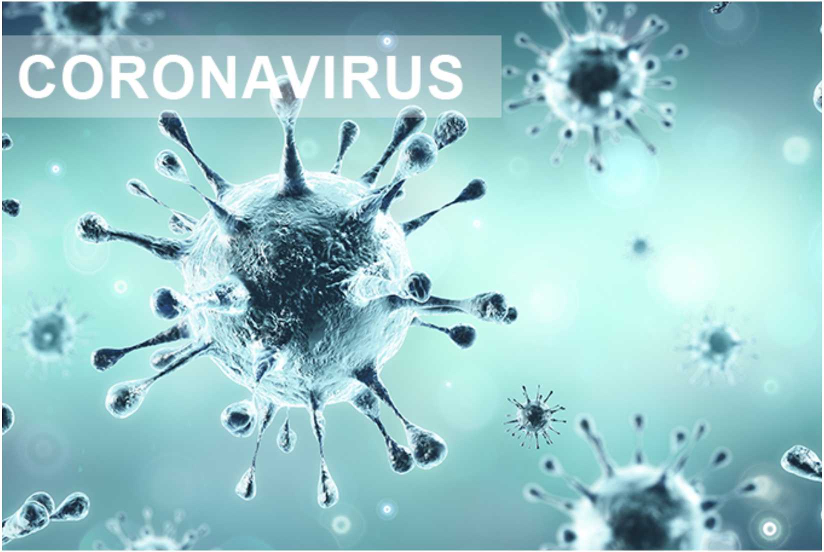
Online estimation methods for Covid-19 death rates using hospital data

Pierre-Yves Boëlle[1], Anna Bonnet[2], **Felix Cheysson**[2], Charlotte Dion[2], Olivier Lopez[2], Maud Thomas[2]

[1] Institut Pierre-Louis d'Epidémiologie et de Santé Publique
[2] Sorbonne Université, LPSM

Séminaire de Modélisation Aléatoire du Vivant
Mercredi 3 mars 2021

Context



Some facts about the first wave

Overview of the pathway for hospitalised Covid-19 patients (Courtejoie and Dubost, 2020):

- SI-VIC database (*système d'information pour le suivi des victimes*) to monitor hospital admissions in the event of exceptional sanitary situations.
- Overall mortality rate: 19%; halved between early March and mid June.
- 17% for women, 21% for men; 2% for < 40 y.o., 33% for > 80 y.o.
- Median age for deceased individuals: 81 years.

Modelling motivation

Covid-19 Dataset from SI-VIC database: all hospitalisation for Covid-19 patients in AP-HP hospitals.

dt.first	dt.last	outcome	sex	age	hospital
2020-03-17	2020-04-05	rad	F	45	Robert Debré
2020-03-14	2020-03-25	rad	F	29	Robert Debré
2020-03-18	2020-03-29	dc	H	80	St Antoine
2020-03-11	2020-03-15	dc	H	62	St Louis
2020-03-04	2020-03-09	dc	F	72	Pitié Salpétrière
2020-03-16	2020-03-20	dc	H	92	Raymond Poincaré

Motivation: We wish to model the risk of death of a patient hospitalised for Covid-19, with respect to covariates, in an online framework.

Offline *vs* online framework

Machine learning terminology.

- **Offline** (or *batch learning*): Build a model from the whole dataset.
- **Online**: Train the model as the data comes in.
  - Learn trends in real-time: adapt on-the-fly to new data.
  - Time constraints: no need to re-run the whole algorithm, past observations can be discarded.

A carousel of estimation methods

1. Logistic regression
2. Survival analysis
3. Non-parametric estimation
4. Perspectives: maximum weighted likelihood estimation

A carousel of estimation methods

1. Logistic regression
2. Survival analysis
3. Non-parametric estimation
4. Perspectives: maximum weighted likelihood estimation

Logistic regression

For individual i, let E_i denote the date of admission and U_i the outcome of hospitalisation:

$$
\begin{aligned}
U_i &= 1, && \text{if the individual } i \text{ dies,}\\
U_i &= 0, && \text{if the individual } i \text{ lives.}
\end{aligned}
$$

