Questão 4

Lucas Lacerda Oliveira

Análise de Dados Categorizados

Questão 4

Iniciamos criando a base de dados com o script abaixo:

```
dados <- array(
  c(11, 42, 43, 169, 14, 20, 104, 132, 8, 2, 196, 59),
  dim = c(2, 2, 3),
  dimnames = list(
    "escoteiro" = c("sim", "não"),
    "delinquente" = c("sim", "não"),
    "nível socioeconômico" = c("baixo", "médio", "alto")
  )
)
```

Com o script abaixo definimos as tabelas marginais de acordo com seus níveis.

- Tabela marginal escoteiro x deliquente

```
dados_ed <- margin.table(dados, c(1, 2))
```

- Tabela marginal nivel x delinquente

```
dados_nd <- margin.table(dados, c(3, 2))
```

- Tabela marginal escoteiro x nivel

```
dados_en <- margin.table(dados, c(1, 3))
```

Alternativa a Abaixo vamos verificar se existe independência multua entre as variáveis. Para isso utilizaremos o teste qui-quadrado o qual obteremos os resultados pela saída da função `chisq.test`.

```
chisq.test(dados_ed)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dados_ed
## X-squared = 6.884, df = 1, p-value = 0.008697
```

```r
chisq.test(dados_en)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dados_en
## X-squared = 172.2, df = 2, p-value < 2.2e-16
```

```r
chisq.test(dados_nd)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dados_nd
## X-squared = 32.826, df = 2, p-value = 7.445e-08
```

Com os testes acima, vemos que para todos os casos temos a rejeição da hipótese nula de independência entre as variáveis das tabelas marginais.

Alternativa b Para testar se as variáveis são condicionalmente independentes temos que verificar se $E \perp D|S$, sendo D a nossa variável resposta. O teste ideal que temos para esse tipo de validação é o teste de Mantel-Haenszel que nos traz a informação da existência de independência que seja estatísticamente significante.

A partir do comando abaixo temos os resultados do teste:

```r
mantelhaen.test(dados, correct = FALSE)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  dados
## Mantel-Haenszel X-squared = 0.0080042, df = 1, p-value = 0.9287
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5970214 1.6009845
## sample estimates:
## common odds ratio
##         0.9776615
```

Podemos ver que temos evidência suficiente para dizer que a hipótese de indepência condicional não é rejeitada.

Alternativa c Para fazer o teste de associação homogênea entre as variáveis, vamos utilizar o teste Breslow-Day. No j-ésimo estrato, dado as margens de uma tabela 2x2 $(m_{1j}, m_{2j}, n_{1j}, n_{2j})$ sub a hipótese de homogeniedade $OR_j = OR$, ou seja, basicamente testamos se as razões de chances são iguais para cada um dos níveis da nossa variável estratificadora. Vamos avaliar a saída do teste abaixo que irá aplicar o teste aos nossos dados.

```r
BreslowDayTest(dados)
```

```
##
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  dados
## X-squared = 0.1518, df = 2, p-value = 0.9269
```

Com o resultado acima podemos ver que a hipótese de homogeniedade entre S, E e D não é rejeitada.

Alternativa d Para ajustar o modelo iremos criar a base de dados necessária

```r
dados <- data.frame(
  "nível" = factor(c(
    rep("baixo", 265),
    rep("médio", 270),
    rep("alto", 265)
  ),
  levels = c("alto", "médio", "baixo")),
  "escoteiro" = as.factor(
    c(
      rep("sim", 11),
      rep("não", 42),
      rep("sim", 43),
      rep("não", 169),
      rep("sim", 14),
      rep("não", 20),
      rep("sim", 104),
      rep("não", 132),
      rep("sim", 8),
      rep("não", 2),
      rep("sim", 196),
      rep("não", 59)
    )
  ),
  "delinquente" = as.factor(c(
    rep("sim", 53),
    rep("não", 212),
    rep("sim", 34),
    rep("não", 236),
    rep("sim", 10),
    rep("não", 255)
  ))
)

head(dados)
```

```
##   nível escoteiro delinquente
## 1 baixo       sim         sim
## 2 baixo       sim         sim
## 3 baixo       sim         sim
## 4 baixo       sim         sim
```

```
## 5 baixo       sim         sim
## 6 baixo       sim         sim
```

```
fit <-
  glm(delinquente ~ ., data = dados, family = binomial("logit"))
summary(fit)
```

```
##
## Call:
## glm(formula = delinquente ~ ., family = binomial("logit"), data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6694  -0.6627  -0.5158  -0.2767   2.5621
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.22139    0.37545  -8.580  < 2e-16 ***
## nívelmédio    1.29371    0.38024   3.402 0.000668 ***
## nívelbaixo    1.83965    0.38421   4.788 1.68e-06 ***
## escoteirosim -0.02252    0.25123  -0.090 0.928579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 591.05  on 799  degrees of freedom
## Residual deviance: 554.79  on 796  degrees of freedom
## AIC: 562.79
##
## Number of Fisher Scoring iterations: 6
```

Resultados

- item 1:

Com os resultados que podemos ver acima temos os betas estimados para nível baixo e médio sendo >0 e isso significa que essas características aumentam a probabilidade de ser classificado como delinquente. Também tivemos um indício de que ser escoteiro reduziria a probabilidade de ser classificado como delinquente, mas o beta é muito próximo a zero e também acabou sendo não significante essa variável.

