Regressão Logística

Propriedades

Geralmente a grande questão a ser respondida nos estudos epidemiológicos é saber qual a relação entre uma ou mais variáveis que refletem a exposição e a doença (efeito). Ou seja, deseja-se saber qual a probabilidade de ocorrência da doença, conhecendo-se como se dá a exposição. A probabilidade da doença varia entre 0 e 1. Para uma dada pessoa, y a doença real é um evento dicotômico, que pode ser entendido como 1 quando a doença ocorre e 0 quando esta não ocorre.

Se desejamos saber se o fumo materno está associado à ocorrência de baixo peso ao nascer, geralmente desejamos controlar o efeito de outras variáveis de confundimento.

No modelo logístico, usamos os valores de uma série de variáveis independentes para predizer a ocorrência da doença (variável dependente). Assim, todas as variáveis consideradas no modelo estão controladas entre si. Como usamos uma série de variáveis independentes, trata-se de um problema multivariável (não confundir como multivariado, termo usado quando se leva em conta uma série de variáveis dependentes (resposta) no modelo ao mesmo tempo e empregado geralmente fora do contexto estatístico na literatura biomédica). A medida de associação calculada a partir do modelo logístico é o odds ratio. Os odds ratio ajustados são obtidos através da comparação de indivíduos que diferem apenas na característica de interesse e que tenham os valores das outras variáveis constantes. O ajuste é apenas estatístico.

A função logística é perfeitamente aplicável aos problemas epidemiológicos porque é uma função que varia também entre 0 e 1. É um função em forma de S alongado. Seu modelo calcula a probabilidade do efeito pela seguinte fórmula:

$$P(X) = \frac{1}{1 + e^{-(\alpha + \Sigma \beta_i X_i)}}$$

Os termos α e β_i neste modelo representam parâmetros desconhecidos que serão estimados com base nos dados amostrais obtidos pelo método da máxima verossimilhança (maximiza a probabilidade de obter o grupo observado de dados). Através do modelo estimamos

$$\hat{\alpha} \qquad \hat{\beta}_i$$

Assim, sabendo os parâmetros α e β_i e conhecendo os valores das variáveis independentes para um indivíduo, podemos aplicar a fórmula acima para calcular a probabilidade de que este indivíduo desenvolva a doença – P (X).

No exemplo abaixo está calculada a regressão logística tendo como variável dependente o baixo peso ao nascer e como variáveis dependentes fumo materno (0,1) –

smoke - e número de consultas pré-natais no primeiro trimestre de gravidez (de 1 a 6) - ftv.

logit low ftv smoke

```
Iteration 0:  Log Likelihood =  -117.336
Iteration 1:  Log Likelihood =-114.58313
Iteration 2:  Log Likelihood =-114.57008
Iteration 3:  Log Likelihood =-114.57008

Logit Estimates                                        Number of obs =    189
                                                       chi2(2)       =   5.53
                                                       Prob > chi2   = 0.0629
Log Likelihood = -114.57008                            Pseudo R2     = 0.0236

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .6977775   .3203178      2.178   0.029        .069966    1.325589
     ftv |  -.1246574   .1554117     -0.802   0.422      -.4292588     .179944
   _cons |  -.9888444    .245176     -4.033   0.000       -1.46938   -.5083083
------------------------------------------------------------------------------
```

No exemplo, a função logística expressa a seguinte probabilidade:

$$P(X) = \frac{1}{1 + e^{-(\alpha + \beta_1 smoke + \beta_2 ftv)}}$$

As estimativas obtidas são α=-0.9888444, β_1 = 0.6977775 e β_2 = -0.1246574. Desse modo, para uma mãe fumante (1), que realizou 3 consultas pré-natais no primeiro trimestre de gravidez, a probabilidade de ter um filho com baixo peso ao nascer, estimada por este modelo é 34%:

$$P(X) = \frac{1}{1 + e^{-(-0.989 + 0.698(1) + (-0.125)(3))}}$$

$$P(X) = 1 / 1 + e^{-(-0.666)} = 1 / 1 + 1.946$$

$$P(X) = 0.339 = 34\%$$

Uma das grandes vantagens da regressão logística é que cada coeficiente estimado fornece uma estimativa do logaritmo natural (ln) do odds ratio ajustado para todas as variáveis do modelo, permitindo a estimação direta do odds ratio através da exponenciação do coeficiente β_1:

$$\mathbf{OR = e^{\beta_1}}$$

No caso da variável fumo materno, smoke, o coeficiente β_1 estimado pela regressão logística foi de 0.698. Exponenciando-se este coeficiente obtém-se 2.010, que representa uma estimativa ajustado do odds ratio para fumo materno, controlando-se pelo número de consultas de pré-natal.
A função exponencial é uma função contrária ao logaritmo natural. Assim, extraindo-se o logaritmo natural do odds ratio obtém-se o coeficiente β_1.

O intervalo de confiança de 95 para o OR é calculado de forma análoga:

IC 95% (OR)= $e^{[\beta_1 \pm 1.96 \times \text{erro padrão}(\beta_1)]}$

IC 95% (OR)= $e^{[0.698 \pm 1.96 \times 0.320]}$

IC 95% (OR)= $e^{[0.698 \pm 1.96 \times 0.320]}$

IC 95% (OR)= $e^{[0.698 \pm 0.627]}$

Limite inferior do IC 95% (OR)= $e^{0.071}$ = 1.074

Limite superior do IC 95% (OR)= $e^{1.325}$ = 3.76

Há dois comandos para realização da regressão logística no Stata: **logit** para se obter os coeficientes do modelo (α e β_i) e **logistic** para se obter os odds ratio.

Vamos analisar o banco de dados LBW (fatores de risco para o baixo peso ao nascer em Massachusetts, 1986). Observe as variáveis deste arquivo:

id	número de identificação do paciente
low	baixo peso ao nascer 0=não 1=sim
age	idade materna em anos completos (14-45)
lwt	peso no início da gravidez em libras (80-250)
race	raça 1=branco 2=negro 3=outra
smoke	fumo materno 0=não 1=sim
ptl	número de partos prematuros anteriores (0-3)
ht	hipertensão materna 0=não 1=sim
ui	irritabilidade uterina 0=não 1=sim
ftv	número de consultas no pré-natal no primeiro trimestre da gravidez (0-6)
bwt	peso ao nascer em gramas (709-4990)

A variável resposta é low. Vamos rodar um modelo de regressão logística simples, com apenas 1 fator de risco. Vamos utilizar a variável categórica dicotômica fumo materno (smoke).

Variável dicotômica

logit low smoke

```
Iteration 0:  Log Likelihood =  -117.336
Iteration 1:  Log Likelihood = -114.9123
Iteration 2:  Log Likelihood = -114.9023

Logit Estimates                                   Number of obs =    189
                                                  chi2(1)       =   4.87
                                                  Prob > chi2   = 0.0274
Log Likelihood =  -114.9023                       Pseudo R2     = 0.0207

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .7040592   .3196386      2.203   0.028       .0775791    1.330539
   _cons |  -1.087051   .2147299     -5.062   0.000      -1.507914   -.6661886
------------------------------------------------------------------------------
```

logistic low smoke

```
Logit Estimates                                       Number of obs =    189
                                                      chi2(1)       =   4.87
                                                      Prob > chi2   = 0.0274
Log Likelihood =  -114.9023                           Pseudo R2     = 0.0207

------------------------------------------------------------------------------
     low | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   2.021944   .6462912      2.203   0.028       1.080668    3.783083
------------------------------------------------------------------------------
```

O coeficiente β_1 = 0.7040592 para smoke, assim:

OR= e $^{0.7040592}$ = 2.02

Como se pode observar acima o fumo materno é um fator de risco para o baixo peso ao nascer nesta população.

