Lista 2 - MLG

Mariana Rodrigues Fontenelle

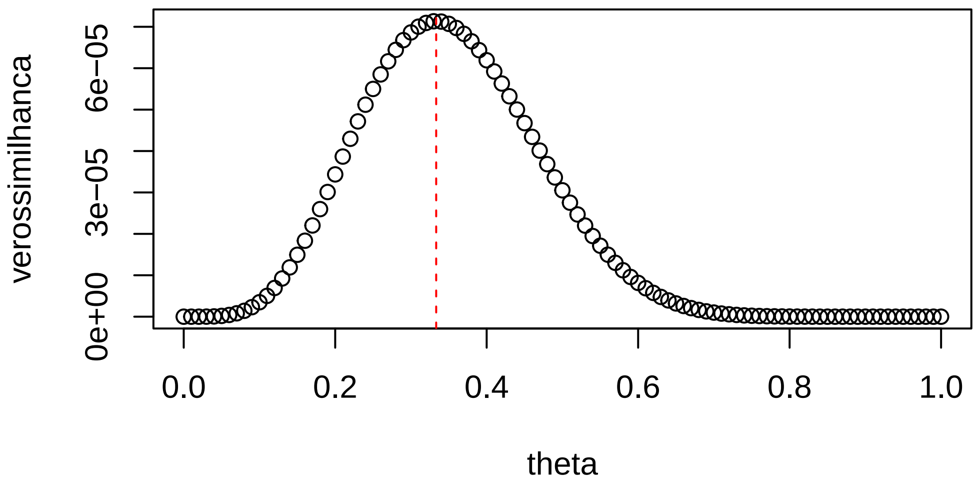
1

Para um indivíduo i, considere a variavel aleatória Yi Bernoulli(θ), tal que 0 θ 1 e a probabilidade do evento sucesso Yi = 1. Um estudo foi desenvolvido com este tipo de variável aleatória sendo medida independentemente para cada indivíduo participante. A seguinte amostra aleatoria foi observada: y1 = 1, y2 = 0, y3 = 0, y4 = 1, y5 = 0, y6 = 0, y7 = 0, y8 = 1, y9 = 0, y10 = 0, y11 = 0, y12 = 1, y13 = 0, y14 = 0, y15 = 1. Responda os itens a seguir:

(b) Use o R para fazer o grafico da função de verossimilhança escrita na letra (a). Você deve plotar "valores de θ versus função de verossimilhança (θ deve estar no eixo horizontal do gráfico). Considere os valores de θ definidos pelo comando theta = seq(0,1,0.01). Qual valor de θ esta associado ao ponto de máximo da curva obtida no gráfico?

$$f(y_1, ...y_n; \theta_i) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} = \theta^5(1-\theta)^{10}$$

```
theta <- seq(0, 1, 0.01)
yi <- c(1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1)
n <- length(yi)
verossimilhanca <- theta^sum(yi) * (1 - theta)^(n - sum(yi))
plot(x = theta, y = verossimilhanca)
abline(v = mean(yi), col = 'red', lty = 2)
```



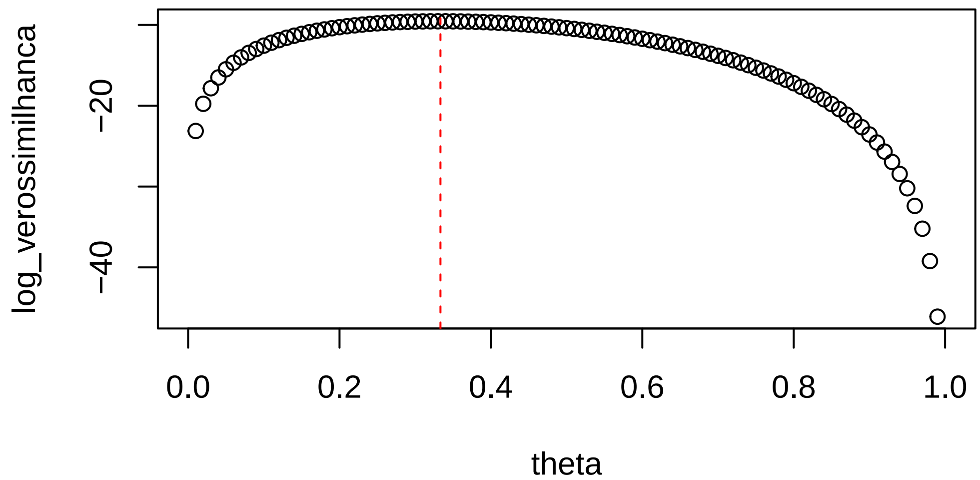
Estimador de Máxima Verossimilhança:

```
mean(yi)
```

```
[1] 0.3333333
```

