintentional missing data
  - observed: observed values
  - complete: observed and missing values

Modeling accounting for data collection

Slide 7

- Notation
  - complete: $y = (y_1, \ldots, y_N)$
  - y_i may be a vector which components are y_{ij}
  - $I = (I_1, \ldots, I_N)$
  - I_i may be a vector which components are I_{ij}
  - if $I_{ij} = 1$, y_{ij} is observed
  - if $I_{ij} = 0$, y_{ij} is missing
  - $\text{obs} = (i, j) : I_{ij} = 1$
  - $\text{mis} = (i, j) : I_{ij} = 0$
  - previously on this course $y = y_{\text{obs}}$
  - more complex patterns, than what can be presented with indicator $0/1$, are possible

Modeling accounting for data collection

Slide 8

- *Stability*
  - process of measuring does not change the values of the data
- *Stable unit treatment*
  - treatment applied to any particular unit does have no effect on outcomes for the other units
  - counter-example: agricultural experiment that test several fertilizers on closely spaced plots
- Without stability assumptions notation is more complicated

Modeling accounting for data collection

- Complete-data likelihood $p(y, I|\theta, \phi)$
  where ϕ is inclusion-vector parameters
- Observed-data likelihood $p(y_{\text{obs}}, I|\theta, \phi) = \int p(y, I|\theta, \phi) dy_{\text{mis}}$
- Super-population inference $p(\theta, \phi|x, y_{\text{obs}}, I)$
- Finite-population inference $p(y_{\text{mis}}|x, y_{\text{obs}}, I, \theta, \phi)$
  - if all units in a finite population are observed, then finite-population inference is exact, but there is still uncertainty on super-population
  - eg. if the length of all students in the class is measured, we know the observations for all, but if the measurement device is inaccurate, the true lengths are still uncertain
- Posterior predictive distributions
  - future complete in theory easy
  - future observed data needs to take inclusion mechanism in to account

Slide 9

Modeling accounting for data collection

- *Ignorability*
  $$\text{if } p(\theta|x, y_{\text{obs}}, I) = p(\theta|x, y_{\text{obs}})$$
  no need to model the data collection process and thus it is on *ignorable*
  - very useful property
  - ignorability assumed in the examples earlier in the course
- Sufficient conditions to ensure ignorability
  - *missing at random* $p(I|x, y, \phi) = p(I|x, y_{\text{obs}}, \phi)$
    - given ϕ missing depends only on x and y_{obs}
    - holds also for deterministic inclusion which depends only on x
  - *distinct parameters* $p(\phi|x, \theta) = p(\phi|x)$
    - parameters of the inclusion process and data generating process are independent
- Known vs. unknown designs

Slide 10

Sample surveys

- Simple random sampling from a finite population
  - eg. hours per week used for studying this course by the students
  - assume exchangeability of units
  - *strongly ignorable* $p(I|x, y, \phi) = p(I|x)$ and known
  - eg. finite-population mean

Slide 11

$$\bar{y} = \frac{n}{N}\bar{y}_{\text{obs}} + \frac{N-n}{N}\bar{y}_{\text{mis}}$$

and

$$\bar{y}|\bar{y}_{\text{obs}} \sim t_{n-1}\left(\bar{y}_{\text{obs}}, \left(\frac{1}{n} - \frac{1}{N}\right) s^2_{\text{obs}}\right)$$

Sample surveys

- *Stratified sampling*
  - N units divided in J strata
  - a simple random sample of size n_j is drawn from each strata $j = 1, \ldots, J$
  - *ignorable* given J vector of indicator variables $x_1, \ldots, x_J$, which describe which unit belongs to which strata
  - hierarchical model often natural choice for this kind of data

Slide 12

  - eg. SAT-example and meta-analysis

Sample surveys

- *Cluster sampling*
  - N units divided in K clusters
  - first sample J clusters and then sample from each cluster n_j units
  - *ignorable* given indicator variables and number of units in each cluster
  - analysis similar to stratified sampling, except that it has to be taken account that not all stratum are observed

Slide 13

  - eg. in SAT-example there could have been more schools, but only some of the schools were observed

Sample surveys

- Unequal probabilities of selection
  - sample survey is *ignorable* given probabilities for selecting each unit

Slide 14

Designed experiments

- An *experiment* involves the assignment of controlled *treatments* to units
- Often only one treatment can be assigned to one unit
  - missing data is outcome given other treatments
- *Design of experiment* important

Slide 15

Designed experiments

- Completely randomized experiments
  - selection of treatments completely random
  - or units to be treated will be divided randomly in equal sized groups
  - assume stable treatment
  - unit can get only on treatment
  - selection of treatments is *known* and *ignorable*

Slide 16

Designed experiments

- *Latin square*
  - example of more complex experiment
  - each row and column has equal amount of treatments
  - balancing reduces variability
  - eg. field divided in 5x5 plots and 5 different fertilizer amounts tested
  - assume stable treatment

Slide 17

  - *ignorable* if the coordinates of the plots included as explanatory variables
    - not needed in completely randomized designed
    - this way more relevant and accurate
  - note, it is also possible to include other explanatory variables, such as distance from the river, etc.

