A LATENT VARIABLE REPRESENTATION OF COUNT DATA MODELS TO ACCOMMODATE SPATIAL AND TEMPORAL DEPENDENCE

Application to Predicting Crash Frequency at Intersections

Marisol Castro

Chandra Bhat & Rajesh Paleti

28 Octubre 2016

Conteos

- Variable dependiente
  - Discreta
  - No negativa
  - Sin límite superior
- Ejemplos
  - Tasa de motorización
  - Viajes generados por hogar
  - Accidentes de tránsito

Modelos de Conteos

- Poisson
- Binomial negativa
- Binomial
- Logarítmica
- Zero-inflated, Hurdle

- Más usada para conteos
- Restricción varianza = media
- Permite sobredispersión
- Permite sub-dispersión
- Problemas de estimación

Modelos univariados:

- Sin correlación temporal
- Sin correlación espacial

- Permite sobre y sub dispersión
- Problemas de estimación
- Controla el "exceso" de ceros

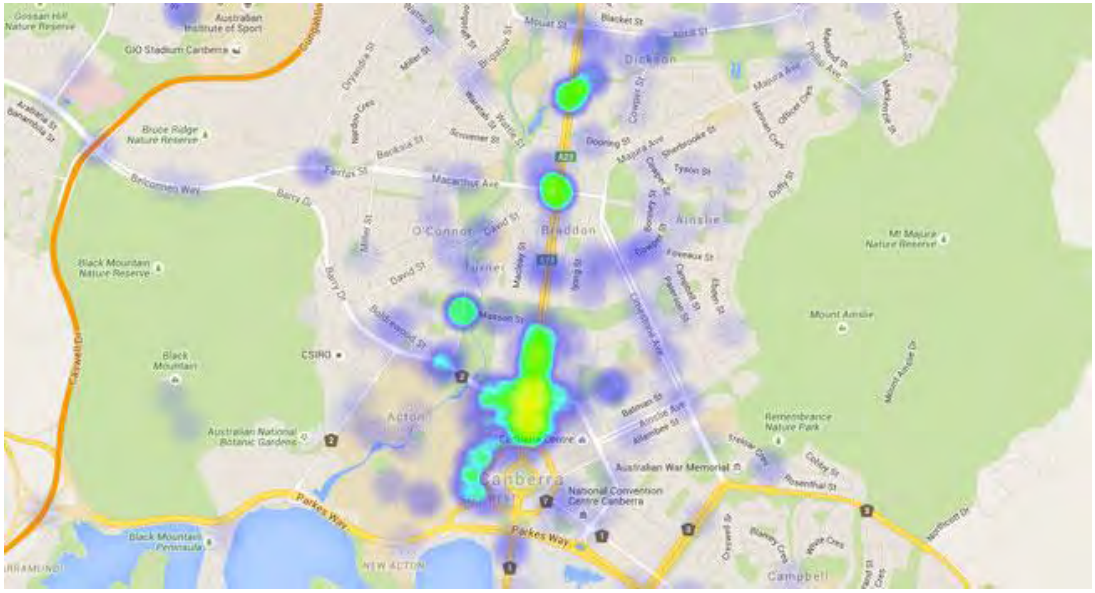
- **Modelos de conteos que <u>sí</u> incluyen correlaciones**
  - Poisson y binomial negativa multivariada
  - Mixture methods: términos aleatorios incluidos en la parametrización de la media

- Problemas con proceso de estimación: incluir correlación espacial y temporal es difícil/imposible

¿Correlación Temporal y Espacial?

- Correlación temporal

- Correlación espacial
  - Efectos no observados
  - Efectos de variables observadas (spillover effects)



Accidentes de ciclistas en Canberra

Alcance del Estudio

- Formular un nuevo modelo para conteos
  - Modelo ordinal
  - Incorpora estructuras flexibles de correlación espacial y temporal

