Generalized Linear Models: Models for Count Data

MA4270 Project by Chong Zhen Jie (A0201613Y)

1 Introduction

We recall that in the case of linear regression, we are trying to predict continuous real-valued quantities. Consider the data set $\mathcal{D} = \{(\mathbf{x}_t, y_t)\}_{t=1}^n$ with n independent data samples. Specifically, we assume a linear relationship between the output variable $y_t \in \mathbb{R}$ and the input vector $\mathbf{x}_t = [x_{t0}, x_{t1}, \ldots, x_{td}]^\top \in \mathbb{R}^{d+1}$, where $x_{t0} = 1$. In other words, the linear regression model takes the form

$$y_t = \boldsymbol{\theta}^\top \mathbf{x}_t + z_t, \tag{1}$$

where $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_d]^\top \in \mathbb{R}^{d+1}$ is a parameter to be estimated. The model assumes a zero-mean Gaussian noise $z_t \sim \mathrm{N}(0, \sigma^2)$. As such, y_t follows a Gaussian conditional distribution with constant variance specified by

$$y_t | \mathbf{x}_t; \boldsymbol{\theta}, \sigma^2 \sim \mathrm{N}(\boldsymbol{\theta}^\top \mathbf{x}_t, \sigma^2). \tag{2}$$

However, the assumption of a linear regression model may not be suitable in every scenario. We outline two examples:

- Some output variables exist in the form of *count data*. For instance, if we want to predict the number of road accidents per month in a country, it makes sense that the output variable only takes counting numbers, or non-negative integer values $y_t \in \mathbb{Z}_{\geq 0}$. Hence, a reasonable assumption for a discrete conditional distribution of y_t would be the Poisson, with some parameter $\lambda_t > 0$. Then, $\mathrm{Var}(y_t|\mathbf{x}_t) = \lambda_t$ depends on λ_t and may not be constant for all t. Also, if $\lambda_t = \mathbb{E}(y_t|\mathbf{x}_t) = \boldsymbol{\theta}^\top \mathbf{x}_t$, then there is no guarantee that λ_t will only take non-negative values because $\boldsymbol{\theta}$ and θ_0 are unconstrained.
- Similar arguments apply for categorical output variables. In particular, the binary case involves predicting binary labels $y_t \in \{0, 1\}$. It would then be more appropriate to assume that y_t follows a Bernoulli conditional distribution.

In this report, we will overview a broad class of statistical models to address the above problem. These models are known as the *generalized linear models* (GLMs), which were formulated by Nelder and Wedderburn in 1972 [9]. The GLMs generalize the linear regression model by providing us with a unified framework to model output variables with various non-Gaussian conditional distributions, including those mentioned above.

Due to space limitations, we choose to present models for count data as the main topic of this report. All methods presented in this report assume that different data samples are statistically independent from each other.

2 GLMs for Count Data

In this section, we define the components that make up a GLM. We then introduce Poisson regression, and explain the motivation behind extending it to negative binomial regression.

2.1 Components of a GLM

In general, GLMs consist of the three components below.

(i) **Random component**: This component identifies the output variable $y_t \in \mathbb{R}$. Its distribution is the conditional version of the distributions from the *exponential dispersion family*, which includes the Gaussian, Poisson and Bernoulli, among many others. The conditional distribution has the form

$$f(y_t|\mathbf{x}_t) \coloneqq f(y_t|\mathbf{x}_t; \psi_t, \phi) = \exp\left[\frac{y_t\psi_t - b(\psi_t)}{a(\phi)} + c(y_t, \phi)\right]. \tag{3}$$

The *natural parameter*, ψ_t, and the *dispersion parameter*, ϕ, are typically known. The functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are also known. Distributions of the above form have certain mathematical properties that make them tractable. For example, the conditional mean and conditional variance of y_t is shown in Appendix A to be

$$\mathbb{E}(y_t|\mathbf{x}_t) = b'(\psi_t) \qquad \text{and} \qquad \text{Var}(y_t|\mathbf{x}_t) = b''(\psi_t)a(\phi). \tag{4}$$

Remark 1. For the rest of the report, we will make the dependence on the distribution parameters implicit for the simplicity of notations as shown in (3), unless otherwise stated.

(ii) **Systematic component**: This component specifies the input vector $\mathbf{x}_t = [x_{t0}, x_{t1}, \ldots, x_{td}]^\top \in \mathbb{R}^{d+1}$ where $x_{t0} = 1$, and a linear predictor of the form

$$\boldsymbol{\theta}^\top \mathbf{x}_t, \tag{5}$$

for some $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_d]^\top \in \mathbb{R}^{d+1}$.

(iii) **Link function**: For some monotone, differentiable function g, the GLM can be represented by

$$g\big[\mathbb{E}(y_t|\mathbf{x}_t)\big] = \boldsymbol{\theta}^\top \mathbf{x}_t. \tag{6}$$

This function g is called the *link function*. If $\boldsymbol{\theta}^\top \mathbf{x}_t = g\big[\mathbb{E}(y_t|\mathbf{x}_t)\big] = \psi_t$, then g is called a *canonical* link function. The use of canonical link functions in GLMs has some desirable properties. An example is briefly mentioned in Section 3.2 and shown in Appendix F. We observe that for GLMs, a linear relationship between the transformed conditional expectation $g\big[\mathbb{E}(y_t|\mathbf{x}_t)\big]$ and the input vector $\mathbf{x}_t$ is assumed. It should be noted that the choice of the link function g is not unique. In practice, there is no clear theoretical choice of the best link function, and this depends on statistical considerations. We will not discuss this in the report.

Remark 2. The linear regression model is an example of a GLM that we have learnt previously. We review its link function and how its conditional distribution takes the form in (3) in Appendix B.

2.2 Overview of Poisson Regression

Poisson Regression Models.

When the output variable is in the form of count data, a simple discrete distribution to model counting numbers would be the Poisson. Because count data only takes non-negative values, the conditional mean $\mathbb{E}(y_t|\mathbf{x}_t)$ must also be non-negative. We can take the natural logarithm of $\mathbb{E}(y_t|\mathbf{x}_t)$ to allow it to take any real-valued quantities, like the linear predictor $\boldsymbol{\theta}^\top \mathbf{x}_t$. It follows that a possible link function is the log function. The corresponding GLM evaluates to

$$\begin{aligned} g\big[\mathbb{E}(y_t|\mathbf{x}_t)\big] = \log\big[\mathbb{E}(y_t|\mathbf{x}_t)\big] &= \boldsymbol{\theta}^\top \mathbf{x}_t, \\ \mathbb{E}(y_t|\mathbf{x}_t) = \exp(\boldsymbol{\theta}^\top \mathbf{x}_t) &= \lambda_t, \end{aligned} \tag{7}$$

where $\lambda_t > 0$ is the parameter of the conditional Poisson distribution for y_t. This GLM is called the Poisson loglinear model, or the Poisson regression model. It is shown in Appendix C that

$$\psi_t = \log\big[\mathbb{E}(y_t|\mathbf{x}_t)\big], \qquad b(\psi_t) = \exp(\psi_t), \qquad a(\phi) = 1, \qquad c(y_t, \phi) = -\log(y_t!). \tag{8}$$

Overdispersion in Poisson Model.

