Universidade de Lisboa

Faculdade de Ciências

Departamento de Estatística
e Investigação Operacional

MODELOS COM EXCESSO DE ZEROS E MODELOS DE DUAS PARTES – A SUA UTILIZAÇÃO NO ESTUDO DA SCHISTOSOMOSE

Daniel Vigário Olivença

Mestrado em Estatística

2011

Universidade de Lisboa

Faculdade de Ciências

Departamento de Estatística
e Investigação Operacional

MODELOS COM EXCESSO DE ZEROS E MODELOS DE DUAS PARTES – A SUA UTILIZAÇÃO NO ESTUDO DA SCHISTOSOMOSE

Daniel Vigário Olivença

Dissertação orientada por: Patrícia de Zea Bermudez
Luzia Gonçalves

Mestrado em Estatística

2011

AGRADECIMENTOS

Os meus sinceros agradecimentos vão para todos aqueles que directa ou indirectamente contribuíram para a realização desta tese.

Um muito especial agradecimento às minhas orientadoras, as professoras Patrícia Bermudez e Luzia Gonçalves. Obrigado pela sabedoria, tempo, críticas e pela muita paciência que tiveram comigo.

Um grande agradecimento à minha família, em especial à minha mãe. Sem o apoio deles este trabalho não teria começado.

Sem o apoio da Rita Duarte este trabalho nunca teria chegado ao fim.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

RESUMO

Para modelar variáveis que apresentam um grande número de zeros existem os modelos com excesso de zeros e os modelos de duas partes, que têm sido amplamente usados numa grande variedade de áreas de estudo. Estes modelos serão aplicados ao estudo, numa população Angolana, da distribuição da Schistosomose, uma parasitose superada em frequência apenas pela malária.

Os resultados sugerem que o modelo GLM (Generalized Linear Models) binomial negativa, o modelo binomial negativa com excesso de zeros e o modelo de duas partes com binomial negativa são preferíveis aos modelos com distribuição de Poisson. Os resultados obtidos mostram que não existem diferenças significativas entre os modelos com binomial negativa. Escolheu-se o modelo binomial negativa com excesso de zeros pois existem duas fontes possíveis para os zeros: os indivíduos que não estão infectados (zeros verdadeiros) e os indivíduos que, apesar estarem infectados, não apresentam ovos na urina (zeros falsos). Assim, a estrutura deste modelo é mais concordante com a situação estudada.

Palavras-chave:
Modelos com excesso de zeros, modelos de duas partes, binomial negativa, Poisson, Schistosomose urinária.

ABSTRACT

The zero inflated and the Hurdle models are widely applied in a large variety of fields of study for modeling variables that exhibit a large number of zeros. These models are here used for studying the number of cases of Schistosomiasis observed in some regions of Angola. The Schistosomiasis is the second most frequent human parasitic disease after the malaria.

The results suggest that the Generalized Linear Model Negative Binomial, the Zero Inflated Negative Binomial and the Hurdle Negative Binomial models are preferable to the Poisson models. No significant differences were found between the various negative binomial models. The Zero Inflated Negative Binomial model was chosen because the sample contains two possible sources of zeros: either the individuals did not have the infection or they were already sick, although no eggs were found in the urine. Therefore, the structure of this model is the most similar to the case being studied.

Key Words:
Zero inflated models, Hurdle models, negative binomial, Poisson, urinary Schistosomiasis

SIGLAS

GLM	Modelos lineares generalizados (Generalized Linear Models).
WHO	Organização Mundial de Saúde (World Health Organization).
ZA	Modelos de duas partes (Zero Altered Models), também conhecidos por modelos de barreira (Hurdle Models).
ZANB	Modelo de duas partes com binomial negativa (Zero Altered Negative Binomial).
ZAP	Modelo de duas partes com Poisson (Zero Altered Poisson).
ZI	Modelos com excesso de zeros (Zero Inflated Models)
ZINB	Binomial negativa com excesso de zeros (Zero Inflated Negative Binomial).
ZIP	Poisson com excesso de zeros (Zero Inflated Poisson).

1 - INTRODUÇÃO

A Schistosomose é considerada a segunda parasitose com maior impacto na saúde pública, a seguir à malária. Estimam-se em 120 milhões o número de indivíduos infectados (Velez, 2010) e 600 milhões em risco de contrair a infecção (Figueiredo, 2008).

Os estudos sobre esta doença centram-se especialmente na população infantil, pois esta é a mais afectada. Apesar disso, a população adulta sofre consideravelmente devido à Schistosomose ser uma doença negligenciada nos países com baixos recursos económicos (Figueiredo, 2008). Um conhecimento aprofundado da distribuição desta doença é importante para ajudar à adopção de medidas que combatam e previnam a doença.

Nesta tese será estudada a distribuição da carga parasitária de *Schistosoma haematobium* numa população de indivíduos com idades entre os 15 e os 75 anos.

A carga parasitária desta doença não pode ser estudada directamente contando parasitas, pois estes alojam-se nas veias do hospedeiro. Assim, no caso do *Schistosoma haematobium*, tenta-se determinar a carga parasitária através do número de ovos presente numa amostra de 10 ml de urina.

Em estudos anteriores (Figueiredo, 2008 e Cardoso, 2010), nos quais a variável de estudo também é o número de ovos de *Schistosoma haematobium* em 10 ml de urina, a sua distribuição é caracterizada por um elevado número de zeros e uma grande dispersão de valores não nulos, sendo as contagens mais pequenas as mais frequentes.

Para modelar dados com grande número de zeros é possível usar os modelos com excesso de zeros (Zero Inflated Models (ZI)) e os modelos de duas partes (Zero Altered Models (ZA)), também conhecidos por modelos de barreira (Hurdle Models). Estes modelos podem ser usados com várias distribuições. Neste trabalho serão usadas as distribuições de Poisson e binomial negativa (também conhecida como distribuição de Polya ou de Pascal). Os modelos lineares generalizados (GLM) são também aplicados com as distribuições Poisson e binomial negativa.

Estes modelos têm sido aplicados em muitas áreas, tais como na economia (Ground e Koch, 2008), na biologia (Gonzales-Barron et al., 2010), na medicina (Rose et al., 2006; Kahama et al., 1999; Kallestrup et al., 2005), na psiquiatria (Bethell et al., 2010), na ecologia (Potts e Elith, 2006; Zuur et al., 2009), entre outras.

O principal objectivo deste trabalho é aplicar os modelos ZI e ZA e averiguar como estes modelam o número de ovos na urina. Os modelos lineares generalizados aparecem com a principal função de servir de termo de comparação com os modelos ZI e ZA.

A hipótese inicial é que um modelo com a distribuição binomial negativa será preferível a um modelo com distribuição de Poisson devido à grande dispersão de valores. Espera-se também que os modelos ZI e ZA tenham uma melhor prestação que os modelos lineares generalizados, devido ao grande número de zeros presente na amostra.

Esta tese está estruturada da seguinte forma: inicialmente faz-se uma caracterização da população e da amostra; descreve-se a doença, o parasita e realiza-se uma análise preliminar dos dados, dando especial ênfase à análise das covariáveis. De seguida, é apresentado um pequeno resumo dos modelos a aplicar. Por fim, descreve-se a aplicação destes modelos e apresentam-se as respectivas conclusões.

2 – DADOS, POPULAÇÃO E AMOSTRA

Os dados foram recolhidos pela Dr.ª Jacinta Teresa Figueiredo para a realização da sua tese de mestrado em Parasitologia Médica no Instituto de Higiene e Medicina Tropical - UNL que consta da bibliografia do presente trabalho. Assim, parte significativa da informação contida neste capítulo baseia-se na tese da Dr.ª Jacinta (Figueiredo, 2008).

No seguimento da tese anteriormente referida, surgiu a necessidade de um tratamento estatístico mais aprofundado, principalmente no que se refere à distribuição do número de ovos em 10 ml de urina, uma vez que esta variável apresenta um grande número de zeros e uma grande dispersão de valores.

A população da qual foi retirada a amostra é constituída por indivíduos angolanos com idades compreendidas entre os 15 e os 75 anos, maioritariamente a viver em comunidades pobres, não tendo acesso a água potável nem a instalações sanitárias.

Estas comunidades estão localizadas em três províncias de Angola: Luanda, Bengo e Kwanza Sul. A amostra foi recolhida em cinco localidades pertencentes a estas três províncias: Luanda, Bom Jesus, Cambalo, Lundy e Macedónia (Figueiredo, 2008).

Figura 1 - Mapa de Angola com as províncias onde se realizou a recolha de dados. (fonte: Figueiredo, 2008)

Existem muitos indivíduos a residir nas zonas de recolha da amostra que, devido à instabilidade que se viveu em Angola durante vários anos, provêm de outras províncias nas quais a Schistosomose urinária é endémica.

A recolha da amostra foi previamente combinada com os chefes locais e realizada em locais de culto ou em escolas durante as férias escolares.

Inicialmente, foi realizada uma pequena sessão de sensibilização dirigida à população alvo, com o objectivo de informar os residentes das zonas acima indicadas sobre a importância do parasita e de explicar o objectivo do estudo.

Os indivíduos foram seleccionados para a amostra com base na sua comparência e concordância com as condições do estudo. De seguida, foram sujeitos a um inquérito com entrevista e procedeu-se a uma explicação de como deveria ser feita a recolha do material biológico (foram facultados dois frascos, um para a recolha de fezes e outro para a urina). A análise da urina foi feita pelo método de filtração através de filtros Millipore e observada com um microscópio óptico. Todos os indivíduos com amostras contendo ovos foram medicados com *Praziquantel* segundo as orientações da Organização Mundial da Saúde (WHO, 2002).

Os indivíduos que apresentavam queixas mais severas foram encaminhados para o Hospital Américo Boavida, em Luanda, para realizarem ecografias e cistoscopias. Em nove destes indivíduos foram detectados ovos nos exames realizados no hospital, embora não tenham sido detectados no exame por filtração da urina.

3 - REVISÃO DA LITERATURA

3.1 - A SCHISTOSOMOSE

Existem várias formas de *Schistosoma*: *S. haematobium*, *S. mansoni*, *S.japonicum*, *S. intercalatum* e *S. mekongi* (WHO, 1979 in Velez, 2010), que causam vários tipos de Schistosomose.

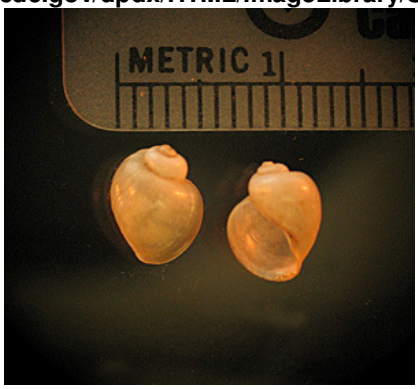
Este trabalho incide sobre a Schistosomose urinária (ou esquistossomose urinária), causada pelo parasita *Schistosoma haematobium*. Estima-se que afecte 120 milhões de pessoas em todo o mundo (Velez, 2010).

Apesar de esta doença ser considerada um grande problema de saúde pública, especialmente em países com poucos recursos económicos, o tratamento é pouco dispendioso. O seu custo varia entre os 9 e os 18 cêntimos de dólar por dose.(http://www.stanford.edu/class/humbio103/ParaSites2005/Praziquantel/Praziquantel.htm).

Figura 2 - Ovo de *Schistosoma.*
(fonte: http://www.dpd.cdc.gov/dpdx/HTML/ImageLibrary/Schistosomiasis_il.htm)

O parasita *Schistosoma haematobium* tem um ciclo de vida, durante o qual assume várias formas. Os ovos são expelidos do organismo através da urina ou das fezes que, ao entrarem em contacto com a água doce, chocam em miracídeos. Estes procuram o vector da doença, um caracol aquático (*Bulinus truncatus*), dando-se a multiplicação da forma assexuada, o esporocisto. Este evolui para a forma multicelular de cercária e abandona o molusco, procurando um hospedeiro.

Figura 3 - *Bulinus truncatus,* vector do *Schistosoma haematobium.*
(fonte: http://www.dpd.cdc.gov/dpdx/HTML/ImageLibrary/Schistosomiasis_il.htm)



As cercárias entram no organismo por qualquer parte do corpo. Podem passar através da pele até encontrar vasos sanguíneos nos quais entram e se deslocam. Viajam até aos pulmões onde se fixam e evoluem para a forma jovem. Deslocam-se para o fígado onde o macho atinge a forma adulta. A fêmea só matura na presença de um macho. Esta instala-se numa fenda no corpo do macho (canal ginecóforo), ao qual a espécie deve o seu nome. A designação *Schistosoma* vem do grego *schistos*=fenda e *soma*=corpo (Colley, 1996 e Belo, 1999 in Cardoso, 2010).

Figura 4 - Cercárias.
(fonte: http://www.scienceinschool.org/2009/issue11/schistosomiasis)

O casal viaja até às veias e artérias perto da bexiga e aí se fixa, iniciando a produção de ovos. Alguns ovos mantêm-se no sistema circulatório e acumulam-se no fígado, podendo provocar uma inflamação.

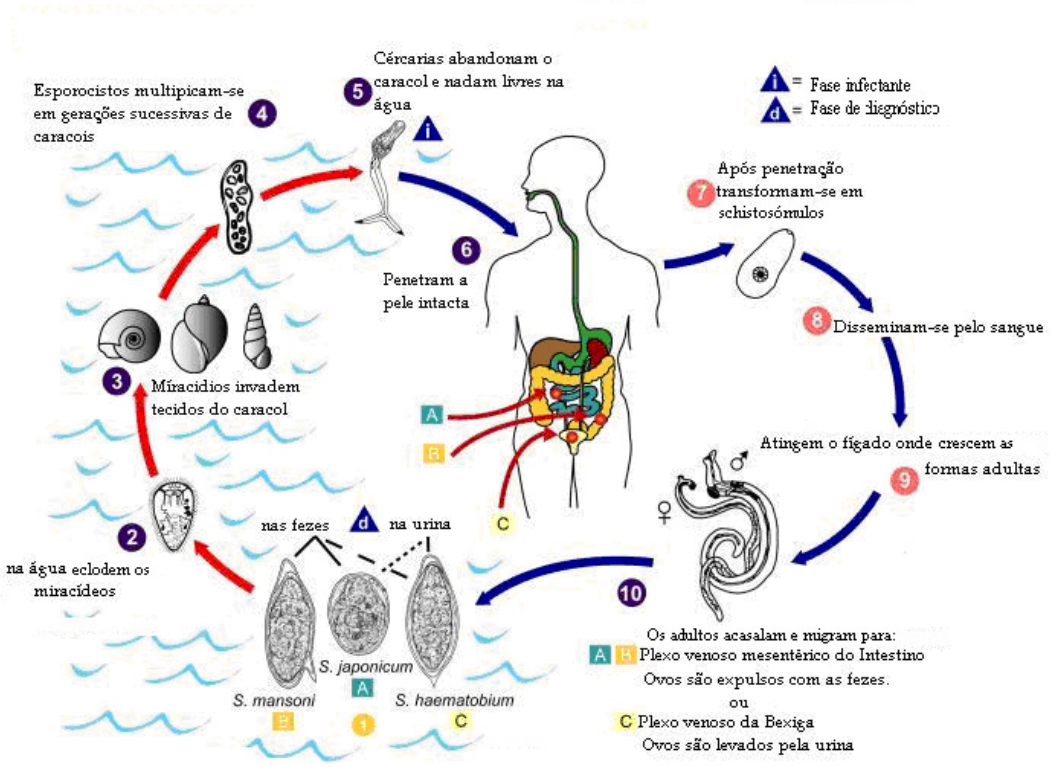
Figura 5 - Macho com fêmea de *Schistosoma haematobium.*
(fonte: http://www.stanford.edu/class/humbio103/ParaSites2004/Schisto/website.html)

Ao serem expelidos, se os ovos entrarem em contacto com a água doce reiniciam o ciclo. Este verme pode viver no organismo humano por um período de 20 a 30 anos (Warren *et al.*, 1974 in Cardoso, 2010).

As consequências da Schistosomose urinária são várias. Devido à sua localização nas veias próximas da bexiga, as lesões patológicas localizam-se sobretudo nos órgãos genitourinários e rins, o que causa dores ao urinar e dores abdominais. Na ausência de tratamento (Figueiredo, 2008) podem evoluir para complicações graves e irreversíveis, incluindo cancro da bexiga.

As crianças são particularmente susceptíveis de serem infectadas devido às actividades que as colocam em contacto com a água doce (ex.: brincadeiras na água, transportar água, entre outras). Também importantes são as consequências no desenvolvimento do indivíduo - problemas no crescimento, menor desenvolvimento cognitivo e problemas de concentração devido ao mal--estar causada pelo parasita. Pode ainda causar anemia, debilidade crónica e morte prematura (WHO, 2002).

Figura 6 - Ciclo de vida do *Schistosoma.*
(fonte:http://www.dpd.cdc.gov/dpdx/HTML/ImageLibrary/Schistosomiasis_il.htm)



3.2 - MODELOS

Seja Y_i a variável em estudo, resultante de um processo de contagem ou, por outras palavras, seja Y_i o número de vezes que um determinado acontecimento ocorre. Seja y_i o valor da variável Y_i para o *i*-ésimo indivíduo da amostra.

Suponhamos que existem *m* covariáveis $X_i = (X_{1i}, X_{2i}, ..., X_{mi})$ e que $x_i = (x_{1i}, x_{2i}, ..., x_{mi})$ representa os valores das covariáveis correspondentes ao *i*-ésimo elemento da amostra.

Para modelar dados com grande número de zeros é possível usar os modelos ZI e ZA. Estes modelos podem ser usados com várias distribuições. Como a variável de interesse (número de ovos em 10 ml de urina) toma valores não negativos, as distribuições de Poisson e binomial negativa (também conhecida como distribuição de Polya ou de Pascal) podem ser apropriadas para proceder à sua modelação.

A distribuição de Poisson foi a primeira a ser utilizada com modelos ZI e ZA, sendo muito natural em casos de contagens.

A distribuição binomial negativa também é muito usada em conjugação com os modelos ZI e ZA e tem a particularidade de "lidar" bem com a grande dispersão de valores, característica presente nos dados analisados neste trabalho.

Tabela 1 - Modelos usados e suas características.

Modelo	Tem em consideração o excesso de zeros.	Tem em consideração a grande amplitude de valores dos dados.
GLM Poisson	não	não
GLM binomial negativa	não	sim
Poisson com excesso de zeros	sim	não
Binomial Negativa com excesso de zeros	sim	sim
Modelo de duas partes com Poisson	sim	não
Modelo de duas partes com binomial negativa	sim	sim

Neste trabalho iremos usar seis tipos de modelos para modelar o número de ovos em 10 ml de urina. Os modelos são GLM Poisson, GLM binomial negativa, Poisson com excesso de zeros (ZIP), binomial negativa com excesso de zeros (ZINB), modelo de duas partes com Poisson (ZAP) e modelo de duas partes com binomial negativa (ZANB). As siglas usadas correspondem às designações na língua inglesa dos modelos que serão apresentadas seguidamente.

3.2.1 GLM POISSON

Este é o modelo mais natural para modelar dados de contagens.

A sua função de massa de probabilidade é dada por:

$$f(y_i \mid \mu) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}, y_i = 0,1,2,... \quad (1)$$

onde μ $(\mu > 0)$ é o valor esperado da população à qual y_i pertence.

Para o *i*-ésimo indivíduo da amostra, se existirem *m* covariáveis representadas por $X_{i1}, X_{i2}, ..., X_{im}$, $\beta_1, \beta_2, ..., \beta_m$ são os coeficientes das covariáveis a estimar em cada modelo e α é o valor da ordenada na origem.

Usando uma função de ligação (log link) podemos calcular o número esperado de ovos na urina de cada indivíduo da amostra da seguinte forma:

$$\mu_i = e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_m x_{im}} \quad (2)$$

Estimativas para α e $\beta_1, \beta_2, ..., \beta_m$ são calculados por máxima verosimilhança, ou seja, maximizando a função de log verosimilhança

$$LL = \sum_i \left(y_i(\alpha + \beta_1 x_{i1} + ... + \beta_m x_{im})\right) - \sum_i e^{\alpha+\beta_1 x_{i1}+...+\beta_m x_{im}} - \sum_i \ln\left(\Gamma(y_i + 1)\right) \quad (3)$$

Os estimadores de máxima verosimilhança para todos os modelos com distribuição de Poisson usados neste trabalho podem ser encontrados em Cameron et al. (1998) in Zuur et al. (2009).

Nesta distribuição $E[Y_i] = Var[Y_i] = \mu_i$.

Como está indicado na Tabela 1, esta distribuição não "lida" com excesso de zeros nem com sobre-dispersão dos dados.

3.2.2 - GLM BINOMIAL NEGATIVA

Para que a distribuição de Poisson seja mais flexível, permite-se que o seu valor médio μ seja substituído por uma variável aleatória λ (Cook, 2009) que usualmente segue uma Gamma(θ,β) tal que $E[\lambda]=\theta\beta=\mu$, de onde se obtém que $\beta=\frac{\mu}{\theta}$. Assim,

$$P(Y=y_i \mid \lambda)=\frac{1}{\Gamma(\theta)\beta^{\theta}}\int_0^{+\infty}\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\lambda^{\theta-1}e^{-\lambda/\beta}d\lambda=$$

$$=\frac{1}{y_i!\Gamma(\theta)\beta^{\theta}}\int_0^{+\infty}\lambda^{\theta+y_i-1}e^{-\lambda\left(1+1/\beta\right)}d\lambda=$$

$$=\frac{1}{\Gamma(y_i+1)\Gamma(\theta)\beta^{\theta}}\Gamma(\theta+y_i)\left(\frac{\beta}{\beta+1}\right)^{\theta+y_i}=$$

$$=\frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)}\left[\frac{1}{\beta^{\theta}}\left(\frac{\beta}{\beta+1}\right)^{\theta}\right]\left(\frac{\beta+1-1}{\beta+1}\right)^{y_i}=$$

$$=\frac{\Gamma(\theta+y_i)}{y_i!\Gamma(\theta)}\left(\frac{1}{\beta+1}\right)^{\theta}\left(1-\frac{1}{\beta+1}\right)^{y_i} \qquad (4)$$

Como $\beta=\frac{\mu}{\theta}$, substituindo em (4) obtemos

$$f(y_i \mid \mu,\theta)=\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)y_i!}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(1-\frac{\theta}{\mu+\theta}\right)^{y_i},\ y_i=0,1,2,... \qquad (5)$$

que é a função de massa de probabilidade de uma binomial negativa onde μ é o valor médio da população e θ o parâmetro de forma. Como no modelo anterior pode-se usar (2) para modelar μ_i para cada indivíduo da amostra, usando as covariáveis. θ não depende das covariáveis e será mais um parâmetro a estimar. Obtém-se assim

$$f(y_i \mid \mu_i,\theta)=\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)y_i!}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(1-\frac{\theta}{\mu_i+\theta}\right)^{y_i},\ y_i=0,1,2,... \qquad (6)$$

Dois modelos dizem-se encaixados (do inglês *nested*) se os dois têm os mesmos termos ou, um dos modelos tem pelo menos, mais um termo que o outro (Zuur et al., 2009). Como a binomial negativa pode ser construída a partir da Poisson, se as covariáveis usadas nos dois modelos forem as mesmas ou o modelo com binomial negativa apresentar pelo menos mais uma covariável que o modelo com Poisson então pode-se afirmar que a binomial negativa e a Poisson estão encaixadas (usando a terminologia de Paulino, 2011).

Neste caso $E[Y_i]=\mu_i$ e $Var[Y_i]=\mu_i+\frac{\mu_i^{\,2}}{\theta}$.

3.2.3 - POISSON COM EXCESSO DE ZEROS (ZIP)

Este modelo é a mistura de uma distribuição Binomial (Bernoulli) com uma Poisson.

A sua principal característica é que pode produzir zeros de duas fontes. A Binomial produz um zero com probabilidade π. Com probabilidade $1-\pi$ é produzido um valor pela Poisson. Por sua vez, a Poisson produz um zero com probabilidade $e^{-\mu}$. Assim, juntando os dois últimos parágrafos obtém-se o primeiro ramo da equação (8).

O segundo ramo corresponde à função massa de probabilidade da Poisson multiplicada pela probabilidade de se obter um valor desta distribuição ($1-\pi$).

Como nos modelos anteriores podemos usar (2) para calcular μ_i para cada indivíduo da amostra, usando as covariáveis.

Para calcular π usa-se a regressão logística (Zuur et al., 2009). Assim, tem-se para o indivíduo *i*

$$\pi_i = \frac{e^{\alpha+\delta_1 z_{i1}+\ldots+\delta_n z_{in}}}{1+e^{\alpha+\delta_1 z_{i1}+\ldots+\delta_n z_{in}}} \qquad (7)$$

Como as covariáveis usadas para calcular a probabilidade de obter zeros na binomial podem ser ou não ser as mesmas que se usam para calcular μ_i, foram aqui denotadas por *z* e a sua quantidade por *n*. Os seus coeficientes foram designados por δ.

Obtemos assim a função de probabilidade:

$$f(y_i \mid \pi_i, \mu_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i} & , y_i = 0 \\ (1-\pi_i)\dfrac{e^{-\mu_i}\mu_i^{\,y_i}}{y_i!} & , y_i = 1,2,3,\ldots \end{cases} \qquad (8)$$

Resumindo, existem duas fontes de zeros, os zeros que provêm da Binomial e os zeros que provêm da Poisson.

Para este modelo temos que $E[Y_i] = \mu_i(1-\pi_i)$ e $Var[Y_i] = (1-\pi_i)(\mu_i + \pi\mu_i^{\,2})$.

O modelo ZIP pode ser construído a partir da Poisson. Se as covariáveis usadas nos dois modelos forem as mesmas ou se o ZIP tiver, pelo menos, mais uma covariável que a Poisson pode-se afirmar que o modelo ZIP e a Poisson estão encaixados.

