"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518, USA), Vol.12, Issue 9, 1-13 | Science Signpost Publishing

Juncture Models: Trivariate Nonlinear Regression Models and Application to In-Hospital Quality of Care

Manasi Sheth[1], N. R. Chaganty[2],

1. *University of Wisconsin, Whitewater, WI, USA*
2. *Old Dominion University, Norfolk, VA, USA*

Abstract: Mortality, readmission, and length of stay (LOS) are commonly used measures for quality of hospital care. These measures are correlated with each other, so we propose a new way to combine these three variables to evaluate hospital care quality of care. In this paper, we investigate the joint probability distribution of these three variables using what we named as the Juncture Distribution Models and characteristics of the Juncture models.

Key words: Bernoulli, Poisson, Mortality, Readmission, Length of Stay, Juncture Models

1 Introduction

Outcome measures for quality of care are increasingly used to monitor and compare hospital performance, with the aim to identify areas of improvement. Three outcome measures that are commonly used in various countries to evaluate quality of care in hospitals are in-hospital mortality, re-admissions and long length of stay (LOS) (See Bottle (2013)). However, these outcomes are interrelated with each other and will affect the interpretation of hospital outcomes. For example, patients who die in hospital cannot be readmitted, so that in theory hospitals may have low readmission rates due to relatively high mortality rate. Such mechanisms need to be understood for outcomes measures to be able to reflect quality of care.

A previous study among Medicare patients hospitalized for acute myocardial infarction, heart failure or pneumonia did not find an association for mortality and readmission, except a weak association for mortality and readmission for heart failure patients (See Krumholz (2013)). However, this study only examined hospital-level and not patient-level associations. Both of types of associations need to be considered to evaluate whether interventions to reduce mortality may raise readmission rates or whether the two measures have shared underlying processes.

Furthermore, to obtain a comprehensive picture of quality - especially for between-hospital comparisons - it is attractive to jointly report outcome measures as e.g. mortality may be very low in some patient groups but re-admissions are common. In addition, combining multiple outcome measures has the advantage of an increased number of events per hospital, resulting in more reliable estimates (lower statistical uncertainty) and thereby more reliable comparisons of hospitals (See Bottle (2013)).

Hospital LOS is often used as a measure of a post-medical procedure outcome, as a guide to the benefit of a treatment of interest, or as an important risk factor for adverse events. Therefore, understanding hospital LOS variability is always an important healthcare focus. Hospital LOS data can be treated as count data, with discrete and non-negative values, typically right skewed, and often exhibiting excessive zeros. In this paper, we aim to (1) disentangle the relationship between mortality, readmission and long LOS both at patient and hospital level, and 2) to develop a measure to jointly report the three outcomes for a more comprehensive, unambiguous and more reliable estimate of hospital - specific quality of care to be used with a hospital over time and between-hospital comparisons.

Until now, each of the three variables were investigated individually by a variety of researchers and investigators. There were few papers that discussed the associations between two of three variables, however, there were no statistical papers that studied all three outcomes together. In this paper, we present the performance of the Juncture (Bernoulli, Bernoulli, Poisson) models as well as Juncture (Bernoulli, Bernoulli, Negative Binomial) models using real data. Data repository is as shown below.

Real-life count data often exhibit two (related) characteristics: over-dispersion and k-inflation. Over-dispersion refers to an excess of variability in the data (i.e., the variance exceeds the mean), while the k-inflation refers an excess of count value of k (See Cameron and Trivedi (1998)). In the presence of over-dispersion, the Poisson model is not adequate and can lead to biased estimates and unreliable standard errors estimates and can be modelled by Negative Binomial Distribution (See Cameron and Trivedi (1998)). In a subsequent paper, Zero-Inflated Poisson (ZIP) and Doubly - Inflated Poisson (DIP) distribution models using real data will be presented.

In this paper, we will first present a gentle introduction to the Juncture Models with some distributional properties. Further, we will outline the most commonly used method of estimation for the parameters of the distribution models, i.e. maximum likelihood estimation (MLE). For this section, we will assume that the data consists of independent counts of our outcome variables and co-variates that are predictors.

