Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Comparação de métodos de estimação do risco relativo em estudos epidemiológicos

Agda Jéssica de Freitas Galletti

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília, como parte dos requisitos para a obtenção do título de Bacharel em Estatística.

Brasília
2015

Agda Jéssica de Freitas Galletti

Comparação de métodos de estimação do risco relativo em estudos epidemiológicos

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília, como parte dos requisitos para a obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dra.
Joanlise Marco de Leon Andrade

Brasília
2015

À minha mãe,
mulher guerreira, que me ama e me apoia
incondicionalmente todos os dias.

Agradecimentos

Os agradecimentos aqui manifestados certamente não atingirão a todos que de certa forma estiveram ou estão presentes na minha vida, mas gostaria de expressar meus sinceros sentimentos de gratidão a cada pessoa que retribuiu minha amizade com amor e carinho.

A vida acadêmica se apresentou cheia de descobertas, realizações, amadurecimento e conquista de novas amizades, o que foi possível com meu ingresso na Universidade de Brasília. Por isso, além de agradecê-la, devo também essa oportunidade à Deus que iluminou-me durante essa caminhada, aos departamentos de Matemática e Estatística, pela preocupação com minha formação, ao meu professor Berçott e aos meus padrinhos Saulo e Mariana que me apoiaram e acreditaram em meu potencial.

Agradeço à professora Ma. Ana Maria Redolfi Gandulfo por apresentar-me a um mundo acadêmico repleto de oportunidades, possibilitar a troca de experiências com diferentes pesquisadores e ajudar-me a evoluir como pessoa e profissional.

Agradeço a todos os professores do curso do departamento de estatística, que foram tão importantes na minha vida acadêmica, em especial aos professores, Dra. Joanlise Marco de Leon Andrade, a qual admiro e respeito, agradeço por sua orientação, disponibilidade, ajuda, compreensão, paciência, amizade, pelo incentivo e principalmente por acreditar em minha capacidade. Dra. Ana Maria Nogales, agradeço pelo apoio, carinho, amizade, incentivos e inúmeras oportunidade de conhecimento. Dr. Eduardo Yoshio Nakano, agradeço por sempre estar disponível, pelo apoio e empolgação transmitida a mim na realização das atividades do PIBIC e da graduação. Dra. Juliana Betini Fachini, Ma. Maria Teresa Leão e Dra. Maria Amélia Biagio, agradeço pelo carinho, atenção, sorriso e abraços calorosos. Dr. George Freitas von Borries e Dr. Bernardo Borba Andrade agradeço por aceitarem estar na minha banca, pela paciência, por serem tão amáveis, pelo

bom humor, pelos sorrisos e gargalhadas, mesmo quando se fazem de durões. Dr. Donald Matthew Pianto, Dr. Lucio José Vivaldi e Dr. Jhames Matos Sampaio, agradeço por me cativarem com imensa simpatia e pela cordialidade.

Agradeço aos funcionários do departamento de estatística, pela cordialidade, profissionalismo, eficiência e alegria. Gostaria de destacar a Tathyanna e a Verinha, que se mostraram grandes amigas.

Agradeço aos amigos e grupos de amigos que fiz pelo caminho, destacando LabMat, em especial agradeço à Alexia, à Andressa, ao Caio, ao Elisson, à Jéssica, à Maísa, ao Celso, à Gabriela, à Priscila, à Raíssa e ao Weverton pelos bons momentos, pelas viagens, pelos estudos, por me suportarem até quando não dava mais, por sempre me perdoarem, pela imensa alegria, pelos eventos de alto nível, pelos momentos de gordice, pelo apoio e imensa amizade. Amostragem, agradeço especialmente ao Alex, ao Quintino e ao Guilherme por me acolherem, compartilharem tanto conhecimento, pela disponibilidade, pelo respeito, por sempre estarem por perto, pela consideração e pela amizade. Bonde da estatística, agradeço em primeiro lugar a Geiziane, que diariamente motiva-nos com seu esforço e garra, por nos unir em estudos dominicais no velho espaço Chiarini. Agradeço também, à Mariana, ao Mateus, ao Rodrigo, à Márcia, à Andressa e ao Gui, membros fundadores, pelos momentos de estudo, pelas madrugadas mal dormidas, pelo apoio, pela consideração e pelas gordices. Agradeço também aos nossos substitutos, Adolfo, Pablo, Pedro Rangel, Patrícia, Japa (Lucas) Cadú e Isa, por fazerem meu último semestre o mais especial de todos.

Andressa, agradeço por apresentar a estatística, por me acompanhar nas aventuras de um curso novo, de um estágio massa e de entrevistas bem e malsucedidas, por fazer parte de quase todos meus grupos de amizade e por continuar aqui depois de tantos abandonos. Érica e Rhayssa, a velocidade com que nos identificamos foi quase a mesma que nos vimos distantes só nos restando saudades do dia a dia de convivência, obrigada por cada momento, pela parceria, pela linda e sincera amizade que guardarei para sempre. Bianca, Davi e Maria Gabriela, agradeço por cada gesto de carinho, amizade e companheirismo que me transmitiram.

Agradeço à Edilene, Camila e Marilene, por serem amigas/irmãs, por me amarem sem importar a distância e quantidade de tempo sem nos encontrarmos, pelos bons momentos vividos, pelas experiências conquistadas, pelos dramas e loucuras adolescentes, pela

companhia inseparável e pela cumplicidade.

Por fim, agradeço à minha família por todo amor compartilhado e à minha maravilhosa mãe, que está sempre presente, com tanto afeto e incentivos.

Resumo

O presente trabalho tem como objetivo a comparação de um método direto de estimação do risco relativo em estudos transversais via regressão log binomial com um método indireto via regressão logística, propostos por Andrade e Carabin (2011) e Santos et. al. (2008), respectivamente. As medidas de associação mais utilizadas em pesquisas epidemiológicas são o risco relativo (RR) e a razão de chances (RC). Essas medidas estimam a magnitude da associação entre um fator de risco (proteção) e um desfecho de interesse. A regressão logística é um dos modelos mais utilizados para se estimar a RC, no entanto, o RR é uma medida de mais fácil interpretação. Uma das dificuldades encontradas por pesquisadores na estimação do RR via regressão log binomial, é que nem sempre há convergência das estimativas. Santos et. al. (2008) apresentaram um método que estima por aproximação o RR e Andrade e Carabin (2011) apresentam um método que minimiza consideravelmente os problemas de convergência do modelo log binomial, que estima o RR de forma direta, por meio do algoritmo adaptativo de barreiras. A comparação dos métodos baseou-se na geração de 1000 amostras de tamanho igual a 500 a fim de se avaliar os métodos para diferentes tamanhos de amostras. Assim, observou-se nas simulações do tipo I que o método indireto subestima o RR à medida que os valores dos parâmetros e o intervalo da variável contínua decrescem, chegando a 0% de cobertura, apesar da alta precisão. Nas simulações do tipo II ambos os modelos forneceram resultados próximos quanto ao viés e à variabilidade das estimativas do RR. Entretanto, como um todo, o método direto apresentou menor VRPM e melhor cobertura, enquanto que, o método indireto se mostrou ligeiramente melhor apenas na estimação do RR da variável dicotômica. Portanto, a estimação do RR via método direto, utilizando-se o método implementado por Andrade e Carabin (2011), se mostrou mais eficiente, quando os dados são gerados de uma log binomial. No entanto, mais estudos são necessários.

Palavras-chave: Risco relativo, Modelo Log Binomial, Modelo Logístico, Epidemiologia.

Sumário

Introdução

A epidemiologia pode ser definida como o estudo da distribuição e determinantes da frequência de uma doença (Rothman, 2012).

O evento de interesse em uma pesquisa epidemiológica é denominado *desfecho de interesse*, que pode ser o surgimento de uma doença ou sintoma, óbito ou outra condição relacionado à saúde. O *fator de risco* é uma variável que desejamos avaliar se tem alguma associação com o desfecho. Logo, os indivíduos que apresentam ou vivenciam o potencial fator de risco são denominados *expostos*.

A *incidência* de algum desfecho de interesse refere-se ao número de novos eventos ou casos que ocorrem em uma população de indivíduos em risco durante um determinado período de tempo. A *prevalência* representa a proporção de indivíduos em uma população que apresenta o desfecho de interesse, incluindo casos novos e casos preexistentes, em um determinado período de tempo.

Pesquisas podem ser observacionais, quando o grupo de interesse é somente observado, ou experimentais, quando o grupo é submetido a procedimentos ou sofrem algum tipo de intervenção que modifique o curso natural do fenômeno estudado durante o período de estudo.

Os estudos epidemiológicos analíticos avaliam a associação entre a exposição e o desfecho dos quais pode-se destacar três tipos de delineamentos: estudos transversais, estudos de coorte e estudos do tipo caso controle.

Em *estudos de coorte* o pesquisador seleciona um grupo de indivíduos expostos e um grupo de não expostos, acompanhando-os para comparar a incidência de doença em cada grupo.

A seleção dos grupos em estudos de coorte pode se dar de duas formas. Na primeira,

os indivíduos livres do desfecho são selecionados de acordo com a exposição. Na segunda, conhecida como estudo de coorte de base populacional, uma amostra é selecionada de uma população definida antes da exposição ocorrer ou ser identificada.

Em *estudos do tipo caso-controle* seleciona-se indivíduos com base no desfecho (se caso ou controle) e assim investiga-se a exposição vivenciada no passado por meio de registros ou entrevistas.

Estudos transversais (ou de prevalência) podem ser utilizados para fornecer uma descrição do estado da saúde da população ou de um grupo e auxiliar na elaboração de políticas públicas com baixo custo quando são de base populacional. Nesses estudos mede-se a exposição e o desfecho em um mesmo período de tempo.

As medidas de associação mais utilizadas em pesquisas epidemiológicas são o risco relativo (RR) e a razão de chances (RC). Essas medidas estimam a magnitude da associação entre um fator de risco e um desfecho de interesse. Em estudos de caso controle, apenas RCs podem ser estimadas. Em estudos transversais e de coorte pode-se estimar RCs ou RRs.

A escolha da estimação do RR se deve ao fato de tal medida ser de mais fácil interpretação. Além disso, a RC fornece uma boa estimativa do RR quando o desfecho de interesse é raro, uma das características dos estudos do tipo caso controle.

Em estudos transversais, os desfechos de interesse são geralmente comuns, ou seja a prevalência é relativamente alta. Nesses casos a RC não fornece boa aproximação para o RR. Barros e Hirakata (2003), Localio et al (2007) e Newcombe (2006) observam que a RC superestima o RR quando o resultado de interesse é maior que 10%. Portanto, quanto maior a frequência de desfecho de interesse, maior o viés da estimativa do RR pela RC.

Modelos de regressão log binomial e as aproximações por modelos logísticos, de Poisson e de Cox podem ser utilizados na estimação do RR. O modelo de regressão de Poisson pode fornecer estimativas de probabilidades de ocorrência do desfecho maiores que 1 e raramente observa-se problemas de convergência, mas não estima RR diretamente. As aproximações via modelos de Cox e de Poisson fornecem intervalos de confiança mais largos que os estimados pelo modelo log binomial. O modelo log binomial requer um algoritmo apropriado para o cálculo numérico do estimador de máxima verossimilhança (EMV)

A regressão logística tem bastante apelo em estudos epidemiológicos. Comparações

do modelo log binomial com as aproximações de Poisson e de Cox já foram realizadas (Blizzard e Hosmer (2006) e Chen et al (2014)).

