Stat 225

Rima Izem

Areal Data Analysis

CAR, summary

Model (regression)

$$Y_i | y_j, j \neq i, x_i \sim N(\sum_j b_{ij} y_j + x_i \beta, \sigma_i^2)$$

- Condition for this model to be valid: $(I - B)^{-1}\Delta$ is symmetric and full rank.
- If $B = \lambda W$, then the proximity structure or graph has to be symmetric (e.g. distance based works, k nn does not always work).
- If W row standardized, Δ has to change accordingly to satisfy symmetry. If W is symmetric ($w_{i,j} = w_{j,i}$), then homoscedastic model can be used.
- Parameters are optimized by max likelihood

SAR and CAR, comparison

SAR model (review)

- SAR Model,
$$Y = CY + \epsilon$$
where $\epsilon \sim MN(0, D)$. So, ϵ_i' are *independent* with mean 0 and variance σ_i^2.
- ϵ induces a distribution on Y,
$Y \sim$

SAR and CAR, comparison

SAR model (review)

- SAR Model,

$$Y = CY + \epsilon$$

  where $\epsilon \sim MN(0, D)$. So, ϵ_i' are *independent* with mean 0 and variance σ_i^2.
- ϵ induces a distribution on Y, $Y \sim MN(0, (I - C)^{-1}D((I - C)^{-1})')$.
- $Cov(\epsilon, Y) =$

SAR and CAR, comparison

SAR model (review)

- SAR Model,

$$Y = CY + \epsilon$$

  where $\epsilon \sim MN(0, D)$. So, ϵ_i' are *independent* with mean 0 and variance σ_i^2.
- ϵ induces a distribution on Y, $Y \sim MN(0, (I - C)^{-1} D((I - C)^{-1})')$.
- $Cov(\epsilon, Y) = D((I - C)^{-1})'$

SAR and CAR, comparison

SAR model (review)

- SAR Model,

$$Y = CY + \epsilon$$

  where $\epsilon \sim MN(0, D)$. So, ϵ_i' are *independent* with mean 0 and variance σ_i^2.
- ϵ induces a distribution on Y, $Y \sim MN(0, (I - C)^{-1} D(((I - C)^{-1})')$.
- $Cov(\epsilon, Y) = D(((I - C)^{-1})'$ (data from one spatial unit is correlated with residuals from another spatial unit)

SAR and CAR, comparison

CAR model

- Proper CAR model, could be re-written as

$$Y = BY + \epsilon;$$

where $\epsilon \sim N(0, \Delta(I - B)')$, and $\Delta = diagonal(\sigma_1^2, \ldots, \sigma_n^2)$.
- Y induces a distribution on ϵ.
- $Cov(\epsilon, Y) =$

SAR and CAR, comparison

CAR model

- Proper CAR model, could be re-written as

$$Y = BY + \epsilon;$$

  where $\epsilon \sim N(0, \Delta(I - B)')$, and
  $\Delta = diagonal(\sigma_1^2, \ldots, \sigma_n^2)$.
- Y induces a distribution on ϵ.
- $Cov(\epsilon, Y) = \Delta$.

SAR and CAR, comparison

CAR model

- Proper CAR model, could be re-written as

$$Y = BY + \epsilon;$$

  where $\epsilon \sim N(0, \Delta(I - B)')$, and
  $\Delta = diagonal(\sigma_1^2, \ldots, \sigma_n^2)$.
- Y induces a distribution on ϵ.
- $Cov(\epsilon, Y) = \Delta$. (data from one spatial unit is uncorrelated with residuals from other spatial units)

SAR and CAR

can we go from one to the other?

- The two models are equivalent iff

SAR and CAR

can we go from one to the other?

- The two models are equivalent iff

$$(I - B)^{-1}\Delta = (I - C)^{-1}D((I - C)^{-1})'$$

  (equal covariances when joint distribution is well defined)
- Consequence:

SAR and CAR

can we go from one to the other?

- The two models are equivalent iff

$$(I - B)^{-1}\Delta = (I - C)^{-1}D((I - C)^{-1})'$$

  (equal covariances when joint distribution is well defined)
- Consequence: any (general) SAR model could be written as a (general) CAR model, but converse is not true.
- WARNING: this does not mean that any SAR model with $C = \lambda\tilde{W}$ could be written as a CAR model with $B = \lambda\tilde{W}$.

