Modelo Autologístico: Alguns Detalhes Na Estimação E Inferência

Elias Teixeira Krainski e Paulo Justiniano Ribeiro Junior

14 de maio de 2007

Resumo

O modelo autologístico tem sido muito utilizado para modelagem de dados de incidencia de doenças em plantas. Neste trabalho, consideramos a correção de viés de David Firth para a função de pseudo-verossimilhança. Também avaliamos o método de Monte Carlo inferência sobre os parâmetros e para estimação da variância das estimativas. Fizemos um estudo de simulação para comparar esses métodos.

1 Introdução

A modelagem de dados binários de imagens, presença de doenças em plantas, mapas epidemiológicos, etc., é vasto campo de pesquisa com freqüêntes avanços na busca de formas de modelagem e principalmente de metodologias de estimação. As duas abordagens mais comuns estão associadas ao modelo de regressão logística. Para dados de áreas geográficas, em dados epidemiológicos por exemplo, o mais comum é considerar a dependência espacial como um efeito aleatório. Isso é feito considerando utilizando esse efeito aleatório espacial como um campo aleatório markoviano. No caso de dados de *lattices* regulares, dados de imagens e doenças de plantas por exemplo, o mais comum é considerar o modelo autologístico. Neste caso, a probabilidade de uma planta estar contaminada é modelada em função da presença ou ausência de sucesso em plantas vizinhas.

O modelo autologístico foi proposto por (BESAG 1972). Posteriormente, (**?**) considerou um constexto mais amplo de auto-modelos. A estimação do modelo autologístico por máxima verossimilhança é intratável na prática. (BESAG 1977) sugeriu maximizar a pseudo-verossimilhança - PL, de *pseudo-likelihood*. A função de PL é simplesmente o produto da densidade das distribuições condicionais de cada observação. Nessas condições a estimação dos coeficientes de regressão do modelo é obtida simplesmente utilizando o método comum de estimação em modelos lineares generalizado, considerando a função de PL como se fosse a função de verossimilhança dos dados.

A estimação utilizando PL não apresenta problemas para a estimação dos coeficientes de regressção do modelo. Porém as estimativas de variância dos coeficientes são subestimadas, induzindo erros ao se fazer inferêencia. Uma proposta para resolver esse problema, é utilizar um procedimento de reamostragem (GUMPERTZ & RISTAINO 1997). Esse procedimento bootstrap só é possível via amostrador de gibbs, onde cada amostra é obtida de sua distribuição condicional.

Alguns outros métodos de estimação foram propostos para a estimação do modelo autologístico, tais como: método de PL generalizado, aproximações para a verossimilhança exata via monte carlo e aproximação estocástica por monte carlo via cadeia de markov, (**?**, **?**), (**?**, **?**) e (**?**, **?**). Todos esses métodos foram omparados com o método PL. O método de PL é o mais simples, porém inadequado para se fazer inferência sobre os parâmetros, por subestimar suas variâncias. O método de PL generalizado não é muito complicado, mas produz resultados similares ao método PL simples. Os demais métodos tem custo computacional elevado, devido ao método de simulaąo necessário para a obtençao de realizações do processo.

Todos os métodos de estimação citados envolvem a maximização da função de PL. Esta função é a mesma para um modelo de regressão logística, considerando dados independentes. Porém, pode não existir estimativas(**?**, **?**), fato é bastante comum em regressão logística. Isso é devido à separação completa ou quase-completa, fato que ocorre quando há uma covariável que prediz perfeitamente a resposta. A separaçõ quase-completa pode ocorrer se, por exemplo, todas as plantas não tenham vizinhas doentes ou tenham apenas uma planta vizinha doente. Além disso, todas as plantas doentes tem uma planta vizinha doente. Neste caso, uma tabela de freqüências 2×2 apresentará uma casela nula. Este fato é um método simples para detectar o problema de separação. Duas caselas nulas indica separação completa. Os métodos de simulação propostos para a estimação do modelo autologístico envolvem Uma solução para esse problema é o uso da correção de viés de David Firth, (Firth 1992) e (Heinze & Schemper 2002).

