Lista 4

```
library(tidyverse)
library(ggplot2)
library(knitr)
library(kableExtra)
```

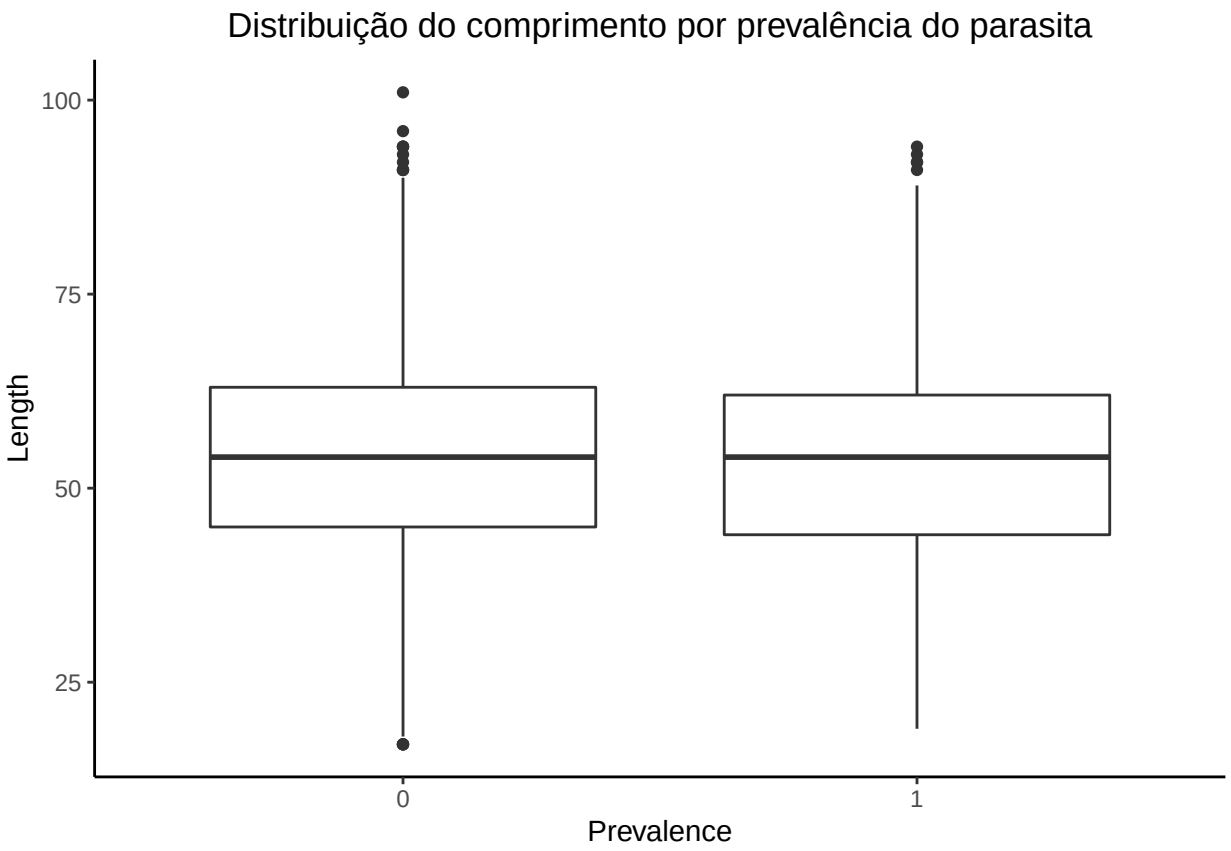
Banco de dados ParasiteCod

```
ParasiteCod <- read.csv('ParasiteCod.txt', sep = '\t')
ParasiteCod$fArea <- factor(ParasiteCod$Area)
ParasiteCod$fYear <- factor(ParasiteCod$Year)
ParasiteCod$Prevalence <- factor(ParasiteCod$Prevalence)
```

O modelo entregue para análise utiliza as variáveis *Length*, *Area* e *Year* para explicar a prevalência do parasita.

