IV Congresso Anual
Sociedade Portuguesa de Estatística
Funchal, 9 a 15 de Junho de 1996

O Aparecimento de Valores Omissos no Contexto da Regressão Logística

Carla Mónica Santos Dias Pereira
C.E.M.P.A., Universidade Portucalense

Sumário: Um dos problemas mais complexos na análise estatística de dados é o problema da ocorrência de valores omissos. Ele pode surgir em várias situações oriundas de diversos factores tais como o tipo de estudo realizado, o processo de amostragem utilizado e o objectivo do inquérito elaborado. Neste trabalho tenta-se alertar para esta situação mencionando-se algumas alterações que poderão ocorrer quando se procede ao ajustamento de um modelo de regressão logística. São ainda apresentadas algumas soluções nomeadamente pela aplicação de algumas técnicas a um exemplo no contexto da medicina.

Abstract: One of the most complex problems in data analysis is that of missing values occurrence. It may happen in several situations as consequence of different causes such as the type of study, the sampling procedures and the goal of the inquiry. In this work, we emphasise this problem by focusing the changes occurred when we fit a logistic regression model. Some solutions are also presented including the practical application of some techniques in the domain of medicine.

Classificação da AMS: 62J02

1. INTRODUÇÃO

Ocorrências frequentes associadas ao aparecimento de valores omissos são por exemplo as seguintes situações:

- falta de entendimento das questões inseridas em determinado inquérito (caso em que este é por exemplo enviado via postal).
- falta de conhecimentos necessários para responder à questão colocada.
- falta de verbas necessárias à obtenção de determinada informação.
- efeitos externos que ocorrem na realização de determinadas experiências e que invalidam a informação recolhida, tais como a avaria de determinado instrumento de medida.
- medo associado à penalização derivada da opção por determinada resposta.
- ocorrência de morte de um indivíduo que se encontrava inserido em determinado estudo ou ensaio clínico.

O que podemos ter a certeza é que qualquer que seja a origem do aparecimento de valores omissos, a análise estatística estará seriamente condicionada. Este facto deverá fazer com que haja um cuidado extremo no desenvolvimento de técnicas apropriadas e

no aproveitamento de toda a informação disponível. Antes que nos lembremos de excluir todos os casos associados à ocorrência de valores omissos em pelo menos uma das variáveis, há que ter bom senso e,entender com clareza qual o objecto da análise, saber de onde são provenientes e o que representam os dados, avaliar quais as causas das omissões, analisar até que ponto é que o desconhecimento destes valores irá influenciar os resultados, determinar qual a proporção de valores omissos relativamente à dimensão da amostra recolhida, verificar se as omissões incidem sobre variáveis com pouca importância para o modelo, estudar a variabilidade com base nas restantes observações e proceder apenas à exclusão dos casos necessários, em cada fase da análise.

Citando Little & Rubin (1987) o que acontece frequentemente é que os *packages* estatísticos excluem as unidades que contêm valores omissos para alguma das variáveis intervenientes na análise. Esta estratégia é normalmente inapropriada já que o investigador geralmente está interessado em inferir sobre a população alvo e não sobre uma pequena parte dela que proporciona respostas para as variáveis relevantes na análise. Por vezes é mais fácil preocuparmo-nos com a tomada de passos que evitem o aparecimento desta situação do que posteriormente enfrentarmos a análise, condicionada pela omissão destes valores. Devemos ainda ter cuidado com a forma como codificamos o conjunto de observações omissas. Chatfield (1995) apresenta algumas situações que alertam para o perigo de uma má codificação.

Para avaliarmos a importância dos valores omissos para o problema em análise, podemos construir uma variável com duas categorias, 0 se estamos perante uma observação omissa e 1 caso contrário, e testar se existem diferenças significativas na variável resposta. Se for esse o caso, devemos tentar encontrar soluções para a redução do número de omissões. Sejam quais forem os métodos utilizados, deveremos, sempre que possível, efectuar uma comparação dos resultados obtidos por mais do que um processo. Se estes forem idênticos então podemos ter confiança nos resultados. Se são diferentes teremos de investigar as razões de tal mudança e avaliar qual se aproxima mais da realidade.

