Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

INCIDÊNCIA DE DENGUE EM UMA CIDADE DA COSTA MEXICANA: UM ESTUDO PREDITIVO

CE225 - Modelos Lineares Generalizados

Eduardo Elias Ribeiro Junior

Curitiba, 19 de novembro de 2014

Contents

1 Introdução

Para investigar a incidência de dengue numa determinada cidade da costa mexicana foram coletadas características de 196 indivíduos, escolhidos aleatoriamente em dois setores da cidade. As características coletadas foram *idade*, idade do entrevistado, *nível*, nível sócio-econômico (nivel=1, nível alto; nivel=2, nível médio; nivel=3, nível baixo), *setor*, setor da cidade onde mora o entrevistado (setor=1, setor 1; setor=2, setor 2) e *caso*, se o entrevistado contraiu (caso=1) ou não (caso=0) a doença recentemente.

O principal objetivo do estudo é tentar prever ou explicar a probabilidade de um indivíduo contrair a doença (variável *caso*=1) dadas as variáveis explicativas *idade*, *nível* e *setor*, porém relação de influência destas variáveis explicativas na variável resposta *caso* também serão estudadas.

2 Materiais e Métodos

Relações onde uma variável, ou um conjunto de variáveis, são utilizadas para explicar outra podem ser analisados, em estatística, com o auxílio da teoria de modelos de regressão. Neste caso, dada a natureza binária da variável resposta *caso* (0 ou 1), serão utilizados os conceitos de modelos lineares generalizados, propostos em 1972 por Nelder e Wedderburn.

Em regressão para dados binários a distribuição Binomial é a principal alternativa como componente aleatório do modelo, o componente sistemário é dado pela combinação linear das variáveis explicativas e para função de ligação trabalharemos com as funções: `logit`, `probit`, `complemento log-log` e `cauchit`. A definição do modelo teórico com a características citadas é descrito abaixo:

$$
\begin{aligned}
Y_i &\sim Binomial(1, \pi_i) \\
g(\pi_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}
\end{aligned}
$$

Sendo Y a variável resposta, $x_{1i}, x_{2i}, \ldots, x_{ni}$ as i-ésimas realizações das respectivas variáveis explicativas $X_1, X_2, \ldots, X_n$ e $g(\pi_i)$ a função de ligação, que assume as expressões conforme tabela 1. Perceba que particularizamos a distribuição para o atual problema fixando o primeiro parâmetro da distribuição Binomial em 1 (resultando em uma Bernoulli). Este foi fixado em um, pois devido a variável *idade* ter sido coletada em anos não houve indivíduos com o mesmo conjunto de covariáveis.

Table 1: Funções de Ligação

Ligação	Logit	Probit	Complemento log-log	Cauchit
$\mathbf{g}(\pi)$	$ln\left(\frac{\pi}{1-\pi}\right)$	$P^{-1}(Z^{1} \leq \pi)$	$ln[-ln(1-\pi)]$	$P^{-1}(C^{2} \leq \pi)$

$^1 Z \sim N(0,1)$ $^2 C \sim Cauchy(0,1)$

Com as funções de ligação listadas acima temos opções para testar diferentes modelos e compará-los, o que será discutido posteriormente.

3 Modelagem Estatística

Nesta seção serão abordados os tópicos para modelagem de dados binários. Exploração dos dados, seleção de variáveis, especificação da função de ligação, definição do modelo e análise de diagnóstico serão temas apresentados e discutidos a seguir.

3.1 Análise Descritiva e Explorátória

Para observar o comportamento e particularidades das variáveis em estudo, nesta seção, serão discutidos alguns gráficos descritivos das variáveis coletadas no estudo. Como primeira visualização são exibidos, na figura 1, gráficos univariados de: *caso*, *idade*, *nível* e *setor*.

