Regressão Logística

- ✓A regressão linear geralmente é adequada quando a variável resposta é quantitativa.

- ✓O modelo logístico é semelhante ao modelo de regressão linear. No entanto, neste modelo a variável resposta é binária ou dicotômica, ou seja, assume dois valores, 0 e 1, “fracasso” e “sucesso”, respectivamente.

- ✓Esses valores podem ser atribuídos arbitrariamente a partir da observação de uma resposta qualitativa.

Exemplos:

✓Em epidemiologia estamos interessados se alguém tem ou não tem determinada doença. Se for uma doença do coração, então os preditores tais como idade, peso, pressão arterial sistólica, número de cigarros fumados e nível de colesterol são relevantes.

✓Em computação, podemos desejar saber se uma máquina completou ou não determinada tarefa em um tempo estabelecido;

✓Em marketing nós podemos desejar saber se alguém comprará ou não um carro na chegada de um novo ano. Aqui os preditores tais como renda anual, número de dependentes, valor da prestação do financiamento da casa, e assim por diante, são preditores relevantes;

✓Em educação nós desejamos saber somente se alguém passa ou não passa num teste;

✓Em processos de produção, um produto pode ser ou não barrado pelo controle de qualidade;

✓Em recursos humanos, um candidato a um posto de trabalho pode ser contratado ou não;

✓Além de possibilitar a classificação dos fenômenos ou indivíduos em categorias específicas, a regressão logística tem como objetivo estimar a probabilidade de ocorrência de determinado evento, ou seja, os resultados da variável dependente devem permitir interpretações em termos de probabilidade e não apenas de classificações em categorias.

Problemas quando a variável resposta é binária

1. Os erros não tem distribuição normal. Cada erro pode assumir um dos dois valores:

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

$$Y_i = 1 \Rightarrow \varepsilon_i = 1 - \beta_0 - \beta_1 X_i$$
$$Y_i = 0 \Rightarrow \varepsilon_i = -\beta_0 - \beta_1 X_i$$

2. Variâncias heterogêneas.

3. Restrição na função resposta. Como a função resposta representa probabilidades quando a variável resposta é binária, então:

$$\boxed{0 \le E(Y) = \pi \le 1}$$

Interpretação da função de resposta quando a variável resposta é binária

Considere Y_i uma v.a. Bernoulli com distribuição de probabilidade:

$$Y_i = 1 \rightarrow P(Y_i = 1) = \pi_i$$

$$Y_i = 0 \rightarrow P(Y_i = 0) = 1 - \pi_i$$

Pela definição de valor esperado, obtemos:

$$E(Y_i) = \pi_i$$

Assim, a resposta média, quando a variável resposta é uma variável binária (1 ou 0), sempre representa a probabilidade de $Y = 1$, para o nível da variável preditora X_i.

✓Note, portanto que o modelo linear não é apropriado para fazer as inferências quando a variável resposta é binária.

✓Geralmente, quando a variável resposta é binária, existe evidência empírica considerável, indicando que a forma da função resposta dever ser suposta não linear. Uma função monotonicamente crescente (ou decrescente) em forma de S é, na maioria das vezes, empregada.

Um modelo onde as probabilidades 0 e 1 são encontradas assintoticamente, como mostra a figura a seguir, é de modo geral, mais apropriado.

