CLASSIFICAÇÃO E ANÁLISE DE DADOS

Métodos e Aplicações III - CLADMAp III

Editores

Helena Bacelar-Nicolau

Fernanda Sousa

Carlos Marcelo

Ana Sousa Ferreira

Paulo Infante

Adelaide Figueiredo

CLASSIFICAÇÃO E ANÁLISE DE DADOS

MÉTODOS E APLICAÇÕES III - CLADMAp III

Editores
Helena Bacelar-Nicolau
Fernanda Sousa
Carlos Marcelo
Ana Sousa Ferreira
Paulo Infante
Adelaide Figueiredo

Associação Portuguesa de
Classificação e Análise de Dados

Título
Classificação e Análise de Dados – Métodos e Aplicações III

Editores
Helena Bacelar-Nicolau (Universidade de Lisboa)
Fernanda Sousa (Universidade do Porto)
Carlos Marcelo (Instituto Nacional de Estatística)
Ana Sousa Ferreira (Universidade de Lisboa)
Paulo Infante (Universidade de Évora)
Adelaide Figueiredo (Universidade do Porto)

Impressão
Instituto Nacional de Estatística
Av. António José de Almeida
1000-043 LISBOA

1.ª Edição
Lisboa, Abril de 2019
ISSN 2183-8801
Depósito legal 454535/19
Tiragem: 200 exemplares

Todos os direitos reservados. Nenhuma parte desta publicação pode ser reproduzida por processo mecânico, eletrónico ou outro sem autorização escrita dos editores.

Avaliação do efeito do desenho de amostragem em modelos de regressão logística

Ana Laura Carreiras[1] · Paulo Infante[2] · Anabela Afonso[3] · Maria Filomena Mendes[4]

Resumo As amostras complexas resultam da combinação de vários métodos de amostragem para a seleção de uma amostra representativa da população. Uma das estratégias mais usadas para corrigir as estimativas obtidas com base no pressuposto da amostra complexa é considerar que a amostra é proveniente de um esquema de amostragem aleatória simples usando os pesos normalizados corrigidos pelo efeito do desenho (*deff*). Neste trabalho, analisa-se o impacto de uma estimação incorreta do *deff* na significância das variáveis e no seu efeito em modelos de regressão logística.

Palavras-chave: Amostras Complexas, Efeito do Desenho, *Jackknife*, Pesos, Regressão Logística.

1 Introdução

Em muitos estudos é necessário usar delineamentos de amostragem complexos para selecionar uma amostra representativa da população. Uma amostra complexa possui pelo menos uma das seguintes características: estratos, conglomerados, probabilidades de seleção desiguais e ajustamentos para compensar as não respostas e outras pós-estratificações (Lavrakas, 2008). Alguns autores ainda tratam este tipo de amostras sob a suposição de amostragem aleatória simples, ignorando o desenho de amostragem. Esta abordagem pode produzir incorreções, tanto para as estimativas, como para as respetivas variâncias, comprometendo os resultados e as conclusões da pesquisa (Osborne, 2011).

Contudo, por motivos de confidencialidade, muitas vezes não é facultada a

[1] Mestrado em Modelação Estatística e Análise de Dados, *analaura.carreiras@gmail.com*
[2] CIMA/IIFA e DMAT/ECT, Universidade de Évora, *pinfante@uevora.pt*
[3] CIMA/IIFA e DMAT/ECT, Universidade de Évora, *aafonso@uevora.pt*
[4] CIDEHUS e DSOC/ECS, Universidade de Évora, *mmendes@uevora.pt*

informação relativa às variáveis associadas ao desenho de amostragem. Algumas vezes são fornecidos os pesos de replicação que contêm informação sobre o desenho e salvaguardam a confidencialidade (Sturgis, 2004). Mas nem todas as técnicas estatísticas que se pretendem utilizar e que estão implementadas em *software*, permitem a utilização destes pesos.

Existem várias estratégias para corrigir as estimativas obtidas, com base no pressuposto da amostra ser proveniente de um esquema de amostragem aleatória simples (Osborne, 2011). A mais usual é ponderar as observações pelos pesos normalizados corrigidos pelo efeito do desenho (*deff*). Deste modo, quando o valor do *deff* é superior a 1 então a amostra aleatória correspondente deverá ter uma dimensão menor, para se obter a mesma precisão nas estimativas da obtida com a amostra complexa.