Se o problema de separação ocorre nos dados, as estimativas de PS não podem ser obtidas e torna-se necessário uma modificação nos algorítmos para os demais métodos. Porém, utilizando os outros métodos de estimação, o problema de separação pode ocorrer em algum momento. A chance disso ocorrer aumenta se a incidência é baixa e ou se a *lattice* é pequena.

Neste artigo avaliamos duas propostas para a estimação e inferência do modelo autologístico: 1) O uso da correção de viés; e 2) O uso do método de monte carlo. Na seção **??** apresentamos o modelo autologístico e os métodos de estimaąo. Na secção **??** aplicamos os métodos a um conjunto de dados da literatura e na seção **??** descrevemos o estudo de simulaçao feito e os resultados obtidos. As conclusões são feitas na seção **??**

2 Modelo autologístico

O modelo de regressão autologístico é um modelo de regressão logística utilizando como covariáveis a informação do *status* das observações vizinhas. A expressão do modelo pode ser dada por

$$\begin{array}{lcl} logit(p_{ij}) & = & eta_{ij} = \beta_0 + \gamma_1(y_{i,j-1} + y_{i,j+1}) + \gamma_2(y_{i-1,j} + y_{i+1,j}) \\ p_{ij} & = & exp\{eta_{ij}\}1 + expeta_{ij} \end{array} \; lcl \qquad (1)$$

em que $y_{i,j-1}$ e $y_{i,j+1}$ são os vizinhos na mesma linha; $y_{i,j-1}$ e $y_{i,j+1}$ são os vizinhos na linha adjacente; β_0 é uma constante desconhecida; γ_1 e γ_2 são coeficientes que indicam a importância do *status* das plantas vizinhas.

Esse modelo possui uma interpretação bastante conhecida. Por exemplo, se $exp\gamma_1 = 2$, significa que uma planta tem o dobro de chance de ficar doente se houver uma planta vizinha doente na mesma linha em relação a uma planta que não tem planta vizinha doente na mesma linha. O fato de considerar-mos duas covariáveis de vizinhança, diferenciando o efeito na linha e entre linha, torna o modelo mais flexível para avalizar se há efeito direcional. Em muitas culturas é razoável considerar essa abordagem, pois o espaçamento entre linhas é diferente do espaçamento dentro das linhas.

A função de verossimilhança pode ser escrita na forma

$$\underline{exp(Q(y))}$$

$\sum_D exp(Q(z))(2)$em que $Q(y)$ é uma soma nas densidades em cada y_{ij} e D é o conjunto de todas as configurações possíveis. Este denominador é praticamente intratável para *lattices* de grandes dimensões. Uma possibilidade para estimação dos parâmetros utilizando o método de máxima verossimilhança é utilizar áproximação utilizando um método baseado em monte carlo via algoritmo de Newton-Raphson ou aproximação estocástica utilizando monte carlo via cadeias de markov(**?**, **?**).

Um método simples para estimar os parâmetros do modelo é maximizar a função de pseudo-verossimilhança, que é o produto das distribuições condicionais, ou, mais eficientemente o seu logaritmo

$\sum_i^I \sum_j^J p_{ijij}^y (1 - p_{ij})^{(1-y)}(3)$em que p_{ij} é dado pela equação 1, I e J são o total de linhas e colunas da *lattice*.

O método de PS não apresenta grandes problemas na estimação dos parâmetros do modelo, porém não é adequado para estimar a variância das estimativas. Uma alterativa é utilizar (GUMPERTZ & RISTAINO 1997) propõe o uso de reamostragem para estimar a variância das estimativas.

3 Aplicação a dados de Pimentas de Sino

4 Estudo de simulação

5 Conclusões

A metodologia bootstrap, consiste em gerar amostras dos dados originais para estimar a quantidade de interesse com estas amostras. Porém, devido a configuração dos dados, a reamostragem só é possível utilizando o algoritmo amostrador de Gibbs. A idéia é de retirar amostras da distribuição de cada observação y_{ij} condicionando a todas os outras $y_{(ij)}$.

