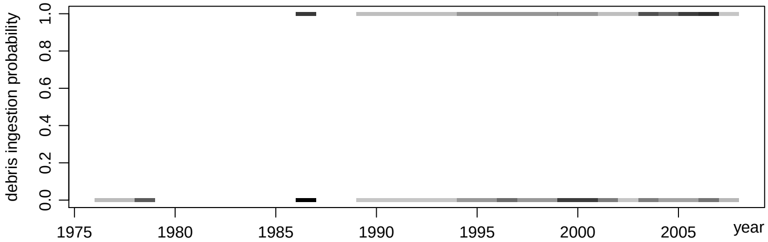
M-estimation when data values are not completely known

Marco Cattaneo

Department of Mathematics
University of Hull

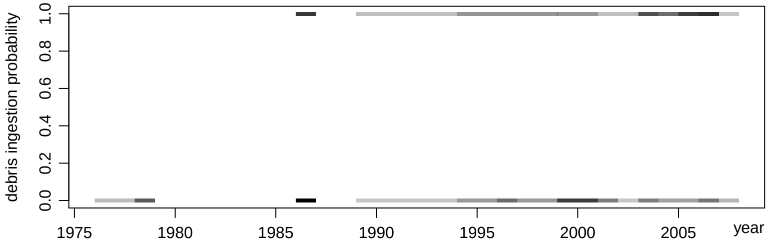
CFE-CMStatistics 2015, London, UK
13 December 2015

example: logistic regression with interval-censored covariate



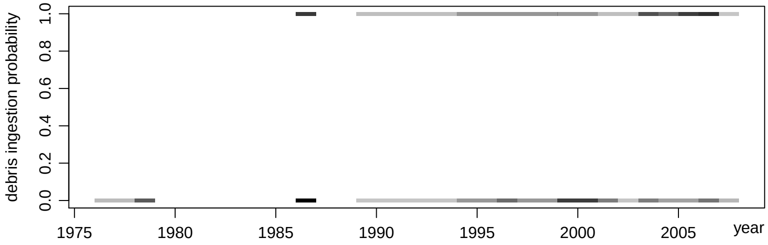
- for 468 green turtles, the data describe the presence or absence of marine debris in the gastrointestinal system at the time of death, which is interval-censored (Schuyler et al., 2014)

example: logistic regression with interval-censored covariate



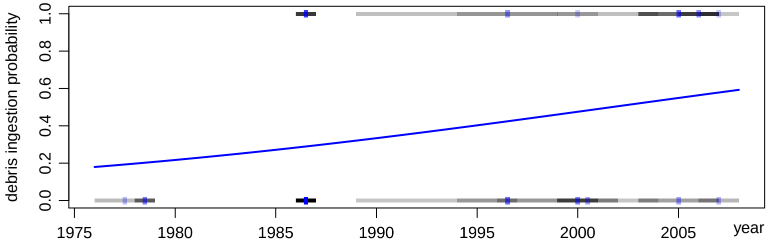
- for 468 green turtles, the data describe the presence or absence of marine debris in the gastrointestinal system at the time of death, which is interval-censored (Schuyler et al., 2014)
- the main question is if the probability of debris ingestion increased over time: we decide to use logistic regression to answer it

example: logistic regression with interval-censored covariate



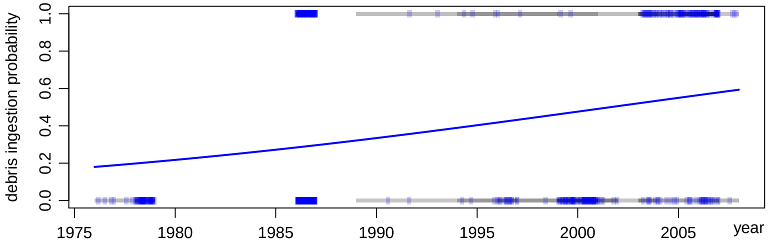
- how can we deal with incomplete/censored data?

