
# CE073 - Análise de Dados Categóricos

## Trabalho No. 4

Luiz Henrique Barretta Francisco - GRR20213026

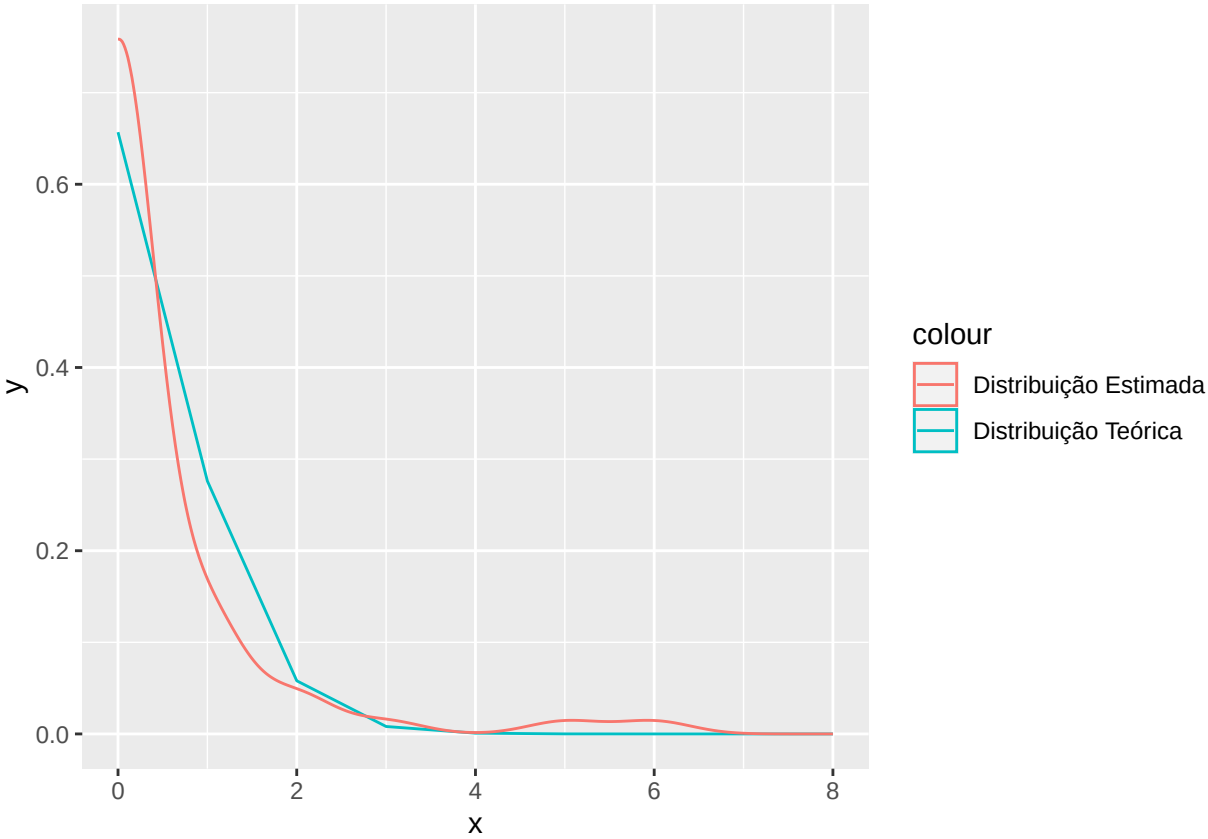
novembro/2023

**3-a) Faça um gráfico comparando as contagens observadas de estimadas com a função de probabilidade Poisson. Comente o ajuste do modelo de Poisson; inflação zero parece ser um problema?**

```
library(tidyverse)
Stormchaser <- read.csv(file = "http://leg.ufpr.br/~lucambio/ADC/Stormchaser.csv")

lambda <- mean(Stormchaser$tornadoes)
x <- seq(0, 8)
pois_teo <- data.frame(x = x, y = dpois(x, lambda))

ggplot()+
  geom_line(aes(x = x, y = y, color = "Distribuição Teórica"), data = pois_teo)+
  geom_density(aes(x = tornadoes, color = "Distribuição Estimada"), data = Stormchaser)
```




Do gráfico, concluímos que a inflação de zeros não parece ser um problema, pois as curvas são semelhantes e a distribuição estimada contempla a quantidade de zeros na amostra.

**3-b) Que suposições diferentes os modelos hurdle e ZIP fazem em relação à causa da contagem zero de tornados?**

Os modelos Hurdle e Zero-Inflated Poisson (ZIP) são ambos utilizados para lidar com a presença excessiva de zeros em dados de contagem, como é comum em registros de ocorrências de tornados. No entanto, esses modelos fazem diferentes suposições em relação à causa desses zeros.

O modelo Hurdle pressupõe que existem dois processos distintos que governam a ocorrência de zeros e de valores positivos. O primeiro processo é um modelo de decisão binária que determina se um evento ocorre ou não. Neste contexto, seria a decisão de se ocorrerá ou não um tornado durante uma perseguição de tempestade.

O segundo processo é um modelo de contagem, como o modelo Poisson, que descreve a distribuição dos valores positivos, ou seja, o número de tornados quando pelo menos um ocorre. Portanto, o Hurdle modela a probabilidade de zero tornadoes separadamente da distribuição condicional dos valores positivos, tratando os zeros como uma categoria distinta.

