Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Regressão Logística Multinomial Classificação de Contraceptivo

André G. P. de Oliveira
Crystiane Fernanda de Souza
Douglas de Paula Nestlehner

Abril, 2022

Sumário

Capítulo 1

Banco de Dados

1.1 Descrição do Banco e Objetivo

O banco de dados utilizado nesse trabalho foi retirado do link `https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice`, ele contém 10 variáveis e 1473 observações.

1.2 Variáveis

As variáveis presentes no conjunto de dados são as seguintes:

1. Idade da mulher (numérica);
2. Nível educação da mulher (categórica) 1= mais baixo, 2, 3, 4 = mais alto;
3. Nível educação do esposo (categórica) 1= mais baixo, 2, 3, 4 = mais alto;
4. Número de filhos da mulher (numérica);
5. Religião da mulher (binária) 0=Não-islâmica, 1=Islâmica;
6. Se a mulher trabalha ou não (binária) 0=Sim, 1=Não;
7. Ocupação do marido (categórica) 1, 2, 3, 4;
8. Nível da qualidade de vida (categórica) 1=mais baixo, 2, 3, 4 = mais alto;
9. Exposição a mídia (binária) 0=Boa, 1=Não boa;
10. Tipo de contraceptivo utilizado (categórica) 1 = Não usa (N), 2 = Uso de curto prazo(CP), 3 = Uso de longo prazo (LP);

A variável resposta é o tipo de contraceptivo utilizado com todas as outras sendo consideradas preditoras.

Nesse sentido, o objetivo desse trabalho é classificar qual tipo de contraceptivo a mulher usa. Nesse caso, como temos um problema multinomial (ou multiclasse), podemos aplicar uma regressão logística multinomial (também chamada de regressão *softmax*).

1.3 Análise Descritiva

Como dito anteriormente, a base de dados utilizada contem 10 variáveis (sendo uma a variável resposta) e 1473 observações. Na Tabela 1.1 temos representadas as quatro primeiras e últimas linhas da base de dados.

Tabela 1.1: Base de dados.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
1	45	1	3	10	1	1	3	4	0	1
2	43	2	3	7	1	1	3	4	0	1
3	42	3	2	9	1	1	3	3	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1470	39	3	3	8	1	0	1	4	0	3
1471	33	3	3	4	1	0	2	2	0	3
1472	17	3	3	1	1	1	2	4	0	3

Para as variáveis quantitativas de Idade da mulher (X_1) e Número de Filhos (X_4) calculamos algumas medidas descritivas, representadas na Tabela 1.2, no intuito de verificar possíveis anomalias.

Tabela 1.2: Medidas descritivas das variáveis X_1 e X_4.

	Idade_Mulher	Num_Filhos
Min.	16.000	0.000
1st Qu.	26.000	1.000
Median	32.000	3.000
Mean	32.544	3.262
3rd Qu.	39.000	4.250
Max.	49.000	16.000

Não encontramos nada de anormal, porém achamos interessante observar o comportamento/frequência delas, o qual esta representado na Figura 1.1:

