

UNIVERSIDADE ABERTA



UNIVERSIDADE
AbERTA
www.uab.pt

**Preferência pelo tipo de parto em Portugal, uma abordagem
multivariada**

Diogo Gonçalo Sousa Barrios

Mestrado em Estatística, Matemática e Computação

2021

UNIVERSIDADE ABERTA



**Preferência pelo tipo de parto em Portugal, uma abordagem
multivariada**

Diogo Gonçalo Sousa Barrios

Mestrado em Estatística, Matemática e Computação

Dissertação de mestrado orientada pela Professora Doutora Catarina Sofia da Costa Nunes
Duarte e pelo Professor Doutor Humberto Machado

2021

Resumo

Introdução: O aumento da taxa de cesarianas, tem sido levada nos últimos anos cada vez mais em consideração pela investigação científica, por causa de ser mais onerosa ao erário público, e tendo maior risco tanto para a mãe como para o bebé. Neste estudo pretende-se efetuar uma pesquisa multifatorial, através de técnicas estatísticas multivariadas, que possa explicar a preferência da grávida pelo parto por via cesariana.

Métodos: Assente numa amostra de 245 grávidas (com 1 omissa, dividida em 2 grupos, G0 - *Preferência pelo parto vaginal* e G1 - *Preferência pelo parto cesariana*, sendo o $n_0 = 190$ e o $n_1 = 54$ respetivamente), inquiridas no Centro Materno-Infantil do Norte, pertencente ao Centro Hospitalar do Porto, foi executada, inicialmente uma análise exploratória dos dados, seguindo-se de uma análise de componentes principais nas questões com natureza ordinal (escala de likert), com propósito de reduzir a complexidade da informação numa estrutura mais simples. Posteriormente foi desenvolvida uma regressão logística com todas as variáveis, incluindo as componentes criadas pela análise de componentes principais, para identificar os possíveis fatores preditivos para a preferência do tipo de parto.

Resultados: A análise de componentes principais, identificou 3 componentes para cada grupo, sendo denominadas por, *vantagens do parto para a Mãe*, *medo do parto* e *preocupação com o bebé*. Pela regressão logística foram apurados fatores como *idade* ($p < 0.05$), *obtenção de informação junto do médico familiar/obstetra* ($p < 0.05$), *obtenção de informação junto do marido/parceiro/amigo* ($p < 0.05$), *vantagens do parto para a Mãe* ($p < 0.01$), *medo do parto* ($p < 0.01$) e *preocupação com o bebé* ($p < 0.05$), que são significativos para opção pelo tipo de parto.

Conclusão: Nesta pesquisa, com atenção à assimetria dos tamanhos dos grupos e sendo a amostra total apenas do sector público, foram identificados fatores específicos que podem ajudar na explicação da problemática, tais como a *idade* (*Odds Ratio* ($\hat{\beta}$) = 1.12, $p < 0.05$), *medo do parto* (*Odds Ratio* ($\hat{\beta}$) = 1.04, $p < 0.01$) e a *preocupação com o bebé* (*Odds Ratio* ($\hat{\beta}$) = 1.02, $p < 0.01$).

Palavras-Chave: cesariana, grávida, preferência pelo parto, análise componentes principais, regressão logística

Abstract

Purpose: The increase in the rate of cesarean sections, has been increasingly taken into account by scientific research in recent years, because it is more costly to the public purse, and has a greater risk for both the mother and the baby. In this study, we intend to carry out a multifactorial research, using multivariate statistical techniques, which may explain the preference of pregnant by cesarean delivery.

Methods: Based on a sample of 245 pregnant (with 1 missing, divided into 2 groups, G0 - *Preference for vaginal delivery* and G1 - *Preference for cesarean delivery*, with $n_0 = 190$ and $n_1 = 54$ respectively), inquired at the Centro Materno-Infantil do Norte, belonging to the Centro Hospitalar do Porto, initially an exploratory analysis of the data was performed, followed by a principal component analysis in the questions of ordinal nature (likert scale), with the purpose of reducing the complexity of the information in a simpler structure. Subsequently, a logistic regression with all variables was developed, including the components created by the principal component analysis, to identify possible predictive factors for the type of delivery preference.

Results: The principal components analysis, identified 3 components for each group, which were denominated as: *advantages of childbirth for the Mother*, *fear of childbirth* and *concern with the baby*. The logistic regression identified *age* ($p < 0.05$), *obtaining information from the family doctor/obstetrician* ($p < 0.05$), *obtaining information from husband/partner/friend* ($p < 0.05$), *advantages of childbirth for the Mother* ($p < 0.01$), *fear of childbirth* ($p < 0.01$) and *concern with the baby* ($p < 0.05$), that are significant for choosing the type of delivery.

Conclusion: In this work, taking into account that the sample was only from a public sector and that the size of the groups was asymmetric, specific factors were identified that can help to explain the reasons behind the preference for the type of delivery, such as *age* (*Odds Ratio* ($\hat{\beta}$) = 1.12, $p < 0.05$), *fear of childbirth* (*Odds Ratio* ($\hat{\beta}$) = 1.04, $p < 0.01$) and the *concern for the baby* (*Odds Ratio* ($\hat{\beta}$) = 1.02, $p < 0.01$).

Keywords: cesarean, pregnant, preference for childbirth, principal component analysis, logistic regression

Dedicatória

Dedico esta dissertação à minha Mãe, por todo o apoio que me deu, desde o ensino obrigatório, para que nunca desistisse dos meus sonhos.

À minha esposa que me apoiou e esteve sempre a meu lado dando ânimo mesmo em dias mais difíceis.

Ao meu filho, pois foi através do nascimento dele, que senti curiosidade sobre este tema.

Aos meus irmãos que me vêm como um exemplo a seguir e aos meus sogros que sempre me apoiaram nesta caminhada.

Agradecimentos

Agradeço à orientadora Professora Doutora Catarina Sofia da Costa Nunes pela disponibilidade e paciência perante todo o desenvolvimento desta dissertação, como também à coorientação do Professor Doutor Humberto Machado.

Publicações e Comunicações

Parte do trabalho e resultados desenvolvidos nesta dissertação, foram aceites, apresentados oralmente e publicados na ata **XXIV Congresso da Sociedade Portuguesa de Estatística**, que ocorreu em Amarante nos dias 6 a 9 de Novembro de 2019, intitulado *Preferência pelo tipo de parto em Portugal: análise em componentes principais*, que encontra-se no Anexo II

ÍNDICE

Índice de Figuras	ix
Índice de Tabelas	x
Índice de Abreviaturas	xii
1 Introdução	1
1.1 Enquadramento	2
1.2 Objetivos	3
1.3 Estrutura	4
2 Análise Exploratória dos Dados	5
3 Análise de Componentes Principais	13
3.1 Enquadramento	14
3.2 Metodologia	15
3.2.1 Preparação dos dados para a Análise de Componentes Principais . .	15
3.2.2 Extração das componentes	16
3.2.3 Rotação e interpretação das componentes	21
3.3 Implementação e Resultados	24
3.3.1 G0 - Preferência pelo Parto Vaginal	25
3.3.2 G1 - Preferência pelo Parto Cesariana	35
4 Regressão Logística	44
4.1 Enquadramento	45
4.2 Metodologia	47
4.2.1 Ajustamento do Modelo de Regressão Logística Múltipla	47

4.2.2	Teste de Significância dos Coeficientes Estimados	49
4.2.3	Método de Seleção das Covariáveis	50
4.2.4	Avaliação da Qualidade do Ajustamento do Modelo Final	52
4.2.5	Diagnóstico do Modelo Final	56
4.3	Implementação e Resultados	58
5	Conclusão	73
	Bibliografia	75
	Anexo I	82
	Anexo II	86
	Anexo III	93
	Anexo IV	94
	Anexo V	98

ÍNDICE DE FIGURAS

2.1	Histograma da $Q3_A$ - <i>Semanas de Gestação</i> e da $Q3_B$ - <i>Idade</i>	7
2.2	Gráfico Q-Q da $Q3_A$ - <i>Semanas de Gestação</i> e da $Q3_B$ - <i>Idade</i>	8
2.3	Caixa de bigodes da $Q3_A$ - <i>Semanas de Gestação</i> e da $Q3_B$ - <i>Idade</i>	8
2.4	$Q3_C$ - <i>Nível de Escolaridade por grupos</i>	9
2.5	Fontes de informação	10
2.6	G0 - <i>Porque preferiria um parto vaginal?</i>	11
2.7	G1 - <i>Porque preferiria um parto cesariana?</i>	12
3.1	<i>Gráfico de escarpa</i> das Componentes Principais do G0	30
3.2	<i>Gráfico de escarpa</i> das Componentes Principais do G1	40
5.3	Visualização das variáveis contínuas do modelo L_{10}	63
5.5	Descriminação do modelo ajustado L_{10} aos dados, com uma AUROC = 81.73, $n = 244$	67
5.6	Diagnóstico dos resíduos do modelo logístico L_{10}	68
5.1	Histogramas antes e depois de escalados, da $PC1$ - <i>vant_parto</i> , $PC2$ - <i>preo_bebe</i> e $PC3$ - <i>medo_parto</i> , referente ao G0	99
5.2	Histogramas antes e depois de escalados, da $PC1$ - <i>vant_parto</i> , $PC2$ - <i>medo_parto</i> e $PC3$ - <i>preo_bebe</i> , referente ao G1	100
5.4	Ponto de corte ótimo, $c = 0.224$	101

ÍNDICE DE TABELAS

2.1	Tabela de contingência da $Q3_D1$ - Já teve algum Parto? e $Q3_D2$ - Se sim, já teve alguma cesariana?	6
3.3	Matriz de correlação do G0	26
3.4	Matriz de correlação anti-imagem do G0	27
3.5	Matriz de correlação do G0 (sem Q1B5 e Q1B10)	28
3.6	Matriz de correlação anti-imagem do G0 (sem Q1B5 e Q1B10)	28
3.7	Extração das Componentes Principais do G0	30
3.8	Matriz não rotacionada das Componentes Principais do G0	31
3.9	Matrizes rotacionadas das Componentes do G0: Conjunto completo e reduzido	34
3.12	Matriz de correlação do G1	36
3.13	Matriz de correlação anti-imagem do G1	37
3.14	Matriz de correlação do G1 (sem o Q1C10, Q1C9 e Q1C4)	37
3.15	Matriz de correlação anti-imagem do G1 (sem Q1C10, Q1C9 e Q1C4)	38
3.16	Extração das Componentes Principais do G1	40
3.17	Matriz não rotacionada das Componentes Principais do G1	41
3.18	Matriz rotacionada das componentes do G1	43
5.2	Ajustamento do modelo logístico L_1 com as covariáveis $p < 0.25$	61
5.4	Modelo logístico preliminar L_4	62
5.7	Modelo logístico preliminar L_{10}	64
5.10	Avaliação da qualidade do Ajustamento do modelo L_{10}	69
5.12	Tabela dos casos influentes ou <i>outliers</i> do modelo L_{10}	70
5.13	Sumário do modelo final L_{31}	71
2.2	Estatísticas descritivas da $Q3_A$ - Semanas de Gestação e $Q3_B$ - Idade	90

2.3	Tabela do $Q3_C$ - Nível de Escolaridade, por grupos valores aboslutos e percentuais	90
2.4	Tabela das Fontes de Informação	91
2.5	Tabela da $Q1B$ - Porque preferiria um parto vaginal?	91
2.6	Tabela da $Q1C$ - Porque preferiria um parto cesariana?	92
3.1	Tabela de frequências das questões $Q1B1$ à $Q1B10$	94
3.2	Tabela de omissos por ordem decrescente das questões $Q1B1$ à $Q1B10$	94
3.10	Tabela de frequências das items $Q1C1$ à $Q1C10$	95
3.11	Tabela de omissos por ordem decrescente das items $Q1C1$ à $Q1C10$	95
5.1	Tabela de contingência entre $Q1A$ e as variáveis dicotômicas	102
5.3	Exclusão de variáveis com $p > 0.05$ do modelo logístico L_1	103
5.5	$\Delta\%\hat{\beta}$ do L_4 em relação ao L_1	103
5.6	Inclusão das variáveis com $p > 0.25$ no modelo logístico L_4 , com a exclusão das variáveis $Q2_5$, $Q3_A$ e $Q3_C$	103
5.8	$\Delta\%\hat{\beta}$ do L_{10} em relação ao L_4 , com a inclusão da variável $preo_bebe$	104
5.9	Inclusão das interações ao modelo logístico L_{10}	104
5.11	Percentis das probabilidades estimadas	105
5.14	$\Delta\%\hat{\beta}$ do L_{31} em relação ao L_{10} , com a exclusão dos casos 198, 203, 215, 240	105

ÍNDICE DE ABREVIATURAS

OMS	Organização Mundial de Saúde
OCDE	Organização para a Cooperação e Desenvolvimento Económico
SNS	Serviço Nacional de Saúde
CNRTC	Comissão Nacional para a Redução da Taxa de Cesari- ana
CMIN	Centro Materno-Infantil do Norte
CHP	Centro Hospitalar do Porto
IQR	Intervalo Interquartil
ACP	Análise de Componentes Principais
AF	Análise Fatorial
KMO	Critério de Kaiser–Meyer–Olkin
MSA	Medida de Adequação da Amostra
D	Desviância
AIC	Critério de Informação Akaike
BIC	Critério de Informação Bayesiana
AUROC	Área sob a Curva da Característica Operatória do Rece- tor
Grupos:	
G0	Preferência pelo parto vaginal
G1	Preferência pelo parto cesariana

Questões:

Q1A	Tipo de Parto
Q1B1	Não deixa Cicatriz
Q1B2	Não quero perder a primeira hora de vida do meu bebê
Q1B3	Tenho medo
Q1B4	Não quero que o meu bebê nasça através de uma cirurgia
Q1B5	Já tive um parto vaginal
Q1B6	Conheço mais pessoas que tiveram um parto vaginal
Q1B7	Quero ter vários filhos e sei que uma cesariana pode limitar o número
Q1B8	A recuperação pós-parto é mais rápida
Q1B9	O tempo de internamento hospitalar é menor
Q1B10	Já tive uma cesariana
Q_B_1	Optaria por anestesia
Q1C1	Tenho medo da dor
Q1C2	Conseguo programar quando será o parto
Q1C3	O meu bebê sofre menos
Q1C4	Já tive uma cesariana
Q1C5	Após uma cesariana é mais fácil retomar a vida sexual
Q1C6	Conheço mais pessoas que tiveram uma cesariana
Q1C7	É mais seguro para a minha saúde
Q1C8	Tenho medo
Q1C9	Já tive um parto vaginal
Q1C10	A recuperação pós-parto é mais rápida
Q1C11	Outro
Q_C_1	Que tipo de anestesia preferia
Q2_1	Através do meu médico família/Obstetra
Q2_1.1	Através de um médico obstetra/de família ou Curso de Preparação para o Parto
Q2_2	Através do meu marido/parceiro
Q2_3	Através de um familiar

Q2_4	Através de um amigo
Q2_5	Ainda não pensei
Q2_6	Através de internet
Q2_7	Através de televisão
Q2_8	Através de um curso/livros
Q2_9	Através de jornais/revistas
Q2_10	Outro
Q3_A	Tempo de Gestação
Q3_B	Idade
Q3_C	Nível de Escolaridade
Q3_D1	Já teve algum parto
Q3_D2	Se sim, já teve alguma cesariana

Variáveis Transformadas:

Q2_2_4	Através do meu marido/parceiro/amigo
Q2_6_7	Através da internet/televisão
Q2_8_9	Através de um curso/livros/jornais/revistas

Variáveis criadas pela ACP:

vant_parto	Vantagens do parto para a Mãe
preo_bebe	Preocupação com o bebê
medo_parto	Medo do parto

CAPÍTULO 1

INTRODUÇÃO

1.1 Enquadramento

O parto por via cesariana foi introduzido na prática clínica como um procedimento *life-saving* tanto para a mãe como para o bebé, sendo que nos últimos 30 anos, têm vindo a aumentar anualmente nos países desenvolvidos e em vias de desenvolvimento [30]. Porém vários estudos indicam que esta prática clínica, incorre em sérios riscos à saúde da mãe e do recém-nascido [47][26].

Em 1985 a *OMS* (Organização Mundial de Saúde), publicou uma recomendação para uma taxa ideal entre os 10% e 15%. Todavia, a taxa de cesarianas tem sido muito superior ao longo do tempo, assim a OMS em 2015, emitiu uma nota, onde já não apresentava nenhuma taxa de cesarianas específica, apenas informava que este tipo de parto fosse praticado quando houvesse necessidade por motivos médicos [36].

Nos últimos 3 anos têm sido publicados cada vez mais estudos sobre a preferência pelo tipo de parto pela grávida. Nos Estados Unidos da América, 90% das grávidas com pelo menos um parto via cesariana, efetuaram um segundo parto da mesma maneira, apesar de preferirem o parto vaginal [2]. Na China, foram revelados fatores como a idade e o baixo nível de educação, para o aumento da taxa de cesarianas [33]. Na Irlanda e nos Emirados Árabes Unidos, para a diminuir a taxa de cesariana e para uma melhor decisão por parte da mulher grávida (devido à falta de conhecimento dos riscos para a saúde), foi dado como soluções, aconselhamento precoce e formação sobre os dois tipos de parto [1][45].

Na publicação das estatísticas da saúde da *OCDE* (Organização para a Cooperação e Desenvolvimento Económico) de 2019 [35], cujo Portugal faz parte, a tendência de cesarianas tem sido de um aumento na generalidade, isto é, no ano 2000, os países que fazem parte desta organização, apresentavam uma taxa média de cesarianas pouco acima dos 20%, e em 2017 situava-se num valor de 28,1%.

Ao nível nacional, de acordo com os dados PORDATA [39], a taxa de cesarianas em 2000 era de 28%, sendo o seu máximo em 2009 com 36,7%. Para fazer face a esta situação em 2013, o *SNS* (Serviço Nacional de Saúde), criou a *CNRTC* (Comissão Nacional para a Redução da Taxa de Cesariana), para implementar medidas de atenuação e redução da taxa de cesarianas, resultando, ano após ano numa tendência decrescente, e em 2017 apresentava um valor de 33%. De acordo com os dados PORDATA [40], quando analisada a taxa de

cesarianas, pelas unidades hospitalares portuguesas nas categorias públicas e privadas, estas tornam-se discrepantes, isto é, em 2000, nos hospitais públicos era de 26% e nos hospitais privados era de 57%, 17 anos depois, os valores são de 28% e 65% respetivamente. Constatando assim, que mais de metade das cesarianas a nível nacional, pertencem aos hospitais privados. Todavia na Argentina (Buenos Aires), num estudo da preferência pelo tipo de parto pela grávida, tanto nos hospitais públicos ou privados não foi verificado qualquer evidência [29].

Por conseguinte, existem poucos estudos sobre as razões das taxas de cesarianas apresentarem tais valores em Portugal. Assim foi efetuado um estudo, sobre a preferência do tipo de parto pelas grávidas [10], de forma a encontrar evidências que poderiam explicar o aumento ano após ano da taxa de parto por via cesariana. O trabalho teve como base um inquérito que encontra-se no Anexo I desta dissertação, a 245 mulheres grávidas no *CMIN* (Centro Materno-Infantil do Norte), do *CHP* (Centro Hospitalar do Porto), onde uma das conclusões do trabalho foi que as grávidas que já tiveram cesariana, preferem este tipo de parto [7], contudo a análise realizada aos dados foi apenas uma análise univariada e bivariada, sendo que possivelmente existe potencial de aprofundar mais este trabalho usando técnicas estatísticas multivariadas.

1.2 Objetivos

Dado que as razões para o aumento das taxas de parto via cesariana são multifatoriais e ainda não totalmente compreendidas, além de que, pela literatura clínica sobre esta problemática existem indícios de causas prováveis, como o aumento de número de gestações, preocupações com a responsabilidade médica, casos de negligência e conveniência de marcação da data e hora do parto[10], e com base no estudo acima mencionado, o objetivo deste trabalho é aprofundar a procura de fatores, que possam contribuir na explicação desta problemática, através da análise estatística multivariada, e eventualmente criar um modelo de previsão satisfatório, por forma a ajudar as autoridades de saúde na afinação de estratégias de mitigação que possam ser eficazes na diminuição da taxa de cesarianas.

1.3 Estrutura

A dissertação que se segue está estruturada em 5 capítulos.

O segundo capítulo é composto por uma revisão dos dados usando uma análise estatística univariada e bivariada de forma a extrair uma base de informação relevante para o aprofundamento através de técnicas estatísticas multivariadas.

O terceiro capítulo, aplicar-se-á uma análise de componentes principais (*ACP*) às questões de escala de Likert, de forma a extrair informação mais simples, relevante e concentrada.

No quarto capítulo, será criado um modelo logístico, com intuito de ser possível encontrar fatores que poderão ser preditivos na preferência pelo tipo de parto pelas grávidas.

Por fim no quinto capítulo, será efetuada uma conclusão da dissertação, apresentar-se-á de forma sucinta o trabalho realizado e resultados, como deixar-se-á algumas questões para futuras investigações.

A informação estatística apresentada nos capítulos seguintes, foi produzida com auxílio do software estatístico gratuito e de código aberto, R, versão 3.6.1.

CAPÍTULO 2

ANÁLISE EXPLORATÓRIA DOS DADOS

Com base nos dados extraídos pelo inquérito do Anexo I [7], o tamanho da amostra é de 245 grávidas (244 dados válidos e 1 omissos), sendo que explicitamente na primeira questão (*Q1A - tipo de parto*), divide a amostra em dois grupos, o *G0 - Preferência pelo parto vaginal*, com 190 casos ou 77,9% do total da amostra e o *G1 - Preferência pelo parto cesariana*, com 54 casos ou 22,1% do total da amostra.

De notar um desfazamento entre o tamanho dos grupos, tendo o maioritário uma representação superior a dois terços da amostra. Esta desproporcionalidade dos grupos é melhor entendida pela tabela 2.1, sendo que 103(42,21%) grávidas já tiveram pelo menos 1 parto e destas, 42(40,78%) grávidas tiveram pelo menos 1 parto via cesariana. Pelo teste do qui-quadrado de Pearson (H_0 : Não há associação entre *Q3_D2* e *Q1A*, $\chi^2(1, N = 103) = 12.34$, $p < 0.01$ ¹), existe evidência estatística para rejeitar a H_0 , ou seja, as grávidas que já tiveram pelo menos 1 parto e foi cesariana, preferem que o novo parto seja da mesma forma.

Pela tabela 2.2 do Anexo III, a variável *Q3_A - Semanas de Gestação* apresenta uma Tabela 2.1: Tabela de contingência da *Q3_D1 - Já teve algum Parto?* e *Q3_D2 - Se sim, já teve alguma cesariana?*

		<i>Q3_D1 - Já teve algum Parto?</i>	
		Sim	
<i>Q3_D2 - Se sim, já teve alguma cesariana?</i>	¹ Sim	42(40,78%)	
	Não	61(59,22%)	
Total		103(42,21%)	

¹ Teste Qui-Quadrado de Pearson, $\chi^2(1, N = 103) = 12.34$, $p < 0.01$

média de $\approx 25 \pm 8$ semanas. Através do teste de Shapiro-Wilks (H_0 : *Q3_A* tem uma distribuição normal, $W_{obs} = 0.955 > W_{(0.05;244)} = 0.947$, $p < 0.01$), mostra evidência estatística para rejeitar a H_0 , portanto a *Q3_A* não segue uma distribuição normal como mostra as figuras 2.1a e 2.2a. Observando por grupos, na variável *Q3_A*, o *G0*, mostra uma média $\approx 26 \pm 8$ semanas, enquanto que o *G1*, mostra uma média $\approx 23 \pm 8$ semanas. Através do teste de Mann-Whitney (H_0 : $\eta_{Q3_{AG0}} = \eta_{Q3_{AG1}}$, $U = 6002$, $p = 0.03$), existe evidência estatística

¹Para todos os testes estatísticos apresentados na dissertação será usado por defeito o nível de significância $\alpha = 0.05$

Figura 2.1: Histograma da $Q3_A$ - *Semanas de Gestação* e da $Q3_B$ - *Idade*



para rejeitar a H_0 . Assim sendo, a diferença entre as medianas dos grupos é estatisticamente significativa, sendo possível visualizar pela figura 2.3a.

No que diz respeito à variável $Q3_B$ - *Idade*, pela tabela 2.2 do Anexo III, apresenta uma média $\approx 32 \pm 5$ anos. Através do teste de Shapiro-Wilks (H_0 : $Q3_B$ tem uma distribuição normal, $W_{obs} = 0.978 > W_{(0.05;244)} = 0.947, p < 0.01$), mostra evidência estatística para rejeitar a H_0 , portanto a $Q3_B$ não segue uma distribuição normal, como mostra as figuras 2.1b e 2.2b. Por grupos, na variável $Q3_B$, o G_0 , apresenta uma média $\approx 31 \pm 5$ anos, por sua vez o G_1 , apresenta uma média $\approx 33 \pm 5$ anos. Através do teste de Mann-Whitney (H_0 : $\eta_{Q3_B_{G_0}} = \eta_{Q3_B_{G_1}}, U = 4202, p = 0.05$), existe evidência estatística para rejeitar a H_0 . Assim sendo, a diferença entre as medianas dos grupos é estatisticamente significativa, sendo possível visualizar pela figura 2.3b.