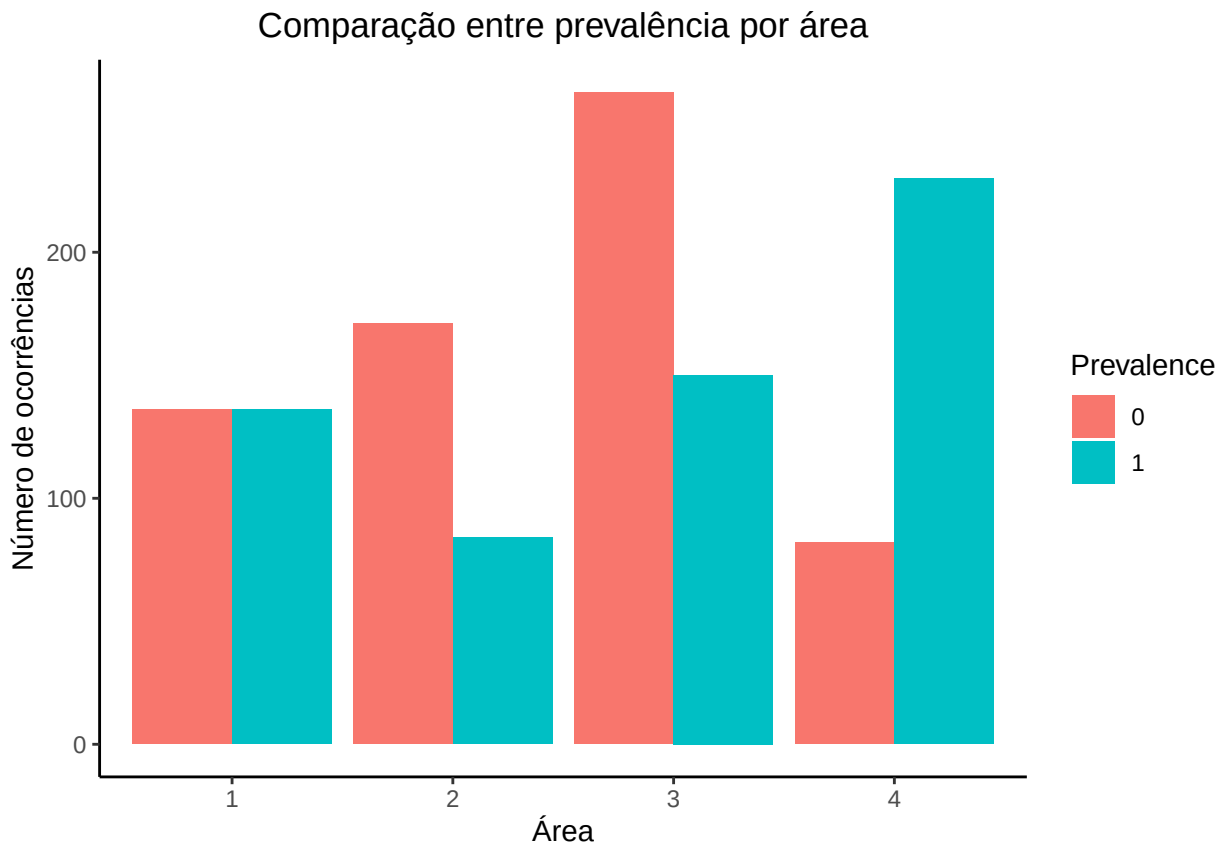
Através do gráfico abaixo vemos que a distribuição da variável *Length* é parecida em casos com e sem prevalência do parasita.

```
ggplot(ParasiteCod) +
  geom_boxplot(aes(x = Prevalence, y = Length)) +
  theme_classic() +
  ggtitle('Distribuição do comprimento por prevalência do parasita') +
  theme(plot.title = element_text(hjust = 0.5))
```



Outra variável presente no modelo é a área e, no gráfico abaixo vemos que para área tem uma proporção diferente de prevalência. Os valores de proporção estão na tabela seguinte.

```
ParasiteCod %>%
  group_by(Prevalence, fArea) %>%
  summarise(contagem = n()) %>%
  ggplot(aes(x = fArea, y = contagem, fill = Prevalence)) +
  geom_col(position = "dodge") +
  theme_classic() + ylab('Número de ocorrências') + xlab('Área') +
  ggtitle('Comparação entre prevalência por área') +
    theme(plot.title = element_text(hjust = 0.5))
```



