Lista 1

Modelos lineares generalizados

Paulo Ricardo Seganfredo Campana

19 de dezembro de 2023

Questão 1. Mostre que a distribuição Poisson Truncada, com função de probabilidade pertence à família exponencial $f(y) = \exp[\phi\{y\theta - b(\theta)\} + c(y, \phi)]$.

$$
\begin{aligned}
f(y;\lambda) &= \frac{\lambda^y}{(e^\lambda - 1)y!}, \quad \text{em que } y = 1, 2, \cdots, \ \lambda : 0 \\
&= \exp\left\{\ln\left(\frac{\lambda^y}{(e^\lambda - 1)y!}\right)\right\} \\
&= \exp\left\{\ln(\lambda^y) - \ln(e^\lambda - 1) - \ln y!\right\} \\
&= \exp\left\{1[y \ln \lambda - \ln(e^\lambda - 1)] - \ln y!\right\}
\end{aligned}
$$

$$
\phi = 1, \quad \theta = \ln \lambda, \quad c(y, \phi) = -\ln y!
$$

$$
b(\theta) = \ln(e^\lambda - 1) = \ln(e^{e^\theta} - 1)
$$

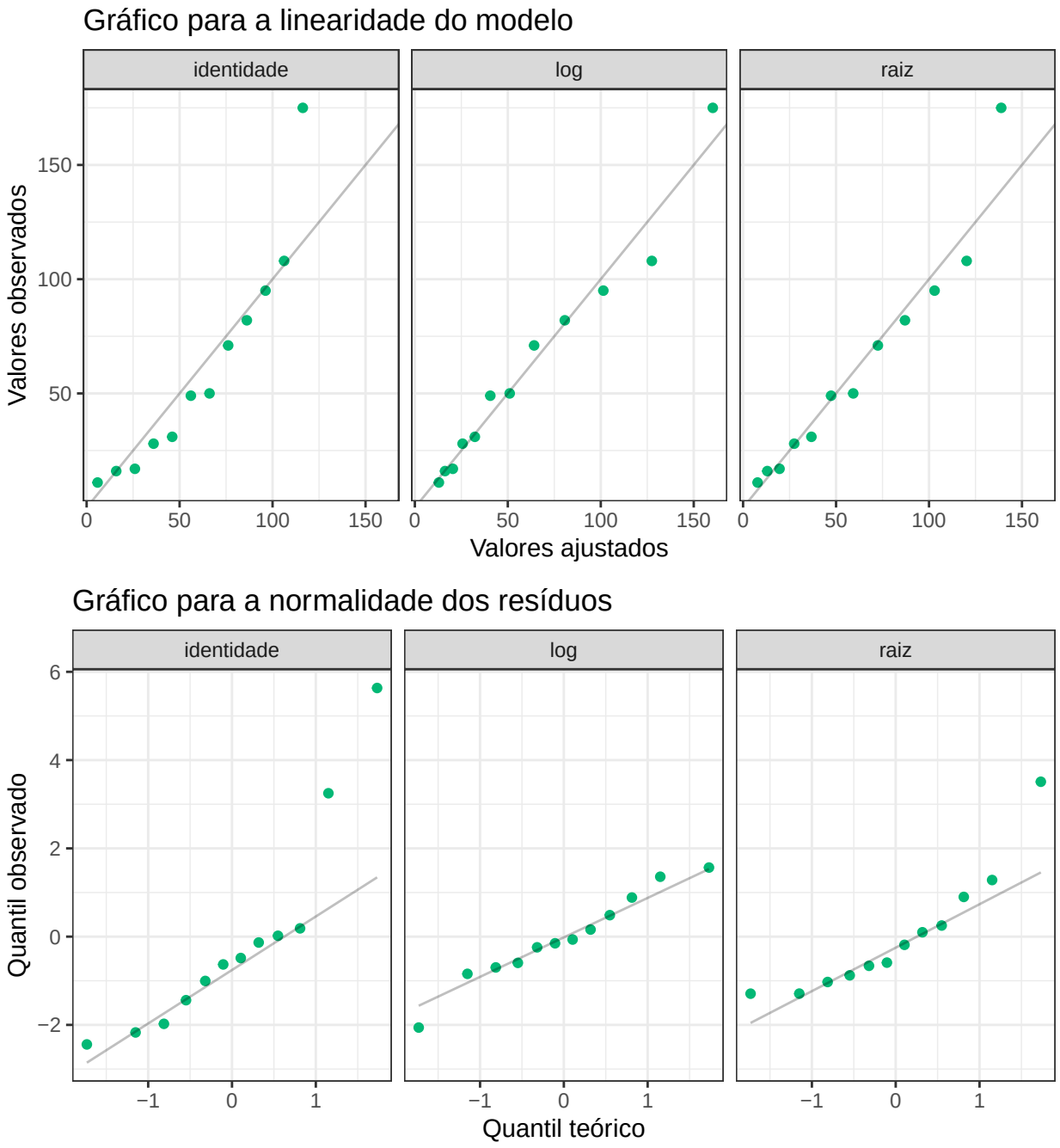
Questão 2. Descrevemos na tabela abaixo o número de bactérias sobreviventes em amostras de um produto alimentício segundo o tempo (em minutos) de exposição do produto a uma temperatura de 300 ºF.

```
bactérias <- data.frame(
    número = c(175, 108, 95, 82, 71, 50, 49, 31, 28, 17, 16, 11),
    tempo = 1:12
)
```

