CE225 - Modelos Lineares Generalizados

Cesar Augusto Taconeli

11 de julho, 2018

Aula 14 - Regressão para dados de contagens com superdispersão

Introdução

Introdução

- O problema da superdispersão, na análise de dados de contagens, se caracteriza por uma dispersão nos dados superior à especificada pelo modelo adotado.

- Como casos mais usuais, temos:
  - $Var(y_i|\boldsymbol{x}_i) > \frac{\pi_i(1-\pi_i)}{m_i}$, para a distribuição binomial;
  - $Var(y_i|\boldsymbol{x}_i) > \mu_i$, para a distribuição Poisson.

- Causas e consequências de superdispersão são discutidas na sequência.

Causas de superdispersão

- Variabilidade entre as unidades amostrais (experimentais) não acomodada no modelo;

- Correlação entre as unidades amostrais (devido a fatores não observados ou não incorporados no modelo);

- Delineamento amostral envolve clusters;

- Omissão de variáveis não observadas.

Consequências da superdispersão

- Embora as estimativas pontuais dos parâmetros ainda sejam consistentes, os erros padrões são incorretos e subestimados (por não incorporar a dispersão extra);

- Os testes de hipóteses são "super otimistas" (inflacionando a probabilidade do erro do tipo I);

- As variações nas deviances para modelos encaixados também são incorretas (inflacionadas), induzindo à escolha de modelos demasiadamente complexos;

- Assim, as inferências produzidas pelo modelo ficam comprometidas por não se levar em conta a superdispersão.

Análise de dados com superdispersão

- Dentre as principais estratégias para se lidar com superdispersão, destacam-se:
  - Assumir um modelo em dois estágios para a resposta (modelos de mistura);
  - Assumir alguma forma mais geral para a variância da distribuição, possivelmente incluindo parâmetros adicionais (modelos de quase verossimilhança);
  - Para problemas com elevada frequências de contagens iguais a zero, utilizar modelos que acomodem o excesso de zeros.

Modelos de mistura

Modelos de mistura

- Inicialmente vamos tratar do problema da superdispersão associado ao modelo de Poisson.

- Vamos considerar a hetergeneidade não observada (causa da superdispersão) introduzida ao modelo de Poisson como uma variável aleatória ν que multiplica a média $(y|\mu, \nu \sim Poisson(\mu\nu))$.

- Integrando com relação à variável ν, obtemos a distribuição marginal de y.

Modelos de mistura

- No contexto de regressão, retomando o modelo log-linear, temos:

$$y_i|\boldsymbol{x}_i, \nu_i \sim Poisson(\mu_i \nu_i); \tag{1}$$

$$\begin{aligned} \mu_i \nu_i &= exp(\beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}) \nu_i \\ &= exp(\beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta} + log(\nu_i)) \\ &= exp((\beta_0 + u_i) + \boldsymbol{x}_i' \boldsymbol{\beta}), \end{aligned} \tag{2}$$

em que $u_i = log(\nu_i)$.

- Assumimos que os termos ν_i são independentes e identicamente distribuídos, com média $E[\nu_i] = 1$ e variância $Var[\nu_i] = \sigma^2_\nu$.
- Diferentes formas podem ser especificadas para a distribuição de ν_i, dando origem a diferentes modelos de mistura.

Modelos de mistura: Poisson-Gama (Binomial negativa)

- Vamos considerar $y|\nu \sim Poisson(\mu\nu)$:

$$f(y|\nu) = \frac{e^{-\mu\nu}(\mu\nu)^y}{y!}, \quad y = 0, 1, 2, ...; \mu > 0 \tag{3}$$

e ν uma variável aleatória com função densidade de probabilidade $g(\nu)$, com $E[\nu] = 1$ e $Var[\nu] = \sigma_\nu^2$.