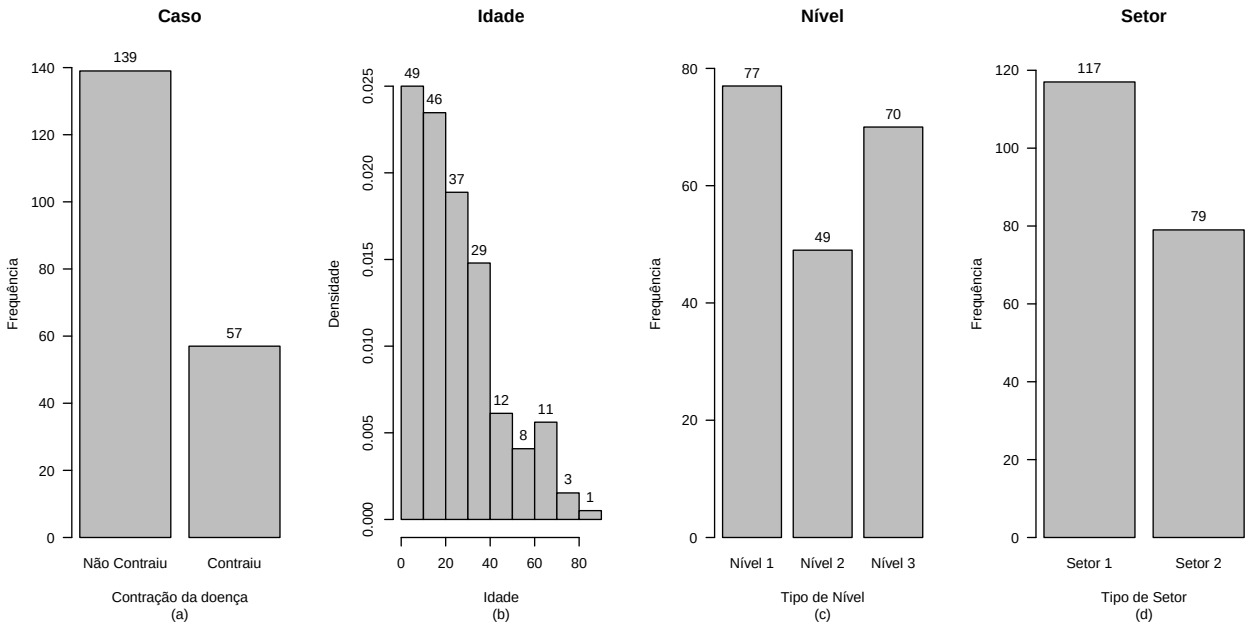

Figure 1: Variáveis Coletadas no Estudo

Na figura 1 (a) temos o gráfico de frequência para contração de dengue, note que tivemos um desbalanceamento acentuado, com a frequença dos indivíduos que não contraíram a doença quase duas vezes e meia maior do que os que contraíram. Para a figura 1 (b) temos o histograma das idades dos indivíduos estudados com forte assimetria a direita, estando aproximadamente 75% das observações abaixo de 35 anos de idade. Para os dois últimos gráficos, figura 1 (c) e figura 1 (d), temos exibidas as frequências observadas das categorias de cada variável. Em (c) as categorias, nível 1, 2 e 3, são referente ao nível sócio-econômico do indivíduo e em (d) as categorias, setor 1 e 2, referenciam o setor onde o entrevistado reside. Nesses gráficos o desbalanceamento entre as frequências nas categorias não é tão intenso quanto o observado na figura 1 (a), mas também é presente.

Retomando o objetivo inicial, onde se deseja verificar a influência das variáveis explicativas *idade*, *nível* e *setor* na contração de dengue, é apresentado na figura 2 um conjunto de gráficos uni e bivariados estratificados pela contração ou não da doença.

