Analyzing Large Sets of Spatial Binary Data Using a GEE - like Approach, and Some Extensions

Samuel D. Oman
Dept of Statistics, Hebrew University of Jerusalem

Outline

Part 1

- Analyzing Spatially Distributed Binary Data Using Independent-Block Estimating Equations (2007, *Biometrics*)

- Joint with

  Victoria Landsman, Dept of Statistics, Hebrew Univ

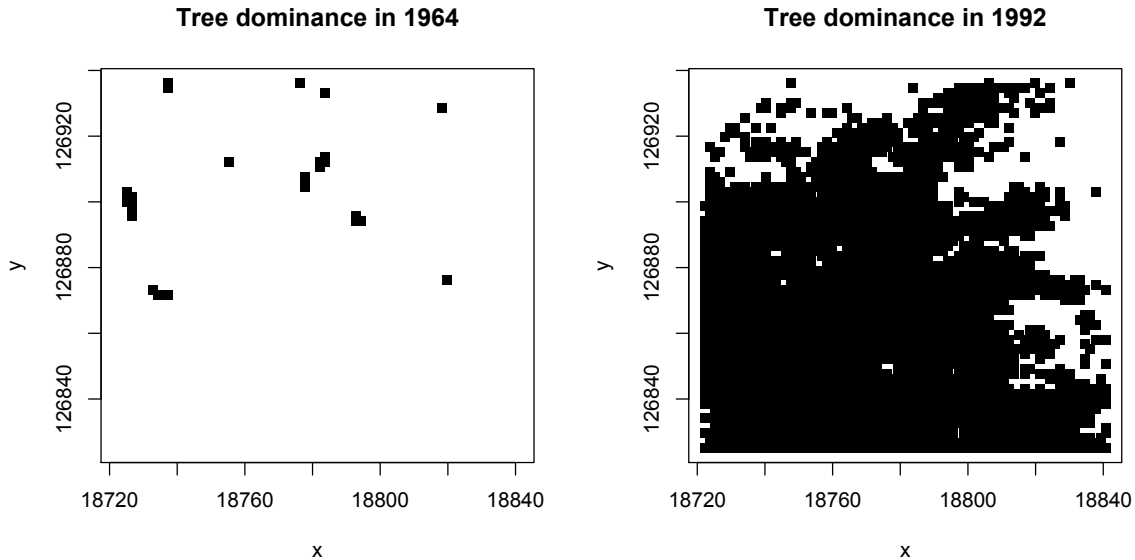
  Yohay Carmel, Dept of Evolution, Systematics and Ecology, Hebrew Univ

  Ronen Kadmon, Dept of Evolution, Systematics and Ecology, Hebrew Univ

Part 2

- Some possible extensions / open questions

The Problem



- Research area: 1,125 $\times$ 1,200 m^2 region of natural forest in Mt. Meron
- $75 \times 80 = 6{,}000$ pixels (sites) of size $15m \times 15m$
- Black: trees (woody vegetation more than $2.5m$ high) are the predominant growth form at site
- **Goal**: identifying factors affecting changes in tree cover

Variables

- Spatially dependent binary responses

$$Y_i = \begin{cases} 1 & \text{if trees dominant in 1992 at site } s_i \\ 0 & \text{if not} \end{cases}$$

$i = 1, \ldots, 6,000$

- x_i = vector of covariates:
  - Tree dominance in 1964
  - Environmental factors - height above sea level, aspect & slope (which affect solar radiation and exposure to rain), distance to nearest river bed
  - Anthropogenic factors - grazing type and intensity, distance to nearest road

Remarks

- Original data: $Y = (\text{tree.prop}, \text{shrub.prop}, \text{herb.prop})$

- Yohay C, Ronen K (1999, *Plant Ecology*): regression of each proportion, using sites at least 5 cells (75 m) apart, with $x, y, xy, \ldots$ as explanatory variables; cross-validation on sites not used

- With Ronit Nirel, Heb U Stats Consulting Service (2001, *Ecological Applications*): transform Y to ordinal variable; compare (polychotomous logistic) regression results using subset and full data set

- $\left.\begin{array}{l}\text{Ronit: interested?}\\ \text{Viki: topic for MA thesis?}\end{array}\right\}$ Me: Yes !