- A distribuição marginal de y, resultante dessa mistura, fica dada por:

$$h(y) = \int f(y|\nu)g(\nu)d\nu. \tag{4}$$

Modelos de mistura: Poisson-Gama (Binomial negativa)

- O caso mais conhecido de mistura para o modelo Poisson corresponde ao caso da distribuição Gama.

- Seja $Z = \mu\nu$, $Y|z \sim Poisson(Z)$ e $Z \sim Gama(\mu, \phi)$:

$$g(z; \mu, \phi) = \frac{z^{\phi-1}}{\Gamma(\phi)} \left(\frac{\phi}{\mu}\right)^{\phi} e^{-\frac{\phi z}{\mu}}, \quad z > 0, \mu > 0, \phi > 0. \tag{5}$$

Modelos de mistura: Poisson-Gama (Binomial negativa)

- A função de probabilidade marginal de y, usando a equação 4, fica dada por:

$$h(y;\mu,\phi) = \frac{\Gamma(\phi + y)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\mu}{\mu+\phi}\right)^y \left(\frac{\phi}{\mu+\phi}\right)^\phi, \quad y = 0, 1, 2, ..., \qquad (6)$$

que corresponde à distribuição **binomial negativa**, com média μ e parâmetro de dispersão ϕ.

- Neste caso, $E(Z) = \mu$ e $Var(Z) = \mu^2/\phi$, de tal forma que $E(Y) = \mu$ e $Var(Y) = \mu + \mu^2/\phi$.
- Assim, a distribuição binomial negativa é uma alternativa ao modelo de Poisson, na presença de superdispersão, uma vez que $Var(Y) = \mu + \mu^2/\phi > \mu$.

Modelos de mistura: Poisson-Gama (Binomial negativa)

- Podemos ajustar um modelo de regressão com família binomial negativa usando a função `glm.nb` do pacote `MASS`. O modelo fica especificado por:

$$y_i|\boldsymbol{x}_i \sim BN(\mu_i, \phi); \tag{7}$$

$$g(\mu_i) = \boldsymbol{x'}_i\boldsymbol{\beta}. \tag{8}$$

- Como funções de ligação mais usuais, temos a logarítmica ($g(\mu_i) = log(\mu_i)$), raiz quadrada ($g(\mu_i) = \sqrt{(}\mu_i)$) e identidade ($g(\mu_i) = \mu_i$).
- Diversas outras distribuições podem ser consideradas como alternativas à Gama na mistura com a Poisson. Vamos tratar da mistura Poisson-Normal.

Modelos de mistura: Poisson-Normal

- Vamos considerar $log(\nu) \sim (0, \sigma^2)$. Assim:

$$\begin{aligned} \mu_i \nu_i &= exp(\boldsymbol{x}_i' \beta + log(\nu_i)) \\ &= exp(\boldsymbol{x}_i' \beta + \sigma \epsilon_i), \end{aligned} \tag{9}$$

com $\epsilon_i \sim N(0, 1)$. A distribuição marginal é obtida integrando em relação a ϵ:

$$\begin{aligned} h(y_i | \boldsymbol{x}_i, \beta, \sigma) &= \int_{-\infty}^{\infty} f(y_i | \boldsymbol{x}_i, \epsilon_i) g(\epsilon_i) d\epsilon_i \\ &= \int_{-\infty}^{\infty} f(y_i | \boldsymbol{x}_i, \epsilon_i) \frac{1}{\sqrt{2\pi}} e^{-\epsilon_i^2/2} d\epsilon_i \\ &= \int_{-\infty}^{\infty} exp(-e^{\boldsymbol{x}_i' \beta + \sigma \epsilon})(-e^{\boldsymbol{x}_i' \beta + \sigma \epsilon_i})^{y_i} \frac{1}{y_i!} \frac{1}{\sqrt{2\pi}} e^{-\epsilon_i^2/2} d\epsilon_i. \end{aligned} \tag{10}$$

Modelos de mistura: Poisson-Normal

- Diferentemente da mistura Poisson-Gama, não há forma fechada para $h(y_i|\boldsymbol{x}_i, \beta, \sigma)$ no caso da mistura Poisson-Normal.

