Probabilistic Modeling for Underreported Count Data

With Applications to College Campus Sexual Assault Reporting

Casey Bradshaw (cb3431)

December 17, 2020

Introduction

Underreported count data arises in many disciplines. The general setup is that some unknown number of events, z, occurs, but only a subset of those events is recorded, such that the recorded number of events is $x \leq z$. It is common to model the recording mechanism as a binomial process. That is, we imagine a latent variable z to be the true number of events occurring. Then, each event $\{1, 2, ..., z\}$ is independently recorded with probability p. This generates x, the observed number of events. This framework may be appropriate in contexts such as studying the true number of cases of an underreported disease, number of traffic accidents, or endangered animal populations remaining in the wild. In this report we will consider the application to data concerning sexual assaults reported by colleges and universities in the United States (refer to Data Glossary in Appendix 1). These institutions are required to report annual crime and fire safety statistics for their campuses and adjacent areas, and sexual assaults in particular are widely believed to be subject to underreporting. Among other uses, inference for the true counts and the reporting probabilities can provide helpful indicators of whether an increase in the reported number of assaults is mainly due to an increase in the true number of assaults (bad) or an increase in the reporting probability (good).

In Section I, we consider annual data on sexual assaults aggregated across all schools. We propose a probabilistic model (Model 1) for the process generating the true and observed counts, and perform posterior inference on the true number of assaults and the reporting probabilities via Gibbs sampling. In Section II, we consider modelling the (disaggregated) sexual assault data consisting of all the reports from individual schools. We propose a new model (Model 2) for this data, introducing several school-level features as covariates to account for the heterogeneity of schools in the sample. Inference for this model is carried out via a Metropolis-within-Gibbs scheme.

Section I

We begin by considering a straightforward model for underreported count data, which we will call Model 1. Refer to Appendix 0 for a graphical model diagram.

Let z_i be the true count for subject i, x_i the recorded count, and p_i the reporting probability for subject i. We assume the recorded counts follow a binomial distribution: $x_i | z_i \sim \text{Binom}(z_i, p_i)$.

Next, we suppose the true counts are generated by a Poisson process: $z_i | \lambda_i \sim \text{Poisson}(\lambda_i)$.

Furthermore, we assume $p_i \sim \text{Beta}(\alpha, \beta)$, and $\lambda_i \sim \text{Gamma}(a, b)$.

Note that in a completely generic case, this model has an identifiability problem, specifically with regard to λ and p. If for instance we choose non-informative priors for λ and p, we cannot distinguish between scenarios with large underlying counts but low reporting probabilities, and scenarios with low underlying counts but high reporting probabilities. We can remedy this problem either through the use of validation data, or by using domain knowledge to assign more informative priors. We will focus on the latter for the applications in this paper, as validation data is not readily available.

For this model, we can use a Gibbs sampler to approximate draws from the posterior distribution, $p(z, \lambda, p; a, b, \alpha, \beta)$. To see this, first note that we can characterize the distribution of the unobserved counts, $u_i = z_i - x_i$ for $i = 1, 2, \ldots, n$, as $u_i | x_i, p_i, \lambda_i \sim \text{Poisson}((1 - p_i)\lambda_i)$. This allows us to sample z_i from its complete conditional by drawing a sample from $\text{Poisson}((1 - p_i)\lambda_i)$ and adding x_i. Refer to Appendix 2 for proof of these details. Next, due to the choice of a gamma prior for λ_i and a Beta prior for p_i, conjugacy allows us to easily sample from the complete conditionals for λ_i and p_i as well. Refer to Appendix 3 for details of this sampling scheme.

We now consider applying Model 1 to aggregated annual data on the number of sexual assaults reported by colleges and universities in the US. For each year from 2014 to 2018, the total number of assaults reported by all schools is our data point x_i. (Reporting methodology changed in 2014, so data from earlier years is not compatible.) The total number of reported assaults has increased every year in this time period (around 10% annualized growth), while the total number of students enrolled in-person has decreased slightly.