Esse trabalho tem como objetivo a comparação de um método direto de estimação do RR em estudos transversais via regressão log binomial com um método indireto via regressão logística, propostos por Andrade e Carabin (2011) e Santos et. al. (2008), respectivamente.

Capítulo 1

Metodologia

1.1 Medidas de associação

Um dos principais propósitos de estudos epidemiológicos envolve a identificação de fatores de risco ou de proteção para desfechos de interesse. Para tanto, utiliza-se medidas de associação, tais como o risco relativo (RR) ou a razão de chances (RC), que estimam a magnitude dos efeitos de covariáveis sobre o desfecho de interesse.

Por exemplo, um pesquisador reúne uma série de informações de pacientes de uma clínica que trata de câncer de pulmão e deseja saber se existem fatores que potencializam o desenvolvimento ou a incidência do câncer. Para isso, ele avalia a existência de associação entre o desfecho e as variáveis explicativas, comparando os pacientes que apresentam com os que não apresentam certas características, como o hábito de fumar, o sexo, a idade, etc.

Por simplicidade, considera-se a situação em que se tem apenas uma variável explicativa binária X, tal que

$$X = \begin{cases} 1, & \text{se a exposição ocorreu (presença de fator de risco em potencial)} \\ 0, & \text{caso contrário.} \end{cases}$$

Supõe-se ainda que o desfecho de interesse seja uma variável resposta dicotômica Y definida por:

$$Y = \begin{cases} 1, & \text{se o desfecho ocorreu} \\ 0, & \text{caso contrário.} \end{cases}$$

Define-se π_1 como a probabilidade (ou risco) de indivíduos apresentarem o desfecho

entre os expostos e π_0 a probabilidade (ou risco) de indivíduos apresentarem o desfecho entre os não expostos. Em contraste, a chance seria definida como a razão entre a probabilidade de ocorrência do evento e sua probabilidade complementar $\pi_i/(1-\pi_i), i = 0, 1$, em que 0 representa o grupo de não expostos e 1, o de expostos.

O RR representa, portanto, a razão entre o risco da ocorrência do desfecho entre os expostos e o risco da ocorrência do desfecho entre os não expostos. Em contraste, a RC como a razão entre a chance da ocorrência do desfecho entre os expostos e a chance da ocorrência do desfecho entre os não expostos.

O RR é expresso por:

$$RR = \frac{\pi_1}{\pi_0}, \tag{1.1}$$

enquanto que a RC é expressa por:

$$RC = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}. \tag{1.2}$$

Na população, para ambas medidas, o valor 1 indica a não existência de associação ente exposição e desfecho. Um resultado superior a 1 indica associação positiva, ou seja, o risco (se RR) ou a chance (se RC) de ocorrência do desfecho é superior entre os expostos. Já um resultado inferior a 1 implica em menor risco ou chance de ocorrência do desfecho entre expostos.

Quando a variável explicativa é quantitativa ou se tem interesse em avaliar mais de uma covariável pode-se utilizar modelos de regressão como descrito a seguir.

1.2 Modelos de regressão para respostas binárias

Assumindo k variáveis explicativas e uma amostra com n observações independentes do par (x_{ij}, y_i), $i = 1, 2, ..., n$ e $j = 1, 2, ..., k$, tal que y_i e x_{ij}, são respectivamente, o valor da variável resposta dicotômica e o valor da j-ésiA variável resposta Y segue uma distribuição Bernoulli, que pode assumir os valores 0 ou 1, representando a ausência ou a presença do evento de interesse, respectivamente. Por outro lado, as variáveis explicativas podem ser quantitativas ou qualitativas. ma variável explicativa para o i-ésimo indivíduo.

A esperança condicional de Y dado X de um modelo de regressão para respostas

binárias e k variáveis explicativas é denotada por:

$$E(Y = 1|X = x_j) = \pi(x_j),$$

tal que $\pi(x_j)$ é a probabilidade dos indivíduos apresentarem a doença dado o valor da k-ésima variável explicativa e $x'_j = \begin{pmatrix} 1 & x_1 & x_2 & ... & x_k \end{pmatrix}$.

O modelo é representado por uma função não linear, mas a relação entre a transformação da função de probabilidade, denominada função de ligação $g(\pi(x_j))$, e os preditores é linear.

Tal relação é expressa por:

$$g(\pi(x_j)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = x'_j \beta,$$

tal que $\beta' = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & ... & \beta_k, \end{pmatrix}$.

O parâmetro β_j, para $j = 1, 2, ..., k$ retorna a taxa de aumento ou diminuição da função ligação para cada varição unitária da variável explicativa fixando-se as demais covariáveis. O sinal indica se a curva cresce ($\beta_j > 0$) ou decresce ($\beta_j < 0$). Quando $\beta_j = 0$, $\pi(x_j)$ é igual para cada observação, isto é, o desfecho independe do valor de x_j.

Os parâmetros desconhecidos são estimados pelo método de máxima verossimilhança, em que os valores são obtidos de modo que maximizem a função de verossimilhança, dada por:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_j)^{y_i} [1 - \pi(x_j)]^{1-y_i}.$$

Por sua vez a log verossimilhança e expressa por

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} y_i \ln[\pi(x_j)] + (1 - y_i) \ln[1 - \pi(x_j)]. \tag{1.3}$$

Para se obter os estimadores de máxima verossimilhança (EMV) dos parâmetros desconhecidos, as derivadas parciais da log verossimilhança são igualadas à zero. Porém essas expressões são não-lineares, fazendo-se necessária a aplicação de algum método iterativo, como por exemplo, o método iterativo de mínimos quadrados ponderados, o método de escore de Fisher ou o método de Newton Raphson.

1.2.1 Regressão Log binomial

Sob o modelo log binomial, a esperança condicional de Y dado X é denotada por:

$$E(Y = 1|X = x_j) = \pi(x_j) = e^{x'_j\beta}, \tag{1.4}$$

com função de ligação logarítmica expressa por:

$$g(x) = ln[\pi(x_j)] = x'_j\beta.$$

Os EMV de β são obtidos solucionando-se a seguinte expressão para o j-ésimo parâmetro:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_{ij} \frac{y_i - \pi(x_j)}{1 - \pi(x_j)} = 0. \tag{1.5}$$

A matriz de variâncias e covariâncias de $\hat{\beta}$ pode ser estimada pelo inverso da informação de Fisher observada tal como

$$\widehat{Var}(\hat{\beta}) = -\left[\frac{\partial^2 l(\beta)}{\partial \beta}\right]^{-1} = -\left[\sum_{i=1}^{n} x_{ij} x'_{ij} \frac{\pi(x_{ij})(1 - y_i)}{(1 - \pi(x_{ij}))^2}\right]^{-1}. \tag{1.6}$$

Há, porém, outras maneiras de se estimar $\widehat{Var}(\hat{\beta})$, descritas em Blizzard e Hosmer (2006), incluindo o uso da informação de Fisher esperada no lugar da observada e de estimadores robustos. No entanto, as simulações realizadas em tal estudo não mostraram diferenças significativas entre os métodos.

Os softwares usualmente utilizam o método iterativo de mínimos quadrados ponderados, como uma adaptação do algoritmo de Newton-Raphson para se resolver a Equação 1.5. No entanto, tal algoritmo é inadequado pois não utiliza restrições, o que causa problemas de convergência. Quando a convergência falha, comumente utiliza-se o estimador de quasi-verossimilhança, mas ele também nem sempre garante a convergência.

Por isso, a estimação do risco relativo apresentada por Andrade e Carabin (2011) deu-se pelo método de máxima verossimilhança via otimização restrita, através do *algoritmo adaptativo de barreira* (Nocedal e Wright, 2006), fornecendo probabilidades dentro do intervalo $[0, 1]$, com rara falha de convergência.

A estimativa do RR pode ser expressa por:

$$\widehat{RR} = \frac{E(Y = 1|x_j = a)}{E(Y = 1|x_j = b)} = \exp((a - b)'\beta).$$

Quando a k-ésima variável explicativa é dicotômica, então

$$\widehat{RR}_j = \exp(\beta_j).$$

O intervalo de confiança é expresso por:

$$IC_{1-\alpha}(\widehat{RR}) : exp\left[(log\widehat{RR}) \pm z_{\frac{\alpha}{2}}\widehat{EP}(log\widehat{RR})\right].$$

1.2.2 Regressão Logística

Sob o modelo de regressão logística, a esperança condicional de Y dado X é denotada por:

$$E(Y|X = x_j) = \pi(x_j) = \frac{\exp(x_j'\beta)}{1 + \exp(x_j'\beta)}, \tag{1.7}$$

com função de ligação logarítmica expressa como a seguir:

$$g(x) = ln\left[\frac{\pi(x_j)}{1 - \pi(x_j)}\right] = x_j'\beta.$$

O EMV de β_j é obtido solucionando-se a Equação 1.8 para o j-ésimo parâmetro.

$$\sum x_j[y - \pi(x_j)] = 0. \tag{1.8}$$

A matriz de variâncias e covariâncias de $\hat{\beta}$ pode ser estimada pelo inverso da informação de Fisher observada tal com

$$\widehat{Var}(\hat{\beta}) = -\left[\frac{\partial^2 l(\beta)}{\partial\beta}\right]^{-1} = -\left[-\sum_{j=1}^{n} x_j x_j'\pi(x_j)(1 - \pi(x_j))\right]^{-1}. \tag{1.9}$$

Santos et al (2008) descrevem métodos de estimação indireta do RR via modelo logístico. A fim de facilitar a compreensão dos métodos a serem apresentados, são considerados n indivíduos com a exposição dicotômica X_1 (0 = não expostas e 1 = expostas) e uma covariável contínua X_2.

O método condicional consiste em atribuir um valor padrão para a covariável contínua, em geral, a média. O RR estimado é dado por:

$$\widehat{RR} = \frac{1 + exp(-\hat{\beta}_0 - \hat{\beta}_2\bar{X}_2)}{1 + exp(-\hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2\bar{X}_2)},$$

tal que $\bar{X}_2$ é a média da variável X_2.

No método marginal, estima-se o risco para cada combinação de valores da variável dependente. Nesse caso o RR estimado é dado por:

$$\widehat{RR} = \frac{\frac{1}{n}\sum_i (1/(1 + exp(-(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_{2i}))))}{\frac{1}{n}\sum_i (1/(1 + exp(-(\hat{\beta}_0 + \hat{\beta}_2 X_{2i}))))}.$$

Pode-se usar tanto o método delta como o método bootstrap para se obter o erro padrão para o risco relativo. No presente trabalho optou-se por estimar o erro padrão pelo primeiro método com a finalidade de se obter o intervalo de confiança para o RR.

O método delta é uma técnica comum entre as utilizadas para funções de distribuições assintóticas de variáveis aleatórias, tendo como base a aproximação de uma série de Taylor. Assim, o intervalo de confiança é expresso por:

$$IC_{1-\alpha}(\widehat{RR}) : exp\left[(log\widehat{RR}) \pm z_{\frac{\alpha}{2}}\widehat{EP}(log\widehat{RR})\right].$$

Capítulo 2

Resultados

Avaliou-se a qualidade da estimação dos modelos log binomial e logístico em simulação tomando-se como base os cenários estabelecidos por Blizzad e Hosmer (2006). O software R (2015), foi utilizado em todas as simulações. Os pacotes prLogistic (Santos et al,2008) e LBReg(Andrade e Carabin, 2006) também foram necessários.