2. SOLUÇÕES

2.1. Ao nível da amostragem

As omissões encontradas poderão ter origem no tipo de amostragem utilizado. Mas se é o estatístico que procede à recolha dos dados então poderá recorrer à utilização de determinados tipos de amostragem que o poderão auxiliar no controlo deste mecanismo. Considerações sobre os diversos tipos de amostragem poderão ser encontradas em Little & Rubin (1987), Cochran (1977) e Vach (1994). De destacar a amostragem probabilística, a dupla amostragem geralmente designada por *double-sampling* ou *two-phase sampling* bem como as metodologias, designadas em Vach (1994) por, *surrogate covariates* e *subsampling of non-responders* .

Existem vários tipos de amostragem que nos permitem contornar este problema no caso em que é o próprio estatístico a elaborar o inquérito e a planear a recolha da amostra. Infelizmente não é isso que acontece na maior parte dos problemas pelo que teremos de nos preocupar com a construção de técnicas apropriadas.

2.2. Remoção de unidades ou de variáveis

Se os valores omissos se encontram concentrados em variáveis pouco significativas para o modelo então uma solução apropriada poderá ser simplesmente a sua remoção. Por outro lado se a percentagem de valores omissos é desprezável, relativamente à dimensão da amostra, e se não conseguimos aceder a nenhuma informação acerca destes valores, então poderemos pensar na exclusão destes casos.

2.3. Inserção de uma nova categoria

Tal como referem Tabachnick & Fidell (1989), por vezes a ausência de resposta a uma determinada questão poderá constituir um bom preditor do comportamento. O facto de não se responder a determinada questão poderá, como qualquer outra resposta, constituir um indicador de atitude. Poderá então ser importante analisarmos esse facto podendo para tal criar-se uma variável dicotómica como aquela que foi referida anteriormente, ou simplesmente adicionar-se uma nova categoria à variável em questão, que englobe os indivíduos que causaram o aparecimento de valores omissos. Este novo grupo é evidentemente muito heterogéneo. Se for pouco numeroso não é grave caso contrário Ghiglione & Matalon (1993) aconselham que se proceda a uma subdivisão num número de categorias mais homogéneas.

2.4. Processos de estimação

Os métodos de estimação dos parâmetros do modelo de regressão logística podem ser divididos em duas classes, os métodos designados por *ad-hoc* e os métodos consistentes. Vach (1994) conclui que os métodos que constituem a primeira classe possuem fracas propriedades já que conduzem a estimadores inconsistentes e à obtenção de variâncias e intervalos de confiança duvidosos. É por esta razão que se desenvolvem os chamados métodos consistentes onde se determinam propriedades assintóticas dos estimadores. No entanto, e ao contrário do que acontece na outra classe, existem enormes dificuldades associadas à sua implementação. Os métodos *ad-hoc* possuem essa grande vantagem já que há sempre possibilidade de se recorrer ao uso de *softwares* estatísticos. Além disso, este tipo de problema não provém sempre da investigação correspondente , mas de questões práticas. Por estas razões, e porque se pretende obter métodos que permitam soluções práticas e simples faremos aqui apenas referência a alguns dos métodos *ad-hoc*.

A forma mais simples de estimação consiste na substituição dos valores omissos por um ´palpite´ com base na experiência adquirida. Contudo, na prática o conhecimento à priori é geralmente bastante escasso. Na falta de mais informação, poderemos optar pela substituição pela média. Podemos ainda melhorar o procedimento se aproveitarmos a informação respeitante a outras variáveis.

