UNIVERSIDADE FEDERAL DE JUIZ DE FORA

Anna Claudia Mancini da Silva Carneiro

ANÁLISE DE DADOS LONGITUDINAIS ATRAVÉS DE MODELOS MARGINAIS

Juiz de Fora, julho de 2012

Anna Claudia Mancini da Silva Carneiro

ANÁLISE DE DADOS LONGITUDINAIS ATRAVÉS DE MODELOS MARGINAIS

Trabalho de Conclusão de Curso apresentado para obtenção de Diploma de Graduação em Estatística pela Universidade Federal de Juiz de Fora.

Orientador: Prof. Ph.D. Marcel de Toledo Vieira

Juiz de Fora, julho de 2012

Anna Claudia Mancini da Silva Carneiro

ANÁLISE DE DADOS LONGITUDINAIS ATRAVÉS DE DADOS LONGITUDINAIS

Trabalho de Conclusão de Curso apresentado para obtenção de Diploma de Graduação em Estatística pela Universidade Federal de Juiz de Fora, UFJF.

Aprovado em 10 de julho de 2012.

BANCA EXAMINADORA

_____

Prof. Ph.D. Marcel de Toledo Vieira – UFJF

_____

Prof. Ph.D. Ronaldo Rocha Bastos – UFJF

_____

Prof. Dr. Clécio da Silva Ferreira – UFJF

AGRADECIMENTOS

À minha mãe, minha avó, meu avô, meu padrasto, meu irmão e minha tia, por serem meu alicerce e meu porto seguro, por me ensinarem com tanto amor e carinho, por estarem ao meu lado em todas as situações, boas ou ruins, sempre me apoiando e por me incentivarem nessa etapa tão cheia de dúvidas e ansiedade.

À minha mãe e à minha avó em especial, porque eu devo a elas tudo o que sou e no que eu acredito, porque são mulheres extraordinárias e meus exemplos. Não há palavras para descrever o tamanho do meu amor por vocês. Vou levá-las para sempre comigo!

À Camila, Franciele, Carolina e Isabela, que muito me ajudaram durante esses anos, seja nos estudos, seja nas dificuldades, e com quem passei momentos inesquecíveis de muita alegria.

Ao Pedro, por estar comigo nos anos de fundamental importância da minha vida, por ser mais que meu namorado, ser meu melhor amigo, meu companheiro, por me apoiar e incentivar sempre, por ter sempre uma palavra de carinho, um sorriso de estímulo e um gesto de amor.

Aos meus professores, que contribuíram para o meu desenvolvimento acadêmico; especialmente ao Marcel, meu professor orientador, que, com sua paciência e ânsia de dividir seus conhecimentos, foi fundamental para a conclusão deste trabalho.

À Universidade Federal de Juiz de Fora, por me proporcionar a oportunidade de estudar em uma instituição de ensino superior de qualidade.

É com grande alegria e alívio que concluo este curso, ansiosa para enfrentar as surpresas e os desafios que me esperam no futuro.

RESUMO

Pesquisas longitudinais são aquelas em que as mesmas (ou diferentes) variáveis são medidas para os mesmos indivíduos em, pelo menos, dois pontos distintos do tempo, permitindo assim estudar a mudança de comportamento, por exemplo, de um mesmo indivíduo ou variações entre indivíduos ao longo do tempo, consistindo também em fontes de dados para estudos sobre mudanças demográficas e socioeconômicas, estudos epidemiológicos e ambientais, entre outros. O objetivo da presente monografia é apresentar uma metodologia muito utilizada para a análise de dados longitudinais, os modelos marginais, a partir de uma aplicação aos dados provenientes de uma pesquisa britânica do tipo painel de base domiciliar, a *British Household Panel Survey* (BHPS). Além do ajuste de modelos marginais, apresentaremos uma breve análise exploratória dos dados longitudinais.

Palavras-chave: Dados longitudinais; modelos marginais; estimação; satisfação.

ABSTRACT

Longitudinal studies are those in which the same (or different) variables are measured for the same individuals in at least two different points of time, thus allows to study the change in behavior, for example, for the same individual or variations between individuals over time, consisting of sources of data for studies on demographic changes and socio-economic, epidemiological and environmental studies, among others. The purpose of this monograph is to present a methodology often used for the analysis of longitudinal data, the marginal models, through an application to data from the British Household Panel Survey (BHPS). In addition to the marginal model fitting procedures, we present a brief exploratory analysis of longitudinal data.

Keywords: Longitudinal data; marginal models; estimation; satisfaction.

LISTA DE QUADROS E FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Pesquisas longitudinais são aquelas em que as mesmas (ou diferentes) variáveis são medidas para os mesmos indivíduos em, pelo menos, dois pontos distintos do tempo. Dessa forma, consistem em uma importante fonte para estudos sobre mudanças demográficas e socioeconômicas, estudos epidemiológicos e ambientais, entre outros.

A análise dos dados longitudinais permite estudar a mudança de comportamento, por exemplo, de um mesmo indivíduo ou variações entre indivíduos ao longo do tempo. Os resultados provenientes de análises desse tipo em estudos de cunho social ou ambiental auxiliam as organizações competentes a tomarem decisões corretas para solucionar seus problemas.

Apesar de ser uma fonte eficiente de informações, os dados longitudinais apresentam dificuldades em sua coleta e análise. Durante uma pesquisa longitudinal, um indivíduo pode não responder a uma determinada pergunta em uma determinada ocasião da pesquisa, e na seguinte responder. Ou pode, simplesmente, com o passar do tempo, optar por não participar mais da pesquisa, o que acontece com certa frequência. Além disso, as pesquisas longitudinais são normalmente estudos de grande porte, sendo muito custosos para os financiadores.

Sob o ponto de vista metodológico, a análise de dados longitudinais é um pouco mais complexa que a análise de dados transversais, que são coletados em apenas uma ocasião, ou no caso de repetição da pesquisa, coletados para amostras diferentes. Para dados longitudinais, é preciso levar em consideração os efeitos do tempo sobre as respostas do indivíduo, além da própria correlação entre as respostas para o mesmo indivíduo ao longo do tempo.

O objetivo da presente monografia é apresentar uma técnica muito usada para a análise de dados longitudinais, os modelos marginais, a partir de uma aplicação a dados provenientes de uma pesquisa longitudinal real. Na Seção 2, descreveremos a *British Household Panel Survey* (BHPS), pesquisa britânica do tipo painel de base domiciliar realizada quase anualmente, fonte dos dados utilizados no nosso estudo. Na Seção 3, nossa análise começará pela parte exploratória, com a apresentação de tabelas e gráficos, alguns específicos para dados longitudinais, de forma breve, já que esse não é o principal objetivo da monografia. Em seguida, na Seção 4, faremos a análise longitudinal, de fato, através dos modelos marginais. As considerações finais acerca dos resultados obtidos encerram este trabalho, na Seção 5.

2 *BRITISH HOUSEHOLD PANEL SURVEY*

A *British Household Panel Survey* (BHPS) é uma pesquisa longitudinal de base domiciliar do Reino Unido. Este estudo vem sendo conduzido anualmente, desde 1991, e cerca de 5.500 domicílios e 10.300 pessoas foram selecionadas inicialmente em 250 setores postais (áreas) da Grã-Bretanha. Com o objetivo de atualizar a amostra da BHPS, foram adicionados 1.500 domicílios da Escócia e 1.500 do País de Gales em 1999 e 2.000 domicílios da Irlanda do Norte em 2001. Atualmente, a amostra da BHPS contém aproximadamente 10.000 domicílios de todo o Reino Unido.

Para a coleta dos dados, a BHPS adota um esquema de amostragem estratificada e conglomerada em múltiplos estágios. Na primeira ocasião da pesquisa, em 1991, 250 setores postais foram selecionados como unidades primárias de amostragem (UPAs), com reposição e com probabilidades de seleção proporcionais ao tamanho, utilizando amostragem sistemática. Inicialmente a população foi dividida em 18 estratos regionais. Dentro de cada estrato, as UPAs foram classificadas e divididas em estratos principais de tamanho aproximado à proporção de chefes de domicílios que possuíam empregos profissionalizados ou ocupavam posições gerenciais. Dentro desses estratos principais, as UPAs foram reclassificadas de acordo com a proporção de indivíduos com idade suficiente para aposentadoria. Os estratos principais foram divididos, então, em dois estratos menores, representando áreas rurais, em que as UPAs foram selecionadas proporcionalmente à população empregada na agricultura, e áreas urbanas, cujas UPAs foram sorteadas proporcionalmente a suas populações de indivíduos que moram sozinhos nos domicílios e de indivíduos com idade suficiente para aposentadoria. As unidades secundárias de amostragem (USAs) utilizadas foram os endereços, selecionadas também por amostragem sistemática. Em endereços com até 3 domicílios presentes, todos os domicílios foram incluídos, enquanto nos endereços que apresentavam mais de 3 domicílios, um procedimento de seleção aleatória foi utilizado para sortear apenas 3 domicílios. Todos os residentes nos domicílios com idade igual ou superior a 16 anos foram considerados participantes da pesquisa. Para maiores informações sobre o desenho amostral da BHPS, consultar Vieira (2009) e Salgueiro, Smith e Vieira (2011).

Como a BHPS é uma pesquisa longitudinal e sua população alvo é dinâmica, tenta-se entrevistar os mesmos indivíduos todos os anos, e caso haja a saída desses indivíduos de seus domicílios para a formação de novas famílias, a pesquisa busca entrevistar todos os membros adultos das novas famílias formadas, incluindo os respondentes originais.

O banco de dados da BHPS inclui pesos longitudinais para indivíduos que responderam à pesquisa em cada ocasião incluindo a última ocasião, que levam em conta diferentes probabilidades de seleção, não resposta na primeira ocasião da pesquisa (1991) e abandonos do painel. Os pesos longitudinais são pesos amostrais ajustados em cada ocasião para considerar na ocasião anterior a ausência de respostas por recusa do respondente, ou qualquer outra forma possível de abandono, na presente ocasião; ou seja, consideram eventuais perdas entre cada par de ocasiões adjacentes até a ocasião atual da pesquisa e o desenho amostral inicial.

O questionário básico da BHPS abrange questões de interesses sociais e políticos, incluindo a composição da família, sua mobilidade residencial e suas condições de habitação, escolaridade, saúde, valores socioeconômicos, comportamento no mercado de trabalho, entre outras. Há um questionário variável com questões que não precisam ser feitas anualmente (geralmente questões provocadas por mudanças políticas), e questões para obter dados retrospectivos dos membros das famílias (incluindo perguntas sobre casamento e fertilidade, riquezas e bens, medidas adicionais de saúde, qualidade de vida e outras, relativas ao período anterior à primeira ocasião da pesquisa). Em 1994, foi introduzido na pesquisa um questionário auto-administrado para crianças com idade entre 11 e 15 anos. Para os adultos, o questionário conta com um pesquisador para aplicá-lo, e o tempo para completá-lo é, em média, 45 minutos. Para famílias com apenas um membro, o questionário é mais curto.