2 2. Methodology Section

2.1 Review of Pertinent Distributions

2.1.1 Bernoulli

The Bernoulli distribution is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$. If Y is a random variable with a Bernoulli distribution, then:

$$P(Y = y) = \begin{cases} p, & \text{if } y = 1 \\ q, & \text{if } y = 0 \end{cases}$$

The Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted (so n would be 1 for such a binomial distribution with the parameters n and p). The mean or the expected value of Y is given by p and variance by pq. (insert reference)

2.1.2 Poisson

The most widely used and the most basic model that explicitly considers the nonnegative integer-valued aspect of the count outcome variable is the Poisson regression model [insert reference here]. Let Y_i, $i = 1, ..., n$, be random variables for the number of occurrences of the event of interest and its realizations $y_i = 0, 1, 2, ...$ Poisson distribution assumes that the outcome variable Y_i is independently distributed with

$$P(Y_i = y_i) = \frac{\exp -\mu_i \mu^{y_i}}{y_i!}, y_i = 0, 1, 2, ... \tag{1}$$

and the mean parameter (i.e. the mean number of events per period) is given by μ_i. For Poisson distribution, the mean and variance are assumed to be equal.

2.1.3 Negative Binomial

The Poisson regression model can be generalized by introducing an unobserved heterogeneity term for observation i. The subjects are assumed to differ randomly in a manner that is not fully accounted for by the observed covariates. This is formulated as:

$$\begin{aligned} E(Y_i|X_i = x_i, \tau_i) &= \mu_i \tau_i \\ &= \exp x_i' \beta + \epsilon_i, \end{aligned}$$

where the unobserved heterogeneity term τ_i is independent of the vector of the predictor variables x_i. Then the conditional distribution of Y_i on $X_i = x_i$ is Poisson with conditional mean and conditional variance $\mu_i \tau_i$:

$$f(Y_i = y_i|X_i = x_i, \tau_i) = \frac{\exp -\mu_i \tau_i (\mu_i \tau_i)^{y_i}}{y_i!}, y_i = 0, 1, 2, ...$$

The negative binomial distribution is derived as a gamma mixture of Poisson random variables (insert reference). Most real-life count data is overdispersed. Negative Binomial Distribution is commonly used to account for overdispersed data. The most commonly used Negative Binomial distribution is the NB2 model with the following distribution:

$$g(y|\mu, \eta) = \frac{\Gamma(y + \eta^{-1})}{\Gamma(y+1)\Gamma(\eta^{-1})} \left(\frac{\eta^{-1}}{\mu + \eta^{-1}}\right)^{\eta^{-1}} \left(\frac{\mu}{\mu + \eta^{-1}}\right)^{y},$$

where $\eta > 0$ and $y = 0, 1, 2, \ldots$. The mean and variance of NB2 model is μ and $\mu(1 + \eta\mu)$, respectively. If we replace η^{-1} with $\eta^{-1}\mu$, then we get the NB1 model which has the following distribution:

$$g(y) = \frac{\Gamma(y + \mu\eta^{-1})}{\Gamma(y+1)\Gamma(\mu\eta^{-1})} \left(\frac{\eta^{-1}}{1 + \eta^{-1}}\right)^{\mu\eta^{-1}} \left(\frac{1}{1 + \eta^{-1}}\right)^{y},$$

where $\eta > 0$ and $y = 0, 1, 2, \ldots$. The mean and variance of NB1 model is μ and $\mu(1 + \eta)$, respectively.

Cameron and Trivedi (1998) discussed the NBr parameterization where $\tau = \mu^{2-r}\eta^{-1}$ and $\xi = \mu^{r-1}\eta$, where $1 \leq r \leq 2$. For $r = 1$ and $r = 2$, Negative Binomial Distribution (τ, ξ) or NB (τ, ξ) is

$$g(y|\tau, \xi) = \frac{\Gamma(\tau + y)}{\Gamma(\tau)\Gamma(y + 1)} \frac{\xi^y}{(1 + \xi)^{\tau+y}}, \text{ for } y = 0, 1, 2, \ldots$$

where, τ is a real positive number and $\xi > 0$. Note that $\xi = 1/(\pi - 1)$ where π is a Bernoulli parameter. The mean is $\mu = \tau\xi$ and variance $\sigma^2 = \tau\xi(1 + \xi)$. One could also consider the generalized NBr model in which case r must be estimated.