a) Realize o ajuste da poisson com as possíveis funções de ligação e decida, entre elas, qual a função de ligação é melhor. Justifique sua escolha.

```
bac_fit1 <- glm(número ~ tempo, bactérias, family = poisson(link = identity))
bac_fit2 <- glm(número ~ tempo, bactérias, family = poisson(link = log     ))
bac_fit3 <- glm(número ~ tempo, bactérias, family = poisson(link = sqrt    ))
```

```
data.frame(
    y          = bactérias$número,
    identidade = fitted(bac_fit1),
    log        = fitted(bac_fit2),
    raiz       = fitted(bac_fit3)
) |>
    tidyr::pivot_longer(-y) |>
    ggplot(aes(x = value, y = y)) +
    facet_grid(cols = vars(name)) +
    geom_point(color = "#02b875") +
    geom_abline(alpha = 0.25) +
    labs(
        title = "Gráfico para a linearidade do modelo",
        x = "Valores ajustados",
        y = "Valores observados"
    )
data.frame(
    identidade = boot::glm.diag(bac_fit1)$rd,
    log        = boot::glm.diag(bac_fit2)$rd,
    raiz       = boot::glm.diag(bac_fit3)$rd
) |>
    tidyr::pivot_longer(tidyr::everything()) |>
    ggplot(aes(sample = value)) +
    facet_grid(cols = vars(name)) +
    geom_qq(color = "#02b875") +
    geom_qq_line(alpha = 0.25) +
    labs(
        title = "Gráfico para a normalidade dos resíduos",
        x = "Quantil teórico",
        y = "Quantil observado"
    )
```



Pelos gráficos acima, a função de ligação log tem melhor performance na estimação do número de bactérias e seus resíduos sofrem menos desvios da normalidade para valores extremos.