- Também neste caso a função de ligação mais usual é a logarítmica, podendo ser substituída, eventualmente, pela raiz quadrada ou identidade.

Modelos de mistura - beta binomial

- Super dispersão pode ser verificada também para dados binários agrupados, situação que "sugere" a distribuição binomial.

- Há duas fontes principais de superdispersão para dados binários neste contexto:
  - Observações para uma particular configuração de covariáveis apresentam probabilidade de sucesso que varia devido a fatores não observados (**heterogeneidade**);
  - Os eventos binários em cada grupo de m_i observações são correlacionados.

- Uma das formas de lidar com superdispersão para dados binários grupados é usando um modelo resultante de mistura (particularmente, o modelo beta-binomial).

Modelo beta binomial

- O modelo beta-binomial resulta de uma mistura das distribuições beta e binomial.
- A especificação do modelo pode ser descrita da seguinte forma:
  - Dado π, $s = ny \sim binomial(n, \pi)$;
  - π tem distribuição Beta.
- A distribuição Beta tem suporte no intervalo $(0, 1$ com função densidade de probabilidade dada por:

$$f(\pi; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\alpha_1 - 1}(1 - \pi)^{\alpha_2 - 1}, \quad 0 < \pi < 1, \quad \alpha_1 > 0, \quad \alpha_2 > 0. \tag{11}$$

Modelo beta binomial

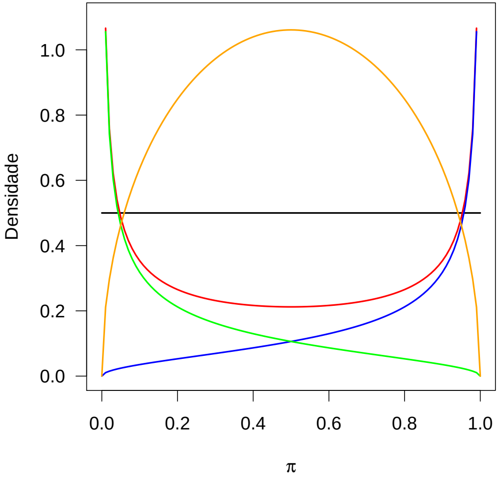

Figura 1: Distribuição beta para diferentes valores dos parâmetros.

Modelo beta binomial

- Considere:

$$\mu = \frac{\alpha_1}{\alpha_1 + \alpha_2}; \quad \theta = \frac{1}{\alpha_1 + \alpha_2}, \tag{12}$$

de tal forma que

$$E(\pi) = \mu; \qquad Var(\pi) = \mu(1-\mu)\frac{\theta}{1+\theta}. \tag{13}$$

- Como resultado da mistura, marginalmente $s = ny$ tem distribuição *beta-binomial*, com função de probabilidades:

$$P(S = s; n, \mu, \theta) = \binom{n}{s} \frac{\left[\prod_{k=0}^{s-1}(\mu + k\theta)\right]\left[\prod_{k=0}^{n-s-1}(1-\mu+k\theta)\right]}{\prod_{k=0}^{n-1}(1+k\theta)}, s = 0, ..., n. \tag{14}$$

Modelo beta binomial

- A distribuição beta-binomial pode assumir formas diversas, muitas vezes diferentes da distribuição binomial.

- A distribuição binomial pode ser verificada como caso particular da beta-binomial, quando $\theta \rightarrow 0$ e, consequentemente, $var(\pi) \rightarrow 0$.

- Para a proporção binomial $y = s/n$, temos, para a distribuição beta-binomial:

$$E(y) = \mu; \quad var(y) = \left[1 + (n-1)\frac{\theta}{1+\theta}\right]\frac{\mu(1-\mu)}{n}. \tag{15}$$

Modelo beta binomial

- No ajuste do modelo beta-binomial, normalmente consideramos θ constante para todas as observações.

