Semiparametric Zero-Inflated Regression Models

Eric S. Roemmele

Department of Statistics, University of Kentucky

Oral Exam Presentation
April 16^{th}, 2018

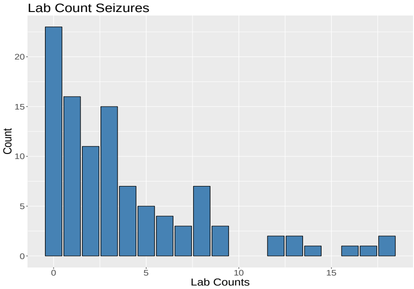
Outline of Topics

1. Motivational Data: Meth Lab Seizures
2. Introduction to Zero-Inflated Regression
3. Semiparametric Extension to ZIP
   - Overview and Estimation
   - Theoretical Results
   - Inference
   - Preliminary Simulations
   - Application to Meth Lab Seizures Data
4. Spatial ZIP
5. Conclusions & Future Directions

Outline of Topics

Meth Labs Counts in KY

- The outcome of interest is the number of clandestine meth lab seizures in each county of Kentucky for the year of 2011.
- Overall, the outcome is highly right-skewed, with about 20% zero observations, which is the most frequent count.
- The median number of lab seizures is 3.
- Thus if we considered our response Poisson distributed, we would expect about $e^{-3} \times 120 \approx 2$ zero observations.
- **Motivation** - How can we define a process in which excessive zeros arise?



Data Description Cont.

- Here are a summary of the counts beyond 20:

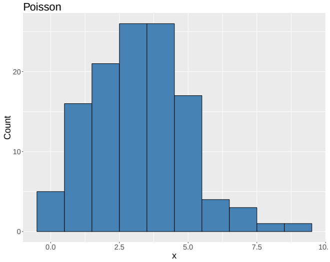
Count Range	Number
20-29	5
30-39	3
50-59	4
70-79	4
90-99	1
$>$100	2

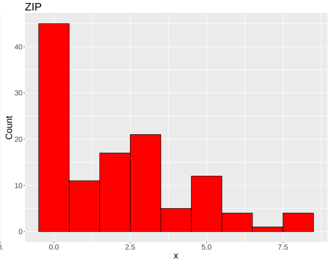
- Observe *over-dispersion* - variance exceeds mean
- Predictors include
  - The amount of pseudophedrine sold (in grams) per 100 people.
  - Socioeconomic variables such as median age, median income, percent poverty, etc.

Outline of Topics

Problems with Excessive Zero Counts

- Suppose $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^{n}$ is observed, where the response Y_i is a discrete random variable, and $\boldsymbol{X}_i$ are covariates.
- Typically, we would model such a process by Poisson or negative binomial regression.
- However, the behavior of zero counts in the observed Y_i can create difficulties.
- For example, data may structurally exclude zeros (i.e., *zero-truncation*), or have different generating processes for the zero and non-zero counts.
- Moreover, data can exhibit *zero-inflation* - excessive zeros relative to the assumed count distribution.





Zero-Generating Process

- How do zeros arise in the Y_is?
- Two common approaches - *Hurdle Models* and *Zero-Inflated Models*.
  - Both employ a mixture structure
- **Hurdle Models** - Zeros come from a different process than the positive counts
  - $f(y|\pi,\mu) = \pi I\{y=0\} + (1-\pi)I\{y \in \mathbb{Z}^+\}\Big[p(y|\mu)/(1-p(0|\mu))\Big]$ where $p(y|\mu)$ is some pmf.
  - i.e., Distribution of the response is a mix between a degenerate component at zero and a left truncated response at zero.
- **Zero-Inflated Models** - Define a latent process which says zeros come from two states - a degenerate and random state (i.e., zero comes from a count distribution)
- In this presentation, we'll focus on the latter of the aforementioned.

ZI Regression Model Definition

Definition

Let the discrete random variable Y_i be a count variable of interest and $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ be vectors of predictor variables measured for each subject i, where $i = 1, \ldots, n$. Let $p\,(y|\mu, \theta)$ be a pmf function with mean μ and dispersion/scale/heterogeneity parameter θ. The **ZI pmf** f is given by

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \mu, \theta) = \pi_i I\{y_i = 0\} + (1 - \pi_i)p(y|\mu_i, \theta)\ , \tag{1}$$

where $0 \leq \pi_i \leq 1$. Parameterizing the count distribution in terms of its mean, μ_i, we relate this quantity to the predictor vector as

$$\log(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta},$$

while the mixing proportions are typically modeled as

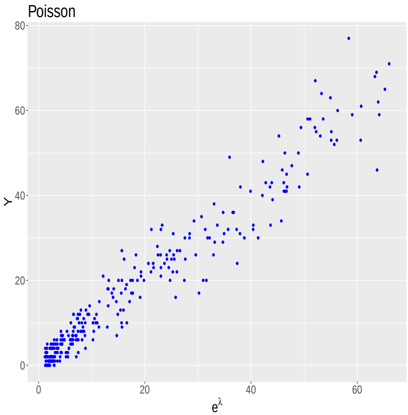
$$\text{logit}(\pi_i) = \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\alpha},$$

where the predictors in $\boldsymbol{z}_i$ may be uncoupled from those predictors in $\boldsymbol{x}_i$.