MODELO

Modelo Ordinal

$$y_q^* = \boldsymbol{\beta}' \boldsymbol{x}_q + \varepsilon_q$$

$$y_q = k \quad \text{if} \quad \psi_{k-1} < y_q^* < \psi_k$$

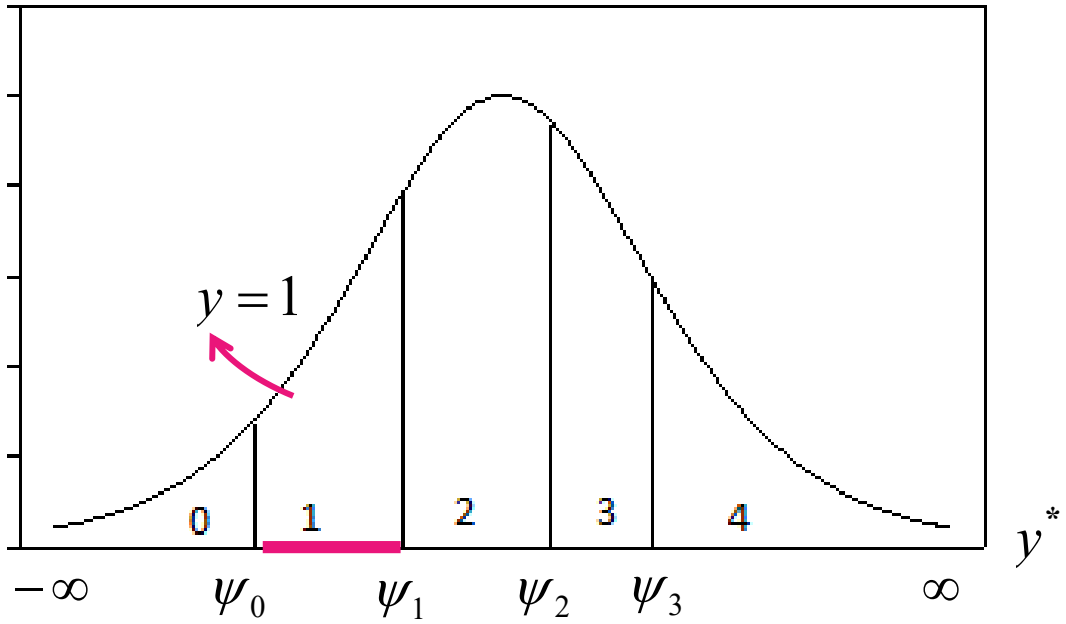
Variable latente

Variable dependiente

q = 1, 2, …, N
unidad de análisis

k = 0, 1, …, K
Nivel (ordinal)

Umbrales $-\infty < \psi_1 < \psi_2 < \ldots < \psi_{k-1} < \infty$



Modelo Ordinal

$$y_q^* = \boldsymbol{\beta}' \boldsymbol{x}_q + \varepsilon_q$$

$$y_q = k \quad \text{if} \quad \psi_{k-1} < y_q^* < \psi_k$$

Variable latente

$q = 1, 2, \ldots, N$
unidad de análisis

Variable dependiente

$k = 0, 1, \ldots, K$
Nivel (ordinal)

Umbrales $-\infty < \psi_1 < \psi_2 < \ldots < \psi_{k-1} < \infty$

- Parametrización de los umbrales $\psi_{qk} = f_k(\boldsymbol{z}_q)$
  - Umbrales (mapping) puede variar entre observaciones

$$\psi_{qk} = \Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k} \frac{\lambda_q^l}{l!} \right) + \alpha_k$$

donde $\lambda_q = e^{\gamma' z_q}$

Φ^{-1} función inversa de la distribución normal univariada

α_k constante asociada al umbral k

γ parámetro a estimar

z_q variables asociadas al umbral

Incluyendo correlaciones…

- Correlación es incorporada a través de la variable latente y^*

- Modelo

$$y^*_{qt} = \delta \sum_{q'=1}^{Q} w_{qq'} y^*_{q't} + \boldsymbol{\beta}'_q \boldsymbol{x}_{qt} + \varepsilon_{qt}$$

$q = 1, 2, \ldots, N$ unidad de análisis

$t = 1, 2, \ldots, T$ periodo de tiempo

- $w_{qq'}$ matriz espacial (weight matrix) $\rightarrow \boldsymbol{W}$ ($N \times N$)
- δ parámetro espacial (escalar)
- $\beta_q \sim N(b, \Omega)$ parámetro aleatorio, ($L \times 1$) $\rightarrow \boldsymbol{\beta}_q = \boldsymbol{b} + \tilde{\boldsymbol{\beta}}_q, \ \tilde{\boldsymbol{\beta}}_q \sim MVN_L(0, \boldsymbol{\Omega})$
- $\varepsilon_{qt} \sim N(0, \Lambda)$ término de error

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \ldots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \ldots & \rho^{T-3} \\ \vdots & \vdots & \vdots & 1 & \ldots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \rho^{T-4} & \ldots & 1 \end{bmatrix}$$