Exponential family

Example 1: Binomial distribution

- Binomial distribution: count of *successes* in n independent trials, with probability of success in each trial is p. Notation $B(n, p)$.
- In spatial unit i, we could observe a variable with distribution $B(n_i, p_i)$. Example: unemployment count.
- Binomial and normality: as $n \longrightarrow \infty$,

Exponential family

Example 1: Binomial distribution

- Binomial distribution: count of *successes* in n independent trials, with probability of success in each trial is p. Notation $B(n, p)$.
- In spatial unit i, we could observe a variable with distribution $B(n_i, p_i)$. Example: unemployment count.
- Binomial and normality: as $n \rightarrow \infty, B(n, p)/n \approx N(p, \mathbf{p(1-p)/n})$.

Exponential family

Example 2: Poisson

- Poisson distribution: count of *successes* for an infinite number of independent trials, with small probability of success such that the rate of success is a constant λ. Notation: $P(\lambda)$.
- For spatial unit i: $P(\lambda_i)$. Example: disease or death count, for a rare disease.
- Binomial and poisson, if $n \rightarrow \infty$ and $p \rightarrow 0$ with $np \rightarrow \lambda$, then

Exponential family

Example 2: Poisson

- Poisson distribution: count of *successes* for an infinite number of independent trials, with small probability of success such that the rate of success is a constant λ. Notation: $P(\lambda)$.
- For spatial unit i: $P(\lambda_i)$. Example: disease or death count, for a rare disease.
- Binomial and poisson, if $n \longrightarrow \infty$ and $p \longrightarrow 0$ with $np \longrightarrow \lambda$, then $B(n, p) \approx P(\lambda)$.
- Poisson and normality: as $\lambda \longrightarrow \infty$, $P(\lambda)$

Exponential family

Example 2: Poisson

- Poisson distribution: count of *successes* for an infinite number of independent trials, with small probability of success such that the rate of success is a constant λ. Notation: $P(\lambda)$.
- For spatial unit i: $P(\lambda_i)$. Example: disease or death count, for a rare disease.
- Binomial and poisson, if $n \to \infty$ and $p \to 0$ with $np \to \lambda$, then $B(n, p) \approx P(\lambda)$.
- Poisson and normality: as $\lambda \to \infty, P(\lambda) \approx N(\lambda, \lambda)$.

Exponential family

A p.d.f (probability density function) belongs to the exponential family if it could be written as

$$
\begin{aligned}
f(x|\theta) &\propto exp(\sum_{i=1}^{k} h_i(\theta)\chi_i(x)) \\
\text{i.e. } \log(\text{p.d.f.}) &\propto \text{sum of (fct. of the parameters)} + \\
&\quad \text{(fct. of the data)}
\end{aligned}
$$

Some examples,

- Gaussian family, parameters $\theta = (\mu, \sigma^2)$
- Binomial family, parameters $\theta = (n, p)$(number of trials, probability of success)
- Poisson family, parameter $\theta = \lambda$ rate of success.

GLM and spatial modelling

- Data Y_i from a distribution with parameter θ_i
- Usual glm:
  - Link function g(specific to a distribution), such that $\eta_i = g(\theta_i)$. Example: g is the logit, or log link.
  - η_i is a linear combination of some covariates.
- Spatial GLM: η_i follows a Spatial model with mean which could depend on covariates.

Usual GLM

Example, log link

$$\begin{aligned} Y_i &\sim \text{Poisson}(\lambda_i) \\ log \lambda_i = X_i^T \beta \end{aligned}$$

Example, logit link

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ logit(p_i) = log \frac{p_i}{1 - p_i} &\sim X_i^T \beta \end{aligned}$$

Disease mapping

Basic (epidemiology) Model

$$Y_i \sim \text{Poisson}(E_i \theta_i)$$

- Y_i is the observed count.
- E_i is called the expected count in region i. It is usually assumed to be the age standardized risk. It depends on size and demographic structure in region i.
- θ_i is a region specific relative risk. It accounts for additional multiplicative risk associated with region i, not already accounted for by E_i. It is usually assumed to be random.