2.2 Definitions and Data Format

For our motivating example, we consider three interrelated quality of care variables, Y_1, Y_2, and Y_3. Y_1 can be assumed to be in-hospital mortality which can be defined as death in hospital during the index admission; Y_2 can be assumed to be readmission which is unplanned (emergency) readmission to the same hospital within 30 days of discharge; and Y_3 is LOS, i.e. length of stay in the same hospital. Here, Y_2 can be considered as (1) having EVER been readmitted to the hospital, and (2) readmission occurence per patient.

Usually, the data is in the following format (See Table 1) where outcome variable is a trivariate response from the three variables mentioned above. This could apply to any other trivariate outcome response where the first variable is a binary response, second variable is a binary response and the third variable is a count variable. Furthermore, the second column is of either continuous or categorical covariates such as age, gender, race, ethnicity, health condition(s), etc. And, the final column is that of frequncy or number of patients who have these reported characteristics and reported trivariate combination of outcome measures.

2.3 Juncture Models

To obtain a comprehensive picture of in-hospital quality of care, it is important to report all three (or more) outcome measures such as in-hospital mortality, readmissions and length of stay (LOS). Thus, we need to establish an appropriate model for jointly reporting (Y_1, Y_2, Y_3). Since Y_1 and Y_2 can be assumed to be binary variable for mortality and readmission, and Y_3 which is a count variable can be modeled initially by Poisson distribution.

Juncture models are Multivariate Non-Linear Regression Models, specifically Trivariate Non-linear Regression models that can be applied to in-hospital quality of care. Juncture models for discrete outcomes will become a valuable tool in the analysis in follow-up data. These models are applicable in the following setting: when the focus is two binary outcomes and a discrete count outcome and we wish to account for the effect of covariates measured with error. Because the Juncture models are capable to provide valid inferences in setting where simpler statistical tools fail to do so, and their wide range of application, we decided to formulate these models and will try to apply these to many advances in the juncture modeling field. Currently, there is no single monograph or text dedicated to this type of models that seem to be available.

Table 1: General Layout of Raw Data with Covariates

BBP Response	Covariates			frequency
$(0,0,y_1)$	x_{11}	$\ldots$	x_{1l}	$n_{(0,0,y_1)}$
$(0,0,y_2)$	x_{21}	$\ldots$	x_{2l}	$n_{(0,0,y_2)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$(0,0,y_{m_1})$	x_{n1}	$\ldots$	x_{nl}	$n_{(0,0,y_{m_1})}$
$(0,1,y_1)$	x_{11}	$\ldots$	x_{1l}	$n_{(0,1,y_1)}$
$(0,1,y_2)$	x_{21}	$\ldots$	x_{2l}	$n_{(0,1,y_2)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$(0,1,y_{m_2})$	x_{n1}	$\ldots$	x_{nl}	$n_{(0,1,y_{m_2})}$
$(1,0,y_1)$	x_{11}	$\ldots$	x_{1l}	$n_{(1,0,y_1)}$
$(1,0,y_2)$	x_{21}	$\ldots$	x_{2l}	$n_{(1,0,y_2)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$(1,0,y_{m_3})$	x_{n1}	$\ldots$	x_{nl}	$n_{(1,0,y_{m_3})}$
$(1,1,y_1)$	x_{11}	$\ldots$	x_{1l}	$n_{(1,1,y_1)}$
$(1,1,y_2)$	x_{21}	$\ldots$	x_{2l}	$n_{(1,1,y_2)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$(1,1,y_{m_4})$	x_{n1}	$\ldots$	x_{nl}	$n_{(1,1,y_{m_4})}$

2.3.1 (Bernoulli, Bernoulli, Poisson) Distribution

The joint probability mass function of $\mathbf{Y} = (Y_1, Y_2, Y_3)$, where $\mathbf{Y}$ is distributed as BBP, is as presented in Table 2. Here, $f(k, \lambda)$ is a Poisson distribution and the

