UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
Programa de Pós-Graduação em Epidemiologia

COMPARAÇÃO ENTRE A ESTIMATIVA DE RAZÃO DE CHANCES GERADA PELO MODELO DE ODDS PROPORCIONAIS COM A RAZÃO DE CHANCES GENERALIZADA

Álvaro Vigo

Profª Drª Jandyra Maria Guimarães Fachel

ORIENTADORA

Porto Alegre – RS, Brasil

Dezembro, 2004

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
Programa de Pós-Graduação em Epidemiologia
Porto Alegre – RS, Brasil

TESE

apresentada em 17/12/2004 como parte dos requisitos para obtenção do Grau de

DOUTOR EM EPIDEMIOLOGIA

Jandyra Maria Guimarães Fachel
PPG-EPI – UFRGS (Orientadora)

BANCA EXAMINADORA

Bruce Bartholow Duncan
PPG-EPI – UFRGS

Sotero Serrate Mengue
PPG-EPI – UFRGS

João Riboldi
PPG-EPI – UFRGS

José Francisco Soares
ICEx – UFMG

Armando Mario Infante
IMECC – UNICAMP

A Stela e ao Pedro.

"All models are wrong, some are useful."

GEORGE E. P. BOX

"The Monte Carlo method is now the most powerful and commonly used technique for analysing complex problems."

REUVEN Y. RUBINSTEIN

"Toda a noite, toda a noite,
Toda a noite sem pensar...
Toda a noite sem dormir
E sem tudo isso acabar."

FERNANDO PESSOA

AGRADECIMENTOS

Ao terminar de escrever este trabalho, carrego uma dívida de gratidão com muitas pessoas. À minha orientadora, Professora Jandyra Fachel, com quem tive o privilégio de conviver desde a graduação e devo grande parte da minha formação, expresso meu profundo agradecimento e admiração.

Ao Programa de Pós-Graduação em Epidemiologia, seus professores e funcionários. Este trabalho não seria viabilizado sem o suporte intelectual, humano e computacional do PPG-EPI. Em especial, sou muito grato aos professores Bruce Duncan e Maria Inês Schmidt que, além de muitos ensinamentos, têm me guiado pelo fascinante e desafiador caminho da Epidemiologia do Diabetes. Não poderia, claro, deixar de agradecer ao Professor Sotero Mengue, pelos ensinamentos e apoio incondicional.

Aos colegas e amigos do curso que, em conjunto, contribuíram para criar uma atmosfera amigável e estimulante, favorável ao estudo.

Ao Departamento de Estatística, pelo suporte recebido.

Finalmente, agradeço aos amigos e familiares, que conferiram atenção indispensável para a realização do curso. Em particular, tenho uma dívida impagável com a minha esposa Stela e meu filho Pedro, por saberem suportar minha ausência física tantas vezes. A vocês, dedico não apenas este trabalho, mas meu amor.

ABSTRACT

Ordinal outcomes are very common in medical and epidemiological research and must be analyzed by means of methods that consider the ordered structure of the categories. The proportional odds model has been used more frequently to describe the relation between an ordinal outcome and the predictors, but the generalized odds ratio can be also useful. Monte Carlo simulations confirm that in contingency tables with an outcome with three ordered categories and a dichotomous explanatory factor, the estimates obtained using the proportional odds model are very close to the generalized odds ratio and have the same efficiency. The methods are illustrated by means of the data of the Brazilian Study of Gestational Diabetes, to investigate the association of the ambient temperature and body mass index (BMI) with the classification of the hyperglycaemia in the pregnancy. There is evidence of an antagonistic interaction between temperature and BMI (P=0.0268). For obese individuals, the odds ratio of classify an individual as diabetic, for an ambient temperature greater or equal to 25 ºC, in relation of a temperature under 25 ºC, is equal to OR=1.94 (95% CI: 1.56-2.41), while for non-obese individuals, the odds ratio is OR=3.03 (95% CI: 2.18-4.23). The results produced by means of the generalized odds ratio and the proportional odds model are identical, as shown in the Monte Carlo study.

RESUMO

Desfechos ordinais são muito comuns em pesquisas médicas e epidemiológicas e devem ser analisados mediante métodos que considerem a estrutura ordenada das categorias. O modelo de odds proporcionais tem sido usado com maior freqüência para descrever a relação entre um desfecho ordinal e os preditores, mas a razão de chances generalizada também pode ser útil. As simulações Monte Carlo deste trabalho confirmam que, em tabelas de contingência com um desfecho com três categorias ordenadas e um fator explanatório dicotômico, as estimativas da razão de chances produzidas pelo modelo de odds proporcionais e da razão de chances generalizada são equivalentes e têm a mesma eficiência. Os métodos são ilustrados mediante os dados do Estudo Brasileiro de Diabetes Gestacional, para investigar a associação da temperatura ambiente e do índice de massa corporal (IMC) com a classificação da hiperglicemia na gravidez. Os resultados evidenciam que existe interação do tipo antagônica entre temperatura e IMC (P = 0,0268). Para indivíduos obesos, a chance de classificar um indivíduo como diabético, para uma temperatura ambiente ≥ 25 ºC, em relação a temperatura < 25 ºC, é igual a RC=1,94 (IC 95%: 1,56-2,41), enquanto que para indivíduos não obesos, a razão de chances é RC=3,03 (IC 95%: 2,18-4,23). Esta comparação empírica mostra que a razão de chances generalizada é equivalente à estimativa da razão de chances do modelo de odds proporcionais.

ÍNDICE

1 INTRODUÇÃO

Desfechos ordinais são muito comuns em pesquisas médicas e epidemiológicas, mas métodos de análise que incorporam a estrutura ordenada das categorias ainda têm sido pouco utilizados. Em muitas situações, o desfecho ordinal representa os níveis de uma escala de medida usual, como a severidade da dor (nenhuma, moderada, severa). Em outros casos, por razões práticas ou porque a variável contínua subjacente não pode ser diretamente observada, a estrutura ordinal surge da categorização de uma ou mais variáveis contínuas, como, por exemplo, na classificação dos valores da glicose plasmática para o diagnóstico de diabetes melito e seus estágios pré-clínicos (glicemia normal, tolerância à glicose diminuída, diabetes melito).

O modelo de odds proporcionais (1;2) é comumente usado para descrever a relação entre um desfecho ordinal e os preditores. Apesar de ser pouco conhecida, a razão de chances generalizada (3) também pode ser uma medida de associação útil. Comparações empíricas (4) sugerem que para um desfecho com *k* categorias ordenadas e um preditor dicotômico, as estimativas da razão de chances geradas pelo modelo de odds proporcionais e pela razão de chances generalizada são praticamente idênticas.

Contudo, nenhum estudo investigou a equivalência destas estimativas. A hipótese de pesquisa postula que, quando as exigências dos modelos de odds proporcionais e da razão de chances generalizada estão atendidas, as estimativas da razão de chances produzidas por estes modelos são equivalentes. Na presença de violações da suposição de odds proporcionais e e da aderência dos dados à

distribuição de probabilidade Tipo-C Normal, seriam esperadas discrepâncias maiores entre as estimativas.

Como na regressão logística tradicional, quando o preditor possui mais do que dois níveis, o modelo de odds proporcionais determina mais de uma razão de chances e estes casos não são abordados neste trabalho. Os objetivos do trabalho são explicitados na próxima seção, enquanto que a estrutura do trabalho é descrita na seção 1.2.

1.1 OBJETIVOS

O objetivo principal deste trabalho é comparar a estimativa da razão de chances generalizada com aquela gerada pelo modelo de odds proporcionais, em tabelas de contingência com preditor binário e desfecho com *k* categorias ordenadas. A equivalência das estimativas da razão de chances é avaliada quando as exigências dos modelos estão ou não estão satisfeitas. Também se deseja ilustrar a aplicação e potencialidade dos modelos e, mais importante, a interpretação dos resultados, mediante dados reais do Estudo Brasileiro de Diabetes Gestacional (EBDG).

Para tanto, como a razão de chances generalizada não possui forma explícita, um estudo de simulação Monte Carlo foi planejado e conduzido para estudar a similaridade entre as estimativas da razão de chances produzidas por estes modelos. A questão da pesquisa e aspectos metodológicos envolvidos estão detalhados nos artigos correspondentes (Capítulo 3) e também no projeto de pesquisa (Anexo A). A próxima seção descreve sucintamente a organização trabalho.

1.2 ESTRUTURA DO TRABALHO

A estrutura do trabalho contempla uma revisão da literatura, dois artigos a serem submetidos para publicação, considerações finais e anexos. O Capítulo 2 apresenta uma revisão dos principais modelos para descrever a relação funcional entre um desfecho ordinal e um conjunto de preditores ou variáveis explanatórias, com destaque para os modelos de odds proporcionais e da razão de chances generalizada.

Os dois artigos compõem o Capítulo 3. O primeiro mostra os resultados do estudo de simulação Monte Carlo realizado para comparar as estimativas de razão de chances geradas pelos modelos de odds proporcionais e da razão de chances generalizada. O segundo artigo tem como objetivo básico ilustrar uma aplicação dos modelos de odds proporcionais e da razão de chances generalizada e comparar empiricamente suas estimativas de razão de chances. Eles foram usados para estimar a associação entre a temperatura ambiente e obesidade com as concentrações de glicose durante um teste oral de tolerância à glicose, utilizando dados da linha de base do EBDG.

O Capítulo 4 exibe algumas considerações finais importantes sobre os métodos examinados, indicando também aspectos que ainda devem ser pesquisados.

Os anexos contêm informações importantes organizadas em três partes (A, B e C). O Anexo A contém o projeto de pesquisa que deu origem ao trabalho. Sua primeira versão foi submetida à Comissão de Pós-Graduação como parte dos requisitos para seleção e ingresso no Programa de Pós-Graduação em Epidemiologia. Após as apresentações nos Seminários de Pesquisa do PPG-EPI, foi reformulado e ganhou o formato atual. O Anexo B, composto por duas seções, mostra as rotinas computacionais para o ajuste dos modelos aos dados do EBDG. Por fim, o Anexo C

contempla as rotinas computacionais desenvolvidas para este estudo de simulação Monte Carlo. Estão sendo apresentados apenas os programas para as simulações do casos em que o parâmetro de associação é $\psi = 1$, com categorização do desfecho simétrica suave (Anexo C1) e simétrica concentrada (Anexo C2). Para as demais situações, os programas são idênticos, bastanto alterar o valor de ψ, os nomes dos arquivos de dados e os caminhos que identificam as pastas onde estão armazenadas as informações necessárias e onde são gravados os resultados.

2 REVISÃO DA LITERATURA

A pesquisa científica caracteriza-se essencialmente como um processo iterativo de acumulação de conhecimento que envolve interpretação de dados e a subseqüente formulação de hipóteses, modelos e teorias. A interpretação dos fenômenos em questões de pesquisa envolve sua observação e transformação dos dados em informação. Assim, mediante estudos observacionais ou experimentais cuidadosamente planejados, o pesquisador deseja medir a incerteza contida nos dados, extraindo a informação necessária para sua interpretação. A escolha da ferramenta estatística adequada é vital para gerar com precisão a informação desejada, de tal forma que as conclusões sejam válidas e possam embasar modificações nos modelos e teorias vigentes.

Apesar da crescente tendência em "quantificar" os fenômenos sob investigação, em virtude da sua natureza ou dificuldade, algumas características não podem ser medidas em uma escala quantitativa, gerando a necessidade de novos procedimentos de análise e/ou do estudo da adequação ou eficácia dos mesmos. Por exemplo, a severidade de um sintoma ou de uma doença, classe social e nível de instrução dificilmente poderiam ser medidos quantitativamente e, usualmente, são observados através de categorias.

Variáveis cuja escala de medida (5) consiste de um conjunto de categorias disjuntas são denominadas variáveis categóricas. Elas surgem nas mais variadas áreas do conhecimento, tais como ciências sociais, epidemiologia, psicologia, ecologia, medicina, etc. Variáveis categóricas para as quais não existe uma ordem natural dos seus níveis ou categorias são ditas nominais. Em muitas variáveis categóricas, entanto, existe uma ordem natural dos seus níveis, expressando (em

ordem crescente ou decrescente) a intensidade de um fenômeno observável. Estas variáveis são chamadas de categóricas ordenadas.

Estudos com resposta categórica ordenada têm aparecido com grande freqüência na literatura médica e epidemiológica. Em muitas situações, o desfecho ordinal representa os níveis de uma escala de medida usual, como a severidade da dor (nenhuma, moderada, severa). Em outros casos, por razões práticas ou porque a variável contínua subjacente não pode ser diretamente observada, a estrutura ordinal surge da categorização de uma ou mais variáveis contínuas. Para extrair eficientemente a informação contida nos dados é vital utilizar métodos de análise que considerem a ordem natural das categorias. A escolha da metodologia geralmente depende dos objetivos e do tipo de delineamento, bem como da observância das exigências do método. Dentre as principais ferramentas de análise destacam-se os modelos para resposta ordinal, os quais permitem avaliar o impacto dos fatores explanatórios ou experimentais sobre o desfecho.

2.1 MODELOS PARA RESPOSTA ORDINAL

A modelagem usualmente permite identificar variáveis que produzem impactos significativos sobre a resposta e estimar a magnitude e direção das associações. A principal motivação para os modelos propostos é a possibilidade de existir uma resposta variável, latente e contínua, usualmente não observável. Assim, os dados observados são uma categorização de uma variável aleatória contínua subjacente, não observável diretamente.

Um caso conhecido é o dos bioensaios, onde a variável latente corresponde ao nível de tolerância de uma droga, sobre o qual se assume uma distribuição contínua na população. O nível de tolerância não pode ser observado diretamente,

mas sua elevação é manifestada através do crescimento da probabilidade de sobrevivência. A existência de uma variável latente não é fundamental para a validade do modelo, mas, se existe, a interpretação dos parâmetros se torna clara e direta (1).

A variável latente, representada por Z, tem função de distribuição $F_{\eta}(z) = F(z - \eta)$, onde η é um parâmetro de posição, tal que para o vetor de covariáveis $\mathbf{x}$, $\eta(\mathbf{x}) = \boldsymbol{\beta}^t \mathbf{x}$. Assim, a distribuição condicional de $Z|\mathbf{x}$ é dada por $F(z - \boldsymbol{\beta}^t \mathbf{x})$. A variável contínua subjacente Z não pode ser observada, mas regiões de valores de Z são conhecidas através da categorização em intervalos reais $(-\infty, \theta_1]$, $(\theta_1, \theta_2], \cdots, (\theta_{k-1}, +\infty)$, onde os θ_j são parâmetros desconhecidos. Isso induz a variável aleatória Y, que assume os valores $Y = j$ se e somente se $Z \in (\theta_{j-1}, \theta_j]$. Assim, as probabilidades acumuladas são definidas por $\gamma_j(\mathbf{x}) = P[Y \leq j|\mathbf{x}] = F(\theta_j - \boldsymbol{\beta}^t \mathbf{x})$.

A distribuição de probabilidade da variável latente Z é um aspecto essencial dos distintos modelos. Usualmente, qualquer distribuição unimodal simétrica pode ser postulada, produzindo resultados similares. No entanto, a suposição de uma distribuição logística tem se mostrado bastante adequada, principalmente pela facilidade de cálculo. A escolha da distribuição logística para a variável latente conduz ao modelo para resposta ordinal usado com maior freqüência, introduzido por S.H. Walker e D.B. Duncan (6) e, mais tarde, denominado *modelo de odds proporcionais* (*proportional odds model*) (1;2).

Para ilustrar, considere a variável dependente Y, cujas categorias de resposta ordenadas são denotadas por $1, 2, \ldots, k$ e $\mathbf{x} = (x_1, x_2, \ldots, x_p)^t$ é o vetor p-dimensional de preditores. Sejam, também, $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \ldots, \pi_k(\mathbf{x})$ as probabilidades

das *k* categorias de resposta ordenadas, quando o vetor de covariáveis assume o valor $\mathbf{x}$; ou seja,

$$P[Y = j|\mathbf{x}] = \pi_j(\mathbf{x}); \quad \forall\ j = 1,2,\cdots,k\ .$$

Então, utilizando a linguagem dos modelos lineares generalizados (2), a função de ligação $log\,it\,\{\gamma_j(\mathbf{x})\}$ produz um modelo linear nos parâmetros, definido por

$$log\,it\,\{\gamma_j(\mathbf{x})\} = log\,\frac{\gamma_j(\mathbf{x})}{1-\gamma_j(\mathbf{x})} = \theta_j - \boldsymbol{\beta}^t\mathbf{x}\ , \qquad 1 \le j < k\ ,$$

onde os parâmetros desconhecidos θ_j satisfazem $\theta_1 \le \theta_2 \le \cdots \le \theta_{k-1}$, $\theta_0 \equiv -\infty$ e $\theta_k \equiv +\infty$. O vetor de parâmetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^t$ representa os coeficientes de regressão a serem estimados, enquanto que o vetor $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_{k-1})^t$ contém os pontos de corte desconhecidos. Equivalentemente, o modelo pode ser escrito na forma

$$P[Y \le j|\mathbf{x}] = \frac{exp\{\theta_j - \boldsymbol{\beta}^t\mathbf{x}\}}{1 + exp\{\theta_j - \boldsymbol{\beta}^t\mathbf{x}\}};\ \forall 1 \le j < k\ ,$$

tal que para os valores de $\mathbf{x}_1$ e $\mathbf{x}_2$ distintos, o quociente

$$\frac{\dfrac{\gamma_j(\mathbf{x}_1)}{1-\gamma_j(\mathbf{x}_1)}}{\dfrac{\gamma_j(\mathbf{x}_2)}{1-\gamma_j(\mathbf{x}_2)}} = \frac{exp\{\theta_j - \boldsymbol{\beta}^t\mathbf{x}_1\}}{exp\{\theta_j - \boldsymbol{\beta}^t\mathbf{x}_2\}} = exp\{\boldsymbol{\beta}^t(\mathbf{x}_2 - \mathbf{x}_1)\}$$

pode ser interpretado como a chance relativa de observar o desfecho em uma categoria menor ou igual a j, entre dois indivíduos com valores diferentes para os preditores.

No modelo de odds proporcionais o vetor dos coeficientes de regressão $\boldsymbol{\beta}$ não depende da categoria j, o que implica que a relação funcional entre as covariáveis $\mathbf{x} = (x_1, x_2, \ldots, x_p)^t$ e a variável resposta Y independe da categoria j. Na

prática, esta suposição pode ser avaliada através de um teste baseado em escores. Se os dados não evidenciam a adequação desta suposição de linhas paralelas, outros modelos estão disponíveis para descrever a dependência de uma variável resposta ordinal e um conjunto de preditores.

Alternativamente, outras suposições sobre a distribuição condicional de $Z|\mathbf{x}$ são possíveis. Por exemplo, a escolha da distribuição normal $N(\boldsymbol{\beta}^t\mathbf{x},1)$ conduz ao modelo de probitos

$$\gamma_j(\mathbf{x}) = P[Y \leq j|\mathbf{x}] = \Phi(\theta_j - \boldsymbol{\beta}^t\mathbf{x}),$$

e, portanto,

$$\Phi^{-1}(\gamma_j(\mathbf{x})) = \theta_j - \boldsymbol{\beta}^t\mathbf{x},$$

onde $\Phi(.)$ denota a função de distribuição de uma variável aleatória com distribuição normal padrão.

Na existência de razões que levem a acreditar que a distribuição subjacente é assimétrica, as funções de ligação *log-log* e *complementar log-log* podem ser usadas (7). Assim, postulando a distribuição do valor extremo (também conhecida como distribuição de Gumbel ou de Gompertz) para a distribuição condicional de $Z|\mathbf{x}$, a função de ligação complementar log-log gera o modelo

$$log(-log(1-\gamma_j(\mathbf{x}))) = \theta_j - \boldsymbol{\beta}^t\mathbf{x};\ \ j = 1,2,\cdots k-1,$$

também conhecido como *modelo de riscos proporcionais* (*proportional hazard model*) para tempo discreto (1;2).

Recentemente, em parte devido às facilidades computacionais, diversos modelos para resposta ordinal têm sido usados. Um resumo dos principais métodos foi apresentado por C. V. Ananth e D. G. Kleinbaum (8), os quais são brevemente descritos a seguir.

No modelo de odds proporcionais, se a probabilidade $\gamma_j(\mathbf{x}) = P[Y \leq j|\mathbf{x}]$ é substituída por $\pi_j(\mathbf{x}) = P[Y = j|\mathbf{x}]$, o modelo resultante pode ser escrito como

$$log \frac{P[Y = j|\mathbf{x}]}{P[Y > j|\mathbf{x}]} = \theta_j - \boldsymbol{\beta}^t \mathbf{x}; \;\; j = 1,2,\cdots,k,$$

sendo chamado de *modelo de razão-continuação* (*continuation-ratio model*). Este modelo freqüentemente é mais útil quando o desfecho obedece alguma estrutura hierárquica ou aninhada (2).

Uma aparente limitação do modelo de odds proporcionais é a suposição de linhas paralelas, nem sempre atendida na prática. Em tais casos, uma alternativa menos exigente é o ajuste do *modelo de odds proporcionais parciais*, *restrito* ou *não restrito* (*constrained or unconstrained partial-proportional odds model*). O modelo de odds proporcionais parciais não restrito postula que um subconjunto de q dos p $(q < p)$ preditores apresenta odds não proporcionais. Se Y é resposta ordinal com k categorias e $\mathbf{x}$ o vetor *p*-dimensional de preditores, o modelo especifica que

$$P[Y \leq j|\mathbf{x}] = \frac{exp\{-\theta_j - \boldsymbol{\beta}^t \mathbf{x} - \boldsymbol{\gamma}_j^t \mathbf{t}\}}{1 + exp\{-\theta_j - \boldsymbol{\beta}^t \mathbf{x} - \boldsymbol{\gamma}_j^t \mathbf{t}\}}; \;\; j = 1,2,\cdots,k$$

ou, equivalentemente,

$$log \frac{P[Y \leq j|\mathbf{x}]}{1 - P[Y \leq j|\mathbf{x}]} = -\theta_j - \boldsymbol{\beta}^t \mathbf{x} - \boldsymbol{\gamma}_j^t \mathbf{t}; \;\; j = 1,2,\cdots k,$$

onde $\mathbf{t}$ é um vetor *q*-dimensional composto pelos $q < p$ preditores para os quais a suposição de odds proporcionais não é assumida a priori ou precisa ser testada; $\boldsymbol{\gamma}_j$ é o vetor $(q \times 1)$ de coeficientes de regressão associadas às componentes do vetor $\mathbf{t}$, com $\boldsymbol{\gamma}_1 = 0$. O modelo de odds proporcionais parciais restrito, por sua vez, especifica que

$$P[Y \leq j|\mathbf{x}] = \frac{exp\{-\theta_j - \boldsymbol{\beta}^t\mathbf{x} - \Gamma_j\boldsymbol{\gamma}^t\mathbf{t}\}}{1+exp\{-\theta_j - \boldsymbol{\beta}^t\mathbf{x} - \Gamma_j\boldsymbol{\gamma}^t\mathbf{t}\}}; \quad j = 1,2,\cdots,k\,,$$

onde os Γ_j são escalares pré-especificados e fixos, com $\Gamma_1 = 0$.

Um modelo comumente usado é o *modelo de estereótipo* (*stereotype model*) introduzido por A. Anderson (9), no qual os coeficientes de regressão $\boldsymbol{\beta}_j$ são modelados mediante a imposição de uma relação linear do tipo

$$\boldsymbol{\beta}_j = -\phi_j\boldsymbol{\beta}\,, \quad j = 1,2,\cdots,k\,,$$

onde $\boldsymbol{\beta} = (\beta_1,\beta_2,\cdots,\beta_p)^t$. Os parâmetros desconhecidos ϕ_j podem ser vistos como "escores" atribuídos à categoria de resposta j e precisam ser estimados. É conveniente salientar que como $\boldsymbol{\beta}_k = 0$, então $\phi_k = 0$ e, para evitar as dificuldades de identificabilidade, toma-se $\phi_1 = 1$. Assim, o modelo de estereótipo especifica que

$$P[Y = j|\mathbf{x}] = \frac{exp\{\theta_j - \phi_j\boldsymbol{\beta}^t\mathbf{x}\}}{\sum_{s=1}^{k} exp\{\theta_s - \phi_s\boldsymbol{\beta}^t\mathbf{x}\}}; \quad j = 1,2,\cdots,k\,.$$

Outro modelo importante é o *modelo de odds proporcionais não paramétrico* proposto por T. Hastie e R. Tibshirani (10). Este modelo pertence à classe dos modelos aditivos generalizados (11) e especifica que

$$log\,it\{P[Y \leq j|\mathbf{x}]\} = \theta_k - \sum_{i=1}^{p} f_i(x_i); \quad j = 1,2,\cdots,k\text{-}1,\ E[f_i(x_i)] = 0, \forall\, i\,.$$

Ainda no contexto de regressão não paramétrica, para o modelo de respostas acumuladas $Y \leq j|\mathbf{x}$, G. Kauermann e G. Tutz (12) substituíram a componente sistemática $\boldsymbol{\beta}^t\mathbf{x}$ por uma componente de alisamento $\lambda(\mathbf{x})$, produzindo o modelo semi-paramétrico

$$P[Y \leq j|\mathbf{x}] = F(\theta_j - \lambda(\mathbf{x})); \quad j = 1,2,\cdots,k\text{-}1,$$

onde $F(.)$ é uma função de distribuição especificada e $\lambda(\mathbf{x})$ é uma função não paramétrica desconhecida. Adicionalmente, os pontos de corte θ_j podem ser substituídos por funções de alisamento (*smooth functions*), conduzindo ao modelo

$$P[Y \leq j|\mathbf{x}] = F(\lambda_j(\mathbf{x})); \quad j = 1,2,\cdots,k\text{-}1,$$

onde as funções de alisamento satisfazem as restrições $\lambda_j(.) \leq \lambda_{j+1}(.)$, $j = 1,2,\cdots,k\text{-}2$.

Similarmente, G. Kauermann e G. Tutz (12) também estenderam o modelo de razão-continuação para um modelo semi-paramétrico, que assume a forma

$$P[Y = j \mid Y > j, \mathbf{x}] = F(\theta_j - \lambda(\mathbf{x})); \quad j = 1,2,\cdots,k-1,$$

onde $F(.)$ é a função de distribuição de uma variável aleatória com distribuição logística. Também neste modelo os pontos de corte θ_j podem ser substituídos por funções de alisamento, conduzindo ao modelo

$$P[Y = j \mid Y > j, \mathbf{x}] = F(\lambda_j(\mathbf{x})); \quad j = 1,2,\cdots,k-1.$$

Uma forma unificada para diferentes modelos logísticos para resposta ordinal foi apresentada por L. Fu e D.G. Simpson (13), sendo denominada *regressão logística simultânea* (*simultaneous logistic regression*). Nesta classe estão incluídos casos especiais como, por exemplo, os modelos de odds proporcionais e de razão-continuação. No entanto, este procedimento é potencialmente útil para modelar dados ordinais correlacionados, como ocorre em alguns estudos longitudinais. Também para este contexto, P.J. Lindsey e J. Kaufmann (14) apresentaram um modelo geral para analisar dados não balanceados de estudos com medidas repetidas com resposta ordinal. O método é bastante flexível e consiste em construir, mediante a transformação de Laplace, a função de distribuição de uma mistura de distribuições, incorporando-a ao modelo de razão-continuação ou de odds proporcionais, por exemplo.

Em determinadas situações pode ser importante avaliar o comportamento simultâneo entre um conjunto de variáveis resposta categóricas e um vetor de covariáveis $\mathbf{x}$. P. McCullagh e J.A. Nelder (2) e G.F.G.Glonek e P. McCullagh (15) introduziram a classe de modelos de regressão logística chamada de *transformação logística multivariada* (*multivariate logistic transform*), que pode ser útil quando as respostas são discretas, ordinais ou nominais. A forma geral da transformação é $\boldsymbol{\eta} = \mathbf{C}^t \, log(\mathbf{L}\boldsymbol{\pi})$, onde $\mathbf{C}$ e $\mathbf{L}$ são matrizes pré-especificadas.

Os modelos de regressão logística multivariados podem ser escritos por $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, onde $\mathbf{X}$ é uma matriz de constantes com dimensão $(k \times p)$ e $\boldsymbol{\beta}$ é um vetor *p*-dimensional de parâmetros desconhecidos. Estudos longitudinais com resposta categórica constituem uma importante aplicação do modelo logístico multivariado. Para ilustrar, considere o caso de duas variáveis de resposta Y_A e Y_B, com r e s categóricas ordenadas, respectivamente. Seja $\gamma_{ij} = P[Y_A \leq i, Y_B \leq j]$, tal que poderia ser considerado o modelo de regressão

$$M: \quad log\, it\{\gamma_{is}\} = \theta_i + \boldsymbol{\beta}_i^t \mathbf{x}; \qquad log\, it\{\gamma_{rj}\} = \phi_j + \boldsymbol{\beta}_j^t \mathbf{x}; \qquad log\left(\frac{\gamma_{ij}\left(1 - \gamma_{is} - \gamma_{rj} + \gamma_{ij}\right)}{\left(\gamma_{is} - \gamma_{ij}\right)\left(\gamma_{rj} - \gamma_{ij}\right)} \right) = \alpha .$$

De fato, é melhor considerá-lo como um conjunto de três modelos separados. O primeiro e segundo modelos descrevem, respectivamente, a relação entre a distribuição marginal de Y_A e de Y_B com o vetor de covariáveis $\mathbf{x}$, enquanto que o último corresponde ao conjunto de modelos da *razão cruzada global* (*global cross-ratio*) .