Os dados relevam também, pela tabela 2.3 do Anexo III, as habilitações literárias pela variável $Q3_C$ - *Nível de Escolaridade*, 53% das mulheres grávidas detêm um nível de escolaridade superior, das quais 35% detêm a Licenciatura (que corresponde à mediana) e 18% detêm um grau superior a esta (i.e., Mestrado ou Doutorado). No entanto, pelo teste do qui-quadrado de Pearson (H_0 : Não há associação entre a $Q3_C$ e $Q1A$, $\chi^2(4, N = 243) = 5.96, p = 0.202$), não existe evidência estatística para rejeitar a H_0 , logo, o nível de escolaridade nada influencia a preferência pelo tipo de parto. Por grupos, na variável $Q3_C$, como mostra a figura 2.4, evidencia o mesmo padrão, portanto uma maior concentra-

Figura 2.2: Gráfico Q-Q da $Q3_A$ - *Semanas de Gestação* e da $Q3_B$ - *Idade*

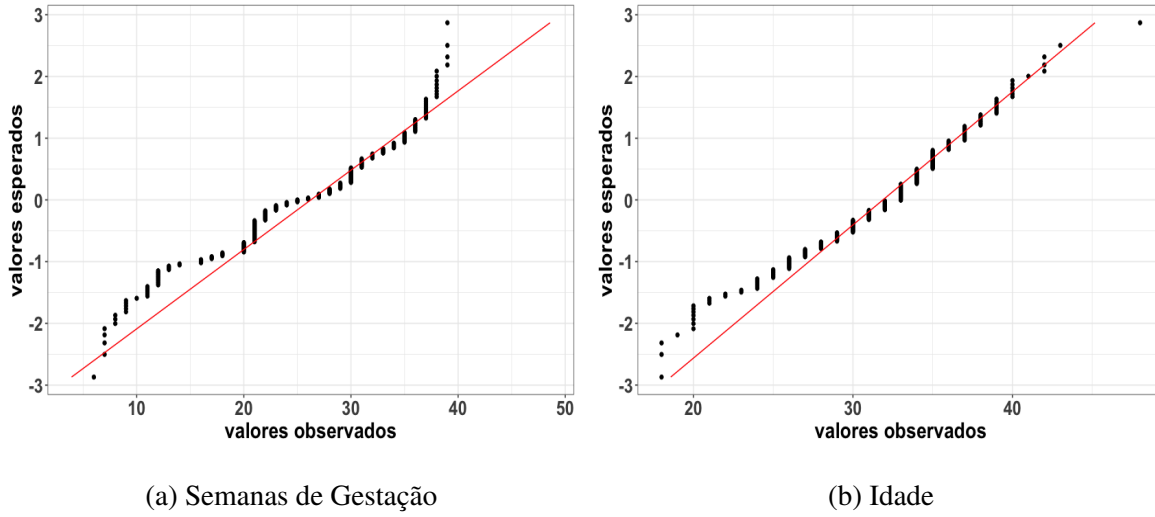
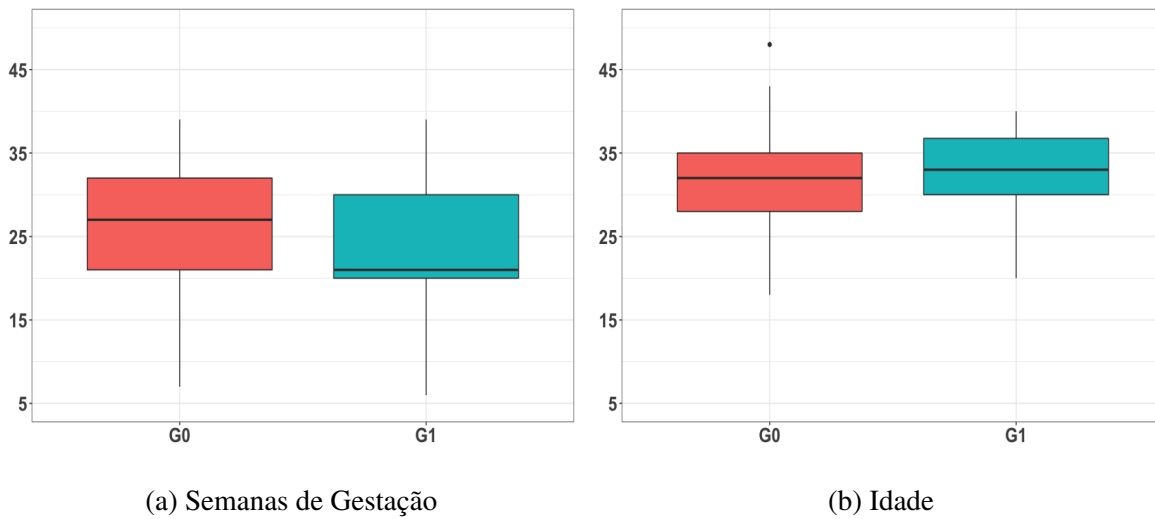
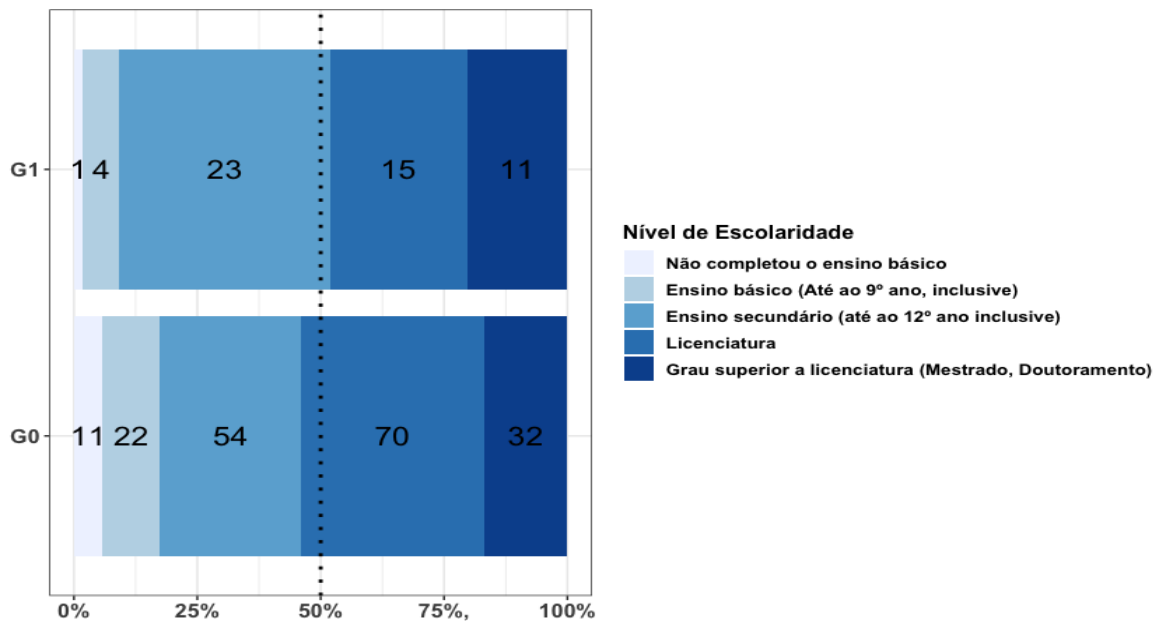


Figura 2.3: Caixa de bigodes da $Q3_A$ - *Semanas de Gestação* e da $Q3_B$ - *Idade*



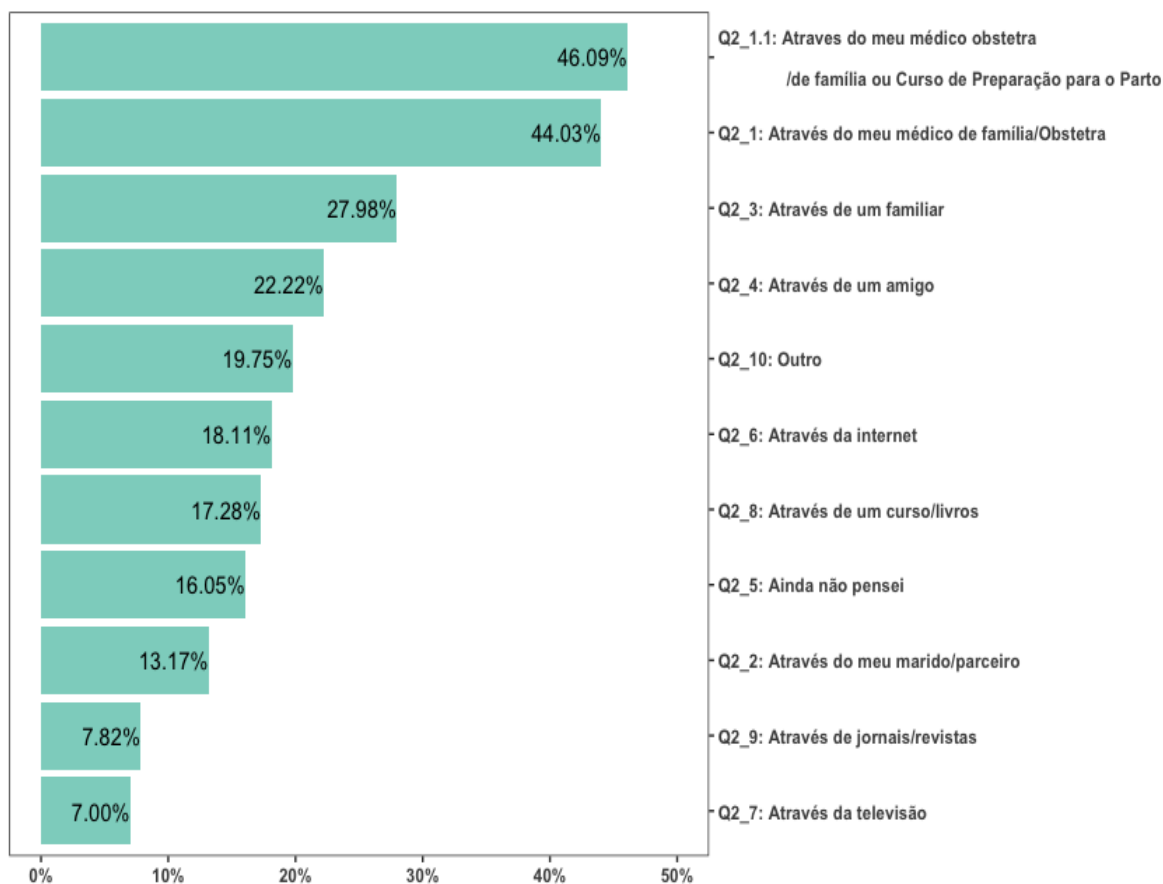
ção de mulheres grávidas com nível de escolaridade de ensino superior, apesar disso, a mediana do *G1* é o Ensino secundário (até ao 12ºano inclusive), enquanto que a mediana do *G0* é Licenciatura. Todavia pelo teste de Mann-Whitney ($H_0: \eta_{Q3_CG0} = \eta_{Q3_CG1}$, $U = 4999.5$, $p = 0.81$), não existe evidência estatística para rejeitar a H_0 , ou seja, as medianas dos grupos são iguais.

Figura 2.4: *Q3_C* - Nível de Escolaridade por grupos



Não obstante, pela tabela 2.4 do Anexo III, a obtenção de informação para auxiliar a decisão na preferência pelo tipo de parto pela grávida, verifica-se que ambos os grupos procuram informar-se junto do seu médico de família/obstetra ou curso de preparação para o parto que representa 46%, como mostra a figura 2.5. Foi averiguado, pelo teste do qui-quadrado de Pearson (H_0 : Não há associação entre a *Q2_1* e *Q1A*, $\chi^2(1, N = 107) = 4.31$, $p = 0.04$), que a *Q2_1* - *Através do meu médico de família/Obstetra*, é estatisticamente significativa, portanto, existe evidência estatística para rejeitar a H_0 , isto é, a obtenção da informação através do médico de família/Obstetra ajuda à decisão na preferência pelo tipo de parto.

Figura 2.5: Fontes de informação

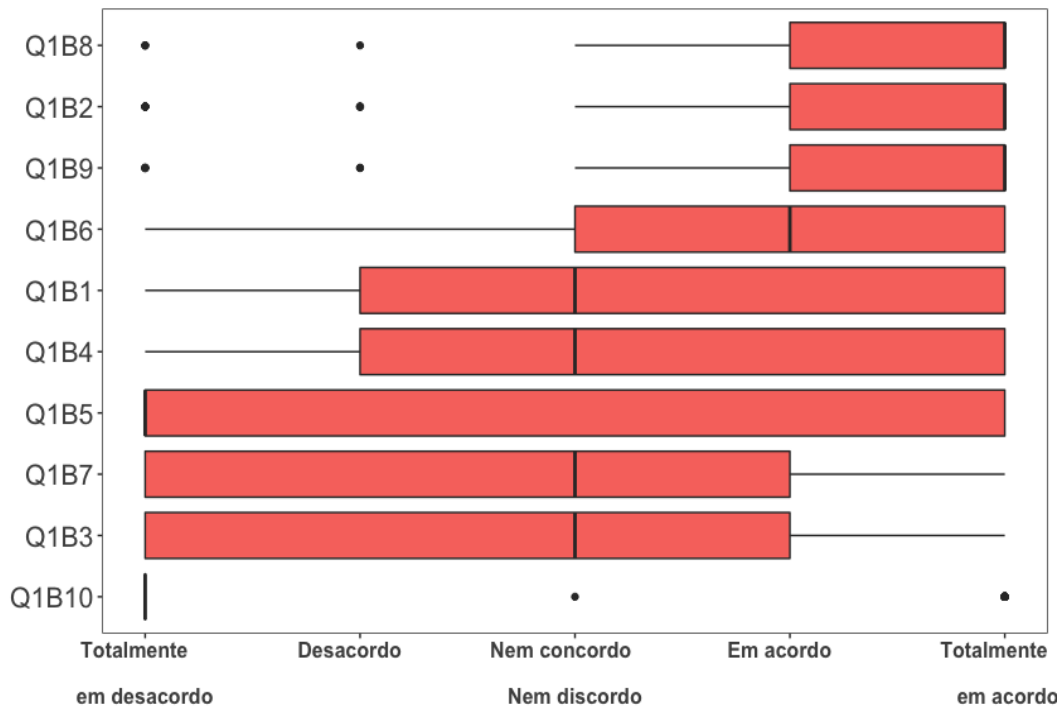


Pelas tabelas 2.5 e 2.6 do Anexo III, onde constam as razões da preferência pelo tipo de parto, através de um grupo de 10 questões de escala de Likert para cada grupo, em que o mínimo 1 - "Totalmente desacordo" e o máximo 5 - "Totalmente de acordo", sendo que o valor "Não sei" foi dado como omissos.

No grupo G0, pela tabela 2.5, é verificado uma unanimidade na mediana no máximo da escala 5 - "Totalmente de acordo", com um intervalo interquartil(IQR) de 1 nas variáveis Q1B8- A recuperação pós-parto é mais rápida, Q1B2 - Não quero perder a primeira hora de vida do meu bebê e Q1B9 - O tempo de internamente hospitalar é menor. Em contrapartida as variáveis Q1B10 - Já tive uma cesariana e Q1B5 - Já tive um parto vaginal, mostram uma mediana no mínimo da escala, 1 - "Totalmente desacordo", com uma IQR de 0 e 4 respectivamente, como é apresentado na figura 2.6. Como anteriormente já havia sido aferido, na amostra existe mais grávidas para ter o primeiro parto e mesmo as que já tiveram pelo

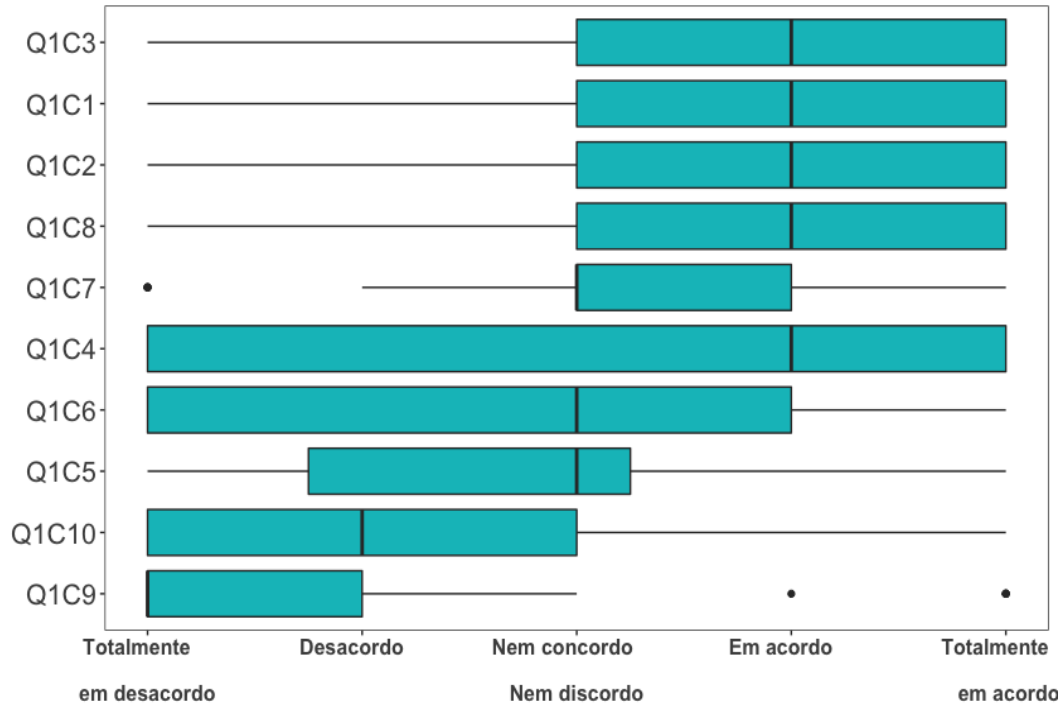
menos 1 parto, este foi vaginal. No entanto, quando o grupo *G0* foi confrontado com a questão *Q_B_1* - *Optaria por anestesia*, 74% admitem anestesia epidural, porém 16% ainda não sabem.

Figura 2.6: *G0* - Porque preferiria um parto vaginal?



No grupo *G1*, pela tabela 2.6, não apresenta nenhuma variável com a mediana no máximo da escala, contudo apresenta as variáveis *Q1C3* - *O meu bebé sofre menos*, *Q1C1* - *Tenho medo da dor*, *Q1C2* - *Consigo programar quando será o parto* e *Q1C8* - *Tenho medo*, com a mediana na escala 4 - "De acordo" e *IQR* de 2, enquanto que a variável *Q1C9* - *Já tive um parto vaginal*, concentrada na mediana do mínimo da escala, 1 - "Totalmente desacordo", com um *IQR* de 1, como é apresentado na figura 2.7. Relativamente às 4 primeiras variáveis com a mesma mediana e *IQR*, originaram algumas pistas pela preferência pelo tipo de parto via cesariana (i.e., medo do parto), quanto à *Q1C9* possivelmente poderá estar relacionada com a situação anteriormente descrita, que as mulheres grávidas que já tiveram pelo menos 1 parto cesariana, preferem ter o novo parto pela mesma via. Igualmente quando o grupo *G1* foi confrontado com a questão *Q_C_1* - *Optaria por anestesia*, 57% admitem anestesia epidural, 15% anestesia geral, contudo 19% ainda não sabem.

Figura 2.7: G1 - Porque preferiria um parto cesariana?



Sumariamente, as grávidas no *G0* em relação ao *G1*, são estatisticamente mais novas e com um maior período de gestação, tendo um nível de escolaridade maioritariamente superior. Na generalidade mostra uma tendência pela obtenção de informação junto do médico de família/Obstetra para a decisão na preferência pelo tipo de parto, contudo quando analisadas as razões individuais para a preferência sobre o tipo de parto é dada muita informação vaga. Por um lado temos a recuperação pós-parto ser mais rápida, o tempo de internamento ser menor e o desejo de assistir ao nascimento do bebé. Por outro lado temos destacado o medo do parto, a conveniência de agendamento e o possível sofrimento do bebé ser menor.

Sendo este capítulo apenas uma revisão da análise univariada de um trabalho já anteriormente feito [10], urge a necessidade de procurar uma informação mais assertiva e possivelmente mais conclusiva através de técnicas multivariadas como a análise de componentes principais (*ACP*), que terá como aplicação, a redução da dimensionalidade, nas questões de escala de Likert em componentes principais. Posteriormente através da regressão logística pretende-se identificar fatores preditivos na preferência pelo tipo de parto via cesariana.

CAPÍTULO 3

ANÁLISE DE COMPONENTES PRINCIPAIS

A aplicação da análise de componentes principais, nas questões de escala de Likert do inquérito do Anexo I. A *Q1B1* à *Q1B10* para o *G0* e *Q1C1* à *Q1C10* para o *G1*, tem como objetivo, reduzir a complexidade da informação através da extração de componentes onde sintetizem a informação das questões anteriores por grupo.

3.1 Enquadramento

A Análise de Componentes Principais (*ACP*) é uma técnica estatística multivariada que transforma linearmente um conjunto de variáveis originais num pequeno conjunto de componentes, que por sua vez representarão as variâncias das variáveis originais [46].

Desta forma com um conjunto reduzido de componentes será mais fácil a sua interpretação e a sua utilização noutras técnicas estatísticas (análise de variâncias, regressão múltipla/logística), em vez de utilizar todas as variáveis originais [11]. Por conseguinte, a *ACP*, no geral tem como objetivos: (1) a redução de dados e (2) interpretação das componentes [22].

A ideia original da *ACP* foi concretizada por Pearson no início do século XX e desenvolvida independentemente por Hotelling [12]. Esta técnica é usada de forma transversal em todas as formas de análise de dados, desde a neurociência até à computação gráfica, devido à sua simplicidade aliada à sua metodologia não-paramétrica de extrair informação relevante de um conjunto de dados opacos [48].

Existe alguma ambiguidade na literatura sobre se a *ACP* está ou não englobada numa técnica de análise estatística multivariada [17], dado que a *ACP* é um método da análise fatorial exploratória (onde também está associada outra técnica estatística multivariada, a análise fatorial comum, que normalmente também é designada por análise fatorial *AF* [49]). No entanto é justificado que são equivalentes nos seus objetivos, ou seja, a *ACP* e a *AF* são úteis em reduzir variáveis originais para apenas algumas componentes, mas diferentes na sua definição matemática, isto é, na variância que está a ser analisada [49]. Enquanto que a *ACP* analisa toda a variância, a *AF* apenas analisa a variância partilhada, excluindo a variância única e de erro.

Por fim são apresentadas pelo menos duas preocupações. Em primeiro lugar, esta técnica estatística multivariada ainda não têm critérios científicos universais para testar a so-

lução dada, ao passo que por exemplo na análise discriminante, na regressão logística e na análise de variância multivariada, a solução é avaliada no quão bem prevê a associação de novos dados ao grupo correto. Em segundo, o próprio processo metodológico, isto é, após a extração das componentes, existem várias possibilidades de rotação destas, sendo todas contabilizadas para a mesma quantidade de variância das variáveis originais, mas com definições ligeiramente diferentes entre as rotações, conseqüentemente a escolha final entre todas as alternativas depende da avaliação do investigador na sua interpretabilidade e utilidade científica [49].

3.2 Metodologia

3.2.1 Preparação dos dados para a Análise de Componentes Principais

Para que a aplicabilidade da *ACP* seja frutífera, é necessário que haja alguma relação simétrica entre os dados, de forma a ser possível gerar componentes. É através de medidas empíricas que é averiguado a fatorabilidade da matriz de correlações (\mathbf{R} , com dimensão $(n \times n)$), ou matriz de covariância ($\mathbf{\Sigma}$ com dimensão $(n \times n)$).

Caso $\mathbf{R}_{(n \times n)}$ apresente no geral valores pequenos (< 0.30) ou se os dados apresentarem correlações parciais muito elevadas (> 0.70), é uma indicação que os dados não são os melhores para aplicar a *ACP* [17]. Também é necessário verificar o determinante de $\mathbf{R}_{(n \times n)}$, ($|\mathbf{R}_{(n \times n)}|$) sendo usado para determinar o inverso ($\mathbf{R}_{(n \times n)}^{-1} = 1/|\mathbf{R}_{(n \times n)}| \implies |\mathbf{R}_{(n \times n)}| \neq 0$), que por sua vez a matriz $\mathbf{R}_{(n \times n)}^{-1}$ é usada no cálculo dos valores próprios (λ_i), contudo caso $|\mathbf{R}_{(n \times n)}| = 0$, será uma matriz singular, não sendo possível obter resultados através da *ACP* [46].

Doravante, caso $\mathbf{R}_{(n \times n)}$ apresente poucas correlações ≥ 0.30 , então existe a possibilidade de aferir a adequação da *ACP* de duas formas [44]:

1. Teste de esfericidade de Bartlett

$$H_0 : \mathbf{R}_{(n \times n)} = \lambda \mathbf{I}_{(n \times n)} \quad \text{ou} \quad H_0 : \lambda_1 = \lambda_2 = \dots = \lambda_m$$

$$- \left[n - 1 - \frac{1}{6}(2m + 5) \right] \ln |\mathbf{R}_{(n \times n)}| \sim \chi^2 \left(\frac{1}{2} m(m - 1) \right) \quad m = 1, 2, \dots, n \quad (3.1)$$

Fornecer uma medida sumária para testar a presença de correlações entre as variáveis, se o teste for significativo ($p < 0.05$), indica que existem correlações suficientes entre as variáveis para proceder à *ACP*, todavia, este teste é sensível ao tamanho da amostra [44].

2. Kaiser-Meyer-Olkin (*KMO*) e Medida de Adequação da Amostra (*MSA*)

$$KMO = \frac{\sum \sum r_{ij}^2}{\sum \sum r_{ij}^2 + \sum \sum b_{ij}^2} \quad i \neq j \quad (3.2)$$

$$MSA_j = \frac{\sum r_i^2}{\sum r_i^2 + \sum b_i^2} \quad i \neq j \quad (3.3)$$

O *KMO* quantifica o grau de relação entre as variáveis e adequação à *ACP*, sendo r_{ij} , a matriz de correlação bivariada e b_{ij} , a matriz de correlação anti-imagem. É pretendido que os coeficientes b_{ij} estejam perto de zero, sendo estes uma estimativa das correlações entre as componentes, assim será pressuposto que as componentes são ortogonais [44]. A classificação do resultado do *KMO* é a seguinte [44]:

<i>KMO</i>	<i>ACP</i>
0.9 - 1	Muito Bom
0.8 - 0.9	Bom
0.7 - 0.8	Médio
0.6 - 0.7	Razoável
0.5 - 0.6	Mau
< 0.5	Inaceitável

Através da *MSA*, que mede a proporção da variância comum entre as variáveis, identifica possíveis variáveis a retirar, estipulando o limite mínimo de ≥ 0.5 ($\sum r_i^2 \geq \sum b_i^2$), de forma a preparar uma base para a *ACP* [17].

3.2.2 Extração das componentes

Após validação dos dados, segue a questão de qual é o ótimo número de componentes a extrair através da *ACP*, porque é de interesse produzir as melhores combinações

De modo que não seja uma solução nula e tenha uma solução para \mathbf{a}'_1 então é necessário que $|\mathbf{\Sigma} - \lambda \mathbf{I}| = 0$, seja uma matriz singular se e só se λ for um valor próprio de $\mathbf{\Sigma}$ [44]. Desde logo, vemos que [44]:

$$Var(Y_1) = Var(\mathbf{a}'_1 \mathbf{X}) = \mathbf{a}'_1 \mathbf{\Sigma} \mathbf{a}_1 \quad (3.10)$$

Resolvendo a equação (3.9) a $\mathbf{\Sigma}$, fica [44]:

$$\mathbf{a}'_1 (\lambda \mathbf{I} \mathbf{a}_1) = \lambda (\mathbf{a}'_1 \mathbf{I} \mathbf{a}_1) = \lambda (\mathbf{a}'_1 \mathbf{a}_1) = \lambda_1 \quad (3.11)$$

Por fim, $Y_1 = \mathbf{a}'_1 \mathbf{X}$, terá o coeficiente \mathbf{a}_1 correspondente ao vetor próprio \mathbf{e}_1 associados ao maior valor próprio (λ_1) da matriz covariância $\mathbf{\Sigma}$ [44].

A segunda componente principal será ortogonal à primeira que reterá toda e a possível variação restante de forma a produzir a segunda melhor combinação linear, ou seja [44],

- Y_2 , é a combinação linear $\mathbf{a}'_2 \mathbf{X}$ que maximiza $Var(Y_2) = Var(\mathbf{a}'_2 \mathbf{X})$ sujeito à restrição $\mathbf{a}'_2 \mathbf{a}_2 = 1$ e $\mathbf{a}'_2 \mathbf{a}_1 = 0$.

