MATH 640: Group 1 Final Paper, Spring 2020

Bayesian Analysis of Poisson Regression with Lognormal Unobserved Heterogeneity: Revisited

Georgetown University
Washington, DC

Georgetown University
Washington, DC

Georgetown University
Washington, DC

1 Introduction

The Poisson generalized linear model is commonly used to simulate count and rate data with the PDF and link function defined as:

$$P(y_i|\lambda_i) = \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!} \tag{1}$$

$$\log(\lambda_i) = x_i'\beta \tag{2}$$

Where β is the n-dimensional set of regression coefficients. Poisson models often perform poorly in practice because they assume that the response variable is Poisson distributed. As a consequence the Poisson model requires that the mean and variance of the response variable be equal, that is:

$$Var(y_i) = E(y_i) = \lambda_i \tag{3}$$

Since real-world data is often overdispersed, this requirement is generally not met. An alternative to Poisson regression is the Negative Binomial regression model where $\lambda_i = \exp(x_i'\beta)$, $\nu \sim Gamma(\phi, \phi)$, and $Y_i|\lambda_i, \nu \sim Poisson(\lambda_i \cdot \nu)$. This relaxes the requirement of equivalence between the mean and variance, and the relationship between them becomes:

$$Var(Y) = E(Y) + \alpha \times E(Y)^p \tag{4}$$

Where α is a scalar constant and p is generally 1 or 2. While the Negative Binomial model relaxes the restrictive assumption of the Poisson model, it retains a parametric link between the variance with the response, which is undesirable. The Negative Binomial model additionally requires that the dispersion of the data follow a Gamma distribution, which may not hold in practice.

In this paper, we implement an alternative to the Poisson and Negative Binomial regression models. This model, the Poisson Regression with Lognormal Unobserved Heterogeneity (PRLUH), is specified in E. Tsionas (2010). The advantage of this model, as discussed in Winkelmann (2008), is that it can leverage a result from the Central Limit Theorem to account for dispersion that results from independent elements omitted from the model. This result states that the sum of independent random variables converges to a normal distribution in the limit. The model leverages this result by incorporating a normally distributed error term within the link function. This model should therefore outperform Poisson and Negative Binomial models in situations where omitted variables contribute to dispersion in the data.

With this model we then attempt to recreate the results of a regression analysis conducted in E. Tsionas (2010). This analysis examined data on the research productivity of 70 different pharmaceutical companies found in Wang (1998). It regressed the number of patents (Y) on several measures of firm research and development (R&D) expenditures. A histogram of the data and of a simulated Poisson data set (n=10000, λ=mean(Y)) is shown in Figures 1 through 3 in Appendix A to demonstrate the inappropriate nature of the simple Poisson model for this data.

2 Methods

In an effort to recreate Tsionas's results, we derive a Gibbs Sampler with a Metropolis-Hastings step for the unnamed posterior conditional density of the λ_i hyperparameters. Let X be the R&D predictor values of the data set and y be the number of patents as a response. Also let β be the coefficients of the Poisson GLM from the data set, σ be the standard deviation of the dispersion term, and λ be the Poisson parameter. Finally, let $\underline{\beta}$ be the mean of the β coefficients, $\underline{H}$ be the prior precision matrix for the multivariate lognormal (so that its inverse is our covariance matrix), and let $\underline{n}$ be an imaginary sample size. Thus, Tsionas defines the full posterior distribution to be

$$\begin{aligned} p(\beta, \sigma, \lambda|y, X) \sim \sigma^{-(n+\underline{n}+1)} \left\{ \prod_{i=1}^{n} \exp(-\lambda_i)\lambda) i^{y_i - 1} \right\} \\ \times \exp\left[-\frac{1}{2\sigma^2} \{ (\log\lambda - X\beta)'(\log\lambda - X\beta) + \underline{q} \} \right] \\ \times \exp\left[-\frac{1}{2}(\beta - \underline{\beta})'\underline{H}(\beta - \underline{\beta}) \right] \end{aligned} \tag{5}$$

and defines the parameters' conditional posterior distributions to be

$$\beta|\sigma, \lambda, y, X \sim MVN[(X'X + \sigma^2 \underline{H})^{-1}(X' \log \lambda + \sigma^2 \underline{H}\underline{\beta}), \sigma^2(X'X + \sigma^2\underline{H})^{-1}] \tag{6}$$

$$\frac{1}{\sigma^2}\left[(\log \lambda - X\beta)'(\log \lambda - X\beta) + \underline{q}\right] \sim \chi^2(n + \underline{n}) \tag{7}$$

$$p(\lambda_i|\beta, \sigma, y, X) \sim \exp(\lambda_i)\lambda_i^{y_i - 1} \exp\left[-\frac{1}{2\sigma^2}(\log \lambda_i - x_i'\beta)^2\right] \tag{8}$$

