Identificação de Ineficiências de Licitações Públicas por Meio do Modelo de Mistura de Regressões Binomiais Negativas.

Aluno: Emerson Pazeto
Orientador: Prof. Dr. Alessandro Sarnaglia
Apresentação: XV Semana de Estatística - UFES

Novembro 2024

Sumário

Sumário

1. Introdução

Na atualidade, a sociedade tem exigido cada vez mais eficiência dos gastos públicos. Em particular, um **indicador** de ineficiência na celebração dos contratos licitatórios é o **número de aditivos**.

Esses dados podem ser baixados abertamente no portal do SIASG, entretanto são disponibilizados de maneira limitada por acesso. Houve necessidade de implementação de um robô para coleta e agregação dos mesmos.

Principal problema encontrado na modelagem do número de aditivos: **excesso de zeros**. Veja a Figura 1.

1. Introdução

Na atualidade, a sociedade tem exigido cada vez mais eficiência dos gastos públicos. Em particular, um **indicador** de ineficiência na celebração dos contratos licitatórios é o **número de aditivos**.

Esses dados podem ser baixados abertamente no portal do SIASG, entretanto são disponibilizados de maneira limitada por acesso. Houve necessidade de implementação de um robô para coleta e agregação dos mesmos.

Principal problema encontrado na modelagem do número de aditivos: **excesso de zeros**. Veja a Figura 1.

1. Introdução

Na atualidade, a sociedade tem exigido cada vez mais eficiência dos gastos públicos. Em particular, um **indicador** de ineficiência na celebração dos contratos licitatórios é o **número de aditivos**.

Esses dados podem ser baixados abertamente no portal do SIASG, entretanto são disponibilizados de maneira limitada por acesso. Houve necessidade de implementação de um robô para coleta e agregação dos mesmos.

Principal problema encontrado na modelagem do número de aditivos: **excesso de zeros**. Veja a Figura 1.

1. Introdução

Na atualidade, a sociedade tem exigido cada vez mais eficiência dos gastos públicos. Em particular, um **indicador** de ineficiência na celebração dos contratos licitatórios é o **número de aditivos**.

Esses dados podem ser baixados abertamente no portal do SIASG, entretanto são disponibilizados de maneira limitada por acesso. Houve necessidade de implementação de um robô para coleta e agregação dos mesmos.

Principal problema encontrado na modelagem do número de aditivos: **excesso de zeros**. Veja a Figura 1.

1. Introdução

Na atualidade, a sociedade tem exigido cada vez mais eficiência dos gastos públicos. Em particular, um **indicador** de ineficiência na celebração dos contratos licitatórios é o **número de aditivos**.

Esses dados podem ser baixados abertamente no portal do SIASG, entretanto são disponibilizados de maneira limitada por acesso. Houve necessidade de implementação de um robô para coleta e agregação dos mesmos.

Principal problema encontrado na modelagem do número de aditivos: **excesso de zeros**. Veja a Figura 1.

1. Introdução

Na atualidade, a sociedade tem exigido cada vez mais eficiência dos gastos públicos. Em particular, um **indicador** de ineficiência na celebração dos contratos licitatórios é o **número de aditivos**.

Esses dados podem ser baixados abertamente no portal do SIASG, entretanto são disponibilizados de maneira limitada por acesso. Houve necessidade de implementação de um robô para coleta e agregação dos mesmos.

Principal problema encontrado na modelagem do número de aditivos: **excesso de zeros**. Veja a Figura 1.

1. Introdução

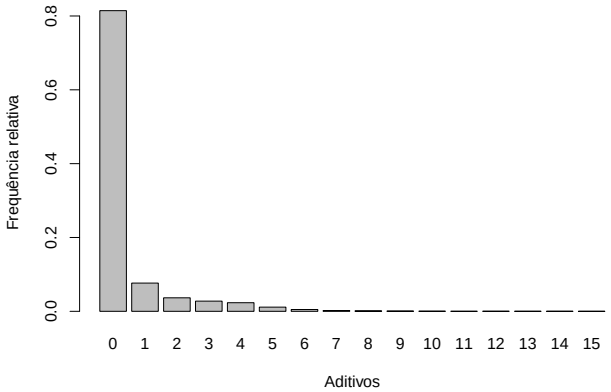


Figura 1: Número de aditivos contratuais em processos licitatórios federais de 2011 a 2019.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Alternativas iniciais de modelagem (podem ser ajustadas via função `glm` do `R`):

- Contagem: Regressão Poisson;
- Contagem + excesso de zeros (sobredispersão): Regressão Binomial Negativa (NB);

Os modelos acima não discriminam entre contratos com "poucos" e "muitos" aditivos. **Ideia:** Regressões Poisson inflada de zeros (ZIP) e Binomial Negativa inflada de zeros (ZINB). Podemos usar a função `gamlss` no `R` (pacote `gamlss`).

Problema

No nosso contexto, regressões infladas de zero ainda são restritivas, pois criam uma classe de contratos com obrigatoriamente zero aditivos, que seriam os eficientes. Mas podem existir contratos "eficientes" com mais de zero aditivos.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

1. Introdução

Solução

Considerar uma mistura de regressões binomiais negativas (Mix-NB). Idealmente, essa estratégia discriminará os contratos em dois gruposm caracterizados por distribuições: com alta probabilidade em zero (eficientes); e com locação e dispersão alta (ineficientes).

Inferência:

- Estimação e discriminação entre contratos "eficientes" e "ineficientes": **Algoritmo EM**;
- Diagnóstico de Ajuste: **KS bootstrap dos Resíduos Quantílicos Randomizados (RQR)**;
- Significância dos parâmetros: **Intervalos de Confiança bootstrap**.

O modelo Mix-NB será comparado com os modelos de regressão Poisson, NB, ZIP e ZINB.

Sumário

2. Metodologia

Os dados são formados pela variável resposta Y_i e covariáveis X_i e Z_i.

Os dados serão ajustados aos modelos:

- de regressão Poisson (caso especial de um Modelo Linear Generalizado, GLM) e NB (GLM se o parâmetro de dispersão fixo);
- de regressão ZIP e ZINB (casos especiais de Modelos Aditivos Generalizados de Locação Escala e Forma, GAMLSS);
- de Mistura de Regressões de Binomiais Negativas (proposta deste trabalho).

2. Metodologia

Os dados são formados pela variável resposta Y_i e covariáveis X_i e Z_i.

Os dados serão ajustados aos modelos:

- de regressão Poisson (caso especial de um Modelo Linear Generalizado, GLM) e NB (GLM se o parâmetro de dispersão fixo);
- de regressão ZIP e ZINB (casos especiais de Modelos Aditivos Generalizados de Locação Escala e Forma, GAMLSS);
- de Mistura de Regressões de Binomiais Negativas (proposta deste trabalho).

2. Metodologia

Os dados são formados pela variável resposta Y_i e covariáveis X_i e Z_i.

Os dados serão ajustados aos modelos:

- de regressão Poisson (caso especial de um Modelo Linear Generalizado, GLM) e NB (GLM se o parâmetro de dispersão fixo);
- de regressão ZIP e ZINB (casos especiais de Modelos Aditivos Generalizados de Locação Escala e Forma, GAMLSS);
- de Mistura de Regressões de Binomiais Negativas (proposta deste trabalho).

2. Metodologia

Os dados são formados pela variável resposta Y_i e covariáveis X_i e Z_i.

Os dados serão ajustados aos modelos:

- de regressão Poisson (caso especial de um Modelo Linear Generalizado, GLM) e NB (GLM se o parâmetro de dispersão fixo);
- de regressão ZIP e ZINB (casos especiais de Modelos Aditivos Generalizados de Locação Escala e Forma, GAMLSS);
- de Mistura de Regressões de Binomiais Negativas (proposta deste trabalho).

Sumário

2.1 GLM

Modelos de regressão para variáveis resposta na família exponencial (FE).

Dizemos que Y pertence à FE se sua distribuição pode ser escrita na forma:

$$f(y|\theta,\phi) = \exp\left(\frac{[y\theta - b(\theta)]}{a(\phi)} + c(y,\phi)\right),$$

em que θ é o parâmetro natural e ϕ é o parâmetro de escala e $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções.

Um GLM pode ser definido por meio da **função de ligação** $g(\mu_i) = \eta_i = X_i'\beta$. Dizemos que η_i é o preditor linear e β é o vetor de coeficientes. Se $g(\cdot)$ é tal que $\eta_i = g(\mu_i) = \theta_i$, dizemos que $g(\cdot)$ é a **ligação canônica**.

2.1 GLM

Modelos de regressão para variáveis resposta na família exponencial (FE).

Dizemos que Y pertence à FE se sua distribuição pode ser escrita na forma:

$$f(y|\theta,\phi) = \exp\left(\frac{[y\theta - b(\theta)]}{a(\phi)} + c(y,\phi)\right),$$

em que θ é o parâmetro natural e ϕ é o parâmetro de escala e $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções.

Um GLM pode ser definido por meio da **função de ligação** $g(\mu_i) = \eta_i = X_i'\beta$. Dizemos que η_i é o preditor linear e β é o vetor de coeficientes. Se $g(\cdot)$ é tal que $\eta_i = g(\mu_i) = \theta_i$, dizemos que $g(\cdot)$ é a **ligação canônica**.

2.1 GLM

Modelos de regressão para variáveis resposta na família exponencial (FE).

Dizemos que Y pertence à FE se sua distribuição pode ser escrita na forma:

$$f(y|\theta,\phi) = \exp\left(\frac{[y\theta - b(\theta)]}{a(\phi)} + c(y,\phi)\right),$$

em que θ é o parâmetro natural e ϕ é o parâmetro de escala e $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções.

Um GLM pode ser definido por meio da **função de ligação** $g(\mu_i) = \eta_i = X_i'\beta$. Dizemos que η_i é o preditor linear e β é o vetor de coeficientes. Se $g(\cdot)$ é tal que $\eta_i = g(\mu_i) = \theta_i$, dizemos que $g(\cdot)$ é a **ligação canônica**.

2.1 GLM

Modelos de regressão para variáveis resposta na família exponencial (FE).

Dizemos que Y pertence à FE se sua distribuição pode ser escrita na forma:

$$f(y|\theta,\phi) = \exp\left(\frac{[y\theta - b(\theta)]}{a(\phi)} + c(y,\phi)\right),$$

em que θ é o parâmetro natural e ϕ é o parâmetro de escala e $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções.

Um GLM pode ser definido por meio da **função de ligação** $g(\mu_i) = \eta_i = X_i'\beta$. Dizemos que η_i é o preditor linear e β é o vetor de coeficientes. Se $g(\cdot)$ é tal que $\eta_i = g(\mu_i) = \theta_i$, dizemos que $g(\cdot)$ é a **ligação canônica**.

2.1 GLM - Regressão Poisson

Suponha $Y_i \sim \text{Pois}(\mu_i)$. Sua distribuição pode ser escrita na forma da FE como

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\{[y_i \log(\mu_i) - \mu_i] + [-\log(y_i!)]\},$$

em que o parâmetro natural é $\theta_i = \log(\mu_i)$, $a(\phi) = 1$, $b(\theta_i) = \mu_i = e^{\theta_i}$ e $c(y, \phi) = -\log(y_i!)$.

Portanto, a ligação canônica é $g(u) = \log(u)$. O modelo de Regressão Poisson é obtido ao utilizar a ligação canônica $\log(\mu_i) = X_i'\beta$.

2.1 GLM - Regressão Poisson

Suponha $Y_i \sim \text{Pois}(\mu_i)$. Sua distribuição pode ser escrita na forma da FE como

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\{[y_i \log(\mu_i) - \mu_i] + [-\log(y_i!)]\},$$

em que o parâmetro natural é $\theta_i = \log(\mu_i)$, $a(\phi) = 1$, $b(\theta_i) = \mu_i = e^{\theta_i}$ e $c(y, \phi) = -\log(y_i!)$.

Portanto, a ligação canônica é $g(u) = \log(u)$. O modelo de Regressão Poisson é obtido ao utilizar a ligação canônica $\log(\mu_i) = X_i'\beta$.

2.1 GLM - Regressão Poisson

Suponha $Y_i \sim \text{Pois}(\mu_i)$. Sua distribuição pode ser escrita na forma da FE como

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\{[y_i \log(\mu_i) - \mu_i] + [-\log(y_i!)]\},$$

em que o parâmetro natural é $\theta_i = \log(\mu_i)$, $a(\phi) = 1$, $b(\theta_i) = \mu_i = e^{\theta_i}$ e $c(y, \phi) = -\log(y_i!)$.

Portanto, a ligação canônica é $g(u) = \log(u)$. O modelo de Regressão Poisson é obtido ao utilizar a ligação canônica $\log(\mu_i) = X_i'\beta$.

2.1 GLM - Regressão Poisson

Suponha $Y_i \sim \text{Pois}(\mu_i)$. Sua distribuição pode ser escrita na forma da FE como

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\{[y_i \log(\mu_i) - \mu_i] + [-\log(y_i!)]\},$$

em que o parâmetro natural é $\theta_i = \log(\mu_i)$, $a(\phi) = 1$, $b(\theta_i) = \mu_i = e^{\theta_i}$ e $c(y, \phi) = -\log(y_i!)$.

Portanto, a ligação canônica é $g(u) = \log(u)$. O modelo de Regressão Poisson é obtido ao utilizar a ligação canônica $\log(\mu_i) = X_i'\beta$.

2.1 GLM - Regressão NB

Suponha $Y_i \sim \text{NB}(\mu_i, \omega)$. Sua distribuição é dada por

$$f(y_i|\mu_i,\omega) = \frac{\Gamma(y_i+\omega)}{\Gamma(y_i+1)\Gamma(\omega)}\left(\frac{\omega}{\mu_i+\omega}\right)^{\omega}\left(\frac{\mu_i}{\mu_i+\omega}\right)^{y_i} \tag{1}$$

$$= \exp\left\{y_i \log\left(\frac{\mu_i}{\mu_i+\omega}\right) - \left[-\omega\log\left(\frac{\omega}{\mu_i+\omega}\right)\right] + c(y_i,\omega)\right\}, \tag{2}$$

em que $c(y_i,\phi) = \log(\Gamma(y_i+\omega)) - \log(\Gamma(y_i+1)) - \log(\Gamma(\omega))$, $\theta = \log\left(\frac{\mu_i}{\mu_i+\omega}\right)$.

A distribuição na Equação 2 só pertence a FE se ω é fixo. A regressão NB é reparametrizada com $\alpha = \omega^{-1}$ e caracterizada pela ligação (não canônica) $g(\mu_i) = \log(\mu_i) = X_i'\beta$ (para comparar com o modelo Poisson). O modelo NB converge para o Poisson se $\alpha \to 0$.

Estimação: método iterativo com α fixo (o que corresponde a um GLM) em cada iteração e atualização de α entre iterações.

2.1 GLM - Regressão NB

Suponha $Y_i \sim \text{NB}(\mu_i, \omega)$. Sua distribuição é dada por

$$f(y_i|\mu_i,\omega) = \frac{\Gamma(y_i+\omega)}{\Gamma(y_i+1)\Gamma(\omega)}\left(\frac{\omega}{\mu_i+\omega}\right)^{\omega}\left(\frac{\mu_i}{\mu_i+\omega}\right)^{y_i} \tag{1}$$

$$= \exp\left\{y_i \log\left(\frac{\mu_i}{\mu_i+\omega}\right) - \left[-\omega\log\left(\frac{\omega}{\mu_i+\omega}\right)\right] + c(y_i,\omega)\right\}, \tag{2}$$

em que $c(y_i,\phi) = \log(\Gamma(y_i+\omega)) - \log(\Gamma(y_i+1)) - \log(\Gamma(\omega))$, $\theta = \log\left(\frac{\mu_i}{\mu_i+\omega}\right)$.