Os métodos de regressão constituem uma alternativa e são mais objectivos que os anteriores. No entanto por vezes ajustam-se "bem demais" pelo facto dos valores terem sido calculados com base nas restantes variáveis. Mas o problema maior desta técnica centra-se na existência de variáveis que sejam boas preditoras; se nenhuma das restantes está associada com aquela que possui omissões, então a opção pela regressão torna-se tão boa quanto a simples substituição pela média.

Todas estas técnicas poderão ser úteis no tratamento dos valores omissos, no entanto, nunca nos poderemos esquecer que elas só terão algum interesse se o efeito da utilização destes métodos for inferior ao efeito proveniente da falta de informação.

3. ALTERAÇÕES NA CONSTRUÇÃO DO MODELO

A forma como tratamos os valores omissos poderá trazer consequências sob vários aspectos da modelização tais como:

- inserção de diferentes variáveis no modelo.
- modificações nos valores dos coeficientes estimados.
- não verificação da hipótese de linearidade no *logit* para as variáveis contínuas.
- rejeição do modelo na fase do diagnóstico.

4. EXEMPLO

O conjunto de dados que deram origem ao exemplo agora apresentado, foram fornecidos pelo Doutor Altamiro Pereira da Faculdade de Medicina do Porto e baseiam-se nos resultados obtidos para um inquérito realizado com o propósito de se determinar os factores que influenciam o diagnóstico de asma, variável que passaremos a designar por ASMA. A análise efectuada baseou-se no estudo de 14 variáveis sobre uma amostra de 363 indivíduos. O processo de recolha foi realizado via postal, facto que originou o aparecimento de uma série de valores omissos nomeadamente na única variável contínua, que representa o número de anos ao longo dos quais persistem dificuldades respiratórias, NUMANO. A incidência de omissões nesta variável é da ordem dos 20% (72 valores omissos) enquanto que nas restantes foram observados no máximo 10 valores omissos. Consequentemente decidiu-se desprezar estes últimos casos concentrando-se a nossa atenção apenas nos 72 que reportam à variável NUMANO. Contudo, antes de nos preocuparmos com o tratamento destes valores foi realizado o teste do Qui-Quadrado para se avaliar o impacto de NUMANO na variável resposta ASMA tendo-se obtido:

		ASMA	
		0 (não)	1 (sim)
VOMISSO	0	60	11
	1	144	148

p-value (com correcção de Yates) < 0,00000

As questões inseridas no inquérito, deram origem ao seguinte conjunto de variáveis independentes:

NUMANO- Número de anos ao longo dos quais persistem as dificuldades respiratórias (contínua)

CLASSE- Classe social (nominal)

FUM- Já fumou? se sim quantos cigarros fuma, em média, diariamente? (ordinal)

EXPIROM- Possui um expirómetro? (dicotómica)

RESPP- Teve respiração com pieira, nos últimos 12 meses? (ordinal)

FREQP- Frequência de pieira, nos últimos 12 meses (ordinal)
INTP- Intensidade de pieira, nos últimos 12 meses (ordinal)
FRIOP- Teve pieira apenas com frio, nos últimos 12 meses? (dicotómica)
FREQTOSSE- Com que frequência, nos últimos 12 meses, é que acordou a tossir ou com o tórax apertado? (ordinal)
MED- Foi ao médico nos últimos 12 meses? (dicotómica)
REMED- Fez alguma medicação nos últimos 12 meses? Regular ou não? (ordinal)
TRATALT- Fez algum outro tratamento alternativo nos últimos 12 meses? (dicotómica)
SEXO- Sexo (dicotómica)
TAREF- Teve dias em que não foi capaz de cumprir as suas tarefas diárias? (dicotómica)

4.1. Considerações acerca dos procedimentos utilizados

Com efeito, para que pudesse ser realizada uma comparação entre os resultados obtidos para diferentes métodos, utilizaram-se os seguintes procedimentos:

1. Ajustamento de uma distribuição → não se obtiveram resultados significativos;
2. Eliminação dos indivíduos onde ocorreu pelo menos uma omissão;
3. Estimação através da Regressão Linear;
4. Substituição pela média.