Variável contínua

Vamos realizar os mesmo procedimentos para a variável age, quantitativa contínua.

logit low age

```
Iteration 0:  Log Likelihood =  -117.336
Iteration 1:  Log Likelihood =-115.96259
Iteration 2:  Log Likelihood =-115.95598
Iteration 3:  Log Likelihood =-115.95598

Logit Estimates                                       Number of obs =    189
                                                      chi2(1)       =   2.76
                                                      Prob > chi2   = 0.0966
Log Likelihood = -115.95598                           Pseudo R2     = 0.0118

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |  -.0511529   .0315138     -1.623   0.105      -.1129188    .0106129
   _cons |   .3845819   .7321251      0.525   0.599      -1.050357    1.819521
------------------------------------------------------------------------------
```

logistic low age

```
Logit Estimates                                       Number of obs =    189
                                                      chi2(1)       =   2.76
                                                      Prob > chi2   = 0.0966
Log Likelihood = -115.95598                           Pseudo R2     = 0.0118

------------------------------------------------------------------------------
     low | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .9501333   .0299423     -1.623   0.105       .8932232    1.010669
------------------------------------------------------------------------------
```

O coeficiente para idade materna é –0.051. O odds ratio para cada ano de incremento na idade é 0.95, ou seja, a cada ano de idade materna, há uma redução de 5% no risco de baixo peso ao nascer. Mas como o intervalo de confiança incluiu o 1, a idade materna não é um fator de risco para o baixo peso ao nascer nesta população. Observe que a idade está modelada como variável contínua. Na logística as variáveis independentes podem ser quantitativas ou categóricas (0 - 1).

O odds ratio de uma variável contíua representa uma média dos odds nos diversos níveis desta variável. Pode-se também calcular o odds ratio para incrementos de idade maior do que 1 ano. Por exemplo, se quisermos calcular o incremento no risco associado a um aumento de 10 anos na idade, tem-se:

OR= $e^{\beta_1 \times 10} = e^{-0.0511529 \times 10} = e^{-0,511529}$ = 0.59957787

O Stata tem um comando para realizar este cálculo:

display exp(_b[age]*10)

.59957787

Portanto o risco de baixo peso cai quase à metade, a cada incremento de 10 anos na idade materna.

O programa calcula dois testes de hipótese para avaliar a significância da variável no modelo: o teste de Wald e o teste da razão de verossimilhanças. O valor z do teste de Wald é obtido dividindo-se o coeficiente β_1 pelo seu erro padrão. No exemplo da idade, -0.0511529/ 0.0315138 = -1.623, o que dá uma probabilidade (valor de p) de 0.105. Este teste segue a distribuição normal. Portanto, se o valor de z do teste der maior que 1.96, diz-se que a variável é significante. O segundo teste é calculado utilizando-se a seguinte fórmula:

TRV = - 2 [log da verossimilhança do modelo com a constante – log da verossimilhança do modelo com a variável (age)]

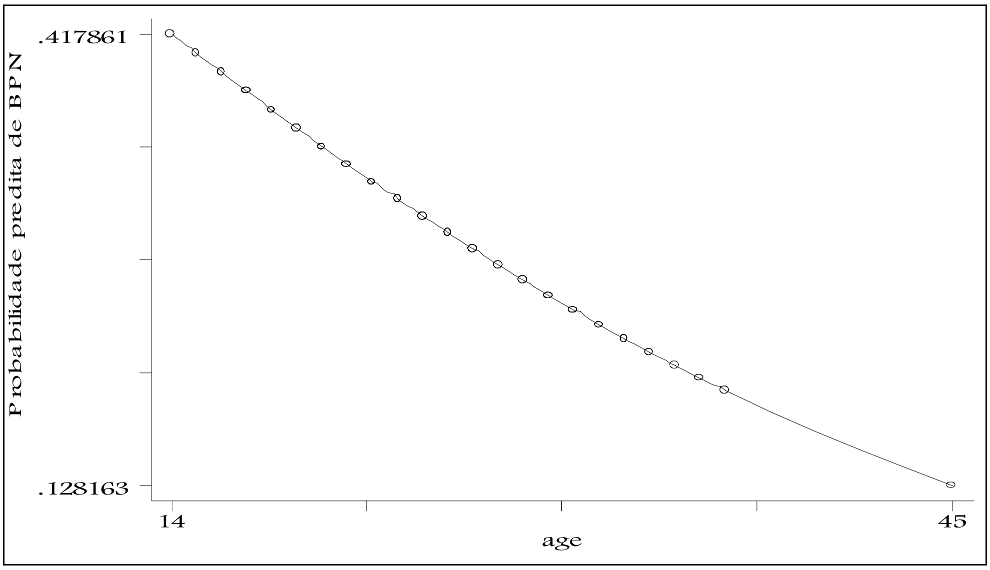
O modelo sem a constante é aquele ajustado no passo interativo 0 (iteration 0)

