Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Cientifica
Departamento de Estatística

Relatório - Parte 2
Exercício 1

Guilherme Pazian RA:160323
Henrique Capatto RA:146406
Hugo Calegari RA:155738
Leonardo Uchoa Pedreira RA:156231

Professor: Caio Lucidius Naberezny Azevedo

Campinas-SP, 20 de Junho de 2017

Introdução

Em 1990, realizou-se um estudo para compreender a influência de certos fatores na quantidade de infecções auditivas de recrutas americanos. As covariáveis avaliadas foram faixa etária (15-19 anos, 20-24 anos, 25-29 anos), sexo (masculino ou feminino), local onde nada (piscina ou praia), frequência com que pratica natação (ocasional ou freqüente) e o número de infecções relatadas pelo recruta. Neste contexto, o objetivo desta análise é avaliar o impacto de cada fator e suas possíveis interações, na quantidade de infecções auditivas.

Análise descritiva

Pode-se observar, pelas tabelas 1, 2, 3 e 4, as seguintes características:

- A quantidade de recrutas que têm ao menos uma infecção, dentre aqueles que nadam frequentemente, é menor do que para o grupo que nada ocasionalmente;
- A quantidade de recrutas que têm pelo menos uma infecção, para aqueles que costumam nadar na praia, é menor do que para o grupo que não costumam nadar na praia;
- A quantidade de recrutas que têm pelo menos uma infecção, para a faixa etária 15-19, é maior do que para os grupos das demais faixas. Além disso, o número de recrutas que têm pelo menos uma infecção na faixa etária 20-24 é maior do que a faixa etária 25-29;
- A quantidade de recrutas homens com ao menos uma infecção é maior do que a quantidade de recrutas mulheres com ao menos uma infecção;

Avaliar o comportamento do número de infecções para diferentes fatores e possíveis interações é de fundamental importância para a futura modelagem. A seguir, são apresentados os gráficos de perfis para verificar as possíveis interações entre fatores (que, em geral, será a combinação entre dois destes). Por estes gráficos, é razoável considerar as interações entre a faixa etária e as covariáveis: hábito de nadar, local onde nada e sexo; interações entre sexo e as covariáveis: local onde nada e o hábito de nadar e a interação entre o local onde costuma nadar e o hábito de nadar.

Pelas figuras 4 e 5 é possível notar que o comportamento dos perfis, ao considerar o hábito de nadar, é muito semelhante para o sexo e para local onde se costuma nadar. Consequentemente, pode-se cogitar em ausência de interação entre os fatores hábito de nadar, sexo e local onde se costuma nadar (sem interação de segunda ordem). Por este motivo, parece razoável não considerar no modelo interação entre três covariáveis

Hábito de nadar	Contagem do número de infecções
Frequente	140
Ocasional	258

Tabela 1: Tabela com os totais de infecções para os diferentes níveis de hábito de nadar.

Local	Contagem do número de infecções
Praia	155
Piscina	243

Tabela 2: Tabela com os totais infecções para os diferentes níveis do lugar onde se costuma nadar.

Faixa etária	Contagem do número de infecções
15-19	223
20-24	92
25-29	83

Tabela 3: Tabela com as quantidades totais de infecções para os diferentes níveis de faixa etária.

Sexo	Contagem do número de infecções
Feminino	131
Masculino	267

Tabela 4: Tabela com as quantidades do número de infecções para os sexos.

Hábito de nadar	Local	Contagem do número de infecções
Frequente	Praia	59
Frequente	Piscina	81
Ocasional	Praia	96
Ocasional	Piscina	162

Tabela 5: Tabela com as quantidades do número de infecções para as combinações dos diferentes hábitos de nadar e local onde se costuma nadar.

Hábito de nadar	Sexo	Contagem do número de infecções
Frequente	Feminino	47
Frequente	Masculino	93
Ocasional	Feminino	84
Ocasional	Masculino	174

Tabela 6: Tabela com as quantidades do número de infecções para as combinações dos diferentes hábitos de nadar e os sexos.

Hábito de nadar	Faixa etária	Contagem do número de infecções
Frequente	15-19	87
Frequente	20-24	33
Frequente	25-29	20
Ocasional	15-19	136
Ocasional	20-24	59
Ocasional	25-29	63

Tabela 7: Tabela com as quantidades do número de infecções para as combinações dos diferentes hábitos de nadar e as diferentes faixas etárias.

Local	Faixa etária	Contagem do número de infecções
Praia	15-19	90
Praia	20-24	20
Praia	25-29	45
Piscina	15-19	133
Piscina	20-24	72
Piscina	25-29	38

Tabela 8: Tabela com as quantidades do número de infecções para as combinações dos diferentes locais onde se costuma nadar e as diferentes faixas etárias.

Local	Sexo	Contagem do número de infecções
Praia	Feminino	81
Praia	Masculino	74
Piscina	Feminino	50
Piscina	Masculino	193

Tabela 9: Tabela com as quantidades do número de infecções para as combinações dos diferentes locais onde se costuma nadar e os sexos.

Sexo	Faixa etária	Contagem do número de infecções
Feminino	15-19	64
Feminino	20-24	37
Feminino	25-29	30
Masculino	15-19	159
Masculino	20-24	55
Masculino	25-29	53

Tabela 10: Tabela com as quantidades do número de infecções para as combinações dos sexos as diferentes faixas etárias.

