ME720 - Modelos Lineares Generalizados

Profa. Larissa Avila Matos

Exemplo: Pós-graduação

Um estudo analisa os fatores que influenciam a decisão de um aluno se inscrever para uma pós-graduação. Pergunta-se aos alunos de graduação se é improvável (`unlikely`), pouco provável (`somewhat likely`) ou muito provável (`very likely`) dele se candidatar ao progrma de pós-graduação. Portanto, nossa variável de resposta possui três categorias.

Dados sobre a escolaridade dos pais, se a instituição de graduação é pública ou privada, e o GPA atual também é coletado.

Os pesquisadores têm motivos para acreditar que as "distâncias" entre esses três pontos não são iguais. Por exemplo, a "distância" entre improvável e um tanto provável pode ser menor do que a distância entre pouco provável e muito provável.

```
dados <- read.dta("ologit.dta")
# download dos dados em https://stats.idre.ucla.edu/stat/data/ologit.dta
head(dados)
```

```
            apply pared public  gpa
1     very likely     0      0 3.26
2 somewhat likely     1      0 3.21
3        unlikely     1      1 3.94
4 somewhat likely     0      0 2.81
5 somewhat likely     0      0 2.53
6        unlikely     0      1 2.59
```

Esse conjunto de dados hipotético possui:

- uma variável com três níveis denominada `apply`, com os níveis: `improvável`, `pouco provável` e `muito provável`, codificadas 1, 2 e 3, respectivamente, que usaremos como nossa variável reposta;
- três variáveis que usaremos como preditores:
  - `pared`: variável binária, indicando se pelo menos um dos pais tem graduação;
  - `public`: variável binária, em que 1 indica que a instituição de graduação é pública e 0 privada; e
  - `gpa`: média da nota do aluno.

Vamos começar com as estatísticas descritivas dessas variáveis.

```
lapply(dados[, c("apply", "pared", "public")], table)
```

```
$apply

       unlikely somewhat likely     very likely
            220             140              40

$pared

  0   1
337  63

$public

  0   1
343  57
```

```
ftable(xtabs(~ public + apply + pared, data = dados))
```

```
                      pared   0   1
public apply
0      unlikely              175  14
       somewhat likely        98  26
       very likely            20  10
1      unlikely               25   6
       somewhat likely        12   4
       very likely             7   3
```

```
summary(dados$gpa)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.900   2.720   2.990   2.999   3.270   4.000
```

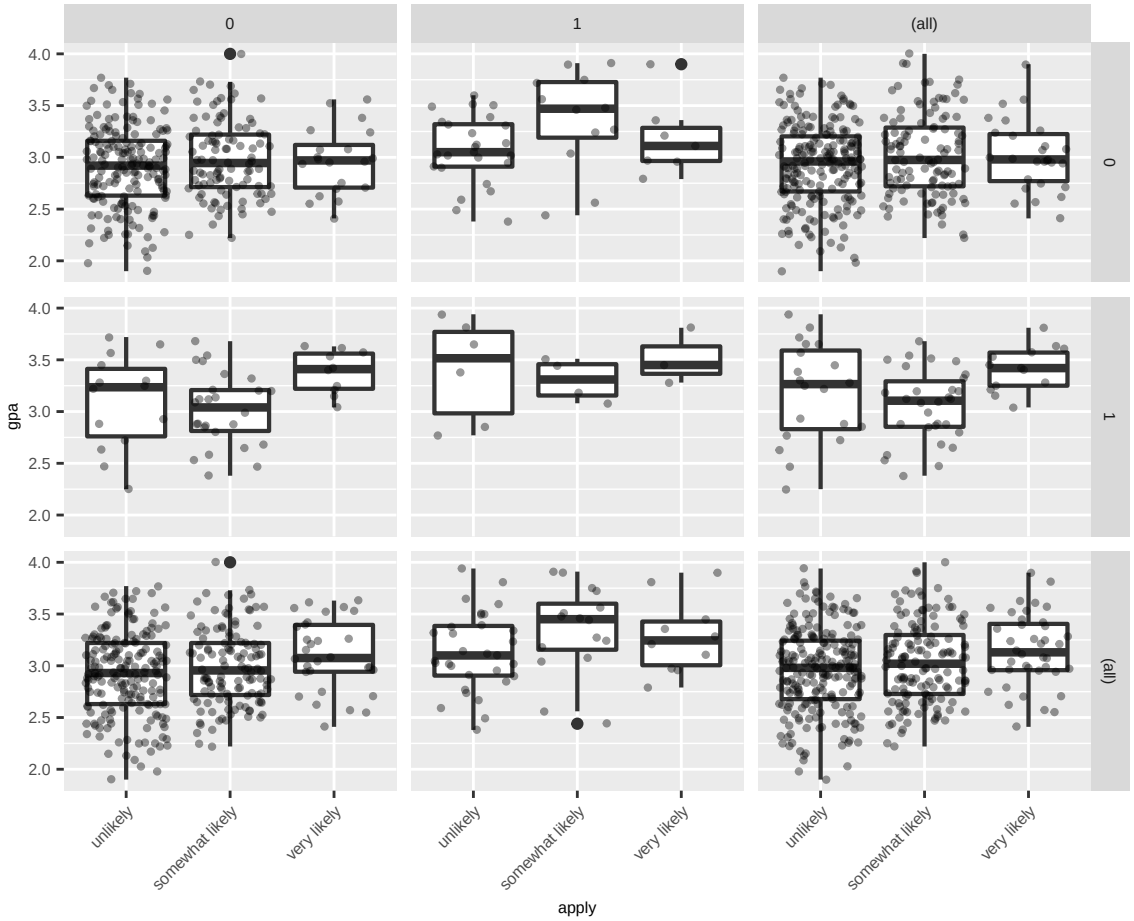
```
sd(dados$gpa)
```

```
[1] 0.3979409
```