O *deff* será descrito no ponto seguinte, mas é de referir que depende do número de observações e dos pesos de amostragem. Contudo, nem sempre é fornecido o valor do *deff* e este, usualmente, varia consoante a variável em estudo (Sturgis 2004). Baseando-nos no facto de não nos ser fornecido nenhum tipo de informação sobre o desenho da amostragem temos de estimar o *deff* para corrigir os pesos.

Assim, o objetivo deste trabalho prende-se com a avaliação das consequências de uma estimação incorreta do efeito do desenho na significância das variáveis e no seu efeito sobre a resposta em modelos de regressão logística. Para tal, foram utilizadas duas abordagens: 1) comparar os vários modelos ajustados quando se consideram as observações ponderadas pelos pesos normalizados corrigidos por vários valores para o *deff* e pelos pesos de replicação; 2) considerar o modelo ajustado com base nos pesos de replicação e comparar as estimativas obtidas para os parâmetros, e respetivos erros padrão, das variáveis significativas quando se consideram os pesos normalizados corrigidos com diferentes valores para o *deff*. Esta segunda abordagem permite verificar qual o impacto de uma incorreta estimação do *deff* na significância dos parâmetros associados às variáveis que sabemos serem significativas.

2 Efeito do desenho e pesos

Ao contrário do que acontece numa amostra aleatória simples, numa amostra complexa nem todos os indivíduos da amostra representam o mesmo número de indivíduos da população, sendo essa informação dada por pesos. Existem vários tipos de pesos, como o peso do delineamento, o peso da pós estratificação, o peso da população, entre outros.

Em modelação estatística e análise dos dados é preciso incluir os pesos, sendo o investigador alertado para isso aquando da cedência dos dados. Mas nem sempre o investigador está familiarizado com os pesos disponibilizados, o que dificulta a identificação dos pesos a utilizar na análise que pretende realizar, bem como se estes devem ou não ser alvo de alguma transformação. Por outro lado, alguns

investigadores optam por não considerar os pesos, abordagem que pode produzir incorreções, tanto para as estimativas, como para as respetivas variâncias, enviesando os resultados e as conclusões (Osborne, 2011).

Para se tentar evitar estes erros, como já foi referido, aplica-se o efeito do desenho. Este efeito é uma medida que quantifica a perda ou o ganho de precisão na estimação devido ao uso de uma amostra complexa em vez de uma amostra aleatória simples e é definido pelo quociente entre a estimativa da variância determinada pelo plano amostral complexo e a estimativa da variância obtida por uma amostra aleatória simples do mesmo tamanho (Kish, 1965).

Seja $\hat{\theta}$ um estimador para θ, o *deff* é definido por:

$$deff(\hat{\theta}) = \frac{Var_{amostra\ complexa}(\hat{\theta})}{Var_{amostra\ aleatória\ simples}(\hat{\theta})}.$$

Sem termos as variáveis associadas ao desenho do desenho nem os pesos de replicação, a solução é utilizar o efeito do desenho aplicando-o aos pesos normalizados. Seja w_i o peso de amostragem do indivíduo i na amostra, o peso normalizado w_i' e o peso corrigido w_i^* são definidos, respetivamente, por:

$$w_i' = \frac{w_i}{\sum_{j=1}^{n} w_j} n \quad \text{e} \quad w_i^* = \frac{w_i'}{\sqrt{deff}}.$$

Esta correção leva à estimação de erros padrão muito próximos dos que se obteriam considerando as variáveis do desenho.

3 Material e métodos

3.1 Dados amostrais

Os dados utilizados para esta análise de sensibilidade aos valores do *deff* pertencem ao Inquérito à Fecundidade 2013, realizado no âmbito de um protocolo entre a Fundação Francisco Manuel dos Santos e o Instituto Nacional de Estatística.

Selecionou-se um conjunto de 27 variáveis com características sociodemográficas dos indivíduos e, a partir destas, ajustaram-se modelos de regressão logística à variável dicotómica tem filhos *vs.* não tem filhos (Hosmer et al., 2013).

Por motivos de confidencialidade, as variáveis associadas ao desenho não foram facultadas, mas, em contrapartida, foram fornecidos os pesos de replicação que contêm toda a informação necessária para se obter uma estimativa do erro padrão do estimador do parâmetro pelo estimador de variância do tipo *Jackknife* (Lohr, 2010).

