

Computer Workshop: Binary logistic regression

The aims of this workshop are:

- Fit and interpret logistic regression models.
- Calculate predicted probabilities according to the fitted model.
- Fit a logistic regression model with an interaction.
- Use likelihood ratio test to compare two nested models.

Download data, open Stata, and set up do file

- Download the Stata dataset **crime2013-14_binary.dta** to a suitable destination. Remember where you saved these files, as we will use this as our “Working Directory” for the rest of the workshop.
- Open Stata and a new do-file (we always recommend using a do-file so that you have a record of your code and can easily re-run the model).
- Set up the do-file by typing the following in the first few rows:

```
capture log close
```


- type the path to your working directory between the quotation marks, e.g.

```
cd "C:\statistics\binarylogit"
```

```
log using "NCRM_binary logit.log", text replace
```

```
use "crime2013-14_binary.dta", clear
```



Finally, click on the  icon in the toolbar to “do” all of the commands that you have typed into the do-file so far. Some output should then appear in the results window.

- Use *describe* to get a feel for the dataset.

In this workshop, we will study the association between a binary response variable and a set of predictors using Binary Logistic Regression. For doing so, we will use a dataset **crime2013-14_binary.dta**, extracted from the Crime Survey for England and Wales, 2013-2014¹. Our aim is to determine whether there is an association between the having been a victim of crime in the last 12 months (**bcsvictim**, 1=Yes, 0=No) and some socio-demographic characteristics of the respondent. The dataset includes the following variables:

¹ Office for National Statistics, University of Manchester. Cathie Marsh Institute for Social Research (CMIST). UK Data Service. (2016). *Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset*. [data collection]. UK Data Service. SN: 8011, <http://doi.org/10.5255/UKDA-SN-8011-1>

VARIABLE	DESCRIPTION
caseid	Case identifier (9 digits)
sex	Gender
agegrp7	Age grouped
educat3	Education
rural2	Type of area 2004: urban/rural
bcsvictim	Experience of any crime in the previous 12 months
tenure	Housing tenure

Producing Descriptive Tables

We will now find out how to construct tables to display categorical data using Stata. First we will start by displaying the frequencies of the variables of interest.

- `ssc install fre // this will install a useful user-written tabulating function to your Stata`
- `fre sex-tenure // show frequencies for all variables in the variable list between and including sex and tenure`

Now, let's study the relationship between the response variable and each one of the potential predictors by producing some cross tabulations and chi-square tests of independence. Option 'row' calculates the table percentages by rows, and 'col' by columns.

- `tab sex bcsvictim, chi2 row`
- `tab agegrp7 bcsvictim, chi2 row`
- `tab educat3 bcsvictim, chi2 row`
- `tab rural2 bcsvictim, chi2 row`

Take a look at the crosstables you have produced and think what they tell you about the potential associations between levels of crime and the explanatory variables.

Fitting Logistic Regression with a Single predictor variable

Now we will carry out a logistic regression analysis. Let's start by modelling the probability of having been a victim of crime, **bcsvictim**, using place of residence, **rural2**, as the only predictor:

- `logit bcsvictim i.rural2`

NOTE: It is not strictly necessary to identify a variable with only two categories as a categorical variable (using the i. notation) in Stata if it is coded as 0, 1. Variables with more than two categories need to be identified as such.

The output from the logistic regression analysis should now appear in the results window:

```
. logit bcsvictim i.rural2
```

```
Iteration 0:    log likelihood =  -3819.306
Iteration 1:    log likelihood = -3803.5848
Iteration 2:    log likelihood = -3803.495
Iteration 3:    log likelihood = -3803.495
```

```
Logistic regression               Number of obs   =      8,802
                                LR chi2(1)        =      31.62
                                Prob > chi2        =      0.0000
Log likelihood = -3803.495        Pseudo R2      =      0.0041
```

bcsvictim	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
rural2					
Rural	-.410481	.0753201	-5.45	0.000	-.5581056 - .2628564
_cons	-1.597351	.0325917	-49.01	0.000	-1.66123 -1.533472

The parameter estimates are included in the column **B**. The equation of this model is:

$$\log\left(\frac{p}{1-p}\right) = -1.597 - 0.410 * rural2$$

Interpretation of the coefficients. Logit scale.