Por outro lado, o modelo ZIP assume que existem duas fontes distintas de zeros: os zeros genuínos, que ocorrem quando não há tornado presente, e os zeros inflados, que representam a probabilidade adicional de um evento ser contado como zero devido a alguma razão específica. Assim, o ZIP combina um modelo de excesso de zeros, semelhante ao Hurdle, com um modelo de contagem Poisson para os valores positivos. Em resumo, enquanto o Hurdle aborda os zeros através de dois processos separados, o ZIP incorpora a ideia de zeros genuínos e inflados em um único modelo.

**3-c) Existem muitos tipos diferentes de tempestades, mas os caçadores de tempestades procuram especificamente aquelas que parecem ter potencial para o desenvolvimento de tornados. Com isso em mente, descreva o que o parâmetro representaria em um modelo inflado de zeros.**

No contexto de um modelo ZIP, o parâmetro representa a probabilidade de um evento observado como zero ser resultado de uma explicação adicional para a ausência de tornados, além daquela considerada pelo modelo de contagem principal. Esse parâmetro inflacionado captura a probabilidade de zeros extras, denominados zeros inflados, que ocorrem devido a fatores específicos que não são abordados pelo modelo de contagem padrão. No caso dos caçadores de tempestades, esse parâmetro poderia refletir a probabilidade adicional de não encontrar tornados durante uma perseguição, mesmo quando as condições meteorológicas indicam um potencial para o seu desenvolvimento. Em outras palavras, ele quantifica a contribuição específica de fatores não modelados que podem levar a contagens de zero, apesar das condições propícias para a formação de tornados.

**3-d) Ajuste os modelos ZIP e hurdle, usando apenas umintercepto para ambas as porções de média e probabilidade dos modelos. Em cada modelo, estime ambos os parâmetros com intervalos de confiança de 95% e interprete os resultados.**

```
library(gamlss)
zip <- gamlss(tornadoes ~ 1, family = ZIP, data = Stormchaser)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 130.8232
## GAMLSS-RS iteration 2: Global Deviance = 121.2745
## GAMLSS-RS iteration 3: Global Deviance = 117.1938
## GAMLSS-RS iteration 4: Global Deviance = 116.67
## GAMLSS-RS iteration 5: Global Deviance = 116.6409
## GAMLSS-RS iteration 6: Global Deviance = 116.6395
## GAMLSS-RS iteration 7: Global Deviance = 116.6393
```

```
summary(zip)
```

```
## ******************************************************************
## Family:  c("ZIP", "Poisson Zero Inflated")
##
## Call:  gamlss(formula = tornadoes ~ 1, family = ZIP, data = Stormchaser)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4056     0.2463   1.647    0.104
##
```

```
## -------------------------------------------------------------------
## Sigma link function:  logit
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9454     0.3497   2.703  0.00869 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -------------------------------------------------------------------
## No. of observations in the fit:  69
## Degrees of Freedom for the fit:  2
##       Residual Deg. of Freedom:  67
##                       at cycle:  7
##
## Global Deviance:     116.6393
##             AIC:     120.6393
##             SBC:     125.1076
## *******************************************************************
```

```
confint(zip)
```

```
##                        2.5 %    97.5 %
## mu.(Intercept)    -0.0770693 0.8883051
## sigma.(Intercept)  0.2600000 1.6307957
```

```
hurdle <- pscl::hurdle(tornadoes ~ 1, data = Stormchaser, dist = "poisson")
summary(hurdle)
```

```
##
## Call:
## pscl::hurdle(formula = tornadoes ~ 1, data = Stormchaser, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.4492 -0.4492 -0.4492 -0.4492  5.9632
##
## Count model coefficients (truncated poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4077     0.2459   1.658   0.0973 .
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2809     0.2919  -4.389 1.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 10
## Log-likelihood: -58.32 on 2 Df
```

```
confint(hurdle)
```

```
##                          2.5 %     97.5 %
## count_(Intercept) -0.07429027  0.8897780
## zero_(Intercept)  -1.85297876 -0.7088889
```

Para o modelo Zero-Inflated Poisson (ZIP), os resultados indicam que o parâmetro de interceptação para o modelo de contagem (mu) é estimado em 0.4056, mas não é estatisticamente significativo ao nível de significância de 0.05 ($p = 0.104$). Isso sugere que, na ausência de outros preditores, a contagem média de

tornados durante as perseguições de tempestades não é estatisticamente diferente de zero. O parâmetro de interceptação para o componente inflacionado de zeros (sigma) é estimado em 0.9454, sendo estatisticamente significativo ($p = 0.00869$), indicando a presença de zeros inflados. A interpretação do parâmetro sigma é complicada, mas em termos gerais, valores positivos indicam uma maior probabilidade de zeros inflados.