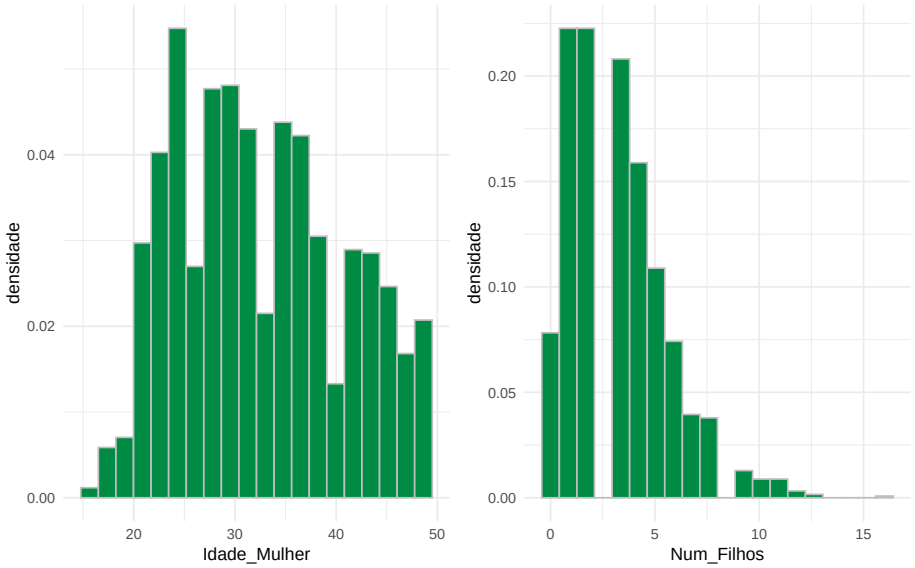

Figura 1.1: Histogramas das variáveis quantitativas.

Também achamos interessante observar as frequências das variáveis qualitativas, representadas na Figura 1.2:

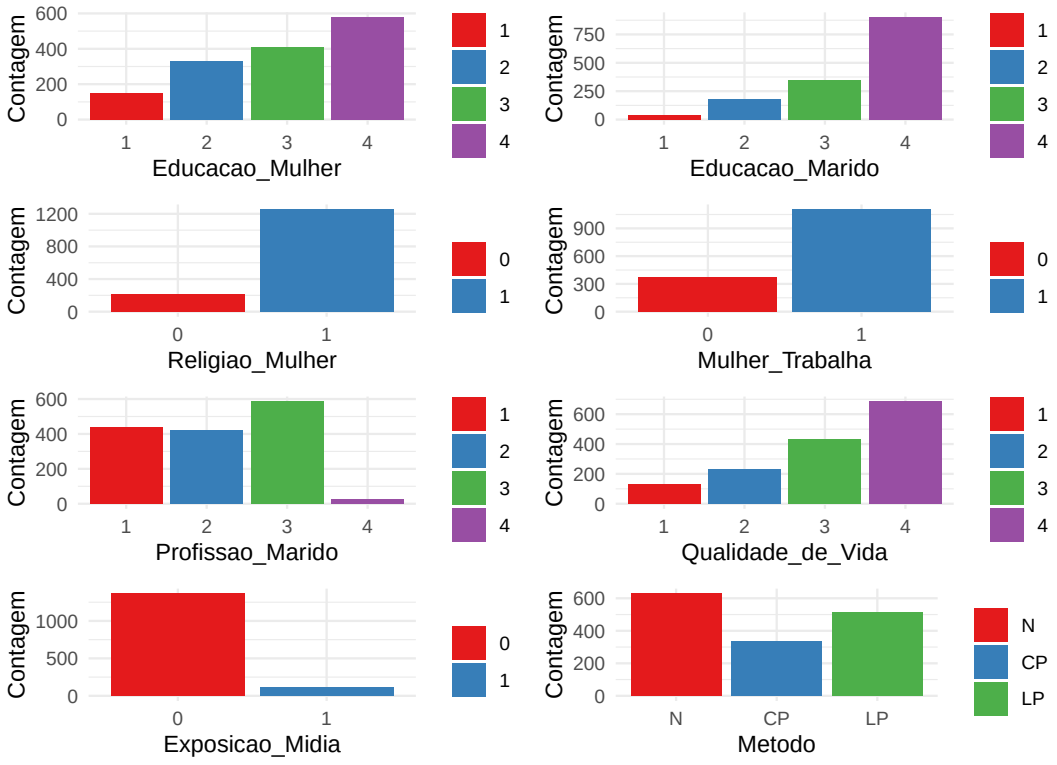

Figura 1.2: Histogramas das variáveis qualitativas.