ZI Reg Continued

- Did zero come from the degenerate or count component?
- Ex: Number of visits to a physician in a year
  - A patient could have zero visits because they were never ill, and thus had no reason to visit a doctor (strategic zero).
  - Or, maybe the patient was ill, but didn't have insurance, or followed alternative medicine (incidental zero).
- The most common choices for $p(\cdot)$ are the Poisson and negative binomial (NB).
  - NB - $p(y|\mu,\theta) = \frac{\Gamma(\theta+y)}{y!\Gamma(\theta)}\left(\frac{\mu}{\theta+\mu}\right)^y\left(\frac{\theta}{\theta+\mu}\right)^\theta$
  - The expectation and variance are μ and $\mu+\mu^2\theta$, respectively.
  - The NB is commonly used to characterize *over-dispersion* - the variability is increasing with the mean.
- For this presentation, we'll focus on the Poisson pmf

NB vs Poisson



Literature Review - Applications

- Highly utilized across various disciplines
- **Manufacturing** - ZI Poisson (ZIP) regression was first developed by Lambert (1992) to characterize the number of defects in a switchboard.
  - *Perfect State* - defects are impossible (degenerate)
  - *Imperfect State* - defects are possible (Poisson)
- **Ecology** - Martin et al. (2005) modeled woodland bird patch occupancy
  - *True Zero* - Bird species does not naturally occupy that site (degenerate)
  - *False Zero* - Bird species occurs at the site, but was not detected during study period (Poisson)
- **Insurance** - Yip and Yau (2005) discuss zero claims in automobile insurance
  - *Incidental Zero* - Did not have any issues regarding their car (degenerate)
  - *Strategic Zero* - Had issue, but did not file the claim since it was small (Poisson)

Optimization

- The observed log-likelihood for ZIP Regression is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\alpha} | \boldsymbol{y}, \boldsymbol{x}) = \sum_{y_i = 0} \log(\boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\alpha} + \exp(-e^{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}})) + \sum_{y_i > 0} (y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})) \\ - \sum_{i=1}^{n} \log(1 + \exp(\boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\alpha})) - \sum_{i=1}^{n} \log(y_i!)$$

- Could always use a gradient-based method (Newton-Raphson, IRLS, etc.)
- We'll employ the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977)

EM Algorithm for ZIP Regression

EM Algorithm for ZIP

Define

$$R_i = \begin{cases} 1 & Y_i \text{ is from the zero state} \\ 0 & Y_i \text{ is from the Poisson state} \end{cases}$$

Then, $R_i \sim \text{Bern}(\pi_i)$ and

$$Y_i | R_i \sim \begin{cases} \text{Poisson}(\mu_i) & R_i = 0 \\ 0 & R_i = 1 \end{cases}$$

Then, the log-likelihood for the complete data $(\boldsymbol{Y}, \boldsymbol{R})$ is

$$\begin{aligned}
\ell_c(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \log\left(f_{R_i}(r_i) f_{Y_i|R_i}(y_i)\right) \\
&= \sum_{i=1}^{n} \log(f_{R_i}(r_i)) + \sum_{i=1}^{n} \log(f_{Y_i|R_i}(y_i)) \\
&= \sum_{i=1}^{n} \left[r_i \text{logit}(\pi_i) - \log\left(1 + \exp(\pi_i)\right)\right] + \sum_{i=1}^{n} I\{r_i = 0\}\left[y_i \log(\mu_i) - \mu_i - \log(y_i!)\right] \\
&\propto \sum_{i=1}^{n} \left(r_i \boldsymbol{z}^{\mathrm{T}} \boldsymbol{\alpha} - \log(1 + \exp(\boldsymbol{z}^{\mathrm{T}} \boldsymbol{\alpha}))\right) + \sum_{i=1}^{n} (1 - r_i)\left(y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})\right) \\
&= \ell_c(\boldsymbol{\alpha}) + \ell_c(\boldsymbol{\beta})
\end{aligned}$$

EM Algorithm for ZIP Regression Cont.

EM Algorithm for ZIP Cont.