- Para a função de ligação, podemos considerar qualquer alternativa aplicada à análise de dados binários.

- Em particular, podemos considerar função de ligação logito:

$$log(\frac{\mu_i}{1-\mu_i}) = \boldsymbol{x}_i\boldsymbol{\beta}, \quad i = 1, 2, ..., n. \tag{16}$$

Modelos de quase verossimilhança

Modelos de quase verossimilhança

- Considere novamente as equações de verossimilhança para modelos lineares generalizados:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right), \quad j = 0, 1, 2, ..., p, \tag{17}$$

em que $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}$.

- Observe que as equações de verossimilhança dependem da distribuição de y_i somente através da média (μ_i) e da função de variância $V(\mu_i)$.
- Os demais momentos da distribuição (momentos de maior ordem) não afetam os valores de $\hat{\beta}$ nem a covariância assintótica, dentre outros.

Modelos de quase verossimilhança

- A abordagem via quase verossimilhança baseia-se nesse fato, requerendo apenas que se especifique a média e a variância da distribuição, através:
  - Do preditor linear $g^{-1}(\mu) = \eta_i = \sum_j \beta_j x_{ij}$ e da função de ligação g;
  - Da função de variância $V(\mu_i)$.
- Observe que sob esta abordagem não é necessário especificar a distribuição de probabilidades de y.

Modelos de quase verossimilhança

- As estimativas de quase verossimilhança são obtidas através da solução das equações de verossimilhança para MLGs (17) substituindo μ_i e $v(\mu_i)$ conforme as especificações do modelo.
- A simples introdução de um parâmetro de dispersão (ϕ) multiplicando $V(\mu)$ pode acomodar a superdispersão, tendo o seguinte impacto no ajuste do modelo:
  - O parâmetro de dispersão fatora (é eliminado) das equações de estimação (em 17)), não tendo efeito nas estimativas pontuais dos parâmetros $\beta's$.
  - A matriz de covariâncias assintótica de $\hat{\beta}$ fica multiplicada por ϕ (consequentemente, os erros padrões ficam multiplicados por $\sqrt{\phi}$).

Modelos de quase verossimilhança

- A introdução de um parâmetro adicional de dispersão requer sua estimação.
- Dentre as estimativas para ϕ podemos considerar o seguinte estimador consistente:

$$\hat{\phi} = \frac{X^2}{n-p}, \tag{18}$$

em que

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \tag{19}$$

Modelo quase Poisson

- Para o modelo quase Poisson assumimos a seguinte relação média-variância:

$$V(\mu_i) = \phi\mu_i. \tag{20}$$

- O parâmetro de dispersão ϕ é estimado conforme (18).
- Resumindo, o modelo quase Poisson equivale ao ajuste do modelo Poisson em que os erros padrões dos $\hat{\beta}'s$ ficam multiplicados por $\sqrt{\phi}$.

Modelo quase binomial

- O modelo quase binomial baseia-se na seguinte função de variância:

$$V(\pi_i) = \phi \frac{\pi_i(1-\pi_i)}{m_i}, \tag{21}$$

para a proporção y_i.

- A estimativa do parâmetro de dispersão ϕ baseia-se na estatística:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{(\hat{\pi}_i(1-\hat{\pi}))/m_i}. \tag{22}$$

- Novamente, para o modelo quase binomial, a matriz de covariância assintótica do modelo binomial $Var(\beta)$ fica multiplicada por ϕ.

Modelos de quase verossimilhança

- Importante destacar que, embora os modelos quase Poisson e quase Biomial sejam extensões dos modelos originais, resultantes da introdução de um parâmetro de dispersão, formas mais gerais são permitidas para a variância.

- Na modelagem de dados de contagem, como alternativa ao modelo quase Poisson pode-se especificar, por exemplo, a variância como $\phi\mu^2$;

- Na modelagem de dados binários agrupados, como alternativa ao modelo quase binomial pode-se especificar, por exemplo, a variância como $\phi\pi^2(1-\pi)^2$.