Também podemos examinar a distribuição da variável `gpa` em todos os níveis da variavél `apply` (variável resposta) e discriminados pelas variáveis `public` e `pared`.

Isso cria um grid 2x2 com um boxplot para `gpa` para cada nível de aplicação (`apply`), para valores específicos de `pared` e `public`. Para ver melhor os dados, adicionamos os dados brutos nos boxplots.

```
ggplot(dados, aes(x = apply, y = gpa)) + geom_boxplot(size = .75) + geom_jitter(alpha = .4, size = 0.8) +
  facet_grid(pared ~ public, margins = TRUE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1, size=6),
        plot.title=element_text(size=7),plot.subtitle=element_text(size=6),
        axis.title.x=element_text(size=6),axis.title.y=element_text(size=6),
        axis.text.y=element_text(size=6),strip.text = element_text(size=6))
```



Métodos para análise

Você pode considerar alguns métodos para a análise:

1. Regressão logística cumulativa: nosso foco;
2. Regressão linear simples ordinal: Essa análise é problemática porque as suposições do modelo são violadas;
3. Regressão logística multinomial: é semelhante a fazer regressão logística ordenada, exceto que se assume que não há ordem para as categorias da variável de resultado (ou seja, as categorias são nominais). A desvantagem dessa abordagem é que as informações contidas no pedido são perdidas; e
4. Regressão probit cumulativa: é muito, muito semelhante à execução de uma regressão logística cumulativa. A principal diferença está na interpretação dos coeficientes.

Regressão logística cumulativa

Primeiramente, para o ajuste do modelo de regressão logística cumulativa, usamos a função `polr` do pacote `MASS`.

A função `polr` usa a interface de fórmula padrão do `R` para especificar um modelo de regressão com resultado seguido por seus preditores. Podemos especificar `Hess = TRUE` para que o modelo retorne a matriz de informação observada da otimização (matriz Hessiana), que é usada para obter os erros padrão.

Além disso, na função `polr` o modelo de regressão logística cumulativo é parametrizado como

$$logito(\mathbb{P}(Y \leq j)) = log\left(\frac{\mathbb{P}(Y \leq j)}{\mathbb{P}(Y > j)}\right) = \alpha_j - \beta_1 \boldsymbol{x}_1 - \beta_2 \boldsymbol{x}_2 - \ldots - \beta_p \boldsymbol{x}_p.$$

Então, ajustando esse modelo, temos

```
fit <- polr(apply ~ pared + public + gpa, data = dados, Hess=TRUE)
summary(fit)
```

```
Call:
polr(formula = apply ~ pared + public + gpa, data = dados, Hess = TRUE)

Coefficients:
          Value Std. Error t value
pared   1.04769     0.2658  3.9418
public -0.05879     0.2979 -0.1974
gpa     0.61594     0.2606  2.3632

Intercepts:
                           Value   Std. Error t value
unlikely|somewhat likely    2.2039  0.7795     2.8272
somewhat likely|very likely 4.2994  0.8043     5.3453

Residual Deviance: 717.0249
AIC: 727.0249
```