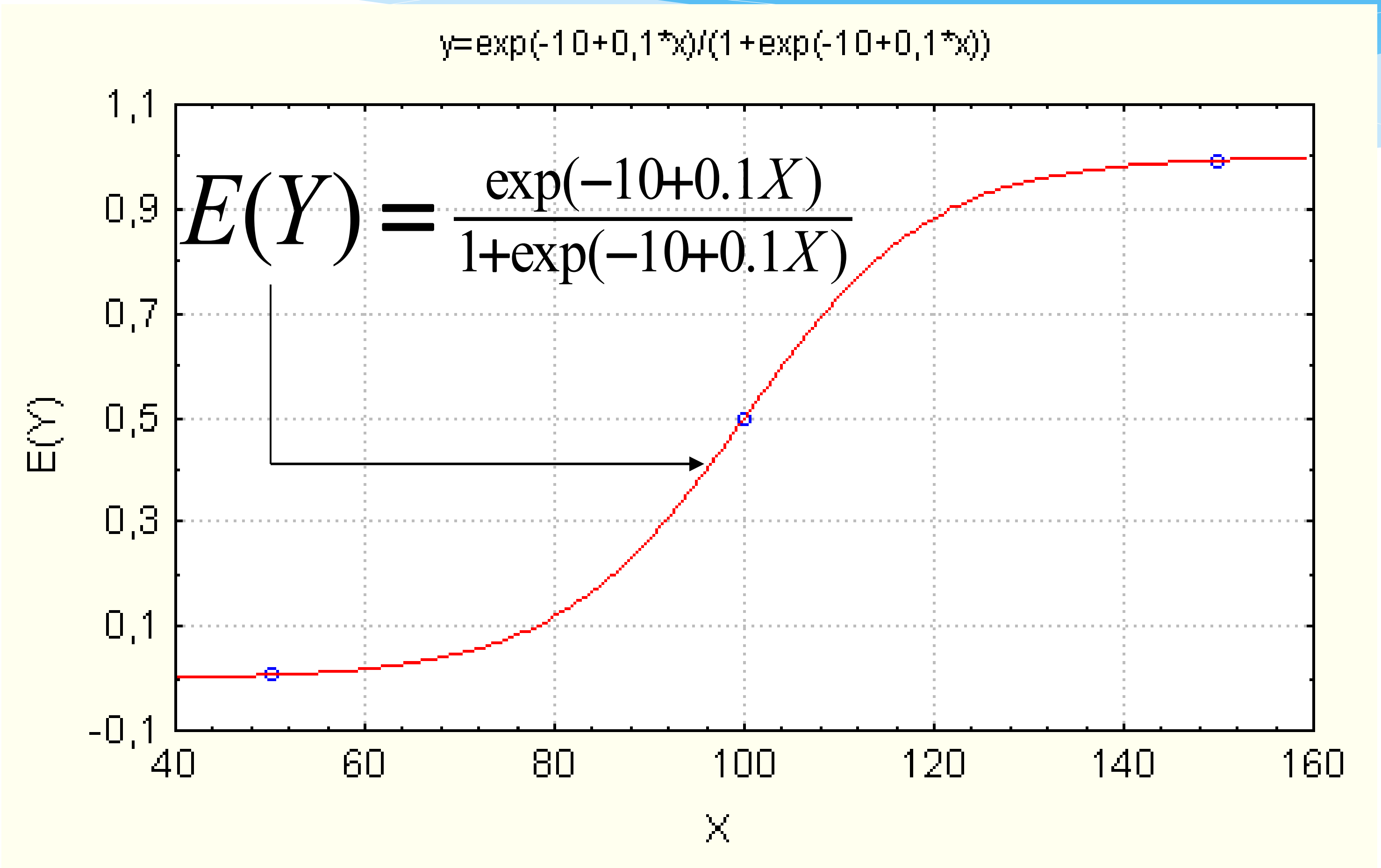


Figura: Função resposta logística

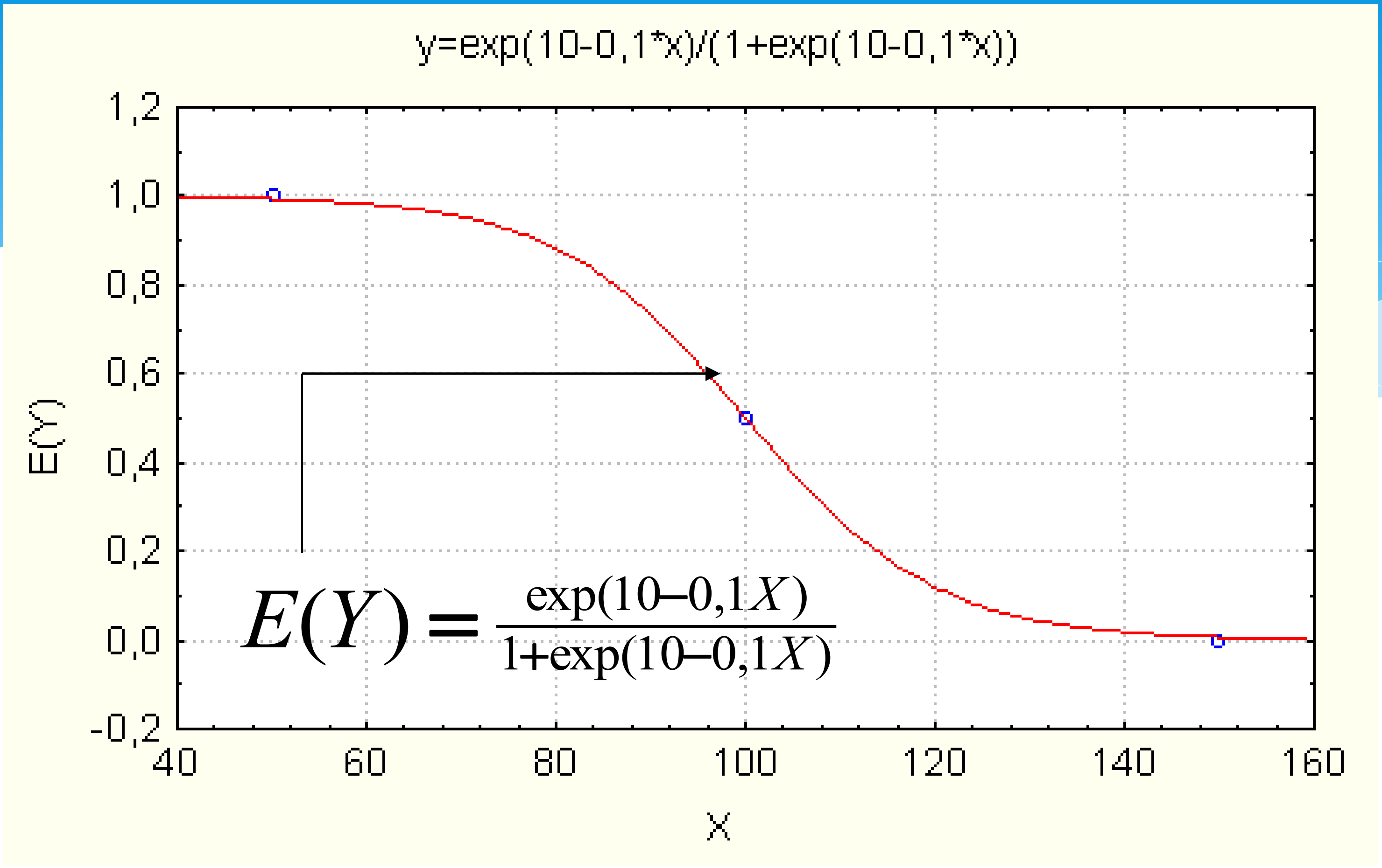


Figura: Função resposta logística

Função resposta logística com uma única variável preditora

- ✓ Considerações teóricas e práticas sugerem que quando a variável resposta é binária, a forma da função resposta será frequentemente curvilínea.

- ✓ Nas duas figuras anteriores, temos 2 funções respostas adequadas para uma variável resposta binária. Elas tem assíntotas em 0 e 1 e, assim, estão de acordo com a restrição 3.

- ✓ As funções respostas das figuras são denominadas *funções logísticas*, cuja expressão é:

$$E(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

Ou equivalentemente:

$$E(Y) = \frac{e^{(\beta_0+\beta_1 X)}}{1+e^{(\beta_0+\beta_1 X)}}$$

em que β_0 e β_1 são os coeficientes de regressão estimados e ***e*** é a base dos logaritmos naturais e sendo a variável Bernoulli tem se:

$$E(Y_i) = \pi_i \quad = \mathrm{P}(Y_i = 1)$$

- ✓ O termo $e^{\beta_0+\beta_1 x}$ é chamado de razão de chances (odds ratio).