Table 2: PMF of IID Trivariate Variables of BBP Distribution

Cases	Y_1	Y_2	Y_3	Joint Prob.	case MGF
1	0	0	k	$q_1q_2f(k,\lambda)$	$\exp\lambda(\exp^t -1)$
2	0	1	k	$q_1p_2f(k,\lambda)$	$\exp\lambda(\exp^t -1)$
3	1	0	k	$p_1q_2f(k,\lambda)$	$\exp\lambda(\exp^t -1)$
4	1	1	k	$p_1p_2f(k,\lambda)$	$\exp\lambda(\exp^t -1)$

sum of all the probabilities must equal to 1. Also, k is a known count data value. The moment generating function of (Y_1, Y_2, Y_3), that is, if $Z = \theta_1 Y_1 + \theta_2 Y_2 + \theta_3 Y_3$ which is distributed as (Bernoulli, Bernoulli, Poisson) is given by the following set of equations:

$$\begin{aligned} M_z(t) &= constant * M_{Y_1}(t) M_{Y_2}(t) M_{Y_3}(t) \\ &= constant * (p_1 + q_1 \exp t) * \\ &\quad (p_2 + q_2 \exp t)(\exp\lambda(\exp t - 1)) \end{aligned}$$

Here, $\theta_1 + \theta_2 + \theta_3 = 1$. Assume, θ_i's are known mixing parameters (See Casella (2002)).

The marginal probability mass function of the (Y_1, Y_2, Y_3)

Table 3: Marginal PMF of IID Trivariate Variables of BBP Distribution

Cases	Y_1	Y_2	Y_3	Marginal Probability
1	0	0		q_1q_2
2	0	1		q_1p_2
3	1	0		p_1q_2
4	1	1		p_1p_2
1	0		k	$q_1f(k,\lambda)$
2	1		k	$p_1f(k,\lambda)$
3		0	k	$q_2f(k,\lambda)$
4		0	k	$p_2f(k,\lambda)$

2.3.2 (Bernoulli, Bernoulli, Negative Binomial) Distribution

The joint probability mass function of $\mathbf{Y} = (Y_1, Y_2, Y_3)$, where $\mathbf{Y}$ is distributed as BBNB, is as presented in Table 4. Here, $g(y)$ is a Negative Binomial distri-

Table 4: PMF of IID Trivariate Variables of BBNB Distribution

Cases	Y_1	Y_2	Y_3	Joint Prob.	case MGF
1	0	0	k	$q_1q_2g(k)$	$(\frac{p}{1-(1-p)e^t})^r$
2	0	1	k	$q_1p_2g(k)$	$(\frac{p}{1-(1-p)e^t})^r$
3	1	0	k	$p_1q_2g(k)$	$(\frac{p}{1-(1-p)e^t})^r$
4	1	1	k	$p_1p_2g(k)$	$(\frac{p}{1-(1-p)e^t})^r$

bution and the sum of all the probabilities must equal to 1. Also, k is a known count data value. The moment generating function of (Y_1, Y_2, Y_3), that is, if $Z = \theta_1Y_1 + \theta_2Y_2 + \theta_3Y_3$ which is distributed as (Bernoulli, Bernoulli, Negative Binomial) is given by the following set of equations:

$$\begin{aligned} M_z(t) &= constant * M_{Y_1}(t)M_{Y_2}(t)M_{Y_3}(t) \\ &= constant * (p_1 + q_1 \exp t) * \\ &(p_2 + q_2 \exp t)((\frac{p}{1-(1-p)e^t})^r) \end{aligned}$$

Here, $\theta_1 + \theta_2 + \theta_3 = 1$. Assume, θ_i's are known mixing parameters.

2.4 Methods of Estimation

2.4.1 Maximum Likelihood Estimation

BBP The likelihood function L using a BBP model is given by the following equation:

$$
\begin{aligned}
L(\boldsymbol{y}|p_1, p_2, \lambda) = & \prod_{\{i:y_{1i}=0,y_{2i}=0,y_{3i}=k\}} (1-p_1)(1-p_2)f(k,\lambda) \\
& \prod_{\{i:y_{1i}=1,y_{2i}=0,y_{3i}=k\}} p_1(1-p_2)f(k,\lambda) \\
& \prod_{\{i:y_{1i}=0,y_{2i}=1,y_{3i}=k\}} (1-p_1)p_2f(k,\lambda) \\
& \prod_{\{i:y_{1i}=1,y_{2i}=1,y_{3i}=k\}} p_1p_2f(k,\lambda)
\end{aligned}
$$