Iterate from $k = 0, 1, \ldots$ til convergence :

1. **E-Step** Update posterior memberships

$$r_i^{(k)} = \mathbb{P}(R_i = 1 | y_i, \boldsymbol{\theta}^{(k)})$$

2. **M-Step** Maximize $\ell_c(\boldsymbol{\theta})$ with $R_i = r_i^{(k)}$ by the following:
   1. Maximize $\ell_c(\boldsymbol{\beta})$, which is equivalent to running a Poisson regression of y_i on $\boldsymbol{x}_i$ with weights $1 - r_i^{(k)}$. Call this $\boldsymbol{\beta}^{(k+1)}$
   2. Maximize $\ell_c(\boldsymbol{\alpha})$, which is equivalent to running logistic regression of $r_i^{(k)}$ on $\boldsymbol{z}_i$. Call this $\boldsymbol{\alpha}^{(k+1)}$.

Outline of Topics

Outline of Topics

Semiparametric Extension

- The (linear) ZIP and ZINB models are great at modeling zero-inflated data,
- Count data is commonly heteroskedastic, so the assumption of globally linear mixing proportions may be too strong.
- Thus, we propose a semiparametric extension to the ZIP model.
  - Same set-up as before, but we now let $\pi(\boldsymbol{z})$ be an arbitrary, smooth function of continuous covariates.
  - For simplicity, let $\boldsymbol{z}$ be one dimensional, although the theory can be extended to higher dimensions analogously.
  - If the dimension of $\boldsymbol{z}$ is high (above 2 or 3), then one needs to be cognizant of the "curse of dimensionality" (Bellman, 1961)

Literature Review - Semiparametric ZIP Modeling

- Lam, Xue, and Cheung (2006) developed a partially linear model for the Poisson mean, λ, with $\log(\lambda) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + m(T)$, where a single predictor T is modeled nonparametrically.
  - Made inference about m by sieve method.
  - He, Xue, and Shi (2010) later extended this with $\text{logit}(\pi) = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\alpha} + k(T)$
- Liu and Chan (2011) modeled jellyfish abundance by a GAM in both the count and degenerate state.
  - Allowed for constraints that a covariate can affect the zero-inflation state proportionally to the Poisson mean state
  - Employed penalized regression splines
- Feng and Zhu (2011) incorporated smoothing into the Poisson state, with a random intercept, to assess side effects of medication longitudinally.
  - Used Monte Carlo EM (MC-EM) algorithm.

Our Approach

- Since the ZIP regression model is a mixture model, we take inspiration from the mixtures-of-regression literature
- Quandt and Ramsey (1978) first proposed "switching regressions" where

$$\begin{aligned} y_i &= \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 + \epsilon_{1i} && \text{with probablity } \lambda \\ y_i &= \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_2 + \epsilon_{2i} && \text{with probablity } 1 - \lambda \end{aligned}$$

  where $\epsilon_{1i} \sim \mathcal{N}(0, \sigma_1^2)$ and $\epsilon_{2i} \sim \mathcal{N}(0, \sigma_2^2)$
- Jordan and Jacobs (1994) later proposed the "hierarchical-mixture-of-experts" model where the λ depend on covariates.
- There has been recent work in semiparametric mixtures of (normal) regressions by Young and Hunter (2010), and Huang and Yao (2012), where the mixing proportions are modeled locally.
- Cao and Yao (2012) employed a semiparametric mixture of binomial regressions to model rainfall in Edmonton.

Identifiability

- *Identifiability* is a key concern in the mixture setting.
- Let $\mathcal{F} = \{f(y|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ be a statistical model.
- We say $\mathcal{F}$ is identifiable if

$$f(y|\boldsymbol{\theta}_1) = f(y|\boldsymbol{\theta}_2)$$

  implies that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.
- Li (2012) proved identifiability of the ZIP regression when $\log(\lambda) = \beta_0 + \beta_1 x$ and logit$(\pi) = \pi(x)$.
- Under more general conditions, we can prove the semiparametric ZIP model is identifiable with the aid of Wang, Yao, and Huang (2014), who studied identifiability in mixtures of GLMs .

Local Likelihoods

- To learn $\pi(\cdot)$, we'll use local or smoothed likelihood method (Loader, 1999)
- Assume $Y_i|Z_i = z_i \sim f(y_i|\theta(z_i))$
- For z close to z_0, we assume $\theta(z)$ can be approximated by a polynomial.
- More formally, when $|z - z_0| < h$,

$$\begin{aligned}\theta(z) &\approx a_0 + a_1(z - z_0) + \frac{1}{2}a_2(z - z_0)^2 + \cdots + \frac{1}{d}(z - z_0)^d \\ &= \boldsymbol{a}^{\mathrm{T}} A(z - z_0)\end{aligned}$$

  where $\boldsymbol{a} = (a_0, \ldots, a_d)^{\mathrm{T}}$ and $A(v) = \left(1, v, \ldots, \frac{1}{d}v^d\right)^{\mathrm{T}}$ is a polynomial basis.
- At a fixed point z_0 in the range of z, define the smoothed log-likelihood at z_0 as

$$\ell^s_{z_0}(\boldsymbol{a}) = \sum_{i=1}^{n} w_i \log(f(y_i|\boldsymbol{a}^{\mathrm{T}} A(z - z_0))) \;, \tag{2}$$

  where $w_i = h^{-1}K\left(\frac{z_i - z_0}{h}\right)$ and $K(\cdot)$ is a kernel function with bandwidth h.
- Let $\widehat{\boldsymbol{a}}$ be the maximizer of (2).
- Then, we estimate $\widehat{\theta}(z_0) = \widehat{a}_0$.