No contexto em que são consideradas simultaneamente duas variáveis respostas, ambas ordinais, e um conjunto de covariáveis $\mathbf{x} = \left(x_1, x_2, \ldots, x_p\right)^t$, J. R. Dale (16) apresentou o modelo estatístico chamado de razão cruzada global (*global cross-ratio model)* que permite estimar distintas razões de chances, chamadas de

razão cruzada global (*global cross-ratio*), doravante chamada de razão de chances global. Para definir o modelo, considere o vetor aleatório discreto $Y(\mathbf{x}) = (Y_1(\mathbf{x}), Y_2(\mathbf{x}))^t$ que representa as características de resposta que dependem do vetor de covariáveis $\mathbf{x}$, qualitativo ou quantitativo. A componente $Y_1(\mathbf{x})$ pode assumir valores nas categorias ordenadas rotuladas como $1,2,\cdots,r$ e as probabilidades acumuladas são dadas por $\eta_{ix} = P[Y_1(\boldsymbol{x}) \leq i];\ i = 1,2.\cdots,r$. Da mesma forma, $Y_2(\mathbf{x})$ assume os valores $1,2,\cdots,c$, tal que $\xi_{jx} = P[Y_2(\boldsymbol{x}) \leq j];\ j = 1,2.\cdots,c$. Essas probabilidades acumuladas obedecem as restrições $\eta_{0x} = \xi_{0x} = 0$, $\eta_{rx} = \xi_{cx} = 1$ e podem ser representadas pelos vetores $\boldsymbol{\eta}_x = (\eta_{1x}, \eta_{2x}, \cdots, \eta_{r-1,x})^t$ e $\boldsymbol{\xi}_x = (\xi_{1x}, \xi_{2x}, \cdots, \xi_{c-1,x})^t$. A função de distribuição do vetor bidimensional $Y(\mathbf{x})$ é definida por

$$F_{ij}(\mathbf{x}; \boldsymbol{\theta}_x) = P[Y_1(\mathbf{x}) \leq i; Y_2(\mathbf{x}) \leq j],$$

onde $\boldsymbol{\theta}_x = (\boldsymbol{\eta}_x, \boldsymbol{\xi}_x, \boldsymbol{\psi}_x)$ e $\boldsymbol{\psi}_x$ é uma matriz de ordem $(r-1)\times(c-1)$ cujos elementos representam as diferentes razão de chances (*global cross-ratio matrix*). O espaço de distintos valores de $Y(\mathbf{x})$ pode ser dividido em quatro quadrantes; isto é,

$$[Y_1 \leq i, Y_2 \leq j], \quad [Y_1 \leq i, Y_2 > j], \quad [Y_1 > i, Y_2 \leq j], \quad [Y_1 > i, Y_2 > j].$$

Assim, com a dupla dicotomia definida no ponto de corte bivariado (i, j), a tabela de contingência $r \times c$ é "transformada" em uma tabela 2×2 e a razão de chances global (16) é escrita como

$$\psi_{ij} = \frac{P[Y_1 \leq i, Y_2 \leq j] \times P[Y_1 > i, Y_2 > j]}{P[Y_1 > i, Y_2 \leq j] \times P[Y_1 \leq i, Y_2 > j]} = \frac{F_{ij}(\mathbf{x}; \boldsymbol{\theta}_x) \times (1 - \eta_{ix} - \xi_{jx} + F_{ij}(\mathbf{x}; \boldsymbol{\theta}_x))}{(\xi_{jx} - F_{ij}(\mathbf{x}; \boldsymbol{\theta}_x)) \times (\eta_{jx} - F_{ij}(\mathbf{x}; \boldsymbol{\theta}_x))}.$$

A equação acima pode ser resolvida para a função de distribuição $F_{ij}(\mathbf{x}; \boldsymbol{\theta}_x)$ através dos vetores $\boldsymbol{\eta}_x, \boldsymbol{\xi}_x$ e da matriz $\boldsymbol{\psi}_x$, produzindo

$$F_{ij}(\mathbf{x}; \boldsymbol{\eta}_x, \boldsymbol{\xi}_x, \boldsymbol{\psi}_x) = \begin{cases} \tfrac{1}{2}(\psi_{ijx} - 1)^{-1}\left[1 + (\eta_{ix} + \xi_{jx})(\psi_{ijx} - 1) - S(\eta_{ix}, \xi_{jx}, \psi_{ijx})\right], & se \ \ \psi_{ijx} \neq 1 \\ \eta_{ix}\xi_{jx}, & se \ \ \psi_{ijx} = 1 \end{cases}$$

para todo $i = 1,2,\cdots,r-1$ e $j = 1,2,\cdots,c-1$, onde

$$S(\eta,\xi,\psi) = \sqrt{[1+(\eta+\xi)(\psi-1)]^2 + 4\psi(1-\psi)\eta\xi}\,.$$

A distribuição definida por $F_{ij}(\mathbf{x};\boldsymbol{\eta}_x,\boldsymbol{\xi}_x,\boldsymbol{\psi}_x)$ é denominada modelo de razão de chances global (*global cross-ratio model*) para $Y(\mathbf{x}) = (Y_1(\mathbf{x}), Y_2(\mathbf{x}))^t$, e as probabilidades acumuladas F_{ijx} dependem do vetor de covariáveis $\mathbf{x}$ somente através dos parâmetros η_{ix}, ξ_{jx} e ψ_{ijx}. No entanto, na forma como foi definida, é um modelo saturado com $rc-1$ parâmetros, que é igual ao número de graus de liberdade da tabela. Assim, a dimensão do espaço de parâmetros deve ser reduzida mediante a imposição de restrições apropriadas. Como ψ_{ijx} assume valores em $(0,+\infty)$, na estimação dos parâmetros é conveniente modelar o logaritmo da razão de chances, ou seja,

$$log\,(\psi_{ijx}) = \Delta + \alpha_{ia} + \beta_{ja} + \delta_{ij} - \boldsymbol{\gamma}^t\mathbf{x}\,,$$

para $i = 1,2,\cdots,r-1$, $j = 1,2,\cdots,c-1$ e o índice a representa a associação, com as restrições de unicidade especificando que $\alpha_{r-1,a} = \beta_{c-1,a} = 0$; $\delta_{i,c-1} = 0$ para $i = 1,2,\cdots,r-1$ e $\delta_{r-1,j} = 0$ para $j = 1,2,\cdots,c-1$. Assim, por exemplo, a igualdade $\boldsymbol{\gamma} = \mathbf{0}$ indica que a associação não depende dos preditores $\mathbf{x}$ e, se todos os δ_{ij} são nulos, não há interação entre as respostas Y_1 e Y_2.

No contexto de medidas de associação, K. Pearson e D. Heron (17) haviam introduzido uma superfície bivariada contínua, cujo parâmetro de associação não depende dos pontos de corte das marginais. Posteriormente, K. V. Mardia chamou este *modelo de distribuição tipo-contingência* ou *Tipo-C* (*contingency-type distribution or C-Type distribution*) ou de superfície de associação constante (18;19). Para o caso em que o desfecho é univariado, J.M.G. Fachel (3) apresentou um

modelo para estimar a *razão de chances generalizada*, que é o parâmetro ψ da distribuição Tipo-C. Assim, supondo que os dados foram gerados por uma superfície subjacente contínua, o modelo produz uma estimativa única da associação em tabelas de contingência $r \times c$.

Apesar da diversidade de modelos para resposta ordinal, neste trabalho são explorados aspectos da razão de chances generalizada e do modelo de odds proporcionais. Resultados empíricos (4) sugerem que, para tabelas de contingência 2 x *k*, as razões de chances estimadas mediante este modelos são muito similares, mas não foi formalmente demonstrado, o que se pretende explorar no presente trabalho. Nas próximas seções são apresentados aspectos formais dos modelos de odds porporcionais e da razão de chances generalizada.

2.2 MODELO DE ODDS PROPORCIONAIS

Considere que Y representa um desfecho observado mediante k categorias ordenadas e x é um fator explanatório, que pode ser discreto ou contínuo. Quando o fator explanatório assume o valor x, a probabilidade condicional de se observar o desfecho em uma categoria menor ou igual a j é $\gamma_j(x) = P\left(Y \leq j \mid x\right)$. O modelo de odds proporcionais especifica que

$$log\ \frac{P\left(Y \leq j \mid x\right)}{1 - P\left(Y \leq j \mid x\right)} = log\ \frac{\gamma_j(x)}{1 - \gamma_j(x)} = \theta_j - \beta\, x; \quad \forall 1 \leq j < k, \qquad [1]$$

onde $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{k-1}$ e β são parâmetros desconhecidos que devem ser estimados.

Uma exigência importante do modelo é a independência entre o coeficiente de regressão β e as categorias do desfecho, também chamada de suposição de odds proporcionais ou de linhas paralelas.

Na sua forma mais simples, para um determinado ponto de corte *j* o modelo possui as mesmas suposições do modelo de regressão logística para resposta dicotômica, ou seja, postula que o logito da probabilidade $P\left(Y \leq j \mid x\right)$ está linearmente relacionado com x e que não existe interação entre os fatores explanatórios.

Neste trabalho estamos particularmente interessados no contexto em que há apenas um fator explanatório dicotômico. Assim, sem perda de generalidade, considere que x representa a exposição $(x=1)$ ou não exposição $(x=0)$ a um determinado fator de risco. Se a suposição de linhas paralelas estiver satisfeita, então a razão de chances associada ao evento $Y \leq j \mid x=1$, em relação ao evento $Y \leq j \mid x=0$, é

$$RC_{OP} = \frac{\dfrac{\gamma_j(1)}{1-\gamma_j(1)}}{\dfrac{\gamma_j(0)}{1-\gamma_j(0)}} = \frac{exp\left\{\theta_j - \beta \times 1\right\}}{exp\left\{\theta_j - \beta \times 0\right\}} = exp\left\{-\beta\right\}, \qquad [2]$$

que claramente não depende da categoria j do desfecho, $\forall\, j = 1,2,\cdots,k$. Isto significa que, para qualquer categoria do desfecho ordenado, a chance de observar uma resposta em uma categoria menor ou igual a j para um indivíduo exposto é $exp\{-\beta\}$ vezes a chance de um indivíduo não exposto, $\forall\, j = 1,2,\cdots,k-1$.

A suposição de linhas paralelas pode ser avaliada mediante um teste baseado em escores, mas, como o teste é pouco conservador, em algumas situações o resultado pode não ser confiável (20;21).

Os parâmetros do modelo podem ser estimados pelo método da máxima verossimilhança, utilizando o procedimento de mínimos quadrados iterativamente reponderados para resolver o sistema das equações de verossimilhança. Estes métodos encontram-se extensamente discutidos na literatura (2;10;22) e foram implementados na rotina PROC LOGISTIC do programa SAS, usada para o ajuste do modelo (20).

2.3 RAZÃO DE CHANCES GENERALIZADA

A razão de chances generalizada (3) é uma medida de associação ainda pouco conhecida, e postula que é possível determinar uma estimativa única da razão de chances para medir a associação entre um preditor e um desfecho, dispostos em uma tabela de contingência *r x c* .

Para definir o modelo, considere que S e T são variáveis aleatórias contínuas com função de distribuição conjunta $H(s,t) = P(S \le s, T \le t)$ e funções de distribuição marginais $F(s) = P(S \le s)$ e $G(t) = P(T \le t)$, respectivamente. As marginais contínuas podem ser dicotomizadas nos pontos arbitrários s e t, gerando a tabela de contingência 2 x 2, mostrada na Tabela 1. Assim, a razão de chances é definida por

$$\psi = \frac{H(s,t)[1 - F(s) - G(t) + H(s,t)]}{[F(s) - H(s,t)]\,[G(t) - H(s,t)]} \qquad [3]$$

e, de maneira equivalente, pode ser escrita como

$$(\psi - 1)\,[H(s,t)]^2 - [1 + [F(s) + G(t)](\psi - 1)]H(s,t) + \psi\, F(s)\, G(t) = 0;\ \psi > 0\,. \qquad [4]$$

Tabela 1. Distribuição de probabilidade conjunta para a tabela de contingência resultante da dicotomização das marginais S e T nos valores arbitrários s e t.

Categorias da variável S	Categorias da variável T		Total
	$T \le t$	$T > t$	
$S \le s$	$H(s,t)$	$F(s) - H(s,t)$	$F(s)$
$S > s$	$G(t) - H(s,t)$	$1 - F(s) - G(t) + H(s,t)$	$1 - F(s)$
Total	$G(t)$	$1 - G(t)$	1

A equação [4] possui uma única raiz (23), dada por

$$H(s,t) = \begin{cases} \dfrac{S(s,t) - \sqrt{[S(s,t)]^2 - 4\psi(\psi-1)F(s)G(t)}}{2(\psi-1)}, \text{ se } \psi \neq 1 \\ F(s)G(t), \text{ se } \psi = 1 \end{cases} \qquad [5]$$

onde $S(s,t) = 1 + (\psi - 1)[F(s) + G(t)]$.

A função $H(s,t)$ é denominada função de distribuição Tipo-C, onde ψ é um parâmetro desconhecido que representa a associação entre as variáveis. Para quaisquer valores s e t que dicotomizam as distribuições marginais, a razão de chances é constante, produzindo uma única estimativa para o parâmetro ψ. Por esta razão, ψ é chamado de *razão de chances generalizada*, pois generaliza a razão de chances calculada originalmente em tabelas de contingência 2 x 2, para tabelas *r* x *c*.

Em tabelas de contingência *r x c*, o parâmetro ψ pode ser estimado pelo método da máxima verossimilhança (3;18). Contudo, para evitar problemas numéricos ao avaliar a função de verossimilhança na vizinhança do valor $\psi = 1$, é

conveniente reescrever a equação [5] mediante a série de Taylor da expansão binomial (3). Assim, fazendo $\lambda = \psi - 1$, para $\psi \neq 1$,

$$H(s,t) = \frac{1}{2\lambda}\left[\left(1 + \lambda\left[F(s) + G(t)\right]\right) - \left[\left[1 + \lambda\left[F(s) + G(t)\right]\right]^2 - 4\lambda(\lambda+1)F(s)G(t)\right]^{\frac{1}{2}} \right] \qquad [6]$$

ou, equivalentemente,

$$H(s,t) = \frac{1}{2}\left[F(s) + G(t)\right] + \frac{1}{2\lambda}\left[1 - \left[1 + 2\lambda\left[F(s) + G(t) - 2F(s)G(t)\right] + \lambda^2\left[F(s) - G(t)\right]^2\right]^{\frac{1}{2}}\right]. \qquad [7]$$

Definindo $z = 1 + 2\lambda\left[F(x) + G(y) - 2F(x)G(y)\right] + \lambda^2\left[F(x) - G(y)\right]^2$ e escrevendo o termo $\left[1+z\right]^{\frac{1}{2}}$ da equação [7] em série de Taylor, segue que

$$H(s,t) = F(s)G(t) + \frac{\lambda U}{1+\lambda V} + \frac{\lambda^3 U^2}{(1+\lambda V)^3} + \frac{2\lambda^5 U^3}{(1+\lambda V)^5} + \frac{5\lambda^7 U^4}{(1+\lambda V)^7} + \frac{14\lambda^9 U^5}{(1+\lambda V)^9} + \frac{42\lambda^{11} U^6}{(1+\lambda V)^{11}} + \frac{132\lambda^{13} U^7}{(1+\lambda V)^{13}} + \cdots \qquad [8]$$

onde $U = F(s)G(t)[1-F(s)][1-G(t)]$ e $V = F(s)[1-G(t)] + [1-F(s)]G(t)$, para $\left|2\lambda\left[F(s) + G(t) - 2F(s)G(t)\right] + \lambda^2\left[F(s) - G(t)\right]^2\right| < 1$. Esta expressão alternativa para $H(s,t)$ converge para a expressão definida na equação [7], sendo extremamente útil para evitar problemas numéricos quando o parâmetro ψ assume valor na vizinhança de $\psi = 1$ (3).

A estimação da razão de chances generalizada ψ, pelo método da máxima verossimilhança, utilizando o método escore de Fisher (*Fisher Scoring method)* no processo iterativo, foi implementada em uma rotina computacional em linguagem Delphi, denominada CROSSPSI (24). Esta rotina também incorpora a expansão em série de Taylor para $H(s,t)$ definida na equação [8], para valores de na vizinhança $0{,}98 \leq \psi \leq 1{,}02$.

2.4 SIMULAÇÃO MONTE CARLO

Métodos de simulação constituem uma ferramenta extremamente poderosa nas mais variadas áreas de pesquisa, desenvolvimento e ensino. As simulações em computadores permitem replicar um experimento, fazendo alterações específicas nos parâmetros de interesse. Especialmente em problemas onde técnicas analíticas são inadequadas, a simulação é uma ferramenta poderosa e versátil, mas produz apenas resultados aproximados. Em outras palavras, como não produz resultados exatos, é uma técnica imprecisa (25).

De uma forma geral, a simulação pode ser definida como uma técnica para replicar um experimento associado a um modelo ou sistema. No entanto, a *simulação estocástica* é menos abrangente e envolve a amostragem de variáveis aleatórias com uma distribuição de probabilidade especificada. Como a amostragem de determinada distribuição de probabilidade envolve números aleatórios, a simulação estocástica é comumente chamada de simulação *Monte Carlo*. O termo "Monte Carlo" é uma alusão aos jogos dos cassinos da cidade de Monte Carlo, Mônaco, e foi usado como um código para o trabalho secreto que estava sendo desenvolvido por Von Neumann e Ulam em Los Alamos, durante a Segunda Guerra Mundial (25).

O estimador da razão de chances generalizada não possui uma forma explícita e, portanto, não é possível demonstrar analiticamente sua relação com a razão de chances estimada mediante o modelo de odds proporcionais. Em situações como esta, é bastante comum utilizar estudos de simulação Monte Carlo para comparar a equivalência, precisão ou eficiência de estimadores. Assim, para comparar as estimativas de razão de chances produzidas por estes modelos, é necessário planejar e conduzir um estudo de simulação Monte Carlo.

Nas simulações assumimos uma população com distribuição bivariada e contínua Tipo-C Normal (18;26). O algoritmo para gerar variáveis com esta distribuição, baseado no *procedimento da distribuição condicional* (*conditional distribution approach*), foi proposto por K.V. Mardia (26;27). Este algoritmo foi implementado no programa SAS. Aspectos metodológicos dos geradores de números aleatórios utilizados pelo SAS estão disponíveis na documentação do programa, na seção "Functions and CALL Routines" do capítulo "SAS Language Reference: Dictionary" (20).

2.5 REFERÊNCIAS

(1) McCullagh P. Regression models for ordinal data. Journal of the Royal Statistical Society B 1980; 42(1):109-127.

(2) McCullagh P, Nelder JA. Generalized linear models. Second ed. New York: Chapman and Hall, 1989.

(3) Fachel JMG. The C-type distribution as an underlying model for categorical data and its use in factor analysis. PhD Dissertation. London School of Economics and Political Sciences, University of London, 1986.

(4) Lopes LS, Biasoli PK, Vigo A, Fachel JMG. A razão de chances generalizada e sua comparação com os parãmetros dos modelos de regressão logística ordinal. Livro de Resumos, XIII Salão de Iniciação Científica, UFRGS, p.13. 2001.

(5) Cureton EE. Psychometrics. In: Kruskal WH, Tanur JM, editors. International encyclopedia of statistics. New York: The Free Press, 1978: 764-782.

(6) Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967; 54(1):167-179.

(7) Hastie TJ, Botha JL, Schnitzler CM. Regression with an ordered categorical response. Statistics in Medicine 1989; 8:785-794.

(8) Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. International Journal of Epidemiology 1997; 26(6):1323-1333.

(9) Anderson A. Regression and ordered categorical variables. Journal of the Royal Statistical Society B 1984; 46(1):1-30.

(10) Hastie TJ, Tibshirani RJ. Non-parametric logistic and proportional odds regression. Applied Statistics 1987; 36(2):260-276.

(11) Hastie TJ, Tibshirani RJ. Generalized additive models. New York: Chapman and Hall, 1990.

(12) Kauermann G, Tutz G. Semi- and nonparametric modeling of ordinal data. Journal of Computational and Graphical Statistics 2003; 12(1):176-196.

(13) Fu LM, Simpson DG. Conditional risk models for ordinal response data: simultaneous logistic regression analysis and generalized score tests. Journal of Statistical Planning and Inference 2002; 108(1-2):201-217.

(14) Lindsey PJ, Kaufmann J. Analysis of a longitudinal ordinal response clinical trial using dynamic models. Journal of the Royal Statistical Society Series C-Applied Statistics 2004; 53:523-537.

(15) Glonek GFG, McCullagh P. Multivariate logistic models. Journal of Royal Statistical Society B 1995; 57(3):533-546.

(16) Dale JR. Global cross-ratio models for bivariate, discrete, ordered variables. Biometrics 1986; 42:909-917.

(17) Pearson K, Heron D. On theories of association. Biometrika 1913; 9:159-315.

(18) Mardia KV. Families of bivariate distributions. London: Griffin´s Statistical Monographs & Courses, Griffin, 1970.

(19) Mosteller F. Association and estimation in contingency tables. Journal of the American Statistical Association 1968; 63:1-28.

(20) SAS Institute Inc. SAS OnlineDoc, Version Eight. Cary, NC: SAS Institute Inc., 1999.

(21) Harrel Jr FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer, 2001.

(22) Vigo A. Análise de experimentos industriais com respostas categóricas ordenadas: método de Taguchi e modelo de McCullagh. Dissertação de Mestrado. IMECC, Universidade Estadual de Campinas, 1994.

(23) Plackett RL. A class of bivariate distributions. Journal of the American Statistical Association 1965; 60:516-522.

(24) D´Ávila ER, Fachel JMG. Programa CROSSPSI para calcular o coeficiente de correlação Tipo-C. Livro de Resumos, X Salão de Iniciação Científica, UFRGS, p.25. 1998.

(25) Rubinstein RY. Simulation and the Monte Carlo Method. New York: Wiley, 1981.

(26) Johnson M.E. Multivariate statistical simulation. New York: John Wiley & Sons, Inc., 1987.

(27) Mardia KV. Some contributions to contingency-type bivariate distributions. Biometrika 1967; 54(1):235-249.

3 ARTIGOS

Artigo 1: Estudo de simulação Monte Carlo para comparar as razões de chances estimadas através dos modelos de odds proporcionais e razão de chances generalizada

Artigo 2: Comparação empírica do modelo de odds proporcionais e da razão de chances generalizada para estimar a associação da temperatura ambiente e da obesidade com as glicemias durante um teste oral de tolerância à glicose

3.1 ARTIGO 1

Estudo de simulação Monte Carlo para comparar as razões de chances estimadas através dos modelos de odds proporcionais e razão de chances generalizada

Álvaro Vigo [1,2]

Jandyra M. G. Fachel [1,3]

1. Departamento de Estatística, Universidade Federal do Rio Grande do Sul
2. Doutorando do Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul
3. Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul

Correspondência: Prof. Álvaro Vigo, Departamento de Estatística, UFRGS, Av. Bento Gonçalves, 9500, Prédio 43-111, Bairro Agronomia, 91509-900, Porto Alegre, RS, BRASIL. Telefone: + 55 51 3316-6225 / 3316-6189 FAX: + 55 51 3316-7301

E_mail: vigo@orion.ufrgs.br

Número de palavras: Resumo:101; Texto:2174 + 5 Tabelas + 8 Figuras

Abstract

Ordinal outcomes occur with great frequency in the medical and epidemiological research, and the proportional odds model has been used to describe the relationship between an ordinal response and the predictors. Although little known, the model of the generalized odds ratio can be a powerful alternative. Monte Carlo simulations confirm that in contingency tables with an outcome with three ordered categories and a dichotomous explanatory factor, the estimates produced by the proportional odds model and by the generalized odds ratio are equivalents. Comparing the corresponding mean square errors, these estimators have the same efficiency.

Keywords: ordinal response, ordered endpoints, ordinal outcomes, odds ratio, generalized odds ratio, proportional odds, simulation, Monte Carlo

Resumo

Desfechos ordinais ocorrem com grande freqüência na pesquisa médica e epidemiológica, e o modelo de odds proporcionais tem sido bastante utilizado para descrever a relação entre um desfecho ordinal e os preditores. Embora pouco conhecido, o modelo da razão de chances generalizada pode ser uma alternativa poderosa. Simulações Monte Carlo confirmam que em tabelas de contingência com um desfecho com três categorias ordenadas e um fator explanatório dicotômico, as estimativas da razão de chances produzidas pelo modelo de odds proporcionais e da razão de chances generalizada são equivalentes. Comparando os correspondentes erros quadráticos médios, estes estimadores têm a mesma eficiência.

Palavras chaves: desfecho ordinal, razão de chances generalizada, odds proporcionais, resposta ordinal, simulação, Monte Carlo

Introdução

Desfechos ordinais são bastante comuns na pesquisa médica e epidemiológica. No entanto, a utilização de métodos de análise que incorporam a estrutura ordenada das categorias ainda é pouco freqüente, podendo levar à perda de informação disponível nos dados. Dentre os principais métodos de análise, atenção especial merecem os modelos para resposta ordinal, pois permitem estimar magnitude e direção de efeitos de múltiplos fatores explanatórios. Diversos tipos de modelos estão disponíveis (1;2), mas o modelo de odds proporcionais tem sido utilizado com maior freqüência.

No contexto em que existe apenas um fator explanatório dicotômico, comparações empíricas sugerem que a razão de chances estimada pelo modelo de odds proporcionais é equivalente à razão de chances generalizada (3;4). O objetivo deste trabalho é comparar as estimativas das razões de chances produzidas por estes modelos quando o fator explanatório é dicotômico, mediante um estudo de simulação Monte Carlo. Aspectos das suposições dos modelos também são considerados.

Métodos

O modelo comumente usado para descrever a relação funcional entre um desfecho ordinal e um conjunto de fatores explanatórios foi inicialmente descrito por S.H. Walker e D.B. Duncan (5), sendo, posteriormente, chamado de *modelo de odds proporcionais* por P. McCullagh (2;6).

Neste modelo, considere que Y representa um desfecho observado mediante k categorias ordenadas e x é um fator explanatório, que pode ser discreto ou contínuo. Quando o fator explanatório assume o valor x, a probabilidade condicional

de se observar o desfecho em uma categoria menor ou igual a j é $\gamma_j(x) = P\left(Y \leq j \mid x\right)$. O modelo de odds proporcionais especifica que

$$log\ \frac{P\left(Y \leq j \mid x\right)}{1 - P\left(Y \leq j \mid x\right)} = log\ \frac{\gamma_j(x)}{1-\gamma_j(x)} = \theta_j - \beta\, x; \quad \forall 1 \leq j < k\,, \qquad [1]$$

onde $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{k-1}$ e β são parâmetros desconhecidos que devem ser estimados.

Uma exigência importante do modelo é a independência entre o coeficiente de regressão β e as categorias do desfecho, também chamada de suposição de odds proporcionais ou de linhas paralelas.

Na sua forma mais simples, para um determinado ponto de corte *j*, o modelo possui as mesmas suposições do modelo de regressão logística para resposta dicotômica, ou seja, postula que o logito (*logit*) da probabilidade $P\left(Y \leq j \mid x\right)$ está linearmente relacionado com x e que não existe interação entre os fatores explanatórios.

Neste trabalho estamos particularmente interessados no contexto em que há apenas um fator explanatório dicotômico. Assim, sem perda de generalidade, considere que x representa a exposição $(x=1)$ ou não exposição $(x=0)$ a um determinado fator de risco. Se a suposição de linhas paralelas estiver satisfeita, então a razão de chances associada ao evento $Y \leq j \mid x = 1$, em relação ao evento $Y \leq j \mid x = 0$, é

$$RC_{OP} = \frac{\dfrac{\gamma_j(1)}{1-\gamma_j(1)}}{\dfrac{\gamma_j(0)}{1-\gamma_j(0)}} = \frac{exp\left\{\theta_j - \beta \times 1\right\}}{exp\left\{\theta_j - \beta \times 0\right\}} = exp\left\{-\beta\right\}, \qquad [2]$$

que claramente não depende da categoria j do desfecho, $\forall\, j = 1,2,\cdots,k$. Isto significa que, para qualquer categoria do desfecho ordenado, a chance de observar uma resposta em uma categoria menor ou igual a j para um indivíduo exposto é $exp\{-\beta\}$ vezes a chance de um indivíduo não exposto, $\forall\, j = 1,2,\cdots,k-1$.

A suposição de linhas paralelas pode ser avaliada mediante um teste baseado em escores, mas, como o teste é pouco conservador, em algumas situações o resultado pode não ser confiável (7;8).

Os parâmetros do modelo podem ser estimados pelo método da máxima verossimilhança, utilizando o procedimento de mínimos quadrados iterativamente reponderados para resolver as equações de verossimilhança. Aspectos da estimação dos parâmetros estão extensamente discutido na literatura (2;9;10), e o procedimento PROC LOGISTIC do programa SAS (*Statistical Analysis System*) foi usado para ajustar o modelo (7).

A razão de chances generalizada (11) é uma medida de associação ainda pouco conhecida, e postula uma estimativa única da razão de chances para medir a associação entre um preditor e um desfecho, dispostos em uma tabela de contingência *r x c*. Este modelo de associação supõe que a distribuição subjacente dos dados da tabela de contingência pertence à *família de distribuições bivariadas Tipo-Contingência* ou *Tipo-C* (12;13). A veracidade desta suposição pode ser avaliada mediante um teste de aderência baseado na estatística Qui-quadrado de Pearson.

Para definir o modelo, considere que S e T são variáveis aleatórias contínuas com função de distribuição conjunta $H(s,t) = P(S \le s, T \le t)$ e funções de distribuição marginais $F(s) = P(S \le s)$ e $G(t) = P(T \le t)$, respectivamente. As marginais contínuas podem ser dicotomizadas nos pontos arbitrários s e t, gerando

a tabela de contingência 2 x 2, mostrada na Tabela 1. Assim, a razão de chances é definida por

$$\psi = \frac{H(s,t)[1 - F(s) - G(t) + H(s,t)]}{[F(s) - H(s,t)]\,[G(t) - H(s,t)]} \qquad [3]$$

e, de maneira equivalente, pode ser escrita como

$$(\psi - 1)\,[H(s,t)]^2 - [1 + [F(s) + G(t)](\psi - 1)]H(s,t) + \psi\, F(s)\, G(t) = 0;\ \psi > 0\,. \qquad [4]$$

A equação [4] possui uma única raiz (13), dada por

$$H(s,t) = \begin{cases} \dfrac{S(s,t) - \sqrt{[S(s,t)]^2 - 4\psi(\psi - 1)F(s)G(t)}}{2(\psi - 1)}, \text{ se } \psi \neq 1 \\ F(s)G(t), \text{ se } \psi = 1 \end{cases} \qquad [5]$$

onde $S(s,t) = 1 + (\psi - 1)[F(s) + G(t)]$.

A função $H(s,t)$ é denominada função de distribuição Tipo-C, onde ψ é um parâmetro desconhecido que representa a associação entre as variáveis. Para quaisquer valores s e t que dicotomizam as distribuições marginais, a razão de chances é constante, produzindo uma única estimativa para o parâmetro ψ. Por esta razão, ψ é chamado de *razão de chances generalizada*.

