UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Um alerta sobre o uso de amostras pequenas na regressão logística

Autor: Rodrigo Coster
Orientador: Professora Msc. Vanessa Bielefeldt Leotti Torman
Co-orientador: Professora Dra. Suzi Alves Camey

Porto Alegre, 04 de Dezembro de 2009.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Um alerta sobre o uso de amostras pequenas na regressão logística

Autor: Rodrigo Coster

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Msc.Vanessa Bielefeldt Leotti Torman
Professor Dr. Álvaro Vigo

Porto Alegre, 04 de Dezembro de 2009.

Dedico este trabalho à Maria Eduarda, por ter me ajudado no aprendizado nestes últimos seis anos.

Mom, you said that the biggest problem in the universe is no one helps each other.
Anakin Skywalker, Star Wars: Episode I - The Phantom Menace

Agradecimentos

Agradeço as professoras Vanessa Bielefeldt Leotti Torman e Suzi Alves Camey por terem aceitado me orientar neste trabalho. Agradeço também aos meus amigos por terem me acompanhado e ajudado nestes quatro anos de curso; aos meus pais, Gilberto e Renate Coster, e meu irmão, Alexandre Coster, por sempre terem me apoiado nestes 22 anos e a minha namorada, Greice Helen da Costa Laureano.

Resumo

A regressão logística está cada dia mais presente nas pesquisas, porém, sabe-se que seus estimadores só possuem boas propriedades se o tamanho de amostra for grande. Entretanto, nem sempre o tamanho amostral utilizado nos estudos é o ideal. Uma regra de bolso para o tamanho amostral amplamente conhecida é de que se deve ter pelo menos dez eventos (sucessos ou fracassos, dependo do que for mais raro) para cada variável independente do modelo. Entretanto, o estudo de simulação a partir do qual esta regra foi elaborada, bem como todos os estudos de simulação encontrados em levantamento bibliográfico realizado, verificou o desempenho da regressão logística apenas para estimar os coeficientes do modelo e não as razões de chances. Através de um estudo simulado de três cenários, mostramos o quão perigoso é usar amostras pequenas para estimar a razão de chance, além de alternativas para o cálculo do tamanho de amostra mínimo para cada caso. Concluímos que as regras utilizadas levando em conta a estimação dos coeficientes não garantem boas propriedades na estimação das razões de chances. Em nossas simulações, encontramos vícios maiores na estimação da razão de chance do que do respectivo coeficiente do modelo. Também entre as conclusões, chamamos atenção para os casos de separação e destacamos que variáveis contínuas são preferíveis à variáveis categóricas.

Palavras chave: Regressão logística, vício, razão de chances, tamanho de amostra.

Sumário

Este artigo será submetido ao “Cadernos de Saúde Pública”.

1. Introdução

Por permitir a modelagem de variáveis dicotômicas (usualmente, morreu ou sobreviveu, doente ou saudável), a regressão logística é muito freqüentemente utilizada nos artigos da área médica. Além disso, ela permite estimar razão de chances (RC), medida de associação de especial interesse principalmente em estudos do tipo caso-controle. O método mais utilizado para estimar os parâmetros do modelo é o da máxima verossimilhança, que em certos casos pode resultar em resultados ruins[1]. Casos onde a variável resposta é rara ou a amostra é pequena estão entre as condições nas quais a máxima verossimilhança não é adequada.

Um dos principais problemas conhecidos dos estimadores de máxima verossimilhança para o modelo de regressão logística é que os mesmos são apenas assintoticamente não viciados, ou seja, são não-viciados somente em amostras grandes[2]. Porém, não é raro encontrar estudos que utilizam amostras pequenas, muitas vezes por desconhecimento da teoria ou até por motivos de operacionalização (falta de condições para conseguir cobaias; de pacientes dispostos a experimentar novos tratamentos, entre outras).

Para uma avaliação preliminar da quantidade de artigos da área médica com amostras pequenas, fez-se uma busca por artigos sobre câncer, de língua inglesa e que utilizassem a regressão logística no repositório PubMed, onde foram encontrados 56 artigos, dos quais foi visto o tamanho da amostra de cada modelo ajustado (alguns artigos ajustaram modelos para mais de um desfecho). Através desta busca, obteve-se a tabela de freqüência abaixo, usando como ponto de corte tamanho de amostra 100 e 500. Estes pontos de corte foram escolhidos porque Long, através de um estudo de simulação do viés, concluiu que usar amostras menores que 100 é arriscado, enquanto o aconselhado é acima de 500[2]. É possível perceber que quase a metade dos artigos utilizou amostras abaixo do ponto de corte considerado seguro por Long.

Tabela 1: Distribuição de frequência dos tamanhos amostrais obtidos na busca no Pubmed

	Menor ou igual que 100	Entre 100 e 500	Maior que 500
Freqüência	7	20	33
Percentual	11,67%	33,33%	55%

Uma regra amplamente utilizada para o cálculo de tamanho amostral[3] sugere que se tenha pelo menos 10 casos do evento mais raro do desfecho para cada variável

independente adicionada no modelo final. Vittinghoff e McCulloch inclusive afirmam que o valor 10 pode ser reduzido em certos casos em que o objetivo da pesquisa não é fazer predições, e sim apenas testar a influência de covariáveis[4]. Entretanto, os estudos de simulação que foram procedidos para obter essas conclusões consideraram apenas a estimação dos coeficientes do modelo e não da RC.

Nosso objetivo neste trabalho é mostrar, através de simulações, os problemas encontrados na regressão logística em amostras pequenas na estimação da RC, além de indicar métodos e softwares para o cálculo do tamanho de amostra ideal.

Na próxima sessão detalharemos como foi feito o estudo, quais cenários foram simulados e as informações utilizadas de cada simulação. Em seguida, uma descrição dos resultados e apontamentos dos aspectos mais importantes, e por fim, as conclusões as quais chegamos.

2. Metodologia

A simulação foi baseada em um estudo transversal fictício, onde o desfecho é ter desenvolvido câncer e tem-se dois fatores de exposição. Um dos fatores representa a idade do paciente e possui distribuição normal com média 48 e desvio padrão 10 e sua verdadeira razão de chances é igual a 1,05 (ou seja, seu coeficiente de regressão é log(1,05)). O outro fator é uma variável categórica de 4 categorias, representando nível de fumo, com suas interpretações, proporções na população e razões de chances dadas na Tabela 2. Esta variável categórica entra no modelo de regressão logística através de três variáveis indicadoras auxiliares, de forma que a categoria "Nunca fumou" é a referência.

Tabela 2: Proporções e razão de chances de cada categoria da variável categórica

Quantidade fumada	Proporções	Razão de chances
Mais de 1 maço por dia	10%	5,0
Até 1 maço por dia	25%	3,0
Ex-fumante	50%	1,5
Nunca fumou	15%	1 (Referência)

Assim o modelo de regressão logística considerado foi:

$$\log\left(\frac{\pi(x_1,x_2,x_3,x_4)}{1-\pi(x_1,x_2,x_3,x_4)}\right) = \exp(\beta_0 + 1{,}609x_1 + 1{,}098x_2 + 0{,}405x_3 + 0{,}048x_4)$$

onde $\pi(x_1,x_2,x_3,x_4)$ é a probabilidade de um indivíduo ter câncer dados os seus particulares valores dos fatores.