example: logistic regression with interval-censored covariate



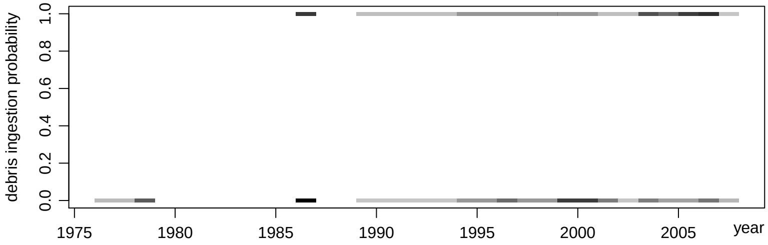
- how can we deal with incomplete/censored data?
  - delete them or make them precise (e.g., interval midpoints) and apply the conventional statistical method: easy, but what does the result mean?

example: logistic regression with interval-censored covariate



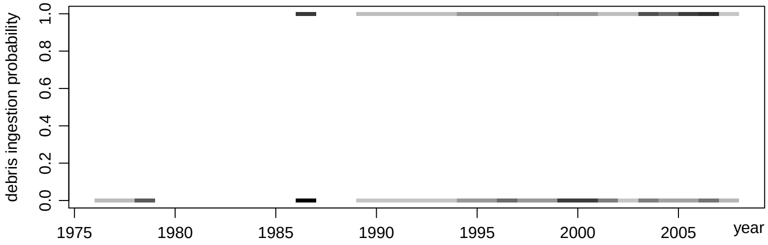
- how can we deal with incomplete/censored data?
  - delete them or make them precise (e.g., interval midpoints) and apply the conventional statistical method: easy, but what does the result mean?
  - impute the precise values and apply the conventional statistical method: the result depends on the imputation assumptions (e.g., uniform on intervals)

example: logistic regression with interval-censored covariate



- how can we deal with incomplete/censored data?
  - delete them or make them precise (e.g., interval midpoints) and apply the conventional statistical method: easy, but what does the result mean?
  - impute the precise values and apply the conventional statistical method: the result depends on the imputation assumptions (e.g., uniform on intervals)
  - apply the conventional statistical method to all compatible precise data sets: why? and how?

example: logistic regression with interval-censored covariate



- how can we deal with incomplete/censored data?
  - delete them or make them precise (e.g., interval midpoints) and apply the conventional statistical method: easy, but what does the result mean?
  - impute the precise values and apply the conventional statistical method: the result depends on the imputation assumptions (e.g., uniform on intervals)
  - apply the conventional statistical method to all compatible precise data sets: why? and how?
  - adapt the conventional statistical method to the case of incomplete/censored data: we will follow this approach

M-estimation: case with completely known data values

- data: $X_1, \ldots, X_n \in \mathcal{X}$ i.i.d.

M-estimation: case with completely known data values

- data: $X_1, \ldots, X_n \in \mathcal{X}$ i.i.d.
- M-estimator minimising the nonparametric MLE of $E[\rho(X_i, \theta)]$:

$$\hat{\theta}(X_1, \ldots, X_n) = \arg\min_{\theta} \sum_{i=1}^{n} \rho(X_i, \theta)$$

M-estimation: case with completely known data values

- data: $X_1, \ldots, X_n \in \mathcal{X}$ i.i.d.
- M-estimator minimising the nonparametric MLE of $E[\rho(X_i, \theta)]$:

$$\hat{\theta}(X_1, \ldots, X_n) = \arg\min_{\theta} \textstyle\sum_{i=1}^{n} \rho(X_i, \theta)$$

- nonparametric/pragmatic approach: under weak regularity conditions (Huber and Ronchetti, 2009),

$$\hat{\theta}(X_1, \ldots, X_n) \xrightarrow[n \to \infty]{\text{a.s.}} \arg\min_{\theta} E[\rho(X_i, \theta)]$$