A partir da Figura 1.2 é possível observar que grande parte das mulheres possuem uma alta escolaridade, fazem parte da religião islâmica, trabalham e possuem uma alta qualidade de vida.

Na Tabela 1.3 temos a contagem dos métodos contraceptivos utilizados.

Tipo de anticoncepcional	Contagem
Não Usa	628
Curto Prazo	333
Longo Prazo	511

Em relação a variável resposta que é o tipo de método contraceptivo utilizado, é possível observar um alto número de mulheres que não utilizam nenhum método contraceptivo.

Capítulo 2

Modelo

2.1 Definição do Modelo

O objetivo neste trabalho é ajustar um Modelo Logístico Multinomial, que é uma extensão do Modelo Logit Binário. Esse modelo será ajustado, pois a variável resposta trata-se sobre o tipo de método contraceptivo utilizado, em que possui 3 níveis, sendo: 1 - Não Usa (N), 2 - Uso de Curto Prazo (CP) e 3 - Uso de Longo Prazo (LP). Separamos 70% das observações para a construção do modelo e 30% para avaliação do poder preditivo.

2.1.1 Distribuição da Variável Resposta

A partir do objetivo, têm que a distribuição da variável resposta, será a Distribuição Multinomial, em que,

$$Y \sim Multinomial(1473, p).$$

2.1.2 Componente Sistemático

Como componente sistemático, teremos as covariáveis: Idade da mulher (X_1), nível de educação da mulher (X_2), nível de educação do marido (X_3), número de filhos da mulher (X_4), religião da mulher (X_5), indicador se a mulher trabalha ou não (X_5), ocupação do marido (X_7), nível de qualidade de vida (X_8) e exposição à mídia (X_9).

2.1.3 Função de Ligação

Como estamos tratando de um Modelo Logístico para uma Resposta Multinomial, a função de ligação canônica será dada por,

$$logito[\pi_k] = ln\left(\frac{\pi_k}{\pi_K}\right) = x'\beta_k$$

2.1.4 Suposições

As suposições necessárias para a definição de um MLG são:

- Ausência de multicolinearidade;
- Ausência de heterocedasticidade;
- Valor esperado dos resíduos igual a zero;
- Relação linear entre o vetor das variáveis explicativas e o logit da variável resposta;
- Confirmação sobre o uso função de ligação certa.

2.2 Seleção de Variáveis

2.2.1 Stepwise

Para aplicação da seleção de variáveis, inicializamos o modelo só com intercepto e buscamos o o modelo com menor AIC. O AIC (Akaike Information Criteria) é uma métrica bastante popular para comparação e seleção de modelos, definida como

$$\text{AIC}(modelo, p) = 2p - 2\log(l(\theta)),$$

em que p é o número de parâmetros e $\log(l(\theta))$ a log-verossimilhança do modelo. Esse critério penaliza o número de parâmetros do modelo e geralmente, analistas escolhem o modelo com base, em partes, no menor AIC ou outros critérios parecidos como o BIC etc.

No *R* a regressão Stepwise é implementada na função `step` que é uma versão menos complexa da função `StepAIC` da biblioteca `MASS`. Essa função trabalha por *default*, em ambas as direções. Para escolhermos o melhor modelo, como já dito, começamos com o modelo só com intercepto e consideramos adicionar todas as variáveis independentes assim como todas as interações dois a dois possíveis.

2.3 Análise de Diagnóstico

Para iniciar a análise de diagnóstico, foi realizado o gráfico do Envelope na Figura 2.1, para verificar se o modelo é adequado.

