UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Carlos Antonio Camilo

Regressão logística em dados da PeNSE 2015 para João Pessoa-PB

Campina Grande - PB

Fevereiro de 2018

Carlos Antonio Camilo

Regressão logística em dados da PeNSE 2015 para João Pessoa-PB

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Gustavo Henrique Esteves

Campina Grande - PB

Fevereiro de 2018

C183r Camilo, Carlos Antonio.
Regressão logística em dados da PeNSE 2015 para João Pessoa - PB [manuscrito] : / Carlos Antonio Camilo. - 2018.
31 p.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2018.
"Orientação : Prof. Dr. Gustavo Henrique Esteves, Departamento de Estatística - CCT."

1. Regressão logística. 2. Razões de chance. 3. PeNSE.

21. ed. CDD 519.5

Carlos Antonio Camilo

Regressão logística em dados da PeNSE 2015 para João Pessoa-PB

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 09 de fevereiro de 2018.

BANCA EXAMINADORA

Gustavo H. E.

Prof. Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Tiago Almeida de Oliveira

Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Mª das Vitórias A. Serafim

Profa. Ms. Maria das Vitórias Alexandre Serafim
Universidade Estadual da Paraíba

A Deus em primeiro lugar, aos meus pais Marineide e Manoel, meus irmãos Luciano, Luciana, Lucineide e Marivaldo, minhas sobrinhas Lauanny, Mikaele e Micheline, aos amigos e familiares dedico com todo carinho e amor.

Agradecimentos

A DEUS pela força de enfrentar os obstáculos e desafios da vida, superando tudo para chegar até o fim dessa trajetória, obrigado DEUS por tudo!

Aos meus pais Marineide e Manoel, aos meus irmãos e sobrinhas pela compreensão do dia a dia.

Aos amigos do ônibus escolar da cidade de Serra Redonda-PB.

Aos meus companheiro(a)s de trabalho da Escola Estadual Dom Adauto do município de Serra Redonda-PB; em especial a minha eterna diretora e grande amiga pessoal Cristiane, por toda compreensão.

Ao meu orientador, Professor Gustavo Esteves por seu comprometimento, empenho e dedicação em suas orientações para comigo.

Aos professores, Tiago e Vitória, por terem aceito o convite de participar da banca e trazer suas contribuições para este trabalho.

Aos professores do Departamento de Estatística da UEPB, no qual temos a honra de ter excelentes professores, fazendo com que o curso cresça a cada dia.

À professora Maria das Vitórias pela contribuição neste trabalho, com o envio de alguns materiais.

A Arlete por ter ajudado muito em algumas disciplinas, com a sua disponibilidade com materiais e tempo para explicar alguns trabalhos.

E não poderia de esquecer da melhor turma que a universidade me deu, obrigado por tudo meus amigos Ângela, Fátima, Vitória, Shirlley, Filipe e Alberto (que infelizmente não pode continuar), sem vocês não teria chegado até aqui. Obrigado pelas risadas, estresses, tardes de estudos na UEPB, enfim, tudo, pois vocês estarão sempre em meu coração.

Por fim, a todos que, de forma direta ou indireta, contribuíram em todos meus momentos nesta universidade.

"Jamais se desespere em meio as sombrias aflições de sua vida, pois das nuvens mais negras cai água límpida e fecunda."

(Provérbio chinês)

Resumo

A regressão logística é uma ferramenta estatística muito útil para uso em determinadas áreas do conhecimento, tendo semelhança com o modelo de regressão linear, se diferenciando por que a variável resposta para o modelo logístico deve ser binária. Tal modelo de regressão permite fazer a estimação da chance associada à probabilidade de ocorrência de determinado evento de interesse. Neste trabalho em particular, o objetivo principal foi verificar a influência de certos comportamentos de risco à saúde na prática de relação sexual em adolescentes escolares usando a regressão logística. Os dados são oriundos da Pesquisa Nacional de Saúde do Escolar – PeNSE – no ano de 2015 para o município de João Pessoa - PB. Os modelos foram ajustados através de métodos computacionais estatísticos utilizando o software R na versão 3.4.1, onde foi possível verificar que as variáveis independentes foram significativas ao nível de 5% de significância estatística. A razão de chances mostrou que os comportamentos de risco à saúde estão mais associados à prática de relação sexual entre os adolescentes. O gráfico de envelope simulado e a análise de diagnóstico do modelo mostraram que os dados foram bem ajustados. Foi possível concluir que na medida que um adolescente faz uso de algum tipo droga, cigarro ou bebida, aumenta sua chance de exposição a relação sexual.

Palavras-chaves: Regressão Logística. Razões de chances. PeNSE.

Abstract

Logistic regression is a very useful statistical tool for use in certain areas of knowledge. It is similar to the linear regression model, with the difference that in the logistic model the response variable must be binary. Such a regression model allows the estimation of the chance associated with the probability of occurrence of a particular event of interest. In this particular study, the main objective was to verify the influence of certain health risk behaviors on the practice of sexual intercourse in school adolescents using logistic regression. The data come from the National School Health Survey (*Pesquisa Nacional de Saúde do Escolar - PeNSE*) in the year 2015 for the João Pessoa-PB city. The models were adjusted by statistical computational methods using software R in the 3.4.1 version, where it was possible to verify that the independent variables were significant at the 5% of statistical significance level. The odds ratio showed that health risk behaviors are more associated with sexual intercourse among adolescents. The simulated envelope graph and the diagnostic analysis of the model showed that the data were well adjusted. It was possible to conclude that as an schoolar teenager makes use of some sort of drug, cigarette or drink, increased is their chance of sexual intercourse exposure.

Key-words: Logistic Regression. Odds ratio. PeNSE.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A adolescência é uma fase da infância para a vida adulta, tendo importantes transformações biológicas, cognitivas, emocionais e sociais. É uma fase da vida marcada pela crescente autonomia e independência em relação à família, bem como pela experimentação de novos comportamentos e vivências (Instituto Brasileiro de Geografia e Estatística, 2015).

A regressão logística é uma técnica estatística utilizada em diversas áreas do conhecimento, como na medicina, em dados financeiros, em dados de seguradora, na economia, dados educacionais, etc, onde a mesma é recomendada em situações em que a variável dependente é dicotômica ou binária. Já as variáveis independentes podem ser categóricas ou não. Esse tipo de regressão permite estimar a probabilidade associada à ocorrência de um determinado evento perante um conjunto de variáveis explanatórias.