Assim, $Y_2 = \mathbf{a}'_2 \mathbf{X}$, é encontrada a partir do coeficiente \mathbf{a}_2 que corresponde ao vetor próprio \mathbf{e}_2 , associado ao segundo valor próprio mais elevado (λ_2) da matriz covariância $\mathbf{\Sigma}$ [22]. Contudo Y_1 e Y_2 não podem estar correlacionadas, ou seja, como referido Y_2 está sujeito à restrição $\mathbf{a}'_2 \mathbf{a}_1 = 0$ [44], isto é,

$$Cov(Y_1, Y_2) = Cov(\mathbf{a}'_1 \mathbf{X}, \mathbf{a}'_2 \mathbf{X}) = E [\mathbf{a}'_2 (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \mathbf{a}_1] = \mathbf{a}'_2 \mathbf{\Sigma} \mathbf{a}_1 = 0 \quad (3.12)$$

Sabendo pela equação (3.9), $\mathbf{\Sigma} \mathbf{a}_1 = \lambda \mathbf{a}_1$, então,

$$\mathbf{a}'_2 \lambda \mathbf{a}_1 = 0 \Leftrightarrow \mathbf{a}'_2 \mathbf{a}_1 = 0 \quad (3.13)$$

portanto \mathbf{a}_2 e \mathbf{a}_1 deverão ser ortogonais.

O processo continua sendo a retenção da variação dos dados cada vez menor até no máximo ser extraído tantas p componentes quantas q variáveis usadas na *ACP* [17].

No geral, seja a matriz de covariância $\mathbf{\Sigma}$ uma matriz simétrica, positiva definitiva associado a um vetor $\mathbf{X}' = [X_1, X_2, \dots, X_q]$, ter pares de valores próprios e vetores próprios $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p)$, onde $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ [22] deste modo,

- Y_p , é a p -ésima combinação linear $\mathbf{e}'_p \mathbf{X}$ que maximiza $Var(Y_p) = Var(\mathbf{e}'_p \mathbf{X})$ sujeito à restrição $\mathbf{e}'_p \mathbf{e}_p = 1$ e $\mathbf{e}'_p \mathbf{e}_k = 0, \forall k < p$.

$$\text{Var}(Y_p) = \mathbf{e}_p' \boldsymbol{\Sigma} \mathbf{e}_p = \lambda_p \quad (3.14)$$

$$\text{Cov}(Y_p, Y_k) = \mathbf{e}_p' \boldsymbol{\Sigma} \mathbf{e}_k = 0 \quad (3.15)$$

As p componentes principais ortogonais têm as variâncias iguais aos valores próprios de $\boldsymbol{\Sigma}$. Sendo $\mathbf{e} = [e_1, e_2, \dots, e_p]$ a matriz dos vetores próprios e \mathbf{Y} o vetor das componentes principais [44] então,

$$\mathbf{Y} = \mathbf{e}' \mathbf{X} \quad (3.16)$$

A matriz de variância-covariância de \mathbf{Y} será [44],

$$\text{Var}(\mathbf{Y}) = \mathbf{e}' \boldsymbol{\Sigma} \mathbf{e} = \boldsymbol{\Lambda} \quad (3.17)$$

Sendo,

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \quad (3.18)$$

Sabendo que \mathbf{e} é uma matriz ortogonal tal que $\mathbf{e}' \mathbf{e} = \mathbf{e} \mathbf{e}' = \mathbf{I}$. Então temos [44],

$$\sum_{j=1}^p \text{Var}(Y_j) = \sum_{j=1}^p \lambda_j = \text{tr}(\boldsymbol{\Lambda}) \quad (3.19)$$

mas,

$$\text{tr}(\boldsymbol{\Lambda}) = \text{tr}(\mathbf{e}' \boldsymbol{\Sigma} \mathbf{e}) = \text{tr}(\boldsymbol{\Sigma} \mathbf{e}' \mathbf{e}) = \text{tr}(\boldsymbol{\Sigma}) = \sum_{j=1}^p \text{Var}(X_j) \quad (3.20)$$

Logo, no máximo, o total da variação das variáveis originais pode ser atribuída às primeiras p componentes principais, conseqüentemente estas componentes principais podem representar as q variáveis originais usados para *ACP* sem perda de muita informação. A j -ésima componente principal explica em proporção da variação total original [44]:

$$\frac{\lambda_j}{\sum_{j=1}^p \lambda_j} = \frac{\lambda_j}{\text{tr}(\boldsymbol{\Lambda})} \quad j = 1, 2, \dots, p \quad (3.21)$$

As primeiras m componentes principais explicam em proporção da variação total [44]:

$$\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^p \lambda_j} = \frac{\sum_{j=1}^m \lambda_j}{\text{tr}(\boldsymbol{\Lambda})} \quad m < p \quad j = 1, 2, \dots, m, p \quad (3.22)$$

Visto que as primeiras m componentes principais irão reter os valores próprios maiores, os restantes valores próprios ficaram muito próximo de zero e a sua contribuição para a explicar a variância restante terá pouco peso [44]. Assim retirar as componentes principais que pouco contribuem originará pouca perda de informação, que por sua vez permite uma redução da dimensionalidade dos dados, tornando os resultados mais simples e de melhor interpretação [44].

Para tal os critérios empíricos para otimizar esta extração, poderão ser [44]:

1. *Valores Próprios*, sendo 1 ou acima é dado como componente significativa, porém se o número de variáveis for menor que 20 a tendência deste é extrair um número conservador de componentes, também foi proposto com base num estudo de várias simulações que um melhor procedimento seria excluir componentes extraídas da matriz de correlações para as quais o seus *valores próprios* fossem < 0.70 [12];
2. *Percentagem da Variância*, quando a informação é menos precisa, é comum escolher todas as componentes que na sua variância cumulativa percentual representam pelo menos 70% [44];
3. *Gráfico de Escarpa*, é usado para identificar visualmente o número ótimo de componentes que poderão ser extraídas antes do total da variância única dominar a estrutura da variância comum, normalmente mostra entre 1 a 3 componentes a mais para consideração de inclusão do que pelo critério dos *Valores Próprios*, mais ainda os resultados são mais óbvios (e confiáveis), quando o tamanho da amostra é maior, nestas situações as comunalidades apresentam valores mais elevados e cada componente tem várias variáveis com *loadings*(pesos) consistentes [17] (será explicado na próxima secção) .

Deste modo, diversas soluções com diferentes números de componentes deve ser examinado antes de definir a estrutura final das variáveis. Caso o número de variáveis originais for relativamente reduzido ($k \leq 30$) ou o número de observações for elevado ($n > 250$) e ainda quando as comunalidades são pelo menos 0.60, tanto o primeiro critério (*Valores Próprios*), como o terceiro (*Gráfico de Escarpa*), resultam em soluções fiáveis quanto ao verdadeiro número de componentes [44].

Loadings(Pesos) e correlações entre variáveis e componentes

O resultado da *ACP* mostra os transformados dos vetores próprios, isto é,

$$\mathbf{e}_j^* = \lambda_j^{\frac{1}{2}} \mathbf{e}_j \quad j = 1, 2, \dots, p \quad (3.23)$$

Sendo que a soma dos quadrados dos seus transformados são iguais ao correspondente valor próprio [44], ou seja,

$$(\mathbf{e}_j^*)' \mathbf{e}_j^* = (\lambda_j^{\frac{1}{2}} \mathbf{e}_j)' (\lambda_j^{\frac{1}{2}} \mathbf{e}_j) = \lambda_j \mathbf{e}_j' \mathbf{e}_j = \lambda_j \quad (3.24)$$

Por seu lado, os elementos das componentes com valores próprios maiores são ponderados de forma significativa. Desta maneira seja $\mathbf{e}^* = [\mathbf{e}_1^*, \mathbf{e}_2^*, \dots, \mathbf{e}_p^*]$ um vetor de transformados de vetores próprios, e a matriz dos vetores próprios e \mathbf{r} a matriz de correlações, então [44]

$$\mathbf{e}^* = \mathbf{e} \mathbf{\Lambda}^{\frac{1}{2}} \quad \mathbf{r} = \mathbf{e}^* (\mathbf{e}^*)' \quad (3.25)$$

Assim, as componentes principais consideram corretamente a proporção de variância explicada pelos dados correspondentes. Os vetores \mathbf{e}_i^* , podem ser compreendidos como sendo os *pesos* das variáveis originais nas componentes principais respectivas, como também medem as correlações entre as componentes principais e as variáveis originais, ou seja, é possível extrair alguma interpretação das componentes principais em função das variáveis que estão correlacionadas [44].

Por fim, a soma dos quadrados dos *pesos* das variáveis para cada componente principal, resulta no valor próprio da componente principal respetiva [44]

$$\lambda_j = \sum_{i=1}^p (e_{ij}^*)^2 \quad (3.26)$$

Enquanto que a soma dos quadrados dos *pesos* das componentes principais para cada variável, resulta na proporção da variação de cada variável explicada pelas componentes principais retidas, as comunalidades [44]

$$h_j = \sum_{i=1}^m (e_{ij}^*)^2 \quad (m \leq p) \quad (3.27)$$

3.2.3 Rotação e interpretação das componentes

Ao conceber empiricamente o número ótimo de componentes extraídas apresentada pela matriz de componentes não rotacionada, os *pesos* apresentados, nos quais apresentam o grau de correspondência entre a variável e a componente, como anteriormente

referido, poderá dar-nos uma interpretação de cada componente, contudo na maioria das vezes não sendo simples esta interpretação é necessário utilizar o método rotacional sobre a matriz referida de forma aproximar a uma estrutura simples [50] e de fácil leitura [17]. Seja, $\mathbf{e}^*_{(n \times p)}$ a matriz de vetores e^*_{ij} antes da rotação, $\mathbf{C}_{(n \times p)}$ a matriz de vetores c^*_{ij} depois da rotação e $\mathbf{G}_{(p \times p)}$, uma matriz ortogonal [44],

$$\mathbf{e}^* = [\mathbf{e}_1^*, \mathbf{e}_2^*, \dots, \mathbf{e}_p^*] \quad \mathbf{C} = [\mathbf{c}_1^*, \mathbf{c}_2^*, \dots, \mathbf{c}_p^*] \quad \mathbf{G} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad (3.28)$$

onde p é o número de componentes principais retidos e

$$\mathbf{C} = \mathbf{e}^* \mathbf{G} \quad (3.29)$$

Para que \mathbf{C} tenha uma estrutura simplificada será necessário as seguintes condições [44]:

- (i) A matriz \mathbf{C} deverá conter pelo menos um zero em cada linha, isto é, cada variável deverá não estar correlacionada com pelo menos uma componente principal depois da rotação [44];
- (ii) A matriz \mathbf{C} deverá conter pelo menos p zeros em cada coluna [44];
- (iii) Na matriz \mathbf{C} em cada par de colunas as variáveis com coeficientes nulos numa coluna não deverão ter na outra coluna coeficientes nulos [44].

A rotação serve para maximizar as correlações mais fortes entre as componentes e as variáveis e minimizar as mais fracas [49], também é descrita com a intenção de redistribuir a variância dos primeiros componentes para os últimos para conseguir uma estrutura simples e teoricamente com um padrão mais significativo [17]. O problema prende-se com o fato de encontrar uma matriz ortogonal \mathbf{G} tal que seja maximizada a seguinte equação,

$$Q = \sum_{j=1}^p \left[\sum_{i=1}^n c_{ij}^4 - \frac{m}{n} \left(\sum_{i=1}^n c_{ij}^2 \right)^2 \right] \quad (3.30)$$

sendo m uma constante que varia conforme o método rotacional usado e c_{ij} os elementos da matriz \mathbf{C} [44].

Existem vários métodos rotacionais englobados em duas categorias:

- (a) *Ortogonalis*, onde destacam-se VARIMAX, QUARTIMAX e EQUIMAX;
- (b) *Oblíquos*, onde destacam-se OBLIMIN, PROMAX, ORTHOBLIQUE.

Contudo o método rotacional mais popular o VARIMAX, foi proposto por Kaiser [44], onde $m = 1$ na equação 3.30, que na sua essência é procedimento de maximização da variância, o objetivo deste método é maximizar a variância dos *pesos* das componentes principais, pretendendo desta forma que para cada componente principal existam alguns *pesos* significativos e todos os restantes fiquem próximos de zero [49].

Interpretação dos *pesos* da matriz rotacionada

Obtida a matriz de componentes rotacionada, é analisado a significância dos *pesos*, visto anteriormente que, o quadrado destes é o total de variância da i -ésima variável que é contabilizada para cada j -ésima componente, $(e_{ij}^*)^2$, logo poderemos avaliar os *pesos* através de um critério de significância prática, desta modo, a classificação sugerida [6]

<i>pesos</i> (variância sobreposta)	Classificação
> 0.71(50%)	Excelente
0.63 - 0.70(40%)	Muito Bom
0.55 - 0.62(30%)	Bom
0.32 - 0.54(20%)	Suficiente
< 0.32(10%)	Fraco

De notar que o valor dos *pesos* é influenciado pela homogeneidade dos valores na amostra, se for suspeito de haver homogeneidade, a interpretação de *pesos* fracos é garantida [49].

Interpretação das componentes e variáveis da matriz rotacionada

Tendo os *pesos* significativos para cada componente principal, são averiguadas as seguintes condições:

- (i) Se existir alguma variável que seja significativa para mais do que uma componente, apelidado de *cross-loading*, torna-se passível de ser retirada [17];

(ii) Se existir alguma comunalidade abaixo de um mínimo empírico 0.50 então essa variável poderá estar a contribuir muito pouco para a explicação das componentes, sendo passível de ser retirada [17];

(iii) Se existir alguma variável sem *pesos* significativos, torna-se candidata a ser retirada [17].

Caso surja alguma situação das referidas é efetuada uma nova aplicação da *ACP*, sendo que, as medidas a tomar poderão ser: (i) retirar a(s) variáv(él)(éis) problemática(s), (ii) efetuar um método rotacional alternativo, (iii) diminuir/aumentar o número de componentes retidas [17].

Por último, após obter-se uma solução de componentes com uma estrutura simples, são criadas legendas para as componentes principais. Elaborado intuitivamente a legenda para cada componente, com base nas legendas das variáveis que compõem a componente principal, por uma ordenação decrescente a partir do *peso* mais significativo [17].

3.3 Implementação e Resultados

A aplicação da *ACP* nas questões de escala de likert (*Q1B1* a *Q1B10* e *Q1C1* a *Q1C10*), estando agrupadas no *G0* e no *G1* respetivamente, apresentadas no questionário no Anexo I, tem como objetivo neste trabalho, sumarizar a informação, identificar características/domínios e originar componentes principais para uma possível futura utilização dos dados em outras técnicas estatísticas multivariadas.

A apresentação dos resultados da aplicação da *ACP*, será feito por grupo, seguindo a metodologia já referida, e onde algumas tabelas e figuras auxiliares estarão presentes no Anexo IV. Foram apresentados os resultados iniciais do trabalho, no *XXIV Congresso da Sociedade Portuguesa de Estatística*, intitulado: **Preferência pelo tipo de parto em Portugal: análise em componentes principais**. No entanto houve uma evolução da implementação que colmatou em novos resultados.

3.3.1 G0 - Preferência pelo Parto Vaginal

Tamanho da Amostra

De acordo com o a tabela 3.1 no Anexo IV, como já havia sido mencionado existem no total 190 casos por item, todavia existem alguns dados omissos por item, no mínimo de 8.95% e no máximo de 41.58%. Sendo que os 3 itens por ordem decrescente que contabilizam mais dados omissos, são: *Q1B10*, *Q1B5* e *Q1B7*, como mostra a tabela 3.2 no Anexo IV. De notar que mesmo com estes dados omissos é mantido e excedido o rácio de casos (5:1) para cada item recomendado [17].

Por último, existe possibilidade de estes dados omissos serem respostas que as grávidas inquiridas não pretenderam responder, deste modo, não foi efetuado qualquer tratamento, pelo simples fato de poderem ser obtidos 5 resultados que podem não espelhar a realidade dos dados iniciais.

Preparação dos dados para ACP

Como mostra a tabela 3.3 as correlações entre todos os itens do *G0* são relativamente fracas (< 0.30), verificando-se apenas a existência de 2 correlações significativas. A tabela 3.4 mostra correlações parciais baixas (< 0.70) sendo um sinal positivo de alguma relação entre os itens. Através do teste de esfericidade de Bartlett ($H_0 : \mathbf{R}_{(n \times n)} = \lambda \mathbf{I}_{(n \times n)}$, $\chi^2(190) = 365.09$, $p < 0.01$), existe evidência estatística para rejeitar a H_0 , por outro lado, o *KMO*, apresenta um valor de 0.574, sendo classificado de mau. O nível de confiabilidade de α de Cronbach é de 0.59 (pobre) [13] e o determinante da matriz de correlações entre os itens do *G0*, tem um valor positivo de 0.139, mostrando que não se trata de uma matriz singular.

A *MSA* representada na diagonal da tabela 3.4, identifica 2 itens abaixo do nível 0.50, *Q1B5* e o *Q1B10*. Como o *Q1B5* tem o valor menor entre os 2 itens será o primeiro a ser retirado e é novamente recalculada a *MSA*. Por sua vez o item *Q1B10* gerou um valor de 0.43 (< 0.50), novamente é recalculada a *MSA* sem os 2 itens mencionados.

Posto isto, pela tabela 3.5 é verificado a inexistência de correlações significativas contudo pela tabela 3.6 as correlações parciais tornam a mostrar valores baixos (< 0.70). Pelo

teste de esfericidade de Bartlett ($H_0 : \mathbf{R}_{(n \times n)} = \lambda \mathbf{I}_{(n \times n)}$, $\chi^2(190) = 273.90$, $p < 0.01$), existe evidência estatística para rejeitar a H_0 , quanto ao *KMO* originou um valor de 0.72, melhorando a classificação da amostra para médio. O nível de confiabilidade α de Cronbach é de 0.72, sendo aceitável e o determinante da matriz de correlações entre os itens do *G0*, registra um valor positivo de 0.228.

Analisando a *MSA* na diagonal da tabela 3.6 mostra que os dados estão preparados para a extração das componentes principais (≥ 0.50), após a retirada da *Q1B5* e da *Q1B10*.

Tabela 3.3: Matriz de correlação do G0

	<i>Q1B1</i>	<i>Q1B2</i>	<i>Q1B3</i>	<i>Q1B4</i>	<i>Q1B5</i>	<i>Q1B6</i>	<i>Q1B7</i>	<i>Q1B8</i>	<i>Q1B9</i>	<i>Q1B10</i>	Correlações ($p < 0.05$)
<i>Q1B1</i>	1										0
<i>Q1B2</i>	0.44	1									0
<i>Q1B3</i>	0.17	0.08	1								0
<i>Q1B4</i>	0.04	0.29	0.19	1							0
<i>Q1B5</i>	0.22	0.06	-0.05	-0.08	1						0
<i>Q1B6</i>	0.30	0.24	0.25	0.31	0.24	1					0
<i>Q1B7</i>	0.28	0.08	0.11	0.19	-0.15	0.29	1				0
<i>Q1B8</i>	0.42	0.05	0.14	0.01	-0.13	0.01	0.28	1			0
<i>Q1B9</i>	0.51	0.40	0.03	0.06	0.20	0.15	0.24	0.31	1		1
<i>Q1B10</i>	-0.09	0.02	0.05	-0.07	-0.05	0.04	-0.05	-0.36	0.04	1	1

Nota: valores a negrito indica correlação significativa

Teste esfericidade de Bartlett, $\chi^2(190) = 365.09$, $p < 0.01$

KMO: 0.574

α de Cronbach: 0.59

$\det(G0) = 0.139$

Tabela 3.4: Matriz de correlação anti-imagem do G0

	<i>Q1B1</i>	<i>Q1B2</i>	<i>Q1B3</i>	<i>Q1B4</i>	<i>Q1B5</i>	<i>Q1B6</i>	<i>Q1B7</i>	<i>Q1B8</i>	<i>Q1B9</i>	<i>Q1B10</i>
<i>Q1B1</i>	0.67									
<i>Q1B2</i>	-0.37	0.58								
<i>Q1B3</i>	-0.11	0.03	0.62							
<i>Q1B4</i>	0.17	-0.28	-0.14	0.55						
<i>Q1B5</i>	-0.22	0.15	0.07	0.1	0.34					
<i>Q1B6</i>	-0.15	-0.08	-0.19	-0.25	-0.28	0.62				
<i>Q1B7</i>	-0.15	0.15	0.04	-0.12	0.26	-0.27	0.61			
<i>Q1B8</i>	-0.35	0.19	-0.12	0.03	0.26	0.08	-0.11	0.52		
<i>Q1B9</i>	-0.22	-0.28	0.08	-0.01	-0.23	0.08	-0.16	-0.24	0.67	
<i>Q1B10</i>	-0.01	0.03	-0.11	0.12	0.17	-0.07	0.01	0.4	-0.19	0.37

Nota: *MSA* está na diagonal, as correlações parciais abaixo da diagonal

Tabela 3.5: Matriz de correlação do G0 (sem Q1B5 e Q1B10)

	<i>Q1B1</i>	<i>Q1B2</i>	<i>Q1B3</i>	<i>Q1B4</i>	<i>Q1B6</i>	<i>Q1B7</i>	<i>Q1B8</i>	<i>Q1B9</i>	Correlações Significativas ao nível .05
<i>Q1B1</i>	1								0
<i>Q1B2</i>	0.42	1							0
<i>Q1B3</i>	0.17	0.13	1						0
<i>Q1B4</i>	0.17	0.31	0.22	1					0
<i>Q1B6</i>	0.27	0.30	0.28	0.33	1				0
<i>Q1B7</i>	0.17	0.11	0.25	0.20	0.30	1			0
<i>Q1B8</i>	0.41	0.10	0.10	0.09	0.06	0.23	1		0
<i>Q1B9</i>	0.5	0.48	0.11	0.15	0.24	0.24	0.36	1	0

Teste esfericidade de Bartlett, $\chi^2(190) = 273.90, p < 0.01$

KMO: 0.729

α de Cronbach: 0.72

$\det(G0) = 0.228$

Tabela 3.6: Matriz de correlação anti-imagem do G0 (sem Q1B5 e Q1B10)

	<i>Q1B1</i>	<i>Q1B2</i>	<i>Q1B3</i>	<i>Q1B4</i>	<i>Q1B6</i>	<i>Q1B7</i>	<i>Q1B8</i>	<i>Q1B9</i>
<i>Q1B1</i>	0.75							
<i>Q1B2</i>	-0.24	0.70						
<i>Q1B3</i>	-0.07	0	0.78					
<i>Q1B4</i>	0.01	-0.23	-0.11	0.75				
<i>Q1B6</i>	-0.13	-0.11	-0.16	-0.2	0.76			
<i>Q1B7</i>	0.04	0.07	-0.16	-0.09	-0.21	0.74		
<i>Q1B8</i>	-0.32	0.15	-0.02	-0.05	0.11	-0.16	0.64	
<i>Q1B9</i>	-0.26	-0.36	0.03	0.06	-0.05	-0.14	-0.22	0.74

Nota: *MSA* está na diagonal, as correlações parciais abaixo da diagonal

Extração das componentes e interpretação dos pesos e correlações entre variáveis e componentes

Existem vários critérios para efetuar a extração das componentes já anteriormente referidos, sendo que serão usados parcimoniosamente na obtenção dos componentes. Como mostra a tabela 3.7 foram usados todos os itens de forma a gerar o mesmo número de componentes, são colocados pela ordem decrescente de variância (*valor próprio*, λ_i), tal que, a primeira componente principal contabiliza o máximo que é possível de variância dos dados originais, 34.11%, a segunda componente principal é escolhida para contabilizar o máximo que é possível do restante da variância 15.78%, que não esteja correlacionada com a primeira componente e assim por diante sendo este um processo iterativo, até ter no máximo tantas componentes principais como variáveis usadas para a *ACP* [12].

Aplicando o critério do *valor próprio*, onde são retidos todas as componentes com valor ≥ 1.0 são obtidos 3 componentes, que é unânime tanto pelo critério *percentagem da variância*, em que a variância acumulativa ≥ 0.60 (62.88%), como pelo critério de *Gráfico de escarpa*, (figura 3.1), no qual são visualmente escolhidas todas as componentes até ao ponto em que o declive fica menos acentuado.

Examinando a tabela 3.8, que contém a matriz não rotacionada das componentes extraídas, temos os *pesos* de cada item para cada componente, isto é, a correlação entre estes, no qual já é verificado alguns itens com correlações fortes para as componentes extraídas, na última coluna, as comunalidades(*h*), sendo passível de usar como um índice de quanta variância de um item é usada na solução, deste modo os itens *Q1B3* ou *Q1B4* tem muito menos em comum com a solução do que o item *Q1B2*, mesmo assim, as comunalidades estão acima do nível normalmente recomendado de 0.50. Na parte inferior da tabela 3.8 estão os *valores próprios* para cada componente e a percentagem do traço, que é representativo do total da variância extraída para cada componente, acumulando 62.89% da variância total.

Tabela 3.7: Extração das Componentes Principais do G0

Componentes	Valores Próprios		
	Total	% Variância	% Acumulada
1	2.72	34.11	34.11
2	1.26	15.78	49.90
3	1.03	12.97	62.88
4	0.76	9.44	72.33
5	0.73	9.12	81.45
6	0.61	7.67	89.13
7	0.46	5.74	94.88
8	0.41	5.12	100

Figura 3.1: Gráfico de esarpa das Componentes Principais do G0

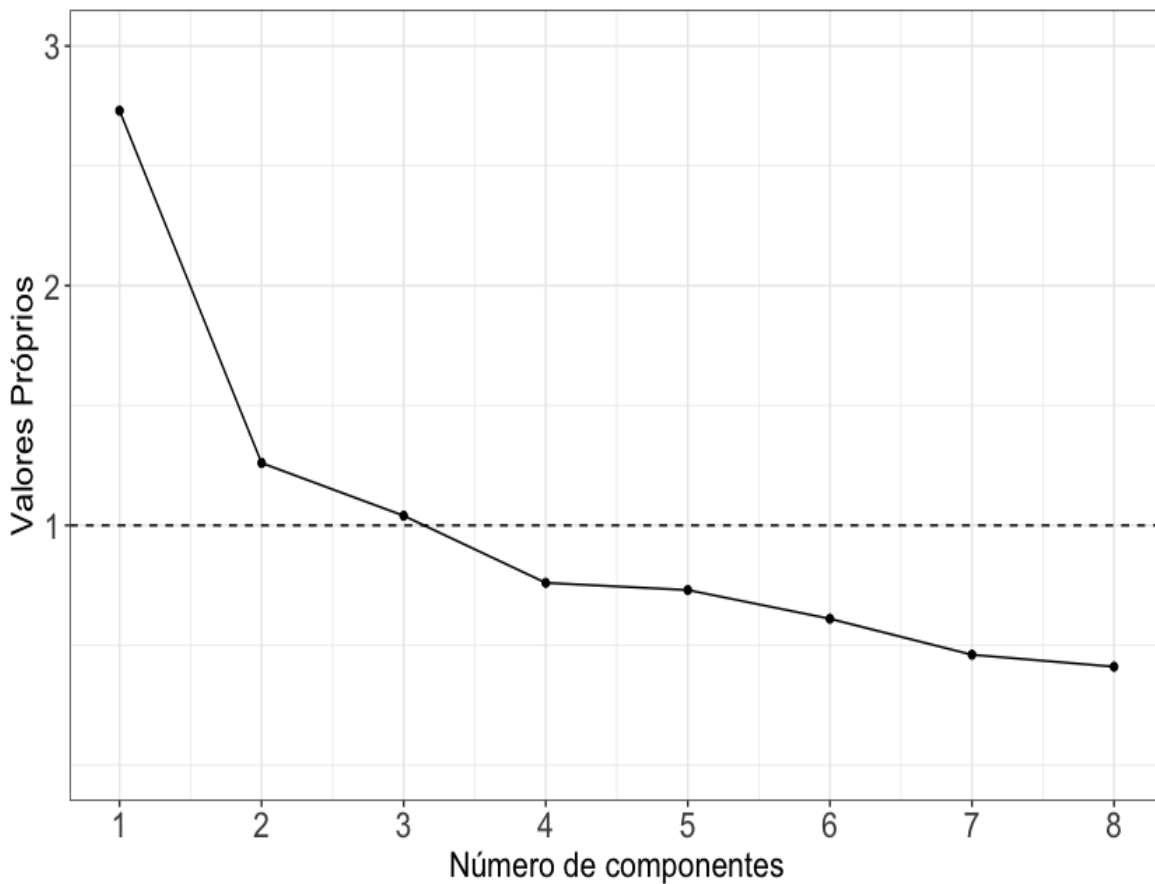


Tabela 3.8: Matriz não rotacionada das Componentes Principais do G0

Items	Componentes			h^*
	1	2	3	
<i>Q1B1</i>	0.72	-0.38	-0.04	0.66
<i>Q1B2</i>	0.66	-0.09	-0.56	0.75
<i>Q1B3</i>	0.43	0.49	0.30	0.52
<i>Q1B4</i>	0.50	0.46	-0.24	0.52
<i>Q1B6</i>	0.59	0.45	-0.11	0.57
<i>Q1B7</i>	0.50	0.28	0.54	0.62
<i>Q1B8</i>	0.49	-0.47	0.51	0.73
<i>Q1B9</i>	0.71	-0.40	-0.10	0.68
				Total
Valores Próprios	2.73	1.26	1.04	5.03
% do traço	34.12	15.79	12.98	62.89

* comunalidades

Análise da matriz de componentes rotacionada

Após efetuar a rotação usando o método rotacional VARIMAX, é obtida a matriz apresentada na parte superior da tabela 3.9. Foram omitidos todos os *pesos* < 0.40 , visto que para o tamanho de amostra do G0 o limite mínimo significativo é ≥ 0.40 [17], quanto aos *pesos* remanescentes foram ordenados pela sua significância por ordem decrescente por componente.