We chose values of hyperparameters for the Gamma distribution on λ to be broadly consistent with the National Crime Victimization Survey (NCVS) analysis which reports that true sexual assault victimizations over this time period ranged from 1.1 to 1.6 per 1000 people [9] [10] [6] [7]. We consider the relevant number of students to be the total number of enrolled students minus the number of students enrolled in only distance-learning programs.

We chose values of hyperparameters for the Beta distribution on p to be broadly consistent with the NCVS estimates that roughly 23-40% of sexual assaults were reported over this time period, as well as earlier research (cite Bureau of Justice Statistics special report) that period roughly 76% of violent crimes occurring at schools were not reported to police, and that sexual assault had a relatively lower rate of reporting than other violent crimes [4].

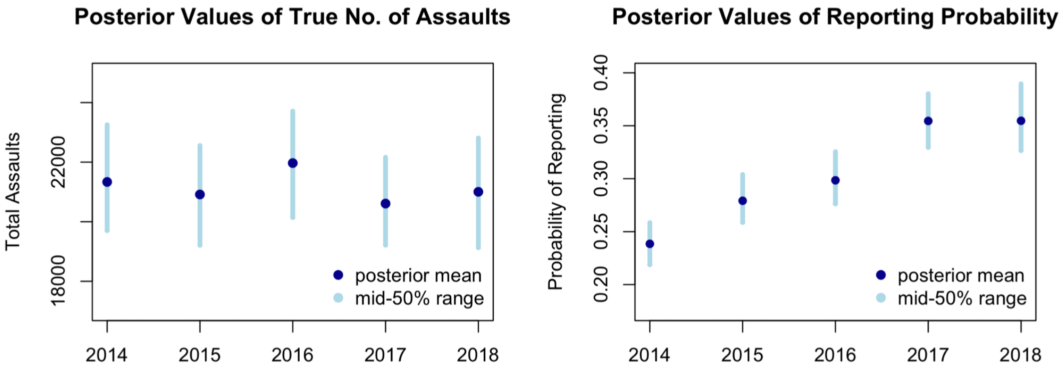


Figure 1: Posterior Summary of z and p under Model 1

Examining the posterior samples for the true number of assaults, we do not see compelling evidence of year-over-year changes. The posterior mean decreases slightly from 2014 to 2018 (a 0.5% annualized decrease), but results fluctuate over the time period in question. The range from 25th quantile to 75th quantile of the posterior samples is highlighted in the above chart. From this, we see that there are certainly plausible scenarios in whichh the true number of assaults remained unchanged or even increased slightly over 2014-2018.

The posteriors for reporting probabilities show more movement. Our posterior samples of the reporting probabilities show a clear upward trend over the time period in question. Considering the mid-50% intervals indicated on the plot, we do see compelling evidence that the reporting probability was meaningfully higher in 2018 than in 2014.

Taken together, these two results suggest that the increase in total sexual assaults reported nationwide over the 2014-2018 period is more attributable to an increase in reporting rates than an increase in the true number of assaults that occurred.

Note that Michigan State University was excluded from the data set due to extreme unreliability of its reporting systems (for which the school was fined over $4 million. In 2018 Michigan State reported over 1000 assaults, many of which were likely misattributed from previous years. For comparison, the total number of assaults reported across all schools in the data set (over 3500) ranged from 5000 to 8000 per year, so inclusion or exclusion of the Michigan State data may warrant a more thoughtful consideration. However, if Michigan State is included in the data, at a high level the story remains the same: the posterior distribution still places more mass on configurations with a small increase in the true number of assaults and a large increase in reporting probability.

Section II

In Section I, we used domain knowledge to explicitly set numerical values for the hyperparameters. Next we turn our attention to modeling the collection of individual schools' reported sexual assault numbers. For this task, we may also be interested in defining p and λ as functions of other relevant characteristics of the individual schools. One such approach, which we will call Model 2, is outlined below. This model is inspired by the approaches in [11] and [1]. In exchange for more flexibility to encode supplementary information about the individual schools, we lose some of the convenience of the first sampling scheme.