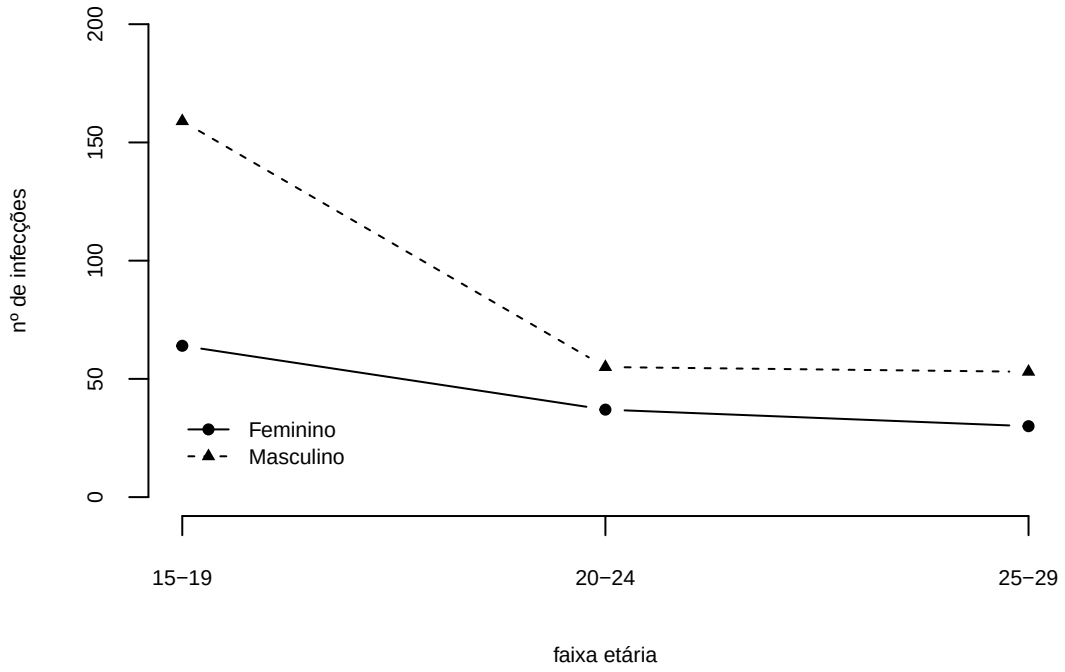

Figura 1: Gráfico de perfil que representa o número de infecções de ouvido para cada sexo. Nota-se que ao mudar de faixa etária há diminuição do número de infecções para ambos os sexos. Observa-se que a redução é maior ao mudar da faixa etária de 15-19 para 20-24 para o sexo masculino. Pelo gráfico, pode-se pensar que existe interação entre as covariáveis faixa etária e sexo.

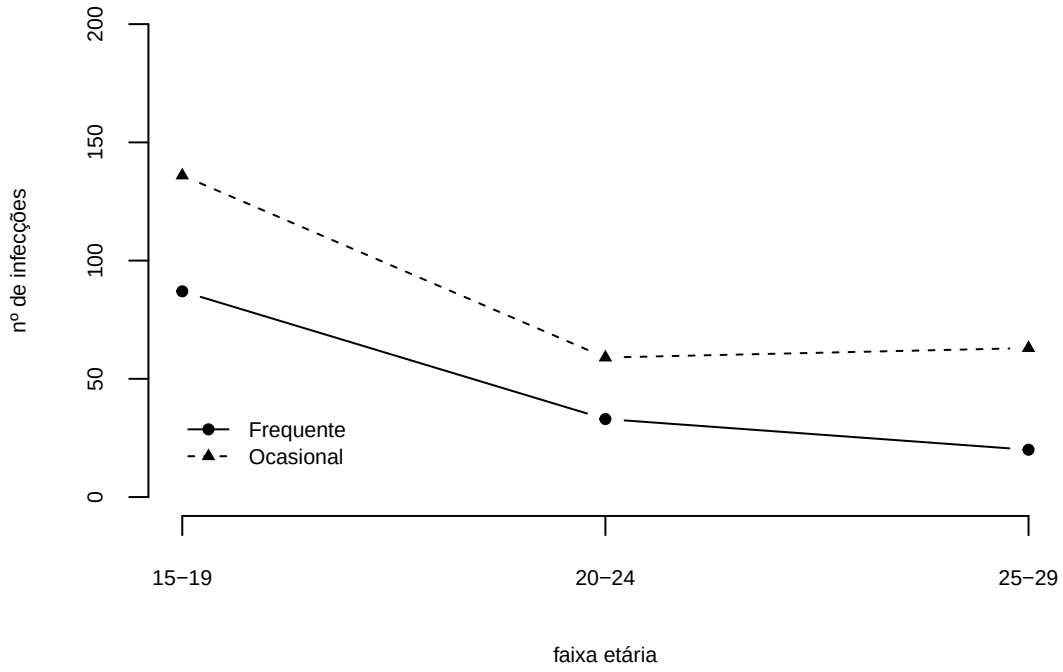

Figura 2: Gráfico de perfil que representa o número de infecções para os grupos que têm hábitos (frequente e ocasional) de nadar diferentes. Para estes grupos, nota-se que o número de infecções diminui. No entanto, para aqueles cujo hábito é ocasional, a quantidade de infecções é superior para diferentes níveis de faixa etária. Portanto, cogita-se que possa existir interação entre o hábito de nadar e a faixa etária.

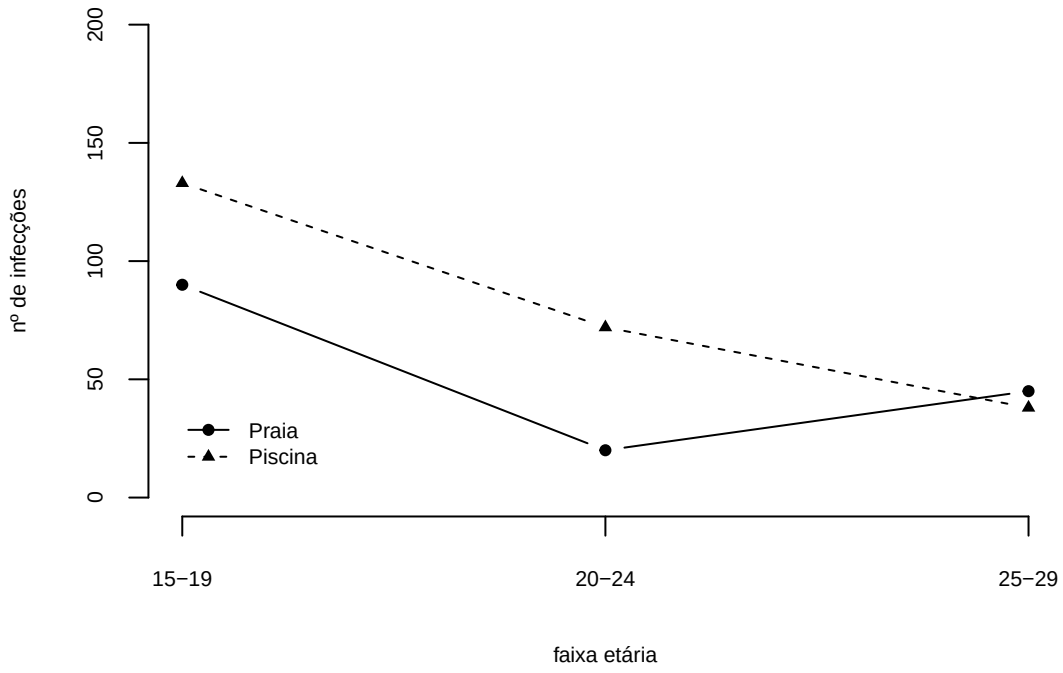

Figura 3: Gráfico de perfil que representa o número de infecções para os indivíduos que costumam nadar na praia ou na piscina. Com exceção da última faixa etária (25-29), o número de infecções para indivíduos que costumam nadar em piscinas é maior para as faixas etárias 15-19 e 20-24. Também observa-se, nas faixas etárias 15-19 e 20-14, que para diferentes lugares onde se costuma nadar, as retas possuem um pequeno desnível que as tornam não paralelas. No entanto, para a última faixa etária, nota-se que o comportamento é diferente para os diferentes locais onde se costuma nadar. Assim, considera-se interação entre a faixa etária e o local onde se costuma nadar.