Note that the distribution of σ^2 may be reparameterized as an inverse gamma, as the reader will find in sections of the supplemental code. Also note that the posterior conditional of λ_i is not recognizable, so we incorporate a Metropolis-Hastings step for the λ_i's in our Gibbs Sampler.

We used the coefficients from a log-link Poisson GLM to initialize the β parameter. We initialized σ^2 by squaring the standard deviation of the residuals from the Poisson GLM, and we initialized the λ parameter using a simple Poisson sample that was itself parametrized by the mean of the fitted values from the Poisson GLM, which is naturally equivalent to the mean of the original data's response value.

Within the Gibbs sampler, we iterated over each of the parameters above as shown. While in the Metropolis-Hastings step, we generated proposal λ_i values by first sampling a proposal variance from a normal distribution and then generating a proposal λ_i value by exponentiating that variance (see equation 10 below). We finally used these proposal λ_i values and the aforementioned normal jumping density to calculate the Metropolis-Hastings ratio and sample for the resulting λ_i. The normal jumping density was parameterized as:

$$h_{proposal} \sim N(\underline{\mu}, \underline{h}) \tag{9}$$

$$\lambda_{proposal} = \exp(h_{proposal}) \tag{10}$$

where $\underline{\mu}$ is the mean of the response values (Y) and $\underline{h} = 10$. Tsionas proposes different values in his paper, and we will address those differences in the Discussion section.

Key metrics for the model were the acceptance rate within the Metropolis-Hastings step, and proximity to Tsionas's results. Target acceptance rate was in the mid 20% range due to the 70 λ_i parameters required to sample from this model.

3 Results

Our Gibbs sampler for a PRLUH model was tested using the same empirical application as the one in Tsionas's paper. The data, taken from the National Bureau of Economic Research and Development, are for 70 pharmaceutical and biomedical companies from 1976. The dependent variable is the number of patent applications. The explanatory variables are a logarithm of each company's R&D budget; a quadratic term of the logarithm of each company's R&D budget, $\log(R\&D)^2$, to account for potential nonlinearities; and the ratio of R&D spending to company sales, to account for the relative intensity of the firm's R&D spending. We checked for potential problems with multicollinearity in the data by calculating the Pearson correlation coefficient between all predictor variables. The correlation between $\log(R\&D)$ and the ratio of R&D spending to company sales was about 0.06 indicating effectively no correlation between those variables.

Our goals for this paper were to first produce a Bayesian model that converged, and second to use that model to reproduce Tsionas's results. Although we believe we initially adhered to the paper's proposed parameters and methods as closely as possible, we encountered a variety of issues throughout our exploration.

One of the key issues with our implementation of Tsionas's model was persistent numerical overflow at the Metropolis-Hastings step. When using Tsionas's stated parameter values we found that in approximately 90% of Metropolis-Hastings samples the ratio would include 'NAN' values due to either $\log(0)$ in the numerator or zeroes in the denominator. We addressed this initially by simply coercing non-numeric outputs for the Metropolis-Hastings ratio to be zero valued (and thus rejected by the sampler). Although this produced a functionally viable model, the results were underwhelming. For reference, Tsionas' published outputs are as follows:

Table 1: Tsionas's published results

(Intercept)	R&DS	lgR&D	lgR&D.sqd	σ
0.393	0.463	0.679	0.062	.545

The results from our most faithful reproduction (described above) of Tsionas's published model produced the confidence intervals in Table 2. Though our confidence intervals do include the published value, this model did not give us great confidence in its predictive power based on the zero coercion that we introduced. Additionally, we observed an overly efficient model that accepted about 64% of λ_i proposals.