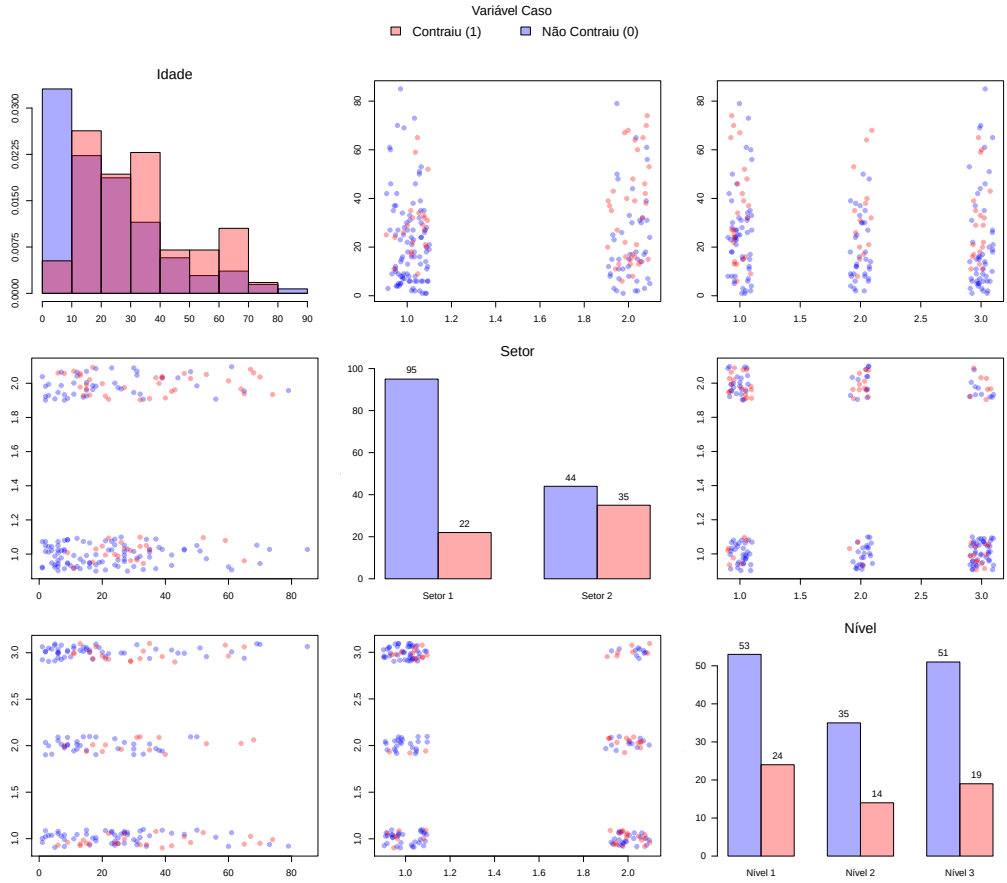

Figure 2: Covariáveis Estratificadas por Contração ou Não de Dengue

Perceba na figura 2 que os gráficos são dispostos em uma matriz de 3 linhas e 3 colunas resultando em 9 gráficos.Os gráficos abaixo da diagonal principal são gráficos bivariados idênticos aos gráficos acima dela, invertendo somente seus eixos. Primeiramente, observando os gráficos univariados da diagonal principal estratificados por *caso*, notamos que o histograma das idades dos indivíduos assume formas distintas dentre os estratos, indicando que esta variável pode ser significativa para explicar a contração de dengue nestes indivíduos. Para o gráfico de frequências do setor onde reside o indivíduo é observado que a disposição das frequências dos indivíduos que contraíram e não contraíram a doença é distinta dentro dos setores 1 e 2, também indicando que esta variável pode auxiliar na explicação da contração da doença. Já para o nível socio-econômico dos indivíduos a diposição das frequências nos estratos $caso = 1$ e $caso = 0$ não parece se distinguir tão evidentemente. Para os gráficos bivariados nenhum padrão sistemático pode ser observado claramente, ou seja, os pontos indicando indivíduos que contraíram ou não a doença parecem se dispor aleatoriamente dentre as combinações das variáveis.

3.2 Seleção de Variáveis

Após o conhecimento adquirido na seção de descrição e exploração dos dados partiremos para seleção de variáveis que será realizado pelo algoritmo *stepwise* considerando como critério de seleção o AIC (Critério de Informação de Akaike)[1].

O algoritmo *step* com critério AIC parte de um modelo especificado e realiza sucessivas atualizações na inclusão ou exclusão de variáveis pertencentes aov modelo até que se atinja o menor AIC possível. Nesse estudo executamos o algoritmo em modelos com as quatro funções de ligação descritas na tabela 1 e em suas três direções: *forward* (passo a frente, iniciando com um modelo nulo e inserindo variáveis, uma a uma, até que se encontre o menor AIC tendo como limete um modelo completo especificado), *backward* (passo a trás, retira variáveis do modelo iniciando com um modelo completo especificado até que se resulte o menor AIC) e *both* ou *stepwise* (passo a passo, iniciando com um modelo completo, retira e insere variáveis sucessivamente até resultar em um modelo com o menor AIC). Consideramos como modelo completo o modelo aditivo com todos os efeitos principais, todas as interações duplas e mais a interação tripla somando ao todo 12 parâmetros. O algoritmo