O banco de dados utilizado em nosso estudo é uma subamostra da BHPS, contendo observações das ocasiões 12 (2002), 13 (2003), 14 (2004) e 15 (2005). Nessas ocasiões, os respondentes foram também questionados sobre seu nível de satisfação material e satisfação com o tempo de lazer. Para quantificar esses níveis de satisfação, os respondentes atribuíam valores às suas respostas, de acordo com a magnitude de sua satisfação ou insatisfação, baseados em uma escala ordinal do tipo Likert de sete pontos, em que 1 equivalia a “Nem um pouco satisfeito” e 7 a “Completamente satisfeito” (Salgueiro, Smith e Vieira, 2011).

Como aplicação motivadora para esta monografia, estudamos a satisfação dos indivíduos com relação ao tempo destinado por eles ao lazer. Nossa variável de interesse, o escore de lazer, foi criada a partir da soma dos escores atribuídos pelos respondentes às perguntas “Quão insatisfeito ou satisfeito você é com...” (i) “sua vida social?”, (ii) “a quantidade de tempo de lazer que você tem?”, (iii) “a forma como você gasta seu tempo de lazer?”, podendo, assim, variar de 3 a 21 (maiores informações sobre o cálculo do escore de lazer podem ser encontradas em Salgueiro, Smith e Vieira, 2011).

Utilizamos outras variáveis em nosso estudo, além do escore de lazer. A maneira como as usamos e os resultados obtidos serão mostrados nos capítulos seguintes.

O banco de dados utilizado neste estudo é uma subamostra da BHPS, contendo dados observados nas ocasiões 12 (2002), 13 (2003), 14 (2004) e 15 (2005). Utilizamos o banco de dados em duas formas: *wide* (curta ou larga) e *long* (longa). A primeira forma apresenta um número maior de variáveis, porque cada indivíduo aparece em apenas uma linha e cada variável medida é codificada de acordo com a ocasião da pesquisa, conforme o quadro abaixo.

Quadro 1. Exemplo de banco de dados na forma curta.

pid	*lscore_l*	*mscore_l*	*nscore_l*	*oscore_l*
1	12	15	19	15
2	19	18	21	21
...	...	...	...	...
n	14	13	15	16

Na forma longa, os indivíduos são repetidos tantas vezes quantas forem as ocasiões da pesquisa. Cada variável é representada em apenas uma coluna e há uma variável representando o tempo, conforme o exemplo que aparece no quadro abaixo.

Quadro 2. Exemplo de banco de dados na forma longa.

pid	*tempo*	*score_l*
1	1	19
1	2	18
1	3	21
1	4	21
2	1	10
2	2	13
2	3	13
2	4	14
...	...	...
n	1	18
n	2	16
n	3	17
n	4	18

As variáveis consideradas no estudo são apresentadas nas tabelas abaixo, com seus nomes, descrições e valores possíveis para o caso das variáveis qualitativas.

Tabela 1. Variáveis do banco de dados na forma curta.

Variável	**Label**	**Tipo**	**Codificação**
pid	Identificação do indivíduo	Contínua	
lage	Idade (em anos) na primeira ocasião	Contínua	
lmastat	Estado civil na primeira ocasião	Nominal	1 – Casado (a) 2 – União estável 3 – Viúvo (a) 4 – Divorciado (a) 5 – Separado (a) 6 – Solteiro (a)
lsex	Sexo na primeira ocasião	Nominal	1 – Masculino 2 – Feminino
lagecat	Idade na primeira ocasião	Ordinal	1 – 16-21 2 – 22-29 3 – 30-39 4 – 40-59 5 – 60 ou mais
lscore_l	Escore de lazer na primeira ocasião	Contínua	
mscore_l	Escore de lazer na segunda ocasião	Contínua	
nscore_l	Escore de lazer na terceira ocasião	Contínua	
oscore_l	Escore de lazer na quarta ocasião	Contínua	

Tabela 2. Variáveis do banco de dados na forma longa.

Variável	Label	Tipo	Codificação
time	Tempo	*	2 – 2002 3 – 2003 4 – 2004 5 – 2005
sex	Sexo	Dummy	0 – Masculino 1 – Feminino
score_l	Escore de lazer	Contínua	
agecat	Idade	Ordinal	1 – 16-21 2 – 22-29 3 – 30-39 4 – 40-49 5 – 50 ou mais
nchild	Número de crianças próprias no domicílio	Discreta	
logfihhmn	Logaritmo da renda domiciliar	Contínua	
jbhrs2	Número de horas trabalhadas na primeira ocasião	Dummy	0 – 15 ou menos, mais de 30 1 – 16-29
qualf	Nível educacional	Ordinal	1 – Curso superior ou mais 2 – Magistério 3 – Ensino médio 4 – Ensino fundamental 5 - Outros
mrjrgsc2	Classificação do último emprego	Nominal	1 – Ocupação profissional 2 – Ocupação gerencial ou técnica 3 – Ocupação qualificada não manual 4 – Ocupação qualificada manual 5 – Ocupação parcialmente qualificada 6 – Ocupação não qualificada 7 – Membro das Forças Armadas
hlstat2	Estado de saúde na primeira ocasião	Dummy	0 – Outros 1 – Bom estado de saúde
mastat	Estado civil	Dummy	0 – Outros 1 – Casado (a)
score_m	Escore de satisfação em relação aos aspectos materiais	Contínua	

* A variável tempo será tratada ao longo desta monografia de duas formas, como discreta ou contínua. Uma discussão sobre a adoção destas duas formas de definição será apresentada no Capítulo 4.

3 ANÁLISE EXPLORATÓRIA DE DADOS LONGITUDINAIS

Para a maioria das técnicas de análise exploratória de dados consideradas, foi necessária a utilização do banco de dados na forma curta, que apresenta uma variável representando a mesma questão para cada ocasião da pesquisa, enquanto que para o ajuste dos modelos marginais de regressão foi necessária a utilização do banco de dados na forma longa, que contém uma variável própria para o tempo.

3.1 APLICAÇÃO

Iniciamos o trabalho com a análise exploratória dos dados (AED) com o objetivo de realizar um estudo descritivo preliminar. Muitas das técnicas de AED podem ser usadas tanto para análises transversais quanto longitudinais e outras são específicas apenas para dados longitudinais. Alguns gráficos, por exemplo, são mais apropriados para dados longitudinais, pois permitem a visualização das mudanças ocorridas nas respostas de um mesmo indivíduo e entre os indivíduos ao longo do tempo.

Os gráficos de dispersão têm a mesma interpretação para ambos os tipos de dados, transversais ou longitudinais. No presente caso, foram utilizados para verificar a relação da variável resposta com o tempo. Da mesma forma, gráficos de barras e gráficos para intervalos de confiança foram considerados no contexto longitudinal.

Para a construção de gráficos e cálculo de estatísticas descritivas usamos o banco de dados na forma curta. A outra forma, longa, foi utilizada para calcularmos correlações e para ajustarmos modelos de regressão, que serão discutidos e analisados mais à frente.

Iniciamos a análise exploratória de dados verificando as estatísticas descritivas da variável contínua *lage*.

Tabela 3. Estatísticas descritivas para a idade em 2002 (*lage*).

Estatística	**Valor**
1º Quartil	33
2º Quartil	42
3º Quartil	50
Média	41,13
Desvio Padrão	11,41
Variância	130,31

A partir da Tabela 3, verificamos uma semelhança entre as medidas de tendência central (média e mediana – representada pelo 2º Quartil), uma variabilidade das respostas grande, sugerindo que há desde jovens a idosos participando da pesquisa, sendo a maioria dos respondentes de "meia idade".

Para continuar o estudo do perfil dos respondentes, construímos uma tabela de dupla entrada, cruzando as variáveis categóricas *lmastat* e *lsex*. Obtivemos uma tabela 6x2, apresentada abaixo, com os percentuais de linhas e colunas e as frequências observadas.

Tabela 4. Tabela de dupla entrada para as variáveis sexo e estado civil na primeira ocasião.

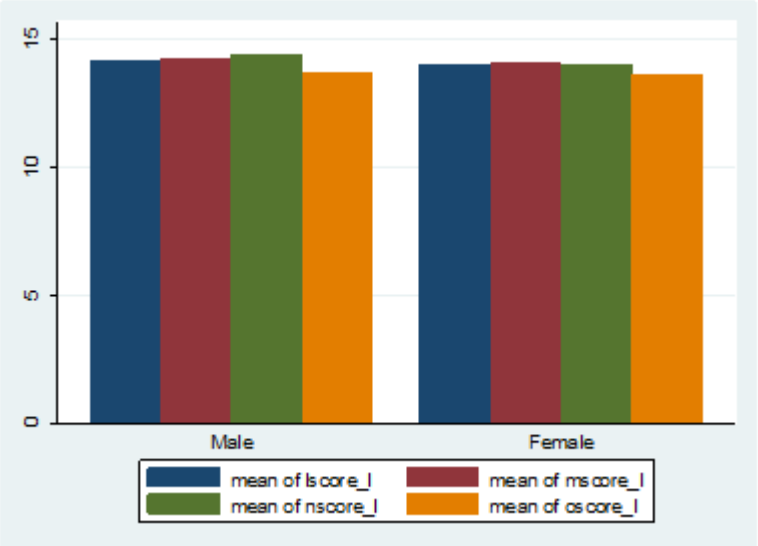
	Sexo		
Estado civil	**Masculino**	**Feminino**	**Total**
Casado	726 52,08% 63,13%	668 47,92% 60,45%	1394 100,00% 61,82%
União estável	144 51,61% 12,52%	135 48,39% 12,22%	279 100,00% 12,37%
Viúvo	6 27,27% 0,52%	16 72,73% 1,45%	22 100,00% 0,98%
Divorciado	46 32,86% 4,00%	94 67,14% 8,51%	140 100,00% 6,21%
Separado	15 39,47% 1,30%	23 60,53% 2,08%	38 100,00% 1,69%
Solteiro	213 55,76% 18,52%	169 44,24% 15,29%	382 100,00% 16,94%
Total	1.150 51,00% 100,00%	1.105 49,00% 100,00%	2.255 100,00% 100,00%

Analisando a Tabela 4, concluímos que a maioria dos respondentes em 2002 eram casados, e a minoria, viúvos. Percebemos, também, que a quantidade de homens e mulheres respondentes foi bem aproximada, considerando as porcentagens totais. Em algumas situações, homens e mulheres apresentaram porcentagens muito discrepantes, como no caso das respostas viúvo (a), divorciado (a) e separado (a), em que as mulheres apresentam a maioria.