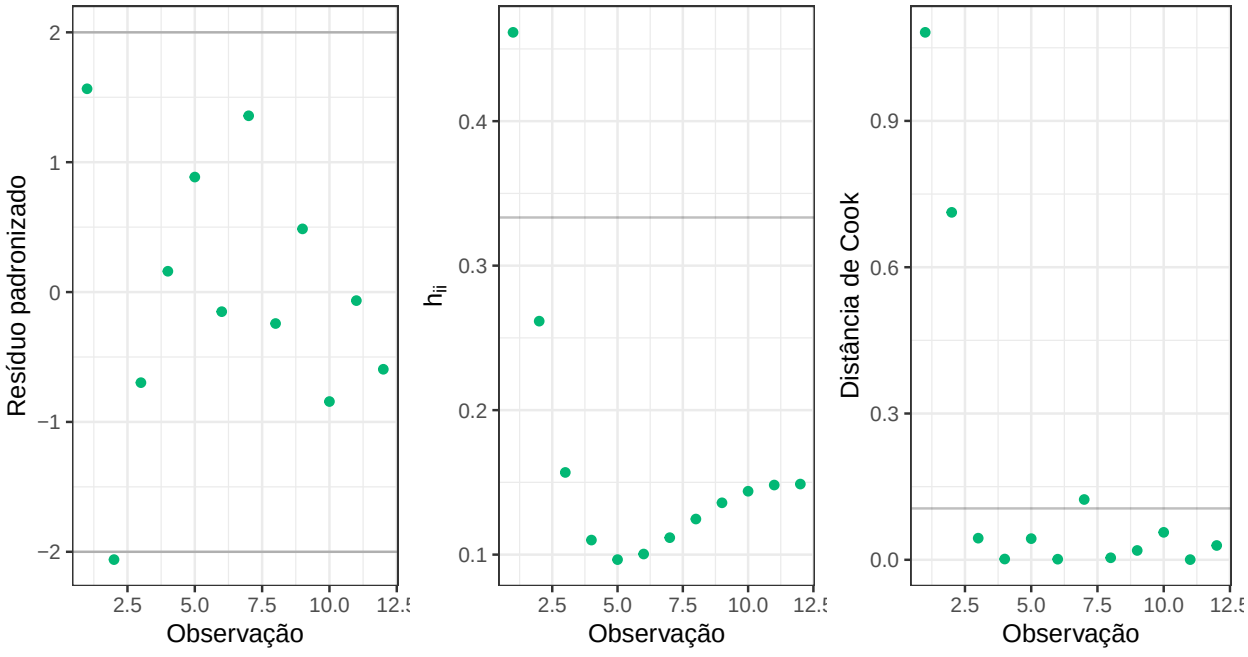
b) Realize uma análise de diagnóstico para o melhor modelo obtido em a). Analise os resultados obtidos.

Coeficiente	Estimativa	Erro padrão	Estatística	p-valor
(Intercepto)	5.306	0.0635	83.58	$< 10^{-16}$
Tempo	-0.229	0.0127	-18.02	$< 10^{-16}$

Os coeficientes do modelo são altamentes significativos atráves o teste de hipótese para os coeficientes, os testes de Lilliefors e Shapiro-Wilk contribuem para a hipótese de normalidade dos resíduos (p-valores 0.883 e 0.916).

```
diag <- boot::glm.diag(bac_fit2)
n <- nrow(bactérias)
p <- ncol(bactérias)

bactérias |>
    ggplot(aes(x = seq_len(n), y = diag$rd)) +
    geom_point(color = "#02b875") +
    geom_hline(alpha = 0.25, yintercept = c(-2, 2)) +
    labs(x = "Observação", y = "Resíduo padronizado")
bactérias |>
    ggplot(aes(x = seq_len(n), y = diag$h)) +
    geom_point(color = "#02b875") +
    geom_hline(alpha = 0.25, yintercept = 2 * p / n) +
    labs(x = "Observação", y = expression(h[i][i]))
bactérias |>
    ggplot(aes(x = seq_len(n), y = diag$cook)) +
    geom_point(color = "#02b875") +
    geom_hline(alpha = 0.25, yintercept = qchisq(0.1, p) / p) +
    labs(x = "Observação", y = "Distância de Cook")
```



Apenas as duas primeiras medidas do número de bactérias possui distância de Cook fora do padrão e h_{ii} alto, indicando que estas observações exercem grande influência na estimativa dos coeficientes por se tratarem de valores muito mais altos que o resto dos dados que se adequão bem ao modelo.