Bandwidth Selection

- Bandwidth selection is a critical component in any kernel method.
- We propose a K-Fold likelihood-based cross-validation.
- For $k = 1, \ldots, K$, partition the data set into $\mathcal{D} = \mathcal{T}_k \cup \mathcal{R}_k$, where $\mathcal{T}_k$ is a training set and $\mathcal{R}_k$ is a test set.
- For a reasonable grid of values $h \in \{h_1, h_2, \ldots, h_t\}$, evaluate

$$CV = \sum_{k=1}^{K} \sum_{l \in \mathcal{T}_k} \log f(y_l | \widehat{\theta}(z_l))$$

  where f is the ZIP pmf
- Advantages of CV - doesn't typically miss features in the data
- Downsides of CV - computationally expensive and noisy

Bandwidth Selection Cont.

- To counter the high variability, it is recommend to repeat the CV process 30 to 50 times, and take the average of the resulting bandwidths.
- Could also consider plug-in bandwidths
  - Ex: Minimum of integrated mean square error (MISE) -

$$MISE(h) = \int \mathbb{E}\Big[\{\widehat{\theta}(z) - \theta(z)\}^2\Big] dz$$

  - Ex : Average Square Error (ASE) -

$$ASE(h) = n^{-1} \sum_{i=1}^{n} \left(\widehat{\theta}(Z_i) - \theta(Z_i)\right)^2$$

- Recent work by Kpotufe and Garg (2013) takes a confidence interval approach to choosing h.

Estimation of the Semiparametric ZIP

- The challenge in estimating our model is that in addition to the mixture structure, the model contains both (global) parametric and local components.
- Therefore, we propose a one-step "back-fitting" algorithm, which alternates between local and global estimation.
- An "EM-like" algorithm is proposed for each step.
- Define $\boldsymbol{\theta}(z_0) = (\pi(z_0), \boldsymbol{\beta}(z_0))$.

Backfitting Procedure

Backfitting Procedure

1. Initial Local Step - For the observed $\mathcal{Z} = \{z_1, \ldots, z_n\}$, maximize the local log-likelihood at each $z_i \in \mathcal{Z}$

$$\ell_1^S(\boldsymbol{\theta}(z_i)) = \sum_{j=1}^{n} K_h(z_j - z_i) \log f(y_i | \boldsymbol{x}_i, z_i, \boldsymbol{\theta}(z_j)) \tag{3}$$

   To do this use an "EM" Algorithm analogous to the parametric EM Algorithm. Call these estimates $\widetilde{\pi}(z_j)$ and $\widetilde{\boldsymbol{\beta}}(z_j)$.

2. Global Step - Given the mixing proportions estimates $\widetilde{\pi}(z_i)$ for $i = 1, \ldots, n$, perform a global estimation of $\boldsymbol{\beta}$ by maximizing

$$\ell_2(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i | \boldsymbol{\beta}, \widetilde{\pi}(z_i), \boldsymbol{x}_i, z_i) \tag{4}$$

   As before, an EM algorithm is implemented for estimation. Call this estimate $\widehat{\boldsymbol{\beta}}$.

3. Final Local Step - Given the global estimate of $\boldsymbol{\beta}$, update the mixing proportions at each z_i by maximizing

$$\ell_3^S(\pi(z_i)) = \sum_{j=1}^{n} K_h(z_j - z_i) \log f(y_i | \pi(z_i), \widehat{\boldsymbol{\beta}}, \boldsymbol{x}_i, z_i) \tag{5}$$

   Again, this is done by an "EM" Algorithm. Call this estimate $\widehat{\pi}(z_i)$.

Outline of Topics

Asymptotic Properties of Estimators

Asymptotic Properties of Estimators

1. Let $\boldsymbol{\theta}(z) = (\pi(z), \boldsymbol{\beta}(z))$. Assume $\sqrt{nh} \to \infty$ and $h \to 0$.
   Then, the estimator at Step 1 has

$$\sqrt{nh}\{\widetilde{\boldsymbol{\theta}}(z) - \boldsymbol{\theta}(z) - \widetilde{b}(z)h^2 + o(h^2)\} \xrightarrow{L} N(0, g^{-1}(z)\mathcal{I}^{-1}(z)v) \tag{6}$$

2. The estimator of $\boldsymbol{\beta}$ at Step 2 has

$$\sqrt{n}\{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\} \xrightarrow{L} N(0, B^{-1}\Sigma B^{-1}) \tag{7}$$

3. Finally, the final estimator of the mixing proportions $\pi(z)$ has

$$\sqrt{nh}\{\widehat{\pi}(z) - \pi(z) - \widehat{b}(z)h^2 + o(h^2)\} \xrightarrow{D} N(0, g^{-1}(z)\mathcal{I}_{\pi}^{-1}(z)v) \tag{8}$$

It can be shown that the asymptotic bias and variance of $\widehat{\pi}(z)$ is smaller than $\widetilde{\pi}(z)$.