→ Eliminação dos indivíduos onde ocorreu pelo menos uma omissão

Na utilização do método *stepwise* será mais proveitoso o cálculo do *p-value* passo a passo. Só se deverá proceder à eliminação dos valores omissos, consoante as variáveis dão entrada no modelo. A utilização do SPSS faz com que todos os casos sejam eliminados à partida.

→ Estimação através da Regressão Linear

Para que fosse implementada esta técnica foram realizados os seguintes passos:

- Avaliação das associações entre a variável NUMANO e as restantes variáveis independentes. Verificou-se a existência de uma associação considerada significativa (*p-value* < 0,15) entre NUMANO e as seguintes variáveis: EXPIROM, RESPP, FREQTOSSE, MED, REMED e SEXO.
- Estimação dos valores omissos por regressão nestas variáveis.

→ Substituição pela média

Os valores omissos em NUMANO foram substituídos pela média, 18, embora se considerasse que esta não era a solução adequada para estes dados. O coeficiente de dispersão indica-nos cerca de 53,8% de dispersão em relação à média.

4.2. Resultados

A tabela seguinte apresenta, para cada procedimento, as variáveis seleccionadas pela ordem a que deram entrada através do método *stepwise*.

2	3	4
EXPIROM	REMED	REMED
REMED	NUMANO	NUMANO
NUMANO	FRIOP	FRIOP
FRIOP	CLASSE	CLASSE
CLASSE	EXPIROM	EXPIROM
FUM		

Os resultados deste método (*p-value* de entrada 0,15 e de saída 0,20) encontram-se na tabela a seguir apresentada. Os valores entre parêntesis indicam-nos o número acumulado de omissões. A variável EXPIROM foi removida por se considerar que não tinha significado biológico para esta análise.

	EXPIROM	REMED	NUMANO	FRIOP	CLASSE	FUM	FREQP	RESPP	FREQ TOSSE	MED	TAREF	INTP	SEX
0	0,0000 (4)	0,0000	0,0000	0,0000	0,0083	0,0011	0,0005	0,0003	0,0146	0,0173	0,0028	0,0013	0,0573
1	0,0000	0,0000 (12)	0,0000	0,0000	0,0001	0,0007	0,0003	0,0000	0,0752	0,0097	0,0006	0,0006	0,0424
2	0,0000	0,0000	0,0000 (76)	0,0000	0,0008	0,0174	0,0358	0,0053	0,7569	0,4397	0,0162	0,0336	0,0697
3	0,0169	0,0000	0,0000	0,0013 (78)	0,0028	0,1019	0,0930	0,3230	0,6389	0,7447	0,1380	0,1225	0,2228
4	0,0193	0,0000	0,0000	0,0013	0,0035 (82)	0,0814	0,1428	0,5505	0,4277	0,7958	0,1829	0,1597	0,3775
5	0,0122	0,0000	0,0000	0,0016	0,0035	0,1302 (82)	0,1454	0,3447	0,3099	0,6819	0,2479	0,2229	0,5622
6	0,0118	0,0000	0,0000	0,0012	0,0053	0,1302	0,1921	0,3226	0,3636	0,5318	0,1661	0,1978	0,3005

* a variável MED não foi incluída por ter sido considerada não significativa, na perspectiva univariada.

De seguida foram determinados os valores da razão das possibilidades para as diversas categorias de cada variável, de forma a se poder avaliar as diferenças encontradas. Tomou-se sempre como grupo de referência, o nível mais baixo.

