PROVA 2 – MODELOS LINEARES GENERALIZADOS

1. (5 pontos por item) Assinale verdadeiro (V) ou falso (F) em cada uma das afirmações relacionadas abaixo, referentes a modelos para dados de contagens. Corrija as afirmações classificadas como falsas.

   (a) O modelo log-linear de Poisson se aplica sempre que tivermos um conjunto de variáveis explicativas discretas;
   FALSO – AULA 16.

   (b) Se estivermos estudando a contagem de sementes que germinam, dentre 15 sementes plantadas para cada uma de 5 condições experimentais, o modelo de Poisson, a princípio, é o mais apropriado.
   FALSO – AULAS 14 E 16.

   (c) O uso da função de ligação logarítmica implica em efeitos multiplicativos para as variáveis explicativas;
   VERDADEIRO – AULA 19.

   (d) Se propormos $y_i \sim Poisson(t_i\lambda_i)$, sendo $t_i$ o tempo de seguimento do i-ésimo indivíduo, devemos incorporar o tempo de seguimento ao modelo somando-o ao preditor linear, na forma: $n_i = x_i^{'}\beta + t_i$;
   FALSO – AULA 16.

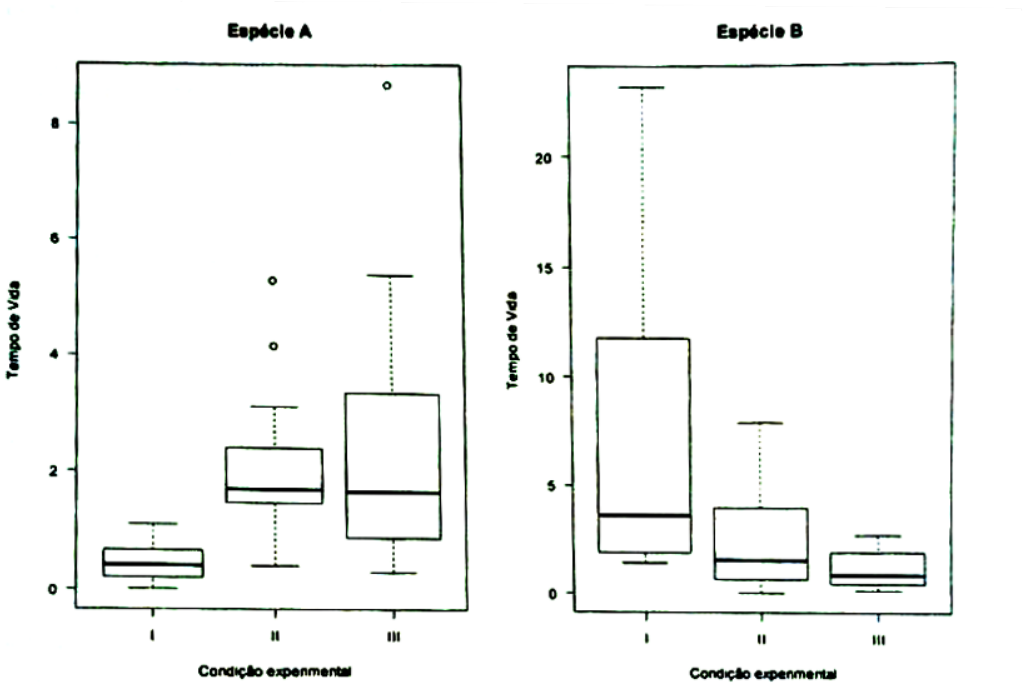
   (e) Na aplicação do modelo log-linear para a análise de tabelas de contingência, o modelo correspondente à hipótese de independência mútua entre as variáveis é o modelo nulo;
   FALSO – AULA 17.

   (f) O problema da superdispersão pode ser causado, dentre outros fatores, por um padrão aleatório na ocorrência dos eventos de interesse ao longo do espaço ou tempo;
   FALSO – AULA 18.