A função prLogisticDelta, do pacote prLogistic, fornece a estimativa do RR e o seu respectivo intervalo de confiança via método delta utilizando o modelo logístico para padronização condicional ou marginal (Santos et al, 2008).

O pacote LBReg (Andrade e Carabin, 2006) baseia-se no algoritmo adaptativo de barreira fornecido pelo constrOptim no lugar do método iterativo de mínimos quadrados ponderados. Ele é um pacote alternativo à função *glm* e está em processo de implementação no software R, parte apresentada por Ananias (2015). As funções lbreg e relrisk, desse pacote, fornecem os resultados da regressão log binomial e as estimativa do risco relativo para as variáveis explicativas e seus intervalos de confiança de simulações.

Dois tipos diferentes de estrutura de simulação foram gerado: o primeiro contém uma variável resposta dicotômica e uma explicativa contínua e o segundo contém uma variável resposta binária e duas explicativas (uma dicotômica e outra contínua).

Para se avaliar os modelos comparou-se o viés relativo percentual médio (VRPM), a média do erro quadrático médio do $log\widehat{RR}$ (MEQM($log\widehat{RR}$)) vezes cem e a cobertura. As duas primeira medidas são expressas, respectivamente, por:

$$VRPM = \frac{100}{m}\sum_{j=1}^{m}\frac{\widehat{RR}_j - RR_j}{RR_j}$$

e

$$MEQM(log\widehat{RR}) = \frac{100}{m}\sum_{j=1}^{m}(log\widehat{RR} - logRR)^2 + \widehat{Var}(log\widehat{RR}),$$

em que m é número de simulações.

A cobertura é porcentagem de intervalos de confiança de 95% obtidos pelas estimativas do RR que contêm o verdadeiro valor do parâmetro.

2.1 Simulação I

A obtenção da simulação do tipo I deu-se a partir do modelo log binomial de tal forma que a variável contínua explicativa X foi gerada de uma distribuição uniforme ($U(-6, a)$). A partir da variável X gerada e dos valores estabelecidos para β_0 e β_1 obtém-se a probabilidade $Pr(Y = 1|X) = \pi(x)$ para cada valor assumido por X, os valores de β_0 e β_1 foram escolhidos de forma que a prevalência, $Pr(Y = 1)$, seja aproximadamente $0,1$ ou $0,2$. Sendo assim, foi possível obter a variável resposta dicotômica Y, isto é, para cada valor de X é associada uma probabilidade de ser obter valor 1 ou 0 ao Y correspondente.

Tabela 2.1: Estrutura dos cenários - Simulação I

Cenário	a*	β_0	β_1	RR_{β_1}	$Pr(Y = 1)$
1	6	-2,30	0,38	1,47	0,22
2	4	-2,30	0,38	1,47	0,12
3	2	-1,20	0,57	1,76	0,20
4	1	-1,20	0,57	1,76	0,13
5	1	-0,69	0,65	1,92	0,21
6	0	-0,69	0,65	1,92	0,13
7	0,50	-0,36	0,71	2,03	0,21
8	-0,50	-0,36	0,71	2,03	0,12

*Limite superior de $X \sim U(0, a)$

A Tabela 2.1 apresenta 8 cenários para os quais foram utilizados valores diferentes de a, β_0 e β_1 para simulações de 1000 amostras de tamanho 500. Estima-se então RRs utilizando os dois métodos, obtendo-se os resultados apresentados na Tabela 2.2.

A distinção entre os cenários, seguindo do primeiro ao oitavo, é caracterizada pela diminuição do valor máximo do intervalo da uniforme, distribuição da variável explicativa X, redução nos valores dos parâmetros, aumento do RR e intercalação da prevalência (ora próximo de $0,1$, ora de $0,2$).

Observa-se que o VRPM do $\widehat{RR}$ estimado pelo modelo log binomial não ultrapassa

Tabela 2.2: Resultados da simulação I para a estimação do RR via modelo log binomial e logístico

	Log Binomial			Logístico		
Cenário	VRPM	MEQM*	Cobertura	VRPM	MEQM*	Cobertura
1	-0,10	0,16	95,30	13,58	2,44	51,10
2	0,82	0,66	94,80	2,35	0,94	95,80
3	-0,08	0,46	94,70	-11,00	1,85	29,80
4	1,21	1,19	95,90	-13,42	2,54	15,20
5	0,11	0,54	96,50	-27,20	10,28	0,00
6	1,27	1,79	95,20	-25,91	9,20	0,00
7	-0,27	0,55	97,20	-36,91	21,38	0,00
8	1,62	2,08	95,00	-33,81	17,28	0,00

$^*MEQM(logRR) * 100$

$1,7\%$ e para os cenários 1, 3 e 7 o viés foi negativo. Em outras palavras, a distância entre a estimativa e o verdadeiro valor é, em módulo, no máximo $1,7\%$. O VRPM positivo indica que, em média, a estimativa do RR é maior que o seu real valor, enquanto que, o VRPM negativo indica que, em média, a estimativa é menor.

Em contraste, o VRPM do $\widehat{RR}$ via modelo logístico apresentou valor em módulo superior a 11%, com exceção do cenário 2, sendo a maior parte com VRPM negativo e nos quatro últimos cenários, essa medida foi inferior -25%. Isto é um indicativo de que a estimativa do RR é, em média, subestimada. Nota-se também que, para esses cenários, a cobertura é igual a zero, ou seja, nenhum dos 1000 intervalos de confiança de 95% gerados pelas estimativas incluiu o verdadeiro valor do RR, enquanto que para os outros quatros cenários, três obtiveram cobertura inferior a 52%, mas o cenário 2 obteve cobertura de aproximadamente 96%. Em contraste, as estimativas obtidas via modelo log binomial obtiveram cobertura maior que $94,7\%$ para todos os cenários.

O erro quadrático médio é uma medida que pode ser divida em duas componentes, a precisão e o viés. Logo, quanto menor essa medida, mais precisa e acurada é a estimativa. No entanto, quando tem-se viés pequeno e alto EQM, as estimativas estão dispersas em torno do verdadeiro valor do parâmetro estimado. Por outro lado, quando tem-se baixa variabilidade e alto VRPM, as estimativas são mais precisas e pouco acuradas. No caso do modelo logístico para os sete cenários, o viés foi alto e a precisão foi baixa.

Em média, o EQM (vezes 100) das estimativas do $logRR$ via regressão log binomial variou entre 0,16 e 2,08, enquanto que, via logística variara entre 0,94 e 21,38. Como o VRPM e o EQM são maiores para o segundo método do que para o primeiro. Com base

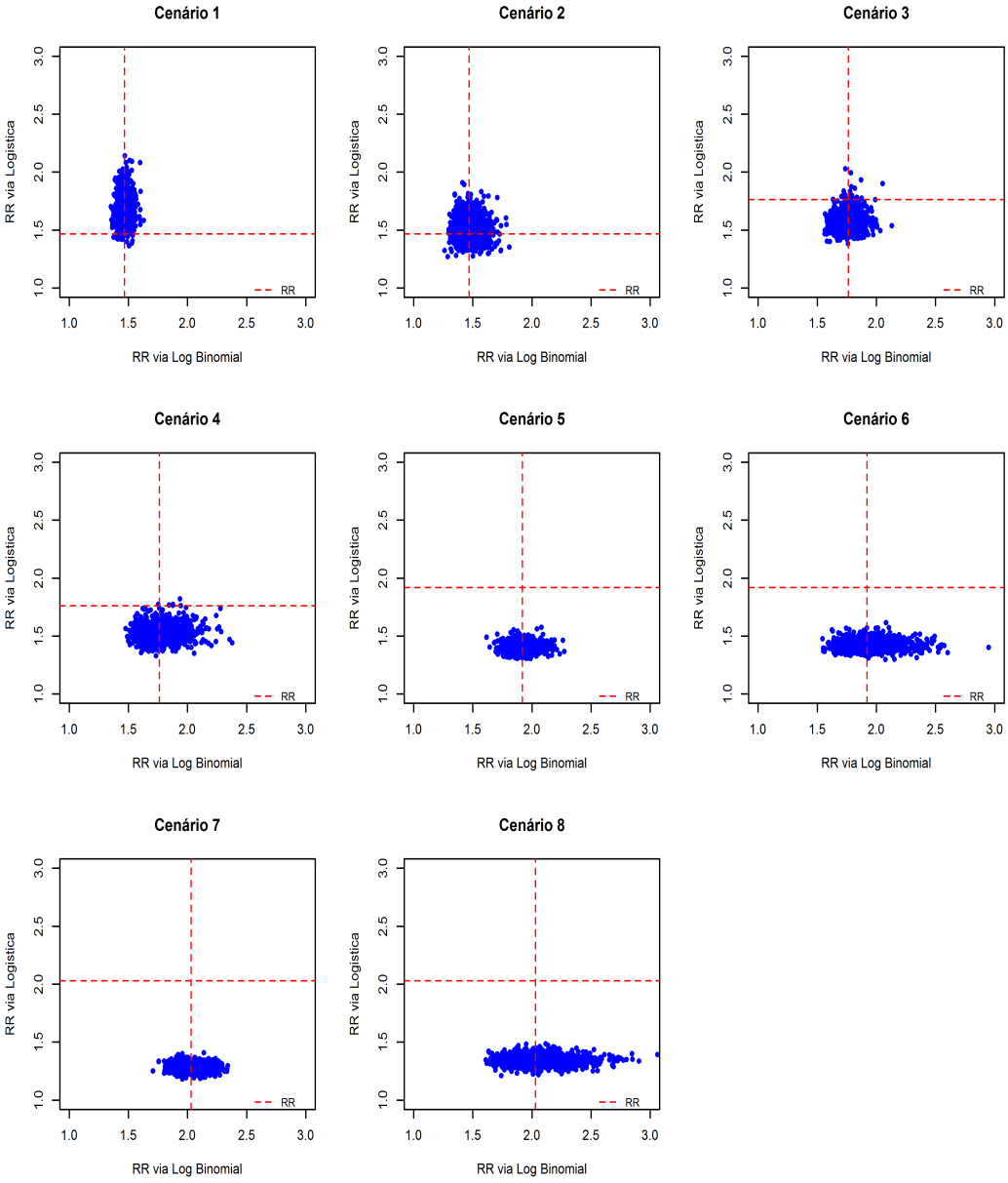


Figura 2.1: Estimativas do RR via modelo log binomial e logístico

na Figura 2.1, de modo geral, observa-se que as estimas obtidas por meio da logística são pouco dispersas. Portanto, apesar de uma boa precisão, essas estimativas são enviesadas e pouco acuradas. Os cenários 1 e 2 são exceção, pois o primeiro tem estimativas mais dispersas e o segundo, melhor acurácia. Em contraste, apesar da maior variabilidade, as estimativas via log binomial possuem viés muito menor em média.