3.2.4 - BINOMIAL NEGATIVA COM EXCESSO DE ZEROS (ZINB)

Este modelo de mistura funciona de uma forma muito similar ao ZIP, bastando substituir a distribuição de Poisson pela binomial negativa. Para melhor entender a fórmula, chama-se à atenção que a probabilidade da binomial negativa originar um zero é $\left(\frac{\theta}{\mu+\theta}\right)^{\theta}$. Assim, de forma similar ao modelo ZIP, o primeiro ramo indica a probabilidade de obtermos um zero, que é a probabilidade de um zero ser gerado pela Binomial (π) mais a probabilidade de ser gerado pela binomial negativa $(1-\pi)\left(\frac{\theta}{\mu+\theta}\right)^{\theta}$. O segundo ramo fornece a probabilidade de obtermos os valores não nulos através da binomial negativa.

Obtém-se μ_i segundo a equação (2) e π_i segundo a equação (7). Obtém-se a função de probabilidade

$$f(y_i \mid \pi_i, \theta, \mu_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta} & , y_i = 0 \\ (1-\pi_i)\frac{\Gamma(y_i+\theta)}{\Gamma(\theta)y_i!}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(1-\frac{\theta}{\mu_i+\theta}\right)^{y_i} & , y_i = 1,2,3,... \end{cases} \qquad (9)$$

Neste modelo $E[Y_i] = \mu_i(1-\pi_i)$ e $Var[Y_i] = (1-\pi_i)\left(\mu_i + \frac{\mu_i^2}{\theta}\right) + \mu_i^2(\pi_i^2 + \pi_i)$

Como o ZINB pode ser construído a partir da binomial negativa, se as covariáveis usadas nos dois modelos forem as mesmas ou se o modelo ZINB tiver pelo menos uma covariável a mais que a binomial negativa, pode-se afirmar que o ZINB e o modelo com binomial negativa estão encaixados. Como a binomial negativa e a Poisson estão encaixadas, tendo atenção à condição das covariáveis, pode-se afirmar que o ZINB e a Poisson estão encaixadas.

3.2.5 - MODELO DE DUAS PARTES COM POISSON (ZAP)

Estes modelos, conhecidos como modelos de barreira (ou Hurdle Models), são designados também de duas partes.

Nestes modelos uma das partes é uma Binomial que origina zeros com probabilidade π. A outra parte do modelo produz, com probabilidade $1-\pi$, o valor de uma Poisson truncada de forma a apenas obtermos valores não nulos.

Ao contrário dos modelos ZI, este tipo de modelos apenas apresenta uma só fonte de zeros.

Obtém-se μ_i segundo a equação (2) e π_i segundo a equação (7). Obtém-se a função de probabilidade

$$f(y_i \mid \pi_i, \mu_i) = \begin{cases} \pi_i & , y_i = 0 \\ (1-\pi_i)\dfrac{e^{-\mu_i}\mu_i^{\,y_i}}{y_i!\left(1-e^{-\mu_i}\right)} & , y_i = 1,2,3,... \end{cases} \qquad (10)$$

Neste modelo $E[Y_i] = \mu\dfrac{(1-\pi_i)}{\left(1-e^{-u_i}\right)}$ e $Var[Y_i] = \dfrac{1-\pi_i}{1-e^{-\mu_i}}\left(\mu_i + \pi_i\mu_i^{\,2}\right) - \left(\dfrac{1-\pi_i}{1-e^{-\mu_i}}\mu_i\right)^2$.

Este modelo não está encaixado com nenhum outro referido anteriormente.

3.2.6 - MODELO DE DUAS PARTES COM BINOMIAL NEGATIVA (ZANB)

Este modelo é em tudo semelhante ao anterior, substituindo a Poisson truncada por uma binomial negativa truncada. Os parâmetros μ_i e π_i são obtidos da forma anteriormente indicada.

A sua massa de probabilidade é dada por:

$$f(y_i \mid \pi_i, \theta, \mu_i) = \begin{cases} \pi_i & , y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma(y_i+\theta)}{\Gamma(\theta)y_i!}\dfrac{{\mu_i}^{y_i}\theta^{\theta}}{(\mu_i+\theta)^{y_i+\theta}}\left[1-\left(\dfrac{\theta}{\mu_i+\theta}\right)^{\theta}\right]^{-1} & , y_i = 1,2,3,... \end{cases} \quad (11)$$

Neste modelo $E[Y_i] = \mu_i \dfrac{(1-\pi_i)}{\left(1-\left(\dfrac{\theta}{\mu_i+\theta}\right)^{\theta}\right)}$, e

$$Var[Y_i] = \frac{(1-\pi_i)}{\left(1-\left(\dfrac{\theta}{\mu_i+\theta}\right)^{\theta}\right)}\left({\mu_i}^2 + \mu_i + \frac{{\mu_i}^2}{\theta}\right) - \left(\frac{(1-\pi_i)}{\left(1-\left(\dfrac{\theta}{\mu_i+\theta}\right)^{\theta}\right)} \times \mu_i\right)^2 .$$

ZANB não está encaixado em nenhum dos modelos referidos anteriormente.

4 - ANÁLISE PRELIMINAR DOS DADOS

4.1 - CARACTERÍSTICAS SOCIO-DEMOGRÁFICAS E COVARIÁVEIS

A exploração da possível distribuição da variável resposta, o número de ovos em 10 ml de urina, foi efectuada através das estatísticas usuais, sendo também apresentados uma caixa-de-bigodes e um gráfico de barras (construídos usando o R 2.12.0) com as frequências relativas para os vários valores da contagem (Figura 7).

A variável resposta foi estudada de duas formas distintas: com e sem os zeros. Inicialmente, no que diz respeito aos zeros, apenas estamos interessados na sua percentagem relativamente ao total da amostra, se estes podem ter apenas uma ou várias proveniências e se estas origens são ou não distintas das origens dos dados não nulos.

As covariáveis usadas neste estudo constam na Tabela 2. Para cada uma foi construído um gráfico de barras no Programa SPSS Statistics 19.0 onde, por cada nível da covariável, é apresentada a percentagem de indivíduos sem ovos na urina, com contagens entre 1 e 50 e com mais de 50 ovos (classificação sugerida pela WHO). Nas figuras 8 a 18 são apresentadas duas caixas-de-bigodes, uma com a totalidade dos dados e outra onde foram omitidos os possíveis *outliers*, obtidas no programa R.

Tabela 2 – Covariáveis e respectivos níveis.

Covariáveis	**Valores e níveis**
Género	0 – Feminino; 1 – Masculino
Idade	Anos
Água canalizada	0 – Tem; 1 – Não tem
Localização do WC	0 – Fora de casa; 1 – Dentro de casa
Contacto com água	1 – Rio; 2 – Lagoa; 3 – Tanque
Conhecimento da doença	0 – Conhece; 1 – Desconhece
Hematúria MAC (presença de sangue na urina)	0 – Negativo 1 – Positivo
Profissão	1 – Funcionário Público; 2- Estudante; 3 – Agricultor; 4 – Trab. doméstico; 5 – Outras actividades
Motivo de contacto com água	1 – Buscar água; 2 – Pescar; 3 – Lavar roupa; 4 – Higiene pessoal; 5 – Nadar
Província (onde o indivíduo vive)	1 – Luanda; 2 – Bengo; 3 – Kwanza Sul
Naturalidade	1 – Luanda, Bengo; 2 - Bié, Huambo, Moxico; 3 – Norte; 4 – Sul

Para comparar o número mediano de ovos em covariáveis com apenas dois níveis (género, água canalizada, localização do WC, conhecimento da doença e hematúria) aplicou-se o teste de Mann-Whitney. Para as covariáveis com mais de dois níveis - idade (após categorização em cinco classes), contacto com água, profissão, motivo, província e naturalidade - utilizou-se o teste de Kruskal-Wallis. O teste de Spearman foi usado para averiguar se existe relação entre as variáveis número de ovos em 10 ml de urina e a idade. Estes testes foram realizados no software SPSS Statistics 19.0.

4.2 - VARIÁVEL DE INTERESSE (OVOS10ML)

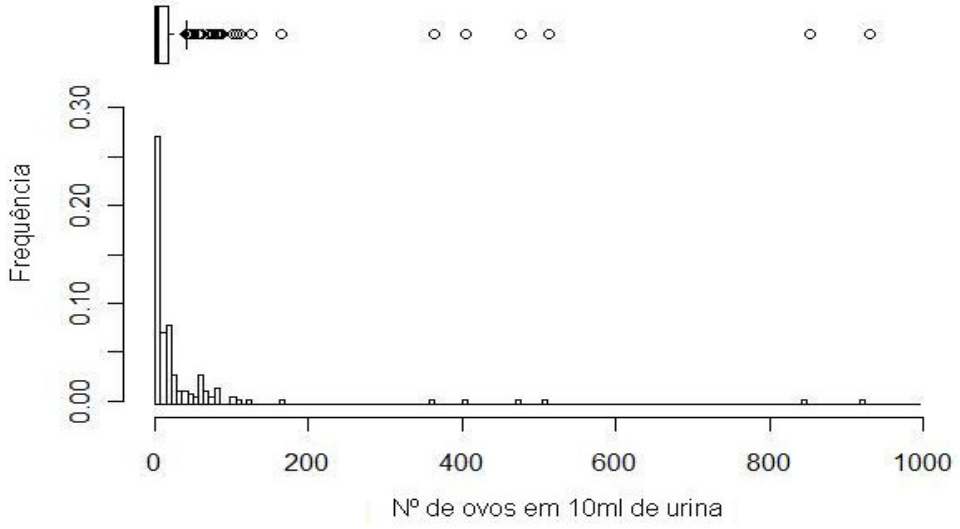
O número de ovos em 10 ml de urina é uma variável que toma valores inteiros não negativos logo espera-se que as distribuições de Poisson e a binomial negativa sejam apropriadas para proceder à sua modelação.

A amostra é formada por 300 indivíduos, dos quais 85 não apresentam ovos no exame de 10 ml de urina, o que corresponde a aproximadamente 28,3% do total. Os zeros têm duas origens possíveis: indivíduos que não estão infectados e indivíduos que por alguma razão estão infectados, mas não apresentam ovos no exame de urina.

Os valores não nulos são 215, 182 dos quais apresentam contagens de ovos entre 0 e 51 (60,7% dos indivíduos da amostra), 22 entre 50 e 100 (7,3% dos indivíduos da amostra) e 11 apresentam contagens maiores ou iguais a 100 (3,7% dos indivíduos da amostra).

O um gráfico de barras e uma caixa-de-bigodes (Figura 7) relativos à variável número de ovos em 10 ml de urina que revelam um grande número de possíveis *outliers*.

Figura 7 - Caixa-de-Bigodes e gráfico de barras da variável número de ovos em 10 ml de urina.



As Tabelas 3 e 4 foram obtidas através do software R. Ambas apresentam as medidas de localização e de dispersão mais usuais: a primeira para a totalidade dos valores da variável resposta (número de ovos em 10 ml de urina) e a segunda para os valores não nulos da mesma variável.

Tabela 3 – Estatísticas da variável número de ovos em 10 ml de urina.

Mínimo	1º quartil	Mediana	3º quartil	Máximo	Média	D. padrão
0	0	2	16	925	24.25	89.306

Tabela 4 - Estatísticas dos valores não nulos da variável n.º de ovos em 10 ml de urina.

Mínimo	1º quartil	Mediana	3º quartil	Máximo	Média	D. padrão
1	2	7	21	925	33.84	104.0069

Existem várias causas para a sobre-dispersão, onde as principais são excesso de zeros, heterogeneidade não controlada na amostra e dependência temporal (Rose et al., 2006).

Como há aproximadamente 28,3% de zeros na amostra, dever-se-á ter em conta a possível existência de sobre-dispersão originada por excesso de zeros.

Nas Tabelas 3 e 4 observa-se que a média é sempre muito inferior ao desvio padrão (e obviamente à variância). A razão entre a variância e a média é de aproximadamente 328,84 para a totalidade dos dados e aproximadamente 319,65 para os valores diferentes de zero. Isto sugere a existência de sobre-dispersão originada por heterogeneidade não controlada na amostra.

Quanto à dependência temporal, neste estudo, esta não parece ter sido uma causa que tenha afectado os dados pois cada indivíduo foi examinado apenas uma vez e, apesar dos dados não terem sido todos recolhidos no mesmo instante, o tempo entre as recolhas foi curto e não existe nenhuma evidência que tal possa afectar a nossa variável de estudo.

Existem diversas formas de estudar a existência de sobre-dispersão sugeridas por Rose et al. (2006). Pode-se investigar a sobre-dispersão causada pela heterogeneidade não controlada comparando os modelos GLM Poisson com o GLM binomial negativa e o ZAP com o ZANB recorrendo ao teste de razão de verosimilhanças (Self e Liang, 1987). Para comparar os modelos GLM com os modelos ZIP e ZINB pode-se recorrer a um teste score (Xiang et al., 2007).

Rose et al. (2006) também sugerem que se verifique se a sobre-dispersão é causada por excesso de zeros, fazendo a comparação do modelo GLM Poisson com os modelos ZIP e ZAP e do GLM binomial negativa com os ZINB e ZANB através do teste de Voung (Voung, 1989).

Na subsecção 5.2.2 será usado o teste de Voung para estudar os modelos GLM binomial negativa, ZINB e ZANB. As restantes sugestões não foram implementadas pois considerou-se que a comparação dos AIC e BIC, análise de gráficos e teste de Voung seria suficiente.

4.3 - RELAÇÃO ENTRE ALGUMAS COVARIÁVEIS E O NÚMERO DE OVOS

Tabela 5 – Tabelas de contingência para a categorização sugerida pela WHO (Organização Mundial de Saúde). Valores p dos testes para o número de ovos em 10 ml de urina segundo os níveis das covariáveis.

Número de ovos	0	1 - 49	>=50	Teste	Valor p
Género				Mann-Whitney	0,354
Feminino	45 (27,1)	100 (60,2)	21 (12,7)		
Masculino	40 (29,9)	82 (61,2)	12 (9,0)		
Idade (anos)	85 (28,3)	182 (60,7)	33 (11,0)	Kruskal-Wallis	0,609
				Spearman : r_s=-0,098	0,090
15 – 24	36 (27,3)	77 (58,3)	19 (14,4)		
25 – 34	18 (26,5)	42 (61,8)	8 (11,8)		
35 – 44	19 (38,0)	29 (58,0)	2 (4,0)		
45 – 54	7 (22,6)	23 (74,2)	1 (3,2)		
>=55	5 (26,3)	11 (57,9)	3 (11,0)		
Água canalizada				Mann-Whitney	0,991
Tem	7 (33,3)	12 (57,1)	2 (9,5)		
Não tem	78 (28,0)	170 (60,9)	31 (11,1)		
Localização do WC				Mann-Whitney	0,628
Dentro de casa	19 (27,9)	44 (64,7)	5 (7,4)		
Fora de casa	66 (28,4)	138 (59,5)	28 (12,1)		
Contacto com água				Kruskal-Wallis	<0,001
Rio	37 (20,7)	119 (66,5)	23 (12,8)		
Lagoa	17 (44,7)	17 (44,7)	4 (10,5)		
Tanque	31 (37,3)	46 (55,4)	6 (7,2)		
Conhece da doença				Mann-Whitney	0,199
Sim	19 (33,9)	32 (57,1)	5 (8,9)		
Não	66 (27,0)	150 (61,5)	28 (11,5)		
Hematúria MAC				Mann-Whitney	<0,001
Negativa	77 (32,4)	144 (60,5)	17 (7,1)		
Positiva	8 (12,9)	38 (61,3)	16 (25,8)		
Profissão				Kruskal-Wallis	0,460
Funcionário Público	4 (21,1)	13 (68,4)	2 (10,5)		
Estudante	11 (23,9)	25 (54,3)	10 (21,7)		
Agricultor	31 (27,7)	73 (65,2)	8 (7,1)		
Trabalhador doméstico.	24 (28,6)	52 (61,9)	8 (9,5)		
Outras actividades	15 (38,5)	19 (48,7)	5 (12,8)		
Motivo de cont. água				Kruskal-Wallis	0,442
Buscar água	57 (31,3)	104 (57,1)	21 (11,5)		
Higiene pessoal	13 (22,0)	40 (67,8)	6 (10,2)		
Lavar roupa	6 (13,6)	32 (72,7)	6 (13,6)		
Pescar	14 (24,1)	41 (70,7)	3 (5,2)		
Nadar	11 (29,7)	21 (56,8)	5 (13,5)		
Província				Kruskal-Wallis	0,267
Luanda	47 (32,2)	85 (58,2)	14 (9,6)		
Bengo	24 (27,9)	52 (60,5)	10 (11,6)		
Kwanza Sul	14 (20,6)	45 (66,2)	9 (13,2)		
Naturalidade					
Luanda Bengo	34 (25,6)	80 (60,2)	19 (14,3)	Kruskal-Wallis	0,043
Bié Huambo Moxico	31 (36,5)	48 (56,5)	6 (7,1)		
Norte	17 (24,6)	46 (66,7)	6 (8,7)		
Sul	3 (23,1)	8 (61,5)	2 (15,4)		

Recorrendo ao software SPSS Statistics 19 construiu-se a Tabela 5 onde se apresentam os resultados da análise das covariáveis e da sua relação com a variável resposta através de tabelas de contingência (frequência absoluta

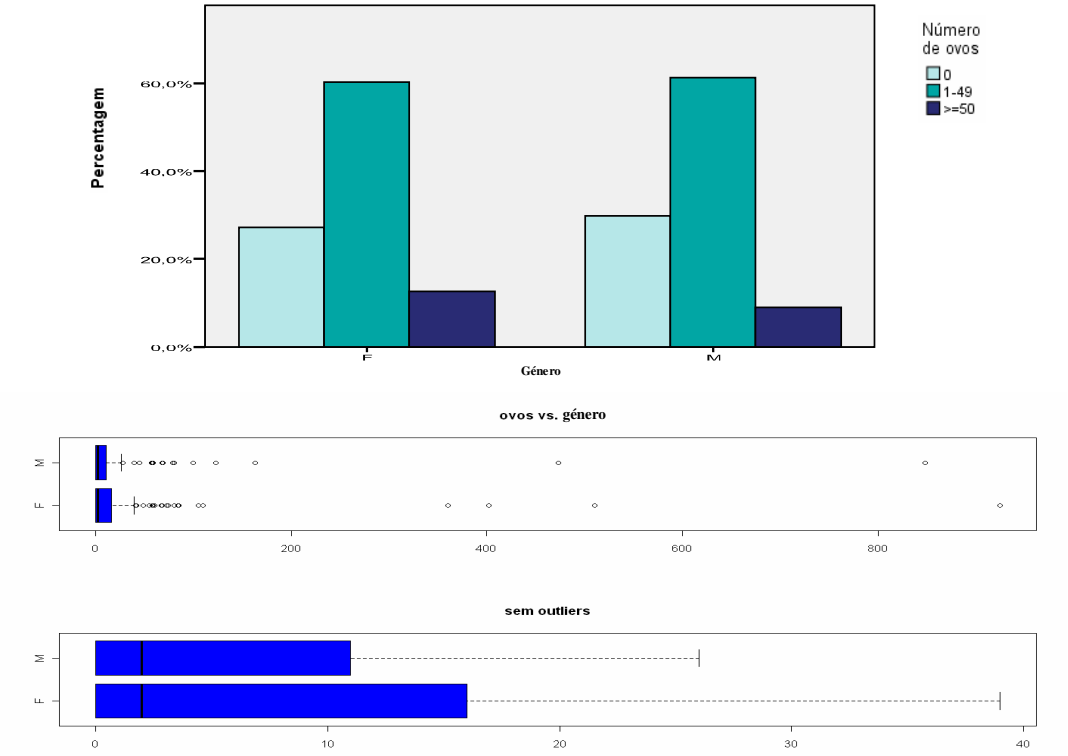
seguida da percentagem relativamente à linha). São também apresentados os valores p dos testes realizados. Nesta tabela, os valores da variável resposta (n.º de ovos em 10 ml de urina) estão agrupados em três níveis (0, 1-49 e >=50). Os testes apresentados foram realizados com esta variável não agrupada.

Os testes sugerem que existem diferenças significativas quanto ao número de ovos por 10 ml de urina nos vários níveis das covariáveis hematúria MAC, contacto com a água e naturalidade. Repare-se que o valor p das duas primeiras covariáveis referidas é <0,001.

A naturalidade apresenta um valor p de 0,043, próximo de 0,05. Optou-se por decidir quanto a diferenças das contagens de ovos nos vários níveis desta covariável depois de um estudo mais aprofundado.

Seguidamente é efectuada uma análise de cada covariável pela mesma ordem apresentada na Tabela 5.

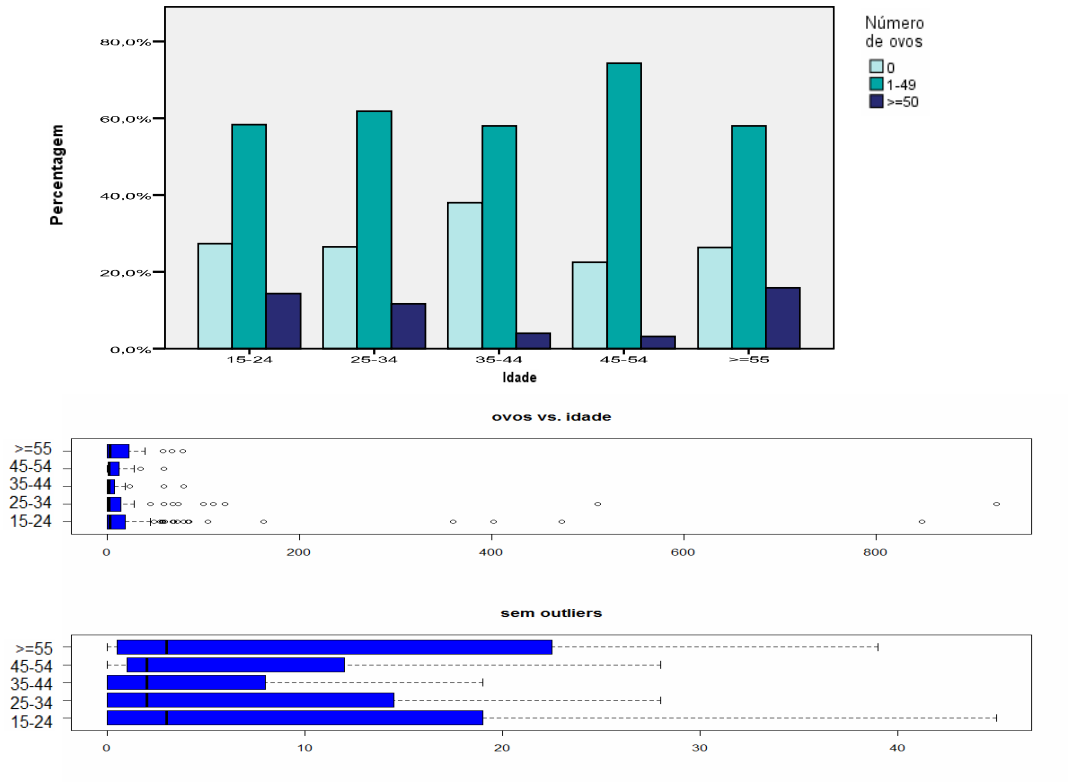
Figura 8 – Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável género.



Os dados sugerem que não existem diferenças significativas entre o número de ovos na urina nos dois sexos. Os valores p obtidos através dos testes de proporções e de Mann-Whitney foram aproximadamente 0,600 e 0,354, respectivamente. Ambos sugerem não existir uma diferença significativa entre os dois sexos.

Na literatura sobre a Schistosomose é referido que a grande maioria das infecções ocorre entre os 4 e 14 anos (WHO, 2002). Nestas idades as crianças entram frequentemente em contacto com a água, não só nas suas brincadeiras, mas também porque estão encarregues de tarefas ligadas a este recurso, como por exemplo, o seu transporte ou da lavagem da roupa.

Figura 9 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável idade.



Os dados só apresentam indivíduos a partir dos 15 anos inclusive, pois foram recolhidos por investigadores que queriam estudar o efeito da Schistosomose na população com, pelo menos, 15 anos. Note-se que a WHO dá ênfase ao controlo da doença nos indivíduos com menos de 15 anos mas também chama à atenção para o facto de esta afectar indivíduos mais velhos (WHO, 2002).