M-estimation: case with completely known data values

- data: $X_1, \ldots, X_n \in \mathcal{X}$ i.i.d.
- M-estimator minimising the nonparametric MLE of $E[\rho(X_i, \theta)]$:

$$\hat{\theta}(X_1, \ldots, X_n) = \arg\min_{\theta} \sum_{i=1}^{n} \rho(X_i, \theta)$$

- nonparametric/pragmatic approach: under weak regularity conditions (Huber and Ronchetti, 2009),

$$\hat{\theta}(X_1, \ldots, X_n) \xrightarrow[n \to \infty]{\text{a.s.}} \arg\min_{\theta} E[\rho(X_i, \theta)]$$

- parametric/idealistic approach: assuming further $X_i \sim P_\theta$ and $E[\rho(X_i, \theta)] < E[\rho(X_i, \theta')]$,

$$\hat{\theta}(X_1, \ldots, X_n) \xrightarrow[n \to \infty]{\text{a.s.}} \theta$$

M-estimation: case with not completely known data values

- data: $S_1, \ldots, S_n \subseteq \mathcal{X}$ i.i.d., with $X_i \in S_i$ unknown

M-estimation: case with not completely known data values

- data: $S_1, \ldots, S_n \subseteq \mathcal{X}$ i.i.d., with $X_i \in S_i$ unknown
- M-estimator minimising the nonparametric MLE of $E[\rho(X_i, \theta)]$:

$$\hat{\theta}(S_1, \ldots, S_n) \stackrel{?}{=} \arg\min_{\theta} \operatorname{co}\left\{ \sum_{i=1}^{n} \rho(x_i, \theta) \;:\; x_i \in S_i \right\}$$

M-estimation: case with not completely known data values

- data: $S_1, \ldots, S_n \subseteq \mathcal{X}$ i.i.d., with $X_i \in S_i$ unknown
- M-estimator minimising the nonparametric MLE of $E[\rho(X_i, \theta)]$:

$$\hat{\theta}(S_1, \ldots, S_n) \stackrel{?}{=} \arg\min_{\theta} \operatorname{co}\left\{\textstyle\sum_{i=1}^{n} \rho(x_i, \theta) \ : \ x_i \in S_i\right\}$$

- nonparametric/pragmatic approach: under weak regularity conditions,

$$\hat{\theta}_{\text{minimax}}(S_1, \ldots, S_n) \xrightarrow[n \to \infty]{\text{a.s.}} \arg\min_{\theta} \overline{E}[\rho(X_i, \theta)],$$

  where $\overline{E}[\rho(X_i, \theta)]$ is the maximum/supremum expectation compatible with the distribution of S_i

M-estimation: case with not completely known data values

- data: $S_1, \ldots, S_n \subseteq \mathcal{X}$ i.i.d., with $X_i \in S_i$ unknown
- M-estimator minimising the nonparametric MLE of $E[\rho(X_i, \theta)]$:

$$\hat{\theta}(S_1, \ldots, S_n) \stackrel{?}{=} \arg\min_{\theta} \operatorname{co}\left\{\textstyle\sum_{i=1}^{n} \rho(x_i, \theta) \ : \ x_i \in S_i\right\}$$

- nonparametric/pragmatic approach: under weak regularity conditions,

$$\hat{\theta}_{\text{minimax}}(S_1, \ldots, S_n) \xrightarrow[n \to \infty]{\text{a.s.}} \arg\min_{\theta} \overline{E}[\rho(X_i, \theta)],$$

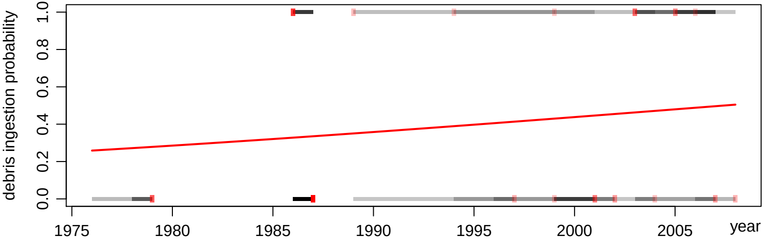
  where $\overline{E}[\rho(X_i, \theta)]$ is the maximum/supremum expectation compatible with the distribution of S_i