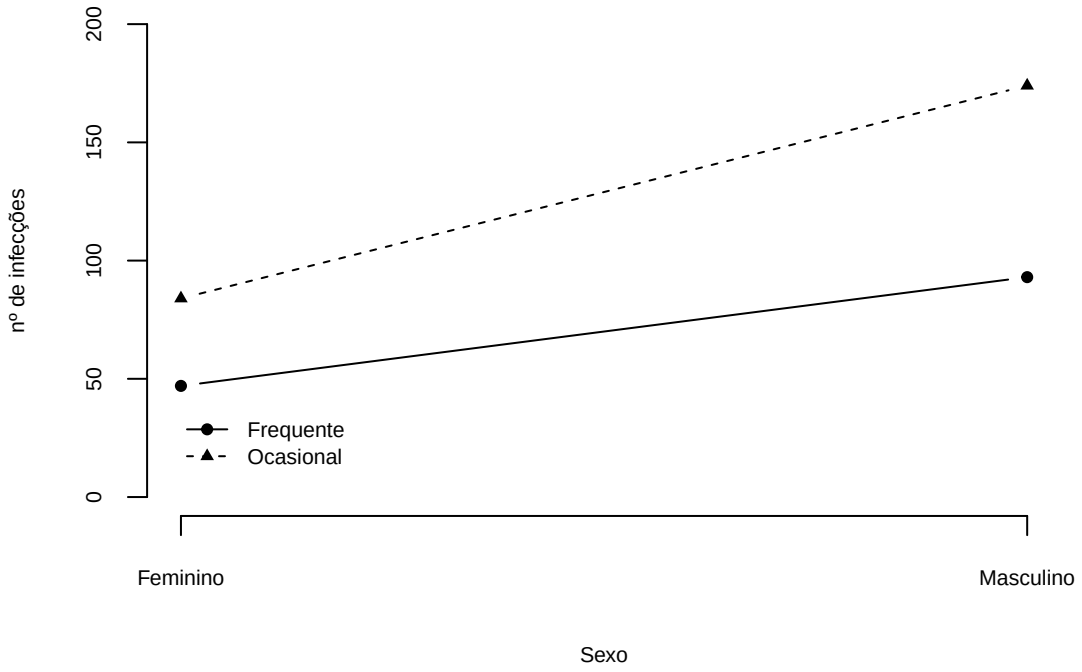

Figura 4: Gráfico de perfil que representa o número de infecções para os indivíduos que tem um determinado hábito de nadar, para os diferentes sexos. Observa-se que as retas não são paralelas e, portanto, possível existência de interação entre sexo e hábito de nadar.

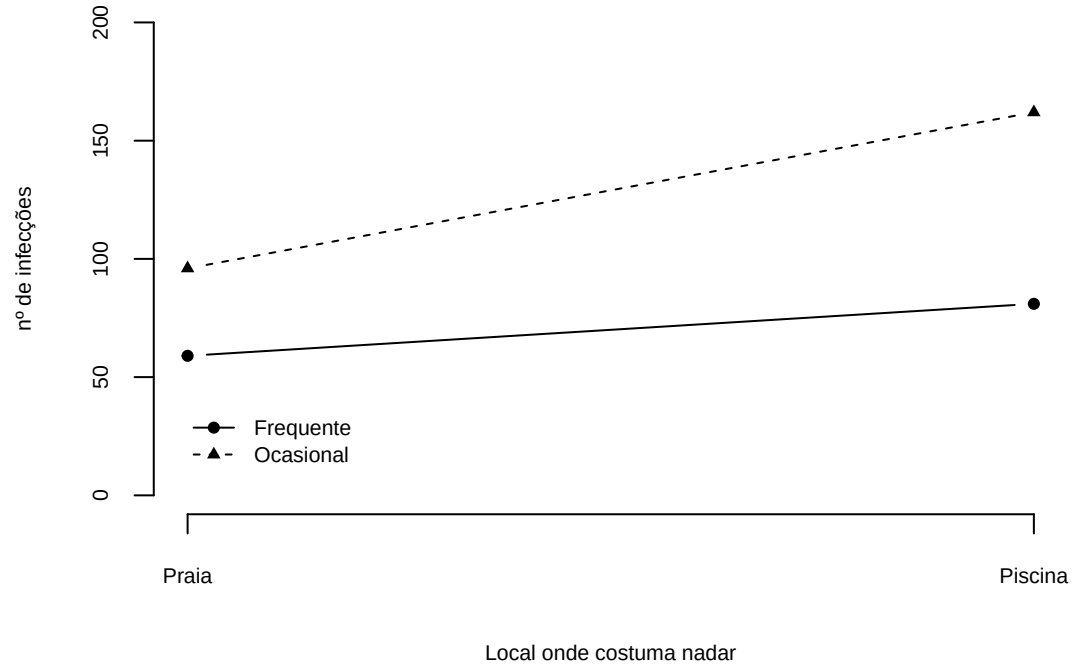

Figura 5: Gráfico de perfil que representa o número de infecções para os indivíduos que tem determinado hábito de nadar e o local que costumam nadar. Como comportamento dos perfis é muito similar ao relatado na figura 4, cogita-se uma possível interação entre local onde costuma nadar e hábito de nadar.

