Trabalho 2

Daniel Krügel

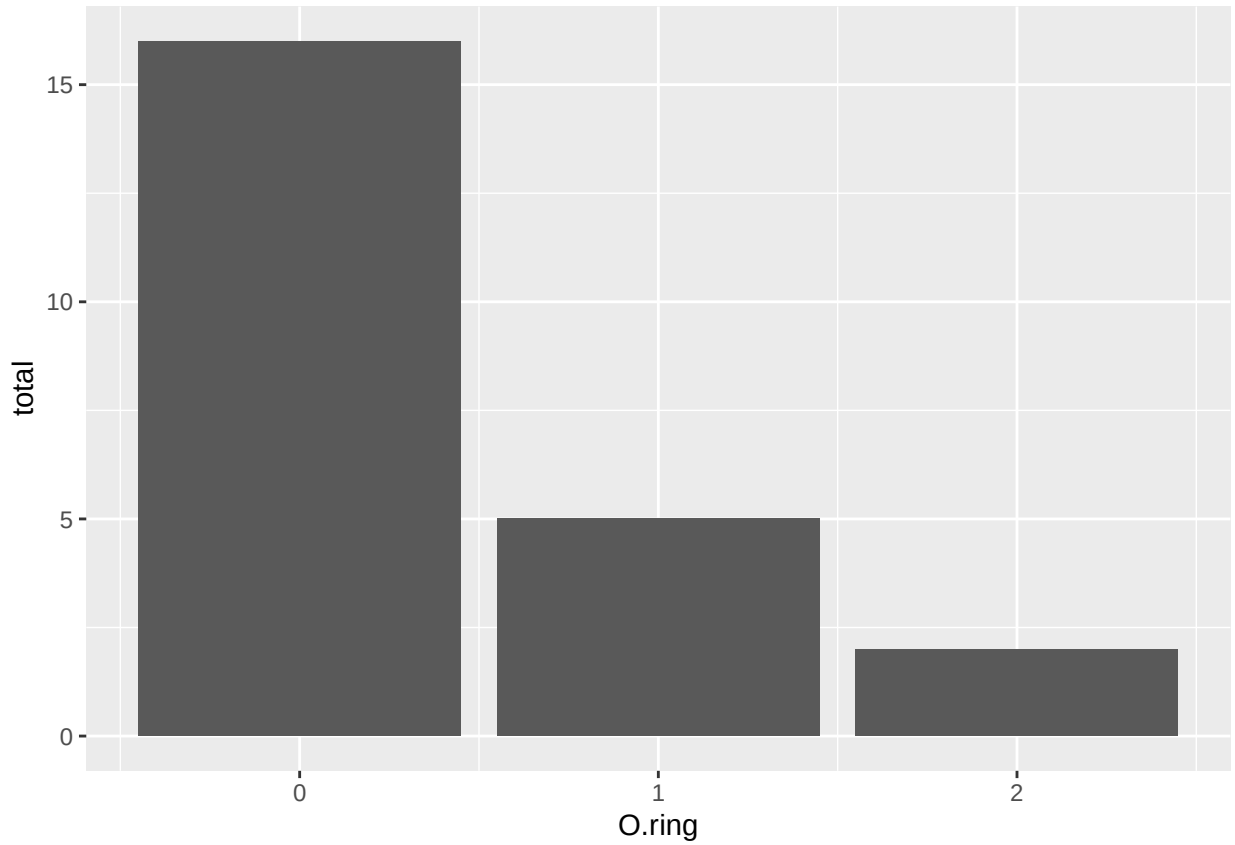
2022-12-11

Questão 4

#Análise descritiva da variável resposta

```
#Carregando dados
challenger <-  read.csv(file = "http://leg.ufpr.br/~lucambio/CE073/20222S/challenger.csv")

#Fazendo um barplot para verificar a distribuição da variável resposta
challenger %>%
  group_by(O.ring) %>%
  summarise(total = n()) %>%
  ggplot(aes(x = O.ring, y = total)) +
  geom_bar(stat = "identity")
```



```r
summary(challenger$O.ring) # Aquele summary para verificar se não há nenhum dado faltante
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3913  1.0000  2.0000
```

Como O.ring é um dado de contagem imaginei que ele poderia fazer parte da família Poisson de distribuições generalizadas, observando o gráfico me pareceu a melhor escolha realmente.