Quase verossimilhança e especificação do modelo

- Para o modelo $\eta_i = g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, as estimativas de quase verossimilhança são as soluções das equações quase escore:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right) \frac{(y_i - \mu_i)}{V(\mu_i)} = \mathbf{0}. \quad (23)$$

- Observe que essas equações são idênticas às equações de verossimilhança apresentadas em 17, substituindo:

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}. \quad (24)$$

Equações de estimação e propriedades da quase verossimilhança

- Se a distribuição de y_i pertence à família exponencial, então as equações apresentadas de fato correspondem a equações de verossimilhança.
- Neste ponto, no entanto, estamos considerando casos que não a distribuição de y_i não pertence à família exponencial.
- O método de quase verossimilhança trata as equações quase escore como as derivadas parciais de uma função de quase log-verossimilhança.

Equações de estimação e propriedades da quase verossimilhança

- Os estimadores de quase verossimilhança (que maximizam a equação de quase verossimilhança) compartilham propriedades similares aos estimadores de máxima verossimilhança:
  - São assintoticamente eficientes (sob especificação correta de μ_i e $V(\mu_i)$;
  - $\hat{\boldsymbol{\beta}}$ tem distribuição assintótica normal com média $\boldsymbol{\beta}$ (não viciado), com matriz de covariâncias aproximada:

$$\boldsymbol{V} = Var(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' [Var(y_i|\boldsymbol{x_i})] \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\right]^{-1}, \tag{25}$$

em que $Var(y_i|\boldsymbol{x_i}) = \phi V(\mu_i)$.

Estimador robusto para $Var(\hat{\beta})$

- Um resultado chave para os estimadores de quase verossimilhança é que eles são consistentes (convergem em probabilidade para β) ainda que $V(\mu_i)$ seja mal especificada.
- No entanto, se a variância não for corretamente especificada ($Var(y_i) \neq \phi V(\mu_i)$), como consequência a matriz de covariâncias de β já não é dada por $Var(\beta)$, conforme descrito em (25).
- Neste caso, mostra-se que a real matriz de covariâncias assintótica fica dada por:

$$Var^*(\hat{\beta}) = \boldsymbol{V}\left[\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)'\left[\frac{Var(y_i|\boldsymbol{x_i})}{[\phi V(\mu_i)]^2}\right]\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)\right]\boldsymbol{V} \tag{26}$$

Estimador robusto para $Var(\hat{\beta})$

- Uma forma de contornar a possível especificação incorreta de $Var(y_i|\boldsymbol{x_i})$ baseia-se na sua estimação empírica, substituindo-a na fórmula de $Var^*(\hat{\beta})$ por $(y_i - \hat{\mu}_i)$, para $i = 1, 2, ..., n$.

- O estimador resultante para para $Var(\hat{\beta})$ é denominado *estimador sanduíche*, uma vez que a evidência empírica encontra-se ensanduichada entre as matrizes de covariância produzidas pelo modelo.

- As inferências são produzidas da maneira usual, baseadas na normalidade assintótica de $\hat{\beta}$ utilizando o estimador sanduíche para $Var^{(}\hat{\beta})$

O problema do excesso de zeros

Modelos para dados inflacionados de zeros

- Outro problema frequente na análise de dados de contagens é o excesso de zeros.
- Imagine que a resposta de interesse se refira à seguinte questão: *Quantas vezes você foi pescar no último mês?*
  - Alguns dos entrevistados de fato foram pescar no último mês, tendo-se contagens maiores que zero para eles;
  - Alguns dos entrevistados, casualmente, não foram pescar no último mês. Talvez sejam pessoas que ocasionalmente saem para pescar, mas não tiveram tempo ou oportunidade para isso no mês passado (*zeros aleatórios*);
  - Alguns dos entrevistados não foram pescar no último mês simplesmente por que eles *nunca* fazem isso (*zeros estruturais*).