3.2 Regressão logística

O ajustamento de modelos de regressão logística foi efetuado com recurso aos pacotes *survey, rms, mfp, EPI e epiR* do programa *R Project* (R Core Team, 2012). Nesta etapa consideraram-se os pesos de replicação (pacote *survey*) e os pesos normalizados corrigidos para o efeito do desenho (restantes pacotes).

Para ajustar os modelos seguimos a seguinte estratégia (Hosmer & Lemeshow, 2013): (1) para o modelo inicial foram selecionadas todas as covariáveis que se revelaram significativas na fase univariada (valor $p < 0,20$); (2) a partir deste modelo foram eliminadas sucessivamente, e por ordem decrescente dos valores p, todas as covariáveis não significativas (valor $p > 0,05$); (3) verificámos se alguma(s) das covariáveis que não foram incluídas no modelo inicial se mostra(m) agora significativa(s) na presença das que estão no modelo, caso em que foram adicionadas ao modelo; (4) avaliámos a forma funcional das covariáveis contínuas, através do alisamento em diagrama de dispersão e ajustamento de um modelo aditivo generalizado (GAM), sendo aplicado o método dos polinómios fracionários em casos de não linearidade; (5) foram testadas as interações que faziam sentido no contexto do estudo (valor $p < 0,05$); (6) foi feita uma análise de resíduos por padrões para pesquisa de observações influentes ou *outliers*, através dos resíduos da desviância, distância de Cook e estatísticas DFBETAS.

A significância das covariáveis e das interações foi testada recorrendo ao teste de Wald modificado (Hosmer e Lemeshow, 2013). A adequabilidade do ajustamento foi feita recorrendo aos testes de bondade de ajustamento de Hosmer e Lemeshow e de Cessie-van Houwelingen (ou teste de Wald no caso de se usarem os pesos de replicação) e a capacidade discriminativa do modelo avaliada pelo valor da AUC da curva ROC.

4 Resultados

4.1 Abordagem 1: comparação dos modelos

A variável resposta dicotómica considerada, tem filhos *vs.* não tem filhos, foi avaliada para todos os indivíduos da amostra e a categoria de referência (não tem filhos) representa 33,2% da dimensão da amostra ($n = 7624$).

Os modelos de regressão logística ajustados apresentam um valor p do teste de Hosmer e Lemeshow (H&L) semelhante, um valor da AUC idêntico, mas o valor de R^2 (Nagelkerque) diminui com o aumento do valor do *deff*, assim como o valor p do teste Cessie Van Howelingen (Tabela 1).

Tabela 1 – Medidas da bondade do ajustamento e capacidade discriminativa dos modelos ajustados com base nos diferentes valores de *deff* e nos pesos de replicação.

Modelos	H&L (p)	R^2	AUC	Valor p
deff = 0,5	0,05	0,88	0,98	0,75*
deff = 1	0,06	0,84	0,98	0,55*
deff = 1,5	0,14	0,81	0,98	0,18*
deff = 2	0,06	0,81	0,98	0,13*
deff = 2,5	0,08	0,79	0,98	0,05*
deff = 3	0,08	0,79	0,98	0,05*
Pesos de replicação	0,11	0,72	0,98	0,93**

* Teste de Cessie van Howelingen, ** Teste de Wald

Com a diminuição do valor do *deff* aumenta o número de variáveis significativas e também o número de interações (Tabela 2). Os modelos ajustados considerando valores do *deff* = 2, 2,5 e 3 têm o mesmo número de variáveis significativas que o modelo ajustado com base nos pesos de replicação, mas diferente número de interações significativas. O nível de significância dos coeficientes varia com o valor considerado para o *deff*. Para alguns coeficientes o nível de significância mantém-se igual em todos os *deff* (por ex., variável A e C) e há também situações em que o nível de significância diminui com o valor do *deff* (por ex., variável B). Este resultado deve-se ao facto de quanto menor o valor do *deff* maior a dimensão da amostra aleatória e, por conseguinte, menor o erro-padrão das estimativas.

Tabela 2 – Significância das variáveis comuns, número de variáveis e de interações significativas nos modelos ajustados para cada *deff* e pesos de replicação.