A fim de estudar diferentes cenários possivelmente encontrados na prática, foram escolhidos três diferentes valores de β_0. O primeiro valor (-5) resulta numa prevalência de

câncer de aproximadamente 13% (evento raro), o segundo valor (-3.9), resulta uma incidência de aproximadamente 30% (evento não raro), e o terceiro valor (-3) resulta em uma incidência de aproximadamente 50% (estudo pareado). Estes valores de incidência foram obtidos através de simulação de uma amostra de tamanho 100000. Para cada valor de β_0 foram considerados 5 tamanhos amostrais diferentes (50, 100, 250, 500 e 1000), totalizando 15 configurações. Para cada configuração, simulou-se 10000 amostras. As simulações foram rodadas no programa R 2.9.2[5], sem a necessidade de carregar pacotes.

A cada amostra simulada, gerava-se apenas valores para o desfecho, e os valores dos fatores de exposição eram mantidos constantes. Uma vez fixados os valores preditores, foi calculada a probabilidade de sucesso (ter câncer), através de:

$$\pi(x_1, x_2, x_3, x_4) = \frac{\exp(\beta_0 + 1{,}609x_1 + 1{,}098x_2 + 0{,}405x_3 + 0{,}048x_4)}{1 + \exp(\beta_0 + 1{,}609x_1 + 1{,}098x_2 + 0{,}405x_3 + 0{,}048x_4)} \quad [1]$$

onde x_1, x_2 e x_3 são as variáveis indicadores da quantidade fumada, representando "Mais de um maço por dia", "Até 1 maço por dia" e "Ex-fumante", respectivamente e variável x_4 é a idade. Os coeficientes das variáveis indicadores, assim como o coeficiente de x_4, são o logaritmo natural das suas razões de chances.

O desfecho então era simulado a partir da distribuição Bernoulli, através das probabilidades de sucesso calculadas. De cada simulação era registrado o valor estimado, vício (diferença entre valor estimado e valor verdadeiro), vício relativo (vício dividido pelo valor verdadeiro), e intervalo de 95% de confiança das razões de chances. Para cada configuração, foi calculado o erro quadrático médio das estimativas, a cobertura e amplitude média dos intervalos e o poder do teste de hipótese bilateral baseado no intervalo, sendo o poder considerado a proporção de intervalos que não continham o valor 1. O erro quadrático médio é a soma entre a variância das estimativas e o quadrado da média dos vícios. Ele é uma medida global de acurácia e precisão de um estimador. A cobertura foi definida como a quantidade de intervalos que continham o verdadeiro valor do parâmetro e a amplitude média é a média aritmética da amplitude dos intervalos de confiança. Alguns gráficos foram feitos para visualizar os resultados encontrados.

Os resultados obtidos foram comparados com cálculos de tamanho amostral e poder a posteriori. Para tais comparações, utilizou-se o programa G*Power 3.1.0[6], que usa os métodos de Hsieh et al. e Demidenko[7,8] para cálculo de tamanho de amostra. Optou-se pelo método de Demidenko por recomendação do manual. O programa só nos permite especificar uma variável do modelo, as outras variáveis são especificadas resumidamente em um parâmetro, como será explicado a seguir. Como variável a ser especificada, foi escolhida a variável dicotômica que representa a categoria 'Mais de um maço por dia' por

possuir menor incidência, maximizando assim o tamanho amostral. Além da incidência da variável, também foi utilizada a sua razão de chances bruta (quando o modelo é ajustado usando apenas essa variável), que para cada cenários foi de aproximadamente 2,71; 2,69 e 2,82. Obtivemos esta RC através do ajuste de uma regressão logística para uma amostra simulada de tamanho 100000. No caso de o interesse ser o ajuste de uma regressão logística múltipla, o programa também permite entrar com o coeficiente de determinação (R^2) do modelo de regressão ajustado da variável principal sendo explicada pelas demais variáveis do modelo. Seja n o tamanho de amostra necessário para estimar a RC na regressão logística simples e n' o tamanho de amostra análogo na regressão logística múltipla. Hsieh[8] concluiu que n' é aproximadamente igual a $n/(1-R^2)$. Ou seja, quando maior o R^2, maior o n'. No caso, foi utilizado $R^2 = 0$, para obter o tamanho de amostra mais otimista possível.

Os três tamanhos amostrais foram estimados para uma probabilidade de erro tipo I de 0,05 e uma probabilidade de erro tipo II de 0,20. O cálculo do poder a posteriori foi feito informando-se os tamanhos amostrais utilizados na simulação. Outros métodos e softwares de cálculo de amostra estão disponíveis[7-9], porém, como a intenção deste artigo não é compará-los, eles não foram utilizados.

3. Resultados

Nas Tabelas 4 a 6 estão os resultados obtidos para as medidas de desempenho da regressão logística ao estimar as razões de chances. Pelas três tabelas podemos notar melhora na qualidade das estimativas dos parâmetros conforme cresce o tamanho da amostra, pois tanto o vício quanto o EQM possuem decréscimos consideraveis. Valores superiores a 1e+307 de vício, vício relativo, EQM e amplitude média foram considerados como infinito e são explicados devido a separação nos dados, que ocorre quando apenas sucessos (ou fracassos) foram observados em alguma categoria da variável categórica ou quando podemos determinar um valor da variável contínua que, abaixo dele só aconteça sucessos e acima dele somente fracassos (ou vice-versa)[1]. Como esperado, esse problema deixa de acontecer conforme cresce o tamanho da amostra e a incidência do desfecho, como mostrado na Tabela 3.

Tabela 3: Quantidade de simulações com separação

Tamanho amostral	50	100	250	500	1000
Cenário 1	6434	4073	956	23	0
Cenário 2	3684	538	2	0	0
Cenário 3	2478	224	2	0	0

Considerando que o modelo possui quatro variáveis independentes, através da regra de Peduzzi et al[3], precisaríamos de um mínimo de 40 sucessos. Considerando as prevalências do desfecho de cada cenário, poderíamos usar erroneamente a regra e concluir que um bom tamanho de amostra em cada cenário é 308, 134 e 80, respectivamente. Como na verdade três de nossas quatro variáveis são provenientes de uma variável categoria, equivocadamente poderíamos pensar que o modelo possui apenas duas variáveis, resultando em tamanhos de amostra iguais a 154, 67 e 40, respectivamente. Primeiramente, percebemos que os tamanhos de amostra obtidos, mesmo pensando corretamente sobre o número de variáveis do modelo, possibilitariam casos de separação nos dados, apesar de não tão freqüentes. Além disso, Peduzzi et al recomendaram o valor mínimo 10 do evento mais raro (sucesso ou fracasso) por variável pois em seu estudo, a partir deste valor o vício relativo dos coeficientes de regressão era menor ou igual a 0,1 (10%). Em nosso estudo, para o primeiro cenário, vê-se que até no tamanho de amostra igual a 1000 obteve-se vício relativo maior que 0,1 para as RCs das variáveis dicotômicas e intercepto. Já no segundo cenário, obteve-se vício relativo menor que 0,1 para todos os tamanhos de amostra para a RC da variável contínua e para tamanho 1000 para as demais RCs. No terceiro cenário o tamanho de amostra 500 já foi suficiente para obter vício relativo menor que 0,1 para algumas RCs.