A Figura 2.1 permite uma melhor compreensão dos resultados destacados, pois pode-se perceber que a nuvem de pontos está melhor distribuída em torno do eixo vertical tracejado (valor real do RR) nos 8 cenários que no eixo horizontal tracejado (valor real do RR). A nuvem concentrada abaixo do eixo horizontal, nos cenários 5 a 8, evidencia

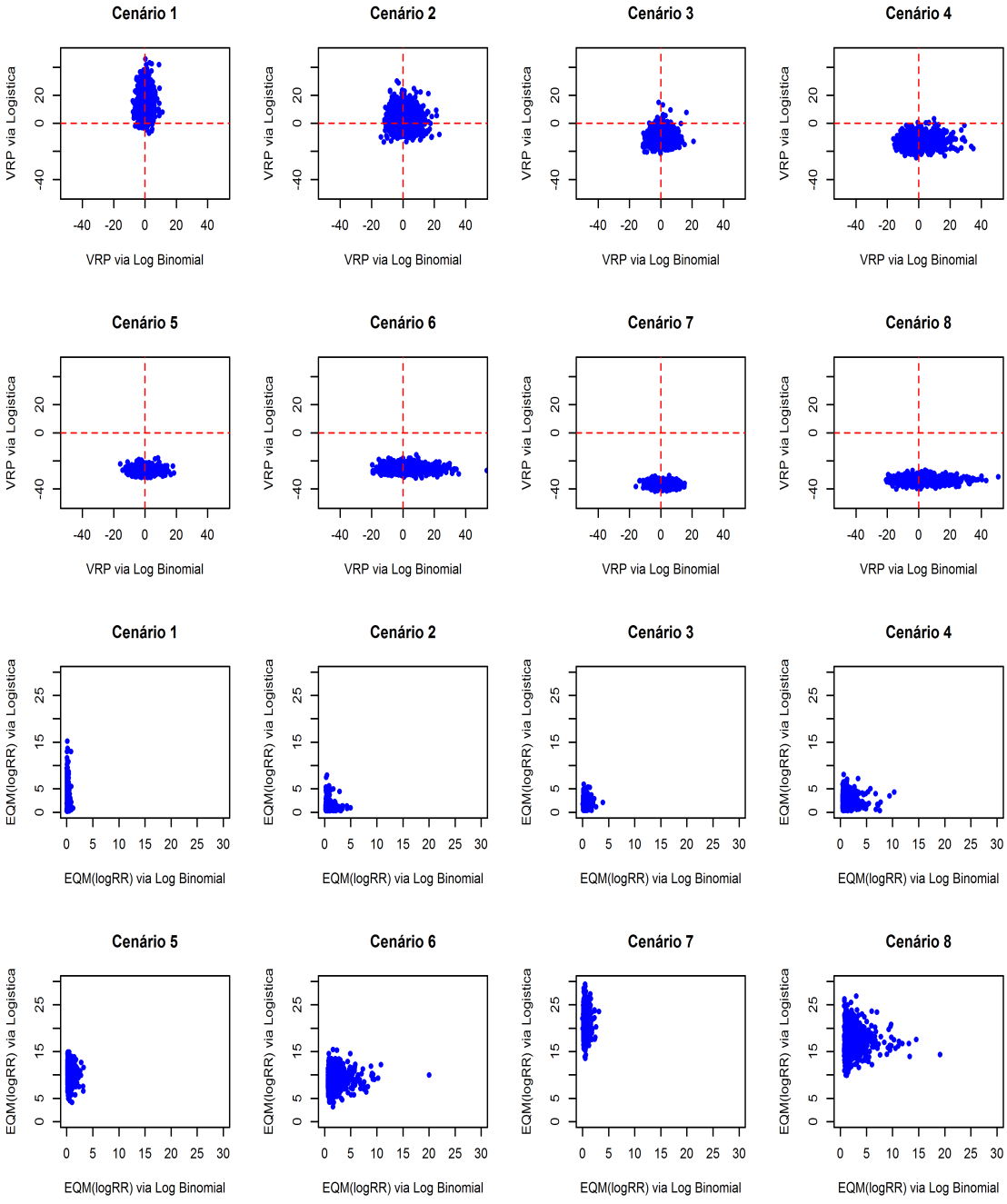


Figura 2.2: VRP e $EQM(\widehat{logRR})$ via modelo log binomial e logístico

que nenhuma estimação assumiu o real valor e que o RR foi subestimado pelo modelo logístico.

2.2 Simulação II

Para se obter a simulação do tipo II, também utilizou-se o modelo log binomial. No entanto, com duas variáveis explicativas. Inicialmente foi gerada a variável dicotômica D,

em que cada valor foi gerado por uma probabilidade p preestabelecida, isto é, simulando-se um ensaio de Bernoulli. A outra variável explicativa U foi obtida por meio da distribuição uniforme no intervalo $(-6 + 2 * D, 2 + 2 * D)$, em que D, é a variável binária gerada que assume 0 ou 1. Por fim, gerou-se Y de forma análoga ao processo anterior, tal que para cada valor de U e de D é associado uma probabilidade de ser obter valor 1, que indica a ocorrência do desfecho de interesse.

Tabela 2.3: Estrutura dos cenários - Simulação II

Cenário	p^*	β_0	β_D	RR_{β_D}	β_U	RR_{β_U}	$Pr(Y = 1)$
1	0,20	-1,20	0,41	1,50	0,18	1,20	0,27
2	0,50	-1,20	0,41	1,50	0,18	1,20	0,35
3	0,20	-1,20	0,69	2,00	0,10	1,11	0,32
4	0,50	-1,20	0,69	2,00	0,10	1,11	0,42

*Probabilidade para gerar $D_i \sim Bernoulli(p)$, em que $i = 1, ..., 4$

A Tabela 2.3 apresenta os 4 cenários trabalhados com a geração de 1000 amostras de tamanho igual a 500, a fim de se obter as estimativas dos RRs para as duas variáveis explicativas. Os resultados de VRPM, MEQM e cobertura para as variáveis D e U para os 2 modelos em 4 cenários são apresentados na Tabela 2.4.

Variações nos cenários envolveram a alternância diferentes probabilidades para a geração da variável explicativa dicotômica D e os valores de β_D e β_U, ocasionando no aumento da prevalência, à medida em que se observa do primeiro ao quarto cenário. β_0 foi o mesmo para todos os cenários.

Tabela 2.4: Resultados da simulação II para estimação do RR via modelo log binomial e logístico

Cenário	$VRPM_D$	$MEQM_D*$	$Cobertura_D$	$VRPM_U$	$MEQM_U*$	$Cobertura_U$
LOG BINOMIAL						
1	0,02	4,67	95,40	0,26	0,22	94,60
2	0,78	3,00	94,90	-0,02	0,15	94,10
3	1,04	3,19	93,40	0,24	0,15	94,40
4	0,70	2,58	95,20	0,12	0,09	94,40
LOGÍSTICO						
5	0,51	5,48	96,20	-1,84	0,20	87,60
6	0,34	3,54	96,30	-1,61	0,17	87,70
7	1,63	3,23	96,30	-0,92	0,13	95,70
8	0,27	2,84	95,20	-0,79	0,10	95,60

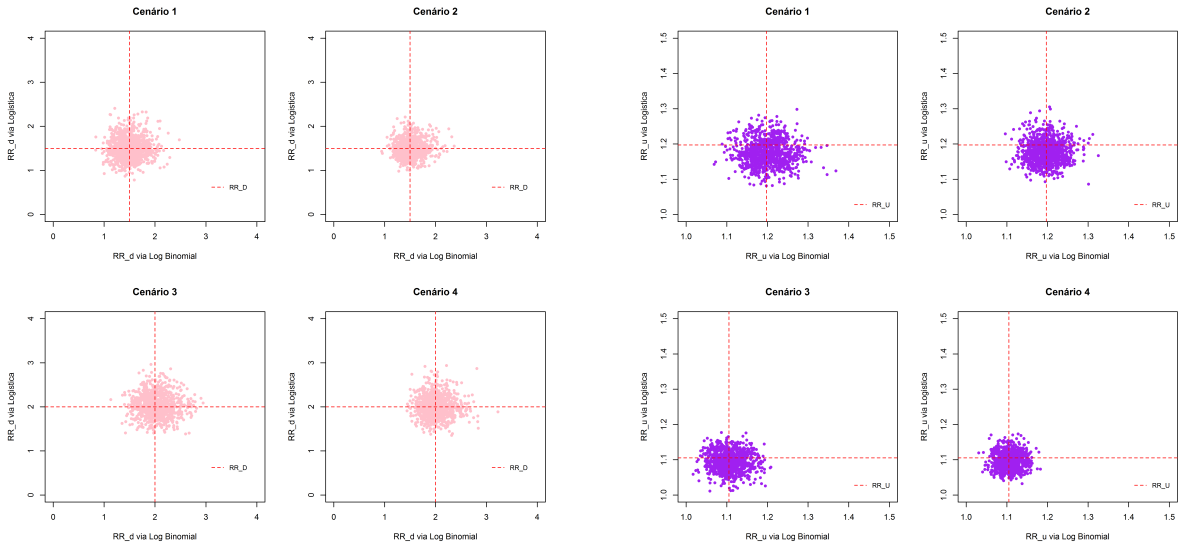
*$MEQM(logRR) * 100$

Os resultados apresentados na Tabela 2.4 para as variáveis D e U são semelhantes para o modelo log binomial e logístico, em que o VRPM não ultrapassa 2% em módulo.

No entanto, o $VRPM_U$ para os 4 cenários e o $VRPM_D$ para 3 dos 4 são menores quando estimados via regressão log binomial, em que o $VRPM_D$ varia entre 0,02% a 1,04% e o $VRPM_U$ varia em módulo entre 0,02% a 0,26% via log binomial. Já para a estimação via modelo logístico, o $VRPM_D$ varia entre 0,27% a 1,63% e o $VRPM_U$ varia em módulo entre 0,79% a 1,84%. Logo, os $\widehat{RR}$s estão mais próximos de verdadeiro valor do parâmetro quando estimados via log binomial.

Os quatros cenários têm $MEQM_D$ menor quando $\widehat{RR}_D$ é obtida pela modelo log binomial. Entretanto, a $Cobertura_D$ é pelo menos 2% maior para as estimativas obtidas pela logística para os cenários 2 a 4, enquanto que as $MQEM_U$s das estimativas são aproximadamente as mesmas para ambos métodos considerando-se cenário a cenário.

A $Cobertura_U$ para o 1° e 2° cenários é inferior a 90% quando a estimação é reduzida via regressão logística. Para os cenários 3 e 4, ambos os modelos forneceram cobertura em torno de 95%. Portanto, pelo menos 85% dos 1000 intervalos de confiança estimados contêm o real valor do parâmetro de interesse. Os baixos valores de $MEQM_U$ e o $VRPM_U$ que reforçam o que pode ser observado por meio dos gráficos na Figura 2.3 indicando que as estimações são acuradas e precisas.



(a) Estimativas do RR_D (b) Estimativas do RR_U

Figura 2.3: Estimativas do RR_D e do RR_U via modelo log binomial e logístico

A Figura 2.3 mostra que os $\widehat{RR}$s para as variáveis explicativas D e U sob os 4 diferentes cenários dos conjuntos de simulação tipo II. As estimativas aparentam estar aleatoriamente distribuídas em torno do real valor do RR. Diante disso, é possível se observar que

ambos modelos possuem boa acurácia, entretanto, o RR referente à variável explicativa U é estimado com maior precisão. A Figura 2.4 que contém os gráficos para VRP e o $EQM(log\widehat{RR})$ das estimativas do RR_D e RR_U também reforça essa ideia.

Portanto, o modelo log binomial, apesar de apresentar resultados semelhantes ao modelo logístico nesse segundo conjunto de simulação, ainda apresenta menor VRPM para ambas as variáveis explicativas, com exceção do cenário 2. Apresenta ainda melhor cobertura no contexto geral, pois a cobertura do modelo log binomial para ambas as variáveis varia em torno de 95%, ao contrário da logística, em que a $Cobertura_U$ somente varia em torno de 95% nos cenários 3 e 4. Os cenários 1 e 2 apresentam cobertura inferior a 88%.