A distribuição na Equação 2 só pertence a FE se ω é fixo. A regressão NB é reparametrizada com $\alpha = \omega^{-1}$ e caracterizada pela ligação (não canônica) $g(\mu_i) = \log(\mu_i) = X_i'\beta$ (para comparar com o modelo Poisson). O modelo NB converge para o Poisson se $\alpha \to 0$.

Estimação: método iterativo com α fixo (o que corresponde a um GLM) em cada iteração e atualização de α entre iterações.

2.1 GLM - Regressão NB

Suponha $Y_i \sim \text{NB}(\mu_i, \omega)$. Sua distribuição é dada por

$$f(y_i|\mu_i,\omega) = \frac{\Gamma(y_i+\omega)}{\Gamma(y_i+1)\Gamma(\omega)}\left(\frac{\omega}{\mu_i+\omega}\right)^{\omega}\left(\frac{\mu_i}{\mu_i+\omega}\right)^{y_i} \tag{1}$$

$$= \exp\left\{y_i \log\left(\frac{\mu_i}{\mu_i+\omega}\right) - \left[-\omega\log\left(\frac{\omega}{\mu_i+\omega}\right)\right] + c(y_i,\omega)\right\}, \tag{2}$$

em que $c(y_i,\phi) = \log(\Gamma(y_i+\omega)) - \log(\Gamma(y_i+1)) - \log(\Gamma(\omega))$, $\theta = \log\left(\frac{\mu_i}{\mu_i+\omega}\right)$.

A distribuição na Equação 2 só pertence a FE se ω é fixo. A regressão NB é reparametrizada com $\alpha = \omega^{-1}$ e caracterizada pela ligação (não canônica) $g(\mu_i) = \log(\mu_i) = X_i'\beta$ (para comparar com o modelo Poisson). O modelo NB converge para o Poisson se $\alpha \to 0$.

Estimação: método iterativo com α fixo (o que corresponde a um GLM) em cada iteração e atualização de α entre iterações.

2.1 GLM - Regressão NB

Suponha $Y_i \sim \text{NB}(\mu_i, \omega)$. Sua distribuição é dada por

$$f(y_i|\mu_i,\omega) = \frac{\Gamma(y_i+\omega)}{\Gamma(y_i+1)\Gamma(\omega)}\left(\frac{\omega}{\mu_i+\omega}\right)^{\omega}\left(\frac{\mu_i}{\mu_i+\omega}\right)^{y_i} \tag{1}$$

$$= \exp\left\{y_i \log\left(\frac{\mu_i}{\mu_i+\omega}\right) - \left[-\omega\log\left(\frac{\omega}{\mu_i+\omega}\right)\right] + c(y_i,\omega)\right\}, \tag{2}$$

em que $c(y_i,\phi) = \log(\Gamma(y_i+\omega)) - \log(\Gamma(y_i+1)) - \log(\Gamma(\omega))$, $\theta = \log\left(\frac{\mu_i}{\mu_i+\omega}\right)$.

A distribuição na Equação 2 só pertence a FE se ω é fixo. A regressão NB é reparametrizada com $\alpha = \omega^{-1}$ e caracterizada pela ligação (não canônica) $g(\mu_i) = \log(\mu_i) = X_i'\beta$ (para comparar com o modelo Poisson). O modelo NB converge para o Poisson se $\alpha \to 0$.

Estimação: método iterativo com α fixo (o que corresponde a um GLM) em cada iteração e atualização de α entre iterações.

2.1 GLM - Regressão NB

Suponha $Y_i \sim \text{NB}(\mu_i, \omega)$. Sua distribuição é dada por

$$f(y_i|\mu_i,\omega) = \frac{\Gamma(y_i+\omega)}{\Gamma(y_i+1)\Gamma(\omega)}\left(\frac{\omega}{\mu_i+\omega}\right)^{\omega}\left(\frac{\mu_i}{\mu_i+\omega}\right)^{y_i} \tag{1}$$

$$= \exp\left\{y_i \log\left(\frac{\mu_i}{\mu_i+\omega}\right) - \left[-\omega\log\left(\frac{\omega}{\mu_i+\omega}\right)\right] + c(y_i,\omega)\right\}, \tag{2}$$

em que $c(y_i,\phi) = \log(\Gamma(y_i+\omega)) - \log(\Gamma(y_i+1)) - \log(\Gamma(\omega))$, $\theta = \log\left(\frac{\mu_i}{\mu_i+\omega}\right)$.

A distribuição na Equação 2 só pertence a FE se ω é fixo. A regressão NB é reparametrizada com $\alpha = \omega^{-1}$ e caracterizada pela ligação (não canônica) $g(\mu_i) = \log(\mu_i) = X_i'\beta$ (para comparar com o modelo Poisson). O modelo NB converge para o Poisson se $\alpha \to 0$.

Estimação: método iterativo com α fixo (o que corresponde a um GLM) em cada iteração e atualização de α entre iterações.

2.1 GLM - Regressão NB

Suponha $Y_i \sim \text{NB}(\mu_i, \omega)$. Sua distribuição é dada por

$$f(y_i|\mu_i,\omega) = \frac{\Gamma(y_i+\omega)}{\Gamma(y_i+1)\Gamma(\omega)}\left(\frac{\omega}{\mu_i+\omega}\right)^{\omega}\left(\frac{\mu_i}{\mu_i+\omega}\right)^{y_i} \tag{1}$$

$$= \exp\left\{y_i \log\left(\frac{\mu_i}{\mu_i+\omega}\right) - \left[-\omega\log\left(\frac{\omega}{\mu_i+\omega}\right)\right] + c(y_i,\omega)\right\}, \tag{2}$$

em que $c(y_i,\phi) = \log(\Gamma(y_i+\omega)) - \log(\Gamma(y_i+1)) - \log(\Gamma(\omega))$, $\theta = \log\left(\frac{\mu_i}{\mu_i+\omega}\right)$.

A distribuição na Equação 2 só pertence a FE se ω é fixo. A regressão NB é reparametrizada com $\alpha = \omega^{-1}$ e caracterizada pela ligação (não canônica) $g(\mu_i) = \log(\mu_i) = X_i'\beta$ (para comparar com o modelo Poisson). O modelo NB converge para o Poisson se $\alpha \to 0$.

Estimação: método iterativo com α fixo (o que corresponde a um GLM) em cada iteração e atualização de α entre iterações.

2.1 GLM - Regressão NB

Suponha $Y_i \sim \text{NB}(\mu_i, \omega)$. Sua distribuição é dada por

$$f(y_i|\mu_i,\omega) = \frac{\Gamma(y_i+\omega)}{\Gamma(y_i+1)\Gamma(\omega)}\left(\frac{\omega}{\mu_i+\omega}\right)^{\omega}\left(\frac{\mu_i}{\mu_i+\omega}\right)^{y_i} \tag{1}$$

$$= \exp\left\{y_i \log\left(\frac{\mu_i}{\mu_i+\omega}\right) - \left[-\omega\log\left(\frac{\omega}{\mu_i+\omega}\right)\right] + c(y_i,\omega)\right\}, \tag{2}$$

em que $c(y_i,\phi) = \log(\Gamma(y_i+\omega)) - \log(\Gamma(y_i+1)) - \log(\Gamma(\omega))$, $\theta = \log\left(\frac{\mu_i}{\mu_i+\omega}\right)$.

A distribuição na Equação 2 só pertence a FE se ω é fixo. A regressão NB é reparametrizada com $\alpha = \omega^{-1}$ e caracterizada pela ligação (não canônica) $g(\mu_i) = \log(\mu_i) = X_i'\beta$ (para comparar com o modelo Poisson). O modelo NB converge para o Poisson se $\alpha \to 0$.

Estimação: método iterativo com α fixo (o que corresponde a um GLM) em cada iteração e atualização de α entre iterações.

Sumário

2.2 GAMLSS

A família GAMLSS permite o ajuste de distribuições com até 4 parêmetros por meio de co-variáveis. São modelos de regressão capazes de explicar características mais complexas da variável resposta. Por exemplo, o **excesso de zeros**.

No GAMLSS, assumimos que a variável resposta satisfaz $Y_i \sim \mathcal{D}(\mu_i, \sigma_i, \nu_i, \tau_i)$, $i = 1, 2, \ldots, n$, em que $\mathcal{D}$ é uma distribuição de, no máximo, quatro parâmetros, em que

$$g_1(\mu_i) = \eta_{i1} = X'_{i1}\beta_1; \qquad g_2(\sigma_i) = \eta_{i2} = X'_{i2}\beta_2;$$

$$g_3(\nu_i) = \eta_{i3} = X'_{i3}\beta_3; \qquad g_4(\tau_i) = \eta_{i4} = X'_{i4}\beta_4.$$

Casos especiais na família GAMLSS incluem distribuições com 1 ($\theta = \mu$), 2 ($\theta = (\mu, \sigma)$), 3 ($\theta = (\mu, \sigma, \nu)$) ou 4 ($\theta = (\mu, \sigma, \nu, \tau)$) parâmetros.

2.2 GAMLSS

A família GAMLSS permite o ajuste de distribuições com até 4 parêmetros por meio de co-variáveis. São modelos de regressão capazes de explicar características mais complexas da variável resposta. Por exemplo, o **excesso de zeros**.

No GAMLSS, assumimos que a variável resposta satisfaz $Y_i \sim \mathcal{D}(\mu_i, \sigma_i, \nu_i, \tau_i)$, $i = 1, 2, \ldots, n$, em que $\mathcal{D}$ é uma distribuição de, no máximo, quatro parâmetros, em que

$$g_1(\mu_i) = \eta_{i1} = X'_{i1}\beta_1; \qquad g_2(\sigma_i) = \eta_{i2} = X'_{i2}\beta_2;$$

$$g_3(\nu_i) = \eta_{i3} = X'_{i3}\beta_3; \qquad g_4(\tau_i) = \eta_{i4} = X'_{i4}\beta_4.$$

Casos especiais na família GAMLSS incluem distribuições com 1 ($\theta = \mu$), 2 ($\theta = (\mu, \sigma)$), 3 ($\theta = (\mu, \sigma, \nu)$) ou 4 ($\theta = (\mu, \sigma, \nu, \tau)$) parâmetros.

2.2 GAMLSS

A família GAMLSS permite o ajuste de distribuições com até 4 parêmetros por meio de co-variáveis. São modelos de regressão capazes de explicar características mais complexas da variável resposta. Por exemplo, o **excesso de zeros**.

No GAMLSS, assumimos que a variável resposta satisfaz $Y_i \sim \mathcal{D}(\mu_i, \sigma_i, \nu_i, \tau_i)$, $i = 1, 2, \ldots, n$, em que $\mathcal{D}$ é uma distribuição de, no máximo, quatro parâmetros, em que

$$g_1(\mu_i) = \eta_{i1} = X'_{i1}\beta_1; \qquad g_2(\sigma_i) = \eta_{i2} = X'_{i2}\beta_2;$$

$$g_3(\nu_i) = \eta_{i3} = X'_{i3}\beta_3; \qquad g_4(\tau_i) = \eta_{i4} = X'_{i4}\beta_4.$$

Casos especiais na família GAMLSS incluem distribuições com 1 ($\theta = \mu$), 2 ($\theta = (\mu, \sigma)$), 3 ($\theta = (\mu, \sigma, \nu)$) ou 4 ($\theta = (\mu, \sigma, \nu, \tau)$) parâmetros.

2.2 GAMLSS

A família GAMLSS permite o ajuste de distribuições com até 4 parêmetros por meio de covariáveis. São modelos de regressão capazes de explicar características mais complexas da variável resposta. Por exemplo, o **excesso de zeros**.

No GAMLSS, assumimos que a variável resposta satisfaz $Y_i \sim \mathcal{D}(\mu_i, \sigma_i, \nu_i, \tau_i)$, $i = 1, 2, \ldots, n$, em que $\mathcal{D}$ é uma distribuição de, no máximo, quatro parâmetros, em que

$$g_1(\mu_i) = \eta_{i1} = X'_{i1}\beta_1; \qquad g_2(\sigma_i) = \eta_{i2} = X'_{i2}\beta_2;$$

$$g_3(\nu_i) = \eta_{i3} = X'_{i3}\beta_3; \qquad g_4(\tau_i) = \eta_{i4} = X'_{i4}\beta_4.$$

Casos especiais na família GAMLSS incluem distribuições com 1 ($\theta = \mu$), 2 ($\theta = (\mu, \sigma)$), 3 ($\theta = (\mu, \sigma, \nu)$) ou 4 ($\theta = (\mu, \sigma, \nu, \tau)$) parâmetros.

2.2 GAMLSS

A família GAMLSS permite o ajuste de distribuições com até 4 parêmetros por meio de covariáveis. São modelos de regressão capazes de explicar características mais complexas da variável resposta. Por exemplo, o **excesso de zeros**.

No GAMLSS, assumimos que a variável resposta satisfaz $Y_i \sim \mathcal{D}(\mu_i, \sigma_i, \nu_i, \tau_i)$, $i = 1, 2, \ldots, n$, em que $\mathcal{D}$ é uma distribuição de, no máximo, quatro parâmetros, em que

$$g_1(\mu_i) = \eta_{i1} = X'_{i1}\beta_1; \qquad g_2(\sigma_i) = \eta_{i2} = X'_{i2}\beta_2;$$

$$g_3(\nu_i) = \eta_{i3} = X'_{i3}\beta_3; \qquad g_4(\tau_i) = \eta_{i4} = X'_{i4}\beta_4.$$

Casos especiais na família GAMLSS incluem distribuições com 1 ($\theta = \mu$), 2 ($\theta = (\mu, \sigma)$), 3 ($\theta = (\mu, \sigma, \nu)$) ou 4 ($\theta = (\mu, \sigma, \nu, \tau)$) parâmetros.

2.2 GAMLSS - Regressão Poisson Inflada de Zeros

Dizemos que $Y \sim \text{ZIP}(\mu, \sigma)$ se existe variável latente $G \sim \text{Ber}(\sigma)$, com $\sigma = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{Pois}(\mu).$$

Assim, a distribuição de Y é dada pela mistura

$$f(y|\mu, \sigma) = (1 - \sigma)\mathbb{1}(y = 0) + \sigma \frac{e^{-\mu}\mu^y}{y!},$$

em que $\mathbb{1}(A)$ é a indicadora do evento A.

A regressão ZIP é obtida quando $Y_i \sim ZIP(\mu_i, \sigma_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$ e $\text{logit}(\sigma_i) = X'_{i2}\beta_2$, em que $\text{logit}(\sigma) = \log\left(\frac{\sigma}{1-\sigma}\right)$.

Assim, o modelo de regressão ZIP é um caso especial da família GAMLSS.

2.2 GAMLSS - Regressão Poisson Inflada de Zeros

Dizemos que $Y \sim \text{ZIP}(\mu, \sigma)$ se existe variável latente $G \sim \text{Ber}(\sigma)$, com $\sigma = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{Pois}(\mu).$$

Assim, a distribuição de Y é dada pela mistura

$$f(y|\mu, \sigma) = (1 - \sigma)\mathbb{1}(y = 0) + \sigma \frac{e^{-\mu}\mu^y}{y!},$$

em que $\mathbb{1}(A)$ é a indicadora do evento A.

A regressão ZIP é obtida quando $Y_i \sim ZIP(\mu_i, \sigma_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$ e $\text{logit}(\sigma_i) = X'_{i2}\beta_2$, em que $\text{logit}(\sigma) = \log\left(\frac{\sigma}{1-\sigma}\right)$.

Assim, o modelo de regressão ZIP é um caso especial da família GAMLSS.