O algoritmo de Gibbs para obter N amostras $\hat{\gamma}$ é da seguinte forma:

1. ajustar o modelo aos dados observados utilizando o método de pseudo-verossimilhança;
2. fazer um ciclo em ordem aleatória pelas plantas atualizando cada planta de acordo com o modelo ajustado dos dados observados e o estado atual dos vizinhos;
3. ajustar o modelo com os re-amostrados e guardar as estimativas dos parâmetros;
4. repetir os passos 2 e 3 N vezes,
5. obter as estimativas de interesse usando os N vetores de parâmetros estimados.

Esse método é bastante utilizado, porém pode apresentar problemas quando a incidência é baixa. Na próxima seção apresentamos o problema que pode ocorrer.

6 Problema

Incialmente consideremos a aplicação do método apresentado na seção anterior ao conjunto de dados de Bell Pepper (GUMPERTZ & RISTAINO 1997). Esse conjunto de dados é um talhão

com 20 linhas de plantas e em cada linha os dados foram agregados em 20 quadrat s. O modelo proposto é:

$$logit(P(Y_{ij} = 1)) = \beta_0 + \gamma_1 R + \gamma_2 C + \gamma_3 dA + \gamma_4 dB \ , \tag{4}$$

em que $R = y_{i,j-1} + y_{i,j+1}$, $C = y_{i-1,j} + y_{i+1,j}$, $dA = y_{i-1,j-1} + y_{i+1,j+1}$ e $dA = y_{i+1,j+1} + y_{i-1,j-1}$.

A análise foi feita utilizando-se o pacote **Rcitrus**, desenvolvido em R (R Development Core Team 2006). Foram feitas 1100 re-amostras bootstrap. Na Figura 6 vizualizamos o traço dos valores obtidos em cada simulação. Chamamos a atenção para o fato de alguns valores estimados serem muito discrepantes.

Nas 1100 simulações observamos 22 valores menores que -5 para o coeficiente associado à covariável de vizinhança na coluna C e cinco valores para o coeficiente associado à covariável de vizihança na primeira diagonal dA. (GUMPERTZ & RISTAINO 1997) fez esta análise utilizando 550 simulações e descartou as primeiras 50 reamostras para fazer a inferência. Se nós descartarmos as primeiras 100 reamostras, teremos o sumário da Tabela 6 para as amostras dos coeficientes.

	(Intercept)	R	C	dA	dB
Min.	-4.1540	-2.3120	-18.3500	-16.8900	-0.9619
1st Qu.	-3.2050	0.9239	-0.7711	0.1457	0.6832
Median	-2.9770	1.2560	-0.2393	0.5366	0.9930
Mean	-2.9930	1.2590	-0.6203	0.4070	1.0050
3rd Qu.	-2.7620	1.5800	0.1912	0.8764	1.3330
Max.	-1.8040	2.8950	1.6440	1.9570	2.9470
Std. Error	0.3321	0.5186	2.5017	1.2272	0.5194
2.5%	-3.6829	0.3000	-1.9442	-0.9353	-0.0455
97.5%	-2.3785	2.2841	0.9410	1.4931	2.0169

Tabela 1: Sumário das dos coeficientes estimados nas 1000 simulações consideradas.

Nota-se na Tabela 6, que ocorreram valores de -18.35 para $\hat{\gamma_2}$ e -16.89 para $\hat{\gamma_3}$. A estimativa do erro-padrão de γ_2 é estimada como sendo 2.5017. Mas, como pode ser visto na Tabela 6, ao desconsiderar os 22 valores discrepantes obtidos, essa estimativa muda para 0.6550!