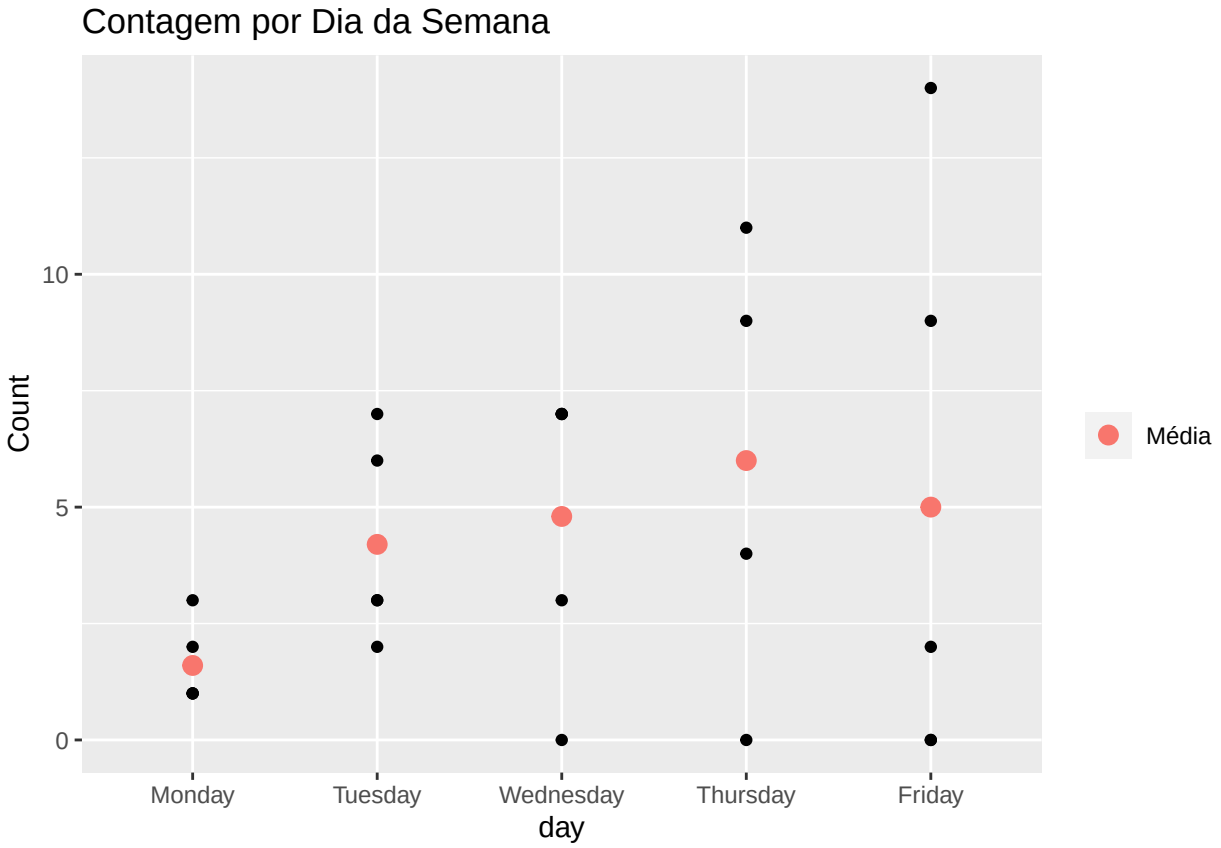
Já para o modelo Hurdle, os resultados revelam que o parâmetro de interceptação para o modelo de contagem (count) é estimado em 0.4077, e embora não seja significativo ao nível de 0.05 ($p = 0.0973$), sugere uma contagem média positiva de tornados durante as perseguições de tempestades. O parâmetro de interceptação para o modelo binomial que modela a probabilidade de zero tornados (zero) é estimado em -1.2809 e é estatisticamente significativo ($p < 0.001$), indicando que há uma probabilidade significativamente menor de encontrar zero tornados do que encontrar pelo menos um tornado durante uma perseguição. Em termos práticos, isso sugere que há uma probabilidade substancial de ocorrerem tornados durante as perseguições, mas com uma probabilidade menor de não ocorrerem.

**13-a) Qual é a população de inferência? Em outras palavras, defina a configuração para a qual se deseja estender as inferências com base nessa amostra.**

A população de inferência neste caso consiste em todas as visitas potenciais à loja da Starbucks em dias da semana. A amostra coletada nos cinco dias da semana ao longo de cinco semanas específicas fornece informações sobre o número de clientes na fila durante esse período específico de tempo. Portanto, a população de inferência é composta por todas as possíveis visitas nesse horário específico, que reflete a extensão das inferências desejadas para entender a diferença de comportamento do número de clientes na fila durante os dias.

**13-b) Construa gráficos de pontos lado a lado dos dados em que o eixo y fornece o número de clientes e o eixo x é o dia da semana. Descreva quais informações esse gráfico fornece sobre o número médio de clientes por dia. Em particular, parece plausível que a verdadeira contagem média seja constante ao longo dos dias? Recomendamos colocar os valores do fator fornecidos em Day em sua ordem cronológica usando a função factor() antes de concluir este gráfico.**

```
starbucks <- read.csv(file = "http://leg.ufpr.br/~lucambio/ADC/starbucks.csv")
starbucks <- starbucks %>% mutate(day = factor(Day, levels = c("Monday", "Tuesday", "Wednesday",
                                                               "Thursday", "Friday")))
medias_dia <- starbucks %>% group_by(day) %>% summarise(mean_count = mean(Count))
ggplot(starbucks)+
  geom_point(aes(x = day, y = Count))+
  geom_point(data = medias_dia, aes(x = day, y = mean_count, color = "Média"), size = 3)+
  labs(color = "", title = "Contagem por Dia da Semana")
```




Conforme mostrado no gráfico acima tanto a média quanto a variância parecem ser maiores com o passar dos dias da semana. Portanto, não é plausível que a média seja igual para todos os dias, porém essa hipótese deve ser testada de forma estatística.