Ascent Properties

- The classic EM Algorithm possess the *ascent property* - the objective function increases at each iteration.
- Can't claim the overall likelihood increases at each iteration, but weaker ascent properties can be established.

Ascent Property (Huang & Yao, 2012)

1. **Asymptotic Ascent** For the "EM" Algorithm in step one, if $nh \to \infty$ and $h \to 0$, it follows

$$\liminf_{n\to\infty} n^{-1}\left[\ell_1^S(\boldsymbol{\theta}^{(k+1)}(z)) - \ell_1^S(\boldsymbol{\theta}^{(k)}(z))\right] \geq 0$$

   in probability.

2. The ascent property in Step 2 follows immediately from the theory of ordinary EM Algorithms
3. For the EM Algorithm in Step 3, the local likelihood will be monotonically increasing at any z; that is, $\ell_3^S(\pi^{(k+1)}(z)) \geq \ell_3^S(\pi^{(k)}(z))$

Outline of Topics

Generalized Likelihood Ratio Test

- We are interested in testing

$$H_0 : \pi(z) \equiv \pi_0$$
$$H_1 : \pi(z) \not\equiv \pi_0$$

- And,

$$H_0 : \pi(z) \in \mathcal{M}_\alpha$$
$$H_1 : \pi(z) \notin \mathcal{M}_\alpha$$

  where $\mathcal{M}_\alpha$ is a parametric family of models.

- Fan, Zhang, and Zhang (2001) argued that the LRT is still a good test provided that the nonparametric quantity is replaced with a good estimator.

$$\lambda_n(h) = \ell(H_1) - \ell(H_0) \tag{9}$$

  where $\ell(H_1)$ and $\ell(H_0)$ is the likelihood under H_0 and H_1.

Inference Cont.

- Furthermore, the limiting distribution should be free of any nuisance parameters ($\boldsymbol{\beta}$ and the true value $\pi(z)$), and should be χ^2 with r_n degrees of freedom under H_0.
- However, calculating r_n can be intractable.
- But, we can employ a parametric bootstrap to estimate the null limiting distribution, and then obtain a bootstrap test.
- **Issues** Negative bootstrap LRT statistics.
  Could be due to:
  - Convergence rates of H_0 are faster than H_1.
  - Non-negligible smoothing bias
- Härdle, Müller, Sperlich, and Werwatz (2004) recommends a bias-adjusted LRT statistic.

Outline of Topics

Simulation Results

- Single covariate in both states was generated from a Unif$(0, 1)$. The true $\boldsymbol{\beta}$ vector is $(.5, 1)$, and the mixing proportions were set as $\pi(x) = .2 + .75\sin(\pi x)$.
- Examined three sample sizes of $n \in \{75, 200, 400\}$
- Also compared three bandwidths of $\{n^{-2/15} \times \widehat{h}, \widehat{h}, 2\widehat{h}\}$, where $\widehat{h}$ is the CV-chosen bandwidth.
- Call these bandwidths under, cv, and over smoothed, respectively.
- Compared $\boldsymbol{\beta}$ by MSE, and π by Root of Average Squared Errors (RASE)

$$\text{RASE} = \sqrt{n^{-1} \sum_{i=1}^{n} \left[\widehat{\pi}(x_i) - \pi(x)\right]^2}$$