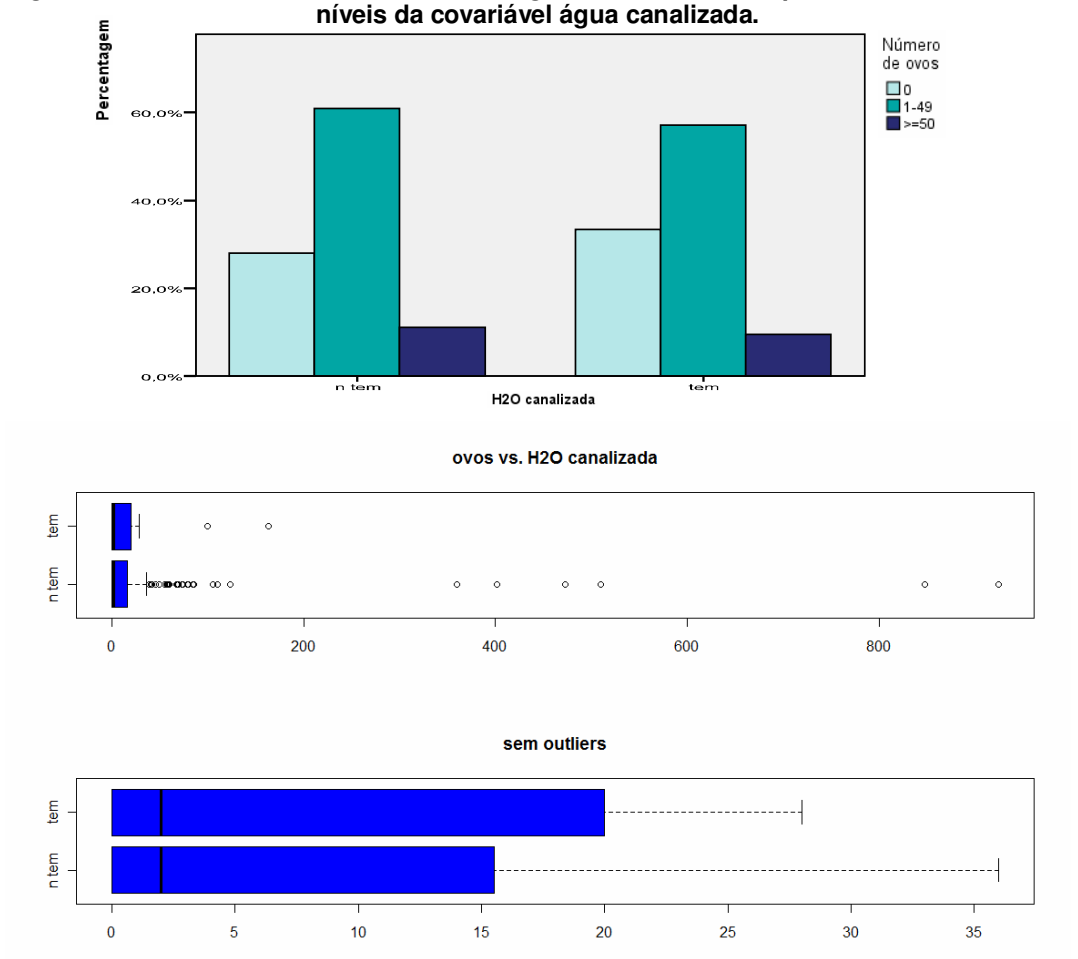
Ao realizar o teste de Kruskal-Wallis obteve-se um valor p de 0,609 e consequentemente não se rejeitou a hipótese nula de igualdade das medianas do número de ovos por 10 ml de urina nas várias classes etárias. Na Tabela 6 observa-se que as medianas do número de ovos são semelhantes em todas as classes da covariável idade.

Tabela 6 – Número mediano de ovos em 10 ml de urina em cada escalão etário.

Idade	Nº de ovos na urina	
	N	Medianas das classes
15-24	132	3
25-34	68	2
35-44	50	2
45-54	31	2
>=55	19	3
Total	300	2

O coeficiente de correlação de Spearman foi -0,098, sendo o valor p obtido de 0,090. Tanto o coeficiente como o valor p obtidos sugerem que não existe correlação significativa entre a covariável idade e o número de ovos existentes na urina.

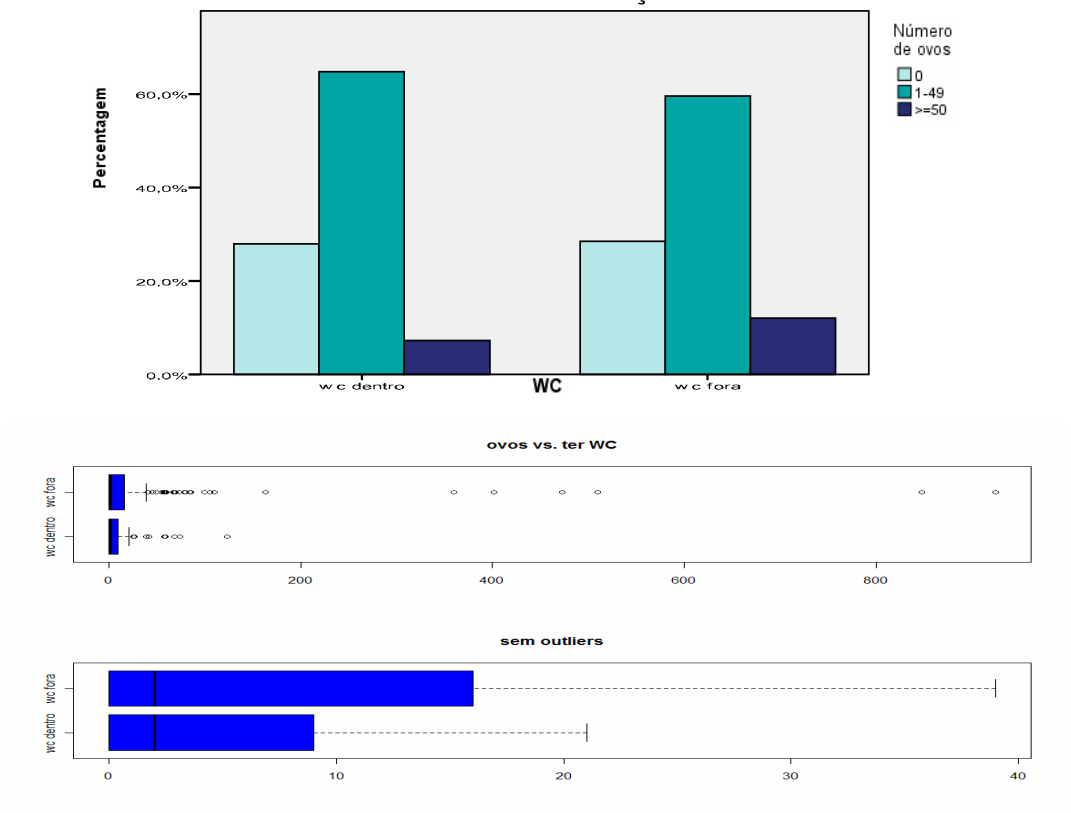
Figura 10 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável água canalizada.



Observa-se que uma grande percentagem dos inquiridos afirma não ter água canalizada (93%). Estes também apresentam valores mais elevados de carga parasitária e têm maior percentagem de infecções (72% contra 66,7% de quem

tem água canalizada). Estes resultados podem ser explicados por a infecção acontecer através do contacto com águas que contêm moluscos e cercárias.

Figura 11 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável localização do WC.

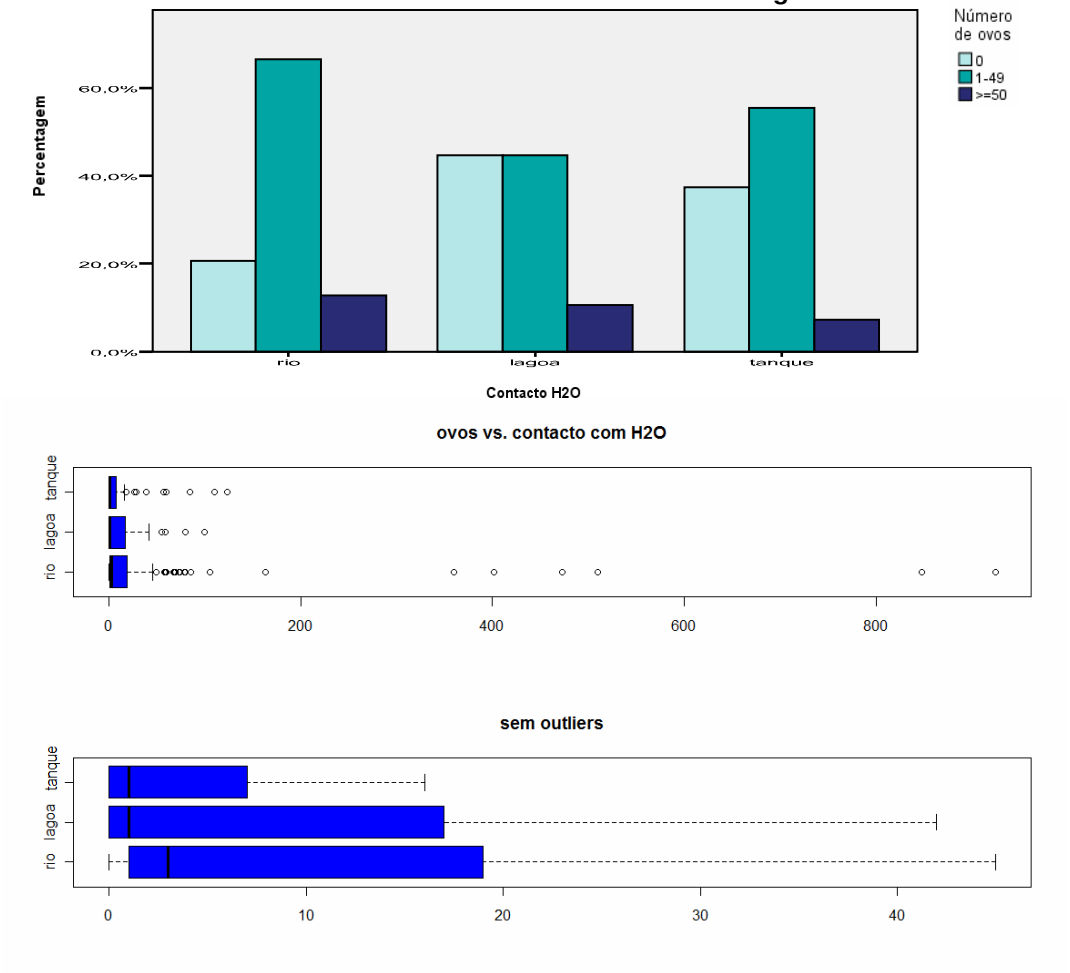


Uma considerável percentagem da amostra afirma não ter WC (77,3%). Segundo o gráfico de barras da figura anterior, estes indivíduos parecem apresentar um maior número de ovos na urina do que aqueles que referem ter WC dentro de casa. Porém, o teste de Mann-Whitney produziu um valor p de 0,628, logo não se rejeita a hipótese nula de que as pessoas com WC dentro (têm WC) e fora (não têm WC) de casa provenham de populações com medianas idênticas. Quanto à percentagem de infecções, esta não é significativamente diferente nos dois grupos de indivíduos (72,1% com WC e 71,6% sem WC).

Uma considerável percentagem dos inquiridos afirma terem contacto com água de rio (59,7%). Destes, 79,3% apresentam ovos na urina (47,4% do total) e valores elevados de carga parasitária.

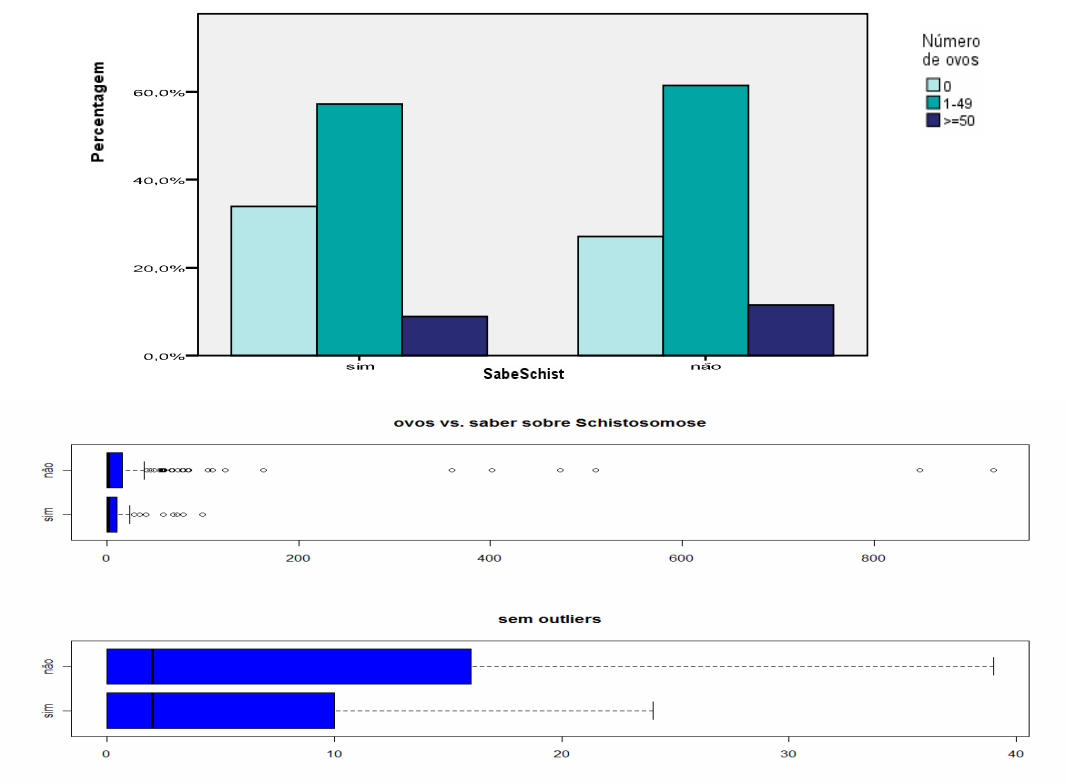
A percentagem de indivíduos infectados que têm contacto com a água de lagoas é de 55.2%. Para os que contactam com água de tanques, a percentagem de infecção registada na amostra é de 62,6%.

Figura 12 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável local de contacto com água.



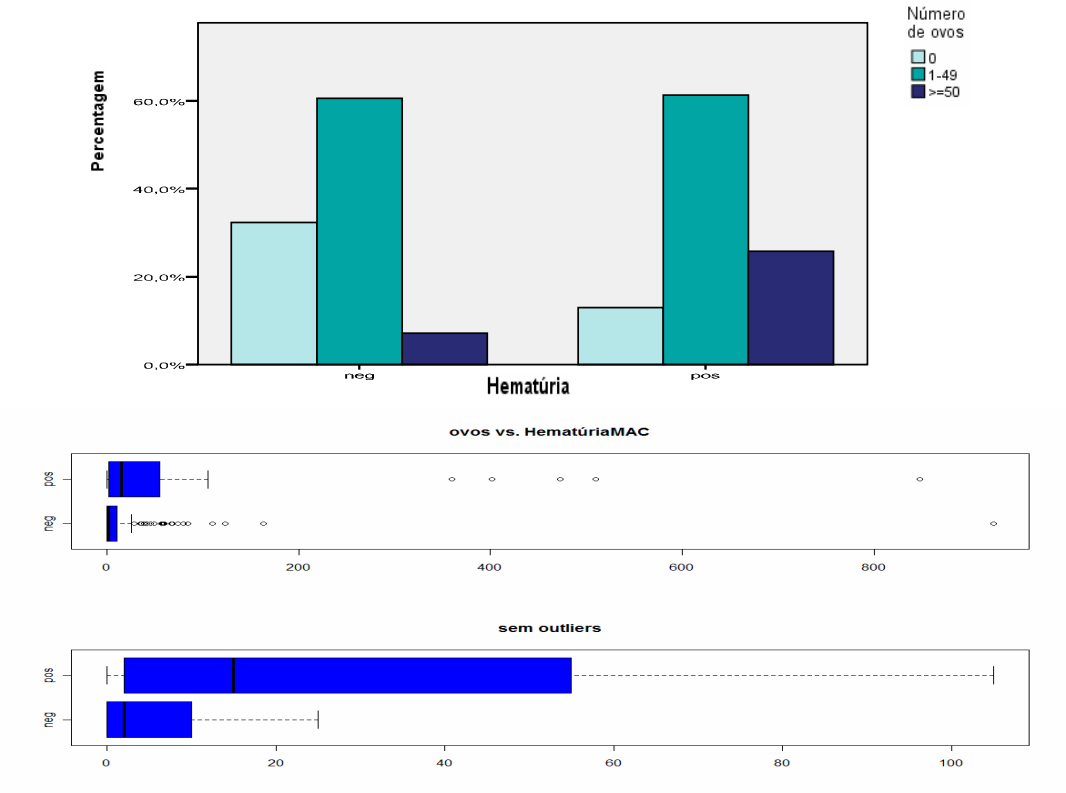
Não existem diferenças significativas quanto à percentagem de infectados entre os indivíduos que afirmam ter contacto com água em lagoas e os que dizem ter contacto com água num tanque. Os primeiros apresentam uma percentagem de infectados ligeiramente inferior, o que poderá parecer estranho. Supõe-se que a água de um tanque seja proveniente da chuva e que esta não tenha contacto com o vector da doença. Porém, os investigadores que recolheram os dados afirmaram que muitas vezes os tanques eram enchidos com água do rio (Figueiredo, 2008). Isto pode explicar a percentagem de infectados entre os utilizadores da água de tanques.

Figura 13 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável saber sobre Schistosomose.



A maioria dos indivíduos da amostra afirma não ter informação sobre a Schistosomose (81,3%) e estes apresentam contagens de ovos na urina superiores aos indivíduos que afirmam estarem informados sobre a doença. A percentagem de infectados em indivíduos que afirmam ter conhecimento da Schistosomose (66,1%) é ligeiramente inferior à percentagem de infectados que afirmam não dispor de informação sobre a doença (73,1%).

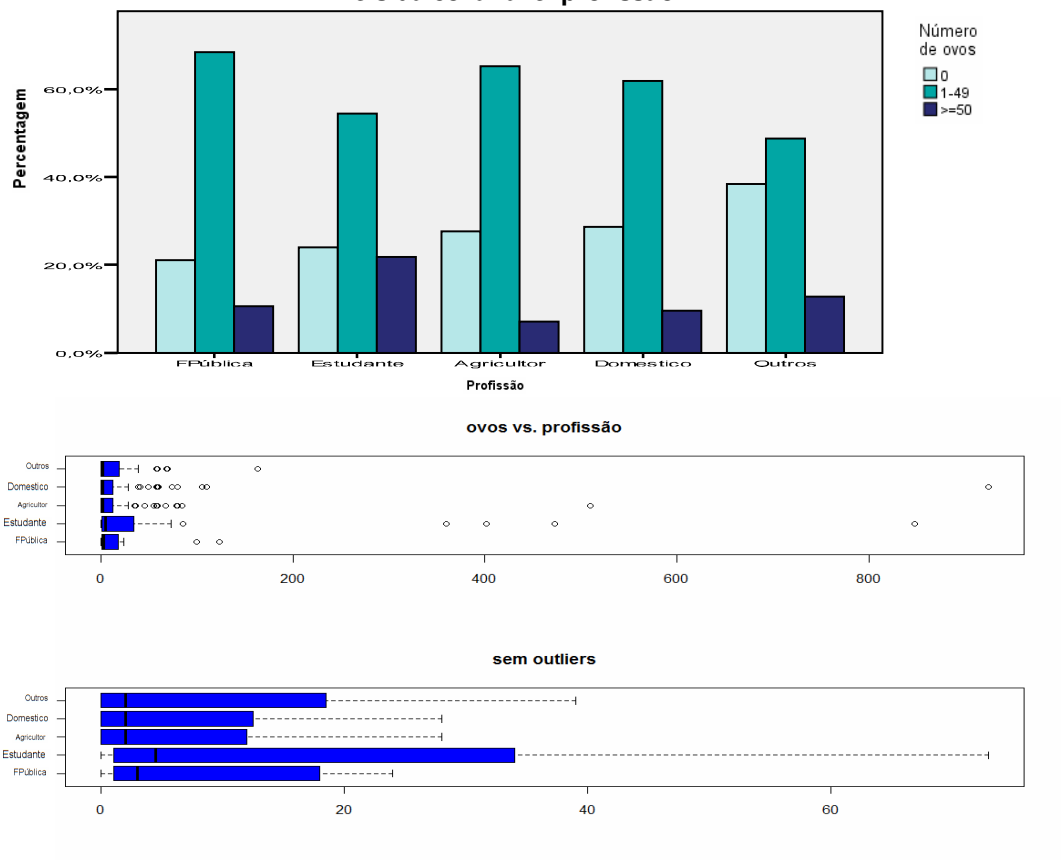
Figura 14 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável hematúria MAC.



Na variável Hematúria MAC está registada a presença de sangue na urina. Esta é a segunda covariável (ver Tabela 5) cujo teste sugere uma relação com a variável resposta (Mann-Whitney, valor p <0,001). É de notar que aproximadamente 90% dos zeros aparecem em indivíduos que não apresentam sangue na urina. Isto pode ser explicado pelo facto dos ovos de *Schistosoma haematobium* possuírem um espigão que causa danos nos tecidos no seu percurso até chegar à urina (Figueiredo, 2008).

Quanto à profissão, os indivíduos que se dedicam à agricultura (37,3%) e trabalhos domésticos (28%) apresentam percentagens de infecção inferiores aos funcionários públicos e estudantes. Isto é surpreendente pois é de esperar que trabalhadores domésticos, em especial os agricultores tenham contacto mais frequente com águas não tratadas.

Figura 15 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável profissão.



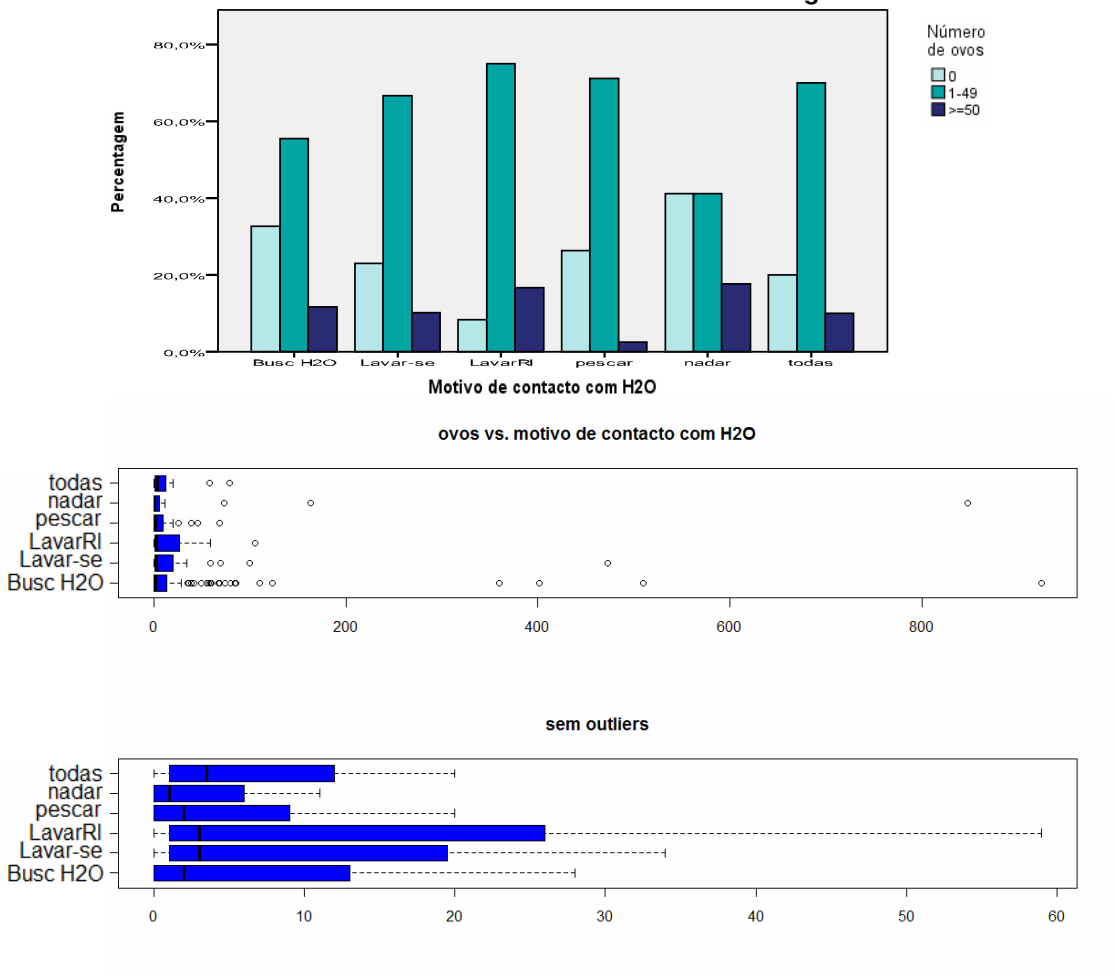
Como se pode ver pelas caixas-de-bigodes apresentadas na Figura 11, os estudantes têm um desvio padrão de 156,7, muito superior ao das restantes profissões. Foi realizado um teste de Levene para confirmar esta hipótese mas o valor p do teste baseado na média é <0,001 e o valor p baseado na mediana é de 0,065. Considerando apenas os indivíduos infectados, obteve-se o mesmo valor p (<0,001) para o teste de Levene baseado na média e 0,093 para o teste baseado na mediana.

Devido à existência de valores muito grandes na amostra, deve-se seguir a indicação do segundo valor p e considerar que não existem diferenças significativas nos valores da variância nas diferentes profissões.

Tabela 7 – Desvios padrão da variável resposta por nível da covariável água canalizada.