Inicialmente, o objetivo principal para esta pesquisa seria fazer uma comparação entre o uso da regressão de Poisson, utilizada no artigo de Sasaki et al. (2015), com a regressão logística, pois tem-se a convicção que a modelagem utilizada no artigo parece estar equivocada, uma vez que a variável resposta utilizada não é proveniente de um processo de contagem, mas sim observações de natureza binária. Porém, por conta dos dados não serem carregados em sua totalidade, não foi possível fazer a comparação que era desejada.

O modelo de regressão de Poisson é recomendado para a análise de dados de contagem, mesmo quando o tempo de observação não é o mesmo para cada unidade amostral. E mesmo quando a resposta de interesse não é originalmente do tipo binário, alguns pesquisadores têm dicotomizado a resposta de modo que a probabilidade de sucesso possa ser ajustada através da regressão logística (PAULA, 2013).

Portanto o presente trabalho partiu para um objetivo similar ao de Sasaki et al. (2015), que foi verificar a associação de fatores de comportamentos de risco à saúde (uso de drogas, fumo e bebidas) com a prevalência na prática de relação sexual em adolescentes no município de João Pessoa-PB com o uso da regressão logística. Os dados foram da Pesquisa Nacional de Saúde do Escolar – PeNSE – do ano de 2015, feita pelo Instituto Brasileiro de Geografia e Estatística (IBGE) com a colaboração dos Ministérios da Educação e da Saúde, feita em todas as capitais brasileiras e algumas cidades do interior, para adolescentes do 9^o ano do Ensino Fundamental II.

Convém relatar que no trabalho original os autores usaram os dados da PeNSE de 2009 para o município de Goiânia-GO, tendo também avaliado a associação de outras variáveis explicativas com a prevalência de prática de relação sexual entre os adolescentes

através da regressão de Poisson, como já mencionado anteriormente.

O presente trabalho está dividido da seguinte forma. O Capítulo 1 apresenta esta introdução. O Capítulo 2 traz a fundamentação teórica, que trata do desenvolvimento do trabalho falando sobre os modelos lineares generalizados, o marco histórico da regressão logística com os principais conceitos dos modelos simples e múltiplo, da seleção do modelo, estimação dos parâmetros e das técnicas de diagnóstico. O Capítulo 3 apresenta os resultados obtidos a partir dos dados utilizados e discussões sobre a análise. Por fim o capítulo 4 trata das principais conclusões sobre o trabalho. Finalmente são apresentadas as referências bibliográficas e um apêndice contendo a rotina do *software* R utilizada na análise.

2 Fundamentação Teórica

O assunto aqui retratado nesta seção, de um modo específico e geral, traz aspectos relacionados aos modelos de regressão logística, dando uma breve revisão sobre os principais conceitos relacionados com os modelos simples e múltiplo.

2.1 Marco Histórico

A regressão logística foi descoberta em meados do século XIX, com intuito de descrever as reações químicas em cursos de autocatálise e no crescimento de populações. Segundo Paula (2013), a regressão logística teve bastante conhecimento desde os anos 50, onde foi a partir deste ano que esse tipo de regressão ficou bastante conhecida entre estatísticos e simpatizantes da área, principalmente através de Cox. A regressão logística é uma das técnicas estatísticas, no qual seu objetivo é produzir um modelo que permita a previsão de valores tomados por uma variável categórica, com frequência binária, a partir de variáveis explicativas contínuas e/ou binárias, através de um conjunto de observações.

2.2 Regressão Logística

Batista (2010) descreve que a regressão logística é uma ferramenta analítica, que se baseia em princípios da regressão múltipla, onde busca predizer a relação entre uma ou mais variáveis conhecidas buscando explicar determinada situação e sua dependência entre as variáveis. A regressão logística, ao contrário da regressão múltipla, trabalha com variáveis métricas, não métricas e categóricas, onde a variável dependente deve ser qualitativa binária, assumindo apenas um entre dois resultados possíveis.

Com um vasto uso nas ciências médicas e sociais, a regressão logística têm êxito por conta das numerosas ferramentas que permitem interpretar de modo aprofundado os resultados obtidos. Esse tipo de regressão é utilizada nas seguintes áreas, de acordo com suas respectivas características:

- Em medicina, a regressão logística permite caracterizar fatores associados de pacientes em boa situação de saúde com relação aos doentes;
- Nas finanças, esse tipo de regressão permite detectar riscos em determinado uso de crédito;
- Na parte de seguros, ela encontra uma maneira de verificar os clientes que sejam sensíveis a determinada política securitária em relação a um dado risco particular;

- Na economia, ela torna possível explicar uma variável discreta.

A variável resposta para a regressão logística deve ser classificada na distribuição Bernoulli, quando elas assumem valores do tipo sucesso ou fracasso; e com n ensaios independentes desta distribuição chega-se à distribuição Binomial. A regressão logística está baseada, de um modo geral, na transformação *logit* para proporções.

2.3 Modelos Lineares Generalizados

Os **Modelos Lineares Generalizados**, conhecidos por MLG, foram introduzidos na área de Estatística por volta do início dos anos 70, tendo como idealizadores Nelder e Wedderburn, e tiveram em sua chegada um impacto muito relevante no desenvolvimento da Estatística Aplicada. O MLG é uma generalização da **Regressão de Mínimos Quadrados** e uma extensão de **Modelos Lineares**. Os MLG's têm uma ideia básica de se ter várias opções para a distribuição da variável resposta, com isso, permite que as distribuições pertençam à família exponencial e traga consigo uma maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear η (PAULA, 2013).

O MLG tem uma ideia básica envolvida, que é expandir as opções de probabilidades dos erros associados ao modelo de regressão, e consequentemente, da variável resposta (PAULA, 2013). Segundo o autor, os precursores desse modelo estatístico propuseram um processo iterativo, com intuito de estimar parâmetros e introduzir o conceito de desvio que tem sido bastante utilizado na parte avaliativa da qualidade do ajuste dos MLG's, no desenvolvimento de resíduos e nas medidas de diagnóstico.

De acordo com Cordeiro e Demétrio (2008), quando é dada uma amostra aleatória de tamanho n, contendo observações independentes entre si, segue que um MLG pode ser resumido da seguinte forma:

i) a variável resposta tem uma distribuição pertencente à família de distribuições exponenciais;

ii) as variáveis explanatórias entram na forma de uma estrutura linear;

iii) e por fim, a função de ligação, que é uma adequação para a função entre os componentes aleatório e sistemático.