[1]Teoria disponível em http://www.yaroslavvb.com/papers/bozdogan-akaike.pdf

em suas diferentes direções e com diferentes funções de ligação resultaram no mesmo conjunto de variáveis, são elas: idade do indivíduo e o setor onde reside. O resultado do algoritmo é coerente com os gráficos apresentados na figura 2, pois as variáveis que mais se diferem dentre as categorias de contração da doença são a *idade* e *setor*. Portanto, no decorrer do estudo seguiremos nossas análise com o modelo definido abaixo.

$$Caso_i \sim Binomial(1, \widehat{\pi}_i)$$
$$g(\widehat{\pi}_i) = \widehat{\beta_0} + \widehat{\beta_1} idade_i + \widehat{\beta_2} setor_i$$

Sendo a *idade*: a idade do individuo em anos e *setor*: uma variável indicadora assumindo 1 quando o setor onde o indivíduo reside é igua a 2 e 0 caso contrário.

3.3 Especificação da Função de Ligação

Com o componente aleatório e sistemático do modelo já definidos conforme discussões anteriores, faremos a escolha da função de ligação nesta subseção. Dentre as funções de ligação definidas na tabela 1, faremos um comparativo conforme medidas descritas na tabela 2 e gráficos apresentados na figura 3.

Table 2: Comparação dos Modelos com Diferentes *Links*

Ligação	gl_s	AIC	Deviance	Area ROC	Pseudo R^2
Logit	3	217.6393	211.6393	0.7254	0.1045
Probit	3	217.3400	211.3400	0.7253	0.1057
Comp. log-log	3	218.1105	212.1105	0.7266	0.1025
Cauchit	3	218.7541	212.7541	0.7274	0.0998

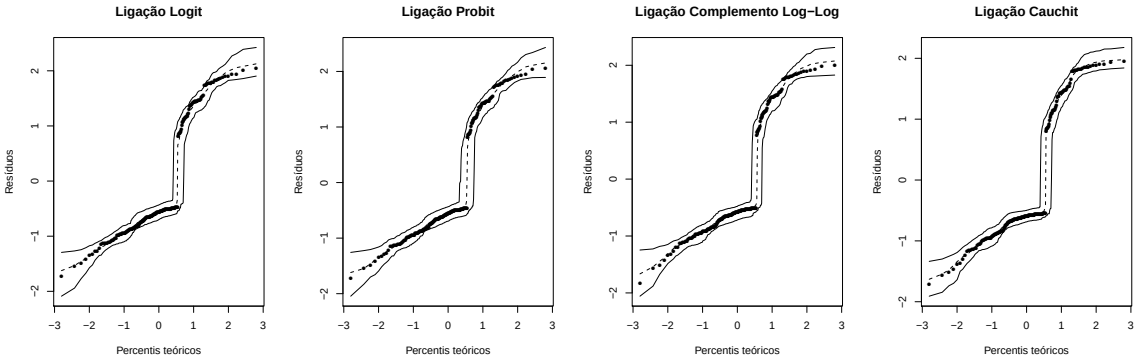

Figure 3: Qanti-Quantil com Envelope Simulado - Comparação de Links

Perceba que os modelos especificados com diferentes funções de ligação apresentaram um comportamento muito parecido. Com base nos gráficos apresentados na figura 3, não há problemas quanto a especificação do modelo nas quatro diferentes funções de ligação propostas, todos os gráficos apresentaram resíduos dentro dos intervalos simulados. Já com base nas medidas de comparação exibidas na tabela 2, nota-se um tímido melhor desempenho das ligações `logit` e `probit` em relação a `complemento log-log` e `cauchit`. A `logit` apresentou um desempenho um pouco abaixo da ligação `probit`, porém, em função da magnitude das medidas comparativas e pela vantagem interpretativa da especificação `logit`, dada em função de razão de chances, esta foi definida no modelo proposto.