2.2 GAMLSS - Regressão Poisson Inflada de Zeros

Dizemos que $Y \sim \text{ZIP}(\mu, \sigma)$ se existe variável latente $G \sim \text{Ber}(\sigma)$, com $\sigma = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{Pois}(\mu).$$

Assim, a distribuição de Y é dada pela mistura

$$f(y|\mu, \sigma) = (1 - \sigma)\mathbb{1}(y = 0) + \sigma \frac{e^{-\mu}\mu^y}{y!},$$

em que $\mathbb{1}(A)$ é a indicadora do evento A.

A regressão ZIP é obtida quando $Y_i \sim ZIP(\mu_i, \sigma_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$ e $\text{logit}(\sigma_i) = X'_{i2}\beta_2$, em que $\text{logit}(\sigma) = \log\left(\frac{\sigma}{1-\sigma}\right)$.

Assim, o modelo de regressão ZIP é um caso especial da família GAMLSS.

2.2 GAMLSS - Regressão Poisson Inflada de Zeros

Dizemos que $Y \sim \text{ZIP}(\mu, \sigma)$ se existe variável latente $G \sim \text{Ber}(\sigma)$, com $\sigma = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{Pois}(\mu).$$

Assim, a distribuição de Y é dada pela mistura

$$f(y|\mu, \sigma) = (1 - \sigma)\mathbb{1}(y = 0) + \sigma \frac{e^{-\mu}\mu^y}{y!},$$

em que $\mathbb{1}(A)$ é a indicadora do evento A.

A regressão ZIP é obtida quando $Y_i \sim ZIP(\mu_i, \sigma_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$ e $\text{logit}(\sigma_i) = X'_{i2}\beta_2$, em que $\text{logit}(\sigma) = \log\left(\frac{\sigma}{1-\sigma}\right)$.

Assim, o modelo de regressão ZIP é um caso especial da família GAMLSS.

2.2 GAMLSS - Regressão NB Inflada de Zeros

Diz-se que $Y \sim \text{ZINB}(\mu, \sigma, \nu)$ se há variável latente $G \sim \text{Ber}(\nu)$, com $\nu = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu, \sigma).$$

Note que, nesse caso, $\sigma = \frac{1}{\alpha}$ na nossa parametrização.

Portanto, a distribuiçao de Y é dada pela mistura

$$f(y|\mu, \sigma, \nu) = (1 - \nu)\mathbb{1}(y = 0) + \nu \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left(\frac{\sigma\mu}{1 + \sigma\mu}\right)^y \left(\frac{1}{1 + \sigma\mu}\right)^y.$$

A regressão ZINB é definida por $Y_i \sim \text{ZINB}(\mu_i, \sigma_i, \nu_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$, $\log(\sigma_i) = X'_{i2}\beta_2$ e $\text{logit}(\nu_i) = X'_{i3}\beta_3$.

O modelo de regressão ZINB também é um membro da família GAMLSS.

2.2 GAMLSS - Regressão NB Inflada de Zeros

Diz-se que $Y \sim \text{ZINB}(\mu, \sigma, \nu)$ se há variável latente $G \sim \text{Ber}(\nu)$, com $\nu = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu, \sigma).$$

Note que, nesse caso, $\sigma = \frac{1}{\alpha}$ na nossa parametrização.

Portanto, a distribuiçao de Y é dada pela mistura

$$f(y|\mu, \sigma, \nu) = (1 - \nu)\mathbb{1}(y = 0) + \nu \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left(\frac{\sigma\mu}{1 + \sigma\mu}\right)^y \left(\frac{1}{1 + \sigma\mu}\right)^y.$$

A regressão ZINB é definida por $Y_i \sim \text{ZINB}(\mu_i, \sigma_i, \nu_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$, $\log(\sigma_i) = X'_{i2}\beta_2$ e $\text{logit}(\nu_i) = X'_{i3}\beta_3$.

O modelo de regressão ZINB também é um membro da família GAMLSS.

2.2 GAMLSS - Regressão NB Inflada de Zeros

Diz-se que $Y \sim \text{ZINB}(\mu, \sigma, \nu)$ se há variável latente $G \sim \text{Ber}(\nu)$, com $\nu = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu, \sigma).$$

Note que, nesse caso, $\sigma = \frac{1}{\alpha}$ na nossa parametrização.

Portanto, a distribuiçao de Y é dada pela mistura

$$f(y|\mu, \sigma, \nu) = (1 - \nu)\mathbb{1}(y = 0) + \nu \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left(\frac{\sigma\mu}{1 + \sigma\mu}\right)^y \left(\frac{1}{1 + \sigma\mu}\right)^y.$$

A regressão ZINB é definida por $Y_i \sim \text{ZINB}(\mu_i, \sigma_i, \nu_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$, $\log(\sigma_i) = X'_{i2}\beta_2$ e $\text{logit}(\nu_i) = X'_{i3}\beta_3$.

O modelo de regressão ZINB também é um membro da família GAMLSS.

2.2 GAMLSS - Regressão NB Inflada de Zeros

Diz-se que $Y \sim \text{ZINB}(\mu, \sigma, \nu)$ se há variável latente $G \sim \text{Ber}(\nu)$, com $\nu = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu, \sigma).$$

Note que, nesse caso, $\sigma = \frac{1}{\alpha}$ na nossa parametrização.

Portanto, a distribuiçao de Y é dada pela mistura

$$f(y|\mu, \sigma, \nu) = (1 - \nu)\mathbb{1}(y = 0) + \nu \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left(\frac{\sigma\mu}{1 + \sigma\mu}\right)^y \left(\frac{1}{1 + \sigma\mu}\right)^y.$$

A regressão ZINB é definida por $Y_i \sim \text{ZINB}(\mu_i, \sigma_i, \nu_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$, $\log(\sigma_i) = X'_{i2}\beta_2$ e $\text{logit}(\nu_i) = X'_{i3}\beta_3$.

O modelo de regressão ZINB também é um membro da família GAMLSS.

2.2 GAMLSS - Regressão NB Inflada de Zeros

Diz-se que $Y \sim \text{ZINB}(\mu, \sigma, \nu)$ se há variável latente $G \sim \text{Ber}(\nu)$, com $\nu = P(G = 1)$, tal que

$$Y|(G = 0) = 0 \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu, \sigma).$$

Note que, nesse caso, $\sigma = \frac{1}{\alpha}$ na nossa parametrização.

Portanto, a distribuiçao de Y é dada pela mistura

$$f(y|\mu, \sigma, \nu) = (1 - \nu)\mathbb{1}(y = 0) + \nu \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left(\frac{\sigma\mu}{1 + \sigma\mu}\right)^y \left(\frac{1}{1 + \sigma\mu}\right)^y.$$

A regressão ZINB é definida por $Y_i \sim \text{ZINB}(\mu_i, \sigma_i, \nu_i)$, com $\log(\mu_i) = X'_{i1}\beta_1$, $\log(\sigma_i) = X'_{i2}\beta_2$ e $\text{logit}(\nu_i) = X'_{i3}\beta_3$.

O modelo de regressão ZINB também é um membro da família GAMLSS.

2.2 GAMLSS - Comentários

Vantagens

- Os modelos de regressão ZIP e ZINB permitem a discriminação dos contratos de licitação em duas categorias: uma com zero aditivos; outra com uma distribuição discreta (Poisson ou NB).
- São facilmente implementados por meio do pacote `gamlss` do `R`.

Desvantagem

A categorização realizada por meio desses modelos restringe um grupo a ter obrigatoriamente zero aditivos, o que não é realista na prática. Não é impossível contratos eficientes terem aditivos.

Assim, uma alternativa mais realista seria considerar uma mistura de regressões NB (Mix-NB). Do que estudamos, não encontramos menção na literatura.

2.2 GAMLSS - Comentários

Vantagens

- Os modelos de regressão ZIP e ZINB permitem a discriminação dos contratos de licitação em duas categorias: uma com zero aditivos; outra com uma distribuição discreta (Poisson ou NB).
- São facilmente implementados por meio do pacote `gamlss` do `R`.

Desvantagem

A categorização realizada por meio desses modelos restringe um grupo a ter obrigatoriamente zero aditivos, o que não é realista na prática. Não é impossível contratos eficientes terem aditivos.

Assim, uma alternativa mais realista seria considerar uma mistura de regressões NB (Mix-NB). Do que estudamos, não encontramos menção na literatura.

2.2 GAMLSS - Comentários

Vantagens

- Os modelos de regressão ZIP e ZINB permitem a discriminação dos contratos de licitação em duas categorias: uma com zero aditivos; outra com uma distribuição discreta (Poisson ou NB).
- São facilmente implementados por meio do pacote `gamlss` do `R`.

Desvantagem

A categorização realizada por meio desses modelos restringe um grupo a ter obrigatoriamente zero aditivos, o que não é realista na prática. Não é impossível contratos eficientes terem aditivos.

Assim, uma alternativa mais realista seria considerar uma mistura de regressões NB (Mix-NB). Do que estudamos, não encontramos menção na literatura.

2.2 GAMLSS - Comentários

Vantagens

- Os modelos de regressão ZIP e ZINB permitem a discriminação dos contratos de licitação em duas categorias: uma com zero aditivos; outra com uma distribuição discreta (Poisson ou NB).
- São facilmente implementados por meio do pacote `gamlss` do `R`.

Desvantagem

A categorização realizada por meio desses modelos restringe um grupo a ter obrigatoriamente zero aditivos, o que não é realista na prática. Não é impossível contratos eficientes terem aditivos.

Assim, uma alternativa mais realista seria considerar uma mistura de regressões NB (Mix-NB). Do que estudamos, não encontramos menção na literatura.

2.2 GAMLSS - Comentários

Vantagens

- Os modelos de regressão ZIP e ZINB permitem a discriminação dos contratos de licitação em duas categorias: uma com zero aditivos; outra com uma distribuição discreta (Poisson ou NB).
- São facilmente implementados por meio do pacote `gamlss` do `R`.

Desvantagem

A categorização realizada por meio desses modelos restringe um grupo a ter obrigatoriamente zero aditivos, o que não é realista na prática. Não é impossível contratos eficientes terem aditivos.

Assim, uma alternativa mais realista seria considerar uma mistura de regressões NB (Mix-NB). Do que estudamos, não encontramos menção na literatura.

2.2 GAMLSS - Comentários

Vantagens

- Os modelos de regressão ZIP e ZINB permitem a discriminação dos contratos de licitação em duas categorias: uma com zero aditivos; outra com uma distribuição discreta (Poisson ou NB).
- São facilmente implementados por meio do pacote `gamlss` do `R`.

Desvantagem

A categorização realizada por meio desses modelos restringe um grupo a ter obrigatoriamente zero aditivos, o que não é realista na prática. Não é impossível contratos eficientes terem aditivos.

Assim, uma alternativa mais realista seria considerar uma mistura de regressões NB (Mix-NB). Do que estudamos, não encontramos menção na literatura.

Sumário

2.3 Modelo Mix-NB

Dizemos que $Y \sim$ Mix-NB$(\mu_0, \alpha_0, \mu_1, \alpha_1, p)$, se existe uma variável latente $G \sim$ Ber(p), com $p = P(G = 1)$, tal que

$$Y|(G = 0) \sim \text{NB}(\mu_0, \tfrac{1}{\alpha_0}) \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu_1, \tfrac{1}{\alpha_1}) \tag{3}$$

O modelo de regressão Mix-NB é definido assumindo que $Y_i \sim$ Mix-NB$(\mu_{i0}, \alpha_0, \mu_{i1}, \alpha_1, p_i)$, $i = 1, 2, \ldots, n$, em que $\log(\mu_{ig}) = X_i'\beta_g$, $g \in \{0, 1\}$, e $\text{logit}(p_i) = Z_i'\gamma$.

A Equação 3 define a **estrutura** de regressão NB em cada grupo $g \in \{0, 1\}$, enquanto que a variável latente G define a lei de **atribuição** a *priori* dos grupos.

Cada grupo g tem o seu próprio vetor de coeficientes β_g e parâmetro de dispersão α_g, enquanto γ é o vetor de coeficientes associado à regressão logística que atribui os grupos a cada observação.

As covariáveis das regressões NB em cada grupo X_i e para atribuição dos grupos Z_i não precisam ser as mesmas.

2.3 Modelo Mix-NB

Dizemos que $Y \sim$ Mix-NB$(\mu_0, \alpha_0, \mu_1, \alpha_1, p)$, se existe uma variável latente $G \sim$ Ber(p), com $p = P(G = 1)$, tal que

$$Y|(G = 0) \sim \text{NB}(\mu_0, \tfrac{1}{\alpha_0}) \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu_1, \tfrac{1}{\alpha_1}) \tag{3}$$

O modelo de regressão Mix-NB é definido assumindo que $Y_i \sim$ Mix-NB$(\mu_{i0}, \alpha_0, \mu_{i1}, \alpha_1, p_i)$, $i = 1, 2, \ldots, n$, em que $\log(\mu_{ig}) = X_i'\beta_g$, $g \in \{0, 1\}$, e $\text{logit}(p_i) = Z_i'\gamma$.

A Equação 3 define a **estrutura** de regressão NB em cada grupo $g \in \{0, 1\}$, enquanto que a variável latente G define a lei de **atribuição** a *priori* dos grupos.

Cada grupo g tem o seu próprio vetor de coeficientes β_g e parâmetro de dispersão α_g, enquanto γ é o vetor de coeficientes associado à regressão logística que atribui os grupos a cada observação.

As covariáveis das regressões NB em cada grupo X_i e para atribuição dos grupos Z_i não precisam ser as mesmas.

2.3 Modelo Mix-NB

Dizemos que $Y \sim$ Mix-NB$(\mu_0, \alpha_0, \mu_1, \alpha_1, p)$, se existe uma variável latente $G \sim$ Ber(p), com $p = P(G = 1)$, tal que

$$Y|(G = 0) \sim \text{NB}(\mu_0, \tfrac{1}{\alpha_0}) \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu_1, \tfrac{1}{\alpha_1}) \tag{3}$$

O modelo de regressão Mix-NB é definido assumindo que $Y_i \sim$ Mix-NB$(\mu_{i0}, \alpha_0, \mu_{i1}, \alpha_1, p_i)$, $i = 1, 2, \ldots, n$, em que $\log(\mu_{ig}) = X_i'\beta_g$, $g \in \{0, 1\}$, e $\text{logit}(p_i) = Z_i'\gamma$.

A Equação 3 define a **estrutura** de regressão NB em cada grupo $g \in \{0, 1\}$, enquanto que a variável latente G define a lei de **atribuição** a *priori* dos grupos.

Cada grupo g tem o seu próprio vetor de coeficientes β_g e parâmetro de dispersão α_g, enquanto γ é o vetor de coeficientes associado à regressão logística que atribui os grupos a cada observação.

As covariáveis das regressões NB em cada grupo X_i e para atribuição dos grupos Z_i não precisam ser as mesmas.

2.3 Modelo Mix-NB

Dizemos que $Y \sim$ Mix-NB$(\mu_0, \alpha_0, \mu_1, \alpha_1, p)$, se existe uma variável latente $G \sim$ Ber(p), com $p = P(G = 1)$, tal que

$$Y|(G = 0) \sim \text{NB}(\mu_0, \tfrac{1}{\alpha_0}) \qquad \text{e} \qquad Y|(G = 1) \sim \text{NB}(\mu_1, \tfrac{1}{\alpha_1}) \tag{3}$$

O modelo de regressão Mix-NB é definido assumindo que $Y_i \sim$ Mix-NB$(\mu_{i0}, \alpha_0, \mu_{i1}, \alpha_1, p_i)$, $i = 1, 2, \ldots, n$, em que $\log(\mu_{ig}) = X_i'\beta_g$, $g \in \{0, 1\}$, e $\text{logit}(p_i) = Z_i'\gamma$.

A Equação 3 define a **estrutura** de regressão NB em cada grupo $g \in \{0, 1\}$, enquanto que a variável latente G define a lei de **atribuição** a *priori* dos grupos.