A key feature of the Poisson distribution is that its variance equal its mean. However, observed count data often have greater conditional variance than its conditional mean. This phenomenon is called *overdispersion*. A common reason could be *unobserved heterogeneity* in the data [6].This means that in the Poisson regression model, $\mathbb{E}(y_t|\mathbf{x}_t)$ may still vary according to some *latent* (unobserved) variable, given that $\mathbf{x}_t$ is fixed. In other words, the variability in $\mathbb{E}(y_t|\mathbf{x}_t)$ is not fully explained by $\mathbf{x}_t$. To account for unobserved heterogeneity, we introduce a latent variable $h_t \coloneqq \exp(\varepsilon_t)$ into the model specified by (7) such that

$$\begin{aligned} \mathbb{E}(y_t|\mathbf{x}_t, \varepsilon_t) = \exp(\boldsymbol{\theta}^\top \mathbf{x}_t + \varepsilon_t) &= \exp(\varepsilon_t)\lambda_t, \\ \mathbb{E}(y_t|\mathbf{x}_t, h_t) &= h_t \lambda_t, \end{aligned} \tag{9}$$

where $y_t|\mathbf{x}_t, h_t \sim \text{Pois}(h_t\lambda_t)$. We prove that overdispersion exists in the above model under certain conditions with the following proposition.

Proposition 1. *Consider the model defined in* (9)*. We assume that h_t has an unknown distribution with variance* $\text{Var}(h_t)$ *and normalized mean* $\mathbb{E}(h_t) = 1$*, and that h_t is independent of $\mathbf{x}_t$. Then, we have that the conditional mean* $\mathbb{E}(y_t|\mathbf{x}_t) = \lambda_t$ *still remains correctly specified but the following holds:*

$$\text{Var}(y_t|\mathbf{x}_t) > \mathbb{E}(y_t|\mathbf{x}_t).$$

Hence, unobserved heterogeneity of this form causes overdispersion.

Proof. By the law of iterated expectations, we have that

$$\mathbb{E}(y_t|\mathbf{x}_t) = \mathbb{E}\Big[\mathbb{E}(y_t|\mathbf{x}_t, h_t)\Big|\mathbf{x}_t\Big] = \lambda_t \mathbb{E}(h_t|\mathbf{x}_t) \overset{(i)}{=} \lambda_t,$$

where (i) uses $\mathbb{E}(h_t) = 1$ and the independence of h_t and $\mathbf{x}_t$ such that $\mathbb{E}(h_t|\mathbf{x}_t) = \mathbb{E}(h_t)$. On the other hand, by the law of iterated variances,

$$\begin{aligned}\mathrm{Var}(y_t|\mathbf{x}_t) &= \mathbb{E}\Big[\mathrm{Var}(y_t|\mathbf{x}_t, h_t)\Big|\mathbf{x}_t\Big] + \mathrm{Var}\Big[\mathbb{E}(y_t|\mathbf{x}_t, h_t)\Big|\mathbf{x}_t\Big] \\ &= \lambda_t \mathbb{E}[h_t|\mathbf{x}_t] + \lambda_t^2 \mathrm{Var}(h_t|\mathbf{x}_t) \\ &\overset{(ii)}{=} \lambda_t + \lambda_t^2 \mathrm{Var}(h_t) > \lambda_t = \mathbb{E}(y_t|\mathbf{x}_t),\end{aligned}$$

where (ii) uses $\mathbb{E}(h_t) = 1$ and the independence of h_t and $\mathbf{x}_t$ similarly. □

Proposition 1 is an important result, showing that in the presence of unobserved heterogeneity, the conditional mean is still correctly specified. This suggests that applying more robust methods based on the Poisson model can still obtain consistent estimates [12]. An example is reviewed in Section 4.2.

Remark 3. In practice, it is more common for overdispersion to occur than the case of *underdispersion* (where the conditional variance is smaller than the conditional mean) [10]. Therefore, we only consider the case of overdispersion in this report.

2.3 Overview of Negative Binomial Regression

Gamma-Poisson Mixture Distribution.

We cannot use the model in (9) directly because of unobserved heterogeneity. Instead, we fit an appropriate probability distribution to h_t, and marginalize the joint distribution $f(y_t, h_t|\mathbf{x}_t)$ over h_t to get a new distribution $f(y_t|\mathbf{x}_t)$ unconditional of h_t. This new distribution derived is called a *compound distribution*, or a *mixture distribution* [8]. In particular, we derive the *gamma-Poisson mixture distribution*, and see how it relates to the negative binomial model. To do that, we perform the following steps:

1. Following the model in (9), its corresponding Poisson conditional distribution with parameter $h_t\lambda_t > 0$ takes the form
$$f(y_t|\mathbf{x}_t, h_t) = \frac{(h_t\lambda_t)^{y_t}\exp(-h_t\lambda_t)}{(y_t!)}, \qquad y_t \in \mathbb{Z}_{\geq 0}. \tag{10}$$

2. Because $h_t = \exp(\varepsilon_t) > 0$, the support of its distribution must be $(0, \infty)$. Thus, the gamma distribution is a possible choice. By Proposition 1, the distribution must satisfy $\mathbb{E}(h_t) = 1$. Hence, the distribution of h_t is Gamma(β, β) with parameter $\beta > 0$:
$$g(h_t) = \frac{\beta^\beta}{\Gamma(\beta)} h_t^{\beta-1}\exp(-\beta h_t), \qquad h_t > 0, \tag{11}$$
where $\mathbb{E}(h_t) = 1$ and $\mathrm{Var}(h_t) = 1/\beta$.

3. Then, the joint distribution of y_t and h_t given $\mathbf{x}_t$ is given by
$$f(y_t, h_t|\mathbf{x}_t) = f(y_t|\mathbf{x}_t, h_t)g(h_t), \tag{12}$$
where we used the independence of h_t and $\mathbf{x}_t$ from Proposition 1 such that $g(h_t|\mathbf{x}_t) = g(h_t)$.