- Viki: use Heagerty and Lele (1998), "Marginal regression models for clustered ordinal measurements", *JASA*

Model: Hierarchical GLIM with Probit Link

- $Z^* = (Z_i^*)$: latent, spatially dependent Gaussian field
- $\{Y_i\}|Z^* \sim \text{Ber}(p_i^*)$ independently,

  $$p_i^* = \Phi(x_i^t\gamma + Z_i^*) = P(Z_i \leq x_i^t\gamma + Z_i^*)$$

- Integrating out Z^*:

  $$p_i \equiv P(Y_i = 1) = P(Z_i \leq \eta_i), \qquad \eta_i = x_i^t\beta$$
  $$\pi_{ij} \equiv P(Y_i = 1, Y_j = 1) = P(Z_i \leq \eta_i, Z_j \leq \eta_j)$$

  where

  $$Z_i \sim N(0,1)$$
  $$\text{cor}(Z_i, Z_j) = \sigma^2\rho^{\|s_i - s_j\|} \qquad (\sigma^2 \leq 1 \text{ for possible nugget effect})$$
  $$\|s_i - s_j\| = \text{distance between sites } s_i \text{ and } s_j$$

- Note that β represents *global* effects

Estimation: Composite Likelihood

- $\theta = (\beta^t, \rho, \sigma^2)^t$: parameter vector
- loglik $\approx \sum_{\|s_i - s_j\| \le d} \log p(y_i, y_j) \equiv \ell$
- $\nabla \ell \equiv U = \sum_{\|s_i - s_j\| \le d} U_{ij}(\theta)$, relatively simple U_{ij}
- $U(\widehat{\theta}) = 0 \Rightarrow \widehat{\theta} \approx N(\theta, V)$, $\quad V = H^{-1}\text{cov}(U)H^{-t} \quad (H = E[(\partial/\partial\theta)U])$

  (assuming regularity; based on Guyon 1995, Lindsay 1988, White 1982)
- "Sandwich Estimator" of V: use window sub-sampling (Sherman, 1996):

  assuming $\quad | R | \, \text{cov}(\mathbf{U}_R) \to \mathbf{\Sigma} \quad$ as region R expands,

  $\widehat{\text{cov}}(U_R) = J^{-1} \sum_j \frac{|R|}{|R_j|} [U_{R_j}(\widehat{\theta})][U_{R_j}(\widehat{\theta})]^t$ for J subwindows $\{R_j\}$ of R
- Like GEE2

BUT it took a lonnnnnng time

- At each Newton - Raphson iteration, computing

  $U = \sum_{\|s_i - s_j\| < d} U_{ij}(\theta)$ and ∇U

  required $\approx 10^6$ evaluations of

  $\pi_{ij} = P(Z_i \leq \eta_i, Z_j \leq \eta_j)$

  for the correlated Normal Z_i and Z_j.

- S-plus functions $\rightarrow$ C-plus

- Me: what about aspect $\times$ slope interaction?

  Viki: tell you tomorrow

- Viki still managed to finish her thesis

Alternative: **I**ndependent **B**lock **E**stimating **E**quations (**IBEE**)

- Let $Y_{N\times 1} = (Y_i)$, $\quad p = (p_i)$, $\quad V = \text{cov}(Y)$
- Want $\widehat{\beta} : D^t V^{-1}[Y - p(\beta)] = 0$ where $D = (\partial p)/(\partial \beta)$, but $N =$ 6,000
- Instead:

  $V \rightarrow$ "Working" covariance matrix $W = \text{diag}(W_{(i)})$ (independent blocks)

$$\begin{aligned} \text{cov}(Y_j, Y_k) &= \int_{-\infty}^{\eta_k}\int_{-\infty}^{\eta_j} \phi(x, y | r(\|s_j - s_k\|))dxdy - p_j p_k \\ &\approx \phi(\eta_j)\phi(\eta_k)\arcsin(\sigma^2 \rho^{\|s_j - s_k\|}) \qquad \text{(Pearson, 1901)} \end{aligned}$$