The coefficients of this model have several interpretations. In the simplest way, analogous to what you do in linear regression, you can see -1.597 as the expected log-odds for the probability of having been a victim of crime in an urban area and -0.410 as the expected decrease in this log-odds given that the respondent lives in a rural area. However, the logit scale is not intuitive and this interpretation does not help us much in understanding the association between these two variables.

Interpretation of the coefficients. Odds scale.

Instead of interpreting the log odds, you can exponentiate both sides of the equation to obtain the odds ratios:

$$\frac{p}{1-p} = e^{-1.597} e^{-0.41 \times rural2}$$

The quantity on the left hand side, $p/(1-p)$ is the odds of **Have** versus **Have not** been a victim of crime. Stata gives you the exponentiated coefficients if you use the command below, so you don't need to calculate them by hand.

➤ `logit bcsvictim i.rural2, or`

```
. logit bcsvictim i.rural2, or
```

```
Iteration 0:    log likelihood =  -3819.306
Iteration 1:    log likelihood = -3803.5848
Iteration 2:    log likelihood =  -3803.495
Iteration 3:    log likelihood =  -3803.495
```

```
Logistic regression                Number of obs    =      8,802
                                   LR chi2(1)         =      31.62
                                   Prob > chi2         =      0.0000
Log likelihood =  -3803.495         Pseudo R2      =      0.0041
```

bcsvictim	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
rural2						
Rural	.6633311	.0499621	-5.45	0.000	.5722922	.7688523
_cons	.202432	.0065976	-49.01	0.000	.1899053	.2157851

The coefficient $e^{-0.41} = 0.663$ is an estimate of the odds ratio (OR) for rural and urban areas. For these data, it tells us that the odds of having been a victim of crime in rural areas are 0.663 times the odds in urban areas. We can also calculate the % change of the odds using $(0.663-1)*100 = -33.7\%$. Therefore, the odds of being a victim of crime are 34% lower in rural than in urban areas. This is one of the most used interpretations for the coefficients of a logistic regression, but be careful when writing down your conclusions about the parameters: we are not saying that the probability of having been a victim is 34% lower in rural than in urban areas!

Standard errors of the coefficients and Confidence Intervals

The standard errors of the estimated coefficients are given in the column **Std. Err.** Stata gave us a 95% confidence interval for the exponentiated coefficients in the last two columns of the table. You could also calculate the 95% confidence interval for the coefficient **B**, using $(B \pm 1.96 * S.E)$ and obtain an approximated confidence interval by exponentiating both lower and upper limits.

Wald chi-square test for the significance of the coefficients

A Wald Chi-Squared test for the significance of the coefficients is given in the columns: **z** (the value of the statistic) and **p>|z|** (the p-value of for the two tail test $H_0: B$ is equal to 0 versus $H_a: B$ is different from 0). If the p-value is less than 0.05, we can conclude that the corresponding coefficient is different from zero in the population at 5% significance level. For one tail tests ($B > 0$ or $B < 0$ is the alternative hypothesis) you should divide p-value by 2. The overall variable significance for variables with more than two levels can be assessed using the *testparm <variable>* command (joint Wald-test).

Predicted probabilities.

