Aplicação da **regressão logística ordinal** em estudos de **lealdade de clientes**. Evidência para a **indústria hoteleira** no Algarve

ISABEL MARIA DA SILVA JOÃO * [ijoao@deq.isel.ipl.pt]

Resumo | Os modelos de regressão logística ordinal têm vindo a ser utilizados na análise de dados cuja variável de resposta se apresenta em categorias com ordenação. Em estudos de satisfação de clientes é usual medir a satisfação dos mesmos, ou a recomendação do serviço a terceiros, com variáveis de natureza ordinal. Este artigo apresenta uma descrição dos modelos de regressão logística ordinal mais usados, o modelo de *odds* proporcionais e o modelo de *odds* proporcionais parciais. Apesar da existência de inúmeros estudos sobre satisfação de clientes e lealdade de clientes na indústria hoteleira, a utilização deste tipo de modelos na área é ainda reduzida. Apresentam-se os aspetos teóricos dos principais modelos de regressão logística ordinal e explicam-se as condições de utilização. A adoção deste tipo de modelos de regressão ordinal é ilustrada com um estudo empírico utilizando um conjunto de respostas de clientes de uma unidade hoteleira da região do Algarve. Foi selecionada a opção de modelação dos dados que se mostrou adequada. Como principais conclusões, de referir que quer a idade dos clientes quer a duração da estadia têm influência na sua intenção de recomendar o serviço.

Palavras-chave | Lealdade dos clientes, Modelos de regressão logística ordinal, Indústria hoteleira.

Abstract | Ordinal logistic regression models have come to be used in the analysis of data whose response variable is of ordinal type meaning that the categories of the variable can be ranked but the distances between categories are unknown. In studies of customers' satisfaction it is usual to measure the satisfaction of the customers or the customers´ recommendation of the service by means of ordinal scales. This paper presents a description of the proportional *odds* model and the partial proportional *odds* model. Although the existence of quite a lot studies on customers´ satisfaction and loyalty of customers in the hotel industry, the use of this type of models in this area is still reduced. The theoretical framework of the main ordinal regression models is presented, as well as their utilization conditions. The use of this sort of models is illustrated with an empirical study using a set of customer's answers of a hotel unit in the Algarve region. It was selected the ordinal regression model that reveal adequate to model the data. As main conclusions to relate is that the age of the customers and the duration of the stay have influence in customers´ intention to recommend the service.

Keywords | Customers' loyalty, Ordinal logistic regression models, Hotel industry.

* **Doutorada em Engenharia e Gestão Industrial** pelo Instituto Superior Técnico, Universidade Técnica de Lisboa. **Investigadora** do CEG-IST, UTL. **Professora Adjunta** do Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa.

1. Introdução

A satisfação dos clientes constitui atualmente uma das maiores prioridades da gestão, fundamentalmente nas empresas que estão empenhadas em alcançar elevados padrões de qualidade dos seus produtos e serviços e maximizar os resultados alcançados junto dos seus clientes (Kotler e Keller, 2006). Danaher e Haddrell (1996) fazem uma revisão dos principais tipos de escalas usados na medição da satisfação de clientes e agrupam-nas, fundamentalmente, em performance, não confirmação de expectativas e satisfação, verificando-se que os dados de natureza ordinal são muito utilizados por investigadores na área dos estudos de qualidade e satisfação dos clientes.

Oh e Parks (1997) efetuaram uma revisão critica sobre satisfação de clientes e qualidade do serviço no setor do alojamento turístico e concluíram que tipicamente a satisfação é medida em escalas de sete, cinco, três e mesmo dois níveis "satisfeito/não satisfeito". Uma escala de dois níveis não permite aferir diferentes graus de satisfação o que limita a sua utilidade (Oliver, 1981). Os níveis da escala de satisfação encontram-se ordenados de acordo com um conjunto de categorias ordinais, onde a atribuição de números às sucessivas categorias da variável a medir representa diferenças em magnitude do tipo "melhor que" ou "pior que" (Stevens, 1946). Numa escala de cinco níveis variando desde "muito insatisfeito" a "muito satisfeito" a atribuição de números de 1 a 5 às várias categorias é feita de forma a preservar a transitividade das categorias: se o valor 3 representa um estado de maior satisfação que o valor 2, e o valor 2 representa um estado de maior satisfação que o valor 1, então a propriedade de transitividade implica que a condição representada pelo valor 3 é melhor que a condição representada pelo valor 1 (Cliff e Keats, 2003, Krantz *et al.*, 1971).