   (g) Ao utilizar a distribuição de Poisson na análise de dados de contagens com superdispersão, os erros padrões dos parâmetros serão subestimados;
   VERDADEIRO – AULA 18.

   (h) Ao usar o modelo de regressão quase-Poisson, considerando $V(\mu_i) = \phi\mu_i$, com $\phi$ a ser estimado, as estimativas pontuais dos $\beta$'$s$ serão idênticas às produzidas pelo modelo de regressão Poisson.
   VERDADEIRO – AULA 18.

2. (40 pontos) Um experimento tem como objetivo comparar os tempos médios de vida de duas espécies de insetos (A e B) submetidos a três condições experimentais distintas (I, II e III). Para isso, foram observados os tempos de vida de 45 insetos de cada espécie, com 15 insetos de cada espécie submetidos a cada condição experimental. Na sequência são apresentados gráficos e algumas medidas descritivas baseadas nos resultados do experimento:



| Espécie A | | | Espécie B | | |
|---|---|---|---|---|---|
| Cond. Exp I | Cond. Exp II | Cond. Exp III | Cond. Exp I | Cond. Exp II | Cond. Exp III |
| $n = 15$ | $n = 15$ | $n = 15$ | $n = 15$ | $n = 15$ | $n = 15$ |
| $\bar{x} = 0,46$ | $\bar{x} = 2,19$ | $\bar{x} = 2,45$ | $\bar{x} = 7,13$ | $\bar{x} = 2,55$ | $\bar{x} = 1,25$ |
| $s^2 = 0,13$ | $s^2 = 2,24$ | $s^2 = 5,20$ | $s^2 = 48,12$ | $s^2 = 5,30$ | $s^2 = 0,78$ |
| $s/\bar{x} = 1,27$ | $s/\bar{x} = 1,46$ | $s/\bar{x} = 1,07$ | $s/\bar{x} = 1,03$ | $s/\bar{x} = 1,11$ | $s/\bar{x} = 1,40$ |

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. Justifique suas especificações.

COMPONENTE ALEATÓRIO:

- Variável resposta: tempo médio de vida (y).
- Distribuição Proposta: $Y_{ij} \sim Gama\,(\mu_{ij},\, \nu)$.

COMPONENTE SISTEMÁTICO

- Preditor linear

$$\eta_{ij} = \beta_0 + \beta_1 espB + \beta_2 cond2 + \beta_3 cond3$$

LIGAÇÃO

- Função de ligação logarítmica.

MODELO RESULTANTE

$$y_{ij} \mid especie_i;\ condicao_j \sim Gama\,(\mu_{ij},\, \nu)$$
$$ln\,(\mu_{ij}) = \beta_0 + \beta_1 espB + \beta_2 cond2 + \beta_3 cond3$$

3. (10 pontos por item) Os dados apresentados na sequencia foram extraídos de um estudo conduzido pelo Instituto de Diabetes e Doenças Digestivas, baseado em 729 mulheres adultas de uma comunidade indígena. O objetivo é identificar fatores relacionados à incidência de diabetes nessa população. Na sequência são apresentadas as seis primeiras linhas da base de dados:

| | Gravidez | Diastólica | imc | Idade | test |
|---|---|---|---|---|---|
| 1 | 6 | 72 | Sobrepeso | 50 | 1 |
| 2 | 1 | 66 | Sobrepeso | 31 | 0 |
| 3 | 8 | 64 | Normal | 32 | 1 |
| 4 | 1 | 66 | Sobrepeso | 21 | 0 |
| 5 | 0 | 40 | Sobrepeso | 33 | 1 |
| 6 | 5 | 74 | Sobrepeso | 30 | 0 |
| 7 | 3 | 50 | Sobrepeso | 26 | 1 |
| 8 | 10 | 42 | Sobrepeso | 29 | 0 |
| 9 | 2 | 70 | Sobrepeso | 53 | 1 |

Descrição das variáveis:

**Gravidez -** Número de vezes que a mulher esteve grávida;

**Diastólica -** Pressão sanguínea diastólica (mm Hg);

**imc -** Índice de massa corporal (Normal ou sobrepeso);

**Idade -** Idade (anos);

**test -** Resultado do teste de diagnóstico de diabetes (0 se negativo, 1 se positivo).

Para a análise desses dados, foi ajustado um modelo de regressão logística, considerando o resultado do teste como variável resposta e as demais variáveis como explicativas. Nenhuma interação foi incluída ao modelo. O quadro apresentado na sequencia contém o resumo do modelo ajustado:

```
> ajuste=glm(test~Gravidez+imc+Idade+Diastólica,family=binomial,data=pima2)
> summary(ajuste)

Call:
glm(formula = test ~ Gravidez + imc + Idade + Diastólica, family = binomial,
    data = pima2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7793  -0.8777  -0.7125   1.1197   2.5376

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.857512   0.653975  -7.428 1.11e-13 ***
Gravidez      0.072874   0.028727   2.537 0.011187 *
imcSobrepeso  2.180638   0.409034   5.331 9.76e-08 ***
Idade         0.030471   0.008772   3.474 0.000513 ***
Diastólica    0.012243   0.007376   1.660 0.096937 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 938.74  on 728  degrees of freedom
Residual deviance: 832.92  on 724  degrees of freedom
  (39 observations deleted due to missingness)
AIC: 842.92

Number of Fisher Scoring iterations: 5
```

Com base no modelo ajustado:

(a) Apresente a equação do modelo ajustado, na escala de probabilidade;

$$\hat{\pi} = \frac{exp\{-4,86 + 0,07x_1 + 2,18x_2 + 0,03x_3 + 0,01x_4\}}{exp\{-4,86 + 0,07x_1 + 2,18x_2 + 0,03x_3 + 0,01x_4\} + 1}$$

(b) Calcule a estimativa da probabilidade de diagnóstico positivo para mulheres com sobrepeso, 40 anos, pressão diastólica de 50mmHg e com uma única gravidez;

```
[1] 0.3158336
```

(c) Estime a razão de chances de diagnóstico de diabetes para mulheres com sobrepeso em relação a mulheres com peso normal (fixadas as demais variáveis);

```
[1] 8.851948
```

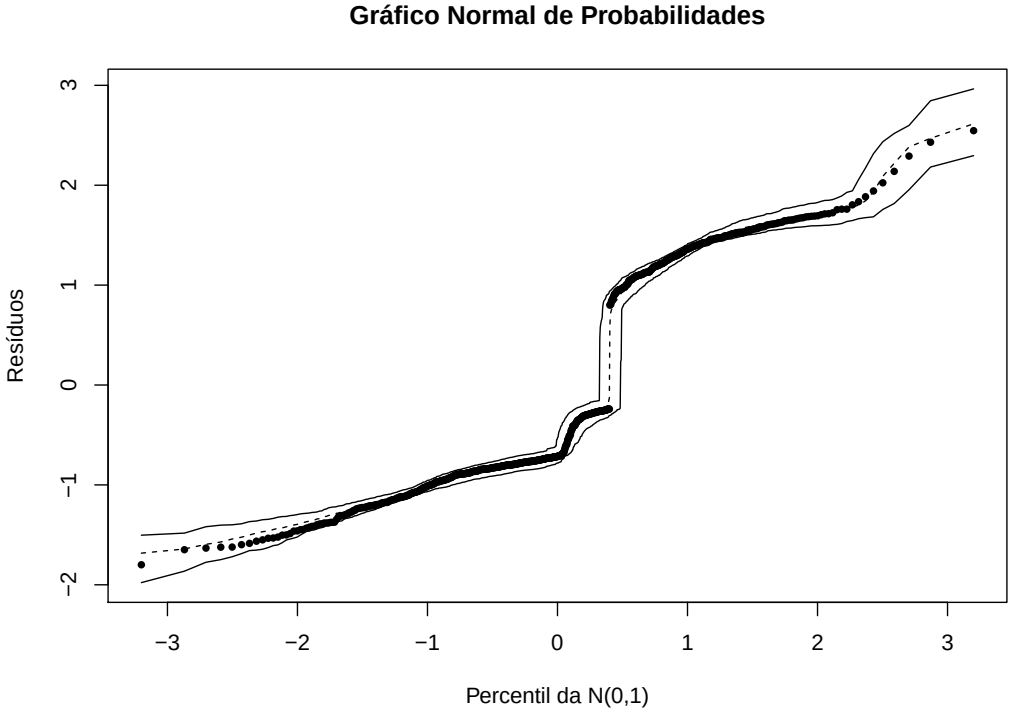
(d) Forneça um intervalo de confiança (95%) para a razão de chances solicitada no item b;