Instead of interpreting the coefficients of the regression, you can use the predicted probabilities under the model to communicate your findings. Note that for the model:

$$\log\left(\frac{p}{1-p}\right) = B_0 + B_1 \times Rural2$$

The probability p can be predicted using

$$p = \frac{e^{B_0 + B_1 \times \text{Rural2}}}{1 + e^{B_0 + B_1 \times \text{Rural2}}}$$

Using our fitted model, for urban areas we have:

$$p_{\text{urban}} = \frac{e^{-1.597}}{1 + e^{-1.597}} = 0.168$$

and, as **rural2** is dichotomous, for rural areas we have:

$$p_{\text{rural}} = \frac{e^{B_0 + B_1}}{1 + e^{B_0 + B_1}} = \frac{e^{-1.597 - 0.41}}{1 + e^{-1.597 - 0.41}} = 0.118$$

Therefore, under this model, the predicted probability of having been a victim of crime is 0.17 in urban areas and 0.12 in rural areas.

Alternatively to calculating the predicted probabilities by hand, you can ask Stata to calculate them for you, using the `margins` command:

➤ `margins rural2`

Logistic Regression with several predictors

Now we will fit a Logistic Regression model to explain the probability of having been a victim of crime (**bcsvictim**) using as predictors: gender (**sex**), age (**agegrp7**), place of residence (**rural2**), and housing tenure (**tenure**).

- `logit bcsvictim i.sex i.agegrp7 i.rural2 i.tenure`
- `logit bcsvictim i.sex i.agegrp7 i.rural2 i.tenure, or`
- `testparm i.sex`
- `testparm i.agegrp7`
- `testparm i.rural2`
- `testparm i.tenure`

Please note the variable gender has a p-value of 0.210, which is larger than 0.05 and therefore it is not statistically significant at the 5% level. After gender is removed from the model, all the other variables were statistically significant at the 5% level (i.e. all p-values were smaller than 0.05).

Note! Please remember that you will have to check whether the dummy variables are significant as a whole. There might be individual categories that are not significantly different from the reference category, but it does not necessarily mean that the entire variable should be removed from the model.

The final model is shown below.

bcsvictim	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agegrp7						
25-34	-.083254	.1143508	-0.73	0.467	-.3073775	.1408695
35-44	-.402513	.1190564	-3.38	0.001	-.6358593	-.1691667
45-54	-.3029218	.1183253	-2.56	0.010	-.5348352	-.0710085
55-64	-.4986378	.1265097	-3.94	0.000	-.7465923	-.2506833
65-74	-.9284313	.1436936	-6.46	0.000	-1.210066	-.6467969
75+	-1.382719	.1655037	-8.35	0.000	-1.7071	-1.058337
rural2						
Rural	-.2867935	.0769305	-3.73	0.000	-.4375744	-.1360125
tenure						
Buying with mortgage	.2489933	.0933833	2.67	0.008	.0659654	.4320212
Rent	.3115792	.0896317	3.48	0.001	.1359043	.4872541
Live for free	.036903	.183952	0.20	0.841	-.3236362	.3974422
_cons	-1.362276	.1236575	-11.02	0.000	-1.60464	-1.119912

An example of model interpretation using odds: The estimated OR for being a victim of crime for those in the oldest age group compared to the youngest (which is the reference group) is 0.251. It means that the odds of being a victim of crime among those aged 75+ are 0.25 times the odds of those aged 16-24, when other variables in the model are controlled for. In other words, the odds of being a victim of crime in the oldest age group were 75% lower than in the youngest age group.

An example of calculating predicted probabilities: Imagine we wanted to calculate the predicted probability of having been a victim of crime for the following scenarios:

- A respondent aged 20 living in an urban area and owning his/her home outright.
- A respondent aged 30 living in a rural area and renting her/his home.

You could do this by using the *margins* command in Stata as shown below. The tables show that the probability on the first scenario is 0.20 (or 20%) and in the second scenario 0.19 (or 19%).