2.4 Regressão Logística Simples

Inicialmente, Paula (2013) considera o modelo logístico linear simples, onde $\pi(x)$ será a probabilidade de "sucesso" sendo x uma variável explicativa qualquer. O modelo

logístico simples é definido por:

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x, \tag{2.1}$$

em que β_0 e β_1 são parâmetros desconhecidos que se pretende estimar. Agora aplicando a função exponencial em ambos os membros da equação (2.1), tem-se o seguinte:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x},$$

dado que:

- e^{β_0} é o risco basal em escala exponencial, e;
- e^{β_1} é o tamanho do efeito de risco associado à variável x, em escala exponencial.

Tomando como exemplo Silva (2014) e Paula (2013), então tem-se uma associação entre uma determinada doença e a ocorrência de um fator particular, por exemplo, sendo amostras independentes, tendo n_1 indivíduos com $(x = 1)$ indicando presença do fator, e n_2 indivíduos com $(x = 0)$ com ausência do fator e $\pi(x)$ a probabilidade de desenvolvimento da doença depois de um certo período fixo. Portanto, a chance de desenvolvimento da doença para um certo indivíduo com a presença do fator será a seguinte:

$$\frac{\pi(1)}{1-\pi(1)} = e^{\beta_0+\beta_1},$$

e por outro lado, a chance de desenvolvimento da doença para um indivíduo com ausência do fator será:

$$\frac{\pi(0)}{1-\pi(0)} = e^{\beta_0}.$$

Portanto, a razão de chances que vai depender do parâmetro desconhecido será:

$$\psi = \frac{\pi(1)\{1-\pi(0)\}}{\pi(0)\{1-\pi(1)\}} = e^{\beta_1}.$$

Continuando o exemplo de Silva (2014) e Paula (2013), agora trabalhando com dois estratos x_1 e x_2 ($x_1 = 0$, para o estrato 1 e $x_2 = 1$, para o estrato 2) sendo n_{11} indivíduos na presença do fator e n_{21} indivíduos na ausência do fator, com esses indivíduos pertencentes a amostra do estrato 1. Para a amostra do estrato 2 teremos n_{12} indivíduos na presença do fator e n_{22} indivíduos na ausência do fator. A probabilidade será dada por $\pi(x_1,x_2)$ para o desenvolvimento da doença, tendo $x_2 = 1$ para a presença do fator e $x_2 = 0$ para a ausência. Consequentemente, terá quatro parâmetros a serem estimados: $\pi(0,0)$, $\pi(0,1)$, $\pi(1,0)$, $\pi(1,1)$. Por consequência, qualquer reparametrização deverá ter no máximo quatro parâmetros.

Considere-se a seguinte reparametrização, então terá:

$$\log\left\{\frac{\pi(x_1,x_2)}{1-\pi(x_1,x_2)}\right\} = \beta_0 + \gamma x_1 + \beta x_2 + \delta x_1 x_2$$

Portanto, γ será o efeito do estrato, β o efeito do fator e δ é a interação entre o fator e o estrato. Concomitante com essa reparametrização, as razões de chances em cada estrato deve-se calcular das seguintes formas:

$$\psi_1 = \frac{\pi(0,1)\{1-\pi(0,0)\}}{\pi(0,0)\{1-\pi(0,1)\}} = e^{\beta},$$

e

$$\psi_2 = \frac{\pi(1,1)\{1-\pi(1,0)\}}{\pi(1,0)\{1-\pi(1,1)\}} = e^{\beta+\delta}.$$

Portanto, tem $(H_0 : \psi_1 = \psi_2)$ sendo a hipótese de homogeneidade das razões de chances que é equivalente à hipótese de não interação $(H_0 : \delta = 0)$. Desta maneira, quando se tem a ausência da interação do fator com o estrato, pode-se dizer que a associação entre o fator e a doença é a mesma nos estratos. A contento, poderá haver efeito de estrato. Por ilustração, rejeita-se a hipótese $H_0 : \delta = 0$. Logo, a chance de desenvolvimento da doença será dado pelo seguinte logaritmo:

$$\log\left\{\frac{\pi(x_1,x_2)}{1-\pi(x_1,x_2)}\right\} = \beta_0 + \gamma x_1 + \beta x_2,$$

cuja interpretação pode ser feita da seguinte maneira: Mesmo não tendo interação entre os dois estratos, com razões de chances constantes, as probabilidades de desenvolvimento da doença podem estar em patamares diferentes, com as probabilidades sendo maiores de um estrato para outro (PAULA, 2013).

2.5 Regressão Logística Múltipla

De acordo com Paula (2013), Considera-se agora o modelo geral da regressão logística, tem-se a seguinte equação:

$$\log\left\{\frac{\pi(X)}{1-\pi(X)}\right\} = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q, \tag{2.2}$$

em que, $X = (1, x_1, ..., x_q)^T$ assume valores de variáveis explicativas. O processo iterativo de mínimos quadrados reponderados afim de obter $\hat{\beta}$ é dado por:

$$\beta^{(m+1)} = (\mathbf{X}^T\mathbf{V}^{(m)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{(m)}\mathbf{z}^{(m)},$$

portanto $\mathbf{V} = \text{diag}\,\{\pi_1(1-\pi_1), \cdots, \pi_n(1-\pi_n)\}$, e $z = \{z_1, ..., z_n\}^T$ é a variável dependente modificada, com

$$z_i = \eta_i + \frac{(y_i - \pi_i)}{\pi_i(1-\pi_i)},$$

além de $m = 0, 1, \cdots$ e $i = 1, ..., n$. Quando os dados são agrupados em k grupos, substitui-se n por k e redefine-se $\mathbf{V} = diag\ \{n_1\pi(1-\pi_i), \cdots, n_k\pi_k(1-\pi_k)\}$ e

$$z_i = \eta_i + \frac{(y_i - n_i\pi_i)}{\{n_i\pi_i(1-\pi_i)\}}.$$

Assintoticamente, em ambos os casos, quando $n \to \infty$ (para a primeira situação) e $\frac{n_1}{n} \to a_1 > 0$ (para a segunda situação) tem-se que

$$\hat{\beta} - \beta \sim \mathbf{N_q}\left[\mathbf{0}, (\mathbf{X^T V X})^{-1}\right].$$

