Departamento de Estatística - ICEx - Universidade Federal de Minas Gerais
Lista de Exercícios 3: EST079 - Modelos Lineares Generalizados

Observação:

- As listas de exercícios de MLG não valem pontos.
  O estudante deve fazer os exercícios apenas como forma de aprendizagem e preparação para as avaliações.
- Algumas questões podem ser resolvidas através do `R`, outras questões devem ser feitas sem auxílio computacional.
  O enunciado do exercício avisará sobre isso.
- Se houver algum banco de dados disponibilizado para esta lista, saiba que ele foi salvo em um computador Linux. Desta forma, pode haver desconfiguração ao abrir o arquivo no "bloco de notas" do Windows. Use o comando `read.table` normalmente no `R` para ler os dados. A `data.frame` criada estará no formato correto.

[Questão 1] O representante de vendas de uma marca de queijo resolveu avaliar o efeito de três fatores sobre o número de unidades do produto vendidas em 1 semana no Mercado Central de Belo Horizonte. Além da resposta Y_i indicando a contagem de vendas na semana i, as seguintes variáveis foram coletadas: X_{1i} é a frequência de exposição do produto nas redes socias durante a semana i, X_{2i} é uma variável binária que indica se na i-ésima semana o produto estava exposto em local de boa visibilidade nas lojas do Mercado Central (1 = boa, 0 = ruim) e, finalmente, X_{3i} é o número de pessoas (dividido por 1000) que estiveram no Mercado Central na semana i. O acompanhamento das vendas foi realizado por um período de 50 semanas e forneceu os resultados disponibilizados no arquivo `dados_Q1_L3_MLG.txt`. Utilizando estes dados, responda:

(a) Use o comando `read.table` para carregar os dados no `R`. A *data frame* criada contém: Y, X_1, X_2 e X_3 nas colunas 1, 2, 3 e 4, respectivamente. Use a função `glm` para ajustar o MLG Poisson (com ligação canônica) para este caso. O preditor linear do modelo deve conter o intercepto e os coeficientes das três covariáveis descritas no enunciado. Mostre a saída computacional obtida com o `summary`. Avalie a significância dos coeficientes. Atenção! informe o valor-p que estiver analisando e o critério para decidir sobre a significância. Note que a questão NÃO está pedindo para você fazer vários ajustes até chegar a um melhor model. Você deve ajustar apenas o modelo que foi solicitado.

(b) Interprete a estimativa obtida em (a) para o coeficiente de X_1, ou seja, explique o que ocorre quando aumentamos em 0.1 a variável X_{1i}. Atenção: a questão NÃO está pedindo para reajustar o modelo com a modificação $X_1 + 0.1$.

(c) Usando comandos do `R` mostre o passo-a-passo das contas da deviance do modelo ajustado em (a). Atenção! implemente a expressão da deviance em seu script e mostre trechos das contas. Não responda usando função do `R` que calcula a deviance automaticamente.

(d) Verifique se a deviance obtida em (c) está entre a média e a mediana da distribuição de probabilidade estabelecida na literatura para a deviance de um MLG. Atenção: mostre o valor da média e da mediana mencionadas.

(e) Construa os gráficos: "resíduos de Pearson vs. valores ajustados" e "resíduos Componente do Desvio vs. valores ajustados". Os resíduos devem ficar no eixo vertical destes gráficos. Trace uma linha horizontal demarcando o patamar do valor zero. Atenção! você deve exibir e interpretar os dois gráficos solicitados.

[Questão 2] Considere os dados disponibilizados no arquivo `dados_Q2_L3_MLG.txt`. A coluna 1 contém uma variável resposta Y_i, que é um índice numérico positivo sobre o nível de criminalidade em 215 cidades de médio porte do Brasil. A coluna 2 contém a covariável X_{1i} representando a proporção da população com renda superior a 1 salário mínimo na cidade i. A coluna 3 contém a variável categórica X_{2i} indicando o nível de investimento da prefeitura na área de segurança (1 = alto, 2 = médio, 3 = baixo). Carregue os dados no `R` e responda:

(a) Faça o histograma da variável resposta Y e o histograma da resposta transformada $\ln(Y)$. Compare o formato dos dois gráficos e avalie se há semelhança deles com a distribuição Normal. Realize o teste de normalidade Shapiro-Wilk para a resposta transformada $\ln(Y)$; escreva as hipóteses, identifique o valor-p e explique o critério de decisão. Você acha que seria adequado ajustar o modelo de regressão linear (supondo normalidade) aos dados com resposta transformada?