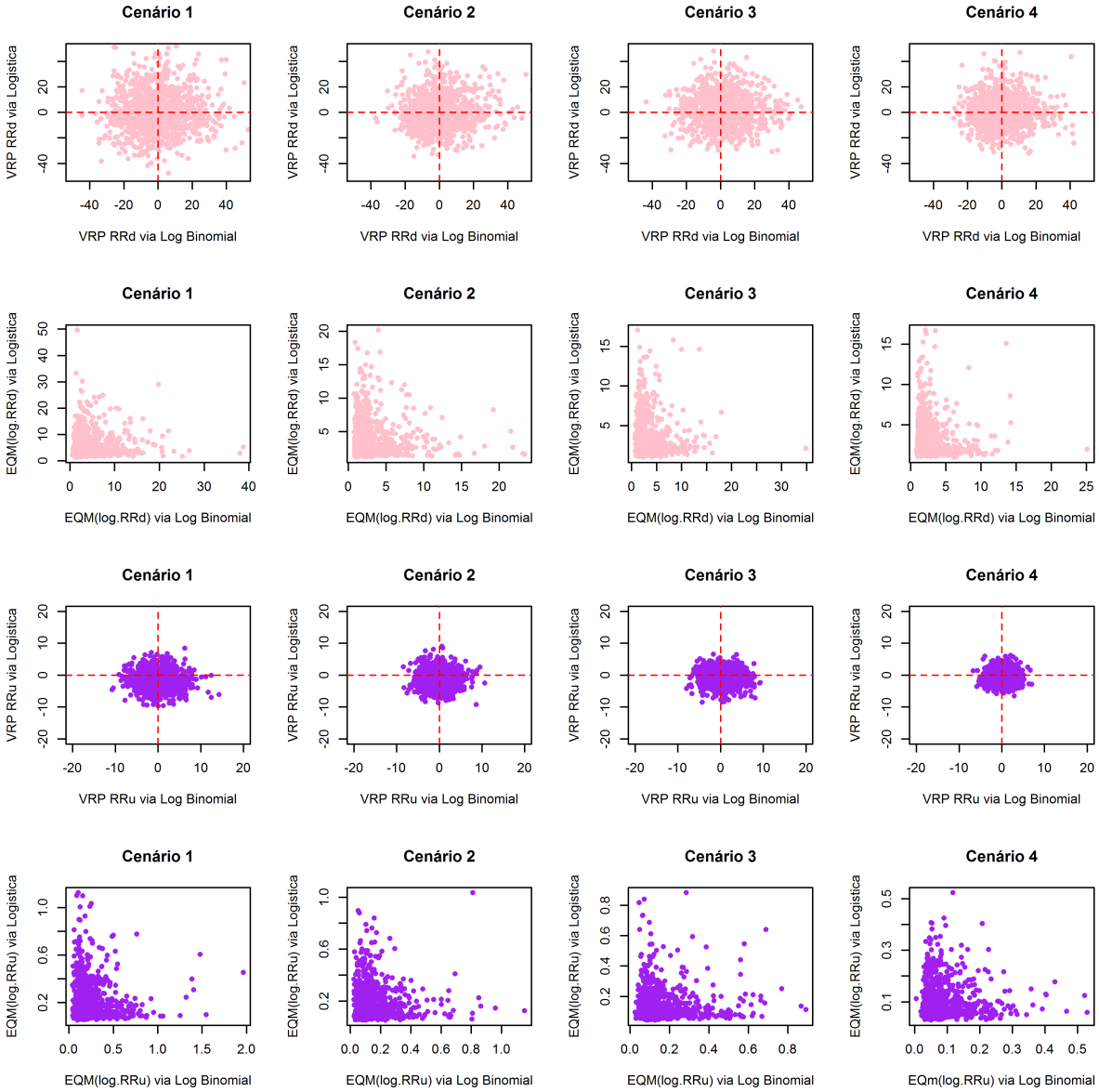


Figura 2.4: VRP e $\widehat{EP}(log\widehat{RR})$ das estimativas $\widehat{RR}_D$ e $\widehat{RR}_U$ via Log Binomial e Logística

2.3 Influência do tamanho da amostra

A observação do comportamento da estimação a partir do tamanho da amostra para os modelos em estudo torna mais evidente o comportamento assintótico das estimativas. Normalmente espera-se que a precisão das estimativa do $log(\widehat{RR})$ aumente quando $n \to \infty$, tal que n é o tamanho da amostra, e que a variância seja mínima.

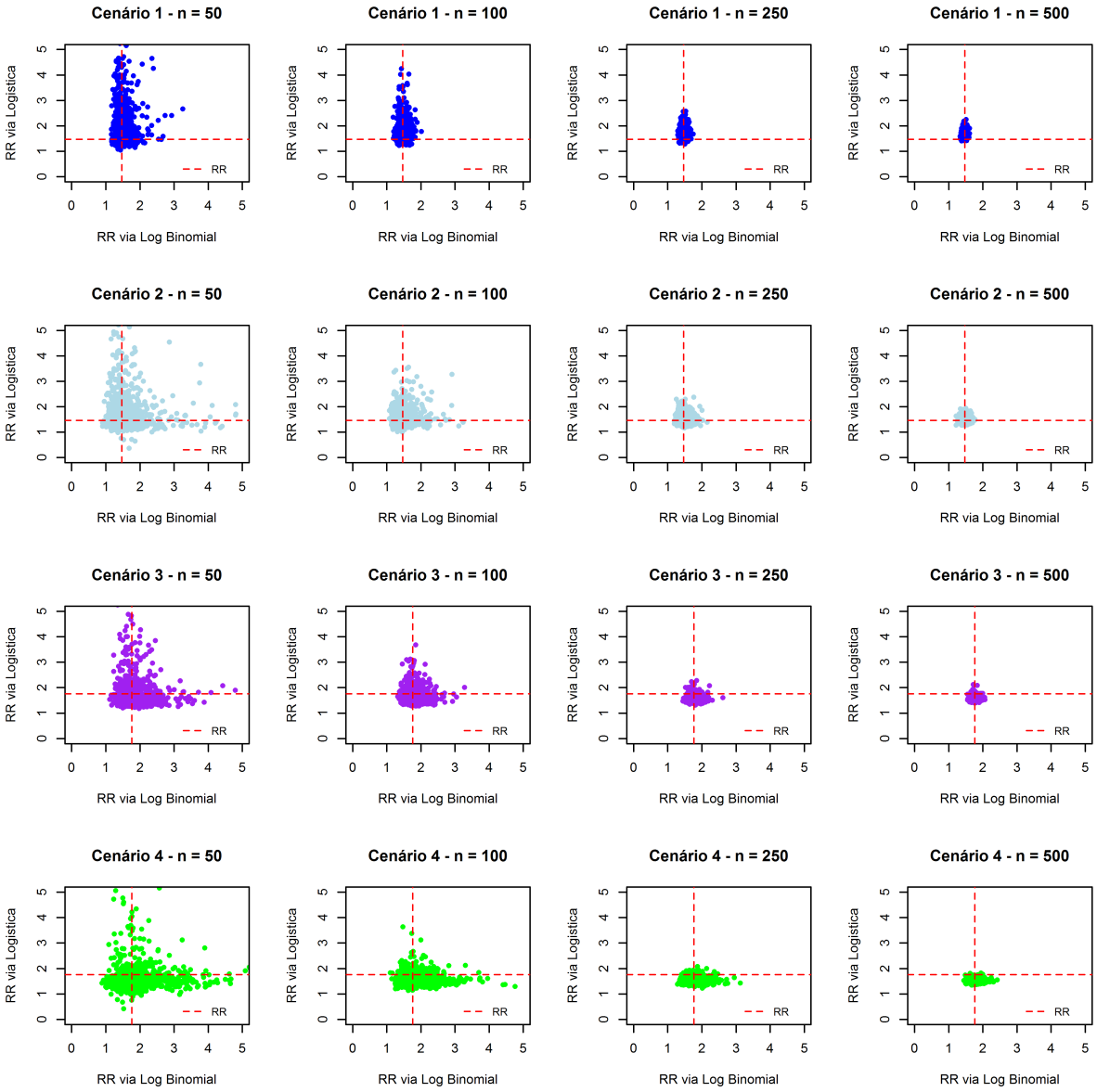


Figura 2.5: $\widehat{RR}$ via modelo log binomial e logístico segundo tamanho da amostra Simulação I para os cenários 1 a 4.

A Tabela 2.5 apresenta os resultados para as simulações do tipo I segundo o tamanho da amostra. Nota-se que, de fato, a MEQM diminui consideravelmente com o aumento da amostra para todos os cenários em ambos os modelos. No entanto, o VRPM diminui

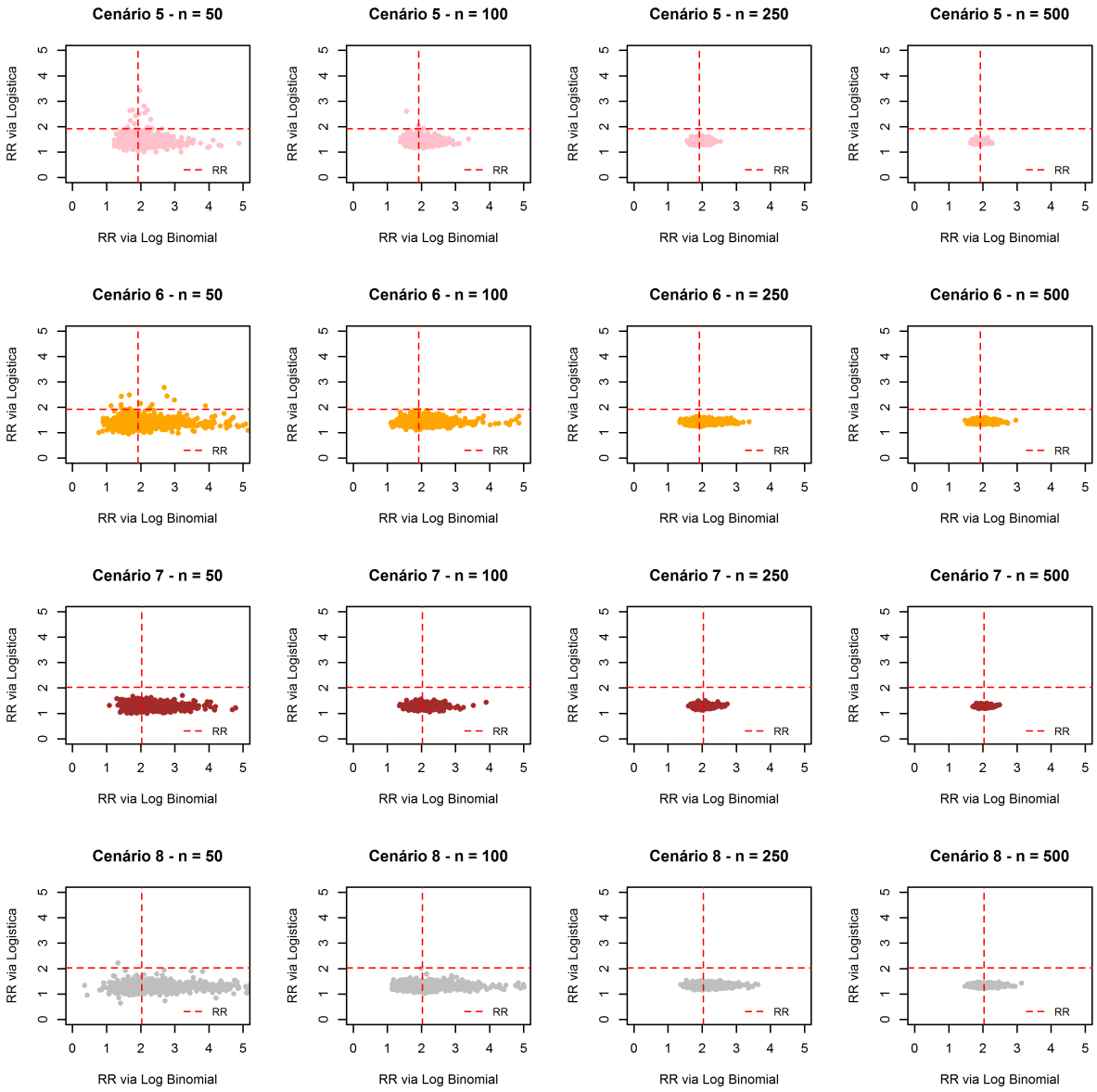


Figura 2.6: $\widehat{RR}$ via modelo log binomial e logístico segundo tamanho da amostra - Simulação I para os cenários 4 a 8

em todos os cenários apenas para o modelo log binomial.