Ao ter $(p-1)$, com $p \leq q$, das $(q-1)$ variáveis explicativas do tipo binário, pode-se ter interpretações importantes para as razões de chances. Ao tomar o exemplo de Paula (2013), com $p = 4$ e onde x_2 ($x_2 = 1$ com a presença do fator, $x_2 = 0$ com a ausência do fator) e x_3 ($x_3 = 1$ presença do fator, $x_3 = 0$ ausência do fator) representando dois fatores. Supondo ainda que $x_4 = x_2x_3$ irá representar a interação entre dois fatores. Neste caso, o modelo fica dado por:

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \sum_{j=5}^{q} x_j\beta_j.$$

A notação será dada da forma seguinte: ψ_{ij} é a razão de chances entre um indivíduo na condição $(x_2 = i, x_3 = j)$ em relação a um indivíduo sob a condição $(x_2 = 0, x_3 = 0)$, com $i, j = 0, 1$, de maneira que os dois indivíduos têm valores observados iguais para as demais $(p-4)$ variáveis explicativas. Nestas condições pode ser facilmente mostrado que

$$\psi_{10} = \exp(\beta_2),\ \psi_{01} = \exp(\beta_3) \text{ e } \psi_{11} = \exp(\beta_2 + \beta_3 + \beta_4).$$

Agora, será testada a hipótese $H_0 : \beta_4 = 0$ (com a ausência de interação) sendo equivalente a hipótese $H_0 : \psi_{11} = \psi_{10}\psi_{01}$ (hipótese de efeito multiplicativo) onde no caso particular, x_3 representa dois estratos ($x_3 = 0$, para o primeiro estrato; $x_3 = 1$, para o segundo estrato). Para o estrato 1 a razão de chances entre a presença e ausência do fator será $\psi_{10} = \exp(\beta_2)$, e para o estrato 2 a razão de chances valerá $\psi_{11}\psi_{01} = \exp(\beta_2 + \beta_4)$. Logo, testa-se $H_0 : \beta_4 = 0$ equivalente ao testar a hipótese de homogeneidade das razões de chances para os dois estratos (PAULA, 2013).

2.6 Seleção do modelo

Quando já se têm um conjunto de covariáveis para serem inclusas no modelo logístico, o bom é encontrar uma melhor maneira de incluir apenas covariáveis e interações com maior importância em um modelo reduzido para poder explicar melhor a probabilidade de sucesso $\pi(x)$. Os métodos que podem resolver problemas envolvendo modelos logísticos são os métodos *forward*, *backward*, *stepwise*, *Akaike*, o teste de *Wald*, o teste da razão

de verossimilhanças, sendo esse último o mais indicado nos casos que ocorrem regressão logística pelo fato de ser obtido pela diferença de duas funções desvio. Porém, é crucial a questão da interpretação dos parâmetros em um modelo logístico, pois se for selecionado de maneira mecânica, pode ser difícil sua interpretação e o modelo pode fica sem sentido. Muitas vezes, não se pode deixar as variáveis consideradas importantes de lado pela sua falta de significância estatística. Dessa forma, para uma seleção de um modelo logístico deve ocorrer por um processo conjugado de seleção estatística de modelos e bom senso (PAULA, 2013).

2.6.1 Método *forward*

De acordo com Paula (2013), esse método é uma variação do método *stepwise*. O início é dado pelo modelo simples $\mu = \beta_0$, e vai ajustando cada variável independente pelo modelo

$$\mu = \beta_0 + \beta_j x_j, \quad j = 1, \cdots, q. \tag{2.3}$$

As hipóteses a serem testadas serão $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$. Com p sendo o menor nível descritivo entre os q testes realizados. Caso $p \leq p_e$, a referida variável entra no modelo.

Supondo por exemplo que X_1 tenha sido escolhida, então ajusta-se os modelos nos seguintes passos

$$\mu = \beta_0 + \beta_1 x_1 + \beta_j x_j, (j = 2, \cdots, q)$$

As hipóteses a serem testadas serão $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$. Sendo p o menor nível descritivo entre os $(q-1)$ testes. Caso $p \leq p_e$, a referida variável entra no modelo. Repete-se esse processo até ocorrer $p > p_e$.

2.6.2 Método *backward*

Para Paula (2013), esse procedimento é iniciado a partir do modelo completo, ou seja, com a inclusão de todas as variáveis explicativas, de acordo com o modelo

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q, \tag{2.4}$$

sendo que em seguida são testadas as seguintes hipóteses $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ com $j = 0, \cdots, q$. Com p sendo o maior nível descritivo entre os q testes. Caso $p > p_s$, a referida variável sai do modelo. Supondo por exemplo que X_1 tenha saído do modelo, reajusta-se o modelo

$$\mu = \beta_0 + \beta_2 x_2 + \cdots + \beta_q x_q,$$

e as hipóteses a serem testadas serão $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, agora com $j = 2, \cdots, q$. Sendo p o maior nível descritivo entre os $(q-1)$ testes. Caso $p > p_s$, a referida variável sai do modelo. O processo é repetido até que ocorra $p \leq p_s$.

2.6.3 Método *stepwise*

Este método é a junção dos métodos *forward* e *backward* e é iniciado com o modelo simples $\mu = \beta_0$, da mesma forma que o *forward* para esse método. O processo de seleção da-se através das variáveis ocorre da seguinte maneira: Quando duas variáveis são incluídas no modelo, verifica-se se a primeira não sai do modelo, o processo é repetido até que nenhuma variável seja incluída ou excluída do modelo. Geralmente se adota $p_e = p_s = 0,20$ como critérios de entrada e saída de variáveis, respectivamente. Sendo um dos métodos que tem uma maior aplicação em regressão logística, (PAULA, 2013).

2.6.4 Teste da razão de verossimilhança

Para Teotonio (2016), a obtenção do teste da razão de verossimilhança se dá por meio de comparação de valores observados da variável resposta com os valores preditos oriundos dos modelos com a variável em questão ou sem ela. A função de verossimilhança se baseia na expressão a seguir

$$D = -2 \ln \left[\frac{L_c}{L_s} \right], \tag{2.5}$$

em que L_c é a verossimilhança do modelo logístico com a inclusão da variável de interesse e L_s é a verosimilhança do modelo logístico sem a inclusão da variável. Ou, escrevendo de outra forma

$$D = 2 \ln(L_s) - 2 \ln(L_c).$$

As hipóteses a serem testadas serão $H_0 : \beta_i = 0$ vs $H_0 : \beta_i \neq 0$. É possível se mostrar que a estatística acima segue distribuição qui-quadrado com 1 grau de liberdade.