O modelo estimado pode ser escrito como:

$$logito(\widehat{\mathbb{P}}(Y \leq 1)) = 2,20 - 1,05 * pared - (-0,06) * public - 0,616 * gpa; \quad \text{e}$$
$$logito(\widehat{\mathbb{P}}(Y \leq 2)) = 4,30 - 1,05 * pared - (-0,06) * public - 0,616 * gpa.$$

Na saída acima, vemos

- `Call`, este é o R lembrando-nos que tipo de modelo executamos, que opções especificamos, etc.
- A seguir, vemos na saída a tabela de coeficiente de regressão usual, incluindo o valor de cada coeficiente, erros padrão e valor da estatística t, que é simplesmente a razão do coeficiente pelo seu erro padrão. Não há teste de significância por default.
- Depois, vemos as estimativas para os dois interceptos.
- Finalmente, vemos o desvio, -2*log-verossimilhança do modelo, bem como a AIC. O desvio e a AIC são úteis para comparação de modelos.

Podemos calcular manualmente o p-valor para os coeficientes do modelo

```
ctable <- coef(summary(fit))
p <- 2*pnorm(abs(ctable[, "t value"]), lower.tail = FALSE)
ctable <- cbind(ctable, "p value" = p)
ctable
```

```
                                Value Std. Error    t value      p value
pared                        1.04769010  0.2657894  3.9418050 8.087072e-05
public                      -0.05878572  0.2978614 -0.1973593 8.435464e-01
gpa                          0.61594057  0.2606340  2.3632399 1.811594e-02
unlikely|somewhat likely     2.20391473  0.7795455  2.8271792 4.696004e-03
somewhat likely|very likely  4.29936315  0.8043267  5.3452947 9.027008e-08
```

Também podemos obter intervalos de confiança para as estimativas dos parâmetros. Se o IC de 95% não conter o 0, a estimativa do parâmetro é estatisticamente significativa.

```
IC=confint(fit)
```

```
Waiting for profiling to be done...
```

```
IC
```

```
            2.5 %    97.5 %
pared   0.5281768 1.5721750
public -0.6522060 0.5191384
gpa     0.1076202 1.1309148
```

```
confint.default(fit) # IC assumindo normalidade
```

```
            2.5 %    97.5 %
pared   0.5267524 1.5686278
public -0.6425833 0.5250119
gpa     0.1051074 1.1267737
```

```
cbind(1.048-1.96*0.266,1.048+1.96*0.266)
```

```
        [,1]    [,2]
[1,] 0.52664 1.56936
```

```
cbind(-0.059-1.96*0.298,-0.059+1.96*0.298)
```

```
         [,1]    [,2]
[1,] -0.64308 0.52508
```

```
cbind(0.616-1.96*0.261,0.616+1.96*0.261)
```

```
        [,1]    [,2]
[1,] 0.10444 1.12756
```

Os ICs para `pared` e `gpa` não incluem 0, mas `public` sim.

As estimativas na saída são fornecidas em unidades de logitos cumulativos ou log chances ordenadas.

Portanto, para o `pared`, diríamos que, para o aumento de uma unidade no `pared` (ou seja, passando de 0 para 1), esperamos um aumento de 1,05 no valor esperado de aplicação na escalad e log de chances, considerando que todas as outras variáveis no modelo são mantidas constantes.

Para o `gpa`, diríamos que, para o aumento de uma unidade no `gpa`, esperaríamos um aumento de 0,62 no valor esperado da variável `apply` na escala de log de chances, uma vez que todas as outras variáveis no modelo são mantidas constantes.

Os coeficientes do modelo podem ser um pouco difíceis de interpretar porque são dimensionados em termos de logs.

Outra maneira de interpretar os modelos de regressão logística é converter os coeficientes para razão de chances (RC). Para obter as RCs e os intervalos de confiança para as RCs, apenas precisamos aplicar a exponencial nas estimativas e nos intervalos de confiança.

```
exp(coef(fit))
```

```
    pared    public       gpa
2.8510579 0.9429088 1.8513972
```

```
exp(cbind(RC = coef(fit), IC))
```

```
              RC     2.5 %   97.5 %
pared  2.8510579 1.6958376 4.817114
public 0.9429088 0.5208954 1.680579
gpa    1.8513972 1.1136247 3.098490
```

Esses coeficientes são chamados de razões de chances proporcionais e nós podemos interpreta-los da mesma maneira que interpretamos as razões de chances de uma regressão logística binária.

