      name: **<unnamed>**
       log: **C:\Users\USER\Downloads\heart disease\Git hub binary logistic regre**
> ssion stroke.smcl
  log type: **smcl**
 opened on: **13 Jan 2026, 12:04:16**

```
1 . use "C:\Users\USER\Downloads\heart disease\stroke.dta"
no; dataset in memory has changed since last saved
r(4);

2 .
3 . tab stroke
```

1 if the patient had a stroke or 0 if not	Freq.	Percent	Cum.
0	**4,861**	**95.13**	**95.13**
1	**249**	**4.87**	**100.00**
Total	**5,110**	**100.00**	

```
4 .
5 . tab smoking_status
```

"formerly smoked", "never smoked", "smokes" or "Unknown". "Unknown" means that t	Freq.	Percent	Cum.
Unknown	**1,544**	**30.22**	**30.22**
formerly smoked	**885**	**17.32**	**47.53**
never smoked	**1,892**	**37.03**	**84.56**
smokes	**789**	**15.44**	**100.00**
Total	**5,110**	**100.00**	

```
6 .
7 . **frequency of stroke by smoking status

8 .
9 . tab stroke smoking_status, row chi2
```

Key
frequency *row percentage*

1 if the patient had a stroke or 0 if not	"formerly smoked", "never smoked", "smokes" or "Unknown". "Unknown" means that t Unknown	formerl..	never s..	smokes	Total
0	**1,497**	**815**	**1,802**	**747**	**4,861**
	30.80	**16.77**	**37.07**	**15.37**	**100.00**
1	**47**	**70**	**90**	**42**	**249**
	18.88	**28.11**	**36.14**	**16.87**	**100.00**

Total	**1,544**	**885**	**1,892**	**789**	**5,110**
	30.22	**17.32**	**37.03**	**15.44**	**100.00**

Pearson chi2(**3**) = **29.1473** Pr = **0.000**

```
10 .
11 . log using "C:\Users\USER\Downloads\heart disease\smoking.smcl"
log file already open
r(604);

12 .
13 . use "C:\Users\USER\Downloads\heart disease\stroke.dta"
no; dataset in memory has changed since last saved
r(4);

14 .
15 . tab stroke
```

1 if the patient had a stroke or 0 if not	Freq.	Percent	Cum.
0	**4,861**	**95.13**	**95.13**
1	**249**	**4.87**	**100.00**
Total	**5,110**	**100.00**	

```
16 .
17 . tab smoking_status
```

"formerly smoked", "never smoked", "smokes" or "Unknown". "Unknown" means that t	Freq.	Percent	Cum.
Unknown	**1,544**	**30.22**	**30.22**
formerly smoked	**885**	**17.32**	**47.53**
never smoked	**1,892**	**37.03**	**84.56**
smokes	**789**	**15.44**	**100.00**
Total	**5,110**	**100.00**	

```
18 .
19 . **frequency of stroke by smoking status

20 .
21 . tab stroke smoking_status, row chi2
```

Key
frequency *row percentage*

1 if the patient had a stroke or 0 if not	"formerly smoked", "never smoked", "smokes" or "Unknown". "Unknown" means that t Unknown	formerl..	never s..	smokes	Total
0	**1,497**	**815**	**1,802**	**747**	**4,861**
	30.80	**16.77**	**37.07**	**15.37**	**100.00**
1	**47**	**70**	**90**	**42**	**249**
	18.88	**28.11**	**36.14**	**16.87**	**100.00**
Total	**1,544**	**885**	**1,892**	**789**	**5,110**
	30.22	**17.32**	**37.03**	**15.44**	**100.00**

Pearson chi2(**3**) = **29.1473** Pr = **0.000**

```
22 .
23 . *** the p-value is 0.000 thus it is statistically significant. 3 stands for th
   > e degrees of freedom. we therefore reject the null hypothesis that the variabl
   > es are independent or that the observed distribution matches the expected dist
   > ribution due to chance alone. there is thus a real significant difference betw
   > een the variables in this population.

24 .
25 . ****Bivariate logistic regression

26 .
27 . logit stroke smoking_status
   no observations
   r(2000);

28 .
29 . ****to ensure my variable is store as string, i'll use the 'describe' command.
   >  then i'll tabulate to see my values and frequencies.

30 .
31 . describe stroke
```

Variable name	Storage type	Display format	Value label	Variable label
stroke	byte	%8.0g		**1 if the patient had a stroke or 0 if not**

```
32 .
33 . ****stroke is note a string variable. it is a numeric variable as it's storage
   >  type is byte.

34 .
35 . tabulate stroke
```

1 if the patient had a stroke or 0 if not	Freq.	Percent	Cum.
0	**4,861**	**95.13**	**95.13**
1	**249**	**4.87**	**100.00**
Total	**5,110**	**100.00**	

```
36 .
37 . describe smoking_status

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
smoking_status  str15   %15s                 "formerly smoked", "never smoked",
                                               "smokes" or "Unknown". "Unknown"
                                               means that t

38 .
39 . ****smoking_status is thus a string variable as the storage type is str15.

40 .
41 . ***convert the string variable to a numeric labeled variable using the 'encode
> ' command.

42 .
43 . encode smoking_status, generate (smoke_num)
variable smoke_num already defined
r(110);

44 .
45 . ***run the logistic regression using the i.prefix for stroke and smoke_num var
> iables as they are categorical variables.

46 .
47 . logistic stroke i.smoke_num

Logistic regression                                     Number of obs =  5,110
                                                        LR chi2(3)    =  28.11
                                                        Prob > chi2   = 0.0000
Log likelihood = -981.13127                             Pseudo R2     = 0.0141

------------------------------------------------------------------------------
      stroke | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    smoke_num |
formerly s.. |   2.735674   .5294594     5.20   0.000     1.872077    3.997652
never smoked |   1.590786    .291642     2.53   0.011     1.110604     2.27858
      smokes |   1.790823   .3886243     2.69   0.007       1.1704    2.740128
             |
       _cons |   .0313961   .0046509   -23.36   0.000     .0234845    .0419731
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

48 .
49 . ****dependent variable stroke does not need the i.prefix.

50 .
51 . logit stroke ib3.smoke_num

Iteration 0:  Log likelihood =  -995.1863
Iteration 1:  Log likelihood = -981.98431
Iteration 2:  Log likelihood = -981.13353
Iteration 3:  Log likelihood = -981.13127
Iteration 4:  Log likelihood = -981.13127

Logistic regression                                     Number of obs =  5,110
                                                        LR chi2(3)    =  28.11
                                                        Prob > chi2   = 0.0000
Log likelihood = -981.13127                             Pseudo R2     = 0.0141
```