Desde que Stevens (1946) propôs a classificação das escalas de mensuração em nominal, ordinal, intervalar e de razão, estabelecendo uma hierarquia de acordo com o tipo de transformações matemáticas que cada nível de mensuração admite, que tem havido grande controvérsia quanto aos tipos de tratamentos estatísticos admissíveis para cada nível de mensuração (Knapp, 1990). Apesar dos dados ordinais serem simples, o seu tratamento estatístico continua objeto de desafio para muitos investigadores (Cliff, 1996; O´Connell, 2006; Hosmer e Lameshow, 2000). O´Connell (2006) refere que historicamente os investigadores têm usado duas abordagens muito diferentes para a análise de dados ordinais, sendo que alguns decidem aplicar modelos paramétricos para analisar dados ordinais, através de modelos de regressão linear múltipla tratando a variável de resposta ordinal como sendo pelo menos uma variável num nível de mensuração intervalar, e outros decidem tratar a variável ordinal como estritamente categórica e aplicar abordagens não paramétricas para melhor compreender os dados. Apesar de ambas as estratégias serem informativas, dependendo da questão em estudo, nenhuma das abordagens é ótima se o objetivo consistir em desenvolver modelos explanatórios para resultados ordinais (Agresti, 1989; Cliff, 1996).

As variáveis ordinais são muitas vezes codificadas com números inteiros consecutivos desde 1 até ao total de categorias. Talvez como resultado deste tipo de codificação exista a tendência de analisar este tipo de variáveis de resposta ordinal recorrendo aos modelos de regressão linear (Long e Freese, 2001). No entanto, com a variável dependente ordinal violam-se os pressupostos do modelo de regressão linear o que pode levar a conclusões incorretas, tal como demonstrado por McKelvy e Zavoina (1975). Desta forma, em modelos de regressão cuja variável dependente é ordinal é melhor utilizar modelos que não consideram o pressuposto de que as distâncias entre as categorias são iguais, uma vez que essa noção de distância é característica de um nível de mensuração intervalar. Por exemplo, numa escala de satisfação de 5 níveis é comum atribuir os números {1,…,5} às categorias {"muito insatisfeito", "insatisfeito",…, "muito satisfeito"}.

Contudo, a métrica que está subjacente à satisfação não é necessariamente a mesma que a métrica linear que relaciona os números de 1 até 5 (i.e. reta real). Em termos substantivos, a diferença entre 1 e 2, "muito insatisfeito" e "insatisfeito" na codificação usada, pode ser diferente da diferença entre 4 e 5, "satisfeito" e "muito satisfeito".

O´Connell (2006) discute um conjunto de métodos que constituem exemplos de modelos de regressão ordinal e são extensões aos modelos de regressão logística para dados cuja resposta é binária. Apesar de existir uma grande diversidade de modelos, a sua aplicação no setor do turismo, e em particular na área de estudos da satisfação de clientes, tem sido pouco explorada o que se pode atribuir à sua complexidade e também às diferenças verificadas no tipo de modelação usada por diferentes tipos de *softwares* comerciais como o SPSS, SAS, ou Stata que usam diferentes técnicas para estimar os modelos de regressão logística ordinal.[1]

Palmer *et al.* (2005) fizeram uma revisão bibliográfica sobre a utilização de ferramentas estatísticas na investigação em turismo e concluíram que os modelos de regressão logística foram utilizados em apenas 3,2% dos estudos. Sendo os estudos de satisfação e lealdade de clientes um subgrupo da investigação que se faz no setor do turismo, pode-se concluir que os modelos de regressão logística constituem um vasto domínio a explorar nesta área.

Neste trabalho apresenta-se uma revisão sumária dos dois principais modelos de regressão logística ordinal, e avalia-se a sua utilidade em estudos de satisfação de clientes na indústria hoteleira.

Na secção 2 apresentam-se os aspetos teóricos dos principais modelos de regressão logística ordinal. Os modelos ordinais considerados neste artigo incluem o modelo de *odds* proporcionais, e o modelo de *odds* proporcionais parciais (não restrito e restrito)[2]

Na secção 3 apresenta-se o problema em análise que consiste em identificar a forma como as características dos clientes determinam a sua lealdade ao serviço prestado. Os dados a utilizar neste trabalho foram retirados de um estudo sobre avaliação da satisfação de clientes na indústria hoteleira realizado em Portugal no Algarve (João, 2009).

O caso escolhido tem como principal objetivo permitir evidenciar de que forma os modelos de regressão logística ordinal podem ser usados nos estudos de satisfação e lealdade de clientes, e mostrar que a sua utilização é simples e que os resultados que se podem obter da análise dos modelos são de grande utilidade para a gestão hoteleira.

2. Modelos de regressão ordinal

Um modelo de regressão ordinal pode ser desenvolvido de várias formas distintas. A abordagem a usar para um modelo de regressão logística ordinal (RLO) é muito semelhante à do modelo de regressão logística binária (RLB). De facto o modelo RLB pode ser visto como um caso especial do modelo ordinal para o qual a variável de resposta somente possui duas categorias.