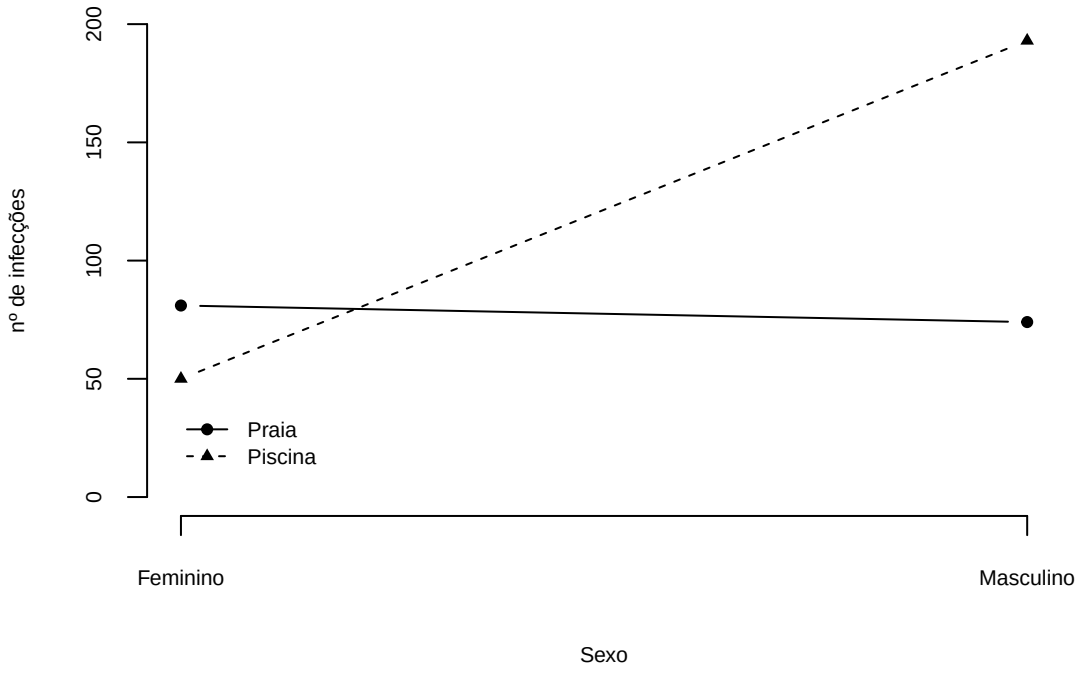

Figura 6: Gráfico de perfil que representa o número de infecções para os sexos, para indivíduos que tem costume de nadar em diferentes lugares. Nota-se que é plusível a interação entre sexo e o lugar onde costuma nadar.

Análise Inferencial e Ajuste de Modelos

Modelo de Poisson

Seja Y_i o número de infecções de ouvido diagnosticadas pelo i-ésimo indivíduo.

$$Y_i \overset{ind.}{\sim} Poisson(\mu_i)$$

O modelo que foi considerado é obtido ao igualar o logarítimo natural da média μ_i para cada indivíduo as covariáveis de interesse.

A estrutura abaixo ilustra um possível modelo (neste caso, o modelo é completo com todas as interações de terceira ordem):

$$ln(\mu_i) = \mu + \alpha O_i + \gamma P_i + \beta_1 (20-24)_i + \beta_2 (25-29)_i + \delta M_i +$$

$$(\alpha\gamma) P_i O_i + (\alpha\beta_1)(20-24)_i O_i + (\alpha\beta_2)(25-29)_i O_i + (\alpha\delta) M_i O_i + (\beta_1\gamma)(20-24)_i P_i +$$

$$(\beta_2\gamma)(25-29)_i P_i + (\delta\gamma) M_i P_i + (\beta_1\delta) M_i (20-24)_i + (\beta_2\delta) M_i (25-29)_i +$$

$$(\alpha\beta_1\gamma)(20-24)_i P_i O_i + (\alpha\beta_2\gamma)(25-29)_i P_i O_i + (\alpha\delta\gamma) M_i P_i O_i + (\beta_1\delta\gamma)(20-24)_i P_i M_i +$$

$$(\beta_2\delta\gamma)(25-29)_i P_i M_i + (\alpha\beta_1\delta\gamma)(20-24)_i P_i O_i M_i + (\alpha\beta_2\delta\gamma)(25-29)_i P_i O_i M_i$$

O_i = 1 se o i-ésimo indivíduo nada ocasionalmente e O_i = 0 se nada frequente.

P_i = 1 se o i-ésimo indivíduo costuma nadar na piscina e P_i = 0 costuma nadar na praia.

$(20-24)_i$ = 1 se o i-ésimo indivíduo pertence à faixa etária 20-24 e $(20-24)_i$ = 0 caso não pertença.

$(25-29)_i$ = 1 se o i-ésimo indivíduo pertence à faixa etária 25-29 e $(25-29)_i$ = 0 caso não pertença.

M_i = 1 se o i-ésimo indivíduo é do sexo masculino e M_i = 0 se do sexo feminino.

Primeiramente será avaliado o modelo completo, isto é, aquele no qual a interação de terceira ordem está presente na especificação das variáveis.