We continue to model the observed counts as binomial conditional upon the true counts and reporting probabilities. Now, we model the reporting probability p_i as a function of covariates $\mathbf{w}_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,K})$. Since p_i is constrained to the unit interval, a logit transform is one natural choice to map p_i to our covariates:

$$\begin{aligned}&\text{logit}(p_i) = \alpha_0 + \alpha * \mathbf{w}_i, \text{ where}\\ &\qquad \alpha_0 \sim \text{Normal}(\mu_0, \sigma_0^2) \text{ and } \alpha_k \sim \text{Normal}(\mu_k, \sigma_k^2)\end{aligned}$$

In the case of the sexual assault data we will use one covariate: the percent of women in the student body for school i.

In Model 2 we also continue to assume the true counts come from a Poisson distribution with parameter λ_i. Here, we express λ_i as a function of covariates $\zeta_i = (\zeta_{i,1}, \zeta_{i,2}, \ldots, \zeta_{i,P})$. Because λ_i must be positive, one natural choice is to map the log of λ_i to our (unconstrained) covariates:

$$\begin{aligned}&\log(\lambda_i) = \beta_0 + \beta * \zeta_i, \text{ where}\\ &\qquad \beta_0 \sim \text{Normal}(\mu_{\beta_0}, \sigma_{\beta_0}^2) \text{ and } \beta_p \sim \text{Normal}(\mu_{\beta_p}, \sigma_{\beta_p}^2)\end{aligned}$$

In the case of the sexual assault data, we will use two covariates: log(number of in-person students enrolled), and the campus urbanization level (urban, suburban, or rural). We treat urbanization as categorical (rather than ordinal or numeric), so the above can be equivalently expressed as: $\log(\lambda_i) = \beta_{0,\text{urbanization[school i]}} + \beta_1 * \zeta_i$.

Refer to Appendix 0 for a graphical model diagram of Model 2.

Inference for Model 2 requires a modification to Gibbs sampling, as we no longer have the conditional conjugacy structure from Model 1. Instead, we follow a Metropolis-within-Gibbs scheme as envisioned in [8]. Details of this sampling procedure are included in Appendix 4.

Examining the posterior samples of $p(z, \alpha, \beta | x, w, \zeta)$, we see that the β_0 coefficients for campus urbanization indicate a slightly lower value for suburban campuses, and little difference between urban and rural. This corresponds to a slightly lower expected number of assaults for suburban schools, all else equal (since $\mathbb{E}(z_i) = \lambda_i$). Note that we also considered a candidate model omitting campus urbanization as a covariate, but its inclusion led to a better predictive likelihood on a held-out data set.

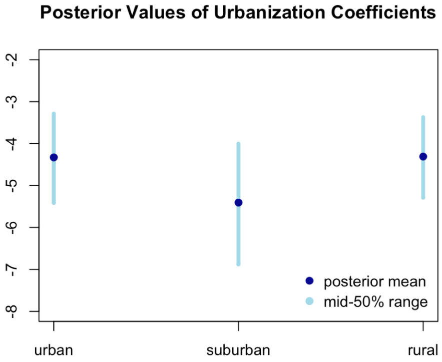


Figure 2: Posterior Samples of beta0 under Model 2

The posterior samples for β_1 center around 0.8, and are nearly all less than 1. Due to the functional form we have assumed for λ_i, a positive value of β_1 means that the expected number of assaults increases as the student population increases. In particular, a positive value of β_1 less than 1 means that that increase is sub-linear.

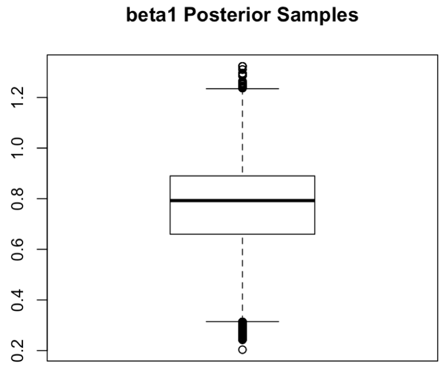


Figure 3: Posterior Samples of beta1 under Model 2

The posterior distributions for α_0 and α_1 do not reveal much information; the posterior samples are distributed very similarly to the assumed priors for the respective variables. If a relationship exists between the gender composition of the student body and the reporting probability, perhaps the relationship is of a different functional form than we have assumed in Model 2, or perhaps this dataset does not contain much information about that relationship.