A RLO pode ser expressa como um modelo de variável latente (Agresti, 2002; Long e Freese, 2001). Considerando a existência de uma variável latente, Y^*, pode-se definir $Y^* = \underline{x}\beta + \varepsilon$ sendo $\underline{x}$ um vetor linha (1^*k) e β um vetor coluna (k^*1) de coeficientes estruturais, e ε corresponde a uma perturbação aleatória com uma distribuição normal reduzida, i.e. $\varepsilon \sim N(0,1)$.

[1] Não é objetivo deste trabalho efetuar sugestões sobre o *software* a usar pois a sua escolha dependerá, fundamentalmente, das preferências dos investigadores, mas fica o alerta de que o software a usar constitui a principal ferramenta de trabalho do investigador e é importante compreender as diferenças de parametrização dos modelos de regressão logística ordinal consoante o pacote de software estatístico que se está a utilizar.

[2] Neste trabalho usou-se o *software* Stata pelo que a formulação dos modelos de regressão aqui apresentados serão consistentes com a abordagem usada nesse software.

Assumindo que a variável latente Y^* é definida como função do conjunto de variáveis explicativas e do erro aleatório, pode-se considerar que esta variável pode tomar um conjunto infinito de valores, os quais podem ser colapsados num conjunto de categorias da variável de resposta Y. A variável latente Y^* vai ter vários pontos de corte (limites): $\alpha_1, \alpha_2, \ldots, \alpha_j$ e o valor da variável observada y estará dentro das regiões definidas por esses pontos de corte. Considere-se, a título de exemplo, a variável ordinal dependente "recomendação do serviço", consistindo num conjunto de cinco categorias ou cinco níveis e variando de "decerto recomenda" a "decerto não recomenda".

Sendo o nível de recomendação do serviço a resposta ordinal, y, variando de 1 a 5, onde 1 = "decerto recomenda", 2 = "provavelmente sim", 3 = "talvez sim/talvez não", 4 = "provavelmente não" e 5 = "decerto não" define-se os pontos de corte de tal forma que $\alpha_1 < \alpha_2 < \alpha_3 .. < \alpha_4$.

$$Y = \begin{cases} 1 \text{ se } y^* \leq \alpha_1 \\ 2 \text{ se } \alpha_1 \leq y^* \leq \alpha_2 \\ 3 \text{ se } \alpha_2 \leq y^* \leq \alpha_3 \\ 4 \text{ se } \alpha_3 \leq y^* \leq \alpha_4 \\ 5 \text{ se } \alpha_4 \leq y^* \leq \infty \end{cases} \quad (1)$$

Pode-se calcular a probabilidade para cada nível de recomendação do serviço. Por exemplo:

$$\begin{aligned} \pi_1 &= P(y=1) = P(y^* \leq \alpha_1) = P(x\beta + \varepsilon \leq \alpha_1) = F(\alpha_1 - x\beta) \\ \pi_2 &= P(y=2) = P(\alpha_1 \leq y^* \leq \alpha_2) = F(\alpha_2 - x\beta) - F(\alpha_1 - x\beta) \\ \pi_3 &= P(y=3) = P(\alpha_2 \leq y^* \leq \alpha_3) = F(\alpha_3 - x\beta) - F(\alpha_2 - x\beta) \\ \pi_4 &= P(y=4) = P(\alpha_3 \leq y^* \leq \alpha_4) = F(\alpha_4 - x\beta) - F(\alpha_3 - x\beta) \\ \pi_5 &= P(y=5) = P(\alpha_4 \leq y^* \leq \infty) = 1 - F(\alpha_4 - x\beta) \end{aligned} \quad (2)$$

Seja Y a variável de resposta com k categorias codificadas de 1, 2, 3, …, k, e o vetor de variáveis explicativas definido por $\underline{x} = (x_1, x_2, \ldots, x_p)$. As k categorias da variável de resposta Y tendo em conta as covariáveis consideradas ocorrem com probabilidades $\pi_1, \pi_2, \ldots, \pi_k$ ou seja $\pi_j = P(Y = j)$, para $j = 1, 2, \ldots, k$.

Também se pode calcular as probabilidades cumulativas usando a fórmula:

$$P(Y \leq j) = F(\alpha_j - x\beta), \text{ com } j = 1, 2, \ldots, J-1 \quad (3)$$

No modelo de regressão logística binária a variável de resposta tem dois níveis considerando-se usualmente 1 = sucesso do evento, 0 = falha do evento. Pode-se prever a probabilidade de sucesso para um conjunto de variáveis explicativas. O modelo de regressão logística pode ser expresso por:

$$\ln(Y') = \text{logit}\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) = \alpha - \beta_1 X_1 - \beta_2 X_2 - \ldots, \beta_p X_p \quad (4)$$