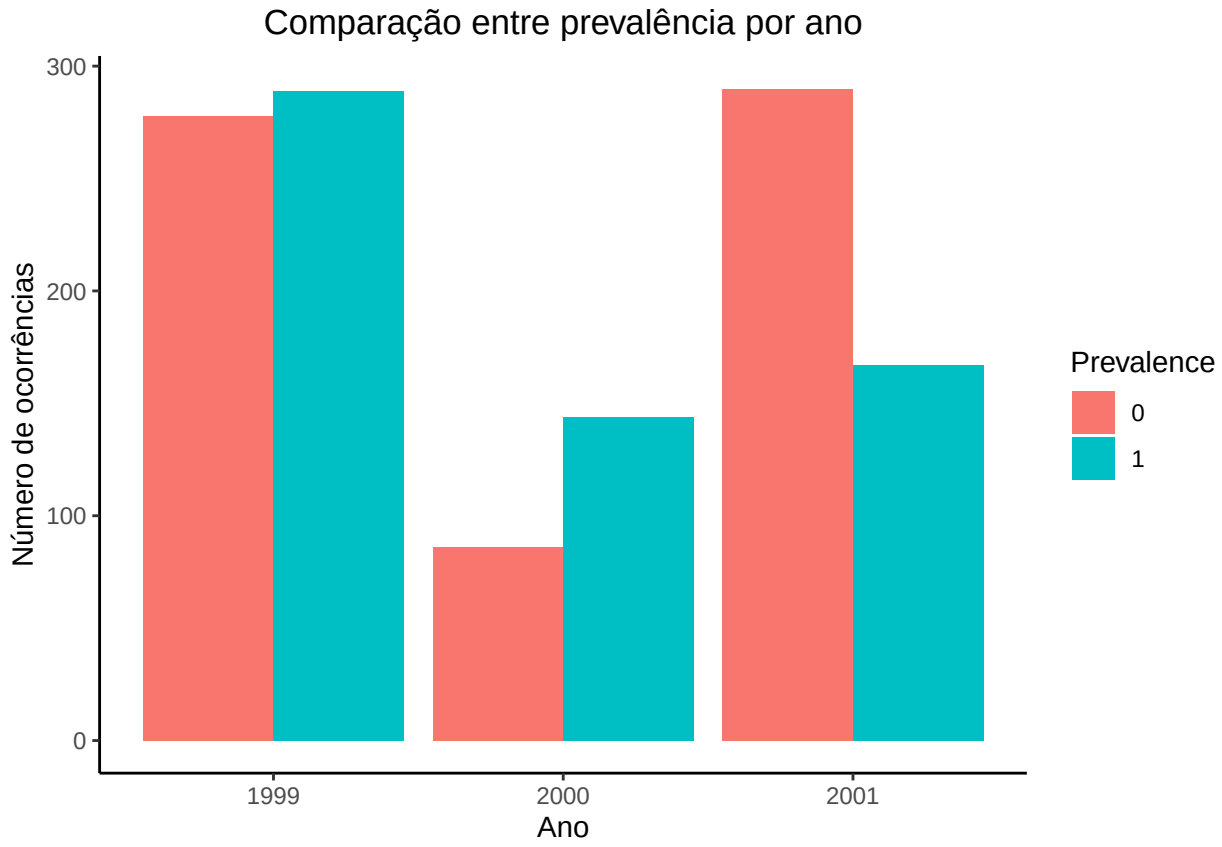

```
ParasiteCod %>%
  group_by(Prevalence, fArea) %>%
  summarise(contagem = n()) %>%
  pivot_wider(id_cols = 'fArea', names_from = 'Prevalence', values_from = 'contagem') %>%
  mutate(`Prevalence proportion` = (`1`/(`1`+`0`))) %>%
  kable(row.names = F,
        caption = "Prevaência do parasita por área",
        booktabs = TRUE, linesep = "") %>%
  add_header_above(c(" ", "Prevaência" = 2, " ")) %>%
  kable_styling(latex_options = "HOLD_position")
```