4. We can now obtain the gamma-Poisson mixture distribution $f(y_t|\mathbf{x}_t)$ by integrating out the latent variable h_t from the joint distribution in (12):

$$\begin{aligned} f(y_t|\mathbf{x}_t) = \int_0^\infty f(y_t|\mathbf{x}_t, h_t)g(h_t)dh_t &= \int_0^\infty \left[\frac{(h_t\lambda_t)^{y_t}\exp(-h_t\lambda_t)}{(y_t!)}\right]\left[\frac{\beta^\beta}{\Gamma(\beta)}h_t^{\beta-1}\exp(-\beta h_t)\right]dh_t \\ &= \frac{\lambda_t^{y_t}\beta^\beta}{(y_t!)\Gamma(\beta)}\int_0^\infty h_t^{y_t+\beta-1}\exp\Big[-h_t(\lambda_t+\beta)\Big]dh_t \\ &\overset{(iii)}{=} \frac{\lambda_t^{y_t}\beta^\beta}{(y_t!)\Gamma(\beta)}\cdot\frac{1}{(\lambda_t+\beta)^{y_t+\beta}}\int_0^\infty z^{y_t+\beta-1}\exp(-z)dz \\ &\overset{(iv)}{=} \frac{\Gamma(y_t+\beta)}{\Gamma(y_t+1)\Gamma(\beta)}\left(\frac{\lambda_t}{\lambda_t+\beta}\right)^{y_t}\left(\frac{\beta}{\lambda_t+\beta}\right)^{\beta}, \end{aligned} \tag{13}$$

where (iii) substitutes $z = h_t(\lambda_t + \beta)$, and (iv) uses properties of the gamma function where

$$\Gamma(y_t+1) = (y_t!) \qquad \text{and} \qquad \Gamma(y_t+\beta) = \int_0^\infty z^{y_t+\beta-1}\exp(-z)dz.$$

We observe that the gamma-Poisson mixture distribution in (13) corresponds to the negative binomial distribution. Then, y_t has conditional mean and conditional variance

$$\mathbb{E}(y_t|\mathbf{x}_t) = \lambda_t \qquad \text{and} \qquad \text{Var}(y_t|\mathbf{x}_t) = \lambda_t(\frac{\lambda_t}{\beta}+1). \tag{14}$$

The term $1/\beta > 0$ is the negative binomial dispersion parameter. The greater the value of $1/\beta$, the greater the overdispersion relative to the Poisson. On the other hand, as $1/\beta$ decreases towards 0, $\text{Var}(y_t|\mathbf{x}_t)$ decreases towards λ_t, converging to the Poisson variance. A formal proof is in Appendix D.

Negative Binomial Models.

As discussed above, the negative binomial model relaxes the Poisson assumption that the conditional mean equals the conditional variance due to the additional dispersion parameter $1/\beta$. Similarly, a common link function for the negative binomial model is the log function such that

$$\begin{aligned} g\big[\mathbb{E}(y_t|\mathbf{x}_t)\big] = \log\big[\mathbb{E}(y_t|\mathbf{x}_t)\big] &= \boldsymbol{\theta}^\top\mathbf{x}_t, \\ \mathbb{E}(y_t|\mathbf{x}_t) &= \exp(\boldsymbol{\theta}^\top\mathbf{x}_t). \end{aligned} \tag{15}$$

For the negative binomial distribution to be considered under the exponential dispersion family, the dispersion parameter $1/\beta > 0$ must be assumed to be a known, fixed constant. Then it is shown in Appendix D that

$$\begin{aligned} \psi_t = \log\left[\frac{\mathbb{E}(y_t|\mathbf{x}_t)}{\beta+\mathbb{E}(y_t|\mathbf{x}_t)}\right], \qquad c(y_t,\phi) &= \log\left[\frac{\Gamma(y_t+\beta)}{\Gamma(y_t+1)\Gamma(\beta)}\right], \\ a(\phi) = 1, \qquad b(\psi_t) &= -\beta\log\big[1-\exp(\psi_t)\big]. \end{aligned} \tag{16}$$

Remark 4. The above model is called the NB2 model, where the variance is quadratic in λ_t as seen in (14). Alternative parameterizations exists, but we only consider the NB2 model as it fits under the GLM framework.

3 Methods of Estimation for GLMs

In this section, we discuss several iterative methods to find the maximum likelihood estimator for $\boldsymbol{\theta}$, and briefly mention a method for model selection. Subsequently, we show some experimental findings on synthetic negative binomial data using both Poisson regression and negative binomial regression.

3.1 Maximum Likelihood Estimation

Consider the data set $\mathcal{D} = \{(\mathbf{x}_t, y_t)\}_{t=1}^n$ with n independent data samples. We can derive the log-likelihood function $\mathcal{L}$ from the general form of the conditional distribution of GLMs in (3):

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \sum_{t=1}^{n} \log f(y_t|\mathbf{x}_t) = \sum_{t=1}^{n} \left[\frac{y_t\psi_t - b(\psi_t)}{a(\phi)} + c(y_t, \phi) \right]. \tag{17}$$

We define the shorthand notations $\ell_t \coloneqq \log f(y_t|\mathbf{x}_t)$, $\mu_t \coloneqq \mathbb{E}(y_t|\mathbf{x}_t)$ and $\eta_t \coloneqq \boldsymbol{\theta}^\top \mathbf{x}_t$. To find the derivative of $\mathcal{L}$ with respect to θ_j, we use the chain rule given by

$$\frac{\partial \ell_t}{\partial \theta_j} = \frac{\partial \ell_t}{\partial \psi_t} \cdot \frac{\partial \psi_t}{\partial \mu_t} \cdot \frac{\partial \mu_t}{\partial \eta_t} \cdot \frac{\partial \eta_t}{\partial \theta_j}.$$

Using the fact that $\mu_t = b'(\psi_t)$ and $\mathrm{Var}(y_t|\mathbf{x}_t) = b''(\psi_t)a(\phi)$ from (4),

$$\frac{\partial \ell_t}{\partial \psi_t} = \frac{y_t - b'(\psi_t)}{a(\phi)} = \frac{y_t - \mu_t}{a(\phi)} \qquad \text{and} \qquad \frac{\partial \mu_t}{\partial \psi_t} = b''(\psi_t) = \frac{\mathrm{Var}(y_t|\mathbf{x}_t)}{a(\phi)}.$$

Also, since $\partial \eta_t / \partial \theta_j = x_{tj}$, the likelihood equation for a GLM is

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = \sum_{t=1}^{n} \frac{\partial \ell_t}{\partial \theta_j} = \sum_{t=1}^{n} \frac{(y_t - \mu_t)x_{tj}}{\mathrm{Var}(y_t|\mathbf{x}_t)} \cdot \frac{\partial \mu_t}{\partial \eta_t} = 0, \qquad j = 0, 1, \ldots, d, \tag{18}$$

where the form of $\partial \mu_t / \partial \eta_t$ depends on the link function g because $\eta_t = g(\mu_t)$ from (6). Similar to ridge regression, we can add a L2 penalty term, $\alpha\|\boldsymbol{\theta}\|_2^2$, to the log-likelihood function in (17), for some parameter $\alpha \geq 0$. For simplicity, we do not consider regularization, or $\alpha = 0$.