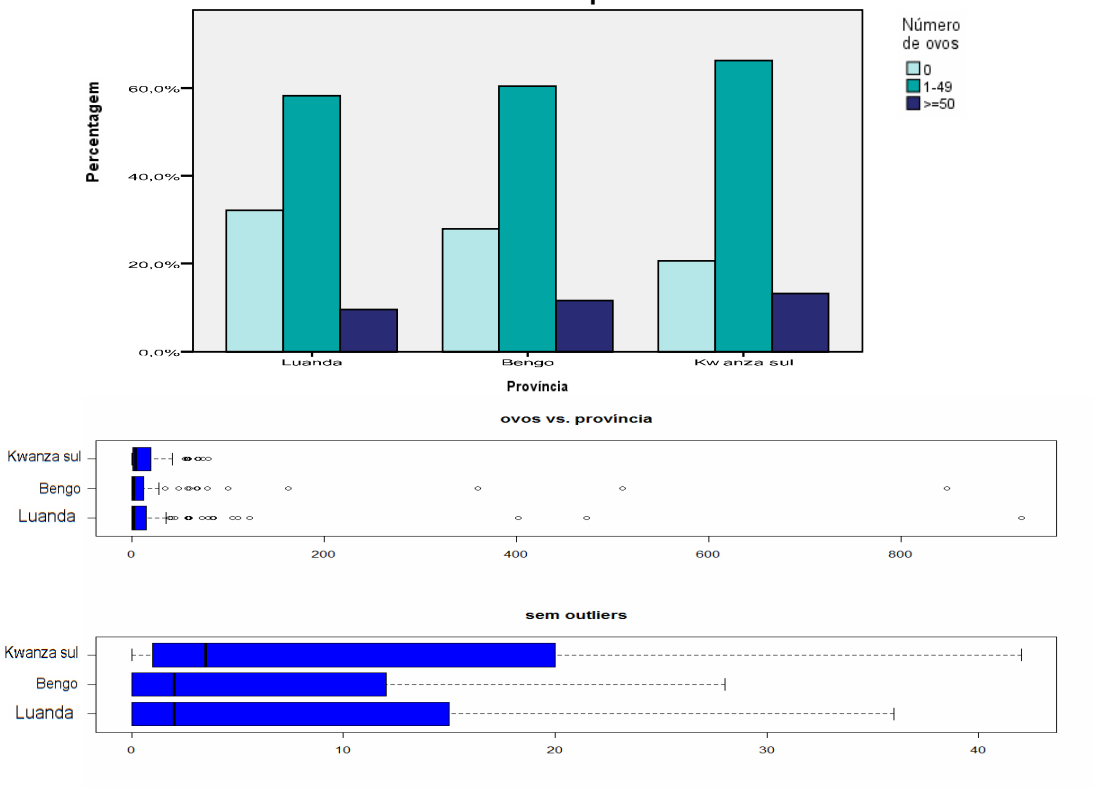
Profissão	Agricultor	F. Publica	Doméstico	Estudante	Outros
D. padrão	50,55	34,12	102,21	156,72	31,29

Figura 16 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável motivo de contacto com água.



Ir buscar água é o motivo que grande parte dos indivíduos da amostra evoca para explicar o contacto com água (54%), o que é coerente com a grande quantidade de inquiridos que dizem não ter água canalizada. Quem afirma lavar roupa apresenta uma grande percentagem de infecções (86,3%), sendo superior aos outros níveis desta covariável.

Figura 17 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável província.



Quanto à província, a grande parte da amostra é proveniente de Luanda (48,7%). A percentagem de infectados é semelhante nas três províncias avaliadas sendo ligeiramente mais pequena em Luanda (Luanda com 67,8%, Bengo com 72,1% e Kwanza sul com 79,4%).

Como foi referido por Figueiredo (2008), muitos dos indivíduos a residir nas zonas de recolha da amostra provêm de outras províncias nas quais a Schistosomose urinária é endémica. Logo a naturalidade é uma covariável importante a considerar.

Foram registadas as províncias de origem dos indivíduos da amostra. Estes resultados são apresentados na Tabela 8.

Tabela 8 – Frequência absoluta da covariável naturalidade e respectiva percentagem.

	Freq. abs.	Percentagem
Bengo	74	24,7%
Benguela	12	4,0%
Bié	38	12,7%
Cabinda	1	0,3%
Huambo	44	14,7%
Huila	1	0,3%
K.Norte	9	3,0%
K.Sul	25	8,3%
Luanda	59	19,7%
Malange	23	7,7%
Moxico	3	1,0%
Uige	9	3,0%
Zaire	2	0,7%
Total	300	100,0%

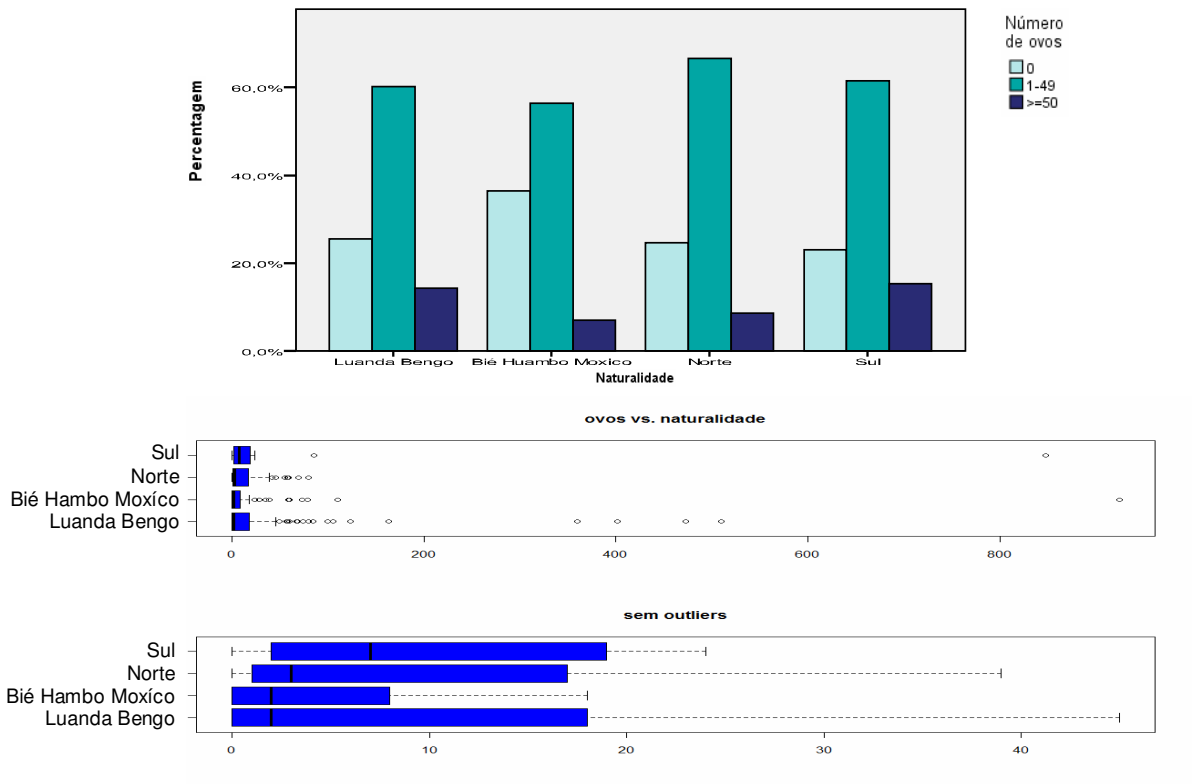
Existem províncias com baixas frequências de inquiridos e assim foi necessário agrupar a covariável. O critério usado para criar quatro grupos foi a proximidade geográfica. Luanda e Bengo foram agrupadas com o número 1, Bié, Huambo e Moxico com o número 2. No norte (3) foram agregados os naturais de Cabinda, Zaire, Uige, Kuanza Norte, Kuanza Sul e Malange. A Sul, agruparam-se as províncias de Benguela e Huila.

É de esperar que a maior parte dos indivíduos da amostra tenha nascido em Luanda, Bengo ou Kwanza Sul pois foi nestas províncias onde se realizou a recolha dos dados.

Relativamente à variável em estudo, a percentagem de zeros é semelhante nos quatro níveis desta covariável (25,6% para Luanda e Bengo, 36,5% para Bié, Huambo e Moxico, 24,6% para o Norte e 23,1% para o Sul).

Para a totalidade dos indivíduos, verifica-se que o Sul apresenta um número médio de ovos superior aos restantes níveis (27,9, 19,3, 13,0 e 78,8). O mesmo se passa relativamente à mediana (2, 2 ,3 e 7). Isto pode ser explicado por apenas treze indivíduos da amostra afirmarem terem nascido nestas províncias. Destes, unicamente 23,1% não apresentam ovos na urina e um destes registos é 848 ovos/10 ml.

Figura 18 - Gráfico de barras e Caixas-de-Bigodes da variável resposta relativamente aos níveis da covariável naturalidade.



A Figura 18 apresenta-se a percentagem de indivíduos sem ovos na urina, com um número de ovos entre 1 e 50 e com mais de 50 ovos. No Buié, Humbo e Moxico, a frequência de indivíduos sem registo de ovos na urina parece ser ligeiramente superior ao observado nos outros agrupamentos de regiões. Essa zona de Angola também apresenta menos indivíduos na categoria que representa maior gravidade (≥50 ovos). Apesar disso, na Figura 18, não se verificam grandes diferenças entre os quatro níveis da covariável naturalidade.

Ao comparar o número de ovos em 10 ml de urina, através do teste de Kruskal-Wallis, conclui-se que, ao nível de significância de 5%, as medianas do número de ovos diferem segundo a zona de naturalidade (valor p = 0,043). Para os indivíduos infectados, através do mesmo teste, obteve-se um valor p = 0,158.

Quanto a covariáveis, é ainda de referir que não foi registado se os indivíduos estavam infectados com HIV. Seria uma covariável interessante de analisar pois existem na literatura evidências que sugerem que a presença do HIV influencia o número de ovos encontrados na urina dos pacientes (Karanja et al., 2005 and Mwanakasale et al.,2003 in Kallestrup, P. et al., 1999).

5 - APLICAÇÃO DOS MODELOS

5.1 – METODOLOGIA

A estimação dos parâmetros de cada modelo foi realizada recorrendo a várias funções e pacotes disponíveis para a linguagem R. Em todos os modelos, os parâmetros são estimados através de máxima verosimilhança e os valores iniciais são calculados através do algoritmo 'glm.fit'.

Na Tabela 9 estão especificados, para cada modelo, as funções e pacotes utilizados.

Tabela 9 – Função e pacote do R para estimar os parâmetros de cada modelo.

Modelo	Função	Pacote
GLM Poisson	glm	---
GLM binomial negativa	glm.nb	MASS
Poisson com excesso de zeros	zeroinfl	PSCL
Binomial Negativa com excesso de zeros	zeroinfl	PSCL
Modelo de duas partes com Poisson	hurdle	PSCL
Modelo de duas partes com binomial negativa	hurdle	PSCL

Para uma descrição mais detalhada destes pacotes pode-se consultar o manual do 'PSCL'. (http://cran.r-project.org/web/packages/pscl/pscl.pdf), do pacote 'MASS' (http://cran.r-project.org/web/packages/MASS/MASS.pdf) e do próprio R onde está descrita a função 'glm' (http://cran.r-project.org/manuals.html).

Para uma descrição de índole mais prática, consultar o artigo "Regression Models for Count Data in R," (Jackman et al., 2008) ou o capítulo 11 do livro "Mixed Effect Models and Extensions in Ecology with R" (Zuur et al., 2009).

Nestes textos o único ponto que necessita de clarificação é a obtenção dos valores preditos. A função "fitted", do pacote 'PCSL', fornece os valores do número de ovos na urina esperados para cada observação da amostra, ou seja, os valores de $E[Y_i]$.

Também é pretendido estimar o número de indivíduos com uma determinada quantidade de ovos na urina. Para isso calcula-se a probabilidade de obtermos um determinado valor de ovos com a função de probabilidade $f(y_i)$ de cada modelo (formulas (1), (5), (8), (9), (10) e (11)) e multiplica-se pelo número de indivíduos da amostra.

Os modelos são construídos inicialmente com todas as covariáveis presentes na Tabela 2. Como a idade e o género são covariáveis muito importantes do ponto de vista epidemiológico, estas são sempre consideradas em todos os modelos.

Quanto às restantes covariáveis, as que não são significativas (com valores p maiores que 0,1) são retiradas e o modelo é ajustado de novo. Este processo decorre até que os coeficientes de todas as covariáveis presentes no modelo sejam significativos, sendo um procedimento semelhante ao método *Stepwise*.

Em 5.2.1 os resultados das estimativas dos parâmetros são apresentados na Tabela 10, seguindo-se uma discussão sobre o seu significado, num contexto epidemiológico.

Para testar se os modelos com covariáveis produzem um melhor ajustamento que os modelos com apenas a ordenada na origem, foi usado um teste de qui-quadrado. Para realizar este teste foi usado o comando 'lrtest' do pacote lmtest da linguagem R. O manual deste pacote pode ser consultado em *http://cran.r-project.org/web/packages/lmtest/lmtest.pdf*.

Em 5.2.2 pretende-se saber qual o modelo que produz um melhor ajustamento aos dados. Para isso foram estudados dois tipos diferentes de dados ajustados: o número de indivíduos que apresenta uma determinada quantidade de ovos em 10 ml de urina e o número de ovos em 10 ml de urina para cada indivíduo da amostra.

Em 5.2.2.1 e 5.2.2.2 estes dois tipos de ajustamentos foram estudados. Para cada modelo apresenta-se um gráfico onde os valores ajustados são sobrepostos com os valores observados e uma tabela onde se acrescentam a estes valores do AIC (Akaike Information Criterion), do BIC (Bayes Information Criterion) e o número de parâmetros usado em cada modelo.

Autores como Rose et al. (2006) usam o teste de qui-quadrado de Pearson para comparar o ajustamento dos modelos. No caso em estudo temos muitos valores observados abaixo de cinco e consequentemente muitos valores ajustados na mesma situação. Assim, para usar-se este teste, seria necessário agrupar num grande número de classes e ao fazer isso, iria distorcer-se o modelo e diminui-se a potência do teste (Weiss, 2008). Optou-se por usar apenas o AIC e o BIC.

Ainda em 5.2.2.1, utilizou-se o teste de proximidade de Vuong (Vuong, 1989), implementado no pacote 'PSCL' do R, para tentar perceber a possível existência de alguma diferença significativa, quanto ao ajustamento aos dados, por parte dos modelos não encaixados (*not nested*), especificamente entre os modelos que utilizam a binomial negativa.

Em 5.2.2.2 comparou-se a média e o desvio padrão, dos valores ajustados para o número de ovos em cada indivíduo da amostra, com os valores observados da média e desvio padrão da variável de interesse. Isto não foi realizado para o outro tipo de valores ajustados pois, como o suporte das distribuições usadas é infinito, recorreu-se à criação de classes de equivalência para os valores mais elevados, o que distorceu o cálculo das estatísticas.

Para os dois tipos de valores ajustados foram calculados e analisados os resíduos. Existem vários tipos de resíduos: os de Pearson não normalizados, de Pearson normalizados e os de desvio (*Deviance*), entre outros. McCullagh and Neder (1989) in Zuur et al. (2009) afirmam que os resíduos de desvio são os mais indicados para dados com as características em estudo. Zuur et al. (2009) argumenta que não se procura normalidade nos resíduos, mas

tendências que assinalem falta de ajustamento dos modelos aos dados. Concordando com este argumento utilizaram-se os resíduos de Pearson normalizados para estudar todos os modelos relativamente ao número de ovos em cada indivíduo da amostra apesar de não serem o tipo de resíduos mais adequado para os modelos utilizados. Esta análise, com as devidas resalvas, encontra-se no Anexo 2.

Os resíduos de Pearson normalizados são obtidos através da seguinte fórmula:

$$\frac{y_i - \hat{y}_i}{\sqrt{\text{var}(Y_i)}} \qquad (12)$$

Onde y_i representa o valor observado na amostra para o *i*-ésimo indivíduo e $\hat{y}_i$ representa a estimativa de um modelo para o *i*-ésimo indivíduo. Para o ao número de ovos em 10 ml de urina de cada indivíduo da amostra, $\hat{y}_i = E[Y_i]$. $\sqrt{\text{var}(Y_i)}$ é o desvio padrão.

Os resíduos de Pearson não normalizados (*raw residuals*) definidos por:

$$y_i - \hat{y}_i \qquad (13)$$

podem também ser analisados como alternativa aos normalizados.

Para o número de indivíduos com determinada quantidade de ovos em 10 ml de urina, $\hat{y}_i = \sum_i f(y_i)$.

Zuur et al. (2009) também recorreram aos resíduos de Pearson para a análise de resíduos.

Os resíduos de desvio foram estudados para o número de ovos em 10 ml de urina de cada indivíduo da amostra nos modelos GLM Poisson e GLM binomial negativa (secção 5.2.2.2). Para os outros modelos deve ser possível calcular e estudar os resíduos de desvio. Porém, não foi encontrada nenhuma referência na literatura acerca da forma como estes se calculam. Acresce-se o facto de que limitações de tempo restringiram a profundidade com a qual a análise de resíduos foi realizada, podendo tal estudo ser realizado no futuro.

Os resíduos de desvio são definidos do seguinte modo:

$$\text{sinal}(y_i - \hat{y}_i)\sqrt{d_i}, i = 1, \ldots, n \qquad (14)$$

$\text{sinal}(y_i - \hat{y}_i) = 0$ se $y_i = \hat{y}_i$, $\text{sinal}(y_i - \hat{y}_i) = 1$ se $y_i > \hat{y}_i$ e $\text{sinal}(y_i - \hat{y}_i) = -1$ se $y_i < \hat{y}_i$, representando d_i a contribuição individual de i para a função desvio.

Para o modelo Poisson com função de ligação logit, tem-se $d_i = 2\left[y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) - y_i + \hat{y}_i \right]$ (Zuur et al, 2009).

Para o modelo binomial com função de ligação logit, tem-se $d_i = 2\left[y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + (1 - y_i)\ln\left(\frac{1-y_i}{1-\hat{y}_i}\right)\right]$ (Zuur et al, 2009).

A análise de resíduos para estes modelos difere bastante da análise realizada para os habituais modelos lineares. Como já foi referido anteriormente, a não normalidade dos resíduos não pode ser interpretada como indicação de um mau ajustamento. O grande número de zeros calculados por um processo separado dos restantes valores (nos modelos ZIP, ZINB, ZAP e ZANB) pode provocar uma tendência relativamente aos correspondentes resíduos. Por outro lado, a sobre-dispersão dos dados favorece o aparecimento de resíduos com valores, em módulo, muito elevados.

Zuur et al. (2009) sugerem que se construam gráficos dos resíduos versus os valores das covariáveis e também dos resíduos versus os valores ajustados. Estes autores afirmam que, para os modelos serem válidos, nos gráficos não deve ser perceptível nenhuma tendência.

Dupont (2002) sugere que para validar um modelo de regressão de Poisson, 95% dos resíduos estejam entre -2 e 2. Este critério também será utilizado nos restantes modelos.

Em todas as análises de resíduos são apresentados os valores da média dos resíduos, da percentagem de resíduos no intervalo (-2,2) e da percentagem de resíduos não negativos. São ainda apresentados gráficos dos resíduos versus os valores observados e versus os valores ajustados.

Apenas para o número de ovos em 10 ml de urina em cada indivíduo da amostra são apresentados os gráficos dos resíduos versus os valores das covariáveis e versus a ordem na amostra, como são sugeridos por Zuur et al. (2009). No que se refere ao outro conjunto de valores ajustados, o número de indivíduos que apresenta uma determinada quantidade de ovos em 10 ml de urina. Na análise deste tipo de valores ajustados é apresentado um histograma dos resíduos para analisar a sua distribuição, sem esperar que esta seja normal.

Para terminar calculou-se, para cada modelo, a probabilidade de apresentar ovos na urina para os nove indivíduos que não foram identificados no exame inicial, mas que foram detectados mais tarde numa biópsia.

5.2 – RESULTADOS E DISCUSSÃO

5.2.1 – COEFICIENTES E COVARIÁVEIS

Na Tabela 10 as covariáveis retiradas correspondem aos espaços assinalados com (---). Para sinalizar se um nível de uma covariável é muito significativo (valor p < 0,001) usa-se a notação ***. Para assinalar p entre 0,001 e 0,01 recorre-se a **. Valores p entre 0,01e 0,05 são identificados por *. Se p estiver entre 0,05 e 0,1 usa-se º. Estes valores p são obtidos através do teste de Wald. A hipótese nula deste teste pressupõe que o parâmetro testado é zero contra a hipótese alternativa do parâmetro ser diferente de zero.

Tabela 10 – Coeficientes atribuídos às covariáveis pelos vários modelos.

Covars vs Modelos	GLM Poisson	GLM Binomial Negativa	ZIP (Zeros)	ZIP (Dist.)	ZINB (Zeros)	ZINB (Dist.)	ZAP (Zeros)	ZAP (Dist. truncada)	ZANB (Zeros)	ZANB (Dist. Truncada)
Ordenada na origem	1,651 ***	3,581 ***	-0,950 *	1,333 ***	-301,97 **	2,602 ***	0,965 *	1,328 ***	0,965 *	0,568
Género(F)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Género (M)	0,458 ***	-0,022	0,162	0,569 ***	-65,752	0,146	-0,200	0,573 ***	-0,200	-0,406
Idade	-0,013 ***	-0,013	0,030	-0,006 ***	4,897	-0,005	-0,005	-0,006	-0,005	-0,029 *
Água Canal. (Sim)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Água Canal. (Não)	---	---	---	---	---	---	---	---	---	---
WC (Dentro)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
WC (Fora)	0,776 ***	---	---	0,920 ***	---	0,670 *	---	0,884 ***	---	0,746 º
Cont. com água (Rio)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Cont. com água (Lagoa)	-0,402 ***	-0,651 º	1,744 ***	-0,102	---	-0,367	-1,801 ***	-0,108 º	-1,801 ***	---
Cont. com água (Tanque)	-1,259 ***	-0,983 ***	0,358	-1,196 ***	---	-1,193 ***	-0.368	-1,199 ***	-0,368	---
Conhec. da doença (Sim)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Conhec. da doença (Não)	1,218 ***	--	---	1,258 ***	---	1,112 ***	---	1,360 ***	---	---
Hematúria (Negativo)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Hematúria (Positivo)	1,322 ***	1,296 ***	-1,163 **	1,183 ***	---	1,276 ***	1,213 **	1,184 ***	1,213 **	1,283 **
Prof. (Func. Público)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Prof. (Estudante)	0,106	---	---	0,335 ***	---	-0,811	---	0,338 ***	---	---
Prof. (Agricultor)	-0,441 ***	---	---	-0,224 **	---	-1,072 *	---	-0,221 **	---	---
Prof. (Trab. doméstico)	0,240 **	---	---	0,607 ***	---	-0,609	---	0,612 ***	---	---
Prof. (Outras)	-0,139 º	---	---	0,186 *	---	-0,804	---	0,194 **	---	---
Motivo (Buscar água)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Motivo(Higiene Pessoal)	-0,300 ***	-0,442	---	-0,456 ***	---	-0,582 º	---	-0,460 ***	---	---
Motivo (Lavar roupa)	-0,878 ***	-0,662 º	---	-1,222 ***	---	-0,342	---	-1,227 ***	---	---
Motivo (Pescar)	-1,263 ***	-0,742*	---	-1,448 ***	---	-0,898 *	---	-1,457 ***	---	---
Motivo (Nadar)	0,053	0,342	---	0,349 ***	---	0,132	---	0,345 ***	---	---
Província (Luanda)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Província (Bengo)	0,223 ***	---	-0,377	0,012	---	---	0,391	-0,010	0,391	---
Província (Kwanza Sul)	0,454 ***	---	-1,321 **	0,290 ***	---	---	1,324 **	-0,287 ***	1,324 **	---
Naturalidade (Luanda Bengo)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)	(a)
Naturalidade (Bié Huambo Moxico)	0,311 ***	---	---	0,528 ***	---	---	---	0,530 ***	---	---
Naturalidade (Norte)	-0,477 ***	---	---	-0,612 ***	---	---	---	-0,615 ***	---	---
Naturalidade (Sul)	0,463 ***	---	---	0,183 ***	---	---	---	0,183 ***	---	---

*** Valor p < 0,001 ** 0,001 < Valor P < 0,01 * 0,01 < Valor p < 0,05 º 0,05 < Valor p < 0,1 (a) Nível de referência

Os níveis de referência são níveis de covariáveis que, como o nome indica, servem de termo de comparação para os outros níveis. Como não têm coeficiente estão assinalados com (a).

Observa-se que os modelos com a distribuição de Poisson usam um maior número de covariáveis que os modelos com a binomial negativa. Por exemplo, considere-se a parte da distribuição truncada dos modelos ZAP e ZANB. O primeiro usa quase todas as covariáveis. O segundo usa apenas as covariáveis hematúria, localização do WC, idade e género. Uma consulta à Tabela 10 confirma que os modelos com binomial negativa são muito mais parcimoniosos.

Isto poderá dever-se à presença de sobredispersão que provoca uma subestimação dos desvios padrão dos parâmetros estimados por máxima verosimilhança nos modelos com Poisson (Navarro et al., 2001; Cameron and Trivedi, 1998 in Rose et al., 2006).

Como a significância de um parâmetro é dada pelo teste de Wald, e como a estatística deste teste depende do desvio padrão[1], a sobre-dispersão poderá estar a influenciar o nível de significância das estimativas das covariáveis nos modelos com distribuição de Poisson.

A título de exemplo foram criados dois modelos GLM Poisson e GLM binomial negativa com as mesmas covariáveis para comparar-se os coeficientes e os seus desvios padrão.

As covariáveis usadas foram localização do WC, conhecimento da doença, hematúria MAC, local de contacto com H_2O, profissão, motivo de contacto com H_2O e naturalidade.

Os resultados são apresentados na Tabela 11, onde se pode observar que os valores do desvio padrão das estimativas dos parâmetros do GLM Poisson são bastante menores que os obtidos para o modelo GLM binomial negativa.

[1] A estatística de teste é $\frac{\text{estimativa do parâmetro}}{\text{desvio padrão da estimativa}}$, quando a estimativa é obtida por máxima verosimilhança e segue uma distribuição normal.

Tabela 11 – Valores dos parâmetros e dos seus desvios padrão para dois modelos, GLM Poisson e GLM binomial negativa com as mesmas covariáveis.