TRV= -2 [-117.336 – (-115.95598)] = **2.76**

Este teste segue a distribuição do Qui-quadrado com 1 grau de liberdade . Portanto, se o seu valor for maior que 3,84, então $p < 0.05$. O programa calcula o valor exato de p, fornecendo 0.0966. O teste da razão de verossimilhanças é mais acurado do que o teste de Wald, sendo preferível o seu uso em amostras de tamanho pequeno ou moderado. Para grandes amostras as duas estimativas fornecem resultados muito próximos um do outro.

Como dito acima, age foi modelada como variável contínua. Um dos poucos pressupostos do modelo logístico é que você só pode modelar uma variável como contínua se houver evidência de linearidade, ou seja de que para cada incremento na idade materna, corresponda um decréscimo ou acréscimo na probabilidade de ocorrência do evento, no caso, de baixo peso ao nascer. É necessário verificar se este pressuposto se aplica no caso da idade. Se se aplicar podemos continuar modelando age como variável contínua. Se não se aplicar, passaremos a modelar age como uma variável categórica, a partir de pontos de corte de significado biológico ou baseados em quartis. Há várias formas de testar este pressuposto. Numa delas, se plota em um gráfico a idade materna versus a probabilidade de ocorrência de baixo peso ao nascer predita pelo modelo. Os comando no Stata são:

predict probbpn
label variable probbpn "Probabilidade predita de BPN"
graph probbpn age, connect(s)



Como se pode verificar no gráfico acima há uma tendência linear de decréscimo na probabilidade de baixo peso ao nascer predita pelo modelo na medida em que aumenta a idade materna. Portanto, a idade materna pode ser modelada como uma variável quantitativa contínua no modelo.

Variável categórica com mais de dois níveis

No caso da variável **race,** categórica, não podemos fazer a regressão sem antes fatorá-la e transformá-la em 3 variáveis dummy (categórica 0 – 1), conforme esquema abaixo:

race	race1	race2	race3
1 – branca	1	0	0
2 – negra	0	1	0
3 – outra	0	0	1

No caso, a variável race1 é a variável dummy para raça branca. Ela assumirá o valor 1 quando a raça for branca e 0 nos demais casos. A variável race2 assumirá o valor 1 quando a raça for negra e 0 nos demais casos e a variável race3 assumirá o valor 1 quando a raça for outra e 0 nos demais casos. Observe que para cada categoria da variável será criada uma variável dummy. O comando é:

tabulate race, generate(race)

A variável race1 não precisa ser utilizada, pois a raça branca é a categoria basal, que servirá como referência para as outras categorias. Vamos fazer a regressão logística simples para race2 e race3.

logistic low race2

```
Iteration 0:  Log Likelihood =  -117.336
Iteration 1:  Log Likelihood =-116.51366
Iteration 2:  Log Likelihood =-116.50935
Iteration 3:  Log Likelihood =-116.50935

Logit Estimates                                       Number of obs =    189
                                                      chi2(1)       =   1.65
                                                      Prob > chi2   = 0.1985
Log Likelihood = -116.50935                           Pseudo R2     = 0.0070

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   race2 |   .5635762   .4325561      1.303   0.193      -.2842181     1.41137
   _cons |  -.8737311     .17184     -5.085   0.000      -1.210531   -.5369309
------------------------------------------------------------------------------
```

logistic low race3

```
Iteration 0:  Log Likelihood =  -117.336
Iteration 1:  Log Likelihood =-116.45064
Iteration 2:  Log Likelihood =-116.44906

Logit Estimates                                       Number of obs =    189
                                                      chi2(1)       =   1.77
                                                      Prob > chi2   = 0.1829
Log Likelihood = -116.44906                           Pseudo R2     = 0.0076

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   race3 |   .4321825   .3233953      1.336   0.181      -.2016606    1.066026
   _cons |  -.9509763   .2019289     -4.709   0.000       -1.34675   -.5552028
------------------------------------------------------------------------------
```

