Produto & Produção, vol. 8, n. 1, p. 13-17, mar. 2005

Modelo de efeitos aleatórios para dados pareados com observações omissas: um exemplo na área de avaliação de gestão ambiental

Mayra Ivanoff Lora
Depto. Eng. Produção – EPUSP

Linda Lee Ho
Depto. Eng. Produção – EPUSP

Julio da Motta Singer
Depto. Estatística – IMEUSP

Apresentamos um modelo apropriado para a análise estatística de dados pareados que pode ser encarado como uma generalização do teste t pareado para a comparação de médias. Ele é suficientemente geral a ponto de permitir a inclusão de covariáveis e respostas omissas. Ilustramos o procedimento com a análise da evolução da gestão ambiental de empresas basileiras baseada em questionários de auto-avaliação respondido por empresas em 1996 e 2001.

Palavras-chave: dados longitudinais; gestão ambiental; respostas omissas; modelo de efeitos aleatórios.

In this paper, we present an appropriate model to analyse paired data. It can be viewed as a generalized paired t-test to compare means since it allows the inclusion of covariables and missing data. This proceeding is illustrated by an example which analyzes the evolution of Brazilian enterprises environmental management, based in self-evaluation questionnaires, which have been answered by the enterprises in two different times: 1996 and 2001.

Keywords: longitudinal data; environmental management; missing data; random effect models.

1 Introdução

Em muitos casos, o objetivo de um estudo é analisar a evolução de determinadas características ao longo do tempo na mesma unidade experimental. Dados com essa estrutura são conhecidos como dados longitudinais. O caso mais simples é aquele conhecido sob a denominação de dados pareados, em que cada unidade de investigação é avaliada em dois instantes. Em geral, para respostas com distribuição gaussiana, esse tipo de dados é analisado por intermédio do conhecido teste t pareado. A aplicação desse teste requer informações completas nos dois instantes da coleta de dados em todas unidades amostrais, condição que muitas vezes não é satisfeita em situações práticas.

Um exemplo deste tipo de problema é um estudo realizado pela ERM Brasil para avaliar a evolução da gestão ambiental das empresas brasileiras entre os anos de 1996 e 2001 por intermédio de um questionário. A primeira parte da pesquisa foi realizada em 1996, por meio de encartes no jornal Gazeta Mercantil, com um total de 311 respostas recebidas. Com o objetivo de verificar a evolução de tal perfil nos últimos 5 anos, a pesquisa foi reeditada, com o mesmo questionário disponível em endereço eletrônico. As 311 empresas que responderam à primeira sondagem tomaram conhecimento da segunda por meio de mensagens enviadas eletronicamente, por carta ou por telefonema, tendo sido recebidas 69 respostas, 18 das quais por empresas que haviam respondido ao primeiro questionário.

O questionário elaborado pela ERM Brasil é composto de 20 perguntas de múltipla escolha, cujas opções correspondiam a uma escala ordinal de 1 a 5, de forma que quanto maior a nota atribuída, melhor o desempenho da empresa no item avaliado por determinada questão. Essas 20 questões referem-se às cinco etapas de desenvolvimento de um Sistema de Gestão Ambiental, segundo a norma ISO 14001, que são:

- Definição da Política de Meio Ambiente: declaração formal da empresa de quais são os seus princípios e intenções em relação a seu desempenho ambiental, incluindo compromisso com a melhoria contínua, a prevenção da poluição e a satisfação aos requisitos regulamentados ou assumidos pela empresa.

- Elaboração do Plano de Ação: determina quais atividades, produtos e serviços do Sistema de Gestão Ambiental devem ser planejados. Para tanto, deve-se: conhecer a Política Ambiental da organização, identificar aspectos ambientais significativos e conhecer requisitos e regulamentações legais relativos a esses aspectos. A partir desses dados são estabelecidos objetivos e metas ambientais (com seus indicadores) que gerarão, então, o programa do Sistema de Gestão Ambiental.
- Implementação do Sistema: compreende a criação do sistema para suportar a implementação dos planos criados na fase anterior. Para tanto, devem ser: definidos os recursos necessários, atribuídas responsabilidades, realizados treinamentos, desenvolvidos sistemas de fluxo de informação na empresa e estabelecidos e mantidos controles para todos os processos e procedimentos para casos de emergência.
- Avaliação Periódica: esta é a fase de verificação, na qual devem-se coletar os dados do processo e então tomar ações corretivas e preventivas, quando necessário. Tem como principal objetivo verificar a eficiência do desempenho ambiental alcançado.
- Revisão do Sistema: a partir de novos fatores, sejam estes gerados por mudanças realizadas em função de não conformidades ou por causas externas, o Sistema de Gestão Ambiental deve ser revisto, corrigindo ou redefinindo a Política Ambiental, as metas e os objetivos.