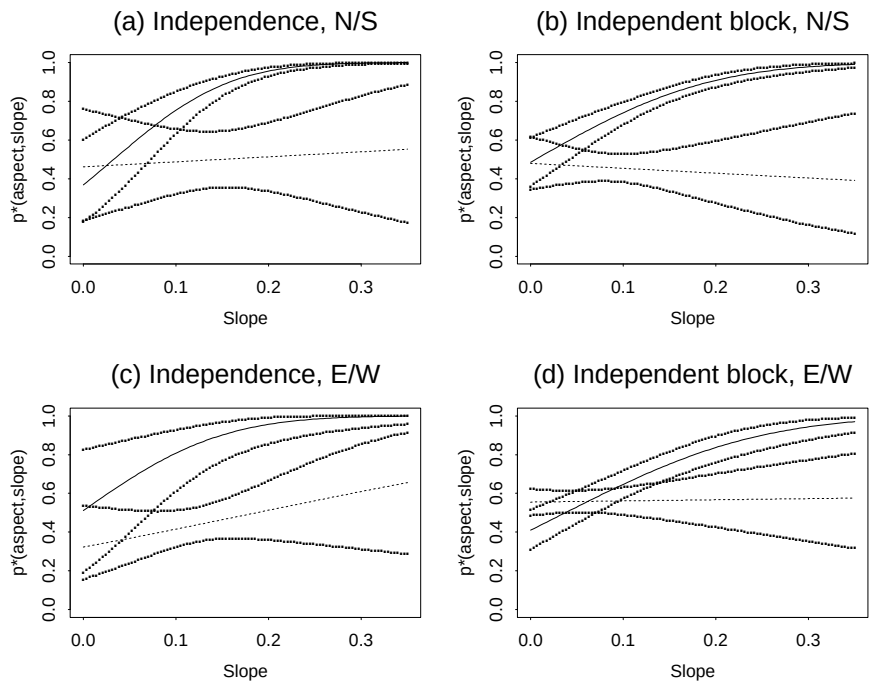
- $\widehat{\sigma}^2, \widehat{\rho}$ from separate estimating equation based on $\{(Y_j - Y_k)^2\}$
- Result: $\widehat{\beta}$ as solution to $U(\beta) \equiv \sum_{\text{blocks } i} G_{(i)}(Y_{(i)} - p_{(i)}) = 0$
  - Estimation of cov($\widehat{\beta}$), asymptotics: as in Haegerty and Lele (1998).
  - Like IEE, but with independent *blocks*.

Estimates (75×80 lattice $\rightarrow 15 \times 16$ - sized blocks; took several minutes)