Com o método rotacional VARIMAX, foram maximizados os *pesos* que apresentaram uma maior correlação com cada componente (sendo que são considerados bons, acima ≥ 0.55 ou 30% de variância sobreposta [6]), e minimizando os *pesos* com uma correlação mais fraca, mantendo as mesmas comunalidades. Foi redistribuída a variância extraída por cada componente, sendo que no total a variância extraída teve um aumento ténue. Contudo o Q1B2 apresenta um *cross-loading*, ou seja, é correlacionado com 2 componentes em simultâneo, tornando complexo a interpretação destas componentes, desta maneira o item mencionado é retirado.

Na parte inferior da tabela 3.9 foi gerada nova matriz através do método rotacional VARIMAX, omitindo o item problemático Q1B2. Desta forma, mantendo a omissão de *pesos* < 0.40 , são apresentados *pesos* com uma significância excelente (acima de 0.71 ou 50% de variância sobreposta [6]), todos os itens com comunalidades ≥ 0.60 representando assim um unanimidade com a solução gerada. Houve uma pequena redistribuição da variância extraída entre as componentes, como também levou a um pequeno aumento no total da variância extraída. Por fim, é obtido uma estrutura simples no qual cada componente é exclusivamente correlacionada com os seus itens.

A interpretação e a nomenclatura das componentes depende do significado das combinações dos itens observados que exibem uma correlação mais forte com a componente, logo uma componente será mais facilmente interpretada quando vários itens têm uma correlação forte com a componente, e estes não se correlacionarem com outras componentes [49].

Assim sendo temos os seguintes componentes:

- Componente 1: *Vantagens do parto para a Mãe*
 - *Q1B1: Não deixa cicatriz;*
 - *Q1B9: O tempo de internamento hospitalar é menor;*
 - *Q1B8: A recuperação pós-parto é mais rápida.*

- Componente 2: *Preocupação com o bebé*
 - *Q1B4: Não quero que o meu bebé nasça através de uma cirurgia;*
 - *Q1B6: Conheço mais pessoas que tiveram um parto vaginal.*

- Componente 3: *Medo do parto*
 - *Q1B7: Quero ter vários filhos e sei que uma cesariana pode limitar o número;*
 - *Q1B3: Tenho medo.*

Tabela 3.9: Matrizes rotacionadas das Componentes do G0: Conjunto completo e reduzido

Items	Componentes			h^*
	1	2	3	
<i>Q1B9</i>	0.79			0.68
<i>Q1B1</i>	0.78			0.66
<i>Q1B8</i>	0.71			0.73
<i>Q1B2</i>	0.51	0.67		0.75
<i>Q1B4</i>		0.67		0.52
<i>Q1B6</i>		0.63		0.57
<i>Q1B7</i>			0.76	0.62
<i>Q1B3</i>			0.66	0.52
				Total
Valores Próprios	2.05	1.59	1.40	5.04
% do traço	25.63	19.88	17.50	62.93
<i>(Q1B2 foi retirada)</i>	1	2	3	h^*
<i>Q1B1</i>	0.80			0.65
<i>Q1B9</i>	0.79			0.70
<i>Q1B8</i>	0.73			0.67
<i>Q1B4</i>		0.77		0.62
<i>Q1B6</i>		0.72		0.60
<i>Q1B7</i>			0.75	0.61
<i>Q1B3</i>			0.74	0.61
				Total
Valores Próprios	1.85	1.34	1.26	4.45
% do traço	26.43	19.14	18.00	63.57

pesos das componentes com valor inferior a 0.4 foram omitidas e as variáveis foram ordenadas pelo peso decrescente para cada componente.

* comunalidades

3.3.2 G1 - Preferência pelo Parto Cesariana

Tamanho da Amostra

De acordo com o a tabela 3.10 no Anexo IV, como já havia sido mencionado existem no total 54 casos por item, todavia existem alguns dados omissos por item, no mínimo de 5.56% e no máximo de 44.40%. Sendo que os 3 itens por ordem decrescente que contabilizam mais dados omissos, são: Q1C9, Q1C5 e Q1C4, como mostra a tabela 3.11 no Anexo IV. De notar que mesmo com estes dados omissos não é atingindo o rácio de casos para cada item recomendado (5:1) [17]. Todavia, a literatura indica que poderá ser aceitável uma amostra pequena ($n < 100$), desde que haja um nível consistente de comunalidades (≥ 0.60) [27].

Por último, também os dados omissos foram abordados da mesma maneira que no G0, portanto, supondo que os dados omitidos são respostas que as grávidas inquiridas não pretenderam responder, deste modo, não foi efetuado qualquer tratamento, pelo simples fato de poderem ser obtidos resultados que podem não espelhar a realidade dos dados iniciais.

Preparação dos dados para ACP

Como mostra a tabela 3.12 as correlações entre todos os itens do G1, ao contrário do G0, são relativamente fortes, verificando a existência de 12 correlações significativas. Pela tabela 3.13, verificam-se correlações parciais elevadas (> 0.70). Através do teste de esfericidade de Bartlett ($H_0 : \mathbf{R}_{(n \times n)} = \lambda \mathbf{I}_{(n \times n)}$, $\chi^2(54) = 413.73$, $p < 0.01$), existe evidência estatística para rejeitar a H_0 , quanto ao KMO, apresenta um valor de 0.197, sendo classificada a amostra de inaceitável. O nível de confiabilidade de α de Cronbach é de 0.59, classificando de pobre, e o determinante da matriz do G1, tem um valor positivo e próximo de 0, significando que podemos estar perante um caso de matriz singular.

Examinando a MSA representada na diagonal da tabela 3.13, verifica-se que todos os itens estão < 0.50 . Como o Q1C10 tem o valor menor entre todos os itens será o primeiro a ser retirado. Após recálculo, a MSA do item Q1C9 gerou um valor de 0.17 (< 0.50), refazendo o processo sem os 2 itens mencionados, o item Q1C4 apresentou um valor de 0.42 (< 0.50). Refaz-se a análise novamente sem os 3 itens.

Na tabela 3.14, verifica-se a existência de 5 correlações significativas, e pela tabela 3.15 as correlações parciais são fracas (< 0.70). Pelo teste de esfericidade de Bartlett ($H_0 : \mathbf{R}_{(n \times n)} = \lambda \mathbf{I}_{(n \times n)}$, $\chi^2(54) = 135.14$, $p < 0.05$), existe evidência estatística para rejeitar a H_0 , quanto ao *KMO*, com um valor de 0.548, a amostra obtém um classificação de mau. O nível de confiabilidade, α de Cronbach é de 0.65, sendo classificado de questionável, e o determinante da matriz das correlações do *G1*, regista de 0.07. Analisando a *MSA* na diagonal da tabela 3.15, verifica-se que os dados estão preparados para a extração das componentes (≥ 0.50), prossegue-se com análise sem os itens *Q1C10*, *Q1C9* e *Q1C4*.

Tabela 3.12: Matriz de correlação do G1

	<i>Q1C1</i>	<i>Q1C2</i>	<i>Q1C3</i>	<i>Q1C4</i>	<i>Q1C5</i>	<i>Q1C6</i>	<i>Q1C7</i>	<i>Q1C8</i>	<i>Q1C9</i>	<i>Q1C10</i>	Correlações ($p < 0.05$)
<i>Q1C1</i>	1										0
<i>Q1C2</i>	-0.10	1									1
<i>Q1C3</i>	-0.17	0.68	1								2
<i>Q1C4</i>	-0.16	0.19	0.10	1							0
<i>Q1C5</i>	0.26	0.19	0.38	0.30	1						0
<i>Q1C6</i>	0.06	0.58	0.65	0.07	0.50	1					2
<i>Q1C7</i>	-0.06	0.57	0.56	0.32	0.52	0.35	1				3
<i>Q1C8</i>	0.87	-0.09	-0.10	-0.22	0.48	0.08	0.10	1			2
<i>Q1C9</i>	0.14	-0.17	-0.04	-0.35	0.09	-0.20	-0.28	0.40	1		1
<i>Q1C10</i>	-0.03	0.27	0.10	-0.39	-0.29	-0.09	-0.06	-0.12	-0.31	1	1

Nota: valores a negrito indica correlação significativa

Teste esfericidade de Bartlett, $\chi^2(54) = 413.73$, $p < 0.01$

KMO: 0.197

α de Cronbach: 0.59

$\det(G1) = 0.209^{-3}$

Tabela 3.13: Matriz de correlação anti-imagem do G1

	Q1C1	Q1C2	Q1C3	Q1C4	Q1C5	Q1C6	Q1C7	Q1C8	Q1C9	Q1C10
Q1C1	0.22									
Q1C2	-0.53	0.22								
Q1C3	-0.55	0.48	0.30							
Q1C4	0.34	-0.77	-0.45	0.14						
Q1C5	-0.08	0.63	0.23	-0.7	0.35					
Q1C6	0.52	-0.87	-0.72	0.79	-0.67	0.19				
Q1C7	0.7	-0.84	-0.73	0.65	-0.55	0.85	0.19			
Q1C8	-0.95	0.52	0.63	-0.3	-0.01	-0.54	-0.71	0.27		
Q1C9	0.74	-0.81	-0.74	0.74	-0.48	0.87	0.9	-0.77	0.09	
Q1C10	0.56	-0.85	-0.65	0.82	-0.52	0.86	0.82	-0.57	0.89	0.08

Nota: *MSA* está na diagonal, as correlações parciais abaixo da diagonal

Tabela 3.14: Matriz de correlação do G1 (sem o Q1C10, Q1C9 e Q1C4)

	Q1C1	Q1C2	Q1C3	Q1C5	Q1C6	Q1C7	Q1C8	Correlações ($p < 0.05$)
Q1C1	1							0
Q1C2	-0.09	1						1
Q1C3	-0.27	0.34	1					1
Q1C5	0.16	0.13	0.38	1				0
Q1C6	-0.02	0.36	0.46	0.26	1			0
Q1C7	-0.02	0.55	0.42	0.51	0.14	1		0
Q1C8	0.79	-0.11	-0.10	0.33	0.01	0.12	1	3

Nota: valores a negrito indica correlação significativa

Teste esfericidade de Bartlett, $\chi^2(54) = 135.14, p < 0.05$

KMO: 0.548

α de Cronbach: 0.65

$\det(G1) = 0.07$

Tabela 3.15: Matriz de correlação anti-imagem do G1 (sem Q1C10, Q1C9 e Q1C4)

	Q1C1	Q1C2	Q1C3	Q1C5	Q1C6	Q1C7	Q1C8
Q1C1	0.51						
Q1C2	-0.13	0.50					
Q1C3	0.27	-0.03	0.70				
Q1C5	0	0.25	-0.19	0.62			
Q1C6	-0.06	-0.36	-0.39	-0.21	0.51		
Q1C7	0.10	-0.59	-0.21	-0.45	0.30	0.52	
Q1C8	-0.78	0.18	-0.07	-0.19	0	-0.14	0.52

Nota: *MSA* está na diagonal, as correlações parciais abaixo da diagonal

Extração das componentes e interpretação dos pesos e correlações entre as variáveis e componentes

Como mostra a tabela 3.16 foram usados todos os itens de forma a gerar o mesmo número de variáveis, onde temos o poder explicativo relativo de cada componente expresso pelo valor próprio. Aplicando o critério do *valor próprio*, onde são retidas todas as componentes com valor ≥ 1 são obtidas 2 componentes, decisão unânime pelo critério *percentagem da variância*, em que a variância acumulativa ≥ 0.60 (63.06%). Pelo critério de *Gráfico de escarpa* (figura 3.2), é verificado que seria melhor extrair pelo menos mais 1 componente do que os outros 2 critérios indicam. Por conseguinte sendo que a terceira componente está muito perto do valor próprio de 1, serão extraídas 3 componentes.

Examinando a tabela 3.17, que contém a matriz não rotacionada das componentes extraídas, temos os *pesos* de cada item para cada componente, isto é, a correlação entre estes. Obtiveram-se correlações significativas entre itens e componentes, por exemplo a segunda componente com uma forte correlação com os itens Q1C8 e Q1C1. Quanto às comunalidades, estão acima do nível normalmente recomendado de 0.50, tendo atenção ao tamanho da amostra são consideradas superficialmente consistentes. Os valores próprios para cada componente e a percentagem do traço estão apresentados na parte inferior da tabela 3.17, que é representado também o total da variância extraída para cada componente, sendo a variância total de 76.02%.

Tabela 3.16: Extração das Componentes Principais do G1

Componentes	Valores Próprios		
	Total	% Variância	% Acumulada
1	2.45	34.95	34.95
2	1.97	28.11	63.06
3	0.91	12.96	76.02
4	0.83	11.85	87.87
5	0.43	6.14	94.01
6	0.25	3.51	97.52
7	0.17	2.48	100.00

Figura 3.2: Gráfico de escarpa das Componentes Principais do G1

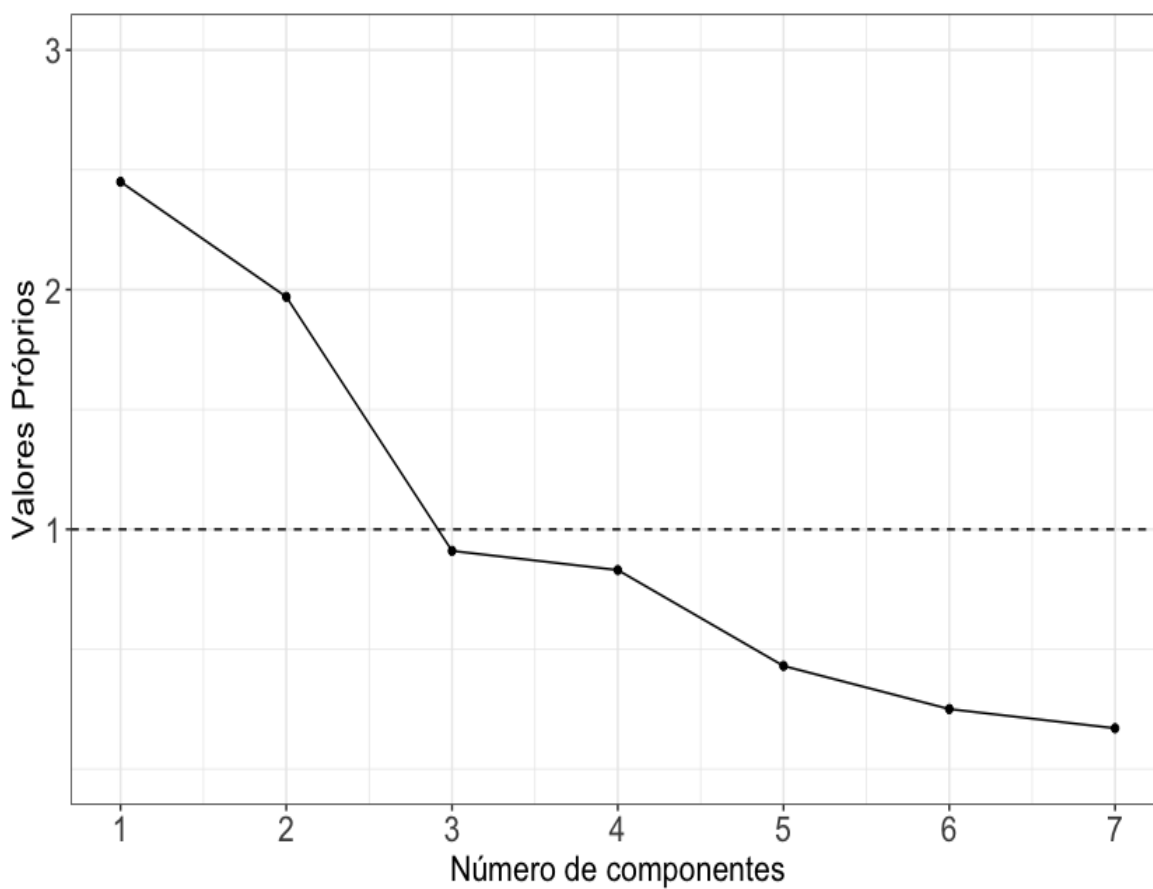


Tabela 3.17: Matriz não rotacionada das Componentes Principais do G1

Items	Componentes			h^*
	1	2	3	
<i>Q1C1</i>	-0.10	0.91	0.09	0.85
<i>Q1C2</i>	0.68	-0.17	-0.22	0.55
<i>Q1C3</i>	0.76	-0.22	0.23	0.68
<i>Q1C5</i>	0.64	0.41	-0.06	0.59
<i>Q1C6</i>	0.61	-0.05	0.71	0.88
<i>Q1C7</i>	0.77	0.12	-0.53	0.89
<i>Q1C8</i>	0.05	0.94	0.06	0.89
				Total
Valores Próprios	2.45	1.97	0.91	5.33
% do traço	34.95	28.11	12.96	76.02

* comunalidades

Análise da matriz de componentes rotacionada

Após efetuar a rotação usando o método rotacional VARIMAX, foi obtida a matriz apresentada na tabela 3.18. Foram omitidos todos os *pesos* < 0.55 , visto que para o tamanho de amostra do G1 o limite mínimo significativo é ≥ 0.55 [17], quanto aos *pesos* remanescentes foram ordenados por ordem da sua significância decrescente por componente.

Da mesma maneira do que foi efetuado no G0, com o método rotacional VARIMAX, foram maximizados os *pesos*, com uma maior correlação com cada componente, sendo que estes estão apresentados na tabela 3.18 (considerados bons (acima ≥ 0.55 ou 30% de variância sobreposta [6]), mantendo as mesmas comunalidades. Foi redistribuída a variância extraída por cada componente, sendo que permaneceu inalterado o total da variância extraída, no entanto a componente 2 fica ligeiramente com mais poder explicativo na solução gerada. Por fim, é obtido uma estrutura simples no qual cada componente é exclusivamente correlacionada com os seus itens.

Ordenando as componentes, temos:

- Componente 1: *Vantagens do parto para a Mãe*
 - *Q1C7: É mais seguro para a minha saúde;*
 - *Q1C2: Consigo programar quando será o parto;*
 - *Q1C5: Após cesariana é mais fácil retomar a vida sexual.*

- Componente 2: *Medo do Parto*
 - *Q1C8: Tenho medo;*
 - *Q1C1: Tenho medo da dor.*

- Componente 3: *Preocupação com o Bebê*
 - *Q1C6: Conheço mais pessoas que tiveram uma cesariana;*
 - *Q1C3: O meu bebê sofre menos.*

Tabela 3.18: Matriz rotacionada das componentes do G1

Ítems	Componentes			h^*
	2	1	3	
<i>Q1C8</i>	0.94			0.85
<i>Q1C1</i>	0.91			0.55
<i>Q1C7</i>		0.94		0.68
<i>Q1C2</i>		0.68		0.59
<i>Q1C5</i>		0.58		0.88
<i>Q1C6</i>			0.93	0.89
<i>Q1C3</i>			0.65	0.89
				Total
Valores Próprios	1.95	1.93	1.44	5.32
% do traço	27.86	27.57	20.57	76.00

pesos das componentes com valor inferior a 0.55 foram omitidas e as variáveis foram ordenadas pelo peso em decrescente para cada componente.

* comunalidades

CAPÍTULO 4

REGRESSÃO LOGÍSTICA

A aplicação da regressão logística aos dados deste trabalho, tem como objetivo identificar possíveis fatores preditivos, que levam a mulher a preferir o tipo de parto por via de cesariana.

4.1 Enquadramento

Muitos fenómenos sociais são discretos ou qualitativos em vez de contínuos ou quantitativos. Na natureza um evento pode ocorrer ou não ocorrer, um indivíduo pode tomar uma determinada decisão em detrimento de outra, portanto eventos binários normalmente tomam uma forma de uma variável *dummy* com valores "0" e "1", sendo normalmente indicado o "0" como grupo de referência [37].

A variável dependente (variável resposta) apenas pode tomar um valor de 0 ou 1, mas os valores previstos para a regressão podem assumir a forma de probabilidades condicionadas aos valores das variáveis independentes (covariáveis). Quanto maior o valor previsto, maior é a probabilidade de um qualquer indivíduo com determinados *scores* nas covariáveis ter a característica/experiência desse evento [37].

As estimativas de uma regressão linear com uma variável resposta dicotómica apresenta alguns problemas necessários a ultrapassar:

- i) *Não linearidade*, com uma variável resposta dicotómica, a regressão linear tem uma representação gráfica em forma de "S", por conseguinte, a variação da covariável junto dos limites máximo (1) ou mínimo (0), faz com que a variável resposta varie de forma residual, enquanto que no meio a mesma variação da covariável tem um efeito maior na variação da variável resposta [37];
- ii) *Aditividade*, caso uma covariável empurre a variável resposta para perto de um dos limites, faz com que os efeitos das outras covariáveis sejam irrelevantes na variável resposta [37];
- iii) *Distribuição não normal*, com uma variável resposta dicotómica, apenas dois resíduos existem para todo o valor da covariável [37];
- iv) *Heterocedasticidade*, a variância não é constante, isto é, junto dos limites os resíduos

dos valores da covariável são pequenos e vão aumentando quando os valores da covariável tendem para o meio [37].

De forma a ser possível analisar a variável dependente dicotômica é necessário efetuar uma transformação logarítmica da regressão linear para eliminar os limites máximos e mínimos, passando a denominar-se por regressão logística. Para tal assume-se que cada caso x_i , $i = 1, 2, \dots, p$ tem uma probabilidade condicionada de ter uma característica/experiência de um evento y , definido por $E(Y|x_i)$, por simplificação, $\pi_{x_i} = E(Y|x_i)$ para representar a média condicional que poderá ser estimada [37][21]:

$$O_{x_i} = \frac{\pi_{x_i}}{1 - \pi_{x_i}} \quad (4.1)$$

$$g_{x_i} = \log(O_{x_i}) = \ln\left(\frac{\pi_{x_i}}{1 - \pi_{x_i}}\right) \quad (4.2)$$

A equação 4.1, representa as chances (O_{x_i}) que expressam a probabilidade de um evento ocorrer $E(Y = 1|x_i)$, relativamente à probabilidade desse mesmo evento não ocorrer $E(Y = 0|x_i)$, tendo um limite mínimo de 0 mas não tem máximo. Enquanto que na equação 4.2 ao efetuar a transformação logarítmica da equação 4.1 deixamos de ter também limite mínimo [37].

Desta forma, a transformação logarítmica apresenta a inexistência de limites mínimos, $\lim_{\pi_{x_i} \rightarrow 0} \log(O_{x_i}) = -\infty$, ou máximos, $\lim_{\pi_{x_i} \rightarrow 1} \log(O_{x_i}) = +\infty$, também existe uma simetria em torno do ponto médio de π_{x_i} , isto é, $\pi_{x_i} = 0.5 \rightarrow \log(0.5/0.5) = 0$, e por fim uma variação em π_{x_i} traduz variações diferentes na equação 4.2. A transformação apresentada na equação 4.2 lineariza a relação não linear entre as covariáveis e a probabilidade da variável resposta [37].

Assim a relação linear da covariável X_i , $i = 1, 2, \dots, p$, com a transformação logarítmica prevista é [37]:

$$g_x = \log(O_x) = \beta_0 + \beta_1 X_1 \quad (4.3)$$

Exponenciando os dois lados da equação 4.3, temos [37]:

$$O_x = e^{\beta_0} * e^{\beta_1 X_1} \Leftrightarrow \frac{\pi_x}{1 - \pi_x} = e^{\beta_0} * e^{\beta_1 X_1} \quad (4.4)$$

Resolvendo a equação 4.4 em ordem a π_x e sabendo que $g_x = \beta_0 + \beta_1 X_1$ fica [37]:

$$\pi_x = \frac{e^{g_x}}{1 + e^{g_x}} \quad (4.5)$$

Esta transformação demonstra a relação não linear entre a covariável X e as probabilidades originais, isto é, uma variação do valor de X , exerce em g_x uma variação proporcional, contudo a variação em π_x já não é proporcional, resultando em grandes/pequenas variações nas probabilidades em níveis pequenos/grandes de X e g_x .

Generalizando para uma seleção de p covariáveis dado pelo vetor $\mathbf{x}' = (x_1, x_2, \dots, x_p)$. Seja a probabilidade condicional $P(Y = 1|\mathbf{x}) = \pi_{\mathbf{x}}$, então a transformação logarítmica da regressão logística múltipla é dada por [21]:

$$g_{\mathbf{x}} = \ln\left(\frac{\pi_{\mathbf{x}}}{1 - \pi_{\mathbf{x}}}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (4.6)$$

Que por sua vez o modelo de regressão logístico é [21],

$$\pi_{\mathbf{x}} = \frac{e^{g_{\mathbf{x}}}}{1 + e^{g_{\mathbf{x}}}} \quad (4.7)$$

4.2 Metodologia

4.2.1 Ajustamento do Modelo de Regressão Logística Múltipla

Na regressão linear é utilizado o método dos mínimos quadrados (OLS), no qual são escolhidos os coeficientes (β') que minimizam a soma dos quadrados dos desvios dos valores observados da variável resposta Y , a partir dos valores previstos pelo modelo. Contudo não é possível replicar o método anteriormente indicado ao modelo logístico em virtude dos resíduos não apresentarem as mesmas propriedades aquando a variável resposta é dicotómica [21].