Taking Columbia University as an example, we see from the reporting data that the number of reported sexual assaults decreased from 19 in 2014 to 9 in 2018. The posterior mean for the latent true number of assaults has also decreased over this time period, but to a much smaller extent, falling from roughly 50 in 2014 to roughly 43 in 2018.

Section III

This is an extension of earlier joint work with Diane Lu (Columbia University). A key weakness of the earlier work was that we were only familiar with inference procedures for models with continuous latent variables. We worked around this by marginalizing out the latent true counts z_i, but since z_i is an important quantity of interest, this was inherently unsatisfying. The sampling schemes proposed in this report allow for sampling of the discrete latent counts. Previous analysis also did not adjust student population to account for students enrolled in 100% distance-learning programs, nor did it include campus urbanization or percent of women in the student body. Adjusting student population to reflect in-person enrollment and including campus urbanization may be promising. For instance, SNHU has roughly 100,000 students, but 90,000 of them are enrolled in distance learning programs. Without accounting for this, the model would struggle to explain only 5 assaults being reported at a school with 100,000 students. Gender composition of the student body, however, does not appear helpful in explaining the differences in sexual assault reporting rates across schools, and there is still much room to improve this aspect of the modeling.

References

[1] Michaela Dvorzak and Helga Wagner. Sparse bayesian modelling of underreported count data. volume 16, pages 24–46. SAGE Publications Sage India: New Delhi, India, 2016.

[2] Peter S Fader and Bruce GS Hardie. A note on modelling underreported poisson counts. 2000.

[3] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[4] Lynn Langton, Marcus Berzofsky, Christopher Krebs, and Hope Smiley-McDonald. Victimizations not reported to the police, 2016-2010 (ncj 238536). 2012.

[5] Elias Moreno and Javier Giron. Estimating with incomplete count data a bayesian approach. *Journal of Statistical Planning and Inference*, 66(1):147–159, 1998.

[6] Rachel E. Morgan and Grace Kena. Criminal victimization, 2016: Revised (ncj 252121). 2018.

[7] Rachel E. Morgan and Jennifer L. Truman. Criminal victimization, 2017 (ncj 252472). 2018.

[8] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[9] Jennifer L. Truman and Lynn Langton. Criminal victimization, 2014 (ncj 248973). 2015.

[10] Jennifer L. Truman and Rachel E. Morgan. Criminal victimization, 2015 (ncj 250180). 2016.

[11] Rainer Winkelmann. Markov chain monte carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics*, 21(4):575–587, 1996.