- parametric/idealistic approach: under additional assumptions,

$$\hat{\theta}_{\text{minimin}}(S_1, \ldots, S_n) \xrightarrow[n \to \infty]{\text{a.s.}} [\theta],$$

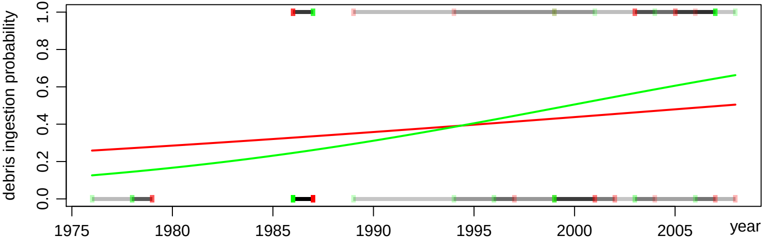
  where $[\theta]$ is the identification region of θ (Manski, 2003): smoothing corrections (of the ε-minimin form) may be needed in case of partial identification

example: logistic regression with interval-censored covariate



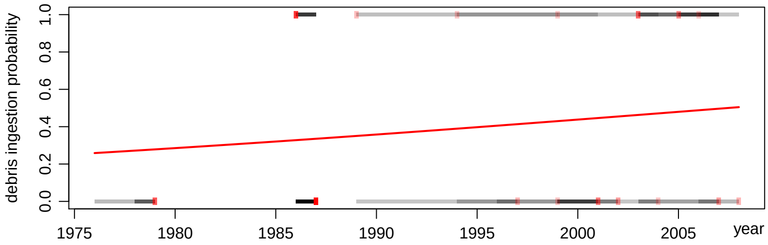
- $\hat{\theta}_{\text{minimax}}(s_1, \ldots, s_n) = \hat{\theta}(x_1, \ldots, x_n)$ when $x_i = \arg\max_{x_i \in s_i} \rho\big(x_i, \hat{\theta}(x_1, \ldots, x_n)\big)$

example: logistic regression with interval-censored covariate



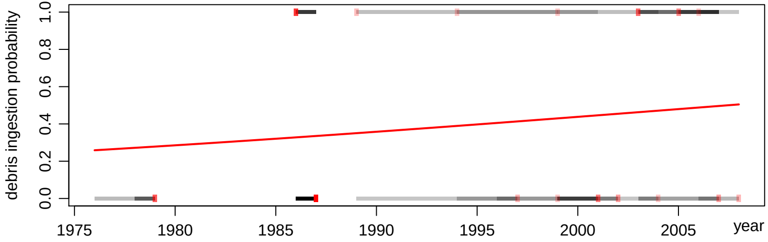
- $\hat{\theta}_{\text{minimax}}(s_1, \ldots, s_n) = \hat{\theta}(x_1, \ldots, x_n)$ when $x_i = \arg\max_{x_i \in s_i} \rho\big(x_i, \hat{\theta}(x_1, \ldots, x_n)\big)$
- the minimax logistic regression with one interval-censored covariate can always be obtained by computing at most two conventional logistic regressions (for the two extreme choices of compatible precise data sets)

example: logistic regression with interval-censored covariate



- $\hat{\theta}_{\text{minimax}}(s_1, \ldots, s_n) = \hat{\theta}(x_1, \ldots, x_n)$ when $x_i = \arg\max_{x_i \in s_i} \rho\big(x_i, \hat{\theta}(x_1, \ldots, x_n)\big)$
- the minimax logistic regression with one interval-censored covariate can always be obtained by computing at most two conventional logistic regressions (for the two extreme choices of compatible precise data sets)
- in the logistic regression with the (unknown) precise covariate, the increase over time of the debris ingestion probability is statistically significant (p-value < 0.001) according to the LR test, because the same is true in the conventional (minimax) logistic regression with worst-case precise data values