O método geral de estimação que leva à função dos mínimos quadrados na regressão logística é chamada de máxima verosimilhança. Este método coloca iterativamente valores aos coeficientes desconhecidos (β') que maximizarão a probabilidade de obter um conjunto de dados semelhante aos dados originais [21].

De acordo com a equação 4.7 indica a probabilidade condicionada de $Y = 1$ dado \mathbf{x} , então a quantidade $1 - \pi_{\mathbf{x}_i}$ indica a probabilidade condicionada de $Y = 0$ dado \mathbf{x} . Assim sendo os pares (x_i, y_i) onde $y_i = 1$, a contribuição a função verosimilhança é $\pi_{\mathbf{x}_i}$, e para os pares (x_i, y_i) onde $y_i = 0$, a contribuição para a mesma função é $1 - \pi_{\mathbf{x}_i}$, temos: [21]

$$\pi_{\mathbf{x}_i}^{y_i} [1 - \pi_{\mathbf{x}_i}]^{1-y_i} \quad (4.8)$$

No entanto como as observações são assumidas como independentes a função verosimilhança é obtida através do produto dos termos da equação 4.8, isto é [21],

$$l(\beta) = \prod_{i=1}^n \pi_{\mathbf{x}_i}^{y_i} [1 - \pi_{\mathbf{x}_i}]^{1-y_i} \quad (4.9)$$

Mas sendo matematicamente mais fácil a transformação logarítmica, logo a equação 4.9, fica [21]:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n \left\{ y_i \ln[\pi_{\mathbf{x}_i}] + (1 - y_i) \ln[1 - \pi_{\mathbf{x}_i}] \right\} \quad (4.10)$$

Assim para encontrar o valor de β que maximiza a equação 4.10, é efetuado a diferenciação $L(\beta)$ a respeito aos coeficientes do vetor $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ e $j = 1, 2, \dots, p$ temos as equações verosimilhança [21]:

$$\sum [y_i - \pi_{\mathbf{x}_i}] \quad (4.11)$$

$$\sum x_{ij} [y_i - \pi_{\mathbf{x}_i}] \quad (4.12)$$

Seja o vetor dos coeficientes estimados $\hat{\beta}$, a solução para estas equações, desta maneira os valores ajustados para o modelo de regressão logística múltipla são $\hat{\pi}_x$ [21]. A partir das segundas derivadas parciais da equação 4.10 são estimadas as variâncias e covariâncias dos coeficientes estimados ($\hat{\beta}$), apresentando de seguinte forma [21],

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = - \sum_{i=1}^n x_{ij}^2 \pi_i (1 - \pi_i) \quad (4.13)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = - \sum_{i=1}^n x_{ij} x_{il} \pi_i (1 - \pi_i), j, l = 1, 2, \dots, p \quad (4.14)$$

onde $\pi_i = \pi_{\mathbf{x}_i}$.

Seja a matriz Informação de Fisher, $I(\beta)_{(p+1) \times (p+1)}$, contendo os termos negativos dados pelas equações 4.13 e 4.14. Então as variâncias e covariâncias dos coeficientes são obtidos através da inversa da matriz $I(\beta)$, $Var(\beta) = I^{-1}(\beta)$ [21].

Posto isto, a matriz de informação útil na discussão do ajustamento do modelo e avaliação do ajustamento é $\hat{I}(\hat{\beta}) = \mathbf{X}'\mathbf{V}\mathbf{X}$, onde \mathbf{X} é uma matriz com dimensão $n \times (p + 1)$, que contém os dados observados e \mathbf{V} é uma matriz diagonal ($n \times n$) com o elemento geral $\hat{\pi}_i(1 - \hat{\pi}_i)$, então [21],

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} \hat{\pi}_1(1 - \hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1 - \hat{\pi}_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\pi}_n(1 - \hat{\pi}_n) \end{bmatrix} \quad (4.15)$$

4.2.2 Teste de Significância dos Coeficientes Estimados

Após estimação dos coeficientes do vetor $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$, é avaliada a significância das covariáveis do modelo. É efetuado um teste de hipóteses de maneira a certificar que as covariáveis são significativas em relação à variável resposta [21],

$$H_0 : \hat{\beta}_p = 0 \quad H_1 : \hat{\beta}_p \neq 0$$

A comparação dos valores observados da variável resposta com os valores previstos, obtidos através dos modelos com e sem a covariável X_i , assim temos o teste rácio da verosimilhança [21],

$$D = -2 \ln \left[\frac{\text{verosimilhança do modelo ajustado}}{\text{verosimilhança do modelo saturado}} \right] \quad (4.16)$$

Usando a equação 4.10 e com a equação 4.16, temos [21],

$$D = -2 \sum_{i=1}^n \left[y_i \ln \left(\frac{\hat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right] \quad (4.17)$$

sendo $\hat{\pi}_i = \hat{\pi}_{x_i}$

A estatística D é chamada de desviância, sendo que é análoga à soma dos quadrados dos resíduos na regressão linear. Por conseguinte, caso pretendamos verificar se uma covariável X_i é significativa, é comparado o valor da estatística D , considerando a inclusão da covariável X_i é [21],

$$\mathbf{G} = D(\text{modelo sem a covariável}) - D(\text{modelo com a covariável}) \sim \chi_{(p)}^2 \quad (4.18)$$

Sendo a estatística \mathbf{G} análoga ao teste F parcial para a regressão linear.

Pode ser usado outro teste de significância (ou como complemento ao anterior), o teste de Wald multivariado é obtido através da expressão seguinte baseado no teste de hipótese [21]:

$$H_0 : \hat{\beta}_{p+1} = 0 \quad H_1 : \hat{\beta}_{p+1} \neq 0$$

$$W = \hat{\beta}' [\hat{Var}(\hat{\beta})]^{-1} \hat{\beta} = \hat{\beta}' [\mathbf{X}' \hat{\mathbf{V}} \mathbf{X}] \hat{\beta} \sim \chi^2_{(p+1)} \quad (4.19)$$

Todavia como existe a necessidade de um passo extra em calcular o vetor-matriz para obter $\hat{\beta}$, não evidencia vantagens em relação ao teste rácio de verosimilhança (equação 4.16) para determinar a significância do modelo.

4.2.3 Método de Seleção das Covariáveis

A principal função dos métodos para selecionar as covariáveis é garantir o "melhor" modelo dentro do problema a ser analisado [21]. Assim para tal é necessário:

- i) Um plano simples para selecionar as covariáveis para o modelo [21];
- ii) Um conjunto de métodos para avaliar a adequação do modelo dado no seu todo, como individualmente [21].

Para a construção do modelo logístico, existem alguns métodos, como o *stepwise*, *seleção propositada das covariáveis* e *seleção dos melhores subconjuntos*. O método mais usado e implementado por defeito em vários softwares estatísticos é o *stepwise* ou *passo a passo*, sendo baseado numa regra fixa da significância estatística dos coeficientes, em que o investigador pouco controla na decisão de incluir/excluir as covariáveis, contudo é um método aconselhável quando o problema é relativamente novo e existe pouca informação sobre as covariáveis e suas associações com a variável resposta [21].

O método *seleção propositada das covariáveis* engloba 7 passos em que o investigador controla todos os passos e toma a decisão de incluir/excluir a covariável. A racionalidade por detrás deste método é a mesma quando o investigador estuda um conjunto de dados e com isso formula um modelo de regressão multivariado [21]. Ambos os métodos (*stepwise* e *seleção propositada*) foram comparados através de simulações [3], que resultou na retenção

de mais covariáveis significativas pela *seleção propositada* em relação ao *stepwise*. Assim sendo aplicar-se-á o método *seleção propositada*, sendo os 7 passos [21]:

Passo 1: Análise univariada de cada covariável

- Na variável categórica a análise pode ser efetuada através de uma tabela de contingência entre a variável resposta e os k níveis da covariável, e para averiguar a sua significância o teste de qui-quadrado de Pearson. Sendo que o teste de qui-quadrado de Pearson é assintoticamente equivalente ao rácio das verosimilhanças [21];
- Na variável contínua a análise pode ser baseada no teste *t de duas amostras*, ou então uma alternativa melhor é através da regressão logística univariada entre a variável resposta e a covariável, de forma obter o erro-padrão estimado e o teste rácio das verosimilhanças. Também nesta situação ambas irão gerar *p-values* equivalentes. [21].

Através desta análise, começam a surgir covariáveis nos quais os seus $p < 0.25$, é recomendável que sejam retidas todas as covariáveis com um $p \leq 0.25$ pelo simples fato de ser um critério de triagem para a seleção inicial das covariáveis [16].

Passo 2: Ajustamento do modelo multivariado engobando todas as covariáveis ($p < 0.25$) identificadas no passo anterior. Neste passo as covariáveis que não contribuem para o modelo ($p < 0.05$) deverão ser eliminadas e o novo modelo deverá ser comparado através da equação 4.16 [21].

Passo 3: Com o modelo ajustado deverá ser efetuado a comparação com o primeiro modelo com todas as covariáveis, mais especificamente uma cuidada atenção que deve ser dada, às covariáveis cujo coeficiente alterou consideravelmente a sua magnitude, isto é, $\Delta\hat{\beta} > 20\%$, sendo a equação dada por [21]:

$$\Delta\hat{\beta}\% = 100 \frac{(\hat{\theta}_i - \hat{\beta}_i)}{\hat{\beta}_i} \quad (4.20)$$

Sendo $\hat{\theta}_i$, o valor estimado do coeficiente do modelo anterior e $\hat{\beta}_i$, o valor estimado do coeficiente do modelo novo para determinada covariável X_j . Caso surja algum coeficiente com $\Delta\hat{\beta} > 20\%$, isto indica que uma ou mais covariáveis excluídas são im-

portantes no sentido que dão-nos um ajustamento necessário do efeito das restantes covariáveis do modelo. Deste modo poderá ser necessário voltar a incluir a variável no modelo e voltar a efetuar o passo 2 e 3 com mais cuidado [21].

Passo 4: Modelo de efeitos preliminares principais. É efetuada a adição das covariáveis restantes ($p > 0.25$), que não passaram a triagem inicial de seleção, uma de cada vez, e é examinada a sua significância com o $p < 0.05$ do teste *wald*, caso seja uma variável categórica com mais de 2 níveis, é efetuado através da equação 4.16. Este passo é essencial na medida em que poderá identificar covariáveis que não estando relacionadas com a variável resposta poderão contribuir na presença de outras covariáveis [21].

Passo 5: Modelo de efeitos principais, neste passo são analisados as covariáveis contínuas, de forma que seja verificada a suposição de que a transformação logarítmica aumente ou diminua linearmente como uma função destas covariáveis. Assim para tal, os métodos mais utilizados são [21]:

- Gráfico de expressão suavizado (*Smooth Lowess Scatter Plot*);
- Polinomiais fraccionados (*Fractional Polynomials*);

Passo 6: Modelo final preliminar, neste passo é explorada a existência de interações entre as variáveis do modelo. É estimada a significância estatística da interação usando a equação 4.16 [21]

Passo 7: Modelo final, é verificada a qualidade do ajustamento como também o diagnóstico dos resíduos de forma analisar casos influentes e/ou *outliers*.

4.2.4 Avaliação da Qualidade do Ajustamento do Modelo Final

Pela qualidade do ajustamento, é pretendido saber se as probabilidades produzidas pelo modelo refletem com exatidão o verdadeiro resultado dos dados [21]. Assumindo que os valores amostrais observados do vetor variável resposta, \mathbf{Y} onde, $\mathbf{Y}' = (y_1, y_2, \dots, y_n)$ e os valores estimados do vetor $\hat{\mathbf{Y}}$, onde $\hat{\mathbf{Y}}' = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$, é ajustado se [21]:

- Medidas sumárias da distância entre y e \hat{y} são pequenas [21];

- A contribuição de cada par (y_i, \hat{y}_i) , $i = 1, 2, 3, \dots, n$ para as medidas sumárias não é sistemática e pequena relativa à estrutura do erro do modelo [21].

Assumindo que o modelo ajustado tem p covariáveis, $\mathbf{x}' = (x_1, x_2, \dots, x_p)$ e seja J , o número de valores distintos de x observados, se alguns sujeitos têm o mesmo valor x , então $J < n$. O número de sujeitos com o mesmo valor $\mathbf{x} = \mathbf{x}_j$ é denotado por padrão de covariável, m_j , $j = 1, 2, \dots, J$, logo $\sum m_j = n$. Todavia inicialmente assumimos que $J \approx n$, sendo o mais comum na prática devido que pelo menos uma variável contínua é encontrada nos modelos. Deste modo, as medidas sumárias são [21]:

- **Estatística Qui-Quadrado de Pearson**

Sabendo que o valor ajustado para o j -ésimo padrão de covariável por \hat{y}_j onde [21],

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \left\{ \frac{e^{\hat{g}_{x_j}}}{1 + e^{\hat{g}_{x_j}}} \right\} \quad (4.21)$$

onde $\hat{g}_{x_j} = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \dots + \hat{\beta}_p x_{jp}$ é a transformação logarítmica estimada. Logo o resíduo de Pearson é [21],

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - \hat{y}_j)}{\sqrt{\hat{y}_j(1 - \hat{\pi}_j)}} \quad (4.22)$$

Assim a estatística sumária baseada na equação 4.22 fica [21],

$$\chi^2 = \sum_{j=1}^J \left[r(y_j, \hat{\pi}_j) \right]^2 \quad (4.23)$$

- **Desviância**

O resíduo da desviância é [21]:

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j \ln \left(\frac{y_j}{\hat{y}_j} \right) + (m_j - y_j) \ln \left(\frac{(m_j - y_j)}{m_j(1 - \hat{\pi}_j)} \right) \right] \right\}^{\frac{1}{2}} \quad (4.24)$$

Para os padrões de covariáveis com $y_j = 0$ o resíduo da desviância é [21],

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j |\ln(1 - \hat{\pi}_j)|} \quad (4.25)$$

E com os padrões de covariáveis $y_j = m_j$ é [21],

$$d(y_j, \hat{\pi}_j) = \sqrt{2m_j |\ln(\hat{\pi}_j)|} \quad (4.26)$$

Então a estatística sumária baseada nos resíduos da desviância é [21],

$$D = \sum_{j=1}^J d(y_j, \hat{\pi}_j)^2 \quad (4.27)$$

- **Teste Hosmer-Lemeshow**

Este teste é efetuado com o agrupamento dos valores das probabilidades estimadas e supõe que $J = n$, assim seguem duas estratégias para agrupar [21]:

1. Criar uma tabela baseada nos percentis das probabilidades estimadas [21];
2. Criar uma tabela baseada nos valores fixos das probabilidades estimadas [21].

Para qualquer das estratégias tomadas, a estatística da qualidade do ajustamento de Hosmer-Lemeshow, \hat{C} , é obtida através da estatística do qui-quadrado de Pearson a partir da tabela $g \times 2$ anteriormente criada de frequências observadas e estimadas [21],

$$\hat{C} = \sum_{k=1}^g \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right] \quad (4.28)$$

Sendo, $o_{1k} = \sum_{j=1}^{c_k} y_j$, $o_{0k} = \sum_{j=1}^{c_k} (m_j - y_j)$, $\hat{e}_{1k} = \sum_{j=1}^{c_k} \hat{y}_j$, $\hat{e}_{0k} = \sum_{j=1}^{c_k} m_j (1 - \hat{\pi}_j)$ e c_k é o número de padrões da covariável no k -ésimo grupo [21]. Outros testes estatísticos que poderiam complementar a avaliação global do ajustamento são: o teste *Stukel*, o teste de *Osius e Rojek* e o teste *decil de risco de Hosmer-Lemeshow*.

- **Tabela de Classificação**

Esta tabela é criada com uma classificação cruzada entre a variável resposta e uma variável binária, no qual os valores são obtidos a partir das probabilidades estimadas. Esta variável binária para ser obtida é criada uma regra, normalmente designada o ponto de corte, c , assim [21],

- Caso a probabilidade estimada $> c$ então fica $y = 1$;
- Caso a probabilidade estimada $< c$ então fica $y = 0$.

O normal é c ter um valor de 0.5, contudo existem pacotes estatísticos que otimizam o valor de c de forma a ser mais eficaz na sensibilidade e especificidade. Todavia a classificação é sensível ao tamanho dos dois grupos criados pela variável dicotómica, dando mais vantagem ao grupo maior, o qual é independente do ajustamento do modelo [21]. Assim, esta tabela é um complemento na avaliação da qualidade do ajustamento, a não ser que de início o investigador tenha pretendido a classificação como um objetivo para a sua análise [21].

- **Área sob a Curva da Característica Operatória do Receptor**

A área sob a curva característica operatória do receptor (*AUROC*), varia entre 0.5 e 1.0, e determina a habilidade do modelo em discriminar os casos nos quais é verificado o resultado de interesse, $y = 1$, entre todos [21]. A *AUROC* pode ser visualizada graficamente entre a probabilidade de detetar um sinal verdadeiro (sensibilidade) e um falso sinal (1-especificidade) para um intervalo de possíveis valores [21].

Podendo ser classificada da seguinte maneira [21]:

<i>AUROC</i>	Classificação
[0.9 – 1.0]	discriminação extremamente boa
[0.8 – 0.9[discriminação excelente
[0.7 – 0.8[discriminação aceitável
]0.5 – 0.7[discriminação pobre
[0.5]	não há discriminação, é perfeitamente aleatório

- **Pseudo R^2**

Designando L_0 , como o modelo nulo, e o intercepto e L_p , como o modelo com o intercepto e as p covariáveis, então [21],

$$R_L^2 = \frac{L_0 - L_p}{L_0} \quad (4.29)$$

Indica quanta inclusão das covariáveis no modelo reduz a variação [31]. Outras variações ao pseudo- R^2 foram apresentadas, como [31]:

– R^2 de Cox & Snell [31]:

$$R_{CS}^2 = 1 - \left(\frac{L_0}{L_p} \right)^{\frac{2}{N}} \quad (4.30)$$

– R^2 de Nagelkerke [31]:

$$R_N^2 = \frac{\left[1 - \left(\frac{L_0}{L_p} \right)^{\frac{2}{N}} \right]}{1 - L_0^{\frac{2}{N}}} \quad (4.31)$$

Baseado em pesquisas às propriedades dos diferentes pseudo- R^2 , foi verificado que a melhor medida é o R_L^2 , porque é conceptualmente parecido ao R^2 da regressão linear na medida que reflete a redução proporcional da quantidade que está a ser minimizada [31].

4.2.5 Diagnóstico do Modelo Final

Caso as suposições da regressão logística sejam violadas, o ajustamento do modelo poderá resultar em pelo menos um efeito problemático [31]:

- a) *Coefficientes enviesados*, isto é, existe uma tendência sistemática em que os coeficientes ou são muito altos (longe de 0) ou muito baixos (perto de 0) [31];
- b) *Estimativas ineficientes*, ou seja, os erros-padrão dos coeficientes mostram grandes valores em relação ao tamanho do coeficiente, fazendo com que surja erro-tipo II [31];
- c) *Inferência estatística inválida*, isto é, a significância estatística dos coeficientes não é precisa [31].

Estes problemas devem-se ao fato de existirem pelo menos uma destas situações:

- 1) *Casos alavancados*, com valores ou muito altos/baixos das covariáveis [31];
- 2) *Outliers*, com valores fora do normal na variável resposta dado os valores das covariáveis [31];
- 3) *Casos influentes*, que agem uma influência desproporcional na estimativa dos parâmetros [31].

Pela regressão logística, é assumido que os erros seguem uma distribuição binomial, portanto caso não seja verificado que os resíduos não são normalmente distribuídos (para pequenas amostras), não existe razão para apreensão sobre as inferências estatísticas[31]. Como indicado, se os erros seguem uma distribuição binomial, conseqüentemente a variância do erro é uma função da média condicional [21]:

$$\text{var}(Y_j|\mathbf{x}_j) = m_j E(Y_j|\mathbf{x}_j) \times [1 - E(Y_j|\mathbf{x}_j)] = m_j \pi_{\mathbf{x}_j} [1 - \pi_{\mathbf{x}_j}] \quad (4.32)$$

Sabendo que a matriz de projeção derivada para a regressão logística com uma aproximação linear dos valores ajustados é [41],

$$\mathbf{H} = \mathbf{V}^{1/2} \mathbf{X} (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{1/2} \quad (4.33)$$

onde $\mathbf{V}_{J \times J}$ é uma matriz diagonal com o elemento geral [21]:

$$v_j = m_j \hat{\pi}_{\mathbf{x}_j} [1 - \hat{\pi}_{\mathbf{x}_j}] \quad (4.34)$$

v_j é o estimador base do modelo da variância de y_i . Na regressão linear os elementos da diagonal na matriz de projeção (4.33) são chamados de valores influentes, no entanto a extensão do conceito para a regressão logística é necessário saber que [21]:

$$h_j = v_j \times b_j \quad (4.35)$$

h_j é o j -ésimo elemento diagonal da matriz de projeção (4.33), e onde [21],

$$b_j = \mathbf{x}'_j (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \mathbf{x}_j \quad (4.36)$$

b_j é a distância ponderada de \mathbf{x}_j à média $\bar{\mathbf{x}}$. Sendo assim, os valores influentes (4.35) são o produto entre v_j e b_j [21]. Doravante casos com $h_j > (k + 1)/N$ (em que k são as covariáveis no modelo), são considerados influentes [31]. Então usando as estatísticas do qui-quadrado de Pearson 4.23 e da desviância 4.27, verificamos que:

- A $\Delta \chi^2$ de Pearson atribuído à retirada de um caso j é [21]:

$$\Delta \chi_j^2 = \frac{r_j^2}{(1 - h_j)} = r_{sj}^2 \quad (4.37)$$

- A ΔD_j da desviância atribuído à retirada de um caso j é [21]:

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{(1 - h_j)} \quad (4.38)$$

substituindo r_j^2 por d_j^2 então [21],

$$\Delta D_j = \frac{d_j^2}{(1 - h_j)} \quad (4.39)$$

- $\Delta \hat{c}_j$ da distância Cook [21],

$$\Delta \hat{c}_j = \frac{r_{sj}^2 h_j}{(1 - h_j)} \quad (4.40)$$

Desta maneira tendo as estatísticas de diagnóstico, a melhor maneira de interpretar, é efetuando representações gráficas entre estas, sendo as principais representações [21]:

- h_j e $\hat{\pi}_j$ [21];
- $\Delta\chi_j^2$ e $\hat{\pi}_j$ [21];
- ΔD_j e $\hat{\pi}_j$ [21];
- $\Delta\hat{c}_j$ e $\hat{\pi}_j$ [21].

Por fim, são analisados visualmente todos os casos influentes ou outliers, e caso surjam, é averiguada a necessidade de os retirar dos dados observados, caso tal aconteça é efetuado novamente o ajustamento do modelo, como também verificado a qualidade do ajustamento. Porventura se y_i ficar desajustado de $m_j\hat{\pi}_j$ então poderá ter acontecido pelo menos uma destas situações [21]:

- O modelo não demonstrou uma boa aproximação da relação correta entre a média condiconada $E(Y|\mathbf{x}_j)$ e \mathbf{x}_j [21];
- Foi omitida, ou não foi medida, uma covariável importante para o modelo [21];
- Não foi efetuada a correção da escala de pelo menos uma das covariáveis [21].

4.3 Implementação e Resultados

Para formular um modelo logístico aos dados analisados neste trabalho, foi necessário efetuar uma cuidada análise às covariáveis:

- Tratamento dos dados omissos por potencial covariável, se for contínua, é efetuada a substituição pela média, se for categórica, é substituída pela mediana;
- No caso de variáveis categóricas politómicas quando divididas por grupo da variável resposta, foram analisadas as células que estão com valores < 10 [17] ou com valores de 0, neste último caso, a solução é colapsar as categorias, no caso anterior a solução poderá ser mais flexível e não colapsar para minimizar a perda de informação. [21].

Foi analisada a variável resposta [Q1A](#), em que apresentou um dado omissos, o caso #93, no qual não indicou a sua preferência pelo tipo de parto, como também não respondeu a praticamente nenhuma outra questão, logo este caso foi removido, mantendo no total com

244 observações válidas, representando 78% do **G0** (grupo de referência) e 22% do **G1**. As covariáveis foram tratadas da seguinte maneira:

- **Variáveis categóricas dicotômicas**, pela tabela 5.1 do Anexo V são apresentadas 3 novas variáveis devido ao colapso de variáveis com células < 10 [17]. Nomeadamente, foram geradas as variáveis: i) **Q2_2_4** - *Através do meu marido/parceiro/amigo*, pelo colapso da **Q2_2** - *Através do meu marido/parceiro* e **Q2_4** - *Através de um amigo*; ii) **Q2_6_7** - *Através da internet/televisão*, pelo colapso da **Q2_6** - *Através da internet* e **Q2_7** - *Através da televisão*; iii) **Q2_8_9** - *Através de um curso/livros/jornais/revistas*, pelo colapso da **Q2_8** - *Através de um curso/livros* e **Q2_9** - *Através de jornais/revistas*;
- **Variáveis categóricas politômicas**, a variável **Q3_C** sendo a única variável politômica dos dados, no qual mostra células com valores < 10 , não foi efetuado qualquer tipo de tratamento com o objetivo de preservar a informação desta variável, gerando um $p = 0.19$, para além do mais o grupo de referência nesta variável é **Q3_C = 0** - *Não completou o ensino básico*;
- **Variáveis contínuas**, a variável **Q3_A** apresentou um $p = 0.03$ e a variável **Q3_B** um $p = 0.05$, ambos os valores gerados pelo teste Mann-Whitney.

As componentes principais criadas pela **ACP** foram transformadas em 3 novas variáveis. Numa primeira fase, foi efetuada a união das componentes de cada grupo com a mesma legenda, originando: 1) **vant_parto** - *Vantagens do Parto*; 2) **preo_bebe** - *Preocupação com o Bebê*; 3) **medo_parto** - *Medo do Parto*. Numa segunda fase os dados das componentes principais tinham pouca interpretação prática, assim sendo, foi efetuado, para cada componente principal de cada grupo, o escalonamento dos postos num intervalo entre 0 e 100, desta forma, foi dado sentido pragmático aos pontos, tendo sido colocadas as fórmulas e respetivas representações das suas distribuições no Anexo V [46].

De acordo com o primeiro passo da *seleção propositada*, foi averiguado se estas novas variáveis seguiam uma distribuição normal, através da aplicação do teste estatístico *Shapiro-Wilk*, que resultou em todas elas, um $p < 0.01$, significando que existe evidência estatística para rejeitar a H_0 . As três novas variáveis como não seguem uma distribuição normal, impossibilita a aplicação do teste *t duas amostras*. Assim aplicando o teste de Mann-Whitney

(teste não paramétrico semelhante ao *teste t duas amostras*) às variáveis em questão, foi aferido que tanto a variável *vant_parto* e a *medo_parto* originou $p < 0.01$ e *preo_bebe* com um $p = 0.26$.