Em tabelas de contingência *r x c*, o parâmetro ψ pode ser estimado pelo método da máxima verossimilhança (11). Contudo, para evitar problemas numéricos ao avaliar a função de verossimilhança na vizinhança do valor $\psi = 1$, é conveniente reescrever a equação [5] mediante a série de Taylor da expansão binomial. Assim, fazendo $\lambda = \psi - 1$, para $\psi \neq 1$,

$$H(s,t) = \frac{1}{2\lambda}\left[(1 + \lambda[F(s) + G(t)]) - \left[[1 + \lambda[F(s) + G(t)]]^2 - 4\lambda(\lambda + 1)F(s)G(t)\right]^{\frac{1}{2}} \right] \qquad [6]$$

ou, equivalentemente,

$$H(s,t)=\frac{1}{2}\left[F(s)+G(t)\right]+\frac{1}{2\lambda}\left[1-\left[1+2\lambda\left[F(s)+G(t)-2F(s)G(t)\right]+\right.\right.$$
$$\left.\left.+\lambda^2\left[F(s)-G(t)\right]^2\right]^{\frac{1}{2}}\right]. \qquad [7]$$

Definindo $z=1+2\lambda\left[F(x)+G(y)-2F(x)G(y)\right]+\lambda^2\left[F(x)-G(y)\right]^2$ e escrevendo o termo $\left[1+z\right]^{\frac{1}{2}}$ da equação [7] em série de Taylor, segue que

$$H(s,t)=F(s)G(t)+\frac{\lambda U}{1+\lambda V}+\frac{\lambda^3U^2}{(1+\lambda V)^3}+\frac{2\lambda^5U^3}{(1+\lambda V)^5}+\frac{5\lambda^7U^4}{(1+\lambda V)^7}+$$
$$+\frac{14\lambda^9U^5}{(1+\lambda V)^9}+\frac{42\lambda^{11}U^6}{(1+\lambda V)^{11}}+\frac{132\lambda^{13}U^7}{(1+\lambda V)^{13}}+\cdots \qquad [8]$$

onde $U=F(s)G(t)[1-F(s)][1-G(t)]$ e $V=F(s)[1-G(t)]+[1-F(s)]G(t)$, para $\left|2\lambda\left[F(s)+G(t)-2F(s)G(t)\right]+\lambda^2\left[F(s)-G(t)\right]^2\right|<1$.

Esta expressão alternativa para $H(s,t)$ converge para a expressão definida na equação [7], sendo extremamente útil para evitar problemas numéricos quando ψ assume valores próximos do valor $\psi=1$ (11).

A estimação da razão de chances generalizada ψ, pelo método da máxima verossimilhança, utilizando o método escore de Fisher (*Fisher scoring method)* no processo iterativo, foi implementada em uma rotina computacional em Delphi, denominada CROSSPSI (14). Esta rotina também incorpora a expansão em série de Taylor para $H(s,t)$ definida na equação [8], para valores de na vizinhança $0{,}98\le\psi\le1{,}02$.

Um estudo de simulação Monte Carlo foi conduzido para comparar as estimativas de razão de chances mediante os modelos de odds proporcionais e da razão de chances generalizada, para o contexto em que o fator explanatório é binário e o desfecho tem três categorias ordenadas. A simulação consistiu em gerar 10.000 amostras de tamanho 500 da *distribuição bivariada Tipo-C normal*, com

parâmetros de associação $\psi = 1$, $\psi = 2$, $\psi = 4$ e $\psi = 10$. O algoritmo usado para gerar os dados com distribuição Tipo-C, baseado no procedimento da distribuição condicional (15;16), foi implementado em linguagem SAS.

Para cada valor do parâmetro ψ, depois de gerar os dados da distribuição conjunta, a primeira distribuição marginal foi dicotomizada na mediana populacional. Esta variável binária representa um possível fator explanatório, indicando, por exemplo, a exposição ou não a um certo fator. A outra distribuição marginal foi categorizada de duas formas simétricas, porém com diferentes concentrações na categoria central. Na primeira categorização, chamada de simétrica suave, 25% da população foi alocada a cada uma das categorias extremas, enquanto que a categoria central contempla 50% da população. Na outra categorização, mais concentrada em torno da média, cada uma das categorias extremas contempla 15% da população, enquanto que a categoria central contempla os 70% restantes. Esta variável representa um desfecho com três categorias ordenadas.

Assim, o delineamento do Estudo Monte Carlo contempla quatro valores do verdadeiro parâmetro ψ (razão de chances) que, combinados com as duas formas de categorização das marginais, definem oito contextos para os quais se deseja comparar as estimativas da razão de chances. Para cada contexto foram geradas 10.000 tabelas de contingência 2 x 3, com tamanho de amostra 500, para as quais são ajustados os modelos.

As simulações foram realizadas com um microcomputador Pentium 4 S333, com 512 MB de memória RAM e sistema operacional Windows 2000 Professional. As rotinas computacionais foram implementadas no programa estatístico SAS – Statistical Analysis System, Versão 8, exigindo considerável tempo de processamento. Com exceção do ajuste do modelo da razão de chances generalizada, realizado mediante a rotina CROSSPSI, Versão 2, e da versão final

dos gráficos, gerada pelo programa S-Plus, Versão 4, todas as análises foram realizadas através do programa SAS.

As comparações das estimativas dos parâmetros são realizadas através do erro quadrático médio entre as estimativas e, também, em relação ao verdadeiro valor de ψ. Os modelos também são comparados quanto ao atendimento da suposição de linhas paralelas (odds proporcionais) e de que os dados se ajustam a uma distribuição subjacente Tipo-C normal (razão de chances generalizada). Este teste de aderência, cuja hipótese nula especifica que os dados da tabela de contingência se ajustam à distribuição Tipo-C normal, é baseado na estatística Qui-Quadrado de Pearson. Como as freqüências esperadas sob a hipótese nula dependem da estimativa da razão de chances generalizada (11), então a distribuição de referência da estatística de teste é qui-quadrado com 1 grau de liberdade. O teste também foi implementado na linguagem do programa SAS.

Resultados

O estudo de simulação mostrou, para os quatro valores de ψ fixados (1, 2, 4 e 10), e em ambas formas de categorização do desfecho (suave e concentrada), que aproximadamente 5% das 10.000 tabelas de contingência com n=500 violaram a suposição de linhas paralelas e/ou não se ajustaram à distribuição Tipo-C normal. A maior taxa de violação (5,76%) ocorreu para o caso $\psi = 10$ e categorização simétrica suave, enquanto que a menor (4,94%) ocorreu para $\psi = 4$ e categorização concentrada. No entanto, as violações das suposições ocorreram simultaneamente e, em raras ocasiões (< 0,20%), para apenas um modelo. De maneira geral, observou-se que a taxa de violação destas suposições é um pouco menor quando o desfecho tem maior concentração em torno da média.

As Tabelas 2 a 5 apresentam as estimativas de ψ para diferentes categorizações do desfecho e, também, considerando toda amostra e separadamente para os casos que atendem a suposição de linhas paralelas e se ajustam à distribuição Tipo-C normal ou violam ao menos uma destas suposições. Para $\psi = 10$ e categorização simétrica concentrada, apenas 9.968 tabelas de contingência foram consideradas, pois o programa CROSSPSI não pode estimar a razão de chances generalizada nas outras 32 tabelas, devido a um erro numérico de ponto flutuante.

Em todas as simulações, as estimativas de ψ mediante os modelos de odds proporcionais são, em média, muito próximas daquelas produzidas pela razão de chances generalizada e as correspondentes estimativas do desvio padrão são praticamente idênticas. No entanto, foram observadas discrepâncias em relação ao verdadeiro valor de ψ, que, em geral, diminuem para a categorização do desfecho simétrica concentrada. O mesmo comportamento é observado para os correspondentes erros quadráticos médios. Quando avaliadas separadamente para tabelas de contingência que satisfazem a suposição de linhas paralelas e se ajustam à distribuição Tipo-C normal, as estimativas de razão de chances produzidas pelos modelos não diferem substancialmente, mas os erros quadráticos médios (e também as variâncias) são visivelmente superiores quando violam no mínimo uma destas suposições. Estes comportamentos, bem como das distribuições empíricas dos estimadores, são ilustrados nas Figuras 1 a 8.

Discussão

Desfechos ordinais ocorrem com grande freqüência na pesquisa médica e epidemiológica. O modelo comumente usado para descrever a relação entre um desfecho ordinal e os preditores é o modelo de odds proporcionais, possivelmente

porque o ajuste de modelos mais sofisticados não está disponível nos programas estatísticos usuais.

As simulações Monte Carlo confirmam que existe uma forte conexão entre o modelo de odds proporcionais e a razão de chances generalizada. Assim, para tabelas de contingência com um desfecho com três categorias ordenadas e um fator explanatório dicotômico, as estimativas da razão de chances produzidas por estes modelos são praticamente idênticas.

Para todos os valores fixados da verdadeira razão de chances ψ, as estimativas pontuais e do desvio padrão são muito similares. Entretanto, o erro quadrático médio é maior quando a suposição de linhas paralelas não é atendida e/ou quando os dados não se ajustam à distribuição Tipo-C normal. A eficiência entre os estimadores da razão de chances, pela razão de chances generalizada e pelo modelo de odds proporcionais, pode ser avaliada mediante a divisão dos correspondentes erros quadráticos médios calculados em relação ao verdadeiro ψ (17). Em todos os contextos considerados na simulação, esta razão é praticamente igual a 1, sugerindo que, quando comparados entre si, os estimadores são igualmente eficientes.

Se o desfecho ordinal surge da categorização de uma variável quantitativa, a forma de categorização parece importante, pois, exceto para $\psi = 10$, aumentando a concentração dos dados na categoria central a estimativa média está mais próxima da verdadeira razão de chances (menor viés), porém com menor precisão.

A equivalência dos modelos, para estimar a razão de chances, permite obter uma estimativa única da associação em tabelas de contingência, através da razão de chances generalizada. Além disso, permite interpretar esta razão de chances generalizada na forma usual, como é feita no modelo de odds proporcionais, podendo vista como uma generalização da razão de chances para tabelas 2 x 2. Na

prática, a razão de chances generalizada poderia ser usada em uma etapa inicial da investigação, para identificar associações relevantes em tabelas *r* x *c*, da mesma forma que a razão de chances é utilizada no caso de tabelas de contingência 2 x 2.

Os resultados são consistentes com aqueles reportados por A.M. Quiroga, sobre a robustez do coeficiente de correlação policórica como medida de associação para variáveis ordinais (18).

Embora as estimativas geradas pelos modelos sejam extremamente similares, chamam atenção as discrepâncias observadas em relação aos verdadeiros valores do parâmetro de associação, fixados nas simulações. Estas diferenças podem ser decorrentes da forma de categorização das marginais, e uma investigação mais profunda precisa ser realizada para avaliar a presença de viés. Na seqüência do estudo o resultado deve ser estendido para tabelas 2 x *k*, $k > 3$, bem como para outras formas de categorização do desfecho e do preditor.

Referências

(1) Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. International Journal of Epidemiology 1997; 26(6):1323-1333.

(2) McCullagh P, Nelder JA. Generalized linear models. Second ed. New York: Chapman and Hall, 1989.

(3) Lopes LS, Biasoli PK, Vigo A, Fachel JMG. A razão de chances generalizada e sua comparação com os parâmetros dos modelos de regressão logística ordinal. Livro de Resumos, XIII Salão de Iniciação Científica, UFRGS, p.13. 2001.

(4) Vigo A, Fachel JMG. Comparação empírica do modelo de odds proporcionais e da razão de chances generalizada para estimar a associação da

temperatura ambiente e da obesidade com as glicemias durante um teste oral de tolerância à glicose. Manuscrito não publicado, 2004.

(5) Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967; 54(1):167-179.

(6) McCullagh P. Regression models for ordinal data. Journal of the Royal Statistical Society B 1980; 42(1):109-127.

(7) SAS Institute Inc. SAS OnlineDoc, Version Eight. Cary, NC: SAS Institute Inc., 1999.

(8) Harrel Jr FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer, 2001.

(9) Hastie TJ, Tibshirani RJ. Non-parametric logistic and proportional odds regression. Applied Statistics 1987; 36(2):260-276.

(10) Vigo A. Análise de experimentos industriais com respostas categóricas ordenadas: método de Taguchi e modelo de McCullagh. Dissertação de Mestrado. IMECC, Universidade Estadual de Campinas, 1994.

(11) Fachel JMG. The C-type distribution as an underlying model for categorical data and its use in factor analysis. PhD Dissertation. London School of Economics and Political Sciences, University of London, 1986.

(12) Mardia KV. Families of bivariate distributions. London: Griffin´s Statistical Monographs & Courses, Griffin, 1970.

(13) Plackett RL. A class of bivariate distributions. Journal of the American Statistical Association 1965; 60:516-522.

(14) D´Ávila ER, Fachel JMG. Programa CROSSPSI para calcular o coeficiente de correlação Tipo-C. Livro de Resumos, X Salão de Iniciação Científica, UFRGS, p.25, 1998.

(15) Johnson M.E. Multivariate statistical simulation. New York: John Wiley & Sons, Inc., 1987.

(16) Mardia KV. Some contributions to contingency-type bivariate distributions. Biometrika 1967; 54(1):235-249.

(17) Azzalini A. Statistical inference based on the likelihood. London: Champman & Hall, 1996.

(18) Quiroga AM. Studies of the polychoric correlation and other correlation measures for ordinal variables. PhD Dissertation. Uppsala University, Sweden, 1992.

Tabela 1. Distribuição de probabilidade conjunta para a tabela de contingência resultante da dicotomização das marginais S e T nos valores arbitrários s e t.

Categorias da variável S	Categorias da variável T		Total
	$T \leq t$	$T > t$	
$S \leq s$	$H(s,t)$	$F(s) - H(s,t)$	$F(s)$
$S > s$	$G(t) - H(s,t)$	$1 - F(s) - G(t) + H(s,t)$	$1 - F(s)$
Total	$G(t)$	$1 - G(t)$	1

Tabela 2. Resultados da simulação Monte Carlo para o parâmetro $\psi = 1$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) e do erro quadrático médio (EQM), em 10.000 amostras de tamanho 500 e separadamente para os casos em que as suposições de linhas paralelas e de que os dados se ajustam à distribuição Tipo-C normal estão ou não satisfeitas, bem como para as categorizações do desfecho simétrica suave e simétrica concentrada.

Estimativa	Simétrica suave			Simétrica concentrada		
	Toda a amostra	Satisfeitas[1]	Não satisfeitas[2]	Toda a amostra	Satisfeitas[1]	Não satisfeitas[2]
n	10.000	9.446	554	10.000	9.465	535
RCG						
Média ($\hat{\psi}_{RCG}$)	0,970	0,970	0,984	0,981	0,981	0,974
Viés	-0,030	-0,030	-0,016	-0,019	-0,019	-0,026
Desvio padrão	0,173	0,173	0,181	0,197	0,198	0,177
MOP						
Média ($\hat{\psi}_{MOP}$)	0,970	0,969	0,984	0,981	0,981	0,974
Viés	-0,030	-0,031	-0,016	-0,019	-0,019	-0,026
Desvio padrão	0,173	0,173	0,181	0,197	0,198	0,177
EQM						
$(\hat{\psi}_{RCG} - \hat{\psi}_{MOP})^2$	$1{,}616 \times 10^{-9}$	$2{,}115 \times 10^{-10}$	$2{,}561 \times 10^{-8}$	$7{,}823 \times 10^{-10}$	$5{,}643 \times 10^{-10}$	$4{,}638 \times 10^{-9}$
$(\hat{\psi}_{RCG} - \psi)^2$	0,031	0,031	0,033	0,039	0,040	0,032
$(\hat{\psi}_{MOP} - \psi)^2$	0,031	0,031	0,033	0,039	0,040	0,032

[1] – Suposição de linhas paralelas e ajuste à distribuição Tipo-C normal satisfeitas

[2] – Suposição de linhas paralelas e/ou ajuste à distribuição Tipo-C normal não satisfeitas

Tabela 3. Resultados da simulação Monte Carlo para o parâmetro $\psi = 2$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) e do erro quadrático médio (EQM), em 10.000 amostras de tamanho 500 e separadamente para os casos em que as suposições de linhas paralelas e de que os dados se ajustam à distribuição Tipo-C normal estão ou não satisfeitas, bem como para as categorizações do desfecho simétrica suave e simétrica concentrada.

Estimativa	Simétrica suave			Simétrica concentrada		
	Toda a amostra	Satisfeitas[1]	Não satisfeitas[2]	Toda a amostra	Satisfeitas[1]	Não satisfeitas[2]
n	10.000	9.493	507	10.000	9.460	540
RCG						
Média ($\hat{\psi}_{RCG}$)	1,924	1,924	1,931	1,965	1,964	1,988
Viés	-0,076	-0,076	-0,069	-0,035	-0,036	-0,012
Desvio padrão	0,353	0,352	0,365	0,409	0,408	0,440
MOP						
Média ($\hat{\psi}_{MOP}$)	1,924	1,924	1,933	1,965	1,964	1,989
Viés	-0,076	-0,076	-0,067	-0,035	-0,036	-0,011
Desvio padrão	0,353	0,352	0,365	0,410	0,408	0,440
EQM						
$(\hat{\psi}_{RCG} - \hat{\psi}_{MOP})^2$	2,601 x 10^{-7}	1,167 x 10^{-7}	2,944 x 10^{-6}	1,229 x 10^{-7}	4,800 x 10^{-8}	1,435 x 10^{-6}
$(\hat{\psi}_{RCG} - \psi)^2$	0,130	0,130	0,137	0,169	0,167	0,193
$(\hat{\psi}_{MOP} - \psi)^2$	0,130	0,130	0,138	0,169	0,167	0,193

[1] – Suposição de linhas paralelas e ajuste à distribuição Tipo-C normal satisfeitas

[2] – Suposição de linhas paralelas e/ou ajuste à distribuição Tipo-C normal não satisfeitas

Tabela 4. Resultados da simulação Monte Carlo para o parâmetro $\psi = 4$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) e do erro quadrático médio (EQM), em 10.000 amostras de tamanho 500 e separadamente para os casos em que as suposições de linhas paralelas e de que os dados se ajustam à distribuição Tipo-C normal estão ou não satisfeitas, bem como para as categorizações do desfecho simétrica suave e simétrica concentrada.

Estimativa	Simétrica suave			Simétrica concentrada		
	Toda a amostra	Satisfeitas[1]	Não satisfeitas[2]	Toda a amostra	Satisfeitas[1]	Não satisfeitas[2]
n	10.000	9.497	503	10.000	9.506	494
RCG						
Média ($\hat{\psi}_{RCG}$)	3,926	3,931	3,836	4,011	4,012	3,984
Viés	-0,074	-0,069	-0,164	0,011	0,012	-0,016
Desvio padrão	0,764	0,765	0,734	0,937	0,938	0,9222
MOP						
Média ($\hat{\psi}_{MOP}$)	3,927	3,931	3,841	4,011	4,013	3,987
Viés	-0,073	-0,069	-0,159	0,011	0,013	-0,013
Desvio padrão	0,764	0,765	0,735	0,937	0,938	0,9223
EQM						
$(\hat{\psi}_{RCG} - \hat{\psi}_{MOP})^2$	2,833 x 10^{-6}	1,239 x 10^{-6}	3,292 x 10^{-5}	9,228 x 10^{-7}	4,640 x 10^{-7}	9,751 x 10^{-6}
$(\hat{\psi}_{RCG} - \psi)^2$	0,589	0,590	0,564	0,879	0,880	0,849
$(\hat{\psi}_{MOP} - \psi)^2$	0,589	0,590	0,565	0,879	0,880	0,850

[1] – Suposição de linhas paralelas e ajuste à distribuição Tipo-C normal satisfeitas

[2] – Suposição de linhas paralelas e/ou ajuste à distribuição Tipo-C normal não satisfeitas

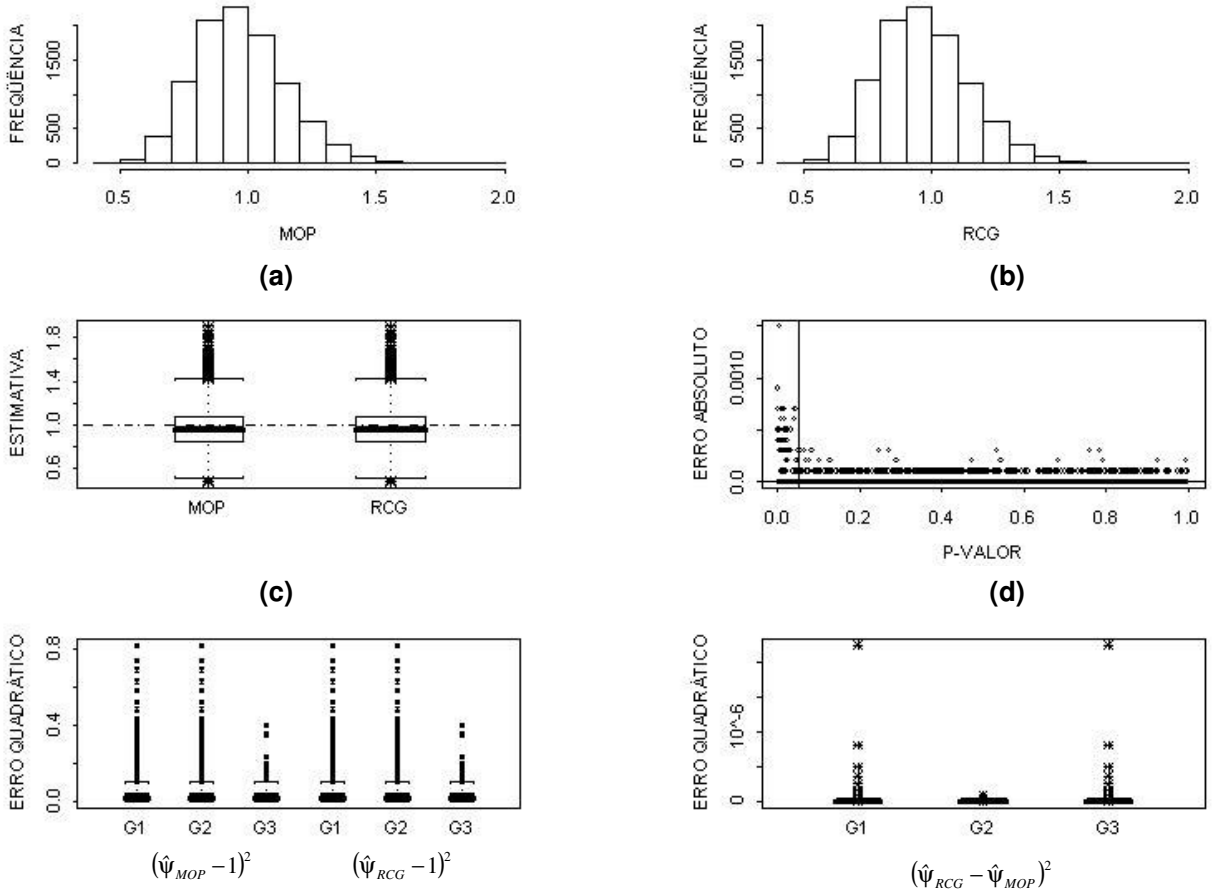
Tabela 5. Resultados da simulação Monte Carlo para o parâmetro $\psi = 10$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) e do erro quadrático médio (EQM), em 10.000 amostras de tamanho 500 e separadamente para os casos em que as suposições de linhas paralelas e de que os dados se ajustam à distribuição Tipo-C normal estão ou não satisfeitas, bem como para as categorizações do desfecho simétrica suave e simétrica concentrada.

Estimativa	Simétrica suave			Simétrica concentrada		
	Toda a amostra	Satisfeitas[1]	Não satisfeitas[2]	Toda a amostra	Satisfeitas[1]	Não satisfeitas[2]
n	10.000	9.424	576	9.968[3]	9.444	524
RCG						
Média ($\hat{\psi}_{RCG}$)	9,996	9,996	10,002	10,376	10,380	10,317
Viés	-0,004	-0,004	0,002	0,376	0,380	0,317
Desvio padrão	2,322	2,319	2,375	3,261	3,261	3,257
MOP						
Média ($\hat{\psi}_{MOP}$)	9,999	9,998	10,019	10,377	10,380	10,323
Viés	-0,001	-0,002	0,019	0,377	0,380	0,323
Desvio padrão	2,322	2,319	2,378	3,261	3,261	3,257
EQM						
$(\hat{\psi}_{RCG} - \hat{\psi}_{MOP})^2$	3,119 x 10^{-5}	1,297 x 10^{-5}	3,292 x 10^{-4}	5,812 x 10^{-6}	3,457 x 10^{-6}	4,826 x 10^{-5}
$(\hat{\psi}_{RCG} - \psi)^2$	5,391	5,377	5,629	10,774	10,779	10,686
$(\hat{\psi}_{MOP} - \psi)^2$	5,393	5,377	5,645	10,773	10,777	10,694

[1] – Suposição de linhas paralelas e ajuste à distribuição Tipo-C normal satisfeitas

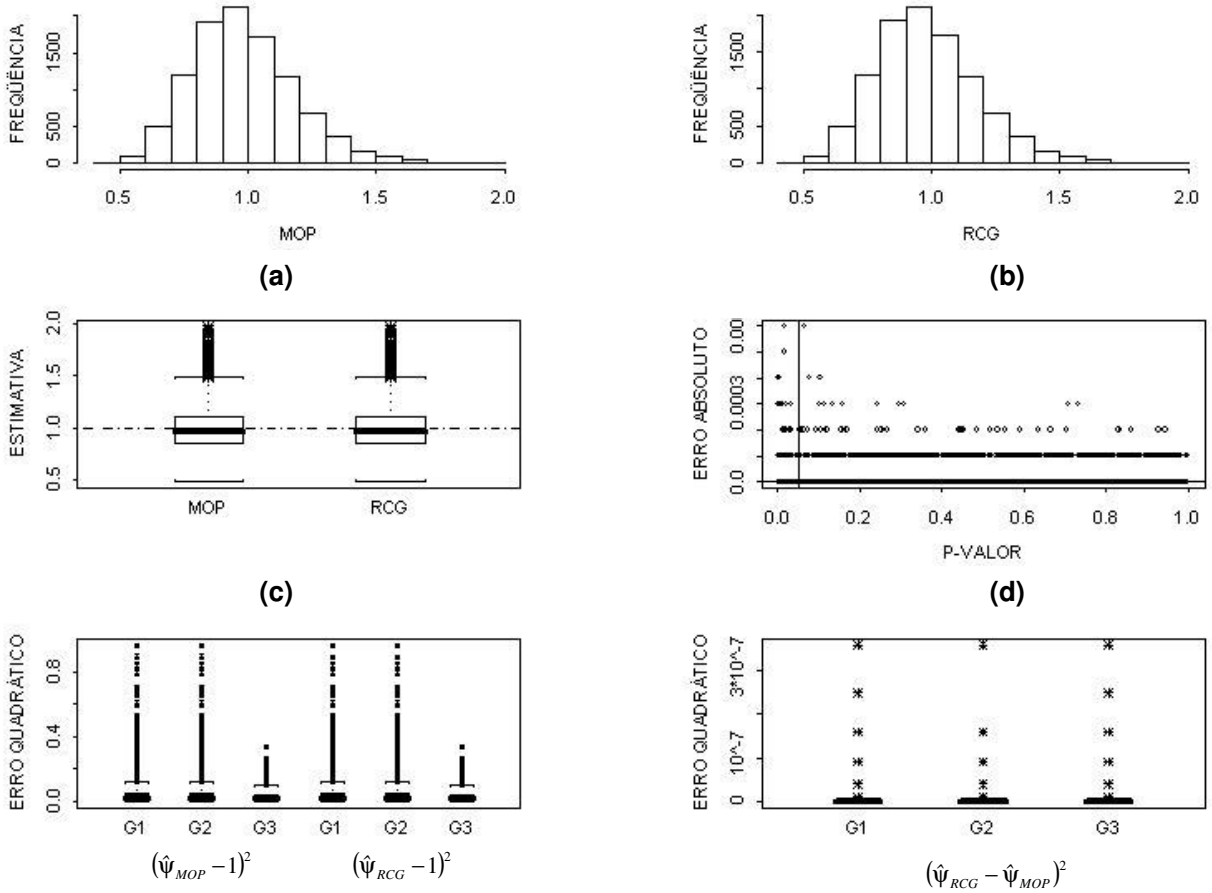
[2] – Suposição de linhas paralelas e/ou ajuste à distribuição Tipo-C normal não satisfeitas

[3] – 32 tabelas de contingência excluídas da análise devido ao erro numérico



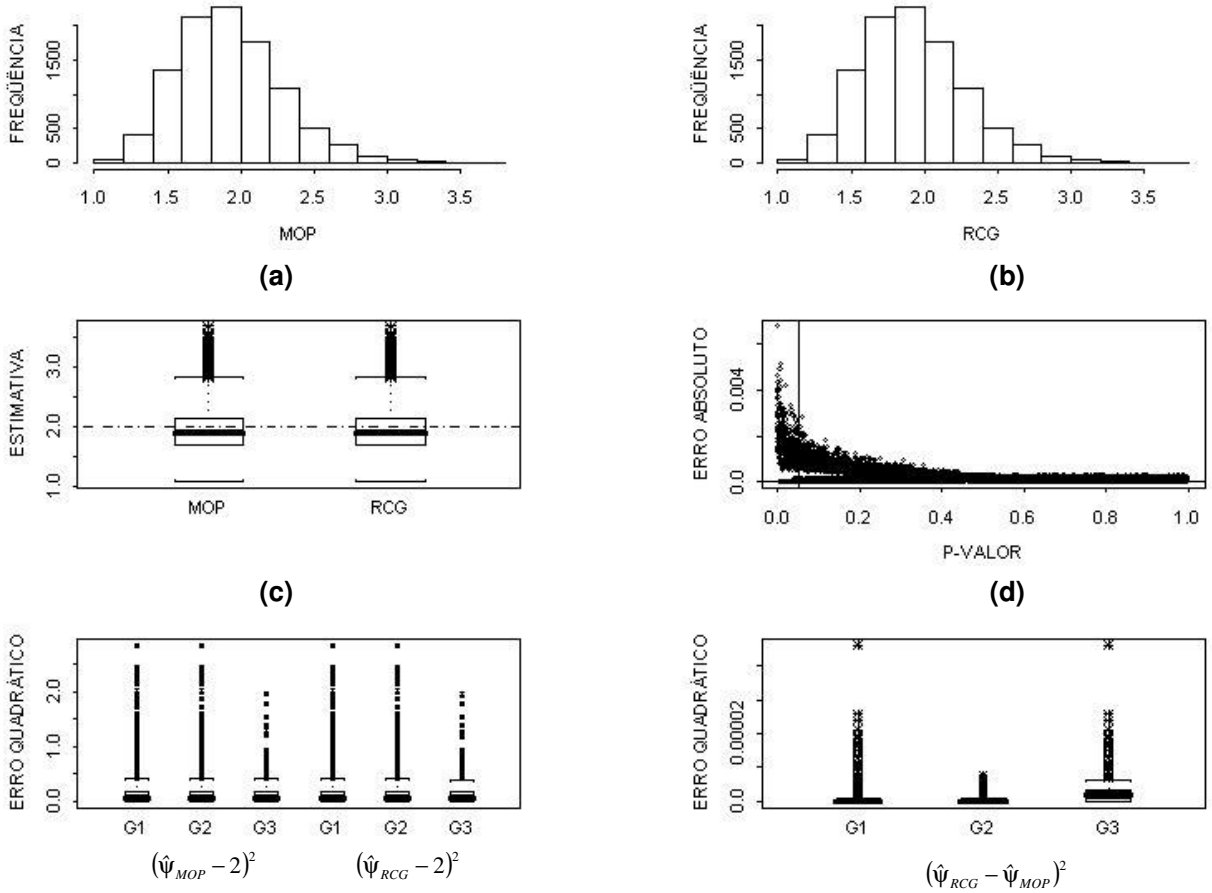
G1: Toda amostra G2: Suposições satisfeitas G3: Suposições não satisfeitas

Figura 1. Comportamento das estimativas da razão de chances $\psi = 1$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) em 10.000 tabelas de contingência com $n = 500$, com categorização simétrica suave para o desfecho: (a) e (b) histogramas das estimativas; (c) diagrama de caixas das estimativas; (d) dispersão entre o erro absoluto das estimativas e o valor P do teste de aderência à distribuição bivariada Tipo-C normal; (e) erro quadrático em relação ao verdadeiro valor $\psi = 1$, para toda amostra e separadamente para casos que satisfazem e não satisfazem as suposições de linhas paralelas e de aderência à distribuição Tipo-C; (f) erro quadrático entre as estimativas produzidas pelos modelos, em toda amostra e de acordo com o atendimento das suposições de linhas paralelas e de aderência à distribuição Tipo-C.