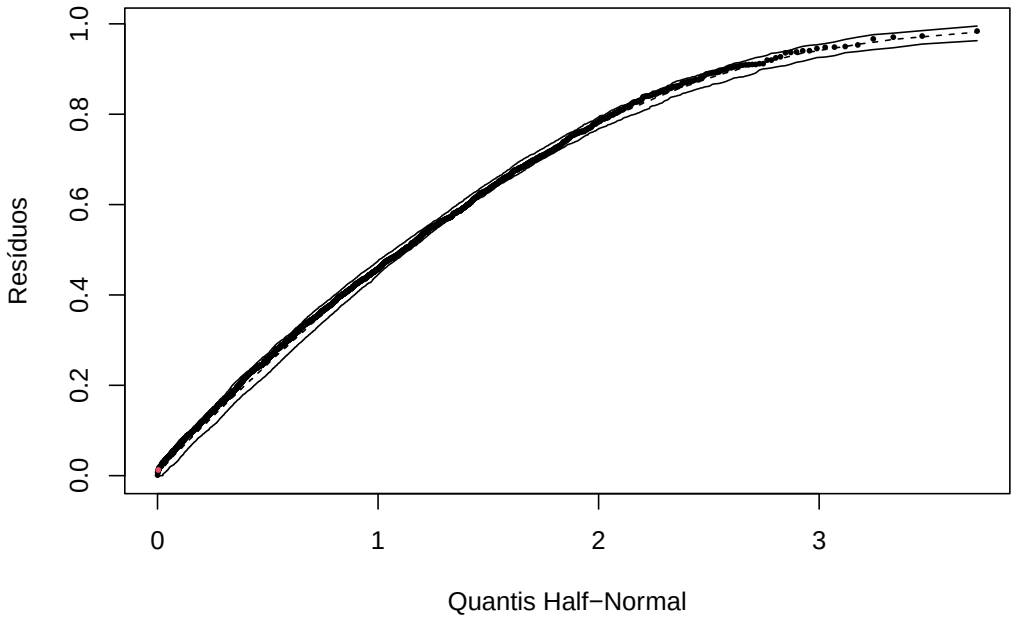

Figura 2.1: Envelope para o modelo ajustado

Nesse contexto, na Figura 2.1 é possível notar que todos os pontos estão dentro do limite, ou seja, todos os pontos estão dentro do envelope. Assim, trazendo indícios de que o modelo foi bem ajustado.

Na Tabela 2.1 temos o Teste da Deviance, em que temos o modelo nulo e o modelo ajustado. Com esse teste, a ideia é comparar a diferença dos desvios de cada modelo e verificar se o modelo que está sendo ajustado é significativo, ou seja, se o modelo ajustado é melhor do que o modelo com apenas o intercepto.

Tabela 2.1: Teste da Deviance.

	Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	Só com intercepto	2058.0000	2198.4836				
2	Ajustado	2024.0000	1859.6658	1 vs 2	34.0000	338.8178	0.0000

Sendo assim, a partir da Tabela 2.1 temos que o valor-p é aproximadamente 0, mostrando então que podemos seguir com esse modelo para as próximas etapas.

A análise de deviance (ANODEV) é uma generalização para a análise de variância, para os modelos lineares generalizados, em que o interesse é testar a significância da inclusão de novos termos. Nesse sentido, será feita a ANODEV para o modelo ajustado, representado na Tabela 2.2, com o intuito de checar a qualidade do modelo ajustado.

Tabela 2.2: ANODEV.

	LR Chisq	Df	Pr(>Chisq)
Educacao_Mulher	60.1695	6.0000	0.0000
S_Idade_Mulher	87.0994	2.0000	0.0000
S_Num_Filhos	11.3849	2.0000	0.0034
Educacao_Marido	20.6137	6.0000	0.0022
Religiao_Mulher	7.9941	2.0000	0.0184
Profissao_Marido	13.1541	6.0000	0.0407
Exposicao_Midia	5.0625	2.0000	0.0796
S_Idade_Mulher:S_Num_Filhos	24.2457	2.0000	0.0000
Educacao_Mulher:S_Num_Filhos	27.3101	6.0000	0.0001

Ao analisar a Tabela 2.2 é possível identificar que ao nível de significância de 5%, que apenas a variável de exposição à mídia (X_9) será não significativa, para explicar a variável resposta que é o tipo de método contraceptivo utilizado. Contudo, como acreditamos que essa covariável é importante para o modelo, seguiremos com o ajuste do modelo considerando todas as variáveis preditoras.