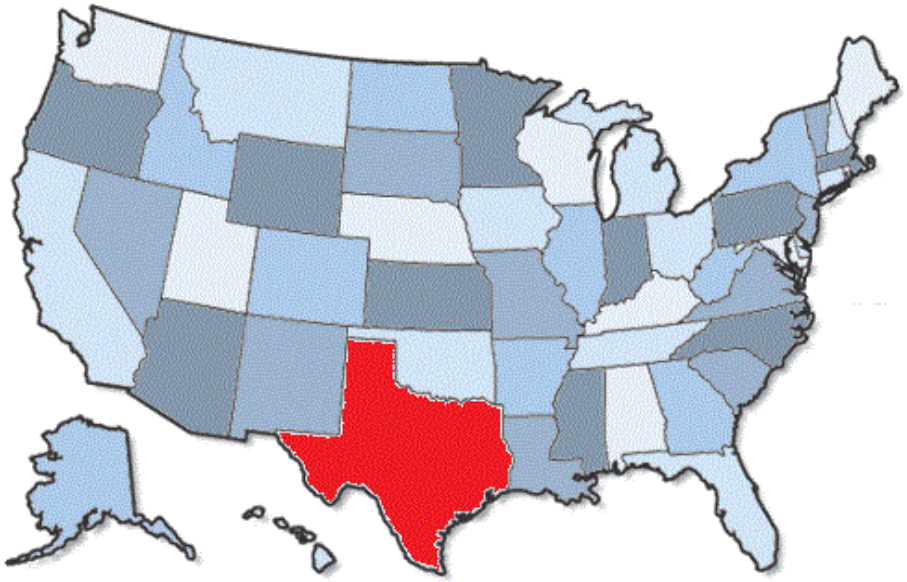
Estructura temporal autoregresiva

$\boldsymbol{\Lambda}$: ($T \times T$)

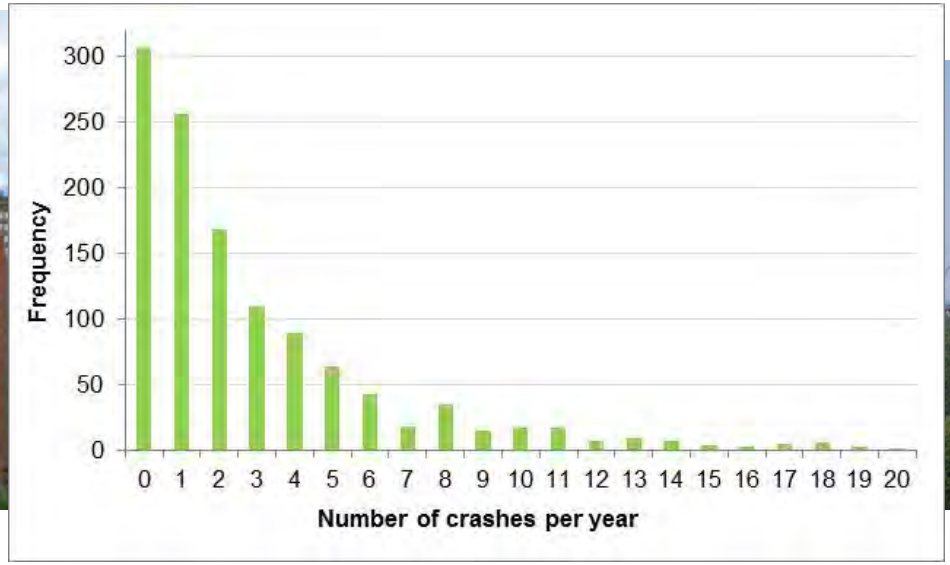
APLICACIÓN DEL MODELO

Análisis de accidentes en intersecciones

Base de Datos



- 170 intersecciones
- 7 años
  - → 1190 observaciones
- 3,503 accidentes
  - Promedio 2,94 por año-intersección



Variables	No Correlation		Spatial and Temporal Effects	
	Estimate	t-stat	Estimate	t-stat
Threshold Variables				
Threshold Specific Constants				
α_1	-0.026	-1.06	0.492	10.30
α_2	-0.166	-5.02	0.708	10.16
α_3	-0.377	-8.81	0.769	8.77
α_4	-0.563	-11.25	0.842	7.93
α_5	-0.745	-13.02	0.901	7.23
α_6	-0.968	-15.18	0.882	6.20
α_7	-1.264	-17.25	0.722	4.65
α_8	-1.430	-18.88	0.809	4.65
α_9	-1.698	-20.81	0.689	3.70
γ - Vector				
Constant	1.223	40.40	2.309	16.09
Approach Roadway Type				
At least one approach roadway is non-city street	0.449	11.39	0.313	9.00
Type of Traffic Control				
Regular signal light	0.837	7.49	0.888	10.82

$$\psi_{q,m_{qt},t} = \Phi^{-1}\left(e^{-\lambda_{qt}}\sum_{l=0}^{m_{qt}}\frac{\lambda_{qt}^{l}}{l!}\right) + \alpha_{k_{qt}}$$