G1: Toda amostra G2: Suposições satisfeitas G3: Suposições não satisfeitas

Figura 2. Comportamento das estimativas da razão de chances $\psi = 1$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) em 10.000 tabelas de contingência com $n = 500$, com categorização simétrica concentrada para o desfecho: (a) e (b) histogramas das estimativas; (c) diagrama de caixas das estimativas; (d) dispersão entre o erro absoluto das estimativas e o valor P do teste de aderência à distribuição bivariada Tipo-C normal; (e) erro quadrático em relação ao verdadeiro valor $\psi = 1$, para toda amostra e separadamente para casos que satisfazem e não satisfazem as suposições de linhas paralelas e de aderência à distribuição Tipo-C; (f) erro quadrático entre as estimativas produzidas pelos modelos, em toda amostra e de acordo com o atendimento das suposições de linhas paralelas e de aderência à distribuição Tipo-C.



G1: Toda amostra G2: Suposições satisfeitas G3: Suposições não satisfeitas

Figura 3. Comportamento das estimativas da razão de chances $\psi = 2$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) em 10.000 tabelas de contingência com $n = 500$, com categorização simétrica suave para o desfecho: (a) e (b) histogramas das estimativas; (c) diagrama de caixas das estimativas; (d) dispersão entre o erro absoluto das estimativas e o valor P do teste de aderência à distribuição bivariada Tipo-C normal; (e) erro quadrático em relação ao verdadeiro valor $\psi = 2$, para toda amostra e separadamente para casos que satisfazem e não satisfazem as suposições de linhas paralelas e de aderência à distribuição Tipo-C; (f) erro quadrático entre as estimativas produzidas pelos modelos, em toda amostra e de acordo com o atendimento das suposições de linhas paralelas e de aderência à distribuição Tipo-C.

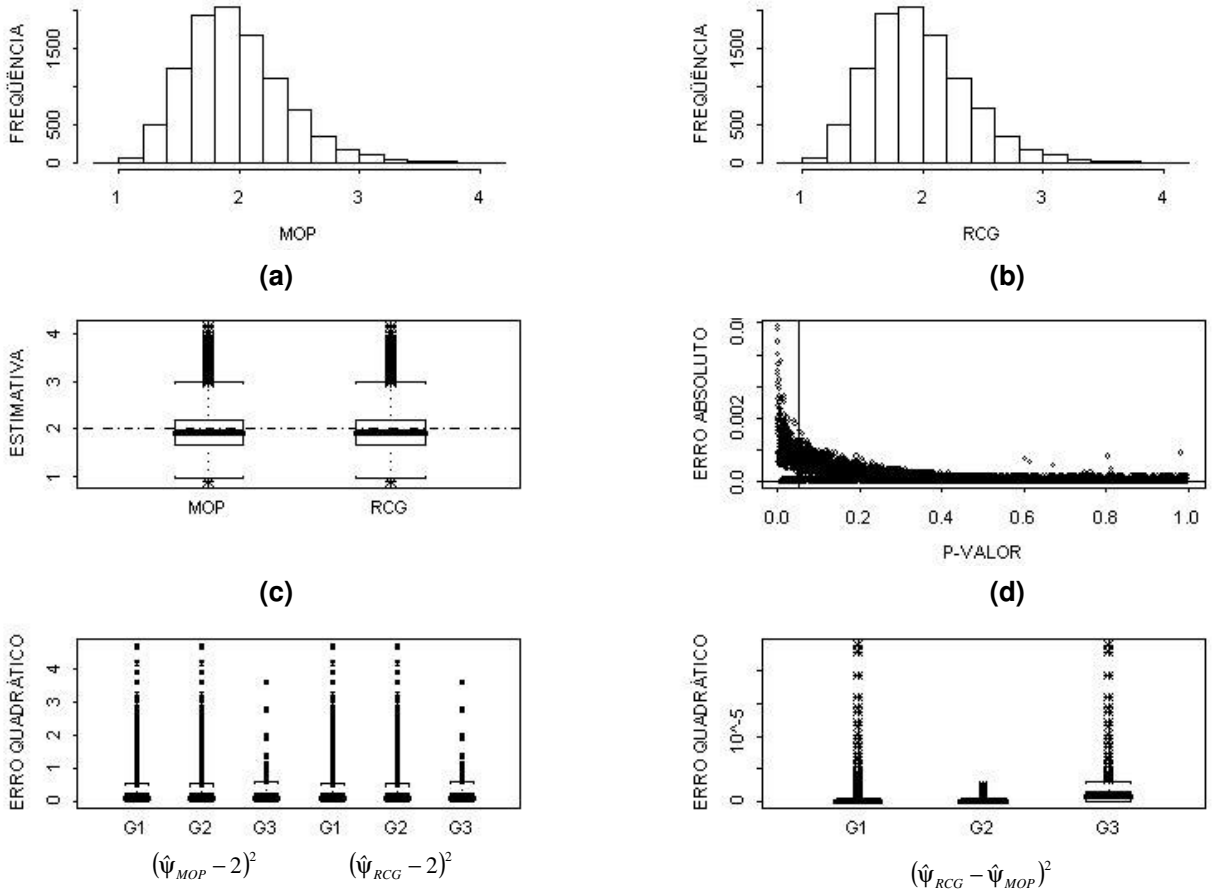


Figura 4. Comportamento das estimativas da razão de chances $\psi = 2$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) em 10.000 tabelas de contingência com $n = 500$, com categorização simétrica concentrada para o desfecho: (a) e (b) histogramas das estimativas; (c) diagrama de caixas das estimativas; (d) dispersão entre o erro absoluto das estimativas e o valor P do teste de aderência à distribuição bivariada Tipo-C normal; (e) erro quadrático em relação ao verdadeiro valor $\psi = 2$, para toda amostra e separadamente para casos que satisfazem e não satisfazem as suposições de linhas paralelas e de aderência à distribuição Tipo-C; (f) erro quadrático entre as estimativas produzidas pelos modelos, em toda amostra e de acordo com o atendimento das suposições de linhas paralelas e de aderência à distribuição Tipo-C.

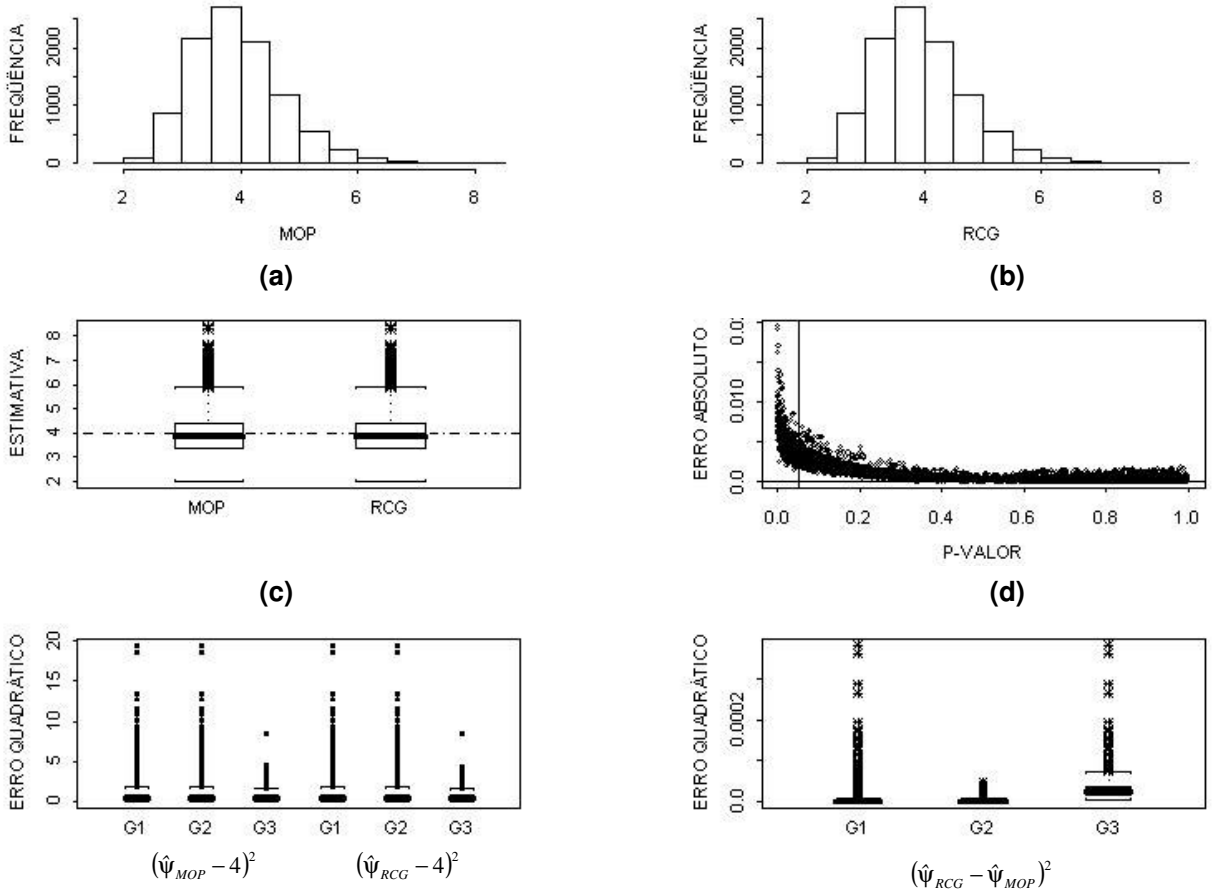


Figura 5. Comportamento das estimativas da razão de chances $\psi = 4$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) em 10.000 tabelas de contingência com $n = 500$, com categorização simétrica suave para o desfecho: (a) e (b) histogramas das estimativas; (c) diagrama de caixas das estimativas; (d) dispersão entre o erro absoluto das estimativas e o valor P do teste de aderência à distribuição bivariada Tipo-C normal; (e) erro quadrático em relação ao verdadeiro valor $\psi = 4$, para toda amostra e separadamente para casos que satisfazem e não satisfazem as suposições de linhas paralelas e de aderência à distribuição Tipo-C; (f) erro quadrático entre as estimativas produzidas pelos modelos, em toda amostra e de acordo com o atendimento das suposições de linhas paralelas e de aderência à distribuição Tipo-C.

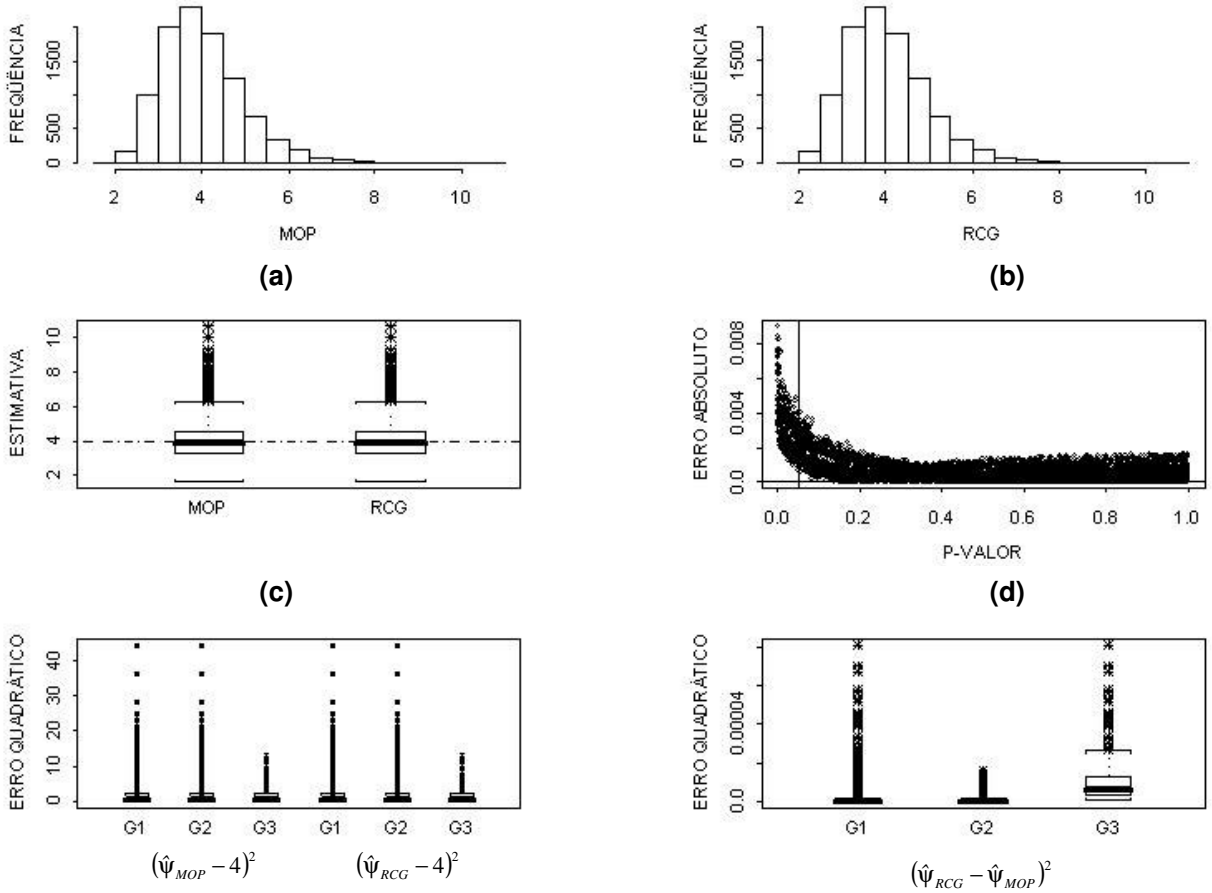


Figura 6. Comportamento das estimativas da razão de chances $\psi = 4$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) em 10.000 tabelas de contingência com $n = 500$, com categorização simétrica concentrada para o desfecho: (a) e (b) histogramas das estimativas; (c) diagrama de caixas das estimativas; (d) dispersão entre o erro absoluto das estimativas e o valor P do teste de aderência à distribuição bivariada Tipo-C normal; (e) erro quadrático em relação ao verdadeiro valor $\psi = 4$, para toda amostra e separadamente para casos que satisfazem e não satisfazem as suposições de linhas paralelas e de aderência à distribuição Tipo-C; (f) erro quadrático entre as estimativas produzidas pelos modelos, em toda amostra e de acordo com o atendimento das suposições de linhas paralelas e de aderência à distribuição Tipo-C.

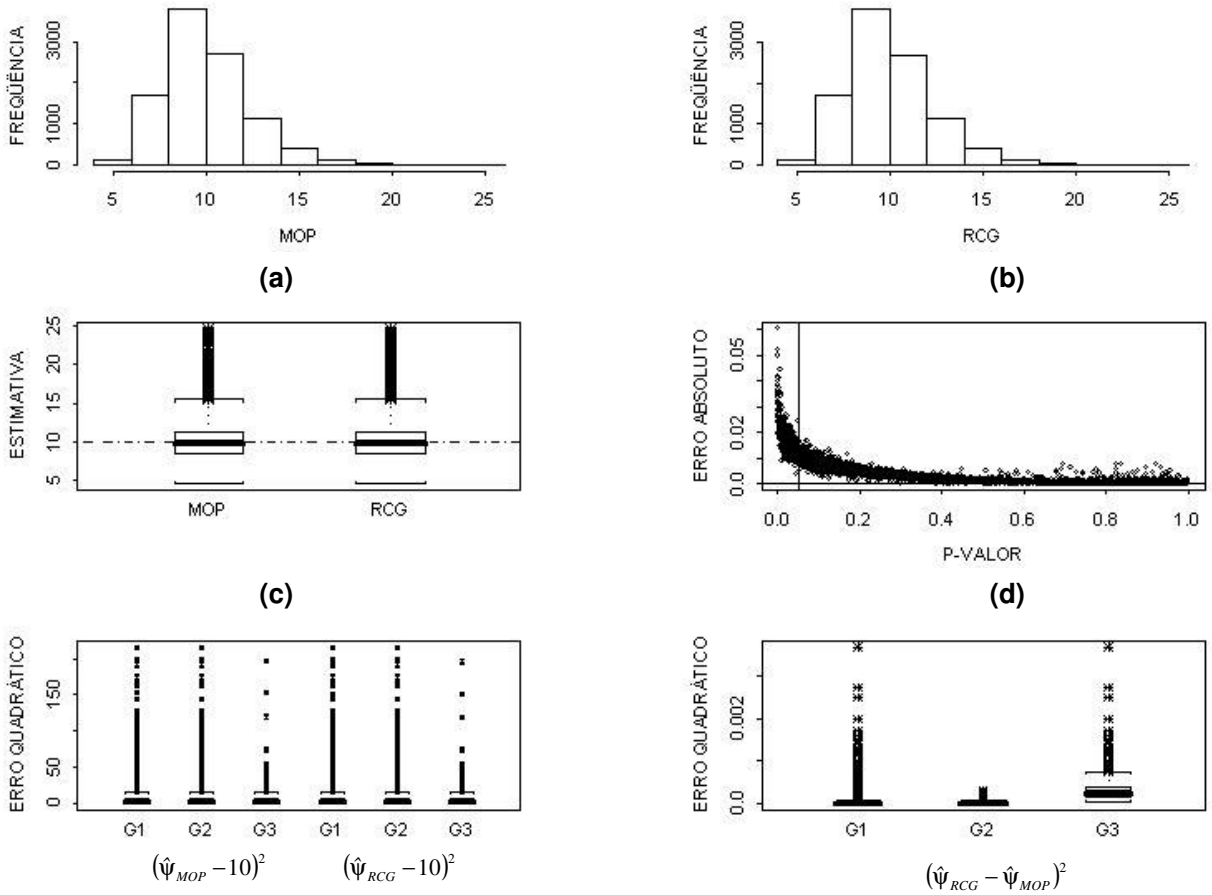


Figura 7. Comportamento das estimativas da razão de chances $\psi = 10$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) em 10.000 tabelas de contingência com $n = 500$, com categorização simétrica suave para o desfecho: (a) e (b) histogramas das estimativas; (c) diagrama de caixas das estimativas; (d) dispersão entre o erro absoluto das estimativas e o valor P do teste de aderência à distribuição bivariada Tipo-C normal; (e) erro quadrático em relação ao verdadeiro valor $\psi = 10$, para toda amostra e separadamente para casos que satisfazem e não satisfazem as suposições de linhas paralelas e de aderência à distribuição Tipo-C; (f) erro quadrático entre as estimativas produzidas pelos modelos, em toda amostra e de acordo com o atendimento das suposições de linhas paralelas e de aderência à distribuição Tipo-C.

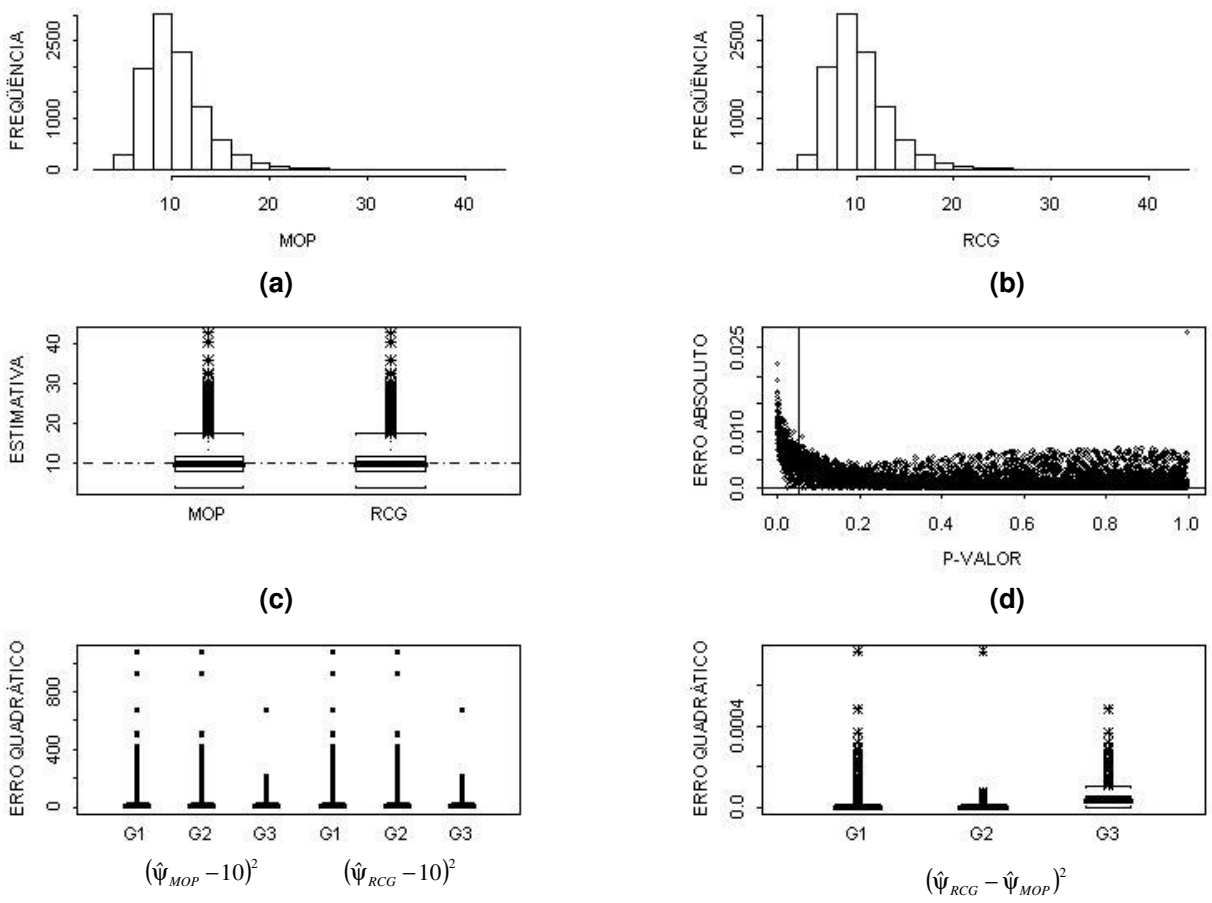


Figura 8. Comportamento das estimativas da razão de chances $\psi = 10$ mediante os modelos de odds proporcionais (MOP) e razão de chances generalizada (RCG) em 10.000 tabelas de contingência com $n = 500$, com categorização simétrica concentrada para o desfecho: (a) e (b) histogramas das estimativas; (c) diagrama de caixas das estimativas; (d) dispersão entre o erro absoluto das estimativas e o valor P do teste de aderência à distribuição bivariada Tipo-C normal; (e) erro quadrático em relação ao verdadeiro valor $\psi = 10$, para toda amostra e separadamente para casos que satisfazem e não satisfazem as suposições de linhas paralelas e de aderência à distribuição Tipo-C; (f) erro quadrático entre as estimativas produzidas pelos modelos, em toda amostra e de acordo com o atendimento das suposições de linhas paralelas e de aderência à distribuição Tipo-C.

3.2 ARTIGO 2

Comparação empírica do modelo de odds proporcionais e da razão de chances generalizada para estimar a associação da temperatura ambiente e da obesidade com as glicemias durante um teste oral de tolerância à glicose

Álvaro Vigo [1,2]

Jandyra M. G. Fachel [1,3]

e colaboradores

1. Departamento de Estatística, Universidade Federal do Rio Grande do Sul
2. Doutorando do Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul
3. Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul

Correspondência: Prof. Álvaro Vigo, Departamento de Estatística, UFRGS, Av. Bento Gonçalves, 9500, Prédio 43111, Bairro Agronomia, 91509-900, Porto Alegre, RS, BRASIL. Telefone: + 55 51 3316-6225 / 3316-6189 FAX: + 55 51 3316-7301

E_mail: vigo@orion.ufrgs.br

Número de palavras: Resumo:201; Texto:2321 + 4 tabelas + 1 Figura

Abstract

Ordinal outcomes are very common in medical and epidemiological research. Models for ordinal response are an important tool to describe the relationship between an ordinal outcome and the explanatory factors. The odds proportional model has been used more frequently, but, even so little known, the generalized odds ratio also can be useful. In this work, the estimate odds ratio generated by the proportional odds model was empirically compared with the generalized odds ratio. The methods are illustrated by means of the data of the Brazilian Study of Gestational Diabetes, to estimate the association of the ambient temperature and the body mass index (BMI) with the glycemia during an oral glucose tolerance test. There is evidence of an antagonistic interaction between temperature and BMI (P=0.0268). For obese individuals, the odds ratio of classify an individual as diabetic, for an ambient temperature greater or equal to 25 ºC, in relation of a temperature under 25 ºC, is equal to OR=1.94 (95% CI: 1.56-2.41), while for non-obese individuals, the odds ratio is OR=3.03 (95% CI: 2.18-4.23). This empirical comparison suggests that the generalized odds ratio is equivalent to the estimate of odds ratio produced by means of the proportional odds model.

Keywords: ordinal outcome, ordered endpoints, generalized odds ratio, proportional odds, odds ratio, oral glucose tolerance test, body mass index, type 2 diabetes

Resumo

Desfechos ordinais são muito comuns em pesquisas médicas e epidemiológicas. Os modelos para resposta ordinal são uma importante ferramenta para descrever a relação entre um desfecho ordinal e os preditores. O modelo de odds proporcionais tem sido usado com maior freqüência, mas, embora pouco conhecida, a razão de chances generalizada também pode ser útil. Neste trabalho, a estimativa da razão de chances do modelo de odds proporcionais foi comparada empiricamente com a razão de chances generalizada. Os métodos são ilustrados mediante os dados do Estudo Brasileiro de Diabetes Gestacional, para estimar a associação da temperatura ambiente e do índice de massa corporal (IMC) com as glicemias durante um teste oral de tolerância à glicose. Os resultados evidenciam que existe interação do tipo antagônica entre temperatura e IMC (P = 0,0268). Para indivíduos obesos, a chance de classificar um indivíduo como diabético, para uma temperatura ambiente ≥ 25 ºC, em relação a temperatura < 25 ºC, é igual a RC=1,94 (IC 95%: 1,56-2,41), enquanto que para indivíduos não obesos, a razão de chances é RC=3,03 (IC 95%: 2,18-4,23). A comparação empírica sugere que a razão de chances generalizada é equivalente à estimativa da razão de chances do modelo de odds proporcionais.

Palavras chaves: desfecho ordinal, razão de chances generalizada, odds proporcionais, teste de tolerância à glicose, índice de massa corporal, diabetes tipo 2

Introdução

Estudos observacionais ou experimentais com resposta categórica ordenada têm aparecido com grande freqüência na literatura médica e epidemiológica. Em muitas situações, o desfecho ordinal representa os níveis de uma escala de medida usual, como a severidade da dor (nenhuma, moderada, severa). Em outros casos, por razões práticas ou porque a variável contínua subjacente não pode ser diretamente observada, a estrutura ordinal surge da categorização de uma ou mais variáveis contínuas.

Para extrair eficientemente a informação contida nos dados é vital utilizar métodos de análise que considerem a ordem das categorias. A escolha da metodologia geralmente depende dos objetivos e do tipo de delineamento, bem como da observância das exigências do método. Dentre as principais ferramentas de análise destacam-se os modelos para resposta ordinal, os quais permitem avaliar o impacto dos fatores explanatórios ou experimentais sobre o desfecho. Diversos modelos estão disponíveis, tais como os modelos *complementar log-log* (1;2), *log-log* (2), de *probitos* (2), de *estereótipo* (3), de *categorias adjacentes* (4), de *odds proporcionais parciais restrito* e *não restrito* (5) ou o *modelo de odds proporcionais não-paramétrico* (6;7).

Neste trabalho são explorados aspectos da *razão de chances generalizada* (8) e do *modelo de odds proporcionais* (1;2). Um estudo de simulação Monte Carlo mostrou que, para tabelas de contingência 2 x 3, as estimativas da razão de chance produzidas por estes modelos são idênticas e têm a mesma eficiência (9).

O objetivo do estudo é apresentar dois procedimentos para descrever a relação entre um desfecho ordinal e os preditores, que permitam estimar a magnitude e a direção dos efeitos. Em particular, deseja-se comparar empiricamente

a estimativa da razão de chances generalizada com aquela gerada pelo modelo de odds proporcionais.

Um exemplo de desfecho ordinal em saúde é a classificação dos valores de glicose plasmática para o diagnóstico de diabetes melito e seus estágios pré-clínicos, mediante as categorias “glicemia normal”, “glicemia de jejum alterada e/ou tolerância à glicose diminuída” e “diabete melito”. Estas categorias podem ser definidas através das glicemias observadas durante um teste oral de tolerância à glicose (TTG), cuja importância para o diagnóstico do diabetes tem sido reafirmada por peritos internacionais (10;11).