From an *i.i.d.* sample $\big((e_1, u_1), \ldots, (e_n, u_n)\big)$, we wish to explain the risk of death as a function of the date of admission of the individual:

$$p_i = \mathbb{P}(U_i = 1 | E_i = e_i).$$

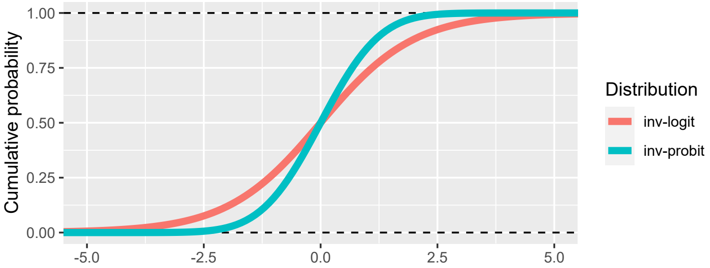
We model the outcome with a logistic regression:

$$
\begin{aligned}
U_i &\overset{\textit{ind.}}{\sim} B(p_i),\\
g(p_i) &= \beta_0 + \beta_1 e_i.
\end{aligned}
$$

Choice of the link function g

- g must be chosen as a map from $(0, 1)$ to $\mathbb{R}$.
- Two usual choices:
  - The *probit* function: $\mathrm{probit}(p_i) = \Phi^{-1}(p_i)$, where $\Phi(x)$ is the CDF of the normal distribution.
  - The *logit* function: $\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.
- The logit function can be easily interpreted in terms of *odds-ratio*:

$$\mathrm{logit}(p_1) - \mathrm{logit}(p_2) = \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right).$$



Maximum likelihood estimation

- Suppose that the data $(U_1, \ldots, U_n)$ is generated from distribution $f_{\theta_0}(y)$ with true parameter θ_0.
- The log-likelihood of the model is written $l_n(\theta) = \frac{1}{n} \sum_{i=1}^n l(\theta; U_i)$.
- For the **logistic regression**,

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^n U_i \log p_i + (1 - U_i) \log(1 - p_i).$$

- Define $\hat{\theta}_n$ as the maximum likelihood estimator of θ.

Theorem: Under regularity conditions, $\hat{\theta}_n$ is consistent, *i.e.* $\hat{\theta}_n \xrightarrow{P} \theta_0$, and is asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right),$$

where $I(\theta_0) = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial}{\partial \theta} \log f_\theta(U) \mid_{\theta=\theta_0}\right)^2\right]$ is called the Fisher information.

Parametric inference

Asymptotic confidence interval

Using the asymptotic normality of the MLE $\hat{\theta}_n$, we build approximate confidence interval for θ_0 for n large:

$$\mathrm{IC}_{1-\alpha}(\theta_0) \approx \left[\hat{\theta}_n + \frac{u_{\alpha/2}}{\sqrt{nI(\theta_0)}}; \hat{\theta}_n + \frac{u_{1-\alpha/2}}{\sqrt{nI(\theta_0)}}\right],$$

with u_a the quantile of order a of the normal distribution.

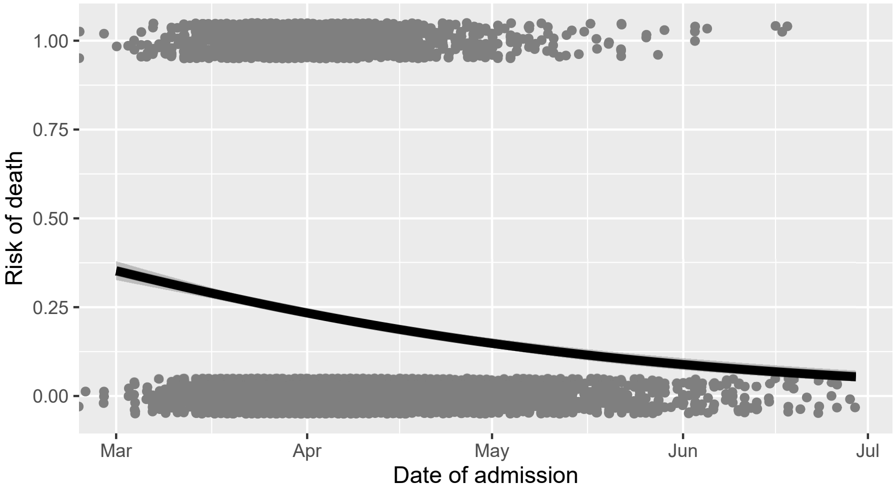
Using R, we find:

$$\mathrm{IC}_{95\%}(\beta_1) = [-0.022; -0.015],$$

or, as an odds-ratio:

$$\mathrm{IC}_{95\%}\left(e^{\beta_1}\right) = [0.978; 0.985].$$

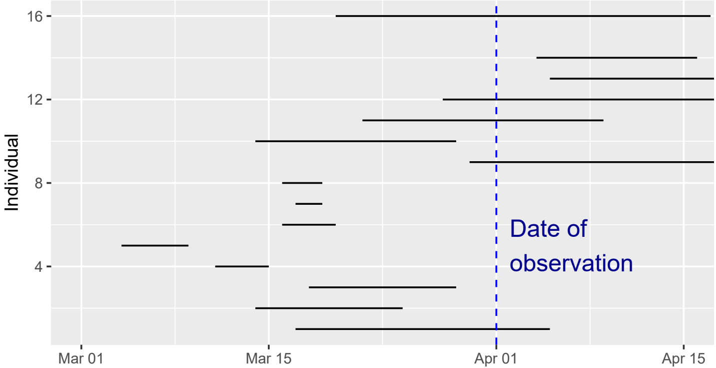
Predicting the risk of death



A carousel of estimation methods

Censored data

- ▶ Analysis of data on times of events in individual life-histories.
- ▷ How to deal with censored data?
- ▷ Modelling events continuously in time, conditioning on past events.
- ▷ *Hazard rate* (and *product-integration*).



Survival analysis

- Survival function and measure:

$$S(t) = \mathbb{P}(T > t), \quad \text{and} \quad S(s,t) = \frac{S(t)}{S(s)}.$$

- Cumulative hazard function and measure:

$$\Lambda(t) = \int_0^t \frac{F(\mathrm{d}s)}{S(s-)}, \quad \text{and} \quad \Lambda(s,u) = \Lambda(s,t) + \Lambda(t,u).$$

- Intuitively:

$$\Lambda(\mathrm{d}t) = \mathbb{P}(T \in \mathrm{d}t \mid T \geq t) = 1 - S(\mathrm{d}t),$$
$$S(\mathrm{d}t) = \mathbb{P}(T \notin \mathrm{d}t \mid T \geq t) = 1 - \Lambda(\mathrm{d}t).$$

This provides the dual relationship:

$$\Lambda(t) = \int_{(0,t]} (1 - S(\mathrm{d}s)), \quad \text{and} \quad S(t) = \prod_{(0,t]} (1 - \Lambda(\mathrm{d}s)).$$

Random censorship model

- Unobservable positive random variables,

$$T_1, \ldots, T_n \sim_{i.i.d.} F; \text{independent of}$$
$$C_1, \ldots, C_n \sim_{i.i.d.} G.$$

- What we observe, for $1 \leq i \leq n$,

$$Y_i = \min(T_i, C_i), \quad \text{and} \quad \delta_i = \mathbb{1}\{T_i \leq C_i\}.$$

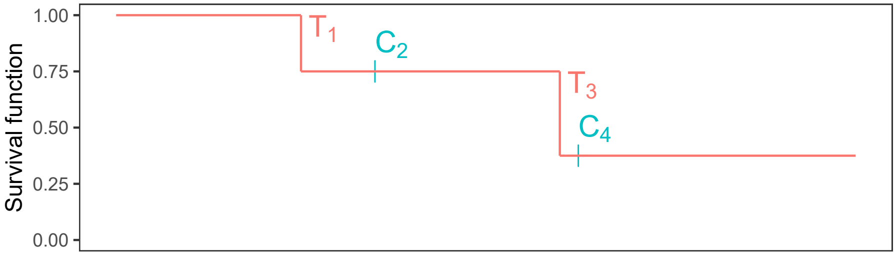
- Nelson-Aalen estimator (Nelson, 1969; Altshuler, 1970; Aalen, 1978):

$$\hat{\Lambda}(\mathrm{d}t) = \frac{\#\{i : Y_i \in \mathrm{d}t, \delta_i = 1\}}{\#\{i : Y_i \geq t\}}$$

then $\hat{\Lambda}(t) = \int_0^t \hat{\Lambda}(\mathrm{d}s)$ and $\hat{S}(t) = \prod_0^t (1 - \hat{\Lambda}(\mathrm{d}s))$.