example: logistic regression with interval-censored covariate



- $\hat{\theta}_{\text{minimax}}(s_1, \ldots, s_n) = \hat{\theta}(x_1, \ldots, x_n)$ when $x_i = \arg\max_{x_i \in s_i} \rho\big(x_i, \, \hat{\theta}(x_1, \ldots, x_n)\big)$
- the minimax logistic regression with one interval-censored covariate can always be obtained by computing at most two conventional logistic regressions (for the two extreme choices of compatible precise data sets)
- in the logistic regression with the (unknown) precise covariate, the increase over time of the debris ingestion probability is statistically significant (p-value < 0.001) according to the LR test, because the same is true in the conventional (minimax) logistic regression with worst-case precise data values
- note however that this reasoning is not valid in general for other tests (e.g., the Wald test)

conclusions

- M-estimation can be adapted to the case of incomplete/censored data

conclusions

- M-estimation can be adapted to the case of incomplete/censored data
- in particular, the minimax M-estimation:

conclusions

- M-estimation can be adapted to the case of incomplete/censored data
- in particular, the minimax M-estimation:
  - generalises the nonparametric/pragmatic approach of finding the best fit to the available data, without assuming the existence of a true model

conclusions

- M-estimation can be adapted to the case of incomplete/censored data
- in particular, the minimax M-estimation:
  - generalises the nonparametric/pragmatic approach of finding the best fit to the available data, without assuming the existence of a true model
  - can often be implemented as a conventional M-estimation with worst-case precise data values, allowing also statistical inferences about the distribution of the true (but not completely known) precise data values

conclusions

- M-estimation can be adapted to the case of incomplete/censored data
- in particular, the minimax M-estimation:
  - generalises the nonparametric/pragmatic approach of finding the best fit to the available data, without assuming the existence of a true model
  - can often be implemented as a conventional M-estimation with worst-case precise data values, allowing also statistical inferences about the distribution of the true (but not completely known) precise data values
  - can be slightly generalised to include also, e.g., minimax Least Quantile of Squares regression (Cattaneo and Wiencierz, 2012, 2014), or minimax Support Vector Regression (Utkin and Coolen, 2011; Wiencierz and Cattaneo, 2015)

conclusions

- M-estimation can be adapted to the case of incomplete/censored data
- in particular, the minimax M-estimation:
  - generalises the nonparametric/pragmatic approach of finding the best fit to the available data, without assuming the existence of a true model
  - can often be implemented as a conventional M-estimation with worst-case precise data values, allowing also statistical inferences about the distribution of the true (but not completely known) precise data values
  - can be slightly generalised to include also, e.g., minimax Least Quantile of Squares regression (Cattaneo and Wiencierz, 2012, 2014), or minimax Support Vector Regression (Utkin and Coolen, 2011; Wiencierz and Cattaneo, 2015)
- by contrast, the parametric/idealistic approach of finding the hypothetical true model can be pursued by minimin M-estimation, but is often severely hindered by partial identification

references

Cattaneo, M., and Wiencierz, A. (2012). Likelihood-based Imprecise Regression. *Int. J. Approx. Reasoning* 53, 1137–1154.

Cattaneo, M., and Wiencierz, A. (2014). On the implementation of LIR: the case of simple linear regression with interval data. *Comput. Stat.* 29, 743–767.

Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd edn. Wiley.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer.

Schuyler, Q., Hardesty, B. D., Wilcox, C., and Townsend, K. (2014). Global analysis of anthropogenic debris ingestion by sea turtles. *Conserv. Biol.* 28, 129–139.

Utkin, L. V., and Coolen, F. P. A. (2011). Interval-valued regression and classification models in the framework of machine learning. In *ISIPTA '11*, eds. F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. SIPTA, 371–380.

Wiencierz, A., and Cattaneo, M. (2015). On the validity of minimin and minimax methods for Support Vector Regression with interval data. In *ISIPTA '15*, eds. T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur. Aracne, 325–332.