No entanto, as condições climáticas no momento da coleta de sangue venoso podem influenciar a concentração de glicose durante um TTG. Mesmo quando a coleta é feita sob protocolo estritamente controlado, a glicemia de 2h pode sofrer variação clinicamente importante em função de temperatura ambiental no momento da coleta. Diversos estudos mostram um aumento clinicamente significativo na glicemia durante um TTG (12-15).

Assim, além de comparar os métodos, deseja-se, também, avaliar a associação da temperatura ambiente e da obesidade com a classificação da hiperglicemia na gravidez, usando os dados da linha de base do Estudo Brasileiro de Diabetes Gestacional.

Métodos

O Estudo Brasileiro de Diabetes Gestacional (EBDG) é um estudo de coorte conduzido em seis capitais do país (Porto Alegre, São Paulo, Rio de Janeiro, Fortaleza, Salvador e Manaus), entre maio de 1991 e agosto de 1995, com o objetivo geral de estudar o diabetes e a intolerância à glicose gestacional em

grávidas com atendimento obstétrico junto ao Sistema Único de Saúde (SUS), em relação a alguns fatores de risco, quanto à prevalência e outros desfechos.

A amostra consiste de 5564 gestantes consecutivas, com idade superior a 20 anos, idade gestacional entre a 21ª e a 28ª semanas e sem histórico de diabetes melito, que realizaram pré-natal pelo SUS em algum dos centros das seis capitais estudadas. Os comitês de ética das instituições locais aprovaram o protocolo do estudo e, após serem informadas sobre a natureza e objetivos do estudo, as pacientes consentiram em participar do mesmo.

Todas as mulheres responderam a um questionário estruturado, realizaram medidas antropométricas padronizadas e foram convidadas a realizar um TTG de 2 horas entre a 24ª e a 28ª semanas de gestação.

O TTG utilizou procedimentos padrões (16). Uma carga de 75 gramas de glicose anidra foi administrada após 12 a 14 horas de jejum. Amostras de jejum, 1h e 2h da veia antecubital foram coletadas em tubos contendo fluoreto e armazenadas a uma temperatura de 4 ºC até a centrifugação. As medidas do plasma foram realizadas mediante métodos enzimáticos de glicose, com coeficiente de variação menor que 5%.

As medidas das glicemias de jejum e de 2h foram utilizadas para a classificação da hiperglicemia na gravidez, categorizada em 3 categorias ordenadas (DIABETES, PRÉ-DIABETES e NORMAL). Pacientes com glicemia de jejum ≥ 126 mg/dl ou glicemia de 2h ≥ 200 mg/dl foram consideradas diabéticas e classificadas na categoria DIABETES, enquanto que na categoria NORMAL foram classificadas pacientes para as quais as glicemias de jejum e de 2h são, respectivamente, menores de 100 mg/dl e 140 mg/dl. A glicemia das demais pacientes foi considerada alterada e, portanto, foram classificadas na categoria PRÉ-DIABETES, um estágio intermediário entre não doente e doente.

A temperatura ambiente (em ºC) foi obtida de fontes oficiais, tendo sido registrada às 9 horas da manhã na estação meteorológica mais próxima do ponto de coleta. Para este estudo, foi dicotomizada nas categorias < 25 ºC e ≥ 25 ºC.

O índice de massa corporal (IMC), definido pela razão do peso (em kg) e o quadrado da altura (em metros), foi dicotomizada nas categorias não obeso (< 25 kg/m^2) e obeso (≥ 25 kg/m^2).

Foram excluídos da análise 25 pacientes pertencentes às etnias oriental e indígena e 556 pacientes devido à falta de informações sobre as glicemias de jejum ou de 2h. Adicionalmente, 563 pacientes foram excluídas devido à falta de informações válidas sobre a temperatura ambiente ou sobre o índice de massa corporal, restando 4420 casos.

Nessa situação, para extrair eficientemente a informação contida nos dados, é importante utilizar métodos de análise que considerem a estrutura ordenada do desfecho. Neste trabalho foram abordados os modelos de odds proporcionais e da razão de chances generalizada, descritos a seguir.

Razão de chances generalizada (RCG)

Considere que S e T são variáveis aleatórias contínuas com função de distribuição conjunta $H(s,t) = P(S \leq s, \mathrm{T} \leq t)$ e funções de distribuição marginais $F(s) = P(S \leq s)$ e $G(T) = P(T \leq t)$, respectivamente. As marginais contínuas podem ser dicotomizadas nos pontos arbitrários s e t, gerando a tabela de contingência 2 x 2, mostrada na Tabela 1. Assim, a razão de chances é definida por

$$\psi = \frac{H(s,t)[1 - F(s) - G(t) + H(s,t)]}{[F(s) - H(s,t)]\,[G(t) - H(s,t)]} \qquad [1]$$

e, de maneira equivalente, pode ser escrita como

$$(\psi - 1)\,[H(s,t)]^2 - [1 + [F(s) + G(t)](\psi - 1)]H(s,t) + \psi F(s)\,G(t) = 0;\ \psi > 0\,. \qquad [2]$$

A equação [2] possui uma única raiz (17), dada por

$$H(s,t)=\begin{cases}\dfrac{S(s,t)-\sqrt{[S(s,t)]^2-4\psi(\psi-1)F(s)G(t)}}{2(\psi-1)}, \text{ se } \psi\neq 1\\ F(s)G(t), \text{ se } \psi=1\end{cases} \qquad [3]$$

onde $S(s,t)=1+(\psi-1)[F(t)+G(t)]$.

A função $H(s,t)$ é denominada função de distribuição Tipo-C, onde ψ é um parâmetro desconhecido que representa associação entre as variáveis. Para quaisquer valores s e t que dicotomizam as distribuições marginais, a razão de chances é constante, produzindo uma única estimativa para o parâmetro ψ (18). Por esta razão, ψ é chamado de parâmetro de associação constante.

Em tabelas de contingência *r* x *c*, o parâmetro ψ pode ser estimado pelo método da máxima verossimilhança (8) e é denominado *razão de chances generalizada*. Contudo, para evitar problemas numéricos ao avaliar a função de verossimilhança na vizinhança do valor $\psi=1$, é conveniente reescrever a equação [3] mediante a série de Taylor da expansão binomial. Assim, fazendo $\lambda=\psi-1$, para $\psi\neq 1$,

$$H(s,t)=\frac{1}{2\lambda}\left[\,(1+\lambda[F(s)+G(t)])-\left[[1+\lambda[F(s)+G(t)]]^2-4\lambda(\lambda+1)F(s)G(t)\right]^{\frac{1}{2}}\,\right] \qquad [4]$$

ou, equivalentemente,

$$\begin{aligned}H(s,t)=\frac{1}{2}[\,F(s)+G(t)\,]+\frac{1}{2\lambda}[\,1-[1+2\lambda[\,F(s)+G(t)-2F(s)G(t)\,]+\\ +\lambda^2[\,F(s)-G(t)\,]^2\,]^{\frac{1}{2}}\,].\end{aligned} \qquad [5]$$

Definindo $z=1+2\lambda[\,F(s)+G(t)-2F(s)G(t)\,]+\lambda^2[\,F(s)-G(t)\,]^2$ e escrevendo o termo $[1+z]^{\frac{1}{2}}$ da equação [5] em série de Taylor, segue que

$$H(s,t) = F(s)G(t) + \frac{\lambda U}{1+\lambda V} + \frac{\lambda^3 U^2}{(1+\lambda V)^3} + \frac{2\lambda^5 U^3}{(1+\lambda V)^5} + \frac{5\lambda^7 U^4}{(1+\lambda V)^7} +$$

$$+ \frac{14\lambda^9 U^5}{(1+\lambda V)^9} + \frac{42\lambda^{11} U^6}{(1+\lambda V)^{11}} + \frac{132\lambda^{13} U^7}{(1+\lambda V)^{13}} + \cdots \qquad [6]$$

onde $U = F(s)G(t)[1-F(s)][1-G(t)]$ e $V = F(s)[1-G(t)] + [1-F(s)]G(t)$, para $\left| 2\lambda [F(s)+G(t)-2F(s)G(t)] + \lambda^2 [F(s)-G(t)]^2 \right| < 1$. Esta expressão alternativa para $H(x,y)$ converge para a expressão definida na equação [5], sendo extremamente útil para evitar problemas numéricos quando ψ assume valores próximos de 1 (8).

A estimação da razão de chances generalizada ψ, pelo método da máxima verossimilhança, utilizando o *método escore de Fisher* (*Fisher scoring method*) no processo iterativo, foi implementada em uma rotina computacional utilizando a linguagem Delphi, denominada CROSSPSI. Esta rotina também incorpora a expressão alternativa [6] quando $0{,}98 \leq \psi \leq 1{,}02$ (19).

Modelo de odds proporcionais

O modelo logístico para desfechos com categorias ordenadas utilizado com maior freqüência foi descrito inicialmente por S.H. Walker e D.B. Duncan (20) e, posteriormente, foi chamado de modelo de odds proporcionais por P. McCullagh (1). Este modelo pode ser útil para descrever a relação funcional entre um desfecho ordinal e um conjunto de fatores explanatórios.

Sem perda de generalidade, considere que Y representa um desfecho medido através de k categorias ordenadas e x é um fator explanatório, que pode ser discreto ou contínuo. Quando o fator explanatório assume o valor x, a probabilidade condicional de se observar o desfecho em uma categoria menor ou igual a j é $\gamma_j(x) = P(Y \leq j \mid x)$. O modelo de odds proporcionais especifica que

$$log\ \frac{P\left(Y \le j \mid x\right)}{1-P\left(Y \le j \mid x\right)} = log\ \frac{\gamma_j(x)}{1-\gamma_j(x)} = \theta_j - \beta\, x;\ \forall 1 \le j < k; \theta_1 \le \theta_2 \le \cdots \le \theta_{k-1}. \qquad [7]$$

Neste modelo está implícita a suposição de que o coeficiente de regressão β não depende das categorias de Y, também chamada de suposição de odds proporcionais ou de linhas paralelas. Na sua forma mais simples, para um determinado ponto de corte *j*, o modelo possui as mesmas suposições do modelo de regressão logística para resposta dicotômica. Isto é, postula que o logito da probabilidade $P\left(Y \le j \mid x\right)$ está linearmente relacionado com x e que não existe interação entre os fatores explanatórios. Assim, se as suposições de linearidade e aditividade estão satisfeitas, então a razão de chances para $P\left(Y \le j\right)$ associada a um incremento de uma unidade no fator explanatório é

$$\psi_{OP} = \frac{\dfrac{\gamma_j(x+1)}{1-\gamma_j(x+1)}}{\dfrac{\gamma_j(x)}{1-\gamma_j(x)}} = \frac{exp\left\{\theta_j - \beta\left(x+1\right)\right\}}{exp\left\{\theta_j - \beta\, x\right\}} = exp\left\{-\beta\left(x+1-x\right)\right\} = exp\left\{-\beta\right\}, \qquad [8]$$

que claramente não depende dos pontos de corte $j, \forall\ j = 1,2,\cdots,k$.

Os parâmetros do modelo podem ser estimados mediante o método de máxima verossimilhança, utilizando o procedimento de mínimos quadrados iterativamente reponderados para resolver as equações de verossimilhança (2;6;21). O ajuste do modelo foi realizado através do programa SAS (*Statistical Analysis System*), mas também está disponível em outros programas estatísticos usuais, tais como SPSS, STATA, S-Plus e R, por exemplo.

Resultados

A Tabela 2 mostra as freqüências observadas nas células de contingência resultantes do cruzamento do desfecho ordinal, que representa a classificação da

hiperglicemia na gravidez (DIABETES, PRÉ-DIABETES e NORMAL), com a temperatura (< 25 ºC, ≥ 25 ºC) e com o índice de massa corporal (< 25 kg/m^2 e ≥ 25 kg/m^2).

Foram ajustados os modelos de odds proporcionais univariáveis, considerando a temperatura ambiente (Modelo 1) e o IMC (Modelo 2) como fatores explanatórios, bem como o modelo que contempla estas variáveis simultaneamente, com interação (Modelo 4) e sem interação (Modelo 3). A diferença entre os valores da estatística –2 log L associados aos modelos multivariáveis com e sem interação é igual a 4,903 que, comparando com a distribuição assintótica de qui-quadrado com 1 grau de liberdade, produz valor P = 0,0268, evidenciando que existe interação entre a temperatura ambiente e o índice de massa corporal. A análise estratificada pelos níveis do IMC é apresentada pelo Modelo 5 (IMC < 25 kg/m^2) e pelo Modelo 6 (IMC ≥ 25 kg/m^2). As estimativas dos parâmetros e outros detalhes do ajuste dos modelos são mostrados na Tabela 3.

A Tabela 4 apresenta as razões de chances estimadas através dos modelos univariáveis, multivariáveis e estratificados pelas categorias do IMC e, também, as respectivas estimativas da razão de chances generalizada. As estimativas pontuais e os intervalos de confiança para a razão de chances generalizada são idênticos às estimativas produzidas pelo correspondente modelo de odds proporcionais. Para todos modelos ajustados a suposição de linhas paralelas e de que os dados da tabela de contingência se ajustam à distribuição bivariada Tipo-C normal estão atendidas (P > 0,05).

A análise estratificada evidencia que, em indivíduos não obesos (IMC < 25 kg/m^2), a chance de classificar a hiperglicemia na gravidez na categoria DIABETES é 203% maior (RC = 3,034; IC 95%: 2,176-4,232) quando a temperatura ambiente no momento da coleta é ≥ 25 ºC, em relação a temperatura ambiente < 25 ºC. Para

indivíduos obesos (IMC ≥ 25 kg/m^2), estima-se um aumento de apenas 94% (RC = 1,940; IC 95%: 1,561-2,410) na chance de classificar a hiperglicemia na gravidez na categoria DIABETES, para dias quentes (≥ 25 ºC) em relação aos dias frios (< 25 ºC). Interpretações idênticas podem ser realizadas para a chance de classificar a hiperglicemia na gravidez nas categorias DIABETES ou PRÉ-DIABETES. A Figura 1 mostra a natureza antagônica dos efeitos da temperatura ambiente no momento da coleta e da obesidade sobre a classificação da hiperglicemia na gravidez.

Discussão

Desfechos ordinais ocorrem com grande freqüência na pesquisa médica e epidemiológica e, para extrair eficientemente a informação contida nos dados, são necessários métodos de análise que incorporem a estrutura ordenada das categorias.

O modelo comumente usado para descrever a relação entre um desfecho ordinal e os preditores é o modelo de odds proporcionais. Um estudo de simulação evidenciou que existe uma forte conexão entre este modelo e a razão de chances generalizada (9). O mesmo comportamento foi observado na comparação empírica no estudo do impacto da temperatura ambiente e do índice da massa corporal sobre as glicemias durante um TTG. Os modelos univariáveis para estimar os efeitos da temperatura (Modelo 1) e do IMC (Modelo 2) e, também, o modelo estratificado pelos níveis do IMC < 25 kg/m^2 (Modelo 5) e IMC ≥ 25 kg/m^2 (Modelo 6), produziram estimativas de razão de chances idênticas aos correspondentes modelos de razão de chances generalizada. Os resultados confirmam, na prática, a equivalência destes modelos para estimar a razão de chances em tabelas de contingência 2 x 3, mostrada no estudo de simulação Monte Carlo (9).

Tanto a temperatura ambiente quanto à obesidade estão associadas com um aumento no risco de diagnóstico do diabetes, mas a interação entre estes fatores revela a presença de efeitos antagônicos. A associação entre a temperatura ambiente e a glicemia é explicada, provavelmente, pelo aumento da arterialização do sangue venoso antecubital, como conseqüência da redistribuição da circulação do antebraço, que minimiza a captação de glicose por músculos, em altas temperaturas (22). Ela é maior em indivíduos não obesos, possivelmente porque, em obesos, a camada superficial de gordura isola o corpo de temperaturas extremas, minimizando as variações da arterialização. Estes resultados são consistentes com aqueles reportados em estudos anteriores (12-15;23).

A associação com a temperatura no momento da coleta foi maior até mesmo do que com a obesidade, um dos principais fatores de risco para o desenvolvimento de diabetes. Esses resultados evidenciam a importância da climatização ambiental no momento da coleta de sangue venoso para dosagem de glicose no TTG e, considerando a vasta área tropical brasileira, podem ter importância na interpretação das prevalências de diabetes e pré-diabetes no Brasil.

Referências

(1) McCullagh P. Regression models for ordinal data. Journal of the Royal Statistical Society B 1980; 42(1):109-127.

(2) McCullagh P, Nelder JA. Generalized linear models. Second ed. New York: Chapman and Hall, 1989.

(3) Anderson A. Regression and ordered categorical variables. Journal of the Royal Statistical Society B 1984; 46(1):1-30.

(4) Agresti A. Categorical data analysis. New York: Wiley, 1990.

(5) Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. International Journal of Epidemiology 1997; 26(6):1323-1333.

(6) Hastie TJ, Tibshirani RJ. Non-parametric logistic and proportional odds regression. Applied Statistics 1987; 36(2):260-276.

(7) Hastie TJ, Tibshirani RJ. Generalized additive models. New York: Chapman and Hall, 1990.

(8) Fachel JMG. The C-type distribution as an underlying model for categorical data and its use in factor analysis. PhD Dissertation. London School of Economics and Political Sciences, University of London, 1986.

(9) Vigo A, Fachel JMG. Estudo de simulação Monte Carlo para comparar as razões de chances estimadas através dos modelos de odds proporcionais e razão de chances generalizada. Manuscrito não publicado 2004.

(10) Genuth S, Alberti KG, Bennett P, Buse J, Defronzo R, Kahn R et al. Follow-up report on the diagnosis of diabetes mellitus. Diabetes Care 2003; 26(11):3160-3167.

(11) World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications. Report of a WHO consultation. Part 1: diagnosis and classification of diabetes mellitus. 1999.

(12) Akanji AO, Bruce M, Frayn K, Hockaday TD, Kaddaha GM. Oral glucose tolerance and ambient temperature in non-diabetic subjects. Diabetologia 1987; 30(6):431-433.

(13) Schmidt MI, Matos MC, Branchtein L, Reichelt AJ, Mengue SS, Iochida LC et al. Variation in glucose tolerance with ambient temperature. Lancet 1994; 344(8929):1054-1055.

(14) Moses R, Griffiths R. Is there a seasonal variation in the incidence of gestational diabetes? Diabet Med 1995; 12(7):563-565.

(15) Akanji AO, Oputa RA. The effect of ambient temperature on glucose tolerance and its implications for the tropics. Trop Geogr Med 1991; 43(3):283-287.

(16) Schmidt MI, Duncan BB, Reichelt AJ, Branchtein L, Matos MC, Costa e Forti et al. Gestational diabetes mellitus diagnosed with a 2-h 75-g oral glucose tolerance test and adverse pregnancy outcomes. Diabetes Care 2001; 24(7):1151-1155.

(17) Plackett RL. A class of bivariate distributions. Journal of the American Statistical Association 1965; 60:516-522.

(18) Mosteller F. Association and estimation in contingency tables. Journal of the American Statistical Association 1968; 63:1-28.

(19) D´Ávila ER, Fachel JMG. Programa CROSSPSI para calcular o coeficiente de correlação Tipo-C. Livro de Resumos, X Salão de Iniciação Científica, UFRGS, p.25. 1998.

(20) Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967; 54(1):167-179.

(21) Vigo A. Análise de experimentos industriais com respostas categóricas ordenadas: método de Taguchi e modelo de McCullagh. Dissertação de Mestrado. IMECC, Universidade Estadual de Campinas, 1994.

(22) Frayn KN, Whyte PL, Benson HA, Earl DJ, Smith HA. Changes in forearm blood flow at elevated ambient temperature and their role in the apparent impairment of glucose tolerance. Clin Sci (Lond) 1989; 76(3):323-328.

(23) Moses RG, Patterson MJ, Regan JM, Chaunchaiyakul R, Taylor NA, Jenkins AB. A non-linear effect of ambient temperature on apparent glucose tolerance. Diabetes Res Clin Pract 1997; 36(1):35-40.

Tabela 1. Distribuição de probabilidade conjunta para a tabela de contingência resultante da dicotomização das marginais S e T nos valores arbitrários s e t.

Categorias da variável S	Categorias da variável T		Total
	$T \leq t$	$T > t$	
$S \leq s$	$H(s,t)$	$F(s) - H(s,t)$	$F(s)$
$S > s$	$G(t) - H(s,t)$	$1 - F(s) - G(t) + H(s,t)$	$1 - F(s)$
Total	$G(t)$	$1 - G(t)$	1

Tabela 2. Freqüências observadas para o cruzamento entre a classificação da hiperglicemia na gravidez e a temperatura ambiente no momento da coleta, para cada nível do índice de massa corporal.

Temperatura ambiente	IMC < 25 kg/m^2			IMC ≥ 25 kg/m^2		
	DIABETES	PRÉ-DIABETES	NORMAL	DIABETES	PRÉ-DIABETES	NORMAL
< 25 ºC	0	53	965	9	178	1335
≥ 25 ºC	2	129	787	8	198	756
Total	2	182	1752	17	376	2091

Tabela 3. Estimativas (IC 95%) dos parâmetros dos modelos de odds proporcionais e correspondentes –2 log L e valor P associado ao teste de linhas paralelas, ajustados aos dados da classificação da hiperglicemia na gravidez (DIABETES, PRÉ-DIABETES, NORMAL) no EBDG.

	UNIVARIÁVEL		MULTIVARIÁVEL		ESTRATIFICADO por IMC	
	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5[(c)]	Modelo 6[(d)]
Intercepto 1	-5,826 (-6,29; -5,36)	-5,814 (-6,28; -5,35)	-6,280 (-6,76; -5,80)	-6,482 (-7,01; -5,96)	-7,549 (-8,96; -6,14)	-5,288 (-5,78; -4,80)
Intercepto 2	-2,259 (-2,39; -2,13)	-2,256 (-2,41; -2,10)	-2,701 (-2,89; -2,51)	-2,904 (-3,18; -2,63)	-2,902 (-3,18; -2,63)	-1,965 (-2,12; -1,81)
Temperatura[(a)]	0,737 (0,56; 0,92)	-	0,802 (0,62; 0,98)	1,108 (0,78; 1,44)	1,110 (0,78; 1,45)	0,663 (0,45; 0,88)
IMC[(b)]	-	0,586 (0,40; 0,77)	0,667 (0,48; 0,86)	0,940 (0,62; 1,26)	-	-
Temperatura * IMC	-	-	-	-0,444 (-0,84; -0,05)	-	-
- 2 log L	3524,549	3552,001	3474,003	3469,100	1190,869	2273,759
Linhas paralelas (valor P)	0,4641	0,0571	0,1494	0,2111	0,3910	0,4995

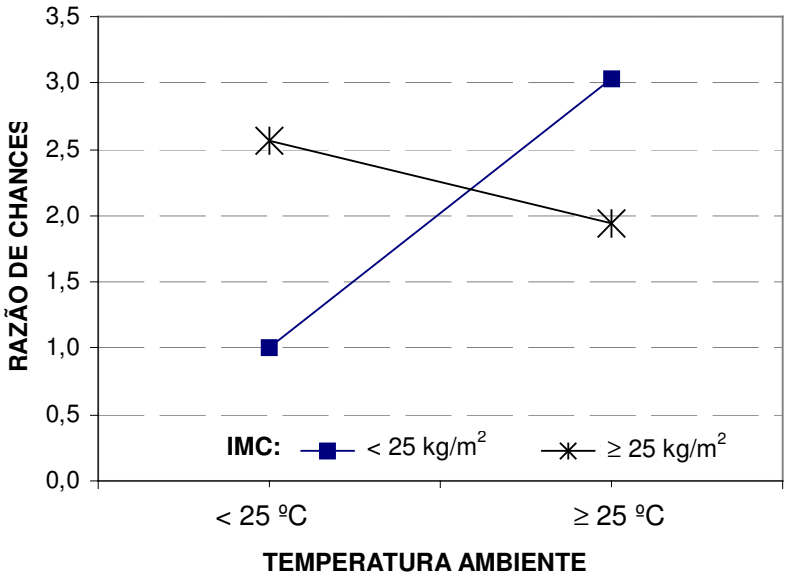
(a) Categoria de referência: < 25 ºC
(b) Categoria de referência: < 25 kg/m^2
(c) IMC < 25 kg/m^2
(d) IMC ≥ 25 kg/m^2

Tabela 4. Estimativas das razões de chances através do modelo de odds proporcionais e razão de chances generalizadas, ajustados aos dados da classificação da hiperglicemia na gravidez (DIABETES, PRÉ-DIABETES, NORMAL) no EBDG.

MODELOS	Odds proporcionais			Razão de chances generalizada		
	RC	IC 95%	P‡	RC	IC 95%	P†
Univariável						
Temperatura [(a)] (Modelo 1)	2,09	1,75-2,50	0,46	2,09	1,75-2,50	0,47
IMC [(b)] (Modelo 2)	1,80	1,49-2,17	0,06	1,80	1,49-2,17	0,06
Multivariável						
Sem interação (Modelo 3)						
Temperatura [(a)]	2,23	1,86-2,67	0,15	-	-	-
IMC [(b)]	1,95	1,61-2,35		-	-	-
Com interação (Modelo 4)						
Temperatura [(a)]	3,03	2,17-4,22	0,21	-	-	-
IMC [(b)]	2,56	1,87-3,51		-	-	-
Temperatura*IMC	0,64	0,43-0,96		-	-	-
Análise Estratificada						
IMC < 25 kg/m^2 (Modelo 5)						
Temperatura [(a)]	3,03	2,18-4,23	0,39	3,03	2,18-4,23	0,39
IMC ≥ 25 kg/m^2 (Modelo 6)						
Temperatura [(a)]	1,94	1,56-2,41	0,50	1,94	1,56-2,41	0,50

† P-valor para o teste de ajustamento da distribuição subjacente Tipo-C normal
‡ P-valor para o teste de linhas paralelas
(a) Categoria de referência: < 25 ºC
(b) Categoria de referência: < 25 kg/m^2

Figura 1. Razões de chances, estimadas pelo Modelo 4, de classificar a hiperglicemia na gravidez nas categorias DIABETES e DIABETES ou PRÉ-DIABETES, para uma temperatura no momento da coleta de sangue venoso maior ou igual a 25 ºC, em relação à temperatura < 25 ºC, considerando os diferentes níveis do índice de massa corporal.



4 CONSIDERAÇÕES FINAIS

Desfechos ordinais são bastante comuns na pesquisa médica e epidemiológica, exigindo ferramentas de análise que considerem a ordem das categorias para extrair eficientemente a informação contida nos dados. Os modelos para resposta categórica ordenada podem ser extremamente úteis para descrever a relação funcional com os preditores, mas ainda têm sido pouco explorados.

O modelo comumente usado para descrever a relação entre um desfecho ordinal e os preditores é o modelo de odds proporcionais. Quando os dados estão dispostos em uma tabela de contingência r x c, o modelo da razão de chances generalizada pode ser uma alternativa importante, pois independentemente do tamanho da tabela de contingência, produz uma estimativa única da associação entre o preditor e o desfecho.

Até o presente momento, a forte conexão entre o modelo de odds proporcionais e o modelo da razão de chances generalizada havia sido verificada apenas empiricamente, como no estudo sobre a associação da temperatura ambiente e da obesidade com a classificação da hiperglicemia na gravidez, discutida na seção 3.2.

O estudo de simulação Monte Carlo realizado neste trabalho mostrou que, para tabelas de contingência com um desfecho com três categorias ordenadas e um fator explanatório dicotômico, as estimativas da razão de chances produzidas por estes modelos são equivalentes. As estimativas pontuais são muito similares, para todos os valores fixados para o parâmetro de associação ψ, mesmo quando não estão satisfeitas as suposições de linhas paralelas e de aderência dos dados a uma distribuição Tipo-C normal. Contudo, o erro quadrático médio aumenta na presença destas violações, apesar das discrepâncias não parecerem relevantes, pois

usualmente são menores do que 10^{-2}. Ainda, utilizando a razão dos erros quadráticos médios em relação ao verdadeiro valor de ψ, em todos os contextos considerados na simulação, os estimadores sob investigação têm a mesma eficiência. Isto sugere que, para tabelas de contingência 2 x 3, a estimativa da razão de chances generalizada é equivalente àquela gerada pelo modelo de odds proporcionais e, também, têm a mesma precisão.

Se o desfecho ordinal surge da categorização de uma variável quantitativa, a forma de categorização parece importante, pois aumentando a concentração dos dados na categoria central, a estimativa média ficou mais próxima da verdadeira razão de chances (menor viés), porém com menor precisão.

A equivalência dos modelos de odds proporcionais e da razão de chances generalizada, para estimar a razão de chances, permite obter uma estimativa única da associação em tabelas de contingência, através da razão de chances generalizada. Além disso, permite interpretar a razão de chances generalizada na forma usual, como é feita no modelo de odds proporcionais, podendo ser vista como uma generalização da razão de chances para tabelas 2 x 2. Na prática, a razão de chances generalizada poderia ser usada em uma etapa inicial da investigação, para identificar associações relevantes em tabelas *r* x *c*, da mesma forma que a razão de chances é utilizada no caso de tabelas de contingência 2 x 2.

Embora as estimativas geradas pelos modelos sejam extremamente similares, chamam atenção as discrepâncias observadas em relação aos verdadeiros valores do parâmetro de associação fixados nas simulações. Estas diferenças podem ser decorrentes da forma de categorização das marginais e uma investigação mais profunda precisa ser realizada para avaliar a presença de viés.

Na seqüência do estudo, o resultado deve ser estendido para outras formas de categorização do desfecho e do preditor e, também, para desfechos com *k* categorias ordenadas.

ANEXOS

O Anexo A apresenta o Projeto de Pesquisa submetido para ingresso no Programa de Pós-Graduação, que tomou a forma atual após as apresentações nos Seminários de Pesquisa.

O Anexo B apresenta as rotinas computacionais necessárias para ajustar os modelos de odds proporcionais e da razão de chances generalizada aos dados do Estudo Brasileiro de Diabetes Gestacional (EBDG).