A razão de chances pode ser interpretada da seguinte forma:

- ✓ **Razão de chances maior que uma unidade:** é o número de vezes que o **sucesso** é **mais provável** do que o **fracasso**, considerando um valor particular de x.
  - ✓ Ex: se a razão de chances é igual a 1,45 por exemplo, significa que o sucesso é 1,45 vezes mais provável que o fracasso, para aquele valor de x, ou seja, 45% maior de ocorrer o sucesso.
- ✓ **Razão de chances menor que uma unidade:** é o número de vezes que o **sucesso é menos provável do que o fracasso**, considerando um valor particular de x.
  - ✓ Ex: se a razão de chances é igual a 0,61 por exemplo, significa que o sucesso é 0,61 vezes menos provável que o fracasso, para aquele valor de x, ou seja, 39% menor de ocorrer o sucesso.

- ✓ O log da razão de chances, $\beta_0 + \beta_1 x$

é uma função linear da variável regressora. Esse termo é chamado de preditor linear, e a inclinação β_1 é a variação no log das chances que resulta do aumento de 1 unidade em x. Isso quer dizer que razão de chances varia de e^{β_1} quando x aumenta de uma unidade.

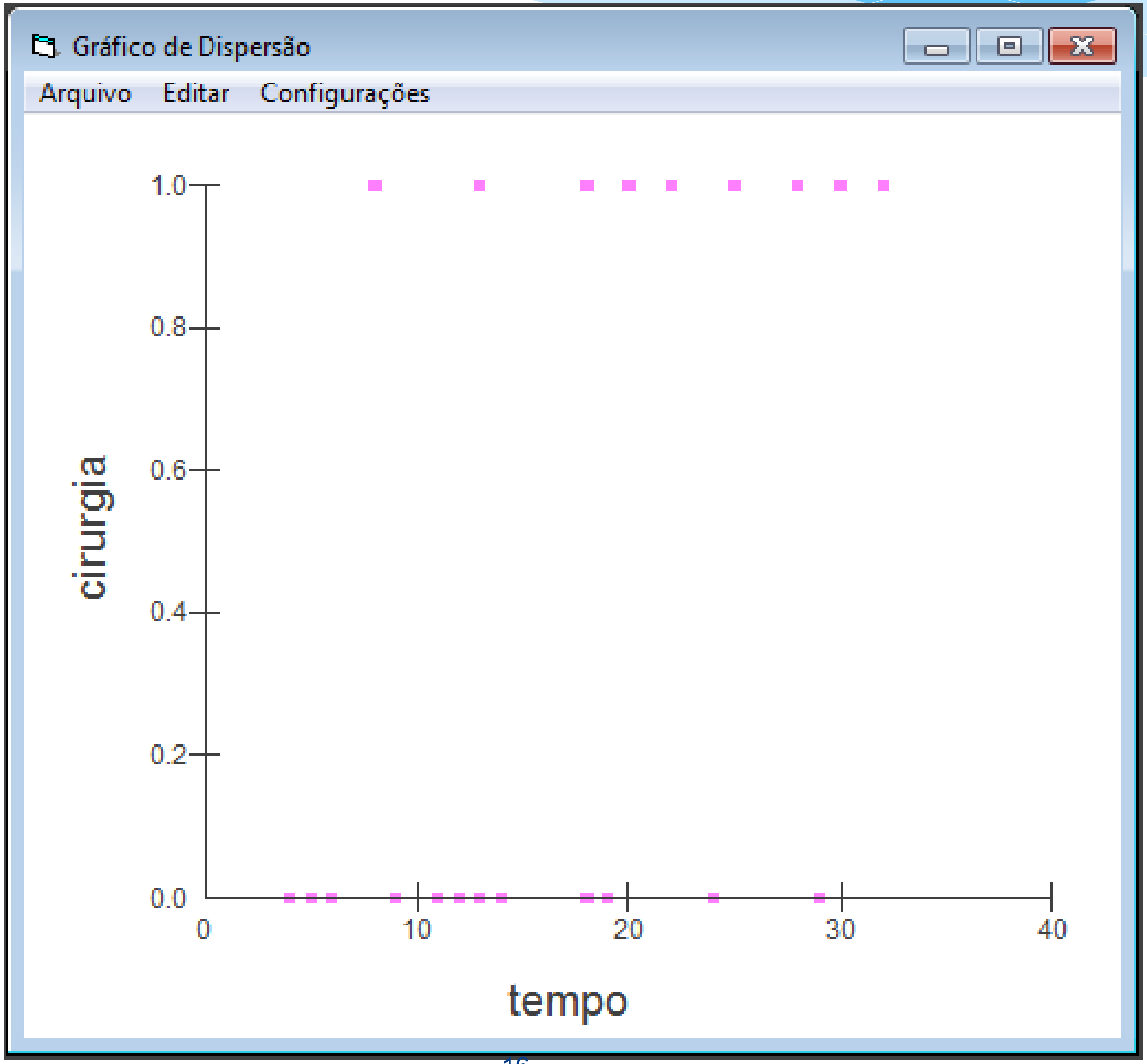
✓A estimação dos parâmetros é feita pelo método da máxima verossimilhança. No entanto, as funções são não lineares, e assim, são necessários métodos numéricos para se encontrar as estimativas de máxima verossimilhança.