2.4 Modelo Ajustado

Por fim, como o modelo ajustado passou pelo crivo da análise de diagnóstico, temos o ajuste final com as estimativas, erro padrão, e valor-p para os tipos de contraceptivos utilizados, de curto e longo prazo, que estão representados na Tabela 2.3.

Tabela 2.3: Sumário do modelo ajustado.

Coeficiente	Categoria "CP"			Categoria "LP"		
	Estim	Erro	P($>\|z\|$)	Estim	Erro	P($>\|z\|$)
(Intercept)	0.2246	0.6802	0.7412	-2.0253	0.8700	0.0199
Educacao_Mulher2	0.1224	0.5633	0.8280	-0.1565	0.3360	0.6415
Educacao_Mulher3	1.1672	0.5441	0.0319	0.2529	0.3460	0.4648
Educacao_Mulher4	2.3476	0.5657	0.0000	1.2853	0.3835	0.0008
S_Idade_Mulher	-0.5375	0.1260	0.0000	-1.0035	0.1164	0.0000
S_Num_Filhos	0.5461	0.3504	0.1191	0.7645	0.2328	0.0010
Educacao_Marido2	-1.4751	0.6201	0.0174	1.8092	0.8005	0.0238
Educacao_Marido3	-1.1172	0.5513	0.0427	1.9758	0.7937	0.0128
Educacao_Marido4	-0.8771	0.5484	0.1098	1.8776	0.7970	0.0185
Religiao_Mulher1	-0.7021	0.2503	0.0050	-0.2700	0.2420	0.2646
Profissao_Marido2	-0.5190	0.2524	0.0397	-0.0152	0.2323	0.9477
Profissao_Marido3	-0.4902	0.2519	0.0517	0.2359	0.2257	0.2958
Profissao_Marido4	-0.3691	0.8459	0.6626	0.7678	0.5509	0.1634
Exposicao_Midia1	-0.9147	0.4990	0.0668	-0.5315	0.3243	0.1012
S_Idade_Mulher:S_Num_Filhos	-0.3400	0.1069	0.0015	-0.4354	0.0946	0.0000
Educacao_Mulher2:S_Num_Filhos	0.4708	0.4007	0.2400	0.2795	0.2647	0.2909
Educacao_Mulher3:S_Num_Filhos	0.2785	0.3659	0.4465	0.0407	0.2610	0.8760
Educacao_Mulher4:S_Num_Filhos	1.3357	0.4006	0.0009	1.0903	0.3084	0.0004

Capítulo 3

Métricas de Eficiência do Modelo e Interpretações

3.1 Interpretação do Modelo

A Figura 3.1 mostra as médias de probabilidade ajustadas com relação às variáveis presentes no modelo.