MSE Comparison

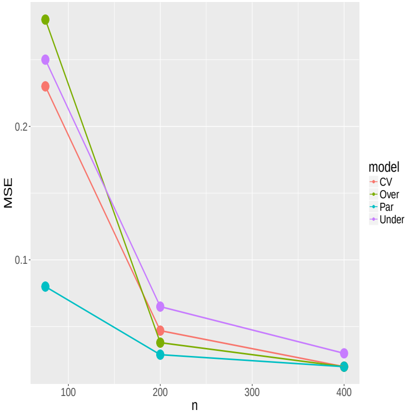


Figure: MSE of β_0

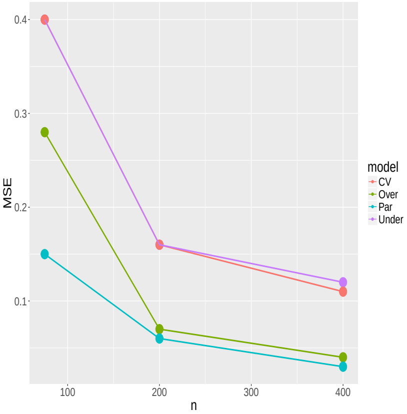


Figure: MSE of β_1

RASE Comparison

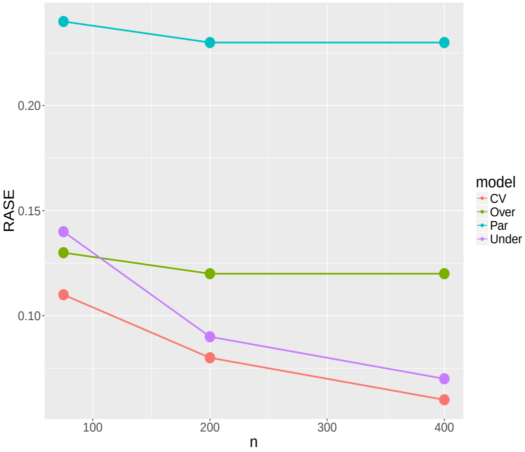


Figure: RASE Comparison

Future Simulations - Proposal

- **Accuracy Comparison**
  - More complicated forms of the mixing proportions $\pi(\cdot)$
  - Accuracy when $\boldsymbol{z}$ is of dimension 2 or 3
  - Examining bandwidth between $\widehat{h}$ and $2\widehat{h}$
- **Bootstrap Examination**
  - Coverage of bootstrap Z-Intervals for $\boldsymbol{\beta}$ and $\pi(\cdot)$
  - Bootstrap bias approximates true bias for π
- **LRT Power**
  - Power of LRT under different bandwidths and sample sizes
  - Examine hypothesis of the form

$$H_0 : \pi_0 + \delta g(z)/\sqrt{nh}$$

where $g(\cdot)$ is a smooth function and $\delta \in [0, 1]$.

- Study conditions in which the proposed model fails.

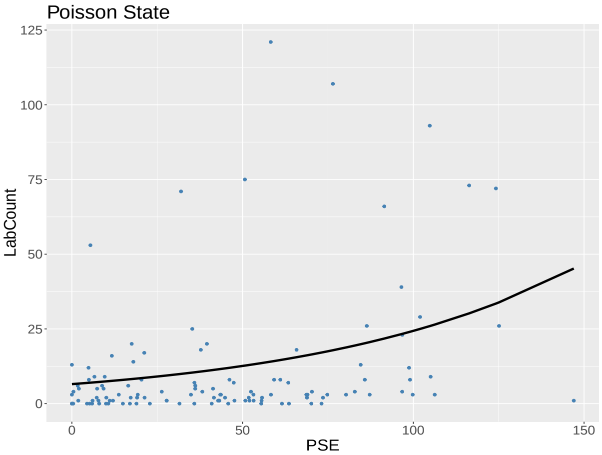
Outline of Topics

Model Fit to Meth Data

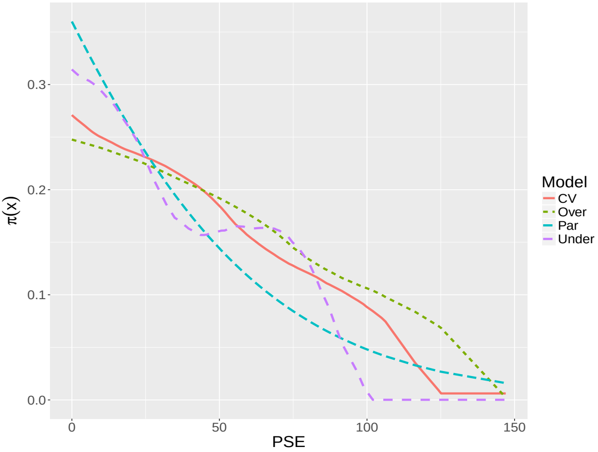
- There are $n = 120$ counties in Kentucky. The main predictor of interest is the amount (in grams) of pseudophedrine sold per 100 people (PSE).
- **Summary of Fits**

Model	h	β_0 (s.e.)	β_1 (s.e)	ℓ_0
Parametric	*	1.88 (.06)	.01 (.001)	-1248.15
Under	28.89	1.88 (.06)	.01 (.001)	-1247.04
CV	54.70	1.88 (.05)	.01 (.001)	-1248.99
Over	75	1.88 (.06)	.01 (.001)	-1249.49

Poisson State Fit

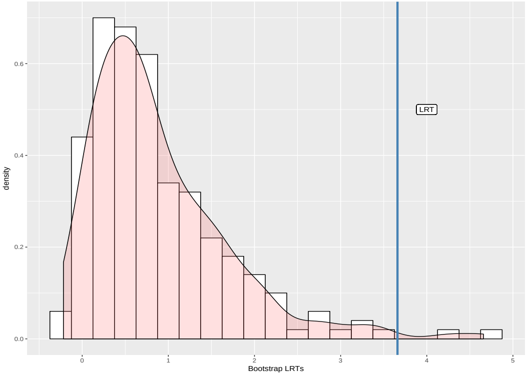


Comparison of Mixing Proportions



LRT Application

- We tested $H_0 : \pi(x) = \pi_0$ versus $H_1 : \pi(x) \neq \pi_0$
- The test statistic is $\lambda_n(h) = 3.66$ where $h = 54.7$ is the CV bandwidth.
- The bootstrap p-value $= .01$ on $m = 200$ bootstrap samples.