Exemplo 1: Um pesquisador esta estudando o efeito do tempo de experiência de um médico sobre a habilidade para completar, dentro de um determinado tempo, uma cirurgia complicada. Vinte e cinco profissionais foram selecionadas para o estudo. A variável preditora, X, corresponde aos anos de experiência, enquanto a variável resposta refere-se ao sucesso na cirurgia (1) ou não (0).

tempo	cirurgia
14	0
29	0
6	0
25	1
18	1
4	0
18	0
12	0
22	1
6	0
30	1
11	0
30	1

tempo	cirurgia
5	0
20	1
13	0
9	0
32	1
24	0
13	1
19	0
4	0
28	1
22	1
8	1

Usando o BIOESTAT



Usando o BIOESTAT

Regressão Logística Múltipla

Imprimir

Regressão Logística (Logit)

No. de 0 (valor zero)	14.0000	56.00 %	Obs: Estes valores referem-se à variável dependente Y
No. de 1 (valor um)	11.0000	44.00 %	

Número de Variáveis Independentes:	Número de Casos	Função Objeto: Máxima Verossimilhança Valor Final
1	25	12.7123

Estimação de Y

Exibir

-2*log(verossimilhança)

Para este modelo	Só Intercepto	Qui-Quadrado	G. Liberdade	(p)
25.4246	34.2965	8.8719	1	0.0029

Equação

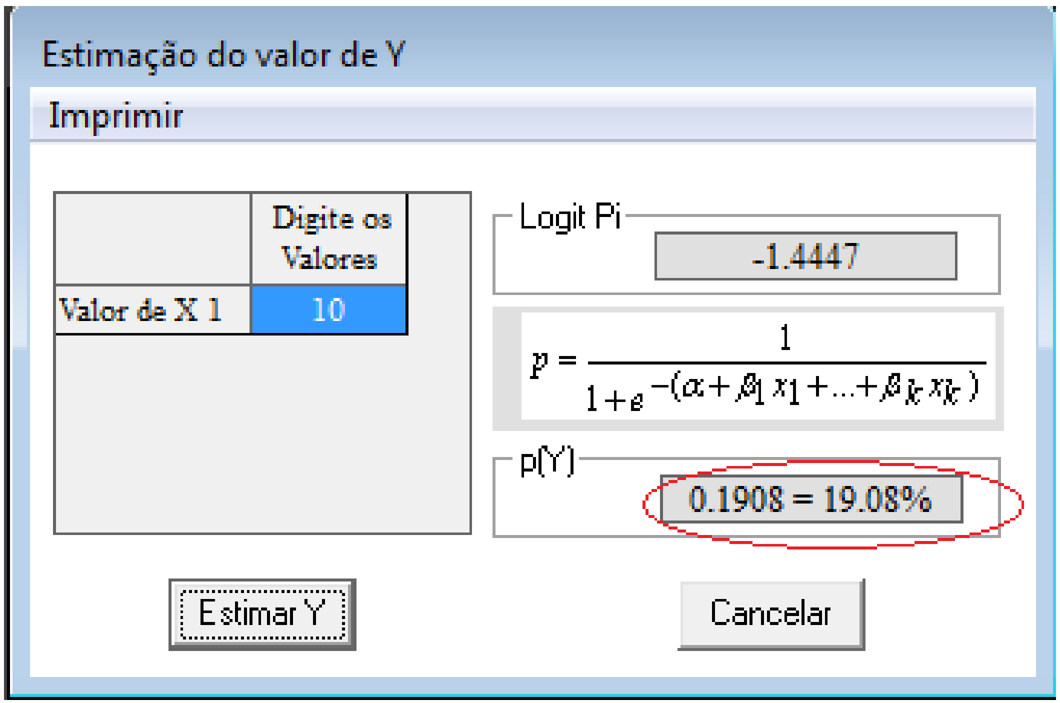
Logit Pi = -3.0597 + (0.1615 X1)

Variáveis	Coeficiente	Erro padrão	Z	p-valor	odds ratio	IC 95%
Intercepto	-3.0597	1.2594	---	---	---	---
X 1	0.1615	0.0650	2.4852	0.0129	1.1753	1.03 a 1.33

Modelo significativo

Um médico com experiência tem 17,5% mais chance de concluir a cirurgia com sucesso do que um médico sem experiência (odds ratio = 1,1753)

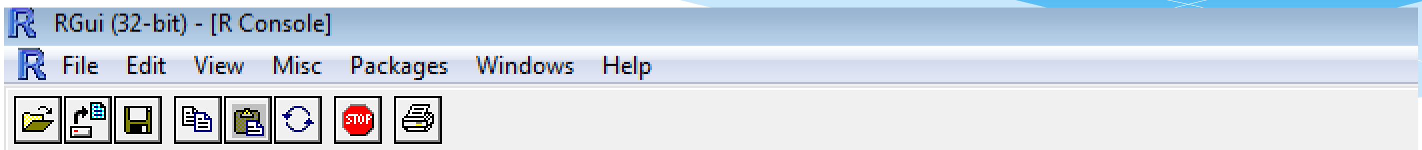
Para obter a probabilidade de realizar a cirurgia com sucesso dado 10 anos de experiência tem-se:



Observação:

Em geral, a razão das chances estimada quando existe uma diferença de *c unidades em X é* $e^{c\beta_1}$. *Exemplo:* desejamos comparar indivíduos com 10 meses e 25 meses de experiência, assim *c=15 meses, a razão das chances é estimada por* $e^{15*0,615}$ *, com resultado 11,3, portanto, isto indica que a chance de uma* pessoa com experiência terminar a tarefa aumenta mais de 11 vezes quando comparado com uma pessoa com pouca experiência.