Na regressão múltipla, é interessante saber se pelo menos uma variável é significativa para o modelo. Sob hipótese nula, teremos p coeficientes iguais a zero, portanto, a estatística D terá distribuição qui-quadrado com p graus de liberdade. Onde o L_c e L_s serão a verossimilhança do modelo com as p variáveis e a verossimilhança do modelo com o intercepto, respectivamente.

2.6.5 Teste de Wald

Esse teste verifica se o parâmetro é estatisticamente significativo para o modelo em estudo. Esse tipo de teste faz testar os parâmetros, colaborando com um grupo de variáveis independentes. A obtenção desse teste dá-se pela comparação entre a estimativa de máxima verossimilhança do parâmetro, $\hat{\beta}_i$, e a estimativa do erro padrão, com a hipótese $H_0 : \beta_i = 0$, (AGRESTI, 2002). A estatística do teste de Wald é dada por

$$W_i = \frac{\hat{\beta}_i}{DP(\hat{\beta}_i)}. \tag{2.6}$$

Essa estatística dada acima segue distribuição expressa da distribuição normal. A interpretação para o teste se dá quando o nível de significância, ou o p valor é menor que o ponto crítico pré definido, e conclui-se que o parâmetro β_i é significativo dentro do modelo.

2.6.6 Procedimento de *Akaike*

Este método é um dos procedimentos mais simples para a seleção de variáveis explicativas. Ele se diferencia dos demais métodos já vistos anteriormente por conta de ser um processo de minimização que não envolve testes estatísticos. Segundo Paula (2013) a ideia básica é selecionar um modelo bem ajustado e que tenha um número reduzido de parâmetros. O logaritmo da função de verossimilhança, denotado por $L(\beta)$, tem seu crescimento com o aumento do número de parâmetros do modelo, uma proposta razoável seria encontrar o modelo com menor valor para a seguinte função

$$AIC = -L(\hat{\beta}) + q, \tag{2.7}$$

onde q denota o número de parâmetros. No caso especial do modelo normal linear pode ser mostrado que o AIC fica expresso na forma

$$AIC = n \log\left(D(y; \hat{\mu}/n)\right) + 2q,$$

quando σ^2 é desconhecido, $D(y; \hat{\mu}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$.

2.7 Estimação dos parâmetros

2.7.1 Estimação dos parâmetros no modelo de regressão logística simples

Suponha que x_i, y_i seja uma amostra independente tendo k pares de observações, onde x_i seja o valor da variável independente da i-ésima observação e y_i seja o valor da variável resposta dicotômica onde $i = 1, 2, 3, \cdots, k$. Deve-se estimar os parâmetros desconhecidos β_0 e β_1 para ajustar o modelo de regressão logística simples, usando a expressão de probabilidade de sucesso do modelo que é dado por

$$\pi_i = \pi(x_i) = P(Y = 1|X = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

Um método a ser utilizado para estimar esses parâmetros é o da máxima verossimilhança, em que a distribuição de probabilidades da variável resposta será dada por $Y_i \sim Ber(\pi_i)$ para o modelo de regressão logística simples. Assim, a função de distribuição de probabilidade conjunta para $y_1, y_2, \cdots, y_k$, tendo que as observações são independentes, será dada por:

$$\prod_{i=1}^{k} f(y_i, \pi_i) = \prod_{i=1}^{k} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \quad y_i \in \{0, 1\},$$

logo a função de verossimilhança será dada por

$$L(\beta) = \prod_{i=1}^{k} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \quad \beta \in \mathbb{R}.$$

Tem-se que o princípio da máxima verossimilhança é estimar o valor de β que maximiza a função $L(\beta)$. Para simplificar um pouco, aplica-se o logaritmo, onde a expressão é definida como

$$\begin{aligned} l(\beta) = \ln\left[L(\beta)\right] &= \ln\left[\prod_{i=1}^{k} \pi_i^{y_i}(1-\pi_1)^{1-y_i}\right] \\ &= \sum_{i=1}^{k}\left[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right] \\ &= \sum_{i=1}^{k}\left[y_i \ln(\pi_i) + \ln(1-\pi_i) - y_i \ln(1-\pi_i)\right] \\ &= \sum_{i=1}^{k}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right]. \end{aligned}$$

Substitui-se as equações, na equação da probabilidade de fracasso dada por:

$$1-\pi_i = 1-\pi(x_i) = P(Y=0|X=x_i) = \frac{1}{1+\exp(\beta_0+\beta_1 x_i)},$$

e a equação da transformação logito da razão de chances dada por:

$$g(x_i) = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_i,$$

tem-se que

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{k}\left[y_i(\beta_0+\beta_1 x_i) + \ln\left(\frac{1}{1+\exp(\beta_0+\beta_1 x_i)}\right)\right] \\ &= \sum_{i=1}^{k}\left[y_i(\beta_0+\beta_1 x_i) - \ln(1+\exp(\beta_0+\beta_1 x_i))\right]. \end{aligned}$$

Agora deve-se encontrar o valor de β que faz a maximização de $l(\beta)$, faz esse processo derivando $l(\beta)$ em relação a cada parâmetro (β_0, β_1), resultando em duas equações

$$\frac{\partial}{\partial\beta_0} l\left(\beta\right) = \sum_{i=1}^{k}\left[y_i - \frac{1}{1+exp(\beta_0+\beta_1 x_i)} exp(\beta_0+\beta_1 x_i)\right]$$

e

$$\frac{\partial}{\partial\beta_1} l\left(\beta\right) = \sum_{i=1}^{k}\left[y_i x_i - \frac{1}{1+exp(\beta_0+\beta_1 x_i)} exp(\beta_0+\beta_1 x_i) x_i\right].$$

Enfim, igualando essas equações a zero, vai gerar os seguintes sistemas

$$\sum_{i=1}^{k}(y_i - \pi_i) = 0 \tag{2.8}$$

$$\sum_{i=1}^{k} x_i(y_i - \pi_i) = 0, \tag{2.9}$$

em que,

$$\pi_i = \frac{exp\,(\beta_0 + \beta_1 x_i)}{1 + exp\,(\beta_0 + \beta_1 x_i)}.$$

Tendo que as equações (2.8) e (2.9) não são lineares em β_1 e β_0, é necessário procurar métodos iterativos para a resolução do problema, esses métodos são disponíveis em vários programas computacionais.