$$\lambda_{qt} = e^{\gamma' z_{qt}}$$

$$\alpha_0 = 0, \quad \alpha_{k_{qt}} = \alpha_K \text{ if } k_{qt} > K,$$

Variables	No Correlation		Spatial and Temporal Effects	
	Estimate	t-stat	Estimate	t-stat
Latent Variables				
Constant	0.000	-	0.000	-
Standard Deviation	-	-	0.640	7.02
Number of Entering Roads				
Three	-0.660	-15.07	-0.950	-13.20
More than four	-0.921	-13.64	-1.276	-13.20
Type of Traffic Control				
Regular signal light	-1.906	-7.35	-3.003	-7.53
Yield sign	-0.912	-10.40	-1.352	-8.55
Standard Deviation	-	-	1.168	6.32
Stop sign	-0.454	-7.80	-0.540	-7.01
Flashing light	0.696	8.69	0.948	9.08
Center stripe / divider	-0.659	-8.83	-0.786	-6.96
Logarithm of daily entering volume (veh/day/10,000)	0.210	6.76	0.374	7.98
Standard Deviation	-	-	0.831	15.60
Flow split imbalance (FSIMB) factor	-0.817	-9.86	-1.042	-9.49
Year-specific Dummy Variables				
Year 2004	-0.210	-2.23	-0.103	-1.09
Year 2005	-0.440	-4.67	-0.291	-2.90
Year 2006	-0.438	-4.69	-0.218	-1.94
Year 2007	-0.441	-4.67	-0.182	-1.52
Year 2008	-0.470	-5.03	-0.250	-2.10
Year 2009	-0.539	-5.85	-0.378	-3.30

$$y_{qt}^{*} = \delta \sum_{q'=1}^{Q} w_{qq'} y_{q't}^{*} + \boldsymbol{\beta}_{q}^{'} \boldsymbol{x}_{qt} + \varepsilon_{qt},$$

δ = 0.422 (9.25)
(parámetro de correlación espacial)

Comentarios finales

- Modelo aplicable a distintos casos de estudio
- (Potencial) importancia de incluir estructuras de correlación
- En el paper hay detalles de:
  - Interpretación del modelo y sus parámetros
  - Método de estimación
  - Elección de la matriz espacial
  - Resultados de la aplicación

GRACIAS!!

Castro, M., Paleti, R., & Bhat, C. R. (2012).

A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections.

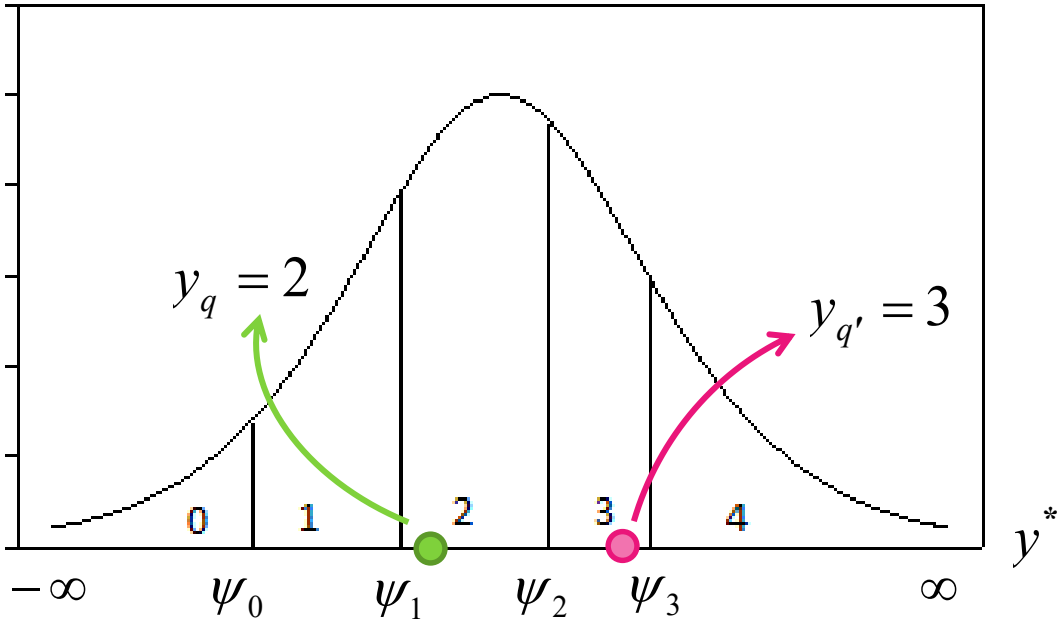
Transportation research part B: methodological, *46*(1), 253-272.