A Figura 1 mostra o vício relativo do coeficiente de regressão de X_3 e da razão de chance para cada tamanho de amostra no cenário de maior prevalência. Vemos que o decrescimento do vício relativo da RC é mais lento que o do coeficiente de regressão. Para tamanhos amostrais de 50 e 100, o erro relativo está em torno de 1e+6 e 1e+5, respectivamente, omitidos para preservar a visibilidade do gráfico.

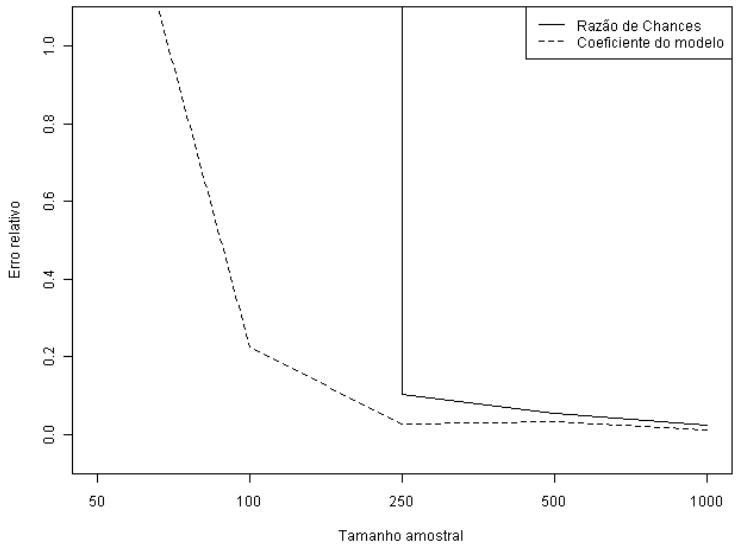


Figura 1: Comparação do erro relativo

Tabela 4: Resultados do cenário 1 – prevalência do desfecho de 13%

RC	Valor verdadeiro	n	Estimado	Vício	Vício relativo	EQM	Cobertura	Poder	Amplitude média
$e^{ß0}$	0,00674	50	0,2775	0,2708	40,1899	116,1453	0,9745	0,1354	Inf
		100	0,0216	0,0149	2,2108	0,0135	0,9708	0,4939	Inf
		250	0,0101	0,0034	0,5026	0,0005	0,9655	0,9045	Inf
		500	0,0086	0,0018	0,2693	0,0001	0,9546	0,9977	Inf
		1000	0,0075	0,0008	0,1140	0,0000	0,9524	1	0,0223
$e^{ß1}$	5	50	Inf	Inf	Inf	Inf	0,9994	0,0112	Inf
		100	2,3E+07	2,3E+07	4,6E+06	1,1E+16	0,9951	0,0548	Inf
		250	3,8E+06	3,8E+06	7,6E+05	3,8E+14	0,9802	0,4259	Inf
		500	1,1E+05	1,1E+05	2,2E+04	1,3E+13	0,9677	0,8657	Inf
		1000	5,7673	0,7673	0,1535	17,5085	0,9591	0,9904	11,0634
$e^{ß2}$	3	50	Inf	Inf	Inf	Inf	0,9918	0,001	Inf
		100	1,1E+07	1,1E+07	3,7E+06	1,1E+15	0,9724	0,0042	Inf
		250	2,1E+06	2,1E+06	7,1E+05	1,0E+14	0,9708	0,1344	Inf
		500	6,1E+04	6,1E+04	2,0E+04	3,4E+12	0,9657	0,6441	Inf
		1000	3,4057	0,4057	0,1352	5,1967	0,9554	0,904	5,9835
$e^{ß3}$	1,5	50	Inf	Inf	Inf	Inf	0,989	0,0035	Inf
		100	5,3E+06	5,3E+06	3,5E+06	2,0E+14	0,9725	0,0092	Inf
		250	1,0E+06	1,0E+06	7,0E+05	2,4E+13	0,9699	0,0058	Inf
		500	3,3E+04	3,3E+04	2,2E+04	1,0E+12	0,9672	0,0836	Inf
		1000	1,7013	0,2013	0,1342	1,2882	0,9567	0,1781	2,9378
$e^{ß4}$	1,05	50	1,1E+36	1,1E+36	1,1E+36	6,1E+75	0,9646	0,1152	Inf
		100	1,0571	0,0071	0,0067	0,0035	0,9528	0,303	0,1505
		250	1,0523	0,0023	0,0022	0,0011	0,9496	0,6738	0,0883
		500	1,0507	0,0007	0,0007	0,0005	0,9525	0,9465	0,0592
		1000	1,0505	0,0005	0,0005	0,0002	0,9482	0,9985	0,0422

Como podemos ver nas Tabelas 4 a 6, as estimativas da razão de chances para a variável contínua apresentam desempenho superior as demais variáveis. Para este parâmetro, apenas no cenário de menor prevalência do desfecho e menor tamanho de amostra se observou valores altos de vício e variabilidade do estimador.

Tabela 5: Resultados do cenário 2 – prevalência do desfecho de 30%

RC	Valor verdadeiro	n	Estimado	Vício	Vício relativo	EQM	Cobertura	Poder	Amplitude média
$e^{ß0}$	0,02024	50	0,4590	0,4388	21,6760	2736,8334	0,9698	0,3994	Inf
		100	0,0423	0,0220	1,0891	0,0223	0,9597	0,7857	Inf
		250	0,0261	0,0058	0,2883	0,0014	0,9483	0,9986	Inf
		500	0,0229	0,0027	0,1320	0,0005	0,9506	1,0000	0,0686
		1000	0,0215	0,0013	0,0642	0,0002	0,9437	1,0000	0,0409
$e^{ß1}$	5	50	2,5E+14	2,5E+14	4,9E+13	6,5E+31	0,9933	0,0668	Inf
		100	9,5E+06	9,5E+06	1,9E+06	6,8E+15	0,9796	0,3605	Inf
		250	3,3E+04	3,3E+04	6,7E+03	1,1E+13	0,9493	0,8154	Inf
		500	5,7171	0,7171	0,1434	15,9667	0,9509	0,9762	11,1899
		1000	5,3412	0,3412	0,0682	6,1020	0,9484	0,9997	6,8184
$e^{ß2}$	3	50	2,0E+07	2,0E+07	6,6E+06	6,0E+15	0,9782	0,0232	Inf
		100	4,3E+06	4,3E+06	1,4E+06	8,4E+14	0,9741	0,1740	Inf
		250	2,3E+04	2,3E+04	7,6E+03	5,4E+12	0,9531	0,6225	Inf
		500	3,3187	0,3187	0,1062	3,7314	0,9525	0,9018	5,3518
		1000	3,1488	0,1488	0,0496	1,5143	0,9496	0,9949	3,3726
$e^{ß3}$	1.5	50	8,5E+06	8,5E+06	5,7E+06	7,9E+14	0,9738	0,0087	Inf
		100	2,0E+06	2,0E+06	1,4E+06	1,8E+14	0,9718	0,0108	Inf
		250	9,9E+03	9,9E+03	6,6E+03	1,0E+12	0,9551	0,1198	Inf
		500	1,6461	0,1461	0,0974	0,8123	0,9538	0,2142	2,4916
		1000	1,5681	0,0681	0,0454	0,3341	0,9541	0,3817	1,5920
$e^{ß4}$	1.05	50	1,0591	0,0091	0,0086	0,0043	0,9540	0,2967	0,1588
		100	1,0545	0,0045	0,0043	0,0017	0,9511	0,5055	0,1105
		250	1,0520	0,0020	0,0019	0,0005	0,9481	0,9342	0,0627
		500	1,0508	0,0008	0,0007	0,0003	0,9471	0,9980	0,0440
		1000	1,0505	0,0005	0,0005	0,0001	0,9482	1,0000	0,0311