Table 2: Standard conditional approach with $\underline{h} = 100$

	(Intercept)	R&DS	lgR&D	lgR&D.sqd	σ^2
50%	-0.312	1.174	0.690	0.297	89.726
2.5%	-1.875	-3.138	-0.253	-0.030	68.586
97.5%	1.438	6.396	1.510	0.646	116.405

At this point, we re-evaluated our goals and concluded that simply reproducing Tsionas's results was less important that building an effective and explainable model. To this end, we sought to remove the need for zero-coercion in the model. Our first attempt was to replace unsatisfactory parameter values not with zeroes, but with their respective means for the Gibbs iteration. While this method seemed promising in concept, we found that under this method the λ_i's collapsed to zeroes after about 100 Gibbs iterations. We eventually discovered that taking the log of the conditional posterior for λ_i (as mentioned in the Methods section) would alleviate our issues here.

In order to bring the efficiency of the model within acceptable bounds per Roberts, Gelman, and Gilks (1997), we reduced the value of $\underline{h}$ from 100 as used by Tsionas to 10 through trial and error. The resulting acceptance rate was then about 26% which we deemed acceptable.

Table 3: Log-conditional approach with $\underline{h} = 10$

	(Intercept)	R&DS	lgR&D	lgR&D.sqd	σ^2
50%	-0.843	-0.413	0.506	-0.185	12.520
2.5%	-2.639	-4.791	-0.643	-0.552	6.443
97.5%	0.972	3.188	1.683	0.168	22.776

Still somewhat dissatisfied with the outputs, we turned to additional parameter tuning and found the $\underline{\mu}$ term to be the next most impactful, as it parameterized the $\lambda_{proposal}$'s mean. Rather than "a small constant" that Tsionas proposes in his paper, we used the mean of the Y values, or approximately 23. We also tried the median of the Y values to compensate for the overdispersion of the responses, but saw worse results than simply using the mean. Table 4 displays the results of these adjustments on the regression coefficients and σ^2.

Table 4: Log-conditional with $\underline{h}$=10, $\underline{\mu}$=mean(Y)

	(Intercept)	R&DS	lgR&D	lgR&D.sqd	σ^2
Tsionas	0.515	0.407	0.854	0.017	5.286
Group's	-0.600	1.277	0.812	0.313	3.110

We can see that the stated parameters provide us with the closest approximation of Tsionas's published results, but they produce an unacceptably low efficiency of 4.5%. We did continue experimenting with variance values at this point, but acceptable efficiency rates produced undesirable parameter values.

4 Discussion

We initially followed Tsionas's work as closely as possible to set up our Gibbs sampler. A multivariate normal distribution was used as the prior for the β coefficients, and the resulting the full conditional posterior distribution of the β coefficients was a multivariate normal distribution as well. The prior of β was a multivariate normal with mean vector zero, and a diagonal precision matrix whose diagonal elements were equal to 10^{-3}.

To generate the latent variable λ_i, for which the full conditional posterior was not available in closed form, a Metropolis-Hastings step was used in the sampler. A normal distribution for $h_i = \log(\lambda_i)$ was used to generate proposals for all λ_i's with each step. In an attempt to address the issue of numerical overflow, we modified the Metropolis-Hastings step such that it utilized the log full conditional posterior for the λ_i's and took the log of the MH ratio. While this addressed the issues related to numerical overflow, it ultimately resulted in the parameters not converging. Thus this approach was not used in the final sampler.

When parameters $\underline{\mu}$ and $\underline{h}$ were set as -2 and 100^2, the mean acceptance rate for all λ_i's was 64.38%. This was considered too high an acceptance rate, per Roberts, Gelman, and Gilks (1997). When $\underline{h}$ was reduced to 10, the mean acceptance rate for all λ_i's was reduced to 26.46%, which was considered acceptable. We will call this model Model A in future references. In general, we were satisfied with the convergence using Model A. Note the qualitative convergence in the running means plots for β parameters and for σ^2 below.