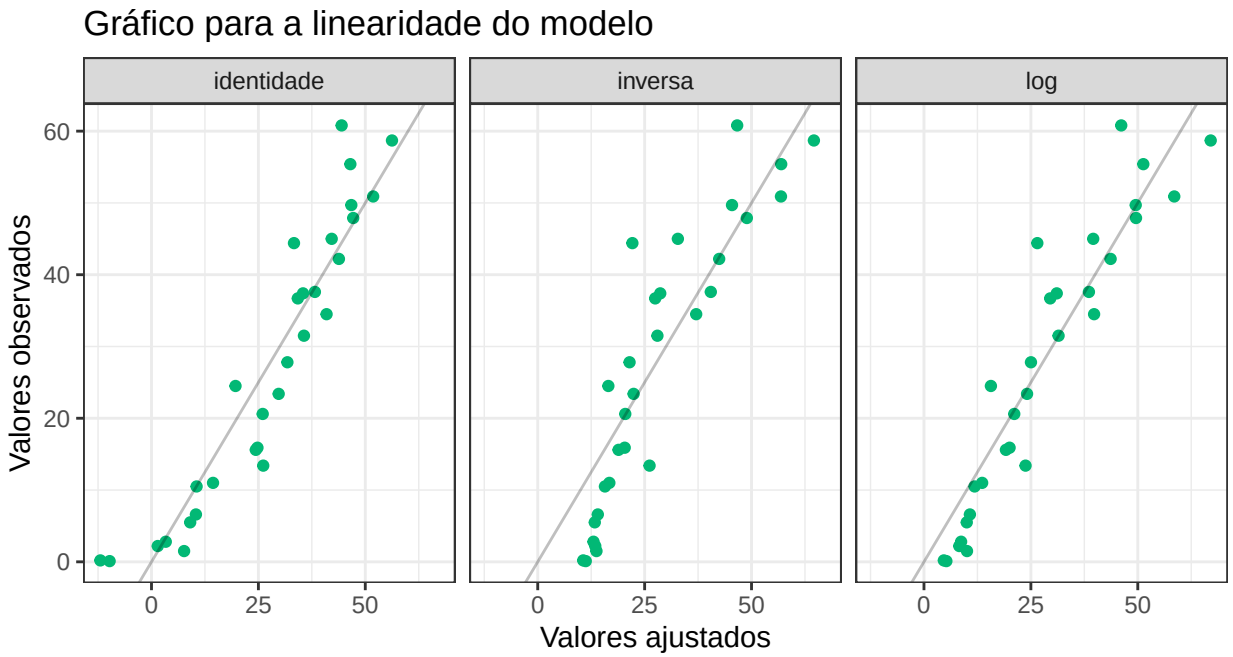
Questão 3. No banco de dados `defects.txt` indica a temperatura do processo de produção, a densidade do produto, taxa de produção e a média do número de defeitos nos produtos (variável resposta).