Cada grupo g tem o seu próprio vetor de coeficientes β_g e parâmetro de dispersão α_g, enquanto γ é o vetor de coeficientes associado à regressão logística que atribui os grupos a cada observação.

As covariáveis das regressões NB em cada grupo X_i e para atribuição dos grupos Z_i não precisam ser as mesmas.

2.3 Modelo Mix-NB

A verossimilhança completa (se os grupos g_i fossem conhecidos) do modelo Mix-NB é dada por

$$L(\theta|y,g) = f(y,g|\theta) = \prod_{i=1}^{n} f(y_i,g_i|\theta) = \prod_{i=1}^{n}[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)p(g_i|\gamma,Z_i)], \tag{4}$$

em que $f(y_i|\beta_{g_i},\alpha_{g_i},X_i)$ é dada pela distribuição NB (Equação 1), com $\log(\mu_i) = X_i'\beta_{g_i}$ e $\omega = \frac{1}{\alpha_{g_i}}$, e $p(g_i|\gamma,Z_i) = P(G_i = g_i) = p_i^{g_i}(1-p_i)^{1-g_i}$, $\text{logit}(p_i) = Z_i'\gamma$. Assim, a log-verossimilhança completa é dada por:

$$l(\theta|y,g) = \log[L(\theta|y,g)] = \sum_{i=1}^{n}\{\log[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)] + \log[p(g_i|\gamma,Z_i)]\}.$$

Se os grupos $g_1,\ldots,g_n$ fossem conhecidos, poderíamos empregar métodos numéricos para obter a estimativa de máxima verossimilhança por meio de $l(\theta|y,g)$. Entretanto, o problema estudado impõe o desconhecimento desses grupos. O algoritmo EM é apropriado nesse caso.

2.3 Modelo Mix-NB

A verossimilhança completa (se os grupos g_i fossem conhecidos) do modelo Mix-NB é dada por

$$L(\theta|y,g) = f(y,g|\theta) = \prod_{i=1}^{n} f(y_i,g_i|\theta) = \prod_{i=1}^{n}[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)p(g_i|\gamma,Z_i)], \tag{4}$$

em que $f(y_i|\beta_{g_i},\alpha_{g_i},X_i)$ é dada pela distribuição NB (Equação 1), com $\log(\mu_i) = X_i'\beta_{g_i}$ e $\omega = \frac{1}{\alpha_{g_i}}$, e $p(g_i|\gamma,Z_i) = P(G_i = g_i) = p_i^{g_i}(1-p_i)^{1-g_i}$, $\text{logit}(p_i) = Z_i'\gamma$. Assim, a log-verossimilhança completa é dada por:

$$l(\theta|y,g) = \log[L(\theta|y,g)] = \sum_{i=1}^{n}\{\log[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)] + \log[p(g_i|\gamma,Z_i)]\}.$$

Se os grupos $g_1,\ldots,g_n$ fossem conhecidos, poderíamos empregar métodos numéricos para obter a estimativa de máxima verossimilhança por meio de $l(\theta|y,g)$. Entretanto, o problema estudado impõe o desconhecimento desses grupos. O algoritmo EM é apropriado nesse caso.

2.3 Modelo Mix-NB

A verossimilhança completa (se os grupos g_i fossem conhecidos) do modelo Mix-NB é dada por

$$L(\theta|y,g) = f(y,g|\theta) = \prod_{i=1}^{n} f(y_i,g_i|\theta) = \prod_{i=1}^{n}[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)p(g_i|\gamma,Z_i)], \tag{4}$$

em que $f(y_i|\beta_{g_i},\alpha_{g_i},X_i)$ é dada pela distribuição NB (Equação 1), com $\log(\mu_i) = X_i'\beta_{g_i}$ e $\omega = \frac{1}{\alpha_{g_i}}$, e $p(g_i|\gamma,Z_i) = P(G_i = g_i) = p_i^{g_i}(1-p_i)^{1-g_i}$, $\text{logit}(p_i) = Z_i'\gamma$. Assim, a log-verossimilhança completa é dada por:

$$l(\theta|y,g) = \log[L(\theta|y,g)] = \sum_{i=1}^{n}\{\log[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)] + \log[p(g_i|\gamma,Z_i)]\}.$$

Se os grupos $g_1,\ldots,g_n$ fossem conhecidos, poderíamos empregar métodos numéricos para obter a estimativa de máxima verossimilhança por meio de $l(\theta|y,g)$. Entretanto, o problema estudado impõe o desconhecimento desses grupos. O algoritmo EM é apropriado nesse caso.

2.3 Modelo Mix-NB

A verossimilhança completa (se os grupos g_i fossem conhecidos) do modelo Mix-NB é dada por

$$L(\theta|y,g) = f(y,g|\theta) = \prod_{i=1}^{n} f(y_i,g_i|\theta) = \prod_{i=1}^{n}[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)p(g_i|\gamma,Z_i)], \tag{4}$$

em que $f(y_i|\beta_{g_i},\alpha_{g_i},X_i)$ é dada pela distribuição NB (Equação 1), com $\log(\mu_i) = X_i'\beta_{g_i}$ e $\omega = \frac{1}{\alpha_{g_i}}$, e $p(g_i|\gamma,Z_i) = P(G_i = g_i) = p_i^{g_i}(1-p_i)^{1-g_i}$, $\text{logit}(p_i) = Z_i'\gamma$. Assim, a log-verossimilhança completa é dada por:

$$l(\theta|y,g) = \log[L(\theta|y,g)] = \sum_{i=1}^{n}\{\log[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)] + \log[p(g_i|\gamma,Z_i)]\}.$$

Se os grupos $g_1,\ldots,g_n$ fossem conhecidos, poderíamos empregar métodos numéricos para obter a estimativa de máxima verossimilhança por meio de $l(\theta|y,g)$. Entretanto, o problema estudado impõe o desconhecimento desses grupos. O algoritmo EM é apropriado nesse caso.

2.3 Modelo Mix-NB

A verossimilhança completa (se os grupos g_i fossem conhecidos) do modelo Mix-NB é dada por

$$L(\theta|y,g) = f(y,g|\theta) = \prod_{i=1}^{n} f(y_i,g_i|\theta) = \prod_{i=1}^{n}[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)p(g_i|\gamma,Z_i)], \tag{4}$$

em que $f(y_i|\beta_{g_i},\alpha_{g_i},X_i)$ é dada pela distribuição NB (Equação 1), com $\log(\mu_i) = X_i'\beta_{g_i}$ e $\omega = \frac{1}{\alpha_{g_i}}$, e $p(g_i|\gamma,Z_i) = P(G_i = g_i) = p_i^{g_i}(1-p_i)^{1-g_i}$, $\text{logit}(p_i) = Z_i'\gamma$. Assim, a log-verossimilhança completa é dada por:

$$l(\theta|y,g) = \log[L(\theta|y,g)] = \sum_{i=1}^{n}\{\log[f(y_i|\beta_{g_i},\alpha_{g_i},X_i)] + \log[p(g_i|\gamma,Z_i)]\}.$$

Se os grupos $g_1,\ldots,g_n$ fossem conhecidos, poderíamos empregar métodos numéricos para obter a estimativa de máxima verossimilhança por meio de $l(\theta|y,g)$. Entretanto, o problema estudado impõe o desconhecimento desses grupos. O algoritmo EM é apropriado nesse caso.

2.3.1 Algoritmo EM

O algoritmo *Expectation-Maximization* (EM) é um método de estimação particularmente útil em modelos com variáveis latentes. Informalmente, a maximização da verossimilhança é substituída por maximizações de verossimilhanças "esperadas".

Intuitivamente, o algoritmo EM alterna iterativamente entre o cálculo da log-verossimilhança completa esperada (Etapa E) e a maximização dessa log-verossimilhança "esperada" (Etapa M). Formalmente, o algoritmo EM pode ser ser descrito da seguinte forma:

1. Inicie com uma estimativa $\hat{\theta}^{(0)}$;
2. Na iteração $t \geq 0$, compute a log-verossimilhança completa esperada $Q(\theta|\hat{\theta}^{(t)}) = E[l(\theta|y, G)]$ com respeito a distribuição *a posteriori* da variável latente $G|(\theta = \hat{\theta}^{(t)}, y_i)$ (Etapa E);
3. Obtenha $\hat{\theta}^{(t+1)} = \text{argmax}\{Q(\theta|\hat{\theta}^{(t)})\}$ (Etapa M);
4. Faça $t = t + 1$ e retorne ao Passo 2 caso não alcance a convergência.

2.3.1 Algoritmo EM

O algoritmo *Expectation-Maximization* (EM) é um método de estimação particularmente útil em modelos com variáveis latentes. Informalmente, a maximização da verossimilhança é substituída por maximizações de verossimilhanças "esperadas".

Intuitivamente, o algoritmo EM alterna iterativamente entre o cálculo da log-verossimilhança completa esperada (Etapa E) e a maximização dessa log-verossimilhança "esperada" (Etapa M). Formalmente, o algoritmo EM pode ser ser descrito da seguinte forma:

1. Inicie com uma estimativa $\hat{\theta}^{(0)}$;
2. Na iteração $t \geq 0$, compute a log-verossimilhança completa esperada $Q(\theta|\hat{\theta}^{(t)}) = E[l(\theta|y, G)]$ com respeito a distribuição *a posteriori* da variável latente $G|(\theta = \hat{\theta}^{(t)}, y_i)$ (Etapa E);
3. Obtenha $\hat{\theta}^{(t+1)} = \text{argmax}\{Q(\theta|\hat{\theta}^{(t)})\}$ (Etapa M);
4. Faça $t = t + 1$ e retorne ao Passo 2 caso não alcance a convergência.

2.3.1 Algoritmo EM

O algoritmo *Expectation-Maximization* (EM) é um método de estimação particularmente útil em modelos com variáveis latentes. Informalmente, a maximização da verossimilhança é substituída por maximizações de verossimilhanças "esperadas".

Intuitivamente, o algoritmo EM alterna iterativamente entre o cálculo da log-verossimilhança completa esperada (Etapa E) e a maximização dessa log-verossimilhança "esperada" (Etapa M). Formalmente, o algoritmo EM pode ser ser descrito da seguinte forma:

1. Inicie com uma estimativa $\hat{\theta}^{(0)}$;
2. Na iteração $t \geq 0$, compute a log-verossimilhança completa esperada $Q(\theta|\hat{\theta}^{(t)}) = E[l(\theta|y, G)]$ com respeito a distribuição *a posteriori* da variável latente $G|(\theta = \hat{\theta}^{(t)}, y_i)$ (Etapa E);
3. Obtenha $\hat{\theta}^{(t+1)} = \text{argmax}\{Q(\theta|\hat{\theta}^{(t)})\}$ (Etapa M);
4. Faça $t = t + 1$ e retorne ao Passo 2 caso não alcance a convergência.

2.3.1 Algoritmo EM

O algoritmo *Expectation-Maximization* (EM) é um método de estimação particularmente útil em modelos com variáveis latentes. Informalmente, a maximização da verossimilhança é substituída por maximizações de verossimilhanças "esperadas".

Intuitivamente, o algoritmo EM alterna iterativamente entre o cálculo da log-verossimilhança completa esperada (Etapa E) e a maximização dessa log-verossimilhança "esperada" (Etapa M). Formalmente, o algoritmo EM pode ser ser descrito da seguinte forma:

1. Inicie com uma estimativa $\hat{\theta}^{(0)}$;
2. Na iteração $t \geq 0$, compute a log-verossimilhança completa esperada $Q(\theta|\hat{\theta}^{(t)}) = E[l(\theta|y, G)]$ com respeito a distribuição *a posteriori* da variável latente $G|(\theta = \hat{\theta}^{(t)}, y_i)$ (Etapa E);
3. Obtenha $\hat{\theta}^{(t+1)} = \text{argmax}\{Q(\theta|\hat{\theta}^{(t)})\}$ (Etapa M);
4. Faça $t = t + 1$ e retorne ao Passo 2 caso não alcance a convergência.

2.3.1 Algoritmo EM

O algoritmo *Expectation-Maximization* (EM) é um método de estimação particularmente útil em modelos com variáveis latentes. Informalmente, a maximização da verossimilhança é substituída por maximizações de verossimilhanças "esperadas".

Intuitivamente, o algoritmo EM alterna iterativamente entre o cálculo da log-verossimilhança completa esperada (Etapa E) e a maximização dessa log-verossimilhança "esperada" (Etapa M). Formalmente, o algoritmo EM pode ser ser descrito da seguinte forma:

1. Inicie com uma estimativa $\hat{\theta}^{(0)}$;
2. Na iteração $t \geq 0$, compute a log-verossimilhança completa esperada $Q(\theta|\hat{\theta}^{(t)}) = E[l(\theta|y, G)]$ com respeito a distribuição *a posteriori* da variável latente $G|(\theta = \hat{\theta}^{(t)}, y_i)$ (Etapa E);
3. Obtenha $\hat{\theta}^{(t+1)} = \text{argmax}\{Q(\theta|\hat{\theta}^{(t)})\}$ (Etapa M);
4. Faça $t = t + 1$ e retorne ao Passo 2 caso não alcance a convergência.

2.3.1 Algoritmo EM

O algoritmo *Expectation-Maximization* (EM) é um método de estimação particularmente útil em modelos com variáveis latentes. Informalmente, a maximização da verossimilhança é substituída por maximizações de verossimilhanças "esperadas".

Intuitivamente, o algoritmo EM alterna iterativamente entre o cálculo da log-verossimilhança completa esperada (Etapa E) e a maximização dessa log-verossimilhança "esperada" (Etapa M). Formalmente, o algoritmo EM pode ser ser descrito da seguinte forma:

1. Inicie com uma estimativa $\hat{\theta}^{(0)}$;
2. Na iteração $t \geq 0$, compute a log-verossimilhança completa esperada $Q(\theta|\hat{\theta}^{(t)}) = E[l(\theta|y, G)]$ com respeito a distribuição *a posteriori* da variável latente $G|(\theta = \hat{\theta}^{(t)}, y_i)$ (Etapa E);
3. Obtenha $\hat{\theta}^{(t+1)} = \text{argmax}\{Q(\theta|\hat{\theta}^{(t)})\}$ (Etapa M);
4. Faça $t = t + 1$ e retorne ao Passo 2 caso não alcance a convergência.

2.3.1 Algoritmo EM

O algoritmo *Expectation-Maximization* (EM) é um método de estimação particularmente útil em modelos com variáveis latentes. Informalmente, a maximização da verossimilhança é substituída por maximizações de verossimilhanças "esperadas".

Intuitivamente, o algoritmo EM alterna iterativamente entre o cálculo da log-verossimilhança completa esperada (Etapa E) e a maximização dessa log-verossimilhança "esperada" (Etapa M). Formalmente, o algoritmo EM pode ser ser descrito da seguinte forma:

1. Inicie com uma estimativa $\hat{\theta}^{(0)}$;
2. Na iteração $t \geq 0$, compute a log-verossimilhança completa esperada $Q(\theta|\hat{\theta}^{(t)}) = E[l(\theta|y, G)]$ com respeito a distribuição *a posteriori* da variável latente $G|(\theta = \hat{\theta}^{(t)}, y_i)$ (Etapa E);
3. Obtenha $\hat{\theta}^{(t+1)} = \text{argmax}\{Q(\theta|\hat{\theta}^{(t)})\}$ (Etapa M);
4. Faça $t = t + 1$ e retorne ao Passo 2 caso não alcance a convergência.