Tabela 1: Prevaência do parasita por área

	Prevaência		
fArea	0	1	Prevalence proportion
1	136	136	0.5000000
2	171	84	0.3294118
3	265	150	0.3614458
4	82	230	0.7371795

O ano da observação é outra variável descritiva do modelo e, novamente vemos no gráfico e na tabela abaixo que a proporção de prevalência varia de ano para ano.

```r
ParasiteCod %>%
  group_by(Prevalence, fYear) %>%
  summarise(contagem = n()) %>%
  ggplot(aes(x = fYear, y = contagem, fill = Prevalence)) +
  geom_col(position = "dodge") +
  theme_classic() + ylab('Número de ocorrências') + xlab('Ano') +
  ggtitle('Comparação entre prevalência por ano') +
    theme(plot.title = element_text(hjust = 0.5))
```




```r
ParasiteCod %>%
  group_by(Prevalence, fYear) %>%
  summarise(contagem = n()) %>%
  pivot_wider(id_cols = 'fYear', names_from = 'Prevalence', values_from = 'contagem') %>%
  mutate(`Prevalence proportion` = (`1`/(`1`+`0`))) %>%
  kable(row.names = F,
        caption = "Prevaência do parasita por Ano",
        booktabs = TRUE, linesep = "") %>%
  add_header_above(c(" ", "Prevaência" = 2, " ")) %>%
  kable_styling(latex_options = "HOLD_position")
```

Tabela 2: Prevaência do parasita por Ano

	Prevaência		
fYear	0	1	Prevalence proportion
1999	278	289	0.5097002
2000	86	144	0.6260870
2001	290	167	0.3654267

Através da tabela abaixo observamos que dentro de cada ano existe variação da proporção de prevalência por área, isso significa que é necessário considerar a interação entre as variáveis no modelo, assim como foi feito.

```
ParasiteCod %>%
  group_by(Prevalence, fArea, fYear) %>%
  summarise(contagem = n()) %>%
  pivot_wider(id_cols = c('fYear', 'fArea'),
              names_from = 'Prevalence', values_from = 'contagem') %>%
  mutate(`Prevalence proportion` = scales::percent(`1`/(`1`+`0`))) %>%
  pivot_wider(id_cols = 'fArea',
              names_from = 'fYear', values_from = 'Prevalence proportion') %>%
  kable(row.names = F,
        caption = "Prevaência do parasita por área e ano",
        booktabs = TRUE, linesep = "") %>%
  kable_styling(latex_options = "HOLD_position")
```

Tabela 3: Prevaência do parasita por área e ano

fArea	1999	2000	2001
1	61.2%	70.9%	10.0%
2	32.65%	36.00%	31.78%
3	33.9%	57.3%	28.7%
4	75.5%	88.0%	65.9%

Ajuste do modelo

O modelo de regressão logística pode ser definido como

$$Y_{ijk} \overset{ind.}{\sim} Bernoulli(\pi_{ijk})$$
$$\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \beta_0 + \beta_1 \text{Comprimento} + \beta_{2j}\text{Area}_j + \beta_{3k}\text{Ano}_k + \beta_{5jk}\text{Area}_j\text{Ano}_k$$