3.2 Numerical Algorithms

Newton-Rhapson Method.

Unlike the linear regression model, many other GLMs do not have closed-form solutions for the likelihood equation in (18). As such, we rely on iterative methods for solving them. A common method would be the *Newton-Rhapson*, which is a root-finding algorithm that iteratively approximates the solutions to the likelihood equation. We define the gradient vector of $\mathcal{L}$ to be

$$\mathbf{u} = \nabla \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \left[\frac{\partial \mathcal{L}}{\partial \theta_0}, \frac{\partial \mathcal{L}}{\partial \theta_1}, \ldots, \frac{\partial \mathcal{L}}{\partial \theta_d} \right]^\top.$$

The vector $\mathbf{u}$ is also called the *score function*. Let $\mathbf{H}$ denote the Hessian matrix with entries $\partial^2\mathcal{L}/(\partial\theta_j\partial\theta_k)$. By the second-order Taylor series expansion,

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) \approx \mathcal{L}(\boldsymbol{\theta}^{(i)}|\mathcal{D}) + \mathbf{u}^{(i)\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)})^\top \mathbf{H}^{(i)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}),$$

where $\mathbf{u}^{(i)}$, $\boldsymbol{\theta}^{(i)}$ and $\mathbf{H}^{(i)}$ denote the values at the i-th iteration for $i = 0, 1, 2, \ldots$ Taking the derivative with respect to $\boldsymbol{\theta}$, we get

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}} \approx \mathbf{u}^{(i)} + \mathbf{H}^{(i)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)}) = \mathbf{0}.$$

Solving the above for $\boldsymbol{\theta}$, we get the following update rule:

$$\mathbf{H}^{(i)}\boldsymbol{\theta}^{(i+1)} = \mathbf{H}^{(i)}\boldsymbol{\theta}^{(i)} - \mathbf{u}^{(i)}. \tag{19}$$

In this maximization problem, if $\mathbf{H}^{(i)}$ is negative definite, then a unique global maximum exists. However, not all GLMs have strictly concave log-likelihood functions [11]. As a result, the algorithm may converge to other local maxima. We will explore another iterative method that provides more stable convergence below.

Remark 5. In high-dimensional settings, assuming the inverse of a Hessian $(\mathbf{H}^{(i)})^{-1}$ exists, computing it can be computationally expensive. By representing the update rule in (19) as a system of linear equations, we can just solve for $\boldsymbol{\theta}^{(i+1)}$ instead.

Fisher Scoring Method.

Fisher scoring is an iterative method modified from the Newton-Rhapson. The Newton-Rhapson uses the Hessian matrix $\mathbf{H}$, where its negative, $-\mathbf{H}$, is called the *observed information*. On the other hand, Fisher scoring uses the expected value of the observed information matrix, or $\mathbf{F} \coloneqq -\mathbb{E}(\mathbf{H}|\mathcal{D})$. The matrix $\mathbf{F}$ is called the *expected information*. By replacing $\mathbf{H}$ with $-\mathbf{F}$ in (19), the update rule for Fisher scoring is

$$\mathbf{F}^{(i)}\boldsymbol{\theta}^{(i+1)} = \mathbf{F}^{(i)}\boldsymbol{\theta}^{(i)} + \mathbf{u}^{(i)}. \tag{20}$$

To compute $\mathbf{F}$, we use a helpful result proven in Appendix E, which holds under certain regularity conditions:

$$(\mathbf{F})_{jk} = -\mathbb{E}\left[\frac{\partial^2\mathcal{L}}{\partial\theta_j\partial\theta_k}\middle|\mathcal{D}\right] = \mathbb{E}\left[\frac{\partial\mathcal{L}}{\partial\theta_j}\cdot\frac{\partial\mathcal{L}}{\partial\theta_k}\middle|\mathcal{D}\right], \tag{21}$$

where the (j,k)-entry of $\mathbf{F}$ can be expressed in a different form. Starting from the right-hand side, we substitute the likelihood equation from (18) such that

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial\mathcal{L}}{\partial\theta_j}\cdot\frac{\partial\mathcal{L}}{\partial\theta_k}\middle|\mathcal{D}\right] &= \mathbb{E}\left[\left(\sum_{t=1}^{n}\frac{(y_t-\mu_t)x_{tj}}{\mathrm{Var}(y_t|\mathbf{x}_t)}\cdot\frac{\partial\mu_t}{\partial\eta_t}\right)^2\middle|\mathcal{D}\right] \\
&\overset{(v)}{=} \mathbb{E}\left[\sum_{t=1}^{n}\frac{(y_t-\mu_t)^2x_{tj}x_{tk}}{\mathrm{Var}^2(y_t|\mathbf{x}_t)}\left(\frac{\partial\mu_t}{\partial\eta_t}\right)^2\middle|\mathcal{D}\right] \\
&\overset{(vi)}{=} \sum_{t=1}^{n}\frac{x_{tj}x_{tk}}{\mathrm{Var}(y_t|\mathbf{x}_t)}\left(\frac{\partial\mu_t}{\partial\eta_t}\right)^2,
\end{aligned} \tag{22}$$

where we use the fact that the conditional expectation only depends on y_t. We have that (v) holds because $\mathbb{E}[(y_t - \mu_t)(y_k - \mu_k)|\mathcal{D}] = \text{Cov}(y_t, y_k|\mathcal{D}) = 0$ for $t \neq k$ by the independence of data samples. As such, some of the cross terms are equal to 0. Then, (vi) holds from the fact that $\mathbb{E}[(y_t - \mu_t)^2|\mathbf{x}_t] = \text{Var}(y_t|\mathbf{x}_t)$.