How do we interpret this representation?

$$y_q^* = \boldsymbol{\beta}'\boldsymbol{x}_q + \varepsilon_q$$
$$y_q = k \quad \textit{if} \quad \psi_{q,k-1} < y_q^* < \psi_{q,k}$$

$$\psi_{qk} = \Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k} \frac{\lambda_q^l}{l!} \right) + \alpha_k \quad \text{where} \quad \lambda_q = e^{\boldsymbol{\gamma}'\boldsymbol{z}_q}$$

- Two sources of variability
  - Latent variable
  - Threshold

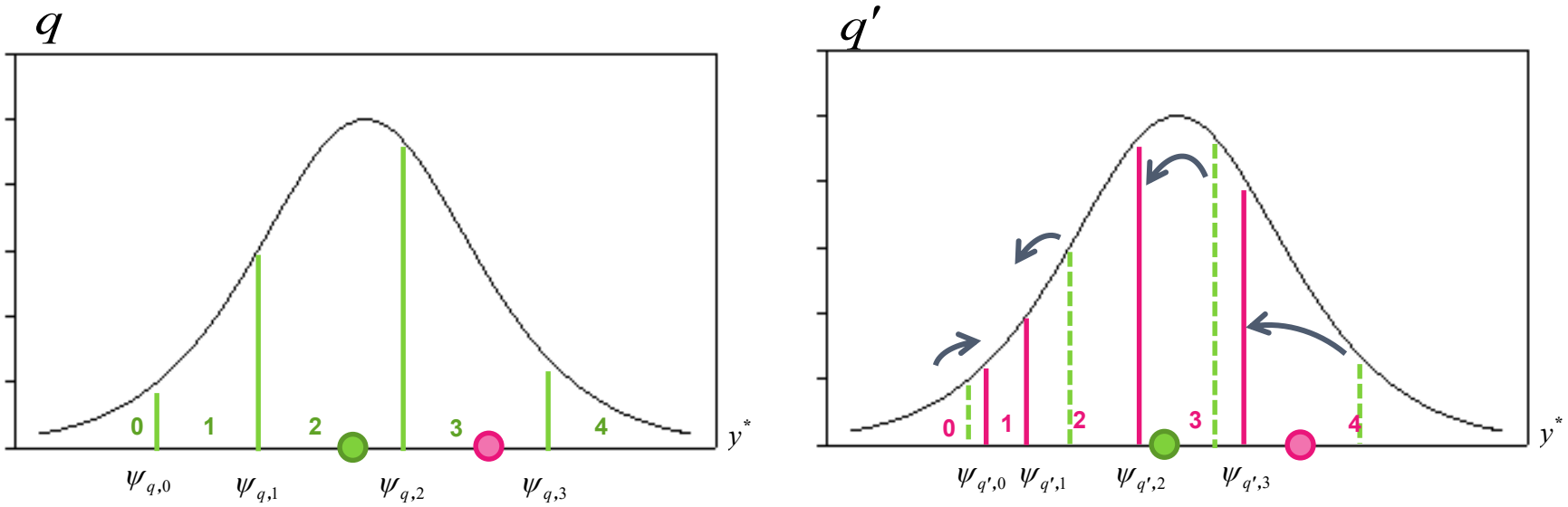


How do we interpret this representation?

$$y_q^* = \boldsymbol{\beta}'\boldsymbol{x}_q + \varepsilon_q$$
$$y_q = k \quad \textit{if} \quad \psi_{q,k-1} < y_q^* < \psi_{q,k}$$

$$\psi_{qk} = \Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k} \frac{\lambda_q^l}{l!} \right) + \alpha_k \quad \text{where} \quad \lambda_q = e^{\boldsymbol{\gamma}'\boldsymbol{z}_q}$$

- Two sources of variability
  - Latent variable → latent long-term propensity
  - Threshold → instantaneous mapping



Why choose this parameterization?

$$y_q^* = \boldsymbol{\beta}'\boldsymbol{x}_q + \varepsilon_q$$
$$y_q = k \quad \textit{if} \quad \psi_{q,k-1} < y_q^* < \psi_{q,k}$$

$$\psi_{qk} = \Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k} \frac{\lambda_q^l}{l!} \right) + \alpha_k \quad \text{where} \quad \lambda_q = e^{\boldsymbol{\gamma}'\boldsymbol{z}_q}$$