Modelos para dados inflacionados de zeros

- Em situações desse tipo, a frequência de zeros, consequente da soma dos zeros aleatórios e estruturais, configura um excedente que não é bem ajustado pelas distribuições Poisson e binomial negativa;

- Nesses casos, é comum se ter uma distribuição bimodal (com uma moda em zero e a outra na contagem não nula mais provável);

- Há diferentes estratégias para lidar com excessos de zeros, dentre as quais vamos destacar as versões inflacionadas dos modelos Poisson e binomial negativa e os modelos de barreira (*hurdle models*).

Modelos Poisson (ZIP) e binomial negativo (ZIBN) inflacionados de zeros

- Os modelos inflacionados de zeros são modelos de mistura de uma distribuição para contagens (Poisson, binomial negativa...) com uma distribuição degenerada em zero (igual a zero com probabilidade um).
- O modelo Poisson inflacionado de zeros (*Zero Inflated Poisson* - ZIP) é definido por:

$$y_i \sim \begin{cases} 0 & \text{com probabilidade } \pi_i \\ Poisson(\mu_i) & \text{com probabilidade } 1-\pi_i \end{cases} \tag{27}$$

- A distribuição de probabilidade não condicional fica dada por:

$$\begin{aligned} P(y_i = 0) &= \pi_i + (1-\pi_i)e^{-\mu_i}, \\ P(y_i = k) &= (1-\pi_i)\frac{e^{-\mu_i}\mu_i^k}{k!}, \quad 0 < \pi_i < 1, \mu_i > 0, k = 1, 2, 3, \ldots \end{aligned} \tag{28}$$

Modelos Poisson (ZIP) e binomial negativo (ZIBN) inflacionados de zeros

- No contexto de regressão, podemos considerar covariáveis para explicar tanto μ_i quanto π_i (que não precisam ser necessariamente as mesmas).
- Para π_i, por se tratar de uma probabilidade, utiliza-se alguma função de ligação para dados binários;
- Para μ_i, parâmetro da Poisson, pode-se considerar função de ligação logarítmica, raiz quadrada,...
- Uma possível configuração para o modelo é apresentada na sequência:

$$\begin{aligned} & y_i|\boldsymbol{x}_i \sim ZIP(\mu_i, \pi_i) \\ & log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_{1i}\boldsymbol{\beta}_1; \quad log(\mu_i) = \boldsymbol{x}_{2i}\boldsymbol{\beta}_2. \end{aligned} \tag{29}$$

Modelos Poisson (ZIP) e binomial negativo (ZIBN) inflacionados de zeros

- A média e a variância do modelo Poisson inflacionado de zeros ficam dadas por:

$$E(y_i) = (1 - \pi_i)\mu_i; \tag{30}$$

$$Var(y_i) = \mu_i(1 - \pi_i)(1 + \pi_i\mu_i). \tag{31}$$

- Observe que $Var(y_i) > E(y_i)$, configurando super dispersão em relação à distribuição Poisson.
- Os vetores de parâmetros $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ são usualmente estimados pelo método da máxima verossimilhança.

Modelos Poisson (ZIP) e binomial negativo (ZIBN) inflacionados de zeros

- Mesmo levando em consideração o excesso de zeros, ainda pode haver superdispersão quanto à parte do modelo correspondente à Poisson.

- Pode-se acomodar isso substituindo a distribuição Poisson pela binomial negativa (*Zero Inflated Negative Binomial* - ZINB) como componente da mistura:

$$y_i \sim \begin{cases} 0 & \text{com probabilidade } \pi_i \\ NegBin(\mu_i, \phi) & \text{com probabilidade } 1 - \pi_i. \end{cases} \tag{32}$$

Modelos Poisson (ZIP) e binomial negativo (ZIBN) inflacionados de zeros

- A média e a variância do modelo ZIBN são dadas, respectivamente, por:

$$E(y_i) = (1 - \pi_i)\mu_i; \tag{33}$$

$$Var(y_i) = \mu_i(1 - \pi_i)[1 + (\phi + \pi_i)\mu_i]. \tag{34}$$

- Observe que a variância da distribuição ZIBN supera a variância da distribuição ZIP, uma vez que $\phi > 0$.

