**Universidade Federal do Paraná – Departamento de Estatística**

**Disciplina CE225 – Modelos Lineares Generalizados**

**Prof. Cesar Augusto Taconeli**

**Prova 1 – 10/10/2014**

|  |
| --- |
| **Notas:**  **1-** Procure ser objetivo em suas respostas. Escreva apenas aquilo que está sendo solicitado, sem se prolongar excessivamente em suas justificativas.  **2-** Seja claro em suas respostas. O uso adequado dos termos e notações matemáticas será considerado na correção. Somente avaliarei **o que você escrever**, não **o que você “pretendia escrever”.**  **3-** A nota bruta obtida será devidamente ajustada para a escala de 0 a 100 pontos usando “regra de três”. |

**Exercício 1 – (20 pontos) Modelos lineares generalizados** configuram extensões dos **modelos lineares com erros normalmente distribuídos**. Apresente, com suas palavras, dois pontos que caracterizam o primeiro como extensão do segundo.

*Nos modelos lineares generalizados são utilizadas distribuições pertencentes à família exponencial de dispersão. O modelo normal faz parte desta família, sendo desta forma um caso particular de um MLG. O algoritmo de estimação dos MLGs é o de mínimos quadrados ponderados, sendo este uma extensão do algoritmo de mínimos quadrados ordinários utilizado na estimação dos modelos lineares normais.*

**Exercício 2 – (10 pontos por item) -** Uma variável aleatória *Y* tem distribuição exponencial de parâmetro se sua função densidade de probabilidade é dada por:

1. Verifique que a distribuição exponencial pertence à família exponencial, expressando-a na forma canônica:
2. Identifique, no contexto de Modelos Lineares Generalizados, a função de ligação canônica e a função de variância correspondentes à distribuição exponencial;
3. Identifique um problema de ordem prática em se utilizar a função de ligação canônica para um MLG com resposta exponencial;

*A função de ligação canônica observada foi a identidade. Como a distribuição exponencial admite somente valores positivos, pode acontecer da função de ligação produzir valores fora do espaço paramétrico da distribuição.*

1. Considere *n* variáveis aleatórias independentes com distribuição Exponencial de parâmetro . Expresse a deviance em termos dos valores observados e dos correspondentes valores ajustados .

**Ajuda:** Definição de deviance:

Sendo e as log-verossimilhanças maximizadas sob o modelo saturado e corrente respectivamente.

**Exercício 3 (30 pontos) –** O diagnóstico de determinado tipo de tumor maligno é realizado por um procedimento cirúrgico bastante invasivo. Um grupo de médicos está estudando um procedimento alternativo, baseado em exames menos invasivos. Os médicos desejam estimar a probabilidade de um paciente ser portador do tumor em questão com base na presença ou ausência de uma proteína X em seu organismo e num escore de predisposição a esse tipo de câncer, que é uma variável numérica que se baseia em uma série de características do paciente, como o histórico familiar.

Para o problema apresentado, proponha um MLG em duas etapas, conforme visto em aula, especificando, num primeiro momento, a distribuição da resposta condicional às covariáveis e, posteriormente, a relação entre a distribuição da resposta e o preditor linear. Não se esqueça de deixar claro quem são as variáveis resposta e explicativas e como são inseridas no modelo.

*Variáveis:*

*Tumor: variável resposta categórica, sendo 1 para portador do tumor e 0 para não portador do tumor;*

*Proteína: variável explicativa categórica, sendo 1 para a presença e 0 para a ausência da proteína;*

*Escore: variável explicativa numérica com o escore de predisposição ao tumor.*

*Distribuição proposta: .*

*Função de ligação: logito (canônica).*

*Modelo resultante: .*

**Exercício 4 (10 pontos por item) –** Um estudo sobre a saúde de médicos produziu um levantamento, referente ao acompanhamento de médicos que atenderam exclusivamente num certo hospital no último ano. Dentre as variáveis contempladas no levantamento, foram destacadas:

* **Queixas –** número de queixas relatadas pelo médico referentes a situações de estresse;
* **Sexo –** (M: masculino; F: feminino);
* **Residência –** o médico está no período de residência? (Y para sim; N para não);
* **Visitas –** número de atendimentos realizados pelo médico no último ano.

