MAE0399 – Análise de Dados e Simulação – primeiro semestre de 2020
Professora: Márcia D´Elia Branco

1) Prove que considerar o modelo de regressão logística dado por

$$p(x) = Prob(Y=1 \mid X) = exp\{\beta 0 + \beta 1\, X\} / (1 + exp\{\beta 0 + \beta 1\, X\})$$

é equivalente a considerar o *logit(p(x))* como uma função linear de X.
Em que $logit(p(x)) = p(x) / (1-p(x))$

2) Suponha um conjunto de dados de estudantes de uma disciplina de pós-graduação e as variáveis X1: horas de estudo, X2: pontuação no exame de admissão e Y= 1, se obteve A na disciplina e Y=0, caso contrário. Foi ajustado um modelo de regressão logística e as estimativas dos coeficientes são b0=-6, b1=0.05 e b3=1 .

(a) Estime a probabilidade de um estudante que estudou 40 horas e possui X2=3.5 obter uma nota A na disciplina.

(b) Para um estudante com a mesma pontuação considerada em (a), quantas horas ele deveria estudar para ter uma probabilidade de 0.5 de obter uma nota A na disciplina?

3) Sobre a interpretação de chance. Dados de *Default***.**

(a) Em média, qual a fração de pessoas com uma chance de 0.37 de não pagar o seu cartão de crédito (inadimplente) não irão de fato pagar?

(b) Suponha que um indivíduo tenha uma probabilidade de 0.16 de ser inadimplente, qual a chance dele não pagar o cartão?

4) Considere o arquivo de dados *Weekly* que faz parte do pacote *ISLR* do **R .** Esses dados possuem 1089 porcentagens de retornos semanais durante 21 anos, de 1990 até 2010.

(a) Faça alguns gráficos resumos dos dados. Consegue identificar algum padrão? Quais?

(b) Ajuste um modelo de regressão logística aos dados completos considerando *Direction* como variável resposta e como preditores as cinco variáveis *lag* mais *volume.* Use a função *summary* para apresentar os resultados. Quais preditores são estatisticamente significantes? Por que? Especifique as hipóteses que estão sendo testadas.

(c) Obtenha a *confusion matrix, a sensibilidade, a especificidade* e a fração total de predições corretas. Comente os resultados.

(d) Ajuste agora um modelo de regressão logística considerando apenas um período de treinamento, de 1990 a 2008. Considere como preditor apenas a variável *lag2.* Use a função *summary* para apresentar e analisar os resultados. Considerando como amostra de teste os anos seguintes (2009 e 2010), obtenha a *confusion matrix , a sensibilidade, a especificidade* e a fração total de predições corretas.

5) Para o conjunto de dados *Default*, apresentado como exemplo em aula, ajuste o modelo de regressão logística considerando como preditores somente as variáveis *balance* e *student.*

(a) Apresente a tabela obtida pela função *summary.* A partir dessa tabela obtenha os intervalos de 95 % de confiança para os coeficientes de regressão associados aos dois preditores.

(b) Obtenha a probabilidade de um estudante ser inadimplente, considerando que ele tenha um *balance* igual a $ 2.000. E para o não estudante, qual é essa probabilidade? O que é mais arriscado oferecer crédito para um estudante ou não estudante? Justifique.

6) Use o conjunto de dados *Boston* do pacote *ISLR* do **R.** Explore o modelo de regressão logística com o objetivo de predizer se um dado *suburb* possuem uma taxa de crime acima ou abaixo da mediana.