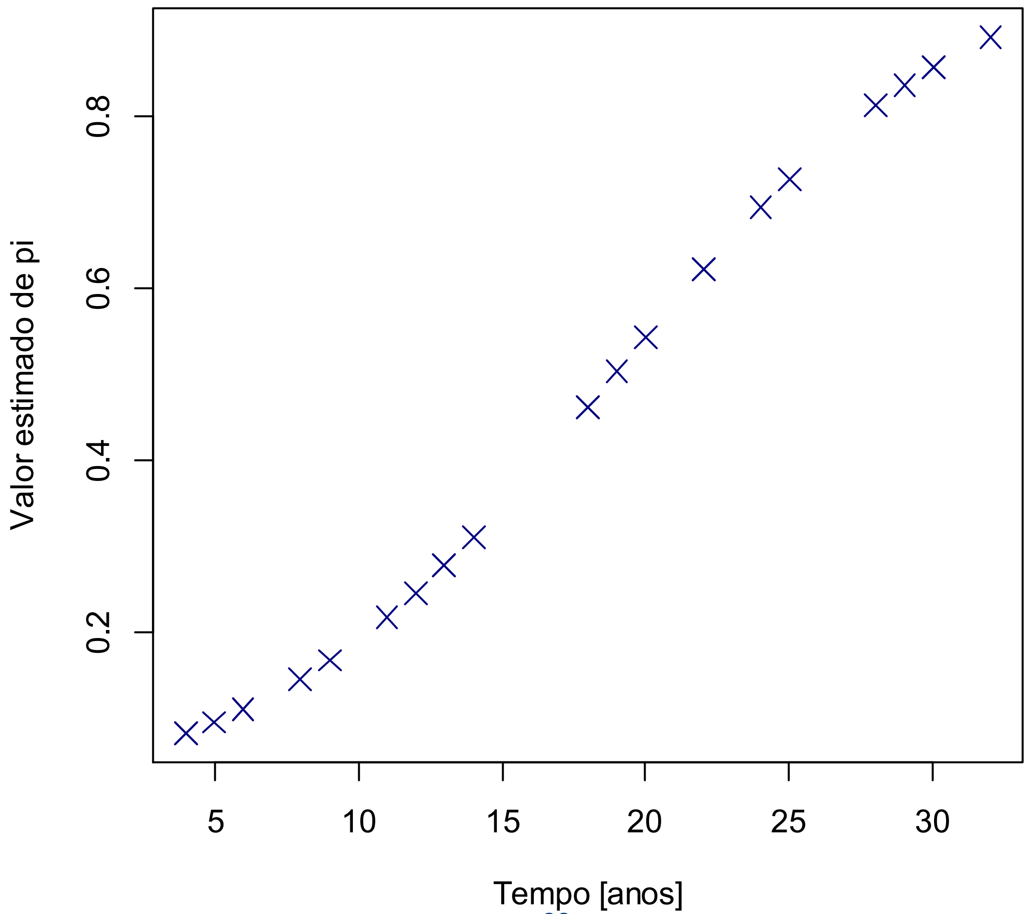
Usando o R




```
> reg<-read.table("clipboard",h=TRUE)
> attach(reg)
> modelo = glm(cirurgia ~ tempo, data = reg, family = binomial(link = 'logit'))
> summary(modelo)

Call:
glm(formula = cirurgia ~ tempo, family = binomial(link = "logit"),
    data = reg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
tempo        0.16149    0.06498   2.485   0.0129 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.296  on 24  degrees of freedom
Residual deviance: 25.425  on 23  degrees of freedom
AIC: 29.425

Number of Fisher Scoring iterations: 4
```

```
>
> #Razão de chances:
> exp(0.16149)
[1] 1.175261
>
>
> # Estimação da probabilidade de um médico com 10 anos de experiência:
> razaopiest<-exp(-3.0597+0.16149*10)
> razaopiest
[1] 0.2357932
> piest<-razaopiest/(1+razaopiest)
> piest
[1] 0.1908031
>
```

```
>
> temp = reg$tempo
> p = modelo$fitted.values
> xlb = expression(paste('Tempo [anos]'))
> #titulo = 'Prob falha x Temperatura'
> plot(temp, p,  col = 'navy', pch = 4, cex = 1.5, ylab = 'Valor estimado de pi', xlab = xlb)
>
```