Por fim, o Anexo C contempla parte das rotinas computacionais desenvolvidas para o estudo de simulação Monte Carlo. O Anexo C1 contém as rotinas, em linguagem SAS, para as simulações do contexto $\psi = 1$ com categorização do desfecho simétrica suave e está dividido em quatro seções: geração da distribuição bivariada Tipo-C Normal e categorização das marginais, preparação dos arquivos de dados para o programa CROSSPSI, importação dos dados gerados no programa CROSSPSI e teste de ajustamento à distribuição Tipo-C Normal e análise descritiva. O Anexo C2 considera o contexto $\psi = 1$ para a categorização simétrica concentrada e contém as mesmas seções do Anexo C1 .

As rotinas computacionais utilizadas para as simulações dos demais valores do parâmetro de associação ($\psi = 2$, $\psi = 4$ e $\psi = 10$) e as duas formas de categorização do desfecho (suave e concentrada) são idênticas aos correspondentes programas dos anexos C1 e C2, alterando apenas o valor de ψ, os nomes dos arquivos de dados e os caminhos que identificam as pastas onde estão armazenadas as informações necessárias e onde são gravados os resultados.

ANEXO A

PROJETO DE PESQUISA

COMPARAÇÃO ENTRE AS ESTIMATIVAS DE RAZÃO DE CHANCES GERADAS PELOS MODELOS DE ODDS PROPORCIONAIS E DA RAZÃO DE CHANCES GENERALIZADA

Autor: Álvaro Vigo

Orientadora: Profª Drª Jandyra M. G. Fachel

Questão da pesquisa

Desfechos ordinais são muito comuns em pesquisas médicas e epidemiológicas, mas métodos de análise que incorporam a estrutura ordenada das categorias ainda têm sido pouco utilizados. Em muitas situações, o desfecho ordinal representa os níveis de uma escala de medida usual, como a severidade da dor (nenhuma, moderada, severa). Em outros casos, por razões práticas ou porque a variável contínua subjacente não pode ser diretamente observada, a estrutura ordinal surge da categorização de uma ou mais variáveis contínuas.

Embora modelos mais elaborados estejam disponíveis, o *modelo de odds proporcionais* (1;2) é comumente usado para descrever a relação entre um desfecho ordinal e os preditores. A *razão de chances generalizada* (3), ainda pouco conhecida, também pode ser um procedimento útil. Contudo, sua interpretação e a identificação das condições necessárias para sua aplicação foram pouco exploradas.

Para um desfecho com *k* categorias ordenadas e um preditor com dois níveis, resultados empíricos (4) sugerem que as estimativas da razão de chances geradas pelo modelo de odds proporcionais e pela razão de chances generalizada são muito similares. Entretanto, nenhum estudo investigou a equivalência entre estas estimativas, nem tão pouco a importância satisfazer a suposição de linhas paralelas (odds proporcionais) e dos dados se ajustarem à distribuição bivariada Tipo-C.

Assim, a demonstração de que a estimativa da razão de chances produzida pela razão de chances generalizada é equivalente àquela do modelo de odds proporcionais, e em quais casos este resultado é válido, é um passo importante para consolidar sua interpretação e disseminar sua utilização.

Entretanto, a razão de chances generalizada não tem forma explícita e, assim, sua comparação com o modelo de odds proporcionais pode ser realizada apenas mediante estudos de simulação Monte Carlo.

Objetivos

O objetivo principal do estudo é comparar a estimativa da razão de chances generalizada com aquela gerada pelo modelo de odds proporcionais, em tabelas de contingência 2 x *k*, onde *k* é o número de categorias ordenadas do desfecho. Em particular, deseja-se comparar a equivalência das estimativas da razão de chances quando as exigências dos modelos (linhas paralelas e aderência à distribuição Tipo-C) estão ou não estão satisfeitas.

Dados reais do Estudo Brasileiro de Diabetes Gestacional (EBDG) devem ser usados para ilustrar a aplicação, potencialidade dos modelos e, mais importante, a interpretação dos resultados.

A hipótese de pesquisa postula que, quando as exigências dos modelos de odds proporcionais e da razão de chances generalizada estão atendidas, as estimativas da razão de chances produzidas por estes modelos são equivalentes, ao menos na prática. Contudo, na presença de violações das suposições dos modelos, podem ocorrer discrepâncias maiores entre as estimativas, aspecto que será investigado.

Métodos

Razão de chances generalizada

Considere que X e Y são variáveis aleatórias contínuas com função de distribuição conjunta $H(x,y) = P(X \leq x, Y \leq y)$ e funções de distribuição marginais

$F(x) = P(X \leq x)$ e $G(y) = P(Y \leq y)$, respectivamente. As marginais contínuas podem ser dicotomizadas nos pontos arbitrários x e y, gerando a tabela de contingência 2 x 2, mostrada na Tabela 1.

Tabela 1. Distribuição de probabilidade conjunta para a tabela de contingência resultante da dicotomização das marginais X e Y nos valores arbitrários x e y.

Categorias da variável X	Categorias da variável Y		Total
	$Y \leq y$	$Y > y$	
$X \leq x$	$H(x,y)$	$F(x) - H(x,y)$	$F(x)$
$X > x$	$G(y) - H(x,y)$	$1 - F(x) - G(y) + H(x,y)$	$1 - F(x)$
Total	$G(y)$	$1 - G(y)$	1

Assim, a razão de chances é definida por

$$\psi = \frac{H(x,y)[1 - F(x) - G(y) + H(x,y)]}{[F(x) - H(x,y)]\,[G(y) - H(x,y)]} \qquad [1]$$

e, de maneira equivalente, pode ser escrita como

$$(\psi - 1)\,[H(x,y)]^2 - [1 + [F(x) + G(y)](\psi - 1)]H(x,y) + \psi F(x)\,G(y) = 0;\ \psi > 0. \qquad [2]$$

A equação [2] possui uma única raiz (5), dada por

$$H(x,y) = \begin{cases} \dfrac{S(x,y) - \sqrt{[S(x,y)]^2 - 4\psi(\psi - 1)F(x)G(y)}}{2(\psi - 1)}, \text{ se } \psi \neq 1 \\ F(x)G(y), \text{ se } \psi = 1 \end{cases} \qquad [3]$$

onde $S(x,y) = 1 + (\psi - 1)[F(x) + G(y)]$.

A função $H(x,y)$ é denominada função de distribuição Tipo-C, onde ψ é um parâmetro desconhecido que representa associação entre as variáveis. Para quaisquer valores x e y que dicotomizam as distribuições marginais, a razão de

chances é constante, produzindo uma única estimativa para o parâmetro ψ (6). Por esta razão, ψ é freqüentemente chamado de razão de chances generalizada.

A estimação da razão de chances generalizada ψ, pelo método da máxima verossimilhança, utilizando o método escore de Fisher no processo iterativo, foi implementada em uma rotina computacional utilizando a linguagem Delphi, denominada CROSSPSI (7). Esta rotina será usada para o ajuste do modelo da razão de chances generalizada no estudo de simulação Monte Carlo.

Modelo de odds proporcionais

O modelo de odds proporcionais por P. McCullagh (1) é o modelo para resposta ordinal utilizado com maior freqüência, e pode ser útil para descrever a relação funcional entre um desfecho ordinal e um conjunto de fatores explanatórios.

Sem perda de generalidade, considere que Y representa um desfecho medido através de k categorias ordenadas e x é um fator explanatório, que pode ser discreto ou contínuo. Quando o fator explanatório assume o valor x, a probabilidade condicional de se observar o desfecho em uma categoria menor ou igual a j é $\gamma_j(x) = P\left(Y \leq j \mid x\right)$. O modelo de odds proporcionais especifica que

$$log\ \frac{P\left(Y \leq j \mid x\right)}{1 - P\left(Y \leq j \mid x\right)} = log\ \frac{\gamma_j(x)}{1 - \gamma_j(x)} = \theta_j - \beta x;\ \forall 1 \leq j < k; \theta_1 \leq \theta_2 \leq \cdots \leq \theta_{k-1}.$$

Está implícita, neste modelo, a suposição de que o coeficiente de regressão β não depende das categorias de Y, também chamada de suposição de odds proporcionais ou de linhas paralelas. O modelo postula que o logito da probabilidade $P\left(Y \leq j \mid x\right)$ está linearmente relacionado com x e que não existe interação entre os fatores explanatórios. Assim, se as suposições de linearidade e

aditividade estão satisfeitas, então a razão de chances para $P(Y \leq j)$ associada a um incremento de uma unidade no fator explanatório é

$$\psi_{OP} = \frac{\dfrac{\gamma_j(x+1)}{1-\gamma_j(x+1)}}{\dfrac{\gamma_j(x)}{1-\gamma_j(x)}} = \frac{exp\{\theta_j - \beta(x+1)\}}{exp\{\theta_j - \beta x\}} = exp\{-\beta(x+1-x)\} = exp\{-\beta\},$$

que claramente não depende dos pontos de corte $j, \forall\ j = 1,2,\cdots,k$.

Os parâmetros do modelo podem ser estimados pelo método de mínimos quadrados iterativamente re-ponderados, extensamente discutido na literatura (2;8;9). O ajuste do modelo pode ser realizado mediante programas estatísticos usuais. No estudo de simulação, será usado o procedimento PROC LOGISTIC do programa SAS (10).

Simulação Monte Carlo

O estimador da razão de chances generalizada não possui uma forma explícita e, portanto, não é possível demonstrar analiticamente sua equivalência com a razão de chances estimada mediante o modelo de odds proporcionais. Em situações como esta, é bastante comum utilizar estudos de simulação Monte Carlo para comparar a equivalência, a precisão ou a eficiência de estimadores (11).

Portanto, para comparar as estimativas de razão de chances produzidas por estes modelos, será conduzido um estudo de simulação Monte Carlo. Devido ao volume de simulações, será abordado o caso particular em que preditor é dicotômico e o desfecho possui 3 categorias ordenadas.

O estudo consiste em gerar 10.000 amostras de tamanho 500 da distribuição bivariada contínua Tipo-C normal, com parâmetros de associação $\psi = 1$, $\psi = 2$,

$\psi = 4$ e $\psi = 10$. O algoritmo básico usado para gerar os dados com distribuição bivariada Tipo-C (12;13) será implementado em linguagem SAS.

Para cada valor do parâmetro ψ, depois de gerar os dados da distribuição conjunta, a primeira distribuição marginal será dicotomizada na mediana populacional. A outra distribuição marginal, que representa o desfecho ordinal, será categorizada de duas formas simétricas, porém com diferentes concentrações em torno da média populacional. Na primeira categorização, chamada de simétrica suave, a categoria central representa 50% da população, enquanto que cada uma das categorias extremas representa 25% da população. Na outra categorização, mais concentrada em torno da média, cada uma das categorias extrema representa 15% da população, enquanto que na categoria central representa os restantes 70%.

Assim, os 4 valores da verdadeira razão de chances ψ, combinados com as duas formas de categorização do desfecho, definem 8 contextos para os quais se deseja comparar as estimativas da razão de chances. Para cada contexto são geradas 10.000 tabelas de contingência 2 x 3, com tamanho de amostra 500, para as quais são ajustados os modelos.

O programa estatístico SAS será usado tanto nas simulações quanto na análise dos resultados, exceto para o ajuste do modelo da razão de chances generalizada, realizado mediante a rotina CROSSPSI, Versão 2.

As comparações das estimativas dos parâmetros são realizadas através do erro quadrático médio entre as estimativas e, também, em relação ao verdadeiro valor de ψ. Os modelos também são comparados quanto ao atendimento da suposição de linhas paralelas (odds proporcionais) e de que os dados se ajustam a uma distribuição subjacente Tipo-C normal (razão de chances generalizada). Este teste de aderência, cuja hipótese nula especifica que os dados da tabela de

contingência se ajustam à distribuição Tipo-C normal, também será implementado no programa SAS.

Aspectos éticos

A principal questão ética concerne à utilização dos dados do Estudo Brasileiro do Diabetes Gestacional (EBDG), usados para ilustrar a aplicação dos modelos. Foi aprovada a solicitação para utilização dos dados e rigorosos procedimentos para garantir a segurança e sigilo foram adotados.

Cronograma básico

PERÍODO	ATIVIDADE
Março a Junho/2001	Revisão da literatura, detalhamento de aspectos teóricos
Julho a Dezembro/2001	Revisão da literatura, apresentação do pré-projeto
Janeiro a Junho/2002	Revisão do pré-projeto, apresentação do projeto
Julho a Dezembro/2002	Simulação Monte Carlo (planejamento do estudo e elaboração de rotinas computacionais)
Janeiro a Dezembro/2003	Elaboração de rotinas computacionais e execução das simulações Aplicação dos modelos aos dados do EBDG
Janeiro a Setembro/2004	Execução das simulações, análise dos dados, artigos e tese
Outubro/2004	Encaminhamento da tese ao PPG
Novembro/2004	Apresentação preliminar
Dezembro/2004	Defesa da tese, publicações

Referências

(1) McCullagh P. Regression models for ordinal data. Journal of the Royal Statistical Society B 1980; 42(1):109-127.

(2) McCullagh P, Nelder JA. Generalized linear models. Second ed. New York: Chapman and Hall, 1989.

(3) Fachel JMG. The C-type distribution as an underlying model for categorical data and its use in factor analysis. PhD Dissertation. London School of Economics and Political Sciences, University of London, 1986.

(4) Lopes LS, Biasoli PK, Vigo A, Fachel JMG. A razão de chances generalizada e sua comparação com os parãmetros dos modelos de regressão logística ordinal. Livro de Resumos, XIII Salão de Iniciação Científica, UFRGS, p.13. 2001.

(5) Plackett RL. A class of bivariate distributions. Journal of the American Statistical Association 1965; 60:516-522.

(6) Mosteller F. Association and estimation in contingency tables. Journal of the American Statistical Association 1968; 63:1-28.

(7) D´Ávila ER, Fachel JMG. Programa CROSSPSI para calcular o coeficiente de correlação Tipo-C. Livro de Resumos, X Salão de Iniciação Científica, UFRGS, p.25. 1998.

(8) Hastie TJ, Tibshirani RJ. Non-parametric logistic and proportional odds regression. Applied Statistics 1987; 36(2):260-276.

(9) Vigo A. Análise de experimentos industriais com respostas categóricas ordenadas: método de Taguchi e modelo de McCullagh. Dissertação de Mestrado. IMECC, Universidade Estadual de Campinas, 1994.

(10) SAS Institute Inc. SAS OnlineDoc, Version Eight. Cary, NC: SAS Institute Inc., 1999.

(11) Rubinstein RY. Simulation and the Monte Carlo Method. New York: Wiley, 1981.

(12) Johnson M.E. Multivariate statistical simulation. New York: John Wiley & Sons, Inc., 1987.

(13) Mardia KV. Some contributions to contingency-type bivariate distributions. Biometrika 1967; 54(1):235-249.

ANEXO B

ROTINAS COMPUTACIONAIS PARA O AJUSTE DOS MODELOS DE ODDS PROPORCIONAIS (MOP) E DA RAZÃO DE CHANCES GENERALIZADA (RCG) AOS DADOS DO EBDG

B1: AJUSTE DO MOP – PROC LOGISTIC

```
options formchar='|----|+|---+=|-/\<>*';
options obs=MAX linesize=80 ps=59 validvarname=upcase;
options nodate nonumber;
*********************************************************************
FILENAME: DMxTEMPxBMI_NO MISSINGS_REVERSO.SAS
Template: DMxTEMPxBMI_NO MISSINGS.SAS

Crosstab of levels of environmental temperature and bmi x DM
Diabtes mellitus: Cutoff points (MIS)
    (See also - TABLE 3 from Kuzuya T. et al Diab Res Clin Pract 55 2002)

NOW EXCLUDING MISSING CASESE FOR UNIVARIATE ANALYSES
(MODELING WITH THE DATASET WITH 4420 CASES)

NOW MODELING P(Y <= DM | x) and
             P(Y <= DM or IFG/IGT | x)

REFERENCES NOW ARE  TEMPBIN = 0 and BMIBIN = 0
*********************************************************************;
libname TEMP V7 "d:\vigo\doutorado\tese\paper1";
data TEMP5;
        set TEMP.TEMP5;
run;

title1 "FREQUENCY OF TEMPERATURE BY CENTER";
proc freq;
        table CENTROA*TEMPBIN / nopercent nocol;
run;

title1 "UNIVARIATE MODEL FOR TEMPERATURE";
proc logistic data=TEMP5 descending;
   class TEMPBIN / param=reference ref=first;
   model DM = TEMPBIN / waldcl waldrl risklimits link=logit;
run;

title1 "UNIVARIATE MODEL FOR BMI";
proc logistic data=TEMP5 descending;
   class BMIBIN / param=reference ref=first;
   model DM = BMIBIN / waldcl waldrl risklimits link=logit;
run;

title1 "BIVARIATE MODEL: TEMPERATURE and BMI without INTERACTION";
* CONSIDERING TEMPERATURE > 25 DEGREES CELCIUS AND  BMI >= 25 KG/M^2 AS REFERENCE;
proc logistic data=TEMP5 descending;
   class BMIBIN TEMPBIN / param=reference ref=first;
   model DM = TEMPBIN BMIBIN / waldcl waldrl risklimits link=logit;
run;

title1 "BIVARIATE MODEL: TEMPERATURE and BMI with INTERACTION";
proc logistic data=TEMP5 descending;
   class TEMPBIN BMIBIN / param=reference ref=first;
   model DM = TEMPBIN BMIBIN TEMPxBMI / waldcl waldrl risklimits link=logit;
run;

title1 "STRATIFIED UNIVARIATE MODEL FOR TEMPERATURE - BMIBIN = 0";
proc logistic data=TEMP5(where=(BMIBIN=0)) descending;
   class TEMPBIN / param=reference ref=first;
   model DM = TEMPBIN / waldcl waldrl risklimits link=logit;
run;

title1 "STRATIFIED UNIVARIATE MODEL FOR TEMPERATURE - BMIBIN = 1";
proc logistic data=TEMP5(where=(BMIBIN=1)) descending;
   class TEMPBIN / param=reference ref=first;
   model DM = TEMPBIN / waldcl waldrl risklimits link=logit;
run;
```

B2: AJUSTE DA RCG – CROSSPSI

TEMPERATURA x DIAGNÓSTICO DO DM

```
Resultados de Tabela de Contingência - sem nome 1
Psi: 2,089587
Desvio padrão: 0,189329
IC 95%: (1,749585;2,495663)

Coeficientes de Correlação Tipo-C:
Tipo-C_M: 0,241293    IC95%: (0,185211; 0,297375)
Tipo-C_P: 0,282285    IC95%: (0,217603; 0,346967)
Tipo-C: 0,266115    IC95%: (0,205060; 0,327169)

Controles do Processo Iterativo:
Derivada do logaritmo: 0,000060
Número de iterações: 4

Valores observados:

10,000000    327,000000    1543,000000
9,000000     231,000000    2300,000000

Valores esperados sob hipótese de distribuição Tipo-C bivariada:

11,524620    325,273400    1543,202000
7,475376     232,726600    2299,798000
```

OBESIDADE x DIAGNÓSTICO DO DM

```
Resultados de Tabela de Contingência - sem nome 1
Psi: 1,796686
Desvio padrão: 0,170769
IC 95%: (1,491306;2,164599)

Coeficientes de Correlação Tipo-C:
Tipo-C_M: 0,193106    IC95%: (0,133098; 0,253114)
Tipo-C_P: 0,226485    IC95%: (0,156737; 0,296233)
Tipo-C: 0,213465    IC95%: (0,147678; 0,279252)

Controles do Processo Iterativo:
Derivada do logaritmo: -0,000004
Número de iterações: 14

Valores observados:

17,000000    376,000000    2091,000000
2,000000     182,000000    1752,000000

Valores esperados sob hipótese de distribuição Tipo-C bivariada:

13,242080    380,195700    2090,562000
5,757920     177,804300    1752,438000
```

ANEXO C

ROTINAS COMPUTACIONAIS PARA DO ESTUDO DE SIMULAÇÃO MONTE CARLO

C1: $\psi = 1$ COM CATEGORIZAÇÃO SUAVE

(25%, 50%, 25%)

C1.1: GERA DISTRIBUIÇÃO TIPO-C NORMAL E CATEGORIZA MARGINAIS

C1.2: PREPARA DADOS PARA O CROSSPSI

C1.3: IMPORTA ESTIMATIVAS DO CROSSPSI E TESTA AJUSTAMENTO

C1.4: ANÁLISE DESCRITIVA

C1.1: GERA DISTRIBUIÇÃO TIPO-C NORMAL E CATEGORIZA MARGINAIS

(SMS_R10PSI1.SAS – SIMÉTRICA SUAVE, $\psi = 1$)

```
* Redirecting and saving the log to the file named "SMS_R1Opsi1.LOG";
proc printto log="c:\vigo\simul\R1OM\psi1\sms\SMS_R1Opsi1.LOG";
run;
*******************************************************************************
*                                                                             *
*  Filename: SMS_R1Opsi1.SAS                                                  *
*  Date: 01/06/2004                                                           *
*                                                                             *
*  Function:                                                                  *
*    1) To generate a random sample of Bivariate Type C Normal distribution   *
*         ALGORITHM: Conditional distribution approach proposed by Mardia (1967)  *
*                    See, also, Johnson (1987, p.193)                         *
*                                                                             *
*    2) To save the full dataset for simulation SMS with 10000 repetitions    *
*                                                                             *
*    3) To Adjust the Proportional Odds Model AND to save the datasets with parameter *
*        estimates                                                            *
*                                                                             *
*    SAMPLE SIZE: 500                                                         *
*    NUMDER OF REPETITIONS: 10000                                             *
*    SIMULATION CASE SMS => X: Symmetric, Y: Mild Symmetric                   *
*    PSI = 1                                                                  *
*                                                                             *
*******************************************************************************;
OPTIONS OBS=MAX;
LIBNAME SMS V7 "c:\vigo\simul\R1OM\psi1\sms";
options formchar="|----|+|---+=|-/\<>*";
options linesize=120 ps=59;
*OPTION MPRINT MLOGIC SYMBOLGEN ;

* Redirecting and saving the output to the file named "SMS_R1Opsi1.LST";
proc printto print="c:\vigo\simul\R1OM\psi1\sms\SMS_R1Opsi1.LST";
run;

*****************************************************************
* Macro to produce random samples from the Bivariate Normal     *
* Type-C distribution and to adjust the Proportional Odds Model *
*****************************************************************;
%macro adjop(rep);
        %do r=1 %to &rep;
                title "SMS_R1Opsi1, REP=&r";

                *>>>>>>>>     PROGRAMMING FROM JOHNSON (1987, P.193)     <<<<<<<*;
                * FIRST, WE NEED TO GENERATE TWO INDEPENDENT RANDOM VARIABLES FROM AN UNIFORM (0,1);

                *** Generating the values of an Uniform (0,1) ***;
                data U1_&r;
                   retain seed 0;
                   do i = 1 to 500;
                          ID=i;                          * USED TO IDENTIFY THE SAMPLE UNIT;
                      call ranuni (seed, U1_&r);
                      output;
                   end;
                   drop seed i;
                   label U1_&r="U1_&r ~ U (0,1)";
                run;
                data U2_&r;
                   retain seed 0;
                   do i = 1 to 500;
                          ID=i;                           * USED TO IDENTIFY THE SAMPLE UNIT;
                          call ranuni (seed, U2_&r);

                      output;
                   end;
                   drop seed i;
                label U2_&r="U2_&r ~ U (0,1)";
                run;
```

```
                * THEN, THE PAIR (U1_r, U2_r) MUST BE TRANSFORMED IN A (Ur, Vr) RANDOM VECTOR
                  WITH TYPE-C UNIFORM DISTRIBUTION (r=1,2,...,10000);

                data R&r;
                     REP = &r;   * USED TO IDENTIFY THE REPETITIONS;
                     SET U1_&r; SET U2_&r;
                          A1=0; A2=0; B=0; D=0;
                          PSI=1;
                          PSISQ=PSI**2;
                          U&r=U1_&r;
                          label U&r="U = U1";
                          A1=U2_&r*(1-U2_&r);
                     A2= PSI + A1*((PSI-1)**2);
                     B = 2*A1*(U1_&r*PSISQ+1-U1_&r) + PSI*(1-2*A1);
                          D = SQRT(PSI)*SQRT((PSI + 4*A1*U1_&r*(1-U1_&r)*((1-PSI)**2)));
                          V&r = (B - (1-2*U2_&r)*D)/(2*A2);
                          label V&r="V=(B-(1-2*U2)*D)/(2*A2)";

                * So, (Ur, Vr) have Type-C Uniform Distribution. Now we must to transform (Ur, Vr)
                  into (Xr, Yr), which have Type-C Normal Distribution (r=1,2,...,10000);
                        X&r=PROBIT(U&r);
                        label X&r="X&r=Probit(U&r)";
                        Y&r=PROBIT(V&r);
                        label Y&r="Y&r=Probit(V&r)";

                * DICHOTOMIZATION OF Xr;
                        if X&r < probit(0.5) then XCAT&r = 1; else XCAT&r = 2;
                * CATEGORIZATION OF Yr;
                        if Y&r < probit(0.25)
                                then YCAT&r = 1;
                                else if Y&r < probit(0.75)
                                                then YCAT&r = 2;
                                                else YCAT&r = 3;
                run;

/*
                *** Table Analysis ***;
                proc freq data=R&r;
                   tables XCAT&r*YCAT&r / NOPERCENT NOROW NOCOL;
                        title "CROSSTABS XCAT&r x YCAT&r";
                run;
*/
                * ADJUSTING THE PROPORTIONAL ODDS MODEL;
                ods output Logistic.CumulativeModelTest=CMT&r
                         (keep=ProbChiSq
                              rename=(ProbChiSq=P_CMT));
                ods output Logistic.ParameterEstimates=PAREST&r
                         (keep=STDERR);

                *** Logistic Regression Analysis ***;
                proc LOGISTIC data=R&r outest=set1est&r;
                   class XCAT&r / param=reference ref=last;
                   model YCAT&r = XCAT&r;
                   title "MODEL for YCAT&r x XCAT&r";
                run;
        %end;
%mend adjop;

*******************************************************************************
*    Macro used to join the datasets named SET1EST1, SET1EST2,...,SET1ESTrep      *
*    which contains the estimative of the parameters INTERCEPT_1, INTERCEPT_2 and *
*    BETA. Note that we changed the original names of the coeficient of regression*
*    associated to the model adjusted to each repetition (r=1,...,rep) using the  *
*    auxiliary datasets S1, S2,...,Srep. The parameter "z" in this macro is used  *
*    just to compose the names of the original coefficients of regression named   *
*    XCAT12, XCAT22,.....,XCATrep2.                                               *
*                                                                                 *
*    NOTE: use z=1 if the reference category is reference=last, or                *
*              z=2 if the reference category is reference=first                   *
*    rep = NUMBER OF REPETITIONS                                                  *
*                                                                                 *
*******************************************************************************;
```

```
%macro xcat(rep,z);
        %do r=1 %to &rep;
                data S&r;
            length _LINK_ $ 8 _TYPE_ $ 8 _STATUS_ $ 11 _NAME_ $ 10 INTERCEPT_1 8 INTERCEPT_2 8 BETA 8 _LNLIKE_ 8;
                        set Work.Set1est&r;
                        BETA=XCAT&r&z;
                        drop XCAT&r&z;
                run;
        %end;
%mend xcat;



*******************************************************************************
*                                                                             *
*    Macro used to select the standard error of BETA and put it in the file with  *
*    just one record, corresponding to the repetition. The standard erros are     *
*    saved on files named PAREST1, PAREST2, ..., PARESTrep, and the standar error *
*    will be saved in the respective datasets STDERR1, STDERR2, ..., STDERRrep    *
*                                                                             *
*******************************************************************************;
%macro stderr(rep);
        %do r=1 %to &rep;
                data STDERR&r;
                        set Work.PAREST&r;
                        if _n_ = 3; /* Excluding the first two lines with INTERCEPT standard error */
                        SE_BETA=STDERR;
                        keep SE_BETA;
                run;
        %end;
%mend stderr;



* Calling macros ADJOP, XCAT and STDERR for 10000 repetitions;
%adjop(10000);
%xcat(10000,1);
%stderr(10000);


*******************************************************************************
* APPENDING DATASETS S1, S2, ..., Srep, which contain the parameter estimates *
* and other information about the model adjusting procedure                   *
*******************************************************************************;
%macro joinS(rep);
        %do r=1 %to &rep;
         proc append
              BASE=SMS_R10psi1_PAR data=S&r;
         run;
        %end;
%mend joinS;

%joinS(10000);

*Excluding the temporary datasets;
proc datasets lib=work nolist;
   delete U1_1-U1_10000;
run;
quit;
proc datasets lib=work nolist;
   delete U2_1-U2_10000;
run;
quit;
proc datasets lib=work nolist;
   delete S1-S10000;
run;
quit;
proc datasets lib=work nolist;
   delete SET1EST1-SET1EST10000;
run;
quit;
```

```
*************************************************************************************************
*  APPENDING DATASETS CMT1, CMT2, ..., CMTrep, which contain the p-values associated to the *
*  Score Test for the Proportional Odds Assumption, obtained using the ODS function         *
*************************************************************************************************;
%macro joinCMT(rep);
        %do r=1 %to &rep;
         proc append
              BASE=CMT data=CMT&r;
         run;
        %end;
%mend joinCMT;

%joinCMT(10000);

*Excluding the temporary datasets Cmt1-Cmt10000;
proc datasets lib=work nolist;
   delete Cmt1-Cmt10000;
run;
quit;

*************************************************************************************
*  APPENDING DATASETS STDERR1, STDERR2, ..., STDERRrep, with the standard error of BETA  *
*************************************************************************************;
%macro joinSTD(rep);
        %do r=1 %to &rep;
         proc append
              BASE=STDERR data=STDERR&r;
         run;
        %end;
%mend joinSTD;

%joinSTD(10000);

*Excluding the temporary datasets STDERR1-STDERR10000 and PAREST1-PAREST10000;
proc datasets lib=work nolist;
   delete STDERR1-STDERR10000 PAREST1-PAREST10000;
run;
quit;
*************************************************************************************************
*  Merging datasets SMS_R1Opsi1_PAR, CMT and STDERR to produce the dataset SMS_R1Opsi1_POM  *
*  with the parameter estimates, -ln likelihood, p-value for the Score Test for the         *
*  Proportional Odds Assumption, and other information about the adjusted models            *
*************************************************************************************************;
data SMS_R1Opsi1_POM;
   merge SMS_R1Opsi1_PAR (in=TABLE1) CMT (in=TABLE2) STDERR (in=TABLE3);
   if TABLE1 and TABLE2 and TABLE3;
run;
*************************************************************************************
*  Saving the dataset with the parameter estimates for simulation SMS_R1Opsi1              *
*************************************************************************************;
data SMS.SMS_R1Opsi1_POM;
     set SMS_R1Opsi1_POM;
         REP=_n_;
     keep REP _STATUS_ INTERCEPT_1 INTERCEPT_2 BETA SE_BETA _LNLIKE_ P_CMT;
run;
*************************************************************************
*  CREATING THE DATASET USED AS AN INPUT IN CROSSPSI SOFTWARE           *
*************************************************************************;
title "CROSSPSI DATASET FOR SMS_R1Opsi1";
%macro mergeCATX(part, first, last);
    data SMS_R1Opsi1_CATX&part;
              set R&first;
         keep ID XCAT&first;
    run;
         %do r = &first+1 %to &last;
         data SMS_R1Opsi1_CATX&part;
              merge SMS_R1Opsi1_CATX&part R&r;
                          by ID;
              keep ID XCAT&first-XCAT&r;
         run;
         %end;
%mend mergeCATX;
```

```
%mergeCATX(1,1,1000);
%mergeCATX(2,1001,2000);
%mergeCATX(3,2001,3000);
%mergeCATX(4,3001,4000);
%mergeCATX(5,4001,5000);
%mergeCATX(6,5001,6000);
%mergeCATX(7,6001,7000);
%mergeCATX(8,7001,8000);
%mergeCATX(9,8001,9000);
%mergeCATX(10,9001,10000);

%macro mergeCATY(part, first, last);
    data SMS_R1Opsi1_CATY&part;
              set R&first;
          keep ID YCAT&first;
    run;
         %do r = &first+1 %to &last;
         data SMS_R1Opsi1_CATY&part;
              merge SMS_R1Opsi1_CATY&part R&r;
                           by ID;
              keep ID YCAT&first-YCAT&r;
         run;
         %end;
%mend mergeCATY;
%mergeCATY(1,1,1000);
%mergeCATY(2,1001,2000);
%mergeCATY(3,2001,3000);
%mergeCATY(4,3001,4000);
%mergeCATY(5,4001,5000);
%mergeCATY(6,5001,6000);
%mergeCATY(7,6001,7000);
%mergeCATY(8,7001,8000);
%mergeCATY(9,8001,9000);
%mergeCATY(10,9001,10000);

* JOINING THESE DATASETS;
%macro joinCAT;
    %do r=1 %to 10;
         data SMS_R1Opsi1_CAT&r;
              merge WORK.SMS_R1Opsi1_CATX&r (in=TABLE1) WORK.SMS_R1Opsi1_CATY&r (in=TABLE2);
              by ID;
              if TABLE1 and TABLE2;
         run;
    %end;
%mend joinCAT;

%joinCAT;

%macro saveCAT;
    %do r=1 %to 10;
    * SAVING CROSSPSI (CATEGORICAL) DATASETS .SAS7BDAT;
         data SMS.SMS_R1Opsi1_CAT&r;
              set SMS_R1Opsi1_CAT&r;
         run;
/*
          * EXPORTING THE CROSSPSI (CATEGORICAL) DATASET IN TAB DELIMITED FORMAT (.DAT);
    * In fact, this kind of file should have .TXT extension;
         proc export data=SMS_R1Opsi1_CAT&r
                    outfile ="c:\vigo\simul\R1OM\psi1\sms\SMS_R1Opsi1_CAT&r.dat"
                    dbms=TAB;
         run;
*/
    %end;
%mend saveCAT;

%saveCAT;
*Excluding the temporary datasets;
proc datasets lib=work nolist;
   delete CMT SMS_R1Opsi1_PAR SMS_R1Opsi1_POM STDERR SMS_R1Opsi1_CAT1-SMS_R1Opsi1_CAT10
          SMS_R1Opsi1_CATX1-SMS_R1Opsi1_CATX10 SMS_R1Opsi1_CATY1-SMS_R1Opsi1_CATY10;
run;
quit;
```

```
************************************************************
*  SAVING THE ORIGINAL DATASETS R1-Rrep                    *
************************************************************;
title "FULL DATASET FOR SMS_R1Opsi1";
***Combine tables by columns (Merge Tables)***;
%macro saveR(rep);
        %do r = 1 %to &rep;
        data SMS.R&r;
             set work.R&r;
             drop A1 A2 B D PSI PSISQ REP XCAT&r YCAT&r;
        run;
        %end;
%mend saveR;
%saveR(10000);

*Excluding the temporary datasets R1-R10000;
proc datasets lib=work nolist;
   delete R1-R10000;
run;
quit;

proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

C1.2: PREPARA DADOS PARA O CROSSPSI

(SMS_R10PSI1_CAT.SAS – SIMÉTRICA SUAVE, $\psi = 1$)