2.1. Modelo de *odds* proporcionais (OP)

O modelo de *logit* cumulativo foi originalmente proposto por Walker e Duncan (1967) e mais tarde chamado de modelo de *odds* proporcionais por McCullagh (1980). A dependência de Y sobre $\underline{x} = (x_1, x_2, \ldots, x_p)$, para o modelo de *odds* proporcionais, pode ser representado da seguinte forma:

$$\left[\pi(Y \leq j \,|\, x_1, x_2, \ldots, x_p)\right] = \left[\frac{e^{\alpha_j + (-\beta_1 X_1 - \beta_2 X_2, \ldots, -\beta_p X_p)}}{1 + e^{\alpha_j + (-\beta_1 X_1 - \beta_2 X_2, \ldots, -\beta_p X_p)}}\right] = \frac{e^{(\alpha_j - \beta X)}}{1 + e^{(\alpha_j - \beta X)}} \text{ com } j = 1, 2, \ldots k \quad (5)$$

No modelo de *odds* proporcionais consideram-se $(k - 1)$ pontos de corte das categorias, sendo que o j-ésimo ponto de corte é baseado na comparação de probabilidades acumuladas. No modelo de *odds* proporcionais trabalha-se com o *logit*, ou seja, com o logaritmo natural dos *odds*. Para estimar o *ln*(*odds*) de estar numa dada categoria ou abaixo dela o modelo de *odds* proporcionais pode ser escrito na seguinte forma:

$$\text{logit}\left[\pi(\underset{\rightarrow}{x})\right]=\ln\left(\frac{\pi_j(\underset{\rightarrow}{x})}{1-\pi_j(\underset{\rightarrow}{x})}\right)=\text{logit}\left[\pi(Y\le j\left|x_1,x_2,...,x_p\right.)\right]=$$

$$=\ln\left[\frac{\pi(Y\le j\left|x_1,x_2,...,x_p\right.}{\pi(Y>j\left|x_1,x_2,...,x_p\right.}\right]=$$

$$=\alpha_j+(-\beta_1X_1-\beta_2X_2,...,-\beta_pX_p) \qquad (6)$$

Onde $\pi_j(\underset{\rightarrow}{x})=\pi(Y\le j\left|x_1,x_2,...,x_p\right.)$ representa a probabilidade de estar na categoria *j* ou abaixo dela, distribuição cumulativa de probabilidades, dado o conjunto de variáveis explicativas considerado.

Pode parecer confuso o facto de o modelo subtrair βX em vez de adicionar. Isto resulta do facto de se calcular a probabilidade de $y \le j$ em vez de $y > j$.

Os α_j são os parâmetros desconhecidos de interseção que satisfazem a condição $\alpha_1 \le \alpha_2 \le \alpha_3 \le \ldots \le \alpha_k$, e $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$ é o vetor dos coeficientes de regressão desconhecidos correspondentes a $x = (x_1, x_2, \ldots, x_p)$.

Por transformação dos *logit* cumulativos podemos obter os *odds* cumulativos assim como as probabilidades cumulativas de estar na categoria *j* ou abaixo dela. Com base no ajuste do modelo a razão de *odds* cumulativos ψ_l para a covariável binária de ordem *l*, representada por x_l pode ser obtida da seguinte forma:

$$\psi_{OP}=\frac{\pi(Y\le y_j\left|x_l^{(1)}\right.)}{\pi(Y\le y_j\left|x_l^{(0)}\right.)}=e^{\left\{-\beta_l(x_l^{(1)}-x_l^{(0)}\right\}} \qquad (7)$$

O modelo assenta no pressuposto de *odds* proporcionais acerca dos $(k - 1)$ pontos de corte. O pressuposto de *odds* proporcionais também é conhecido por pressuposto de regressão paralela que é assumida para cada covariável incluída no modelo. De notar que o vetor dos coeficientes de regressão, β, não depende de *j* o que implica que o modelo assume que a relação entre x_j e *Y* é independente de *j*. McCullagh (1980) denomina este pressuposto de *odds* proporcionais para os $(k - 1)$ pontos de corte, também chamado de pressuposto de regressão paralela, que é assumida para cada covariável incluída no modelo, e daí o nome do modelo, pois assume-se que a razão de *odds* de qualquer variável explicativa é constante ao longo de todas as categorias ordenadas.

Pressuposto de regressão paralela

O pressuposto de regressão paralela pode ser testado através da comparação das estimativas que se obtêm das $(k - 1)$ regressões binárias.

$$P(y \le j \,|\, x) = F(\alpha_j - x\beta_j), \text{ com } j = 1,2,\ldots, k-1 \quad (8)$$

O pressuposto de regressão paralela pode ser testado comparando as estimativas das $(k - 1)$ regressões binárias onde se permite que os β'*s* difiram para as várias opções. O pressuposto da regressão paralela implica que $\beta_1 = \beta_2 = \ldots = \beta_{k-1}$. Tendo em conta o grau pelo qual o pressuposto se verifica os coeficientes $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{k-1}$ devem ser "próximos".