. margins, at(age=1 rural2 = 1 tenure = 1)						
Adjusted predictions			Number of obs		= 8,802	
Model VCE : OIM						
Expression : Pr(bcsvictim), predict()						
at	: agegrp7	=	1			
	: rural2	=	1			
	: tenure	=	1			
	Delta-method					
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]	
_cons	.2038707	.0200705	10.16	0.000	.1645332	.2432082

```
. margins, at(age = 2 rural2 = 2 tenure = 3)
```

```
Adjusted predictions      Number of obs      =      8,802
Model VCE      : OIM
```

```
Expression      : Pr(bcsvictim), predict()
at              : agegrp7      =      2
                  rural2      =      2
                  tenure      =      3
```

	Delta-method				
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]
_cons	.194545	.0154699	12.58	0.000	.1642245 .2248655

If you wanted to see a more general picture of the results using predicted probabilities, you could choose to focus on keeping some of the variables constant while varying the variables of interest. For instance, we might be interested in the effect of housing tenure on the probability of having been a victim of crime. We might choose to hold the values of all other variables ‘as observed’, meaning that Stata calculates the predicted probabilities by keeping the values of age and type of area as they were observed for each respondent, but varies their housing tenure status. The mean of these probabilities becomes the predicted probability for having been a victim of crime given a specific housing tenure status². The code and resulting table is shown below.

```
. /* Predicted probabilities by housing tenure, other variables as observed */
. margins tenure, asobs
```

```
Predictive margins      Number of obs      =      8,802
Model VCE      : OIM
```

```
Expression      : Pr(bcsvictim), predict()
```

	Delta-method				
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]
tenure					
Own outright	.1322728	.0078813	16.78	0.000	.1168256 .1477199
Buying with mortgage	.1628939	.0072158	22.57	0.000	.1487513 .1770365
Rent	.1714159	.006876	24.93	0.000	.1579393 .1848926
Live for free	.1364883	.0192462	7.09	0.000	.0987663 .1742102

Those who own outright or live for free have the lowest probability of being victims of crime (around 13%). The highest risk is among those who are renting (17%) leaving those, who are buying with mortgage somewhere in between (16%).

Fitting a model with an interaction term

Having fitted the model with all predictors, we wonder if there is an interaction between *rural2* and *tenure*, i.e. if the relationship between having been a victim of crime is different between respondents with different housing tenures in the rural and urban areas. In order to include an interaction, in Stata you do not need to calculate a new variable

² See more information here: <https://www.stata-journal.com/article.html?article=st0260>. Stata output shown below.

because the software does it for you. Remember that a single “#” refers only to the interaction term (so in the code below we are including both terms in the model, but testing only the interaction term), and “##” includes both the main effects and the interaction term in the model.

You could run the model with interaction by typing:

```
➤ logit bcsvictim i.ageg i.tenure##b2.rural
➤ logit bcsvictim i.ageg i.tenure##b2.rural,or
➤ testparm i.tenure#i.rural
```

The last command above conducts a joint Wald-test testing whether the interaction effect as a whole is significant in the model. Note that here we only use # meaning that we’re only testing the interaction, not the main effects.

The resulting table is shown below:

```
. logit bcsvictim i.ageg i.tenure##b2.rural,or
```

```
Iteration 0:  log likelihood = -3819.306
Iteration 1:  log likelihood = -3694.0262
Iteration 2:  log likelihood = -3686.4631
Iteration 3:  log likelihood = -3686.4435
Iteration 4:  log likelihood = -3686.4435
```

```
.logistic regression      Number of obs   =      8,802
                        LR chi2(13)         =      265.73
                        Prob > chi2         =      0.0000
.log likelihood = -3686.4435  Pseudo R2      =      0.0348
```