Figure 1: Model A Diagnostics for β Parameters

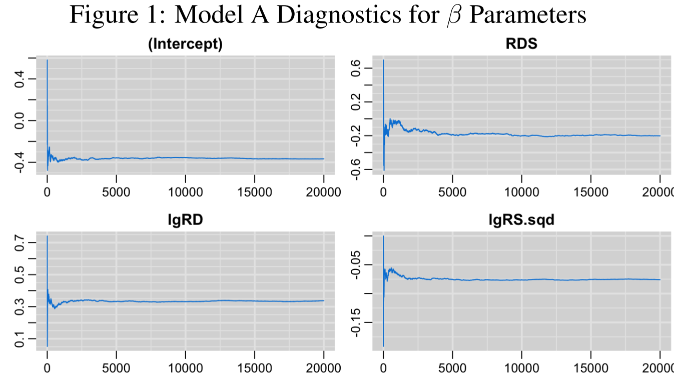
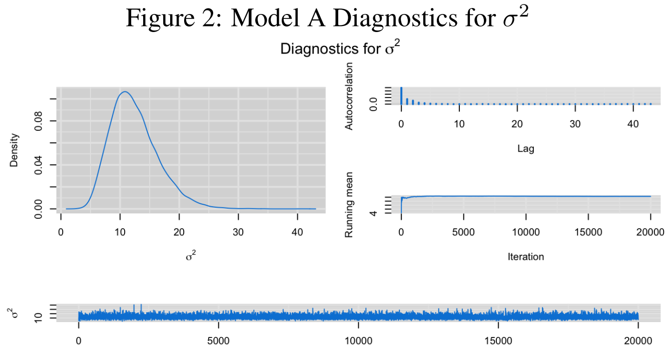


Figure 2: Model A Diagnostics for σ^2



Next, we generated another model in which we also changed $\underline{\mu}$ to be the mean of the Ys. This brought the estimates of the β values closer to Tsionas's (see Table

4 above), but the acceptance rate for the Metropolis-Hastings step was only 4.4%, below the range considered acceptably by Roberts, Gelman, Gilks (1997). This model will be referred to as Model B.

Visual inspection of the trace plots and running mean plots for the β coefficients and σ^2 from the 20,000 retained samples indicated that these parameters had successfully converged in Model A and almost all parameters had successfully converged in Model B. ACF plots for Model A indicated sufficiently low autocorrelation for each parameter. The parameters in Model B showed reasonably high autocorrelation, though this could be mitigated through thinning. In addition, Geweke's diagnostic suggested convergence for all parameters in both models except for the "Intercept" parameter in Model B, which was just slightly below $p = 0.05$. This indicates that we should reject the null hypothesis of constant means between the two tested segments of the "Intercept" parameter in Model B and conclude non-convergence. We ran this model for an additional 70,000 samples (total of 100,000), and used the same burn-in to achieve sufficient convergence from the same qualitative and quantitative metrics, but we should highlight that the acceptance rate remained 4.4%. Qualitative results from the non-converging 20,000 retained samples are shown below.

Figure 3: Model B Diagnostics for β Parameters

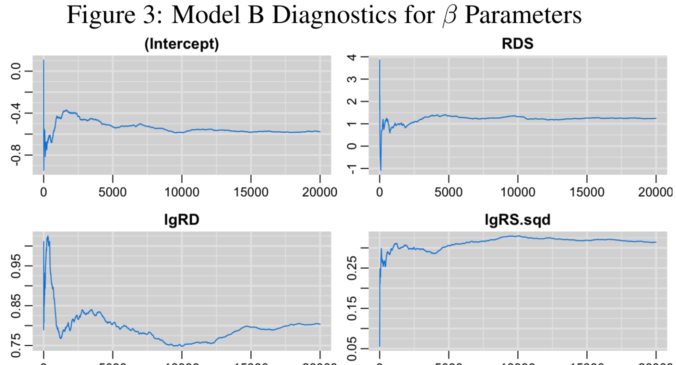
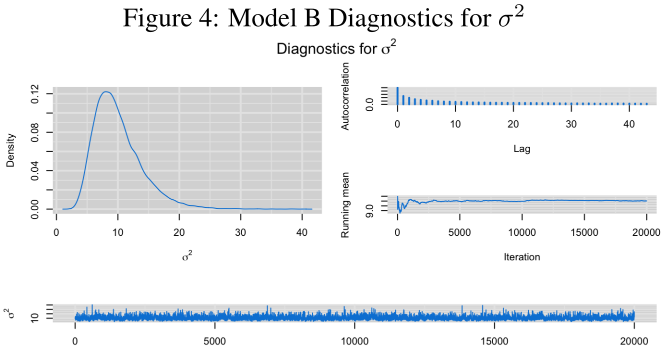
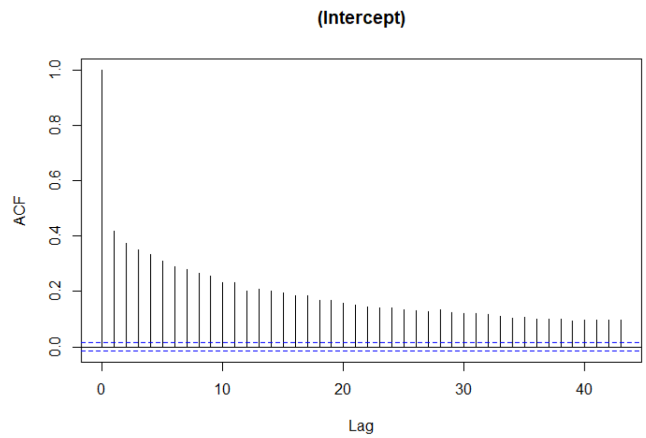