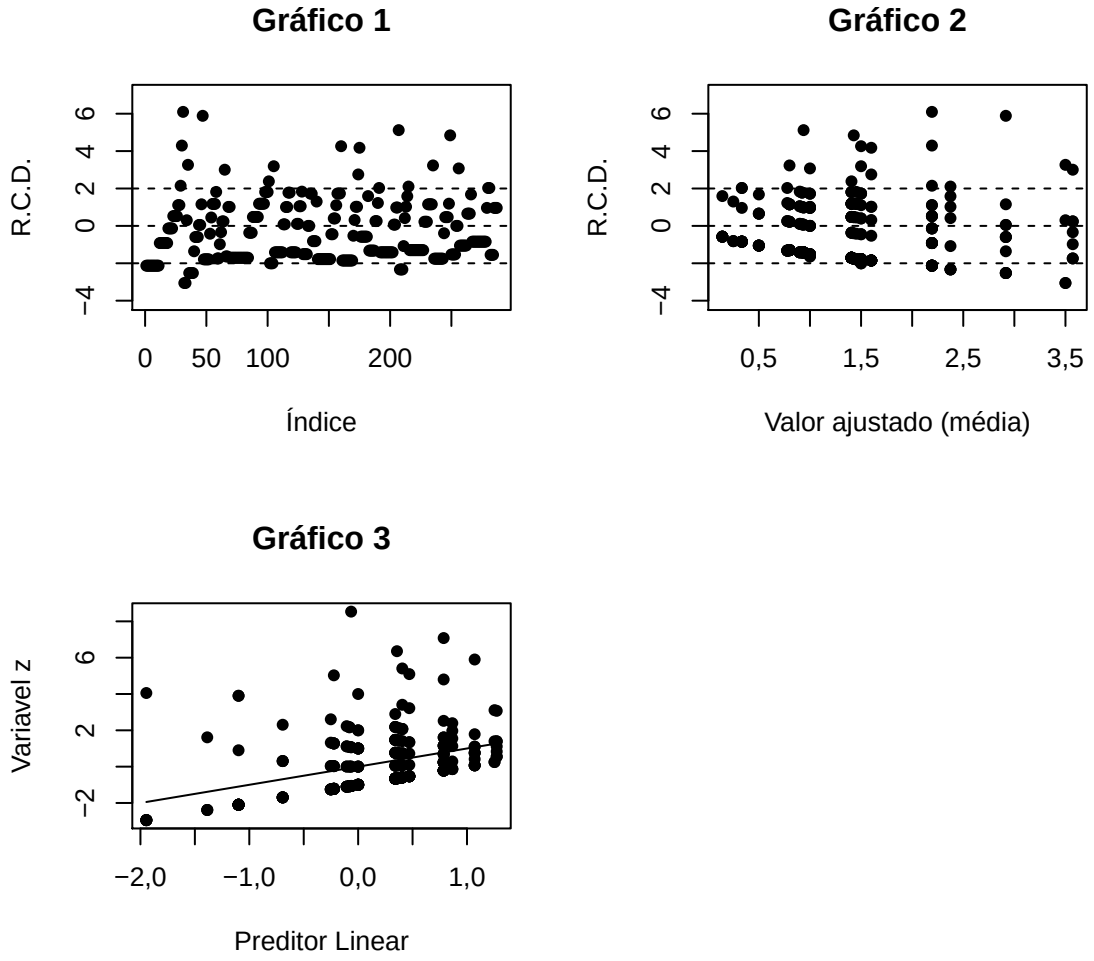

Figura 7: Gráficos para verificar a adequabilidade do ajuste para o modelo completo (note que R.C.D. é o acrônimo para Resíduo Componente do Desvio). Pelo Gráfico 1 nota-se que muitos resíduos estão fora dos limites [-2, 2], ou seja, valores dos resíduos muito extremos. Pelo Gráfico 2, também percebe-se que muitos dos valores ajustados estão fora de [-2,2], característica não procurada para um ajuste razoável. Para o Gráfico 3, observa-se possivelmente problemas com a função de ligação e/ou com o preditor linear, pois muitos pontos não seguem um padrão linear ao redor da reta de referência.

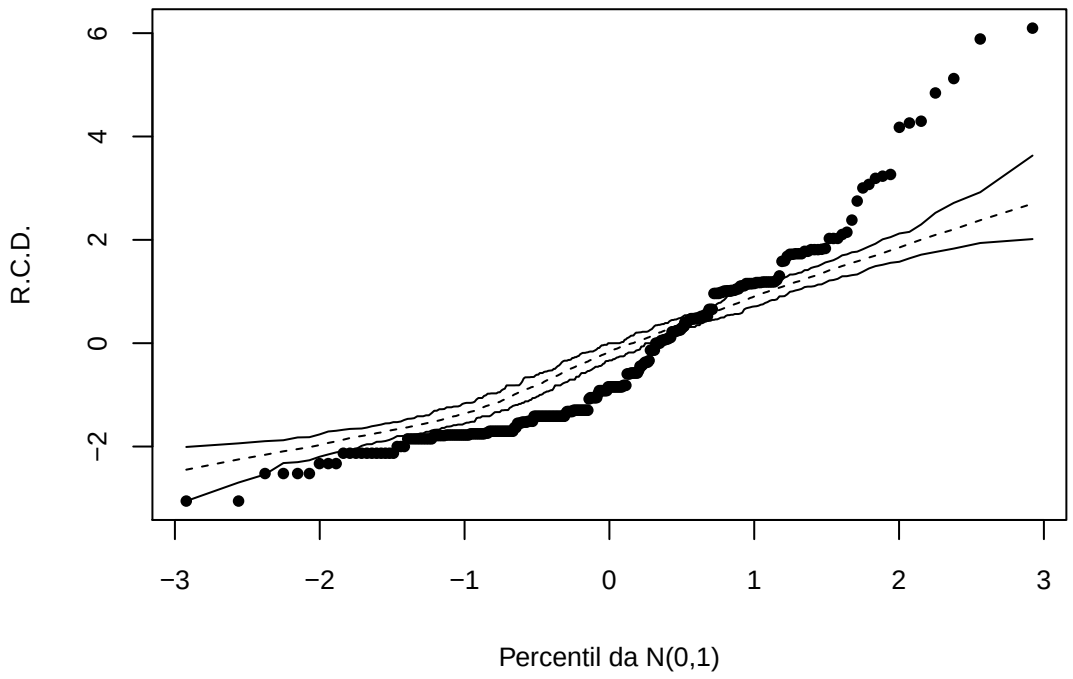

Figura 8: Este gráfico compara os quantís do RCD (ver Paula, 2013) com os quantís de uma distribuição normal padrão, para o modelo completo. Note que muitos dos resíduos estão fora da banda de confiança e bem distantes da linha de referência (tracejada), o que fornece argumentos contra o bom ajuste do modelo. Com isso, pode-se concluir que o ajuste não é razoável para os dados em questão.

Por fim, para este modelo, os valores do desvio observado (ver Paula, 2013) e do AIC (veja Paula, 2013) foram de 703,72, 1124,11, respectivamente.

O modelo ser considerado, agora, é aquele no qual os efeitos principais e as interações de segunda ordem podem ser avaliadas. Pela análise descritiva, notou-se possivelmente a não interação para alguns fatores de segunda ordem. Apesar disso, o modelo com terceira ordem foi avaliado e de seu ajuste, considerou-se que

- A não significância da interação de terceira ordem e de algumas interações de segunda ordem (nota-se que em uma destas não significâncias, foi observada para a interação de hábito de nadar, local onde costuma nadar e sexo - característica já observada na análise descritiva).
- A retirada da interação de terceira ordem é razoável. Agora, tem-se modelos com interações de segunda ordem, que foram avaliados por meio de testes de significância conjunta dos efeitos, por meio de estatísticas $C\beta = M$;

O modelo final obtido levou em consideração todos fatores principais como hábito de nadar, local onde costuma nadar e faixa etária. Algumas interações de primeira ordem foram mantidas, como interações entre hábito de nadar e faixa etária, hábito de nadar e local onde costuma nadar e, por fim, local onde costuma nadar e faixa etária. Para mais informações, veja a tabela 11, na seção Anexos.

Para tal modelo, os valores do desvio e AIC foram 743,78, 1136,18. Nota-se que, embora estes valores sejam maiores em relação ao modelo completo, características de qualidade de ajuste observado pelos gráficos do modelo final se apresentam razoavelmente melhor. É importante lembrar-se que considerar o valor do desvio faz com que características como o preditor e a função de ligação sejam desconsideradas, isto é, o desvio não leva em consideração tais características.