1. Satisfy threshold ordering condition

$$-\infty < \psi_1 < \psi_2 < \ldots < \psi_{k-1} < \infty$$

2. Allows identification for variables common in both x and z

3. Is a generalization of the Poisson model
   - $\alpha_k = 0$
   - $\beta = 0$
   - $\varepsilon \sim N(0,1)$

$$P[y_q = k] = P\left[\Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k-1} \frac{\lambda_q^l}{l!} \right) < y_q^* < \Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k} \frac{\lambda_q^l}{l!} \right) \right]$$

$$= \Phi\left(\Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k} \frac{\lambda_q^l}{l!} \right) \right) - \Phi\left(\Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k-1} \frac{\lambda_q^l}{l!} \right) \right)$$

$$= \frac{e^{-\lambda_q} \lambda_q^k}{k!}$$

Why choose this parameterization?

$$y_q^* = \boldsymbol{\beta}'\boldsymbol{x}_q + \varepsilon_q$$
$$y_q = k \quad \textit{if} \quad \psi_{q,k-1} < y_q^* < \psi_{q,k}$$

$$\psi_{qk} = \Phi^{-1}\left(e^{-\lambda_q} \sum_{l=0}^{k} \frac{\lambda_q^l}{l!} \right) + \alpha_k \quad \text{where} \quad \lambda_q = e^{\boldsymbol{\gamma}'\boldsymbol{z}_q}$$

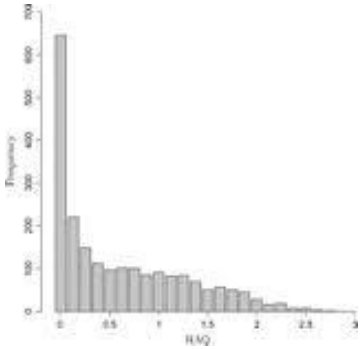
1. Satisfy threshold ordering condition
2. Allows identification for variables common in both x and z
3. Is a generalization of the Poisson model
4. Data does not require an upper bound K

$$y_q = K \quad \textit{if} \quad \psi_K < y_q^*$$

5. The threshold-specific constants α can assign different probability mass to the counts

$$\alpha_k = \alpha_K \quad \forall k \geq K$$

6. Can handle flexible correlation structures



Special Cases

Latent variable representation

$$y_{qt}^{*} = \delta \sum_{q'=1}^{Q} w_{qq'} y_{q't}^{*} + \boldsymbol{\beta}_{\boldsymbol{q}}' \boldsymbol{x}_{\boldsymbol{qt}} + \varepsilon_{qt},$$

$$\beta \sim N(0, \Omega)$$

$$y_{qt} = k_{qt} \quad \text{if} \quad \psi_{q,k_{qt}-1,t} < y_{qt}^{*} < \psi_{q,k_{qt},t}$$

Threshold parameterization

$$\psi_{q,m_{qt},t} = \Phi^{-1}\left(e^{-\lambda_{qt}} \sum_{l=0}^{m_{qt}} \frac{\lambda_{qt}^{l}}{l!} \right) + \alpha_{k_{qt}}$$

$$\lambda_{qt} = e^{\gamma' \boldsymbol{z}_{\boldsymbol{qt}}}$$

$$\alpha_0 = 0, \quad \alpha_{k_{qt}} = \alpha_K \text{ if } k_{qt} > K,$$

- $\rho = 0$ — Lack of time-varying temporal correlation
- $\delta = 0$ — No spatial correlation
- non-diagonal elements of $\Omega = 0$ (diagonal elements non-zero)

  Presence of time-invariant and unobserved heterogeneity effects, but without correlation between these effects
- $\Omega = 0$ — Lack of time-invariance and unobserved heterogeneity
- $\rho = \delta = \Omega = 0$ — Ordered response model without correlation
- $\rho = \delta = \Omega = \alpha = 0$ — Poisson model (without correlation)

Estimation

$$L(\boldsymbol{\theta}) = P(\boldsymbol{y} = \boldsymbol{m}) = \int_{D_{y^*}} \phi_{QT}(\boldsymbol{y}^* \mid \boldsymbol{b}, \boldsymbol{\Sigma}) d\boldsymbol{y}^*,$$

D_{y^*} : multivariate region of the elements of y^*

$$\theta = (\boldsymbol{b}', \boldsymbol{\Omega}', \rho, \delta, \boldsymbol{\gamma}', \boldsymbol{\alpha}')'$$

($QT \times QT$)

- Composite marginal likelihood (CML)
  - Used when the likelihood function is difficult/impossible to evaluate
  - Estimator:
    - is consistent and asymptotically normal distributed
    - loses some asymptotic efficiency relative to a likelihood estimator
  - Maximizes a surrogate likelihood function that compounds much easier-to-compute, lower-dimensional, marginal likelihoods
    - Pairwise CML: bivariate cumulative normal distribution