2.3.1 Algoritmo EM - Modelo Mix-NB

Para o modelo Mix-NB, $G|(\theta, y)$ tem distribuição $p(g|\theta, y) = \frac{f(y,g|\theta)}{f(y|\theta)}$, em que a marginal

$$f(y|\theta) = \sum_{g_1=0}^{1} \cdots \sum_{g_n=0}^{1} [f(y_1, g_1|\theta) \cdots f(y_n, g_n|\theta)] = \prod_{i=1}^{n} \left[\sum_{g_i=0}^{1} f(y_i, g_i|\theta) \right] = \prod_{i=1}^{n} f(y_i|\theta),$$

cujo último termo satisfaz $f(y_i|\theta) = f(y_i, g_i = 0|\theta) + f(y_i, g_i = 1|\theta)$ (veja Equação 4). Definindo $q_{it} = p(1|\theta^{(t)}, y_i)$, podemos mostrar que a log-verossimilhança esperada é

$$\begin{aligned} Q(\theta|\theta^{(t)}) = E(l(\theta|y, G)) &= \sum_{i=1}^{n} \{\log(y_i|\beta_0, \alpha_0, X_i)(1 - q_{it})\} \\ &+ \sum_{i=1}^{n} \{\log(y_i|\beta_1, \alpha_1, X_i) q_{it}\} \\ &+ \sum_{i=1}^{n} \{\log[p(0|\gamma, Z_i)](1 - q_{it}) + \log[p(1|\gamma, Z_i)] q_{it}\} \\ &= Q(\beta_0, \alpha_0|\theta^{(t)}) + Q(\beta_1, \alpha_1|\theta^{(t)}) + Q(\gamma|\theta^{(t)}). \end{aligned} \tag{5}$$

2.3.1 Algoritmo EM - Modelo Mix-NB

Para o modelo Mix-NB, $G|(\theta, y)$ tem distribuição $p(g|\theta, y) = \frac{f(y,g|\theta)}{f(y|\theta)}$, em que a marginal

$$f(y|\theta) = \sum_{g_1=0}^{1} \cdots \sum_{g_n=0}^{1} [f(y_1, g_1|\theta) \cdots f(y_n, g_n|\theta)] = \prod_{i=1}^{n} \left[\sum_{g_i=0}^{1} f(y_i, g_i|\theta) \right] = \prod_{i=1}^{n} f(y_i|\theta),$$

cujo último termo satisfaz $f(y_i|\theta) = f(y_i, g_i = 0|\theta) + f(y_i, g_i = 1|\theta)$ (veja Equação 4). Definindo $q_{it} = p(1|\theta^{(t)}, y_i)$, podemos mostrar que a log-verossimilhança esperada é

$$\begin{aligned} Q(\theta|\theta^{(t)}) = E(l(\theta|y, G)) &= \sum_{i=1}^{n} \{\log(y_i|\beta_0, \alpha_0, X_i)(1 - q_{it})\} \\ &+ \sum_{i=1}^{n} \{\log(y_i|\beta_1, \alpha_1, X_i) q_{it}\} \\ &+ \sum_{i=1}^{n} \{\log[p(0|\gamma, Z_i)](1 - q_{it}) + \log[p(1|\gamma, Z_i)] q_{it}\} \\ &= Q(\beta_0, \alpha_0|\theta^{(t)}) + Q(\beta_1, \alpha_1|\theta^{(t)}) + Q(\gamma|\theta^{(t)}). \end{aligned} \tag{5}$$

2.3.1 Algoritmo EM - Modelo Mix-NB

Para o modelo Mix-NB, $G|(\theta, y)$ tem distribuição $p(g|\theta, y) = \frac{f(y,g|\theta)}{f(y|\theta)}$, em que a marginal

$$f(y|\theta) = \sum_{g_1=0}^{1} \cdots \sum_{g_n=0}^{1} [f(y_1, g_1|\theta) \cdots f(y_n, g_n|\theta)] = \prod_{i=1}^{n} \left[\sum_{g_i=0}^{1} f(y_i, g_i|\theta) \right] = \prod_{i=1}^{n} f(y_i|\theta),$$

cujo último termo satisfaz $f(y_i|\theta) = f(y_i, g_i = 0|\theta) + f(y_i, g_i = 1|\theta)$ (veja Equação 4). Definindo $q_{it} = p(1|\theta^{(t)}, y_i)$, podemos mostrar que a log-verossimilhança esperada é

$$\begin{aligned} Q(\theta|\theta^{(t)}) = E(l(\theta|y, G)) &= \textstyle\sum_{i=1}^{n} \{\log(y_i|\beta_0, \alpha_0, X_i)(1 - q_{it})\} \\ &+ \textstyle\sum_{i=1}^{n} \{\log(y_i|\beta_1, \alpha_1, X_i) q_{it}\} \\ &+ \textstyle\sum_{i=1}^{n} \{\log[p(0|\gamma, Z_i)](1 - q_{it}) + \log[p(1|\gamma, Z_i)] q_{it}\} \\ &= Q(\beta_0, \alpha_0|\theta^{(t)}) + Q(\beta_1, \alpha_1|\theta^{(t)}) + Q(\gamma|\theta^{(t)}). \end{aligned} \tag{5}$$

2.3.1 Algoritmo EM - Modelo Mix-NB

Para o modelo Mix-NB, $G|(\theta, y)$ tem distribuição $p(g|\theta, y) = \frac{f(y,g|\theta)}{f(y|\theta)}$, em que a marginal

$$f(y|\theta) = \sum_{g_1=0}^{1} \cdots \sum_{g_n=0}^{1} [f(y_1, g_1|\theta) \cdots f(y_n, g_n|\theta)] = \prod_{i=1}^{n} \left[\sum_{g_i=0}^{1} f(y_i, g_i|\theta) \right] = \prod_{i=1}^{n} f(y_i|\theta),$$

cujo último termo satisfaz $f(y_i|\theta) = f(y_i, g_i = 0|\theta) + f(y_i, g_i = 1|\theta)$ (veja Equação 4). Definindo $q_{it} = p(1|\theta^{(t)}, y_i)$, podemos mostrar que a log-verossimilhança esperada é

$$\begin{aligned} Q(\theta|\theta^{(t)}) = E(l(\theta|y, G)) &= \textstyle\sum_{i=1}^{n} \{\log(y_i|\beta_0, \alpha_0, X_i)(1 - q_{it})\} \\ &+ \textstyle\sum_{i=1}^{n} \{\log(y_i|\beta_1, \alpha_1, X_i) q_{it}\} \\ &+ \textstyle\sum_{i=1}^{n} \{\log[p(0|\gamma, Z_i)](1 - q_{it}) + \log[p(1|\gamma, Z_i)] q_{it}\} \\ &= Q(\beta_0, \alpha_0|\theta^{(t)}) + Q(\beta_1, \alpha_1|\theta^{(t)}) + Q(\gamma|\theta^{(t)}). \end{aligned} \tag{5}$$

2.3.1 Algoritmo EM - Modelo Mix-NB)

A Etapa M de maximização de *dedde* pode ser empregada maximizando as funções $Q(\beta_0, \alpha_0|\theta^{(t)})$, $Q(\beta_1, \alpha_1|\theta^{(t)})$ e $Q(\gamma|\theta^{(t)})$ na Equação 5 de maneira independente.

As maximizações de $Q(\beta_0, \alpha_0|\theta^{(t)})$ e $Q(\beta_1, \alpha_1|\theta^{(t)})$ correspondem ao ajuste de modelos de regressão NB ponderados com pesos $1 - q_{it}$ e q_{it}, respectivamente, que podem ser implementatos no `R` por meio da função `glm.nb` (pacote `MASS`).

Já a função $Q(\gamma|\theta^{(t)})$ se assemelha à verossimilhança de um modelo de regressão logística, em que as "observações" $q_{it} \in (0, 1)$ e não $\{0, 1\}$. Assim, a maximização dessa função foi implementada usando a função `optim` do `R`

2.3.1 Algoritmo EM - Modelo Mix-NB)

A Etapa M de maximização de *dedde* pode ser empregada maximizando as funções $Q(\beta_0, \alpha_0|\theta^{(t)})$, $Q(\beta_1, \alpha_1|\theta^{(t)})$ e $Q(\gamma|\theta^{(t)})$ na Equação 5 de maneira independente.

As maximizações de $Q(\beta_0, \alpha_0|\theta^{(t)})$ e $Q(\beta_1, \alpha_1|\theta^{(t)})$ correspondem ao ajuste de modelos de regressão NB ponderados com pesos $1 - q_{it}$ e q_{it}, respectivamente, que podem ser implementatos no `R` por meio da função `glm.nb` (pacote `MASS`).

Já a função $Q(\gamma|\theta^{(t)})$ se assemelha à verossimilhança de um modelo de regressão logística, em que as "observações" $q_{it} \in (0, 1)$ e não $\{0, 1\}$. Assim, a maximização dessa função foi implementada usando a função `optim` do `R`

2.3.1 Algoritmo EM - Modelo Mix-NB)

A Etapa M de maximização de *dedde* pode ser empregada maximizando as funções $Q(\beta_0, \alpha_0|\theta^{(t)})$, $Q(\beta_1, \alpha_1|\theta^{(t)})$ e $Q(\gamma|\theta^{(t)})$ na Equação 5 de maneira independente.

As maximizações de $Q(\beta_0, \alpha_0|\theta^{(t)})$ e $Q(\beta_1, \alpha_1|\theta^{(t)})$ correspondem ao ajuste de modelos de regressão NB ponderados com pesos $1 - q_{it}$ e q_{it}, respectivamente, que podem ser implementatos no `R` por meio da função `glm.nb` (pacote `MASS`).

Já a função $Q(\gamma|\theta^{(t)})$ se assemelha à verossimilhança de um modelo de regressão logística, em que as "observações" $q_{it} \in (0, 1)$ e não $\{0, 1\}$. Assim, a maximização dessa função foi implementada usando a função `optim` do `R`

2.3.2 Diagnóstico do Ajuste

Os resíduos de Pearson e *Deviance* têm sido criticado nas aplicações GLM para dados de contagem. Uma alternativa são os Resíduos Quantílicos Randomizados (RQR) (Dunn e Smyth, 1996).

Escrevendo $W_i = (X_i', Z_i')'$, os RQR's são definidos por

$$r_i = \Phi^{-1}(\mathcal{F}(y_i|\hat{\theta}, W_i)), \tag{6}$$

em que $\Phi^{-1}(\cdot)$ denota a função quantílica de uma normal padrão e

$$\mathcal{F}(y_i|\hat{\theta}, W_i) = F(y_i^-|\hat{\theta}, W_i) + u_i[F(y_i|\hat{\theta}, W_i) - F(y_i^-|\hat{\theta}, W_i)]$$

é a distribuição acumulada randomizada, e $F(y_i^-|\hat{\theta}, W_i) = \lim_{y \uparrow y_i} F(y|\hat{\theta}, W_i)$ e $u_i \sim \mathrm{U}(0,1)$, $\forall i$.

Vantagem

Se considerarmos $\hat{\theta} = \theta$ fixo e o modelo estiver corretamente especificado, então os r_i são normalmente distribuídos. Logo, se $\hat{\theta}$ é consistente, então a distribuição de r_i é assintoticamente normal.

2.3.2 Diagnóstico do Ajuste

Os resíduos de Pearson e *Deviance* têm sido criticado nas aplicações GLM para dados de contagem. Uma alternativa são os Resíduos Quantílicos Randomizados (RQR) (Dunn e Smyth, 1996).

Escrevendo $W_i = (X_i', Z_i')'$, os RQR's são definidos por

$$r_i = \Phi^{-1}(\mathcal{F}(y_i|\hat{\theta}, W_i)), \tag{6}$$

em que $\Phi^{-1}(\cdot)$ denota a função quantílica de uma normal padrão e

$$\mathcal{F}(y_i|\hat{\theta}, W_i) = F(y_i^-|\hat{\theta}, W_i) + u_i[F(y_i|\hat{\theta}, W_i) - F(y_i^-|\hat{\theta}, W_i)]$$

é a distribuição acumulada randomizada, e $F(y_i^-|\hat{\theta}, W_i) = \lim_{y \uparrow y_i} F(y|\hat{\theta}, W_i)$ e $u_i \sim \mathrm{U}(0,1)$, $\forall i$.

Vantagem

Se considerarmos $\hat{\theta} = \theta$ fixo e o modelo estiver corretamente especificado, então os r_i são normalmente distribuídos. Logo, se $\hat{\theta}$ é consistente, então a distribuição de r_i é assintoticamente normal.

2.3.2 Diagnóstico do Ajuste

Os resíduos de Pearson e *Deviance* têm sido criticado nas aplicações GLM para dados de contagem. Uma alternativa são os Resíduos Quantílicos Randomizados (RQR) (Dunn e Smyth, 1996).

Escrevendo $W_i = (X_i', Z_i')'$, os RQR's são definidos por

$$r_i = \Phi^{-1}(\mathcal{F}(y_i|\hat{\theta}, W_i)), \tag{6}$$

em que $\Phi^{-1}(\cdot)$ denota a função quantílica de uma normal padrão e

$$\mathcal{F}(y_i|\hat{\theta}, W_i) = F(y_i^-|\hat{\theta}, W_i) + u_i[F(y_i|\hat{\theta}, W_i) - F(y_i^-|\hat{\theta}, W_i)]$$

é a distribuição acumulada randomizada, e $F(y_i^-|\hat{\theta}, W_i) = \lim_{y \uparrow y_i} F(y|\hat{\theta}, W_i)$ e $u_i \sim \mathrm{U}(0,1)$, $\forall i$.

Vantagem

Se considerarmos $\hat{\theta} = \theta$ fixo e o modelo estiver corretamente especificado, então os r_i são normalmente distribuídos. Logo, se $\hat{\theta}$ é consistente, então a distribuição de r_i é assintoticamente normal.

2.3.2 Diagnóstico do Ajuste

Os resíduos de Pearson e *Deviance* têm sido criticado nas aplicações GLM para dados de contagem. Uma alternativa são os Resíduos Quantílicos Randomizados (RQR) (Dunn e Smyth, 1996).

Escrevendo $W_i = (X_i', Z_i')'$, os RQR's são definidos por

$$r_i = \Phi^{-1}(\mathcal{F}(y_i|\hat{\theta}, W_i)), \tag{6}$$

em que $\Phi^{-1}(\cdot)$ denota a função quantílica de uma normal padrão e

$$\mathcal{F}(y_i|\hat{\theta}, W_i) = F(y_i^-|\hat{\theta}, W_i) + u_i[F(y_i|\hat{\theta}, W_i) - F(y_i^-|\hat{\theta}, W_i)]$$

é a distribuição acumulada randomizada, e $F(y_i^-|\hat{\theta}, W_i) = \lim_{y \uparrow y_i} F(y|\hat{\theta}, W_i)$ e $u_i \sim \mathrm{U}(0,1)$, $\forall i$.

Vantagem

Se considerarmos $\hat{\theta} = \theta$ fixo e o modelo estiver corretamente especificado, então os r_i são normalmente distribuídos. Logo, se $\hat{\theta}$ é consistente, então a distribuição de r_i é assintoticamente normal.

2.3.2 Diagnóstico do Ajuste

Neste trabalho, a normalidade dos RQR's é investigada pelo teste Kolmogorov-Smirnov (KS).

A estatística de teste KS é dada por:

$$\text{KS} = \sup_r |F_n(r) - \Phi(r)|, \tag{7}$$

em que $F_n(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(r_i \leq r)$ é a distribuição acumulada empírica dos RQR.

Embora desejemos testar $H_0 : r_i \sim N(0,1)$ (distribuição com parâmetros pré-especificados), os RQR's são produzidos por meio de estimativas dos parâmetros do modelo, o que pode tornar a distribuição da estatística KS imprecisa.

Para contornar esse problema, este trabalho se inspira em Sarnaglia et al. (2018) e utiliza bootstrap (Efron, 1992) para reamostrar a estatística KS sob H_0.

2.3.2 Diagnóstico do Ajuste

Neste trabalho, a normalidade dos RQR's é investigada pelo teste Kolmogorov-Smirnov (KS).