O Stata tem um comando para gerar variáveis dummy automaticamente. Verifique digitando:

xi: logistic low i.race

```
i.race           _Irace_1-3         (naturally coded; _Irace_1 omitted)

Logistic regression                              Number of obs   =        189
                                                 LR chi2(2)      =       5.01
                                                 Prob > chi2     =     0.0817
Log likelihood = -114.83082                      Pseudo R2       =     0.0214

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Irace_2 |   2.327536   1.078613     1.82   0.068     .9385073    5.772385
    _Irace_3 |   1.889234   .6571342     1.83   0.067     .9554577    3.735597
------------------------------------------------------------------------------
```

Estratégias de modelagem

Modelo reduzido

Variáveis medindo efeitos principais

A meta do modelo reduzido é obter o melhor modelo ajustado minimizando o número de variáveis incluídas no modelo, descartando aquelas não significantes, que dão contribuição quase nula para o ajuste. Ficam apenas as variáveis com valor de p menor que 0.05, a não ser que a variável seja biologicamente muito importante e tenha um valor de p próximo a 0.05.
Qualquer mudança biologicamente importante no coeficiente do fator de risco estimado, comparando-se modelos com e sem o fator de risco, indica que a covariável é um fator de confusão e deve permanecer no modelo, mesmo que o seu próprio coeficiente não seja significante.
Inicia-se o processamento realizando o que fizemos acima, a regressão logística simples para cada variável independente. Selecionam-se, depois as variáveis que apresentarem um p no teste de hipótese de pelo menos 0.20 ou menos. Nos dois exemplos acima, todas as variáveis vão entrar no modelo, pois o fumo materno apresentou um p=0.0274, e a idade materna, embora não associada na análise bruta, mostrou um p=0.0966, portanto maior do que 0.20. A raça também apresentou nas suas duas categorias valores de p inferiores a 20%.
Realize agora a regressão linear simples para as outras variáveis do modelo. Vamos fazer apenas o procedimento **logistic**.

```
logistic low lwt
logistic low ptl
logistic low ht
logistic low ui
logistic low ftv
```

Os valores de p para cada variável constam da tabela abaixo:

Variável	Valor de p
age	0.0966
smoke	0.0274
lwt	0.0145
ptl	0.0092
ht	0.0449
ui	0.0243
ftv	0.3792
race2	0.1985
race3	0.1829

Como observado acima, apenas a variável **ftv** não preenche o critério para entrada no modelo, pois apresentou um p > 0.20.

Agora vamos rodar a regressão logística múltipla com todas as variáveis acima, exceto **ftv**.

logit low age smoke lwt ptl ht ui race2 race3

```
Iteration 0:  Log Likelihood =  -117.336
Iteration 1:  Log Likelihood =-101.38735
Iteration 2:  Log Likelihood =-100.72104
Iteration 3:  Log Likelihood =-100.71348
Iteration 4:  Log Likelihood =-100.71348

Logit Estimates                                        Number of obs =    189
                                                       chi2(8)       =  33.25
                                                       Prob > chi2   = 0.0001
Log Likelihood = -100.71348                            Pseudo R2     = 0.1417

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |  -.0270698   .0364526     -0.743   0.458      -.0985156     .044376
   smoke |   .9233492   .4008583      2.303   0.021       .1376813    1.709017
     lwt |  -.0151826   .0069279     -2.192   0.028       -.028761   -.0016041
     ptl |   .5417551   .3462666      1.565   0.118      -.1369149    1.220425
      ht |   1.833696     .69177      2.651   0.008       .4778514     3.18954
      ui |   .7585965   .4593918      1.651   0.099      -.1417949    1.658988
   race2 |   1.263219   .5264677      2.399   0.016       .2313616    2.295077
   race3 |   .8616351   .4391975      1.962   0.050       .0008239    1.722446
   _cons |   .4644033   1.204702      0.385   0.700      -1.896769    2.825576
------------------------------------------------------------------------------
```

O ajuste global do modelo pode ser verificado através do teste da razão de verossimilhanças. Este teste verifica a hipótese de nulidade de que todos os coeficientes no modelo, exceto o da constante são iguais a zero. O resultado de 0.0001 indica que o modelo se ajusta adequadamente aos dados.