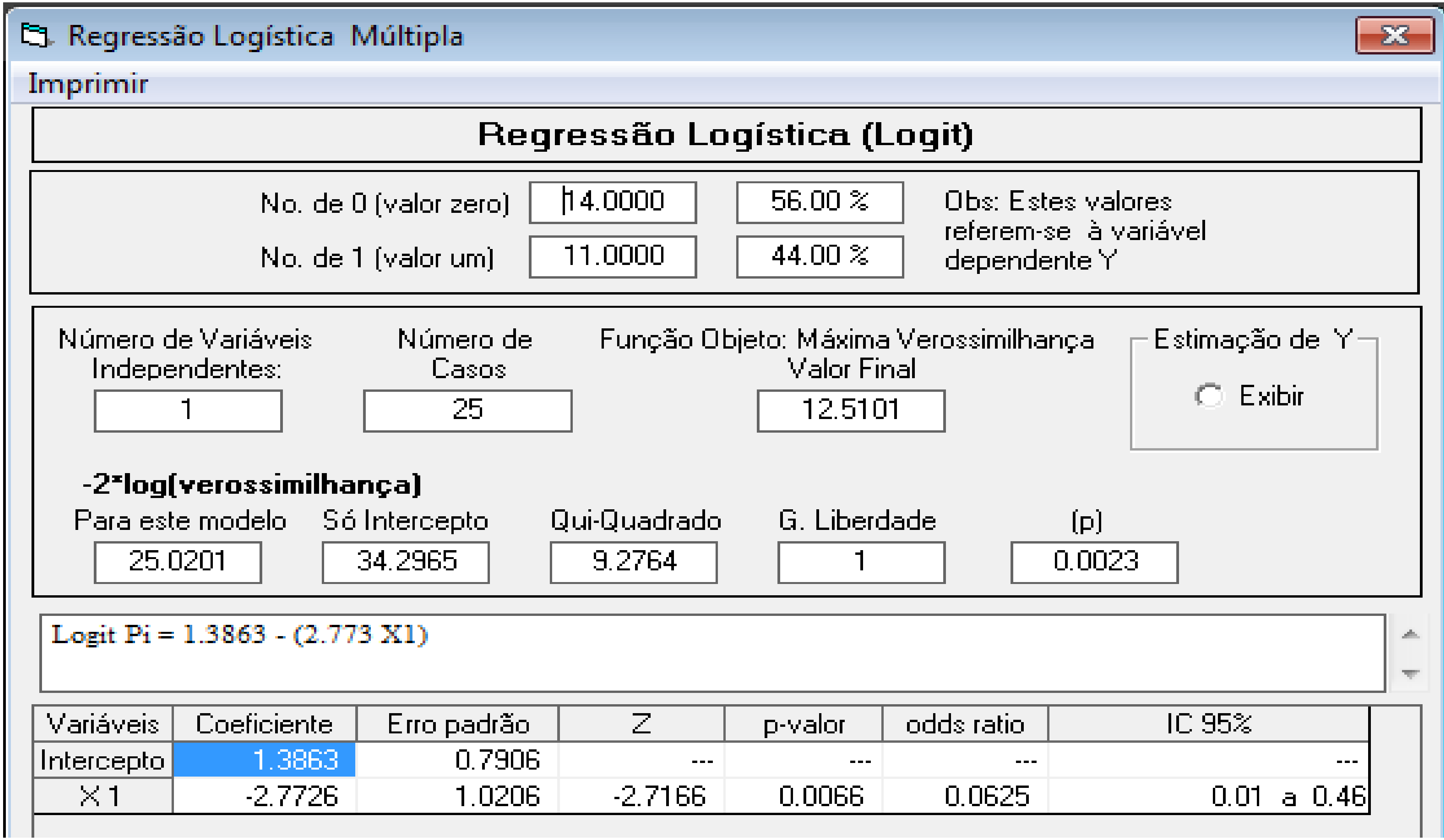


Função resposta logística com variável indicadora

Tal como o modelo de regressão linear, o de regressão logística pode ser generalizado para incluir variáveis explicativas discretas ou nominais, em adição as contínuas.

Exemplo 2: No exemplo 1, ao invés de avaliar o efeito do tempo de experiência de um médico sobre a habilidade para completar, dentro de um determinado tempo, uma cirurgia complicada, vamos considerar como variável preditora o sexo do profissional, codificando 1 para Masculino e 0 para Feminino.

cirurgia	0	0	0	1	1	0	0	0	1	0	1	0	1	0	1	0	0	1	0	1	0	0	1	1	1
sexo	1	1	1	0	0	1	0	1	0	1	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1



A chance da cirurgia ser bem sucedida é de 0,0625 vezes menos provável se o individuo é do sexo masculino (1), ou seja, 93,75% menor de ocorrer sucesso na cirurgia.

Função resposta logística com várias variáveis preditoras

- ✓Testa um variável binária Y em função de duas ou mais variáveis independentes que podem ser binárias ou não.
- ✓O modelo é: $E(Y)=\dfrac{1}{1+e^{-(\beta_0+\beta_1 X+\beta_2 X_2+...+\beta_k X_k)}}$

ou

$$E(Y)=\frac{e^{(\beta_0+\beta_1 X+\beta_2 X_2+...+\beta_k X_k)}}{1+e^{(\beta_0+\beta_1 X+\beta_2 X_2+...+\beta_k X_k)}}$$

Exemplo 3: Um pesquisador esta estudando o efeito do tempo de experiência e do sexo do médico sobre a habilidade para completar, dentro de um determinado tempo, uma cirurgia complicada. Vinte e cinco profissionais foram selecionados para o estudo. As variáveis preditoras foram: experiência em anos e sexo (1-Masculino; 0-feminino), enquanto a variável resposta refere-se ao sucesso na cirurgia (1) ou não (0).

cirurgia	0	0	0	1	1	0	0	0	1	0	1	0	1	0	1	0	0	1	0	1	0	0	1	1	1
tempo	14	29	6	25	18	4	18	12	22	6	30	11	30	5	20	13	9	32	24	13	19	4	28	22	8
sexo	1	1	1	0	0	1	0	1	0	1	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1

Usando o BIOESTAT

Regressão Logística Múltipla

Imprimir

Regressão Logística (Logit)

No. de 0 (valor zero)	14.0000	56.00 %
No. de 1 (valor um)	11.0000	44.00 %

Obs: Estes valores referem-se à variável dependente Y

Número de Variáveis Independentes:	Número de Casos	Função Objeto: Máxima Verossimilhança Valor Final
2	25	10.1372

Estimação de Y

Exibir

-2*log(verossimilhança)

Para este modelo	Só Intercepto	Qui-Quadrado	G. Liberdade	(p)
20.2744	34.2965	14.0221	2	0.0009

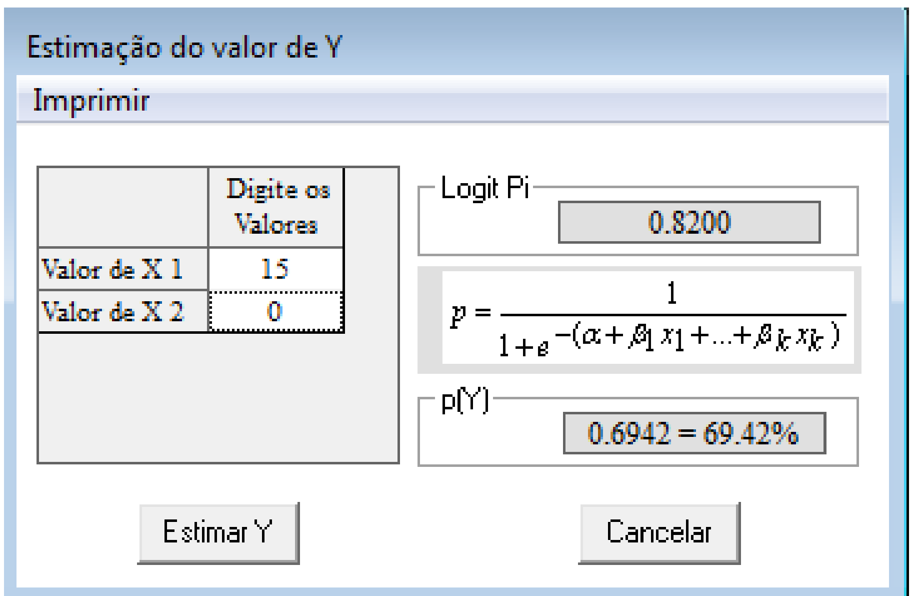
Logit Pi = -1.2635 + (0.1389 X1) - (2.427 X2)