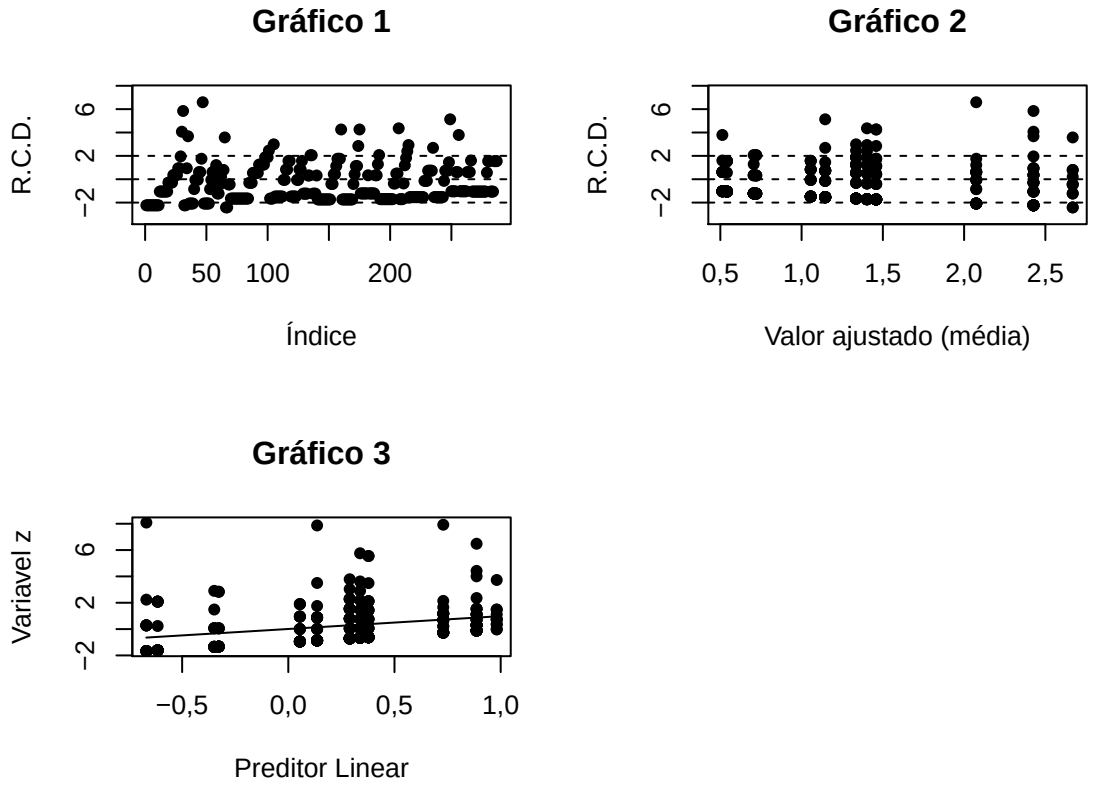

Figura 9: Gráficos para verificar a adequabilidade do ajuste para o modelo reduzido. Pelo Gráfico 1 observa-se que existem muitos pontos que estão fora dos limites -2 e 2, isto é, observações discrepantes. Pelo Gráfico 2, nota-se também muitos pontos fora do intervalo -2 e 2, além de que existem muitos valores ajustados (média) concentrados entre [0,5;1,5] e menos concentrados entre aproximadamente [1,5;2,75]. Isto se deve possivelmente as características dos próprios dados, ou seja, pode-se ter maior concentração de indivíduos cujas características do valor do preditor linear é semelhante e consequentemente, provocar a concentração mencionada. Pelo Gráfico 3, diagnosticou-se que, possivelmente, falta alguma covariável para compor o preditor linear ou há algum problema com a função de ligação.

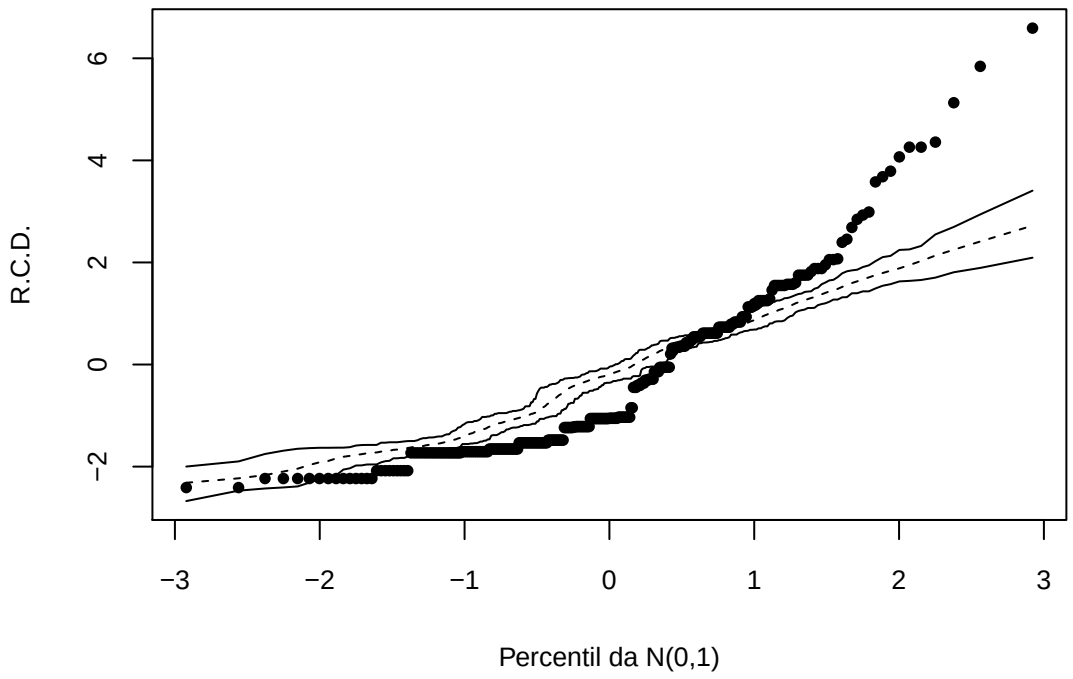

Figura 10: Este gráfico compara os quantís do RCD para o modelo reduzido (ver Paula, 2013). Quanto mais pontos estiverem dentro da banda de confiança (linha cheia) e perto da reta pontilhada (linha de referência), mais indicativos temos de um bom ajuste. Observa-se claramente que o ajuste, portanto, não é razoável.

Para os modelos completo e reduzido, os gráficos de diagnósticos (figuras 7 e 9, respectivamente) e envelopes (figuras 8 e 10, respectivamente) apontam que a falta de ajuste não é muito bem justificada devido à falta de covariáveis no preditor linear, ou pela função de ligação errada. Isto indica que considerar uma estrutura probabilística de Poisson possivelmente não seja a abordagem mais adequada para estes dados.