O próximo passo na modelagem é excluir as variáveis que não apresentaram significância de 0.10 ou mais no modelo completo. No exemplo acima vamos excluir a variável age e aplicar o teste da razão de verossimilhanças comparando-se o modelo cheio com o modelo sem a variável age. Modelos deste tipo, que têm as mesmas variáveis apenas com exceção de algumas que são retiradas, são ditos modelos aninhados. Antes vamos ter que salvar os resultados do modelo cheio com o comando:

lrtest, saving(0)

logit low smoke lwt ptl ht ui race2 race3

```
Iteration 0:  Log Likelihood =  -117.336
Iteration 1:  Log Likelihood =-101.58398
Iteration 2:  Log Likelihood =-100.99797
Iteration 3:  Log Likelihood =-100.99279
Iteration 4:  Log Likelihood =-100.99279

Logit Estimates                                        Number of obs =    189
                                                       chi2(7)       =  32.69
                                                       Prob > chi2   = 0.0000
Log Likelihood = -100.99279                            Pseudo R2     = 0.1393

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .9387268   .3987195      2.354   0.019       .1572509    1.720203
     lwt |  -.0159053   .0068553     -2.320   0.020      -.0293414   -.0024691
```

```
     ptl |   .5032149   .3412323      1.475   0.140      -.1655881    1.172018
      ht |   1.855042   .6951214      2.669   0.008       .4926286    3.217455
      ui |   .7856975   .4564423      1.721   0.085       -.108913    1.680308
   race2 |   1.325719   .5222464      2.538   0.011       .3021351    2.349304
   race3 |   .8970779   .4338846      2.068   0.039       .0466797    1.747476
   _cons |  -.0865495    .951768     -0.091   0.928      -1.951981    1.778882
------------------------------------------------------------------------------
```

lrtest

```
Logit:  likelihood-ratio test                        chi2(1)     =       0.56
                                                     Prob > chi2 =     0.4548
```

O teste acima foi calculado pela mesma fórmula explicada anteriormente. Veja:

TRV = - 2 [-100.99279 – (-100.71348)] = 0.56

Portanto a retirada da variável idade materna do modelo não influiu no ajuste do mesmo. Desse modo, o modelo preferido é o modelo sem a idade materna.

Outra variável que pode ser retirada para verificar a sua contribuição no ajuste é ptl, que tem um valor de p de 0.14.

logit low smoke lwt ht ui race2 race3

```
Iteration 0:  Log Likelihood =  -117.336
Iteration 1:  Log Likelihood =-102.68681
Iteration 2:  Log Likelihood =-102.11335
Iteration 3:  Log Likelihood =-102.10831
Iteration 4:  Log Likelihood =-102.10831

Logit Estimates                                        Number of obs =    189
                                                       chi2(6)       =  30.46
                                                       Prob > chi2   = 0.0000
Log Likelihood = -102.10831                            Pseudo R2     = 0.1298

------------------------------------------------------------------------------
     low |      Coef.   Std. Err.        z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   1.035831   .3925611      2.639   0.008       .2664256    1.805237
     lwt |  -.0167325   .0068034     -2.459   0.014      -.0300669    -.003398
      ht |   1.871416   .6909051      2.709   0.007       .5172672    3.225565
      ui |    .904974   .4475541      2.022   0.043        .027784    1.782164
   race2 |   1.324562   .5214669      2.540   0.011       .3025055    2.346618
   race3 |   .9261969   .4303893      2.152   0.031       .0826495    1.769744
   _cons |   .0562761   .9378604      0.060   0.952      -1.781897    1.894449
------------------------------------------------------------------------------
```

lrtest

```
Logit:  likelihood-ratio test                        chi2(2)     =       2.79
                                                     Prob > chi2 =     0.2479
```

Como pode ser verificado acima, a retirada de ptl do modelo também não alterou o seu ajuste. Portanto um modelo sem ptl e age é o preferível. Este modelo ajustado é considerado o modelo reduzido final.