No que diz respeito à nossa aplicação motivadora, nosso interesse foi verificar o nível de satisfação dos respondentes com relação ao tempo destinado por eles ao lazer. As respostas foram quantificadas em escores: quanto maior o escore de um indivíduo, mais satisfeito o (a) respondente é, conforme discutido na Seção 2.

A primeira análise desse nível de satisfação foi feita após a construção de um gráfico de barras das médias dos escores de lazer de cada ocasião, agrupadas por sexo, conformo a Figura 1 abaixo. Pudemos ver que as médias de escore de cada ocasião são aproximadamente iguais para os dois sexos, parecendo não haver tendência a aumento ou redução ao longo do tempo.

Figura 1. Gráfico de barras da média dos escores de lazer por sexo.



Os diagramas de dispersão dos escores de lazer para as diferentes ocasiões sugerem uma correlação positiva entre estas medidas realizadas ao longo do tempo para os mesmos respondentes. Para validar essa afirmação, calculamos a matriz de correlação mostrada a seguir na Figura 2.

Figura 2. Diagramas de dispersão para os escores de lazer.

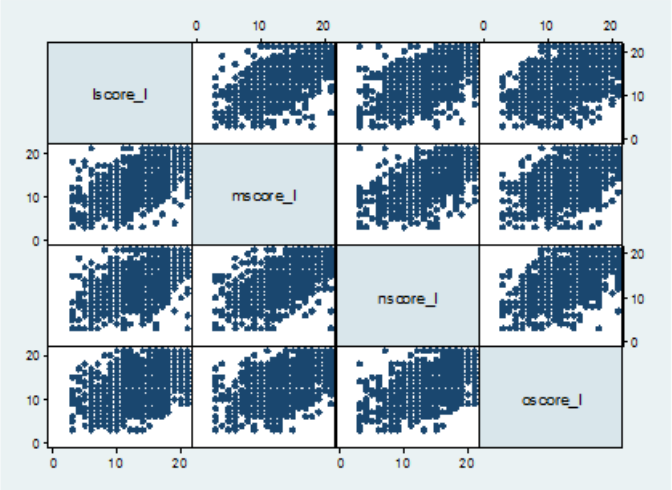


Figura 3. Matriz de correlação para os escores de lazer.

	2002	2003	2004	2005
2002	1,00			
2003	0,65	1,00		
2004	0,61	0,66	1,00	
2005	0,56	0,60	0,64	1,00

De acordo com os resultados incluídos na Figura 3, acima, podemos considerar que a correlação entre as observações para os mesmos indivíduos ao longo do tempo parecem estar apresentando uma leve tendência de decréscimo na medida em que a distância entre as observações aumenta. Entretanto, consideramos que está tendência observada pode ser devida apenas a variação amostral e não ser de fato significativa.

As técnicas até aqui utilizadas foram aplicadas ao banco de dados na forma curta. As que serão apresentadas a partir de agora foram realizadas com o banco de dados em sua forma longa.

Como já foi dito anteriormente, em um estudo longitudinal consideramos a presença de efeitos das correlações intra e entre indivíduos sobre as respostas observadas. Assim, foi de nosso interesse verificar as possíveis mudanças ocorridas nos escores de lazer ao longo do tempo através de estatísticas descritivas.

De acordo com os resultados obtidos, a média longitudinal do escore de lazer, calculada a partir das 4 observações para cada uma das 2255 pessoas, era aproximadamente 14,02. Levando em conta o total de observações, os escores de lazer variavam entre 3 e 21. Cada indivíduo variava seus escores de lazer ao longo do tempo, em média, entre 3,52 e 22,02. Notamos que a variabilidade das respostas para essa questão aumentava quando

aumentava a dimensão de comparação: a variabilidade intra-indivíduo era menor que a entre os indivíduos, que era menor que a total. Isso era esperado, considerando-se que as mesmas pessoas tomadas ao longo do tempo tendem a ser parecidas com o que eram na ocasião anterior, e diferentes das demais quando comparadas no conjunto. Estes resultados mencionados foram retirados da Tabela 5 abaixo.

Seja

$$\tilde{y}_{it} = y_{it} - y_i + y$$

em que y_{it} são as observações da variável de interesse para cada indivíduo *i* no tempo *t*, y_i é a média de y_{it} das observações do grupo *i* e y é a média global da variável de interesse.

O desvio padrão intra-indivíduos é o desvio padrão estimado para $\tilde{y}_{it}$, e o desvio padrão entre os indivíduos é o desvio padrão estimado para as *n* médias y_i.

Tabela 5. Estatísticas descritivas para os escores de lazer.

Variável		Média	Desvio padrão	Mínimo	Máximo	Observações
Escore de lazer	**Total**	14,02	3,59	3,00	21,00	N=9020
	Entre		3,03	3,00	21,00	n=2255
	Intra		1,93	3,52	22,02	T=4

Ainda sob a perspectiva longitudinal, montamos a Tabela 6, que apresenta frequências para dados longitudinais para a variável sexo. Os valores globais são resultantes em termos de pessoa-ano. 4600 pessoas-ano presentes no banco de dados responderam “masculino” para a questão referente ao gênero, e 4420 pessoas-ano responderam “femininos”. Os intervalos de dados são referentes a pessoas, e não pessoas-ano. Existiam 1150 homens e 1105 mulheres no banco de dados. O percentual presente na última coluna é a medida da estabilidade global da variável.

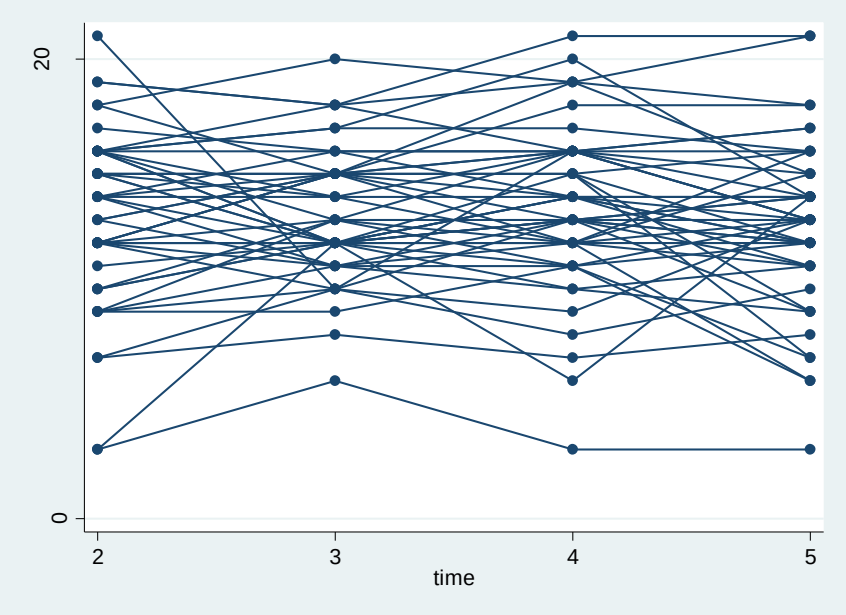
Tabela 6. Tabela de frequências para a variável sexo.

Sexo	Total (Pessoas-ano)		Entre (Pessoas)		Intra
	Frequência	Percentual	Frequência	Percentual	Percentual
Masculino	4600	51,00	1150	51,00	100,00
Feminino	4420	49,00	1105	49,00	100,00
Total	9020	100,00	2255	100,00	100,00

Após analisarmos os perfis de resposta de maneira geral, seguimos o trabalho tentando traçar os perfis individuais de resposta, através de um gráfico específico para isso. Como

havia muitas observações no banco de dados, o gráfico ficaria carregado, dificultando a visualização dos perfis de resposta, assim, selecionamos aleatoriamente 50 indivíduos para serem representados no gráfico de linhas mostrado a seguir.

Figura 4. Gráfico de linhas para os escores de lazer.



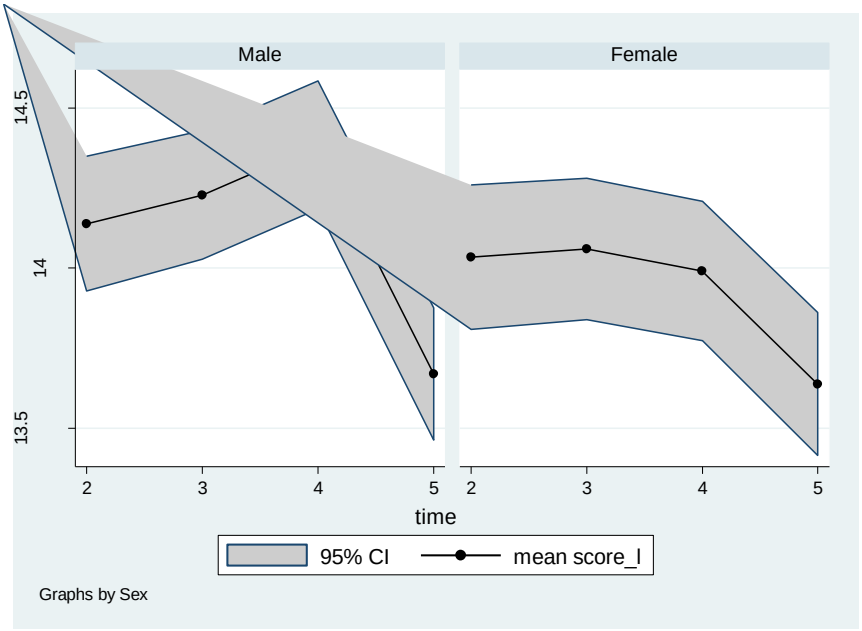
Percebemos que não havia um padrão bem definido para as respostas. O gráfico apenas sugere uma menor variabilidade nos escores de lazer na segunda ocasião.

Para finalizar nossa análise exploratória, construímos gráficos de intervalos de confiança de 95% (Figuras 5 e 6) para os escores médios de lazer, primeiramente agrupados por sexo, e depois por categoria de idade, a partir de resultados apresentados nas Tabela 7 e 8, respectivamente.

Tabela 7. Intervalos de confiança de 95% para os escores médios de lazer por sexo.

Ocasião	Sexo	Média	Desvio padrão	Observações	Lim. Inferior	Lim. Superior
1	Masculino	14,14	3,58	1150	13,93	14,35
1	Feminino	14,03	3,76	1105	13,81	14,26
2	Masculino	14,23	3,41	1150	14,03	14,43
2	Feminino	14,06	3,68	1105	13,84	14,28
3	Masculino	14,38	3,39	1150	14,18	14,58
3	Feminino	13,99	3,62	1105	13,77	14,21
4	Masculino	13,67	3,50	1150	13,46	13,88
4	Feminino	13,64	3,71	1105	13,41	13,86

Figura 5. Gráficos de intervalos de confiança para os escores de lazer por sexo.