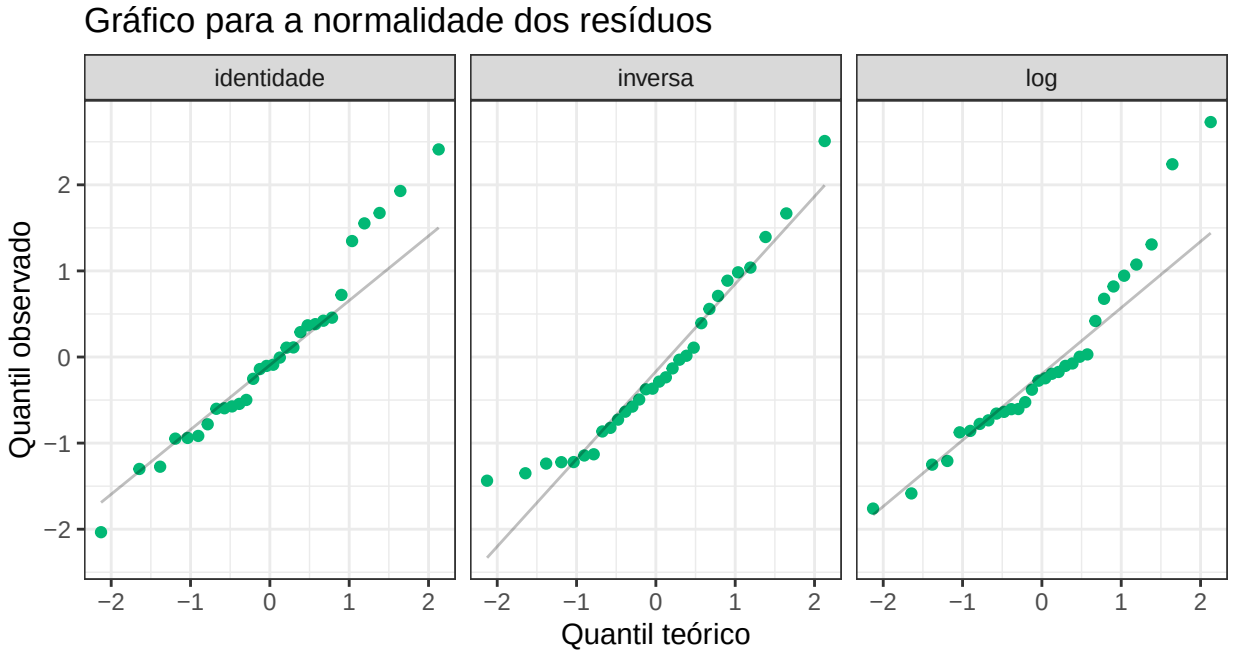
```
defeitos <- read.csv("defects.txt", header = FALSE)
names(defeitos) <- c("temperatura", "densidade", "produção", "defeitos")
```

a) Realize o ajuste da normal com as possíveis funções de ligação e decida, entre elas, qual a função de ligação é melhor. Justifique sua escolha.

```
def_fit1 <- glm(defeitos ~ ., defeitos, family = gaussian(link = identity))
def_fit2 <- glm(defeitos ~ ., defeitos, family = gaussian(link = inverse ))
def_fit3 <- glm(defeitos ~ ., defeitos, family = gaussian(link = log     ))
```

```
data.frame(
    y          = defeitos$defeitos,
    identidade = fitted(def_fit1),
    inversa    = fitted(def_fit2),
    log        = fitted(def_fit3)
) |>
    tidyr::pivot_longer(-y) |>
    ggplot(aes(x = value, y = y)) +
    facet_grid(cols = vars(name)) +
    geom_point(color = "#02b875") +
    geom_abline(alpha = 0.25) +
    labs(
        title = "Gráfico para a linearidade do modelo",
        x = "Valores ajustados",
        y = "Valores observados"
    )
data.frame(
    identidade = boot::glm.diag(def_fit1)$rd,
    inversa    = boot::glm.diag(def_fit2)$rd,
    log        = boot::glm.diag(def_fit3)$rd
) |>
    tidyr::pivot_longer(tidyr::everything()) |>
    ggplot(aes(sample = value)) +
    facet_grid(cols = vars(name)) +
    geom_qq(color = "#02b875") +
    geom_qq_line(alpha = 0.25) +
    labs(
        title = "Gráfico para a normalidade dos resíduos",
        x = "Quantil teórico",
        y = "Quantil observado"
    )
```





A função de ligação inversa superestima bastante a média do número de defeitos para aquelas amostras com baixo número de defeitos, enquanto que a função de ligação identidade subestima essas mesmas observações, predizendo até mesmo um número negativo de efeitos. com a função de ligação log temos um modelo mais linear porém com resíduos positivos altos, mesmo assim, escolho a ligação por log.

b) Selecione as variáveis. Realize uma análise residual para o melhor modelo obtido. O modelo é adequado? Por quê?