Além das respostas às vinte perguntas, foram coletadas, como candidatas para explicar o desempenho ambiental, as seguintes informações das empresas: região geográfica (Sudeste ou outras regiões), período de início de atividades (até 1980 ou a partir de 1981), setor de atividade econômica (indústria química ou outros setores), origem de capital (nacional ou misto e estrangeiro) e número de funcionários (até 99 ou 100 ou mais). No texto elas serão referenciadas como covariáveis. O objetivo do estudo é identificar que covariáveis influenciaram as distribuições de respostas assim como verificar se houve uma evolução no desempenho dessas empresas nos últimos 5 anos.

Mais especificamente, o objetivo é estudar o efeito desses fatores na comparação das médias de três variáveis respostas construídas a partir das notas atribuídas às vinte questões originais. A primeira delas está associada ao grau de definição e elaboração do plano e corresponde à média aritmética das notas atribuídas às questões 1 a 4; a segunda, ao grau de implantação do sistema por meio da média artimética das notas atribuídas às questões 5 a 17 e a terceira resposta se relaciona com a avaliação e revisão do sistema, correspondendo à média aritmética das notas atribuídas às questões 18 a 20.

Na Seção 2 apresentamos detalhes técnicos do modelo de efeitos aletórios empregado na análise, e na Seção 3 discutimos os resultados apresentados, comparando-os com os obtidos através de um modelo semelhante usando apenas as 18 observações completas, porém que permite a comparação das outras covariáveis.

2 Modelos de efeitos aleatórios para análise de dados pareados

O modelo de efeitos aleatórios pode ser escrito como

$$Y_i = X_i\beta + X_i b_i + \varepsilon_i$$

onde:

Y_i é o vetor de respostas da i-ésima unidade experimental, que contém as observações dos t_i instantes, sendo Y_{ik} a observação realizada no k-ésimo instante na i-ésima unidade amostral; i = 1, ..., n e k = 1, ..., t_i;

X_i é a matriz de dimensão t_i x (p+1) de variáveis explicativas, ou covariáveis, cujas colunas contêm os valores dessas variáveis para cada um dos p parâmetros;

β é o vetor dos p parâmetros do modelo;

$b_i \sim N(0, \Sigma_b)$ para i = 1, ..., n é o vetor de efeitos aleatórios associados à i-ésima unidade experimental;

$\varepsilon_i \sim N(0, \sigma^2 I_{t_i})$ para i = 1, ..., n é o vetor de erros casuais associados à i-ésima unidade experimental;

sendo os dois últimos vetores aleatórios independentes.

Para o conjunto de dados deste trabalho, i = 1, ..., 362; k = 1, 2; ou seja, $t_i = 2$, $\forall i$; onde k = 1 representa as observações feitas em 1996 e k = 2 as feitas em 2001;

Como conseqüência temos os seguintes resultados:

$$E(Y_i) = X_i\beta$$
$$Var(Y_i) = X_i \Sigma_b X_i^T + \sigma^2 I_{t_i} = \Sigma_i(\theta)$$
$$Y_i \sim N(X_i\beta, \Sigma_i)$$

onde β é um vetor de parâmetros que define a estrutura de covariância. Esses resultados implicam que as covariâncias entre as observações na mesma unidade experimental podem não ser nulas. Esse modelo incorpora uma possível correlação entre as respostas da mesma empresa nos dois momentos da pesquisa.

Explicitando os componentes do modelo linear utilizado para descrever as respostas das empresas temos que:

$$Y_{ij} = X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + e_{ij} \quad (1)$$

onde:

X_0 = 1: para a constante

X_1 = { 0: para medidas realizadas em 1996; 1: para medidas realizadas em 2001

X_2 = { -1: para medidas realizadas em empresas da região sudeste; 1: para medidas realizadas em empresas de outras regiões

X_3 = { -1: para medidas realizadas em empresas da indústria química; 1: para medidas realizadas em empresas de outros setores de atividades

X_4 = { -1: para medidas realizadas em empresas com até 99 funcionários; 1: para medidas realizadas em empresas com 100 ou mais funcionários

X_5 = { -1: para medidas de empresas com início de atividades até 1980; 1: para medidas de empresas com início de atividades a partir de 1981

X_6 = { -1: para medidas realizadas em empresas de capital nacional; 1: para medidas realizadas em empresas de capital misto ou estrangeiro

e os respectivos componentes do vetor de parâmetros β podem ser interpretados como:

β_1: diferença entre as médias das respostas das pesquisas realizadas em 2001 e 1996;
β_2: contribuição da covariável região geográfica para a média das respostas;
β_3: contribuição da covariável setor de atividades para a média das respostas;
β_4: contribuição da covariável número de funcionários para a média das respostas;
β_5: contribuição da covariável início de atividades para a média das respostas;
β_6: contribuição da covariável tipo de capital para a média das respostas.