#Regressão

Então ajustando o glm usando a função de ligação canônica, temos:

```r
ajuste <- glm(O.ring ~ Temp + Pressure,
              family = poisson(link = "log"),
              data = challenger)

coef(ajuste)
```

```
##  (Intercept)         Temp     Pressure
##  3.533779532 -0.086891983  0.007957437
```

Então a forma linear da regressão logistica fica: eˆ(3.533-0.0868 * Temp+0.00795 * Pressure)

Mas será que as variáveis explicativas são significativas? Para responder vamos fazer uma tabela de variancia usando a função Anova do pacote cars, para que a entrada das covariáveis não altere a significancia delas.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring
##          LR Chisq Df Pr(>Chisq)
## Temp       4.6405  1    0.03123 *
## Pressure   1.4041  1    0.23604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pressure não foi significante com um -2log(lambda) = 1.4041 e com um p-valor de P(A > 1.4041) = 0.236. Enquanto Temp foi significativa porém a hipótese nula de que B_change = 0 não seria rejeitado caso meu nível de significância alpha fosse igual a 0.01.

#A variável Pressure pode ser retirada?

Para responder esta pergunta vamos fazer algumas análises, a primeira vamos ver se há alguma correlação entre elas ou se as variáveis são ortogonais:

```r
cor(challenger$Temp, challenger$Pressure)
```

```
## [1] 0.03981769
```

```r
cor.test(challenger$Temp, challenger$Pressure,
         alternative = c("two.sided"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  challenger$Temp and challenger$Pressure
## t = 0.18261, df = 21, p-value = 0.8569
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3785984  0.4447207
## sample estimates:
##        cor
## 0.03981769
```

O teste pearson para correlação entre os dados deu que a correlação entre os dados é de aproximadamente 0.04 com um p valor de 0.8569, me não me fornecendo indicativos o suficiente para rejeitar a hipótese nula de que a cor entre Temp e Pressure é 0. Portanto não devemos ver resultados diferentes caso optemos por comparar modelos em que a entrada das covariáveis seja diferente ou na realização de um teste de análise das deviances de forma sequencial.

```
anova(ajuste, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: O.ring
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         22     22.434
## Temp      1   5.6004        21     16.834  0.01796 *
## Pressure  1   1.4041        20     15.430  0.23604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ajuste.transposto <- glm(O.ring ~ Pressure + Temp,
                family = poisson(link = "log"),
                data = challenger)

anova(ajuste, ajuste.transposto) # A mudança entre a ordem de entrada não foi significativa
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring ~ Temp + Pressure
## Model 2: O.ring ~ Pressure + Temp
##   Resid. Df Resid. Dev Df   Deviance
## 1        20      15.43
## 2        20      15.43  0 1.7764e-15
```

```
ajuste.real  <- glm(O.ring ~ Temp,
                     family = poisson(link = "log"),
```

```
                      data = challenger)

anova(ajuste.real, ajuste, test = "Chisq") # A remoção do parâmetro não foi significativa a qualidade d
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring ~ Temp
## Model 2: O.ring ~ Temp + Pressure
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     16.834
## 2        20     15.430  1   1.4041    0.236
```

Sendo que o pvalor de $P(A > 1.4041) = 0.236$ a adição da variável não é estatísticamente significativa a nenhum valor abaico de alpha $= 0.05$ Portanto a escolha de permanecer com o modelo mais simples faz sentido. Claro que a base de dados tem apenas 23 observações, então caso o experimento seja refeito, com uma amostra maior possamos indentificar efeitos mais extremos que podem ter passado despercebidos e desenvolver um modelo mais preciso, porém com base neste experimento este é o modelo minimal para estre problema.

##Questão 7

```
# Carregando dados
tb <-  read.csv ( file = "http://leg.ufpr.br/~lucambio/CE073/20222S/placekick.BW.csv")

# Transformando eles em zero e um
for(i in 1:length(tb$Good)){
  if(tb$Good[i] == "Y"){
    tb$Good[i] <- 1
  }else{
    tb$Good[i] <-0
  }
}
 tb$Good <- as.numeric(tb$Good)

 #Realizando o ajuste de modelo
 ajuste7 <- glm(Good ~ Distance,
                family = binomial(link = 'logit'),
                data = tb)
summary(ajuste7)
```

```
##
## Call:
## glm(formula = Good ~ Distance, family = binomial(link = "logit"),
##     data = tb)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6166   0.2785   0.4538   0.7390   1.5099
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.409295   0.294819   18.35   <2e-16 ***
## Distance    -0.106270   0.007026  -15.12   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2104  on 2002  degrees of freedom
## Residual deviance: 1817  on 2001  degrees of freedom
## AIC: 1821
##
## Number of Fisher Scoring iterations: 5
```