A estatística de teste KS é dada por:

$$\text{KS} = \sup_r |F_n(r) - \Phi(r)|, \tag{7}$$

em que $F_n(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(r_i \leq r)$ é a distribuição acumulada empírica dos RQR.

Embora desejemos testar $H_0 : r_i \sim N(0,1)$ (distribuição com parâmetros pré-especificados), os RQR's são produzidos por meio de estimativas dos parâmetros do modelo, o que pode tornar a distribuição da estatística KS imprecisa.

Para contornar esse problema, este trabalho se inspira em Sarnaglia et al. (2018) e utiliza bootstrap (Efron, 1992) para reamostrar a estatística KS sob H_0.

2.3.2 Diagnóstico do Ajuste

Neste trabalho, a normalidade dos RQR's é investigada pelo teste Kolmogorov-Smirnov (KS).

A estatística de teste KS é dada por:

$$\text{KS} = \sup_r |F_n(r) - \Phi(r)|, \tag{7}$$

em que $F_n(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(r_i \leq r)$ é a distribuição acumulada empírica dos RQR.

Embora desejemos testar $H_0 : r_i \sim N(0,1)$ (distribuição com parâmetros pré-especificados), os RQR's são produzidos por meio de estimativas dos parâmetros do modelo, o que pode tornar a distribuição da estatística KS imprecisa.

Para contornar esse problema, este trabalho se inspira em Sarnaglia et al. (2018) e utiliza bootstrap (Efron, 1992) para reamostrar a estatística KS sob H_0.

2.3.2 Diagnóstico do Ajuste

Neste trabalho, a normalidade dos RQR's é investigada pelo teste Kolmogorov-Smirnov (KS).

A estatística de teste KS é dada por:

$$\text{KS} = \sup_r |F_n(r) - \Phi(r)|, \tag{7}$$

em que $F_n(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(r_i \leq r)$ é a distribuição acumulada empírica dos RQR.

Embora desejemos testar $H_0 : r_i \sim N(0,1)$ (distribuição com parâmetros pré-especificados), os RQR's são produzidos por meio de estimativas dos parâmetros do modelo, o que pode tornar a distribuição da estatística KS imprecisa.

Para contornar esse problema, este trabalho se inspira em Sarnaglia et al. (2018) e utiliza bootstrap (Efron, 1992) para reamostrar a estatística KS sob H_0.

2.3.2 Diagnóstico do Ajuste

Partindo de $b = 1$, o procedimento para obter B réplicas bootstrap da estatística KS sob H_0 é:

1. Extraia uma reamostra bootstrap paramétrica iid $y_b^* = (y_{1b}^*, \ldots, y_{nb}^*)'$ de tamanho n do modelo Mix-NB (Equação 3) com $\theta = \hat{\theta}$, e com as covariáveis originais X_i e Z_i;
2. Ajuste o modelo Mix-NB à y_b^* com as covariáveis originais X_i e Z_i para produzir a b-ésima estimativa bootstrap $\hat{\theta}_b^*$ e os RQR's bootstrap $r_b^* = (r_{1b}^*, \ldots, r_{nb}^*)'$ conforme a Equação 6;
3. Usando a Equação 7, compute a estatística KS para os RQR's bootstrap r_b^*, isto é, $\mathrm{KS}_b^* = \sup_r |F_n^*(r) - \Phi(r)|$, em que $F_n^*(r) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(r_{ib}^* \leq r)$ é a acumulada empírica dos RQR's bootstrap;
4. faça $b = b + 1$ e retorne ao Passo 2 enquanto $b \leq B$.

O p-valor bootstrap do teste KS é

$$p^* = \frac{1 + \sum_{b=1}^{B} \mathbb{1}(\mathrm{KS}_b^* > \mathrm{KS}_{\mathrm{orig}})}{B + 1}.$$

2.3.2 Diagnóstico do Ajuste

Partindo de $b = 1$, o procedimento para obter B réplicas bootstrap da estatística KS sob H_0 é:

1. Extraia uma reamostra bootstrap paramétrica iid $y_b^* = (y_{1b}^*, \ldots, y_{nb}^*)'$ de tamanho n do modelo Mix-NB (Equação 3) com $\theta = \hat{\theta}$, e com as covariáveis originais X_i e Z_i;
2. Ajuste o modelo Mix-NB à y_b^* com as covariáveis originais X_i e Z_i para produzir a b-ésima estimativa bootstrap $\hat{\theta}_b^*$ e os RQR's bootstrap $r_b^* = (r_{1b}^*, \ldots, r_{nb}^*)'$ conforme a Equação 6;
3. Usando a Equação 7, compute a estatística KS para os RQR's bootstrap r_b^*, isto é, $\mathrm{KS}_b^* = \sup_r |F_n^*(r) - \Phi(r)|$, em que $F_n^*(r) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(r_{ib}^* \leq r)$ é a acumulada empírica dos RQR's bootstrap;
4. faça $b = b + 1$ e retorne ao Passo 2 enquanto $b \leq B$.

O p-valor bootstrap do teste KS é

$$p^* = \frac{1 + \sum_{b=1}^{B} \mathbb{1}(\mathrm{KS}_b^* > \mathrm{KS}_{\mathrm{orig}})}{B + 1}.$$

2.3.2 Diagnóstico do Ajuste

Partindo de $b = 1$, o procedimento para obter B réplicas bootstrap da estatística KS sob H_0 é:

1. Extraia uma reamostra bootstrap paramétrica iid $y_b^* = (y_{1b}^*, \ldots, y_{nb}^*)'$ de tamanho n do modelo Mix-NB (Equação 3) com $\theta = \hat{\theta}$, e com as covariáveis originais X_i e Z_i;
2. Ajuste o modelo Mix-NB à y_b^* com as covariáveis originais X_i e Z_i para produzir a b-ésima estimativa bootstrap $\hat{\theta}_b^*$ e os RQR's bootstrap $r_b^* = (r_{1b}^*, \ldots, r_{nb}^*)'$ conforme a Equação 6;
3. Usando a Equação 7, compute a estatística KS para os RQR's bootstrap r_b^*, isto é, $\mathrm{KS}_b^* = \sup_r |F_n^*(r) - \Phi(r)|$, em que $F_n^*(r) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(r_{ib}^* \leq r)$ é a acumulada empírica dos RQR's bootstrap;
4. faça $b = b + 1$ e retorne ao Passo 2 enquanto $b \leq B$.

O p-valor bootstrap do teste KS é

$$p^* = \frac{1 + \sum_{b=1}^{B} \mathbb{1}(\mathrm{KS}_b^* > \mathrm{KS}_{\mathrm{orig}})}{B + 1}.$$

2.3.2 Diagnóstico do Ajuste

Partindo de $b = 1$, o procedimento para obter B réplicas bootstrap da estatística KS sob H_0 é:

1. Extraia uma reamostra bootstrap paramétrica iid $y_b^* = (y_{1b}^*, \ldots, y_{nb}^*)'$ de tamanho n do modelo Mix-NB (Equação 3) com $\theta = \hat{\theta}$, e com as covariáveis originais X_i e Z_i;
2. Ajuste o modelo Mix-NB à y_b^* com as covariáveis originais X_i e Z_i para produzir a b-ésima estimativa bootstrap $\hat{\theta}_b^*$ e os RQR's bootstrap $r_b^* = (r_{1b}^*, \ldots, r_{nb}^*)'$ conforme a Equação 6;
3. Usando a Equação 7, compute a estatística KS para os RQR's bootstrap r_b^*, isto é, $\mathrm{KS}_b^* = \sup_r |F_n^*(r) - \Phi(r)|$, em que $F_n^*(r) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(r_{ib}^* \leq r)$ é a acumulada empírica dos RQR's bootstrap;
4. faça $b = b + 1$ e retorne ao Passo 2 enquanto $b \leq B$.

O p-valor bootstrap do teste KS é

$$p^* = \frac{1 + \sum_{b=1}^{B} \mathbb{1}(\mathrm{KS}_b^* > \mathrm{KS}_{\mathrm{orig}})}{B + 1}.$$

2.3.2 Diagnóstico do Ajuste

Partindo de $b = 1$, o procedimento para obter B réplicas bootstrap da estatística KS sob H_0 é:

1. Extraia uma reamostra bootstrap paramétrica iid $y_b^* = (y_{1b}^*, \ldots, y_{nb}^*)'$ de tamanho n do modelo Mix-NB (Equação 3) com $\theta = \hat{\theta}$, e com as covariáveis originais X_i e Z_i;
2. Ajuste o modelo Mix-NB à y_b^* com as covariáveis originais X_i e Z_i para produzir a b-ésima estimativa bootstrap $\hat{\theta}_b^*$ e os RQR's bootstrap $r_b^* = (r_{1b}^*, \ldots, r_{nb}^*)'$ conforme a Equação 6;
3. Usando a Equação 7, compute a estatística KS para os RQR's bootstrap r_b^*, isto é, $\text{KS}_b^* = \sup_r |F_n^*(r) - \Phi(r)|$, em que $F_n^*(r) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(r_{ib}^* \leq r)$ é a acumulada empírica dos RQR's bootstrap;
4. faça $b = b + 1$ e retorne ao Passo 2 enquanto $b \leq B$.

O p-valor bootstrap do teste KS é

$$p^* = \frac{1 + \sum_{b=1}^B \mathbb{1}(\text{KS}_b^* > \text{KS}_{\text{orig}})}{B + 1}.$$

Sumário

Sumário

3.1 Coleta e Tratamento da Dase de Dados

Dados dos contratos coletados do portal do SIASG e de Unidades Federativas (UF) do portal BuscaCEP por meio de robôs no R. Foram obtidos dados complementares de variáveis sócio-econômicas em outras fontes, por exemplo: **IPEA IBGE (Sidra)**, entre outras.

Período: de janeiro de 2011 a dezembro de 2019. Totalizando dados de 256942 contratos na base de dados bruta.

Tratamento preliminar com exclusão: de variáveis sem importância para o estudo; de registros repetidos; de contratos com vigência maior que 52 semanas (> 1 ano); e de casos com valores faltantes para as demais variáveis ou com campos permutados.

Neste trabalho, restrigimos a análise a contratos da modalidade "Pregão" e com origem no "SISRP" (Sistema de Registro de Preços).

3.1 Coleta e Tratamento da Dase de Dados

Dados dos contratos coletados do portal do SIASG e de Unidades Federativas (UF) do portal BuscaCEP por meio de robôs no R. Foram obtidos dados complementares de variáveis sócio-econômicas em outras fontes, por exemplo: **IPEA IBGE (Sidra)**, entre outras.

Período: de janeiro de 2011 a dezembro de 2019. Totalizando dados de 256942 contratos na base de dados bruta.

Tratamento preliminar com exclusão: de variáveis sem importância para o estudo; de registros repetidos; de contratos com vigência maior que 52 semanas (> 1 ano); e de casos com valores faltantes para as demais variáveis ou com campos permutados.

Neste trabalho, restrigimos a análise a contratos da modalidade "Pregão" e com origem no "SISRP" (Sistema de Registro de Preços).

3.1 Coleta e Tratamento da Dase de Dados

Dados dos contratos coletados do portal do SIASG e de Unidades Federativas (UF) do portal BuscaCEP por meio de robôs no R. Foram obtidos dados complementares de variáveis sócio-econômicas em outras fontes, por exemplo: **IPEA IBGE (Sidra)**, entre outras.

Período: de janeiro de 2011 a dezembro de 2019. Totalizando dados de 256942 contratos na base de dados bruta.

Tratamento preliminar com exclusão: de variáveis sem importância para o estudo; de registros repetidos; de contratos com vigência maior que 52 semanas (> 1 ano); e de casos com valores faltantes para as demais variáveis ou com campos permutados.

Neste trabalho, restrigimos a análise a contratos da modalidade "Pregão" e com origem no "SISRP" (Sistema de Registro de Preços).

3.1 Coleta e Tratamento da Dase de Dados

Dados dos contratos coletados do portal do SIASG e de Unidades Federativas (UF) do portal BuscaCEP por meio de robôs no R. Foram obtidos dados complementares de variáveis sócio-econômicas em outras fontes, por exemplo: **IPEA IBGE (Sidra)**, entre outras.

Período: de janeiro de 2011 a dezembro de 2019. Totalizando dados de 256942 contratos na base de dados bruta.

Tratamento preliminar com exclusão: de variáveis sem importância para o estudo; de registros repetidos; de contratos com vigência maior que 52 semanas ($>$ 1 ano); e de casos com valores faltantes para as demais variáveis ou com campos permutados.

Neste trabalho, restrigimos a análise a contratos da modalidade "Pregão" e com origem no "SISRP" (Sistema de Registro de Preços).

3.1 Coleta e Tratamento da Dase de Dados

Dados dos contratos coletados do portal do SIASG e de Unidades Federativas (UF) do portal BuscaCEP por meio de robôs no R. Foram obtidos dados complementares de variáveis sócio-econômicas em outras fontes, por exemplo: **IPEA IBGE (Sidra)**, entre outras.

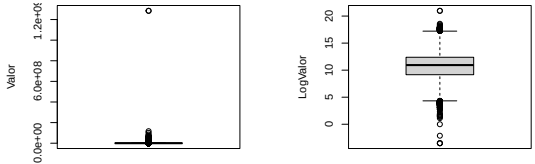
Período: de janeiro de 2011 a dezembro de 2019. Totalizando dados de 256942 contratos na base de dados bruta.

Tratamento preliminar com exclusão: de variáveis sem importância para o estudo; de registros repetidos; de contratos com vigência maior que 52 semanas ($>$ 1 ano); e de casos com valores faltantes para as demais variáveis ou com campos permutados.

Neste trabalho, restrigimos a análise a contratos da modalidade "Pregão" e com origem no "SISRP" (Sistema de Registro de Preços).

3.1 Coleta e Tratamento da Dase de Dados

Vale ressaltar que, para evitar pontos de influência, aplicou-se transformações logarítmicas nas variáveis Valor e Vigência do Contrato, atenuando a assimetria de suas distribuições. Veja a Figura 2.



(a) Valor (b) LogValor

Figura 2: Distribuições das variáveis Valor e LogValor.

O tratamento resultou em 29953 registros e 17 variáveis, sendo 16 covariáveis e a variável resposta (Aditivos).

3.1 Coleta e Tratamento da Dase de Dados

A Tabela 1 apresenta as variáveis no banco de dados após tratamento e seus respectivos significados.

Tabela 1: 1: Variáveis Aleatórias Utilizadas no Modelo.

SIGLA	VARIÁVEL
ADITIVOS	Número de Aditivos dos Contratos
AEF	Abandono Ensino Fundamental
AEM	Abandono Ensino Médio
GAEd	Grau de Abertura Econômica defasado
GR	Gini de Renda
LogValor	Logaritmo Neperiano do Valor do Contrato
LogVigencia	Logaritmo Neperiano do Valor da Vigência do Contrato
PIBPCE	Produto Interno Bruto Estadual Per Capita
RGPECPd	Razão de Gasto Público Estadual com Educação e Cultura Pelo PIB defasado
RGPESSPd	Razão de Gasto Público Estadual com Saúde e Saneamento Pelo PIB defasado
RGPEPd	Razão do Gasto Público Estadual Pelo PIB defasado
RIPEPd	Razão do Investimento Público Estadual Pelo PIB
RRPEPd	Razão da Receita Pública Estadual Pelo PIB defasado
TxCrPIBpC	Taxa de Crescimento do PIB Per Capita defasado
TH	Taxa de Homicídio
TM	Taxa de Mortalidade
TN	Taxa de Natalidade

Sumário

3.2 Análise e Discussão - Análise exploratória

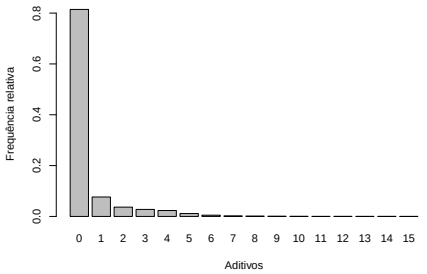


Figura 3: Número de Aditivos Contratuais.