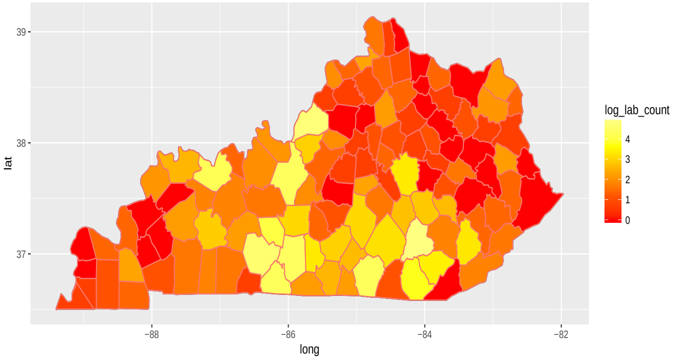
Multivariate Analysis

- Based on a previous analysis, median family income and median age were found to affect the count component, but not the zero inflation probability.
- Thus, we fit the both the parametric and semiparametric models with covariates:
  - Poisson - median income (scaled by 1000), median age, and PSE
  - Zero - PSE
- Model Comparison

Model	h	β_0	β_1	β_2	β_3	ℓ_0
Parametric	*	4.85 (.42)	-.03 (.01)	-.05 (.01)	.01 (.001)	-1224.22
Under	28.89	4.85 (.45)	-.03 (.01)	-.05 (.01)	.01 (.001)	-1223.11
CV	54.7	4.85 (.40)	-.03 (.01)	-.05 (.01)	.01 (.001)	-1225.03
Over	75	4.85 (.45)	-.03 (.01)	-.05 (.01)	.01 (.001)	-1225.57

Outline of Topics

Heat Map for KY



Summary by ADD Level

- Summary by ADD (Area Development District) Level -

ADD	n_j	Mean	Median	# of Zeros
Barren River	10	35.9	23	0
Big Sandy	5	2.2	0.0	3
Bluegrass	17	3.71	1.0	3
Buffalo Trace	5	.80	0.0	3
Cumberland Valley	8	28.63	15.0	1
FIVCO	5	4.00	3.0	0
Gateway	5	.80	0.0	3
Green River	7	14.00	5.0	1
Kentucky River	8	6.75	3.5	1
KIPDA	7	19.86	6.0	1
Lake Cumberland	10	17.30	12.5	0
Lincoln Trail	8	12.50	5.0	1
Northern Kentucky	8	3.00	2.5	1
Pennyrile	9	4.33	3.0	3
Purchase	8	2.38	1.5	2

- Could consider clustering at ADD Level

Spatial ZIP - Literature Review

- Agarwal, Gelfand, and Citron-Pousty (2002) employed a conditionally autoregressive prior (CAR) to model abundance of isopod nest burrows.
- Neelson, Ghosh, and Loebs (2013) also developed a CAR prior for the Poisson hurdle model to study ER visits in Durham County, NC.
- Hoef and Jansen (2007) applied the CAR random effect and a AR(1) for the spatio-temporal correlation to investigate harbor seal abundance.
- All are estimated through a Bayesian approach.

CAR Model with Proposed Extension

- Homogeneous CAR Gaussian model due to Cressie (1993)
- Let $Y(\boldsymbol{s})$ be the outcome at lattice location $\boldsymbol{s}$.
- Then,

$$Y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j)\}_{j\neq i} \sim \mathcal{N}\left(\mu_i + \rho \sum_{j=1}^{n} c_{ij}\Big(y(\boldsymbol{s}_j) - \mu_j\Big), \tau^2\right)$$

  where $c_{ii} = 0$, $c_{ij} = 0$ unless sites i and j are spatial neighbors.
- Besag (1974) then derived (using Brook's Lemma) the joint distribution of $\boldsymbol{Y}$ is

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \tau^2(I - \rho C)^{-1})$$

CAR Prior for ZIP

- Spatial-ZIP with CAR - Employ mixed model

$$\log(\mu_i) = \log(\mu(\boldsymbol{s}_i)) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \delta_{1i}$$
$$\text{logit}(\pi_i) = \text{logit}(\pi(\boldsymbol{s}_i)) = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha} + \delta_{2i}$$

  where $\boldsymbol{\delta}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2(I - \rho_k C)^{-1})$ for $k = 1, 2$.
- The above assumes the spatial process for the degenerate and count component are independent.
- Neelson et al. (2013) relaxed this by introducing a bivariate CAR for the two spatial processes.
- Let $\boldsymbol{\psi}_i = (\delta_{1i}, \delta_{2i})^{\mathrm{T}}$
- The likelihood is

$$L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\psi}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i, \boldsymbol{\psi}_i) f_{\boldsymbol{\psi}_i}(\boldsymbol{\psi}_i)$$

  which is difficult to optimize from a frequentist perspective.
- Thus, MCMC is utilized, typically with "flat" priors on the parameters, to make inferences.