The log-likelihood function is given by

$$
\begin{aligned}
\ell = logL(\boldsymbol{\theta}|\boldsymbol{y}) = & \sum_{\{i:y_{1i}=0,y_{2i}=0,y_{3i}=k\}} \log\{q_1q_2f(k,\lambda)\} \\
& + \sum_{\{i:y_{1i}=0,y_{2i}=1,y_{3i}=k\}} \log\{q_1p_2f(k,\lambda)\} \\
& \sum_{\{i:y_{1i}=1,y_{2i}=0,y_{3i}=k\}} \log\{p_1q_2f(k,\lambda)\} \\
& + \sum_{\{i:y_{1i}=1,y_{2i}=1,y_{3i}=k\}} \log\{p_1p_2f(k,\lambda)\}
\end{aligned}
$$

We maximize the joint log-likelihood with respect to the parameter vectors to obtain their maximum-likelihood estimates (MLEs), denoted by $(\hat{p_1}, \hat{p_2}, \hat{\lambda})$. These solutions are found iteratively via the Newton Raphson algorithm. Standard errors are obtained by taking the square roots of the diagonal elements of the inverse observed information matrix.

BBNB The likelihood function L using a BBNB model is given by the following equation:

$$
\begin{aligned}
L(\boldsymbol{y}|p_1, p_2, \mu, r) = & \prod_{\{i:y_{1i}=0,y_{2i}=0,y_{3i}=k\}} (1-p_1)(1-p_2)g(k) \\
& \prod_{\{i:y_{1i}=1,y_{2i}=0,y_{3i}=k\}} p_1(1-p_2)g(k) \\
& \prod_{\{i:y_{1i}=0,y_{2i}=1,y_{3i}=k\}} (1-p_1)p_2g(k) \\
& \prod_{\{i:y_{1i}=1,y_{2i}=1,y_{3i}=k\}} p_1p_2g(k)
\end{aligned}
$$

The log-likelihood function is given by

$$
\begin{aligned}
l = logL(\boldsymbol{\theta}|\boldsymbol{y}) = & \sum_{\{i:y_{1i}=0,y_{2i}=0,y_{3i}=k\}} \log\{q_1 q_2 g(k)\} \\
& + \sum_{\{i:y_{1i}=0,y_{2i}=1,y_{3i}=k\}} \log\{q_1 p_2 g(k)\} \\
& \sum_{\{i:y_{1i}=1,y_{2i}=0,y_{3i}=k\}} \log\{p_1 q_2 g(k)\} \\
& + \sum_{\{i:y_{1i}=1,y_{2i}=1,y_{3i}=k\}} \log\{p_1 p_2 g(k)\}
\end{aligned}
$$

We maximize the joint log-likelihood with respect to the parameter vectors to obtain their maximum-likelihood estimates (MLEs), denoted $(\hat{p_1}, \hat{p_2}, \hat{\mu}, \hat{r})$. These solutions are found iteratively via the Newton Raphson algorithm. Standard errors are obtained by taking the square roots of the diagonal elements of the inverse observed information matrix.

2.5 BBP (p_1, p_2, λ) Regression Model

Suppose that data consisting of count responses $\boldsymbol{y_i}, i = 1, \ldots, n$, for each subject i are independently distributed as Juncture BBP Model with parameters $(p_{1i}, p_{2i}, \lambda_i)$. The BBP (p_1, p_2, λ) Regression can be modeled under the assumptions

$$
\begin{aligned}
\text{logit}(p_{1i}) &= \boldsymbol{A_i \alpha}, \\
\text{logit}(p_{2i}) &= \boldsymbol{B_i \beta}, \\
\log(\lambda_i) &= \boldsymbol{G_i \gamma}
\end{aligned}
$$

where $\boldsymbol{G_i}, \boldsymbol{B_i}$ are subject-specific covariates, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are regression parameters, and $\boldsymbol{\mu_i} = \tau_i \xi_i$ Here, $\boldsymbol{\alpha}$s and $\boldsymbol{\beta}$s will have interpretation in terms of the binomial probabilities and, $\boldsymbol{\gamma}$'s have interpretation in terms of the Poisson mean, μ_i.