Seguindo o processo de *seleção propositada*, foi efetuado um ajustamento do modelo logístico com todas as covariáveis com um $p \leq 0.25$, ou seja, foram inseridas: a *Q2_1*, a *Q2_3*, a *Q2_5*, a *Q2_2_4*, a *Q3_A*, a *Q3_B*, a *vant_parto*, a *medo_parto* e a *Q3_C* com 5 níveis (sendo o grupo de referência *Q3_C* = 0 -*Não completou o ensino básico*), como se pode verificar na tabela 5.2. Em relação ao modelo nulo existe uma melhoria da desviância (*D*), apresentado pelo $G = 60.46$, $p < 0.01$, diminuindo de $D_0 = 257.94$ para $D_1 = 197.47$, isto é, inserir as variáveis ao modelo foi melhor do que não ter nenhuma variável. Por sua vez o critério de informação Akaike (*AIC*), que mede a quantidade perdida de informação, gerando um valor menor caso tenha minimizado o desperdício de informação, significando maior qualidade do modelo. Originou um valor no modelo nulo de $AIC_0 = 259.94$ passando para $AIC_1 = 223.47$, uma melhoria significativa.

Todavia e apenas como complemento, pelo critério de informação bayesiana (*BIC*), tendo este critério uma penalização superior ao *AIC* na introdução de novas variáveis ao modelo. O $BIC_0 = 263.44$ e o $BIC_1 = 268.94$, um valor $BIC > 0$, faz com que as variáveis tenham suporte suficiente serem inseridas no modelo logístico [42].

Pela tabela 5.2 do Anexo V, com o propósito de descobrir um modelo com todas as variáveis incluídas significativas, foram retiradas 3 variáveis: a *Q2_5*, a *Q3_A* e a *Q3_C*, ao modelo logístico L_1 , apresentado na tabela 5.3 do Anexo V, onde mostra a retirada progressiva, individualizada e por ordem decrescente das variáveis que apresentaram um $p > 0.05$, gerando o modelo logístico preliminar L_4 , apresentado na tabela 5.4. Obteve um ligeiro aumento, $G = 5.41$, $p = 0.24$, e a desviância de $D_1 = 197.47$ para uma $D_4 = 203.53$, contudo o intuito inicial era apenas de retirar as variáveis com $p > 0.05$ de forma a ajustar o modelo.

Deste modo, no seguimento das exclusões efetuadas, na tabela 5.5 do anexo V, mostra a $\Delta\% \hat{\beta}$ entre o modelo atual L_4 em relação ao modelo anterior L_1 , sendo que a variável *Q3_B* obteve uma $\Delta\% \hat{\beta} = 28.98\%$, ou seja, havia uma sobre-estimação do coeficiente estimado da *Q3_B* no modelo L_1 . Portanto, com as exclusões das variáveis já anteriormente indicadas, houve um ajustamento estatístico ao efeito da *Q3_B*.

Doravante foram incluídas individualmente as variáveis: a *Q2_1.1*, a *Q2_6_7*, a *Q2_8_9*,

Tabela 5.2: Ajustamento do modelo logístico L_1 com as covariáveis $p < 0.25$

Variável	$\hat{\beta}$	Erro-Padrão	z	p	$e^{\hat{\beta}}$	$e^{\hat{\beta}}$ 95% I.C.	
Intercepto	-2.200	1.800	-1.224	0.22	0.11	0.00	3.28
<i>Q2_1</i>	-0.802	0.411	-1.950	0.05	0.45	0.20	0.99
<i>Q2_3</i>	-0.904	0.489	-1.848	0.06	0.41	0.15	1.02
<i>Q2_5</i>	0.016	0.528	0.030	0.98	1.02	0.36	2.82
<i>Q2_2_4</i>	-0.727	0.469	-1.550	0.12	0.48	0.19	1.18
<i>Q3_A</i>	-0.017	0.022	-0.808	0.42	0.98	0.94	1.03
<i>Q3_B</i>	0.089	0.037	2.428	0.02	1.09	1.02	1.18
<i>Q3_C</i> = 1	0.772	1.223	0.631	0.53	2.16	0.25	47.49
<i>Q3_C</i> = 2	1.631	1.123	1.451	0.15	5.10	0.80	101.19
<i>Q3_C</i> = 3	0.773	1.127	0.686	0.49	2.17	0.33	43.08
<i>Q3_C</i> = 4	1.080	1.164	0.928	0.35	2.95	0.41	60.89
<i>vant_parto</i>	-0.045	0.010	-4.704	< 0.01	0.96	0.94	0.97
<i>medo_parto</i>	0.029	0.008	3.451	< 0.01	1.03	1.01	1.05

$D:197.47$ $G:60.46(p < 0.01)$ $AIC:223.47$ $BIC:268.94$

a *Q2_10*, a *Q3_D1* e a *preo_bebe*, no modelo logístico preliminar L_4 , que na triagem de seleção inicial obtiveram um $p > 0.25$ como mostra a tabela 5.6 do anexo V. Foi averiguado que nenhuma das variáveis apresentou um valor significativo a ($p < 0.05$), no entanto como a variável *preo_bebe* apresentou um $p = 0.05$ e ajustamento do modelo significativo a $p = 0.05$, foi inserida no modelo logístico, originando o modelo L_{10} que é apresentado na tabela 5.7. O modelo L_{10} apresenta uma melhoria ligeira, com $G = 3.88$, $p = 0.05$, e a desviância de $D_4 = 203.53$ para uma $D_{10} = 199.65$, como também o valor $AIC_4 = 217.53$ para um valor de $AIC_{10} = 215.66$.

Tal como foi feito anteriormente aquando da exclusão de variáveis, com a inclusão da

variável *preo_bebe* houve uma $\Delta\% \hat{\beta}$ do novo modelo L_{10} , contudo como mostra a tabela 5.8 do Anexo V, foi verificado que não existem variações $> 20\%$.

Tabela 5.4: Modelo logístico preliminar L_4

Variável	$\hat{\beta}$	Erro-Padrão	z	p	$e^{\hat{\beta}}$	$e^{\hat{\beta}}$ 95% I.C.	
Intercepto	-0.857	1.248	-0.687	0.49	0.42	0.04	4.87
<i>Q2_1</i>	-0.854	0.374	-2.286	0.02	0.43	0.20	0.87
<i>Q2_3</i>	-0.933	0.463	-2.017	0.04	0.39	0.15	0.94
<i>Q2_2_4</i>	-0.845	0.436	-1.940	0.05	0.43	0.18	0.98
<i>Q3_B</i>	0.069	0.035	1.979	0.05	1.07	1.00	1.15
<i>vant_parto</i>	-0.046	0.010	-4.887	< 0.01	0.96	0.94	0.97
<i>medo_parto</i>	0.031	0.008	4.002	< 0.01	1.03	1.02	1.05

$D:203.53$ $G:5.41(p = 0.25)$ $AIC:217.53$ $BIC:242.01$

Em consenso com os passos da seleção propositada, foi analisada a suposição de linearidade da transformação logarítmica do modelo L_{10} em relação às covariáveis contínuas incluídas (*Q3_B*; *vant_parto*; *medo_parto*; *preo_bebe*), as suas representações gráficas são mostradas pela figura 5.3, assim apura-se que apresentam uma linearidade em relação às chances logarítmicas.

Com o modelo preliminar L_{10} , foram examinadas todas as possíveis interações entre as variáveis deste modelo, como revela a tabela 5.9 do Anexo V. Nenhuma interação foi significativa a $p < 0.05$, mas a interação $Q2_1 * Q2_2_4$ apresenta-se como significativa, $p < 0.10$, sendo o ajustamento também significativo, $p < 0.10$.

A análise à qualidade do ajustamento do modelo L_{10} , está sumariamente descrito na tabela, 5.11, tendo R^2 apresentado um valor de 0.226, enquanto que o $R_N^2 = 0.326$.

Pelo teste de Hosmer-Lemeshow baseado nos deciles e baseado na equação (4.28), o valor de $\hat{C} = 5.91$, com 8 graus de liberdade indicou um $p = 0.66$, provando uma boa quali-

Figura 5.3: Visualização das variáveis contínuas do modelo L_{10}

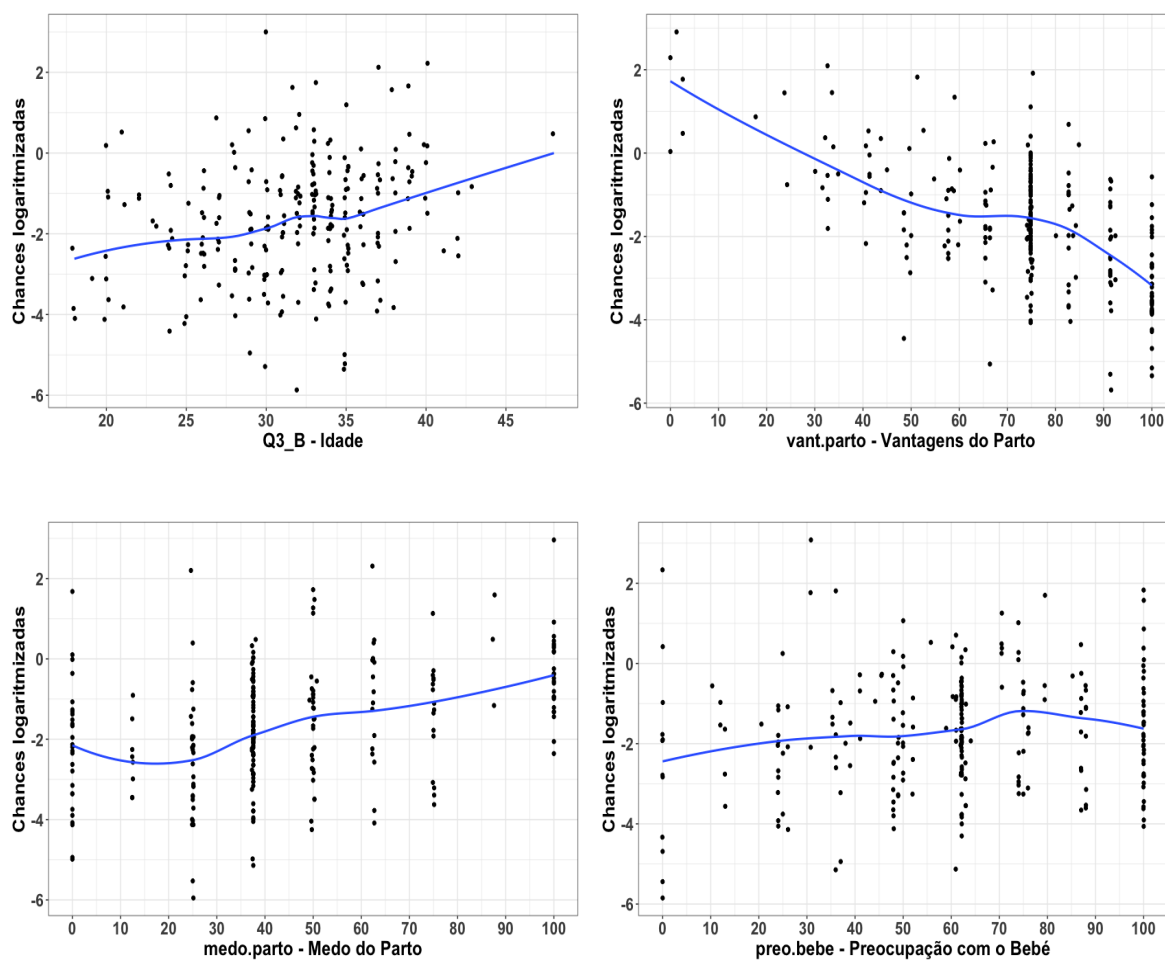


Tabela 5.7: Modelo logístico preliminar L_{10}

Variável	$\hat{\beta}$	Erro-Padrão	z	p	$e^{\hat{\beta}}$	$e^{\hat{\beta}}$ 95% I.C.	
Intercepto	-1.492	1.290	-1.157	0.25	0.23	0.02	2.80
<i>Q2_1</i>	-0.900	0.380	-2.371	0.02	0.41	0.19	0.84
<i>Q2_3</i>	-1.007	0.469	-2.147	0.03	0.37	0.14	0.88
<i>Q2_2_4</i>	-0.897	0.441	-2.035	0.04	0.41	0.16	0.94
<i>Q3_B</i>	0.074	0.035	2.130	0.03	1.08	1.01	1.16
<i>vant_parto</i>	-0.050	0.010	-5.094	< 0.01	0.95	0.93	0.97
<i>medo_parto</i>	0.030	0.008	3.728	< 0.01	1.03	1.02	1.05
<i>preo_bebe</i>	0.013	0.007	1.926	0.05	1.01	1.00	1.03

$D:199.65$ $G:3.88(p = 0.05)$ $AIC:215.66$ $BIC:243.63$

dade de ajustamento. De forma a confirmar a qualidade do ajustamento pela tabela 5.10 do Anexo V, mostra a tabela de contingência de deciles de Hosmer-Lemeshow e a análise passa pelo seguinte cálculo,

$$\frac{|y_1 - \hat{y}_1|}{\sqrt{\hat{y}_1}} \sim N(\mu, \sigma) \Leftrightarrow \frac{|7.00 - 4.40|}{\sqrt{4.40}} = 1.24$$

assim a diferença padronizada se foi corretamente ajustada, não deverá exceder o valor 1.96 (no máximo 2.0), ou seja, segue uma distribuição normal, desta forma existe uma validação complementar do ajustamento do modelo [21].

A tabela de classificação 5.11, com um ponto de corte otimizado a $c = 0.224$, como pode ser verificado na figura 5.4 do Anexo V, apresenta uma sensibilidade a 72,22% e especificidade a 72,10%, com uma classificação correta de 72,13%, deste modo, além de uma boa qualidade de ajustamento também tem uma razoável classificação, atendendo que existe uma acentuada assimetria no tamanho das amostras entre os grupos ($n_0 = 190$ e $n_1 = 54$). Quanto ao valor da área sob a curva da característica operatória do receptor foi de $AUROC = 81.73$ com um $I.C.(AUROC,0.95) = (75.27, 88.18)$, fornecendo assim uma classificação de discrimi-

nação excelente, que poderá ser melhor visualizada no conjunto de gráficos da figura 5.5, é mostrado pelos histogramas que o G0 apresenta uma assimetria negativa acentuada, e o G1 é mais disperso, sendo esta dispersão melhor verificada na figura 5.5b.

O diagnóstico do modelo ajustado foi iniciado pela visualização da figura 5.3, desta maneira, foi mais acessível a verificação de casos influentes nas figuras 5.6a e 5.6b e de *outliers* pelas figuras, 5.6c, 5.6d. Então foram averiguados e extraídos os casos pelos seguintes critérios [21]:

- Casos influentes:
 - $h > 0.10$
 - $\Delta\hat{c} > 0.10$

- *Outliers*:
 - $\Delta\chi^2_{Pearson} \leq -3.0$ ou $\Delta\chi^2_{Pearson} \geq 3.0$;
 - $\Delta D \leq -2.0$ ou $\Delta D \geq 2.0$;

Sendo assim, originou a tabela 5.12, onde foram extraídos 11 casos potenciais de análise. Por conseguinte, a regra de exclusão de casos dos dados originais, foi a validação nos 2 critérios acima referidos para cada tipo de situação (casos influentes ou *outliers*). Entretanto apenas 4 casos foram identificados pela regra de exclusão: 198; 203; 215; 240, sendo *outliers*. Após retirada dos casos foi criado novo modelo logístico L_{31} , na tabela 5.13 foi sumarizada toda análise a este modelo.

Desta maneira, com a exclusão dos 4 casos mencionados, o ajustamento do modelo ficou significativo a $p < 0.05$, como também o *AIC* e o *BIC* tiveram uma melhoria para 187.96 e 215.81, respetivamente. O R^2 e o R_N^2 tiveram igualmente uma melhoria para 0.300 e 0.413, respetivamente e a estatística de Hosmer & Lemeshow reforçou a boa qualidade do ajustamento ($\hat{C} = 4.24, p = 0.84$).

Os $\hat{\beta}$ são todos significativos a $p < 0.05$, sendo que a sua $\Delta\%\hat{\beta}$ em relação ao modelo L_{10} estava subestimada como pode ser verificada na tabela 5.14 do anexo V, havendo um ajustamento $\Delta\%\hat{\beta} > 20\%$ em 3 variáveis (*Q2_2_4*, *Q3_B* e *preo_bebe*).

As chances originadas por este modelo, tabela 5.13, revelam algumas particularidades interessantes, nomeadamente, existe uma diminuição das chances em preferir um parto via

cesariana quando as mulheres grávidas obtêm informação, acentuando aqui um factor indireto de relacionamento, quanto maior a proximidade da relação, maior a confiança no tipo de parto que a mulher grávida prefere.

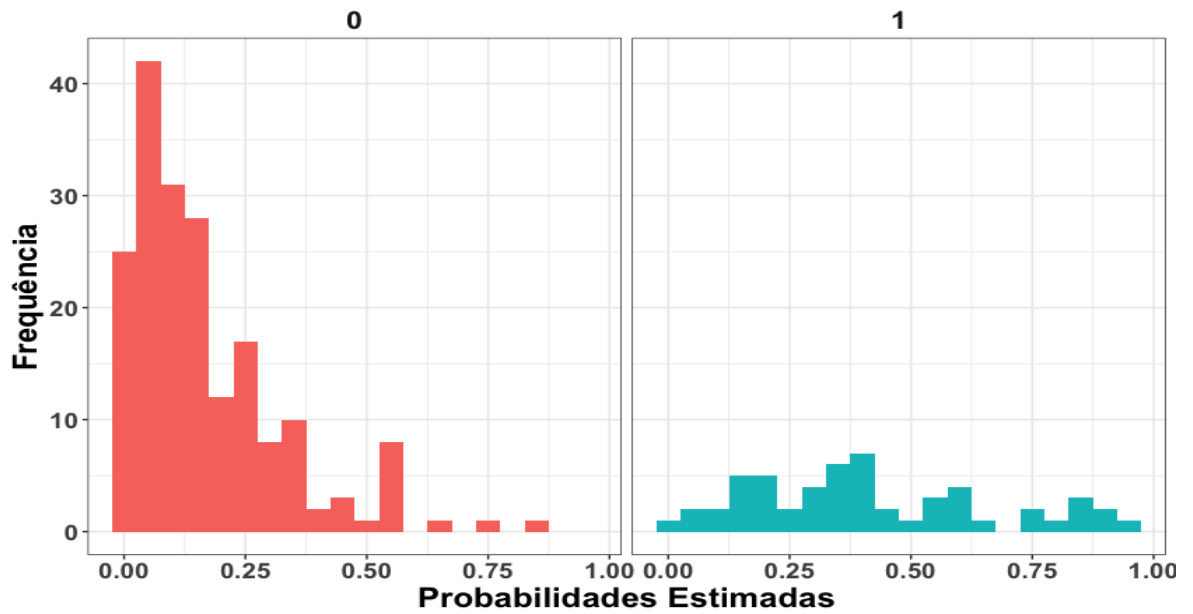
Enquanto que na covariável [Q2_1](#), quando obtêm uma informação especializada (médico obstetra/família), as chances diminuem 61% de preferir um parto via cesariana com um $I.C_{(e^{\beta},0.95)} = (14\%,83\%)$, no caso da covariável [Q2_2_4](#), quando obtêm uma informação do seu marido/parceiro/amigo, as chances diminuem 72% de preferir um parto via cesariana com um $I.C_{(e^{\beta},0.95)} = (28\%,90\%)$. Todavia, na covariável [Q3_B](#), por cada ano de idade, aumenta as chances em 12% de preferir um parto via cesariana com um $I.C_{(e^{\beta},0.95)} = (4\%,22\%)$.

Por último, a preditividade classificatória do modelo logístico L_{31} , obteve melhorias na sua sensibilidade para 76,00%, na sua especificidade para 76,84%, obtendo também uma razoabilidade classificativa correta de 76,66%. A área sob a curva característica do receptor foi de $AUROC = 88.73$ com um $I.C_{(AUROC,0.95)} = (84.12,93.35)$, mantendo a classificação excelente.

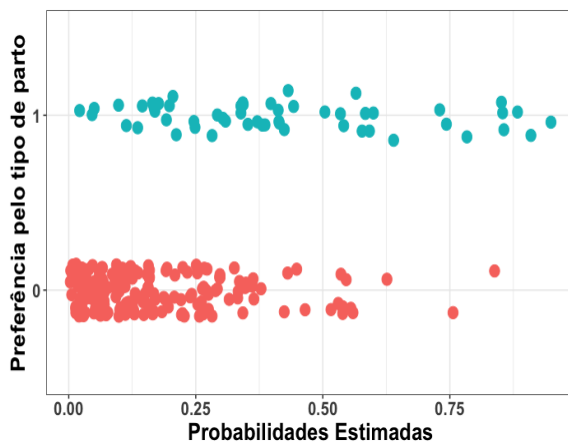
Terminando esta análise regressiva logística, foi apurado que os fatores que mais contribuem para que as mulheres grávidas escolham tendencialmente o parto via cesariana são: a Idade ([Q3_B](#)), o medo do parto (*medo_parto*) e a preocupação com o bebé (*preo_bebe*). Sendo estes dois últimos fatores mais latos, estando neles englobados:

- Medo do Parto:
 - Medo do parto em si;
 - Medo da dor.
- Preocupação com o bebé:
 - Conhecem outros casos de parto por via cesariana;
 - O bebé sofre menos.

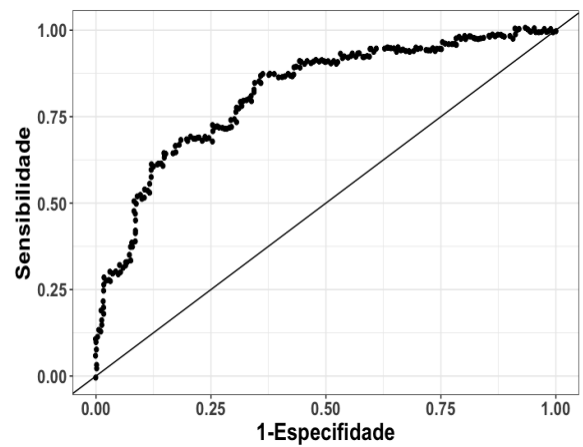
Figura 5.5: Discriminação do modelo ajustado L_{10} aos dados, com uma AUROC = 81.73, $n = 244$



(a) Histograma por Grupo

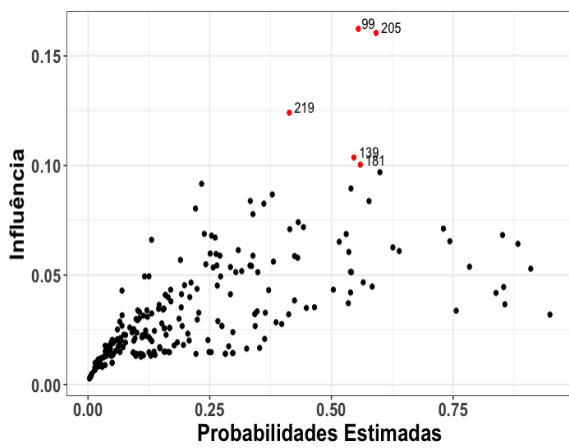


(b) Dispersão por Grupo

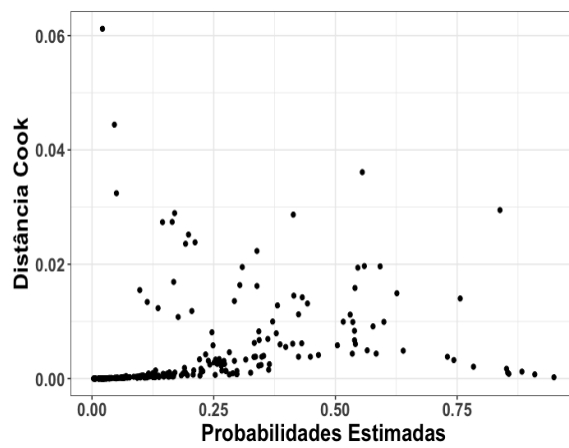


(c) Curva Característica do Receptor

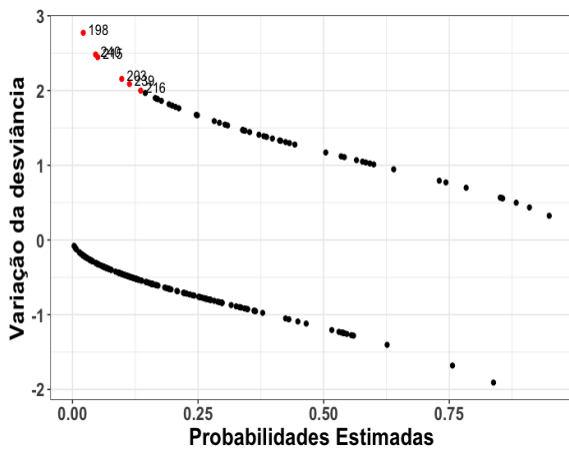
Figura 5.6: Diagnóstico dos resíduos do modelo logístico L_{10}



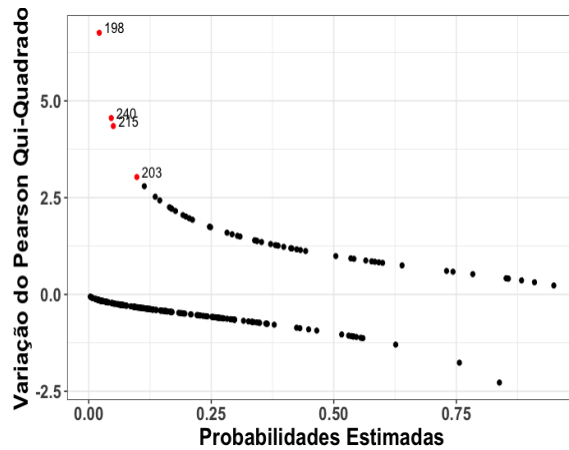
(a) Leverage



(b) Distância Cook



(c) Δ Desviância



(d) $\Delta\chi^2$ Pearson

Tabela 5.10: Avaliação da qualidade do Ajustamento do modelo L_{10}

	Modelo Base L_1	Modelo Intermédio L_4		Modelo Final L_{10}	
	Estatística	Δ	p	Δ	p
$-2LL$	-98.74	3.03	0.42	-1.94	0.05
AIC	223.47	-5.94		-1.88	
BIC	268.94	-26.93		1.62	
Medidas Sumárias do modelo L_{10}					
R^2	0.226				
R_{CS}^2	0.212				
R_N^2	0.326				
	Estatística	$g.l.$	p		
Hosmer & Lemeshow χ^2	5.91	8	0.66		
Matriz de Classificação do modelo L_{10}					
		Observados			
Estimados	1 - Cesariana	0 - Vaginal		Total	
1 - Cesariana	39	53		92	
0 - Vaginal	15	137		152	
Total	54	190		244	
Sensibilidade: 72,22%		Especificidade: 72,10%		Classificação Correta: 72,13%	

Tabela 5.12: Tabela dos casos influentes ou *outliers* do modelo L_{10}

Casos	99	139	181	198	203	205	215	216	219	239	240
<i>Q1A</i>	0	0	0	1	1	1	1	1	1	1	1
<i>Q2_1</i>	1	1	0	0	1	1	0	0	0	1	0
<i>Q2_3</i>	0	0	0	1	0	0	0	0	0	0	0
<i>Q2_2_4</i>	1	1	0	1	0	0	1	0	1	0	0
<i>Q3_B</i>	34	33	48	21	29	21	29	32	34	29	30
<i>vant_parto</i>	0.00	65.42	74.25	74.82	74.82	2.57	92.43	92.43	60.14	74.82	74.82
<i>medo_parto</i>	0.00	100.00	25.00	37.57	37.57	0.00	37.57	30.78	85.26	37.57	0.00
<i>preo_bebe</i>	73.99	100.00	87.00	49.62	50.00	100.00	62.16	75.00	0.00	62.16	0.00
\hat{y}	0.56	0.55	0.56	0.02	0.10	0.59	0.05	0.14	0.41	0.11	0.05
$\Delta\chi^2$	-1.12	-1.10	-1.13	6.76*	3.03*	0.83	4.35*	2.52	1.19	2.80	4.56*
ΔD	-1.27	-1.26	-1.28	2.77*	2.16*	1.02	2.45*	2.00*	1.33	2.09*	2.48*
h	0.16 [†]	0.10	0.10	0.01	0.01	0.16 [†]	0.01	0.02	0.12 [†]	0.01	0.02
$\Delta\hat{c}$	0.04	0.02	0.02	0.06 [†]	0.02	0.02	0.03	0.01	0.03	0.01	0.04

[†] Casos influentes

* *Outlier*

Tabela 5.13: Sumário do modelo final L_{31}

Modelo Base L_1		Modelo Intermédio L_4		Modelo Intermédio L_{10}		Modelo Final L_{31}	
Estatística		Δ	p	Δ	p	Δ	p
-2LL	-98.74	3.03	0.42	-1.94	0.05	-13.85	0.02
<i>AIC</i>	223.47	-5.94		-1.88		-27.69	
<i>BIC</i>	268.94	-26.93		1.62		-27.82	
Medidas Sumárias do modelo L_{31}							
R^2			0.300				
R_{CS}^2			0.264				
R_N^2			0.413				
		Estatística	<i>g.l.</i>	p			
Hosmer & Lemeshow χ^2		4.24	8	0.84			
Modelo Logístico L_{31}							
Variável	$\hat{\beta}$	Erro-Padrão	z	p	$e^{\hat{\beta}}$	$e^{\hat{\beta}}$ 95% I.C.	
Intercepto	-2.758	1.412	-1.953	0.05	0.06	0.00	0.98
<i>Q2_1</i>	-0.934	0.414	-2.256	0.02	0.39	0.17	0.86
<i>Q2_3</i>	-1.183	0.516	-2.290	0.02	0.31	0.10	0.81
<i>Q2_2_4</i>	-1.263	0.500	-2.526	0.01	0.28	0.10	0.72
<i>Q3_B</i>	0.112	0.039	2.878	0.04	1.12	1.04	1.22
<i>vant_parto</i>	-0.062	0.011	-5.460	< 0.01	0.94	0.92	0.96
<i>medo_parto</i>	0.037	0.009	4.154	< 0.01	1.04	1.02	1.06
<i>preo_bebe</i>	0.021	0.008	2.630	0.01	1.02	1.00	1.04
Matriz de Classificação do modelo L_{31}							
		Observados					
Estimados	1 - Cesariana		0 - Vaginal		Total		
1 - Cesariana	38		44		82		
0 - Vaginal	12		146		158		
Total	50		190		240		
Sensibilidade: 76,00%		Especificidade: 76,84%		Classificação Correta: 76,66%			

CAPÍTULO 5

CONCLUSÃO

Nesta dissertação, o que inicialmente foi planeado, era identificar potenciais fatores para o aumento da taxa de cesarianas pela preferência do tipo de parto pela grávida. Assim sendo, através da aplicação de técnicas estatísticas multivariadas, foi com sucesso, expostos potenciais fatores que podem promover e potencializar a preferência pelo parto via cesariana, tais como: idade, o medo do parto e da dor, e a preocupação com o bem-estar do bebê.