(2×2)

$$L_{CML}(\boldsymbol{\theta}) = \left(\prod_{g=1}^{QT-1} \prod_{g'=g+1}^{QT} P\left([\boldsymbol{y}]_g = [\boldsymbol{k}]_g, [\boldsymbol{y}]_{g'} = [\boldsymbol{k}]_{g'} \right) \right)$$

$\frac{QT(QT-1)}{2}$ times

  - Standard Errors $\boldsymbol{V}_{CML}(\hat{\boldsymbol{\theta}}) = [\boldsymbol{G}(\hat{\boldsymbol{\theta}})]^{-1} = [\boldsymbol{H}(\hat{\boldsymbol{\theta}})]^{-1} \boldsymbol{J}(\hat{\boldsymbol{\theta}}) [\boldsymbol{H}(\hat{\boldsymbol{\theta}})]^{-1}$

$$\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}}) = -\left[\sum_{g=1}^{QT-1} \sum_{g'=g+1}^{QT} \frac{\partial^2 \log L_{CML,gg'}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]_{\hat{\boldsymbol{\theta}}} \qquad \boldsymbol{J}(\hat{\boldsymbol{\theta}}) = \frac{\tilde{W}}{D} \left[\sum_{d=1}^{D} \left[\frac{1}{N_d} \left([\boldsymbol{s}_{CML,d}(\boldsymbol{\theta})][\boldsymbol{s}_{CML,d}(\boldsymbol{\theta})]' \right)_{\hat{\theta}} \right] \right]$$

Estimation

- Spatial correlation $y_{qt}^{*} = \delta \sum_{q'=1}^{Q} w_{qq'} y_{q't}^{*} + \boldsymbol{\beta}_{q}' \boldsymbol{x}_{qt} + \varepsilon_{qt}$
  - Spatial weight matrix
    - $1/d_{ij}^{p}$ *p = 1, 2, 3, …*
    - $1/exp(d_{ij})$
    - 1 if $d_{ij} < U$
  - Distance bands
    - Explore different bands
    - Minimize the variance $tr[\boldsymbol{V}_{CML}(\hat{\boldsymbol{\theta}})]$
      - 2, 3, 4, 5, 6, 7, and 11.81 miles

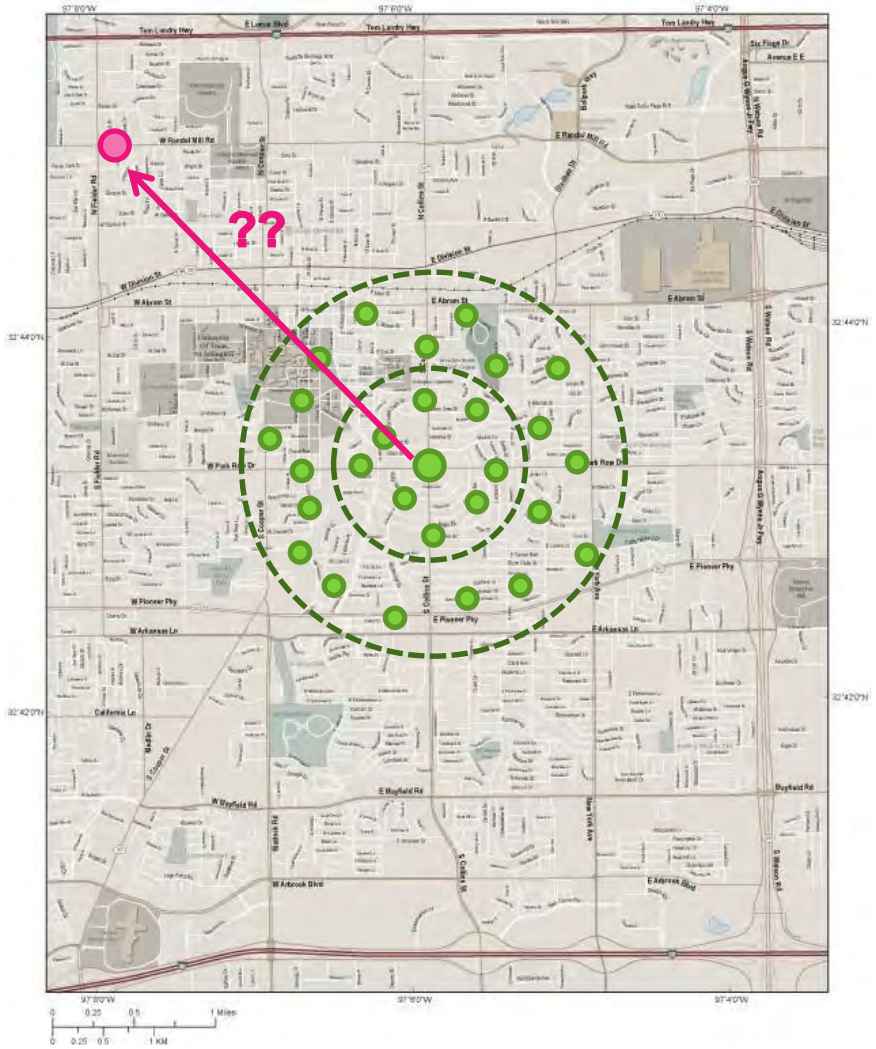
- Models
  - No correlation, $\rho = \delta = \Omega = 0$ → *27 parameters*
  - Temporal correlation, $\rho, \Omega \neq 0, \delta = 0$ → *30 parameters*
  - Temporal and spatial correlation, $\rho, \Omega, \delta \neq 0$ → *31 parameters*
  - Best model ?
    - ADCLR ~ χ^2