Kaplan-Meïer estimator (Kaplan and Meier, 1958)

$$1 - \hat{F}_n(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i:n]}}{n-i+1}\right)^{\mathbb{1}\{Y_{i:n} \leq x\}}$$



Methods for studying the Kaplan-Meïer estimator:

- as an empirical process (*e.g.* Donsker theorem);
- through martingale methods (*e.g.* Glivenko-Cantelli theorem);
- (Gill, 1993; Stute, 1995).

IPCW estimator (Satten and Datta, 2001)

Define the Inverse-Probability-of-Censoring Weighted estimator of F by weighing the e.c.d.f. by the inverse of the probability that the failure time T is observed:

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}\{t_i \leq t\}\delta_i}{1 - \hat{G}(t_i)},$$

$$\hat{G}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}\{t_i \leq t\}\bar{\delta}_i}{1 - \hat{F}(t_i)}.$$

Back to Covid-19

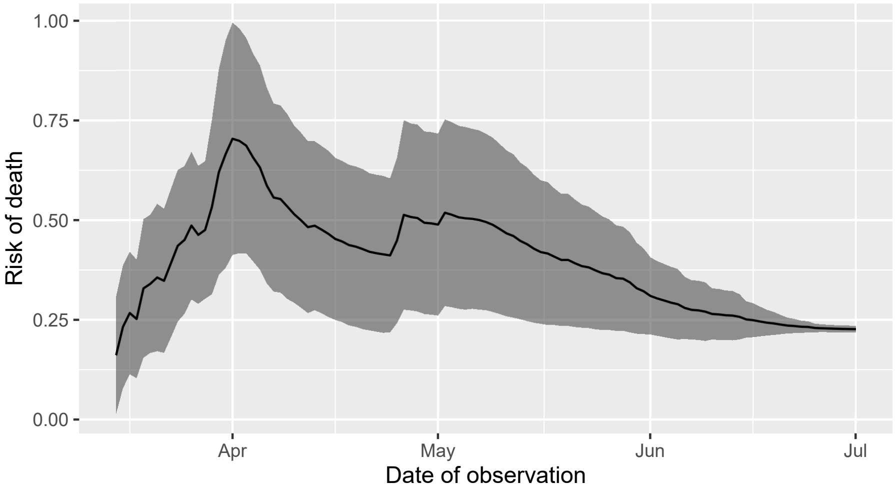
- Recall that we are interesting in estimating $p = \mathbb{P}(U = 1)$, with (T_i, C_i, U_i) i.i.d. and $T_i \perp\!\!\!\perp C_i$.
- Since

$$\mathbb{E}\left[\frac{\delta_1 U_1}{1 - G(Y_1-)}\right] = \mathbb{E}\left[\frac{U_1}{1 - G(T_1-)}\mathbb{E}[\mathbb{1}\{T_1 \leq C_1\} \mid T_1, U_1]\right]$$
$$= \mathbb{E}\left[\frac{U_1}{1 - G(T_1-)}(1 - G(T_1-))\right]$$
$$= \mathbb{P}(U_1 = 1).$$

- Define, for a given date of observation:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i U_i}{1 - \hat{G}(Y_i-)}.$$

Online estimation of risk of death



A carousel of estimation methods

Kernel density estimation

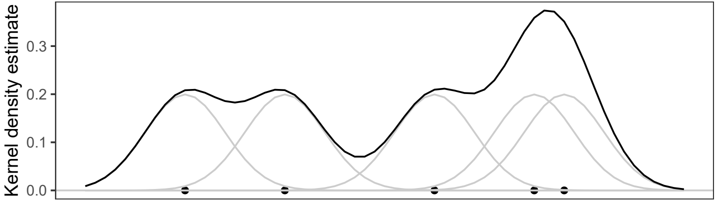
- Introduced by (Rosenblatt, 1956) to extend the histogram.
- For a general kernel (positive, symmetric, integrates to 1) and a bandwidth parameter h, define:

$$K_h(x) = \frac{1}{h} K\Big(\frac{x}{h}\Big).$$

- Kernel density estimator:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big).$$

- Trade-off in convergence speed: bias $\mathcal{O}(h^2)$ *vs.* variance $\mathcal{O}(1/\sqrt{nh})$.