As estimativas obtidas por meio do modelo logístico apresentam diminuição da MEQM e do VRPM para os quatro primeiros cenários. Para os outros quatros, a MEQM diminui pouco e o VRPM mantém-se próximo e bastante alto para os quatro tamanhos de amostra.

Tais considerações também podem ser observadas nas Figuras 2.5 e 2.6. Portanto, poderia-se dizer que as estimativas fornecidas pelo log binomial são assintoticamente mais consistentes ao se comparar com as fornecidas pela logística.

Tabela 2.5: Resultados da simulação I para estimação do RR via modelos log binomial e logístico considerando o tamanho da amostra

	LOG BINOMIAL			LOGÍSTICO		
n	VRPM	MEQM*	Cobertura	VRPM	MEQM*	Cobertura
Cenário 1						
50	2,41	2,88	95,90	63,86	39,91	97,00
100	0,52	1,04	95,60	20,39	9,31	91,80
250	-0,15	0,35	95,00	14,38	3,60	76,50
500	-0,35	0,16	94,70	13,14	2,36	55,50
Cenário 2						
50	646,74	923,64	94,20	Inf		96,20
100	3,54	4,36	95,20	8,70	7,35	97,00
250	1,88	1,51	94,50	3,49	2,10	95,20
500	0,69	0,67	94,40	2,08	0,92	95,30
Cenário 3						
50	5,47	7,55	95,30	$2{,}76{*}10^{42}$	Inf	76,40
100	1,94	2,79	95,20	-6,96	4,54	70,20
250	0,79	0,99	95,50	-10,35	2,32	50,60
500	0,44	0,47	95,50	-10,98	1,86	28,30
Cenário 4						
50	42,98	1315,67	94,90	$1{,}37{*}10^{118}$	Inf	73,20
100	7,06	8,44	95,20	-11,91	5,07	62,20
250	2,35	2,73	94,00	-12,83	2,91	39,10
500	0,93	1,17	95,60	-13,38	2,54	16,40
Cenário 5						
50	12,93	11,47	96,80	$1{*}10^{10}$	Inf	22,10
100	1,79	3,25	95,90	-26,79	11,09	6,70
250	0,28	1,11	96,20	-27,16	10,47	0,00
500	0,03	0,57	95,70	-27,21	10,29	0,00
Cenário 6						
50	$2{,}87{*}10^{5}$	225,47	94,70	-27,95	15,12	27,80
100	9,26	12,14	95,70	-26,50	11,08	8,80
250	2,24	3,63	94,50	-26,22	9,71	0,00
500	1,53	1,75	95,40	-25,99	9,27	0,00
Cenário 7						
50	8,77	11,35	95,40	-37,87	24,62	3,10
100	0,88	3,42	94,60	-37,14	22,45	0,00
250	-0,14	1,21	95,20	-37,04	21,75	0,00
500	-0,33	0,57	94,30	-36,94	21,42	0,00
Cenário 8						
50	$1{,}10{*}10^{36}$	$5{,}41{*}10^{3}$	93,90	-36,91	25,42	11,40
100	8,81	13,47	94,50	-33,81	17,28	0,00
250	3,00	4,58	95,20	-34,09	17,87	0,00
500	1,08	2,07	95,00	-33,77	17,22	0,00

*$MEQM(logRR) * 100$

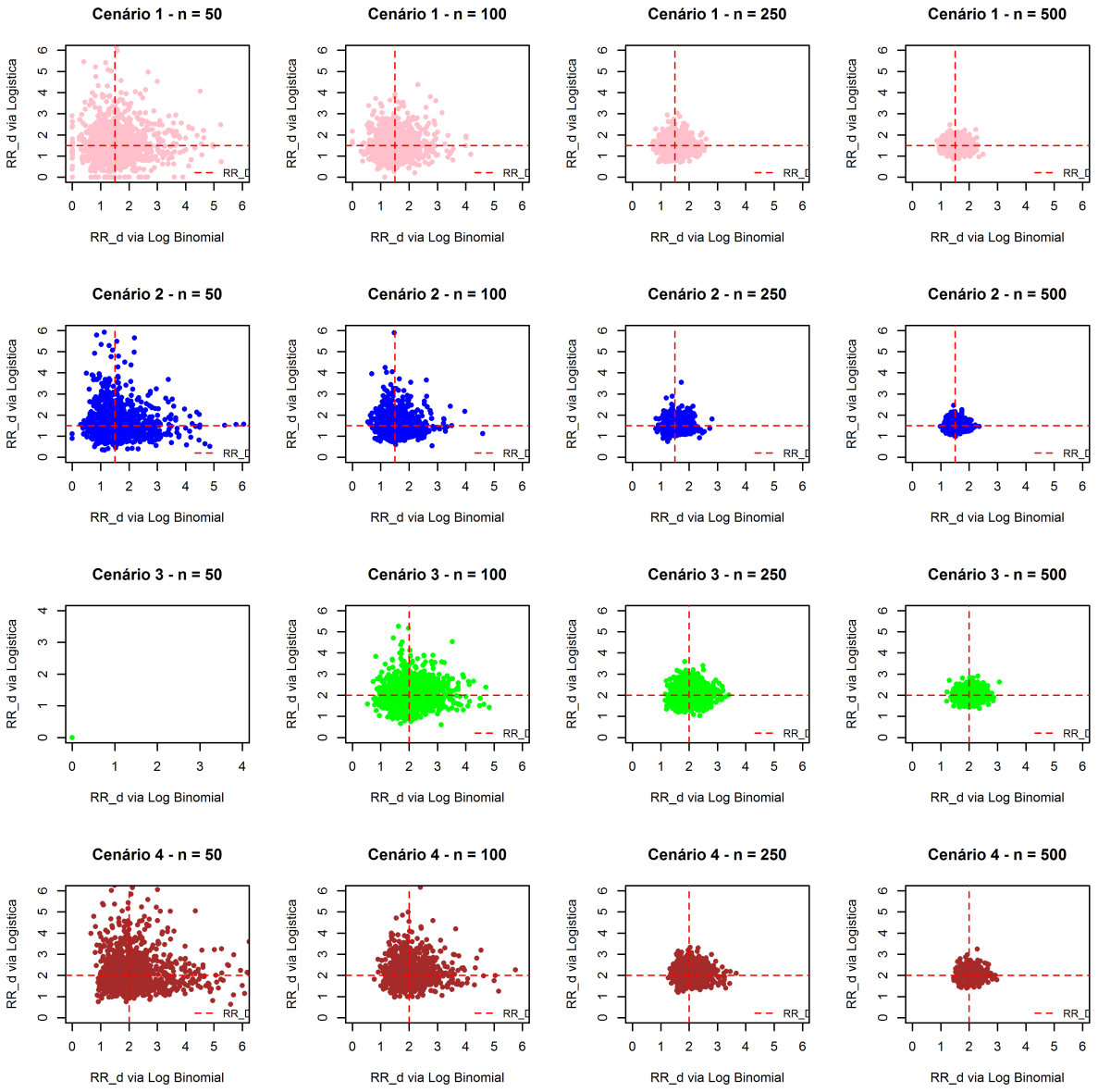


Figura 2.7: $\widehat{RR}_D$ via modelo log binomial e logístico segundo tamanho da amostra - simulação II

Os resultados obtidos para as simulações do tipo II para diferentes tamanhos de amostras estão apresentados na tabela 2.7. Observa-se que à medida que o tamanho da amostra aumenta, as $MEQM_D$ e $MEQM_U$ diminuem para ambos os métodos. No entanto, a $MEQM_D$ dos $\widehat{RR}_D$s obtidos por meio da regressão logística é maior que a dos obtidos pela log binomial. Já, a $MEQM_U$ das estimativas obtidas por meio da regressão logística é próxima à das obtidas pela log binomial.

O aumento do tamanho da amostra acarreta na diminuição do $VRPM_D$ dos $\widehat{RR}$s. Essa medida é maior para os quatros cenários em que o RR é estimado pelo modelo log binomial. Todavia, o $VRPM_U$ apenas diminui quando o RR é estimado pelo modelo