A Figura 3 mostra que, entre 2 e 4, observa-se um discreto platô, indicando que a distribuição dos dados pode ser bem explicada por um mistura.

Temos que $\bar{X} = 0.4572$ e $S^2 = 1.5005$, respectivamente, o que indica característica de sobredispersão, pois $S^2 > \bar{X}$.

Conclusão: Além da vantagem de permitir a discriminação entre dois grupos sem as restrições dos modelos ZIP e ZINB, a combinação dos fatos acima consolida a motivação para a utilização do modelo Mix-NB.

3.2 Análise e Discussão - Análise exploratória

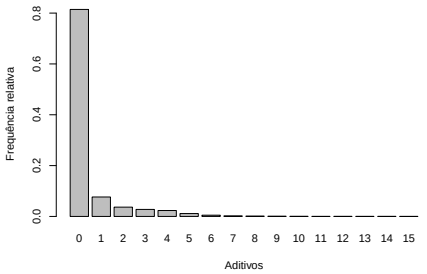


Figura 3: Número de Aditivos Contratuais.

A Figura 3 mostra que, entre 2 e 4, observa-se um discreto platô, indicando que a distribuição dos dados pode ser bem explicada por um mistura.

Temos que $\bar{X} = 0.4572$ e $S^2 = 1.5005$, respectivamente, o que indica característica de sobredispersão, pois $S^2 > \bar{X}$.

Conclusão: Além da vantagem de permitir a discriminação entre dois grupos sem as restrições dos modelos ZIP e ZINB, a combinação dos fatos acima consolida a motivação para a utilização do modelo Mix-NB.

3.2 Análise e Discussão - Análise exploratória

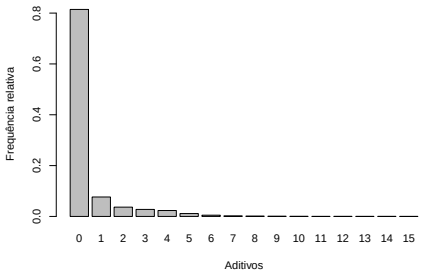


Figura 3: Número de Aditivos Contratuais.

A Figura 3 mostra que, entre 2 e 4, observa-se um discreto platô, indicando que a distribuição dos dados pode ser bem explicada por um mistura.

Temos que $\bar{X} = 0.4572$ e $S^2 = 1.5005$, respectivamente, o que indica característica de sobredispersão, pois $S^2 > \bar{X}$.

Conclusão: Além da vantagem de permitir a discriminação entre dois grupos sem as restrições dos modelos ZIP e ZINB, a combinação dos fatos acima consolida a motivação para a utilização do modelo Mix-NB.

3.2 Análise e Discussão - Análise exploratória

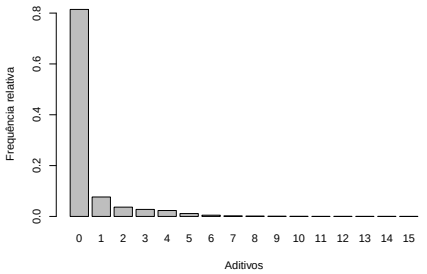


Figura 3: Número de Aditivos Contratuais.

A Figura 3 mostra que, entre 2 e 4, observa-se um discreto platô, indicando que a distribuição dos dados pode ser bem explicada por um mistura.

Temos que $\bar{X} = 0.4572$ e $S^2 = 1.5005$, respectivamente, o que indica característica de sobredispersão, pois $S^2 > \bar{X}$.

Conclusão: Além da vantagem de permitir a discriminação entre dois grupos sem as restrições dos modelos ZIP e ZINB, a combinação dos fatos acima consolida a motivação para a utilização do modelo Mix-NB.

3.2 Análise e Discussão - PCA

A fim de obter um modelo final mais parciomonioso, aplicou-se a técnica de Análise de Componentes Principais (PCA) com 14 variáveis.

As variáveis LogValor e LogVigencia foram incluídos diretamente no modelo devido à sua relevância, por isso não foram incluídas no PCA.

Aplicação do PCA resultou em 3 Componentes Principais (PC) capazes de explicar $\sim 93\%$ da variabilidade total das covariáveis.

Assim, nos modelos ajustados, em vez das covariáveis, foram utilizadas as 3 PC's obtidas.

3.2 Análise e Discussão - PCA

A fim de obter um modelo final mais parciomonioso, aplicou-se a técnica de Análise de Componentes Principais (PCA) com 14 variáveis.

As variáveis LogValor e LogVigencia foram incluídos diretamente no modelo devido à sua relevância, por isso não foram incluídas no PCA.

Aplicação do PCA resultou em 3 Componentes Principais (PC) capazes de explicar $\sim 93\%$ da variabilidade total das covariáveis.

Assim, nos modelos ajustados, em vez das covariáveis, foram utilizadas as 3 PC's obtidas.

3.2 Análise e Discussão - PCA

A fim de obter um modelo final mais parciomonioso, aplicou-se a técnica de Análise de Componentes Principais (PCA) com 14 variáveis.

As variáveis LogValor e LogVigencia foram incluídos diretamente no modelo devido à sua relevância, por isso não foram incluídas no PCA.

Aplicação do PCA resultou em 3 Componentes Principais (PC) capazes de explicar $\sim 93\%$ da variabilidade total das covariáveis.

Assim, nos modelos ajustados, em vez das covariáveis, foram utilizadas as 3 PC's obtidas.

3.2 Análise e Discussão - PCA

A fim de obter um modelo final mais parciomonioso, aplicou-se a técnica de Análise de Componentes Principais (PCA) com 14 variáveis.

As variáveis LogValor e LogVigencia foram incluídos diretamente no modelo devido à sua relevância, por isso não foram incluídas no PCA.

Aplicação do PCA resultou em 3 Componentes Principais (PC) capazes de explicar $\sim 93\%$ da variabilidade total das covariáveis.

Assim, nos modelos ajustados, em vez das covariáveis, foram utilizadas as 3 PC's obtidas.

3.2 Análise e Discussão

Neste trabalho, foram considerados os cenários de modelagem especificados na Tabela 2.

Tabela 2: Covariáveis Incorporadas ao Preditor Linear por Componente do Modelo.

Cenário	**Média** - $\log(\mu_i)$	**Probabilidade** - $\text{logit}(p_i)$
1	Só Intercepto ($X_i \sim 1$)	Só Intercepto ($Z_i \sim 1$)
2	Só Intercepto ($X_i \sim 1$)	Intercepto + covariáveis ($Z_i \sim .$)
3	Intercepto + covariáveis ($X_i \sim .$)	Só Intercepto ($Z_i \sim 1$)
4	Intercepto + covariáveis ($X_i \sim .$)	Intercepto + covariáveis ($Z_i \sim .$)

3.2 Análise e Discussão - KS Bootstrap

A Tabela 3 apresenta os p-valores dos teste KS bootastrap com $B = 200$ réplicas empregadas para testar a normalidade dos RQR's de cada modelo considerado.

Tabela 3: Resultado (p-valor bootstrap) do teste KS.

	Modelo				
Cenário	Pois	NB	ZIP	ZINB	Mix-NB
1: $X_i \sim 1$ e $Z_i \sim 1$	0,00498	0,00498	0,00498	**0,05473**	**0,12935**
2: $X_i \sim 1$ e $Z_i \sim .$	0,00498	0,00498	**0,16915**	**0,28358**	**0,28358**
3: $X_i \sim .$ e $Z_i \sim 1$	0,00498	**0,27861**	0,00498	0,02488	**0,14925**
4: $X_i \sim .$ e $Z_i \sim .$	0,00498	**0,23881**	0,00995	**0,96517**	**0,28358**

A Tabela 3 mostra que os únicos que têm aderência satisfatória aos dados são os modelos NB, ZIP, ZINB e Mix-NB.

Isso indica que a sobredispersão é uma característica importante dos dados e, por isso, precisa ser modelada adequadamente.

3.2 Análise e Discussão - KS Bootstrap

A Tabela 3 apresenta os p-valores dos teste KS bootastrap com $B = 200$ réplicas empregadas para testar a normalidade dos RQR's de cada modelo considerado.

Tabela 3: Resultado (p-valor bootstrap) do teste KS.

	Modelo				
Cenário	**Pois**	**NB**	**ZIP**	**ZINB**	**Mix-NB**
1: $X_i \sim 1$ e $Z_i \sim 1$	0,00498	0,00498	0,00498	**0,05473**	**0,12935**
2: $X_i \sim 1$ e $Z_i \sim .$	0,00498	0,00498	**0,16915**	**0,28358**	**0,28358**
3: $X_i \sim .$ e $Z_i \sim 1$	0,00498	**0,27861**	0,00498	0,02488	**0,14925**
4: $X_i \sim .$ e $Z_i \sim .$	0,00498	**0,23881**	0,00995	**0,96517**	**0,28358**

A Tabela 3 mostra que os únicos que têm aderência satisfatória aos dados são os modelos NB, ZIP, ZINB e Mix-NB.

Isso indica que a sobredispersão é uma característica importante dos dados e, por isso, precisa ser modelada adequadamente.

3.2 Análise e Discussão - KS Bootstrap

A Tabela 3 apresenta os p-valores dos teste KS bootastrap com $B = 200$ réplicas empregadas para testar a normalidade dos RQR's de cada modelo considerado.

Tabela 3: Resultado (p-valor bootstrap) do teste KS.

	Modelo				
Cenário	**Pois**	**NB**	**ZIP**	**ZINB**	**Mix-NB**
1: $X_i \sim 1$ e $Z_i \sim 1$	0,00498	0,00498	0,00498	**0,05473**	**0,12935**
2: $X_i \sim 1$ e $Z_i \sim .$	0,00498	0,00498	**0,16915**	**0,28358**	**0,28358**
3: $X_i \sim .$ e $Z_i \sim 1$	0,00498	**0,27861**	0,00498	0,02488	**0,14925**
4: $X_i \sim .$ e $Z_i \sim .$	0,00498	**0,23881**	0,00995	**0,96517**	**0,28358**

A Tabela 3 mostra que os únicos que têm aderência satisfatória aos dados são os modelos NB, ZIP, ZINB e Mix-NB.

Isso indica que a sobredispersão é uma característica importante dos dados e, por isso, precisa ser modelada adequadamente.

3.2 Análise e Discussão - KS Bootstrap

A Tabela 3 apresenta os p-valores dos teste KS bootastrap com $B = 200$ réplicas empregadas para testar a normalidade dos RQR's de cada modelo considerado.

Tabela 3: Resultado (p-valor bootstrap) do teste KS.

	Modelo				
Cenário	**Pois**	**NB**	**ZIP**	**ZINB**	**Mix-NB**
1: $X_i \sim 1$ e $Z_i \sim 1$	0,00498	0,00498	0,00498	**0,05473**	**0,12935**
2: $X_i \sim 1$ e $Z_i \sim .$	0,00498	0,00498	**0,16915**	**0,28358**	**0,28358**
3: $X_i \sim .$ e $Z_i \sim 1$	0,00498	**0,27861**	0,00498	0,02488	**0,14925**
4: $X_i \sim .$ e $Z_i \sim .$	0,00498	**0,23881**	0,00995	**0,96517**	**0,28358**

A Tabela 3 mostra que os únicos que têm aderência satisfatória aos dados são os modelos NB, ZIP, ZINB e Mix-NB.

Isso indica que a sobredispersão é uma característica importante dos dados e, por isso, precisa ser modelada adequadamente.

3.2 Análise e Discussão - Desempenho e Parcimônia

O desempenho e a parcimônia do modelo foram avaliados pelo Critério de Informação Bayesiano (BIC) definido por

$$\text{BIC} = k\log(N) - 2\log(f(y|\hat{\theta})),$$

em que k é quantidade de parâmetros, N é tamanho da amostra e $f(y|\hat{\theta})$ a verossimilhança máxima. A Tabela 4 apresenta os valores do BIC para os modelos que passaram no teste KS.

Tabela 4: Valor do BIC.

	Modelo				
Cenário	Pois	NB	ZIP	ZINB	Mix-NB
1: $X_i \sim 1$ e $Z_i \sim 1$	66736	47576	45219	47254	47250
2: $X_i \sim 1$ e $Z_i \sim .$	66736	47576	45221	44276	43917
3: $X_i \sim .$ e $Z_i \sim 1$	57969	44945	45014	44850	43300
4: $X_i \sim .$ e $Z_i \sim .$	57969	44945	44662	43955	**43249**

Assim, o BIC aponta para o modelo Mix-NB no Cenário 4.

3.2 Análise e Discussão - Desempenho e Parcimônia

O desempenho e a parcimônia do modelo foram avaliados pelo Critério de Informação Bayesiano (BIC) definido por

$$\text{BIC} = k \log(N) - 2 \log(f(y|\hat{\theta})),$$

em que k é quantidade de parâmetros, N é tamanho da amostra e $f(y|\hat{\theta})$ a verossimilhança máxima. A Tabela 4 apresenta os valores do BIC para os modelos que passaram no teste KS.

Tabela 4: Valor do BIC.

Cenário	Modelo				
	Pois	**NB**	**ZIP**	**ZINB**	**Mix-NB**
1: $X_i \sim 1$ e $Z_i \sim 1$	66736	47576	48219	47254	47250
2: $X_i \sim 1$ e $Z_i \sim .$	66736	47576	45221	44276	43917
3: $X_i \sim .$ e $Z_i \sim 1$	57969	44945	45914	44850	43300
4: $X_i \sim .$ e $Z_i \sim .$	57969	44945	44662	43955	**43249**

Assim, o BIC aponta para o modelo Mix-NB no Cenário 4.

3.2 Análise e Discussão - Desempenho e Parcimônia

O desempenho e a parcimônia do modelo foram avaliados pelo Critério de Informação Bayesiano (BIC) definido por

$$\text{BIC} = k\log(N) - 2\log(f(y|\hat{\theta})),$$

em que k é quantidade de parâmetros, N é tamanho da amostra e $f(y|\hat{\theta})$ a verossimilhança máxima. A Tabela 4 apresenta os valores do BIC para os modelos que passaram no teste KS.

Tabela 4: Valor do BIC.

	Modelo				
Cenário	**Pois**	**NB**	**ZIP**	**ZINB**	**Mix-NB**
1: $X_i \sim 1$ e $Z_i \sim 1$	66736	47576	48219	47254	47250
2: $X_i \sim 1$ e $Z_i \sim .$	66736	47576	45221	44276	43917
3: $X_i \sim .$ e $Z_i \sim 1$	57969	44945	45914	44850	43300
4: $X_i \sim .$ e $Z_i \sim .$	57969	44945	44662	43955	**43249**

Assim, o BIC aponta para o modelo Mix-NB no Cenário 4.

3.2 Análise e Discussão - Resultados

A Tabela 5 apresenta as proporções de contratos classificados como "eficientes" (Grupo 0) e "ineficientes" (Grupo 1):

Tabela 5: Proporções de contratos classificados em cada grupo.

Grupo	0	1
Proporação	0.904	0.096

É possível notar uma proporção de contratos "eficientes" relativamente alta no período considerado $\approx 90.4\%$.

Aproximadamente 9.6% dos contratos foram classificados com o "ineficientes". Isso mostra que é necessário se aprofundar nos fatores determinantes para a ineficiência encontrada.

3.2 Análise e Discussão - Resultados

A Tabela 5 apresenta as proporções de contratos classificados como "eficientes" (Grupo 0) e "ineficientes" (Grupo 1):

Tabela 5: Proporções de contratos classificados em cada grupo.

Grupo	0	1
Proporação	0.904	0.096

É possível notar uma proporção de contratos "eficientes" relativamente alta no período considerado $\approx 90.4\%$.