Se levarmos em consideração apenas a possibilidade de ocorrer separação nos dados, podemos definir tamanhos amostrais mínimos para cada cenário como sendo, respectivamente, entre 500 e 1000, 100 e 250, 50 e 100. Entretanto, o que se faz na prática é calcular o tamanho amostral mínimo para detectar no mínimo certa RC, e garantir certa probabilidade de erro tipo I e poder para o teste de hipótese de interesse. Assim, considerando 5% de erro tipo I e poder de 80% para o teste do coeficiente de $X_{1,}$ em nossa simulação temos que o tamanho amostral mínimo para os três cenários está nos intervalos 500 e 1000, 250 e 500, 250 e 500, respectivamente. Os tamanhos amostrais para cada cenário obtidos através do programa G*Power foram 526, 377 e 374, condizentes com as

simulações. Vale ressaltar que, caso utilizássemos outra variável como base para os cálculos teríamos obtidos tamanhos amostrais maiores, pois elas apresentam razões de chance mais próximas de um do que a variável utilizada. Apesar do cálculo do tamanho de amostra considerando as probabilidades de erro tipo I e II não levar em conta a possibilidade de amostras com separação, vemos que os tamanhos amostrais obtidos têm pouca probabilidade de conduzir a casos desse tipo. Entretanto, vemos que estes tamanhos de amostras calculados não garantem baixos vícios relativos, talvez porque fizemos a pressuposição otimista de que $R^2=0$.

Tabela 6: Resultados do cenário 3 – prevalência do desfecho de 50%

RC	Valor verdadeiro	n	Estimado	Vício	Vício relativo	EQM	Cobertura	Poder	Amplitude média
$e^{ß0}$	0,04979	50	0,2395	0,1897	3,8100	10,9776	0,9581	0,3315	Inf
		100	0,0898	0,0400	0,8027	0,0499	0,9523	0,7255	Inf
		250	0,0620	0,0122	0,2452	0,0065	0,9476	0,9857	0,2623
		500	0,0558	0,0060	0,1204	0,0022	0,9519	1,0000	0,1459
		1000	0,0525	0,0028	0,0553	0,0009	0,9475	1,0000	0,0855
$e^{ß1}$	5	50	2,6E+14	2,6E+14	5,3E+13	4,4E+31	0,9762	0,0649	Inf
		100	3,3E+06	3,3E+06	6,6E+05	1,9E+15	0,9692	0,4240	Inf
		250	1,9E+04	1,9E+04	3,8E+03	3,9E+12	0,9564	0,8361	Inf
		500	5,6718	0,6718	0,1344	13,6280	0,9513	0,9879	10,4167
		1000	5,3029	0,3029	0,0606	5,3308	0,9490	1,0000	6,3973
$e^{ß2}$	3	50	1,3E+07	1,3E+07	4,2E+06	2,7E+16	0,9732	0,1447	Inf
		100	4,4E+05	4,4E+05	1,5E+05	2,4E+14	0,9558	0,3627	Inf
		250	3,4381	0,4381	0,1460	6,6248	0,9515	0,7176	7,0905
		500	3,2103	0,2103	0,0701	2,3175	0,9502	0,9519	4,2676
		1000	3,0944	0,0944	0,0315	1,0023	0,9513	0,9982	2,8057
$e^{ß3}$	1.5	50	4,2E+06	4,2E+06	2,8E+06	1,4E+15	0,9758	0,0231	Inf
		100	1,8E+05	1,8E+05	1,2E+05	4,0E+13	0,9636	0,0783	Inf
		250	1,6574	0,1574	0,1049	1,2363	0,9518	0,1606	2,9950
		500	1,5841	0,0841	0,0561	0,4442	0,9501	0,3088	1,8627
		1000	1,5386	0,0386	0,0257	0,2034	0,9496	0,5347	1,2480
$e^{ß4}$	1.05	50	1,0579	0,0079	0,0075	0,0043	0,9511	0,2679	0,1627
		100	1,0538	0,0038	0,0036	0,0013	0,9498	0,6093	0,0973
		250	1,0519	0,0019	0,0018	0,0005	0,9496	0,9537	0,0599
		500	1,0507	0,0007	0,0006	0,0002	0,9523	0,9995	0,0410
		1000	1,0504	0,0004	0,0003	0,0001	0,9471	1,0000	0,0279

Uma maneira que poderia ser utilizada de resolver o problema de separação é amostrar um desfecho em específico até que se tenha pelo menos um em cada variável, porém este método deve ser feito com cuidado. Se aumentarmos apenas o desfecho na categoria do fator que desejamos estaremos mudando a verdadeira distribuição dos dados, fazendo com que nossa amostra não seja mais representativa da população. Outros métodos para resolver casos de separação podem ser encontrados em Heinze[10] e Rousseeuw[11].

Na Tabela 7 mostramos o quão perigoso é realizar amostras com restrições, no caso foi imposto um número mínimo de sucessos e fracassos para as variáveis. Para produzirmos a tabela foi utilizado o cenário onde o desfecho era raro (13%) e o tamanho amostral 50. Fizemos a proporção de pessoas que desenvolveram e que não desenvolveram câncer para todos os casos e apenas para os casos em que cada variável tinha pelo menos um sucesso ou fracasso. A diferença entre as probabilidades estimadas torna-se maior nas categorias com menos frequência, como é o caso de "Mais de 1 maço por dia" e "Nunca fumou".

Tabela 7: Fixando o número de sucessos e fracassos

	Todos os casos		Pelo menos um de cada desfecho	
Categoria	Com câncer	Sem câncer	Com câncer	Sem câncer
Mais de 1 maço por dia	0,2648	0,7352	0,3324	0,6676
Até 1 maço por dia	0,1863	0,8137	0,2034	0,7966
Ex-fumante	0,1152	0,8848	0,1213	0,8787
Nunca fumou	0,0868	0,9132	0,1663	0,8337

A Figura 2 mostra o que acontece com o vício relativo absoluto (eixo vertical) conforme cresce o número mínimo de sucessos ou fracassos por variável (eixo horizontal). Quando temos o valor 0 no eixo horizontal, temos a análise de todas as simulações, quando temos o valor 1, analisamos apenas as simulações em que todas as variáveis categóricas possuem pelo menos um de cada desfecho, quando temos o valor 2, consideramos apenas as simulações em que todas as variáveis categóricas possuem pelo menos dois de cada desfecho, e assim por diante. Para sua construção, foi utilizado o mesmo cenário que para a Tabela 7 (cenário 1). Podemos ver que há uma queda até um determinado valor, e após esse valor o vício relativo começa a crescer. Embora estranho, esse comportamento pode ser facilmente explicado. Cada cenário possui um número esperado de desfechos para suas variáveis categóricas, e conforme o número mínimo desejado de desfechos se aproxima do

menor destes valores, vemos uma melhora nas estimativas, e conforme ele vai se distanciando, há um aumento no vício.