(b) Aplique a função `glm` para ajustar o MLG Gama (com ligação `log`) para os dados sem transformação na variável resposta. O preditor linear do modelo deve conter o intercepto e os coeficientes relacionados às duas covariáveis descritas no enunciado. Mostre a saída computacional obtida via `summary`. Avalie a significância dos coeficientes. Atenção! informe o valor-p que estiver analisando e o critério para decidir sobre a significância. Note que a questão NÃO está pedindo para você fazer vários ajustes até chegar a um melhor model. Você deve ajustar apenas o modelo que foi solicitado.

(c) Usando o resultado em (b), interprete o impacto da variável X_2. Além de informar as condições da análise, você deve dizer claramente quem é impactado por X_2, qual é a magnitude do impacto e se esse impacto representa aumento ou diminuição.

(d) Verifique se a deviance, obtida no ajuste em (b), está entre a média e a mediana da distribuição de probabilidade estabelecida na literatura para a deviance de um MLG. Atenção: informe claramente o valor da média e da mediana mencionadas.

(e) Construa os gráficos: "resíduos de Pearson vs. valores ajustados" e "resíduos Componente do Desvio vs. valores ajustados". Os resíduos devem ficar no eixo vertical dos gráficos. Trace uma linha horizontal demarcando o patamar do valor zero. Atenção! você deve exibir e interpretar os dois gráficos solicitados.

[Questão 3] Considere aqui o conjunto de dados `dados_Q3_L3_MLG.txt`. Estes dados foram coletados em um estudo para avaliar o possível efeito cancerígeno de um produto químico. As observações são referentes a 400 camundongos; alguns deles receberam a mesma dosagem do produto químico (grupo tratamento) e foram acompanhados por 90 dias (verificando o aparecimento ou não de um tumor). Os demais animais não receberam o produto químico (grupo controle) e foram acompanhados pelo mesmo período. O conjunto de dados contém as variáveis:

- Coluna 1: $Y =$ variável resposta (0 = sem tumor e 1 = presença de tumor após 90 dias).
- Coluna 2: $X_1 =$ indicadora de grupo (0 = controle e 1 = tratamento).
- Coluna 3: $X_2 =$ indicadora de gênero (0 = fêmea e 1 = macho).

Responda os ítens a seguir.

(a) Carregue os dados no `R` e faça uma separação em duas partes da *data frame* obtida. A Parte I contém as observações das linhas 1 a 300. A Parte II contém as observações das linhas 301 a 400. Aplique a função `glm` para ajustar a regressão logística para os dados da Parte I. O preditor linear do modelo deve conter o intercepto e os coeficientes relacionados às duas covariáveis descritas no enunciado. Mostre a saída computacional via `summary`. Avalie a significância dos coeficientes. Atenção! informe o valor-p que estiver analisando e o critério para decidir sobre a significância. Note que a questão NÃO está pedindo para você fazer vários ajustes até chegar a um melhor model. Você deve ajustar apenas o modelo que foi solicitado.

(b) Usando as estimativas dos coeficientes obtidas em (a), calcule as probabilidades de ocorrência de tumor após 90 dias para os camundongos alocados na Parte II da separação dos dados. Mostre todas as probabilidades estimadas!

(c) Considerando o limiar de probabilidade igual a 0.5, utilize o resultado em (b) para classificar cada um dos 100 camundongos alocados na Parte II dos dados em "sem tumor" e "com tumor". Usando as classificações geradas neste exercício e as observações reais de $Y_{301}, Y_{302}, \ldots, Y_{400}$, calcule a sensibilidade e a especificidade do método de classificação baseado no modelo logístico ajustado neste problema. Mostre o passo a passo das contas e o resultado final! Atenção! não use funções prontas do `R` que calculam a sensibilidade ou especificiade automaticamente.