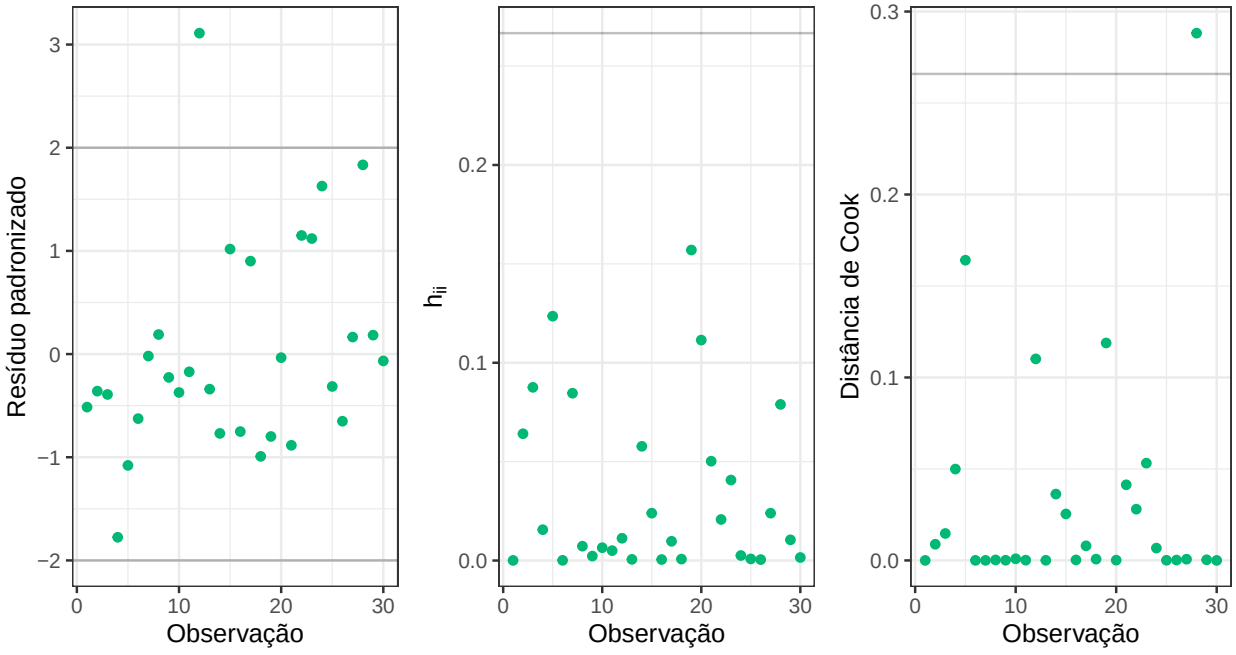
A iteração *stepwise* chegou em um modelo apenas com intercepto e a variável temperatura, porém o intercepto não era significante pelo teste t, então selecionei apenas a temperatura.

```r
def_fit <- glm(defeitos ~ temperatura - 1, defeitos, family = gaussian(link = log))
```

$$\ln \hat{\mu} = 1.376 \times \text{Temperatura}$$

```r
diag <- boot::glm.diag(def_fit)
n <- nrow(defeitos)
p <- ncol(defeitos)

defeitos |>
    ggplot(aes(x = seq_len(n), y = diag$rd)) +
    geom_point(color = "#02b875") +
    geom_hline(alpha = 0.25, yintercept = c(-2, 2)) +
    labs(x = "Observação", y = "Resíduo padronizado")
defeitos |>
    ggplot(aes(x = seq_len(n), y = diag$h)) +
    geom_point(color = "#02b875") +
    geom_hline(alpha = 0.25, yintercept = 2 * p / n) +
    labs(x = "Observação", y = expression(h[i][i]))
defeitos |>
    ggplot(aes(x = seq_len(n), y = diag$cook)) +
    geom_point(color = "#02b875") +
    geom_hline(alpha = 0.25, yintercept = qchisq(0.1, p) / p) +
    labs(x = "Observação", y = "Distância de Cook")
```



Uma observação teve distância de Cook muito mais alta doque as demais, a mesma possui o maior número médio de defeitos dos dados, já a observação com resíduo acima de 3 ocorre por um número mais elevado da média de defeitos para uma temperatura razoável.