	(Intercept)	R	C	dA	dB
Min.	-4.1540	-2.3120	-2.3790	-2.7120	-0.9619
1st Qu.	-3.2020	0.9212	-0.7395	0.1517	0.6839
Median	-2.9710	1.2510	-0.2211	0.5381	0.9909
Mean	-2.9910	1.2510	-0.2668	0.4753	1.0060
3rd Qu.	-2.7570	1.5700	0.1981	0.8765	1.3330
Max.	-1.8040	2.8950	1.6440	1.9570	2.9470
Std. Error	0.3329	0.5178	0.6550	0.6029	0.5161
2.5%	-3.6832	0.2987	-1.6387	-0.8900	-0.0136
97.5%	-2.3644	2.2829	0.9471	1.4934	2.0189

Tabela 2: Sumário das dos coeficientes estimados nas 978 simulações sem valores discrepantes consideradas.

O procedimento ad hoc de utilizar apenas observações não discrepantes pode ser útil para corrigir o problema de valores estimados muito discrepantes, porém, convém estudar a causa do problema.

A regressão logística é comumente utilizada em diversas areas do conhecimento e é comum o problema de separação completa e quase-completa (Heinze & Schemper 2002) (Zorn 2005). Uma das formas de identificar a ocorrência dese problema, é a identificação de estimativa discrepante do coeficiente da variável causadora do problema de separação, associada a valor muito grande do erro-padrão do coeficiente. Se a covariável é discreta, uma maneira muito simples de identificar, é utilizando uma tabela de contingência. Neste caso o problema é identificado pela ocorrência de zero em pelo menos uma das caselas. Na próxima seção apresentamos o método de estimação via verossimilhança penalizada que resolve o problema de separação.

7 Metodologia

(Firth 1993) propos um método para reduzir o viés das estimativas de máxima verossimilhança. Esse método é baseado em uma penalização feita através da priori invariante de Jeffreys (Firth 1992). Esta penalização na função de verossimilhança remove o termo $O(1/n$ do vício assintótico dos coeficientes estimados e encontra estimativas finitas para os coeficientes e erros-padrões.

8 Aplicação

O procedimento de Firth está implementado para a regressão logística no pacote **brlr**, em R. Esse procedimento foi incluido como opção no pacote **Rcitrus**. O modelo ajustado na seção anterior, foi também ajustado usando o método de verossimilhança penalizada, em logar do método de máxima verossimilhança. Para comparar os resultados, utilizamos o mesmo número de simulações e a mesma semente.

Na Figura 8, notamos que não há valores discrepantes. nas estimativas dos coeficientes e na Tabela 8.

	(Intercept)	R	C	dA	dB
Min.	-3.8610	-2.0970	-3.2170	-2.4780	-0.7457
1st Qu.	-3.0660	0.9196	-0.6288	0.1848	0.6520
Median	-2.8620	1.1840	-0.1683	0.5437	0.9505
Mean	-2.8730	1.2060	-0.2344	0.4839	0.9600
3rd Qu.	-2.6640	1.4880	0.1901	0.8439	1.2550
Max.	-1.7830	2.6160	1.6380	1.8860	2.5950
Std. Error	0.3015	0.4589	0.6338	0.5500	0.4694
2.5%	-3.4869	0.3561	-1.5340	-0.8289	0.0097
97.5%	-2.3122	2.0922	0.8677	1.4014	1.9051

Tabela 3: Sumário das dos coeficientes estimados pelo método de verossimilhança penalizada nas 1000 simulações consideradas.

9 Conclusão

A solução para o problema bastante comum que ocorre em regressão logística é aplicada na metodologia de estimação do modelo autologístico.

Referências

BESAG, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data, *Journal of the Royal Statistics Society, Series B* (34): 75–83.

BESAG, J. (1977). Efficiency of pseudo likelihood estimators for simple gaussian fields, *Biometrika* (64): 616–618.

Firth, D. (1992). *Advances in GLIM and Statistical Modelling*, New York: Springer, chapter Bias reduction, the Jeffreys prior and GLIM, pp. 91–100.

Firth, D. (1993). Bias reduction of maximum likelihood estimates, *Biometrika* .

GUMPERTZ, M. L. ; GRAHAM, J. M. & RISTAINO, J. B. (1997). Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: Effects of soil variables on disease presence, *Journal of Agricultural, Biological and Environmental Statistics* **2**(2): 131–156.

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

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. *http://www.R-project.org

Zorn, C. (2005). A solution to separation in binary response models, *Political Analysis* **13**: 157–170.