2.7.2 Estimação dos parâmetros no modelo de regressão logística múltipla

A obtenção para as estimativas do vetor $\beta = (\beta_0, \beta_1, ..., \beta_q)$, será feito utilizando o método da máxima verossimilhança dos parâmetros encontrados no modelo e a matriz de covariância, dada por

$$L[\beta_0, \beta_1, ..., \beta_q | (x_i; \mu_i; y_i)] = \sum_{i=1}^{n} y_i g(X) - \mu_i \ln\left(1 + e^{g(X)}\right), \tag{2.10}$$

em que

$$g(X) = \ln\left(\frac{\pi_(x)}{1 - \pi_(x)}\right).$$

Agora fazendo a derivação de (2.10) com relação ao parâmetro β_0, tem-se

$$\frac{\partial}{\partial \beta_0} L(\beta_0, \beta_1, ..., \beta_q) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \mu_i \frac{e^{g(X)}}{1 + e^{g(X)}}$$

e, de modo geral, para um β_j qualquer

$$\frac{\partial}{\partial \beta_j} L(\beta_0, \beta_1, ..., \beta_q) = \sum_{i=1}^{n} y_i x_j - \sum_{i=1}^{n} \mu_i x_j \frac{e^{g(X)}}{1 + e^{g(X)}}.$$

Consequentemente, deve-se igualar a zero e fazer a substituição dos parâmetros pelos estimadores, daí obtêm-se as seguintes equações

$$\sum_{i=1}^{n} y_i(1 + e^{g(X)}) - \sum_{i=1}^{n} \mu_i e^{g(X)} = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} y_i x_i(1 + e^{g(X)}) - \sum_{i=1}^{n} \mu_i x_i e^{g(X)} = 0,$$

cuja solução será dada de maneira análoga à estimação dos parâmetros do modelo de regressão logística simples, que é utilizando processos iterativos.

Para calcular as probabilidades ajustadas, logo após obter as estimativas do modelo, basta calcular

$$\hat{\pi}_i = \frac{e^{\hat{g}(X_i)}}{1 + e^{\hat{g}(X_i)}},$$

em que

$$g(X_i) = \hat{\beta}_0 + \hat{\beta_{1i}} + ... + \hat{\beta_{qi}}.$$

2.7.3 Interpretação dos parâmetros

Para Agresti (2002) na regressão logística a interpretação dos parâmetros se dá quando uma variável independente é dicotômica, ou seja, fazendo comparação da probabilidade do evento não ocorrer com a probabilidade do evento ocorrer. Faz-se necessário a definição da razão de chances (*Odds Ratio*), que é dada pelo quociente da probabilidade do evento ocorrer pela probabilidade do evento não ocorrer.

O logaritmo da razão de chances será dada pela seguinte equação

$$g(x_i) = \ln\left[\frac{\pi_i}{1-\pi_i}\right],$$

caso haja a codificação da variável X como 0 ou 1, a chance da resposta quando tiver $x = 1$ será definida por $\pi(1)/\left[1-\pi(1)\right]$ e quando o $x = 0$, será definida por $\pi(0)/\left[1-\pi(0)\right]$.

A razão de chances pode ser denotada por OR, então essa razão poderá ser definida por

$$\begin{aligned} OR &= \frac{\pi(1)/\left[1-\pi(1)\right]}{\pi(0)/\left[1-\pi(0)\right]} \\ &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{(\beta_0+\beta_1)}}\right)/\left(\frac{1}{1+e^{(\beta_0+\beta_1)}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{(\beta_0)}}\right)/\left(\frac{1}{1+e^{(\beta_0)}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \end{aligned}$$

Portanto, o logaritmo da razão de chances será dado por

$$\ln(OR) = \ln\left[e^{\beta_1}\right] = \beta_1.$$

Por causa de sua fácil interpretação no modelo de regressão logística, a razão de chances é uma medida de grande interesse. Para Paula (2013) a distribuição assimétrica de β_1 é devido ao seu limite ter tendência a zero. Onde as inferências são baseadas na distribuição do logaritmo da razão de chances, $\ln(OR) = \beta_1$, no qual têm tendência para seguir distribuição normal, mesmo sendo para amostras pequenas.

2.8 Técnicas de diagnóstico

Na análise de um ajuste de regressão tem uma etapa bastante importante que é a verificação de supostos afastamentos das suposições feitas para um determinado modelo, de modo especial, para o componente aleatório e para a parte sistemática do modelo, além da existência de observações discordantes com alguma interferência de modo desproporcional ou inferencial nos resultados obtidos pelo ajuste. Etapa esta, conhecida como análise de diagnóstico, que teve seu início com a análise de resíduos para poder detectar a presença

de pontos anormais e fazer uma avaliação da distribuição proposta para a variável resposta, (PAULA, 2013).

A análise de resíduos e diagnóstico detecta problemas como:

- presença de observações com pontos aberrantes (pontos discrepantes);
- inadequação das pressuposições para as médias ou erros aleatórios;
- colinearidade entre as colunas da matriz do referido modelo;
- forma funcional do modelo inadequado;
- presença de observações influentes (observações dominantes).

2.8.1 Gráfico meio-normais com envelope simulado

Segundo Moral (2013) o gráfico meio-normal com envelope simulado é bastante útil para a verificação do ajuste do modelo, esse tipo de técnica envolve os modelos lineares generalizados. Esse tipo de gráfico permite verificar a adequação do modelo, mesmo que os resíduos não tenham uma aproximação da distribuição normal.

Quando o modelo está bem ajustado, os resíduos estarão dispersos aleatoriamente entre os limites do envelope, caso tenha alguns pontos fora do envelope, a presença desses pontos podem indicar problemas de ajuste.

3 Aplicação

Os dados são oriundos da Pesquisa Nacional de Saúde do Escolar - PeNSE, realizada no ano de 2015, sob coordenação dos Ministérios da Saúde e Educação, sendo sua execução feita pelo Instituto Brasileiro de Geografia e Estatística - IBGE. Os dados analisados neste trabalho são referentes aos adolescentes do município de João Pessoa, capital do estado da Paraíba.