ADCLRT: 1646

ADCLRT: 356

Estimation

- Spatial correlation $y_{qt}^{*} = \delta \sum_{q'=1}^{Q} w_{qq'} y_{q't}^{*} + \boldsymbol{\beta}_{q}' \boldsymbol{x}_{qt} + \varepsilon_{qt}$
  - Spatial weight matrix
    - $1/d_{ij}^{p}$ *p = 1, 2, 3, …*
    - $1/exp(d_{ij})$
    - 1 if $d_{ij} < U$
  - Distance bands
    - Explore different bands
    - Minimize the variance $tr[\boldsymbol{V}_{CML}(\hat{\boldsymbol{\theta}})]$
      - 2, 3, 4, 5, 6, 7, and 11.81 miles



Estimation

- Spatial correlation $y^*_{qt} = \delta \sum_{q'=1}^{Q} w_{qq'} y^*_{q't} + \boldsymbol{\beta}'_{\boldsymbol{q}} \boldsymbol{x}_{\boldsymbol{qt}} + \varepsilon_{qt}$
  - Spatial weight matrix
    - $1/d_{ij}^{p}$ *p = 1, 2, 3, …*
    - $1/exp(d_{ij})$
    - 1 if $d_{ij} < U$
  - Distance bands
    - Explore different bands
    - Minimize the variance $tr[\boldsymbol{V}_{CML}(\hat{\boldsymbol{\theta}})]$
      - 2, 3, 4, 5, 6, 7, and 11.81 miles

- Models
  - No correlation, $\rho = \delta = \Omega = 0$
  - Temporal correlation, $\rho, \Omega \neq 0, \delta = 0$
  - Temporal and spatial correlation, $\rho, \Omega, \delta \neq 0$
  - Best model ?
    - ADCLR ~ χ^2

ρ *not significant*

Variable	No correlation		Spatial and Temporal Effects	
	Estimate	Standard error	Estimate	Standard error
Number of Entering Roads				
Three	-51.07	4.31	-54.44	5.63
More than four	-60.95	4.07	-75.08	4.39
Approach Roadway Type Combination				
At least one approach roadway is a non-city street	133.44	20.20	85.22	15.63
Type of Traffic Control				
Regular signal light	-2.85	33.24	-7.87	36.31
Yield sign	-59.45	5.04	-69.50	6.78
Stop sign	-34.79	4.78	-37.50	5.83
Flashing light	69.45	11.94	99.40	21.43
Center stripe / divider	-47.36	5.53	-53.71	7.42
Logarithm of daily entering volume (veh/day/10,000) – 10% increase	1.75	0.31	5.06	1.09
Flow split imbalance (FSIMB) factor – 0.1 increase	-6.88	0.98	-9.74	1.65

Variables	Sample share
Number of Entering Roads	
Three	24.6
Four	71.6
More than four	3.8
Roadway Alignment	
All approaches are straight with no vertical grade	95.7
At least one approach has horizontal curvature or vertical grade	4.3
Approach Roadway Type Combination	
All approach roadways are city streets	94.4
At least one approach roadway is a non-city street	5.6
Type of Traffic Control	
Regular signal light	52.8
Yield sign	15.3
Stop sign	12.4
Flashing light	7.3
Center stripe / divider	4.6
No traffic control or minimal traffic control	7.6

Descriptive Statistics

	Minimum	Maximum	Mean	Std. Dev.
Total daily entering volume (vehicles/day)	2,866	193,178	35,222	33,784
Flow split imbalance (FSIMB) factor	0.00	0.97	0.43	0.25
Distance between intersections (miles)	0.05	11.81	4.42	2.29