Regression smoothing

- From the i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$, estimate the non-parametric regression model:

$$Y_i = m(X_i) + \varepsilon_i,$$

  where $m(x) = \mathbb{E}[Y \mid X = x]$.
- The Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964):

$$\hat{m}_h(x) = n^{-1} \frac{\sum_{k=1}^n K_h(x - X_i) Y_i}{n^{-1} \sum_{k=1}^n K_h(x - X_i)}.$$

- Same bias-variance trade-off: $\mathcal{O}(h^2) + \mathcal{O}(1/\sqrt{nh})$.
- (Härdle, 1991).

Choice of bandwidth parameter h

- For optimal speed of convergence, choose $h \sim n^{-1/5}$, then $MSE(\hat{m}_h(x)) = \mathcal{O}(n^{-4/5})$.
- In practice,

$$Var(\hat{m}_h(x)) \propto K, f(x), \sigma^2(x),$$
$$Bias^2(\hat{m}_h(x)) \propto K, m''(x), m'(x), f'(x), f(x).$$

- Idea: find the bandwidth h that minimises *a* distance between the unknown curve m and the estimator $\hat{m}_h$.
- Cross-validation score:

$$CV(h) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{m}_{h,i}(X_i))^2 w(X_i),$$

where $w(\cdot)$ allows to drop observations at the boundary of X.

- A sequence of bandwidths based on $CV(h)$ is asymptotically optimal.

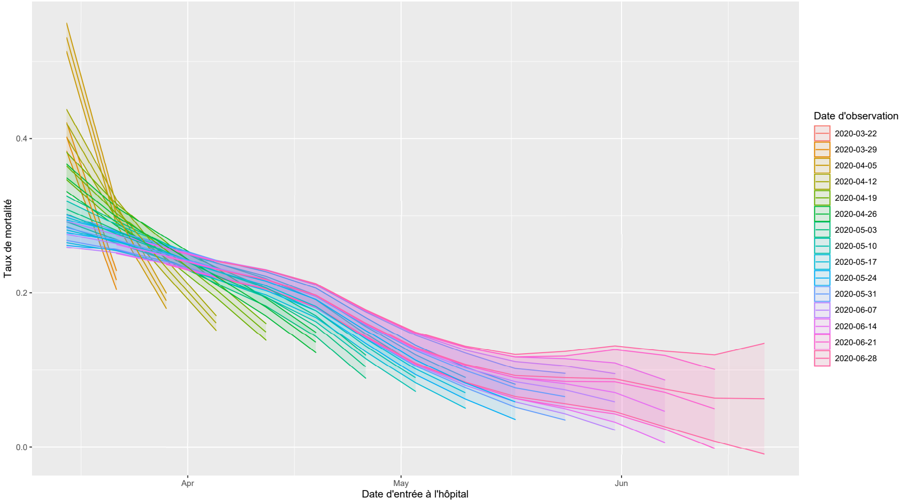
Back to Covid-19

- Recall that we are interesting in estimating $p = \mathbb{P}(U = 1)$, with (T_i, C_i, U_i) i.i.d. and $T_i \perp\!\!\!\perp C_i$.
- In presence of censoring, define the following Nadaraya-Watson estimator:

$$\hat{p}_h(e) = n^{-1} \frac{\sum_{k=1}^{n} K_h(e - E_i)\delta_i U_i / (1 - \hat{G}(Y_i -))}{n^{-1} \sum_{k=1}^{n} K_h(e - E_i)}.$$

- Studied partially by (Guessoum and Ould-Said, 2009).