Figure 4: Model B Diagnostics for σ^2



Note that the ACF plot for "Intercept" is representative of all β parameters from Model B, thought it was the only one that failed the Geweke diagnostic test.

Figure 5: Model B Autocorrelation for β "Intercept"



In summary, Model A achieved acceptable efficiency, though the results differed noticeably from Tsionas's published results. The results of Model B were closer to Tsionas's results, but the acceptance rate for the Metropolis-Hastings step was below what is considered acceptable.

5 Conclusions

As stated above, the results of our Gibbs sampler with an MH step were ultimately different from the published results in Tsionas's paper. Regardless, it is clear that the parameter values generated by the sampler converged, and thus the Bayesian implementation of a Poisson regression model with lognormal unobserved heterogeneity is possible, given that the model is properly specified.

This exercise in constructing a Gibbs sampler was instructive in introducing some of the challenges of Bayesian analysis when working with Markov Chain Monte Carlo methods to test latent variable regression models for count data. Numerical overflow in the calculation of posterior values that involve logarithms and exponentiation was one such challenge, and working with the log-conditional posterior and log-MH ratios offer one potential workaround. Another challenge came from specifying the values of model parameters. Altering the parameters of the priors even slightly made large differences in the acceptance rate of the λ_is as well as the final values obtained for the β coefficients and σ^2. Additionally, the lack of reproducible code made constructing the Gibbs sampler, especially the Metropolis-Hastings step, a considerably more challenging exercise.

6 Replication Code

Our replication code can be found in the file `Group1_replication_code.R` alongside this manuscript.

7 Citations

Brannas, Kurt. "Conditional Heteroskedasticity in Some Common Count Data Models for Financial Time Series Data." *SSRN Electronic Journal*, 2002. https://doi.org/10.2139/ssrn.336440.

Cameron, A. Colin, and Pravin K. Trivedi. "12 Count Data Models for Financial Data." In *Handbook of Statistics*, 14:363–91. Statistical Methods in Finance. Elsevier, 1996. https://doi.org/10.1016/S0169-7161(96)14014-1.

Gelman, A., W. R. Gilks, and G. O. Roberts. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms." *The Annals of Applied Probability* 7, no. 1 (February 1, 1997). https://doi.org/10.1214/aoap/1034625254.

Tsionas, Efthymios G. "Bayesian Analysis of Poisson Regression with Lognormal Unobserved Heterogeneity: With an Application to the Patent-RD Relationship." *Communications in Statistics - Theory and Methods* 39, no. 10 (May 12, 2010): 1689–1706. https://doi.org/10.1080/03610920802491774.

Wang, Peiming, Iain M. Cockburn, and Martin L. Puterman. "Analysis of Patent Data: A Mixed-Poisson-Regression-Model Approach." *Journal of Business & Economic Statistics* 16, no. 1 (January 1998): 27. https://doi.org/10.2307/1392013.

Winkelmann, Rainer. *Econometric Analysis of Count Data*. 5th ed. Berlin: Springer, 2008.