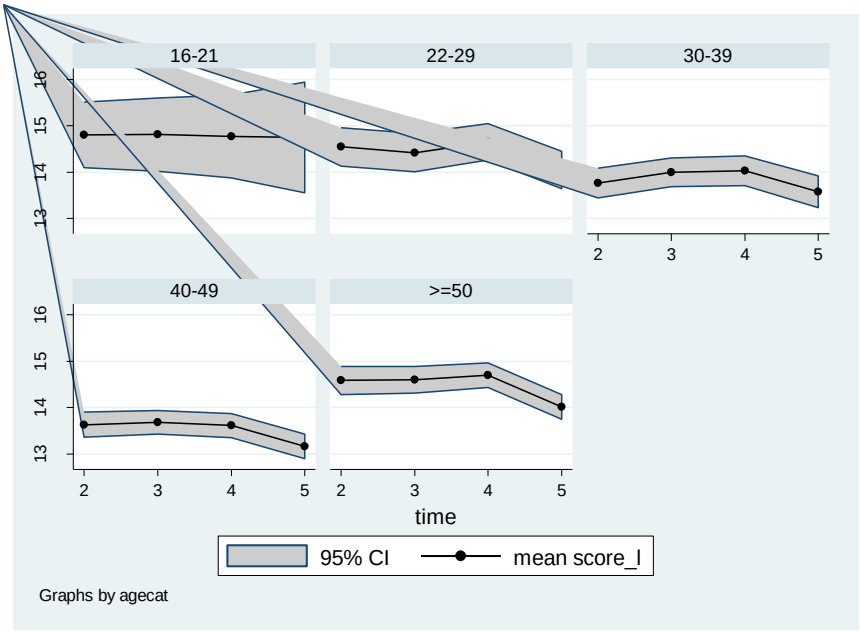
Analisando a Tabela 7, percebemos que as médias de escores de lazer para homens e mulheres durante as primeiras ocasiões eram aproximadas. A semelhança foi diminuída na terceira ocasião, em que a média para homens aumentou, enquanto a média para mulheres sofreu pouca mudança. Na quarta ocasião, os valores voltaram a ser parecidos, mas com uma maior diminuição na média para os homens. Os desvios-padrão tiveram pouca variação, tanto se considerados ao longo do tempo quanto por sexo em cada ocasião.

Na figura 5, notamos uma variação maior na média dos escores para os homens, com mudanças abruptas ao longo do tempo. Nas três primeiras ocasiões os homens apresentaram médias de escore maiores do que as mulheres, situação que foi invertida na última ocasião. Isso pode ser interpretado como o fato de que, durante os anos da pesquisa, os homens estavam mais satisfeitos com o tempo dedicado por eles ao lazer do que as mulheres.

Tabela 8. Intervalos de confiança de 95% para os escores de lazer por idade (em categorias).

Ocasião	Categoria	Média	Desvio padrão	Observações	Lim. Inferior	Lim. Superior
1	16-21	14,81	3,63	106	14,10	15,51
1	22-29	14,54	3,64	313	14,13	14,95
1	30-39	13,76	3,69	535	13,44	14,08
1	40-49	13,63	3,60	724	13,68	13,90
1	>=50	14,58	3,70	577	14,28	14,89
2	16-21	14,81	3,52	79	14,02	15,60
2	22-29	14,42	3,54	296	14,01	14,83
2	30-39	13,99	3,47	510	13,69	14,30
2	40-49	13,68	3,45	735	13,42	13,93
2	>=50	14,60	3,64	635	14,31	14,89
3	16-21	14,77	3,50	61	13,87	15,67
3	22-29	14,65	3,26	279	14,26	15,04
3	30-39	14,03	3,48	468	13,71	14,35
3	40-49	13,61	3,57	754	13,35	13,87
3	>=50	14,70	3,48	693	14,43	14,96
4	16-21	14,74	3,53	35	13,55	15,93
4	22-29	14,04	3,28	266	13,64	14,45
4	30-39	13,58	3,59	436	13,23	13,92
4	40-49	13,16	3,65	770	12,90	13,43
4	>=50	14,02	3,61	748	13,75	14,28

Figura 6. Gráficos de intervalos de confiança para os escores de lazer por idade (em categorias).



A Tabela 8 e a Figura 6 sugerem que os valores encontrados para a média e o desvio padrão dos escores de lazer agrupados por idade se parecem com os obtidos para os dados agrupados por sexo. Observamos que os respondentes mais satisfeitos com o tempo dedicado ao lazer são os jovens (entre 16 e 21 anos e entre 22 e 29 anos) e os idosos (maiores de 50 anos), pois apresentaram médias de escore de lazer mais altas. A tendência para as faixas etárias, com exceção dos jovens entre 16 e 21 anos, foi parecida, principalmente entre a terceira e quarta ocasião, em que houve um decréscimo mais acentuado das médias. A faixa etária que apresentou a maior variabilidade de repostas foi a dos jovens, mantendo a média dos escores de lazer aproximadamente constante. As pessoas entre 30 e 49 anos obtiveram as menores médias dos escores de lazer, o que pode ter sido causado pelo fato de que a maioria das pessoas nessa faixa trabalha fora de casa, dedicando menos tempo ao lazer.

4 MODELOS MARGINAIS

Os modelos marginais constituem uma importante ferramenta para a análise de dados longitudinais: permitem a análise simultânea das relações entre a variável resposta e as variáveis preditoras em diferentes pontos do tempo. Além disso, estes modelos possibilitam o estudo da variabilidade da variável resposta entre e intra-indivíduos ao longo do tempo. Para a produção do conteúdo desta seção foram considerados Twisk (2003) e Vieira (2009).

Como as respostas de um mesmo indivíduo normalmente não são independentes, uma estrutura de correlação para as medidas da variável resposta deve ser assumida inicialmente para corrigir a falta de independência. As possíveis estruturas de correlação que podem ser consideradas serão apresentadas na Subseção 4.1.

Um modelo marginal é da forma

$$Y_{it} = \underline{x}_{it}\underline{\beta} + \varepsilon_{it} \qquad (1)$$

com *i* variando de 1 a *N* e *t* variando de 1 a *T*, onde Y_{it} são as respostas do indivíduo *i* no tempo *t*, $\underline{x}_{it}$ é um vetor 1xq com as q variáveis preditoras fixas, $\underline{\beta}$ é o vetor qx1 de coeficientes para as variáveis preditoras (que representam o efeito de cada uma das covariáveis na variável resposta) e ε_{it} é o resíduo para o indivíduo *i* no tempo *t*, com E(ε_{it})=0.

Outra abordagem de modelagem de dados longitudinais tem com base os modelos de efeitos aleatórios, apresentados a seguir,

$$Y_{it} = \underline{x}_{it}\underline{\beta} + u_i + v_{it} \qquad (2)$$

onde, u_i são os efeitos aleatórios permanentes (ou fatores específicos individuais não observados) e v_{it} são os efeitos aleatórios transitórios. Nesta monografia não consideramos os modelos de efeitos aleatórios.

Diferentes estruturas de correlação resultam em estimativas de coeficientes diferentes, como mostraremos na Subseção 4.1.1. Os métodos adotados para a estimação dos coeficientes dos modelos marginais são robustos quanto à escolha de uma estrutura de correlação inadequada (ou seja, diferente da verdadeira). Entretanto, recomenda-se mesmo assim bastante cautela nesta escolha, uma vez que a adoção de diferentes estruturas de correlação pode levar a diferentes conclusões acerca das relações entre as variáveis.

Apresentaremos as pressuposições desses modelos, os possíveis métodos de estimação de parâmetros, interpretações dos coeficientes e demais resultados nos tópicos seguintes.

4.1 METODOLOGIA

4.1.1 Estruturas de Correlação

Consideramos abaixo os cinco principais tipos de estruturas de correlação que podem ser adotadas para o ajuste dos modelos marginais. A escolha do tipo de correlação pode ser baseada na estrutura de correlação dos dados observados para a variável resposta – a estrutura que mais se assemelhar à mencionada deve ser utilizada.

Na estrutura *independente* de correlação as medidas subsequentes têm correlação igual a zero. Ela se apresenta da seguinte maneira:

Figura 7. Exemplo de estrutura de correlação independente.

	t_1	t_2	t_3	t_4	t_5
t_1	1	0	0	0	0
t_2	0	1	0	0	0
t_3	0	0	1	0	0
t_4	0	0	0	1	0
t_5	0	0	0	0	1

A estrutura *permutável* mostrada abaixo assume que a correlação entre as medidas consecutivas é a mesma independentemente da amplitude do intervalo de tempo.

Figura 8. Exemplo de estrutura de correlação permutável.

	t_1	t_2	t_3	t_4	t_5
t_1	1	ρ	ρ	ρ	ρ
t_2	ρ	1	ρ	ρ	ρ
t_3	ρ	ρ	1	ρ	ρ
t_4	ρ	ρ	ρ	1	ρ
t_5	ρ	ρ	ρ	ρ	1

Uma estrutura *m-dependente* considera que as correlações entre as observações *t* unidades de tempo a frente são iguais, as correlações entre as observações *t+1* unidades de tempo a frente são iguais, e assim sucessivamente, com *t* variando de 1 a m, m variando de acordo com o interesse do estudo. Utilizamos como exemplo uma estrutura de correlação *3-dependente*.

Figura 9. Exemplo de estrutura de correlação 3-dependente.

	t_1	t_2	t_3	t_4	t_5
t_1	1	ρ_1	ρ_2	ρ_3	0
t_2	ρ_1	1	ρ_1	ρ_2	ρ_3
t_3	ρ_2	ρ_1	1	ρ_1	ρ_2
t_4	ρ_3	ρ_2	ρ_1	1	ρ_1
t_5	0	ρ_3	ρ_2	ρ_1	1

Quando a correlação entre as observações consecutivas é igual a um valor ρ, a correlação entre as observações espaçadas em dois intervalos de tempo é igual a ρ^2, a correlação entre as observações espaçadas em *t* intervalos de tempo é igual a ρ^t, a estrutura de correlação é chamada *autorregressiva*, com a seguinte forma:

Figura 10. Exemplo de estrutura de correlação autorregressiva.

	t_1	t_2	t_3	t_4	t_5
t_1	1	ρ^1	ρ^2	ρ^3	ρ^4
t_2	ρ^1	1	ρ^1	ρ^2	ρ^3
t_3	ρ^2	ρ^1	1	ρ^1	ρ^2
t_4	ρ^3	ρ^2	ρ^1	1	ρ^1
t_5	ρ^4	ρ^3	ρ^2	ρ^1	1

Por fim, uma estrutura de correlação *desestruturada* assume que todas as correlações são diferentes.