Com a aplicação da *ACP*, reduziu-se um conjunto de informação complexa, com razoabilidade, a uma informação mais simples e pragmática, desvendando fatores que de outra maneira não seriam acessíveis. O uso da regressão logística mostrou-se determinante, de modo a identificar quais os fatores preditivos que influenciam quer negativa ou positivamente a preferência pelo tipo de parto por via cesariana.

Apesar do trabalho efetuado apresentar fatores potenciais da preferência pelo tipo de parto, pode não representar toda a população, dado que apenas inclui grávidas que estavam em consulta no sector de saúde público, onde a taxa de cesariana é metade do sector de saúde privado. No entanto seria pertinente, recolher uma amostra semelhante no privado de forma a estudar outros potenciais fatores e/ou fundamentar os já determinados.

Embora devam ser razões puramente clínicas a determinar o tipo de parto e a preferência da grávida não deva ser determinante, é essencial que esta esteja bem informada sobre os riscos. No sector de saúde privado, a preferência da grávida pode ter maior peso na escolha do clínico, dada ansiedade associada ou à logística hospitalar, sendo ainda mais relevante compreender os motivos desta preferência.

Conclui-se que, pela *ACP*, as grávidas decidem a sua preferência pelo tipo de parto por via de cesariana, com base na segurança para a sua saúde, na programação do parto, no medo, e no conhecimento de outros partos pela mesma via. Por sua vez, pela regressão logística, foram identificados fatores de risco para a preferência pelo tipo de parto por via cesariana, como o medo e a idade, e fatores preventivos para a preferência pelo tipo de parto vaginal como a obtenção de informação junto do médico de família/obstetra, familiares ou amigos.

BIBLIOGRAFIA

- [1] Rami H Al-Rifai, Iffat Elbarazi, Nasloon Ali, Tom Loney, Abderrahim Oulhaj, and Luai A Ahmed. Knowledge and preference towards mode of delivery among pregnant women in the United Arab Emirates: The mutaba'ah study. *International Journal of Environmental Research and Public Health*, 18(1):36, 2021. doi:[10.3390/ijerph18010036](https://doi.org/10.3390/ijerph18010036).
- [2] Laura B Attanasio, Katy B Kozhimannil, and Kristen H Kjerulff. Women's preference for vaginal birth after a first delivery by cesarean. *Birth*, 46(1):51–60, 2018. doi:[10.1111/birt.12386](https://doi.org/10.1111/birt.12386).
- [3] Zoran Bursac, C Heath Gauss, David Keith Williams, and David W Hosmer. Purposeful selection of variables in logistic regression. *Source code for biology and medicine*, 3(1):3–17, 2008. doi:[10.1186/1751-0473-3-17](https://doi.org/10.1186/1751-0473-3-17).
- [4] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966. doi:https://doi.org/10.1207/s15327906mbr0102_10.
- [5] Norman Cliff and Charles D Hamburger. The study of sampling errors in factor analysis by means of artificial experiments. *Psychological bulletin*, 68(6):430–475, 1967. doi:[10.1037/h0025178](https://doi.org/10.1037/h0025178).
- [6] Andrew L Comrey and Howard B Lee. *A First Course in Factor Analysis. 2nd Edition*. Erlbaum Associates, 1992.
- [7] Inês Coelho da Costa, Catarina S Nunes, and Humberto S Machado. Mode of delivery and labour analgesia: a study of preference in portuguese pregnant women. *Journal of Anesthesia & Clinical Research*, 9(1), 2018. doi:[10.4172/2155-6148.1000849](https://doi.org/10.4172/2155-6148.1000849).

- [8] Anna B Costello and Jason Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, 10(1), 2005. doi:<https://doi.org/10.7275/jyj1-4868>.
- [9] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951. doi:<https://doi.org/10.1007/BF02310555>.
- [10] Maria Inês Cardoso Coelho da Costa. Tipo de parto em Portugal: Análise da preferência. *Dissertação Mestrado Integrado de Medicina*. Universidade do Porto, 2018.
- [11] George H Dunteman. *Principal components analysis*. Number 69 in Quantitative Applications in the Social Sciences. Sage, 1989.
- [12] Brian S Everitt and Graham Dunn. *Applied multivariate data analysis*. Wiley Online Library, 2001.
- [13] Darren George and Paul Mallery. *IBM SPSS Statistics 26 step by step: A simple guide and reference*. Routledge, 2019.
- [14] Richard L Gorsuch. *Factor analysis Lawrence Erlbaum Associates*. 1983.
- [15] Narinder Kaur Gosall and Gurpal Singh Gosall. *The doctor's guide to critical appraisal*. PasTest Ltd, 2012.
- [16] Sander Greenland. Modeling and variable selection in epidemiologic analysis. *American journal of public health*, 79(3):340–349, 1989. doi:[10.2105/ajph.79.3.340](https://doi.org/10.2105/ajph.79.3.340).
- [17] Joseph F Hair, William C Black, Barry J Babin, and Rolph E Anderson. *Multivariate data analysis: Pearson new international edition*. Pearson International Edition, 2014.
- [18] Harry H Harman. *Modern factor analysis*. University of Chicago press, 1976.
- [19] Nathanael A Heckert, James J Filliben, C M Croarkin, B Hembree, William F Guthrie, P Tobias, and J Prinz. Handbook 151: Nist/sematech e-handbook of statistical methods. 2002. URL: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm>.

- [20] John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965. doi:<https://doi.org/10.1007/BF02289447>.
- [21] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
- [22] Richard Arnold Johnson and Dean W Wichern. *Applied Multivariate Statistical Analysis, 6th Edn, New International Edition*. Pearson Education International, 2014.
- [23] Ian T Jolliffe. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(2):160–173, 1972. doi:<https://doi.org/10.2307/2346488>.
- [24] Henry F Kaiser. A second generation little jiffy. *Psychometrika*, 35(4):401–415, 1970. doi:<https://doi.org/10.1007/BF02291817>.
- [25] Henry F Kaiser and John Rice. Little jiffy, mark iv. *Educational and psychological measurement*, 34(1):111–117, 1974. doi:<https://doi.org/10.1177/001316447403400115>.
- [26] DPhil; Kurinczuk Jennifer J. MSc MD; Spark Patsy BSc; Brocklehurst Peter MBChB Knight, Marian MBChB. Cesarean delivery and peripartum hysterectomy. *Obstetrics & Gynecology*, 111(1):97–105, 2008. doi:[10.1097/01.AOG.0000296658.83240.6d](https://doi.org/10.1097/01.AOG.0000296658.83240.6d).
- [27] Robert C MacCallum, Keith F Widaman, Shaobo Zhang, and Sehee Hong. Sample size in factor analysis. *Psychological methods*, 4(1):84–99, 1999. doi:<https://doi.org/10.1037/1082-989X.4.1.84>.
- [28] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947. doi:[10.1214/aoms/1177730491](https://doi.org/10.1214/aoms/1177730491).
- [29] Agustina Mazzoni, Fernando Althabe, Laura Gutierrez, Luz Gibbons, Nancy H Liu, Ana María Bonotti, Gustavo H Izbizky, Marta Ferrary, Nora Viergue, Silvia I Vigil, et al. Women’s preferences and mode of delivery in public and private hospitals: a prospective cohort study. *BMC pregnancy and childbirth*, 16(1):1–8, 2016. doi:[10.1186/s12884-016-0824-0](https://doi.org/10.1186/s12884-016-0824-0).

- [30] Liu N Bonotti A Gibbons L Sánchez A Belizán J Mazzoni A, Althabe F. Women's preference for caesarean section: a systematic review and meta-analysis of observational studies. *BJOG: an international journal of obstetrics & gynaecology*, 118(4):391–399, 2011. doi:[10.1111/j.1471-0528.2010.02793.x](https://doi.org/10.1111/j.1471-0528.2010.02793.x).
- [31] Scott Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002.
- [32] Mehmet Mendes and Akin Pala. Type i error rate and power of three normality tests. *Pakistan Journal of Information and Technology*, 2(2):135–139, 2003. doi:[10.3923/itj.2003.135.139](https://doi.org/10.3923/itj.2003.135.139).
- [33] Yanhong Ming, Rong Huang, Wei Zhou, Bin Wang, Hongping Yu, and Jun Zhang. Is age and socioeconomic status associated with preference for birth mode in nulliparous women in china? *Archives of gynecology and obstetrics*, 300(1):33–40, 2019. doi:<https://doi.org/10.1007/s00404-019-05140-w>.
- [34] Ganesh R Naik. *Advances in Principal Component Analysis: Research and Development*. Springer, 2017. doi:<https://doi.org/10.1007/978-981-10-6704-4>.
- [35] OECD. *Health at a Glance 2019*. 2019. [acedido a 12 de Agosto 2020]. URL: <https://www.oecd-ilibrary.org/content/publication/4dd50c09-en>.
- [36] World Health Organization et al. Who statement on caesarean section rates. Technical report, World Health Organization, 2015. [acedido a 20 de Julho de 2020]. URL: https://www.who.int/reproductivehealth/publications/maternal_perinatal_health/cs-statement/en/.
- [37] Fred C Pampel. *Logistic regression: A primer*, volume 132. Sage publications, 2000.
- [38] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi:<https://doi.org/10.1080/14786440109462720>.
- [39] PORDATA. Cesarianas nos hospitais (%). [acedido a 12 de Agosto 2020]. URL: [https://www.pordata.pt/Portugal/Cesarianas+nos+hospitais+\(percentagem\)-1985](https://www.pordata.pt/Portugal/Cesarianas+nos+hospitais+(percentagem)-1985).

- [40] PORDATA. Partos nos hospitais públicos por tipo e partos nos hospitais privados por tipo. [accedido a 14 de Agosto 2020]. URL: <https://www.pordata.pt/DB/Portugal/Ambiente+de+Consulta/Tabela>.
- [41] Daryl Pregibon et al. Logistic regression diagnostics. *Annals of statistics*, 9(4):705–724, 1981. doi:<https://doi.org/10.1214/aos/1176345513>.
- [42] Adrian E Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995. doi:<https://doi.org/10.2307/271063>.
- [43] Nornadiah Mohd Razali, Yap Bee Wah, et al. *r. Journal of statistical modeling and analytics*, 2(1):21–33, 2011.
- [44] Elizabeth Reis. *Estatística multivariada aplicada*. Edições Sílabo, 2001.
- [45] Gillian Ryan, Kate CO Doherty, Declan Devane, Fionnuala McAuliffe, and John Morrison. Questionnaire survey on women’s views after a first caesarean delivery in two tertiary centres in Ireland and their preference for involvement in a future randomised trial on mode of birth. *BMJ open*, 9(10):e031766, 2019. doi:<http://dx.doi.org/10.1136/bmjopen-2019-031766>.
- [46] Randall E Schumacker. *Using R with multivariate statistics*. Sage Publications, 2015.
- [47] K.S. Joseph Maureen Heaman Reg Sauve Michael S. Kramer Shiliang Liu, Robert M. Liston et al. Maternal mortality and severe morbidity associated with low-risk planned cesarean delivery versus planned vaginal delivery at term. *CMAJ*, 176(4):455–460, 2007. doi:<https://doi.org/10.1503/cmaj.060870>.
- [48] Jonathon Shlens. A tutorial on principal component analysis. Technical report, 2014.
- [49] BG Tabachnick and LS Fidell. *Using Multivariate Statistics, 6th Edn, New International Edition*. Pearson Education International, 2013.
- [50] Louis Leon Thurstone. Multiple-factor analysis; a development and expansion of the vectors of mind. 1947.

- [51] Sigmund Tobias and James E Carlson. Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivariate behavioral research*, 4(3):375–377, 1969. doi:[10.1207/s15327906mbr0403_8](https://doi.org/10.1207/s15327906mbr0403_8).
- [52] William R Zwick and Wayne F Velicer. Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432–442, 1986. doi:<https://doi.org/10.1037/0033-2909.99.3.432>.

Anexo I

Questionário - Preferência entre parto Vaginal e Cesariana

Este questionário anónimo é composto por 3 partes. Na 1ª, caso tenha respondido Parto vaginal na alínea A, responda somente à alínea B. Caso tenha respondido Cesariana, responda somente à alínea C.

1.

A. Caso pudesse escolher, que tipo de parto escolheria?

- Parto vaginal
 Cesariana

B. Porque preferiria um parto vaginal?

	Totalmente em desacordo		Totalmente de acordo		
• Não deixa cicatriz	1	2	3	4 5	Não sei
• Não quero perder a primeira hora de vida do meu bebé	1	2	3	4 5	Não sei
• Tenho medo	1	2	3	4 5	Não sei
• Não quero que o meu bebé nasça através de uma cirurgia	1	2	3	4 5	Não sei
• Já tive um parto vaginal	1	2	3	4 5	Não sei
• Conheço mais pessoas que tiveram um parto vaginal	1	2	3	4 5	Não sei
• Quero ter vários filhos e sei que uma cesariana pode limitar o n ^o	1	2	3	4 5	Não sei
• A recuperação pós-parto é mais rápida	1	2	3	4 5	Não sei
• O tempo de internamento hospitalar é menor	1	2	3	4 5	Não sei

Outro:

B.1. Optaria por anestesia?

- Sim, anestesia epidural
 Sim, anestesia local
 Não
 Não sei

C. Porque preferiria um parto por cesariana?

	Totalmente em desacordo		Totalmente de acordo		
• Tenho medo da dor	1	2	3	4 5	Não sei
• Consigo programar quando será o parto	1	2	3	4 5	Não sei
• O meu bebé sofre menos	1	2	3	4 5	Não sei
• Já tive uma cesariana	1	2	3	4 5	Não sei
• Após uma cesariana é mais fácil retomar a vida sexual	1	2	3	4 5	Não sei
• Conheço mais pessoas que tiveram uma cesariana	1	2	3	4 5	Não sei
• É mais seguro para a minha saúde	1	2	3	4 5	Não sei
• Tenho medo	1	2	3	4 5	Não sei
• Consigo programar quando será o parto	1	2	3	4 5	Não sei

Outro:

C.1. Que tipo de anestesia preferiria?

- Anestesia epidural
 Anestesia local
 Anestesia geral
 Não sei

2. De que modo obteve a informação que a ajudou a decidir as suas preferências? (selecione as alíneas que se aplicam)

- | | | | |
|--|--------------------------|-----------------------------|--------------------------|
| Através do meu médico obstetra/ de família | <input type="checkbox"/> | Através da internet | <input type="checkbox"/> |
| Através do meu marido/parceiro | <input type="checkbox"/> | Através da televisão | <input type="checkbox"/> |
| Através de um familiar | <input type="checkbox"/> | Através de um curso/livros | <input type="checkbox"/> |
| Através de um amigo | <input type="checkbox"/> | Através de jornais/revistas | <input type="checkbox"/> |
| Ainda não pensei | <input type="checkbox"/> | Outro: | <input type="checkbox"/> |

3. Dados pessoais:

- Tempo de gestação: _____ semanas
- Idade: _____
- Nível de escolaridade:
 - Não completou o ensino básico
 - Ensino básico (até ao 9º ano, inclusive)
 - Ensino Secundário (até ao 12º ano, inclusive)
 - Licenciatura
 - Mestrado
 - Doutoramento
- Já teve algum parto? Sim Não
Se sim, já teve alguma cesariana? Sim Não

Obrigado por responder a este questionário.

Anexo II

Preferência pelo tipo de parto em Portugal: análise em componentes principais

Diogo Barrios

Universidade Aberta, Portugal, 1600761@estudante.uab.pt

Humberto S. Machado

Centro Hospitalar e Universitário do Porto, Serviço de Anestesiologia, Porto, Portugal, humbertomachado@chporto.min-saude.pt

Catarina S. Nunes

Universidade Aberta, Departamento de Ciências e Tecnologia, Delegação do Porto, Portugal, CatarinaS.Nunes@uab.pt

Palavras-chave: Análise em Componentes Principais, análise de consistência

Resumo: Portugal tem historicamente uma das taxas mais altas de cesariana da Europa, que tem vindo a aumentar. Um dos fatores frequentemente associado a este aumento, é a preferência da grávida por esta via. O medo da dor, muitas vezes associado ao parto vaginal, é apontado como um dos motivos que leva as grávidas a preferirem uma cesariana eletiva. O objetivo deste estudo foi verificar quais os motivos associados com a preferência pelo tipo de parto em Portugal. Num trabalho anterior foram recolhidos questionários anónimos a 245 mulheres grávidas [1], este questionário averiguava qual o tipo de parto preferido e explorava os motivos para esta preferência. A primeira secção do questionário avaliava a preferência pelo tipo de parto e incluía uma pergunta de resposta fechada sobre o tipo de parto preferido: "Parto vaginal" ou "Cesariana". Consoante a sua opção, as inquiridas eram direccionadas para uma de duas alíneas, em que se exploravam os motivos pela sua preferência. Quanto aos motivos da preferência pelo tipo de parto, cada alínea era composta por dez afirmações e as respostas eram medidas com recurso a uma escala de 5 pontos de Likert, em que 1 correspondia a "totalmente em desacordo" e 5 a "totalmente de acordo" existindo também a hipótese de seleccionar a resposta "não sei" e no final acrescentar outros motivos em resposta aberta. Neste trabalho, foi explorada a aplicação de Análise em Componentes Principais (ACP) aos 10 motivos/itens explorados para a preferência pelo tipo de parto considerando os dois grupos: G1- "Preferência por parto vaginal" e G2- "Preferência por cesariana". Com o objectivo de identificar as componentes mais importantes que influenciam a preferência pelo tipo de parto. A análise da consistência interna através do coeficiente alfa de Cronbrach da escala global dos 10 itens revelou uma fraca consistência interna tanto no grupo G1 como no grupo G2, com valores de 0,55 e 0,6 respetivamente. Calculando a correlação de cada item com o total da escala, excluindo o item em causa, verificou-se que a consistência interna melhoraria significativamente se fossem retirados os itens relativos à existência de um parto prévio em ambos os

grupos, nomeadamente as perguntas "Já tive uma cesariana" e "Já tive um parto vaginal", e o item "A recuperação pós-parto é mais rápida", todos estes 3 itens são comuns aos dois grupos. A ACP foi aplicada à escala com 7 itens. No grupo G1, o teste de Kaiser-Meyer-Olkin (KMO) obteve um valor de 0,74, indicando que existe uma correlação boa entre os itens [2] e teste de esfericidade de Bartlett com $p < 0,05$, leva à rejeição da hipótese da matriz de correlação ser a identidade. Os testes confirmam a adequação dos dados à análise fatorial. Utilizando o critério de Kaiser (mais adequado quando temos menos de 30 itens), tendo sido seleccionados os fatores com valores próprios superiores a 1 [2], os resultados revelaram a existência de 2 componentes principais explicando 53,1% da variância dos resultados. A componente 1 agrega 3 itens, relacionados com menor tempo de internamento e mais tempo com o bebé. A componente 2 agrega os restantes 4 itens mais relacionados com medos tanto da mãe como de prejudicar o bebé. No grupo G2, os resultados foram similares com 2 componentes principais agregando os itens relacionados com a saúde e por outro lado os medos. No entanto, existe um item que precisa de uma análise mais aprofundada "Consigo programar quando será o parto", porque na realidade as pessoas inquiridas não teriam à partida essa hipótese de escolha. No futuro pretende-se cruzar a informação agregada nas componentes principais dos dois grupos e analisar o impacto que estas têm na preferência pelo tipo de parto e incidência de cesarianas em Portugal.

Referências

- [1] Costa, I.C., Nunes C.S., Machado H.S. Mode of Delivery and Labour Analgesia: A Study of Preference in Portuguese Pregnant Women. *Journal of Anesthesia and Clinical Research*, 9: 849, 2018 doi:10.4172/2155-6148.1000849
- [2] Hair, J. F. et al. *Multivariate Data Analysis (7th ed.)*. Pearson Education Limited, London, UK, 2014.

Anexo III

Tabela 2.2: Estatísticas descritivas da Q3_A - Semanas de Gestação e Q3_B - Idade

Sumário	Semanas de Gestação ²			Idade ³		
	Total	G0 - Vaginal	G1 - Cesariana ⁴	Total	G0 - Vaginal	G1 - Cesariana ⁵
Min	6	7	6	18	18	20
1ºQuartil	21	21	20	28.75	28	30
Mediana	25	27	21	33	32	33
Média	25.14	25.76	23.02	31.66	31.31	32.81
3ºQuartil	31.50	32	30	35	35	36.75
Máx	39	39	39	48	48	40
Desvio-Padrão	8.59	8.55	8.55	5.31	5.36	5.02
AIQ	11	11	10	6.25	7	6.75

² *Teste Shapiro-Wilk*, $W_{calculado} = 0.955 > W_{(0.05;244)} = 0.947$, $p < 0.01$

³ *Teste Shapiro-Wilk*, $W_{calculado} = 0.978 > W_{(0.05;244)} = 0.947$, $p < 0.01$

⁴ *Teste de U de Mann-Whitney*, $U = 6002$, $p = 0.03$

⁵ *Teste de U de Mann-Whitney*, $U = 4202$, $p = 0.05$

Tabela 2.3: Tabela do Q3_C - Nível de Escolaridade, por grupos valores absoletos e percentuais

Nível de Escolaridade	Total(%) ⁶	G0(%) ⁷	G1(%) ⁷
Não completou o ensino básico	12(5)	11(6)	1(2)
Ensino básico (Até ao 9º ano, inclusive)	26(11)	22(12)	4(7)
Ensino secundário (até ao 12º ano inclusive)	77(32)	54(29)	23(43)
Licenciatura	85(35)	70(37)	15(28)
Grau superior a licenciatura (Mestrado, Doutoramento)	43(18)	32(17)	11(20)

⁶ *Teste Qui-Quadrado de Pearson*, $\chi^2(4, N = 243) = 5.96$, $p = 0.20$

⁷ *Teste de U de Mann-Whitney*, $U = 4999.5$, $p = 0.81$

Tabela 2.4: Tabela das Fontes de Informação

Fontes de Informação	Total(%)	G0(%) ⁸	G1(%) ⁸	<i>p</i> ⁹
Q2_1 Através do meu médico de família/Obstetra	107(44)	89(0.47)	17(32)	0.04
Q2_1.1 Atraves do meu médico obstetra/de família ou Curso de Preparação para o Parto	112(46)	88(47)	23(43)	0.68
Q2_2 Através do meu marido/parceiro	32(13)	28(15)	4(8)	0.16
Q2_3 Através de um familiar	68(28)	58(31)	10(19)	0.08
Q2_4 Através de um amigo	54(22)	46(24)	8(15)	0.14
Q2_5 Ainda não pensei	39(16)	27(14)	12(23)	0.16
Q2_6 Através da internet	44(18)	36(19)	8(15)	0.49
Q2_7 Através da televisão	17(7)	15(8)	2(4)	0.29
Q2_8 Através de um curso/livros	42(17)	33(17)	9(17)	0.90
Q2_9 Através de jornais/revistas	19(8)	16(8)	3(6)	0.49
Q2_10 Outro	48(20)	37(20)	11(21)	0.88

⁸ Respostas positivas por grupo, (G0, N = 190 e G1, N = 54)

⁹ Teste Qui-Quadrado de Pearson

Tabela 2.5: Tabela da Q1B - Porque preferiria um parto vaginal?