In matrix notation, this corresponds to $\mathbf{F} = \mathbf{X}^\top \mathbf{W}\mathbf{X}$, where $\mathbf{W} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with non-negative diagonal elements $(\partial\mu_t/\partial\eta_t)^2/\text{Var}(y_t|\mathbf{x}_t)$, and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times(d+1)}$ is the input matrix. The update rule then becomes

$$\mathbf{X}^\top \mathbf{W}^{(i)} \mathbf{X}\boldsymbol{\theta}^{(i+1)} = \mathbf{X}^\top \mathbf{W}^{(i)} \mathbf{X}\boldsymbol{\theta}^{(i)} + \mathbf{u}^{(i)}. \tag{23}$$

For GLMs that use canonical link functions, both Newton-Rhapson and Fisher scoring can be shown to be identical. We outline the proof in Appendix F. Unlike the Newton-Rhapson, Fisher scoring has more stable convergence because $\mathbf{F}$ is always positive semi-definite [4]. To verify this, for any $\mathbf{z} \in \mathbb{R}^{d+1}$, we have $\mathbf{z}^\top \mathbf{F}\mathbf{z} = \mathbf{z}^\top \mathbf{X}^\top \mathbf{W}\mathbf{X}\mathbf{z} = (\mathbf{X}\mathbf{z})^\top \mathbf{W}\mathbf{X}\mathbf{z} \geq 0$ because $\mathbf{W}$ is defined to be a diagonal matrix with non-negative diagonal elements.

The log-likelihood can be a stopping criterion for the algorithm, terminating when the absolute change in log-likelihood between iterations falls below a pre-determined tolerance level. An alternative would be the *deviance* statistic, measuring the goodness-of-fit of a model, and is summarized in Appendix G.

3.3 Akaike Information Criterion

The *Akaike information criterion* (AIC) informs us of the quality of a model relative to other models. Let $\hat{\boldsymbol{\theta}}$ be the maximum-likelihood estimate of a model, and k be the number of estimated parameters. Then, the AIC value of a model is

$$\text{AIC} = 2k - 2\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathcal{D}). \tag{24}$$

The AIC rewards goodness-of-fit measured by the log-likelihood function, but also penalizes the model for having many parameters. Hence, we would want to select the model with the minimum AIC.

3.4 Experiments

We now implement both Poisson regression and negative binomial regression using the more stable Fisher scoring method on synthetic negative binomial data in `Python`. The exact formulas for the log-likelihood functions, score functions and expected information matrices are documented in Appendix H. Because the log-likelihood functions involve computing $\log[\Gamma(z)]$ for $z \in \mathbb{R}_{>0}$, we compute it using the `loggamma` function from the `scipy` numerical library. Otherwise, the code is fully implemented using the `numpy` linear algebra library and the `matplotlib` visualization library. The code is included in another `ipynb` file.

Generation of Synthetic Negative Binomial Data.

To construct the data, we can exploit the fact that the negative binomial distribution is derived as a Poisson-gamma mixture [7]. We set the true estimate $\boldsymbol{\theta} = [1, 0.75, -1.25]^\top$. For $n = 500$ points, the $\mathbf{x}_t$ values are chosen at random from $\text{N}(0, 1)$. For simplicity, we fix the dispersion parameter $1/\beta = 2$ and treat it as known when fitting the negative binomial model. Then, we fit a $\text{Gamma}(\beta, \beta)$ distribution to h_t. After computing $\mathbb{E}(y_t|\mathbf{x}_t, h_t) = h_t\lambda_t$ from (9), we use it to draw samples from the Poisson distribution to get y_t.

Experimental Results from Fisher Scoring.

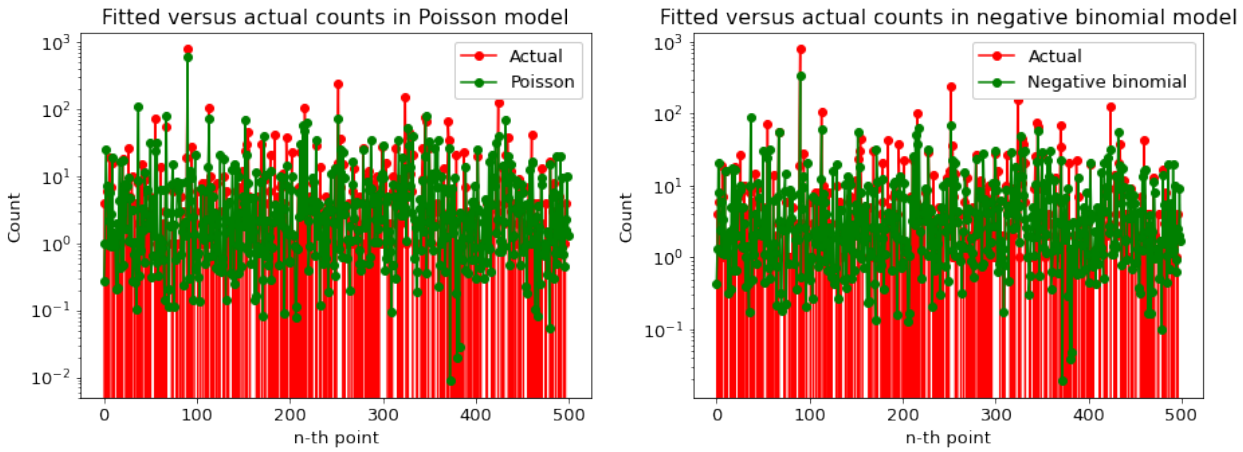

Figure 1: Comparison of fitted values to actual counts for $n = 500$ points.

From the log plots in Figure 1, we see that both the Poisson model and the negative binomial model seem to fit the actual counts in a similar manner. To figure out which model is better, we compare the AIC values.

Model	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	**AIC**
True value	1	0.75	-1.25	-
Poisson	0.9096	0.7135	-1.3124	5986.6
Negative binomial	1.0788	0.7051	-1.1209	2193.2

Table 1: Maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ and AIC values of models.

From Table 1, we see that the estimates for both models are similar to the true value, but the AIC value of the Poisson model is significantly higher, suggesting that the Poisson model is misspecified. These findings tie in with the results from Proposition 1 because of the presence of unobserved heterogeneity in the data.

4 Extensions and Further Results

In this section, we outline some extensions to the above models for count data, and present a brief summary of a published research paper that extends Poisson regression beyond traditional statistical methods.

4.1 Estimation of Dispersion Parameter

Because the negative binomial dispersion parameter $1/\beta$ is usually unknown, we need to estimate it to ensure that the distribution fits under the exponential dispersion family. One possible way is to apply maximum likelihood estimation for both $\boldsymbol{\theta}$ and $1/\beta$ simultaneously. To do that, we alternate between the estimation for $\boldsymbol{\theta}$ and the estimation for $1/\beta$ until both converge. It can be shown that $\boldsymbol{\theta}$ and $1/\beta$ are orthogonal, and their maximum likelihood estimates are asymptotically independent, so the model tends to be relatively stable [3].