3.4 Modelo Proposto

Com os elementos: distribuição Binomial, preditor linear com efeitos principais aditivos de *idade* e *setor* e função de ligação logito vamos escrever o modelo resultante das análises até aqui:

$$Caso_i \sim Binomial(1, \widehat{\pi}_i) \tag{1}$$
$$g(\widehat{\pi}_i) = ln(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}) = \widehat{\beta_0} + \widehat{\beta_1} idade_i + \widehat{\beta_2} setor_i \tag{2}$$

A tabela 3 complementa complementa e equação (2) com os valores estimados dos parâmetros e seus respectivos erros padrão:

Table 3: Resumo das Estimativas para o Modelo Ajustado

Efeito	Parâmetro	Estimativa	E. Erro Padrão	Estatística Z	Pr($>|z|$)
Constante	β_0	-2.1597	0.3439	-6.28	0.0000
Idade	β_1	0.0268	0.0086	3.10	0.0019
Setor	β_2	1.1817	0.3370	3.51	0.0005

Note nesta tabela 3 que marginalmente (considerando a distribuição Z - Normal(0,1)) todos os efeitos são bastante significativos. A interpretação das estimativas dos parâmetros será discutida posteriormente, pois esta se dará em função da razão de chances conforme já mencionado na seção 3.3.

Table 4: Análise de Diferenças de *Deviances*

Modelos	gl_s	*Deviances*	Diferença de *Deviances*	Diferença de gl_s	Valor p
Nulo	195	236.33			
Idade	194	224.32	12.0130	1	0.0005283
Setor | *Idade*	193	211.64	12.6771	1	0.0003702

Na tabela 4 é apresentada a análise de deviances sequenciais, onde são testados os efeitos das variáveis *idade* e *setor*. Primeiramente temos a hipótese de não significância do efeito da variável *idade* expressa por $H_0 : \beta_1 = 0$ onde, pelo teste de razão de verossimilhanças (TRV), obteve-se $p - valor \approx 0.00053$ evidenciando a significância estatística deste efeito. Da mesma forma, para a não significância do efeito da variável *setor* na presença da variável *idade*, $H_0 : \beta_2 = 0$, obteve-se $p - valor \approx 0.0004$. Deste modo, há também evidências de efeito significativo da variável *setor* na preseça da variável *idade*. Com isso validamos a seleção de variáveis realizada na seção 3.2.

3.5 Análise de Diagnóstico

Com o principal objetivo de subsidiar a avaliação da qualidade do modelo, a análise de diagnóstico verificará a adequação da distribuição proposta, da função de ligação, do preditor linear, enfim do modelo de regressão ajustado aos dados.

Com relação a especificação das covariáveis no modelo podemos observar na figura 4 que não há grandes evidências de má especificação das covariáveis no modelo, mesmo sendo observada a assimetria dos resíduos. Note que a interpretação dos gráficos de diagnóstico é mais flexível nestes casos, pois a limitação da variável resposta (suporte 0 ou 1) interfere na interpretação gráfica.

```
## Error in eval(expr, envir, enclos):  object 'caso' not found
```

Para a figura 5 são apresentados outros 3 gráficos que auxiliam na identificação de possíveis fuga de suposições do modelo. No caso apresentado, não temos evidências gráficas para suspeitar de nenhuma suposição não atendida. No gráfico (a) a magnitude dos resíduos não ultrapassa 2 e temos, apenas, uma leve frequência maior de resíduos abaixo de zero. No segundo gráfico (b) representando o resíduo vs. valores ajustados, temos uma disposição aparentemente centrada em zero, novamente lembramos que a natureza da variável resposta dificulta a interpretação. No terceiro e último gráfico deste figura, (c), temos o gráfico quantil-quantil com envelope simulado, onde resíduos dispostos dentro das bandas de confiança representam adequação dos dados ao modelo proposto.