Variável	A	B	C	N.º de variáveis significativas a 5%	N.º de interações significativas a 1%
deff = 0,5	***	*	***	13	14
deff = 1	***	*	***	11	10
deff = 1,5	***	.	***	9	8
deff = 2	***	***	***	8	7
deff = 2,5	***	**	***	8	6
deff = 3	***	**	***	8	6
Pesos de replicação	***	**	***	8	5

Significativa a: . 0,10; * 0,05; ** 0,01; *** <0,001.

A Tabela 2 foi elaborada de forma a permitir ter uma noção da complexidade dos modelos sem torná-la muito densa, dado o número de variáveis e interações envolvidas em alguns dos modelos. Com o objetivo de obter modelos mais parcimoniosos optou-se por considerar um nível de significância inferior para inclusão das interações no modelo (1%). As comparações foram apenas realizadas

para as variáveis que não estão incluídas nas interações, pois os coeficientes e significância das restantes variáveis são influenciados pelas interações em que as mesmas estão envolvidas. As 3 variáveis apresentadas na Tabela 2 são as que não intervêm em interações no modelo ajustado com os pesos de replicação, mas que com o aumento do valor do *deff* e consequente aumento de interações passam também a estar incluídas em interações. As variáveis significativas para o modelo ajustado com os pesos de replicação são 8. Todos os outros modelos incluem essas 8 variáveis. Os modelos ajustados com os pesos normalizados corrigidos pelo *deff* = 0,5, *deff* = 1 e *deff* = 1,5 são os que incluem mais variáveis significativas. As interações significativas para o modelo ajustado com os pesos de replicação foram 5. Todos os outros modelos ajustados com os pesos normalizados corrigidos pela estimação do *deff* incluem essas 5 interações e mais as reportadas na Tabela 2.

O efeito nas estimativas dos parâmetros dos modelos de regressão logística foi medido através de uma comparação da variação percentual entre desvio-padrão, coeficientes e *odd racios* (OR) estimados com base nos modelos com os pesos normalizados corrigidos pelo *deff,* e estimado com base nos pesos de replicação (Figura 1 e Figura 2). Esta análise apenas é possível realizar entre as variáveis comuns às do modelo ajustado com os pesos de replicação. O número de barras é idêntico em todos os valores de *deff* e representa as categorias das variáveis, contemplando também as interações que são idênticas entre os modelos.

Comparando a variação percentual entre o desvio-padrão estimado com base nos modelos com os pesos normalizados corrigidos pelo *deff* e o estimado com base nos pesos de replicação, a sub ou sobrestimação do desvio-padrão é maior nos valores de *deffs* mais reduzidos (Figura 1). Na maior parte dos casos ocorre uma subestimação para *deffs* mais baixos e quando aumenta o valor do *deff* algumas estimativas do desvio-padrão reduziram, o que em alguns casos originou uma sobrestimação dos mesmos.

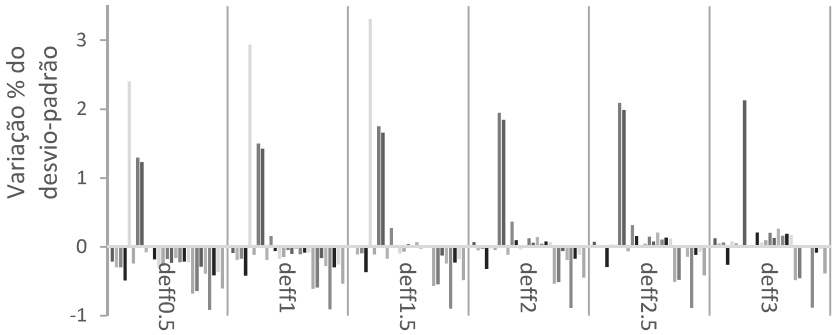

Figura 1 – Variação percentual do desvio-padrão das estimativas dos parâmetros dos modelos ajustados considerando os pesos normalizados corrigidos pelo *deff* e a do modelo ajustado com os pesos de replicação.

Para todos os valores do *deff* considerados, os coeficientes dos modelos ajustados parecem mostrar uma subestimação relativamente aos coeficientes do modelo ajustado com os pesos de replicação (Figura 2). Consequentemente, ocorre maioritariamente uma subestimação dos ORs, pois estes obtém-se recorrendo à exponencial das estimativas dos coeficientes.