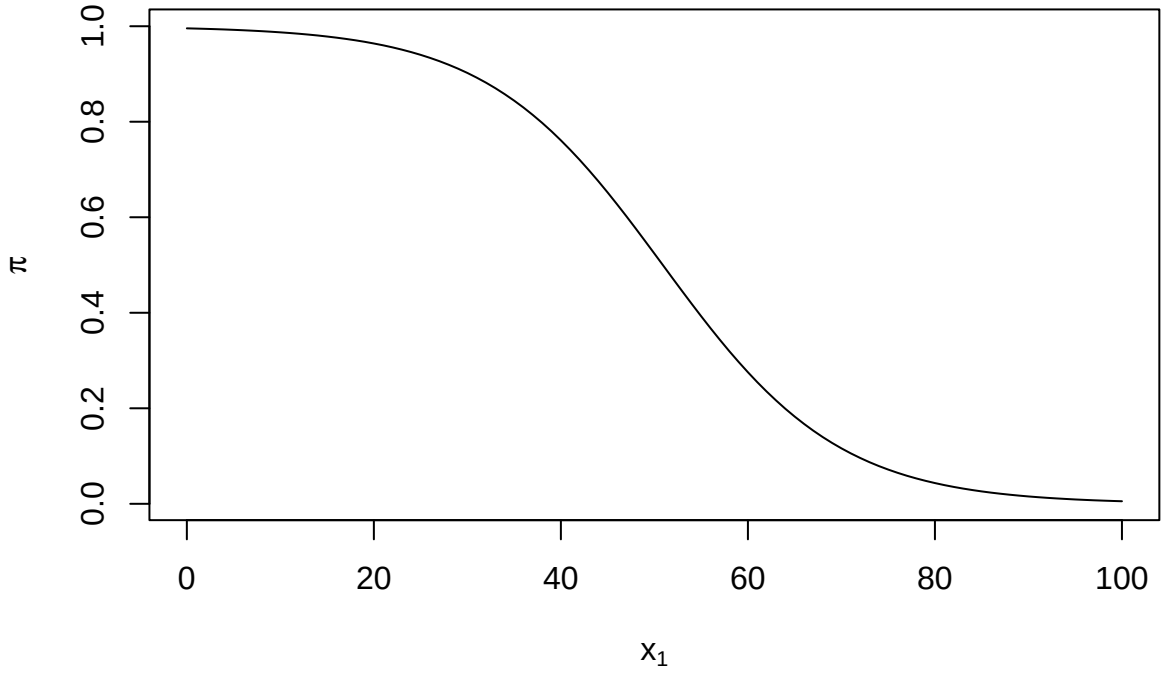
Neste caso ajustando a fórmula Good ~ Distance utilizando a família binomial (Já que a variável resposta é dicotômica) usando a função de ligação logito podemos verificar se o R está modelando a probabilidade de sucesso ou de falha ao olhar na estimativa de Beta1, caso ela seja negativa estamos olhando para uma regressão modelando a probabilidade de FALHA, caso a estimativa seja positiva uma probabilidade de SUCESSO.

Isto se dá pelo formato da curva sigmoidal gerada através da função de ligação, quando a estimativa Beta1 é negativa a curva é decrescente dando a entender que as probabilidades começam altas onde há um peso maior.

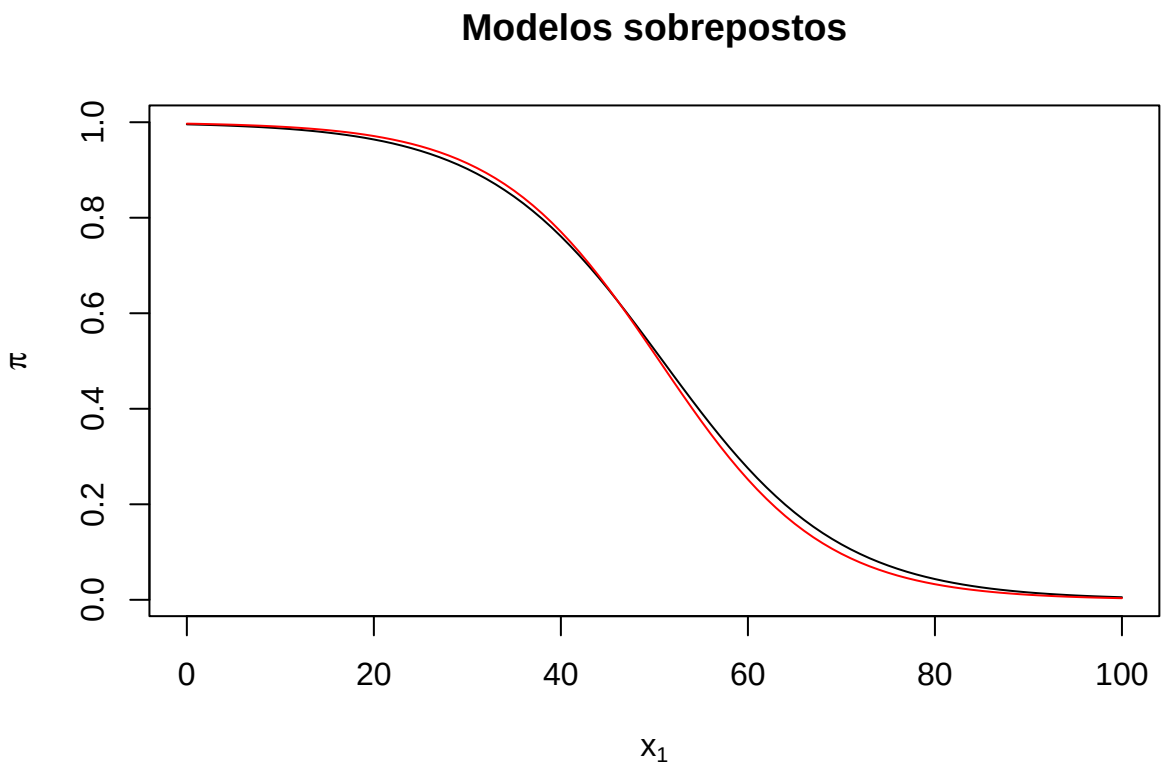
#Curva sigmoide

```
beta0 <- coef(ajuste7)[[1]]
beta1 <- coef(ajuste7)[[2]]
curve ( expr = exp( beta0 + beta1 *x) / (1+ exp( beta0 + beta1 *x)), xlim = c(0, 100) ,
        col = " black ", main = expression (pi == frac (e ^{5.4092-0.1062* x[1]} , 1+e ^{5.4092-0.1062*
        xlab = expression (x [1]) , ylab = expression (pi))
```