em que $i = 1, \ldots, n_{jk}$ representa o parasita i na área $j = 1, 2, 3, 4$ no ano $k = 1, 2, 3$. E π_{ijk} é a probabilidade do parasita i prevalecer na área j no ano k.

```
P12 <- glm(Prevalence ~  Length+fArea * fYear ,
           family = binomial, data = ParasiteCod)
summary(P12)
```

```
##
## Call:
## glm(formula = Prevalence ~ Length + fArea * fYear, family = binomial,
##     data = ParasiteCod)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0922  -0.9089  -0.4545   0.9678   2.2394
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.003226   0.291973   0.011  0.99118
## Length             0.008516   0.004585   1.858  0.06324 .
## fArea2            -1.185849   0.276897  -4.283 1.85e-05 ***
## fArea3            -1.136105   0.231248  -4.913 8.97e-07 ***
## fArea4             0.728736   0.261815   2.783  0.00538 **
## fYear2000          0.383756   0.343877   1.116  0.26444
## fYear2001         -2.655704   0.433542  -6.126 9.03e-10 ***
## fArea2:fYear2000 -0.209035   0.503494  -0.415  0.67802
## fArea3:fYear2000  0.561158   0.443733   1.265  0.20600
## fArea4:fYear2000  0.451582   0.588318   0.768  0.44274
## fArea2:fYear2001  2.595866   0.528472   4.912 9.01e-07 ***
## fArea3:fYear2001  2.403050   0.493512   4.869 1.12e-06 ***
## fArea4:fYear2001  2.115534   0.513489   4.120 3.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1727.8  on 1247  degrees of freedom
## Residual deviance: 1495.2  on 1235  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 1521.2
##
## Number of Fisher Scoring iterations: 4
```

Os fatores de referência são *fYear* = 1999 e *fArea* = 1. *Length* é uma variável contínua e sua média é 53.45. Usaremos a média do comprimento para o cálculo das probabilidades de cada ano e área.

Interpretação da chance de prevalência Para encontrar a probabilidade dos eventos precisamos calcular o inverso da função de ligação utilizando os coeficientes resultantes do treinamento do modelo.

Ex.: A probabilidade de prevalência do parasita quando o ano é **1999** e a área é **1** utilizando os coeficientes 0.0032259, e 0.0085164 é dada por

$$\hat{\pi}_{i11} = \frac{exp\{\beta_0 + \hat{\beta}_1 \text{Comprimento} + \hat{\beta}_{21}\text{Area}_1 + \hat{\beta}_{31}\text{Ano}_1 + \hat{\beta}_{511}\text{Area}_1\text{Ano}_1\}}{(1 + exp\{\beta_0 + \hat{\beta}_1 \text{Comprimento} + \hat{\beta}_{21}\text{Area}_1 + \hat{\beta}_{31}\text{Ano}_1 + \hat{\beta}_{511}\text{Area}_1\text{Ano}_1\})}$$

Então, a probabilidade de prevalência de parasita no ano de 1999 na área 1 com comprimento de 53.45 é 0.61.

Ex.: Para calcular as probabilidades das interações, todas as informações referentes às características da amostra para a qual queremos calcular a probabilidade são utilizadas. Considerando agora a probabilidade de prevalência na área **2** no ano **2000**, para isso precisaremos considerar os seguintes β_s:

- intercepto: 0.0032259
- efeito do comprimento: 0.0085164
- efeito da área 2: -1.1858494
- efeito do ano 2000: 0.3837563
- efeito da interação entre ano e área: -0.209035

Portanto, a probabilidade de prevalência do parasita, considerando comprimento médio, na área **2** no ano **2000** é 37%.

Ex.: Com o auxílio da função predict podemos calcular as probabilidades pra todas combinações de ano e área. A Tabela abaixo apresenta as probabilidades calculadas para as combinações de ano e área, e os resultados nos mostram os efeitos das interações. Por exemplo, a probabilidade em relação ao nível 3 de área e ano 1999 é menor que do nível 1 de área e ano 1999, isso devido ao efeito negativo do nível 3 da variável área. Mas a probabilidade em relação a área 3 e ano de 2001 é maior que a probabilidade da área 1 e o ano de 2001, isso devido ao fato do efeito da interação entre o ano de 2001 e o nível 3 da variável área. A conclusão é a mesma ao observarmos a área 2. Isso quer dizer que os os níveis da variável área tem um efeito diferente dependendo do ano, oque caracteriza uma interação entre área e ano.

Tabela 4: Probabilidade calculada para cada nível de área e ano com Length fixo em 53,45.