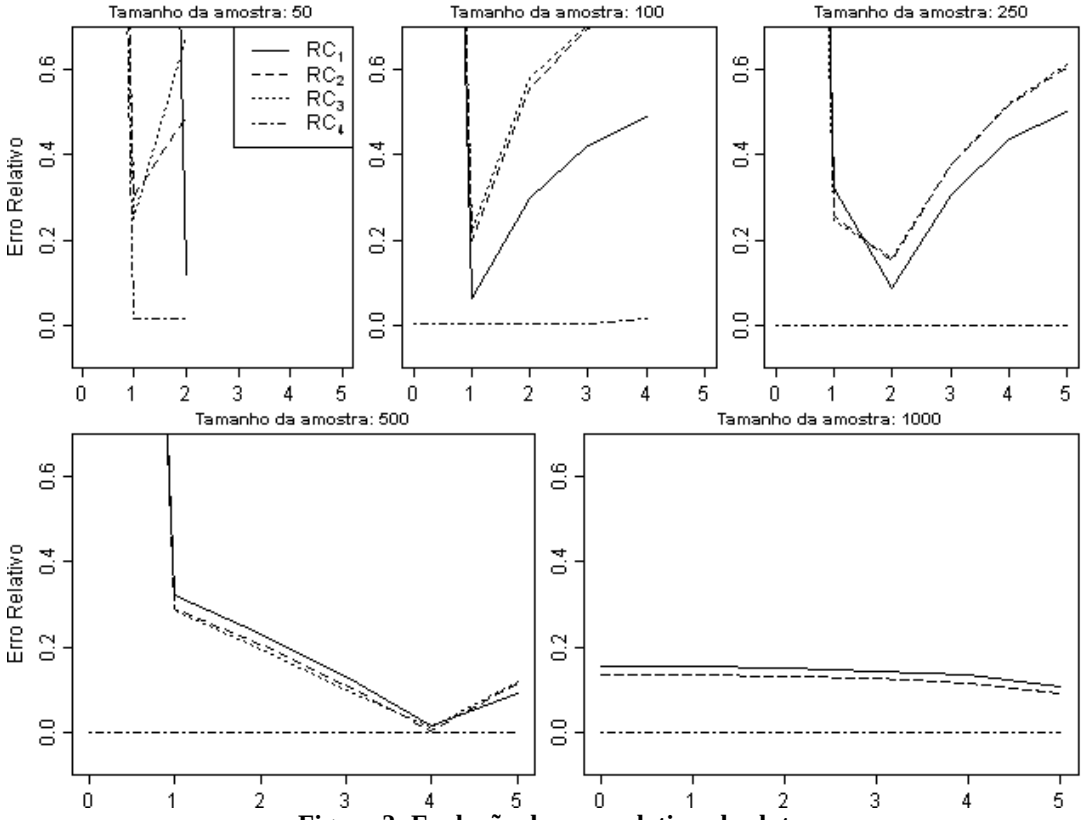


Figura 2: Evolução do erro relativo absoluto

4. Conclusões

Durante o levantamento bibliográfico pudemos notar que todos os estudos realizados para sugerir regras de bolso sobre o tamanho de amostra utilizavam as medidas de acurácia dos coeficientes do modelo de regressão e não das razões de chances, como feito neste trabalho. A importância de se ressaltar isso é de que o vício da razão de chances é muito superior ao do seu coeficiente no modelo. Por exemplo, um coeficiente com erro percentual de 10% e verdadeiro valor igual a 2 produz uma razão de chances com erro percentual 22,14%. Ou seja, se o objetivo é interpretar as razões de chances, essas regras de bolso não devem ser utilizadas.

No momento recomenda-se o cálculo formal do tamanho de amostra necessário. Entretanto, uma dificuldade para este cálculo é que os artigos geralmente não publicam informações suficientes para o mesmo, muitas vezes não deixando claro nem quantas variáveis foram consideradas no modelo final. Não temos conhecimento de artigo que publicou a informação sobre o R^2 do fator principal contra os demais fatores. Estas informações também são importantes caso o leitor queira identificar se há possibilidade de

vício nos resultados. Informar apenas o tamanho amostral (o que não é feito por todos) não é o suficiente, pois vimos neste trabalho que um mesmo tamanho amostral para dois modelos iguais em número de variáveis e diferentes somente na prevalência do desfecho (únicas informações utilizada pelas regras de bolso atuais) possui medidas de acurácia completamente diferentes.

Maneiras de reduzir o vício dos coeficientes do modelo foram propostas por Neme[2], com destaque para o *bootstrap* e alternativas para o estimador de máxima verossimilhança. Porém, o próprio autor avisa que estes métodos podem aumentar a variância e EQM das estimativas, além da complexidade das análises. Métodos similares podem vir a melhorar as estimativas da razão de chances, porém não foram testadas neste trabalho.

Notamos uma superioridade da variável contínua em relação às variáveis categóricas, apresentando em quase todas as combinações erro relativo inferior a 1%. Alertamos sobre isso pois é comum, principalmente na área médica, dicotomizar variáveis contínuas (como por exemplo, acima e abaixo de um determinado valor) a fim de facilitar o entendimento do modelo. Entretanto, também não sabemos se este bom desempenho se mantém para estimar razão de chances entre mais de uma unidade de diferença na variável contínua, algo que posteriormente estudaremos.

Vale também alertar o comportamento dos programas frente a problemas de separação. Apesar de não haver eventos para todos os valores das variáveis categóricas, os programas conseguem estimar um modelo, porém com vícios muito altos. Além de nem todos os programas avisarem deste problema nos dados, alguns programas que possuem mais de uma maneira de fazer regressão logística dão resultados diferentes quando os dados possuem separação.

Apesar destas conclusões terem sido obtidas através da simulação de um estudo transversal (onde se tem a população e são retiradas amostras dela, sem restrições), as conclusões também são válidas para estudos com delineamento caso-controle (onde o número de desfechos é predeterminado), devido a esse delineamento interferir apenas no termo constante do modelo (β_0), que representa a prevalência do desfecho na população.

Continuações deste estudo poderiam ser feitas adicionando mais cenários e variáveis nas simulações, além de verificar se métodos corretivos para os coeficientes funcionam também para a razão de chances.

5. Bibliografia

1. King EN, Ryan TP. A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression. *The American Statistician*. 2002;56(3):163-170.

2. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology*. 2009;9(1):56.

3. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis*1. *Journal of Clinical Epidemiology*. 1996;49(12):1373-1379.

4. Vittinghoff E, McCulloch CE. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *Am. J. Epidemiol.* 2006:kwk052.

5. The R Project for Statistical Computing. Available at: http://www.r-project.org/ [Acessado Novembro 29, 2009].

6. Heinrich-Heine-Universität - Institut für experimentelle Psychologie. Available at: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/ [Acessado Novembro 29, 2009].

7. Demidenko E. Sample size determination for logistic regression revisited. *Statistics in Medicine*. 2007;26(18):3385-3397.

8. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*. 1998;17(14):1623-1634.

9. Self SG, Mauritsen RH. Power/Sample Size Calculations for Generalized Linear Models. *Biometrics*. 1988;44(1):79-86.

10. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine*. 2002;21(16):2409-2419.

