MAE 0560 - Análise de Dados Categorizados
2º Semestre de 2020

Profª Márcia D'Elia Branco

Lista 4

1. A tabela a seguir apresenta os efeitos estimados para um modelo de regressão logística com $y =$ presença de células escamosas no câncer de esôfago (1 = sim, 0 = não). A categoria fumante (f) é igual a 1 para pelo menos um pacote por dia e 0 caso contrário, consumo de álcool (a) é igual ao número de médio de bebidas alcólicas consumidas por dia e raça (r) é igual a 1 para negros e 0 para brancos.

   (a) Para descrever a interação raça-fumante, construa a equação de predição quando $r = 1$ e novamente quando $r = 0$. Encontre a razão de chances condicional estimada para o efeito do fumo para cada caso. Analogamente, construa a equação de predição quando $s = 1$ e novamente quando $s = 0$. Encontre a razão de chances condicional estimada para o efeito de raça em cada caso.

   (b) Na tabela o que os coeficientes de fumante e raça representam? A quais hipóteses seus valores-P se referem?

Variável	Efeito	Valor-P
Intercepto	-7.00	< 0.01
Consumo de álcool	0.10	0.03
Fumante	1.20	< 0.01
Raça	0.30	0.02
Raça x Fumante	0.20	0.04

2. Para a tabela da Pesquisa Social Geral de 2016 abaixo, crie um arquivo com os dados e analise este utilizado regressão logística. Apresente suas análises em um relatório curto, incluindo as saídas editadas em um apêndice.

		Crença em Vida Após a Morte	
Raça	Religião	Sim	Não ou Indeciso
Branco	Protestante	817	250
	Católico	519	194
	Outra	48	9
Preto	Protestante	298	86
	Católico	39	13
	Outra	119	38

3. O arquivo *MBTI* (ver *www.stat.ufl.edu/ aa/intro-cda/data*) classifica uma amostra de pessoas em se elas declaram consumir álcool com frequência (ou não) e nas quatro escalas binárias do teste de personalidade de Myers-Briggs: Extroversão/Introversão (E/I), Sensorial/Intuição (S/N), Razão/Sentimento (T/F) e Julgamento/Percepção (J/P). As 16 combinações de preditores correspondem aos 16 tipos de personalidades. Para os itens abaixo, considere modelos de regressão logística usando os quatro tipos de personalidade como preditores das probabilidades necessárias.

   (a) Quando a proporção amostral de 0.092 que alegou consumir álcool frequentemente é o ponto de corte para contrução de uma tabela de classificação, temos sensitividade = 0.53 e especificidade = 0.66. Explique essas quantidades e mostre que a proporção amostral de classificações corretas foi de 0.65.

   (b) O arquivo *MBTI* também mostra respostas sobre se a pessoa fuma frequentemente. Quando uma tabela de classificação para o modelo contendo os termos dos quatro efeitos principai para predizer o hábito de fumar adota uma proporção amostral de pessoas que fumam com frequência de 0.23 como ponto de corte, temos sensitividade = 0.48 e especificidade = 0.55. A área abaixo da curva ROC é 0.55. O conhecimento do tipo de personalidade ajuda a predizer se a pessoa fuma ou não com frequência? Justifique.

4. Construir as tabelas de classificação para o exemplo 1 da aula 13, considerando $p_c = 0.5$ e $p_c = 0.4$. Desenhar a curva ROC associada, apresentar a médida AUC e interpretar.

5. O arquivo *Students* em *www.stat.ufl.edu/ aa/intro-cda/data* apresenta respostas de alunos de Ciências Sociais da Universidade da Flórida a um questionário contendo perguntas sobre

   - *gender*: gênero com 1 = mulher e 0 = homem
   - *age*: idade
   - *hsgpa*: média ponderada no ensino médio, em uma escala de 0 a 4
   - *cogpa*: média ponderada na graduação, em uma escala de 0 a 4
   - *dhome*: distância da cidade natal ao campus, em milhas
   - *dres*: distãncia de casa ao campus, em milhas
   - *tv*: número médio de horas por semana que assiste televisão
   - *sport*: número médio de horas por semana que realiza atividades físicas
   - *news*: número de vezes por semana que lê o jornal
   - *aids*: número de conhecidos que morreram de AIDS ou que são HIV+
   - *veg*: se é vegetariano com 1 = sim e 0 = não
   - *affil*: afiliação política com 1 = Democrata, 2 = Republicano e 3 = Independente
   - *ideol*: ideologia política com 1 = muito liberal, 2 = liberal, 3 = ligeiramente liberal, 4 = moderado, 5 = ligeiramente conservador, 6 = conservador, 7 = muito conservador
   - *relig*: quão frequentemente atende serviços religiosos com 0 = nunca, 1 = ocasionalmente, 2 = a maiora das semanas, 3 = toda semana
   - *abor*: opinião sobre se o aborto nos primeiros três meses de gestação deveria ser legal com 1 = sim e 0 = não
   - *affirm*: se apoia ações afirmativas com 1 = sim e 0 = não
   - *life*: se crê em vida após a morte com 1 = sim, 2 = não e 3 = indeciso.