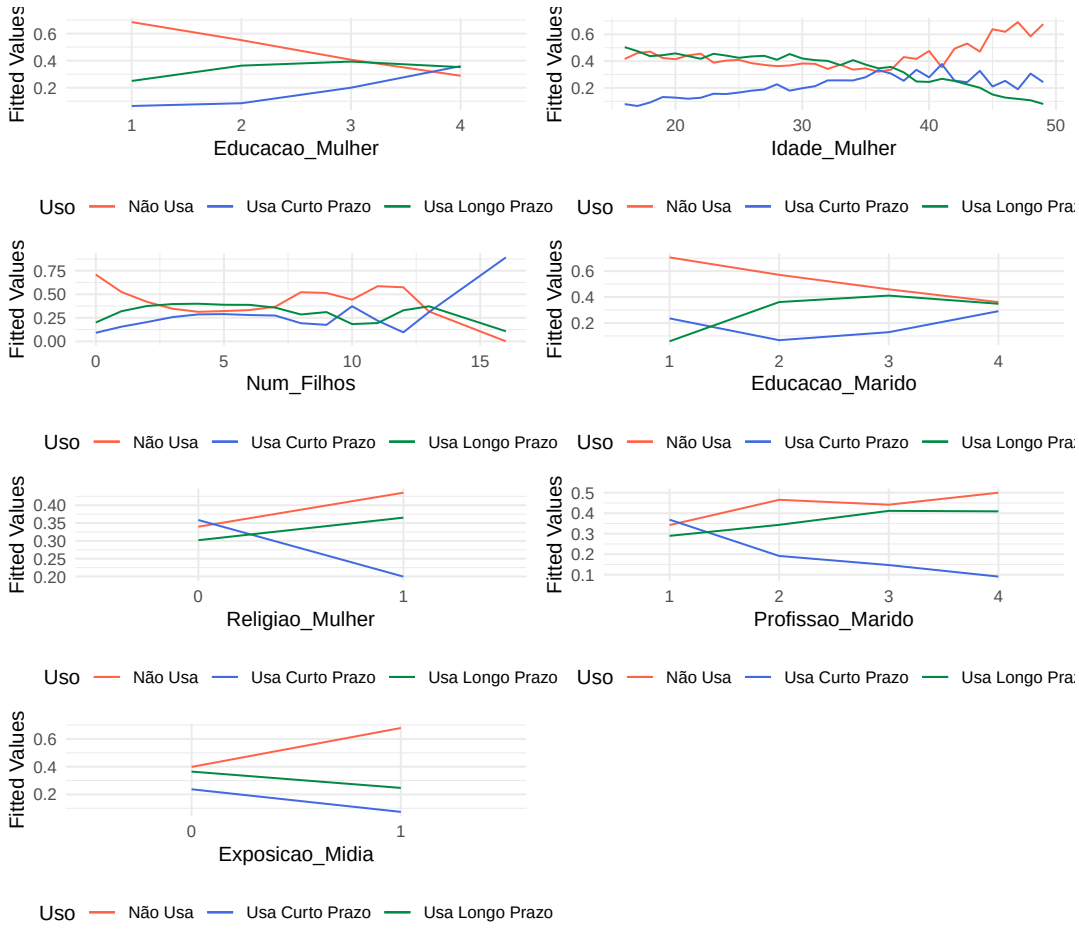

Figura 3.1: Probabilidades ajustadas por cada nível.

Observando alguns dos comportamentos mais marcantes, quanto maior o nível educacional da mulher menor a probabilidade dela não usar anticoncepcionais (reta vermelha), esse padrão também ocorre conforme o nível de educação do marido aumenta. Com relação a idade, temos que mulheres mais velhas tendem a não usar nenhum tipo de contraceptivo, o que faz sentido, visto que a chance de gravidez diminui conforme a idade avança para os 40 ou mais anos.

Por fim, a religião também tem um efeito considerável, com mulheres de religião islâmica possuindo probabilidades bem menores de fazerm uso de anticoncepcionais, em especial os de curto prazo (reta azul). Já a exposição a mídia tem efeito parecido (mulheres com valor 1 são pouco expostas), com mulheres que não tem muito acesso a mídia possuindo chances menores de usar tanto contraceptivos de curto e longo prazo.

Dado que estamos em modelo *log-linear*, podemos calcular as razões de chances (exponencial das *log-odds*). A Tabela 3.1 mostra as *odds* calculadas considerando a que a classe de referência é o não uso de contraceptivos (classe "N").

Tabela 3.1: *Odds* dos coeficientes ajustados.