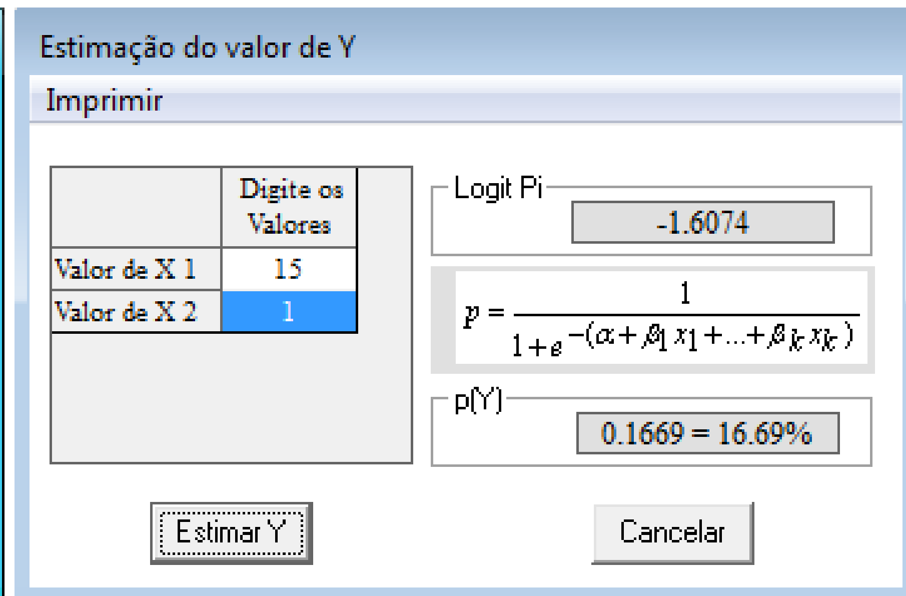
Variáveis	Coeficiente	Erro padrão	Z	p-valor	odds ratio	IC 95%
Intercepto	-1.2635	1.5069	---	---	---	---
X 1	0.1389	0.0705	1.9717	0.0486	1.1490	1.00 a 1.32
X 2	-2.4274	1.1466	-2.1170	0.0343	0.0883	0.01 a 0.84

• A chance da cirurgia ser bem sucedida é de cerca de 1,14 vezes para cada ano de acréscimo na experiência do profissional e de 0,09 vezes menos provável se o individuo é do sexo masculino (1), ou seja, as chances de a cirurgia ser bem sucedida é de 91% menor se esta for realizada por um profissional do sexo masculino.

• Um profissional com 15 anos de experiência e do sexo feminino tem 69,42% de chance de sucesso.

• Um indivíduo do sexo masculino com 15 anos tem 16,69%.





Exemplo 4. Um estudo visou verificar se a aquisição de um produto (SIM (1) e NÃO (0)) estava relacionada com uma campanha publicitária (sim (1) e não(0)) e com o sexo do cliente (feminino (1) e masculino (0)). Os dados amostrados foram:

Aquisição	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0	1	1	0	1	1
Campanha	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	1	1	1	1	0
Sexo	1	0	1	0	0	1	1	0	0	0	1	0	0	1	0	1	1	0	1	1

a) Obtenha o modelo ajustado e faça a interpretação de todos os parâmetros de interesse.

b) Verifique se as duas variáveis estão relacionadas com a aquisição.

c) Estime as probabilidades de aquisição para indivíduos dos sexos feminino e masculino com e sem realização de campanha.

Medidas de avaliação do Modelo Logístico

- ✓Log Verossimilhança
  - ✓A Log Verossimilhança, representado por -2LL, é o logaritmo natural da função de verossimilhança multiplicada por -2, seguindo uma distribuição de qui-quadrado. Seu papel é semelhante ao da estatística F, utilizada na avaliação do modelo linear.

  - ✓Calcula se a Log Verossimilhança, incluindo apenas a constante do modelo, e depois incluindo todas as variáveis independentes.

✓Quanto mais elevada a diferença entre os dois valores, maior o potencial dos coeficientes para estimar probabilidades associadas a ocorrência de determinado evento.

✓Essa diferença serve para testar a hipótese de que todos os coeficientes do modelo logístico são iguais a zero, tal como se verifica na distribuição F.

✓Teste Wald

✓Sua finalidade é aferir o grau de significância de cada coeficiente do modelo logístico. Seu papel é semelhante ao teste t, utilizado na avaliação dos modelos lineares, ou seja, testa a hipótese de que um determinado coeficiente é nulo.

Hipóteses: $H_0: \beta_j = 0$ versus $H_1: \beta_j \neq 0$

Nota: **A estatística de Wald segue uma distribuição normal padrão**

Estatística do Teste:

$$Wald = \left(\frac{\hat{\beta}_1}{ep\left(\hat{\beta}_1\right)} \right) \sim$$

Exercício: O gerente de uma transportadora de veículos esta interessado em aprimorar a sua política de vendas para expandir a base de clientes. Ele acredita que em muitas situações teria condições de realizar contratos a preços mais competitivos se tivesse uma melhor percepção da taxa de risco a que se expõe em cada operação. Recorrendo à sua base de dados, resolveu extrair uma amostra aleatória de 36 elementos para identificar quais são as variáveis que mais contribuem para diferencia-los quanto a ocorrência de sinistros. Com isso, espera poder estimar de forma mais racional o risco que ficará exposto em futuras operações e, consequentemente, conceder descontos mais adequados. Em relação a cada indivíduo, foram levantadas as seguintes informações: idade, estado civil (Solteiro-1; Casado-0), sexo (Feminino-1; Masculino-0) em relação a variável resposta sinistro (Houve sinistro-1; Não houve sinistro-0). Os resultados são apresentados a seguir:

Sinistro	Idade	Estado Civil	Sexo	Sinistro	Idade	Estado Civil	Sexo
0	22	1	1	0	19	0	0
1	24	0	0	0	18	0	0
0	45	1	1	0	21	1	1
0	58	0	1	0	59	0	1
0	27	1	0	1	24	0	0
1	31	0	1	0	56	0	1
1	32	0	1	0	54	0	1
0	30	0	0	0	47	0	1
0	56	0	1	0	40	1	0
0	44	0	1	1	31	1	1
1	21	0	1	0	43	0	0
1	23	0	1	0	35	1	0
0	29	1	0	0	23	1	0
0	20	1	0	1	22	1	1
0	60	1	1	0	21	1	1
1	30	0	1	0	63	0	1
1	22	0	1	0	22	1	0
0	26	1	0	0	26	1	0

Baseando-se neste histórico, pede-se:

a) Esboce um modelo capaz de descrever o relacionamento existente entre a ocorrência de sinistro e as variáveis em estudo;
b) Explicar o significado de cada componente do modelo;
c) Interprete as razões de chance obtidas;
d) Estime a probabilidade de sinistro associada a um cliente do sexo masculino, casado com 25 anos de idade;
e) Para o mesmo cliente citado no item anterior, qual a probabilidade de sinistro se ele for solteiro?
f) Compare os resultados obtidos nos dois itens anteriores e reflita sobre as estratégias que poderiam ser adotadas pela companhia para atrair novos clientes.