(c) Use o R para fazer o grafico da função de log-verossimilhança escrita na letra (a). Você deve plotar "valores de θ versus função de log-verossimilhança (θ deve estar no eixo horizontal do gráfico). Considere os valores de θ definidos pelo comando theta = seq(0,1,0.01). Qual valor de θ esta associado ao ponto de máximo da curva obtida no gráfico?

```
log_verossimilhanca <- log(verossimilhanca)
plot(x = theta, y = log_verossimilhanca)
abline(v = 0.333333, col = 'red', lty = 2)
```




```
log_ver_theta <- data.frame(log_verossimilhanca, theta)
log_ver_theta %>%
  filter(log_verossimilhanca == max(log_verossimilhanca))
```

```
  log_verossimilhanca theta
1           -9.548089  0.33
```

(d) Aplique a função optim do R (conforme ensinado nas aulas) para obter a estimativa de maxima verossimilhança de θ. Voce deverá maximizar a função de log-verossimilhança e considerar a amostra apresentada no enunciado desta questão (use o chute inicial $\theta^{(0)} = 0.5$). Dentro da função optim, selecione o metodo L-BFGS-B para realizar a otimização numérica (especifique limite inferior = 0.0001 e superior = 0.9999 para o L-BFGS-B). Mostre seu script e indique claramente a estimativa final de θ

```
loglik_bernoulli = function(chute,y)
{
  n = length(y);
  out = sum(y) * log(chute) + (n - sum(y)) * log(1 - chute)
  return(out)
}
```

```
chute_inicial <- 0.5
ajuste <- optim(par = chute_inicial, fn = loglik_bernoulli,  y = yi,
                method = "L-BFGS-B", lower = 0.0001,  upper = 0.9999, control = list(fnsca
                hessian = FALSE)
ajuste$par
```

```
[1] 0.3333337
```

2

Seja Y1, Y2,..., Yn uma amostra aleatoria tal que $Yi \sim Poisson(\theta_i)$ com $\theta_i > 0$. Desejamos analisar o impacto do regressor X1i sobre a resposta Yi. Considere o preditor linear $\eta_i = \beta_0 + \beta_1 X_{1i}$. Denotando $\beta = (\beta_0, \beta_1)^T$ e $X_i = (1, X_{1i})^T$, podemos tambem escrever $\eta_i = X_i^T \beta$. No caso Poisson, lembre que a função de ligação é estabelecida por $\theta_i = e^{\eta_i}$. A seguinte amostra deverá ser utilizada para resolver esta questão:

```
Yi <- c(7, 2, 24, 33, 2, 5, 2, 5, 8, 8, 16, 20, 11, 4, 37, 1, 8, 8, 5, 7, 5, 4, 4, 2, 26)
X1i <- c(-0.27, 0.57, -0.74, -0.94, 0.64, 0.86, 0.5, 0.12, -0.36, 0.12, -0.55, -0.81, -0.2
df <- data.frame(Yi, X1i)
```

(a) Estabelecça o chute inicial $\hat{\beta}^{(r=0)} = (1, 1)^T$ a ser utilizado no IWLS. Usando este chute e o resultado amostral fornecido no enunciado, construa no R (exibir o script) a matriz $W^{(0)}$ correspondente a este problema. Mostre a submatriz 5×5 formada pelas linhas 1 a 5 e colunas 1 a 5 da matriz $W^{(0)}$.

```
beta_chute_inicial <- c(1,1)
x1 <- rep(1, 25)
x <- cbind(x1, df$X1i)
eta <- x %*% beta_chute_inicial
theta <- exp(eta)
var <- theta
der_mi_eta <- exp(eta)
W <- matrix(0, 25, 25)
w <- as.matrix(1/var * (der_mi_eta)^2)
for(i in 1:25){
      W[i,i] = w[i]
}
W[1:5,1:5]
```

```
         [,1]     [,2]     [,3]     [,4]     [,5]
```

```
[1,] 2.075081 0.000000 0.00000 0.000000 0.00000
[2,] 0.000000 4.806648 0.00000 0.000000 0.00000
[3,] 0.000000 0.000000 1.29693 0.000000 0.00000
[4,] 0.000000 0.000000 0.00000 1.061837 0.00000
[5,] 0.000000 0.000000 0.00000 0.000000 5.15517
```