Covars vs Modelos	GLM Poisson Covars	Desvio padrão	GLM Binomial Negativa	Desvio Padrão
Ordenada na origem	1,9744 ***	0,080	2,632 ***	0,565
WC (Dentro)	(a)	(a)	(a)	(a)
WC (Fora)	0,598 ***	0,041	0,727 **	0,276
Cont. com água (Rio)	(a)	(a)	(a)	(a)
Cont. com água (Lagoa)	-0,355 ***	0,055	-0,525	0,372
Cont. com água (Tanque)	-1,278 ***	0,042	-1,207 ***	0,290
Conhec. da doença (Sim)	(a)	(a)	(a)	(a)
Conhec. da doença (Não)	1,233 ***	0,042	1,126 ***	0,287
Hematúria (Negativo)	(a)	(a)	(a)	(a)
Hematúria (Positivo)	1,270 ***	0,029	1,267 ***	0,297
Prof. (Func. Público)	(a)	(a)	(a)	(a)
Prof. (Estudante)	-0,007	0,064	-0,886 º	0,529
Prof. (Agricultor)	-0,747 ***	0,064	-1,358 **	0,498
Prof. (Trab. doméstico)	0,195 **	0,063	-0,878 º	0,503
Prof. (Outras)	-0,255 ***	0,071	-0,994 º	0,561
Motivo (Buscar água)	(a)	(a)	(a)	(a)
Motivo(Higiene Pessoal)	-0,340 ***	0,033	-0,638 *	0,299
Motivo (Lavar roupa)	-0,867 ***	0,045	-0,396	0,362
Motivo (Pescar)	-1,048 ***	0,046	-0,861 **	0,319
Motivo (Nadar)	0,325 ***	0,034	0,192	0,390
Naturalidade (Luanda Bengo)	(a)	(a)	(a)	(a)
Naturalidade (Bié Huambo Moxico)	0,210 ***	0,033	0,059	0,287
Naturalidade (Norte)	-0,311 ***	0,044	0,261	0,308
Naturalidade (Sul)	0,476 ***	0,040	-0,169	0,553

*** Valor p < 0,001 ** 0,001 < Valor P < 0,01 * 0,01 < Valor p < 0,05
º 0,05 < Valor p < 0,1 (a) Nível de referência

Observa-se que a parte dos modelos ZI e ZA que toma em consideração o excesso de zeros usa um número bastante menor de covariáveis que a parte relativa às distribuições.

Os dois modelos GLM são diferentes relativamente às covariáveis que são significativas e aos seus coeficientes. Acontece o contrário nos modelos ZIP e ZAP, onde as covariáveis usadas são essencialmente as mesmas e os seus coeficientes são semelhantes.

Nos modelos ZINB e ZANB voltam a aparecer diferenças significativas, onde apenas uma covariável (exceptuando o género e idade, que estão sempre presentes) é comum na parte da distribuição binomial negativa - a hematúria, que em ambos os modelos apresentam coeficientes bastante semelhantes.

Se o coeficiente de um nível de uma covariável for positivo isto quer dizer que, segundo o modelo aplicado, o número de ovos tende a aumentar se o valor do nível aumentar. Assim, se o coeficiente de um nível for significativo e positivo é considerado um factor de risco. Se o coeficiente for negativo, o número de ovos na urina tende a diminuir quando o valor do nível da covariável aumenta. Neste

caso, se o coeficiente de um nível for significativo e negativo este é considerado um factor de protecção.

Como foi referido anteriormente, os modelos com Poisson subestimam o desvio padrão dos estimadores dos parâmetros. Isso faz com que, no teste de Wald, a estatística de teste seja inflacionada e que se considere as estimativas dos parâmetros significativas, ou seja, aumenta o erro de tipo I do teste.

Assim, achou-se prudente só considerar que uma covariável influencia o número de ovos em 10 ml de urina se esta for significativa nos modelos com binomial negativa.

Através da Tabela 10, podemos observar que o coeficiente associado ao género masculino é significativo apenas nos modelos com Poisson, na parte dos modelos que "lida" com a distribuição Poisson ou Poisson truncada, e não na parte dos modelos que "lida" com os zeros. Como coeficiente é sempre positivo, poder-se-ia afirmar que os dados sugerem que pertencer ao género masculino é um factor de risco quanto à Schistosomose. Como esta conclusão apoia-se apenas em modelos com Poisson, decidiu-se proceder a um estudo mais aprofundado. Foi considerado o estudo da covariável género realizado anteriormente. Como os dois sexos apresentaram resultados semelhantes, concluiu-se que a covariável género parece não influenciar significativamente os valores da variável resposta.

A idade é significativa no modelo GLM Poisson e na parte dos valores não nulos do modelo ZANB. Em ambos os casos, o coeficiente é sempre muito próximo de zero, o que sugere que a idade não influencia o número de ovos na urina. A literatura afirma que crianças e idosos são considerados populações de risco (WHO, 2002). Como a amostra só tem indivíduos entre 15 e 75 anos, o efeito da variável idade parece ter sido reduzido.

Quanto a factores de risco nas restantes covariáveis, não ter WC, presença de sangue na urina (hematúria MAC), contacto com água nos rios e não estar informado sobre a doença aparecem como factores de risco. Um estudo dos coeficientes dos vários modelos sugere que nadar e ir buscar água também poderão ser factores de risco, mas isto é menos claro que nas covariáveis referidas anteriormente.

A variável água canalizada não é significativa em nenhum dos modelos, o que provavelmente se deve a que os indivíduos, com acesso a água canalizada da rede pública, tenham os mesmos comportamentos de risco que os indivíduos sem acesso (nadar em rios, pescar, lavar roupa, etc.).

Seguidamente vai ser realizado um teste de qui-quadrado entre modelos com covariáveis e modelos apenas com a ordenada na origem para perceber se os modelos com covariáveis produzem um melhor ajustamento aos dados.

Na Tabela 12 apresentamos os valores da estatística de teste e os valores p obtidos aplicando o teste de qui-quadrado.

Tabela 12 – Resultados do teste de Qui-Quadrado e AIC's dos modelos com covariáveis confrontados com modelos onde apenas se considerou a ordenada na origem.

	GLM Poisson	GLM NB	ZIP	ZINB	ZAP	ZANB
qui-quadrado	9276,5	60,7	8436,0	79,4	8439,6	57
GL	20	9	27	17	27	11
valor p	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001
AIC ordenada na origem	27293,0	2065,7	22804,7	2067,7	22804,7	2030,9
BIC ordenada na origem	27296,7	2067,4	22812,1	2078,8	22812,1	2042,0
AIC covars	18056,5	2023,0	14422,6	2022,2	14419,1	1995,8
BIC covars	18134,3	2058,1	14530,0	2096,3	14526,5	2047,7

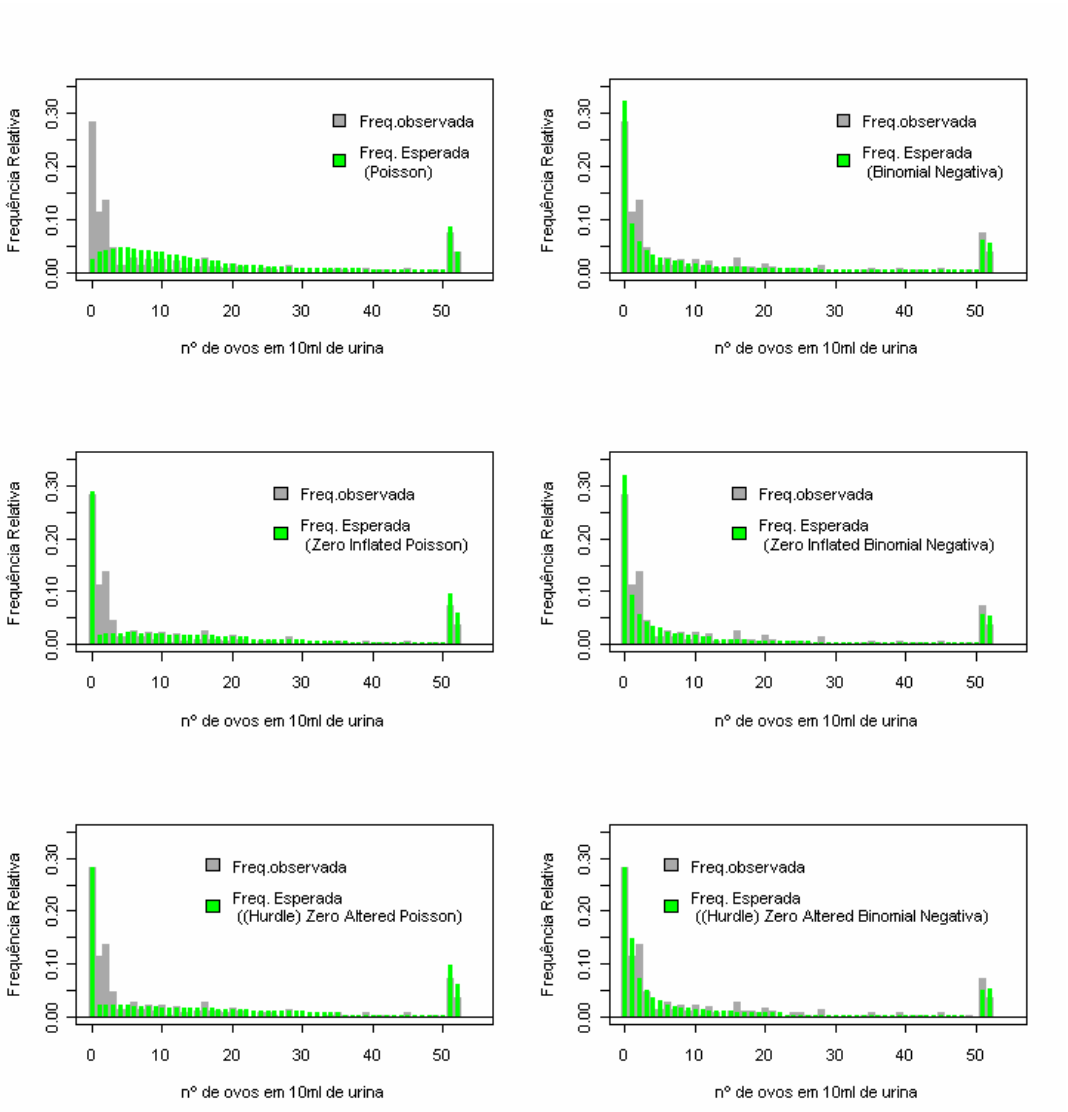
Como se pode constatar em todos os modelos, o uso de covariáveis produziu diferenças significativas no ajustamento dos modelos. Para saber qual o melhor ajustamento, usou-se o AIC e o BIC. Aqui também todos os modelos com covariáveis apresentaram um menor AIC. Quanto ao BIC, o modelo ZINB apresenta valores mais elevados quando se usa as covariáveis mas a diferença é pequena.

5.2.2 – ANÁLISE DE VALORES AJUSTADOS

5.2.2.1 – NÚMERO DE INDIVÍDUOS QUE APRESENTA UMA DETERMINADA QUANTIDADE DE OVOS EM 10 ML DE URINA

Relativamente ao número de indivíduos que apresenta uma determinada quantidade de ovos no exame de urina, na Figura 19 estão representados gráficos que sobrepõem os dados ajustados (barras em verde claro) aos dados recolhidos (barras em cinza) para os seis modelos. As últimas duas barras em cada gráfico representam o número de indivíduos que apresentou contagens entre 50 e 100 ovos e mais de 100 ovos, respectivamente.

Figura 19 - **Gráficos que comparam a frequência relativa do número de indivíduos que apresenta uma determinada quantidade de ovos em 10 ml de urina observados (cinza) com a frequência relativa dos valores estimados por cada modelo (verde claro).**



Os gráficos mostram que o GLM Poisson não é o melhor modelo para estes dados pois não prevê correctamente o número de zeros, ao contrário do ZIP e do ZAP. Todos os modelos baseados na distribuição de Poisson parecem ter dificuldade em modelar o número de indivíduos com contagens reduzidas de ovos.

Os dois primeiros modelos, GLM Poisson e GLM binomial negativa, são modelos que não tratam as observações nulas por um processo separado. Observa-se que o modelo GLM Poisson não se ajusta satisfatoriamente aos dados nulos ao contrário do GLM binomial negativa, que parece ajustar a grande quantidade de zeros de uma forma satisfatória, mesmo comparando com os modelos ZI e ZA.

É importante relembrar que a média de ovos em 10 ml de urina é de 24,25 e variância é de 7975,58. Considerando apenas os valores não nulos dos dados, a média é de 33,84 e a variância é de 10817,44. Tal como foi afirmado atrás, estes valores indicam uma grande dispersão dos dados relativamente à média. Logo a distribuição de Poisson não é uma boa opção para modelar estes dados e será apenas considerada para comparação.

Tabela 13 – Frequências absolutas observadas e frequências estimadas pelos vários modelos, para o número de indivíduos que apresentam uma determinada quantidade de ovos no teste de 10 ml de urina.

Nº de ovos	Amostra	GLM Poisson	GLM Bin. Neg.	ZIP	ZINB	ZAP	ZANB
0	85	7	97	87	96	85	85
1	34	11	27	5	28	7	44
2	41	12	17	6	17	7	22
3	14	13	12	6	13	7	15
4	4	14	10	6	10	7	11
5	4	14	8	7	9	7	9
6	8	13	7	7	7	6	7
7	4	12	6	6	6	6	6
8	7	12	6	6	6	6	5
9	3	11	5	6	5	6	5
10	7	11	5	6	5	5	4
11	1	10	4	5	4	5	4
12	6	10	4	5	4	5	3
13	2	9	3	5	3	5	3
14	0	8	3	5	3	5	3
15	3	7	3	5	3	5	3
16	8	7	3	5	3	5	2
17	3	6	3	5	3	5	2
18	3	6	2	4	2	4	2
19	2	5	2	4	2	4	2
20	5	5	2	4	2	4	2
$21 \leq x \leq 49$	23	60	37	58	36	57	30
$50 \leq x \leq 99$	22	26	18	29	17	29	15
≤ 100	11	11	16	18	16	18	16
	300	300	300	300	300	300	300
Parâmetros		21	10+1	8+21+1	3+16+1+1	8+21+1	8+5+1+1
AIC		18056,5	2021,0	14422,6	2022,3	14419,1	1995,8
BIC		18134,2	2058,1	14530,0	2096,3	14526,5	2047,7

A melhoria do ajustamento dos modelos ZIP e ZAP relativamente ao GLM Poisson é confirmada com os valores dos seus AIC's e dos BIC's. O modelo ZIP apresenta um AIC de 14422,6 e um BIC de 14530,0). No modelo ZAP obtém-se um AIC de 14419,1 e um BIC de 14526,5. Ambos são melhores que o AIC (18056,5) e BIC (18134,2) do modelo GLM Poisson.

Analisando a Figura 19 e a Tabela 13 conclui-se que os modelos baseados na distribuição binomial negativa ajustaram-se melhor que os modelos com Poisson.

Quanto às estimativas para valores não nulos, os modelos com binomial negativa "lidam" melhor com a grande dispersão dos dados que os modelos com Poisson. Apesar disso, todos os modelos tem dificuldade em "lidar" com os valores muito elevados presentes na amostra.

Fumes e Corrente (2010) aplicaram modelos ZINB a dados de um questionário que estuda a frequência alimentar, deparando-se com grande quantidade de zeros e grande dispersão dos dados não nulos. Encontraram dificuldades em ajustar os modelos ZINB a estes dados. Porém referem vários autores que procederam a ajustamentos bem sucedidos destes modelos em amostras com grande percentagem de zeros mas com pequena dispersão de dados não nulos. Isto sugere que existe um limite para a dispersão dos valores não nulos que a binomial negativa consegue comportar. Estes autores afirmam ainda que o referido limite pode ser influenciado pela percentagem de zeros na amostra.

Durante a realização deste trabalho foram construídas um conjunto de amostras artificiais com 5000 indivíduos para estudar o funcionamento destes modelos com amostras de maior dimensão. Tentou-se que estas amostras apresentassem valores e proporções semelhantes aos da amostra.

Verificou-se que nenhuma distribuição, mesmo a binomial negativa com uma amostra de grande dimensão, conseguia simular convenientemente valores com uma dispersão semelhante à observada. Isto sugere que dever-se-ia procurar uma distribuição que conseguisse acomodar maior dispersão ou optar por misturas com mais de duas componentes.

A observação dos gráficos sugere pouca diferença entre o ajustamento feito pelos três modelos com a binomial negativa. Os modelos GLM binomial negativa (2021,0; 2058,1), ZINB (2022,3; 2096,3) e ZANB (1995,8; 2047,7) apresentam valores de AIC e BIC semelhantes.

Deste modo, usou-se o teste de proximidade de Vuong (Vuong (1989)), implementado no pacote PSCL do R, para tentar perceber se existe alguma diferença significativa quanto ao ajustamento aos dados por parte destes modelos. Neste teste, a hipótese nula afirma que não existem diferenças significativa entre os ajustamentos dos modelos. A hipótese alternativa afirma que um dos dois modelos testados tem um ajustamento mais próximo dos dados.

Este teste foi utilizado por permitir a comparação de modelos que não estejam encaixados (*nested*). Chama-se à atenção que o ZINB está encaixado no modelo GLM binomial negativa pois o primeiro pode ser obtido acrescentando componentes ao segundo. O mesmo não se passa entre os modelos GLM binomial negativa e ZANB ou entre os modelos ZINB e ZANB.

Este teste apenas sugere qual dos modelos faz um ajustamento mais próximo aos dados. No entanto, isto não determina qual o melhor ajustamento. Por exemplo, uma recta de regressão poderá ser o ajustamento mais próximo dos dados. Mas se se modelar uma série temporal que exiba um efeito sazonal importante, que se deseja representar, então um modelo que capte esse efeito, mas que apresente resíduos mais elevados, será preferível à recta de regressão.

Tabela 14 – Resultados do teste de proximidade de Vuong para os modelos com distribuição binomial negativa.

Teste de proximidade de Vuong's	GLM NB vs ZINB	GLM NB vs ZANB	ZINB vs ZANB
Estatística	-2,01845	-2,2914	-0,95125
valor p	0,022	0,011	0,171
Ajustamento mais próximo	ZINB	ZANB	ZANB

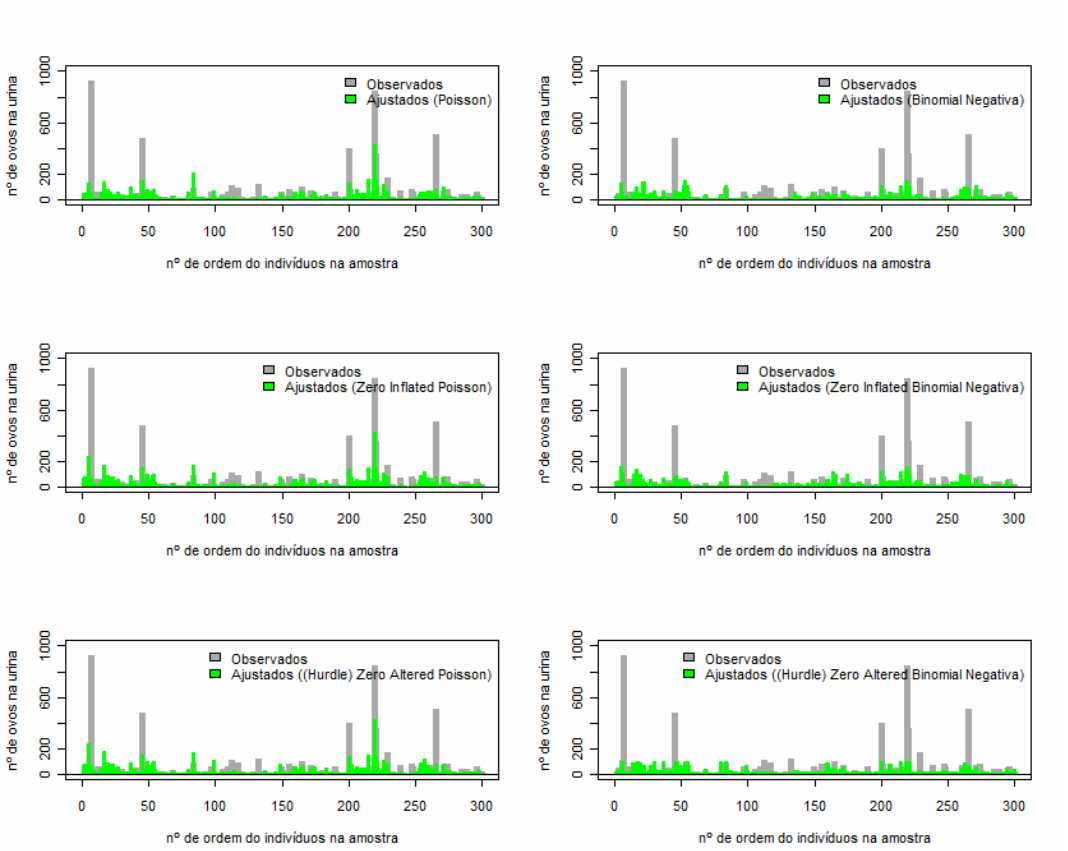
Como se pode observar na Tabela 14 os modelos ZINB e ZANB têm um ajustamento significativamente mais próximo dos dados que o modelo GLM Binomial Negativo. O teste indica que o ajustamento do modelo ZANB é mais próximo dos dados do que o de ZINB, mas que a diferença não é significativa (o valor p de 0,171 sugere que H_0 não seja rejeitada, logo não existe um modelo que se ajuste significativamente melhor).

Apesar destes resultados, é de referir que os coeficientes da parte que trata dos zeros no modelo ZINB são todos não significativos, o que sugere semelhança entre os modelos GLM Binomial Negativo e o modelo ZINB.

5.2.2.2 – NÚMERO DE OVOS EM 10 ML DE URINA PARA CADA INDIVÍDUO DA AMOSTRA

As representação gráficas do número de ovos, em 10 ml de urina, observados e estimados para cada indivíduo pelos vários modelos são apresentado na Figura 20. É evidente a falta de concordância entre os dois conjuntos de barras o que sugere um ajustamento deficiente dos modelos aos dados.

Figura 20 - Gráficos que comparam o número de ovos em 10 ml de urina observados em cada indivíduo da amostra (cinza) com o número de ovos em 10 ml de urina estimados para cada indivíduo (verde claro), em cada modelo.



Na tabela seguinte apresentam-se os valores observados e os valores ajustados pelos seis modelos para os 20 primeiros indivíduos da amostra. Não foi possível apresentar estes valores para todos os indivíduos pois a tabela completa seria demasiado extensa. Uma listagem exaustiva é apresentada em anexo.

Tabela 15 – Número de ovos em 10 ml de urina observados e número de ovos estimados pelos vários modelos, para cada indivíduo da amostra.

Indivíduo	Amostra	GLM Poisson	GLM Bin. Neg.	ZIP	ZINB	ZAP	ZANB
1	0	45	25	69	38	69	20
2	0	51	28	74	40	73	24
3	45	8	12	7	12	7	15
4	1	24	16	26	21	25	19
5	0	128	124	235	154	235	68
6	105	61	55	53	102	53	97
7	925	47	26	71	39	70	21
8	28	16	16	14	21	14	20
9	0	12	27	25	13	24	9
10	59	5	11	5	8	5	16
11	1	5	11	5	8	5	15
12	24	11	21	17	12	16	15
13	41	18	53	13	33	13	92
14	58	59	53	52	101	53	92
15	15	24	53	23	33	23	90
16	7	133	97	173	139	174	81
17	16	38	55	29	102	29	97
18	1	59	39	77	90	77	50
19	28	78	51	87	100	88	85
20	0	41	36	74	32	73	15
Parâmetros		21	10+1	8+21+1	3+16+1+1	8+21+1	8+5+1+1
AIC		18056,5	2021,0	14422,6	2022,3	14419,1	1995,8
BIC		18134,2	2058,1	14530,0	2096,3	14526,5	2047,7

Uma análise da Figura 26 e da Tabela 15 mostram que todos os modelos considerados produzem valores ajustados bastante diferentes dos valores observados.

Para ter-se uma noção da ineficácia dos modelos vamos comparar o número de indivíduos com zeros. O modelo ZIP prevê 3 indivíduos sem ovos e o modelo ZINB apenas 1. Todos os outros modelos não prevêem indivíduos sem ovos. Considerando que existem 85 indivíduos que não apresentam ovos no teste de urina, os resultados são decepcionantes.

Verifica-se que são os modelos com Poisson, principalmente o ZIP e o ZAP, que apresentam um melhor ajustamento. Estes modelos parecem ajustar melhor valores muito elevados, como o observado no indivíduo 219 (ver Figura 26) que apresenta 848 ovos no teste de urina. O modelo GLM Poisson prevê 429 ovos, o ZIP prevê 428 e o ZAP 430, valores mais próximos que dos 148 do modelo GLM binomial negativa, 156 do ZINB e 72 do ZANB.

No seguinte quadro consta uma aproximação do valor da média e do desvio padrão para os valores ajustados nos vários modelos e as mesmas estatísticas para a amostra.

Tabela 16 – Valores da média e desvio padrão da amostra e para os valores ajustados dos vários modelos.

	GLM Poisson	GLM binomial negativa	ZIP	ZINB	ZAP	ZANB	Amostra
Média	24,253	24,011	26,073	23,728	25,986	24,559	24,253
Desvio padrão	37,651	26,750	40,124	27,380	40,267	24,599	89,306

A análise da Tabela 16 releva que todos os modelos têm uma média próxima da média dos valores da amostra (24,253). O modelo Poisson forneceu uma média igual à média da amostra.

Quanto ao desvio padrão, os dados sugerem que nenhum dos modelos conseguiu modelar a grande dispersão dos valores da amostra. Todos os modelos apresentam um desvio padrão bastante inferior ao amostral. Os modelos com distribuição de Poisson apresentarem um desvio padrão mais elevado, logo mais próximo do desvio padrão da amostra.

Duas questões levantaram-se quando estes resultados foram observados. Porquê um ajustamento tão mau? Porque razão estes dados aparentam dar vantagem aos modelos com Poisson, uma vez que na literatura e em todo o trabalho foram os modelos com binomial negativa que tiveram sempre supremacia?

Relativamente à primeira pergunta suspeitou-se que a amostra seria reduzida para se poder, a partir dela modelar de forma satisfatória um conjunto de dados com tão grande dispersão.