4.2 Quasi-Likelihood Methods

In *quasi-likelihood* estimation, we do not assume any specific conditional distribution for the output variable y_t. Instead, we only assume a mean-variance relation where $\text{Var}(y_t|\mathbf{x}_t) = V(\mu_t)$ for some variance function V. In the case of the quasi-Poisson model, we have $V(\mu_t) = \phi\mu_t$, where ϕ is the quasi-Poisson dispersion parameter to allow us to model overdispersion. By substituting this back into the likelihood equation in (18), we obtain the quasi-likelihood estimating equations. There is no proper log-likelihood function here, but under certain conditions, quasi-likelihood estimation may be shown to be asymptotically efficient [1].

4.3 Zero-Inflated Models

In various applications, it is common for count data to have many zeros such that it does not fit standard distributions like the Poisson. Such data are called *zero-inflated* data. An example would be the *bonus hunger* phenomenon in insurance claims [2], where customers self-insure small losses instead of making claims to the insurance company to avoid paying a higher premium the next year. As such, it would be common for customers to have 0 claims. One well-known model would be the zero-inflated Poisson (ZIP) model. ZIP fits a mixture distribution, combining the standard Poisson distribution with additional probability mass on 0 outcomes. This allows flexibility in modeling the probability of 0 outcomes.

4.4 Literature Review: Non-Linear Poisson Regression using Neural Networks

Fallah et al. (2009) [5] proposed an extension of the Poisson regression to be based on the *multi-layer perceptron* (MLP), which is a type of neural network. Due to space limitations, we briefly describe the intuition behind how Poisson regression can be formulated under the MLP model.

A MLP consists of an input layer of nodes, an output layer of nodes and at least one hidden layer of nodes. The input nodes pass their inputs to the nodes in the first hidden layer. The hidden layer sums up the inputs according to some weights and adds a bias term. It then computes an *activation function* and passes the results to the nodes in the next layer. This process continues until we reach the output layer.

We recall that for Poisson regression in (7), there is an assumption of linearity in which $\log(\mu_t)$ is linear in $\mathbf{x}_t$. To introduce non-linearity into the MLP model, we can use non-linear functions like the hyperbolic tangent as the activation function in the hidden layers. Similar to Poisson regression, an exponential function is used in the output layer for predictions. The loss function we are trying to minimize can be the negative log-likelihood function as well. Then, a widely used algorithm to minimize the loss function is *backpropagation*. This involves computing the gradient of the loss function with respect to the weights and biases. Gradient methods like the stochastic gradient descent can then be used to update the weights and biases to minimize loss. For model selection, we can use the AIC to identify the number of nodes in the hidden layers.

Results from this research paper show that when actual data satisfy the linearity assumption mentioned above, the non-linear neural network predictions are slightly less accurate than the standard Poisson regression in terms of the mean squared error. However, the gains in accuracy are significant when the linearity assumption is not satisfied. This suggests that neural networks and Poisson regression can be combined to provide a natural solution to model count data. Other researchers have looked at extending other GLMs like linear regression and logistic regression to neural networks as well.

References

[1] Alan Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics. Wiley, 1 edition, 2015.

[2] Jean-Philippe Boucher, Michel Denuit, and Montserrat Guillen. Number of Accidents or Number of Claims? An Approach with Zero-Inflated Poisson Models for Panel Data. *The Journal of Risk and Insurance*, 76(4):821–846, 2009.

[3] David R. Cox and Nancy Reid. Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1):1–39, 1987.

[4] Eugene Demidenko. *Mixed Models : Theory and Applications with R*. Wiley Series in Probability and Statistics, 893. John Wiley & Sons, 2013.

[5] Nader Fallah, Hong Gu, Kazem Mohammad, Seyyed Ali Seyyedsalehi, Keramat Nourijelyani, and Mohammad Reza Eshraghian. Nonlinear Poisson regression using neural networks: A simulation study. *Neural Computing and Applications*, 18:939–943, 11 2009.

[6] William Greene. *Functional Form and Heterogeneity in Models for Count Data*. Now Publishers Inc, 2007.

[7] Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2 edition, 2011.

[8] Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill Education, 2013.

[9] John A. Nelder and Robert W.M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[10] Kimberly F. Sellers and Darcy S. Morris. Underdispersion models: Models that are "under the radar". *Communications in Statistics - Theory and Methods*, 46, 2017.

[11] Robert W.M. Wedderburn. On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models. *Biometrika*, 63(1):27–32, 1976.

[12] Rainer Winkelmann. *Econometric Analysis Of Count Data*. Springer, 2008.

Appendix

A Properties of Exponential Dispersion Family

Conditional distributions that follow the exponential dispersion family has the form

$$f(y_t|\mathbf{x}_t) = \exp\left[\frac{y_t\psi_t - b(\psi_t)}{a(\phi)} + c(y_t, \phi)\right].$$

Differentiating $\int f(y_t|\mathbf{x}_t)dy_t = 1$ with respect to ψ_t on both sides, we have

$$\begin{aligned}
\int f(y_t|\mathbf{x}_t)\frac{y_t - b'(\psi_t)}{a(\phi)}dy_t &= 0,\\
\mathbb{E}(y_t|\mathbf{x}_t) = \int y_t \cdot f(y_t|\mathbf{x}_t)dy_t &= b'(\psi_t)\int f(y_t|\mathbf{x}_t)dy_t,\\
\mathbb{E}(y_t|\mathbf{x}_t) &= b'(\psi_t),
\end{aligned}$$

which is the conditional mean. If we differentiate $\int f(y_t|\mathbf{x}_t)dy_t = 1$ twice with respect to ψ_t on both sides, we get

$$\begin{aligned}
\int y_t \cdot f(y_t|\mathbf{x}_t)\frac{y_t - b'(\psi_t)}{a(\phi)}dy_t &= b''(\psi_t)\int f(y_t|\mathbf{x}_t)dy_t + b'(\psi_t)\int f(y_t|\mathbf{x}_t)\frac{y_t - b'(\psi_t)}{a(\phi)}dy_t,\\
\int f(y_t|\mathbf{x}_t)\frac{\big[y_t - b'(\psi_t)\big]^2}{a(\phi)}dy_t &= b''(\psi_t),\\
\mathrm{Var}(y_t|\mathbf{x}_t) = \int f(y_t|\mathbf{x}_t)\big[y_t - b'(\psi_t)\big]^2 dy_t &= b''(\psi_t)a(\phi),
\end{aligned}$$

which is the conditional variance.