11. Rousseeuw PJ, Christmann A. Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*. 2003;43(3):315-332.

6. Anexos

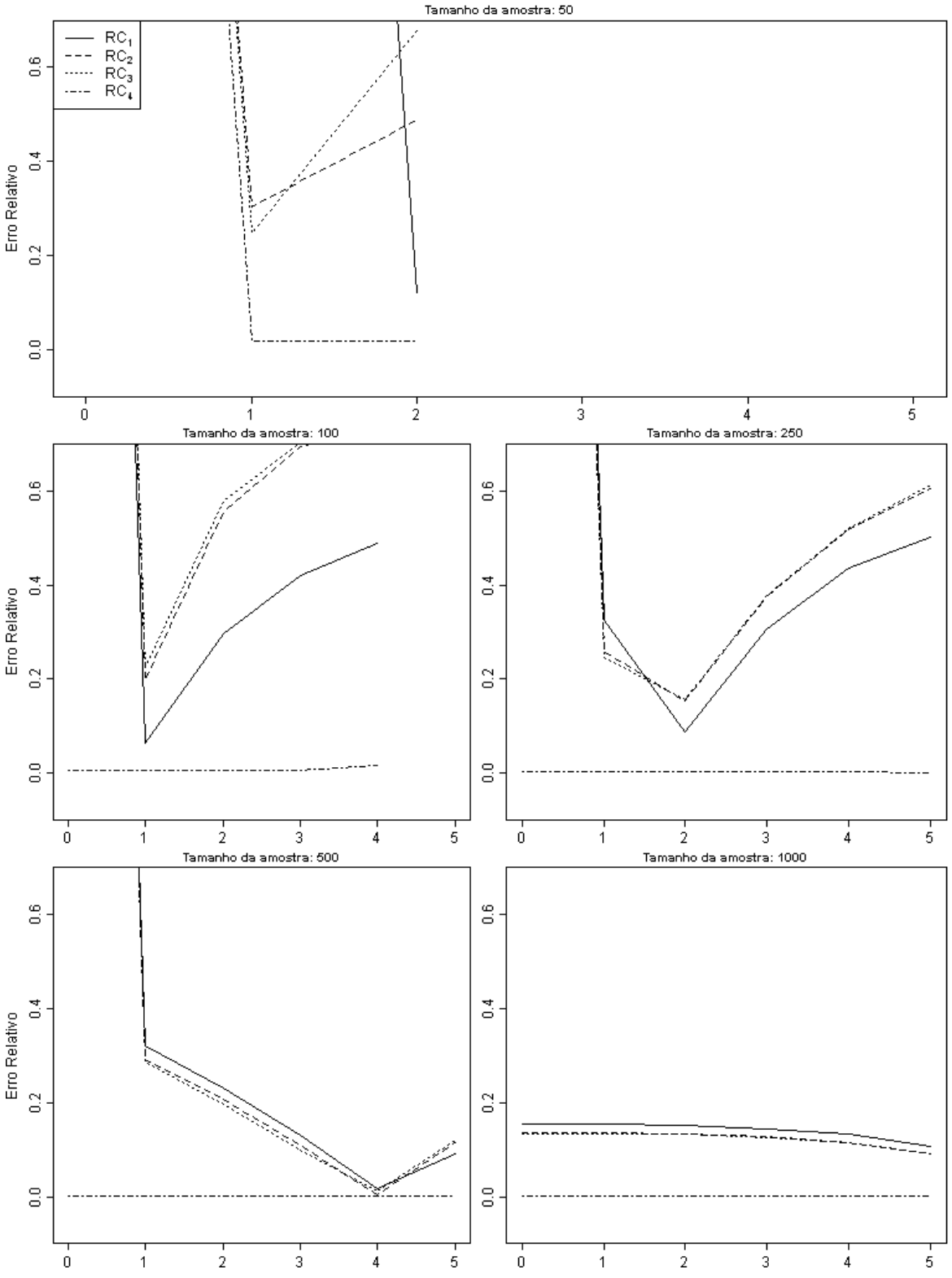
Anexo I – Gráfico do erro relativo absoluto (eixo vertical) contra o número mínimo de eventos (eixo horizontal) do cenário 1.

Anexo II – Gráfico do erro relativo absoluto (eixo vertical) contra o número mínimo de eventos (eixo horizontal) do cenário 2.

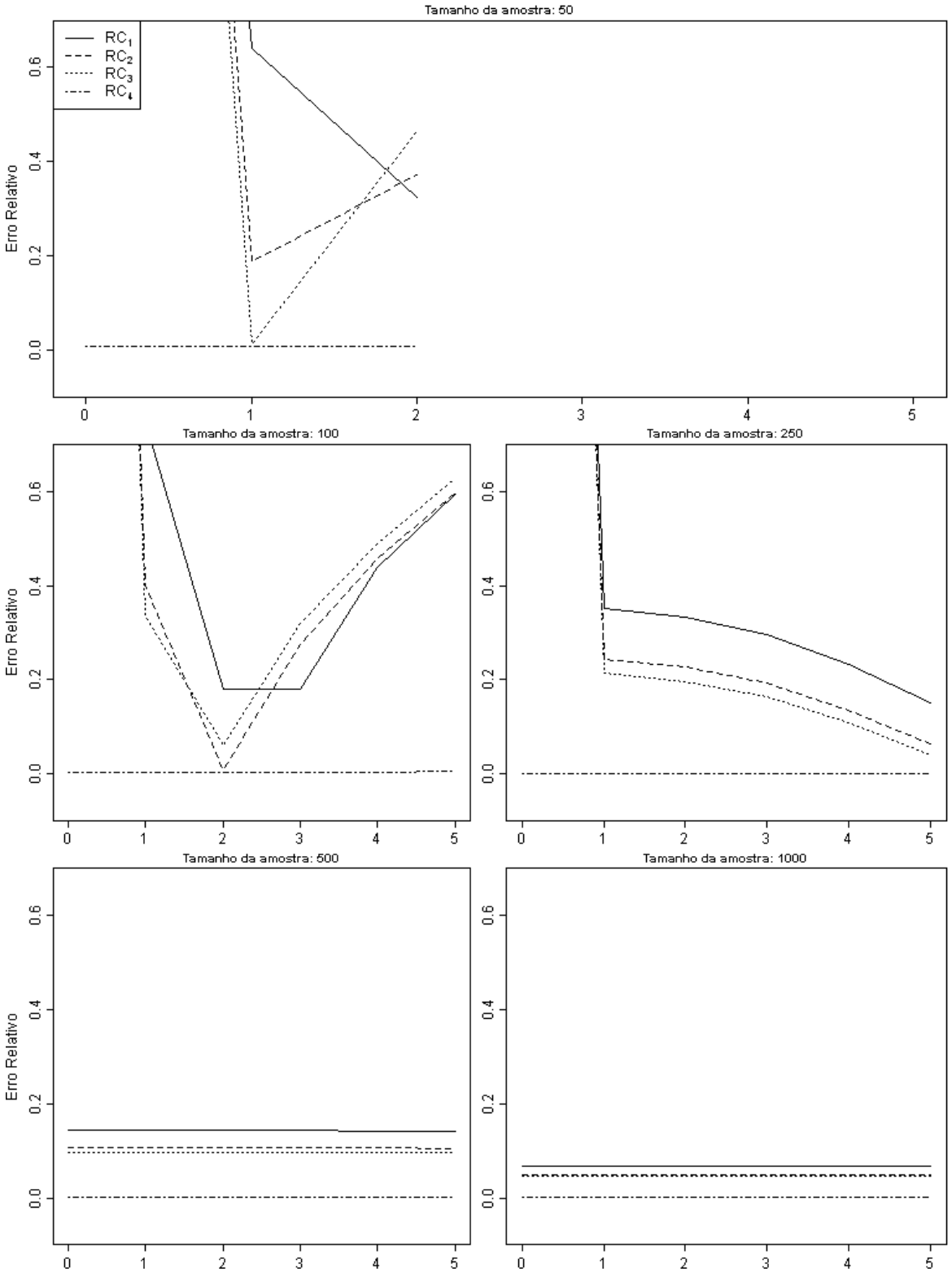
Anexo III – Gráfico do erro relativo absoluto (eixo vertical) contra o número mínimo de eventos (eixo horizontal) do cenário 3.