Interpretando as razões de chances

- Educação dos pais

1. Para estudantes cujos pais frequentaram a faculdade, as chances de serem mais propensos (ou seja, muito ou pouco provável versus improvável) a fazer pós graduação é 2,85 ($e^{1.04769}$) vezes as chances dos estudantes cujos pais não frequentaram a faculdade, mantendo constantes todas as outras variáveis.
2. Para os alunos cujos pais não frequentaram a faculdade, a chance de não se candidatar (ou seja, improvável versus um pouco ou muito provável) é de 2,85 vezes a dos alunos cujos pais foram à faculdade, mantendo constantes todas as outras variáveis.

- Tipo de faculadade

1. Para estudantes de escolas públicas, as chances de serem mais prováveis (ou seja, muito ou pouco provável versus improvável) de se candidatar são $5,71\%$ mais baixas [ou seja, $(1 - 0,943) * 100\%$] do que os alunos de escolas particulares, mantendo constantes todas as outras variáveis.

- GPA

1. Para cada aumento de uma unidade no GPA do aluno, as chances de ser mais provável de se candidatar (muito ou pouco provável versus improvável) são multiplicadas por 1,85 vezes (ou seja, aumentam 85%), mantendo constantes todas as outras variáveis.
2. Para cada diminuição de uma unidade no GPA do aluno, as chances de ser menos provável de se candidatar (improvável versus um pouco ou muito provável) são multiplicadas por 1,85 vezes, mantendo constantes todas as outras variáveis.

Probabilidades previstas

Podemos obter probabilidades previstas, geralmente mais fáceis de entender do que os coeficientes ou as razões de chances. Por exemplo, podemos variar o `gpa` para cada nível de `pared` e `public` e calcular a probabilidade de estar em cada uma das categorias de `apply`. Para isso, criamos um novo conjunto de dados de todos os valores a serem usados na previsão.

```
dados1 <- data.frame(
  pared = rep(0:1, 200),
  public = rep(0:1, each = 200),
  gpa = rep(seq(from = 1.9, to = 4, length.out = 100), 4))

dados1 <- cbind(dados1, predict(fit, dados1, type = "probs"))
head(dados1)
```

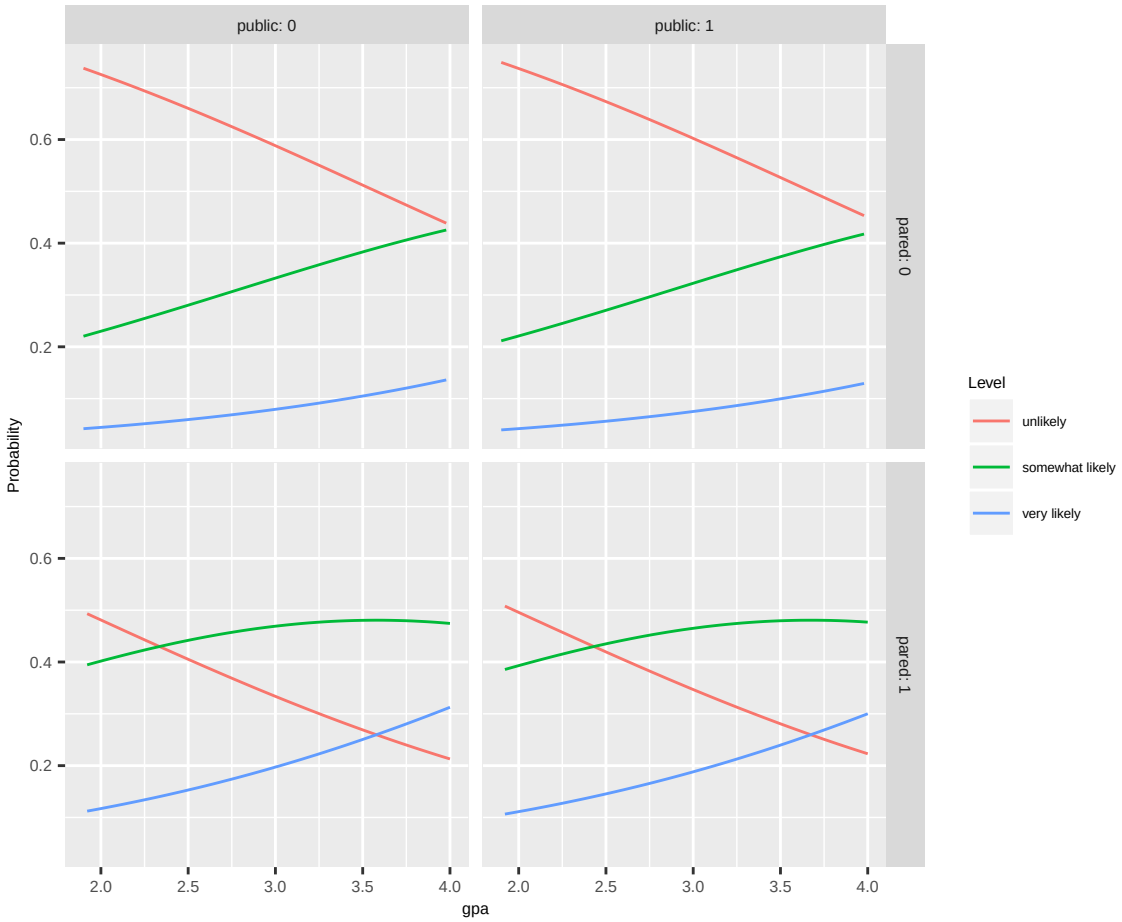
```
  pared public      gpa  unlikely somewhat likely very likely
1     0      0 1.900000 0.7376186       0.2204577  0.04192370
2     1      0 1.921212 0.4932185       0.3945673  0.11221424
3     0      0 1.942424 0.7325300       0.2244841  0.04298593
4     1      0 1.963636 0.4866885       0.3984676  0.11484395
5     0      0 1.984848 0.7273792       0.2285470  0.04407383
6     1      0 2.006061 0.4801630       0.4023098  0.11752712
```