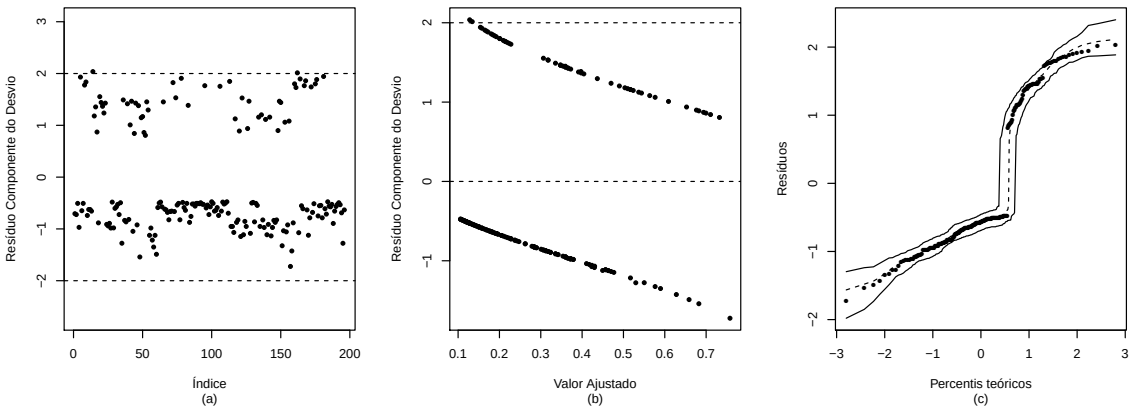

Figure 4: Resíduos vs. Covariáveis e Preditor Linear

Atendidos os pressupostos de adequação do modelo proposto com distribuição, especificação das covariáveis no preditor linear e função de ligação bem ajustadas, verificaremos possíveis obervações influentes a partit da figura 6.

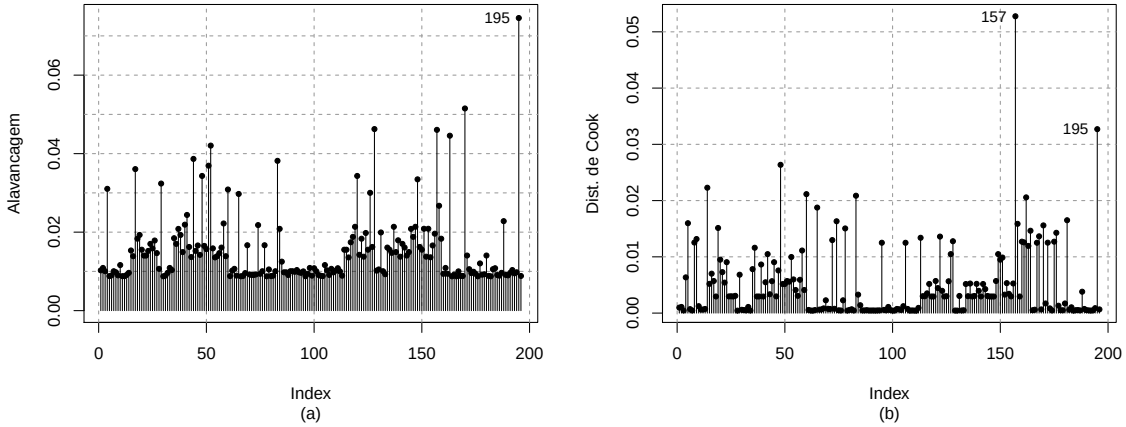

Figure 5: Medidas de Influência

Nos gráficos da figura 6 são apresentadas duas medidas de influência: valores de alavancagem h em (a) e distâncias de cook em (b). Ambos indicam que valores com grandes magnitudes, em relação aos demais, podem se apresentar como observações influentes. Nos dois gráficos destamos 2 observações (#157 e #195) que apresentaram valores muito diferentes dos demais. A observação #157 refere-se a um indivíduo com 79 anos, residente do setor 2 da cidade e que não contraiu a doença recentemente, perceba que para este perfil, segundo o modelo, teríamos uma maior probabilidade de contração da doença (estimada em aproximadamente 0.76). Da mesma forma para o perfil #195: 85 anos, residente do setor 1 da cidade e não apresentando contração da doença recentemente, o modelo estima uma probabilidade de aproximadamente 0.53 de contrair a doença. Foram ajustados modelos sem as variáveis identificadas, porém as estimativas e componentes do modelo não apresentaram diferenças significativas. As observações continuaram presentes na análise.

Com isso, pelas análises de diagnóstico realizadas anteriormente podemos utilizar o modelo proposto para inferência e interpretações.

4 Resultados

Com o modelo especificado e avaliado podemos realizar predições e interpretações. A figura 7 exibe os gráficos provenientes do modelo descrito nas expressões (1) e (2) da seção 3.4.