Appendix 0: Extra Figures

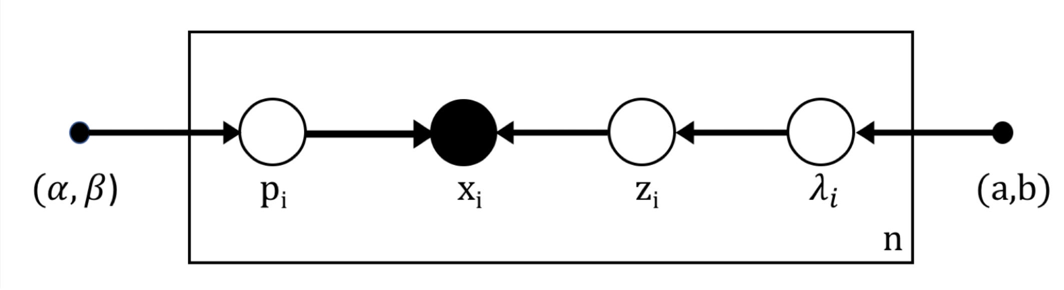


Figure 4: Graphical Model Diagram for Model 1

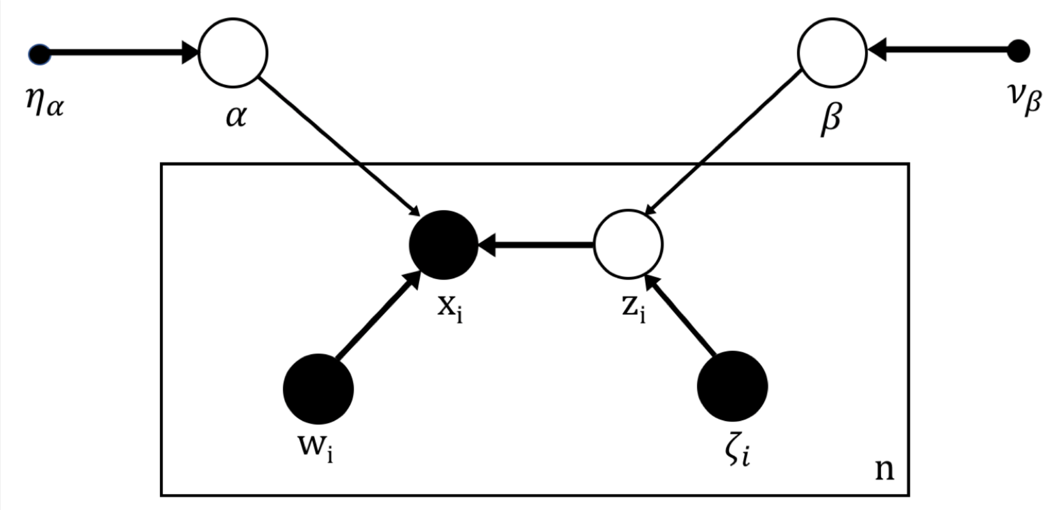


Figure 5: Graphical Model Diagram for Model 2

Appendix 1: Data Glossary

The foregoing analysis pertains to postsecondary schools in the United States which grant academic degrees, reported a positive number of students enrolled in-person in each of the five years 2014-2018, and which submitted campus security reports under the Clery Act in each of these five years. A limited number of US colleges and universities are not subject to these reporting requirements.

Data for number of reported assaults was furnished by the US Department of Education Office of Postsecondary Education, Campus Safety and Security database. All other data is available in the Integrated Postsecondary Education Data System (IPEDS), maintained by the National Center for Education Statistics (NCES), a division of the US Department of Education.

Number of reported assaults- For each school in each year, this is the number of sexual assaults disclosed in the campus security report corresponding to that calendar year

Student population- For each school in each year, this consists of the total number of students enrolled in the Fall, minus the number of students enrolled in 100% distance learning programs.

Percent female- For each school in each year, this is the percent of the student body enrolled in the fall who were classified as female (note that ”male” and ”female” are currently the only supported options in the reporting system)

Geography- Degree of urbanization is categorized on a 12-point scale defined by the NCES (link: `https://nces.ed.gov/programs/edge/docs/LOCALE_CLASSIFICATIONS.pdf` For the purpose of the foregoing analysis, this scale was collapsed into 3 categories corresponding to ”urban”, ”suburban”, and ”rural”.

Appendix 2

Consider decomposing the true count z_i into the number of reported events, x_i, and the number of unreported events, u_i. When x_i is known, the randomness in z_i comes only from u_i. Thus the distribution of $z_i|x_i, \lambda, p$ is the distribution of $u_i|x_i, \lambda, p$, shifted to the right by the observed value x_i.

Claim: The distribution of $u_i|x_i, \lambda, p$ is Poisson with parameter $(1-p)\lambda$ (as referenced in [1] and [2], among others).

First, note that

$$\begin{aligned} p(u, x|\lambda, p) &= \sum_z p(u, x|z, \lambda, p)p(z|\lambda, p) \\ &= p(u, x|z = u + x, \lambda, p)p(z = u + x|\lambda, p) \quad (*) \\ &= \binom{u+x}{x} p^x (1-p)^u \frac{e^{-\lambda}\lambda^{u+x}}{(u+x)!} \\ &= \frac{(u+x)!}{x!u!}(p\lambda)^x((1-p)\lambda)^u \frac{e^{-\lambda}}{(u+x)!} \\ &= \frac{(p\lambda)^x((1-p)\lambda)^u e^{-\lambda}}{x!u!} \end{aligned}$$

Where (*) is due to the fact that $p(u, x|z, \lambda, p) = 0$ for any value of z such that $z \neq u + z$.