**13-c) Usando um modelo de regressão Poisson que permite diferentes contagens médias em dias diferentes, conclua o seguinte:**

```
library(multcomp)
aj1 <- glm(Count ~ day, data = starbucks, family = poisson)
aj2 <- glm(Count ~ 1, data = starbucks, family = poisson)
anova(aj2, aj1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Count ~ 1
## Model 2: Count ~ day
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        24     87.432
## 2        20     72.431  4   15.002 0.004698 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(aj1, type = "II")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Count
##     LR Chisq Df Pr(>Chisq)
```

```
## day   15.002  4   0.004698 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(glht(aj1, linfct = mcp(day = "Tukey")))
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: glm(formula = Count ~ day, family = poisson, data = starbucks)
##
## Quantile = 2.7158
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                            Estimate lwr      upr
## Tuesday - Monday == 0       0.96508 -0.16325  2.09341
## Wednesday - Monday == 0     1.09861 -0.01009  2.20731
## Thursday - Monday == 0      1.32176  0.24113  2.40238
## Friday - Monday == 0        1.13943  0.03631  2.24256
## Wednesday - Tuesday == 0    0.13353 -0.67795  0.94502
## Thursday - Tuesday == 0     0.35667 -0.41601  1.12936
## Friday - Tuesday == 0       0.17435 -0.62950  0.97821
## Thursday - Wednesday == 0   0.22314 -0.52060  0.96688
## Friday - Wednesday == 0     0.04082 -0.73525  0.81689
## Friday - Thursday == 0     -0.18232 -0.91773  0.55308
```

```
round(exp(confint(aj1, calpha = qnorm(0.975))), 2)
```

```
##              2.5 % 97.5 %
## (Intercept)   0.73   2.98
## dayTuesday    1.21   6.31
## dayWednesday  1.41   7.13
## dayThursday   1.80   8.78
## dayFriday     1.47   7.41
```

Com base na análise estatística do modelo de regressão Poisson concluímos que há uma diferença estatisticamente significativa no número médio de clientes na fila em diferentes dias da semana. O teste de deviance mostra que o modelo que inclui o dia como variável explicativa é estatisticamente melhor do que o modelo com apenas a intercepto. Além disso, o teste de Tukey para comparações múltiplas entre os dias revela diferenças significativas entre alguns pares de dias, indicando que certos dias têm uma média de clientes na fila significativamente diferente de outros. A diferença entre quinta-feira e segunda-feira, sexta-feira e segunda-feira são todas estatisticamente significantes pois em seu intercepto de confiança não está completado o valor 0. Portanto, podemos inferir que o dia da semana tem uma influência significativa no número médio de clientes na fila.

**13-d) Discuta por que essas duas maneiras de escrever as hipóteses são equivalentes. Escreva as formas apropriadas do modelo de regressão Poisson para dar suporte ao seu resultado.**

As duas formas de escrever as hipóteses são equivalentes porque estão expressando a mesma ideia de que não há efeito significativo dos dias da semana ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$) versus a alternativa de que pelo menos um dos dias da semana tem um efeito diferente dos demais, ao menos um $\beta_i \neq 0$. A relação entre os

coeficientes da regressão Poisson e as médias ($\mu$) é direta, pois o modelo Poisson assume que a variável de resposta (número de clientes na fila) segue uma distribuição Poisson com média $\mu$.

O modelo de regressão Poisson pode ser expresso da seguinte forma:

$$log(\mu_i) = \beta_0 + \beta_1 Tuesday + \beta_2 Wednesday + \beta_3 Thursday + \beta_4 Friday$$

onde é $\mu_i$ a média do número de clientes na fila no dia i. Nesse contexto, a hipótese nula (H0) é representada por $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, indicando que não há efeito significativo dos dias da semana, e a hipótese alternativa (H1) é representada por Pelo menos um $\beta_i \neq 0$, indicando que pelo menos um dia da semana tem um efeito diferente dos outros.

Portanto, ambas as formas de expressar as hipóteses refletem a mesma ideia e são equivalentes: testam se há diferença significativa nas médias do número de clientes na fila entre os dias da semana.

**19) dados relativos à recorrência de tumores de câncer de bexiga em pacientes que receberam tratamento anterior para remover um tumor primário. Analise esses dados.**