Variáveis	2	3	4
D1REMED	2,2389	3,3111	3,5629
D2REMED	7,3832	13,3565	13,8191
FRIOP	0,3056	0,2213	0,2151
D1CLASSE	**0,3217**	**0,2929**	**0,2937**
D2CLASSE	**0,3475 ↑**	**0,2462 ↓**	**0,2554 ↓**
D3CLASSE	**0,3349**	**0,3314**	**0,3342**
NUMANO *	1,7212	1,5296	1,8112

* valores obtidos para um incremento de 10 anos

A tabela não evidencia grandes alterações a nível de comportamento das variáveis à excepção da variável CLASSE cuja interpretação passa a fazer mais sentido biológico. Esta variável possui as seguintes categorias:

0-classe baixa / 1-classe média / 2-classe alta /
3-domésticas, estudantes, desempregados e reformados

O que acontecia no modelo 2 é que se verificava um aumento no "risco" de diagnóstico de asma quando comparados os indivíduos da classe alta e média relativamente aos da classe baixa i.e. os resultados evidenciavam que o diagnóstico de asma era mais frequente nos indivíduos pertencentes à classe alta, quando comparada com a baixa, do que os da classe média. Ora no modelo 3 e 4 a conclusão já é contrária indo de encontro ao que à partida era esperado. Quanto à última categoria achamos que não fazia sentido fazer qualquer comentário.

	2	3	4
número de indivíduos no modelo	281	337	341
número de covariáveis-classe	248	237	219

→ Medidas globais de ajustamento

Na fase do diagnóstico, devido ás exigências afectas à validação distribucional das estatísticas de teste, os resultados foram agrupados em covariáveis-classe.

	D	*p-value*	X^2	*p-value*	número de graus de liberdade
2	268,6	0,0702	258,6	0,1503	236
3	233,4	0,4068	260,2	0,0775	229
4	251,3	0,0304	265,8	0,0064	211

D- *Deviance* / X^2- Estatística do Qui-Quadrado

⇒ O modelo obtido pelo procedimento 4 é rejeitado.

⇒ O modelo obtido pelo procedimento 2 à partida parece não ser rejeitado tendo, no entanto, que se ter cuidado com as conclusões uma vez que não é satisfeita a hipótese assintótica associada aos resultados distribucionais.

⇒ O modelo obtido pelo procedimento 3 parece ser o que melhor se adapta.

→ Medidas individuais de ajustamento

Neste caso utilizaram-se as estatísticas ΔX_j^2 e ΔD_j como medidas do efeito da remoção da j-ésima covariável-classe do modelo (ver Hosmer & Lemeshow, 1989). Como podemos concluir pela análise do gráfico que passaremos a apresentar verifica-se um aumento do número de classes de covariáveis mal ajustadas quando passamos do modelo 2 para o 3 (o gráfico de ΔX_j^2 é idêntico). Este aumento não é deveras muito significativo já que deixamos de possuir 5 (≈ 2%) classes de covariáveis mal ajustadas para passarmos a um conjunto de 10 (≈4%).

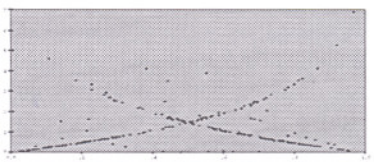

Gráfico de $\hat{\Pi}_j$ versus ΔD_j para o modelo 2

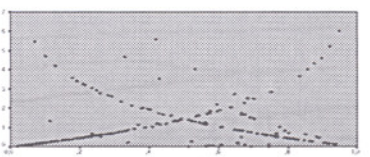

Gráfico de $\hat{\Pi}_j$ versus ΔD_j para o modelo 3

→ Medida de influência

Para se avaliar a existência de observações que alterem significativamente os coeficientes estimados recorreu-se à estatística $\Delta\hat{\beta}_j$ como indicador do efeito provocado na eliminação da j-ésima covariável-classe (ver Hosmer & Lemeshow, 1989). No presente caso, a simples representação desta medida versus os valores das probabilidades estimadas é suficiente para que se possa concluir da não existência de observações influentes em nenhum dos dois modelos. Aliás também os próprios valores do *leverage* são considerados pequenos.