(b) Ainda levando em conta o chute inicial $\widehat{\beta}^{(r=0)} = (1,1)^T$, calcule o vetor $z^{(0)}$ definido no IWLS deste problema relacionado a Poisson. Mostre os valores do seu vetor $z^{(0)}$

```r
der_eta_mi <- 1/theta
y <- df$Yi
z <- eta + (y - theta) * der_eta_mi
z
```

```
           [,1]
 [1,]  3.1033629
 [2,]  0.9860904
 [3,] 17.7652381
 [4,] 30.1382296
 [5,]  1.0279601
 [6,]  1.6383632
 [7,]  0.9462603
 [8,]  1.7513990
 [9,]  3.8583394
[10,]  2.7302384
[11,]  9.6520504
[12,] 15.7291827
[13,]  4.7822927
[14,]  1.3961427
[15,] 36.0000000
[16,]  0.6465970
[17,]  3.0864297
[18,]  2.6830942
[19,]  1.9647577
[20,]  2.2866112
[21,]  1.6642921
[22,]  1.3882812
[23,]  1.5025113
[24,]  1.2489137
[25,] 18.7448528
```

(c) Utilizando $W^{(0)}$ e $z^{(0)}$obtidos em (a) e (b), respectivamente, calcule a estimativa $\beta^{(1)} =$

$(X^T W^{(0)} X)^{-1} X^T W^{(0)} z^{(0)}$ referente a primeira iteração do IWLS. Mostre o script R e o resultado final de $\beta^{(1)}$

```
beta_r_1 <- solve(t(x) %*% W %*% x) %*% t(x) %*% W %*% z
beta_r_1
```

```
        [,1]
x1  6.153008
   -8.109299
```

(d) Use os dados amostrais fornecidos no enunciado e aplique a funcção glm do R para estimar β_0 e β_1. Use o mesmo chute inicial sugerido em (a). Mostre a saída do comando summary e avalie a significancia dos coeficientes. Atenção! sua análise sobre a significância deve indicar claramente a estimativa do coeficiente e o valor-p sob avaliação. Note também que a questão NÃO está pedindo para você realizar vários ajustes para encontrar o melhor modelo. O objetivo aqui e apenas ver se você sabe interpretar a saída computacional solicitada.

```
dados <- data.frame(Yi = df$Yi, x1, x1i = df$X1i)
ajuste_glm <- glm(Yi ~ x1i , data = dados, family = "poisson", start = beta_chute_inicial)
summary(ajuste_glm)
```

```
Call:
glm(formula = Yi ~ x1i, family = "poisson", data = dados, start = beta_chute_inicial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9798     0.0841   23.54   <2e-16 ***
x1i          -1.5115     0.1240  -12.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 199.783  on 24  degrees of freedom
Residual deviance:  25.098  on 23  degrees of freedom
AIC: 123.27

Number of Fisher Scoring iterations: 15
```

```
ajuste_glm$coefficients
```

```
(Intercept)         x1i
   1.979824   -1.511514
```

Como o valor-p do intercepto e de $\beta_0 < 0.05$ podemos considerá-las significativas para o modelo.

(e) A saída computacional investigada em (d) mostra a deviance do modelo ajustado. Calcule esta deviance através do R. Depois disso, avalie se essa deviance é pequena, moderada ou grande. Atenção! você deve mostrar o passo a passo da conta (use a expressão da deviance para o caso Poisson explicada nas aulas). Não aplique uma função pronta do R que calcula a deviance automaticamente.

$$\hat{\mu_i} = g(X_i^T \hat{\beta}) = e^{X_i^T \hat{\beta}}$$

```
mi_hat <- exp(x %*% ajuste_glm$coefficients)
D <- 2*sum(df$Yi*log(df$Yi/mi_hat) - df$Yi + mi_hat)
D
```

```
[1] 25.09765
```

```
n <- nrow(x)
p <- ncol(x)
qchisq(0.95, n-p)
```

```
[1] 35.17246
```

Como a deviance é menor porém próxima do qui quadrado, assumimos que ela é moderada.