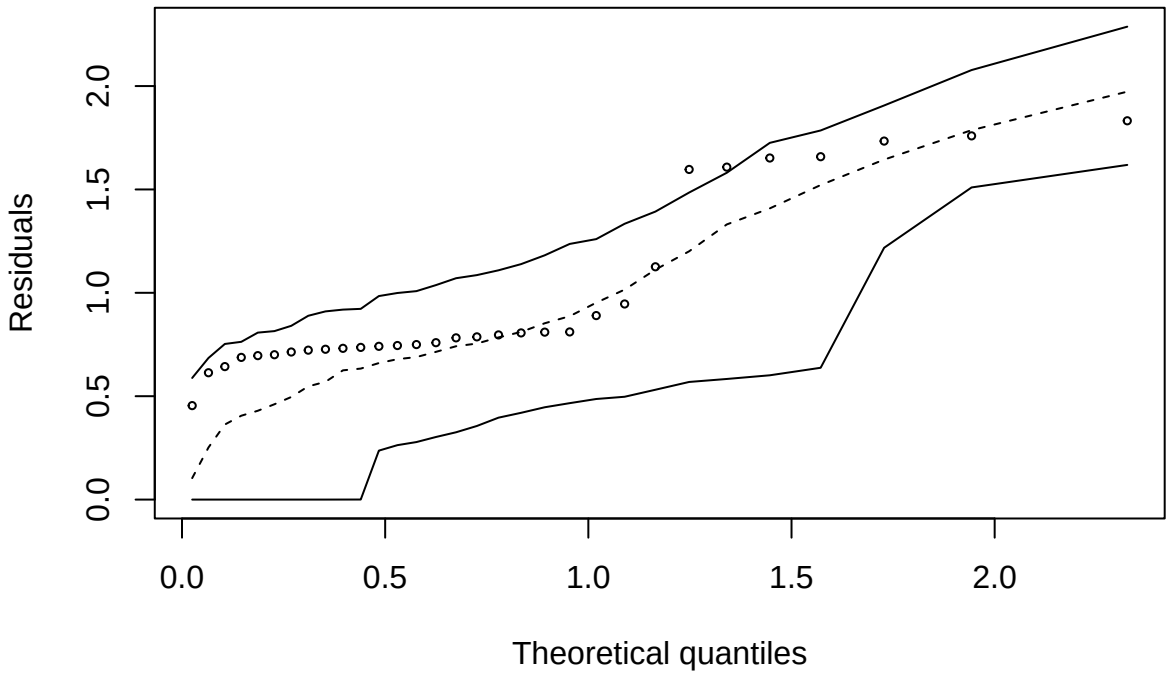
```
BladderCancer = read.csv(file = "http://leg.ufpr.br/~lucambio/ADC/BladderCancer.csv")
BladderCancer <- BladderCancer %>% mutate(size = factor(Size))
aj1 <- glm(Size ~ Tumors*Time, binomial, BladderCancer)
aj2 <- glm(Tumors ~ size, poisson, BladderCancer, offset = Time)
aj3 <- glm(Tumors ~ size, poisson, BladderCancer, offset = log(Time))
aj4 <- MASS::glm.nb(Tumors ~ size, BladderCancer, offset(Time))
aj5 <- MASS::glm.nb(Tumors ~ size, BladderCancer, offset(exp(Time)))
aj6 <- glm(Tumors ~ size, quasipoisson(), BladderCancer, offset = Time)
aj7 <- glm(Tumors ~ size, quasipoisson(), BladderCancer, offset = log(Time))

model_list <- list(aj1, aj2, aj3, aj4, aj5)
aic_values <- sapply(model_list, AIC)
best_model_index <- which.min(aic_values)
aic_values
```

```
## [1] 4.217776e+01 9.550495e+02 1.036868e+02 1.242372e+03 2.122849e+12
```

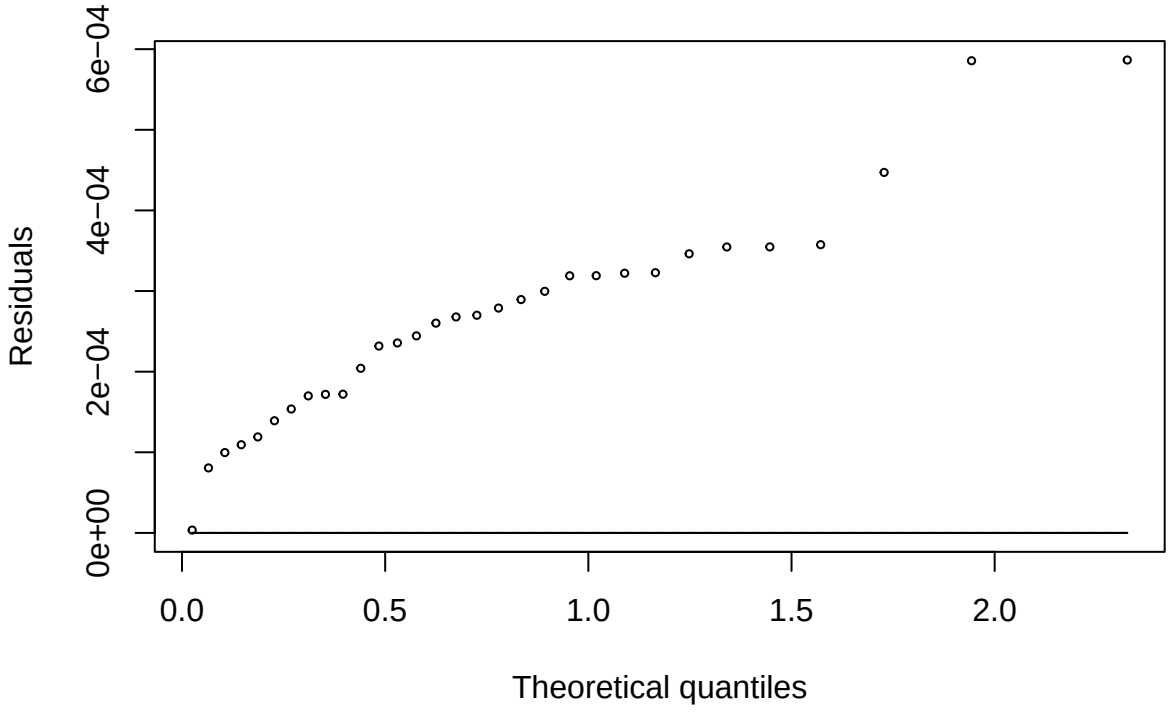
```
best_model <- model_list[[best_model_index]]
hnp::hnp(best_model)
```

```
## Binomial model
```