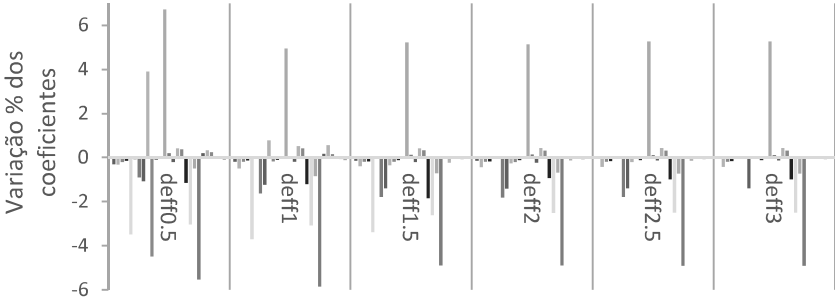

Figura 2 – Variação percentual das estimativas dos coeficientes dos modelos ajustados considerando os pesos normalizados corrigidos pelo *deff* e as do modelo ajustado com os pesos de replicação.

4.2 Abordagem 2: modelo base ajustado com os pesos de replicação

Nesta abordagem consideram-se apenas as variáveis no modelo ajustado a partir dos pesos de replicação e estimam-se os coeficientes deste modelo considerando os pesos normalizados corrigidos pelo *deff*.

Todos os modelos parecem estar ajustados corretamente aos dados, todos apresentam valores de AUC semelhantes e, tal como na abordagem anterior, o valor do coeficiente R^2 diminui com o aumento do valor do *deff* (Tabela 3).

Tabela 3 – Medidas da bondade do ajustamento e capacidade discriminativa dos modelos ajustados com base nos diferentes valores de *deff* e nos pesos de replicação.

Modelos	Hosmer	R^2	AUC	Valor p
deff = 0,5	0,45	0,64	0,93	0,05*
deff = 1	0,45	0,59	0,92	0,05*
deff = 1,5	0,45	0,57	0,91	0,05*
deff = 2	0,45	0,55	0,91	0,05*
deff = 2,5	0,45	0,54	0,91	0,05*
deff = 3	0,45	0,53	0,91	0,05*
Pesos de replicação	0,45	0,46	0,91	0,18**

* Teste de Cessie van Howelingen, ** Teste de Wald

O nível de significância para algumas variáveis pode ser bastante diferente do obtido no modelo dos pesos de replicação sendo que diferem mais quanto menor o *deff* (Tabela 4). O modelo que mostrou menos diferenças em relação ao modelo ajustado com os pesos de replicação foi o obtido quando *deff* = 2 (Tabela 4). A significância das variáveis é comparável apenas para as variáveis que não estão incluídas em interações.

Tabela 4 – Significância das variáveis, das suas interações e categorias nos modelos ajustados.

Variável ou interação	Categoria	*deff=0,5*	*deff=1*	*deff=1,5*	*deff=2*	*deff=2,5*	*deff=3*	Pesos de replicação
1	1	***	***	***	***	***	***	***
2	1	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	2	***	***	***	***	***	***	***
	3	*	.	.	n.s.	n.s.	n.s.	n.s.
3	1	***	***	***	***	**	**	**
4	1	***	***	***	***	***	***	***
5	1	***	***	***	**	**	**	**
	2	***	***	***	***	***	***	***
6	1	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
7	1	***	***	***	***	***	***	***
8		***	***	***	***	***	***	**
2*8	1	***	***	***	***	***	***	***
	2	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	3	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
7*8		***	***	***	***	***	***	***
2*5	1	**	*	*	*	*	.	*
	2	***	***	***	***	***	***	***
	3	.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	4	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	5	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	6	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
2*6	1	***	***	***	***	***	***	***
	2	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	3	**	*	*	*	.	.	n.s.
2*7	1	***	***	***	***	***	***	***
	2	***	***	***	***	***	***	***
	3	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Significativa a: · - 0,10; * - 0,05; ** - 0,01; *** - <0,001.

Verificamos que para valores de *deffs* mais baixos ocorre uma subestimação dos desvios-padrão e para valores de *deffs* mais elevados ocorre maioritariamente uma sobrestimação, embora em algumas categorias das interações ocorra uma

subestimação do desvio-padrão (Figura 3). De referir que a variável 8 é continua e por isso a célula da categoria se encontra em branco.