O pressuposto de *odds* proporcionais pode ser testado utilizando o teste de Wald, que foi desenvolvido por Brant (1990), ou de forma alternativa pelo teste da razão de verosimilhanças (RV), que foi desenvolvido por Wolfe e Gould (1998). O teste da razão de verosimilhanças prova se os coeficientes de todas as variáveis são iguais. Com este teste não se pode determinar se existem coeficientes que para algumas variáveis são idênticos para as várias equações binárias, enquanto os coeficientes de outras variáveis possam diferir. Para tal, o teste de Wald, proposto por Brant (1990), tem grande utilidade pois permite testar o pressuposto de regressão paralela para cada variável individualmente. O teste identifica as variáveis que violam o pressuposto e clarifica como é que esses pressupostos são violados através da análise dos coeficientes estimados para as *j*-1 regressões binárias efetuadas. Pode-se, desta forma, efetuar um teste global onde se verifica se alguma variável viola o pressuposto e também testar

o pressuposto de linhas paralelas para cada variável a título individual.

Muitas vezes, mesmo quando os pressupostos do modelo de *odds* proporcionais são violados, é prática comum ignorar a violação e seguir com o modelo de *odds* proporcionais, o que pode, contudo, levar a resultados incorretos (Williams, 2006).

2.2. Modelo de *odds* proporcionais parciais (OPP)

O modelo de *odds* proporcionais (McCullagh, 1980) pode ser modificado de forma a permitir a existência de *odds* não proporcionais para um conjunto de variáveis explicativas. Como existem muitas situações para as quais o pressuposto de *odds* proporcionais não se verifica, tal levou ao desenvolvimento do modelo de *odds* proporcionais parciais (Peterson e Harrell Jr., 1990). Este é uma extensão ao modelo de *odds* proporcionais e consiste numa modelação mais realista permitindo que algumas variáveis explicativas possam ser modeladas considerando o pressuposto de *odds* proporcionais, enquanto outras são modeladas considerando que esse pressuposto não é válido e, como tal, considerando a existência de alguns parâmetros específicos que são incluídos no modelo e que variam consoante as diferentes categorias que estão a ser comparadas. Num modelo de *odds* proporcionais parciais considera-se a existência de um subconjunto de variáveis explicativas para as quais se assume a não existência de *odds* proporcionais. Existem dois tipos de modelos de *odds* proporcionais parciais, o não restrito e o restrito.

No modelo de *odds* proporcionais parciais não restrito (OPP-NR) considera-se que de todas as variáveis explicativas p, $\underset{\sim}{x}=(x_1,x_2,...,x_p)$, só algumas verificam o pressuposto de *odds* proporcionais. Suponha-se que as primeiras q covariáveis não verificam o pressuposto de *odds* proporcionais. Para uma variável x_a que não verifique o pressuposto de *odds* proporcionais $\alpha_a - \beta_a X_a$ é incrementado de um coeficiente (γ_{aj}) que é o efeito associado a cada j-ésimo *logit* cumulativo ajustado pelas restantes variáveis, $\alpha_a-(\beta_a X_a+\gamma_{aj})$. Para este modelo estimam-se $(k-1)$interseções, p coeficientes β independentes das categorias comparadas e $q(k-1)$ parâmetros *gamma* (γ), os quais se encontram associados a cada covariável e categoria da variável de resposta. O modelo reduz-se ao modelo de *odds* proporcionais se os parâmetros (γ) forem nulos, $\gamma_j = 0$ para todos os j. Para as primeiras q variáveis explicativas, o coeficiente angular depende de j, o que significa que a relação entre $\underset{\sim}{x}$ e Y é dependente da categoria em questão. Os *odds ratio* (OR) são estimados para todas as comparações das categorias da variável de resposta. Para as restantes covariáveis os coeficientes angulares β são independentes de j, e como tal somente um OR é estimado. A forma geral do modelo é semelhante à anterior, mas agora com coeficientes associados a cada categoria da variável de resposta:

$$\text{logit}\left[\pi(Y\le j\left|x_1,x_{2,}...,x_p\right.)\right]=\ln\left[\frac{\pi(Y\le j\left|x_1,x_{2,}...,x_p\right.}{\pi(Y>j\left|x_1,x_{2,}...,x_p\right.}\right] \quad (9)$$

$$\text{logit}\left[\pi(Y\le j\left|x_1,x_{2,}...,x_p\right.)\right]=\alpha_j+\left[-(\beta_1+\gamma_{1j})X_1-...-\right.$$
$$-(\beta_q+\gamma_{qj})X_q-(\beta_{q+1})X_{q+1}-...-$$
$$\left.-(\beta_p)X_p\right],\ \text{com}\ j=1,2,...,k-1 \quad (10)$$