Replicou-se a amostra observada usando o método *Bootstrap* não-paramétrico de forma a obter uma amostra com uma dimensão de 5000 indivíduos. Aplicando os mesmos modelos, verificou-se que o ajustamento do número de ovos por indivíduo mantinha as mesmas características.

Tentou-se ainda atribuir o valor de zero a todos os indivíduos que apresentassem uma probabilidade, maior ou igual a 50%, de ter zero ovos. Caso contrário, foi atribuído ao indivíduo o número de ovos previsto pelo modelo. O número de zeros sofreu um incremento em todos os modelos excepto nos GLM (GLM Poisson 0,GLM NB 0, ZIP 17, ZINB 14, ZAP e ZANB 9) mas mesmo assim ficaram aquém dos 85 zeros registados.

Quanto à segunda pergunta, como os modelos com Poisson usam maior número de covariáveis, conjecturou-se que isso poderia ser a causa de um melhor ajustamento por parte destes modelos. Construíram-se modelos baseados na distribuição binomial negativa onde eram usadas as mesmas covariáveis que nos modelos com Poisson. Não foi detectada uma melhoria significativa nos modelos de binomial negativa modificados.

O número de ovos na urina para cada indivíduo da amostra são essencialmente valores médios, logo são afectados por valores extremos. Como esta amostra apresenta seis valores que somam quase metade da amostra (925, 848, 510, 473, 402 e 510 somados dão 3518 contra 3758, a soma dos restantes dados), estes valores inflacionam os valores médios, daí a escassez de zeros. A grande quantidade de zeros diminui as estimativas para valores grandes.

Retiraram-se os seis maiores valores da amostra e foram novamente construídos os modelos. O problema manteve-se.

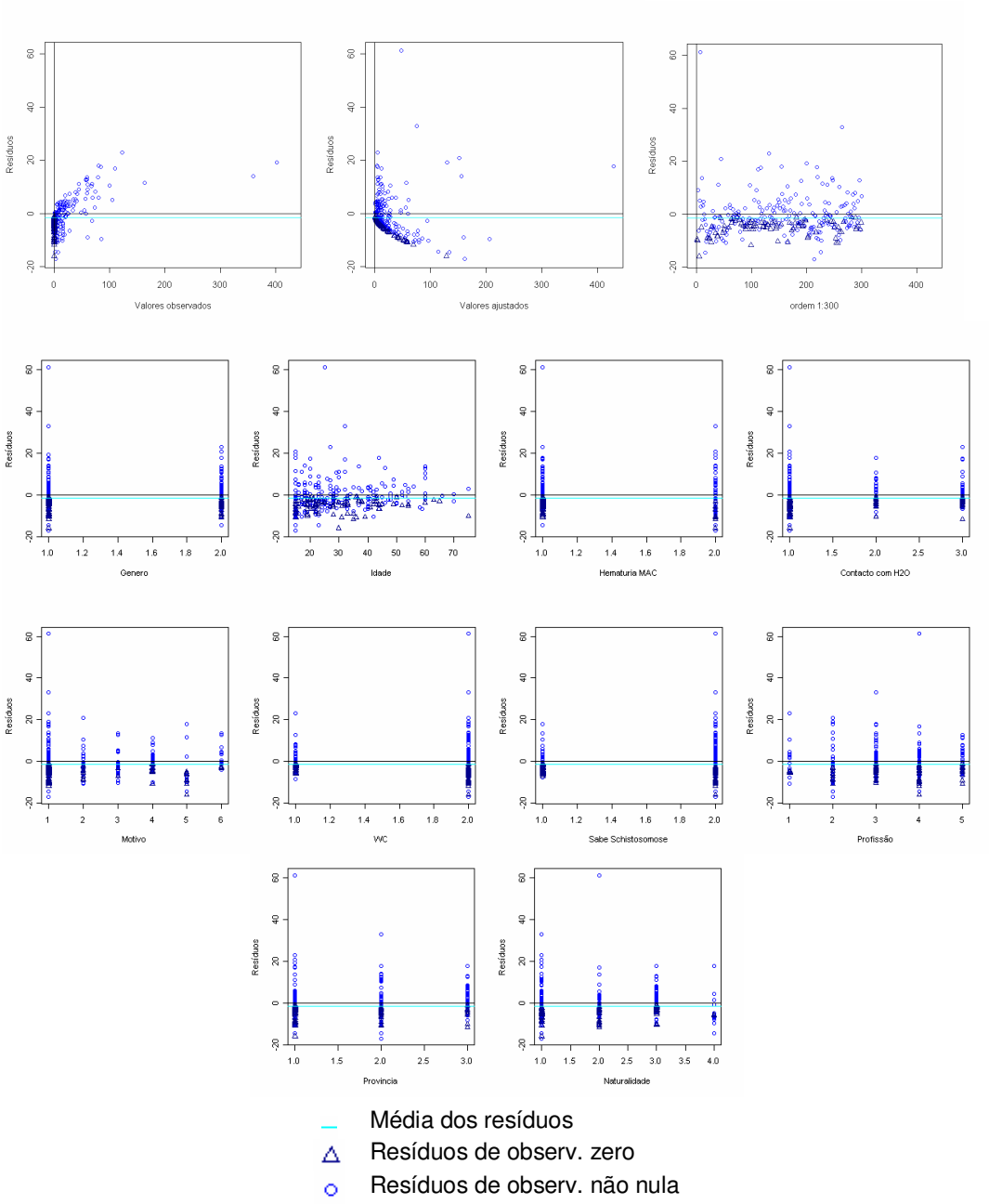
Depois de todas estas tentativas resta concluir que estes modelos não são indicados para prever o número de ovos para os indivíduos da amostra.

Seguidamente será apresentada uma análise dos resíduos de desvio destes valores.

GLM POISSON (GLM POISSON)

Figura 21 - Análise de resíduos de desvio do modelo GLM Poisson.

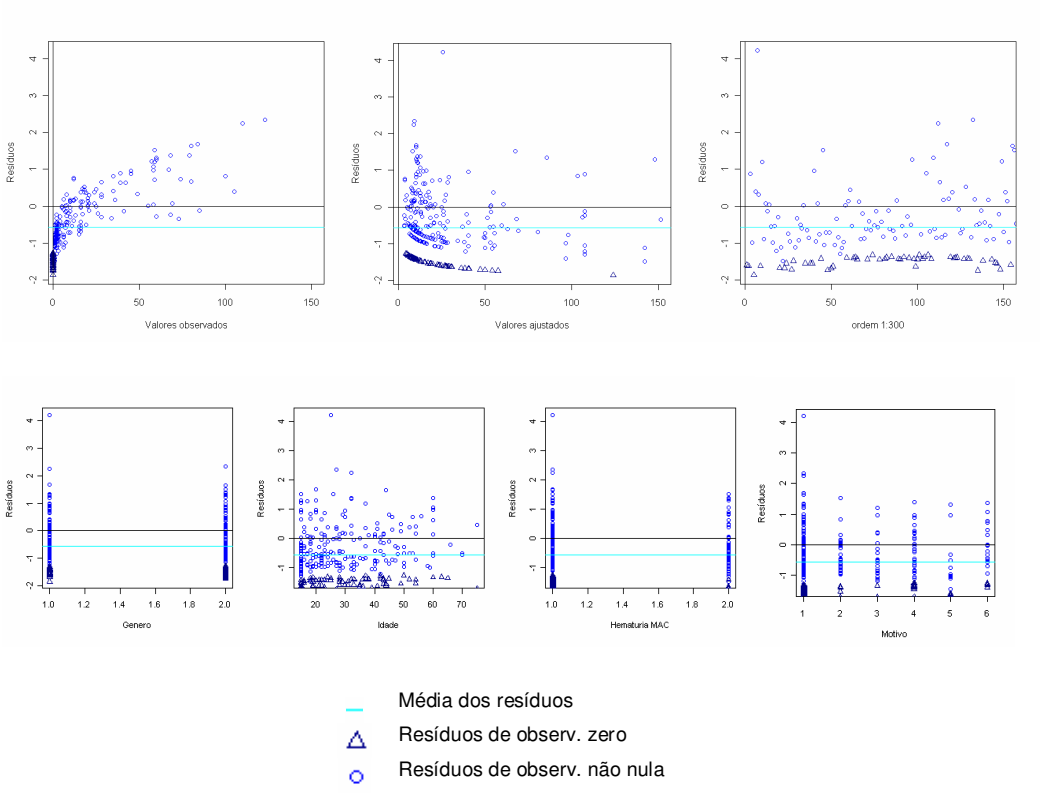
Média dos resíduos
-1,503
Percentagem de resíduos entre -2 e 2
21,3%
Percentagem de resíduos não negativos
27,3%



GLM BINOMIAL NEGATIVA (GLM NB)

Figura 22 – Análise de resíduos d desvio do modelo GLM binomial negativa.

Média dos resíduos -0,571
Percentagem de resíduos entre -2 e 2 99,0%
Percentagem de resíduos não negativos 24,0%



Analisando os gráficos dos resíduos versus os valores observados, estes sugerem que os resíduos não apresentam homocedasticidade. Apesar disso, esta tendência não é tão pronunciada nestes resíduos como nos de Pearson.

Analisando os gráficos dos resíduos versus os valores observados verifica-se que pequenos valores observados (0, 1, 2) têm resíduos negativos. À medida que o número de ovos observado aumenta, também aumenta o seu resíduo.

Quanto à percentagem de resíduos entre -2 e 2, a GLM Poisson apresenta mais de 5% de resíduos com valores fora do intervalo (-2,2), o que não se passa com o GLM binomial negativa. Isto sugere um melhor ajustamento por parte do modelo GLM binomial negativa.

Outra tendência verificada é a pequena percentagem de resíduos não negativos encontrada em todos os modelos (GLM Poisson 27,3% e GLM binomial negativa 24,0%).

Observa-se ainda a tendência de os resíduos dos zeros observados serem exclusivamente negativos. Os modelos têm um suporte não negativo (0, 1, 2...) e os valores ajustamos não são mais que médias. Assim é expectável que, para os zeros, a média seja um valor positivo e nunca negativo. Como os zeros têm valores ajustados positivos, os resíduos $(y_i - \hat{y}_i)$ das observações nulas têm de ser negativos.

5.2.3 – MODELAÇÃO DE INDIVÍDUOS INFECTADOS QUE NÃO APRESENTAM OVOS NA URINA

Como foi referido anteriormente, em nove dos indivíduos os ovos de *Schistosoma haematobium* só foram detectados através de biópsia, não sendo detectados nos exames de urina.

Usando os seis modelos, foi calculada a probabilidade de estes indivíduos terem ovos na urina e os resultados estão registados na Tabela 17.

Tabela 17 – Probabilidade de um indivíduo ter ovos, segundo os modelos ajustados, para nove indivíduos que não apresentam ovos na urina, que foram no entanto detectados através de uma biópsia.

Biópsia	Ovos viáveis	Ovos viáveis	Ovos viáveis	Ovos calcificados	Ovos calcificados	Ovos calcificados	Ovos calcificados	Ovos calcificados	Ovos e cancro
Indivíduo	138	166	179	121	122	143	176	181	135
GLM P	1	1	1	1	1	1	1	0,89	1
GLM NB	0,66	0,76	0,60	0,63	0,62	0,65	0,66	0,62	0,78
ZIP	0,29	0,64	0,33	0,58	0,58	0,63	0,33	0,37	0,96
ZINB	0,64	0,82	0,69	0,74	0,74	0,53	0,75	0,48	0,65
ZAP	0,27	0,64	0,32	0,57	0,56	0,63	0,31	0,61	0,96
ZANB	0,27	0,64	0,32	0,57	0,56	0,63	0,31	0,61	0,96

Considerou-se que um modelo prevê que um indivíduo apresenta ovos no exame de urina se atribuir uma probabilidade superior a 50% de isso acontecer. Se um modelo detectar ovos nestes nove indivíduos, isto pode ser interpretado como um bom desempenho por parte do modelo.

Para esta análise ser coerente devemos ter presente quantos zeros estes modelos prevêem para uma qualquer amostra com dimensão 300 (número de indivíduos sem ovos nos valores estudados em 5.2.2.1.) e quantos zeros eles prevêem para a presente amostra (número de zeros nos valores estudados em 5.2.2.2.).

Tabela 18 – Números de zeros estimados pelos modelos para a presente amostra e para uma qualquer amostra de 300 indivíduos.

	GLM P	GLM NB	ZIP	ZINB	ZAP	ZANB	Amostra
Número de zeros estimados para uma amostra com 300 indivíduos.	7	97	87	96	85	85	85
Número de zeros estimados para a presente amostra.	0	0	17	14	9	9	85

O modelo GLM Poisson dá como certo estes indivíduos apresentarem ovos no exame de urina. Porém, devemos ter em atenção que este modelo prevê muito poucos zeros para qualquer amostra de 300 indivíduos e não prevê qualquer zero na presente amostra.

Note-se que o modelo GLM binomial negativa apresenta uma previsão adequada do número de zeros numa qualquer amostra com 300 indivíduos mas não prevê qualquer zero para a presente amostra. Assim não é

surpreendente que atribua, aos nove indivíduos que estamos a estudar, uma probabilidade superior a 50% de apresentarem ovos no exame de urina.

O modelo ZINB atribui uma probabilidade acima de 50% aos nove indivíduos excepto um, o 181º elemento da amostra.

Os modelos ZAP e ZANB atribuem exactamente as mesmas probabilidades aos nove indivíduos e a três destes atribuem probabilidades inferiores a 50% de apresentarem ovos na urina. Tendo em conta que os modelos prevêem apenas nove zeros na presente amostra, e que três destes zeros são atribuídos a indivíduos infectados pelo parasita, isto sugerem que estes modelos não sejam os melhores para estudar estes indivíduos. O mesmo se pode dizer do modelo ZIP que atribui uma probabilidade superior a 50% apenas a cinco destes indivíduos.

Todo considerado, os dados sugerem que o único modelo capaz de apontar para a identificação da infecção em indivíduos que não apresentam ovos no exame de urina é o modelo ZINB.

6 - CONCLUSÕES

Nos resultados apresentados anteriormente verificou-se que apenas o modelo GLM Poisson não conseguiu ajustar o grande número de zeros da amostra. Todos os outros modelos foram bem sucedidos neste ponto.

Os modelos com distribuição binomial negativa tratam melhor com a grande dispersão dos dados que a distribuição de Poisson. Estes três modelos (GLM binomial negativa, ZINB e ZANB) aparentam apresentar resultados semelhantes relativamente ao ajustamento dos dados. O teste de Voung sugere que existe uma melhor aproximação aos dados por parte dos modelos ZINB e ZANB comparativamente ao modelo GLM binomial negativa.

Os modelos com distribuição de Poisson são menos parcimoniosos que os modelos com distribuição binomial negativa. Destes últimos, o modelo GLM binomial negativa é o mais parcimonioso com 11 parâmetros a estimar, o ZANB tem 15 parâmetros e o ZINB é o que apresenta mais parâmetros, 21 no total. Verificou-se que poder-se-ia retirar algumas covariáveis do modelo ZINB, diminuindo consideravelmente o número de parâmetros a estimar, sem que os valores do AIC e BIC se alterassem significativamente (o valor do BIC até apresentou uma diminuição). Assim, o número de parâmetros deverá ser usado com precaução como critério de escolha entre os modelos com distribuição binomial negativa.

Como é apontado na subsecção 5.2.2.1, todos os modelos com distribuição binomial negativa ajustam-se satisfatoriamente ao número de indivíduos com um dado número de ovos. Quanto aos resíduos destes valores, nenhuma tendência foi detectada na sua análise que sugira um ajustamento deficiente dos modelos. Assim, esta subsecção, sugere uma vantagem dos modelos com binomial negativa. Porém, dentro destes, não aponta uma vantagem clara para nenhum deles.

Na subsecção 5.2.2.2 foram estudadas as estimativas que cada modelo efectuava para cada indivíduo da amostra onde todos os modelos demonstraram performances medíocres face às expectativas. Os modelos com distribuição de Poisson demonstraram uma pequena vantagem devido a conseguirem prever melhor valores elevados de ovos relativamente aos modelos com binomial negativa. Os resíduos destes valores apresentam várias tendências que, apesar de explicáveis, sugerem um deficiente ajuste dos modelos. Esta subsecção sugere que estes modelos não são uma boa escolha para modelar o número de ovos em indivíduos.

Relativamente à modelação de indivíduos infectados, que não apresentam ovos no exame de urina, o modelo ZINB apresentou a melhor prestação, prevendo ovos em oito dos nove indivíduos.

Por estas razões, todos os modelos com distribuição binomial negativa parecem ser as escolhas correctas para modelar estes dados.

Quanto à selecção entre os modelos ZI e ZA, antes de mais, é de referir que o modelo GLM binomial negativa apresentou um ajustamento bom aos dados e similar ao ajustamento dos modelos ZINB e ZANB.

ZANB prevê um número de zeros exactamente igual ao número de zeros da amostra mas isto não pode ser visto como um mérito do modelo. Este tipo de modelos prevê que valor esperado do número de zeros é sempre igual ao número de zeros da amostra. Por isso esta "vantagem" do modelo ZANB deve ser posta em perspectiva.

Poder-se-ia argumentar que, devido a apresentar um AIC e BIC mais baixo, dever-se-ia optar pelo modelo ZANB. O primeiro argumento que refuta esta opção é que os valores destas medidas não são muito diferentes. Outro argumento é de que o modelo ZANB usa o número de zeros da amostra para prever o número de zeros no modelo. Isto quer dizer que, neste caso, aproximadamente 28,3% da amostra tem um ajustamento perfeito à partida. Assim, os modelos ZA partem com uma vantagem em termos da qualidade do ajustamento.

Um argumento para preferir o ZANB e o ZINB é o seu AIC inferior, mas mesmo assim este é muito próximo do AIC do modelo GLM binomial negativa para se poder tomar a decisão com base apenas neste critério. Quanto ao BIC, este favorece o GLM binomial negativa.

Segundo Rose et al. (2006), se o objectivo for apenas realizar previsões, é indiferente escolher entre modelos ZI e ZA desde que as previsões sejam semelhantes. O mesmo não acontece se o objectivo for realizar inferência, onde a estrutura do modelo é importante e a existência ou não de várias fontes de zeros deve ser levada em conta.

Um dos objectivos deste trabalho é realizar inferência. Assim, devemos atender ao número e natureza das fontes de zeros. O modelo GLM binomial negativa deve ser usado quando a fonte de zeros é a mesma que a fonte dos valores não nulos. O modelo ZINB deve ser utilizado quando existe um processo que apenas origina zeros, mas o processo que origina os valores não nulos também pode produzir zeros. Por fim, o modelo ZANB deve ser seleccionado no caso de existir um processo que apenas produz zeros e outro processo que apenas produz valores não nulos.

No caso em estudo, os zeros na contagem de ovos de *Schistosoma haematobium* na urina podem ter duas origens. O indivíduo pode não estar infectado (processo que só produz zeros). Por outro lado, um indivíduo pode estar infectado e não lhe detectarem ovos no exame originando zeros que são provenientes do mesmo processo que produz os valores não nulos. Como exemplo, refira-se a existência de nove indivíduos que, apesar de não terem apresentado ovos na urina, estes foram detectados posteriormente numa biópsia.

Assim, a estrutura do modelo ZINB é a mais concordante com a situação estudada.

Resumindo:

O modelo GLM binomial negativa tem o melhor BIC dos modelos e é o mais parcimonioso.

O modelo ZANB apresenta o menor AIC e o teste de Voung sugere que faz um ajustamento mais próximo dos dados que o GLM binomial negativa.

O teste de Voung sugere que o modelo ZINB faz um ajusta-se melhor aos dados do que o GLM binomial negativa, tem a estrutura mais adequada para lidar com a variável em estudo e é o único modelo que parece conseguir sinalizar a infecção em indivíduos infectados que não revelaram ovos na urina.

Estas informações sugerem que o modelo ZINB é o mais adequado para modelar estes dados.

Em trabalhos futuros sugere-se que se dedique mais atenção à análise de resíduos, principalmente à análise dos resíduos de desvio em modelos ZI e ZA. Na análise da bibliografia foram encontradas poucas referências à utilização de resíduos nos modelos ZI e ZA, principalmente se não foram de Pearson. Também foram encontrados poucas referências sobre a interpretação dos resultados obtidos.

Ainda relativamente aos resíduos, Franciscon (2004) refere os quantis residuais. Estes poderão ser adequados para os estudo destes dados. Infelizmente, devido a restrições de tempo não foi possível aprofundar o estudo deste tipo de resíduos.

Também será interessante aplicar estes modelos a dados sem restrições relativamente à idade. Lembra-se que os dados utilizados provinham de indivíduos com quinze ou mais anos de idade. Outros conjuntos de dados relativamente a este parasita observados no decorrer deste trabalho apresentavam uma distribuição que sugeria uma adequação mais forte do modelo ZINB.

Como já foi referido na secção 4.2, em estudos futuros seria interessante recolher informação que permita analisar a relação da Schistosomose com o HIV.

Por fim, seria interessante aplicar a estes dados, ou a similares, modelos de mistura onde não só os zeros fossem tratados por uma distribuição à parte, mas também os valores elevados tivessem um tratamento semelhante.

7 - BIBLIOGRAFIA

Basáñez, M. G., Marshall, C., Carabin, H., Gyorkos, T., Joseph, L. (2004). Bayesian statistics for parasitologists. *Trends in parasitology*, volume 20, n.º 2, pp. pp. 85 - 91.

Basu, M., Maji, A. K., Chakraborty, A., Banerjee, R., Mullick, S., Saha, P., Das, S., Kanjilal, S. D., Sengupta, S. (2010). Genetic association of Toll-like-receptor 4 and tumor necrosis factor-a polymorphisms with Plasmodium falciparum blood infection levels. *Infection, Genetics and Evolution,* nº 10, pp. 686 – 696

Bethell, J., Rhodes, A. E., Bondy, S.J, Lou W. Y. and Guttmann A. (2010). Repeat self-harm: application of hurdle models. *The British Journal of Psychiatry* 196, pp. 243 – 244

Cardoso, Sheila (2010). Schistosomose urinária e helmintoses intestinais: contribuição para o estudo clínico-epidemiológico e da respostaimune humoral na comunidade angolana. *IHMT: HMM - Dissertações de Mestrado*. Instituto de Higiene e Medicina Tropical. Universidade Nova de Lisboa.

Cook, J. (2009). Notes on the Negative Binomial Distribution. Retirado a 20 de Abril de 2011 da World Wide Web: http://www.johndcook.com

Cox, J. L., Heyse, J. F., Tukey, J. W. (2000). Efficacy Estimates from Parasite Count Data That Include Zero Counts. *Experimental Parasitology*, nº 96**,** pp.1 – 8.

Do, Chuong B., Batzoglu, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, nº 26, pp. 897 - 899.

Dominik, H., Barbour, A. D., Torgerson, P. R. (2009). Compound processes as model for clumped parasite data. *Mathematical Biosciences*, volume 222, nº 1, pp. 27 – 35.

Dupont, W. (2002). *Statistical modeling for biomedical researchers: A simple introduction to the analysis of complex data.* Cambridge University Press. Cambridge.

Franciscon, L. (2004).*Quantis Residuais*. Retirado a 20 de Setembro de 2011 da Word Wide Web http://www.est.ufpr.br/rt/fern04.pdf
Última actualização : Setembro de 2006

Figueiredo, Jacinta T. (2008). *Contribuição para o estudo da epidemiologia de morbilidade da Schistosomose vesical na população adulta de Angola. Províncias de Luanda, Bengo e Kwanza Sul.* Instituto de Higiene e Medicina Tropical. Universidade Nova de Lisboa.

Fomes, G., Corrente, J. (2010). Modelos inflacionados de zeros: aplicação na análise de um questionário de frequência alimentar. *Revista Brasileira Biometria,* volume 28, pp. 24 - 38.

Gonzales-Barron, U., Kerr, M., Sheridan, J.J., Butler, F. (2010). Count data distributions and their zero-modified equivalents as a framework for modeling microbial data with a relatively high occurrence of zero counts. *International Journal of Food Microbiology* 136, pp. 268 – 277.

Ground, M., Koch, S. F. (2008). Hurdle models of alcohol and tobacco expenditure in south African households. *South African Journal of Economics,* volume 76, nº 1, pp. 132 – 143.

Jackman,S., Kleiber ,C., Zeileis, A. (2008). Regression Models for Count Data in R. Journal of Statistical Software. *American Statistical Association*, volume 27(i08).

McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. New York, John Wiley & Sons, inc.

Morrison, David A. (2004). Technical variability and required sample size of helminth egg isolation procedures: revisited.
Parasitol Res, nº 94, pp. 361 – 366

Navarro, A., Utzet, F., Puig, P., Caminal, J., Martín, M. (2001). La distribución binomial negativa frente a la de Poisson en el análisis de fenómenos recurrentes. *Gac Sanit*, nº 15, pp. 447 - 52.

Paulino, C. D. (2011), *Glossário Inglês - Português de Estatística*. Sociedade Portuguesa de Estatística – Associação Brasileira de Estatística
Retirado a 5 de Junho de 2011 da World Wide Web: http://glossario.spestatistica.pt/
Última actualização : 1 de Junho de 2011

Potts, J. M. and Elith, J. (2006). Comparing species abundance models.
Ecological Modelling, volume 199, pp. 153 – 163.

Kahama, A. I., Odek, A. E., Kihara, R. K., Vennervald, B. J., Kombe, Y., Nkulila, T., Hatz, C. F., Ouma, J. O., Deelder, A. M. (1999). Urine circulation soluble egg antigen in relation to egg counts, hematuria, and urinary tract pathology before and after treatment in children infected with *Schistosoma haematobium* in Kenya. *American Journal of Tropical Medicine and Hygiene*, nº 61(2), pp. 215 - 219.

Kallestrup, P., Zinyama, R., Gomo, E., Butterworth, A.E., van Dam, G.J., Erikstrup, C., Ullum, H. (2005). *Schistosomiasis and HIV-1 infection in rural Zimbabwe: implications of coinfection for excretion of eggs.* Copenhagen, Center of Inflammation and Metabolism, Department of Infectious Diseases.