(f) Calcule (use o R) a estatística de Pearson generalizada para o ajuste em (d). Em seguida compare o resultado com algum limiar adequado que deve ser obtido a partir da distribuição qui-quadrado. Atenção! mostre o passo-a-passo da conta e nao use função pronta do R que calcula a estatística.

```
pearson <- sum((df$Yi - ajuste_glm$fitted.values)^2 / ajuste_glm$fitted.values)
pearson
```

```
[1] 26.68479
```

```r
qchisq(0.95, n - p)
```

```
[1] 35.17246
```

3

Um experimento de laboratório foi desenvolvido para estudar um tipo de larva que ataca lavouras de café. O estudo considerou 52 recipientes, os quais receberam (cada um) 10 larvas. As larvas em cada recipiente foram submetidas (ao mesmo tempo) a uma dose de um produto químico A (variável X1) e uma dose de um produto químico B (variável X2). Os dados estao disponíveis no arquivo dados_Q3_L2_MLG.txt disponibilizado para esta lista de exercícios. A variável resposta do estudo é o número de larvas que morreram no recipiente após 1 hora da aplicação dos dois produtos químicos. Carregue os dados no R com a função read.table. A dataframe obtida tera a variável resposta na coluna 1, X1 na coluna 2 e X2 na coluna 3. Responda os itens a seguir (você poderá usar o R para responder).

```r
dados <- read.table("dados_Q3_L2_MLG.txt", col.names = c("Yi", "X1", "X2"))
n <- 10
```

(a) Aplique a função glm do R para estimar os coeficientes $\beta_0, \beta_1, \beta_2$do modelo MLG apropriado a estes dados (use obrigatoriamente o link canonico). Mostre a saída do comando summary e avalie a significância dos coeficientes. Atenção! sua análise sobre a significância deve indicar claramente o valor estimado do coeficiente e o valor-p sob avaliação. Note também que a questão NÃO está pedindo para você realizar vários ajustes para encontrar o melhor modelo. O objetivo aqui é apenas ver se você sabe qual é o MLG (Normal, Binomial, Bernoulli, Poisson, Gama, etc.) apropriado e como interpretar o resultado.

```r
modelo <- glm(cbind(Yi,n- Yi) ~ X1 + X2, data = dados, family = "binomial")
summary(modelo)
```

```
Call:
glm(formula = cbind(Yi, n - Yi) ~ X1 + X2, family = "binomial", 
    data = dados)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.0911     0.2191  -4.980 6.37e-07 ***
X1            1.5420     0.3487   4.422 9.77e-06 ***
X2            0.1722     0.2990   0.576    0.565    
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 72.906  on 51  degrees of freedom
Residual deviance: 50.796  on 49  degrees of freedom
AIC: 193.97

Number of Fisher Scoring iterations: 4
```

Foi usada o MLG Poisson por se tratar de uma contagem do número de larvas que morreram no período. Apenas β_0, β_1 apresentaram significância para $\alpha = 0.05$.

(b) Interprete o impacto da covariável X1 quando seu valor e aumentado em 0.1 (zero ponto um). Atenção! explique claramente qual elemento do modelo sofrera influência direta deste aumento em X1.

A variável resposta Yi sofrerá influência direta do aumento de X1.

Desconsiderando x2 por ser não significativo:

$$
\begin{aligned}
&log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1 x_1 \\
&log\left(\frac{\theta(x_1+0.1)}{1-\theta(x_1+0.1)}\right) = \beta_0 + \beta_1(x_1+0.1) \\
&log\left(\frac{\theta(x_1+0.1)}{1-\theta(x_1+0.1)}\right) = \beta_0 + \beta_1 x_1 + 0.1\beta_1 \\
&log\left(\frac{\theta(x_1+0.1)}{1-\theta(x_1+0.1)}\right) = log\left(\frac{\theta}{1-\theta}\right) + 0,1\beta_1 \\
&log\left(\frac{\theta(x_1+0.1)}{1-\theta(x_1+0.1)}\right) - log\left(\frac{\theta}{1-\theta}\right) = 0,1\beta_1 \\
&log\left(\frac{\frac{\theta(x_1+0.1)}{1-\theta(x_1+0.1)}}{\frac{\theta}{1-\theta}}\right) = 0,1\beta_1 \\
&\frac{\frac{\theta(x_1+0.1)}{1-\theta(x_1+0.1)}}{\frac{\theta}{1-\theta}} = e^{0,1\beta_1}
\end{aligned}
$$

$e^{0,1\beta_1}$ é o fator multiplicativo para Y ao incrementarmos x1 em 0,1 unidade.