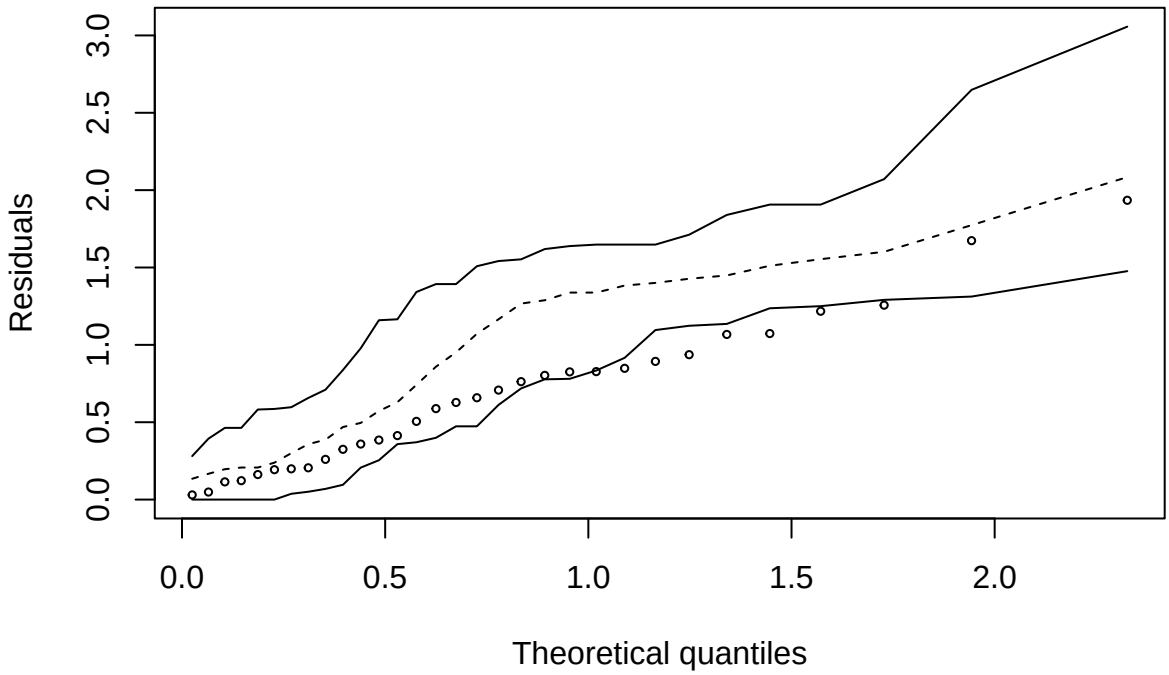

```
hnp::hnp(aj6)
```

```
## Quasi-Poisson model
```




```
hnp::hnp(aj7)
```

```
## Quasi-Poisson model
```




```
summary(best_model)
```

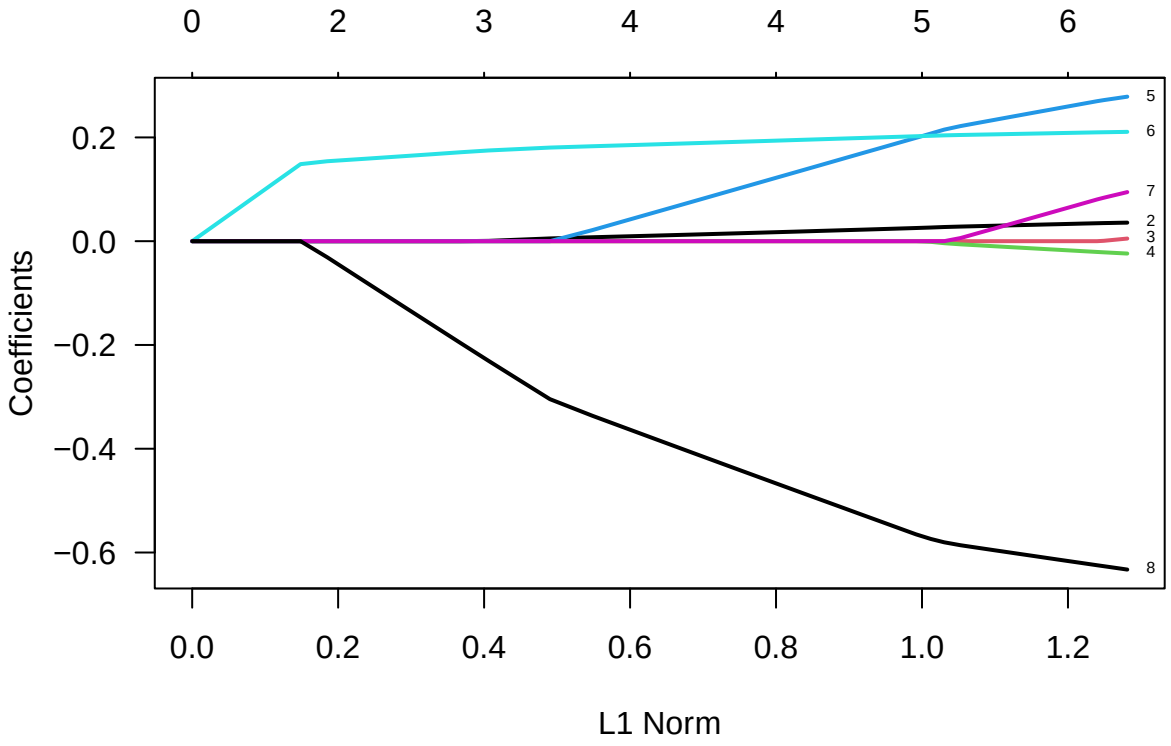
```
##
## Call:
## glm(formula = Size ~ Tumors * Time, family = binomial, data = BladderCancer)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.04578    1.85789  -1.639    0.101
## Tumors       1.72410    1.15683   1.490    0.136
## Time         0.10296    0.11344   0.908    0.364
## Tumors:Time -0.08896    0.07127  -1.248    0.212
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37.351  on 30  degrees of freedom
## Residual deviance: 34.178  on 27  degrees of freedom
## AIC: 42.178
##
## Number of Fisher Scoring iterations: 4
```