B Linear Regression Revisited

Let $\mu_t \coloneqq \mathbb{E}(y_t|\mathbf{x}_t)$. By comparing the linear predictor in (5) and the conditional mean of the linear regression model in (2), it is easy to see that the corresponding link function is the identity link function, where $g(\mu_t) = \mu_t = \boldsymbol{\theta}^\top \mathbf{x}_t$. Then, the Gaussian conditional distribution of y_t follows the exponential dispersion family because

$$\begin{aligned}
f(y_t|\mathbf{x}_t) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_t - \mu_t)^2}{2\sigma^2}\right]\\
&= \exp\left[\frac{y_t\mu_t - \frac{1}{2}\mu_t^2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y_t^2}{2\sigma^2}\right],
\end{aligned}$$

where

$$\psi_t = \mu_t, \qquad b(\psi_t) = \frac{1}{2}\psi_t^2, \qquad a(\phi) = \sigma^2, \qquad c(y_t, \phi) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y_t^2}{2\sigma^2}.$$

C Poisson Conditional Distribution

Let $\mu_t \coloneqq \mathbb{E}(y_t|\mathbf{x}_t)$. The Poisson conditional distribution of y_t follows the exponential dispersion family because

$$\begin{aligned} f(y_t|\mathbf{x}_t) &= \frac{\mu_t^{y_t}\exp(-\mu_t)}{(y_t!)} \\ &= \exp\big[y_t\log(\mu_t) - \mu_t - \log(y_t!)\big], \end{aligned}$$

where

$$\psi_t = \log(\mu_t), \qquad b(\psi_t) = \exp(\psi_t), \qquad a(\phi) = 1, \qquad c(y_t,\phi) = -\log(y_t!).$$

D Negative Binomial Conditional Distribution

Let $\mu_t \coloneqq \mathbb{E}(y_t|\mathbf{x}_t)$. The negative binomial conditional distribution of y_t follows the exponential dispersion family because

$$\begin{aligned} f(y_t|\mathbf{x}_t) &= \frac{\Gamma(y_t+\beta)}{\Gamma(y_t+1)\Gamma(\beta)}\left(\frac{\mu_t}{\mu_t+\beta}\right)^{y_t}\left(\frac{\beta}{\mu_t+\beta}\right)^{\beta} \\ &= \exp\left[y_t\log\left(\frac{\mu_t}{\mu_t+\beta}\right) + \beta\log\left(\frac{\beta}{\mu_t+\beta}\right) + \log\left(\frac{\Gamma(y_t+\beta)}{\Gamma(y_t+1)\Gamma(\beta)}\right)\right], \end{aligned}$$

where

$$\begin{gathered} \psi_t = \log\left(\frac{\mu_t}{\beta+\mu_t}\right), \qquad c(y_t,\phi) = \log\left[\frac{\Gamma(y_t+\beta)}{\Gamma(y_t+1)\Gamma(\beta)}\right], \\ a(\phi) = 1, \qquad b(\psi_t) = -\beta\log\big[1-\exp(\psi_t)\big]. \end{gathered}$$

To show that the above distribution converges to the Poisson as $1/\beta \to 0$ or $\beta \to \infty$, we have that

$$f(y_t|\mathbf{x}_t) = \frac{\mu_t^{y_t}}{\Gamma(y_t+1)}\cdot\frac{\Gamma(y_t+\beta)}{\Gamma(\beta)(\mu_t+\beta)^{y_t}}\cdot\frac{1}{(\frac{1}{\beta}\mu_t+1)^{\beta}}.$$

Taking the limits, we get

$$\lim_{\beta\to\infty} f(y_t|\mathbf{x}_t) = \frac{\mu_t^{y_t}}{\Gamma(y_t+1)}\cdot 1\cdot\frac{1}{\exp(\mu_t)} = \frac{\mu_t^{y_t}\exp(-\mu_t)}{(y_t!)},$$

where we used the gamma function $\Gamma(y_t+1) = (y_t!)$ and a known result $\lim_{\beta\to\infty}(\frac{\mu_t}{\beta}+1)^{\beta} = \exp(\mu_t)$. By considering the simplest form in which β is an integer parameter of the negative binomial, the middle term evaluates to 1 because

$$\lim_{\beta\to\infty}\frac{\Gamma(y_t+\beta)}{\Gamma(\beta)(\mu_t+\beta)^{y_t}} = \lim_{\beta\to\infty}\frac{\prod_{j=1}^{y_t}(\beta+j-1)}{\prod_{j=1}^{y_t}(\beta+\mu_t)} = \lim_{\beta\to\infty}\frac{\prod_{j=1}^{y_t}(1+\frac{j-1}{\beta})}{\prod_{j=1}^{y_t}(1+\frac{\mu_t}{\beta})} = 1.$$

E Alternative Form of the Expected Information Matrix

We derive an alternative form of the entries in the expected information matrix $\mathbf{F}$, which holds under certain regularity conditions:

$$(\mathbf{F})_{jk} = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial\theta_j \partial\theta_k}\middle|\mathcal{D}\right] = \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial\theta_j}\cdot\frac{\partial \mathcal{L}}{\partial\theta_k}\middle|\mathcal{D}\right].$$

Define $L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{t=1}^{n} f(y_t|\mathbf{x}_t)$ to be the likelihood function such that $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \log L(\boldsymbol{\theta}|\mathcal{D})$. We have that

$$\begin{aligned}
-\frac{\partial^2 \mathcal{L}}{\partial\theta_j \partial\theta_k} = -\frac{\partial^2 (\log L)}{\partial\theta_j \partial\theta_k} &= -\frac{\partial}{\partial\theta_j}\left(\frac{1}{L}\cdot\frac{\partial L}{\partial\theta_k}\right) \\
&= -\frac{1}{L}\cdot\frac{\partial^2 L}{\partial\theta_j\partial\theta_k} + \frac{1}{L^2}\cdot\frac{\partial L}{\partial\theta_j}\cdot\frac{\partial L}{\partial\theta_k} \\
&= -\frac{1}{L}\cdot\frac{\partial^2 L}{\partial\theta_j\partial\theta_k} + \frac{\partial \mathcal{L}}{\partial\theta_j}\cdot\frac{\partial \mathcal{L}}{\partial\theta_k}.
\end{aligned}$$

By taking the expectation on both sides, the first term in the last step becomes

$$\begin{aligned}
\mathbb{E}\left[-\frac{1}{L}\cdot\frac{\partial^2 L}{\partial\theta_j\partial\theta_k}\middle|\mathcal{D}\right] &= -\int\cdots\int \frac{1}{L}\cdot\frac{\partial^2 L}{\partial\theta_j\partial\theta_k}\cdot L\; dy_1\cdots dy_n \\
&= -\frac{\partial^2}{\partial\theta_j\partial\theta_k}\int\cdots\int L\; dy_1\cdots dy_n \\
&= -\frac{\partial^2}{\partial\theta_j\partial\theta_k} = 0.
\end{aligned}$$

We used the fact that $L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{t=1}^{n} f(y_t|\mathbf{x}_t)$ by the independence of data samples, so each corresponding integral is equal to 1. The result then follows.