Exemplo 4: O controler de uma transportadora tem observado que menos de 40% dos funcionários submetidos a um programa de treinamento voltado para redução dos custos reagem positivamente. Interessado em aprimorar a política de pessoal, solicitou um estudo para identificar as causas desse baixo desempenho, inclusive porque o próprio treinamento já estava sob uma relação custo/benefício desvantajosa. Para tanto, extraiu-se uma amostra aleatória constituída de 36 empregados em relação aos quais foram consideradas as seguintes variáveis: número de anos de escolaridade, idade e sexo (1 = masculino e 0 = feminino). O fenômeno que está sob análise é a reação de cada componente da amostra e será codificada como 1 (um), quando positiva, e como 0 (zero), quando negativa, conforme consta no quadro no slide a seguir.

Dados do exemplo de Aplicação de Regressão Logística Múltipla

Reação	idade	escolaridade	sexo	Reação	idade	escolaridade	sexo
0	22	6	1	1	19	9	0
1	38	12	1	0	18	4	1
0	36	12	0	1	22	12	1
0	58	8	0	0	23	6	1
1	37	12	1	1	24	12	1
0	31	12	0	1	50	12	1
1	32	10	1	1	20	12	1
1	54	12	1	0	47	12	0
0	60	8	1	1	34	12	0
1	34	12	0	1	31	12	1
0	45	12	0	1	43	12	0
1	27	12	0	1	35	8	1
0	30	8	1	0	23	8	0
0	20	4	1	1	34	12	0
0	30	8	1	1	51	16	0
1	30	8	1	0	63	12	1
1	22	12	1	1	22	12	1
1	26	8	1	1	29	12	1

Quadro resumo do ajuste da regressão logística múltipla

Parâmetro	Estimativa	Erro-Padrão	Wald	Sig.	Exp(β)	IC de 95% Exp(β)	
						LI	LS
intercepto	-10,5877	4,4945	5,5493	**0,0185**	0,000025	--	--
Idade - β_1	-0,2314	0,1022	5,1265	**0,0236**	0,7934	0,6494	0,9694
Escolaridade - β_2	1,6425	0,6356	6,6779	**0,0098**	**5,1681**	1,4870	17,9619
Sexo - β_3	4,5705	2,1358	4,5794	**0,0324**	96,5924	1,4687	6352,42

Pelo teste de Wald, ao nível de significância de 5%, há evidencias de que todas as variáveis contribuem para explicar a reação de funcionários aos treinamentos em referencia.

Com base no quadro resumo do ajuste (slide anterior) tem-se que o modelo ajustado é:

$$\pi_i = P(Y_1 = 1 \mid X = x_i) = \frac{\exp(-10,5877 - 0,2314x_{1i} + 1,6425x_{2i} + 4,5705x_{3i})}{1 + \exp(-10,5877 - 0,2314x_{1i} + 1,6425x_{2i} + 4,5705x_{3i})}$$

Em que x1: representa a idade; x2: número de anos de escolaridade; x3: sexo do empregado (1 = masculino e 0 = mulher).

No conjunto de variáveis, a que alcançou maior nível de significância foi o número de anos de escolaridade de cada empregado. Dessa forma pode-se afirmar que as chances de o empregado apresentar reações positivas após o treinamento se elevam por um fator de 5,1681 (exp(1,6425)), a cada ano de acréscimo na escolaridade. Obviamente, o efeito sobre a probabilidade vai depender do nível em que ela se encontre em função das demais variáveis.

A variável sexo também exerce um efeito positivo, pois a estimativa do coeficiente associado a variável sexo é positivo. Conclui-se imediatamente que se o funcionário for do sexo masculino, as chances de se obterem reações favoráveis ao programa de redução de custos aumentam de forma considerável. Mais precisamente, elas se modificam por um fator 96,59.

Já a variável idade exerce um efeito contrário, pois a estimativa do coeficiente associado a variável idade é negativo, conclui-se que quanto mais elevada a idade do funcionário, menor é a probabilidade de que ele venha a reagir de forma favorável ao treinamento.

Simplificando-se um pouco mais o modelo ajustado, podemos chegar na seguinte expressão:

$$\pi_i = P\left(Y_1 = 1 \mid X = x_i\right) = \frac{1}{1 + e^{-(-10{,}5877 - 0{,}2314 x_{1i} + 1{,}6425 x_{2i} + 4{,}5705 x_{3i})}}$$

Agora vamos estimar a probabilidade de uma pessoa do sexo feminino reagir positivamente ao treinamento, tendo 22 anos de idade e apenas 6 de escolaridade (x1 = 22, x2 = 6 e x3 = 0):

$$\pi_i = P\left(Y_1 = 1 \mid X = x_i\right) = \frac{1}{1 + e^{-(-10{,}5877 - 0{,}2314 * 22 + 1{,}6425 * 6 + 4{,}5705 * 0)}} = 0{,}0029$$

Agora vamos estimar a probabilidade de uma pessoa do sexo masculino reagir positivamente ao treinamento, tendo 22 anos de idade e apenas 6 de escolaridade (x1 = 22, x2 = 6 e x3 = 1):

$$\pi_i = P\left(Y_1 = 1 \mid X = x_i\right) = \frac{1}{1 + e^{-(-10{,}5877 - 0{,}2314 * 22 + 1{,}6425 * 6 + 4{,}5705 * 1)}} = 0{,}2222$$