O modelo de *odds* proporcionais parciais restrito (OPP-R), proposto por Peterson e Harrell Jr. (1990), aplica-se quando existe uma relação linear entre cada OR dos pontos de corte específicos e a variável de resposta. Neste caso, as restrições, representadas pelos parâmetros *gamma* (γ), são inseridas como parâmetros no modelo de forma a incorporar essa linearidade. Para uma determinada covariável o coeficiente γ não depende dos pontos de corte mas é multiplicado por um coeficiente *tau* (τ), que corresponde a um escalar fixo que vai tomar a forma de restrição alocada aos parâmetros. Assim, para uma determinada covariável x_m considera-se que γ_m não depende dos pontos de corte mas é multiplicado por γ_j para cada j-ésimo *logit*.

A forma geral do modelo é:

$$\text{logit}\left[\pi(Y \le j \left| x_1, x_{2,} ..., x_p \right.)\right] = \ln\left[\frac{\pi(Y \le j \left| x_1, x_{2,} ..., x_p \right.)}{\pi(Y > j \left| x_1, x_{2,} ..., x_p \right.)}\right] \quad (11)$$

$$\text{logit}\left[\pi(Y \le j \left| x_1, x_{2,} ..., x_p \right.)\right] = \alpha_j + \left[\tau_j \left(-(\beta_1 + \gamma_1)X_1 - ... - \right.\right.$$
$$\left.\left. -(\beta_q + \gamma_q)X_q\right)\right] - (\beta_{q+1})X_{q+1} - ... -$$
$$-(\beta_p)X_p, \text{ com} j = 1,2,...,k-1 \quad (12)$$

O modelo de *odds* proporcionais parciais restrito é de usar quando exista algum tipo de tendência linear entre cada OR dos pontos de corte e a variável de resposta.

3. Modelização da recomendação do serviço com regressão logística ordinal. Resultados

Os dados utilizados foram retirados de um estudo referente à medição da satisfação de clientes na indústria hoteleira (João, 2009).

Os dados foram recolhidos a partir de 390 inquéritos de satisfação de clientes com o serviço prestado por uma unidade hoteleira do Algarve. Os dados reportam ao período compreendido entre junho e setembro de 2008.

A variável dependente a estudar é a "recomendação do serviço a terceiros" que é uma medida da lealdade dos clientes.

A lealdade envolve mais do que simplesmente efetuar uma compra ou mesmo compras repetidas. A lealdade não deve ser confundida com a repetição da compra, pois tal pode apenas representar inércia do cliente. A lealdade representa um nível de compromisso com a organização através do apoio ativo que estes clientes lhe proporcionam (Hill, 1996).

De acordo com Reichheld (2006), a questão considerada primordial consiste em verificar se o cliente é um defensor da organização e, como tal, para este autor a questão chave é a "recomendação a terceiros".

A variável dependente "recomendação do serviço a terceiros" foi classificada em cinco categorias (1 = "decerto recomenda", 2 = "provavelmente sim", 3 = "talvez sim/talvez não", 4 = "provavelmente não" e 5 = "decerto não"). A variável dependente é ordinal e os vários níveis podem ser ordenados de acordo com o grau de recomendação. Para o estudo consideraram-se três variáveis explicativas binárias que foram codificadas da seguinte forma: "número de vezes hospedado no hotel" (primeira vez – 1, não é a primeira vez – 0), "idade" (inferior a 55 anos – 1, igual ou superior a 55 anos – 0), "estadia" (superior a sete noites – 1, inferior ou igual a sete noites – 0).

Em primeiro lugar, ajusta-se o modelo de *odds* proporcionais considerando o total das três variáveis explicativas atrás descritas. O Quadro 1 apresenta os resultados do ajuste para o modelo com as três variáveis independentes.

Antes de interpretar os resultados para o modelo, examina-se o pressuposto de *odds* proporcionais realizando o teste de Wald (Brant, 1990) para testar o pressuposto da regressão paralela (*odds* proporcionais) para o modelo com as três variáveis explicativas, e testar cada uma das variáveis independentes (Quadro 2). Efetuou-se ainda o teste de razão de verosimilhanças, RV (Wolfe e Gould, 1998).

Para o modelo o pressuposto de *odds* proporcionais é verificado pois $\chi^2_9 = 2{,}45$ com $p = 0{,}982$. Examinando os testes para cada variável independente, observa-se que o pressuposto de regressão paralela se verifica para todas as variáveis, pelo que o modelo de *odds* proporcionais é adequado para modelação dos dados.

No Quadro 3 apresentam-se os resultados dos *logit*s e dos *odds* cumulativos (coeficientes exponenciados).