Anexo IV – Códigos da simulação

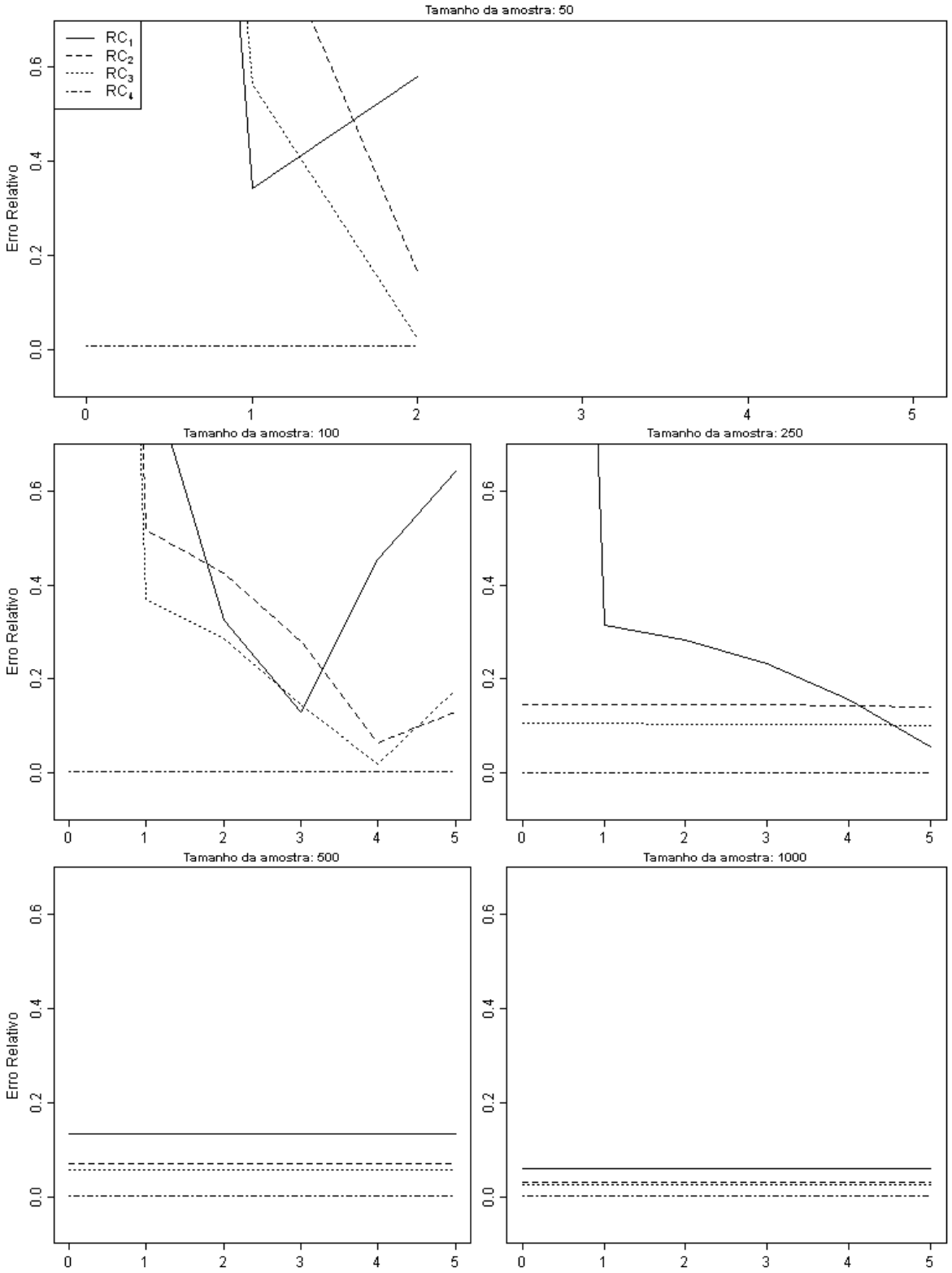
Anexo I


Tamanho da amostra: 50
Tamanho da amostra: 100
Tamanho da amostra: 250
Tamanho da amostra: 500
Tamanho da amostra: 1000
RC1
RC2
RC3
RC4
Erro Relativo
0.0
0.2
0.4
0.6
0
1
2
3
4
5

Anexo II


Tamanho da amostra: 50
Tamanho da amostra: 100
Tamanho da amostra: 250
Tamanho da amostra: 500
Tamanho da amostra: 1000
RC1
RC2
RC3
RC4
Erro Relativo
0.0
0.2
0.4
0.6
0
1
2
3
4
5

Anexo III


Tamanho da amostra: 50
Tamanho da amostra: 100
Tamanho da amostra: 250
Tamanho da amostra: 500
Tamanho da amostra: 1000
Erro Relativo
RC1
RC2
RC3
RC4