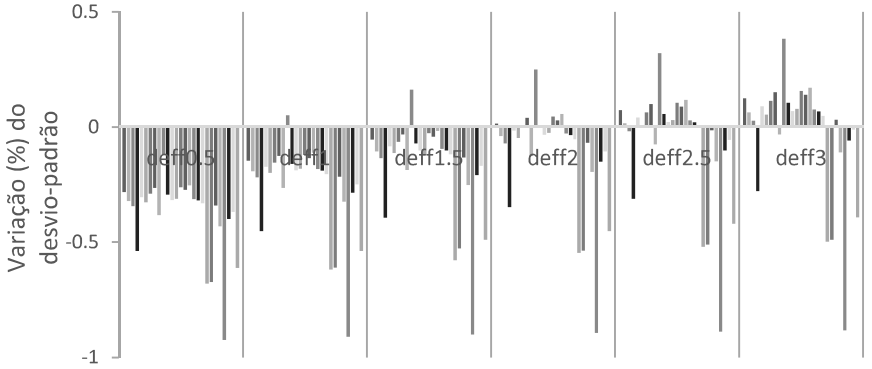

Figura 3 – Variação percentual do desvio-padrão das estimativas dos parâmetros dos modelos ajustados considerando os pesos normalizados corrigidos pelo *deff* e a do modelo ajustado com os pesos de replicação.

As estimativas pontuais dos parâmetros, bem como dos ORs, não são afetadas pelo valor dos pesos considerados, uma vez que os modelos ajustados são idênticos.

Refira-se, por fim, que para todos os modelos o agrupamento das categorias das variáveis foi sempre idêntico, concluindo-se que uma incorreta estimação do valor do *deff* não trará nenhuma influência ao agrupamento das categorias. Tal também se concluiu na primeira abordagem.

5 Considerações Finais

Neste trabalho estudou-se o efeito de vários valores de *deff* no ajustamento dos modelos de regressão logística e na significância das variáveis e suas interações, com vista a alertar para as consequências de uma incorreta estimação do valor do *deff*, bem como para a sua omissão no processo de inferência.

Os modelos resultantes podem diferir bastante consoante o valor do *deff* considerado, quer nas estimativas dos coeficientes, quer nos seus desvios-padrão. Contudo, o número de variáveis a incluir no modelo aumenta com a diminuição do valor do *deff*, havendo uma tendência para a subestimação do desvio-padrão com valores de *deff* mais reduzidos. As estimativas dos coeficientes e dos OR's são enviesadas sempre que são incluídas no modelo novas variáveis por consequência da diminuição do *deff*. Os intervalos de confiança dos ORs parecem não ser

sensíveis a uma incorreta estimação do valor do *deff,* principalmente se a estimação deste for abaixo do valor de *deff* correto. Quando se ajusta o modelo obtido com os pesos de replicação, as estimativas obtidas para os coeficientes nos modelos ajustados com pesos normalizados corrigidos pelo *deff* são idênticas.

Conclui-se, assim, que uma estimação incorreta do valor do *deff* afeta a significância das variáveis e das interações entre as variáveis, mas não parece afetar a capacidade discriminativa nem a bondade do ajustamento.

A terminar, refira-se que o investigador poderá estar perante um modelo que se parece ajustar aos dados e com uma boa capacidade discriminativa, mas que, no entanto, não explica convenientemente o evento em estudo. A discussão de resultados e a sua teorização sobre o fenómeno em estudo pode ser condicionada por uma não inclusão do efeito do delineamento (*deff*) ou por uma incorreta estimação desse efeito.

Agradecimentos

Anabela Afonso e Paulo Infante são membros do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2019), financiado pela Fundação para a Ciência e Tecnologia (FCT).

Referências

KISH, L. (1965). *Survey sampling*, John Wiley & Sons, New York.

LAVRAKAS, P. J. (2008). *Encyclopedia of survey research methods*, SAGE Publications, Thousand Oaks.

LORH, S. L. (2010). *Sampling: Design and Analysis*, Second Edition, Michelle Julet, Boston.

OSBORNE, J. W. (2011). Best practices in using large, complex samples: the importance of using appropriate weights and design effect compensation, *Practical Assessment, Research & Evaluation*, 16, 1-7.

STURGIS, P. (2004). Analysing complex survey data: clustering, stratification and weights, *Social Research Update,* 43, Autumn Issue.

HOSMER, D., LEMESHOW, S. & STURDIVANT, R. (2013). *Applied logistic regression*, 3rd Edition, Wiley, New York.