Ridout, M., Demétrio, C. G. B., Hinde, J. (1998). Models for count data with many zeros. *Proceedings of the XIXth International Biometric Conference*, pp. 179 - 192.

Rose, C. E., Martin, S. W., Wannemuehler, K. A. and Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics* 16, pp. 463 – 481.

Self, S. G., Liang , K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association*, vol. 82, nº 398, pp. 605 – 610.

Verez, Ângela (2010). *Bioecologia e caracterização molecular de Bulinus globosus de Angola.* Lisboa. Instituto de Higiene e Medicina Tropical.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and non-nested Hypotheses, *Econometrica*, volume 57, nº. 2, pp. 307 - 333.

Weiss, Jack (2008). *Lectures.* Retirado a 11 de Janeiro de 2011 da Word Wide Web : http://www.unc.edu/courses/2008fall/ecol/563/001/docs/lectures/lecture13.htm Última actualização : 13 de Outubro de 2008

World Health Organization, Expert Committee on the Control of Schistosomiasis (2002). *Prevention and control of schistosomiasis and soil-transmitted helminthiasis: report of a WHO expert committee.* Geneva, Switzerland, World Health Organization.

Xiang, L., Lee, A. H., Yau, K. K. W., McLachlan, G. J. (2007). A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistical Medecine,* 23, pp. 1608 – 1622.

Zuur, A.F., Leno, E. N., Walker, N. J., Saveliev, A. A. Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R.* Statistics for Biology and Health. Springer.

ANEXO 1 – NÚMERO DE OVOS EM 10 ML DE URINA OBSERVADOS E NÚMERO DE OVOS ESTIMADOS PELOS VÁRIOS MODELOS, PARA CADA INDIVÍDUO DA AMOSTRA

Indivíduo	amostra	GLM POIS	GLM NB	ZIP	ZINB	ZAP	ZANB
1	0	45	25	69	38	69	20
2	0	51	28	74	40	73	24
3	45	8	12	7	12	7	15
4	1	24	16	26	21	25	19
5	0	128	124	235	154	235	68
6	105	61	55	53	102	53	97
7	925	47	26	71	39	70	21
8	28	16	16	14	21	14	20
9	0	12	27	25	13	24	9
10	59	5	11	5	8	5	16
11	1	5	11	5	8	5	15
12	24	11	21	17	12	16	15
13	41	18	53	13	33	13	92
14	58	59	53	52	101	53	92
15	15	24	53	23	33	23	90
16	7	133	97	173	139	174	81
17	16	38	55	29	102	29	97
18	1	59	39	77	90	77	50
19	28	78	51	87	100	88	85
20	0	41	36	74	32	73	15
21	6	55	142	76	61	76	34
22	1	36	142	49	39	49	64
23	0	43	38	76	33	75	17
24	0	40	18	36	21	36	18
25	2	31	13	45	19	44	10
26	3	27	38	44	20	44	24
27	0	58	50	56	56	56	69
28	0	12	13	13	12	12	17
29	2	17	50	14	18	14	69
30	34	28	68	22	22	22	97
31	1	18	12	20	26	20	18
32	6	25	18	21	18	21	26
33	2	5	8	4	14	4	9
34	20	26	19	21	18	21	27
35	0	26	19	21	18	21	27
36	15	96	69	87	66	88	99
37	2	26	19	21	18	21	27
38	0	35	19	36	18	36	27
39	0	35	19	36	18	36	27
40	45	5	11	4	11	4	13
41	1	15	14	13	28	13	24
42	21	25	18	21	18	21	26
43	2	4	14	5	4	4	15
44	0	1	9	1	3	1	9
45	473	152	68	151	77	152	69
46	3	54	55	41	84	41	99
47	10	22	55	18	28	18	99
48	0	25	54	18	32	18	69
49	9	79	50	95	56	95	69
50	0	17	40	21	14	20	18
51	0	10	29	9	11	9	27
52	6	43	107	52	35	52	38
53	73	55	151	85	44	85	99
54	6	79	50	95	56	95	69

55	2	38	107	35	39	36	99
56	2	25	54	18	32	18	69
57	1	16	15	19	13	19	16
58	8	9	8	11	8	11	8
59	0	13	8	18	11	17	10
60	10	3	8	3	3	3	9
61	18	10	9	12	8	11	10
62	2	7	7	10	9	9	10
63	0	2	9	2	2	2	8
64	0	5	8	6	5	6	7
65	0	2	10	2	2	2	12
66	10	5	8	6	6	6	8
67	1	2	11	1	2	2	13
68	26	27	39	25	21	25	35
69	1	9	10	9	6	9	8
70	2	6	9	7	6	6	9
71	0	3	10	3	4	3	12
72	4	3	10	3	4	3	11
73	2	5	10	4	6	4	11
74	0	1	6	0	2	1	3
75	2	2	7	2	3	2	5
76	8	5	10	5	6	5	8
77	2	2	8	3	3	3	7
78	0	3	11	2	4	2	12
79	16	58	40	43	36	43	95
80	1	5	9	5	4	5	6
81	0	7	10	8	7	7	20
82	0	6	8	7	7	7	13
83	85	206	107	166	118	167	99
84	2	6	9	8	4	8	21
85	0	10	7	15	11	15	11
86	2	10	10	10	12	10	28
87	1	17	10	19	12	19	29
88	12	6	9	7	6	7	9
89	0	16	10	12	12	12	20
90	0	11	10	11	12	10	22
91	3	14	10	18	12	18	21
92	0	11	8	15	11	15	12
93	1	11	8	16	11	15	13
94	4	10	7	15	10	14	11
95	0	6	9	7	7	7	16
96	3	7	9	7	7	7	18
97	60	10	11	8	10	8	23
98	0	5	10	4	7	4	27
99	0	70	31	108	40	107	65
100	0	17	10	19	12	19	29
101	0	3	6	3	0	3	7
102	1	5	7	6	7	6	11
103	0	16	10	12	12	12	20
104	0	11	8	15	11	15	12
105	39	14	10	18	12	18	21
106	0	7	6	10	7	9	5
107	2	7	10	8	8	8	22
108	0	10	10	10	12	10	20
109	60	14	10	18	12	18	21
110	2	12	9	16	11	16	16
111	1	6	10	6	12	6	20
112	110	12	9	16	11	16	16
113	28	10	10	10	12	10	20
114	1	21	9	28	14	28	12

115	10	4	7	4	8	3	8
116	1	4	8	4	8	3	8
117	84	5	10	5	8	5	21
118	13	13	9	17	11	16	17
119	0	5	8	7	7	6	12
120	18	12	10	9	25	9	13
121	0	12	9	9	24	9	11
122	0	11	9	9	24	8	10
123	0	6	8	6	10	5	9
124	6	9	7	8	22	8	8
125	1	8	6	7	22	7	6
126	2	11	9	9	24	8	10
127	0	13	10	9	25	9	13
128	18	10	8	8	23	8	8
129	0	8	9	7	9	7	12
130	0	8	10	7	9	7	13
131	1	12	9	9	25	9	12
132	123	5	9	4	13	3	7
133	6	1	4	1	3	1	11
134	1	2	3	2	5	2	13
135	0	16	58	12	9	12	31
136	3	9	9	8	9	8	18
137	21	28	26	31	24	31	28
138	0	5	13	3	8	3	4
139	9	6	24	5	5	5	10
140	1	11	12	8	20	8	18
141	12	6	9	5	11	5	16
142	2	1	7	1	3	1	10
143	0	4	11	4	3	4	23
144	1	5	11	4	4	3	16
145	0	15	17	15	24	15	8
146	0	11	20	13	8	13	10
147	17	7	26	6	12	6	15
148	0	53	47	78	48	77	18
149	57	24	11	26	12	26	23
150	2	12	11	12	11	12	17
151	24	8	13	3	19	3	14
152	20	11	16	7	15	7	15
153	2	17	27	11	28	11	19
154	0	8	24	7	9	7	17
155	80	4	10	5	6	4	7
156	59	18	8	20	14	19	11
157	2	7	6	7	10	7	11
158	25	13	12	25	18	24	14
159	55	47	53	58	62	58	84
160	4	47	53	58	62	58	84
161	35	12	10	17	20	16	11
162	12	8	9	10	15	10	9
163	1	4	21	3	9	3	31
164	3	69	34	47	118	46	47
165	100	28	29	20	37	20	34
166	0	56	41	70	89	70	40
167	17	5	8	5	11	5	27
168	1	2	3	1	4	1	15
169	0	5	9	6	12	5	36
170	10	6	13	6	6	6	13
171	69	31	61	26	24	26	62
172	16	35	13	46	19	46	13
173	0	19	29	20	13	20	11
174	11	59	37	62	95	61	17

175	3	24	24	17	42	17	9
176	0	15	12	8	28	8	3
177	1	1	11	0	2	1	7
178	2	17	20	18	22	18	14
179	0	6	7	3	14	3	7
180	28	5	18	4	8	4	9
181	0	1	8	1	2	1	14
182	7	10	9	8	10	7	9
183	8	38	21	47	34	47	11
184	8	21	20	22	35	22	15
185	0	10	7	6	11	6	8
186	2	12	12	11	12	11	14
187	8	19	14	15	16	15	20
188	2	2	7	2	3	2	6
189	0	5	5	4	7	3	4
190	58	6	14	4	12	4	15
191	21	8	12	7	8	7	8
192	2	14	16	12	11	12	12
193	0	11	10	7	8	7	8
194	0	9	8	6	7	6	5
195	7	13	27	10	20	10	14
196	2	15	17	12	11	12	14
197	0	2	9	1	3	1	7
198	0	3	9	2	3	2	7
199	5	13	16	15	17	15	6
200	402	130	107	138	118	139	99
201	80	38	28	44	40	44	25
202	0	41	23	66	37	66	16
203	0	36	27	43	39	42	23
204	2	35	26	42	39	42	22
205	36	18	27	19	25	18	23
206	19	76	60	56	48	56	75
207	16	51	63	48	49	48	83
208	3	22	23	29	23	29	17
209	1	20	21	28	22	28	15
210	12	38	28	44	40	44	25
211	49	46	27	48	39	48	26
212	3	34	20	41	35	41	15
213	16	43	16	49	21	49	23
214	3	162	107	144	118	145	103
215	2	3	11	2	9	2	18
216	10	12	18	14	21	14	15
217	0	24	28	21	25	21	27
218	60	160	106	143	118	144	100
219	848	429	148	428	156	430	72
220	360	156	103	142	117	143	95
221	15	86	98	79	88	80	86
222	1	27	23	32	23	33	19
223	13	26	22	32	23	32	17
224	0	10	19	9	21	9	13
225	1	28	16	37	33	37	10
226	2	114	40	108	44	108	20
227	6	72	40	90	44	90	20
228	8	22	26	20	24	20	24
229	163	56	40	78	44	78	20
230	2	14	18	19	11	19	9
231	1	3	13	2	4	2	19
232	3	10	13	8	12	8	18
233	16	11	8	8	10	8	8
234	12	2	4	2	5	2	8

235	10	12	11	11	14	11	13
236	5	4	5	3	7	3	15
237	0	3	6	2	7	2	19
238	68	12	11	12	14	11	14
239	3	12	11	17	19	16	19
240	1	10	13	8	12	8	18
241	12	3	4	3	4	3	8
242	2	3	3	3	4	3	6
243	2	3	5	2	5	2	15
244	2	3	4	3	5	3	9
245	0	4	5	3	5	3	12
246	1	12	15	12	16	12	30
247	79	10	13	12	16	11	24
248	58	6	13	5	16	5	26
249	0	4	6	3	7	3	16
250	2	18	10	13	14	13	12
251	0	13	8	17	13	17	8
252	2	13	12	14	11	13	16
253	0	42	23	49	35	49	13
254	0	59	23	84	36	84	14
255	20	17	16	24	24	23	9
256	1	21	25	20	24	20	22
257	0	58	31	115	40	115	12
258	0	13	8	17	13	17	8
259	16	71	80	69	96	69	44
260	67	14	16	16	20	16	10
261	0	23	20	30	22	30	15
262	1	52	97	43	88	43	84
263	35	20	17	28	21	28	12
264	0	23	27	21	25	21	26
265	510	75	86	75	84	75	67
266	19	8	10	7	11	7	12
267	0	5	7	5	9	5	8
268	5	13	17	11	14	11	28
269	5	1	5	0	2	1	12
270	4	8	23	5	10	5	40
271	68	94	107	80	61	80	57
272	8	1	5	1	3	1	10
273	2	21	21	26	18	26	11
274	74	24	24	27	19	27	13
275	25	24	25	28	20	28	14
276	2	17	28	16	20	16	18
277	3	16	26	15	20	15	15
278	2	16	27	16	20	16	16
279	16	19	17	18	11	18	16
280	9	6	13	5	6	5	13
281	1	8	15	7	7	7	12
282	0	1	4	1	2	1	4
283	2	5	7	4	5	3	8
284	42	10	15	9	12	9	13
285	17	5	14	5	9	5	10
286	2	18	15	22	14	22	12
287	7	3	7	2	4	2	8
288	39	8	14	11	16	11	12
289	0	1	6	1	2	1	6
290	0	16	13	21	13	20	9
291	0	11	9	17	12	17	5
292	1	8	13	9	8	9	9
293	20	5	14	4	8	4	13
294	6	25	49	16	50	16	46

295	3	6	9	5	9	5	6
296	59	6	19	6	11	6	9
297	0	17	15	17	12	17	12
298	2	24	25	28	20	28	14
299	0	5	15	4	4	4	13
300	20	3	20	1	8	2	43

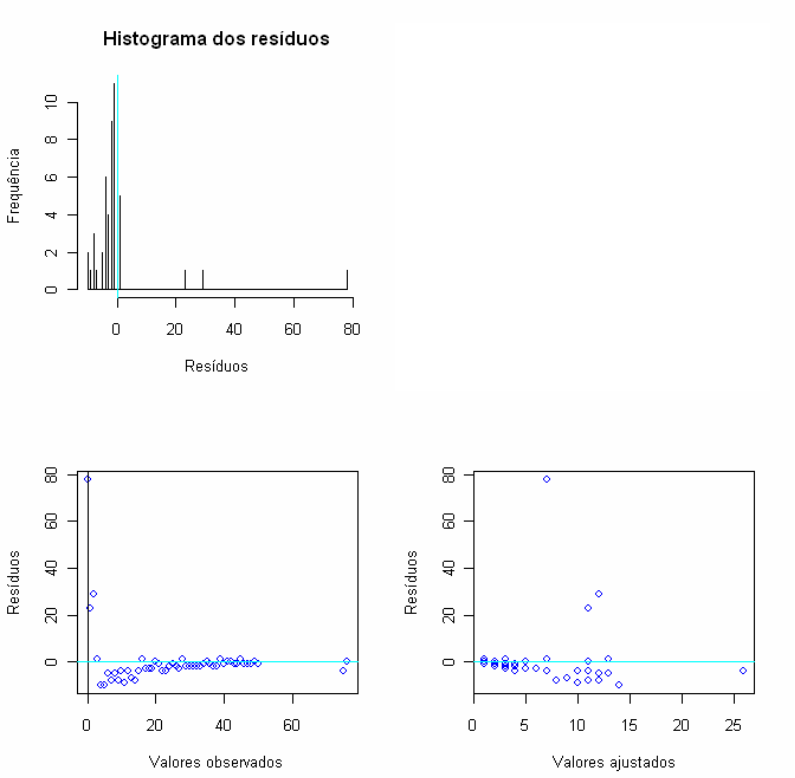
ANEXO 2 – ANÁLISE DOS RESIDUOS DE PEARSON.

A2 - 1 - NÚMERO DE INDIVÍDUOS QUE APRESENTA UMA DETERMINADA QUANTIDADE DE OVOS EM 10 ML DE URINA

Segue-se uma análise dos resíduos de Pearson não normalizados de todos os modelos.

GLM POISSON (GLM POISSON)

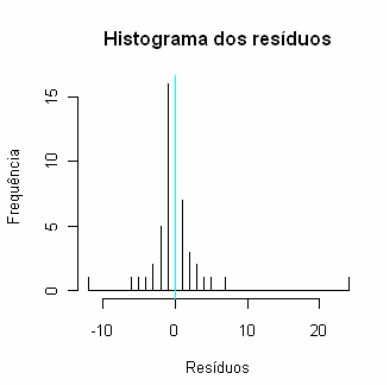
Figura 23 – Análise de resíduos do modelo GLM Poisson.



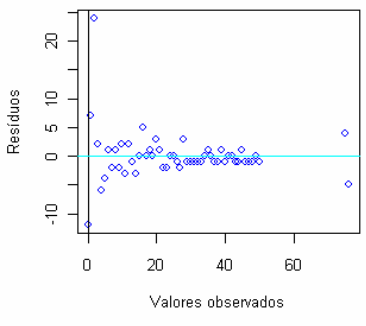
Média dos resíduos 0
Percentagem de resíduos entre -2 e 2 41,5%
Percentagem de resíduos não negativos 26,4%

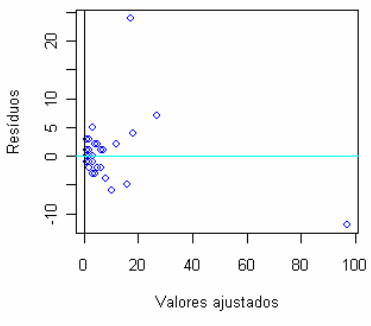
GLM BINOMIAL NEGATIVA (GLM NB)

Figura 24 – Análise de resíduos do modelo GLM binomial negativa.



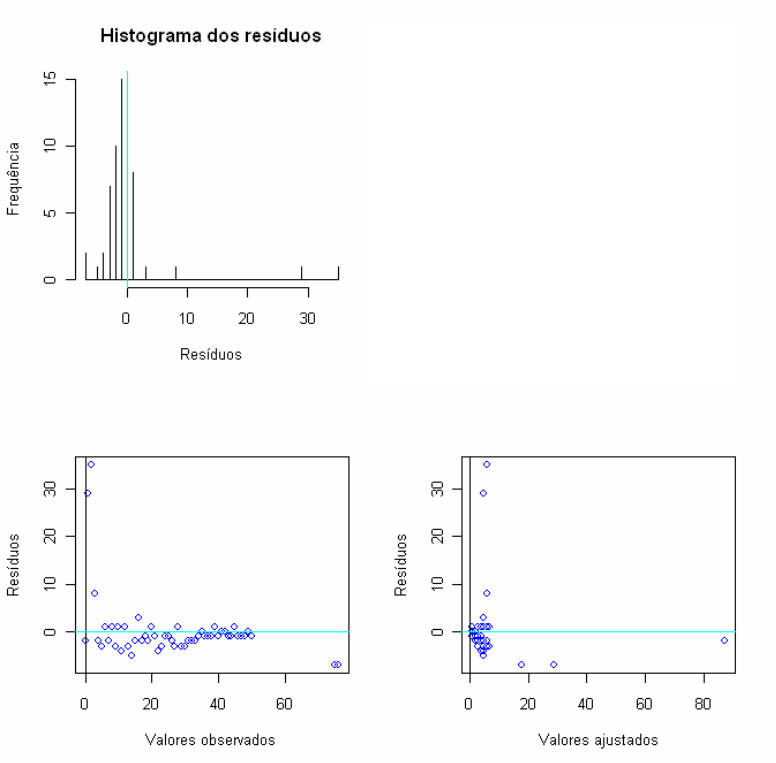
Média dos resíduos
0
Percentagem de resíduos entre -2 e 2
62,3%
Percentagem de resíduos não negativos
49,1%





POISSON COM EXCESSO DE ZEROS (ZIP)

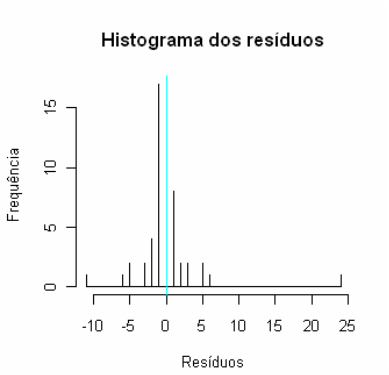
Figura 25 – Análise de resíduos do modelo Poisson com excesso de zeros.



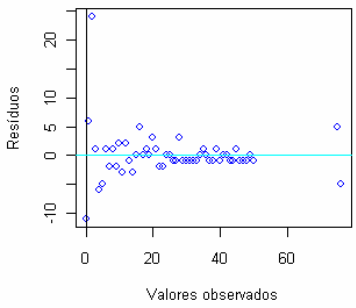
Média dos resíduos
0
Percentagem de resíduos entre -2 e 2
50,9%
Percentagem de resíduos não negativos
30,2%

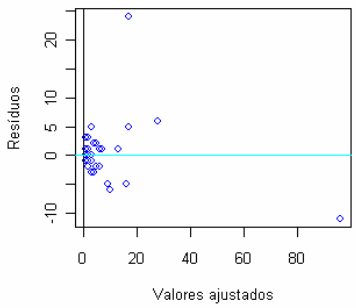
BINOMIAL NEGATIVA COM EXCESSO DE ZEROS (ZINB)

Figura 26 – Análise de resíduos do modelo binomial negativa com excesso de zeros.



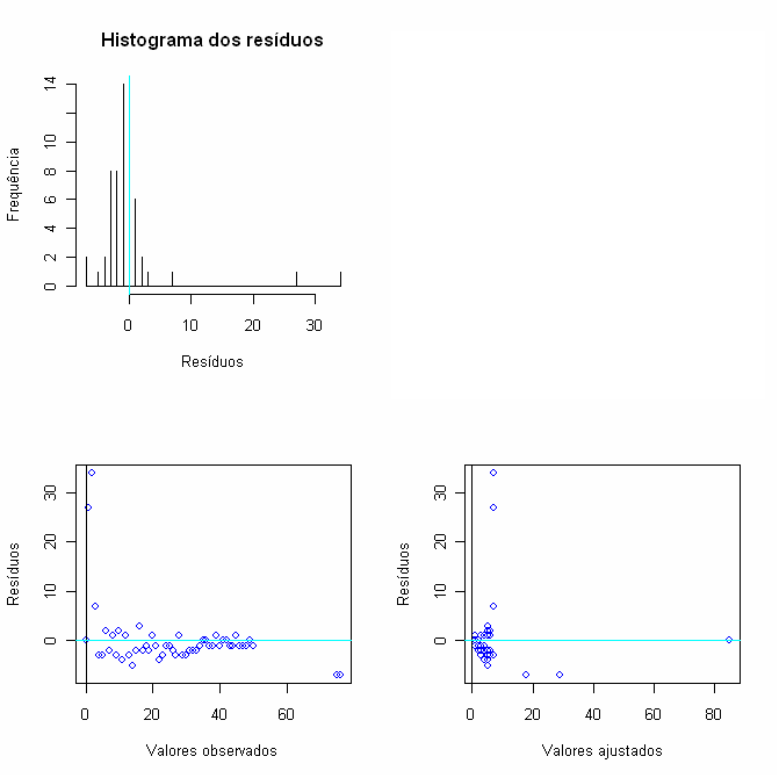
Média dos resíduos
0
Percentagem de resíduos entre -2 e 2
66,0%
Percentagem de resíduos não negativos
49,1,3%





MODELO DE DUAS PARTES COM POISSON (ZAP)

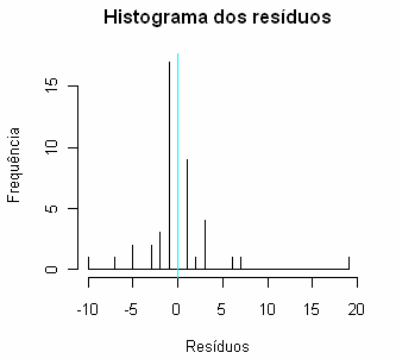
Figura 27 – Análise de resíduos do modelo de duas partes com Poisson.



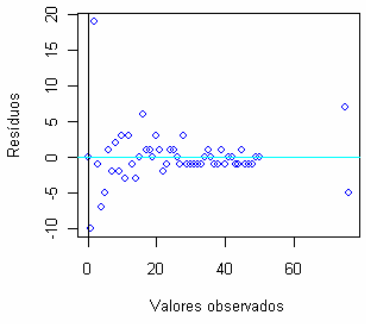
Média dos resíduos
0
Percentagem de resíduos entre -2 e 2
49,1%
Percentagem de resíduos não negativos
34,0%

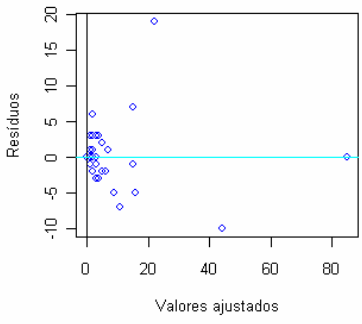
MODELO DE DUAS PARTES COM BINOMIAL NEGATIVA (ZANB)

Figura 28 – Análise de resíduos do modelo de duas partes com binomial negativa.



Média dos resíduos -0,019
Percentagem de resíduos entre -2 e 2 67,9%
Percentagem de resíduos não negativos 50,9%





Não é de esperar que os resíduos de Pearson não normalizados satisfaçam a condição que Dupont (2002) sugere que para a sua validação (95% dos resíduos dentro do intervalo (-2,2)). Apesar disso, todos os modelos com binomial negativa apresentam uma maior percentagem de resíduos dentro do intervalo (-2,2) que os modelos com Poisson.

Também se observa que a percentagem de resíduos não negativos nos modelos de Poisson é pequena se compararmos com a mesma percentagem nos modelos de binomial negativa. Estes últimos apresentam percentagens muito próximas de 50%. Os modelos de Poisson apresentam os seguintes valores: 26,4% (GLM Poisson), 30,2% (ZIP) e 34,0% (ZAP)

Os modelos ZA exibem maior percentagem de resíduos dentro do intervalo (-2,2), seguidos pelos modelos ZI e GLM.