Figura 11. Exemplo de estrutura de correlação desestruturada.

	t_1	t_2	t_3	t_4	t_5
t_1	1	ρ_1	ρ_2	ρ_3	ρ_4
t_2	ρ_1	1	ρ_5	ρ_6	ρ_7
t_3	ρ_2	ρ_5	1	ρ_8	ρ_9
t_4	ρ_3	ρ_6	ρ_8	1	ρ_{10}
t_5	ρ_4	ρ_7	ρ_9	ρ_{10}	1

4.1.2 Métodos de Estimação

No contexto dos modelos de regressão para dados transversais, os parâmetros dos modelos são estimados através da comparação de indivíduos com valores específicos de *x*, com os demais casos com outros valores. Além disso, uma comparação das respostas para um indivíduo ao longo do tempo é conduzida adicionalmente, permitindo que *x* possa variar com o tempo. Quando se ajusta um modelo de regressão longitudinal, cada indivíduo pode ser visto como seu próprio controle. Além disso, há muita variabilidade entre indivíduos causada por atributos não observados, que são ocultados quando se ajusta modelos transversais. Em outras palavras, os coeficientes dos modelos marginais têm dupla interpretação: uma para a relação entre os indivíduos e outra para a relação intra-indivíduos.

Os métodos de estimação de parâmetros considerados para o ajuste dos modelos nesta monografia são os de Máxima Verossimilhança e de Máxima Pseudo-Verossimilhança, que serão apresentados ao longo desta subseção.

Para maiores informações sobre as expressões da variância e da covariância, consultar Vieira (2009).

A correlação entre as respostas para o mesmo indivíduo ao longo do tempo é denotada por:

i. $\text{CORR}(Y_{it}, Y_{it'}) = \rho$

ou seja, a correlação populacional *intra-indivíduo*, podendo variar de 0 a 1.

Para seguirmos com a estimação dos parâmetros dos modelos marginais, assumimos nesta monografia que as observações são igualmente espaçadas no tempo, e que o número de indivíduos entrevistados (*N*) é 'grande' em relação ao número de observações por indivíduo (*T*). Também consideramos que a amostra foi selecionada em uma ocasião e as mesmas unidades amostrais retornaram de cada uma das $T-1$ ocasiões subsequentes da pesquisa, e que não há dados faltantes (todos os Y_i têm o mesmo tamanho *T*).

Suponha que $\underline{Y_i} = (Y_{i1}, \dots, Y_{iT})'$ tem como distribuição uma normal multivariada *T*-dimensional, denotada por

$$\underline{Y_i} \sim N_T \left[\underline{\mu_i}\left(\underline{\beta}\right), \Sigma\right] \qquad (3)$$

onde Σ é uma matriz de covariâncias positivamente definida.

Sejam $\underline{y_i}$ vetores $T\times 1$ com os valores observados para a variável resposta para cada indivíduo *i* em cada ocasião. Assumimos que toda a população finita é observada, e $Y_1, \dots, Y_N$ são independentes. Definimos a função de densidade conjunta (ou função de probabilidade de massa) do censo para todas as observações como o produto das densidades marginais normais. Assim,

$$f\left(\underline{y}_1, \dots, \underline{y}_N; \underline{\beta}\right) = (2\pi)^{-N_0/2}|\Sigma|^{-N/2} e^{-\sum_{i=1}^{N}\left[\underline{y_i}-\underline{\mu_i}\left(\underline{\beta}\right)\right]\Sigma^{-1}\left[\underline{y_i}-\underline{\mu_i}\left(\underline{\beta}\right)\right]/2} \qquad (4)$$

com Σ representando a matriz de covariâncias (assumida como conhecida) e $\sum$ sendo o símbolo padrão para o somatório. A equação apresentada em (4) é chamada de função de verossimilhança do censo, denotada por $l_N\left[\underline{\beta}\right]$.

O vetor de parâmetros $\underline{\beta}$ é estimado maximizando o logaritmo da verossimilhança do censo com respeito a $\underline{\beta}$, que é

$$L_N\left[\underline{\beta}\right] = -\frac{1}{2}\left[N_0 \log 2\pi + N \log|\Sigma| + \sum_{i=1}^{N}\left[\underline{y_i} - \underline{\mu_i}\left(\underline{\beta}\right)\right]\Sigma^{-1}\left[\underline{y_i} - \underline{\mu_i}\left(\underline{\beta}\right)\right]\right]. \qquad (5)$$

O estimador de máxima verossimilhança do censo $\underline{\hat{\beta}}_N$ de $\underline{\beta}$ é obtido, então, minimizando o expoente de (4), ou seja,

$$\sum_{i=1}^{N}\left[\underline{y}_i-\underline{\mu}_i\left(\underline{\beta}\right)\right]'\Sigma^{-1}\left[\underline{y}_i-\underline{\mu}_i\left(\underline{\beta}\right)\right] \qquad (6)$$

com respeito a $\underline{\beta}$. A expressão (6) é a soma dos quadrados da distância multivariada generalizada de $\underline{Y}_i$ a $\underline{\mu}_i\left(\underline{\beta}\right)$. Supondo que foi observada apenas $\underline{y}_i$ para a unidade *i* na amostra *s*, {1, ..., *n*}. A expressão (6) pode ser estimada como

$$\frac{N}{n}\sum_{i=1}^{n}\left[\underline{y}_i-\underline{\mu}_i\left(\underline{\beta}\right)\right]'\Sigma^{-1}\left[\underline{y}_i-\underline{\mu}_i\left(\underline{\beta}\right)\right]. \qquad (7)$$

O estimador de máxima verossimilhança de $\underline{\beta}$ é obtido minimizando a expressão (7). Alternativamente podemos resolver o seguinte sistema de equações:

$$\sum_{i=1}^{n}\left[\underline{y}_i-\underline{\mu}_i\left(\underline{\beta}\right)\right]'\Sigma^{-1}\frac{\partial\underline{\mu}_i\left(\underline{\beta}\right)}{\partial\underline{\beta}}=0 \qquad (8)$$

conhecido como equações pseudo-escore para $\underline{\beta}$. Seja $X_i=\left(\underline{x}_{i1}^{'},\ldots,\underline{x}_{iT}^{'}\right)'$ a matriz *T*x*q* com as covariáveis para o indivíduo *i*. Quando $\underline{\mu}_i\left(\underline{\beta}\right)=X_i\underline{\beta}$, temos:

$$\frac{\partial\underline{\mu}_i\left(\underline{\beta}\right)}{\partial\underline{\beta}}=X_i. \qquad (9)$$

As equações pseudo-escore têm uma solução definida para $\underline{\hat{\beta}}(\Sigma)$ dada por

$$\underline{\hat{\beta}}(\Sigma)=\left(\sum_{i=1}^{n}X_i^{'}\Sigma^{-1}X_i\right)^{-1}\sum_{i=1}^{n}X_i^{'}\Sigma^{-1}\underline{y}_i. \qquad (10)$$

Em geral Σ é desconhecida e sua estimação é discutida em detalhes por Vieira e Skinner (2008) e Vieira (2009). Nesta monografia, consideramos uma abordagem que tem como base a substituição de Σ por uma matriz de covariância de trabalho *T*x*T*, como aquelas apresentadas nos exemplos apresentados na sub-seção anterior, denotada por V. Assim, o estimador passa a ser $\underline{\hat{\beta}}(\mathrm{V})$, com solução definida por

$$\underline{\hat{\beta}}(\mathrm{V})=\left(\sum_{i=1}^{n}X_i^{'}\mathrm{V}^{-1}X_i\right)^{-1}\sum_{i=1}^{n}X_i^{'}\mathrm{V}^{-1}\underline{y}_i. \qquad (11)$$

Como discutido anteriormente, diferentes estruturas de V resultam em diferentes estimativas de $\underline{\beta}$. Seja R uma matriz de correlação *T*x*T* correspondente a V, temos:

ii. $\mathrm{R} = \mathrm{D_V^{-1}} \mathrm{V} \mathrm{D_V^{-1}}$

iii. $\mathrm{V} = \mathrm{D_V} \mathrm{R} \mathrm{D_V}$

iv. $\mathrm{D_V} = [\mathrm{diag(V)}]^{1/2}$

A matriz diag(V) pode ser obtida substituindo os elementos fora da diagonal de V por zero. Na prática, a escolha de R pode ser feita depois de calculadas as estimativas empíricas das correlações e pode assumir diferentes padrões, como os mostrados na Subseção 4.1.1. Esta será a estratégia adotada na Subseção 4.2.

Um caso especial de $\underline{\hat{\beta}}(\mathrm{V})$ é quando V é uma matriz identidade *T*x*T*, ou seja, as respostas repetidas de um determinado indivíduo são independentes. Neste caso,

$$\underline{\hat{\beta}}(I) = \left(\sum_{i=1}^{n} X_i^{'} X_i\right)^{-1} \sum_{i=1}^{n} X_i^{'} \underline{y}_i, \quad (12)$$

que é equivalente ao estimador de Mínimos Quadrados Ordinários (MQO).

Para qualquer escolha da matriz constante V, $\underline{\hat{\beta}}(\mathrm{V})$ é estimador consistente de $\underline{\beta}$, generalização que depende apenas da especificação correta da média (Vieira, 2009), ou seja, de $\mu(\underline{x}_{it}, \underline{\beta})$.

Para desenhos amostrais complexos, envolvendo estratificação, conglomeração e seleção em múltiplos estágios, que violam a suposição de que os dados são selecionados por amostragem aleatória simples, a estimação de $\underline{\beta}$ é diferente.

Assumimos que $\underline{Y_i}$ é definida da mesma maneira que a já apresentada, e distribuída conforme (3). A função de densidade conjunta do censo e a função de verossimilhança do censo são definidas como (4), $\underline{y_i}$ e $\underline{X_i}$ vetor e matriz definidos como anteriormente.

Consideramos que o estimador de máxima verossimilhança do censo $\underline{\hat{\beta}_N}$ para o parâmetro $\underline{\beta}$ pode ser obtido pela minimização do expoente da função de verossimilhança do censo com respeito a $\underline{\beta}$.

Vamos considerar as observações $\underline{y_i}$ para as i unidades na amostra s. N é assumido desconhecido para desenho amostral complexo, assim, a expressão (6) é estimada por

$$\sum_{i=1}^{N} w_i \left[\underline{y_i} - \underline{\mu_i}\left(\underline{\beta}\right)\right]' \Sigma^{-1} \left[\underline{y_i} - \underline{\mu_i}\left(\underline{\beta}\right)\right] \qquad (13)$$

onde w_i são os pesos amostrais. O estimador de máxima pseudo verossimilhança (PMV) de $\underline{\beta}$ é obtido minimizando-se (13) ou resolvendo-se o seguinte sistema de equações:

$$\sum_{i=1}^{n} w_i \left[\underline{y_i} - \underline{\mu_i}\left(\underline{\beta}\right)\right]' \Sigma^{-1} \frac{\partial \underline{\mu_i}\left(\underline{\beta}\right)}{\partial \underline{\beta}} = 0. \qquad (14)$$

Da mesma forma, quando $\underline{\mu_i}\left(\underline{\beta}\right) = X_i\underline{\beta}$ temos (9).