Coeficiente	CP	LP
(Intercepto)	1.2518	0.1320
Educacao_Mulher2	1.1302	0.8552
Educacao_Mulher3	3.2129	1.2878
Educacao_Mulher4	10.4607	3.6157
S_Idade_Mulher	0.5842	0.3666
S_Num_Filhos	1.7265	2.1479
Educacao_Marido2	0.2288	6.1053
Educacao_Marido3	0.3272	7.2124
Educacao_Marido4	0.4160	6.5380
Religiao_Mulher1	0.4955	0.7634
Profissao_Marido2	0.5951	0.9849
Profissao_Marido3	0.6125	1.2661
Profissao_Marido4	0.6913	2.1550
Exposicao_Midia1	0.4006	0.5877
S_Idade_Mulher:S_Num_Filhos	0.7118	0.6470
Educacao_Mulher2:S_Num_Filhos	1.6012	1.3225
Educacao_Mulher3:S_Num_Filhos	1.3212	1.0416
Educacao_Mulher4:S_Num_Filhos	3.8028	2.9751

De todos os valores calculados, chamamos atenção para algumas variáveis. No geral,

- Quanto maior o nível educacional da mulher, maior as chances dela utilizar contraceptivos de curto prazo, começando de 113% a mais de chance para um nível acima da referência e terminando com 1047% a mais para as mulheres com nível educacional maior. Com relação a longo prazo (LP), o padrão é semelhante com

o nível educacional aumentando a chance de uso. Ressaltamos que esta variável se confunde tanto com a Educação do Marido, e geralmente estão relacionadas, mas como são variáveis muito importantes para termos informação do meio social de onde a mulher vive, não achamos razoável tirar elas.

- A religião também tem um efeito considerável no uso, pois quando a mesma é 1 (significando religião islâmica), a chance da mulher utilizar anticoncepcionais também caem.

- A variável Exposição Mídia tem efeito semelhante a da religião, com mulheres pouco expostas (pelos critérios definidos pelos coletadores da base de dados) apresentando menor uso tanto a curto como longo prazo.

- Por fim, a variável idade de mulher tem efeito de decréscimo conforme a idade cresce, o que é esperado pois conforme a mulher é mais velha o risco de gravidez diminui e ela não precisa fazer tanto uso de anticoncepcionais.

3.2 Predição e métricas

Utilizando o modelo ajustado para predição, consideramos o conjunto de dados com 30% dos dados. Ressaltamos que na regressão logística multinomial a observação i é classificada no grupo j onde o π_{ij} for maior (ou seja uma observação com probabilidades preditas = (0.1,0.5,0.4) fará parte do segundo grupo, por exemplo), comumente não se faz uso de regras de corte via curva ROC etc.

A Tabela 3.2 concentra a matriz de confusão de valores preditos contra reais.

Tabela 3.2: Matriz de confusão.

	Preditos		
Reais	N	CP	LP
N	133	15	47
CP	39	31	32
LP	42	23	80

Porém, a partir da Tabela 3.2 fica difícil avaliar a qualidade de predição, desse modo podemos calcular algumas das métricas comuns em classificação. Considerando as seguintes medidas:

- Sensibilidade: Definida como o número de observações da classe classificadas corretamente divididas pelo número de elementos da classe.

- Especificade: Definida como o percentual de observações que não são dessa classe classificadas corretamente como não parte da mesma.

- VPP (verdadeiros preditos positivos): Porcentagem de valores atribuídos a classe que realmente fazem parte da mesmsa.
- VPN (verdadeiros preditos negativos): Porcentagem de valores atribuídos corretamente que não são da classe em questão que foram classificados em outras classes.

Mais detalhes sobre essas medidas e como calculamos as mesmas no `R` podem ser encontradas em [1].