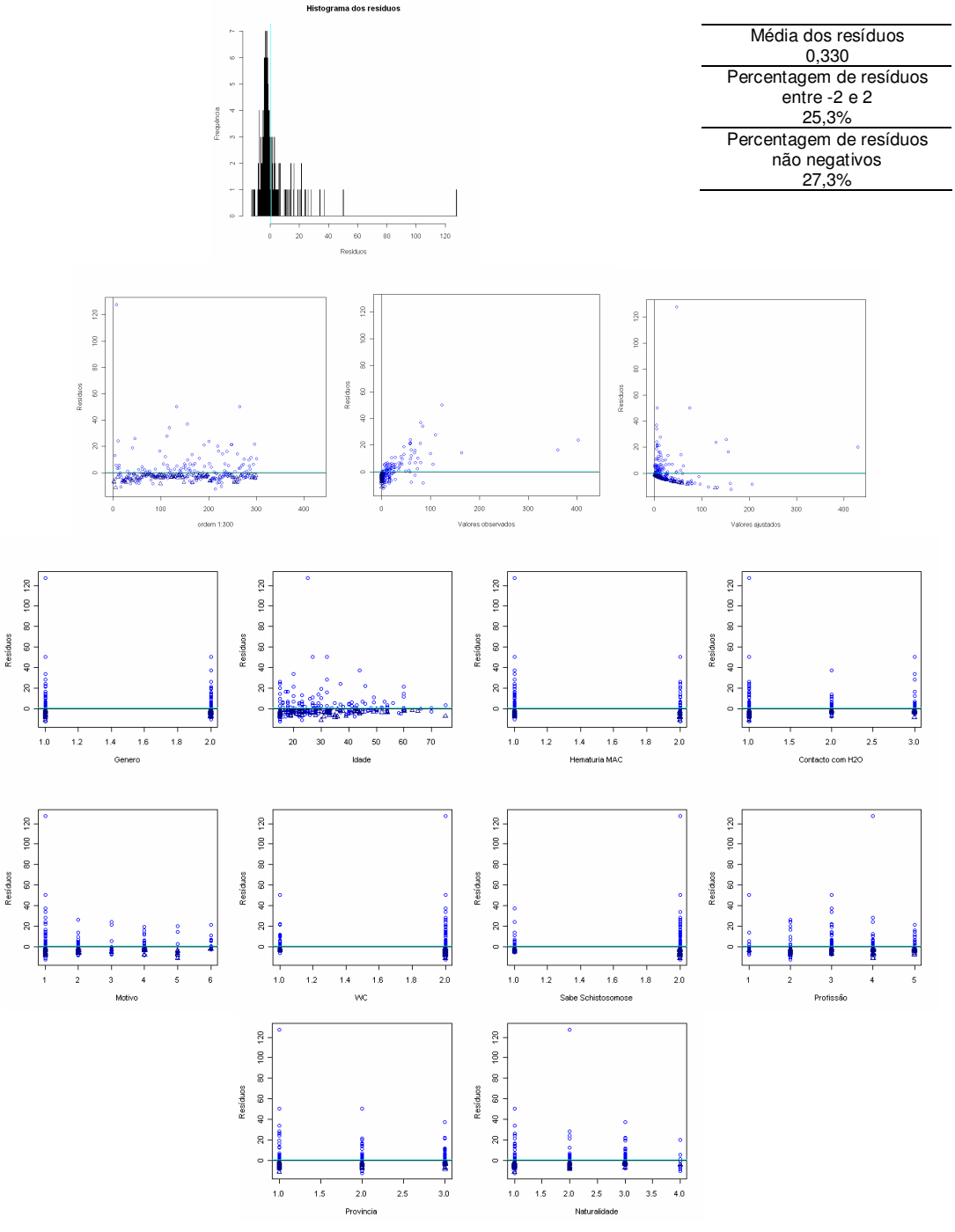
Esta análise parece favorecer os modelos com binomial negativa e dentro destes, o modelo ZANB.

A2 - 2 - NÚMERO DE OVOS EM 10 ML DE URINA PARA CADA INDIVÍDUO DA AMOSTRA

Seguidamente será apresentada uma análise dos resíduos de Pearson normalizados para estes valores.

GLM POISSON (GLM POISSON)

Figura 29 – Análise de resíduos do modelo GLM Poisson.

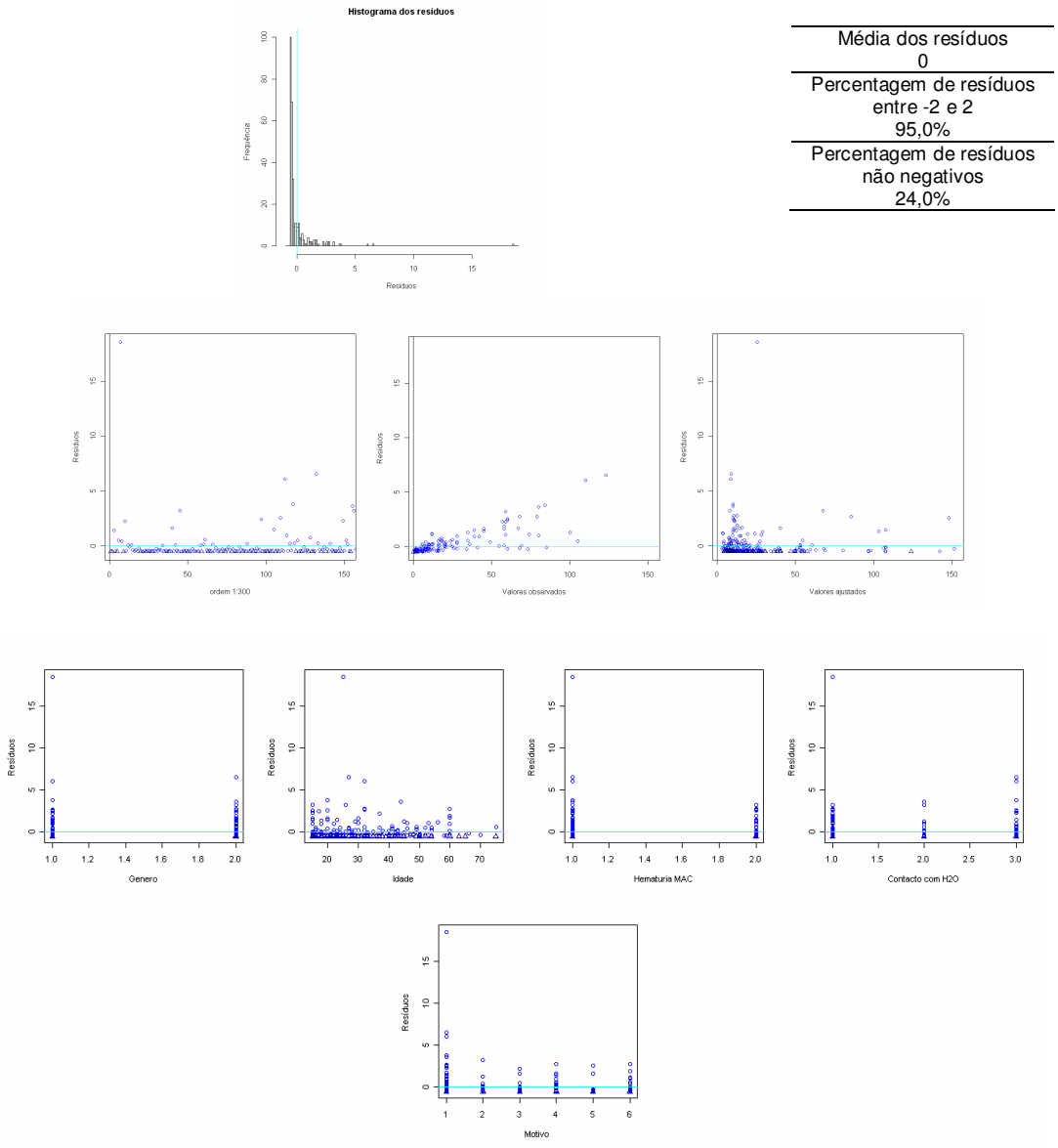


Média dos resíduos 0,330
Percentagem de resíduos entre -2 e 2 25,3%
Percentagem de resíduos não negativos 27,3%

Média dos resíduos Resíduos de observ. zero Resíduos de observ. não nula

GLM BINOMIAL NEGATIVA (GLM NB)

Figura 30 – Análise de resíduos do modelo GLM binomial negativa.

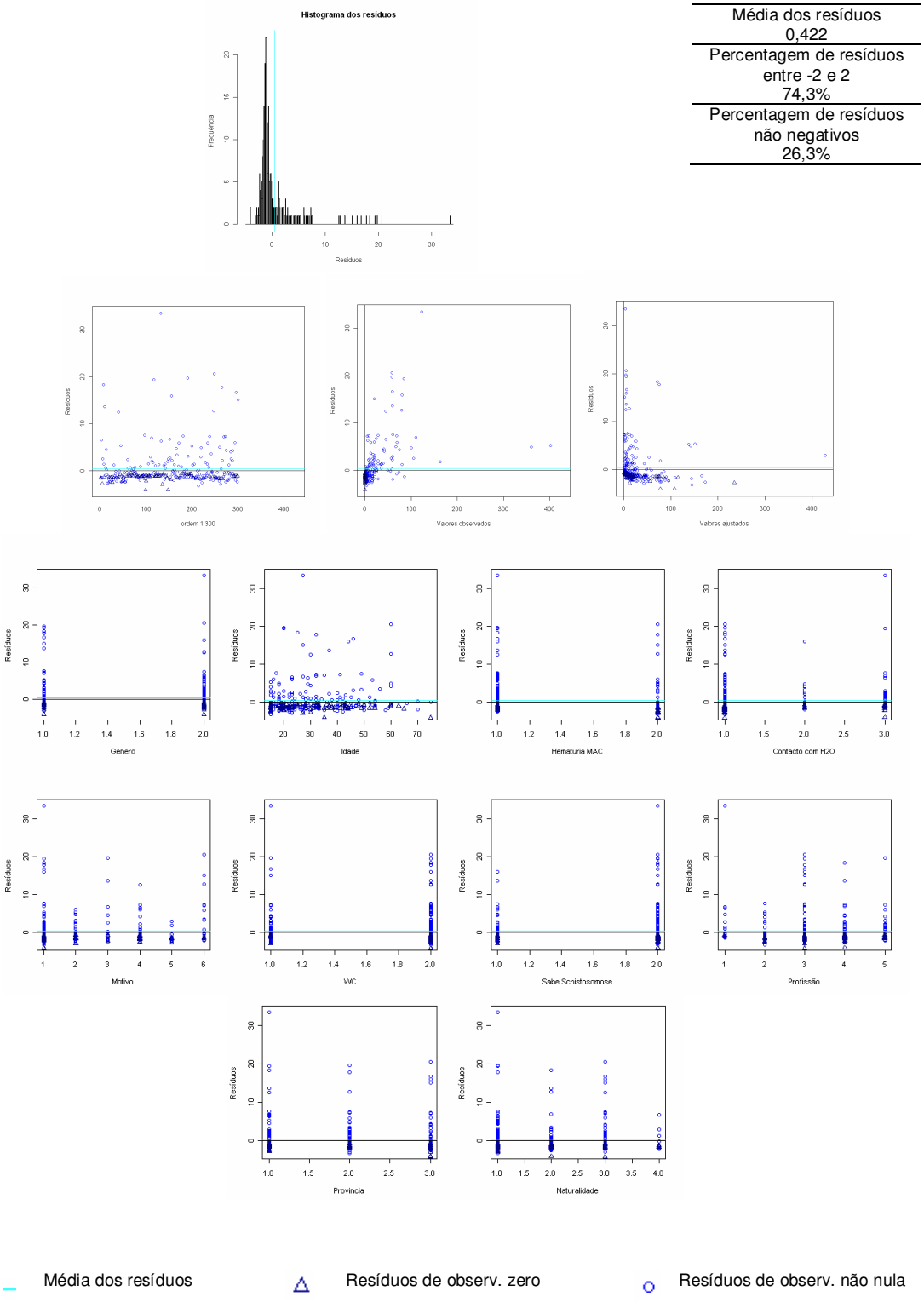


Média dos resíduos 0
Percentagem de resíduos entre -2 e 2 95,0%
Percentagem de resíduos não negativos 24,0%

Média dos resíduos Resíduos de observ. zero Resíduos de observ. não nula

POISSON COM EXCESSO DE ZEROS (ZIP)

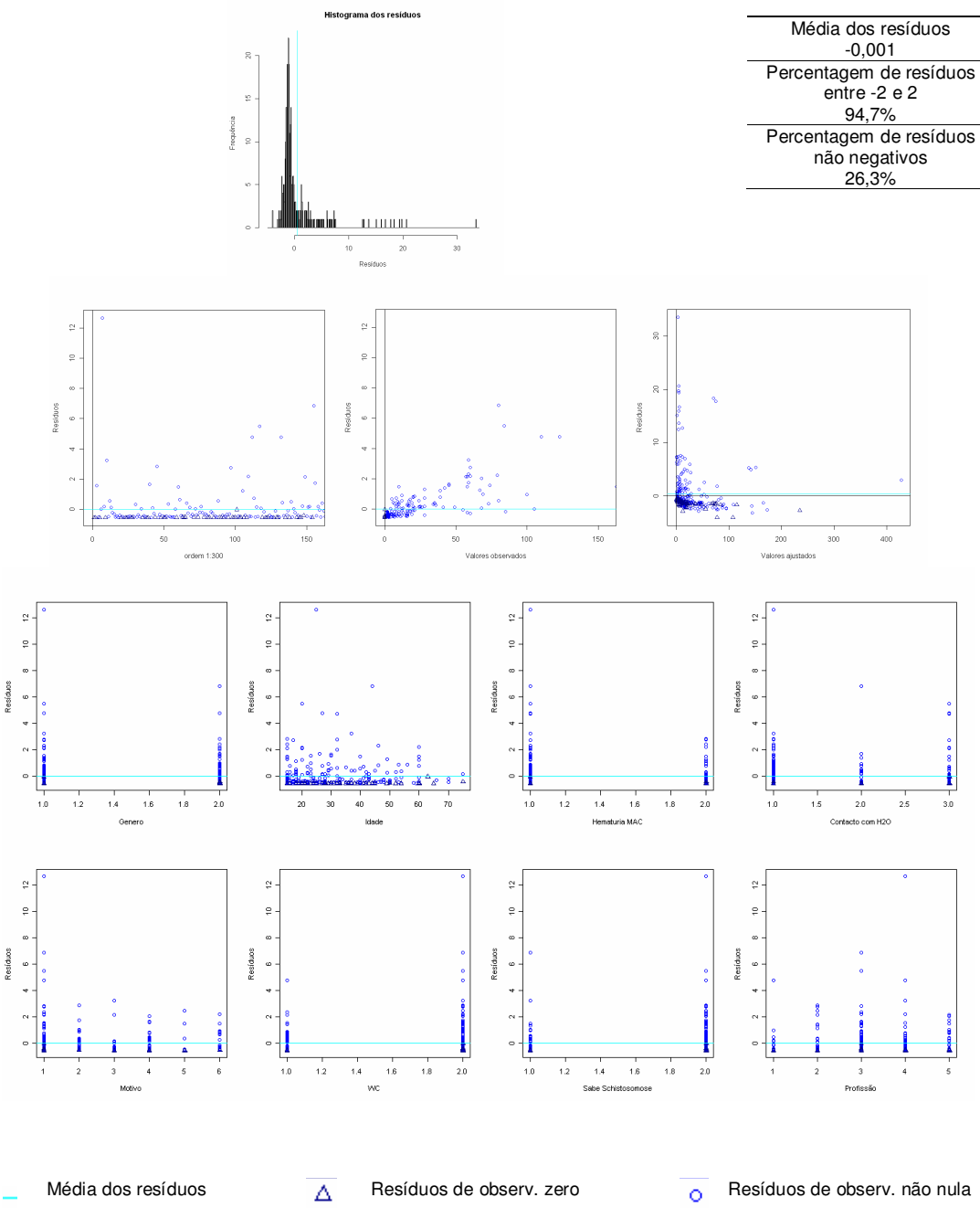
Figura 31 – Análise de resíduos do modelo Poisson com excesso de zeros.



Média dos resíduos
0,422
Percentagem de resíduos entre -2 e 2
74,3%
Percentagem de resíduos não negativos
26,3%

BINOMIAL NEGATIVA COM EXCESSO DE ZEROS (ZINB)

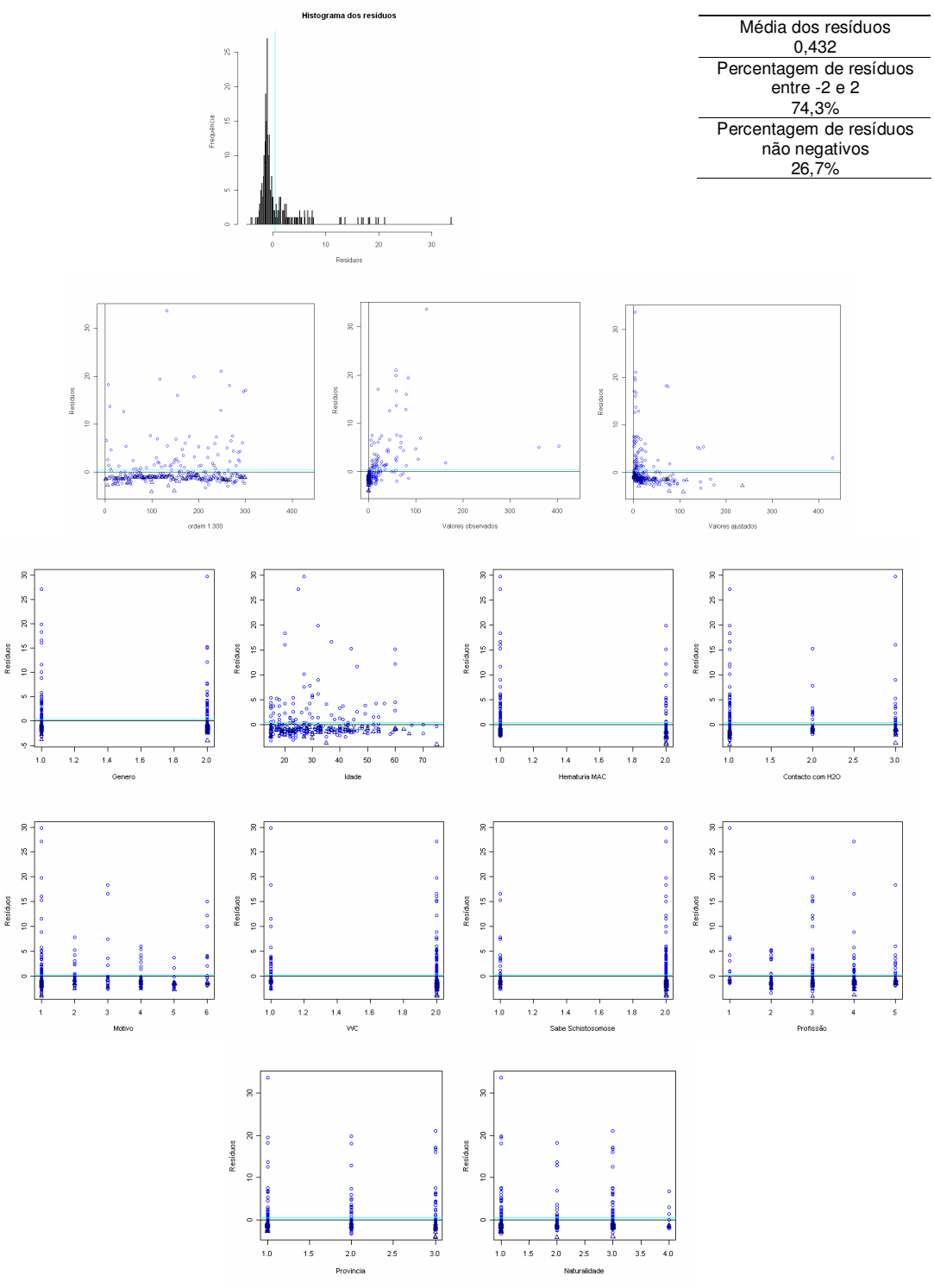
Figura 32 – Análise de resíduos do modelo binomial negativa com excesso de zeros.



Média dos resíduos -0,001
Percentagem de resíduos entre -2 e 2 94,7%
Percentagem de resíduos não negativos 26,3%

MODELO DE DUAS PARTES COM POISSON (ZAP)

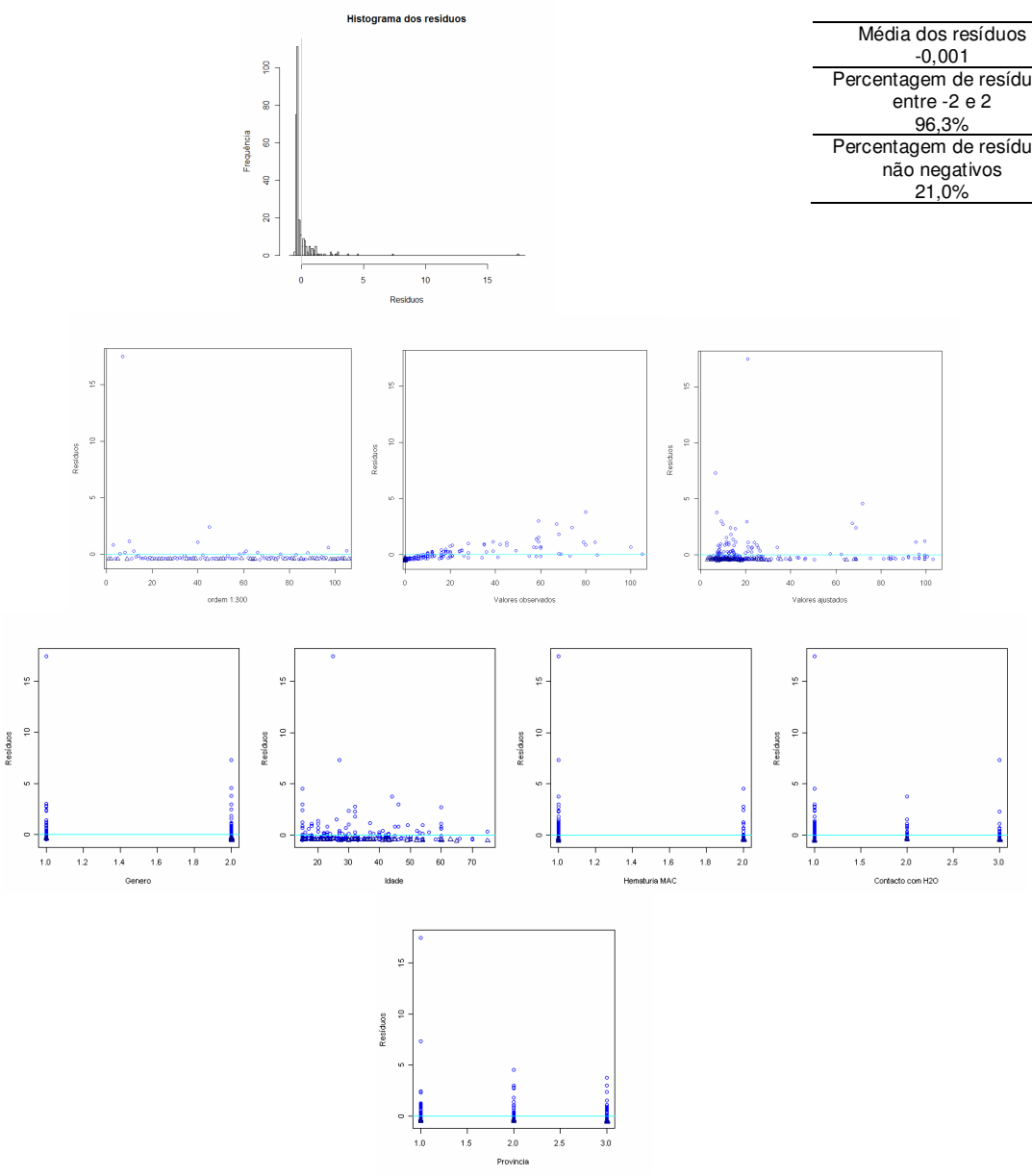
Figura 33 – Análise de resíduos do modelo de duas partes com Poisson.



Média dos resíduos
0,432
Percentagem de resíduos entre -2 e 2
74,3%
Percentagem de resíduos não negativos
26,7%

MODELO DE DUAS PARTES COM BINOMIAL NEGATIVA (ZANB)

Figura 34 – Análise de resíduos do modelo de duas partes com binomial negativa.



Média dos resíduos -0,001
Percentagem de resíduos entre -2 e 2 96,3%
Percentagem de resíduos não negativos 21,0%

Todos os modelos com a distribuição de Poisson apresentam mais de 5% de resíduos com valores fora do intervalo (-2,2), o que não se verifica com os modelos de distribuição binomial negativa.

O modelo ZINB tem uma percentagem de 94,7% de resíduos dentro do intervalo. O modelo ZANB apresenta 96,3% dos resíduos no intervalo e o modelo GLM binomial negativa tem 95% dos resíduos no intervalo.

Nos modelos com distribuição binomial negativa foi os resíduos negativos apresentarem valores pequenos em módulo (todos menores que 0,5) enquanto os resíduos positivos não apresentaram esse comportamento. Em todos os modelos os resíduos positivos apresentam uma maior amplitude que os resíduos negativos mas é nos modelos com distribuição binomial negativa onde esta diferença é mais evidente.

A análise destes resíduos reforça a conclusão que os modelos com distribuição binomial negativa têm uma melhor performance que os que usam a distribuição de Poisson. Para escolher entre os três modelos com a distribuição binomial negativa, a análise dos resíduos não se mostrou muito conclusiva.