2.6 BBNB (p_1, p_2, τ, ξ) Regression Model

Suppose that data consisting of responses $\boldsymbol{y_i}, i = 1, \ldots, n$, for each subject i are independently distributed as Juncture BBNB Model with parameters $(p_{1i}, p_{2i}, \tau_i, \xi_i)$. The BBNB (p_1, p_2, τ, ξ) Regression can be modeled under the assumptions

$$
\begin{aligned}
\text{logit}(p_{1i}) &= \boldsymbol{A_i \alpha}, \\
\text{logit}(p_{2i}) &= \boldsymbol{B_i \beta}, \\
\log(\tau_i) &= \boldsymbol{G_i \gamma_1}, \\
\log(\xi_i) &= \boldsymbol{H_i \gamma_2}
\end{aligned}
$$

where $\boldsymbol{G_i}, \boldsymbol{B_i}$ are subject-specific co-variates, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are regression parameters, and $\boldsymbol{\mu_i} = \tau_i \xi_i$ Here, $\boldsymbol{\alpha}$s and $\boldsymbol{\beta}$s will have interpretation in terms of the binomial probabilities and, $\boldsymbol{\gamma}$'s have interpretation in terms of the NB mean, μ_i.

2.7 Correlated Data

For this article, we look how the correlation or covariance plays a role in determining what type of relationship exists between the bi-variate outcomes' distribution.

2.7.1 Theorem 1

Let (Y_1, Y_2, Y_3) be distributed as (Bernoulli, Bernoulli, Poisson) Model. If (Y_1, Y_2, Y_3) are correlated , then Y_2 can be modelled as a non-linear function of Y_1 and Y_3 can be a nonlinear function of Y_1. That is, (a) $Y_2 = g_1(\alpha_1 Y_1)$ and (b) $Y_3 = g_2(\alpha_2 Y_1)$. Note: Here, g's are non-linear functions of the variables.

Proof: For part (a), say g_1 is a linear. Then, $g_1(\beta x) = \beta g_1(x)$. For Bernoulli to Bernoulli correlation, we have left hand side to be $f(\beta\alpha x) = p^{\beta\alpha x} * (1-p)^{1-\beta\alpha x}$. But right hand side is $\beta f(\alpha x) = \beta p^{\alpha x}(1-p)^{1-\alpha x}$. Left hand side and right hand side are clearly different. Thus, g_1 must be non-linear function. For part (b), say g_2 is a linear.

We can apply similar reasoning to part (b). Assume g_2 is linear. Then, $g_2(\beta x) = \beta g_2(x)$. For Bernoulli to Poisson correlation, we see that the left hand side is $g_1(\beta\alpha x) = \exp \lambda\lambda^{\beta\alpha x}/(\beta\alpha x)!$. However, the right hand side is equal to $\beta g_2(\alpha x) = \beta Poi(\lambda)$. Thus, it is evident that the function g_2 is non-linear function.

2.7.2 Theorem 2

Let (Y_1, Y_2, Y_3) be distributed as (Bernoulli, Bernoulli, Negative Binomial). If (Y_1, Y_2, Y_3) are correlated , then Y_2 can be a non-linear function of Y_1 and Y_3 can be a nonlinear function of Y_1. That is, (a) $Y_2 = g_1(\alpha_1 Y_1)$ and (b) $Y_3 = g_3(\alpha_3 Y_1)$.

Proof: For part (b), say g_3 is a linear. Then, $f(\beta x) = \beta f(x)$. For Bernoulli to negative binomial correlation, Using

$$g(y|\tau, \xi) = \frac{\Gamma(\tau + y)}{\Gamma(\tau)\Gamma(y+1)} \frac{\xi^y}{(1+\xi)^{\tau+y}}, \text{ for } y = 0, 1, 2, \ldots$$

it is clear that $f(\beta x) \neq \beta f(x)$. Thus, g_3 must be non-linear.

3 Illustration

It is evident from the data that in longitudinal studies, typically a wealth of information is recorded for each patient, and interest often lies in complex associations between these different pieces of information.