Modelos Poisson (ZIP) e binomial negativo (ZIBN) inflacionados de zeros

- Assim como para o modelo ZIP, também para o ZIBN podemos incluir covariáveis tanto no ajuste de μ quanto no ajuste de π (na verdade, a biblioteca `gamlss` permite ainda incluir covariáveis para σ).
- Uma especificação usual para o modelo ZIBN é a seguinte:

$$\begin{aligned} & y_i|\boldsymbol{x}_i \sim ZIBN(\mu_i, \sigma, \pi_i) \\ & log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_{1i}\boldsymbol{\beta}_1; \quad log(\mu_i) = \boldsymbol{x}_{2i}\boldsymbol{\beta}_2. \end{aligned} \tag{35}$$

- A estimação dos parâmetros do modelo, usualmente, se dá pelo método da máxima verossimilhança.

Modelos de barreira (ou modelos *zero alterados*)

- Uma abordagem alternativa é modelar a inflação de zeros usando um modelo em duas partes:
    - **Parte 1:** A primeira parte pode ser um modelo logístico ou probito que indica se a resposta é zero ou positiva;
    - **Parte 2:** Condicional a um resultado positivo na parte 1, a parte 2 usa um modelo truncado (com probabilidade nula) em zero (com as demais probabilidades ajustadas tal que a soma das probabilidades, para o modelo resultante, seja um).
- Diferentemente dos modelos inflacionados, os modelos de barreira acomodam tanto inflação quanto deflação de zeros.

Modelos de barreira (ou modelos *zero alterados*)

- Suponha que a primeira parte do modelo defina probabilidades $P(y_i = 0) = \pi_i$ e $P(y_i > 0) = 1 - \pi_i$ e que a parte positiva de Y_i segue uma distribuição discreta $f(y_i; \mu_i)$ truncada em zero.
- Como resultado, temos a seguinte distribuição para y_i:

$$
\begin{aligned}
P(y_i = 0) &= \pi_i, \\
P(y_i = k) &= (1 - \pi_i)\frac{f(k; \mu_i)}{1 - f(0; \mu_i)}, \quad k = 1, 2, 3, \dots
\end{aligned}
\tag{36}
$$

- Para o caso em que a distribuição na parte 2 é a Poisson, temos:

$$
\begin{aligned}
P(y_i = 0) &= \pi_i, \\
P(y_i = k) &= (1 - \pi_i)\frac{e^{-\mu_i}\mu_i^k/k!}{1 - e^{-\mu_i}}, \quad , 0 < \pi_i < 1, \mu_i > 0, k = 1, 2, 3, \dots
\end{aligned}
\tag{37}
$$

Modelos de barreira (ou modelos *zero alterados*)

- A média e a variância da distribuição Poisson zero alterada, descrita em (37), ficam dadas por:

$$E(y_i) = (1 - \pi_i)\frac{\mu_i}{1 - e^{-\mu_i}}; \tag{38}$$

$$Var(y_i) = (1 + \mu_i)E(y_i) - [E(y_i)]^2. \tag{39}$$

Modelos de barreira (ou modelos *zero alterados*)

- Também para o caso de modelos de barreira, podemos incluir covariáveis nas duas partes do modelo.

- Uma especificação usual para o modelo de regressão Poisson zero alterado é a seguinte:

$$\begin{aligned} & y_i|\boldsymbol{x}_i \sim ZAP(\mu_i, \pi_i) \\ & log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_{1i}\boldsymbol{\beta}_1; \quad log(\mu_i) = \boldsymbol{x}_{2i}\boldsymbol{\beta}_2. \end{aligned} \tag{40}$$

- Um modelo de regressão binomial negativa zero alterado pode ser definido trocando a distribuição Poisson truncada pela binomial negativa