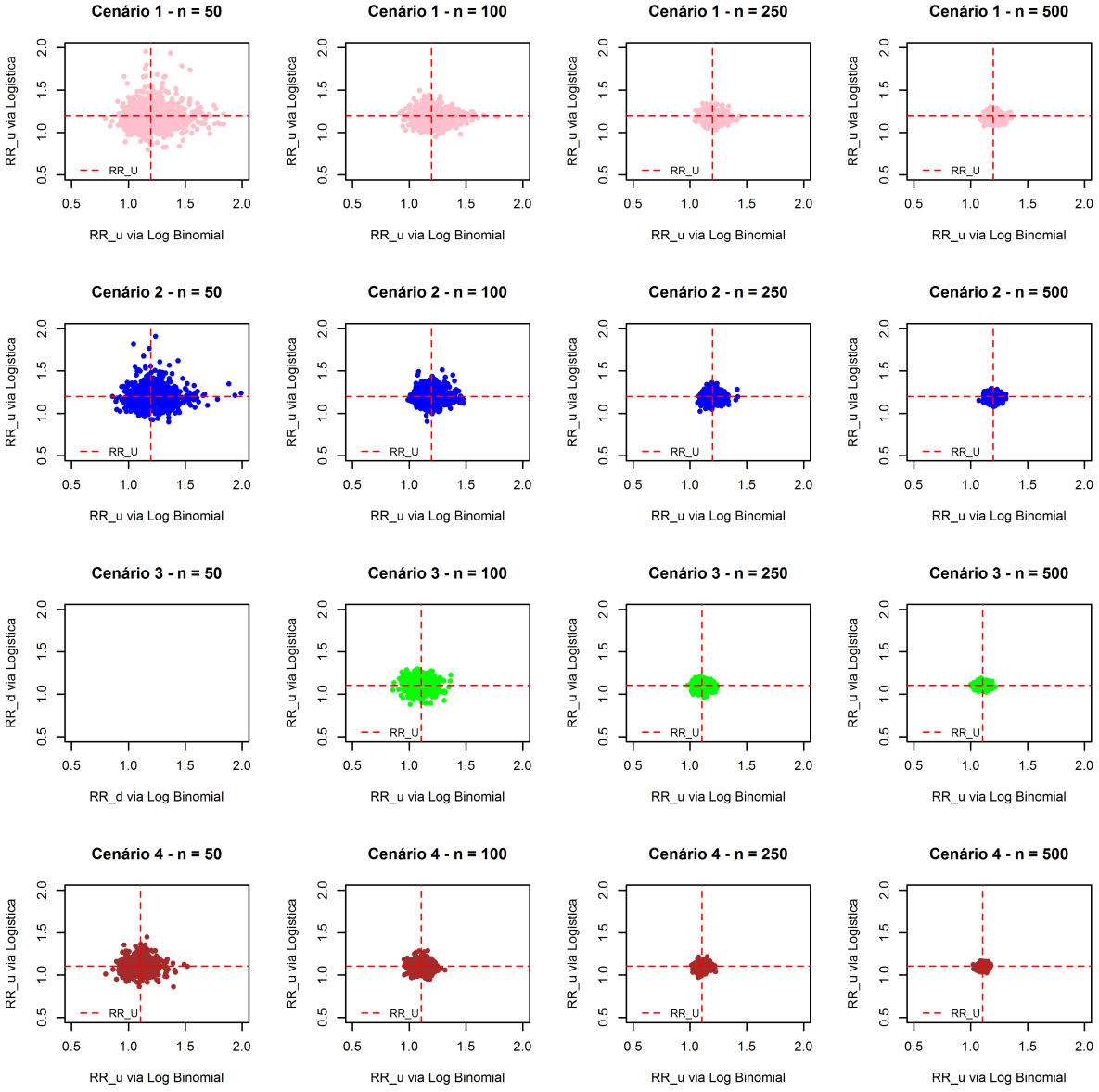


Figura 2.8: $\widehat{RR}_U$ via modelo log binomial e logístico segundo tamanho da amostra - simulação II

log binomial, para o modelo logístico o viés aumenta em módulo para os cenários 1,2 e 4, indicando que tal método produz estimativas mais precisas e menos acuradas que a estimação via log binomial.

As figuras 2.7 e 2.8 evidenciam que à medida que a amostra aumenta, menor é a variabilidade entre as estimativas. Pode-se observar também que para a variável contínua uniforme as estimativas por ambos os modelos são precisas.

Inicialmente pretendeu-se observar também as estimações do RR para amostras de tamanho 25. No entanto, para a simulação do tipo I não consegui-se obter nenhum resultado, diferentemente do método log binomial (resultados na Tabela 2.6). Não se

Tabela 2.6: Resultados da simulação I para amostra de tamanho 25 via regressão log binomial

Cenário	VRPM	MEQM	Cobertura
1	$2{,}77*10^{3}$	1511,196	95,7
2	$8{,}89*10^{11}$	$2{,}03*10^{4}$	97,4
3	$7{,}51*10^{5}$	$7{,}36*10^{3}$	96,2
4	$4{,}26*10^{50}$	Inf	97,0
5	$1{,}93*10^{33}$	$1{,}12*10^{4}$	96,3
6	$1{,}27*10^{30}$	$1{,}15*10^{5}$	97,6
7	$6{,}48*10^{14}$	$4{,}24*10^{3}$	96,0
8	$5{,}27*10^{20}$	Inf	97,1

realizou simulações do tipo I para esse tamanho de amostra.

Por fim, faz-se necessário destacar que para as simulações do tipo I e II, foram geradas amostras de tamanho 50 para ambos métodos. Porém, simulações de amostras de tamanho 25 só foram possíveis para o modelo log binomial. Nota-se que alguns $VRPM$s e $MEQM$s com valores exorbitantes, principalmente quando a amostra é de tamanho 25, e mesmo assim, com uma cobertura superior a 93% em alguns casos. Provavelmente tal fenômeno deve-se à uma estimativa ou algumas estimativas discrepantes que influenciaram presença de tais medidas, que são valores médios.

Tabela 2.7: Resultados da simulação II para estimação do RR via modelos log binomial e logístico pelo tamanho da amostra

	LOG BINOMIAL			LOGÍSTICO		
n	VRPM.D	MEQM.D	Cobertura.D	VRPM.D	MEQM.D	Cobertura.D
Cenário 1						
50	8.65	4827.17	94.60	6.16	Inf	98.20
100	2.77	228.14	95.00	3.80	Inf	97.00
250	1.13	9.49	95.00	1.22	10.71	97.10
500	1.09	4.61	95.10	0.21	5.14	97.00
Cenário 2						
50	$1.85 * 10^3$	838.88	94.70	$4.82 * 10^7$	Inf	98.00
100	5.06	16.70	94.90	6.40	18.98	97.80
250	2.04	6.00	95.00	2.73	7.23	97.40
500	1.41	2.89	95.70	0.56	3.33	97.90
Cenário 3						
50	-	-	-	7.50	Inf	97.50
100	4.95	17.87	93.60	4.75	18.52	96.70
250	1.22	6.35	94.20	1.26	6.91	95.70
500	0.59	3.19	94.10	-0.30	3.36	96.80
Cenário 4						
50	498.15	125.16	96.30	18894151.14	Inf	98.00
100	4.50	14.19	94.40	7.56	15.79	96.80
250	2.76	5.37	95.20	2.30	5.58	96.20
500	1.17	2.59	95.90	0.81	2.77	96.80
	LOG BINOMIAL			LOGÍSTICO		
n	VRPM.U	MEQM.U	Cobertura.U	VRPM.U	MEQM.U	Cobertura.U
Cenário 1						
50	1.98	2.87	94.80	-1.27	2.33	98.30
100	0.56	1.24	94.00	-1.61	0.87	96.10
250	0.29	0.43	94.50	-1.75	0.34	94.80
500	0.05	0.22	94.80	-1.94	0.20	87.40
Cenário 2						
50	1.34	1.92	93.00	-0.53	1784.21	98.30
100	0.34	0.77	92.60	-1.16	0.79	96.10
250	0.24	0.30	93.50	-1.34	0.31	92.20
500	-0.02	0.14	94.90	-1.63	0.17	87.70
Cenário 3						
50	-	-	-	-1.22	1.66	98.40
100	0.38	0.82	93.60	-0.77	0.68	98.20
250	0.21	0.31	94.20	-0.84	0.25	97.80
500	0.17	0.15	94.70	-0.80	0.13	95.80
Cenário 4						
50	0.51	1.04	94.30	-0.55	836.33	99.40
100	0.15	0.44	94.20	-0.99	0.48	98.40
250	0.08	0.18	95.90	-0.89	0.19	96.20
500	-0.07	0.09	93.50	-1.03	0.10	94.80

*** Não foi possível obter-se estimativas para o cenário 3 com amostra de tamanho 50 via log binomial**

Capítulo 3

Considerações Finais

Nesse estudo comparou-se os métodos de estimação do RR implementados por Andrade e Carabin (2011) e por Santos et. al. (2008), via regressão log binomial e logística, respectivamente. Para isso, realizou-se simulações com duas estruturas de simulações com base em cenários estabelecidos por Blizzar e Hosmer (2006).

A comparação dos métodos baseou-se na geração de 1000 amostras de tamanho igual a 500 a fim de se avaliar os métodos para grandes amostras. Assim, observou-se nas simulações do tipo I que o método indireto subestima o RR à medida que os valores dos parâmetros e o intervalo da variável contínua decrescem, chegando a 0% de cobertura, apesar da alta precisão. Nas simulações do tipo II ambos os modelos forneceram resultados próximos quanto ao viés e à variabilidade das estimativas do RR. Entretanto, como um todo, o método direto apresentou menor VRPM e melhor cobertura, enquanto que, o método indireto se mostrou ligeiramente melhor apenas na estimação do RR da variável dicotômica.

A utilização de diferentes cenários para a comparação das estimativas do RR, evidencia que, apesar do aumento na precisão, o VRPM das estimativas via método indireto é maior, especialmente para as simulações do tipo I. Percebeu-se também que a cobertura para esse método diminuiu à medida que o tamanho da amostra cresce, devido à menor dispersão das estimativas para a simulação do tipo I e para simulação do tipo II, em especial para a variável contínua. Constata-se também que, ao se trabalhar com uma amostra pequena com o tipo de cenários apresentados na simulação do tipo I, os métodos apresentam maior viés e dependendo do tamanho, apenas o método direto conseguiria obter as estimativas do RR.

Estudos como o realizado têm sido mais frequentes nos últimos anos. Barros e Hirakata (2003) sugerem modelos diferentes da regressão logística para se analisar dados com desfechos binários em estudos transversais, como o uso da regressão de Cox, Poisson e log binomial com alguns ajustes. Deddens e Petersen (2008) recomendam não se estimar RR via regressão logística para estudos de desfecho comum, sugerem que se utilize regressão log binomial e, em situações em que não há convergência, pode-se ainda utilizar o método COPY, o de Poisson robusta ou mínimos quadrados não lineares. Nijem et. al. (2005) comparam os métodos de estimação do RR via regressão logística, log binomial, de Cox e de Cox com variância robusta, ressaltando preferência pelo modelo de Cox com variância robusta. Esses são alguns exemplos de trabalhos que apresentam métodos alternativos ao logístico para a estimação do RR.

As comparações de métodos de estimação do RR em estudos epidemiológicos encontradas na literatura, como alguns desses destacados no parágrafo anterior, por vezes chegam à conclusão de que a estimação por meio do modelo log binomial é adequada, no entanto, quando deparam-se com problemas de convergência optam por algum outro método. Williamson et. al. (2013) exploram esse tema e recomendam aos pesquisadores que não simplesmente abandonem a sua decisão de usar o modelo log binomial mas que considerem um exame mais cuidadoso de possíveis causas. Diante disso, Andrade e Carabin (2011) implementaram um método para diminuir o problema de convergência desse modelo.

Seria interessante analisar se o VRPM aumentará em módulo se o tamanho das amostras fosse superior à 500, o que acarretaria em menor cobertura e subestimação do RR. A subestimação do RR pode ocasionar na não identificação de algumas variáveis como fatores de risco para um determinado desfecho de interesse.

Portanto, a estimação do RR via método direto, utilizando-se o método implementado por Andrade e Carabin (2011), se mostrou mais eficiente, quando os dados são gerados de uma log binomial. No entanto, mais estudos são necessários.

Referências Bibliográficas

AGRESTI, A. **Introduction to Categorical Data Analysis**. John Wiley, 1996.

ANANIAS, M. C. **Implementação computacional do modelo log binomial**. Monografia apresentada ao Departamento de Estatística da Universidade de Brasília como requisitos para obtenção do grau Bacharel em Estatística - Departamento de Estatística, Universidade de Brasília, Brasília, Dezembro de 2015.

ANDRADE, B.B.; CARABIN, H. On the estimation of relative risk via log binomial regression. **Revista Brasileira de Biometria**, São Paulo, v.29,n.1, p.25-46, 2011.

BARROS, A.J.D., HIRAKATA, V.N. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. **BMC Medical Research Methodology**, 3:21-33, 2003.

BLIZZARD, L.; HOSMER, D. W. Parameter estimation and goodness-of-¯t in log binomial regression. **Biom. J., Weinheim**, v.48, n.1, p.5-22, 2006.