Designed experiments

- *Sequential design*
  - randomized experiment, where selection probability of a treatment for unit i depend on randomization or outcomes of the previously treated units
  - *ignorable* given all variables used to decide treatment choice, including time and previous outcomes

Slide 18

Designed experiments

- Smallest sufficient set of explanatory variables
  - It is possible to find out the *adequate summary* for designed randomized experiment and minimal analysis based on that
  - often additional information available, which is useful to include
    - eg. in agricultural example, the distance from the river

Slide 19

Modeling accounting for data collection

- Complete randomization vs. systematic designs
  - both can *ignorable*
- Complete randomization
  - possible to estimate the effect of leaving explanatory x out of the model
  - easier posterior predictive checking
  - less sensitive to modeling assumptions for y given x
  - smaller chance to "cheat" accidentally or intentionally
- Systematic designs
  - due to balancing smaller posterior uncertainty about interesting quantities (but the sensitivity to the modeling assumptions)

Slide 20

Modeling accounting for data collection – example

Slide 21

- Complete randomization vs. systematic designs
  - example: poll students passing through Exactum entrance hall between 10am and 11am
  - two ways to model
    (1) *nonignorable* since probability that a student i is included in the sample depends on non-observed routes of $N - n$ students
    (2) *ignorable* since probability that a student i is included in the sample is dependent on indicator variable x_i, which states whether student pass the lobby at given time
  - in case (1) need to model I given y
  - in case (2) infer distribution of y given x no data if if $x = 0$
  - in both cases the result is sensitive to prior assumptions, unless n/N almost 1

Observational studies

Slide 22

- *observational studies*
  - treated units are observed, but treatments are not controlled
  - eg. in SAT-example experiment was designed if students getting coaching were selected randomly
    - in observational study, students would have decided themselves, whether to participate in coaching
  - in a good observational study
    - background information on units known
    - sufficient amount of independent units with both treatments
    - study designed without reference to the outcome of the analysis
    - amount of unintentional missing data is minimized (eg. nonresponse)
    - the analysis takes account of the information used in the design
  - most difficult part is the background information
    - eg. in SAT-example what are the previous grades and motivation

Observational studies

- *Observational studies*
  - given enough explanatory variables, we may assume that observational study is *ignorable*
  - note that, causal inference is different and more difficult

Slide 23

Censoring and truncation

- Weigh an object 100 times on a scale, with known measurement distribution $N(\theta, 1)$ and we observe 91 values

1. Data missing completely at random with known probability of missingness
   - inclusion independent of y, and thus *ignorable*
2. Data missing completely at random with unknown probability of missingness π
   - *ignorable* if π and θ a priori independent
   - *nonignorable* if π and θ a priori dependent and now we need a joint-model for π and θ
3. Censored data: values over 200kg are reported as “too heavy´´
   - 9 censored measurements contribute additional information
   - need a joint-model, where y_{mis} unknown and need to integrate over it
4. Censored data with unknown censoring point
   - same as 3., but censoring point in the joint model is unknown

Slide 24

Censoring and truncation

- Weigh an object 100 times on a scale, with known measurement distribution $N(\theta, 1)$ and we observe 91 values

5. Truncated data: over 200kg values not observed, and know information about how many of such values
   - truncation point truncates the observed-data likelihood
   - can be analysed as censored data, where N unknown and prior $p(N) \propto 1/N$

Slide 25

5. Truncated data with unknown truncation point
   - same as 5. but truncation point in the joint model unknown
   - with non-informative prior on truncation point, marginal posterior distribution same as in case 1.

Censoring and truncation

- Censored data is common in survival analysis
  - when comparing alternative medical treatments, when the experiment end, some of the patients are still alive and thus their survival time is censored with known censoring point
  - similar case is lifetime estimates for engineered devices (like light-bulbs and hard discs)

Slide 26