Online non-parametric estimation of risk of death



A carousel of estimation methods

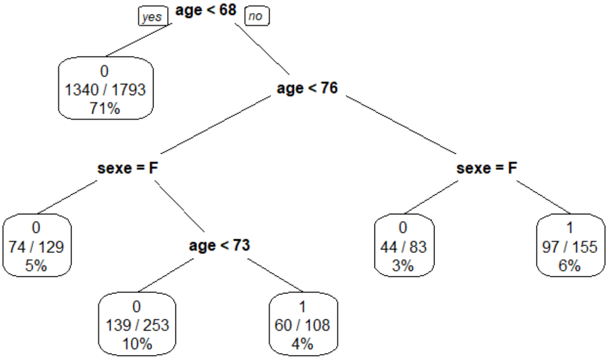
▷ The Nadaraya-Watson is a method of moment estimator: we ought to do better with methods based on the likelihood.
▷ We would like to add more covariates to estimation: curse of dimensionality for non-parametric estimation.
▷ The full likelihood for the random censorship model is not straightforward to work with.
▷ Define as an estimator of p the value that maximises

$$\hat{l}_n(\theta; e) = \frac{1}{n} \sum_{i=1}^{n} K_h(e - E_i) \frac{\delta_i}{1 - \hat{G}(Y_i-)} l(\theta; U_i),$$

where $l(\theta; U_i) = U_i \log p + (1 - U_i) \log(1 - p)$.
▷ Idea: show that $\hat{l}_n(\theta; e)$ is a consistent estimator of $\mathbb{E}[l(\theta; U) \mid E = e]$.

Projected advantages of this approach

- Easy to implement, by weighting the input of existing maximum likelihood estimation algorithms.
- Able to extend the likelihood by including covariates in the model.
- Determine categories of individuals at risk by integrating decision trees into the estimation method.



For Further Reading I

- Aalen, Odd (1978). "Nonparametric inference for a family of counting processes". In: *The Annals of Statistics*, pp. 701–726.
- Altshuler, Bernard (1970). "Theory for the measurement of competing risks in animal experiments". In: *Mathematical Biosciences* 6, pp. 1–11.
- Courtejoie, Noémie and Claire-Lise Dubost (2020). *Parcours hospitalier des patients atteints de la Covid-19 lors de la première vague de l'épidémie*. Tech. rep. 67. Les dossiers de la DREES.
- Gill, Richard D (1993). *Lectures on Survival Analysis*. Ed. by Pierre Bernard. Lectures on Probability Theory. Springer-Verlag Berlin Heidelberg, pp. 1–127. doi: `10.1007/BFb0073871`.
- Guessoum, Zohra and Elias Ould-Said (2009). "On nonparametric estimation of the regression function under random censorship model". In: *Stat. Decis.* 26.3, pp. 159–177. issn: 0721-2631. doi: `10.1524/stnd.2008.0919`.

For Further Reading II

- Härdle, Wolfgang (1991). *Smoothing techniques: with implementation in S*. New York: Springer-Verlag. isbn: 9781461287681.
- Kaplan, E L and Paul Meier (1958). "Nonparametric estimation from incomplete samples". In: *J. Am. Stat. Assoc.* 53.282, pp. 457–481.
- Nadaraya, Elizbar A (1964). "On estimating regression". In: *Theory of Probability & Its Applications* 9.1, pp. 141–142.
- Nelson, Wayne (1969). "Hazard plotting for incomplete failure data". In: *Journal of Quality Technology* 1.1, pp. 27–52.
- Rosenblatt, Murray (1956). "Remarks on Some Nonparametric Estimates of a Density Function". In: *The Annals of Mathematical Statistics* 27.3, pp. 832 –837. doi: 10.1214/aoms/1177728190.
- Satten, Glen A and Somnath Datta (2001). "The Kaplan-Meier Estimator as an Weighted Average". In: *Am. Stat.* 55.3, pp. 207–210. doi: 10.1198/000313001317098185.The.

For Further Reading III

Stute, Winfried (1995). "The statistical analysis of Kaplan-Meier integrals". In: *Lect. Notes-Monograph Ser.* Vol. 27, pp. 231–254. isbn: 9780444500793. doi: 10.1214/lnms/1215452223.

Watson, Geoffrey S (1964). "Smooth regression analysis". In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372.

Some notations

- Suppose that the data $(Y_1, \dots, Y_n)$ is generated from distribution $f_{\theta_0}(y)$ with true parameter θ_0.
- The log-likelihood of the model is written

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_i(\theta).$$

- For the **logistic regression**,

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} Y_i \log p_i + (1 - Y_i) \log(1 - p_i).$$

- Define $\hat{\theta}_n$ as the maximum likelihood estimator of θ.