Next, note that:

$$\begin{aligned} p(u|x, \lambda, p) &= \frac{p(u, x|\lambda, p)}{\sum_{u=0}^{\infty} p(u, x|\lambda, p)} \\ &= \frac{\frac{1}{x!u!}(p\lambda)^x((1-p)\lambda)^u e^{-\lambda}}{\sum_{u=0}^{\infty} \frac{1}{x!u!}(p\lambda)^x((1-p)\lambda)^u e^{-\lambda}} \\ &= \frac{\frac{1}{u!}((1-p)\lambda)^u}{\sum_{u=0}^{\infty} \frac{1}{u!}((1-p)\lambda)^u} \\ &= \frac{\frac{1}{u!}((1-p)\lambda)^u}{e^{(1-p)\lambda}} \\ &= \frac{1}{u!}((1-p)\lambda)^u e^{(1-p)\lambda} \end{aligned}$$

Which is indeed the pmf of a Poisson random variable with parameter $(1-p)\lambda$. Thus, to sample from $z_i|x_i, \lambda, p$, we can sample $u_i|x_i, \lambda, p \sim \text{Poisson}((1-p)\lambda)$ and add this to the known value of x_i.

Appendix 3: Model 1 Sampling and Diagnostics

As explained in Appendix 2, we can sample z_i from its complete conditional by drawing a sample from Poisson($(1 - p_i)\lambda_i$) and adding x_i.

To sample values for the reporting probabilities, we consider the complete conditional for p_i:

$$p(p_i|z, x, p_{-i}, \lambda; \alpha, \beta) \propto p(x_i|z_i, p_i)p(p_i; \alpha, \beta)$$

Due to conjugacy, we recognize that

$$p(p_i|x, y, p_{-i}, \lambda; \alpha, \beta) \sim \text{Beta}(\hat{\alpha}_i, \hat{\beta}_i)$$

where $\hat{\alpha}_i = \alpha + x_i$ and $\hat{\beta}_i = \beta + z_i - x_i$

Likewise for λ_i:

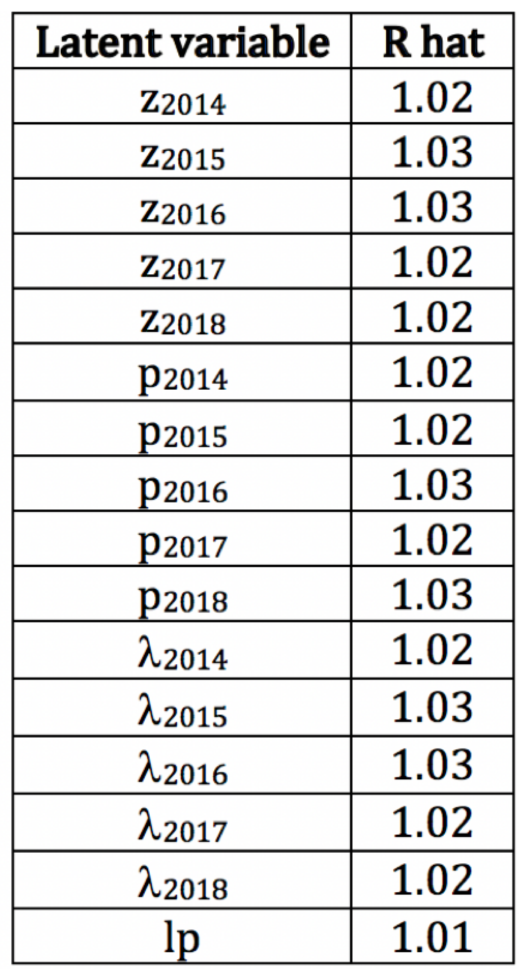
$$p(\lambda_i|x, z, \lambda_{-i}; a, b) \propto p(z_i|\lambda_i)p(\lambda_i; a, b)$$