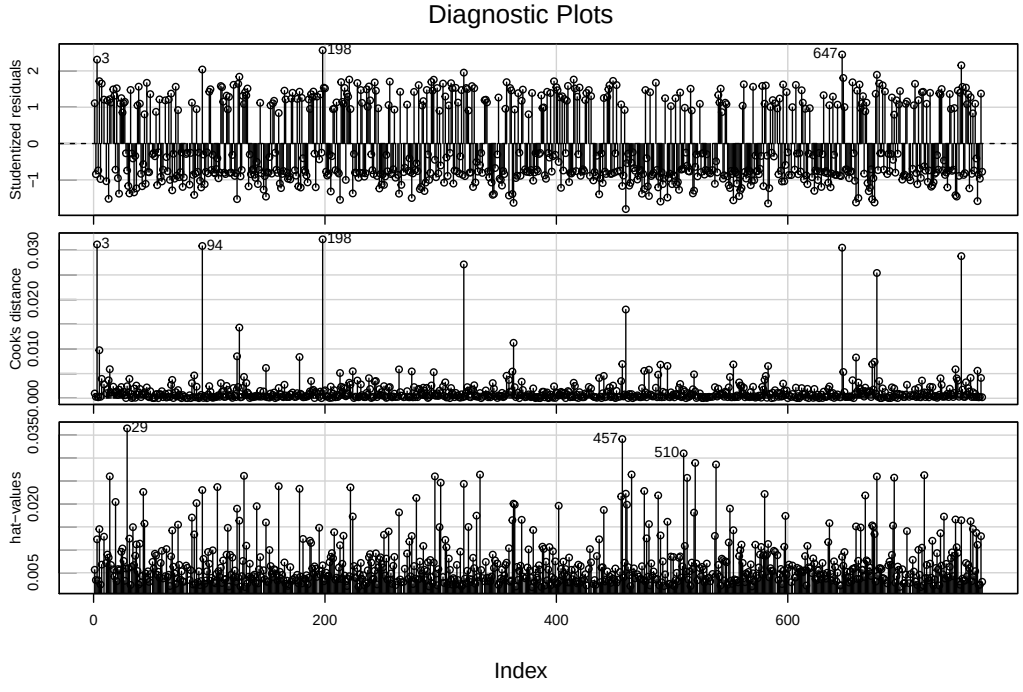
```
    2.5 %    97.5 %
3.970711 19.733744
```

(e) Estime a razão de chances de diagnóstico de diabetes para indivíduos com k + 10 anos em relação a indivíduos com k anos (fixadas as demais variáveis);

```
[1] 1.356232
```

(f) Para efeito de diagnóstico do ajuste, são apresentados o gráfico qqplot para resíduos (com envelopes simulados) e gráficos de resíduos studentizados, distância de Cook e dos valores da diagonal da matriz H versus o índice das observações. Com base nos gráficos apresentados, e nos resultados produzidos pelo summary, avalie a qualidade do ajuste.




De acordo com o qqplot, o modelo está bem ajustado. As observações 3, 94 e 198 apresentaram distância de cook mais elevada, de forma que pode-se considerar a remoção dessas observações para o ajuste de um novo modelo e comparar se houve mudanças consideráveis.