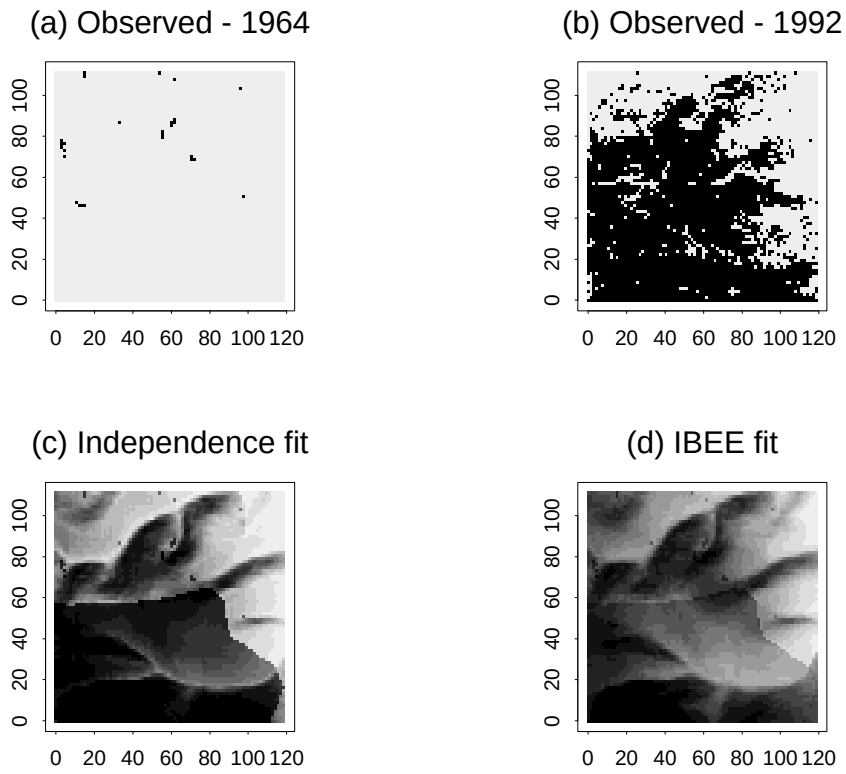
	Coefficient estimate			Standard error			p-value		
Variable	Ind	Cl	Ibee	Ind	Cl	Ibee	Ind	Cl	Ibee
tree.64	1.16	1.35	0.55	0.27	0.27	0.27	0.00	0.00	0.05
alt	2.12	2.20	1.65	0.53	0.38	0.40	0.00	0.00	0.00
str.dist	-2.44	-3.16	-3.18	1.76	1.60	1.62	0.16	0.05	0.05
rd.dist	0.96	0.85	0.48	0.60	0.49	0.44	0.11	0.09	0.27
g.spor	-0.57	-0.61	-0.06	0.27	0.29	0.43	0.04	0.04	0.89
g.mod	-1.00	-0.96	-0.77	0.32	0.33	0.24	0.00	0.00	0.00
c.spor	0.53	0.54	-0.23	0.27	0.28	0.19	0.05	0.05	0.23
c.mod	-0.94	-0.92	-0.61	0.28	0.25	0.25	0.00	0.00	0.02
c.heavy	-1.66	-1.70	-0.79	0.44	0.39	0.23	0.00	0.00	0.00
ns.aspct	0.48	0.20	0.00	1.05	1.05	0.61	0.65	0.85	1.00
ew.aspct	0.98	0.99	-0.71	1.16	1.08	0.33	0.40	0.36	0.03
slope	4.27	2.91	0.94	4.17	4.16	1.85	0.31	0.48	0.61
ns.slope	-19.10	-17.65	-14.87	6.33	6.65	4.82	0.00	0.00	0.00
ew.slope	11.87	12.67	11.72	8.02	7.89	2.90	0.14	0.11	0.00

- In general, $\widehat{\beta}_i^{\text{ind}} \approx \widehat{\beta}_i^{\text{cl}} \neq \widehat{\beta}_i^{\text{ibee}}$
- $\text{avg}\{\text{s.e.}(\widehat{\beta}_i^{\text{cl}})/\text{s.e.}(\widehat{\beta}_i^{\text{ind}})\} = 0.95$; $\text{avg}\{\text{s.e.}(\widehat{\beta}_i^{\text{ibee}})/\text{s.e.}(\widehat{\beta}_i^{\text{ind}})\} = 0.73$

Estimates and confidence envelopes for $P(Y = 1 \mid \text{aspect, slope, } x_{\text{others}} = \text{mean})$



- N/S: Upper curves for N aspect, lower for S.
  Northern aspect $\Rightarrow$ less exposure to sun $\Rightarrow$ less evaporation of morning dew $\Rightarrow$ more trees.
- E/W: Upper curves for W aspect, lower for E.
  Western aspect $\Rightarrow$ more exposure to rain (prevailing winds during rainfall are from W and S-W) $\Rightarrow$ more trees.



Estimated probabilities

Asymptotic Efficiencies (one explanatory variable)

- $W = \text{diag}(p_i[1-p_i])^{1/2}, \quad u = \frac{\partial}{\partial\beta}p = (x_i\phi(\beta x_i)), \quad R = \text{cor}(Y)$
- $\widehat{\beta}$ solution to $\quad c^t W^{-1}[Y - p(\beta)] = 0 \quad$ for $\quad c \in \Re^N \quad \Rightarrow$