Para estimar β e Σ_i de maneira simultânea, métodos interativos, como o de Newton-Raphson, Scoring de Fisher ou EM (ver ANDREONI, 1989) são utilizados. As interações do método são interrompidas quando, segundo algum critério, a diferença entre as estimativas de dois passos consecutivos atinge um valor predeterminado.

Obtidas as estimativas de β, há interesse em testar as seguintes hipóteses:

1) $\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 > 0 \end{cases}$ para testar se o desempenho ambiental das empresas respondentes em 2001 foi melhor do que em 1996;

2) $\begin{cases} H_0 : \beta_p = 0 \\ H_1 : \beta_p \neq 0 \end{cases}$ para testar se existe influência do p-ésimo parâmetro na resposta $(2 \leq p \leq 6)$.

A estrutura da matriz de covariância usada para o modelo é a simétrica composta:

$$\sum_i = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 + \sigma_1^2 \end{bmatrix}$$

que implica que as variâncias das observações realizadas em 1996 e 2001 são iguais.

3 Resultados e conclusões

Na Tabela 1 estão resumidas as respostas utilizadas com as respectivas questões e etapas de desenvolvimento do Sistema de Gestão Ambiental correspondentes. A utilização das médias aritméticas como variáveis respostas dá mais suporte à suposição de normalidade do modelo utilizado.

O modelo descrito foi implementado utilizando o procedimento Proc Mixed do pacote estatístico *Statistical*

Tabela 1 – Definição das variáveis respostas

Etapas do desenvolvimento de um Sistema de Gestão Ambiental avaliadas	Questões
Definições iniciais e elaboração do plano de ação (etapas 1 e 2)	1 a 4
Implementação do sistema (etapa 3)	5 a 17
Avaliação e revisão do sistema (etapas 4 e 5)	18 a 20

Analysis System (SAS), que possibilita a análise de dados não balanceados em relação ao tempo, *i.e.* em que o número de observações nos dois períodos não é igual.

Testes das hipóteses de interesse sob o modelo completo (1) sugerem a rejeição das hipóteses nulas $\beta_p = 0$, p = 1, 4, e 6 para as três variáveis respostas. Ou seja, as covariáveis ano da pesquisa (β_1), número de funcionários (β_4) e tipo de capital (β_6) são significativas. Em função deste resultado um modelo mais simples, porém que expressa uma possível interação entre número de funcionários e tipo de capital, expresso em (2) foi avaliado.

$$Y_{ij} = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_6 X_6 + \beta_{46} X_4 X_6 \qquad (2)$$

Como a hipótese nula H_0: $\beta_{46} = 0$ não é rejeitada para as três variáveis respostas, o modelo expresso em

$$Y_{ij} = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_6 X_6 \qquad (3)$$

foi considerado como modelo final. Estimativas (est.) dos parâmetros significativos com os respectivos desvios padrão (dp) e níveis descritivos (p) estão resumidos na Tabela 2. Além dos parâmetros do modelo, a estimativa da covariância com seu respectivo erro padrão e nível descritivo também estão dispostos na Tabela 2.

Os níveis descritivos dados pelo teste da razão de verossimilhanças (LRT) dispostos na Tabela 2 indicam a adequação do modelo final.

Analisando os resultados obtidos, pode-se notar que houve uma melhora na auto-avaliação das empresas respondentes da pesquisa de 2001 em relação à de 1996, além de existir influência tanto do número de funcionários quanto do tipo de capital na resposta das empresas. Com a estimativa obtida para o parâmetro β_4, pode-se dizer que o desempenho ambiental das empresas com 100 ou mais funcionários é melhor do que o daqueles com até 99 funcionários para qualquer uma das 3 variáveis respostas analisadas; e a estimativa do parâmetro β_6 mostra que, para qualquer uma das variáveis respostas, o desempenho das empresas de capital nacional foi pior do que aquele de empresas com capital misto ou estrangeiro.

A análise usual dos dados, seria baseada apenas nas 18 observações (total de empresas que responderam às duas pesquisas) e em conseqüência não permitiria a inclusão de todas as covariáveis simultaneamente. Essa análise é baseada em um modelo semelhante ao proposto inicialmente, considerando apenas as 18 respostas completas e apenas uma covariável. Mais especificamente, o modelo teria a forma (1) com uma única covariável.