As equações pseudo escore têm então uma solução determinada para $\underline{\hat{\beta}}(\Sigma)_{PMV}$ dada por:

$$\underline{\hat{\beta}}(\Sigma)_{PMV} = \left(\sum_{i=1}^{n} w_i X_i' \Sigma^{-1} X_i\right)^{-1} \sum_{i=1}^{n} w_i X_i' \Sigma^{-1} \underline{y_i}. \qquad (15)$$

Quando Σ é substituída por V, análogo ao que já foi apresentado,

$$\underline{\hat{\beta}}(\mathrm{V})_{PMV} = \left(\sum_{i=1}^{n} w_i X_i' \mathrm{V}^{-1} X_i\right)^{-1} \sum_{i=1}^{n} w_i X_i' \mathrm{V}^{-1} \underline{y_i}. \qquad (16)$$

O estimador $\underline{\hat{\beta}}(\mathrm{V})$ definido em (11) é um caso especial de $\underline{\hat{\beta}}(\Sigma)_{PMV}$ quando os pesos amostrais são constantes. Assim, (16) é equivalente a

$$\underline{\hat{\beta}}(\mathrm{V})_{PMV} = \left(\sum_{i=1}^{n} w_i X_i^{*'} \mathrm{V}^{-1} X_i^{*}\right)^{-1} \sum_{i=1}^{n} w_i X_i^{*'} \mathrm{V}^{-1} \underline{y_i}^{*} \qquad (17)$$

onde $X_i^* = \left(\underline{x_{i1}^*}', \ldots, \underline{x_{iT}^*}'\right)'$ com $\underline{x_{it}^*} = \sqrt{w_i}\underline{x_{it}}$, e $\underline{y_i^*} = (y_{i1}^*, \ldots, y_{iT}^*)'$ com $y_{it}^* = \sqrt{w_i} y_{it}$.

Quando V é uma matriz identidade TxT , a expressão (16) é equivalente ao estimador de MQO para $\underline{\beta}$ ponderado pelos pesos amostrais w_i:

$$\underline{\hat{\beta}}(\mathrm{I})_{PMV} = \left(\sum_{i=1}^{n} w_i X_i' X_i\right)^{-1} \sum_{i=1}^{n} w_i X_i' \underline{y_i}. \qquad (18)$$

O estimador $\underline{\hat{\beta}}_{PMV}$, supondo que V é uma matriz constante, é aproximadamente não-viesado para qualquer escolha de V. Para maiores detalhes consultar Vieira (2009).

O estimador da variância semi-robusto para $\underline{\hat{\beta}}_{PMV}$ é dado por:

$$var_r\left[\underline{\hat{\beta}}_{PMV}(\mathrm{V})\right] = \left[\sum_{i=1}^{n} w_i X_i^{'} V^{-1} X_i\right]^{-1}\left[\sum_{i=1}^{n} w_i^2 \left(X_i' V^{-1} \underline{\hat{\varepsilon}}_i\right)\left(X_i' V^{-1} \underline{\hat{\varepsilon}}_i\right)'\right]\left[\sum_{i=1}^{n} w_i X_i^{'} V^{-1} X_i\right]^{-1} \qquad (19)$$

onde $\underline{\hat{\varepsilon}}_i = \underline{y}_i - X_i \underline{\hat{\beta}}_{PMV}(V)$. Para informações sobre como obter esta expressão, ver Vieira (2009). Este estimador não considera o plano amostral como um todo, mas considera as probabilidades de seleção através da consideração dos pesos amostrais. O caso particular desse estimador é considerado quando os pesos amostrais são definidos como uma constante.

4.2 APLICAÇÃO

O nosso principal interesse na aplicação foi estudar as mudanças ocorridas nos escores de lazer dos indivíduos participantes da pesquisa ao longo do tempo. Para isso, ajustamos diversos modelos de regressão, entre eles os modelos marginais. Os resultados dos ajustes são apresentados nesta subseção, com as devidas explicações e interpretações.

Embora o nosso banco de dados seja proveniente de uma pesquisa longitudinal, inicialmente ajustamos um modelo para dados transversais, escolhendo a variável *score_l* como variável dependente e as variáveis *sex*, *agecat*, *nchild*, *logfihhmn*, *jbhrs2*, *qualf*, *mrjrgsc2*, *hlstat2*, *mastat* e *score_m* como variáveis independentes. Esse modelo foi ajustado com o objetivo de dar suporte à determinação da estrutura de correlação (para os resíduos) mais adequada para a utilização no ajuste do modelo de regressão linear múltipla para dados longitudinais (modelo marginal).

Neste modelo, as interações entre a variável de tempo e as demais variáveis tiveram que ser especificadas, pois como o modelo é adequado apenas para dados transversais, ele não considera as mudanças que podem ocorrer nas variáveis ao longo do tempo. O método de estimação de parâmetros utilizado é o de Mínimos Quadrados Ordinários (MQO; Charnet et al, 1999), que procura minimizar a soma dos resíduos do modelo elevados ao quadrado.

Na Tabela 9 são apresentadas as estimativas dos coeficientes do modelo ajustado, seus erros padrão, as estatísticas t e os valores p relativos ao teste T de Student, e os intervalos de confiança de 95%.

Tabela 9. Coeficientes do modelo de regressão para dados transversais.

Variável	Coeficiente	Erro padrão	t	Valor p	Lim. inferior	Lim. superior
sex	-0,15	0,15	-1,05	0,293	-0,44	0,13
time=3 x *sex*	-0,08	0,21	-0,38	0,705	-0,48	0,33
time=4 x *sex*	-0,35	0,21	-1,72	0,086	-0,76	0,05
time=5 x *sex*	0,03	0,21	-0,17	0,867	-0,44	0,37
agecat	-0,27	0,07	-3,89	0,000	-0,41	-0,13
time=3 x *agecat*	0,08	0,10	0,81	0,416	-0,12	0,28
time=4 x *agecat*	0,09	0,10	0,85	0,395	-0,11	0,28
time=5 x *agecat*	0,05	0,10	-0,53	0,596	-0,25	0,15
nchild	-0,54	0,08	-6,97	0,000	-0,69	-0,39
time=3 x *nchild*	0,02	0,11	0,20	0,840	-0,19	0,24
time=4 x *nchild*	0,04	0,11	0,39	0,699	-0,18	0,26
time=5 x *nchild*	0,05	0,11	-0,42	0,674	-0,27	0,18
logfihhmn	-1,28	0,34	-3,74	0,000	-1,95	-0,61
time=3 x *logfihhmn*	0,51	0,47	1,08	0,281	-0,42	1,43
time=4 x *logfihhmn*	0,33	0,43	0,76	0,450	-0,52	1,17
time=5 x *logfihhmn*	-0,25	0,48	-0,53	0,600	-1,18	0,68
jbhrs2	-0,01	0,20	-0,03	0,979	-0,40	0,39
time=3 x *jbhrs2*	0,13	0,29	0,47	0,641	-0,43	0,69
time=4 x *jbhrs2*	0,26	0,29	0,91	0,365	-0,30	0,82
time=5 x *jbhrs2*	0,06	0,29	0,21	0,833	-0,50	0,62
qualf	0,16	0,06	2,89	0,004	0,05	0,27
time=3 x *qualf*	-0,09	0,08	-1,13	0,258	-0,24	0,07
time=4 x *qualf*	-0,08	0,08	-0,95	0,343	-0,23	0,08
time=5 x *qualf*	0,04	0,08	0,47	0,641	-0,12	0,19
mrjrgsc2	0,06	0,05	1,18	0,239	-0,04	0,17
time=3 x *mrjrgsc2*	0,05	0,08	0,68	0,495	-0,10	0,20
time=4 x *mrjrgsc2*	0,04	0,08	0,54	0,586	-0,11	0,19
time=5 x *mrjrgsc2*	0,00	0,08	-0,01	0,990	-0,15	0,15
time=3	-1,47	1,80	-0,82	0,411	-4,98	2,04
time=5	2,14	1,81	1,18	0,237	-1,41	5,69
hlstat2	0,16	0,14	1,20	0,232	-0,10	0,43
time=3 x *hlstat2*	-1,95	0,20	-1,02	0,307	-0,57	0,18
time=4 x *hlstat2*	0,01	0,20	0,07	0,945	-0,36	0,39
time=5 x *hlstat2*	0,00	0,20	0,00	0,999	-0,38	0,37
mastat	0,17	0,18	0,93	0,354	-0,18	0,51
time=3 x *mastat*	-0,05	0,25	-0,20	0,843	-0,54	0,44
time=4 x *mastat*	-0,09	0,25	-0,36	0,721	-0,57	0,40
time=5 x *mastat*	0,04	0,25	0,15	0,880	-0,45	0,52
time=4	-0,74	1,65	-0,45	0,653	-3,99	2,50
score_m	0,66	0,02	28,16	0,000	0,62	0,71
time=3 x *score_m*	-0,04	0,03	-1,09	0,274	-0,10	0,03
time=4 x *score_m*	-0,03	0,03	-0,86	0,392	-0,09	0,04
time=5 x *score_m*	-0,10	0,03	-2,91	0,004	-0,16	-0,03
constante	9,21	1,29	7,13	0,000	6,69	11,75

Prob > F = 0,000
R-squared = 0,2897

De acordo com o valor p do teste F apresentado no fim da tabela, concluiu-se que a hipótese nula do teste de significância do modelo foi rejeitada, o que indica que pelo menos um coeficiente do modelo é diferente de zero.

O coeficiente de determinação, apesar de muito pequeno, está próximo do que observamos com frequência em aplicações que envolvem o ajuste de modelos de regressão para dados sócio-econômico-demográficos.

Figura 12. Diagramas de dispersão para os resíduos padronizados do modelo de regressão linear para dados transversais.