$$\pi = \frac{e^{5.4092 - 0.1062 x_1}}{1 + e^{5.4092 - 0.1062 x_1}}$$




```
curve ( expr = exp( beta0 + beta1 *x) / (1+ exp( beta0 + beta1 *x)), xlim = c(0, 100) ,
        col = " black ", main =  "Modelos sobrepostos",
        xlab = expression (x [1]) , ylab = expression (pi))
curve ( expr = exp( 5.8121 - 0.1150 *x) / (1+ exp( 5.8121 - 0.1150 *x)), xlim = c(0, 100), col = 'red',
```



Como ambos os estudos são parecidos analisando uma variável em função de outra e ambas idealmente seguem as mesmas distribuições de probabilidade na prática é como se os dados fornecidos no conjunto placekick.BW fossem uma repetição de um experimento. (Acredito que seja algo por este pensamento, meu conhecimento de futebol é bem fraco, não entendi a maior parte das variáveis).

Questão 19

#Carregando os dados e reordenando os níveis

```
#Carregando dados
data  <-  read.csv(file = "http://leg.ufpr.br/~lucambio/ADC/healthcare_worker.csv")

#Reordenando os níveis dos grupos de ocupação para que a parcela de menor contato ser o grupo controle
class(data$Occup.group)
```

```
## [1] "character"
```

```
data$Occup.group <- as.factor(data$Occup.group)
data$Occup.group <- relevel(data$Occup.group, ref = "No patient contact")
levels(data$Occup.group)
```

```
## [1] "No patient contact" "Exposure prone"     "Fluid contact"
## [4] "Lab staff"          "Patient contact"
```

Ajuste

```
#Ajustando o GLM da família Binomial com a função de ligação LOGITO
ajuste19 <- glm(Hepatitis/Size ~ Occup.group,
                family = binomial(link = 'logit'),
                weights = Size,
                data = data)

#Análise das deviances
car::Anova(ajuste19)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Hepatitis/Size
##             LR Chisq Df Pr(>Chisq)
## Occup.group    3.735  4     0.4431
```

Nenhum dos grupos foi significativo a um alpha de 0.05, vamos partir para uma análise das razões de chances para ver se realmente essa análise bate com o resultado.

#Odds Ratio

```
as.data.frame(exp(1*ajuste19$coefficients[1:5]))
```

```
##                            exp(1 * ajuste19$coefficients[1:5])
## (Intercept)                                         0.006410256
## Occup.groupExposure prone                           0.354545455
## Occup.groupFluid contact                            0.428432956
## Occup.groupLab staff                                0.883018868
## Occup.groupPatient contact                          0.252427184
```

Nenhuma das Odds Ratio apresenta um grande aumento, o mais destacado é o grupo responsável pelo manejamento de líquidos em laboratório.

#Por que?

Consultando o aluno Adam Domingues de Enfermagem da PUC e a aluna Eduarda Santini de Farmácia da UFPR e alguns dos motivos para que os grupos onde se foi estudado a contaminação de Hepatite C não serem significativos para o aumento da contaminação é de que a doença não tem vacina, ao contrário da variaçãoo B da doença que e obrigatória para agentes de saúde segundo o programa nacional de saúde (PNI), portanto há um cuidado muito grande para a prevenção com materiais de proteção. A exceção seria justamente nos laboratórios,que apresentam o maior risco de descuido e o maior número de ponto de falhas, o que explica a sua odds ratio ser levemente maior do que a dos outros grupos.