Questões	Mediana	IQR
Q1B1 Não deixa cicatriz	3.00	3.00
Q1B2 Não quero perder a primeira hora de vida do meu bebê	5.00	1.00
Q1B3 Tenho medo	3.00	3.00
Q1B4 Não quero que o meu bebê nasça através de uma cirurgia	3.00	3.00
Q1B5 Já tive um parto vaginal	1.00	4.00
Q1B6 Conheço mais pessoas que tiveram um parto vaginal	4.00	2.00
Q1B7 Quero ter vários filhos e sei que uma cesariana pode limitar o número	3.00	3.00
Q1B8 A recuperação pós-parto é mais rápida	5.00	1.00
Q1B9 O tempo de internamento hospitalar é menor	5.00	1.00
Q1B10 Já tive uma cesariana	1.00	0.00

Tabela 2.6: Tabela da Q1C - Porque preferiria um parto cesariana?

Questões	Mediana	IQR
Q1C1 Tenho medo da dor	4.00	2.00
Q1C2 Consigo programar quando será o parto	4.00	2.00
Q1C3 O meu bebé sofre menos	4.00	2.00
Q1C4 Já tive uma cesariana	4.00	4.00
Q1C5 Após uma cesariana é mais fácil retomar a vida sexual	3.00	1.50
Q1C6 Conheço mais pessoas que tiveram uma cesariana	3.00	3.00
Q1C7 É mais seguro para a minha saúde	3.00	1.00
Q1C8 Tenho medo	4.00	2.00
Q1C9 Já tive um parto vaginal	1.00	1.00
Q1C10 A recuperação pós-parto é mais fácil	2.00	2.00

Anexo IV

Tabela 3.1: Tabela de frequências das questões Q1B1 à Q1B10

Escala Likert	<i>Q1B1</i> (%)	<i>Q1B2</i> (%)	<i>Q1B3</i> (%)	<i>Q1B4</i> (%)	<i>Q1B5</i> (%)	<i>Q1B6</i> (%)	<i>Q1B7</i> (%)	<i>Q1B8</i> (%)	<i>Q1B9</i> (%)	<i>Q1B10</i> (%)
1	29(15.30)	7(3.68)	48(25.30)	40(21.00)	68(35.80)	22(11.60)	49(25.80)	3(1.58)	5(2.63)	88(46.30)
2	18(9.47)	4(2.11)	30(15.8)	21(11.00)	10(0.53)	7(3.68)	20(10.50)	1(0.53)	3(1.58)	0.00
3	39(20.50)	14(7.37)	38(20.00)	26(13.70)	1(0.53)	32(16.80)	32(16.08)	18(9.47)	21(11.00)	1(0.53)
4	30(15.80)	41(21.60)	15(7.89)	23(12.10)	6(3.16)	30(15.80)	8(4.21)	34(17.90)	34(17.90)	0.00
5	53(27.90)	107(56.30)	26(13.70)	52(27.40)	52(27.40)	75(39.50)	30(15.80)	112(59.00)	98(51.6)	22(11.60)
Omissos	21(11.00)	17(8.95)	33(17.40)	28(14.70)	62(32.60)	24(12.60)	51(26.80)	22(11.60)	29(15.30)	79(41.60)

Tabela 3.2: Tabela de omissos por ordem decrescente das questões Q1B1 à Q1B10

Questões	(%)	Legenda
<i>Q1B10</i>	41.58	Já tive uma cesariana
<i>Q1B5</i>	32.63	Já tive um parto vaginal
<i>Q1B7</i>	26.84	Quero ter vários filhos e sei que uma cesariana pode limitar o número
<i>Q1B3</i>	17.37	Tenho medo
<i>Q1B9</i>	15.26	O tempo de internamento hospitalar é menor
<i>Q1B4</i>	14.74	Não quero que o meu bebé nasça através de uma cirurgia
<i>Q1B6</i>	12.63	conheço mais pessoas que tiveram um parto vaginal
<i>Q1B8</i>	11.58	A recuperação pós-parto é mais rápida
<i>Q1B1</i>	11.05	Não deixa cicatriz
<i>Q1B2</i>	8.95	Não quero perder a primeira hora de vida do meu bebé

Tabela 3.10: Tabela de frequências das items Q1C1 à Q1C10

Escala Likert	Q1C1(%)	Q1C2(%)	Q1C3(%)	Q1C4(%)	Q1C5(%)	Q1C6(%)	Q1C7(%)	Q1C8(%)	Q1C9(%)	Q1C10(%)
1	8(14.80)	4(7.41)	6(11.10)	14(25.90)	8(14.80)	12(22.20)	5(9.26)	8(14.80)	21(38.90)	13(24.10)
2	0	5(9.26)	1(1.85)	1(1.85)	5(9.26)	8(14.80)	3(5.56)	3(5.56)	2(3.70)	8(14.80)
3	12(22.20)	12(22.20)	9(16.70)	1(1.85)	11(20.40)	9(16.70)	17(31.70)	9(16.70)	1(1.85)	7(13.00)
4	10(18.50)	9(16.70)	9(16.70)	1(1.85)	5(9.26)	5(9.26)	4(7.41)	7(13.00)	1(1.85)	4(7.41)
5	21(38.90)	19(35.20)	17(31.50)	16(29.60)	3(5.56)	9(16.70)	9(16.70)	19(35.20)	5(9.26)	1(1.85)
Omissos	3(5.56)	5(9.26)	12(22.20)	21(38.90)	22(40.70)	11(20.40)	16(29.60)	8(14.80)	24(44.40)	21(38.90)

Tabela 3.11: Tabela de omissos por ordem decrescente das items Q1C1 à Q1C10

items	(%)	Legenda
Q1C9	44.40	Já tive um parto vaginal
Q1C5	40.70	Após uma cesariana é mais fácil retomar a vida sexual
Q1C4	38.90	Já tive uma cesariana
Q1C10	38.90	A recuperação pós-parto é mais rápida
Q1C7	29.60	É mais seguro para a minha saúde
Q1C3	22.20	O meu bebé sofre menos
Q1C6	20.40	Conheço mais pessoas que tiveram uma cesariana
Q1C8	14.80	Tenho medo
Q1C2	9.26	Conseguo programar quando será o parto
Q1C1	5.56	Tenho medo da dor

Anexo V

Escalonamento dos postos das componentes principais[46]:

- Ponderação das componentes de ambos os grupos G_0 e G_1 :

No caso do grupo G_0 , pela tabela 3.9, temos:

$$- \mathbf{PC}_1 G_1 = 0.8Q1B1 + 0.79Q1B9 + 0.73Q1B8$$

$$- \mathbf{PC}_2 G_1 = 0.77Q1B4 + 0.72Q1B6$$

$$- \mathbf{PC}_3 G_1 = 0.75Q1B7 + 0.74Q1B3$$

No caso do grupo G_1 , pela tabela 3.18, temos:

$$- \mathbf{PC}_1 G_2 = 0.94Q1C7 + 0.68Q1C2 + 0.58Q1C5$$

$$- \mathbf{PC}_2 G_2 = 0.94Q1C8 + 0.91Q1C1$$

$$- \mathbf{PC}_3 G_2 = 0.93Q1C6 + 0.65Q1C3$$

- Obtenção do desvio-padrão $\sigma_{\mathbf{PC}_i G_j}$ e da média $\mu_{\mathbf{PC}_i G_j}$, do vetor da componente principal \mathbf{PC}_i do grupo G_j dentro do intervalo [m,M]:

$$\sigma_{\mathbf{PC}_i G_j} = M / [\max(\mathbf{PC}_i G_j) - \min(\mathbf{PC}_i G_j)] \quad (1)$$

$$\mu_{\mathbf{PC}_i G_j} = m - [\min(\mathbf{PC}_i G_j) * \sigma_{\mathbf{PC}_i G_j}] \quad (2)$$

- Através de transformação linear com o uso das equações 1 e 2 é calculado o vetor da componente escalada \mathbf{S}_i de cada grupo G_j :

$$\mathbf{S}_i G_j = \mu_{\mathbf{PC}_i G_j} + (\sigma_{\mathbf{PC}_i G_j} * \mathbf{PC}_i G_j) \quad (3)$$

sendo $i = 1, 2, 3$ e $j = 1, 2$

Figura 5.1: Histogramas antes e depois de escalados, da *PC1 - vant_parto*, *PC2 - preo_bebe* e *PC3 - medo_parto*, referente ao G0

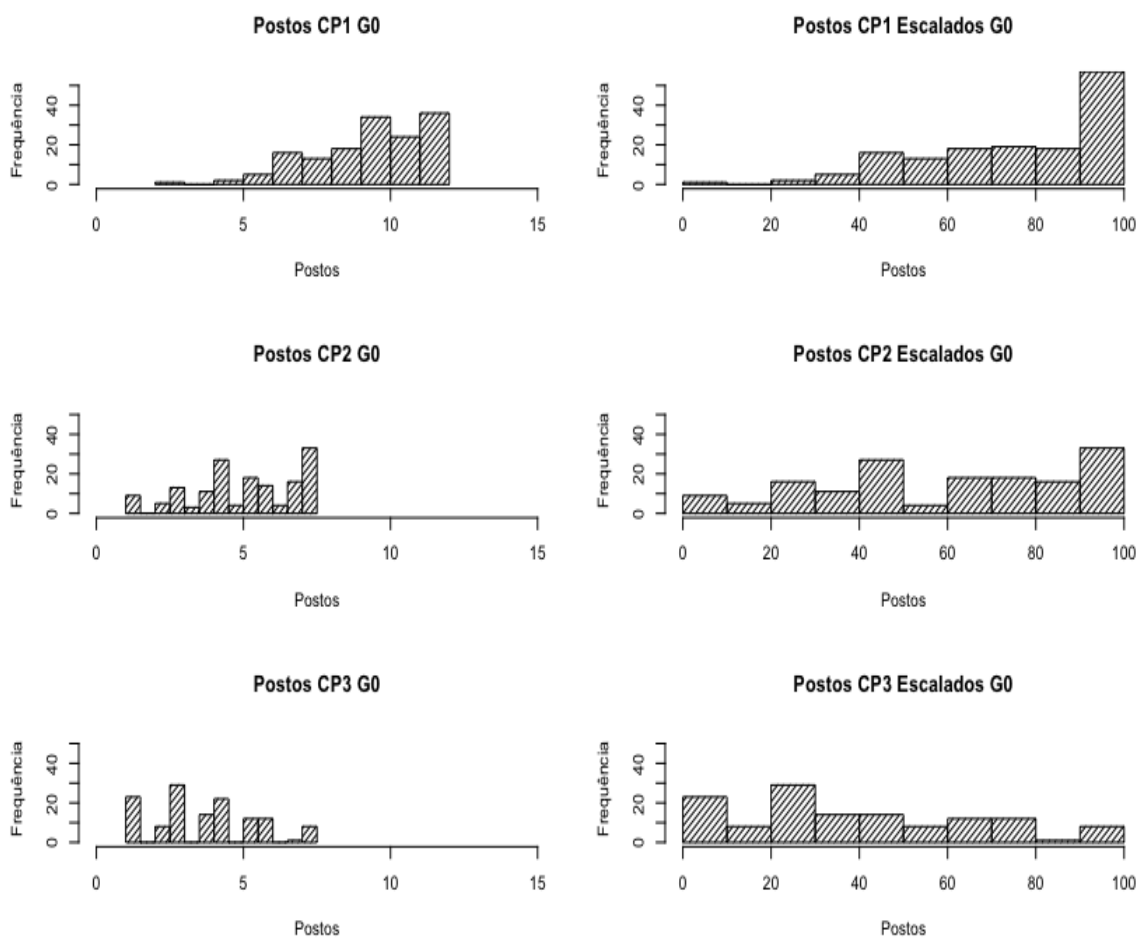


Figura 5.2: Histogramas antes e depois de escalados, da $PC1$ - *vant_parto*, $PC2$ - *medo_parto* e $PC3$ - *preo_bebe*, referente ao G1

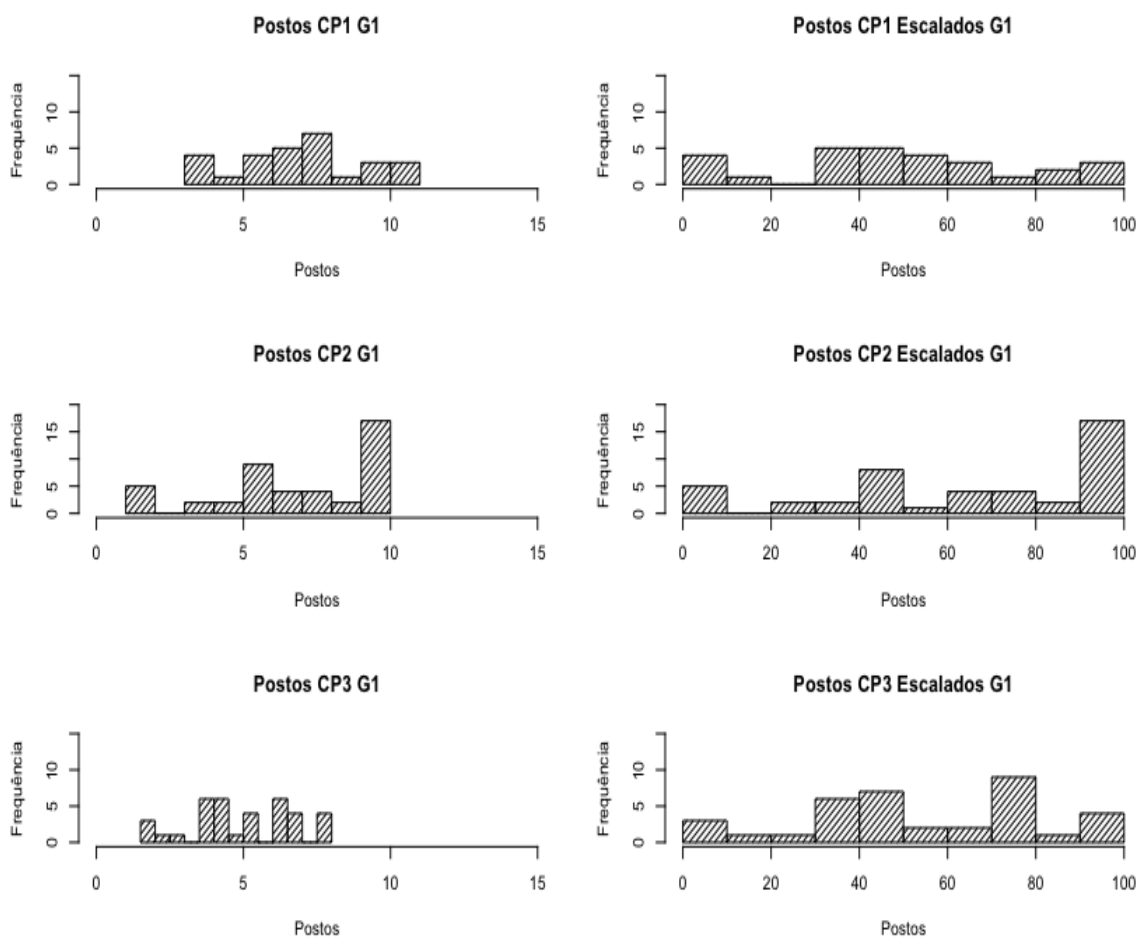


Figura 5.4: Ponto de corte ótimo, $c = 0.224$

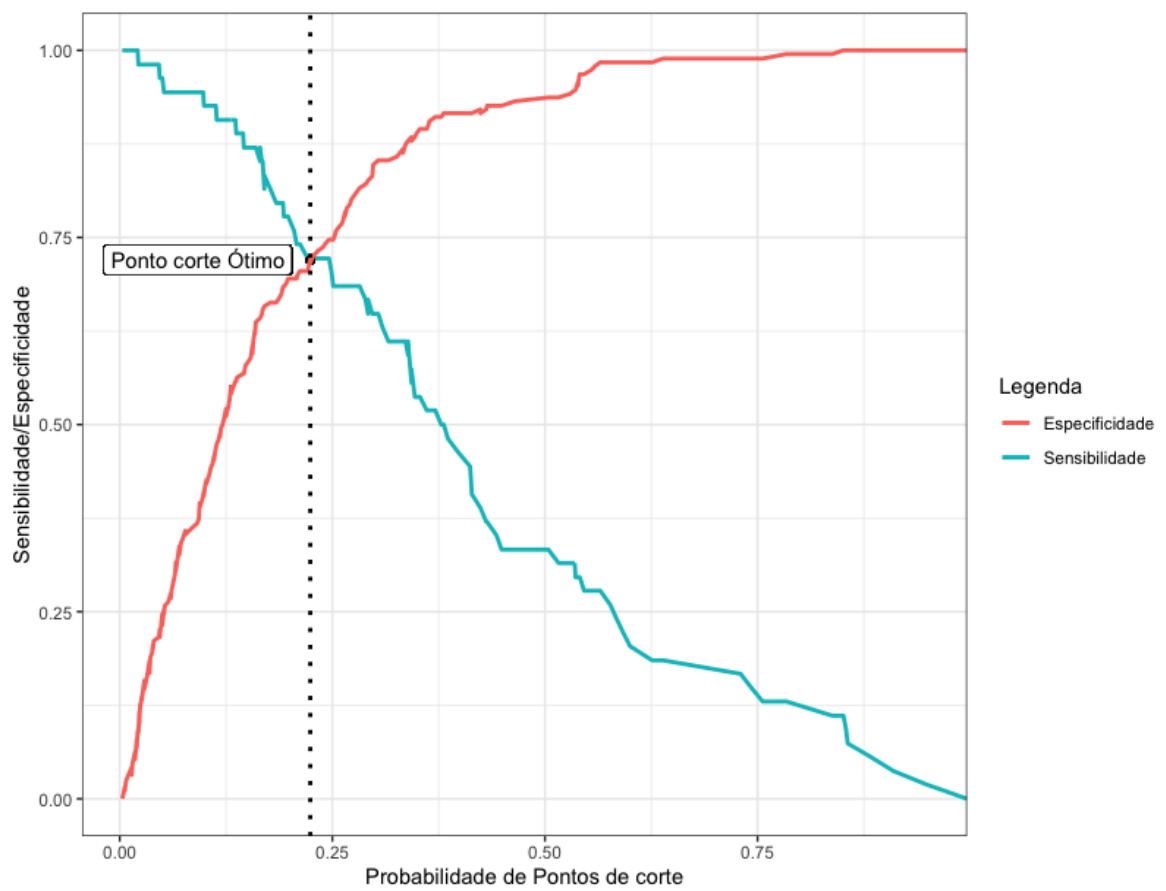


Tabela 5.1: Tabela de contingência entre Q1A e as variáveis dicotômicas

	Q1A		p^*
	0 - Parto Vaginal	1 - Parto Cesariana	
<u>Q2_1</u>			0.04
0	100	90	
1	37	17	
<u>Q2_1.1</u>			0.63
0	102	88	
1	31	23	
<u>Q2_2_4</u>			0.02
0	124	66	
1	44	10	
<u>Q2_3</u>			0.08
0	132	58	
1	44	10	
<u>Q2_6_7</u>			0.43
0	149	41	
1	45	9	
<u>Q2_5</u>			0.16
0	163	27	
1	42	12	
<u>Q2_8_9</u>			0.95
0	152	38	
1	43	11	
<u>Q2_10</u>			0.88
0	153	37	
1	43	11	
<u>Q3_D1</u>			0.86
0	110	80	
1	32	22	

* *Teste Qui-Quadrado de Pearson*

Tabela 5.3: Exclusão de variáveis com $p > 0.05$ do modelo logístico L_1

Modelo Logístico: Variável excluída	D	G	p	AIC	BIC
	197.47	60.46	< 0.01	223.47	268.94
L_2 : $Q2_5$	197.47	0.00	0.98	221.47	263.44
L_3 : $Q3_A$	198.13	0.65	0.42	220.13	258.59
L_4 : $Q3_C$	203.53	5.41	0.25	217.53	242.01

Tabela 5.5: $\Delta\% \hat{\beta}$ do L_4 em relação ao L_1

Variável	$\hat{\beta}_4$	$\hat{\beta}_1$	$\Delta\% \hat{\beta}$
$Q2_1$	-0.854	-0.802	-0.06%
$Q2_3$	-0.933	-0.904	-0.03%
$Q2_2_4$	-0.845	-0.727	-13.96%
$Q3_B$	0.069	0.089	28.98%
$vant_parto$	-0.046	-0.045	-0.02%
$medo_parto$	0.031	0.029	-0.06%

Tabela 5.6: Inclusão das variáveis com $p > 0.25$ no modelo logístico L_4 , com a exclusão das variáveis $Q2_5$, $Q3_A$ e $Q3_C$

Modelo Logístico: Variável incluída	Variável				Modelo Logístico				
	$\hat{\beta}$	Erro-Padrão	z	p	D	G	p	AIC	BIC
					203.53	5.41	0.25	217.53	242.01
L_5 : $Q2_1.1$	-0.449	0.390	-1.152	0.25	202.19	1.34	0.25	218.19	246.16
L_6 : $Q2_6_7$	0.068	0.468	0.143	0.89	203.51	0.02	0.89	219.51	247.49
L_7 : $Q2_8_9$	0.01	0.452	0.012	0.99	203.53	0.00	0.99	219.53	247.51
L_8 : $Q2_10$	-0.610	0.489	-1.246	0.21	201.92	1.61	0.20	201.92	245.90
L_9 : $Q3_D1$	-0.353	0.394	-0.897	0.37	202.72	0.81	0.37	218.72	246.70
L_{10} : $preo_bebe$	0.013	0.007	1.926	0.05	199.66	3.88	0.05	215.66	243.63

Tabela 5.8: $\Delta\% \hat{\beta}$ do L_{10} em relação ao L_4 , com a inclusão da variável *preo_bebe*

Variável	$\hat{\beta}_{10}$	$\hat{\beta}_4$	$\Delta\% \hat{\beta}$
<i>Q2_1</i>	-0.900	-0.854	-5.11%
<i>Q2_3</i>	-1.007	-0.933	-7.35%
<i>Q2_2_4</i>	-0.897	-0.845	-5.80%
<i>Q3_B</i>	0.074	0.069	-6.76%
<i>vant_parto</i>	-0.050	-0.046	-8.00%
<i>medo_parto</i>	0.030	0.031	3.33%

Tabela 5.9: Inclusão das interações ao modelo logístico L_{10}

Modelo Logístico: Interação incluída	Variável				Modelo Logístico				
	$\hat{\beta}$	Erro-Padrão	z	p	D	G	p	AIC	BIC
					199.66	3.88	0.05	215.66	243.63
L_{11} : <i>Q2_1*Q2_3</i>	-0.495	1.005	-0.492	0.62	199.40	0.25	0.62	217.40	248.88
L_{12} : <i>Q2_1*Q2_2_4</i>	-2.117	1.246	-1.700	0.09	195.88	3.77	0.05	213.88	245.36
L_{13} : <i>Q2_1*Q3_B</i>	0.051	0.074	0.683	0.49	199.18	0.47	0.49	217.18	248.66
L_{14} : <i>Q2_1*vant_parto</i>	-0.024	0.018	-1.349	0.18	197.81	1.85	0.17	215.81	247.28
L_{15} : <i>Q2_1*medo_parto</i>	0.03	0.015	0.200	0.84	199.62	0.04	0.84	217.62	249.09
L_{16} : <i>Q2_1*preo_bebe</i>	-0.020	0.015	-1.381	0.17	197.76	1.90	0.17	215.76	247.23
L_{17} : <i>Q2_3*Q2_2_4</i>	0.670	0.978	0.716	0.47	199.14	0.51	0.47	217.14	248.62
L_{18} : <i>Q2_3*Q3_B</i>	-0.038	0.084	-0.456	0.65	199.45	0.21	0.65	217.45	248.92
L_{19} : <i>Q2_3*vant_parto</i>	-0.021	0.023	-0.936	0.35	198.73	0.92	0.34	216.73	248.21
L_{20} : <i>Q2_3*medo_parto</i>	-0.025	0.020	-1.244	0.21	198.02	1.64	0.20	216.02	247.49
L_{21} : <i>Q2_3*preo_bebe</i>	0.030	0.019	1.561	0.12	196.99	2.67	0.10	214.99	246.46
L_{22} : <i>Q2_2_4*Q3_B</i>	-0.066	0.080	-0.816	0.41	199.01	0.65	0.42	217.01	248.48
L_{23} : <i>Q2_2_4*vant_parto</i>	0.028	0.022	1.287	0.20	197.93	1.73	0.19	215.93	247.40
L_{24} : <i>Q2_2_4*medo_parto</i>	0.000	0.016	-0.002	0.99	199.66	0.00	0.99	217.66	249.13
L_{25} : <i>Q2_2_4*preo_bebe</i>	0.005	0.016	0.334	0.74	199.54	0.11	0.74	217.54	249.02
L_{26} : <i>Q3_B*vant_parto</i>	0.000	0.000	-0.039	0.97	199.65	0.00	0.97	217.65	249.13
L_{27} : <i>Q3_B*medo_parto</i>	0.000	0.001	0.502	0.62	199.40	0.25	0.61	217.40	248.88
L_{28} : <i>Q3_B*preo_bebe</i>	0.000	0.001	0.345	0.73	199.54	0.12	0.73	217.54	249.10
L_{29} : <i>vant_parto*medo_parto</i>	0.000	0.000	-1.486	0.14	197.33	2.33	0.13	215.33	246.80
L_{30} : <i>vant_parto*preo_bebe</i>	0.000	0.000	-1.044	0.30	198.57	1.09	0.30	216.57	248.04

Tabela 5.11: Percentis das probabilidades estimadas

Decil	Prob.	Y = 1		Y = 0		n	Pr o b.*
		y	\hat{y}	y	\hat{y}		
1	0.0243	1.00	0.41	24.00	24.59	25	0.02
2	0.0496	1.00	0.89	23.00	23.11	24	0.04
3	0.0908	1.00	1.65	24.00	23.35	25	0.07
4	0.118	2.00	2.50	22.00	21.50	24	0.10
5	0.156	2.00	3.25	22.00	20.75	24	0.14
6	0.208	7.00	4.40	18.00	20.60	25	0.18
7	0.272	3.00	5.63	20.00	17.37	23	0.24
8	0.353	9.00	8.19	17.00	17.81	26	0.31
9	0.539	12.00	10.47	12.00	13.53	24	0.44
10	0.949	16.00	16.62	8.00	7.38	24	0.69

* Probabilidade média

Tabela 5.14: $\Delta\%\hat{\beta}$ do L_{31} em relação ao L_{10} , com a exclusão dos casos 198, 203, 215, 240

Variável	$\hat{\beta}_{31}$	$\hat{\beta}_{10}$	$\Delta\%\hat{\beta}$
	casos excluídos	todos os casos	
<i>Q2_1</i>	-0.934	-0.900	-3.64%
<i>Q2_3</i>	-1.183	-1.007	-14.88%
<i>Q2_2_4</i>	-1.263	-0.897	-28.98%
<i>Q3_B</i>	0.112	0.074	-33.93%
<i>vant_parto</i>	-0.062	-0.050	-19.35%
<i>medo_parto</i>	0.037	0.030	-18.92%
<i>preo_bebe</i>	0.021	0.013	-38.10%