Vamos pedir então o comando logistic para analisar os odds ratio do modelo final:

logistic low smoke lwt ht ui race2 race3

```
Logit Estimates                                         Number of obs =    189
                                                        chi2(6)       =  30.46
                                                        Prob > chi2   = 0.0000
Log Likelihood = -102.10831                             Pseudo R2     = 0.1298

------------------------------------------------------------------------------
     low | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   2.817447    1.10602      2.639   0.008        1.30529    6.081411
     lwt |   .9834068   .0066905     -2.459   0.014       .9703806    .9966078
      ht |   6.497492    4.48915      2.709   0.007       1.677437     25.1678
      ui |   2.471868   1.106295      2.022   0.043       1.028174    5.942703
   race2 |   3.760537   1.960996      2.540   0.011       1.353245    10.45017
   race3 |   2.524889   1.086685      2.152   0.031       1.086161    5.869353
------------------------------------------------------------------------------
```

Modelo passo a passo automático

Regressão logística passo a passo (stepwise) com seleção para trás (backward selection)

Stepwise com backward selection 0.10

xi: sw logistic low age lwt i.race smoke ptl ht ui ftv, pr (.1)

```
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)
                      begin with full model
p = 0.7048 >= 0.1000  removing ftv
p = 0.4577 >= 0.1000  removing age
p = 0.1403 >= 0.1000  removing ptl

Logistic regression                               Number of obs   =        189
                                                  LR chi2(6)      =      30.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -102.10831                       Pseudo R2       =     0.1298

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ui |   2.471868   1.106295     2.02   0.043     1.028174    5.942703
         lwt |   .9834068   .0066905    -2.46   0.014     .9703806    .9966078
    _Irace_2 |   3.760537   1.960996     2.54   0.011     1.353245    10.45017
    _Irace_3 |   2.524889   1.086685     2.15   0.031     1.086161    5.869353
       smoke |   2.817447    1.10602     2.64   0.008      1.30529    6.081411
          ht |   6.497492    4.48915     2.71   0.007     1.677437     25.1678
------------------------------------------------------------------------------
```

Exercícios

1. Verifique se a variável lwt (peso materno no início da gravidez) pode ser modelada como uma variável quantitativa, plotando os valores preditos pelo modelo com o peso materno.
2. O conjunto de dados lowbwt contém a informação para as amostras de 100 bebês com baixo peso ao nascer, nascidos em Boston, Massachusetts. A variável grmhem é uma variável aleatória dicotômica que indica se um bebê teve hemorragia da matriz germinal. O valor 1 indica que ocorreu hemorragia e 0 que não. Os escores Apgar de cinco minutos dos bebês estão salvos sob a variável apgar5 e os indicadores de toxemia – em que 1 representa o diagnóstico de toxemia durante a gravidez para as mães das crianças e 0 sem tal diagnóstico – sob a variável de nome tox.
   a) Usando a hemorragia da matriz germinal como resposta, ajuste o modelo de regressão logística, tendo como variável explanatória o escore Apgar. Interprete B1, o coeficiente estimado do escore Apgar.
   b) Se determinada criança tem escore Apgar de cinco minutos de 3, qual a probabilidade prevista de que ela tenha hemorragia no cérebro? Qual a probabilidade se o escore é 7?
   c) Ao nível de significância de 0,05, teste a hipótese nula de que o parâmetro β_1 da população é igual a 0. O que você conclui?
   d) Agora ajuste o modelo de regressão com a toxemia. Interprete β_1, o coeficiente estimado da toxemia.
   e) Para uma criança cuja mãe foi diagnosticada com toxemia durante a gravidez, qual a probabilidade prevista dela ter hemorragia da matriz germinal? Qual a probabilidade para uma criança cuja mãe não teve toxemia?
   f) Qual a chance estimada de haver uma hemorragia da matriz germinal em crianças cujas mães foram diagnosticadas com toxemia relativa às crianças cujas mães não foram diagnosticadas?
   g) Construa um intervalo de confiança de 95% para a razão de chances da população. Esse intervalo contém o valor 1? O que isto lhe diz?
   h) Verifique se o sexo e a idade gestacional estão associados com maior chance de hemorragia da matriz germinal.
   i) Rode um modelo ajustado completo com sexo, apgar5, toxemia e idade gestacional.
   j) Rode um modelo stepwise com as mesmas variáveis usando P<0,20 para inclusão e P<0,10 para retenção.
   k) Ao final, quais as suas conclusões? Que fatores estão associados com hemorragia cerebral?