Aproximadamente 9.6% dos contratos foram classificados com o "ineficientes". Isso mostra que é necessário se aprofundar nos fatores determinantes para a ineficiência encontrada.

3.2 Análise e Discussão - Resultados

A Tabela 5 apresenta as proporções de contratos classificados como "eficientes" (Grupo 0) e "ineficientes" (Grupo 1):

Tabela 5: Proporções de contratos classificados em cada grupo.

Grupo	0	1
Proporação	0.904	0.096

É possível notar uma proporção de contratos "eficientes" relativamente alta no período considerado $\approx 90.4\%$.

Aproximadamente 9.6% dos contratos foram classificados com o "ineficientes". Isso mostra que é necessário se aprofundar nos fatores determinantes para a ineficiência encontrada.

3.2 Análise e Discussão - Resultados

A Tabela 5 apresenta as proporções de contratos classificados como "eficientes" (Grupo 0) e "ineficientes" (Grupo 1):

Tabela 5: Proporções de contratos classificados em cada grupo.

Grupo	0	1
Proporação	0.904	0.096

É possível notar uma proporção de contratos "eficientes" relativamente alta no período considerado $\approx 90.4\%$.

Aproximadamente 9.6% dos contratos foram classificados com o "ineficientes". Isso mostra que é necessário se aprofundar nos fatores determinantes para a ineficiência encontrada.

3.2 Análise e Discussão - Resultados

O modelo escolhido possui a seguinte estrutura estimada:

$$
\begin{aligned}
\log \hat{\mu}_{i0} &= -7.365(0.1828) + 0.4127(0.0121)\text{LogValor}_i + 0.2690(0.0463)\text{LogVigencia}_i \\
&\quad - 0.0164(0.0019)\text{PC1}_i + 0.0193(0.0027)\text{PC2}_i + 0.0016(0.0043)\text{PC3}_i, \\
\log \hat{\mu}_{i1} &= -668.58(8.6089) + 0.0299(0.0069)\text{LogValor}_i + 168.57(21.681)\text{LogVigencia}_i \\
&\quad - 0.0037(0.0011)\text{PC1}_i + 0.0040(0.0013)\text{PC2}_i - 0.0002(0.0019)\text{PC3}_i, \\
\text{logit}(\hat{p}_i) &= -8.3545(0.0905) + 0.1336(0.0119)\text{LogValor}_i + 1.3628(0.0162)\text{LogVigencia}_i \\
&\quad - 0.0088(0.0019)\text{PC1}_i + 0.0055(0.0025)\text{PC2}_i - 0.0149(0.0036)\text{PC3}_i,
\end{aligned}
$$

com $\alpha_0 = 5.8999(0.24)(< 0.0001)$ e $\alpha_1 = 0.0017(0.0073)(0.8110)$.

3.2 Análise e Discussão - Resultados

O modelo escolhido possui a seguinte estrutura estimada:

$$\begin{aligned}
\log \hat{\mu}_{i0} &= -7.365(0.1828) + 0.4127(0.0121)\text{LogValor}_i + 0.2690(0.0463)\text{LogVigencia}_i \\
&\quad - 0.0164(0.0019)\text{PC1}_i + 0.0193(0.0027)\text{PC2}_i + 0.0016(0.0043)\text{PC3}_i, \\
\log \hat{\mu}_{i1} &= -668.58(8.6089) + 0.0299(0.0069)\text{LogValor}_i + 168.57(21.681)\text{LogVigencia}_i \\
&\quad - 0.0037(0.0011)\text{PC1}_i + 0.0040(0.0013)\text{PC2}_i - 0.0002(0.0019)\text{PC3}_i, \\
\text{logit}(\hat{p}_i) &= -8.3545(0.0905) + 0.1336(0.0119)\text{LogValor}_i + 1.3628(0.0162)\text{LogVigencia}_i \\
&\quad - 0.0088(0.0019)\text{PC1}_i + 0.0055(0.0025)\text{PC2}_i - 0.0149(0.0036)\text{PC3}_i,
\end{aligned}$$

com $\alpha_0 = 5.8999(0.24)(< 0.0001)$ e $\alpha_1 = 0.0017(0.0073)(0.8110)$.

3.2 Análise e Discussão - Resultados

O modelo escolhido possui a seguinte estrutura estimada:

$$\begin{aligned}
\log \hat{\mu}_{i0} &= -7.365(0.1828) + 0.4127(0.0121)\text{LogValor}_i + 0.2690(0.0463)\text{LogVigencia}_i \\
&\quad - 0.0164(0.0019)\text{PC1}_i + 0.0193(0.0027)\text{PC2}_i + 0.0016(0.0043)\text{PC3}_i, \\
\log \hat{\mu}_{i1} &= -668.58(8.6089) + 0.0299(0.0069)\text{LogValor}_i + 168.57(21.681)\text{LogVigencia}_i \\
&\quad - 0.0037(0.0011)\text{PC1}_i + 0.0040(0.0013)\text{PC2}_i - 0.0002(0.0019)\text{PC3}_i, \\
\text{logit}(\hat{p}_i) &= -8.3545(0.0905) + 0.1336(0.0119)\text{LogValor}_i + 1.3628(0.0162)\text{LogVigencia}_i \\
&\quad - 0.0088(0.0019)\text{PC1}_i + 0.0055(0.0025)\text{PC2}_i - 0.0149(0.0036)\text{PC3}_i,
\end{aligned}$$

com $\alpha_0 = 5.8999(0.24)(< 0.0001)$ e $\alpha_1 = 0.0017(0.0073)(0.8110)$.

3.2 Análise e Discussão - Resultados

O modelo escolhido possui a seguinte estrutura estimada:

$$
\begin{aligned}
\log \hat{\mu}_{i0} &= -7.365(0.1828) + 0.4127(0.0121)\text{LogValor}_i + 0.2690(0.0463)\text{LogVigencia}_i \\
&\quad - 0.0164(0.0019)\text{PC1}_i + 0.0193(0.0027)\text{PC2}_i + 0.0016(0.0043)\text{PC3}_i, \\
\log \hat{\mu}_{i1} &= -668.58(8.6089) + 0.0299(0.0069)\text{LogValor}_i + 168.57(21.681)\text{LogVigencia}_i \\
&\quad - 0.0037(0.0011)\text{PC1}_i + 0.0040(0.0013)\text{PC2}_i - 0.0002(0.0019)\text{PC3}_i, \\
\text{logit}(\hat{p}_i) &= -8.3545(0.0905) + 0.1336(0.0119)\text{LogValor}_i + 1.3628(0.0162)\text{LogVigencia}_i \\
&\quad - 0.0088(0.0019)\text{PC1}_i + 0.0055(0.0025)\text{PC2}_i - 0.0149(0.0036)\text{PC3}_i,
\end{aligned}
$$

com $\alpha_0 = 5.8999(0.24)(< 0.0001)$ e $\alpha_1 = 0.0017(0.0073)(0.8110)$.

3.2 Análise e Discussão - Resultados

O modelo escolhido possui a seguinte estrutura estimada:

$$
\begin{aligned}
\log \hat{\mu}_{i0} &= -7.365(0.1828) + 0.4127(0.0121)\text{LogValor}_i + 0.2690(0.0463)\text{LogVigencia}_i \\
&\quad - 0.0164(0.0019)\text{PC1}_i + 0.0193(0.0027)\text{PC2}_i + 0.0016(0.0043)\text{PC3}_i, \\
\log \hat{\mu}_{i1} &= -668.58(8.6089) + 0.0299(0.0069)\text{LogValor}_i + 168.57(21.681)\text{LogVigencia}_i \\
&\quad - 0.0037(0.0011)\text{PC1}_i + 0.0040(0.0013)\text{PC2}_i - 0.0002(0.0019)\text{PC3}_i, \\
\text{logit}(\hat{p}_i) &= -8.3545(0.0905) + 0.1336(0.0119)\text{LogValor}_i + 1.3628(0.0162)\text{LogVigencia}_i \\
&\quad - 0.0088(0.0019)\text{PC1}_i + 0.0055(0.0025)\text{PC2}_i - 0.0149(0.0036)\text{PC3}_i,
\end{aligned}
$$

com $\alpha_0 = 5.8999(0.24)(< 0.0001)$ e $\alpha_1 = 0.0017(0.0073)(0.8110)$.

3.2 Análise e Discussão - Resultados

Tabela 6: Estimativas, Erros-Padrão e p-Valores para $\log \hat{\mu}_{i0}$.

Variável	**Parâmetro**	**Estimativa**	**Erro-Padrão**	p-**Valor**
Intercepto	β_{00}	-7.3647	0.1828	< 0.0001
LogValor	β_{01}	0.4127	0.0121	< 0.0001
LogVigencia	β_{02}	0.2690	0.0463	< 0.0001
PC1	β_{03}	-0.0164	0.0019	< 0.0001
PC2	β_{04}	0.0193	0.0027	< 0.0001
PC3	β_{05}	0.0016	0.0043	0.7078

$\alpha_0 = 5.8999(0.24)(< 0.0001)$

3.2 Análise e Discussão - Resultados

Tabela 7: Estimativas, Erros-Padrão e p-Valores para $\log \hat{\mu}_{i1}$.

Variável	**Parâmetro**	**Estimativa**	**Erro-Padrão**	***p*-Valor**
Intercepto	β_{10}	-668.58	8.6089	< 0.0001
LogValor	β_{11}	0.0299	0.0069	< 0.0001
LogVigencia	β_{12}	168.57	21.681	< 0.0001
PC1	β_{13}	-0.0037	0.0011	0.0005
PC2	β_{14}	0.0040	0.0013	0.0025
PC3	β_{15}	-0.0002	0.0019	0.9003

$\alpha_1 = 0.0017(0.0073)(0.8110)$

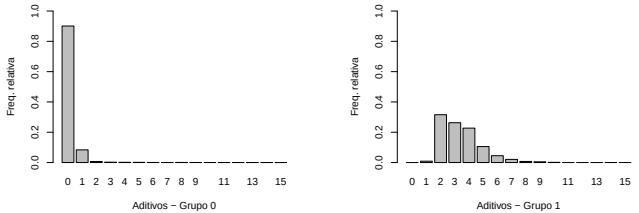
3.2 Análise e Discussão - Resultados

Tabela 8: Estimativas, Erros-Padrão e p-Valores para logit($\hat{p}_i$).

Variável	**Parâmetro**	**Estimativa**	**Erro-Padrão**	**p-Valor**
Intercepto	γ_0	-8.3545	0.0905	< 0.0001
LogValor	γ_1	0.1336	0.0119	< 0.0001
LogVigencia	γ_2	1.3628	0.0162	< 0.0001
PC1	γ_3	-0.0088	0.0019	< 0.0001
PC2	γ_4	0.0055	0.0025	0.0239
PC3	γ_5	-0.0149	0.0036	< 0.0001

3.2 Análise e Discussão - Resultados

A Figura 4 mostra as distribuições dos aditivos contratuais dos classificados no Grupo 0 ("eficientes") e os contratos do Grupo 1 ("ineficientes").



(a) Distribuição no Grupo 0. (b) Distribuição no Grupo 1.

Figura 4: Distribuição dos aditivos nos Grupos 0 e 1.

3.2 Análise e Discussão - Resultados

É possível notar que as distribuições em cada grupo seguem as características esperadas:

- Grupo 0 - alta concentração de zero aditivos;
- Grupo 1 - distribuição de aditivos caracterizada por uma locação mais alta e maior dispersão, o que indicaria a ineficência no planejamento licitatório.

A estimativa do parâmetro α_1 é muito próximo de zero, indicando que nesse grupo,os dados têm distribuição que pode ser aproximada por uma Poisson.

3.2 Análise e Discussão - Resultados

É possível notar que as distribuições em cada grupo seguem as características esperadas:

- Grupo 0 - alta concentração de zero aditivos;
- Grupo 1 - distribuição de aditivos caracterizada por uma locação mais alta e maior dispersão, o que indicaria a ineficência no planejamento licitatório.

A estimativa do parâmetro α_1 é muito próximo de zero, indicando que nesse grupo,os dados têm distribuição que pode ser aproximada por uma Poisson.

3.2 Análise e Discussão - Resultados

É possível notar que as distribuições em cada grupo seguem as características esperadas:

- Grupo 0 - alta concentração de zero aditivos;
- Grupo 1 - distribuição de aditivos caracterizada por uma locação mais alta e maior dispersão, o que indicaria a ineficência no planejamento licitatório.

A estimativa do parâmetro α_1 é muito próximo de zero, indicando que nesse grupo,os dados têm distribuição que pode ser aproximada por uma Poisson.

Sumário

4. Conclusão

Neste trabalho, o número de aditivos contratuais é utilizado como indicativo de ineficiência contratual.

O objetivo é propor metodologia estatística para discriminar contratos "eficientes" de "ineficientes", utilizando essa variável.

Os modelos existentes apresentam limitações para a aplicação considerada neste trabalho.

Para contornar essas limitações, é proposto um modelo de mistura de regressões binomial negativas (Mix-NB).

4. Conclusão

Neste trabalho, o número de aditivos contratuais é utilizado como indicativo de ineficiência contratual.

O objetivo é propor metodologia estatística para discriminar contratos "eficientes"de "ineficientes", utilizando essa variável.

Os modelos existentes apresentam limitações para a aplicação considerada neste trabalho.

Para contornar essas limitações, é proposto um modelo de mistura de regressões binomial negativas (Mix-NB).

4. Conclusão

Neste trabalho, o número de aditivos contratuais é utilizado como indicativo de ineficiência contratual.

O objetivo é propor metodologia estatística para discriminar contratos "eficientes"de "ineficientes", utilizando essa variável.

Os modelos existentes apresentam limitações para a aplicação considerada neste trabalho.

Para contornar essas limitações, é proposto um modelo de mistura de regressões binomial negativas (Mix-NB).

4. Conclusão

Neste trabalho, o número de aditivos contratuais é utilizado como indicativo de ineficiência contratual.

O objetivo é propor metodologia estatística para discriminar contratos "eficientes"de "ineficientes", utilizando essa variável.

Os modelos existentes apresentam limitações para a aplicação considerada neste trabalho.

Para contornar essas limitações, é proposto um modelo de mistura de regressões binomial negativas (Mix-NB).

4. Conclusão

Pelo nosso conhecimento, esse modelo ainda não foi considerado na literatura.

O modelo proposto se mostrou bem ajustado aos dados e apresentou o melhor desempenho comparado aos já existentes na literatura.

Além disso, o modelo foi capaz de discriminar contratos eficientes de ineficientes de maneira mais flexível às alternativas existentes.

4. Conclusão

Pelo nosso conhecimento, esse modelo ainda não foi considerado na literatura.

O modelo proposto se mostrou bem ajustado aos dados e apresentou o melhor desempenho comparado aos já existentes na literatura.

Além disso, o modelo foi capaz de discriminar contratos eficientes de ineficientes de maneira mais flexível às alternativas existentes.

4. Conclusão

Pelo nosso conhecimento, esse modelo ainda não foi considerado na literatura.

O modelo proposto se mostrou bem ajustado aos dados e apresentou o melhor desempenho comparado aos já existentes na literatura.

Além disso, o modelo foi capaz de discriminar contratos eficientes de ineficientes de maneira mais flexível às alternativas existentes.

Referências

P. K. Dunn e G. K. Smyth. Randomized quantile residuals. *Journal of Computational and graphical statistics*, 5(3):236–244, 1996.

B. Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

A. J. Q. Sarnaglia, N. A. J. Monroy, e A. G. da Vitória. Modeling and forecasting daily maximum hourly ozone concentrations using the RegAR model with skewed and heavy-tailed innovations. *Environmental and Ecological Statistics*, 25:443–469, 2018.

OBRIGADO!

Contato: Emerson Pazeto
E-mail: emebompaz@gmail.com
Whatsapp: 27 99775 6194