  (asymptotic) $\text{var}(\widehat{\beta}) = c^t R c/(c^t u)^2$
- $\widehat{\beta}_{\text{opt}} : c = R^{-1}u, \qquad \text{var(opt)} = 1/(u^t R^{-1} u)$
- $\widehat{\beta}_{\text{ind}} : c = u, \qquad \text{var(ind)} = u^t R u/(u^t u)^2$
- $\widehat{\beta}_{\text{cl}}, \widehat{\beta}_{\text{ibee}} : c = \ldots, \qquad \text{var} = \#$
- Compute eff(est) = var(opt)/var(est), especially when x : eff(ind) is low:
  - eff(ind) $= (u^t u)^2/[(u^t R u)(u^t R^{-1} u)]$.
  - If
    $\{(\lambda_i, v_i)\} = \text{eig}(R), \quad \lambda_{\text{max}} \gg \lambda_{\text{min}} \quad$ and $\quad u = (v_{\text{max}} + v_{\text{min}}) \equiv \bar{v}$
  - then
    $\text{eff(ind)} = [(\lambda_{\text{max}} + \lambda_{\text{min}})(\lambda_{\text{max}}^{-1} + \lambda_{\text{min}}^{-1})]^{-1} \approx \lambda_{\text{min}}/\lambda_{\text{max}}$

Numerical Results ($N = 24 \times 24$)

$\beta = 1$, $\quad \text{cov}(Y) = R = (\phi(\beta x_j)\phi(\beta x_k)c(\| s_j - s_k \|))$

If x_i : for $u_i \equiv x_i\phi(\beta x_i)$, $u = \bar{v}$, need eigenvectors of R

So use $\bar{v}$ from R when $\beta = 0$; C = corresponding condition number

$\mathbf{1}$: vector of ones; $\bar{\mathbf{v}}$: unfavorable vector; $\mathbf{r}$: pseudo-random numbers

Cov matrix				Ind	Ibee(k)			Cl(d)		
σ^2	ρ	C	$\mathbf{x}$		2	3	4	1	2	3
			$\mathbf{1}$	0.81	0.81	0.81	0.81	0.79	0.78	0.77
0.8	0.90	225	$\bar{\mathbf{v}}$	0.05	0.27	0.53	0.75	0.15	0.08	0.09
			$\mathbf{r}$	0.85	0.92	0.96	0.97	0.92	0.92	0.89
			$\mathbf{1}$	0.92	0.92	0.92	0.92	0.90	0.88	0.86
0.8	0.61	21	$\bar{\mathbf{v}}$	0.32	0.69	0.84	0.91	0.52	0.40	0.41
			$\mathbf{r}$	0.86	0.93	0.97	0.98	0.92	0.90	0.88
			$\mathbf{1}$	0.98	0.98	0.98	0.97	0.96	0.94	0.93
0.8	0.37	5	$\bar{\mathbf{v}}$	0.70	0.90	0.95	0.97	0.82	0.75	0.74
			$\mathbf{r}$	0.94	0.97	0.98	0.99	0.96	0.94	0.92

- eff(cl(d)) tends to *decrease* with d (Nott and Ryden, 1999 *Biometrika*)

Some remarks

- IBEE seems to "work" for exponential autocorrelation structure, even when $u = \bar{v}$, while CL does not

  Possible insight: time series

  - $\text{cov}(Y) = R$ of AR(1) structure

  - R^{-1} is tri-diagonal $\approx$ block-diagonal

- Ind, CL and IBEE all work more-or-less OK for $u = 1$ and $u =$ vector of random numbers

  Should *always* consider the worst-scenario (*e.g.*, choice of β for simulation study of new regression estimator, . . .)

Possible extensions - 1

Comparing IBEE with other approaches

- IBEE: ad hoc, due to large (?) data set of 6,000 sites
- Likelihood-based approach
  - Invert cov(Y) for IRWLS?
  - Other approximations for cov(Y), cov^{-1}(Y) ?
- Bayesian approach
  - MCMC with CAR?
  - INLA?
- Comparing computational feasibility
- Comparing substantive conclusions

Possible extensions - 2

Ordinal response

Binary: $Y_i = 0, 1$ $\quad P(Y_i = 1 \mid Z^*) = \Phi(x_i^t\gamma + Z_i^*)$

$\Rightarrow E(Y_i) = P(Z_i \leq \eta_i), \quad E(Y_iY_j) = P(Z_i \leq \eta_i, Z_j \leq \eta_j), \quad \eta = x^t\beta$