```
* Redirecting and saving the log to the file named "SMS_R1Opsi1.LOG";
proc printto log="c:\vigo\simul\R1OM\psi1\sms\SMS_R1Opsi1_cat.LOG";
run;

*******************************************************************************
*                                                                             *
*  Filename: SMS_R1Opsi1_CAT.SAS                                              *
*  Date: 01/07/2004                                                           *
*                                                                             *
*  Function: To read the categorical datasets in .SAS7BDAT format and to save them as *
*            .DAT ( in fact, in .TXT format)                                  *
*                                                                             *
*    SAMPLE SIZE: 500                                                         *
*    NUMDER OF REPETITIONS: 10000                                             *
*    SIMULATION CASE SMS => X: Symmetric, Y: Mild Symmetric                   *
*    PSI = 1                                                                  *
*                                                                             *
*******************************************************************************;
OPTIONS OBS=MAX;
LIBNAME SMS V7 "c:\vigo\simul\R1OM\psi1\sms";
options formchar="|----|+|---+=|-/\<>*";
options linesize=120 ps=59;
*OPTION MPRINT MLOGIC SYMBOLGEN ;

* Redirecting and saving the output to the file named "SMS_R1Opsi1.LST";
proc printto print="c:\vigo\simul\R1OM\psi1\sms\SMS_R1Opsi1_cat.LST";
run;



* READING FILES IN .SAS7BDAT;
%macro readCAT;
        %do r=1 %to 10;
                data CAT&r;
                        set SMS.SMS_R1Opsi1_CAT&r;
                RUN;
        %end;
%mend readCAT;

%readCAT;


* SAVING AS .DAT;
%macro saveCAT;
    %do r=1 %to 10;
         * EXPORTING THE CROSSPSI (CATEGORICAL) DATASET IN TAB DELIMITED FORMAT (.DAT);
    * In fact, this kind of file should have .TXT extension;
        proc export data=CAT&r
                  outfile ="c:\vigo\simul\R1OM\psi1\sms\SMS_R1Opsi1_CAT&r..dat"
                  dbms=TAB;
        run;
    %end;
%mend saveCAT;

%saveCAT;

proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

C1.3: IMPORTA ESTIMATIVAS DO CROSSPSI E TESTA AJUSTAMENTO

(SMS_R10PSI1_FIT.SAS – SIMÉTRICA SUAVE, $\psi = 1$)

```
* Redirecting and saving the log to the file named "SMS_R1Opsi1_FIT.LOG";
proc printto log="c:\vigo\simul\R1OM\psi1\sms\SMS_R1Opsi1_FIT.LOG";
run;

*********************************************************************
*                                                                   *
*  Filename: SMS_R1Opsi1_FIT.SAS                                    *
*  Date: 01/07/2004                                                 *
*                                                                   *
*  Function:                                                        *
*    1) To produce the Goddness of Fit test to evaluate if the      *
*       data become from a Bivariate Normal Type C distribution     *
*       ALGORITHM: Fachel (1986, p.59)                              *
*                                                                   *
*    2) CREATE A DATASET NAMED SMS_R1Opsi1_FIT with the main        *
*       results of simulation SMS_R1Opsi1                           *
*                                                                   *
*   Note: THE COEFFICIENT PSI IS ESTIMATED BY THE PLACKETT          *
*         APPROACH (see Johnson (1987, p.193)                       *
*                                                                   *
*    SAMPLE SIZE: 500                                               *
*    NUMBER OF REPETITIONS: 10000                                   *
*    CATEGORIZATION: SMS => X: Symmetric, Y: Mild Symmetric         *
*    TRUEPSI = 1                                                    *
*********************************************************************;

OPTIONS OBS=MAX nodate nonumber;
LIBNAME SMS V7 'c:\vigo\simul\R1OM\psi1\sms';
options formchar='|----|+|---+=|-/\<>*';
options linesize=120 ps=59;
* Redirecting and saving the output to the file named "SMS_R1Opsi1_FIT.LST";
proc printto print="c:\vigo\simul\R1OM\psi1\sms\SMS_R1Opsi1_FIT.LST";
run;

* READING DATASETS SMS_R1Opsi1_CAT1-SMS_R1Opsi1_CAT10;
%macro readCAT;
        %do i=1 %to 10;
                data SMS_R1Opsi1_CAT&i;
                        set SMS.SMS_R1Opsi1_CAT&i;
                run;
        %end;
%mend readCAT;

%readCAT;

* MERGING DATASETS;
%macro mergeCAT;
    data SMS_R1Opsi1_CAT;
            set SMS_R1Opsi1_CAT1;
    run;
        %do j = 2 %to 10;
        data SMS_R1Opsi1_CAT;
             merge SMS_R1Opsi1_CAT SMS_R1Opsi1_CAT&j;
                         by ID;
        run;
        %end;
%mend mergeCAT;
%mergeCAT;

*Excluding the temporary datasets SMS_R1Opsi1_CAT1-SMS_R1Opsi1_CAT10;
proc datasets lib=work nolist;
   delete SMS_R1Opsi1_CAT1-SMS_R1Opsi1_CAT10;
run;
quit;
```

```
*****************************************************************
* Macro used to obtain the frequencies of the contingency tables    *
* XCAT1 x YCAT1, XCAT2 x YCAT2, ..., XCATrep x YCATrep, and to save *
* them in datasets named FREP1, FREP2, ..., FREPrep                 *
*****************************************************************;
%macro freqdata(rep);
        %do r=1 %to &rep;
                proc freq data=SMS_R1Opsi1_CAT;
                        tables XCAT&r*YCAT&r / NOPERCENT NOROW NOCOL out=FREQ&r;
                        title "CROSSTAB XCAT&r x YCAT&r";
                run;

                data F11A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 1 and YCAT&r = 1
                        then do;
                                if (XCAT&r = 1 and YCAT&r = 1);
                                F11 = COUNT;
                        end;
                        if _n_=1;
                        keep REP F11;
                run;

                data F12A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 1 and YCAT&r = 2
                                then do;
                                        if (XCAT&r = 1 and YCAT&r = 2);
                                        F12 = COUNT;
                                end;
                        if _n_=2;
                        keep REP F12;
                run;

                data F13A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 1 and YCAT&r = 3
                        then do;
                                if (XCAT&r = 1 and YCAT&r = 3);
                                F13 = COUNT;
                        end;
                        if _n_=3;
                        keep REP F13;
                run;

                data F21A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 2 and YCAT&r = 1
                        then do;
                                if (XCAT&r = 2 and YCAT&r = 1);
                                F21 = COUNT;
                        end;
                        if _n_=4;
                        keep REP F21;
                run;

                data F22A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 2 and YCAT&r = 2
                        then do;
                                if (XCAT&r = 2 and YCAT&r = 2);
                                F22 = COUNT;
                        end;
                        if _n_=5;
                        keep REP F22;
                run;
```

```
                data F23A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 2 and YCAT&r = 3
                        then do;
                                if (XCAT&r = 2 and YCAT&r = 3);
                                F23 = COUNT;
                        end;
                        if _n_=6;
                        keep REP F23;
                run;

                * MERGING DATASETS F11A, F12A, F13A, F21A, F22A, F23A;
                data FREP&r;
                        merge WORK.F11A&r (in=TABLE1)
                                  WORK.F12A&r (in=TABLE2)
                                  WORK.F13A&r (in=TABLE3)
                                  WORK.F21A&r (in=TABLE4)
                                  WORK.F22A&r (in=TABLE5)
                                  WORK.F23A&r (in=TABLE6)
                        ;
                run;
        %end;
%mend freqdata;

*CALLING FREQDATA MACRO FOR 10000 REPETITIONS;
%freqdata(10000);

*Excluding the temporary datasets R1-R10000 FREQ1-FREQ10000;
proc datasets lib=work nolist;
   delete SMS_R10psi1_CAT FREQ1-FREQ10000;
run;
quit;
*Excluding the temporary datasets;
proc datasets lib=work nolist;
   delete F11A1-F11A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F12A1-F12A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F13A1-F13A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F21A1-F21A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F22A1-F22A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F23A1-F23A10000;
run;
quit;

*******************************************************************************
*  APPENDING DATASETS FREP1, FREP2, ..., FREPrep, which contain the frequencies *
*  associated to the tables for each repetition                                *
*******************************************************************************;
%macro joinFREP(rep);
        %do r=1 %to &rep;
         proc append
              BASE=FREQS data=FREP&r;
         run;
        %end;
%mend joinFREP;

%joinFREP(10000);
```

```
*Excluding the temporary datasets FREP1-FREP10000;
proc datasets lib=work nolist;
   delete FREP1-FREP10000;
run;
quit;

* Saving the dataset with frequencies counts for simulation SMS_R10psi1";
data SMS.SMS_R10psi1_FREQS;
    set FREQS;
    keep REP F11 F12 F13 F21 F22 F23;
run;

* Importing the datasets SMS_R10psi1_PSI1.DAT - SMS_R10psi1_PSI10.DAT"
  produced by CROSSPSI, which contains the Plackett generalized correlation (TIPO_C_P);
%macro importPSI;
        %do r=1 %to 10;
                proc import datafile ="c:\vigo\simul\R10M\psi1\sms\SMS_R10psi1_PSI&r..dat"
                out = PSI&r
                        dbms=tab replace;
                run;
        %end;
%mend importPSI;

%importPSI;

* Appending datasets PSI1-PSI10;
%macro joinPSI;
        %do r=1 %to 10;
         proc append
              BASE=PSI data=PSI&r force;
         run;
        %end;
%mend joinPSI;

%joinPSI;
*Excluding the temporary datasets PSI1-PSI10;
proc datasets lib=work nolist;
   delete PSI1-PSI10;
run;
quit;

data FIT;
        set FREQS;
        set PSI;
* Calculating the expected frequencies under the null hypothesis of the data follow
  a Type C Bivariate Normal Distribution. Algorithm taken from Fachel (1986, p.59);

* CONVERTING THE ESTIMATE OF PSI FROM THE SOFTWARE CROSSPI, USING THE PLACKETT APPROACH;
        PSI = (ARCOS(-TIPO_C_P)/(3.14159265358979323846264338327795 - ARCOS(-TIPO_C_P)))**2;
              /* See Johnson (1987, p.193) */

        F1 = (F11 + F12 + F13)/500;
        G1 = (F11 + F21)/500;
        S11 = 1 + (PSI - 1)*(F1 + G1);
        A11 = S11**2 - 4*PSI*(PSI - 1)*F1*G1;
        P11 = (S11 - SQRT(A11))/(2*(PSI - 1));   /* Expected proportion in the cell (1, 1) */
        E11 = 500 * P11; /* Expected frequency in the cell (1, 1) */

        G2 = (F12 + F22)/500;
        S12 = 1 + (PSI - 1)*(F1 + G1 + G2);
        A12 = S12**2 - 4*PSI*(PSI - 1)*F1*(G1 + G2);
        H12 = (S12 - SQRT(A12))/(2*(PSI - 1));
        P12 = H12 - P11;   /* Expected proportion in the cell (1, 2) */
        E12 = 500 * P12; /* Expected frequency in the cell (1, 2) */

        P13 = F1 - H12; /* Expected proportion in the cell (1, 3) */
        E13 = 500 * P13; /* Expected frequency in the cell (1, 3) */

        P21 = G1 - P11; /* Expected proportion in the cell (2, 1) */
        E21 = 500 * P21; /* Expected frequency in the cell (2, 1) */
```

```
        P22 = G2 - H12 + P11; /* Expected proportion in the cell (2, 2) */
        E22 = 500 * P22; /* Expected frequency in the cell (2, 2) */

        P23 = 1 - F1 - G1 - G2 + H12; /* Expected proportion in the cell (2, 3) */
        E23 = 500 * P23; /* Expected frequency in the cell (2, 3) */

* CHECKING SUMS OF EXPECTED FREQUENCIES;
        ER1 = E11 + E12 + E13;
        ER2 = E21 + E22 + E23;
        ER = ER1 + ER2;

        EC1 = E11 + E21;
        EC2 = E12 + E22;
        EC3 = E13 + E23;
        EC = EC1 + EC2 + EC3;

* CHI SQUARE STATISTICS FOR THE GOODNESS OF FIT TEST;
        FIT_CHISQ = ((F11 - E11)**2)/E11 + ((F12 - E12)**2)/E12 + ((F13 - E13)**2)/E13 +
          ((F21 - E21)**2)/E21 + ((F22 - E22)**2)/E22 + ((F23 - E23)**2)/E23;
                 GL = (2-1)*(3-1) - 1; /*Degrees of freedom */
        P_FIT = 1 - PROBCHI(FIT_CHISQ, GL);

* CREATING AN INDICATOR OF VIOLATION OF THE UNDERLYING TYPE-C BIVARIATE NORMAL DISTRIBUTION;
* IF TYPE_C = 1 MEANS THAT THE SAMPLE CAME FROM A TYPE-C BIVARIATE NORMAL DISTRIBUTION (ALFA 5%);
        if P_FIT > 0.05 then TYPE_C = 1; else TYPE_C = 0;

* CALCULATING THE VARIANCE OF PSI;
        C = 1/(2*(PSI-1)**2);
        dH11 = C*((S11 - 2*(PSI-1)*F1*G1)/SQRT(A11) - 1); /* See Fachel(1986,p.66) */
    dH12 = C*((S12 - 2*(PSI-1)*F1*(G1 + G2))/SQRT(A12) - 1);

    dP11 =  dH11;  /* See Table 3.5 on Fachel(1986, p.65) */
        dP12 =  dH12 - dH11;
        dP21 = - dH11;
        dP13 = - dH12;
        dP22 = - dH12 + dH11;
        dP13 = - dH12;
        dP23 = dH12;

* Expected value of second order differentiation of the logarithm of likelihood with
  respect to PSI  >>>>>  See Fachel(1986, p.67);
        Ed2lnL = 500*(dH11*(-(1/P11)*dP11 + (1/P12)*dP12 + (1/P21)*dP21 - (1/P22)*dP22) +
                       dH12*(-(1/P12)*dP12 + (1/P13)*dP13 + (1/P22)*dP22 - (1/P23)*dP23));
* ASYMPTOTIC VARIANCE OF PSI >>>  See Fachel(1986, p.79);
        VARPSI = - 1 / Ed2lnL;
        SE_PSI = SQRT(VARPSI);
* ASYMPTOTIC VARIANCE OF log(PSI) >>>  See Fachel(1986, p.79-80);
        LNPSI = LOG(PSI);
    VLNPSI = ((1/PSI)**2)*VARPSI;
* IC 95% FOR LN(PSI);
    LOW95_LNPSI = LNPSI - 1.95996*SQRT(VLNPSI);
        UPP95_LNPSI = LNPSI + 1.95996*SQRT(VLNPSI);
* IC 95% FOR PSI;
        LOW95_PSI = EXP(LOW95_LNPSI);
        UPP95_PSI = EXP(UPP95_LNPSI);
run;

* DATASET WITH THE MAIN RESULTS OF SIMULATION;
data SMS_R1Opsi1_FIT;
        set FIT;
        set SMS.SMS_R1Opsi1_POM;   /* READING PARAMETER ESTIMATES FOR THE PROP. ODDS MODEL */;

        LOW95_BETA = BETA - 1.95996*SE_BETA;
        UPP95_BETA = BETA + 1.95996*SE_BETA;
        OR = exp(BETA);
        LOW95_OR = exp(LOW95_BETA);
        UPP95_OR = exp(UPP95_BETA);

* CREATING AN INDICATOR OF THE VIOLATION PROPORTIONAL ODDS ASSUMPTION;
* IF P_ODDS = 1 MEANS THAT THE PROPORTIONAL ODDS ASSUMPTION MAY BE ACCEPT (ALFA 5%);
        if P_CMT > 0.05 then P_ODDS = 1; else P_ODDS = 0;
```

```
        TRUEPSI = 1;   * SETTING THE TRUE VALUE OF PSI ;

*       QUADRATIC ERROR OF PSI IN RELATION TO THE TRUE VALUE OF PSI;
        QERROR_PSI_TRUEPSI = (PSI - TRUEPSI)**2;

*       QUADRATIC ERROR OF PSI IN RELATION TO THE ODDS RATIO ESTIMATED BY THE PROPORTIONAL ODDS MODEL;
        QERROR_PSI_OR = (PSI - OR)**2;

*       QUADRATIC ERROR OF THE ODDS RATIO ESTIMATED BY THE PROPORTIONAL ODDS MODEL IN RELATION TO THE TRUE PSI;
        QERROR_OR_TRUEPSI = (OR - TRUEPSI)**2;

*       ERROR OF PSI IN RELATION TO THE ODDS RATIO ESTIMATED BY THE PROPORTIONAL ODDS MODEL;
        ERROR_PSI_OR = (PSI - OR);

*       ERROR OF PSI IN RELATION TO THE TRUE VALUE OF PSI;
        ERROR_PSI_TRUEPSI = (PSI - TRUEPSI);

*       ERROR OF PSI IN RELATION TO THE TRUE VALUE OF PSI;
        ERROR_OR_TRUEPSI = (OR - TRUEPSI);

*       ABSOLUTE ERROR OF PSI IN RELATION TO THE ODDS RATIO ESTIMATED BY THE PROPORTIONAL ODDS MODEL;
        ABSERROR_PSI_OR = ABS(PSI - OR);

*       ABSOLUTE ERROR OF PSI IN RELATION TO THE TRUE VALUE OF PSI;
        ABSERROR_PSI_TRUEPSI = ABS(PSI - TRUEPSI);

*       ABSOLUTE ERROR OF OR IN RELATION TO THE TRUE VALUE OF PSI;
        ABSERROR_OR_TRUEPSI = ABS(OR - TRUEPSI);

        keep REP CRUZAMENTO FIT_CHISQ GL P_FIT PSI SE_PSI LOW95_PSI UPP95_PSI _STATUS_ INTERCEPT_1
         INTERCEPT_2 BETA SE_BETA _LNLIKE_ P_CMT LOW95_BETA UPP95_BETA OR LOW95_OR UPP95_OR
         TYPE_C P_ODDS QERROR_PSI_TRUEPSI QERROR_OR_TRUEPSI QERROR_PSI_OR ERROR_PSI_TRUEPSI ERROR_OR_TRUEPSI
         ERROR_PSI_OR ABSERROR_PSI_TRUEPSI ABSERROR_OR_TRUEPSI ABSERROR_PSI_OR;
run;

*****************************************************
*       SAVING RESULTS OF SIMULATION SMS_R1Opsi1    *
*****************************************************;
data SMS.SMS_R1Opsi1_FIT;
     set SMS_R1Opsi1_FIT;
run;


proc freq data=SMS_R1Opsi1_FIT;
        tables _STATUS_ P_ODDS*TYPE_C;
run;

*** Print descriptive statistics for analysis variables ***;
      title;
      footnote;
proc means data=SMS_R1Opsi1_FIT fw=12 maxdec=8 mean std min max;
   var PSI OR QERROR_PSI_TRUEPSI QERROR_PSI_OR QERROR_OR_TRUEPSI ERROR_PSI_OR
       ERROR_PSI_TRUEPSI ERROR_OR_TRUEPSI ABSERROR_PSI_TRUEPSI ABSERROR_OR_TRUEPSI ABSERROR_PSI_OR; ;
   attrib _all_ label='';
run;
```

C1.4: ANÁLISE DESCRITIVA

(SMS_R10PSI1_DES.SAS – SIMÉTRICA SUAVE, $\psi = 1$)

```
*Redirecting and saving the log to the file named "SMS_R10PSI1_DES.LOG";
proc printto log="c:\vigo\simul\R10M\PSI1\SMS\SMS_R10PSI1_DES.LOG";
run;
* Redirecting and saving the output to the file named "SMS_R10PSI1_DES.LST";
proc printto print="c:\vigo\simul\R10M\PSI1\SMS\SMS_R10PSI1_DES.LST";
run;
*********************************************************************
*  Filename: SMS_R10PSI1_DES.SAS                                    *
*  Date: 08/27/2004                                                 *
*                                                                   *
*  Function:                                                        *
*    1) To produce boxplots for the estimative of PSI and OR, and   *
*       also for the error and the quadratic error in relation of   *
*       the true value of PSI and among the estimatives using       *
*       CROSSPSI and PO MODEL                                       *
*                                                                   *
*    SAMPLE SIZE: 500                                               *
*    NUMBER OF REPETITIONS: 10000                                   *
*    CATEGORIZATION: SMS                                            *
*    TRUEPSI = 1                                                    *
*********************************************************************;
OPTIONS OBS=MAX;

LIBNAME SMS V7 'c:\vigo\simul\R10M\PSI1\SMS';
options formchar='|----|+|---+=|-/\<>*';
options linesize=120 ps=59 nocenter nodate nonumber;

data TEMP1;
        set SMS.SMS_R10PSI1_FIT;
        if P_ODDS = 1 and TYPE_C = 1
    then SUPOS = 1; * CASES SATISFYING SUPOSITIONS OF BOTH MODELS;
    else SUPOS = 0;
        * OTHER DEFINITION, NOW WITH 4 CATEGORIES;
    if P_ODDS = 1 and TYPE_C = 1
    then SUPOS4 = 1;    * CASES SATISFYING SUPOSITIONS OF BOTH MODELS;
    else if P_ODDS = 1 and TYPE_C = 0
         then SUPOS4 = 2;    *CASES NOT SATISFYING PARALEL LINES;
         else if P_ODDS = 0 and TYPE_C = 1
                       then SUPOS4 = 3;    * CASES NOT FITTING TO TYPE-C DISTRIBUTION;
                            else SUPOS4 = 4;    * CASES NOT SATISFYING BOTH CONDITIONS;
run;

* SUMMARY DESCRIPTIVES - MODEL SUPOSITIONS;
*** Table Analysis ***;
title;
proc freq data=TEMP1 ORDER=INTERNAL;
   tables TYPE_C*P_ODDS SUPOS SUPOS4/ NOPERCENT NOROW NOCOL;
run;

* SUMMARY DESCRIPTIVES - OVERALL;
*proc means data=TEMP1 fw=8 maxdec=4 mean std n min max;
proc means data=TEMP1 n mean std min max;
   var PSI OR QERROR_PSI_OR QERROR_PSI_TRUEPSI QERROR_OR_TRUEPSI;
   attrib _all_ label='';
run;
* SUMMARY DESCRIPTIVES - BY MODEL SUPOSITION;
proc means data=TEMP1 n mean std min max;
   var PSI OR QERROR_PSI_OR QERROR_PSI_TRUEPSI QERROR_OR_TRUEPSI;
   class SUPOS;
   attrib _all_ label='';
run;

proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

C2: $\psi = 1$ COM CATEGORIZAÇÃO CONCENTRADA

(15%, 70%, 15%)

C2.1: CATEGORIZA MARGINAIS

C2.2: PREPARA DADOS PARA O CROSSPSI

C2.3: IMPORTA ESTIMATIVAS DO CROSSPSI E TESTA AJUSTAMENTO

C2.4: ANÁLISE DESCRITIVA

C2.1: CATEGORIZA MARGINAIS

(SSS_R10PSI1.SAS – SIMÉTRICA CONCENTRADA, $\psi = 1$)

```
* Redirecting and saving the log to the file named "SSS_R1Opsi1.LOG";
proc printto log="c:\vigo\simul\R1OM\psi1\SSS\SSS_R1Opsi1.LOG";
run;
*********************************************************************************
*                                                                               *
*  Filename: SSS_R1Opsi1.SAS                                                    *
*  Date: 01/13/2004                                                             *
*  Template: SMS_R1Opsi1.SAS                                                    *
*                                                                               *
*  Function: To Adjust the Proportional Odds Model AND to save the datasets with      *
*            parameter estimates                                                *
*                                                                               *
*    SAMPLE SIZE: 500                                                           *
*    NUMDER OF REPETITIONS: 10000                                               *
*    SIMULATION CASE SSS => X: Symmetric, Y: Strong Symmetric                   *
*    PSI = 1                                                                    *
*********************************************************************************;
OPTIONS OBS=MAX;
LIBNAME SMS V7 "c:\vigo\simul\R1OM\psi1\sms";
LIBNAME SSS V7 "c:\vigo\simul\R1OM\psi1\SSS";
options formchar="|----|+|---+=|-/\<>*";
options linesize=120 ps=59;
*OPTION MPRINT MLOGIC SYMBOLGEN ;

* Redirecting and saving the output to the file named "SSS_R1Opsi1.LST";
proc printto print="c:\vigo\simul\R1OM\psi1\SSS\SSS_R1Opsi1.LST";
run;

*****************************************************************
*         Macro to adjust the Proportional Odds Model           *
*****************************************************************;
%macro adjop(rep);
        %do r=1 %to &rep;
                title "SSS_R1Opsi1, REP=&r";
                data R&r;
                        set SMS.R&r;
                    * DICHOTOMIZATION OF Xr;
                             if X&r < probit(0.5) then XCAT&r = 1; else XCAT&r = 2;
                        * CATEGORIZATION OF Yr;
                            if Y&r < probit(0.15)
                                   then YCAT&r = 1;
                                    else if Y&r < probit(0.85)
                                                   then YCAT&r = 2;
                                                   else YCAT&r = 3;
                 run;
/*
                *** Table Analysis ***;
                proc freq data=R&r;
                   tables XCAT&r*YCAT&r / NOPERCENT NOROW NOCOL;
                        title "CROSSTABS XCAT&r x YCAT&r";
                run;
*/
                * ADJUSTING THE PROPORTIONAL ODDS MODEL;
                ods output Logistic.CumulativeModelTest=CMT&r
                         (keep=ProbChiSq
                              rename=(ProbChiSq=P_CMT));
                ods output Logistic.ParameterEstimates=PAREST&r
                         (keep=STDERR);

                *** Logistic Regression Analysis ***;
                proc LOGISTIC data=R&r outest=set1est&r;
                   class XCAT&r / param=reference ref=last;
                   model YCAT&r = XCAT&r;
                   title "MODEL for YCAT&r x XCAT&r";
                run;
        %end;
%mend adjop;
```