Anexo IV

```
##
# Funções principais
##

## Função para criar o vetor de probabilidade de sucesso de Y e a matrix de X
probs <- function(n=100,betas=c(-5,log(c(5,3,1.5,1.05)))) {
 x1 <- rep(c(1,0),times=c(n*.1,n*.9))  # x1: Incidencia: 10%
 x2 <- rep(c(0,1,0),times=c(n*.1,floor(n*.25),ceiling(n*.65)))  # x2: Incidencia: 25%
 x3 <- rep(c(0,1,0),times=c(floor(n*.35),n*.5,ceiling(n*.15)))  # x3: Incidencia: 50%
 x4 <- rnorm(n,48,10) # x4: Normal(48,10)
 t1 <- cbind(1,x1,x2,x3,x4) %*% cbind(betas)  # Multiplicando o X pelos betas
 cbind(exp(t1)/(1+exp(t1)),cbind(1,x1,x2,x3,x4))  # Retornando as probabilidades de sucesso e os X
}

## Função para criar o objeto simul, que possui as simulações.
simullog <- function(n=100,rep=10000,betas=c(-5,log(c(5,3,1.5,1.05))),prob=probs(n=n,betas=betas),...) {
 t <- Sys.time()
 n <- nrow(prob) # Tamanho da Amostra
 y <- replicate(rep,rbinom(n,1,prob[,1])) # Simula as amostras

 # Função para modelar a regressão logistica (exata quando e=TRUE), e retornar as informações necessárias
 simul <- function(y,x,e=F) {
  if (e == F) { # Confere se a regressão exata foi solicitada (no caso, não)
   m <- glm(y ~ x,family=binomial) # Gera o modelo
   o <- as.vector(confint.default(m)) # Gera os intervalos de confiança
   names(o) <-
c('OD_X0_i','OD_X1_i','OD_X2_i','OD_X3_i','OD_X4_i','OD_X0_s','OD_X1_s','OD_X2_s','OD_X3_s','OD_X4_
s')
   return(exp(c(m$coef,o))) # Retorna os coeficientes estimados e seus intervalos de confiança - Para os Odds
  }
  if (e == T) { # Confere se a regressão exata foi solicitada (no caso, sim)
   cat('foi\n')
   m <- elrm(y/1 ~ x,interest=~x)
   return(m)
  }
 }
 out <- list()
 out[[1]] <- apply(y,2,simul,x=prob[,-(1:2)],...)  # Faz as simulações e armazena ela
 out[[2]] <- prob[,-(1:2)] # Matriz de X
 out[[3]] <- y # Matriz de Y
 out[[4]] <- c(n=n,rep=rep,betas=betas)
 out[[5]] <- c(Sys.time(),t)
 names(out) <- c('simul','x','y','param','tempo') # Dando nome aos bois
 class(out) <- 'simul' # Definindo a classe simul (para o print)
 return(out) # Retornando o objeto
}

## Função para identificar casos especiais da variavel X
identsimul <- function(simul,cond=0,inv=F,o=1) {
 p <- floor(c(.1,.35,.85)*simul$param[1]) # Variavel para auxiliar na localização dos X
 y_t1 <- simul$y # Cria variavel temporaria y_t1 (sucessos)
 y_t2 <- 1- simul$y # Cria variavel temporaria y_t2 (Fracassos)
 temp_3 <-
t(cbind(colSums(y_t1[1:p[1],]),colSums(y_t1[(p[1]+1):p[2],]),colSums(y_t1[(p[2]+1):p[3],]),colSums(y_t1[(p[3]+1
):simul$param[1],]))) # Conta os sucessos por X
 temp_4 <- diff(c(0,p,simul$param[1])) - temp_3 # Conta os fracassos
 temp_1 <- apply(temp_3,2,function(x) all(x >= cond)) # Armazena quais simulações possuem, pelo menos, 1
casela com a frequência menor que a desejada (Padrão = 0, separação completa)
 temp_2 <- apply(temp_4,2,function(x) all(x >= cond)) # Armazena quais simulações possuem, pelo menos, 1
casela com a frequência menor que a desejada (Padrão = 0, separação completa)
 out <- temp_1 & temp_2 # Combina os resultados
 if (inv) out <- !out # Inverte caso solicitado
 if (o == 2) {
  out <- cbind('1'=rowMeans(temp_3[,out]),'0'=rowMeans(temp_4[,out]))/diff(c(0,p,simul$param[1])) # Faz a
proporção média de Y por X
 }
 out # Retorna o objeto
}
```

```
## print para a classe 'simul'
print.simul <- function(simul,cond=0,inv=F,pr=T) {
 p <- simul$param # Armazena os parâmetros
 simul_res <- simul$simul
 if ((all(cond == 0)) && (inv == F) && (pr == T)) {
  cat(sprintf('\nResultado das simulações:\n n: %i \n Rep: %i \n Prevalência: %.2f \n Tempo: %.2f%s \nBeta
(Odds)\n B0 %.5f (%f) \n B1 %.6f (%f)\n B2 %.6f (%f)\n B3 %.6f (%f)\n B4 %.6f
(%f)\n',p[1],p[2],mean(simul$y),simul$tempo[1]-simul$tempo[2],attr(simul$tempo[1]-
simul$tempo[2],'unit'),p[3],exp(p[3]),p[4],exp(p[4]),p[5],exp(p[5]),p[6],exp(p[6]),p[7],exp(p[7]))) # Mostra
informações sobre a simulação (parâmetros)
 }
 if (length(cond) == 1) {  esc <- identsimul(simul,cond,inv) # Vetor com as simulações que respeitam as
condições
  if (pr) { cat(sprintf('\nResumo solicitado:\n Condição: Todas caselas são %s que %.0f.\n Total utilizado:
%i\n\n',ifelse(inv,'menor','maior ou igual'),cond,sum(esc))) } # Indica a condição e quantas simulações a
atende
 }
 else { esc <- cond # Vetor com as simulações desejadas
  if (pr) { cat(sprintf('\nResumo solicitado:\n Condição: Pré-seleção\n Total utilizado: %i\n\n',sum(esc))) } #
Indica a condição e quantas simulações a atende
 }
 o <- vector()
 if (sum(esc) > 0) {
  for (i in 0:4) { o <- rbind(o,c('Cobertura'=sum(simul_res[i+6,esc] < exp(p[i+3]) & exp(p[i+3]) <
simul_res[i+11,esc])/sum(esc),'Poder'=sum(!(simul_res[i+6,esc] < 1 & 1 <
simul_res[i+11,esc]))/sum(esc),'Amplitude'=mean(simul_res[i+11,esc] - simul_res[i+6,esc]))) } # Informações
sobre cobertura, poder e intervalo de confiança
  vicio <- simul_res[1:5,esc] - exp(p[3:7]) # Vicio de cada simulação
  if (sum(esc) > 1) {
   o <- cbind('VE'=rowMeans(simul_res[1:5,esc]),'Vicio'=rowMeans(vicio),'Vicio %'=rowMeans(vicio) /
exp(p[3:7]),'EQM'=rowMeans(vicio^2) + diag(var(t(simul_res[1:5,esc]))),o) # Matriz de informações
  }
  else { o <- cbind('VE'=simul_res[1:5,esc],'Vicio'=vicio,'Vicio %'=vicio / exp(p[3:7]),'EQM'=NA,o) } # Matriz de
informações
  rownames(o) <- paste('Odds_',0:4,sep='') # Dando nome aos bois [2]
  if (pr) {   print(o)  } # Imprime a matriz
 }
 else if (pr) { cat('Nenhuma simulação se encaixa\n') } # Caso nenhuma simulação repeite as condições,
imprime esse aviso.
 return(invisible(o))
}

####
# Simulando
####

set.seed(3141592)

simulacoes <- list()
simulacoes[[1]] <- list()
simulacoes[[1]][[1]] <-  simullog(n=50)
simulacoes[[1]][[2]] <-  simullog(n=100)
simulacoes[[1]][[3]] <-  simullog(n=250)
simulacoes[[1]][[4]] <-  simullog(n=500)
simulacoes[[1]][[5]] <-  simullog(n=1000)
simulacoes[[2]] <- list()
simulacoes[[2]][[1]] <-  simullog(n=50,betas=c(-3.9,log(c(5,3,1.5,1.05))))
simulacoes[[2]][[2]] <-  simullog(n=100,betas=c(-3.9,log(c(5,3,1.5,1.05))))
simulacoes[[2]][[3]] <-  simullog(n=250,betas=c(-3.9,log(c(5,3,1.5,1.05))))
simulacoes[[2]][[4]] <-  simullog(n=500,betas=c(-3.9,log(c(5,3,1.5,1.05))))
simulacoes[[2]][[5]] <-  simullog(n=1000,betas=c(-3.9,log(c(5,3,1.5,1.05))))
simulacoes[[3]] <- list()
simulacoes[[3]][[1]] <-  simullog(n=50,betas=c(-3,log(c(5,3,1.5,1.05))))
simulacoes[[3]][[2]] <-  simullog(n=100,betas=c(-3,log(c(5,3,1.5,1.05))))
simulacoes[[3]][[3]] <-  simullog(n=250,betas=c(-3,log(c(5,3,1.5,1.05))))
simulacoes[[3]][[4]] <-  simullog(n=500,betas=c(-3,log(c(5,3,1.5,1.05))))
simulacoes[[3]][[5]] <-  simullog(n=1000,betas=c(-3,log(c(5,3,1.5,1.05))))

# Fim \o/
```