Para o modelo de *odds* proporcionais a interpretação dos *odds* cumulativos é independente dos pontos de corte (existem quatro pontos de corte, uma vez que a variável dependente tem cinco níveis ordinais), pois eles são constantes para todos os níveis da variável de resposta (regressões paralelas).

Quadro 1 | Modelo de regressão logística ordinal

Número de observações = 355						
LR $\chi^2(3)$ = 12,27			Prob > χ^2 = 0,0065			
Log verosimilhança = -281,5165			Pseudo R^2 = 0,0213			
Recomend	Coeficiente	Erro padrão	z	P>|z|	[95% intervalo de confiança]	
Estadia	0,57658	0,24930	2,31	0,021	0,08796	1,06520
N vezes	0,36707	0,28724	1,28	0,201	-0,19592	0,93005
Idade	0,63025	0,25917	2,43	0,015	0,12229	1,13821
/corte 1	2,00170	0,33720			1,3408	2,66261
/corte 2	3,43933	0,37857			2,69734	4,18132
/corte 3	4,25961	0,43093			3,41500	5,10421
/corte 4	4,76218	0,48396			3,81364	5,71073

Fonte: elaboração própria.

Quadro 2 | Teste de Wald e teste RV

Coeficientes estimados para *j*–1 regressões binárias				
	y > 1	y > 2	y > 3	y > 4
Estadia	0,567912	0,801137	0,367274	0,202949
N vezes	0,374822	0,412310	0,724938	0,952957
Idade	0,620093	0,744424	0,916453	0,750573
cons	-1,998916	-3,695021	-4,644916	-5,133934
Teste para o pressuposto de regressão paralela				
	χ^2	p > χ^2	Graus de liberdade	
Todas	2,45	0,982	9	
Estadia	1,56	0,669	3	
N vezes	0,34	0,952	3	
Idade	0,32	0,957	3	
Teste de razão de verosimilhanças (RV)				
χ^2_9 = 2,23		Prob > χ^2 = 0,9873		

Fonte: elaboração própria.

Quadro 3 | Coeficientes de ajuste do modelo de regressão logística ordinal

Número de observações = 355						
Fator de variação em *odds*: >m *versus* <=m						
Recomend	β	z	P>|z|	e^{β}	$e^{\beta s_x}$	s_x
Estadia	0,57658	2,313	0,021	1,7799	1,3347	0,5007
N vezes	0,36707	1,278	0,201	1,4435	1,1804	0,4518
Idade	0,63025	2,432	0,015	1,8781	1,3666	0,4955
β = coeficiente						
z = valor de z para teste de β=0						
P>|z| = valor de prova para teste z						
e^{β} = variação do fator em *odds* para aumento de uma unidade em X						
$e^{\beta s_x}$= variação em *odds* para o aumento do desvio padrão em X						
s_x = desvio padrão de X						

Fonte: elaboração própria.

Por análise do Quadro 3, pode-se concluir que a variável N_vezes (número de vezes hospedado no hotel) não é significativa para um nível de significância de 5%.

O *odds ratio* $OR_{Estadia}$ = 1,7799 é a razão de chances proporcional da comparação da estadia superior a uma semana com a estadia menor ou igual a uma semana sobre a variável dependente

"recomendação do serviço", considerando que as restantes variáveis do modelo são mantidas constantes. Deste modo, para a estadia maior que uma semana, a chance da categoria 5 *versus* as categorias combinadas 1, 2, 3 e 4 é 1,7799 vezes mais alta que para a estadia menor ou igual a uma semana. Do mesmo modo, para a estadia maior que uma semana, a chance das categorias combinadas 5 e 4 *versus* as categorias combinadas 1, 2 e 3 é 1,7799 vezes mais alta que para a estadia menor ou igual a uma semana, e assim sucessivamente.

O *odds ratio* OR_{Idade} = 1,8781 é a razão de chances proporcional da comparação da idade inferior a 55 anos com a idade igual ou superior a 55 anos sobre a variável dependente "recomendação do serviço", considerando que as restantes variáveis do modelo são mantidas constantes. Deste modo, para a idade inferior a 55 anos, a chance da categoria 5 *versus* as categorias combinadas 1, 2, 3 e 4 é 1,8781 vezes mais alta que para a idade igual ou superior a 55 anos. Do mesmo modo, para a idade inferior a 55 anos, a chance das categorias combinadas 5 e 4 *versus* as categorias combinadas 1, 2 e 3 é 1,8781 vezes mais alta que para a estadia menor ou igual a uma semana, e assim sucessivamente.

Assim, poder-se-á concluir que, relativamente à duração da estadia, se verifica que para estadias mais prolongadas a recomendação a terceiros é mais baixa do que para estadias mais curtas. Da mesma forma se verifica que para clientes mais novos (com idades inferiores a 55 anos) a recomendação a terceiros é mais baixa do que para clientes com idades iguais ou superiores a 55 anos.