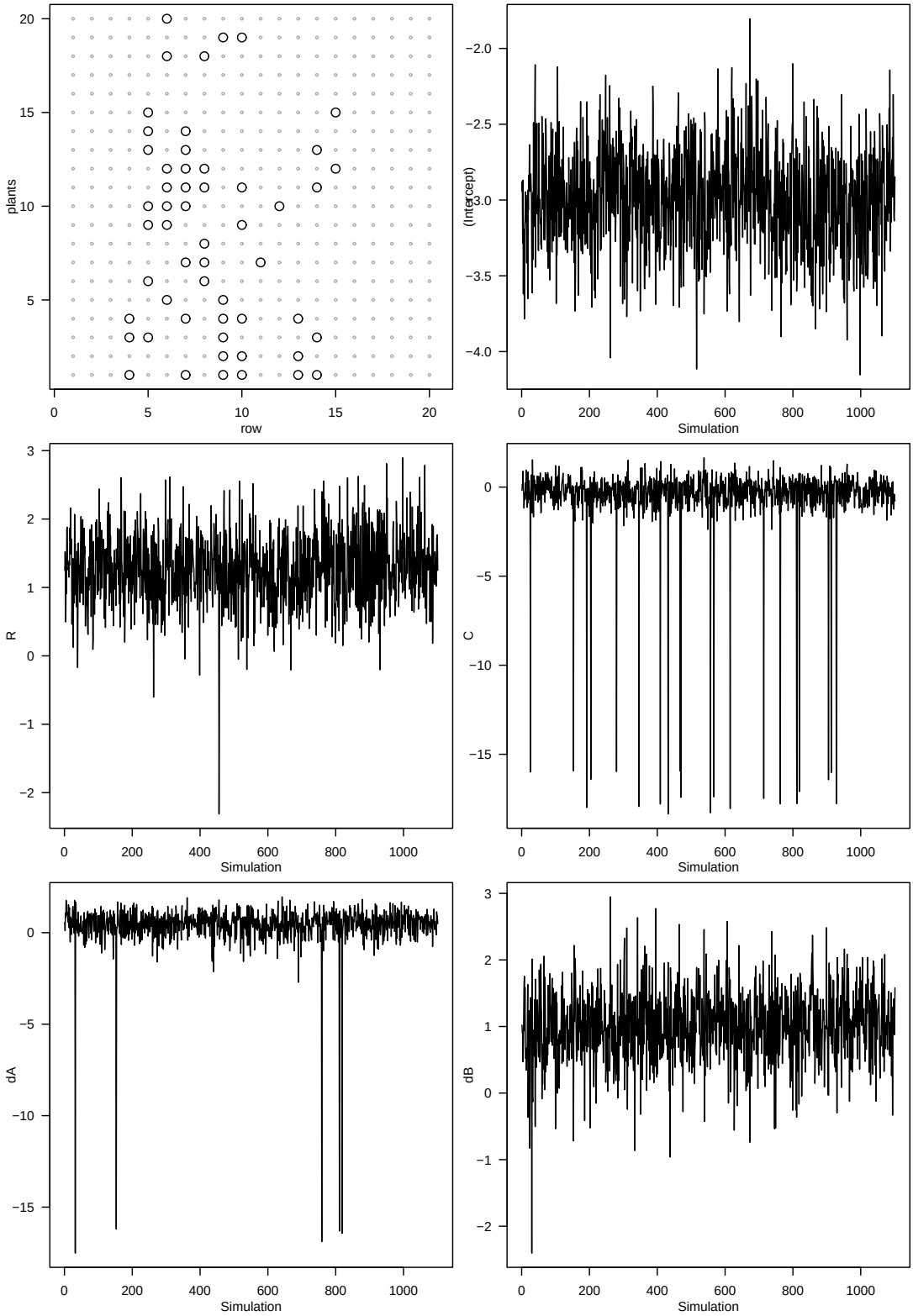

Figura 1: Mapa dos dados de Bell Pepper (superior esquerdo) e traç dos valores estimados com 1100 reamostragens