stroke	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
smoke_num						
Unknown	-.464228	.1833321	-2.53	0.011	-.8235523	-.1049038
formerly s..	.5421499	.1648598	3.29	0.001	.2190306	.8652692
smokes	.1184472	.1918704	0.62	0.537	-.2576119	.4945063
_cons	-2.996843	.1080095	-27.75	0.000	-3.208537	-2.785148

```
52 .
53 . describe smoke_num

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
smoke_num       long    %15.0g     smoke_num
                                              "formerly smoked", "never smoked",
                                                "smokes" or "Unknown". "Unknown"
                                                means that t

54 .
55 . tab smoke_num

      "formerly
smoked", "never
       smoked",
    "smokes" or
     "Unknown".
"Unknown" means
         that t |      Freq.     Percent        Cum.
----------------+-----------------------------------
        Unknown |      1,544       30.22       30.22
formerly smoked |        885       17.32       47.53
   never smoked |      1,892       37.03       84.56
         smokes |        789       15.44      100.00
----------------+-----------------------------------
          Total |      5,110      100.00

56 .
57 . **** the category for the never smoked individuals is 3. i want to use them as
   >  my reference category for the logistic regression showing the relationship be
   > tween stroke and smoking status.

58 .
59 . logit stroke ib3.smoke_num

Iteration 0:  Log likelihood =  -995.1863
Iteration 1:  Log likelihood = -981.98431
Iteration 2:  Log likelihood = -981.13353
Iteration 3:  Log likelihood = -981.13127
Iteration 4:  Log likelihood = -981.13127

Logistic regression                                     Number of obs =  5,110
                                                        LR chi2(3)    =  28.11
                                                        Prob > chi2   = 0.0000
Log likelihood = -981.13127                             Pseudo R2     = 0.0141
```

stroke	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
smoke_num						
Unknown	-.464228	.1833321	-2.53	0.011	-.8235523	-.1049038
formerly s..	.5421499	.1648598	3.29	0.001	.2190306	.8652692
smokes	.1184472	.1918704	0.62	0.537	-.2576119	.4945063
_cons	-2.996843	.1080095	-27.75	0.000	-3.208537	-2.785148

```
60 .
61 . ***to specify the odds ratio, i will use the 'or' command. Alternatively, i ca
   > n also use the logistic command

62 .
63 . logit stroke ib3.smoke_num, or

Iteration 0:  Log likelihood =  -995.1863
Iteration 1:  Log likelihood = -981.98431
Iteration 2:  Log likelihood = -981.13353
Iteration 3:  Log likelihood = -981.13127
Iteration 4:  Log likelihood = -981.13127

Logistic regression                                     Number of obs =  5,110
                                                        LR chi2(3)    =  28.11
                                                        Prob > chi2   = 0.0000
Log likelihood = -981.13127                             Pseudo R2     = 0.0141
```

stroke	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
smoke_num						
Unknown	.6286202	.1152462	-2.53	0.011	.4388699	.9004112
formerly s..	1.7197	.2835094	3.29	0.001	1.244869	2.375646
smokes	1.125747	.2159976	0.62	0.537	.7728951	1.639689
_cons	.0499445	.0053945	-27.75	0.000	.0404157	.06172

Note: **_cons** estimates baseline odds.

```
64 .
65 . logistic stroke ib3.smoke_num

Logistic regression                                     Number of obs =  5,110
                                                        LR chi2(3)    =  28.11
                                                        Prob > chi2   = 0.0000
Log likelihood = -981.13127                             Pseudo R2     = 0.0141
```

stroke	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
smoke_num						
Unknown	.6286202	.1152462	-2.53	0.011	.4388699	.9004112
formerly s..	1.7197	.2835094	3.29	0.001	1.244869	2.375646
smokes	1.125747	.2159976	0.62	0.537	.7728951	1.639689
_cons	.0499445	.0053945	-27.75	0.000	.0404157	.06172

Note: **_cons** estimates baseline odds.

```
66 .
67 . *** Odds Ratio (OR) = 1: The odds are the same in both groups thus there are n
   > o association. OR > 1: The odds are higher in the exposed group than in the re
   > ference group. OR < 1: The odds are lower in the exposed group than in the ref
   > erence group.

68 .
69 . **** With our reference group as never smoked, the odds of stroke are 1.72 tim
   > es as high in the formerly smoked compared to the never smoked and this repres
   > ents a 72% increase in the odds. the value is 0.001 thus this observation is n
   > ot due to chance.

70 .
71 . log using "C:\Users\USER\Downloads\heart disease\Git hub binary logistic regre
   > ssion stroke.smcl"
   log file already open
   r(604);

72 . save "C:\Users\USER\Downloads\heart disease\stroke.dta", replace
   file C:\Users\USER\Downloads\heart disease\stroke.dta saved

73 . exit
```