Ordinal: $Y_i = 0, 1, \ldots, c$ $\quad P(Y_i \geq k \mid Z^*) = \Phi(x_i^t\gamma - \alpha_k^* + Z_i^*)$

for $\quad \alpha_1^* < \cdots < \alpha_c^*$

Then if $U_i^{(k)} \equiv 1_{\{Y_i \geq k\}}$,

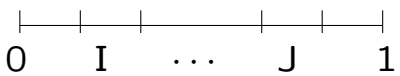
- $E(U_i^{(k)}) = P(Z_i \leq \eta_i - \alpha_k)$
- $E(U_i^{(k)}U_i^{(\ell)}) = E(U_i^{(\ell)}))$ if $k < \ell$
- $E(U_i^{(k)}U_j^{(\ell)}) = P(Z_i \leq \eta_i - \alpha_k, Z_j \leq \eta_j - \alpha_\ell)$

Possible extensions - 3

Latent-scale correlograms

- When $Y_i = 1_{(Z_i \le \eta_i)}$, how check the form of $\text{cov}(Z_i, Z_j) = r(\|s_i - s_j\|)$?
- Data: $\{(Y_i, Y_j) : \|s_i - s_j\| = d\} \Longrightarrow$ sample covariance at distance d
- But: $\text{cov}(Y_i, Y_j)$ depends on p_i, p_j and is on *wrong scale*
- Solution:

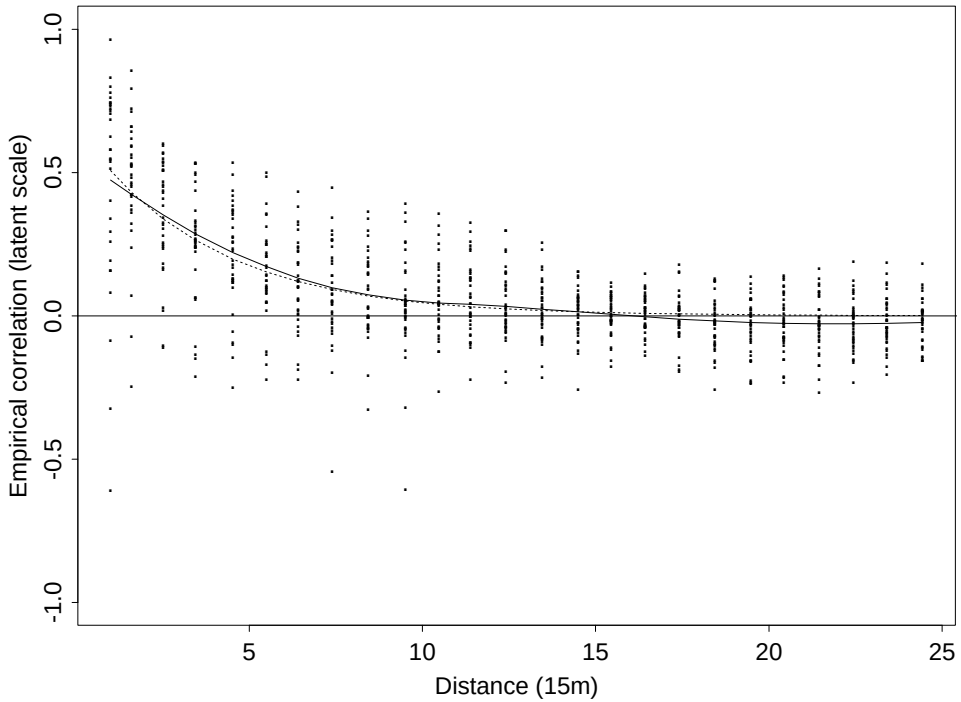
  Pearson: $\text{cov}(Y_i, Y_j) \approx \phi(\eta_i)\phi(\eta_j) \arcsin(r(d))$ $\qquad \widehat{p}_i = \Phi(\widehat{\eta}_i)$ from probit fit

  $$c(I, J, d) = \text{sample cov}\{(Y_i, Y_j) : \widehat{p}_i \in I,\ \widehat{p}_j \in J,\ \|s_i - s_j\| = d\}$$

  $$\Rightarrow \sin\left[\frac{c(I,J,d)}{\phi_I \phi_J}\right] \approx r(d), \qquad \phi_K = \text{mean}\{\phi(\widehat{\eta}_k) : \Phi(\widehat{\eta}_k) \in K\}$$

For the Meron vegetation data



Solid line is lowess fit, dashed line is least-squares fit $0.66(0.77)^d$

For a Poisson response?