O quadro abaixo apresenta algumas linhas da base de dados.

|  |
| --- |
| Id Queixas Sexo Residência Visitas  1 2 F Y 2014  2 3 M N 3091  3 1 M Y 879  4 1 M N 1780  5 11 M N 3646  6 1 M N 2690 |

O objetivo da análise é ajustar um MLG para explicar o número de queixas com base nas demais variáveis. Para isso, considerou-se a distribuição de Poisson para o número de queixas e função de ligação logaritmo. O quadro apresentado na página seguinte descreve os resultados referentes ao modelo ajustado. Com base nele, responda os seguintes itens:

1. Escreva a equação do modelo ajustado na escala da resposta (ou seja, com relação ao número médio de queixas);
2. Estime o número médio de queixas para médicos com o seguinte perfil: Residente, do sexo masculino, atendendo 2000 visitas ao ano;

*Estima-se que o numero médio de queixas para médicos com esse perfil seja de 1,8405 no ano.*

1. Quantos médicos compõem a base de dados? Justifique.

*Verifica-se que o modelo ajustado possui 4 parâmetros e 40 graus de liberdade para a deviance. Com isso, conclui-se que a base de dados possui 44 médicos.*

1. Avalie a qualidade do ajuste com base na deviance. Teste a qualidade do ajuste ao nível de significância de 5%;

*Comparando o valor da deviance de 50,739 com o quantil 5% da distribuição qui-quadrado com 40 graus de liberdade que é 55,759, não se rejeita a hipótese nula de que o modelo está bem ajustado, ao nível de significância de 5%.*

1. Faça um breve relato dos resultados do modelo com base nas estimativas e nos respectivos testes. Usando suas palavras, identifique fatores que estão relacionados a uma maior (ou menor) frequência de queixas;

*Verifica-se no modelo que o sexo masculino está associado com um maior número de queixas, embora seja recomendável desconsiderar esse resultado devido a pequena significância da variável. Médicos residentes tendem a apresentar uma menor quantidade de queixas que os médicos não residentes. Um maior número de visitas está associado com um maior número de queixas.*

1. O modelo ajustado poderia ser apenas o ponto de partida para a proposta/ajuste de outros modelos. Cite duas alterações que você faria no modelo ajustado visando a obtenção de um novo modelo, que talvez proporcionasse melhor ajuste.

**Nota –** Considere, neste último item, que os dados disponíveis são apenas esses, não havendo a possibilidade de selecionar mais médicos e/ou variáveis.

*Algumas possibilidades:*

* *Remover as variáveis não significativas. Primeiramente remover a variável sexo e, caso a residência continue não significativa ao nível de 5% depois dessa remoção, considerar removê-la também.*
* *Alterar a função de ligação (neste caso para a raiz quadrada) é uma possibilidade, embora não costume apresentar resultados muito diferentes.*
* *Outra possibilidade é ajustar um modelo com outra distribuição da família exponencial, no caso com a distribuição Binomial Negativa.*

**Boa Prova!**

|  |
| --- |
| > ajuste1=glm(Queixas~Sexo+Residência+Visitas,family=poisson,data=dados)  > summary(ajuste 1)  Call:  glm(formula = Queixas ~ Sexo + Residência + Visitas, family = poisson, data = dados)  Deviance Residuals:  Min 1Q Median 3Q Max  -1.9514 -0.9058 -0.3792 0.6189 1.9395  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (intercept) -0.7611853 0.4072631 -1.869 0.0616 .  SexoM 0.0780814 0.2076261 0.376 0.7069  ResidênciaY -0.3046352 0.1736236 -1.755 0.0793 .  Visitas 0.0007989 0.0001456 5.487 4.1e-08 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for poisson family taken to be 1)  Null deviance: 89.477 on 43 degrees of freedom  Residual deviance: 50.739 on 40 degrees of freedom  AIC: 181.52  Number of Fisher Scoring iterations: 5 |