(a) Apresente todos os passos de um método de seleção de modelos tal como seleção intencional para adoção de um modelo para prever *abor* quando as covariáveis cogitadas são *ideol*, *relig*, *news*, *hsgpa* and *gender*.

(b) Utilizando uma ferramenta automática tal como *stepAIC* ou *bestglm* no *R*, construa um modelo para predizer *abor* considerando todas as variáveis binárias e quantitativas como variáveis explicativas.

(c) Com $y = veg$ e todas as variáveis binárias e quantitativas do arquivo, mostre que o teste da razão de verossimilhanças para testar $H_0 = \beta_0 = \ldots, = \beta_p$ tem valor-P < 0.001, porém o método de inclusão passo a frente utilizando o testes de Wald com valor-P de 0.05 como critério de seleção adota o modelo nulo. Explique como isso pode acontecer.

6. A tabela a seguir consiste em uma tabela de contingência $2 \times 2 \times 6$ para $y =$ se foi admitido na graduação da Universidade da Califórnia, Berkeley, no Outono de 1973, pelo gênero do candidato e pelos seis maiores departamentos.

	Aprovado, Homem		Aprovado, Mulher	
Departamento	Sim	Não	Sim	Não
1	512	313	89	19
2	353	207	17	8
3	120	205	202	391
4	138	279	131	244
5	53	138	94	299
6	22	351	24	317
Total	1198	1493	557	1278

(a) Ajuste um modelo de regressão logística que tem o Departamento como única variável explicatória para y. Utilize os resíduos padronizados para discutir a qualidade do ajuste.

(b) Quando adicionamos o efeito de gênero, a razão de chances condicional entre aprovação e gênero é 0.90. A tabela marginal, colapsada sobre o Departamento, tem razão de chances 1.84. Explique o que faz com que essas associações sejam tão distintas.

7. Suponha que $y = 0$ quando $x = 0, 10, 20, 30$ e $y = 1$ quando $x = 70, 80, 90, 100$.

(a) Explique, intuitivamente, o motivo de termos $\hat{\beta} = \infty$ para o modelo $\text{logit}[P(Y = 1)] = \alpha + \beta x$. Apresente $\hat{\beta}$ e seu desvio padrão obtido com seu software de preferência.

(b) Adicione duas observações em $x = 50$, sendo uma com $y = 1$ e outra com $y = 0$. Obtenha $\hat{\beta}$ e seu desvio padrão. Você acha que essas estimativas estão corretas? Por quê? O que acontece se trocarmos essas duas observações por $y = 1$ quando $x = 49.9$ e $y = 0$ quando $x = 50.1$?

8. Considere o exercício 4 com $y = veg$. Encontre um modelo de regressão logística para o qual pelo menos uma das estimativas de máxima verossimilhança é infinita. Explique a característica dos dados que causa isso. Apresente e interprete os resultados do ajuste de um modelo usando a regressão logística e inferência Bayesiana.

9. Um estudo investigou características associadas com y = se o paciente com câncer apresentou remissão (1 = sim, 0 = não). Uma covaríavel importante foi a *marcagem*, que mede a atividade proliferativa das células após o paciente ter recebido uma injeção de trimidina tritriatada. A tabela a seguir apresenta a saída do ajuste de um modelo probito. Interprete as estimativas dos parâmetros

   (a) encontrando o valor de marcagem no qual a probabilidade estimada de remissão é igual a 0.5;

   (b) encontrando a diferença entre as probabilidades estimadas de remissão nos quartis inferior e superior de marcagem, 14 e 28;

   (c) usando o modelo de variável latente normal correspondente;

   (d) usando características da curva de resposta da distribuição acumulada da normal.

	Estimativa	Erro Padrão	Valor Z	$\Pr(>\|z\|)$
Intercepto	-2.31777	0.76060	-3.047	0.00231
Marcagem	0.08785	0.03293	2.668	0.00763