Foram ajustados 7 diferentes modelo de regressão e ao final, foi comparado o Half Normal Plot do modelo com menor AIC, e com os outros dois que utilizarm um método de Quasi-Poisson. A escolha se deu pelo modelo do ajuste 1, que pelos gráficos de HNP é o que se comporta melhor se mantendo dentro das bandas de confiança.

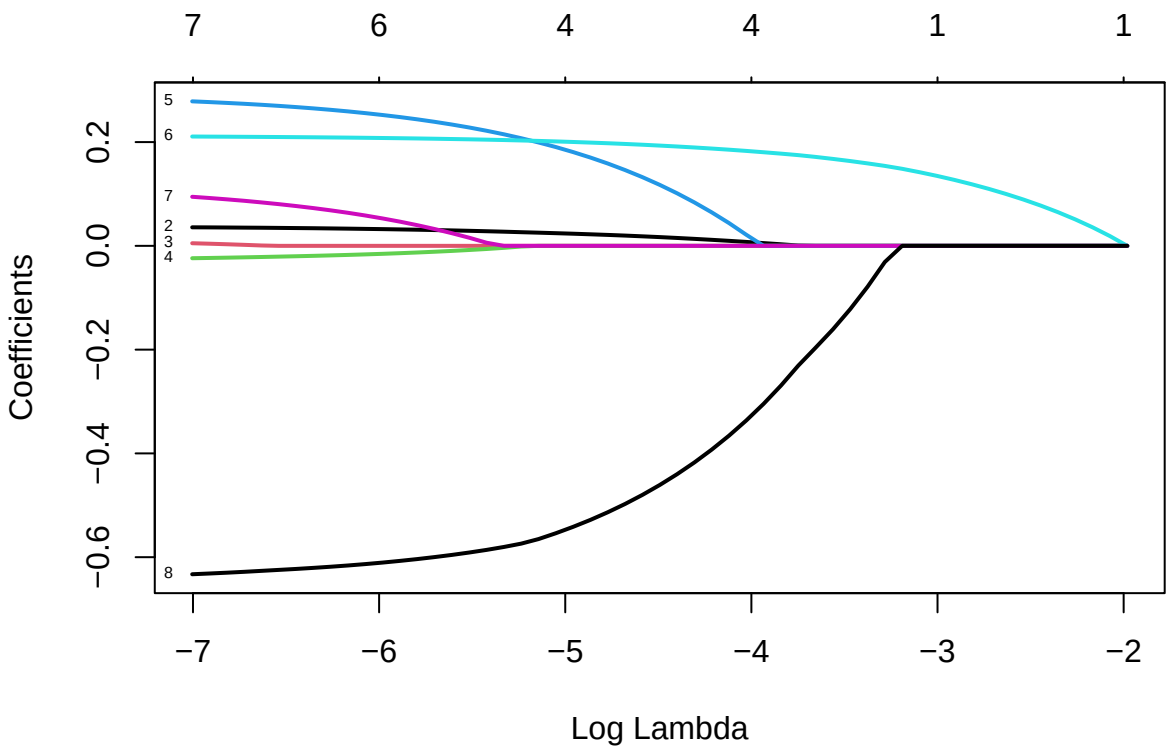
O teste de significância para os coeficientes indica que, a um nível de significância de 0.05, nenhum dos coeficientes é estatisticamente significativo. Com base na interpretação desse modelo, não encontramos evidências suficientes para concluir que o tamanho do tumor primário, o número de tumores ou o tempo de acompanhamento estão significativamente relacionados à probabilidade de recorrência de tumores de bexiga após o tratamento inicial.