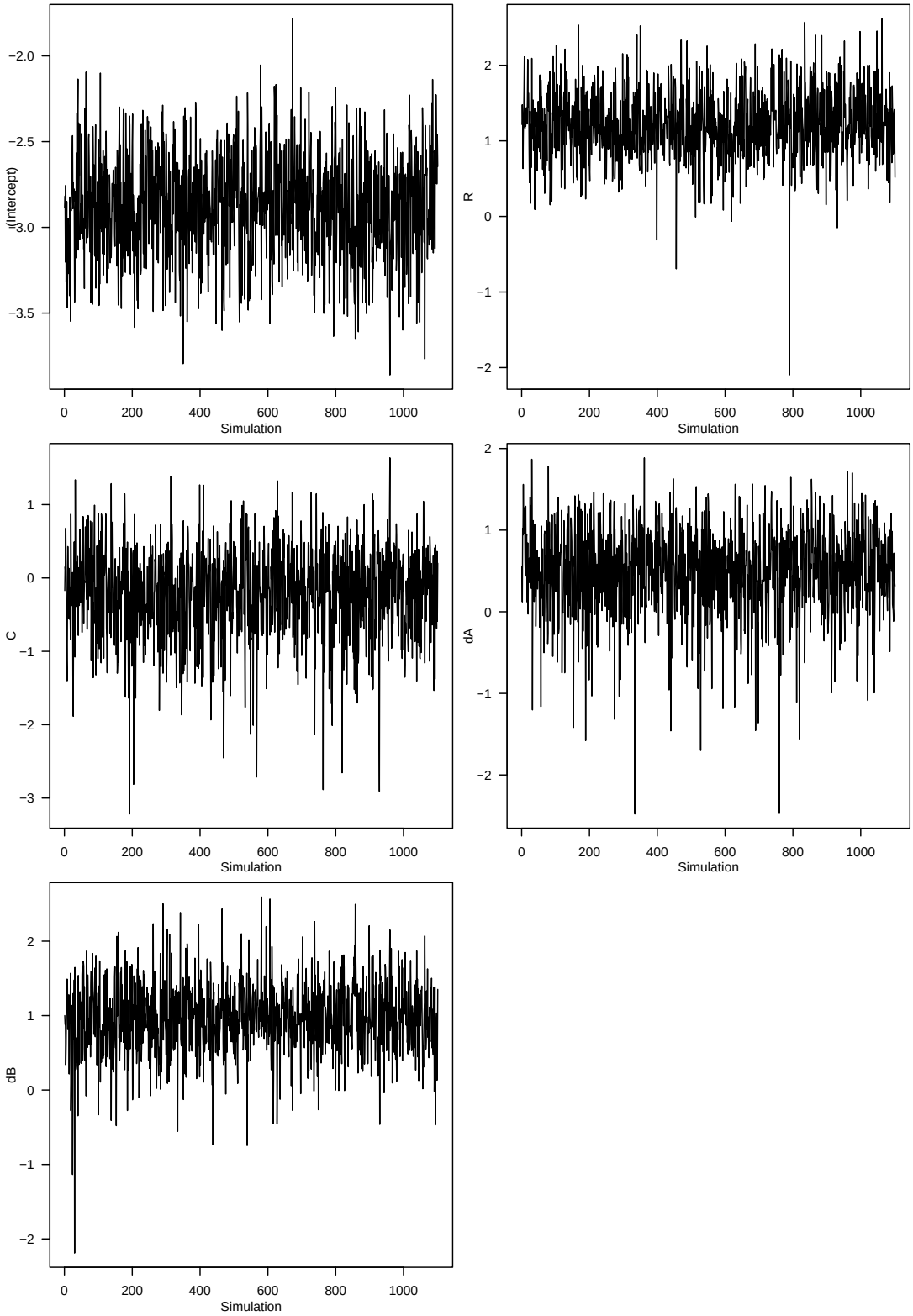

Figura 2: Traç dos valores estimados em 1100 reamostragens, utilizando o método de verossimilhança penalizada.