A pesquisa PeNSE teve sua primeira realização no ano de 2009, a segunda edição saiu em 2012 e a terceira em 2015, sendo essa última a referência para este referido trabalho. A referida pesquisa foi realizada em todo o Brasil, com 102.301 alunos do 9^o ano do Ensino Fundamental II, onde os dados foram coletados entre os meses de abril e setembro de 2015. Para o município de João Pessoa, referência deste trabalho, participaram da PeNSE 68 escolas e 89 turmas de estudantes, dos quais tinham 3.036 alunos matriculados na referida série. Foi trabalhado com 2.417 observações, onde os dados para a análise foram ajustados através do programa estatístico R (R Core Team, 2017).

A variável dependente para o estudo em questão foi a mesma utilizada por Sasaki et al. (2015), que foi atividade sexual, só que neste trabalho foram usados os dados do município de João Pessoa-PB. As variáveis independentes foram os dados de característica para os comportamentos de risco à saúde (consumo de cigarro, álcool e drogas).

Para maiores informações acerca dos dados utilizados, é possível consultar o site do IBGE [1].

3.1 Resultados e discussões

Para chegar aos resultados da pesquisa, foi utilizado o *software* estatístico (R Core Team, 2017) na versão 3.4.1, em que foi ajustado os dados pelo referido programa. Na Tabela 1, tem-se a descrição das características para as variáveis demográficas e socioeconômicas dos participantes da pesquisa PeNSE 2015. Sendo a maioria dos alunos do sexo feminino (52,67%) com faixa etária de 14 anos (48,53%) e da cor parda (43,79%), sendo que a maioria dos adolescentes moravam com a mãe (90,72%) e (62,01%) com o pai.

O ajuste para o modelo logístico encontra-se na Tabela 2, em que pode-se verificar que há evidências do efeito para todas as variáveis independentes em estudo, inclusive o intercepto, ao nível de 5% de significância, ou seja, todas elas são significativas para o modelo logístico estudado. Isso mostra que o estudante ao submeter a quaisquer usos

[1] <https://ww2.ibge.gov.br/home/estatistica/populacao/pense/default.shtm>

Tabela 1 – Características demográficas e socioeconômicas dos adolescentes do município de João Pessoa, PeNSE 2015.

Variáveis	Tamanho das observações (% válidos)
Demográfica	
Sexo (n=2417)	
Feminino	1273(52,67%)
Masculino	1144(47,33%)
Idade (n=2388)	
$\leq$ 13 anos	569(23,83%)
14 anos	1159(48,53%)
$\geq$ 15 anos	660(27,64%)
Raça/cor (n=2416)	
Branca	782(32,37%)
Preta	270(11,18%)
Amarela	190(7,86%)
Parda	1058(43,70%)
Indígena	116(4,80%)
Socioeconômicas	
Mora com a mãe (n=2414)	
Sim	2190 (90,72%)
Não	224 (9,28%)
Mora com o pai (n=2414)	
Sim	1497(62,01%)
Não	917(37,99%)

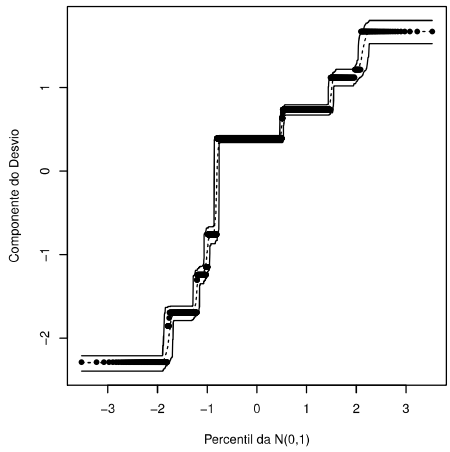
de bebida alcoólica, do fumo ou de drogas ilícitas, ele tem um aumento significativo na chance da prática de relação sexual na adolescência.

Na Figura 1, tem-se o gráfico de envelope simulado, que de acordo com o mesmo, não observa-se nenhum indícios de que a distribuição utilizada para este modelo pareça inadequada. Como o envelope é bastante útil para verificar a qualidade do ajuste, logo pode-se observar que o referido modelo é adequado.

Tabela 2 – Análise de variância para o resultado do uso de algum comportamento de risco à saúde em relação a atividade sexual.

	Estimativas	Erro padrão	Valor z	Pr (>|z|)
Intercepto	-1,101	0,162	-6,799	<0,001
Álcool	1,371	0,133	10,296	<0,002
Cigarro	1,022	0,148	6,921	<0,004
Droga	1,243	0,190	6,552	<0,005

Figura 1 – Envelope simulado para o modelo de regressão logística ajustado.



Na Tabela 3, tem as razões de chances oriundas da análise de regressão logística relacionada da associação da variável dependente (atividade sexual) e das variáveis independentes, em adolescentes do município de João Pessoa-PB, da PeNSE de 2015. Como pode-se verificar na referida tabela, a medida que um adolescente bebe, ele tem $3,94$ vezes mais chances de ter tido relação sexual comparado com o adolescente que não fez o uso de algum tipo de bebida alcoólica. Para quem fuma, tem $2,78$ vezes mais chances de ter tido relação sexual comparado com quem não fuma e por fim, a chance para aquele adolescente que fez o uso de algum tipo de droga, é $3,46$ mais de ter tido relação sexual comparado com quem não usou algum tipo de drogas.

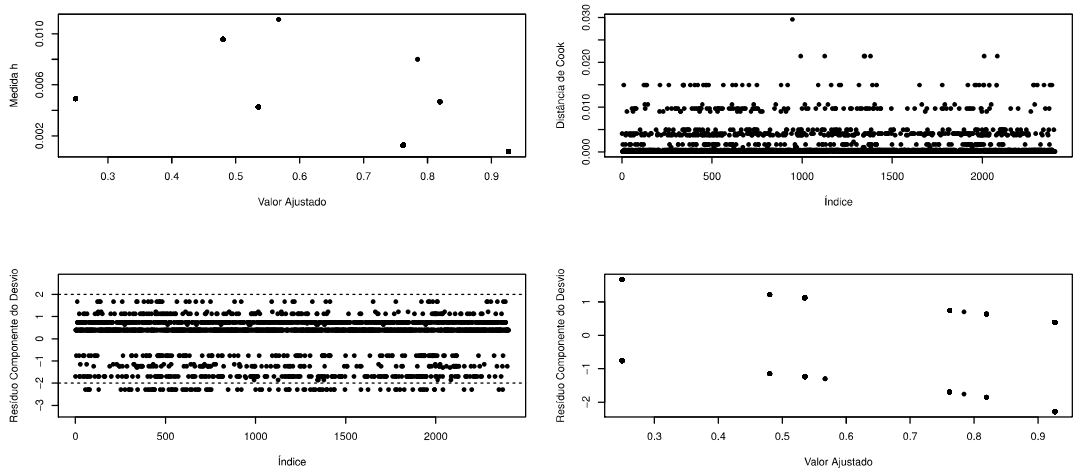
Tabela 3 – Resultado da análise de regressão logística entre a variável dependente (Atividade Sexual) e variáveis independentes, para os adolescentes do município de João Pessoa-PB, PeNSE 2015.