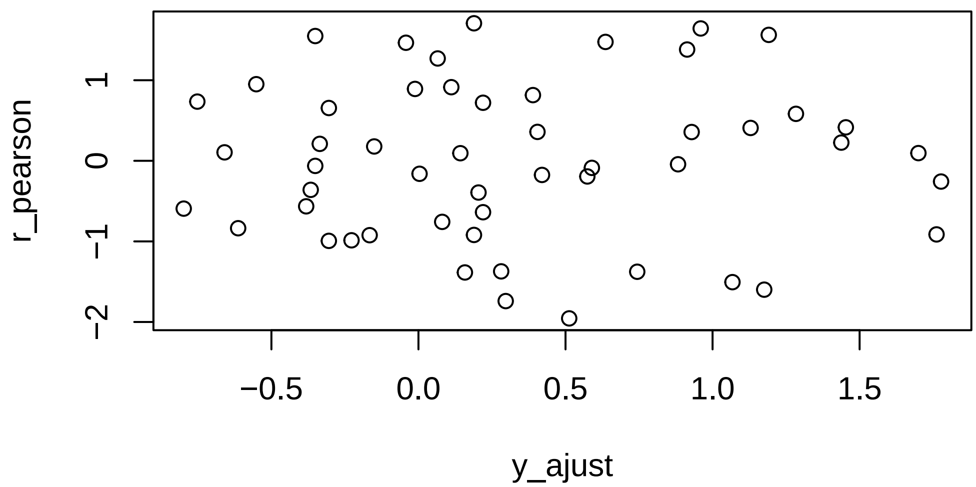
```r
exp(0.1 * modelo$coefficients[2])
```

```
      X1
1.166726
```

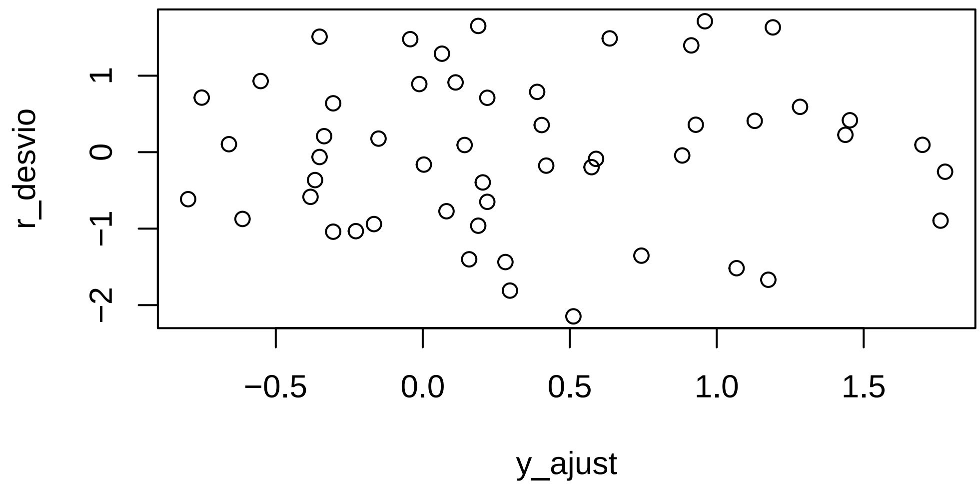
A odds aumenta em 16,6% em relação ao x1 sem incremento.

(c) Baseando-se no ajuste obtido em (a), construa os seguintes gráficos: "resíduos de Pearson vs. Valores ajustados" e "resíduos componente do desvio vs. valores ajustados". Interprete os dois graficos. Atenção! mostre os comandos que usou para obter os resíduos e os valores ajustados.

```
y_ajust <- modelo$coefficients[[1]] + modelo$coefficients[[2]]*dados$X1 + modelo$coefficie
r_pearson <- residuals(modelo, type = "pearson")
r_desvio <- residuals(modelo, type = "deviance")
plot(y_ajust, r_pearson)
```




```
plot(y_ajust, r_desvio)
```