Os níveis descritivos (p) para as 5 análises realizadas dessa maneira estão resumidos na Tabela 3.

Os resultados obtidos permitiram detectar diferenças significantes entre as respostas médias obtidas nos dois anos da pesquisa, para as três variáveis respostas. No entanto, ao contrário da análise anterior, apenas em alguns casos foi possível detectar um efeito significativo das covariáveis.

O modelo completo, além de permitir o uso de todas as observações (o que pode reduzir o viés), permite incluir as cinco covariáveis simultaneamente e também avaliar possíveis interações entre elas.

Tabela 2 – Estimativas obtidas para as covariáveis significativas

Parâmetro		Etapas 1 e 2			Etapa 3			Etapas 4 e 5		
		est.	d.p.	p	est.	d.p.	p	est.	d.p.	p
β_0	Intercepto	3,48	0,08	0,0001	3,21	0,08	0,0001	2,86	0,10	0,0001
β_1	Ano da pesquisa	0,75	0,14	0,0001	0,86	0,13	0,0001	0,92	0,16	0,0001
β_4	Nº de funcionários	0,17	0,06	0,0057	0,21	0,06	0,0002	0,24	0,07	0,0015
β_6	Tipo de capital	0,18	0,08	0,0224	0,18	0,07	0,0131	0,19	0,09	0,0452
σ_1^2	Covariância Modelo (LRT)	0,51	0,25	0,0426 0,1601	0,33	0,22	0,1390 0,2268	0,74	0,31	0,0159 0,0686

Tabela 3 – Níveis descritivos para os 5 modelos baseados em apenas 18 observações

	Etapas 1 e 2	Etapa 3	Etapas 4 e 5
Ano pesquisa	0,001	0,002	0,004
Região geográfica	0,370	0,404	0,772
Ano pesquisa	0,001	0,002	0,001
Setor de atividades	0,258	0,148	0,345
Ano pesquisa	0,003	0,008	0,017
Nº funcionários	0,092	0,088	0,019
Ano pesquisa	0,001	0,004	0,009
Início de atividades	0,932	0,853	0,778
Ano pesquisa	0,001	0,001	0,001
Tipo de capital	0,168	0,023	0,112

Referências

ANDREONI, S. *Modelos de Efeitos Aleatórios para Análise de Dados Longitudinais não Balanceados em Relação ao Tempo.* São Paulo, 1989. Dissertação (Mestrado em Estatística) – Departamento de Estatística do Instituto de Matemática e Estatística, USP.

BANAS AMBIENTAL. São Paulo, ano II, n. 12, jun. 2001. Suplemento especial.

BANAS QUALIDADE. São Paulo, ano X, n. 108, maio 2001.

ELIAN, S. N.; OKAZE, S. M. Relatório de análise estatística sobre o projeto: *Mudanças de desempenho em atividades motoras da vida diária em idosos participantes de um programa de educação física*. São Paulo, IME-USP, 1998. (RAE – CEA – 98P07).

GAZETA MERCANTIL. São Paulo. Gestão Ambiental – Compromisso da Empresa. Vol. 1 a 8. 20.03.1996 a 08.05.1996. Suplemento especial.

GILBERT, M. J. *ISO 14001/BS7750:* Sistemas de Gerenciamento Ambiental. São Paulo: IMAM, 1995.

HOJDA, R. G. *ISO 14001 – Sistemas de Gestão Ambiental.* São Paulo, 1997. Dissertação (Mestrado em Engenharia de Produção) – Departamento de Engenharia de Produção da Escola Politécnica, USP.

LEVY, P. S.; LEMESHOW, S. *Sampling on Populations:* Methods and Applications. 3. ed. USA: Wiley Inter Science, 1999.

NETER, J. *et al*. *Applied Linear Statistical Models.* 4. ed. New York: McGraw-Hill, 1996.

SAS TECHNICAL REPORT P-229 SAS/STAT SOFTWARE: Changes and Enhancements, Release 6.07. USA, 1992. Manual de instruções.

SINGER, E. da M. *et al*. *Perfil de Gestão Ambiental:* Programa: Gestão Ambiental – Compromisso da Empresa. São Paulo, 1996. Relatório de pesquisa.

SINGER, J. M. Análise estatística de dados longitudinais. *Textos para discussão – As inovações nas informações sociais e econômicas*, v. 3, ST 110. Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais. Rio de Janeiro: IBGE, 1996. ■