F Simplification for Canonical Link Functions

For GLMs that use the canonical link function, both Newton-Rhapson and Fisher scoring can be shown to be identical. We recall that canonical link functions g satisfy $\psi_t = g(\mu_t) = \eta_t$ by (6). Also, we have the conditional mean $\mu_t = b'(\psi_t)$ from (4). We can simplify $\partial\mu_t/\partial\eta_t$ in the likelihood equation to

$$\frac{\partial\mu_t}{\partial\eta_t} = \frac{\partial\mu_t}{\partial\psi_t} = \frac{\partial b'(\psi_t)}{\partial\psi_t} = b''(\psi_t).$$

Since the conditional variance is $\mathrm{Var}(y_t|\mathbf{x}_t) = b''(\psi_t)a(\phi)$ from (4), the likelihood equation in (18) becomes

$$\frac{\partial\mathcal{L}}{\partial\theta_j} = \sum_{t=1}^{n}\frac{\partial\ell_t}{\partial\theta_j} = \sum_{t=1}^{n}\frac{(y_t-\mu_t)x_{tj}}{\mathrm{Var}(y_t|\mathbf{x}_t)}b''(\psi_t) = \sum_{t=1}^{n}\frac{(y_t-\mu_t)x_{tj}}{a(\phi)} = 0, \qquad j = 0, 1, \ldots, d.$$

Then, the second derivative of $\mathcal{L}$ evaluates to

$$\frac{\partial^2\mathcal{L}}{\partial\theta_j\partial\theta_k} = \sum_{t=1}^{n}\frac{-x_{tj}}{a(\phi)}\cdot\frac{\partial\mu_t}{\partial\theta_k}.$$

This does not depend on y_t so by taking expectations,

$$\frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k} = \mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k}\middle| \mathcal{D}\right].$$

In matrix form, this means that $\mathbf{H} = -\mathbf{F}$. Hence, the update rule under both Newton Rhapson and Fisher scoring are equal when using canonical link functions.

G Deviance Statistic

While fitting a model to the data set $\mathcal{D}$, we need to examine the level of discrepancy between the estimated values of μ_t and the observed values y_t. One way to measure the lack of fit is to compute the *deviance* of a model.

Let $\mathcal{L}(\boldsymbol{\mu}|\mathcal{D})$ be the log-likelihood function expressed in terms of $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^\top$. Suppose our current model has maximum likelihood estimate $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ so the maximum log-likelihood corresponds to $\mathcal{L}(\hat{\boldsymbol{\mu}}|\mathcal{D})$. Then, if we consider all possible models, the model with the perfect fit will naturally have the maximum achievable log-likelihood $\mathcal{L}(\mathbf{y}|\mathcal{D})$, where $\boldsymbol{\mu} = \mathbf{y} = [y_1, \ldots, y_n]^\top$ are the observed values themselves. This model is called the *saturated model*. Hence, the deviance statistic measures the difference between our current model and the saturated model, given by

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\mu}}|\mathcal{D}) &= 2[L(\mathbf{y}|\mathcal{D}) - L(\hat{\boldsymbol{\mu}}|\mathcal{D})] \\ &= 2\sum_{t=1}^{n}\left[\frac{y_t\tilde{\psi}_t - b(\tilde{\psi}_t) - y_t\hat{\psi}_t + b(\hat{\psi}_t)}{a(\phi)}\right], \end{aligned} \tag{25}$$

where $\tilde{\psi}_t$ is the estimate for the saturated model and $\hat{\psi}$ is the estimate for the current model. We observe that $\boldsymbol{\mu}$ depends on ψ_t because $\mu_t = b'(\psi_t)$ from (4). By definition, we have $L(\mathbf{y}|\mathcal{D}) \geq L(\hat{\boldsymbol{\mu}}|\mathcal{D})$ so $D(\mathbf{y}, \hat{\boldsymbol{\mu}}|\mathcal{D}) \geq 0$. Therefore, the greater the deviance, the poorer the fit of the current model. This can be used as a stopping criterion for the numerical algorithms in Section 3.2, comparing the model between iterations.

H Exact Formulas for Model Fitting

We derive the log-likelihood functions, score functions and the expected information matrices for both Poisson regression and negative binomial regression.

H.1 Log-Likelihood Functions

The general form for the log-likelihood function for GLMs is

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \sum_{t=1}^{n}\left[\frac{y_t\psi_t - b(\psi_t)}{a(\phi)} + c(y_t, \phi)\right].$$

For the Poisson model, it has log-likelihood function of the form:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \sum_{t=1}^{n} \big[y_t \log \mu_t - \mu_t - \log(y_t!) \big].$$

For the negative binomial model, it has log-likelihood function of the form:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \sum_{t=1}^{n} \left[y_t \log \left(\frac{\frac{1}{\beta}\mu_t}{1 + \frac{1}{\beta}\mu_t} \right) - \beta \log \left(1 + \frac{1}{\beta}\mu_t \right) + \log \left(\frac{\Gamma(y_t + \beta)}{\Gamma(y_t + 1)\Gamma(\beta)} \right) \right].$$

H.2 Score Functions

Since both the Poisson model and the negative binomial model use the log link function such that $\log \mu_t = \eta_t$, we have $\partial \mu_t / \partial \eta_t = \mu_t$. By substituting in their corresponding variances $\text{Var}(y_t|\mathbf{x}_t)$, it turns out that the elements of the score function for both models share the same form:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta_j} &= \sum_{t=1}^{n} \frac{(y_t - \mu_t) x_{tj}}{\text{Var}(y_t|\mathbf{x}_t)} \cdot \frac{\partial \mu_t}{\partial \eta_t} \\ &= \sum_{t=1}^{n} \frac{(y_t - \mu_t) x_{tj}}{1 + \frac{1}{\beta}\mu_t}, \end{aligned}$$

where $1/\beta = 0$ denotes the expression for the Poisson model, while $1/\beta > 0$ denotes the expression for the negative binomial model.

H.3 Expected Information Matrices

For the expected information matrix $\mathbf{F} = \mathbf{X}^\top \mathbf{W} \mathbf{X}$, it turns out that the diagonal matrix $\mathbf{W}$ has elements of the same form for both models as well:

$$\frac{(\partial \mu_t / \partial \eta)^2}{\text{Var}(y_t|\mathbf{x}_t)} = \frac{\mu_t}{1 + \frac{1}{\beta}\mu_t},$$

where $1/\beta = 0$ denotes the expression for the Poisson model, while $1/\beta > 0$ denotes the expression for the negative binomial model.