Tabela 3.3: Medidas de qualidade para cada classe

	Sensibilidade	Especificidade	VPP	VPN
Class: N	0.6821	0.6721	0.6215	0.7281
Class: CP	0.3039	0.8882	0.4493	0.8097
Class: LP	0.5517	0.7340	0.5031	0.7703

Notamos que no geral, a classe melhor classificada positivamente (em termos de acertos) é a que utilizamos de referência, N (não uso de contraceptivos), com sensibilidade de 68.21% e VPP de 62.15%, isto é, 68.21% das mulheres que realmente não usam contraceptivo foram classificadas corretamente e, de todas as mulheres que nosso modelo classificou como não usuárias de contraceptivo, 62.15% realmente não usam.

A classe intermediária, CP (uso a curto prazo), não é muito bem classificada pelo modelo, muito provavelmente pois é pouco numerosa perto das outras. Apenas 30% das mulheres nessa classe foram classificadas corretamente e de todas que o modelo colocou nessa classe, menos da metade são realmente usuárias de métodos de curto prazo.

Já a classe de longo prazo (LP) possui sensibilidade de 55.17% e VPP de 50.31%, ou seja, de todas as mulheres realmente dessa classe 55.17% o modelo corretamente agrupou, e de todas as mulheres agrupadas por esse modelo nesta classe, 50.31% realmente fazem parte. Seus valores de especifidade são de 73.40% (percentual de mulheres que não usam contraceptivos de longo prazo que o modelo colocou em outras classes) e VPN de 77.03% (percentual de mulheres que o modelo disse não usar a longo prazo que realmente não usam).

Para métricas gerais do modelo, podemos considerar 2,

- Acurácia: Simplesmente quantos acertos o modelo teve, isto é a soma da diagonal principal de 3.2 dividido pelo total de indivíduos.
- Escore F1: Média harmônica entre VPP e VPN (também chamadas de *Precision* e *Recall*) na comunidade de *Machine Learning*.

Tabela 3.4: Métricas gerais

	Valor
Acurácia	0.5520
F1	0.5131

As duas métricas discutidas vão de 0 a 1, porém possuem interpretações diferentes. A acurácia do modelo para o conjunto de testes é de 55.20% (considerando pesos iguais), indicando que pouco mais da metade das observações foram corretamente classificadas. Ressalto que se o classificador fosse aleatório, a Acurácia seria de 33.33%.

Por fim, o escore F1, muito usado nas comunidades de *Machine Learning*, tem valor de 0.5131 e sua interpretação é um pouco mais nebulosa. Em suma, podemos dizer que nosso classificador baseado na regressão logística multinomial é minimamente decente para a predição (um tanto melhor que um classificador aleatório), entretanto o real valor está também na inferência realizada, em especial nos cálculos da razões de chances, ou *odds ratios*.

Capítulo 4

Conclusão

Devido ao fato da variável resposta presente nos dados trabalhados conter 3 classes, fizemos uso do modelo linear generalizado multinomial. Selecionamos as variáveis dentre todas as interações 2 a 2 possíveis a partir do método de *stepwise*, obtendo um modelo com 36 parâmetros, 18 relacionados à cada classe da variável resposta.

A partir disso, fizemos o diagnóstico do modelo ajustado, com o envelope apresentado resultados satisfatórios, não indicando que o modelo está mal ajustado. Feito isso, partimos para a inferência e interpretação dos coeficientes calculados, calculando as probabilidades preditas e razões de chances (*odds ratio*) para cada variável e comparando se as informações fazem sentido e avaliando o impacto do valor.

Por fim, separamos 30% dos dados não considerados para o ajuste do modelo e calculamos as métricas usuais de classificação, como a Acurácia, Escore F1, Sensibilidade etc. Interpretando essas medidas, podemos concluir que o ajuste tem poder preditivo decente, (55% de acurácia e 52% de escore F1), embora a vantagem desse modelo comparados a técnicas focadas em poder preditivo (como a *Random Forest*) é a interpretabilidade dos parâmetros e a inferência feita.

Referências Bibliográficas

[1] Max Kuhn. The `caret` Package. `https://topepo.github.io/caret/measuring-performance.html#measures-for-predicted-classes`. Accessado em: 2022-04-16.