```
*******************************************************************************
*                                                                             *
*    Macro used to join the datasets named SET1EST1, SET1EST2,...,SET1ESTrep  *
*    which contains the estimative of the parameters INTERCEPT_1, INTERCEPT_2 and *
*    BETA. Note that we changed the original names of the coeficient of regression*
*    associated to the model adjusted to each repetition (r=1,...,rep) using the  *
*    auxiliary datasets S1, S2,...,Srep. The parameter "z" in this macro is used  *
*    just to compose the names of the original coefficients of regression named   *
*    XCAT12, XCAT22,....,XCATrep2.                                            *
*                                                                             *
*    NOTE: use z=1 if the reference category is reference=last, or            *
*              z=2 if the reference category is reference=first               *
*    rep = NUMBER OF REPETITIONS                                              *
*******************************************************************************;
%macro xcat(rep,z);
        %do r=1 %to &rep;
                data S&r;
            length _LINK_ $ 8 _TYPE_ $ 8 _STATUS_ $ 11 _NAME_ $ 10 INTERCEPT_1 8 INTERCEPT_2 8 BETA 8 _LNLIKE_ 8;
                        set Work.Set1est&r;
                        BETA=XCAT&r&z;
                        drop XCAT&r&z;
                run;
        %end;
%mend xcat;

*******************************************************************************
*                                                                             *
*    Macro used to select the standard error of BETA and put it in the file with  *
*    just one record, corresponding to the repetition. The standard erros are     *
*    saved on files named PAREST1, PAREST2, ..., PARESTrep, and the standar error *
*    will be saved in the respective datasets STDERR1, STDERR2, ..., STDERRrep    *
*                                                                             *
*******************************************************************************;
%macro stderr(rep);
        %do r=1 %to &rep;
                data STDERR&r;
                        set Work.PAREST&r;
                        if _n_ = 3; /* Excluding the first two lines with INTERCEPT standard error */
                        SE_BETA=STDERR;
                        keep SE_BETA;
                run;
        %end;
%mend stderr;

* Calling macros ADJOP, XCAT and STDERR for 10000 repetitions;
%adjop(10000);
%xcat(10000,1);
%stderr(10000);

*******************************************************************************
* APPENDING DATASETS S1, S2, ..., Srep, which contain the parameter estimates *
* and other information about the model adjusting procedure                   *
*******************************************************************************;
%macro joinS(rep);
        %do r=1 %to &rep;
         proc append
              BASE=SSS_R10psi1_PAR data=S&r;
         run;
        %end;
%mend joinS;
%joinS(10000);

proc datasets lib=work nolist;
   delete S1-S10000;
run;
quit;

proc datasets lib=work nolist;
   delete SET1EST1-SET1EST10000;
run;
quit;
```

```
*************************************************************************************************
*  APPENDING DATASETS CMT1, CMT2, ..., CMTrep, which contain the p-values associated to the *
*  Score Test for the Proportional Odds Assumption, obtained using the ODS function         *
*************************************************************************************************;
%macro joinCMT(rep);
        %do r=1 %to &rep;
         proc append
              BASE=CMT data=CMT&r;
         run;
        %end;
%mend joinCMT;

%joinCMT(10000);

*Excluding the temporary datasets Cmt1-Cmt10000;
proc datasets lib=work nolist;
   delete Cmt1-Cmt10000;
run;
quit;

*************************************************************************************
*  APPENDING DATASETS STDERR1, STDERR2, ..., STDERRrep, with the standard error of BETA *
*************************************************************************************;
%macro joinSTD(rep);
        %do r=1 %to &rep;
         proc append
              BASE=STDERR data=STDERR&r;
         run;
        %end;
%mend joinSTD;

%joinSTD(10000);

*Excluding the temporary datasets STDERR1-STDERR10000 and PAREST1-PAREST10000;
proc datasets lib=work nolist;
   delete STDERR1-STDERR10000 PAREST1-PAREST10000;
run;
quit;
*************************************************************************************************
*  Merging datasets SSS_R1Opsi1_PAR, CMT and STDERR to produce the dataset SSS_R1Opsi1_POM  *
*  with the parameter estimates, -ln likelihood, p-value for the Score Test for the         *
*  Proportional Odds Assumption, and other information about the adjusted models            *
*************************************************************************************************;
data SSS_R1Opsi1_POM;
   merge SSS_R1Opsi1_PAR (in=TABLE1) CMT (in=TABLE2) STDERR (in=TABLE3);
   if TABLE1 and TABLE2 and TABLE3;
run;
*************************************************************************************
*  Saving the dataset with the parameter estimates for simulation SSS_R1Opsi1              *
*************************************************************************************;
data SSS.SSS_R1Opsi1_POM;
     set SSS_R1Opsi1_POM;
         REP=_n_;
     keep REP _STATUS_ INTERCEPT_1 INTERCEPT_2 BETA SE_BETA _LNLIKE_ P_CMT;
run;
*************************************************************************
*  CREATING THE DATASET USED AS AN INPUT IN CROSSPSI SOFTWARE           *
*************************************************************************;
title "CROSSPSI DATASET FOR SSS_R1Opsi1";
%macro mergeCATX(part, first, last);
    data SSS_R1Opsi1_CATX&part;
              set R&first;
         keep ID XCAT&first;
    run;
         %do r = &first+1 %to &last;
         data SSS_R1Opsi1_CATX&part;
              merge SSS_R1Opsi1_CATX&part R&r;
                          by ID;
              keep ID XCAT&first-XCAT&r;
         run;
         %end;
%mend mergeCATX;
```

```
%mergeCATX(1,1,1000);
%mergeCATX(2,1001,2000);
%mergeCATX(3,2001,3000);
%mergeCATX(4,3001,4000);
%mergeCATX(5,4001,5000);
%mergeCATX(6,5001,6000);
%mergeCATX(7,6001,7000);
%mergeCATX(8,7001,8000);
%mergeCATX(9,8001,9000);
%mergeCATX(10,9001,10000);

%macro mergeCATY(part, first, last);
    data SSS_R1Opsi1_CATY&part;
              set R&first;
          keep ID YCAT&first;
    run;
         %do r = &first+1 %to &last;
         data SSS_R1Opsi1_CATY&part;
              merge SSS_R1Opsi1_CATY&part R&r;
                          by ID;
              keep ID YCAT&first-YCAT&r;
         run;
         %end;
%mend mergeCATY;

%mergeCATY(1,1,1000);
%mergeCATY(2,1001,2000);
%mergeCATY(3,2001,3000);
%mergeCATY(4,3001,4000);
%mergeCATY(5,4001,5000);
%mergeCATY(6,5001,6000);
%mergeCATY(7,6001,7000);
%mergeCATY(8,7001,8000);
%mergeCATY(9,8001,9000);
%mergeCATY(10,9001,10000);

* JOINING THESE DATASETS;
%macro joinCAT;
    %do r=1 %to 10;
         data SSS_R1Opsi1_CAT&r;
              merge WORK.SSS_R1Opsi1_CATX&r (in=TABLE1) WORK.SSS_R1Opsi1_CATY&r (in=TABLE2);
              by ID;
              if TABLE1 and TABLE2;
         run;
    %end;
%mend joinCAT;

%joinCAT;

%macro saveCAT;
    %do r=1 %to 10;
    * SAVING CROSSPSI (CATEGORICAL) DATASETS .SAS7BDAT;
         data SSS.SSS_R1Opsi1_CAT&r;
              set SSS_R1Opsi1_CAT&r;
         run;
/*
          * EXPORTING THE CROSSPSI (CATEGORICAL) DATASET IN TAB DELIMITED FORMAT (.DAT);
    * In fact, this kind of file should have .TXT extension;
         proc export data=SSS_R1Opsi1_CAT&r
                    outfile ="c:\vigo\simul\R1OM\psi1\SSS\SSS_R1Opsi1_CAT&r.dat"
                    dbms=TAB;
         run;
*/
    %end;
%mend saveCAT;
%saveCAT;

proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

C2.2: PREPARA DADOS PARA O CROSSPSI

(SSS_R10PSI1_CAT.SAS – SIMÉTRICA CONCENTRADA, $\psi = 1$)

```
* Redirecting and saving the log to the file named "SSS_R1Opsi1.LOG";
proc printto log="c:\vigo\simul\R1OM\psi1\SSS\SSS_R1Opsi1_cat.LOG";
run;

*******************************************************************************
*                                                                             *
*  Filename: SSS_R1Opsi1_CAT.SAS                                              *
*  Date: 01/13/2004                                                           *
*                                                                             *
*  Function: To read the categorical datasets in .SAS7BDAT format and to save them as *
*            .DAT ( in fact, in .TXT format)                                  *
*                                                                             *
*    SAMPLE SIZE: 500                                                         *
*    NUMDER OF REPETITIONS: 10000                                             *
*    SIMULATION CASE SSS => X: Symmetric, Y: Strong Symmetric                 *
*    PSI = 1                                                                  *
*                                                                             *
*******************************************************************************;
OPTIONS OBS=MAX;
LIBNAME SSS V7 "c:\vigo\simul\R1OM\psi1\SSS";
options formchar="|----|+|---+=|-/\<>*";
options linesize=120 ps=59;
*OPTION MPRINT MLOGIC SYMBOLGEN ;

* Redirecting and saving the output to the file named "SSS_R1Opsi1.LST";
proc printto print="c:\vigo\simul\R1OM\psi1\SSS\SSS_R1Opsi1_cat.LST";
run;


* READING FILES IN .SASS7BDAT;
%macro readCAT;
        %do r=1 %to 10;
                data CAT&r;
                        set SSS.SSS_R1Opsi1_CAT&r;
                RUN;
        %end;
%mend readCAT;

%readCAT;


* SAVING AS .DAT;
%macro saveCAT;
    %do r=1 %to 10;
         * EXPORTING THE CROSSPSI (CATEGORICAL) DATASET IN TAB DELIMITED FORMAT (.DAT);
    * In fact, this kind of file should have .TXT extension;
        proc export data=CAT&r
                  outfile ="c:\vigo\simul\R1OM\psi1\SSS\SSS_R1Opsi1_CAT&r..dat"
                  dbms=TAB;
        run;
    %end;
%mend saveCAT;

%saveCAT;

proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

C2.3: IMPORTA ESTIMATIVAS DO CROSSPSI E TESTA AJUSTAMENTO

(SSS_R10PSI1_FIT.SAS – SIMÉTRICA CONCENTRADA, $\psi = 1$)

```
* Redirecting and saving the log to the file named "SSS_R10PSI1_FIT.LOG";
proc printto log="c:\vigo\simul\R10M\PSI1\SSS\SSS_R10PSI1_FIT.LOG";
run;

*********************************************************************
*                                                                   *
*  Filename: SSS_R10PSI1_FIT.SAS                                    *
*  Date: 01/13/2004                                                 *
*                                                                   *
*  Function:                                                        *
*    1) To produce the Goddness of Fit test to evaluate if the      *
*       data become from a Bivariate Normal Type C distribution     *
*       ALGORITHM: Fachel (1986, p.59)                              *
*                                                                   *
*    2) CREATE A DATASET NAMED SSS_R10PSI1_FIT with the main        *
*       results of simulation SSS_R10PSI1                           *
*                                                                   *
*   Note: THE COEFFICIENT PSI IS ESTIMATED BY THE PLACKETT          *
*         APPROACH (see Johnson (1987, p.193)                       *
*                                                                   *
*    SAMPLE SIZE: 500                                               *
*    NUMBER OF REPETITIONS: 10000                                   *
*    CATEGORIZATION: SSS => X: Symmetric, Y: Strong Symmetric       *
*    TRUEPSI = 1                                                    *
*                                                                   *
*********************************************************************;

OPTIONS OBS=MAX nodate nonumber;
LIBNAME SSS V7 'c:\vigo\simul\R10M\PSI1\SSS';
options formchar='|----|+|---+=|-/\<>*';
options linesize=120 ps=59;
* Redirecting and saving the output to the file named "SSS_R10PSI1_FIT.LST";
proc printto print="c:\vigo\simul\R10M\PSI1\SSS\SSS_R10PSI1_FIT.LST";
run;

* READING DATASETS SSS_R10PSI1_CAT1-SSS_R10PSI1_CAT10;
%macro readCAT;
        %do i=1 %to 10;
                data SSS_R10PSI1_CAT&i;
                        set SSS.SSS_R10PSI1_CAT&i;
                run;
        %end;
%mend readCAT;

%readCAT;

* MERGING DATASETS;
%macro mergeCAT;
    data SSS_R10PSI1_CAT;
              set SSS_R10PSI1_CAT1;
    run;
        %do j = 2 %to 10;
        data SSS_R10PSI1_CAT;
             merge SSS_R10PSI1_CAT SSS_R10PSI1_CAT&j;
                          by ID;
        run;
        %end;
%mend mergeCAT;
%mergeCAT;

*Excluding the temporary datasets SSS_R10PSI1_CAT1-SSS_R10PSI1_CAT10;
proc datasets lib=work nolist;
   delete SSS_R10PSI1_CAT1-SSS_R10PSI1_CAT10;
run;
quit;
```

```
*****************************************************************
* Macro used to obtain the frequencies of the contingency tables    *
* XCAT1 x YCAT1, XCAT2 x YCAT2, ..., XCATrep x YCATrep, and to save *
* them in datasets named FREP1, FREP2, ..., FREPrep                 *
*****************************************************************;
%macro freqdata(rep);
        %do r=1 %to &rep;
                proc freq data=SSS_R1OPSI1_CAT;
                        tables XCAT&r*YCAT&r / NOPERCENT NOROW NOCOL out=FREQ&r;
                        title "CROSSTAB XCAT&r x YCAT&r";
                run;

                data F11A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 1 and YCAT&r = 1
                        then do;
                                if (XCAT&r = 1 and YCAT&r = 1);
                                F11 = COUNT;
                        end;
                        if _n_=1;
                        keep REP F11;
                run;

                data F12A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 1 and YCAT&r = 2
                                then do;
                                        if (XCAT&r = 1 and YCAT&r = 2);
                                        F12 = COUNT;
                                end;
                        if _n_=2;
                        keep REP F12;
                run;

                data F13A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 1 and YCAT&r = 3
                        then do;
                                if (XCAT&r = 1 and YCAT&r = 3);
                                F13 = COUNT;
                        end;
                        if _n_=3;
                        keep REP F13;
                run;

                data F21A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 2 and YCAT&r = 1
                        then do;
                                if (XCAT&r = 2 and YCAT&r = 1);
                                F21 = COUNT;
                        end;
                        if _n_=4;
                        keep REP F21;
                run;

                data F22A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 2 and YCAT&r = 2
                        then do;
                                if (XCAT&r = 2 and YCAT&r = 2);
                                F22 = COUNT;
                        end;
                        if _n_=5;
                        keep REP F22;
                run;
```

```
                data F23A&r;
                        set FREQ&r;
                        REP=&r;
                        if XCAT&r = 2 and YCAT&r = 3
                        then do;
                                if (XCAT&r = 2 and YCAT&r = 3);
                                F23 = COUNT;
                        end;
                        if _n_=6;
                        keep REP F23;
                run;

                * MERGING DATASETS F11A, F12A, F13A, F21A, F22A, F23A;
                data FREP&r;
                        merge WORK.F11A&r (in=TABLE1)
                                  WORK.F12A&r (in=TABLE2)
                                  WORK.F13A&r (in=TABLE3)
                                  WORK.F21A&r (in=TABLE4)
                                  WORK.F22A&r (in=TABLE5)
                                  WORK.F23A&r (in=TABLE6)
                        ;
                run;
        %end;
%mend freqdata;

*CALLING FREQDATA MACRO FOR 10000 REPETITIONS;
%freqdata(10000);


*Excluding the temporary datasets R1-R10000 FREQ1-FREQ10000;
proc datasets lib=work nolist;
   delete SSS_R10PSI1_CAT FREQ1-FREQ10000;
run;
quit;
*Excluding the temporary datasets;
proc datasets lib=work nolist;
   delete F11A1-F11A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F12A1-F12A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F13A1-F13A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F21A1-F21A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F22A1-F22A10000;
run;
quit;
proc datasets lib=work nolist;
   delete F23A1-F23A10000;
run;
quit;


*****************************************************************************
*  APPENDING DATASETS FREP1, FREP2, ..., FREPrep, which contain the frequencies *
*  associated to the tables for each repetition                              *
*****************************************************************************;
%macro joinFREP(rep);
        %do r=1 %to &rep;
         proc append
              BASE=FREQS data=FREP&r;
         run;
        %end;
%mend joinFREP;
```

```
%joinFREP(10000);
*Excluding the temporary datasets FREP1-FREP10000;
proc datasets lib=work nolist;
   delete FREP1-FREP10000;
run;
quit;

* Saving the dataset with frequencies counts for simulation SSS_R10PSI1";
data SSS.SSS_R10PSI1_FREQS;
    set FREQS;
    keep REP F11 F12 F13 F21 F22 F23;
run;

* Importing the datasets SSS_R10PSI1_PSI1.DAT - SSS_R10PSI1_PSI10.DAT"
  produced by CROSSPSI, which contains the Plackett generalized correlation (TIPO_C_P);
%macro importPSI;
        %do r=1 %to 10;
                proc import datafile ="c:\vigo\simul\R10M\PSI1\SSS\SSS_R10PSI1_PSI&r..dat"
                out = PSI&r
                        dbms=tab replace;
                run;
        %end;
%mend importPSI;

%importPSI;

* Appending datasets PSI1-PSI10;
%macro joinPSI;
        %do r=1 %to 10;
         proc append
              BASE=PSI data=PSI&r force;
         run;
        %end;
%mend joinPSI;

%joinPSI;
*Excluding the temporary datasets PSI1-PSI10;
proc datasets lib=work nolist;
   delete PSI1-PSI10;
run;
quit;

data FIT;
        set FREQS;
        set PSI;
* Calculating the expected frequencies under the null hypothesis of the data follow
  a Type C Bivariate Normal Distribution. Algorithm taken from Fachel (1986, p.59);

* CONVERTING THE ESTIMATE OF PSI FROM THE SOFTWARE CROSSPI, USING THE PLACKETT APPROACH;
        PSI = (ARCOS(-TIPO_C_P)/(3.14159265358979323846264338327795 - ARCOS(-TIPO_C_P)))**2;
             /* See Johnson (1987, p.193) */

        F1 = (F11 + F12 + F13)/500;
        G1 = (F11 + F21)/500;
        S11 = 1 + (PSI - 1)*(F1 + G1);
        A11 = S11**2 - 4*PSI*(PSI - 1)*F1*G1;
        P11 = (S11 - SQRT(A11))/(2*(PSI - 1));   /* Expected proportion in the cell (1, 1) */
        E11 = 500 * P11; /* Expected frequency in the cell (1, 1) */

        G2 = (F12 + F22)/500;
        S12 = 1 + (PSI - 1)*(F1 + G1 + G2);
        A12 = S12**2 - 4*PSI*(PSI - 1)*F1*(G1 + G2);
        H12 = (S12 - SQRT(A12))/(2*(PSI - 1));
        P12 = H12 - P11;   /* Expected proportion in the cell (1, 2) */
        E12 = 500 * P12; /* Expected frequency in the cell (1, 2) */

        P13 = F1 - H12; /* Expected proportion in the cell (1, 3) */
        E13 = 500 * P13; /* Expected frequency in the cell (1, 3) */

        P21 = G1 - P11; /* Expected proportion in the cell (2, 1) */
        E21 = 500 * P21; /* Expected frequency in the cell (2, 1) */
```

```
        P22 = G2 - H12 + P11; /* Expected proportion in the cell (2, 2) */
        E22 = 500 * P22; /* Expected frequency in the cell (2, 2) */

        P23 = 1 - F1 - G1 - G2 + H12; /* Expected proportion in the cell (2, 3) */
        E23 = 500 * P23; /* Expected frequency in the cell (2, 3) */

* CHECKING SUMS OF EXPECTED FREQUENCIES;
        ER1 = E11 + E12 + E13;
        ER2 = E21 + E22 + E23;
        ER = ER1 + ER2;

        EC1 = E11 + E21;
        EC2 = E12 + E22;
        EC3 = E13 + E23;
        EC = EC1 + EC2 + EC3;

* CHI SQUARE STATISTICS FOR THE GOODNESS OF FIT TEST;
        FIT_CHISQ = ((F11 - E11)**2)/E11 + ((F12 - E12)**2)/E12 + ((F13 - E13)**2)/E13 +
          ((F21 - E21)**2)/E21 + ((F22 - E22)**2)/E22 + ((F23 - E23)**2)/E23;
                 GL = (2-1)*(3-1) - 1; /*Degrees of freedom */
        P_FIT = 1 - PROBCHI(FIT_CHISQ, GL);

* CREATING AN INDICATOR OF VIOLATION OF THE UNDERLYING TYPE-C BIVARIATE NORMAL DISTRIBUTION;
* IF TYPE_C = 1 MEANS THAT THE SAMPLE CAME FROM A TYPE-C BIVARIATE NORMAL DISTRIBUTION (ALFA 5%);
        if P_FIT > 0.05 then TYPE_C = 1; else TYPE_C = 0;

* CALCULATING THE VARIANCE OF PSI;
        C = 1/(2*(PSI-1)**2);
        dH11 = C*((S11 - 2*(PSI-1)*F1*G1)/SQRT(A11) - 1); /* See Fachel(1986,p.66) */
    dH12 = C*((S12 - 2*(PSI-1)*F1*(G1 + G2))/SQRT(A12) - 1);

    dP11 =  dH11;  /* See Table 3.5 on Fachel(1986, p.65) */
        dP12 =  dH12 - dH11;
        dP21 = - dH11;
        dP13 = - dH12;
        dP22 = - dH12 + dH11;
        dP13 = - dH12;
        dP23 = dH12;

* Expected value of second order differentiation of the logarithm of likelihood with
  respect to PSI  >>>>>  See Fachel(1986, p.67);
        Ed2lnL = 500*(dH11*(-(1/P11)*dP11 + (1/P12)*dP12 + (1/P21)*dP21 - (1/P22)*dP22) +
                        dH12*(-(1/P12)*dP12 + (1/P13)*dP13 + (1/P22)*dP22 - (1/P23)*dP23));
* ASYMPTOTIC VARIANCE OF PSI >>>  See Fachel(1986, p.79);
        VARPSI = - 1 / Ed2lnL;
        SE_PSI = SQRT(VARPSI);
* ASYMPTOTIC VARIANCE OF log(PSI) >>>  See Fachel(1986, p.79-80);
        LNPSI = LOG(PSI);
    VLNPSI = ((1/PSI)**2)*VARPSI;
* IC 95% FOR LN(PSI);
    LOW95_LNPSI = LNPSI - 1.95996*SQRT(VLNPSI);
        UPP95_LNPSI = LNPSI + 1.95996*SQRT(VLNPSI);
* IC 95% FOR PSI;
        LOW95_PSI = EXP(LOW95_LNPSI);
        UPP95_PSI = EXP(UPP95_LNPSI);
run;


* DATASET WITH THE MAIN RESULTS OF SIMULATION;
data SSS_R10PSI1_FIT;
        set FIT;
        set SSS.SSS_R10PSI1_POM;   /* READING PARAMETER ESTIMATES FOR THE PROP. ODDS MODEL */;

        LOW95_BETA = BETA - 1.95996*SE_BETA;
        UPP95_BETA = BETA + 1.95996*SE_BETA;
        OR = exp(BETA);
        LOW95_OR = exp(LOW95_BETA);
        UPP95_OR = exp(UPP95_BETA);

* CREATING AN INDICATOR OF THE VIOLATION PROPORTIONAL ODDS ASSUMPTION;
* IF P_ODDS = 1 MEANS THAT THE PROPORTIONAL ODDS ASSUMPTION MAY BE ACCEPT (ALFA 5%);
        if P_CMT > 0.05 then P_ODDS = 1; else P_ODDS = 0;
```

```
        TRUEPSI = 1;   * SETTING THE TRUE VALUE OF PSI ;

*       QUADRATIC ERROR OF PSI IN RELATION TO THE TRUE VALUE OF PSI;
        QERROR_PSI_TRUEPSI = (PSI - TRUEPSI)**2;

*       QUADRATIC ERROR OF PSI IN RELATION TO THE ODDS RATIO ESTIMATED BY THE PROPORTIONAL ODDS MODEL;
        QERROR_PSI_OR = (PSI - OR)**2;

*       QUADRATIC ERROR OF THE ODDS RATIO ESTIMATED BY THE PROPORTIONAL ODDS MODEL IN RELATION TO THE TRUE PSI;
        QERROR_OR_TRUEPSI = (OR - TRUEPSI)**2;

*       ERROR OF PSI IN RELATION TO THE ODDS RATIO ESTIMATED BY THE PROPORTIONAL ODDS MODEL;
        ERROR_PSI_OR = (PSI - OR);

*       ERROR OF PSI IN RELATION TO THE TRUE VALUE OF PSI;
        ERROR_PSI_TRUEPSI = (PSI - TRUEPSI);

*       ERROR OF PSI IN RELATION TO THE TRUE VALUE OF PSI;
        ERROR_OR_TRUEPSI = (OR - TRUEPSI);

*       ABSOLUTE ERROR OF PSI IN RELATION TO THE ODDS RATIO ESTIMATED BY THE PROPORTIONAL ODDS MODEL;
        ABSERROR_PSI_OR = ABS(PSI - OR);

*       ABSOLUTE ERROR OF PSI IN RELATION TO THE TRUE VALUE OF PSI;
        ABSERROR_PSI_TRUEPSI = ABS(PSI - TRUEPSI);

*       ABSOLUTE ERROR OF OR IN RELATION TO THE TRUE VALUE OF PSI;
        ABSERROR_OR_TRUEPSI = ABS(OR - TRUEPSI);

        keep REP CRUZAMENTO FIT_CHISQ GL P_FIT PSI SE_PSI LOW95_PSI UPP95_PSI _STATUS_ INTERCEPT_1
         INTERCEPT_2 BETA SE_BETA _LNLIKE_ P_CMT LOW95_BETA UPP95_BETA OR LOW95_OR UPP95_OR
         TYPE_C P_ODDS QERROR_PSI_TRUEPSI QERROR_OR_TRUEPSI QERROR_PSI_OR ERROR_PSI_TRUEPSI ERROR_OR_TRUEPSI
         ERROR_PSI_OR ABSERROR_PSI_TRUEPSI ABSERROR_OR_TRUEPSI ABSERROR_PSI_OR;
run;

*****************************************************
*       SAVING RESULTS OF SIMULATION SSS_R1OPSI1    *
*****************************************************;
data SSS.SSS_R1OPSI1_FIT;
     set SSS_R1OPSI1_FIT;
run;


proc freq data=SSS_R1OPSI1_FIT;
        tables _STATUS_ P_ODDS*TYPE_C;
run;


*** Print descriptive statistics for analysis variables ***;
      title;
      footnote;
proc means data=SSS_R1OPSI1_FIT fw=12 maxdec=8 mean std min max;
   var PSI OR QERROR_PSI_TRUEPSI QERROR_PSI_OR QERROR_OR_TRUEPSI ERROR_PSI_OR
       ERROR_PSI_TRUEPSI ERROR_OR_TRUEPSI ABSERROR_PSI_TRUEPSI ABSERROR_OR_TRUEPSI ABSERROR_PSI_OR; ;
   attrib _all_ label='';
run;
```

C2.4: ANÁLISE DESCRITIVA

(SSS_R10PSI1_DES.SAS – SIMÉTRICA CONCENTRADA, $\psi = 1$)

```
*Redirecting and saving the log to the file named "SSS_R10PSI1_DES.LOG";
proc printto log="c:\vigo\simul\R10M\PSI1\SSS\SSS_R10PSI1_DES.LOG";
run;
* Redirecting and saving the output to the file named "SSS_R10PSI1_DES.LST";
proc printto print="c:\vigo\simul\R10M\PSI1\SSS\SSS_R10PSI1_DES.LST";
run;
*********************************************************************
*  Filename: SSS_R10PSI1_DES.SAS                                    *
*  Date: 08/27/2004                                                 *
*                                                                   *
*  Function:                                                        *
*    1) To produce boxplots for the estimative of PSI and OR, and   *
*       also for the error and the quadratic error in relation of   *
*       the true value of PSI and among the estimatives using       *
*       CROSSPSI and PO MODEL                                       *
*                                                                   *
*    SAMPLE SIZE: 500                                               *
*    NUMBER OF REPETITIONS: 10000                                   *
*    CATEGORIZATION: SSS                                            *
*    TRUEPSI = 1                                                    *
*                                                                   *
*********************************************************************;
OPTIONS OBS=MAX;

LIBNAME SSS V7 'c:\vigo\simul\R10M\PSI1\SSS';
options formchar='|----|+|---+=|-/\<>*';
options linesize=120 ps=59 nocenter nodate nonumber;

data TEMP1;
        set SSS.SSS_R10PSI1_FIT;
        if P_ODDS = 1 and TYPE_C = 1
    then SUPOS = 1; * CASES SATISFYING SUPOSITIONS OF BOTH MODELS;
    else SUPOS = 0;
        * OTHER DEFINITION, NOW WITH 4 CATEGORIES;
    if P_ODDS = 1 and TYPE_C = 1
    then SUPOS4 = 1;    * CASES SATISFYING SUPOSITIONS OF BOTH MODELS;
    else if P_ODDS = 1 and TYPE_C = 0
         then SUPOS4 = 2;    *CASES NOT SATISFYING PARALEL LINES;
         else if P_ODDS = 0 and TYPE_C = 1
                       then SUPOS4 = 3;    * CASES NOT FITTING TO TYPE-C DISTRIBUTION;
                            else SUPOS4 = 4;    * CASES NOT SATISFYING BOTH CONDITIONS;
run;

* SUMMARY DESCRIPTIVES - MODEL SUPOSITIONS;
*** Table Analysis ***;
title;
proc freq data=TEMP1 ORDER=INTERNAL;
   tables TYPE_C*P_ODDS SUPOS SUPOS4/ NOPERCENT NOROW NOCOL;
run;

* SUMMARY DESCRIPTIVES - OVERALL;
*proc means data=TEMP1 fw=8 maxdec=4 mean std n min max;
proc means data=TEMP1 n mean std min max;
   var PSI OR QERROR_PSI_OR QERROR_PSI_TRUEPSI QERROR_OR_TRUEPSI;
   attrib _all_ label='';
run;
* SUMMARY DESCRIPTIVES - BY MODEL SUPOSITION;
proc means data=TEMP1 n mean std min max;
   var PSI OR QERROR_PSI_OR QERROR_PSI_TRUEPSI QERROR_OR_TRUEPSI;
   class SUPOS;
   attrib _all_ label='';
run;
proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```