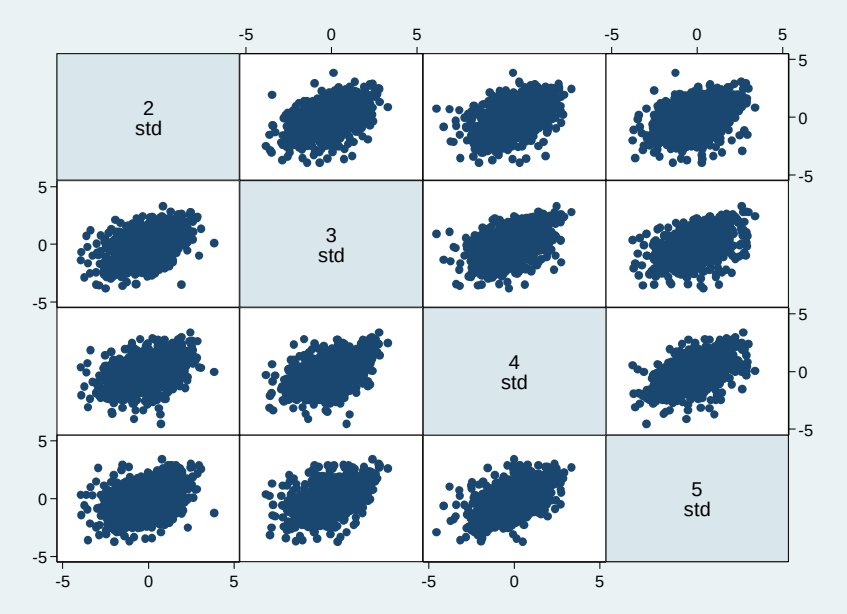


Figura 13. Matriz de correlação para os resíduos padronizados (std) do modelo de regressão linear para dados transversais.

	std2002	std2003	std2004	std2005
std2002	1,0000			
std2003	0,5202	1,0000		
std2004	0,4674	0,5309	1,0000	
std2005	0,4230	0,4719	0,5384	1,0000

A Figura 12 e a Figura 13 indicam correlação positiva. Mesmo com uma aparente redução nos valores estimados para as correlações apresentadas na Figura 13, consideramos que, devido à ordem de magnitude desta redução ser pequena (podendo, então ser causada por erros amostrais), a estrutura da matriz de correlação para os resíduos mais indicada parece ser a permutável.

Como o objetivo desta monografia não é o estudo de modelos de regressão para dados transversais, seguimos com o ajuste dos modelos de regressão para dados longitudinais.

Sendo assim, iniciamos a seguir nosso exercício de modelagem com o ajuste de dois modelos marginais de regressão simples para dados longitudinais, ambos tendo como variável dependente *score_l*. Como variáveis preditoras foram consideradas respectivamente *time*, definida como categórica (Tabela 10), e *logfihhmn*, que é uma variável contínua (Tabela 11). Após o ajuste dos modelos foram estimadas matrizes de correlação para os resíduos ao longo do tempo, considerando uma estrutura permutável, e as mesmas são apresentadas nas Figuras 14 e 15. Os modelos a seguir foram ajustados com o objetivo de explorar as possíveis alternativas de modelos a serem adotados.

Tabela 10. Coeficientes do modelo de regressão simples para dados longitudinais – variável independente: *time*.

score_l	**Coeficiente**	**Erro padrão**	**z**	**P > |z|**	**Lim. inferior**	**Lim. superior**
time=3	0,06	0,07	0,89	0,375	-0,07	0,19
time=4	0,10	0,07	1,58	0,114	-0,03	0,23
time=5	-0,43	0,07	-6,58	0,000	-0,56	-0,30
constante	14,09	0,08	186,77	0,000	13,94	14,24

Wald chi2(3) = 84,65
Prob > chi2 = 0,000

Figura 14. Matriz de correlação do modelo de regressão simples para dados longitudinais – variável independente: *time*.

	2002	2003	2004	2005
2002	1,0000			
2003	0,6179	1,0000		
2004	0,6179	0,6179	1,0000	
2005	0,6179	0,6179	0,6179	1,0000

Os valores p das estimativas dos coeficientes apresentados na Tabela 10 sugerem que o coeficiente para 2005 e o intercepto foram significativos, ao contrário dos coeficientes para 2003 e 2004. A Figura 14 indica uma correlação linear positiva intra-indivíduos, ou seja, uma correlação positiva para os resíduos para os mesmos indivíduos ao longo do tempo.

A seguir apresentamos os resultados para o ajuste do modelo que tem como variável explicativa a variável contínua *logfihhmn*.

Tabela 11. Coeficientes do modelo de regressão simples para dados longitudinais – variável independente: *logfihhmn*.

score_l	Coeficiente	Erro padrão	z	P > |z|	Lim. inferior	Lim. superior
logfihhmn	-0,29	0,13	-2,14	0,032	-0,55	-0,02
constante	15,01	0,47	32,14	0,000	14,10	15,93

Wald chi2(1) = 4,59
Prob > chi2 = 0,0321

Figura 15. Matriz de correlação do modelo de regressão simples para dados longitudinais – variável independente: *logfihhmn*.

	2002	2003	2004	2005
2002	1,0000			
2003	0,6148	1,0000		
2004	0,6148	0,6148	1,0000	
2005	0,6148	0,6148	0,6148	1,0000

A Figura 15 possui valores positivos, indicando uma correlação linear positiva intra-indivíduos. Neste novo modelo, apresentado na Tabela 11, tanto a constante quanto o intercepto foram significativos. Este modelo indica que, considerando as observações realizadas ao longo do tempo, um indivíduo que possui uma unidade a mais no logaritmo da renda, tende a ter um valor para o escore de lazer 0,29 unidades menor.

Seguimos com o ajuste do modelo com a variável *logfihhm* como variável preditora considerando agora o estimador do erro padrão (EP) semi robusto, calculado de acordo com a expressão (19). Os resultados, apresentados na Tabela 12, foram praticamente os mesmos do modelo anterior, com mudanças apenas na estatística Wald e no valor p do teste.

Tabela 12. Coeficientes do modelo de regressão simples para dados longitudinais com EP semi robusto – variável independente: *logfihhm*.

score_l	Coeficiente	EP semi-robusto	z	P > |z|	Lim. inferior	Lim. superior
logfihhmn	-0,29	0,13	-2,14	0,032	-0,55	-0,02
constante	15,01	0,47	32,14	0,000	14,10	15,93

Wald chi2(1) = 4,60
Prob > chi2 = 0,0320

Figura 16. Matriz de correlação do modelo de regressão simples para dados longitudinais com EP semi robusto – variável independente: *logfihhmn*.

	2002	2003	2004	2005
2002	1,0000			
2003	0,6148	1,0000		
2004	0,6148	0,6148	1,0000	
2005	0,6148	0,6148	0,6148	1,0000

Continuando, ajustamos um modelo de regressão para dados longitudinais incluindo um termo quadrático, mantendo o cálculo do erro padrão semi-robusto. A variável independente considerada foi *time*. Concluímos que o termo quadrático é requerido nesse caso, como mostra a Tabela 13, o que está de acordo com os resultados apresentados na Tabela 10. A Figura 17 sugere correlação linear positiva intra-indivíduo, apresentando uma estrutura permutável.

Tabela 13. Coeficientes do modelo de regressão para dados longitudinais com termo quadrático.

score_l	**Coeficiente**	**EP semi-robusto**	**z**	**P > |z|**	**Lim. inferior**	**Lim. superior**
time	0,92	0,16	5,77	0,000	0,61	1,23
timesq	-0,15	0,02	-6,61	0,000	-0,19	-0,10
constante	12,82	0,27	47,87	0,000	12,29	13,34

Wald chi2(2) = 72,32
Prob > chi2 = 0,000

Figura 17. Matriz de correlação do modelo de regressão para dados longitudinais com termo quadrático.

	2002	2003	2004	2005
2002	1,0000			
2003	0,6176	1,0000		
2004	0,6176	0,6176	1,0000	
2005	0,6176	0,6176	0,6176	1,0000

Os resultados apresentados nas Tabelas 14, 15 e 16 foram devidos a ajustes de modelos de regressão múltipla para dados longitudinais, os modelos marginais, considerando mais de uma variável explicativa. As variáveis explicativas utilizadas foram as mesmas do ajuste do modelo para dados transversais, além das variáveis *time* e *timesq*, significativas no

modelo apresentado na Tabela 13. Consideramos uma estrutura de correlação permutável, de início, que indicou correlação linear positiva intra-indivíuo (resultados na Tabela 14).

Tabela 14. Coeficientes do modelo marginal com estrutura de correlação permutável.

score_l	**Coeficiente**	**EP semi-robusto**	**z**	**P > |z|**	**Lim. inferior**	**Lim. superior**
sex=1	-0,28	0,12	-2,28	0,022	-0,52	-0,04
agecat=2	-0,01	0,23	-0,04	0,967	-0,46	0,44
agecat=3	-0,32	0,25	-1,25	0,210	-0,82	0,18
agecat=4	-0,69	0,25	-2,76	0,006	-1,18	-0,20
agecat=5	-0,55	0,26	-2,09	0,037	-1,06	-0,03
nchild	-0,46	0,06	-7,36	0,000	-0,58	-0,34
logfihhmn	-0,89	0,18	-4,62	0,000	-1,20	-0,48
jbhrs2=1	0,09	0,17	0,53	0,599	-0,24	0,42
qualf=2	0,24	0,15	1,56	0,120	-0,06	0,54
qualf=3	0,33	0,19	1,70	0,090	-0,05	0,71
qualf=4	0,49	0,19	2,53	0,011	0,11	0,86
qualf=5	0,52	0,21	2,49	0,013	0,11	0,92
mrjrgsc2=2	0,15	0,24	0,63	0,527	-0,32	0,62
mrjrgsc2=3	0,44	0,26	1,71	0,087	-0,06	0,95
mrjrgsc2=4	0,35	0,27	1,32	0,186	-0,17	0,88
mrjrgsc2=5	0,48	0,28	1,73	0,083	-0,06	1,03
mrjrgsc2=6	0,05	0,46	0,11	0,913	-0,86	0,96
mrjrgsc2=7	4,22	0,29	14,58	0,000	3,65	4,79
hlstat2=1	0,13	0,11	1,17	0,241	-0,09	0,34
mastat=1	0,17	0,12	1,39	0,166	-0,07	0,40
score_m	0,49	0,01	32,77	0,000	0,46	0,52
time	0,30	0,16	1,86	0,063	-0,02	0,62
timesq	-0,06	0,02	-2,40	0,016	-0,10	-0,01
constante	9,36	0,81	11,49	0,000	7,76	10,95

Wald chi2(22) = 6560.85
Prob > chi2 = 0,000

Figura 18. Matriz de correlação do modelo marginal com estrutura de correlação permutável.

	2002	2003	2004	2005
2002	1,0000			
2003	0,5099	1,0000		
2004	0,5099	0,5099	1,0000	
2005	0,5099	0,5099	0,5099	1,0000

Considerando um nível de significância de 5%, a maior parte dos coeficientes presentes na Tabela 14 foi significativa. No caso de variáveis categóricas, basta que pelo

menos uma categoria da variável seja significativa para o modelo para que ela continue no ajuste do modelo, que é o que acontece com a variável *mrjrgsc2*.

Assumindo uma estrutura de correlação independente, os resultados do ajuste do modelo marginal são apresentados a seguir, na Tabela 15.