3.1 Example Data Set

The average length of stay (LOS) in a U.S. emergency department (ED) is approximately 4.2 hours. The admitted patients require hospital admission face much longer ED stays, with boarding times (time from decision to admit until leaving the ED). Increased stays are often due to boarding, diagnostic tests and staffing shortages (See Miazgowski et al. (2023)).

This study or the data set has several components: patients, emergency department visits, ambulatory visit, readmission registry, discharges, providers, and emergency department unique ID. It contains admission date, discharge date, provider ID, etc. The dataset had calculated the expected LOS however, as per my knowledge, it had only calculated it as a single outcome measure for hospital management purposes. Here, we show how one can model the three variables together, i.e. mortality, re-admission and length of stay (LOS).

Our interest here is to study the length of stay along with mortality and readmissions, as defined above, for the patient population with varying characteristics or factors.

3.1.1 Inferential Objectives for Raw Outcomes Measures

The maximum value of negative log-likelihood is 2958.445 with $\hat{p_1} = 0.276(\hat{se} = 0.016)$, $\hat{p_2} = 0.142(\hat{se} = 0.013)$ and $\hat{\lambda} = 9.985(\hat{se} = 0.115)$ (See Table 5) using BBP Juncture Model. Using BBNB Juncture Model, the maximum value of negative loglikelihood value is 2888.342, a significant improvement by introducing one more parameter. The estimates are $\hat{p_1} = 0.276(\hat{se} = 0.016)$, $\hat{p_2} = 0.142(\hat{se} = 0.013)$, $\hat{\tau} = 12.617(\hat{se} = 0.1.533)$ and $\hat{\xi} = 0.791(\hat{se} = 0.097)$ (See Table 5). That is, even without co-variates, the BBNB distribution performs better as compared to the Poisson model due to certain counts being over-dispersed.

Table 5: MLE Parameter Estimates for Raw Counts (W.o. covariates)

	BBP Model		BBNB Model	
Parameter	Est.	Std.Err.	Est.	Std.Err.
p_1	0.276	0.016	0.276	0.016
p_2	0.142	0.013	0.142	0.013
λ	9.985	0.115	–	–
τ	–	-	12.618	1.533
ξ	–	-	0.791	0.097
NegLL	2958.445	-	2888.342	-

3.1.2 Effect of Covariates on Trivariate Outcomes Measures

The most common type of research questions in medical studies, in general, is to estimate or test for the effect of a set of covariates in some outcomes of interest. For example, for hospital management data, we would like to know how age and gender of a patient has an effect in the average of the three outcome profiles of the patients. The answer to such questions requires one to postulate a suitable statistical model that relates the co-variate(s) to the outcomes of interest. Depending on the nature of the outcome(s), several types of statistical models can be made available including BBP and BBNB regression models.

As seen from the tables 6 and 7, one can expect the BBNB regression model to perform better than the BBP regression model on this data. There is still a need an improvement in the BBNB regression models and that will be discussed more in a subsequent paper.

Table 6: MLE Parameter Estimates for BBP Regression Model

Parameters	Estimates	Std Err	p-value
Logit Link for p_1			
Intercept	2.398	0.567	< 0.0001
Age	-0.070	0.011	< 0.0001
Gender	0.695	0.173	< 0.0001
Logit Link for p_2			
Intercept	9.458	1.080	< 0.0001
Age	-0.045	0.024	0.0605
Gender	1.071	0.058	< 0.0001
Log Link for λ			
Intercept	4.167	311.664	0.9893
Age	21.015	6.131	0.0006
Gender	-16.957	96.768	0.8611
NegLL	3564.446		

Table 7: MLE Parameter Estimates for BBNB Regression Model

Parameters	Estimates	Std Err	p-value
Logit Link for p_1			
Intercept	-1.764	0.574	0.003
Age	0.012	0.105	0.909
Gender	0.135	0.169	0.424
Logit Link for p_2			
Intercept	0.993	0.695	0.153
Age	-0.049	0.013	0.0002
Gender	-0.035	0.007	< 0.0001
Log Link for μ			
Intercept	9.945	108.178	0.9268
Age	1.813	2.284	0.4273
Gender	1.747	29.164	0.9522
Log Link for ξ			
Intercept	9.945	4.120	0.016
Age	1.813	0.083	< 0.0001
Gender	1.747	1.844	0.344
NegLL	2947.146		