Variáveis	Razão de Chances (I.C. 95%)	Teste de Wald
Consumo de alcóol		
Não	1	<0,001
Sim	3,94 (3,03; 5,11)	
Uso de cigarro		
Não	1	<0,001
Sim	2,78 (2,08; 3,71)	
Consumo de drogas		
Não	1	<0,001
Sim	3,46 (2,39; 5,02)	

Na Figura 2, tem-se os gráficos da análise de resíduos, pode-se observar que no primeiro gráfico do lado esquerdo tem o gráfico de $\hat{h}_{ii}$ contra valores ajustados, notando que os mesmos não ultrapassam o limite, sendo assim, portanto nenhum ponto para esse gráfico será considerado ponto de alavanca. No primeiro gráfico do lado esquerdo abaixo,

têm-se o gráfico de resíduos, em que mostra que a maioria dos pontos estão no intervalo [2,-2], pouquíssimos estão no intervalo [2,-3]. No segundo gráfico do lado direito abaixo têm os valores ajustados, ele corrobora com o segundo gráfico da Figura do lado esquerdo. O primeiro gráfico do lado esquerdo acima, tem-se o gráfico de influência L_{Di}, em que pode-se observar que todos os pontos estão próximos uns dos outros, indicando com isso que não tem pontos aberrantes. Isto leva a crer que o modelo foi bem ajustado.

Figura 2 – Gráficos de análise de resíduo.



4 Conclusão

O intuito inicial, infelizmente, para esta pesquisa não foi alcançado, para que fosse feita uma comparação do uso indevido da análise de regressão de Poisson no artigo de Sasaki et al. (2015) como o uso que deveria ser o correto, no qual seria usar a análise de regressão logística. Mas como não foi possível carregar os dados da PeNSE do ano de 2009 na sua totalidade, não foi possível comparar os resultados das duas análises.

Com o uso do modelo de regressão logístico, para dados da PeNSE 2015 no município de João Pessoa-PB, foi possível verificar como o uso de algum tipo de comportamento de risco à saúde (consumo de drogas, álcool e do cigarro) faz com que um adolescente tenha um risco aumentado da prática de relação sexual. O modelo foi ajustado usando técnicas estatísticas computacionais no (R Core Team, 2017), onde verificou-se que as variáveis independentes foram significativas ao nível de 5% de significância. Pela parte gráfica aplicada ao modelo, através da análise dos gráficos de resíduos e do gráfico de envelope, pode-se ver que o modelo ajustado foi adequado. Logo após todo estudo feito a partir das análises e do modelo ajustado, pode-se afirmar que realmente houve efeito do uso de drogas, uso de bebidas e do uso de fumo em alguma prática de relação sexual feita por algum adolescente do 9^0 ano do Ensino Fundamental II para o município de João Pessoa-PB.

Referências

AGRESTI, A. *Categorical data analysis*. 2ª edição. ed. [S.l.]: Hoboken: John Wiley & Sons, 2002. v. 710 p. ISBN 0-471-36093-7. Citado 2 vezes nas páginas 19 e 23.

BATISTA, A. A. S. *Análise da qualidade de vida no trabalho utilizando um modelo de regressão logística*. Dissertação (Mestrado) — Universidade Tecnológica do Paraná, 2010. Citado na página 13.

CORDEIRO, G. M.; DEMéTRIO, C. *Modelos Lineares Generalizados e Extensões*. [S.l.], 2008. Citado na página 14.

Instituto Brasileiro de Geografia e Estatística. Pesquisa nacional de saúde do escolar. *IBGE*, v. 132, 2015. Citado na página 11.

MORAL, R. A. *Modelagem estatística e ecológica de relações tróficas em pragas e inimigos e naturais*. Dissertação (Mestrado) — Escola superior de Agricultura "Luiz de Queiroz- USP, 2013. Citado na página 24.

PAULA, G. de. *Modelos de Regressão com Apoio Computacional*. São Paulo, 2013. Citado 11 vezes nas páginas 11, 13, 14, 15, 16, 17, 18, 19, 20, 23 e 24.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>. Citado 2 vezes nas páginas 25 e 29.

SASAKI, R. S. A. et al. Prevalência de relação sexual e fatores associados em adolescentes escolares de goiânia, brasil. *Ciência & Saúde Coletiva*, 2015. Citado 3 vezes nas páginas 11, 25 e 29.

SILVA, A. J. *Estudo teórico sobre modelos lineares generalizados com aplicação a dados genéticos*. 2014. Citado na página 15.

TEOTONIO, M. de L. A. *Uso da regressão logística para avaliar o efeito de antissépticos em cirurgia hospitalar*. 2016. Citado na página 19.

APÊNDICE A – Rotina utilizada na análise

Este apêndice apresenta o *script* do programa R utilizado nas análises apresentadas neste trabalho.

```
## Carregando os dados já salvos
load("dados_Carlos.RData")

## Ajustando o modelo de regressão logística para sex ~ beb + fum + dro
fit.model <- glm(sex ~ beb + fum + dro, data=dados.Carlos,
family=binomial(link="logit"))
fit.model

## Verificando o ajuste do modelo
summary(fit.model)

## Tabela de ANOVA
anova(fit.model)

## Calculando razões de chance
library(epiDisplay)
logistic.display(fit.model)

## Diagnóstico do modelo
attach(dados.Carlos)
source("funcoes_gilberto/diag_bino.R", encoding="latin1")
dados.Carlos
summary(dados.Carlos)

## Envelope simulado
source("funcoes_gilberto/envel_bino.R")
```