————————MODIFICADO ATÉ AQUI————————-

Com o apoio da análise descritiva, um possível argumento para a falta de ajuste é a maior quantidade de contagem de infecções de ouvido para determinadas covariáveis, ou seja, variabilidade muito grande na distribuição das contagens de infecções. Note, por exemplo, as tabelas 1, 2 e 3 que representam a contagem do número de infecções de ouvido em relação às variáveis hábito de nadar, local onde costuma nadar e faixa etária. Observa-se que para um determinado nível sobressai-se um determinado valor de contagem.

Com isso, pode-se pensar que existe uma superdispersão dos dados que não é captada pelo modelo proposto e ajustado. Logo, uma possível alternativa seria utilizar um modelo que contemple a superdispersão. Neste caso, será avaliado o ajuste com o modelo binomial negativo.

Modelo Binomial negativo

Observamos, nos modelos anteriores, que nenhum deles foi adequado para os dados, uma vez que as análises de resíduo indicaram um mal ajuste dos modelos do tipo Poisson. Possivelmente, um dos aspectos à influenciar esse mal ajuste dos modelos Poisson é a característica da variável resposta, a qual apresenta média igual a 1,3868 e variância igual a 5,4688, ou seja, a variância observada na variável resposta é quase quatro vezes o valor observado para a média da variável resposta. Uma das imposições do modelo de Poisson é que a média da variável resposta tem a mesma magnitude do valor de sua respectiva variância, porém isso não é observado nos dados referente à variável resposta, de maneira que temos uma situação de superdispersão, ou seja, a variável resposta apresenta variância maior do que aquela imposta pelo modelo probabilístico.

Dadas as observações feitas quanto ao modelo de Poisson, um modelo de regressão construido a partir de uma distribuição binomial negativa talvez seja mais adequado (veja Azevedo 2017) para os dados em questão.

Temos, portando o modelo:

Seja Y_i o número de infecções de ouvido diagnosticadas pelo i-ésimo indivíduo.

$$Y_i \overset{ind.}{\sim} BN(\mu_i, \phi)$$

Ao se modelar o logaritmo de μ_i tem-se a mesma estrutura da observada em Poisson, como segue:

$$ln(\mu_i) = \mu + \alpha O_i + \gamma P_i + \beta_1(20-24)_i + \beta_2(25-29)_i + \delta M_i +$$

$$(\alpha\gamma)P_iO_i + (\alpha\beta_1)(20-24)_iO_i + (\alpha\beta_2)(25-29)_iO_i + (\alpha\delta)M_iO_i + (\beta_1\gamma)(20-24)_iP_i +$$

$$(\beta_2\gamma)(25-29)_iP_i + (\delta\gamma)M_iP_i + (\beta_1\delta)M_i(20-24)_i + (\beta_2\delta)M_i(25-29)_i +$$

$$(\alpha\beta_1\gamma)(20-24)_iP_iO_i + (\alpha\beta_2\gamma)(25-29)_iP_iO_i + (\alpha\delta\gamma)M_iP_iO_i + (\beta_1\delta\gamma)(20-24)_iP_iM_i +$$

$$(\beta_2\delta\gamma)(25-29)_iP_iM_i + (\alpha\beta_1\delta\gamma)(20-24)_iP_iO_iM_i + (\alpha\beta_2\delta\gamma)(25-29)_iP_iO_iM_i$$

O_i = 1 se o i-ésimo indivíduo nada ocasionalmente e O_i = 0 se nada frequente.
P_i = 1 se o i-ésimo indivíduo costuma nadar na piscina e P_i = 0 costuma nadar na praia.
$(20-24)_i$ = 1 se o i-ésimo indivíduo pertence à faixa etária 20-24 e $(20-24)_i$ = 0 caso não pertença.
$(25-29)_i$ = 1 se o i-ésimo indivíduo pertence à faixa etária 25-29 e $(25-29)_i$ = 0 caso não pertença.
M_i = 1 se o i-ésimo indivíduo é do sexo masculino e M_i = 0 se do sexo feminino.

Os gráficos para verificar a qualidade do ajuste binomial negativo para o modelo completo foram obtidos e avaliados como mostrados a seguir. Observe de imediato a diferença dos gráficos de diagnóstico e de envelope. O valor do desvio e do AIC são,

respectivamente, 269,83 e 923. Note que tais valores são menores do que para o modelo de Poisson completo e o modelo reduzido final.