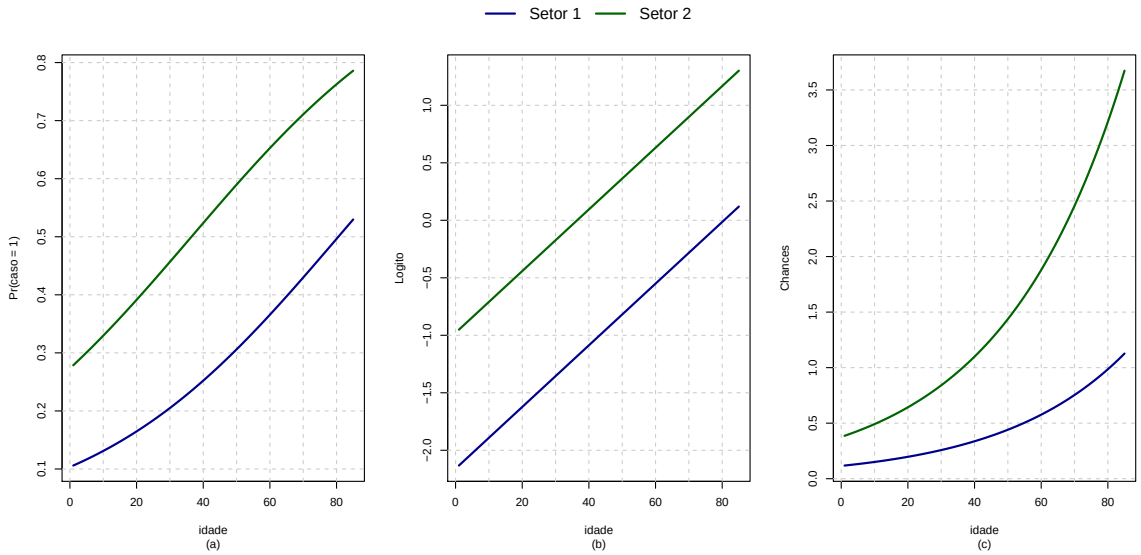

Figure 6: Predicaõ do Modelo Ajustado

Note que na figura 7 são exibidos três gráficos de predição sob o mesmo modelo. Em modelos lineares generalizados temos mais que um gráfico de predição, especialmente neste caso temos dois deles, (a) e (c), que são de extrema importância para interpretações e inferências. Apresentando as diferenças e equivalências dos gráficos temos em (a) o gráfico que representa a predição na escala do parâmetro de interesse, a probabilidade de um indíviduo contrair a doença, com características de *idade* e *setor* descritas. Já para a figura 7 (b) é considerada a escala do logito ($ln(\frac{\hat{\pi}}{1-\hat{\pi}})$) e temos retas paralelas apresentadas, concordante com a especificação do preditor linear do modelo. Finalmente para o gráfico (c) temos a escala das chances apresentada no eixo y, ou seja, exponencial do logito que também é de fundamental interesse no estudo, observe que a chance de contração da doença é diretamente proporcional a idade do indivíduo e indivíduos residentes no setor 2 da cidade tem uma maior chance estimada.

O modelo de regressão binomial com função de ligação logito tem como atrativo, a facilidade da obtençção das razões de chances devido a construção da função de ligação. Neste estudo uma estimativa para a razão de chances entre o setor 1 e setor 2, ajustada para idade do indivíduo, é dada por $exp\{1.1817\} \approx 3.26$, ou seja, estima-se que a chance de contração da doença em indivíduos residentes do setor 2 é aproximadamente 3.26 vezes a chance dos indivíduos residente do setor 1 (observe também o gráfico (c)). De modo análogo, a razão de chances entre indivíduos com 1 ano de diferença, ajustada pelo setor de residência do indivíduo, é estimada por $exp\{0.2681\} \approx 1.31$, ou seja, estima-se que a chance de contração da doença em indivíduos com $x+1$ anos de idade é 1.31 vezes a chance dos indivíduos com x anos.

5 Conclusões

Com base neste estudo foi possível constatar que o setor da cidade tem forte influência na propagação da dengue, ainda foi evidenciado que a idade do indivíduo também tem certa influência. Para o nível socio-ecônomico, variável também coletada no estudo, não foi possível verificar associação com a doença. Ainda para as variáveis significativas verificou-se que interações não foram significaticas, ou seja, o efeito de idade estratificado pelos diferentes setores não foi significativo. O setor 2 da cidade teve uma chance de contração de aproximadamente 3 vezes maior que o setor 1 da cidade. Para a idade do indivíduo também tivemos uma chance maior para indivíduos mais idosos, porém com menor intensidade.