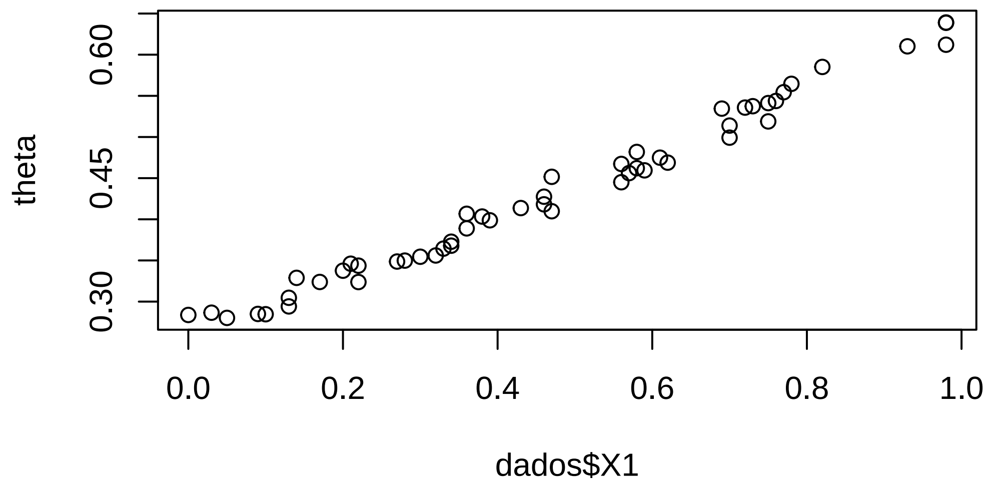


(d) Levando em consideração o resultado do ajuste em (a), obtenha as estimativas das probabilidades de morte da larva em cada recipiente do experimento (mostre o script R). Faça o gráfico: "X1 vs. probabilidade de morte" e "X2 vs. probabilidade de morte". Comente os resultados dos dois gráficos.

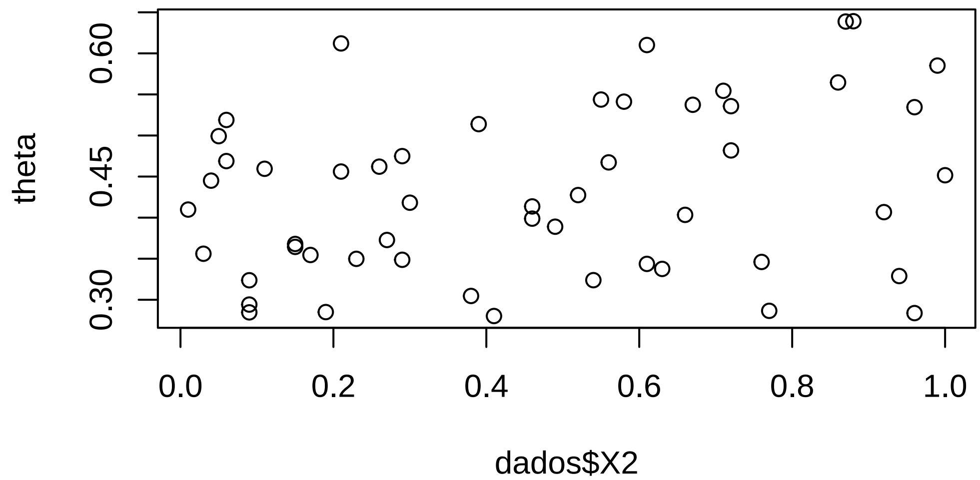
```
theta <- modelo$fitted.values
theta
```

```
        1         2         3         4         5         6         7         8 
0.3889749 0.5645748 0.2941773 0.2865373 0.5851185 0.6101619 0.5189159 0.4672559 
        9        10        11        12        13        14        15        16 
0.3560539 0.4450411 0.3237918 0.2846642 0.3987850 0.5544073 0.2837782 0.4992192 
       17        18        19        20        21        22        23        24 
0.4182048 0.4620742 0.4136614 0.4748793 0.6391317 0.5357181 0.4516117 0.6387343 
       25        26        27        28        29        30        31        32 
0.3288569 0.3238808 0.2850316 0.3679493 0.3460559 0.5344933 0.4098282 0.5138558 
       33        34        35        36        37        38        39        40 
0.4561065 0.4067152 0.6121120 0.3545130 0.3498347 0.3046541 0.3437027 0.3486787 
       41        42        43        44        45        46        47        48 
0.5412220 0.4274519 0.4033556 0.3375498 0.4818166 0.3727690 0.4688499 0.4594864 
       49        50        51        52 
0.5374110 0.5437668 0.3643706 0.2802087 
```

```r
plot(x = dados$X1, y = theta)
```




```r
plot(x = dados$X2, y = theta)
```


theta
0.30
0.45
0.60
0.0
0.2
0.4
0.6
0.8
1.0
dados$X2