Expected count and standardization

External indirect (use other data),

$$E_i = \sum_j n_{ij} r_j$$

r_j is the rate for age strata j, and n_{ij} is the number of people in region i and age strata j.
Internal indirect (use data Y),

$$E_i = \sum_j n_{ij} \frac{\sum_i Y_{ij}}{\sum_i n_{ij}} = \sum_j n_{ij} \hat{r}_j$$

Example: Cape Cod breast cancer

Columns 2 and 3 show age-specific population based on the 1990 census, and the number of newly diagnosed breast cancers in Massachusetts from 1987 to 1994, based on data from the Massachusetts Cancer Registry. Col. 4 shows the age-specific population on Cape Cod.

	Whole State		Upper Cape	
agegroup	1990 pop	cases	1990 pop	exp
5-24	819538	20	12717	0
25-34	552659	768	8881	12
35-44	465950	3619	8601	67
45-54	306719	6014	5430	106
55-64	272295	7357	5809	157
65-74	262749	9723	6189	229
75-84	173447	6919	3604	144
85+	68434	2013	1386	41

Example: contd

By summing up the age-specific expected numbers, we get the overall number of expected cancers in the Upper Cape region. In this particular case, we had a total of 864 breast cancers observed in the Upper Cape between 1986 and 1994. This yields an estimated SMR of 100*864/756=114.

Breast Cancer SIRS by tract for Upper Cape Cod

Tract	Obs.	Exp.	SIR	95% Conf Int
122	41	36.5	112	(81,152)
123	7	3.5	200	(80,412)
124	15	20.6	73	(41,120)
125	36	27.9	129	(90,177)
126	58	59.1	98	(75,127)
127	59	44.4	133	(101,171)
...				
...				
149	54	40.6	133	(100,174)
150	18	26.3	68	(41,108)
151	16	13.0	123	(70,200)
152	17	13.4	127	(74,203)

Disease mapping

Basic Model

$$Y_i \sim \text{Poisson}(E_i \eta_i)$$

- Naive estimate of $\hat{\theta}_i$,

Disease mapping

Basic Model

$$Y_i \sim \text{Poisson}(E_i \eta_i)$$

- Naive estimate of $\hat{\theta}_i$, $\hat{\theta}_i = \frac{Y_i}{E_i}$.
- Problem with naive estimates:

Disease mapping

Basic Model

$$Y_i \sim \text{Poisson}(E_i \eta_i)$$

- Naive estimate of $\hat{\theta}_i$, $\hat{\theta}_i = \frac{Y_i}{E_i}$.
- Problem with naive estimates: estimate is noisy, $Var(\frac{Y_i}{E_i}) = \frac{\lambda_i}{E_i^2}$.
- How to find smoother estimates of θ_i?

Disease mapping

Basic Model

$$Y_i \sim \text{Poisson}(E_i \eta_i)$$

- Naive estimate of $\hat{\theta}_i$, $\hat{\theta}_i = \frac{Y_i}{E_i}$.
- Problem with naive estimates: estimate is noisy, $Var(\frac{Y_i}{E_i}) = \frac{\lambda_i}{E_i^2}$.
- How to find smoother estimates of θ_i? use covariates and spatial model.

Poisson-gamma Model

Distribution of θ_i. If we assume a common gamma distribution,

$$\theta_i \sim \mathcal{G}(\alpha, \beta)$$

then calculations are relatively straightforward (Clayton and Kaldor, Biometrics 1987) because we can integrate θ out of the likelihood in closed form.

Poisson-gamma model (contd)

more precisely:

- Under the gamma distribution, the mean $\mu = \alpha/\beta$ and $\sigma^2 = \alpha/\beta^2$. We may want to fix $\mu \equiv 1$ if we have internally standardized.
- The standard predictions of the random effects in this framework come in closed form (they are also so-called Empirical Bayes estimators)

$$\mathrm{E}(\theta_i|Y) = \frac{Y_i + \alpha}{E_i + \beta} = \frac{E_i(Y_i/E_i) + (\mu/\sigma^2)\mu}{E_i + \mu/\sigma^2}$$
$$= \frac{w_{\text{data}}(Y_i/E_i) + w_{\text{mean}}\mu}{w_{\text{data}} + w_{\text{mean}}}$$

- This is a precision-weighted average also called shrinkage toward the mean.

Poisson-Gamma model, More details

- A problem: what are α and β? These are critical in controlling the amount of shrinkage.
- The Poisson-gamma formulation allows one to integrate θ_i out of the likelihood and give a closed-form likelihood as a function solely of α and β, which can be maximized numerically. (Ex: using `nlm()` in R)
- The gamma representation gives a negative binomial distribution for the counts marginally.