Agora podemos remodelar os dadoscom o pacote 'reshape2 e graficar todas as probabilidades previstas para as diferentes condições.

```
dados2 <- melt(dados1, id.vars = c("pared", "public", "gpa"),
  variable.name = "Level", value.name="Probability")
head(dados2)
```

```
  pared public      gpa    Level Probability
1     0      0 1.900000 unlikely   0.7376186
2     1      0 1.921212 unlikely   0.4932185
3     0      0 1.942424 unlikely   0.7325300
4     1      0 1.963636 unlikely   0.4866885
5     0      0 1.984848 unlikely   0.7273792
6     1      0 2.006061 unlikely   0.4801630
```

```
ggplot(dados2, aes(x = gpa, y = Probability, colour = Level)) +
  geom_line() + facet_grid(pared ~ public, labeller="label_both") +
    theme(plot.title=element_text(size=7),plot.subtitle=element_text(size=6), axis.title.x=element_text(size=6),
        axis.title.y=element_text(size=6),axis.text.x=element_text(size=6), axis.text.y=element_text(size=6),
        text = element_text(size=6),strip.text = element_text(size=6))
```



Função `vlgm`

```
dados$apply <- factor(dados$apply, ordered=T)
fit2 <-  vglm(apply~pared+public+gpa,family=cumulative(parallel=TRUE),data=dados)
summary(fit2)
```

```
Call:
vglm(formula = apply ~ pared + public + gpa, family = cumulative(parallel = TRUE),
    data = dados)

Pearson residuals:
                      Min      1Q Median     3Q    Max
logitlink(P[Y<=1]) -1.671 -1.1309 0.6756 0.8247 1.8059
logitlink(P[Y<=2]) -4.057  0.1806 0.2040 0.4554 0.7532

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1  2.20335    0.78440   2.809  0.00497 **
(Intercept):2  4.29879    0.80915   5.313 1.08e-07 ***
pared         -1.04766    0.26845  -3.903 9.52e-05 ***
public         0.05867    0.28861   0.203  0.83891
gpa           -0.61575    0.26258  -2.345  0.01903 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 717.0249 on 795 degrees of freedom

Log-likelihood: -358.5124 on 795 degrees of freedom

Number of Fisher scoring iterations: 4

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
    pared    public       gpa
0.3507593 1.0604268 0.5402335
```

Referência

- Introduction to SAS. UCLA: Statistical Consulting Group. from https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/ (acessado em 22 de outubro de 2019).
- Paula, G.A. (2013). Modelos de Regressão com Apoio Computacional.
- Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley series in probability and statistics.