	bcsvictim	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
agegrp7							
	25-34	.9243488	.1058255	-0.69	0.492	.7385582	1.156877
	35-44	.6687655	.0797137	-3.38	0.001	.5294374	.8447596
	45-54	.7398933	.0877473	-2.54	0.011	.5864369	.9335057
	55-64	.6077001	.0770391	-3.93	0.000	.4740031	.7791077
	65-74	.3939181	.0567114	-6.47	0.000	.2970712	.5223377
	75+	.2493323	.0413768	-8.37	0.000	.1801031	.3451723
tenure							
	Buying with mortgage	1.617856	.2850028	2.73	0.006	1.14549	2.285011
	Rent	1.809863	.323751	3.32	0.001	1.274622	2.569863
	Live for free	.5653123	.3034854	-1.06	0.288	.1973898	1.61902
rural2							
	Urban	1.636998	.2354918	3.43	0.001	1.234803	2.170193
tenure#rural2							
	Buying with mortgage#Urban	.7364023	.1418696	-1.59	0.112	.5048114	1.074239
	Rent#Urban	.6987747	.1367715	-1.83	0.067	.476136	1.025518
	Live for free#Urban	1.954032	1.103856	1.19	0.236	.6457695	5.912698
	_cons	.1649477	.0271662	-10.94	0.000	.119442	.2277905

The joint Wald-test shows that the p-value is 0.088, which means that the interaction is not significant at the 5% level, but it is significant at 10% level, because $0.05 < 0.088 < 0.1$.

Predicted probabilities can be quite useful in interpreting interaction effects. An interaction between two variables means that the association of one variable with the outcome depends on the values of another variable. Thus, if we want to say something about the interaction between these variables, we have to find how the association between tenure and the outcome differs in rural and urban areas.

By typing:

```
➤ margins tenure#rural, at (age=2)
```

you will get the predicted probabilities for all combinations of housing tenure and urban/rural area of residence for those in age category 25-34, as shown below.

```
. margins tenure#rural2, at(age = 2)
```

```
Adjusted predictions          Number of obs      =      8,802
```

```
Model VCE      : OIM
```

```
Expression   : Pr(bcsvictim), predict()
```

```
at           : agegrp7      =      2
```

	Delta-method					[95% Conf. Interval]
	Margin	Std. Err.	z	P> z		
tenure#rural2						
Own outright#Urban	.1997386	.0174286	11.46	0.000	.1655792	.2338981
Own outright#Rural	.1322979	.0173444	7.63	0.000	.0983035	.1662923
Buying with mortgage#Urban	.229205	.0144091	15.91	0.000	.2009637	.2574463
Buying with mortgage#Rural	.1978652	.0207853	9.52	0.000	.1571267	.2386037
Rent#Urban	.2399225	.0133519	17.97	0.000	.2137533	.2660917
Rent#Rural	.2162693	.0228904	9.45	0.000	.1714048	.2611337
Live for free#Urban	.2161219	.0314928	6.86	0.000	.1543972	.2778467
Live for free#Rural	.0793531	.0382709	2.07	0.038	.0043435	.1543627

In both areas the safest tenure type is owning outright, followed by living for free, buying with mortgage and renting. However, in urban areas the differences are quite small, whereas in rural areas particularly living for free (and to a lesser extent owning outright) are very clearly safer options than renting or buying with mortgage. This could be because the types of crimes typically committed in urban and rural areas are quite different or the type of people living under certain tenure conditions are different in urban and rural areas.

Using likelihood ratio test to test the significance of the interaction term

You may remember that likelihood ratio test can be used to test which of two nested models provides a better fit for our data. Let's test the significance of our interaction effect with a likelihood ratio test.

First we need to run a model without the interaction, but including all the same variables otherwise and save the results using 'estimates store' command:

```
➤ logit bcsvictim i.ageg i.tenure b2.rural,or
➤ estimates store no_interaction
```

Then we run the model with our interaction effect and similarly store the estimates:

- `logit bcsvictim i.ageg i.tenure##b2.rural,or`
- `estimates store with_interaction`

Finally, we can test the two models:

- `lrtest no_interaction with_interaction`

The likelihood ratio test (shown below) provides us a p-value of 0.0677, meaning that the interaction effect was not significant if we are using the conventional $p < 0.05$ threshold, but it is very close.

. lrtest no_interaction with_interaction

```
Likelihood-ratio test          LR chi2(3)  =      7.13
(Assumption: no_interaction nested in with_interac~n) Prob > chi2 =    0.0677
```