- item 2:

A seguir temos as probabilidades de cada uma das coombinações:

Escoteiro e classe baixa

```
exp(-3.22139 + 1.293715 * 0 + 1.83965 * 1 - 0.02252 * 1) / (1 + exp(-3.22139 + 1.293715 * 0 + 1.83965 *
```

```
## [1] 0.197141
```

Não escoteiro e classe baixa

```r
exp(-3.22139 + 1.293715 * 0 + 1.83965 * 1 - 0.02252 * 0) / (1 + exp(-3.22139 + 1.293715 * 0 + 1.83965 *
```

```
## [1] 0.201631
```

Escoteiro e classe média

```r
exp(-3.22139 + 1.293715 * 1 + 1.83965 * 0 - 0.02252 * 1) / (1 + exp(-3.22139 + 1.293715 * 1 + 1.83965 *
```

```
## [1] 0.1245321
```

Não escoteiro e classe média

```r
exp(-3.22139 + 1.293715 * 1 + 1.83965 * 0 - 0.02252 * 0) / (1 + exp(-3.22139 + 1.293715 * 1 + 1.83965 *
```

```
## [1] 0.1273684
```

Escoteiro e classe alta

```r
exp(-3.22139 + 1.293715 * 0 + 1.83965 * 0 - 0.02252 * 0) / (1 + exp(-3.22139 + 1.293715 * 0 + 1.83965 *
```

```
## [1] 0.03840148
```

Não esccoteiro e classe alta

```r
exp(-3.22139 + 1.293715 * 0 + 1.83965 * 0 - 0.02252 * 0) / (1 + exp(-3.22139 + 1.293715 * 0 + 1.83965 *
```

```
## [1] 0.03840148
```

Podemos ver acima que os resultados não são tão influenciados quando estamos tratando do jovem ter participado ou não dos escoteiros e, como citado no item anterior, vemos que a variável escoteiro não tam tanta influência assim sobre a probabilidade de ser classificado como delinquente ou não, a variável resposta acaba sendo melhor explicada pelo nível socioeconômico. Isso explica muito também o caso de não termos idependência marginal mas termos a não rejeição da hipótese de indepedência condicional.

- item 3:

Como vimos no item 2 a probabilidade de um jovem escoteiro ser classificado como delinquente é maior que a de ser classificado como delinquente dado que ele é de classe alta. Podemos calcular o OR para esse caso com $e^{\beta_{media}}$.

```r
exp(1.29371)
```

```
## [1] 3.646289
```

E com isso vemos que a chance de ser classificado como delinquente é cerca de **3,65x** maior para um jovem escoteiro de classe média quando comparado ao de classe alta.

- item 4:

Para fazer análise diagnóstico podemos testar se existe multicolinearidade

```r
car::vif(fit)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## nível     1.176967  2        1.041576
## escoteiro 1.176967  1        1.084881
```

Podemos ver que não há indicios de multicolineariedade no modelo.

Vamos agora analisar os resíduos.

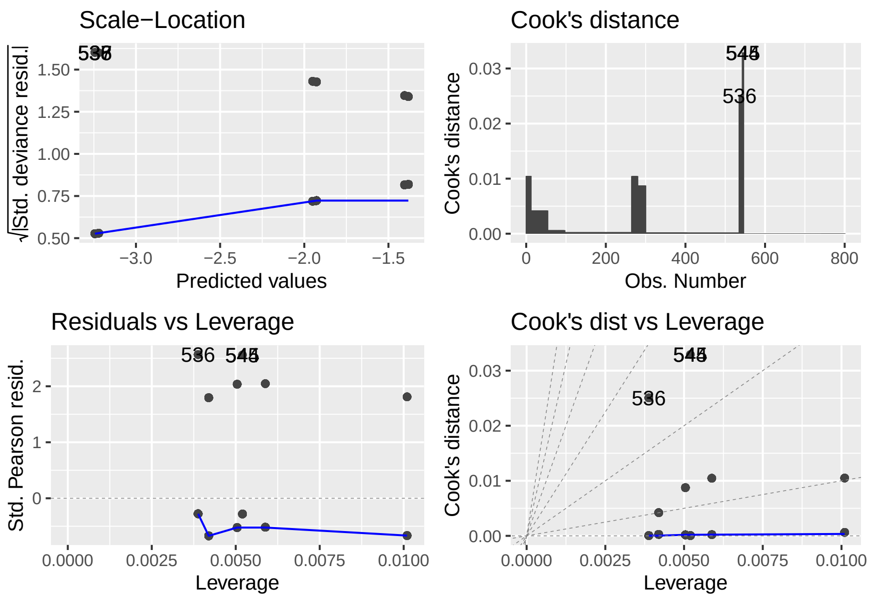

Figure 1: Figura 1: Gráfico de Distância de Cook.

Com os gráficos acimo podemos ver que as observações 536 e 545 muito provavelmente são pontos de influência.