DEDDENS,J. A.; PETERSEN, M. R. Approaches for estimating prevalence ratios. **Occup Environ Med**,65:501-506, 2008

HOSMER, D.W. **Applied Logistic Regression**. John Wiley, 1989.

LOCALIO, A.R., MARGOLIS, D.J., BERLIN, J.A. Relative risks and confidence intervals were easily computed indirectly from multivariate logistic regression. **Journal of Clinical Epidemiology**, 60:874-882, 2007.

MOTTA, V.T. **Bioestatísitca**. 2 ed. Caxias do Sul: EDUCS, 2006.

NEWCOMBE, R.G. A deficiency of the odds ratio as a measure of effect size. **Statistics in Medicine**, 25:4235-4240, 2006.

NIJEM,K; KRISTENSEN, P.; KHATIB, P.; BJERTNESS, E.Application of different statistical methods to estimate relative risk for self-reported health complaints among shoe factory workers exposed to organic solvents and plastic compounds. **Norsk Epidemiologi**, 15 (1): 111-116, 2005.

R Development Core Team (2015). R: A language and environment for statistical computing.**R Foundation for Statistical Computing**, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

ROTHMAN, K.J. **Epidemiology : An Introduction**. 2 ed. New York: Oxford University Press, 2012.

ROTHMAN, K.J.; GREENLAND,S.; LASH, T.L. **Modern Epidemiology**. 3 ed. Philadelphia, PA: Lippincott, Williams e Wilkins, 2008.

SANTOS,C.A.ST et al. Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. **BMC Medical Research Methodology**, p.1-10, 2008.

WILLIAMSON, T.; ELIASZIW, M.; FICK, G. H. Log-binomial models: exploring failed convergence.**Emerging Themes in Epidemiology**, 10:14, 2013.

Apêndice A

Códigos das funções para realizar as simulações

Neste apêndice apresenta-se os códigos desenvolvidos em linguagem R de computação estatística (R Core Team, 2015) das funções para realizar as simulações do tipo I e do tipo II. Os parâmetros das funções permitem que se entre com os valores dos cenários estabelecidos por Blizzard e Hosmer (2006) conforme as tabelas 2.1 e 2.3 do Capítulo 2.

A utilização das funções requer a instalação dos pacotes prLogistic (Santos et. al., 2008) e LBReg (Andrade e Carabin, 2006), este último encontra-se em fase de implementação por Ananias (2015).

```
#####################################################################
####################         Simulacao 1         ####################
#####################################################################
simulacao<- function(a, B0, B1, n, nsim, metodo ){
source('LBReg_new.R')
require(prLogistic)
mensagem<-"Método incorreto!"

modelo.logbinomial<- function (a, B0, B1, n, nsim){
# onde serao guardados os resultados lbreg
Result <- matrix(NA, nrow=nsim, ncol=3)
colnames(Result) <- c('RR', 'VRP','EQM')
rr<-exp(B1)
qntco<-c()
marginal<-c()
```

```
for(t in 1:nsim){
x <- runif(n, -6, a)
y <- rbinom(n, size=1, prob=exp(B0 + B1*x))
xy <- data.frame(x=x,y=y)

reglb <- lbreg(y ~ x, data=xy, start=c(B0,B1))
resultado<-relrisk(reglb)
Result[t,1] <- resultado[2] # estimativa do risco relativo
Result[t,2] <- ((resultado[2] - rr)/rr)*100  #VRP(RR)=(RR*-RR/RR)*100
Result[t,3] <- (((log(resultado[2])-B1)^2)+((log(resultado[6])-
log(resultado[2]))/1.959964)^2)*100 #EQM(logRR)=vies^2+var(logRR)

vrpm <- mean(Result[,2]) #VRPM log binomial
meqm  <- mean(Result[,3]) #MEQM(logRR) log binomial
#ic
if(rr > resultado[4] & rr < resultado[6])
qntco[t]<-1
else qntco[t]<-0

marginal[t]<-sum(y)/n
}
prevalencia<-mean(marginal)
Cobertura<- (sum(qntco)/nsim)*100
resultlbreg<-data.frame(VRPM= vrpm, MEQM=meqm, Cobertura = Cobertura,
Prevalencia=prevalencia)
return(list(Estimacao=Result,Resultado=resultlbreg))
}

modelo.logistico<- function (a, B0, B1, n, nsim){
 # onde serao guardados os resultados
Result <- matrix(NA, nrow=nsim, ncol=3)
colnames(Result) <- c('RR','VRP','EQM')
rr<-exp(B1)
qntco<-c()
marginal<-c()

for(t in 1:nsim){
x <- runif(n, -6, a)
```

```
y <- rbinom(n, size=1, prob=exp(B0 + B1*x))
xy <- data.frame(x=x,y=y)

reglog <- prLogisticDelta(y~ x, dataset=xy, pattern='marginal')$ci
Result[t,1]<-  reglog[1] #estimativa risco relativo
Result[t,2]<- ((reglog[1]-rr)/rr)*100   #VRP
Result[t,3]<- (((log(reglog[1])-B1)^2)+((log(reglog[3])-
log(reglog[1]))/1.959964)^2)*100  #EQM(logRR)

vrpm<-mean(Result[,2]) #VRPM
meqm<-mean(Result[,3]) #MEQM(logRR)
if(rr > reglog[2] & rr < reglog[3])
qntco[t]<-1
else qntco[t]<-0
marginal[t]<-sum(y)/n
}
prevalencia<-mean(marginal)
Cobertura<- (sum(qntco)/nsim)*100
resultlogistic<-data.frame(VRPM=vrpm, MEQM=meqm,Cobertura=Cobertura,
 Prevalencia=prevalencia)
return(list(Estimacao=Result,Resultado=resultlogistic))
}

if (metodo=="logbinomial")
return(modelo.logbinomial(a, B0, B1, n, nsim))
else
if (metodo=="logistico")
return (modelo.logistico(a, B0, B1, n, nsim))
else return(mensagem)
}
```

```
#######################################################################
####################         Simulacao 2         ######################
#######################################################################
simulacao2<-function(prob, B0, Bd, Bu, n, nsim, metodo){
source('LBReg_new.R')
require(prLogistic)
mensagem<-"Método incorreto!"

modelo.logbinomial<-function(prob, B0, Bd, Bu, n, nsim){
 # onde serao guardados os resultados
Result <- matrix(NA, nrow=nsim, ncol=6)
colnames(Result) <- c('RR_D','VRP_D','EQM_D','RR_U','VRP_U','EQM_U')
rrd<-exp(Bd)
rru<-exp(Bu)
qntco<-c()
qntcou<-c()
marginal<-c()

for(t in 1:nsim){
d<- rbinom(n, size=1,prob)
u <- c()

for(b in 1:n){
u[b] <- runif(1, -6 + 2*d, 2 + 2*d)
}
y <- rbinom(n, size=1,prob= exp(B0 + Bd*d + Bu*u))
duy <- data.frame(d=d,u=u ,y=y)

#REGRESSÃO LOG BINOMIAL
reglb <- lbreg(y ~ d + u, data=duy, start=c(B0,Bd,Bu))
resultado<-relrisk(reglb)

#bd
Result[t,1] <- resultado[2] # estimativa do risco relativo
Result[t,2] <- ((resultado[2]-rrd)/rrd)*100 #VRP(RR)=RR*-RR/RR
Result[t,3] <- (((log(resultado[2])-Bd)^2)+((log(resultado[8])-
log(resultado[2]))/1.959964)^2)*100  #EQM(logRR)
#Bu
Result[t,4] <- resultado[3] # estimativa do risco relativo
```

```
Result[t,5] <- ((resultado[3]-rru)/rru)*100 #VRP(RR)=RR*-RR/RR
Result[t,6] <- (((log(resultado[3])-Bu)^2)+((log(resultado[9])-
log(resultado[3]))/1.959964)^2)*100  #EQM(logRR)

#Bd
vrpm <- mean(Result[,2])
meqm  <- mean(Result[,3])
if(rrd > resultado[5] & rrd < resultado[8])
qntco[t]<-1
else qntco[t]<-0
#Bu
vrpm2 <-mean(Result[,5])
meqm2 <-mean(Result[,6])
if(rru > resultado[6] & rru < resultado[9])
qntcou[t]<-1
else qntcou[t]<-0

marginal[t]<-sum(y)/n
}

prevalencia<-mean(marginal)
Cobertura<- (sum(qntco)/nsim)*100
Coberturau<- (sum(qntcou)/nsim)*100
resultlbreg<-data.frame(VRPM.D = vrpm, MEQM.D = meqm, Cobertura.D =
Cobertura, VRPM.U= vrpm2, MEQM.U = meqm2, Cobertura.U = Coberturau,
Prevalencia=prevalencia)
return(list(Estimacao=Result,Resultado=resultlbreg))
}

modelo.logistico<-function(prob, B0, Bd, Bu, n, nsim, metodo){
Result <- matrix(NA, nrow=nsim, ncol=6) # onde serao guardados
colnames(Result) <- c('RR.D','VRP.D','EQM.D','RR.U','VRP.U','EQM.U')
rrd<-exp(Bd)
rru<-exp(Bu)
qntco<-c()
qntcou<-c()
marginal<-c()

for(t in 1:nsim){
```

```
d<- rbinom(n, size=1,prob)
u <- c()

for(b in 1:n){
u[b] <- runif(1, -6 + 2*d, 2 + 2*d)
}
y <- rbinom(n, size=1,prob= exp(B0 + Bd*d + Bu*u))
duy <- data.frame(d=d,u=u ,y=y)

##REGRESSÃO LOGISTICA
reglog <- prLogisticDelta(y~ d+u, dataset=duy, pattern='marginal')$ci
#Bd
Result[t,1] <- reglog[1] #estimativa risco relativo
Result[t,2] <- ((reglog[1]-rrd)/rrd)*100   #VRP
Result[t,3] <- (((log(reglog[1])-Bd)^2) +((log(reglog[5])-log(reglog[1]))/1.959964)^2)*1
#Bu
Result[t,4] <- reglog[2] #estimativa risco relativo
Result[t,5] <- ((reglog[2]- rru)/rru)*100   #VRP
Result[t,6] <- (((log(reglog[2])-Bu)^2)+((log(reglog[6])- log(reglog[2]))/1.959964)^2)*1

#Bd
vrpm <-mean(Result[,2])
meqm <-mean(Result[,3])
if(rrd > reglog[3] & rrd < reglog[5])
qntco[t]<-1
else qntco[t]<-0
#Bu
vrpm2 <- mean(Result[,5]) #vies
meqm2  <- mean(Result[,6])
if(rru > reglog[4] & rru < reglog[6])
qntcou[t]<-1
else qntcou[t]<-0

marginal[t]<-sum(y)/n
}
prevalencia<-mean(marginal)
Cobertura <- (sum(qntco)/nsim)*100
Coberturau <- (sum(qntcou)/nsim)*100
resultlogistic<-data.frame(VRPM.D = vrpm, MEQM.D = meqm, Cobertura.D
```

```
= Cobertura, VRPM.U= vrpm2, MEQM.U = meqm2, Cobertura.U = Coberturau,
Prevalencia=prevalencia)
return(list(Estimacao=Result,Resultado=resultlogistic))
}

if (metodo=="logbinomial")
return(modelo.logbinomial(prob, B0, Bd, Bu, n, nsim))
else
if (metodo=="logistico")
return (modelo.logistico(prob, B0, Bd, Bu, n, nsim))
else return(mensagem)
}
```