Tabela 15. Coeficientes do modelo marginal com estrutura de correlação independente.

score_l	Coeficiente	EP semi-robusto	z	P > |z|	Lim. inferior	Lim. superior
sex=1	-0,31	0,12	-2,64	0,008	-0,55	-0,08
agecat=2	0,05	0,29	0,18	0,860	-0,51	0,61
agecat=3	-0,44	0,30	-1,47	0,141	-1,03	0,15
agecat=4	-0,80	0,30	-2,68	0,007	-1,38	-0,21
agecat=5	-0,69	0,30	-2,28	0,022	-1,29	-0,10
nchild	-0,46	0,07	-6,99	0,000	-0,60	-0,33
logfihhmn	-1,06	0,22	-4,83	0,000	-1,49	-0,63
jbhrs2=1	0,07	0,16	0,46	0,649	-0,24	0,39
qualf=2	0,27	0,16	1,71	0,087	-0,04	0,58
qualf=3	0,31	0,20	1,52	0,128	-0,09	0,70
qualf=4	0,49	0,19	2,55	0,011	0,11	0,87
qualf=5	0,49	0,21	2,35	0,019	0,08	0,90
mrjrgsc2=2	0,18	0,24	0,74	0,459	-029	0,65
mrjrgsc2=3	0,54	0,26	2,07	0,038	0,03	1,04
mrjrgsc2=4	0,39	0,27	1,47	0,143	-0,13	0,92
mrjrgsc2=5	0,57	0,28	2,02	0,043	0,02	1,12
mrjrgsc2=6	0,09	0,46	0,18	0,854	-0,82	0,99
mrjrgsc2=7	4,07	0,29	13,81	0,000	3,49	4,65
hlstat2=1	0,12	0,11	1,11	0,268	-0,09	0,33
mastat=1	0,16	0,13	1,21	0,228	-0,10	0,42
score_m	0,62	0,02	35,01	0,000	0,59	0,66
time	0,19	0,17	1,16	0,247	-0,13	0,52
timesq	-0,04	0,02	-1,58	0,114	-0,08	-0,01
constante	8,38	0,93	9,04	0,000	6,56	10,20

Wald chi2(22) = 6690.42
Prob > chi2 = 0,000

A um nível de significância de 5%, novamente grande parte dos coeficientes da Tabela 15 foi significativa. Os coeficientes significantes neste modelo foram os mesmos do modelo marginal considerando que as observações são correlacionadas. Algumas estimativas de intervalo de confiança mudaram consideravelmente em relação ao modelo anterior, enquanto outras se mantiveram próximas aos valores encontrados anteriormente. Os erros padrão semi robustos calculados nos modelos apresentados nas Tabelas 14 e 15 foram bem parecidos.

Consideramos também uma estrutura de correlação autorregressiva para ajuste do modelo marginal. A matriz de correlação do modelo, apresentada na Figura 18, também apresentou correlação linear positiva intra-indivíduos.

Tabela 16. Coeficientes do modelo marginal com estrutura de correlação autorregressiva.

score_l	**Coeficiente**	**EP semi-robusto**	**z**	**P > |z|**	**Lim. inferior**	**Lim. superior**
sex=1	-0,25	0,12	-2,03	0,042	-0,49	-0,01
agecat=2	-0,06	0,25	-0,24	0,814	-0,55	0,43
agecat=3	-0,44	0,27	-1,62	0,104	-0,96	0,09
agecat=4	-0,84	0,27	-3,12	0,002	-1,36	-0,31
agecat=5	-0,71	0,28	-2,56	0,010	-1,25	-0,17
nchild	-0,49	0,06	-7,78	0,000	-0,61	-0,36
logfihhmn	-0,81	0,19	-4,27	0,000	-1,18	-0,44
jbhrs2=1	0,07	0,17	0,41	0,682	-0,26	0,40
qualf=2	0,29	0,16	1,85	0,064	-0,02	0,60
qualf=3	0,36	0,20	1,81	0,070	-0,03	0,75
qualf=4	0,56	0,19	2,93	0,003	0,19	0,94
qualf=5	0,61	0,21	2,82	0,005	0,18	1,02
mrjrgsc2=2	0,15	0,24	0,61	0,541	-0,33	0,63
mrjrgsc2=3	0,44	0,26	1,69	0,092	-0,07	0,96
mrjrgsc2=4	0,32	0,27	1,19	0,234	-0,21	0,86
mrjrgsc2=5	0,49	0,28	1,73	0,083	0,06	1,05
mrjrgsc2=6	0,03	0,47	0,06	0,949	-0,88	0,94
mrjrgsc2=7	3,99	0,29	13,59	0,000	3,41	4,56
hlstat2=1	0,14	0,11	1,29	0,198	-0,07	0,36
mastat=1	0,20	0,12	1,58	0,114	-0,05	0,44
score_m	0,51	0,01	33,81	0,000	0,48	0,54
time	0,27	0,16	1,62	0,105	-0,06	0,59
timesq	-0,05	0,02	-2,26	0,024	-0,10	-0,01
constante	9,11	0,85	10,77	0,000	7,45	10,77

Wald chi2(22) = 6050,00
Prob > chi2 = 0,000

Figura 19. Matriz de correlação do modelo marginal com estrutura de correlação autorregressiva.

	2002	2003	2004	2005
2002	1,0000			
2003	0,5364	1,0000		
2004	0,2877	0,5364	1,0000	
2005	0,1543	0,2877	0,5364	1,0000

Considerando os resultados da Tabela 16, quase todos os mesmos coeficientes foram significativos comparando-se com os outros modelos marginais ajustados. As estimativas do erro padrão semi-robusto foram bem aproximadas das estimativas dos outros modelos, assim como os intervalos de confiança se assemelham aos do primeiro modelo marginal. Os resultados dos testes de Wald indicam que os coeficientes, quando testados conjuntamente, são significativos para os modelos que foram ajustados.

Como a matriz com os diagramas de dispersão e a matriz de correlação para os resíduos, apresentadas nas Figuras 12 e 13, sugeriam que a forma de especificação da matriz de correlação de trabalho mais adequada era a permutável, interpretamos apenas os coeficientes estimados para o modelo que leva em consideração tal estrutura, apresentados na Tabela 14, que é a que mais se aproxima à estrutura mencionada. Lembrando que o ajuste do modelo para dados transversais apresentado na Tabela 9 foi feito para dar suporte à escolha da estrutura de correlação de trabalho utilizada no ajuste dos modelos marginais através da estrutura de correlação apresentada pelos resíduos do modelo transversal. O coeficiente para a categoria ‘feminino’ da variável *sex*, estimado em -0,28 e significativo ao nível de 5%, sugere que mulheres possuem em média 0,28 unidades a menos no escore de satisfação em relação ao lazer em relação aos homens, que constituem a categoria de referência.

Os coeficientes para as categorias ‘40 a 49 anos’ e ‘maiores de 50 anos’ estimados em -0,69 e -0,55, respectivamente, sugerem que indivíduos presentes nessas categorias possuem em média 0,69 e 0,55 unidades a menos no escore de lazer em relação aos jovens de 16 a 21 anos, que são a categoria de referência. Analogamente, os indivíduos que pertencem às categorias ‘Ensino Fundamental’ e ‘Outros’ da variável *qualf* devem apresentar, em média, 0,49 e 0,52 unidades a mais no escore de lazer do que os indivíduos que possuem curso superior ou mais; assim como membros das Forças Armadas devem apresentar 4,22 unidades a mais no escore de lazer do que indivíduos com alguma ocupação profissional.

Para os coeficientes das variáveis contínuas *nchild*, *logfihhmn* e *score_m*, significativos a um nível de 5%, a interpretação é um pouco diferente. Um determinado indivíduo que possui um filho a mais em relação a outro indivíduo deve possuir, em média, 0,46 unidades a menos no escore de lazer, enquanto um indivíduo que possui uma unidade a mais no logaritmo da renda domiciliar em relação a outro indivíduo possui 0,89 unidades a menos no escore de lazer. O coeficiente estimado para o escore de satisfação em relação aos bens materiais sugere que um indivíduo que possui uma unidade a mais nessa variável em relação a outro individuo possui 0,49 unidades a mais no escore de satisfação em relação ao tempo destinado ao lazer. Como o coeficiente para a variável *time* é positivo, o coeficiente

negativo da variável *timesq* sugere um aumento nas respostas dos indivíduos para o escore de lazer nas ocasiões iniciais, estabilidade nas ocasiões intermediárias e um decréscimo nas ocasiões finais.

5 CONSIDERAÇÕES FINAIS

A proposta desta monografia foi apresentar de maneira breve técnicas de análise exploratória de dados e de modelagem para dados longitudinais. Sob o ponto de vista metodológico, a análise de dados longitudinais é bem mais complexa que a análise de dados transversais, que são coletados em apenas uma ocasião, ou no caso de repetição da pesquisa, coletados para amostras diferentes (pesquisas transversais repetidas). Para dados longitudinais, é preciso levar em consideração os efeitos do tempo sobre as respostas do indivíduo, além da própria correlação entre as respostas para o mesmo indivíduo ao longo do tempo. Concluímos que os modelos de regressão para dados longitudinais, os modelos marginais, são mais adequados para análises longitudinais que os modelos para dados transversais, devido ao fato de que o ajuste do modelo transversal para os dados longitudinais resultou em resíduos correlacionados, que violam o pressuposto de independência do método de MQO.

Uma das principais dificuldades do estudo contido nesta monografia foi a manipulação da base de dados utilizada. A inversão do formato do banco de dados e a recodificação das variáveis foram desafios enfrentados ao longo do estudo, mas que foram superados.

Como trabalho futuro, iremos considerar integralmente o desenho amostral adotado nos ajustes e análises dos modelos marginais.

6 REFERÊNCIAS BIBLIOGRÁFICAS

Charnet, Reinaldo; Freire Clarice A. L.; Charnet, Eugênia M. R.; Bonvino, Heloisa. *Análise de Modelos de Regressão Linear com aplicações*. Campinas-SP: Editora da Unicamp, 1999.

Salgueiro, M. F. R. F., Smith, P. W. F., Vieira, M. D. T. (2011) *A multi-process second-order latent growth curve model for subjective well-being. Quality and Quantity*.

Twisk, J. W. R. (2003) *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. Cambridge: Cambridge University Press.

Vieira, M. D. T. *Analysis of Longitudinal Survey Data*. 1. ed. Saarbrücken: VDM Verlag Dr. Müller. 2009.

Vieira, M. D. T., Bastos, R. R., Souza, A. C., Hippert, H. S. (2011) *On the use of exploratory and confirmatory longitudinal data analysis. Advances and Applications in Statistics*, v. 22, p. 129-156.

Vieira, M. D. T. and Skinner, C. J. (2008). *Estimating Models for Panel Survey Data under Complex Sampling*. *Journal of Official Statistics*, 24, 343-364.