Again due to conjugacy, we recognize that

$$p(\lambda_i|y, x, \lambda_{-i}; a, b) \sim \text{Gamma}(\hat{a}_i, \hat{b}_i)$$

where $\hat{a}_i = a + z_i$ and $\hat{b}_i = b + 1$

```
**Algorithm 1:** Gibbs Sampler
Input:
    • $x_{1:n}$ : observed counts,
    • $(\alpha, \beta)$ : hyperparameters of a Beta distribution
    • (a,b): hyperparameters of a Gamma distribution
Output: sample from the posterior $p(z, p, \lambda|x; a, b, \alpha, \beta)$
Initialize latent variables $(z, p, \lambda)$;
while the sampler has not converged do
    for $i$ in $1 : n$ do
        Sample $u_i \sim$ Poisson($(1 - p_i)\lambda_i$) ;
        Set $z_i = u_i + x_i$ ;
        Sample $p_i \sim$ Beta($\hat{\alpha}_i, \hat{\beta}_i$) ;
        Sample $\lambda_i \sim$ Poisson($\hat{a}_i, \hat{b}_i$) ;
    end
end
return $(z, p, \lambda)$
```



Latent variable	R hat
z_{2014}	1.02
z_{2015}	1.03
z_{2016}	1.03
z_{2017}	1.02
z_{2018}	1.02
p_{2014}	1.02
p_{2015}	1.02
p_{2016}	1.03
p_{2017}	1.02
p_{2018}	1.03
λ_{2014}	1.02
λ_{2015}	1.03
λ_{2016}	1.03
λ_{2017}	1.02
λ_{2018}	1.02
lp	1.01

Figure 6: R hat for latent variables in Model 1

We are satisfied with this check for convergence, as R hat values are comfortably less than 1.1 (an approximate guideline offered by [3]).

Appendix 4: Model 2 Sampling and Diagnostics

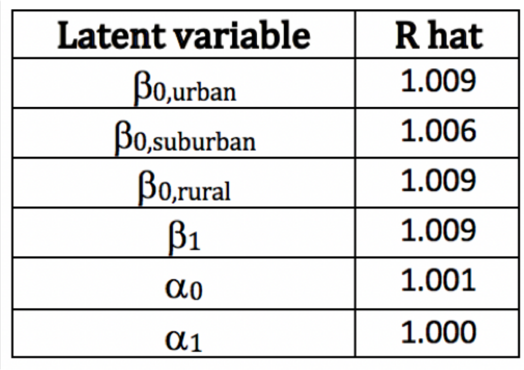
Sampling Scheme

To fit this model, we implement a component-wise Metropolis Hastings update scheme. Since the complete conditionals are available for $z_{1:n}$, these latent variables are updated via a Gibbs step, where we draw a new sample from the complete conditional and accept with probability 1. For the α and β variables, we sample from a proposal distribution and accept or reject according to the usual Metropolis Hastings rules, using the complete conditional (which we know up to a normalizing constant) as the target distribution. Proposal distributions for α and β variables were Normal distributions centered at the previous sample value of the respective variable. Variances of the proposal distributions were tuned to achieve a reasonable acceptance rate, in the 20% to 50% range.

Algorithm 2: Sampler for Model 2

```
Input:
    • $x_{1:n}$ : observed counts,
    • $w_{1:n}, \zeta_{1:n}$ : observed covariates
    • $\eta_\alpha$ : hyperparameters governing $\alpha$
    • $\nu_\beta$ : hyperparameters governing $\beta$
Output: sample from the posterior $p(z, \alpha, \beta | x, w, \zeta; )$
Initialize latent variables $(z, \alpha, \beta)$;
while the sampler has not converged do
    Update $\beta_{0,urban}$ via MH step ;
    Update $\beta_{0,suburban}$ via MH step ;
    Update $\beta_{0,rural}$ via MH step ;
    Update $\alpha_0$ via MH step ;
    Update $\alpha_1$ via MH step ;
    for $i$ in $1 : n$ do
        Sample $u_i \sim \text{Poisson}((1 - p_i)\lambda_i)$ ;
        Set $z_i = u_i + x_i$ ;
    end
end
return $(z, \alpha, \beta)$
```

Convergence Diagnostics



Latent variable	R hat
$\beta_{0,\text{urban}}$	1.009
$\beta_{0,\text{suburban}}$	1.006
$\beta_{0,\text{rural}}$	1.009
β_1	1.009
α_0	1.001
α_1	1.000

Figure 7: R hat for latent variables in Model 2

We are satisfied with this check for convergence, as R hat values are comfortably less than 1.1 (an approximate guideline offered by [3]).