**9) Considere o problema de tentar prever se uma pessoa tem seguro privado com base em seu padrão de uso de serviços de saúde. Isso sugere um modelo de regressão logística para privins, uma variável binária. Use o LASSO para identificar quais das variáveis restantes se relacionam com a probabilidade de uma pessoa ter seguro privado. Interpretar os resultados; em particular, estimar o efeito de cada variável explicativa sobre essa probabilidade.**

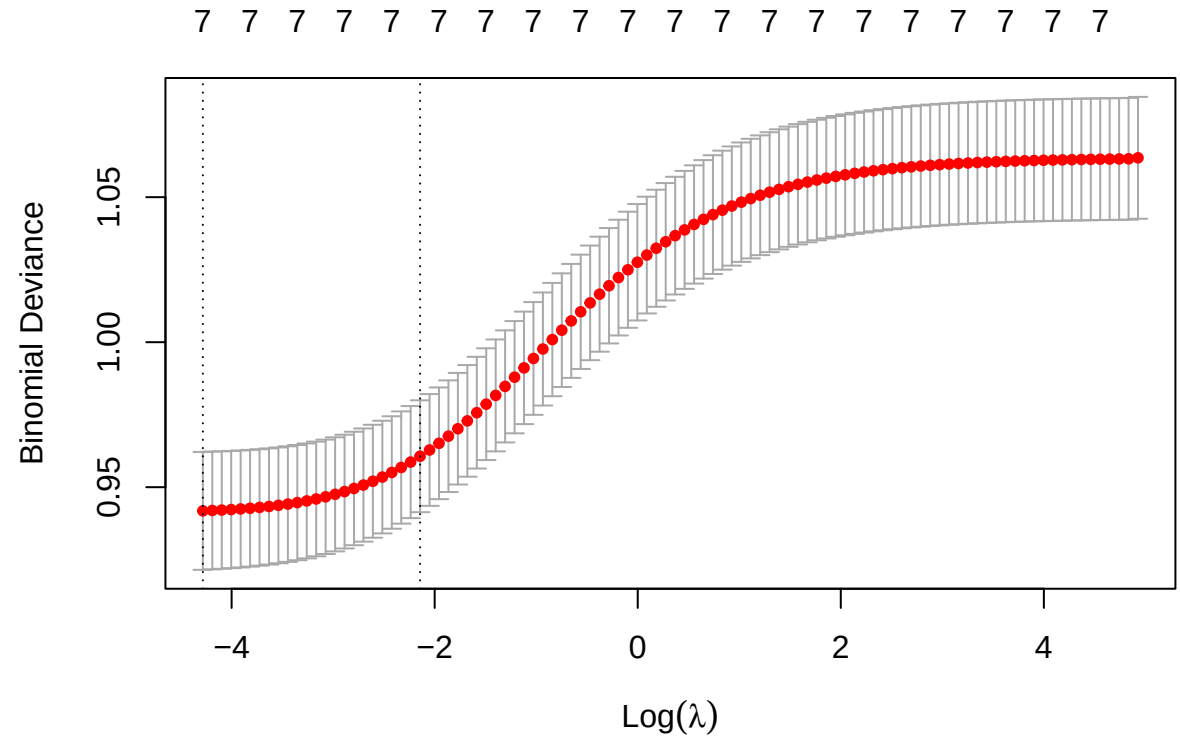
```
library(glmnet)
dt = read.csv( file = "http://leg.ufpr.br/~lucambio/ADC/dt.csv" )
X <- model.matrix(privins ~ ., data = dt)
y <- dt$privins
lasso <- glmnet(y = y , x = X , family = "binomial")
plot(lasso, las = 1, lwd = 2, label=TRUE)
```




```
plot(lasso, xvar="lambda", label=TRUE, lwd = 2, cex = 20)
```




```
cvfit <- cv.glmnet(X, y, family = 'binomial', alpha = 0, nfolds = 20)
plot(cvfit)
```




```
cvfit$lambda.min
```

```
## [1] 0.0137996
```

```r
coef(lasso, s=cvfit$lambda.min)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)      -0.61758404
## (Intercept)       .
## ofp               0.01304904
## hosp              .
## numchron          .
## gender            0.07913725
## school            0.18911465
## health_excellent  .
## health_poor      -0.41148445
```

```r
nonzero_coefs <- which(coef(lasso, s = cvfit$lambda.min) != 0)
selected_coefs <- coef(lasso, s = cvfit$lambda.min)[nonzero_coefs]
selected_coefs
```

```
## [1] -0.61758404  0.01304904  0.07913725  0.18911465 -0.41148445
```

```r
exp(selected_coefs)
```

```
## [1] 0.5392457 1.0131346 1.0823529 1.2081795 0.6626658
```

O conjunto final de variáveis selecionadas inclui “ofp” (número de ofertas para emprego), “gender” (gênero), “school” (anos de escolaridade) e “health_poor” (avaliação da saúde como ruim). Os coeficientes estimados para essas variáveis indicam o efeito que cada uma tem na log-odds da probabilidade de possuir seguro privado.

Interpretando os coeficientes exponenciados, observamos que:

1. Para “ofp”, um aumento de uma unidade está associado a um aumento de aproximadamente 1,3% nas odds de possuir seguro privado.
2. Para “gender” (gênero), ser do sexo masculino está associado a um aumento de aproximadamente 8,2% nas odds de possuir seguro privado.
3. Para “school” (anos de escolaridade), um aumento de uma unidade está associado a um aumento de aproximadamente 20,8% nas odds de possuir seguro privado.
4. Para “health_poor” (avaliação da saúde como ruim), ter uma saúde ruim está associado a uma diminuição de aproximadamente 33,7% nas odds de possuir seguro privado.