- $Y_i \mid Z \sim \text{Pois}(\mu_i)$ independently

  $\log(\mu_i) = \eta_i + Z_i, \quad \eta_i = x_i^t\beta, \quad Z_i \sim N(0, \sigma^2)$

  $\rho_{ij} \equiv \text{cor}(Z_i, Z_j) = \rho(\|s_i - s_j\|)$

- $E(Y_i \mid Z) = e^{\eta_i + Z_i}$

  $$\begin{aligned} \Rightarrow E(Y_i) &= e^{\eta_i} E(e^{Z_i}) \\ &= e^{\eta_i + \sigma^2/2} \\ &= \gamma_i e^{\sigma^2/2}, \quad \gamma_i = e^{\eta_i} \end{aligned}$$

- $$\begin{aligned} E(Y_i Y_j \mid Z) &= E(Y_i \mid Z) E(Y_j \mid Z) \\ &= e^{\eta_i + Z_i} e^{\eta_j + Z_j} \end{aligned}$$

  $$\begin{aligned} \Rightarrow E(Y_i Y_j) &= e^{\eta_i + \eta_j} E(e^{Z_i + Z_j}) \\ &= \cdots = \gamma_i \gamma_j e^{\sigma^2(1+\rho_{ij})} \end{aligned}$$

- Have $E(Y_i) = \gamma_i e^{\sigma^2/2}, \qquad E(Y_i Y_j) = \gamma_i \gamma_j e^{\sigma^2(1+\rho_{ij})}, \quad \gamma_i = e^{x_i^t \beta}$

- So $\bar{Y} = (Y_1 + \cdots + Y_n)/n$

  $\approx e^{\sigma^2/2}(\gamma_1 + \cdots + \gamma_n)/n \approx e^{\sigma^2/2}\bar{\hat{\gamma}}$ where $\hat{\gamma}_i = e^{x_i^t \hat{\beta}}$

  $\Rightarrow \hat{\sigma}^2 = 2\log(\bar{Y}/\bar{\hat{\gamma}})$

- Divide distance axis into bins and let

  $N_d = \{(i,j) : | \, \| s_i - s_j \| - d \, | < \epsilon\}, \quad m_d = | N_d |$

- Then $\underbrace{\frac{1}{m_j} \Sigma_{(i,j) \in N_d} Y_i Y_j}_{V(d)} \approx \underbrace{\frac{1}{m_j} \Sigma_{(i,j) \in N_d} \gamma_i \gamma_j}_{U(d)} e^{\sigma^2(1+\rho_{ij})}$

  $\Rightarrow \hat{\rho}(d) = \hat{\sigma}^{-2} \log[V(d)/\hat{U}(d)] - 1$

- Graph $\hat{\rho}(d)$ vs d

GLIM - in general?

- $Y_i \mid Z \sim \text{Pois}(\mu_i)$ $\qquad$ $Y_i \mid Z \sim f(y \mid \mu_i), \quad \mu_i = E(Y_i \mid Z)$

  $f(y \mid \mu) = \exp\{[y\theta - b(\theta)] + c(y)\}$

  $b'(\theta) = \mu \qquad (\phi = w_i \equiv 1)$

  $\log(\mu_i) = \eta_i + Z_i$ $\qquad$ $g(\mu_i) = \eta_i + Z_i$

  (g canonical : $g(\mu_i) = \theta_i$)

  $\mu_i = g^{-1}(\eta_i + Z_i) \equiv h(\eta_i + Z_i)$

- Then

  $E(Y_i \mid Z) = h(\eta_i + Z_i)$

  $\Rightarrow E(Y_i) = E[h(\eta_i + Z_i)] = \text{function of } (\eta_i, \sigma^2)$

  $\longrightarrow$ (Laplace approx?) estimate of σ^2

  $E(Y_iY_j \mid Z) = h(\eta_i + Z_i)h(\eta_j + Z_j)$

  $\Rightarrow E(Y_iY_j) = E[h(\eta_i + Z_i)h(\eta_j + Z_j)] = \text{function of } (\eta_i, \eta_j, \sigma^2, \rho_{ij})$

  $\longrightarrow$ (Laplace approx?) estimate of ρ_{ij}

Muchas gracias

Merci beaucoup

Thank you

Todah rabah