3.2 Association between Outcomes

Often it is also of interest to investigate the association structure between outcomes. After merging the data sets, we obtain the following correlation coefficient between Readmission and Mortality = -0.242 (t-statistic = -6.854, p-value < 0.0001); between Readmission and LOS = -0.089 (t-statistic = -2.4378, p-value = 0.015); and, between Mortality and LOS = 0.0061 (t-statistic = 0.166, p-value = 0.868). We found opposite results with readmission and mortality as compared to this reference (See [?]) showing a significantly moderate negative correlation between the two variables. Mortality and readmission rates may reflect different processes of care. Early intervention and coordination of care in a hospital may be particularly important for mortality and length of stay, where the outpatient clinic and patient education may be factors influencing chances of re-admissions.

There were 208 re-admissions of the 715 in the merged dataset. The readmission was 208 out of 715 observations. There were about 40 missing values. The mean length of stay is 9.985, median is 10 and the standard deviation is 4.123. Of the 107 mortal patients, only 1 death occurred. The remaining 106 were alive. Of the 107 patients, there were 16 patients from Cardiology, 19 General Medicine, 11 Hospitalist, 29 ICU, 27 Neurology, and 5 Orthopaedics. The average age of the patients was 52.48 , median age 52.39 and standard deviation of 7.89 years. Of those were readmitted, 29 were serviced by Cardiology, 74 were serviced by general Medicine, 18 by Hospitalist, 65 by ICU, 20 Neurology and 2 in Orthopaedics.

If mortality = 1 (death), correlation between readmission and LOS is 0.3111 (p-value = 0.0015). If mortality = 0 (alive), correlation between readmission and LOS is -0.0693 (p-value = 0.0909).

4 Conclusion

Real-life count data often exhibit two (related) characteristics: overdispersion and k-inflation. Overdispersion refers to an excess of variability in the data (i.e., the variance exceeds the mean), while the k-inflation refers an excess of count value of k. In the presence of overdispersion, the Poisson model is not adequate and can lead to biased estimates and unreliable standard errors estimates. Therefore, a negative binomial model for the length of stay is more appropriate in conjunction with the two binary variables, mortality and readmission.

5 Limitations

There are several limitations here. Since we assumed discrete binary nature of the first two variables, i.e. mortality and readmission, it puts limitations how the bounds of the probability is utilized. Furthermore, in this paper, because we were introducing the juncture models we also assumed independence between the three variables (mortality, readmission, and length of stay), it puts limitations since these three variables are actually somewhat correlated as seen above. Furthermore, there is still a need for an improvement in the regression models due to the large standard errors, and this will be furthered investigated in an upcoming paper.

Acknowledgment Dr. Manasi Sheth would like to thank the Departments of Mathematics and Computer Science at the University of Wisconsin - Whitewater for their support. Dr. Chaganty acknowledges the following individuals for their support: Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA, USA.

Data Availability (Source) The data that support the findings of this study are available in Kaggle at https://www.kaggle.com/datasets/neenunair/hospital-dataset.

References

1. A. Bottle et al. Global comparators project international comparison of hospital outcomes using administrative data. Health Serv Res. (2013), (48) 2081–100 .

2. A. C. Cameron and P. K. Trivedi , *Regression Analysis of Count Data*, New York: Cambridge, (1998).

3. G. Casella and R.L. Berger, Statistical Inference (2002), Thompson Learning Inc.

4. H.M. Krumholz et al. Relationship of hospital performance with readmission and mortality rates for patients hospitalized with acute myocardial infarction, heart failure or pneumonia (2013). JAMA 309, 587–93.

5. B. Miazgowski , C. Pakulski, and T. Miazgowski, Length of Stay in Emergency Department by ICD-10 Specific and Non-specific Diagnoses: A Single-Centre Restrospective Study (2023). J. Clin Med. 12(14); 4679.

6. D. Rizopoulos. Joint Models for Longitudinal and Time-to-Event Data with Applications in R (2012). Chapman & Hall, CRC Press Biostatistics Series.