Length	fArea	fYear	Probabilidade
53.45	1	1999	0.61
53.45	2	1999	0.33
53.45	3	1999	0.34
53.45	4	1999	0.77
53.45	1	2000	0.70
53.45	2	2000	0.37
53.45	3	2000	0.57
53.45	4	2000	0.88
53.45	1	2001	0.10
53.45	2	2001	0.31
53.45	3	2001	0.28
53.45	4	2001	0.66

Na figura abaixo são apresentadas as probabilidades de prevalência por comprimento dado a área e o ano.

```
Xn = expand.grid(Length=with(ParasiteCod, seq(min(Length, na.rm = T),
                                              max(Length, na.rm = T),
                                              length.out = 10)),
                 fArea = levels(ParasiteCod$fArea),
                 fYear = levels(ParasiteCod$fYear))
Xn = cbind(Xn, prob = predict(P12, Xn, type="response"))
ggplot(Xn, aes(Length, prob, color=fArea, linetype=fYear)) +
  geom_line() +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Probabilidade predita") +
  theme_classic()
```

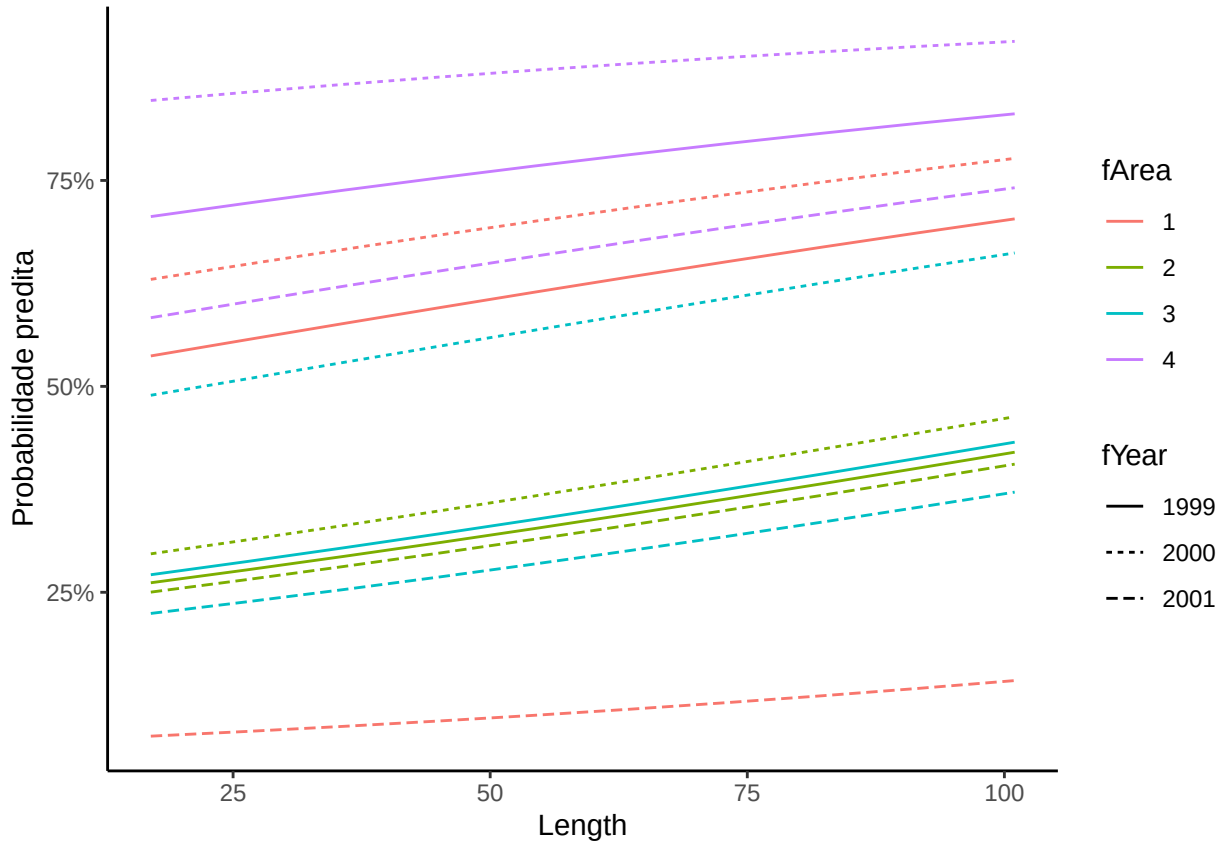


Figura 1: Probabilidade predita dado o comprimento por Área e Ano.