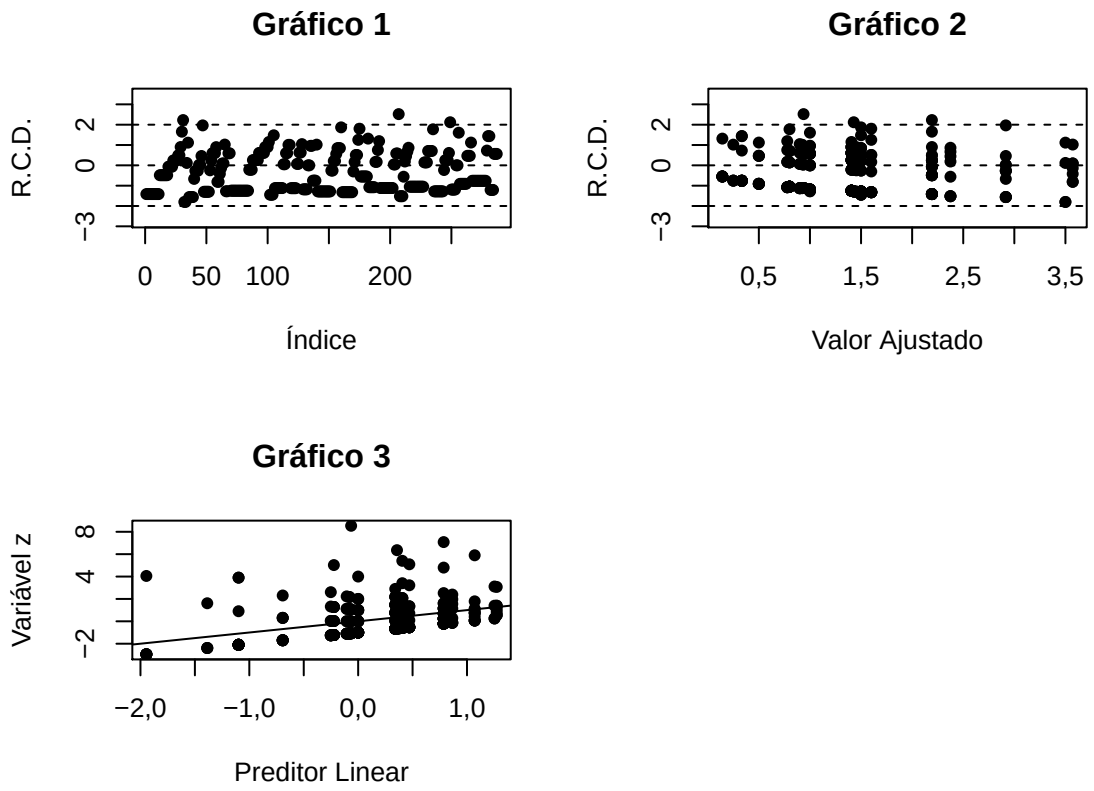

Figura 11: Gráficos para verificar a adequabilidade do ajuste para o modelo completo. Nota-se pelos Gráficos 1 e 2 que são poucos os valores dos resíduos que estão fora do intervalo [-2,2], comportamento diferente em relação ao modelo de Poisson completo. No entanto, quando se observa o Gráfico 3, pode-se notar que o comportamento é muito similar quando se compara com o modelo de Poisson completo. Possivelmente características relacionadas com as covariáveis do preditor linear e/ou a função de ligação influenia nessa tal característica.

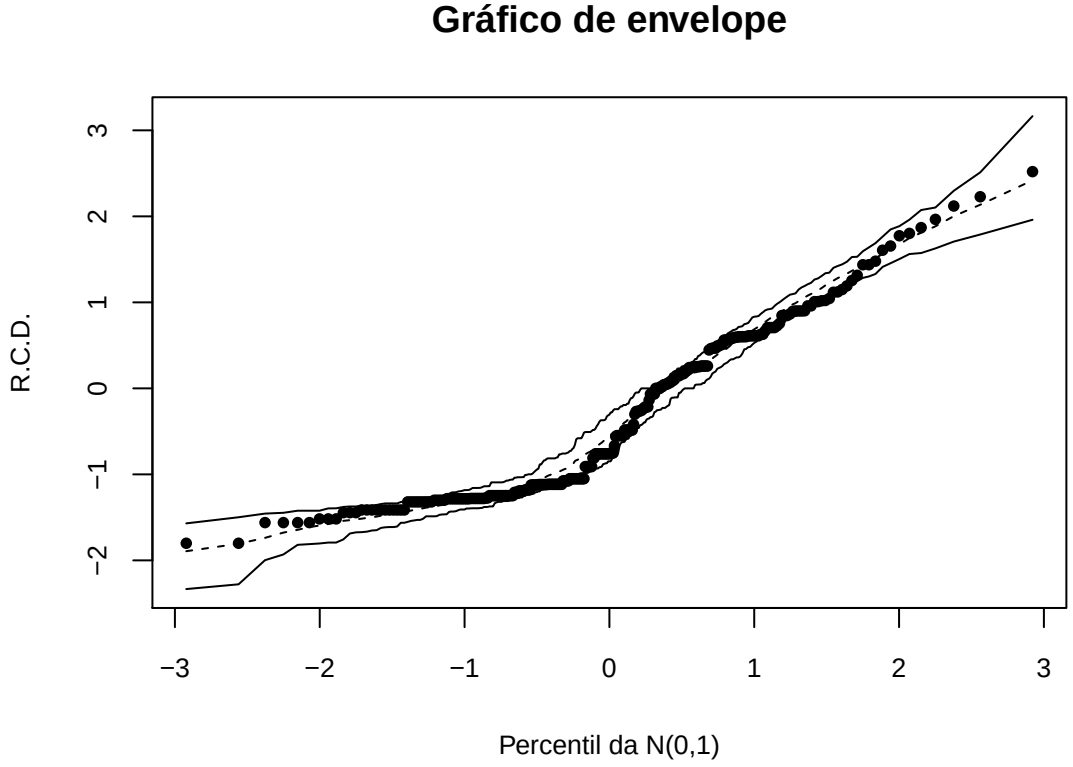

Figura 12: Este gráfico compara os quantis do RCD para o modelo completo. Nota-se que muitos dos resíduos estão dentro da banda de confiança e próximos da linha de referência. Isto evidencia um bom ajuste. Observa-se, além disso, a grande diferença quando comparado com o modelo de Poisson completo.

Conclusão

Constatou-se que o modelo de Poisson, tanto o completo quanto o reduzido, não se ajustaram bem aos dados, segundo o valor do desvio, e das análises de resíduos. Entre os modelos de Poisson, ainda observou-se que o critério de informação (AIC) para o modelo completo é menor comparado com o modelo reduzido final. Foi possível notar que a característica de interesse (número nde infecções de ouvido) apresenta característica de superdispersão, com isso outra abordagem foi utilizada para que se obtivesse uma análise correta dos dados. O novo modelo considerado, baseou-se na binomial negativa, que se ajustou mmelhor tanto no que se diz respeito à análise dos resíduos, assim como nos valores dos desvios e do critério de informação usado.

Bibliografia

1. Azevedo, C. L. N. (2017). Notas de aula sobre análise de dados discretos, (disponível em http://www.ime.unicamp.br/~cnaber/Material_ADD_1S_2017.htm)

2. Paula, G. A. (2013). Modelos de regressão com apoio computacional, versão pré-eliminar, (disponível em http://www.ime.usp.br/~giapaula/texto_2013.pdf)

3. Agresti. A. (2012). Categorical data analysis, terceira edição. New York, John Wiley.

4. Agresti. A. (2007). An introduction to categorical data analysis, segunda edição.

Anexo

	Estimativa	Erro Padrão	Estatística Z	P-valor
Intercepto	0,1356	0,1502	0,9029	0,3666
Hábito (Ocasional)	0,2421	0,1886	1,2838	0,1992
Local (Piscina)	0,1534	0,1904	0,8055	0,4205
Faixa etária 20-24	-0,8041	0,2879	-2,7931	0,0052
Faixa etária 25-29	-0,7509	0,2762	-2,7185	0,0066
Hábito (Ocasional):Faixa etária 20-24	0,0776	0,2612	0,2971	0,7664
Hábito (Ocasional):Faixa etária 25-29	0,7112	0,2930	2,4272	0,0152
Hábito (Ocasional):Local (Piscina)	0,3555	0,2217	1,6038	0,1088
Local (Piscina):Faixa etária 20-24	0,5696	0,2882	1,9763	0,0481
Local (Piscina):Faixa etária 25-29	0,1346	0,2620	0,5137	0,6075

Tabela 11: Tabela resumo para o modelo de Poisson. Note que Hábito é o hábito que costuma nadar, para esta variável a referência é considerada como frequente; Local é o local onde se costuma nadar, cuja referência é considerada praia; Faixa etária, em que a referência é considerada 15-19; e as interações entre essas variáveis como Hábito e faixa etária, hábito e local, e local e faixa etária.