Consistency of the likelihood

Define, for $l_1(\theta) = \log f_\theta(Y_1)$:

$$l(\theta) = \mathbb{E}_{\theta_0}[l_1(\theta)] = \int (\log f_\theta(y)) f_{\theta_0}(y)\mathrm{d}y.$$

Lemma: For any θ,

$$l(\theta) \leq l(\theta_0).$$

If the model is identifiable, then the inequality is strict for $\theta \neq \theta_0$.

Idea of the proof: Remark that the difference

$$l(\theta_0) - l(\theta) = \mathbb{E}_{\theta_0} \log \frac{f_{\theta_0}(Y)}{f_\theta(Y)}$$

is a Kullback-Leibler divergence. Show that it is non-negative (*e.g.* using Jensen's inequality).

Consistency of the likelihood (cont'd)

Define, for $l_1(\theta) = \log f_\theta(Y_1)$:

$$l(\theta) = \mathbb{E}_{\theta_0}[l_1(\theta)] = \int (\log f_\theta(y)) f_{\theta_0}(y) \mathrm{d}y.$$

Theorem: If $l_n(\theta)$ is continuous and has a unique maximum, then $\hat{\theta}_n$ is consistent, *i.e.* $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Idea of the proof: We have the following assertions:

- $\hat{\theta}_n$ is the maximiser of $l_n(\theta)$ (by definition);
- θ_0 is the maximiser of $l(\theta)$ (by lemma);
- $\forall \theta$, $l_n(\theta) \xrightarrow{P} l(\theta)$ (by WLLN).

Fisher information

Define, for a log-likelihood $l(\theta) = f_\theta(y)$, the **Fisher information** function by

$$I(\theta) = \mathbb{E}_\theta\left[(l'(\theta))^2\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(Y)\right)^2\right].$$

Lemma: We have the following:

$$I(\theta) = \mathrm{Var}_\theta\big(l'(\theta)\big), \qquad \text{and } I(\theta) = -\mathbb{E}_\theta\left[l''(\theta)\right].$$

Idea of the proof: We have, by swapping the derivative and the integral:

$$\int \frac{\partial}{\partial\theta} f_\theta(y)\mathrm{d}y = \frac{\partial}{\partial\theta}\int f_\theta(y)\mathrm{d}y = 0,$$

and

$$\int \frac{\partial^2}{\partial^2\theta} f_\theta(y)\mathrm{d}y = \frac{\partial^2}{\partial^2\theta}\int f_\theta(y)\mathrm{d}y = 0.$$

Asymptotic normality

Theorem: Under regularity conditions, we have that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right).$$

Idea of the proof: A Taylor expansion of $l_n'(\hat{\theta}_n)$ around θ_0 gives:

$$0 = l_n'(\hat{\theta}_n) = l_n'(\theta_0) + (\hat{\theta}_n - \theta_0) l_n''(\theta_n^*),$$

for some θ_n^* between θ_0 and $\hat{\theta}_n$.
Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\sqrt{n}\, l_n'(\theta_0)}{l_n''(\theta_n^*)}.$$

Asymptotic normality (cont'd)

For the numerator:

$$\begin{aligned}\sqrt{n}\, l_n'(\theta_0) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n l_i'(\theta_0) - 0\right)\\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n l_i'(\theta_0) - \mathbb{E}_{\theta_0} l_1'(\theta_0)\right)\\ &\to \mathcal{N}\left(0, \mathrm{Var}_{\theta_0}(l_1'(\theta_0)) = I(\theta_0)\right), \qquad \text{by CLT.}\end{aligned}$$

For the denominator:

- For all θ, $l_n''(\theta) \xrightarrow{P} \mathbb{E}_{\theta_0} l_1''(\theta)$ (by WLLN);
- Since $\theta_n^* \in [\theta_0, \hat{\theta}_n]$ and $\hat{\theta}_n \xrightarrow{P} \theta_0$ (by consistency), we have $\theta_n^* \xrightarrow{P} \theta_0$;
- Therefore $l_n''(\theta_n^*) \xrightarrow{P} \mathbb{E}_{\theta_0} l_1''(\theta_0) = -I(\theta_0)$.