Vamos interpretar como as probabilidades estimadas se comportam nos niveis da variável ano para a área fixa

Área 1

Ano 1999

$$\pi_{i11} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto})}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto})}$$

Ano 2000

$$\pi_{i12} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} + 0.3837563)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} + 0.3837563)}$$

Ano 2001

$$\pi_{i13} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} - 2.6557044)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} - 2.6557044)}$$

Logo, $\pi_{i12} > \pi_{i11} > \pi_{i13}$. Área 2

Ano 1999

$$\pi_{i21} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1858494)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1858494)}$$

Ano 2000

$$\pi_{i22} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1858494 + 0.3837563 - 0.209035)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1858494 + 0.3837563 - 0.209035)}$$

Ano 2001

$$\pi_{i23} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1858494 - 2.6557044 + 2.595866)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1858494 - 2.6557044 + 2.595866)}$$

Logo, $\pi_{i22} > \pi_{i21} > \pi_{i23}$. Área 3

Ano 1999

$$\pi_{i31} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1361052)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1361052)}$$

Ano 2000

$$\pi_{i32} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1361052 + 0.3837563 + 0.5611579)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1361052 + 0.3837563 + 0.5611579)}$$

Ano 2001

$$\pi_{i33} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1361052 - 2.6557044 + 2.4030501)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} - 1.1361052 - 2.6557044 + 2.4030501)}$$

Logo, $\pi_{i32} > \pi_{i31} > \pi_{i33}$. Área 4

Ano 1999

$$\pi_{i41} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} + 0.7287359)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} + 0.7287359)}$$

Ano 2000

$$\pi_{i42} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} + 0.7287359 + 0.3837563 + 0.4515817)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} + 0.7287359 + 0.3837563 + 0.4515817)}$$

Ano 2001

$$\pi_{i43} = \frac{\exp(0.0032259 + 0.0085164\text{Comprimeto} + 0.7287359 - 2.6557044 + 2.1155344)}{1 + \exp(0.0032259 + 0.0085164\text{Comprimeto} + 0.7287359 - 2.6557044 + 2.1155344)}$$

Logo, $\pi_{i42} > \pi_{i41} > \pi_{i43}$.

Interpretação da razão de chance O modelo de regressão logística é caracterizado pela seguinte relação entre a probabilidade p e o preditor linear $X\beta$

$$\log\left(\frac{p_i}{1-p_i}\right) = X\beta,$$

em que X é uma matriz de dimensão $n \times p$ com as covariáveis e β o vetor de coeficientes de dimensão p. Podemos calcular os efeitos das covariáveis na razão de $p_i/(1-p_i)$, que chamamos de razão de chances. Logo,

- $\exp(0.08516) = 1.088$, a chance de prevalência do parasita aumenta em 8.8% com o aumento de uma unidade de comprimento.
- $\exp(-1.185849) = 0.3054$, a chance de prevalência do parasita na área 2 é 70% menor que na área 1 para o ano de 1999.
- $\exp(0.383756) = 1.46778$, a chance de prevalência do parasita no ano de 2000 aumenta em 46% em relação ao ano de 1999 para a área 1.
- $\exp(-0.209035) = 0.81136$, a chance de prevalência do parasita na área 2 é 19% menor do que na área 1 em relação ao ano de 2000, mas a chance de prevalência do parasita na área 2 é 1240% ($\exp(2.595866) = 13.4$) maior que na área 1 em relação ao ano de 2001. O que concluímos que claramente a área tem um efeito diferente na prevalência dependendo do ano selecionado.