4. Conclusões

Neste trabalho utilizou-se a regressão logística ordinal para modelar dados de natureza categórica. O método usado permitiu verificar que quer a idade dos clientes, quer a duração da sua estadia, influenciam a sua intenção de recomendar o serviço. Será de todo o interesse para os gestores hoteleiros tentar compreender o motivo pelo qual os clientes mais novos evidenciam menor lealdade, e também o motivo pelo qual uma estadia mais prolongada faz baixar a recomendação. Terá interesse efetuar estudos para vários horizontes temporais e verificar se as variáveis significativas se mantêm ao longo do tempo, e se os seus efeitos são do mesmo tipo. Este tipo de estudos poderá ter interesse para a gestão hoteleira no sentido de desenvolver estratégias que melhorem a lealdade dos clientes.

Bibliografia

Agresti, A., 1989, Tutorial on modelling ordered categorical response data, *Psychological Bulletin*, Vol. 105, pp. 290-301.

Agresti, A., 2002, *Categorical Data Analysis*, 2 ed., John Wiley & Sons, New York.

Brant, R., 1990, Assessing proportionality in the proportional odds model for ordinal logistic regression, *Biometrics*, Vol. 46, pp. 1171-1178.

Cliff, N., 1996, Answering ordinal questions with ordinal data using ordinal statistics, *Multivariate Behavioral Research*, Vol. 3, pp. 331-350.

Cliff, N., and Keats, J., 2003, *Ordinal Measurement in the Behavioral Sciences*, Lawrence Erlbaum Associates, Mahwah, NJ.

Danaher, P., and Haddrell, V., 1996, A comparison of question scales used for measuring customer satisfaction, *International Journal of Service Industry Management*, Vol. 7, pp. 4-26.

Hill, N., 1996, *Handbook of Customer Satisfaction Measurement*, Gower Publishing Limited, England.

Hosmer, D., and Lemeshow, S., 2000, *Applied Logistic Regression*, 2 ed., John Wiley & Sons, Inc., New York.

Knapp, T., 1990, Treating ordinal scales as interval scales: an attempt to resolve the controversy, *Nursing Research*, Vol. 39, pp. 121-123.

Kotler, P., and Keller, K., 2006, *Marketing Management*, 12 ed., Pearson Education, Inc.

Krantz, D., Luce, R., Suppes, P., and Tversky, A., 1971, *Foundations of Measurement: Vol. 1. Additive and Polynomial Representations,* Academic Press, New York.

João, I., 2009, *Um Método Multicritério para Avaliação da Satisfação de Clientes na Indústria Hoteleira*, Tese de doutoramento, IST-UTL, Lisboa, Portugal.

Long, J., and Freese, J., 2001, *Regression Models for Categorical Dependent Variables Using Stata*, Stata Press Publication, TX.

McCullagh, P., 1980, Regression models for ordinal data, *Journal of the Royal Statistical Society: Series B*, Vol. 42, pp. 109-142.

McKelvey, R., and Zavoina, W., 1975, A statistical model for the analysis of ordinal level dependent variables, *Journal of Mathematical Sociology*, Vol. 4, pp. 103-120.

O´Connell, A., 2006, *Logistic Regression Models for Ordinal Response Variables*, Sage Publications, Thousand Oaks, London, New Delhi.

Oh, H., and Parks, S., 1997, Customer satisfaction and service quality: a critical review of the literature and research implications for the hospitality industry, *Hospitality Research Journal*, Vol. 20, pp. 35-64.

Oliver, R., 1981, Measurement and evaluation of satisfaction processes in retail settings, *Journal of Retailing*, Vol. 57, pp. 25-48.

Palmer, A., Sesé, A., and Montaño, J., 2005, Tourism and Statistics: Bibliometric Study 1998–2002, *Annals of Tourism Research*, Vol. 32, pp. 167-178.

Peterson, B., and Harrell Jr., F., 1990, Partial proportional odds models for ordinal response variables, *Applied Statistic*, Vol. 39, pp. 205-217.

Reichheld, D., 2006, *The Ultimate Question: Driving Good Profits and True Growth*, Harvard Business School Press, USA.

Stevens, S., 1946, On the theory of scales of measurement, *Science*, Vol.103, pp. 677-680.

Walker S., and Duncan, D., 1967, Estimation of the probability of an event as a function of several independent variables, *Biometrika*, Vol. 549, pp. 167-179.

Williams, R., 2006, Generalized ordered logit/partial proportional odds models for ordinal dependent variables, *The Stata Journal*, Vol. 6, pp. 58-82.

Wolfe, R., and Gould, W., 1998, An approximate likelihood-ratio test for ordinal response models, *Stata Technical Bulletin*, Vol. 42, pp. 24-27.