Proposed Extension to CAR ZIP

- Extensions to CAR
  - Incorporate smoothing into both components in a partially linear model:

$$\log(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f_2(w) + \delta_{1i}$$
$$\text{logit}(\pi_i) = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha} + f_1(w) + \delta_{2i}$$

  - Integrate a time effect to analyze the 2011 and 2012 data, simultaneously
  - Model mixing proportions as $\text{logit}(\pi_i) = f_1(\boldsymbol{w})$ using tensor splines?
- Also look at simpler analysis with random effect of ADD?
  - Employ the EM algorithm of Hall (2000) incorporating Gaussian quadrature in E-Step?

Outline of Topics

Conclusions

- Zero-Inflated models are a great way to account for excessive zeros in the response.
- The semiparametric ZIP model provides flexibility in modeling zero inflation, and can be a confirmation of the parametric model.
- The Generalized LRT is a promising technique for semiparametric inference.
- The CAR ZIP model is typically employed for spatial zero-inflated data.

Future Directions

- Semiparametric Work
  - Extend to ZINB
  - Refine Bandwidth selection
  - Computationally - EM Algorithms are slow.
- Spatial Model
  - Extend to ZINB
  - More novel spatial methods - scan statistic? (Kulldorff, 1997)

References I

Agarwal, D., Gelfand, A., & Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, *9*(4), 341–255.

Bellman, R. (1961). *Adaptive control processes*. Princeton, NJ: Princeton University Press.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, *36*, 192–236.

Cao, J., & Yao, W. (2012). Semiparametric mixture of binomial regression with a degenerate component. *Statistical Sinica*, *22*, 27–46.

Cressie, N. A. (1993). *Statistics for spatial data*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, *39*(1), 1–38.

Fan, J., Zhang, C., & Zhang, J. (2001). Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics*, *29*(1), 153–193.

Feng, J., & Zhu, Z. (2011). Semiparametric analysis of longitudinal zero-inflated count data. *Journal of Multivariate Analysis*, *102*, 61–72.

References II

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, *56*, 1030 – 1039.

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin,Germany: Springer-Verlag Berlin Heidelberg.

He, X., Xue, H., & Shi, N.-Z. (2010). Sieve maximum likelihood estimation for doubly semiparametric zero-inflated models. *Journal of Multivariate Analysis*, *101*, 2026–2038.

Hoef, J. M. V., & Jansen, J. K. (2007). Space-time zero-inflated count models of harbor seals. *Environmetrics*, *18*, 697–712.

Huang, M., & Yao, W. (2012). Mixture of regression modesl with varying mixing proportions : A semiparametric approach. *Journal of the American Statistical Association*, *107*, 711–424.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchiacal mixtures of experts and the em algorithm. *Neural Computation*, *6*, 181–214.

Kpotufe, S., & Garg, V. K. (2013). Adaptivity to local smoothness and dimension in kernel regression. *Neural Information Processing Systems*.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, *26*, 1481–1496.

References III

Lam, K., Xue, H., & Cheung, Y. (2006). Semiparametric analysis of zero-inflated count data. *Biometrics*, *62*(4), 996–1003.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, *34*(1), 1–14.

Li, C.-S. (2012). Identifiability of zero-inflated poisson models. *Brazilian Journal of Probability and Statistics*, *26*(3), 306–312.

Liu, H., & Chan, K.-S. (2011). Generalized additive models for zero-inflated data with partial constraints. *Scandinavian Journal of Statistics*, *38*(4), 650-665.

Loader, C. (1999). *Local regression and likelihood*. New York City, New York: Springer-Verlag New York.

Martin, T., Wintle, B., Rhodes, J., Kunhert, P., Field, S., Low-Choy, S., . . . Possingham, H. (2005). Zero tolerance ecology : Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, *8*(11), 1235–1246.

Neelson, B., Ghosh, P., & Loebs, P. F. (2013). A spatial poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society, Series A*, *176*(2), 389–413.

References IV

Quandt, R. E., & Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, *73*(364), 730–738.

Wang, S., Yao, W., & Huang, M. (2014). A note on the identifiability of nonparametric and semiparametric mixtures of glms. *Statistics and Probablity Letters*, *93*, 41–45.

Yip, K., & Yau, K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, *36*(2), 153–163.

Young, D., & Hunter, D. (2010). Mixtures of regressions with predictor dependent mixing proportions. *Computational Statistics & Data Analysis*, *54*, 2253-2266.