Banco de dados hdp

```
library(lme4)
hdp <- read.csv("dados/hdp.txt")
hdp <- within(hdp, {
  Married <- factor(Married, levels = 0:1, labels = c("no", "yes"))
  DID <- factor(DID)
  HID <- factor(HID)
})
model <- glmer(remission ~ 1 + CancerStage + (1 + CancerStage | DID), data = hdp,
           family = 'binomial', control = glmerControl(optimizer = 'bobyqa'))
#BOBYQA:Bound Optimization by Quadratic Approximation
# summary(model)
```

Definição do modelo

O modelo ajustado para o banco de dados hdp é na forma

$$
\begin{aligned}
Y_{ijk}|\mathbf{b_i} &\overset{ind.}{\sim} Bernoulli(\pi_{ij}) \\
\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_0 + \beta_{1j} + b_{0i} + b_{1ji} \\
\mathbf{b_i} = \begin{bmatrix} b_{0i} \\ b_{12i} \\ b_{13i} \\ b_{14i} \end{bmatrix} &\sim N(\mathbf{0}, \mathbf{\Sigma}_i) \\
\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_0^2 & & & \\ \sigma_{02} & \sigma_2^2 & & \\ \sigma_{03} & \sigma_{23} & \sigma_3^2 & \\ \sigma_{04} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}&,
\end{aligned}
$$

em que $i = 1, \ldots, 407$ representa o médico, $j = 1, \ldots, 4$ representa o estágio do câncer, $k = 1, \ldots, n_{ij}$ é a repetição, e a matriz $\mathbf{\Sigma}_i$ é simétrica.

Interpretação dos efeitos fixos

Os coeficientes dos efeitos fixos estão apresentados na tabela abaixo.

Tabela 5: Estimativas dos efeitos fixos.

Termo	Estimativa
Intercepto	-0.8033
CancerStageII	-0.6042
CancerStageIII	-1.2709
CancerStageIV	-3.1219

O exponencial das estimativas da tabela acima são interpretadas como a razão de chances populacional, ou seja, A chance populacional de se ter remissão é

- 45,34% menor para o estágio II em relação ao I.
- 71,94% menor para o estágio III em relação ao I.
- 95,59% menor para o estágio IV em relação ao I.

As probabilidades populacionais de remissão para cada estágio estão apresentadas na tabela abaixo.

```
estagio_cancer = expand_grid(CancerStage = sort(unique(hdp$CancerStage)))
prob_rem = predict(model, estagio_cancer, re.form=NA, type="response")
cbind(estagio_cancer, prob_rem) %>%
  kable(digits=4, col.names = c("Estágio do câncer", "Probabilidade de Remissão"),
        caption = "Probabilidade de remissão por estãgio do câncer estimadas.",
        booktabs=TRUE, linesep="") %>%
  kable_styling(latex_options = "HOLD_position")
```

Tabela 6: Probabilidade de remissão por estãgio do câncer estimadas.

Estágio do câncer	Probabilidade de Remissão
I	0.3093
II	0.1966
III	0.1116
IV	0.0194

Pela razão de chances e pela tabela apresentada acima, nota-se que quanto mais avançado o estágio do câncer, menor a chance de remissão.

Exercício

Provar que y_{ij} e y_{ik} estão correlacionados.

Resposta:

Seja $y_i|\mathbf{b} \sim$ Bernoulli(π_i).

$E[y_i|\mathbf{b}] = \pi_\mathbf{i}$ e $g(\pi) = \mathbf{x_i}\boldsymbol{\beta} + \mathbf{z_i}\boldsymbol{b}$.

$$Cov(y_i, y_j) = Cov(E(y_i|\mathbf{b}), \mathbf{E(y_j}|\mathbf{b})) + \mathbf{E(Cov(y_i, y_j}|\mathbf{b})) = \mathbf{Cov}(\pi_\mathbf{i}, \pi_\mathbf{j}) + \mathbf{0}$$

$$Cov(\pi_i, \pi_j) = Cov\left(\frac{\exp(\mathbf{x_i}\boldsymbol{\beta} + \mathbf{z_i}\boldsymbol{b})}{1 + \exp(\mathbf{x_i}\boldsymbol{\beta} + \mathbf{z_i}\boldsymbol{b})}, \frac{\exp(\mathbf{x_j}\boldsymbol{\beta} + \mathbf{z_j}\boldsymbol{b})}{1 + \exp(\mathbf{x_j}\boldsymbol{\beta} + \mathbf{z_j}\boldsymbol{b})}\right)$$