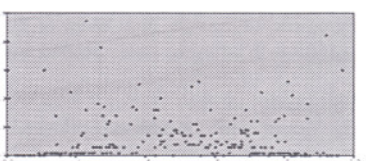

Gráfico de $\hat{\Pi}_j$ versus $\Delta\hat{\beta}_j$ para o modelo 2

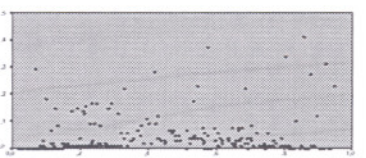

Gráfico de $\hat{\Pi}_j$ versus $\Delta\hat{\beta}_j$ para o modelo 3

5. COMENTÁRIOS

Em face dos resultados encontrados podemos concluir que a forma como tratamos os valores omissos poderá influenciar as conclusões obtidas. Embora este exemplo não seja marcadamente ilustrativo das alterações que poderão ocorrer, por certo muitas análises estarão seriamente condicionadas pelas metodologias utilizadas para a resolução deste problema.

Cada vez mais deveremos preocuparmo-nos com a satisfação das novas exigências que surgem associadas ás práticas estatísticas. È importante que não se tratem os valores omissos como um simples resíduo sem interesse (Ghiglione & Matalon ,1993). Geralmente as análises realizadas apresentam lacunas no que diz respeito ao tratamento destes valores. No entanto, embora se encontrem ainda bastantes limitações, já se começa a revelar um certo interesse no desenvolvimento de técnicas apropriadas bem como na criação de *softwares* estatísticos capazes de lidar com a implementação de novos métodos (como por exemplo no BMDPAM).

Encontramo-nos perante um campo muito vasto e ainda pouco explorado onde os resultados encontrados baseiam-se geralmente na experiência e em resultados afectos a um conjunto de simulações. Contudo a qualidade dos resultados obtidos requer que se dê o máximo de atenção ao processo de recolha dos dados e à forma como estes são recodificados. É claro que se torna importante que nos apercebamos que por mais que sejamos cuidadosos poderemos cair na situação em que nos surgem valores omissos, nomeadamente quando lidamos com grandes conjuntos de dados. No entanto, tal como Chatfield (1995) refere, se o processo de construção do modelo for bem orientado e minucioso, poderemos obter bons resultados.

Perante esta situação deixamos aqui a sugestão para trabalhos futuros salientando no entanto a opinião de Vach (1994) na qual nos alerta para o facto de que, mesmo os bons métodos para o tratamento dos valores omissos nunca nos deverão seduzir a esquecer que estes nunca constituirão uma forma de obtenção de um bom substituto dos dados completos.

BIBLIOGRAFIA

CHATFIELD, C. (1995) - Problem Solving, a statistician´s guide. Second Edition. Chapman & Hall.

COCHRAN, W. (1977) - Sampling Techniques. Third Edition. John Wiley & Sons.

GIBBONS, L. & HOSMER, D.W (1991) - Conditional Logistic Regression With Missing Data. Commun. Statist.-Simula. , 20(1), pp. 109-120.

GHIGLIONE & MATALON (1993) - O Inquérito, teoria e prática. Celta Editora.

HOSMER , D. & LEMESHOW, S. (1980) - Goodness-of-fit Tests for the Multiple Logistic Model. Commun. Statist.- Theor.Meth. , A9(10), 1043-1069

HOSMER , D. & LEMESHOW, S. (1989) - Applied Logistic Regression. John Wiley & Sons.

LITTLE, R. & RUBIN, D. (1987) - Statistical Analysis with Missing Data. Second Edition. John Wiley & Sons.

TABACHNICK, B. & FIDELL, L.(1989) - Using Multivariate Statistics. Harper & Row, Publishers.

VACH, W. (1994) - Logistic Regression With Missing Values in the Covariates. Springer-Verlag.