

# Logistic Regression - Exercise 1

## Solutions

### 1. Descriptive statistics

Rather than go through all of the descriptive statistics and unconditional associations I will pick three variables (-sepsis-, -age- and -attd-) and deal with them.

```
-----
sepsis                                     sepsis (0=no,1=yes)
-----
      type:  numeric (byte)
      label:  ynlbl

      range:  [0,1]                units:  1
unique values: 2                    missing .: 0/254

      tabulation:  Freq.  Numeric  Label
                   181      0      no
                   73      1      yes
```

Things that I check for are:

- missing data - no observations missing
- minimum and maximum - 0 and 1, as expected
- distribution of values - 73 septic calves = a reasonable sample size

```
-----
attd                                       attitude (0=bright, 1=depr.,2=comatose)
-----
      type:  numeric (byte)
      label:  attdlbl

      range:  [0,2]                units:  1
unique values: 3                    missing .: 6/254

      tabulation:  Freq.  Numeric  Label
                   28      0      bright
                   174     1      depressed
                   46      2      comatose
                   6       .
```

Things that I check for are:

- missing data - 6 observations missing
- minimum and maximum - 0 and 2, as expected
- distribution of values - only 28 calves that were "0" = bright and alert This is a fairly small category so the estimate of the effect of this category will be relatively imprecise.

```
-----
age                                       age at admission (in days)
-----
      type:  numeric (byte)

      range:  [1,28]              units:  1
unique values: 27                  missing .: 1/254

      mean:  9.36364
```

```

std. dev: 6.19477
percentiles: 10% 25% 50% 75% 90%
              2   5   8   13  18

```

Things that I check for are:

- missing data - one observations missing (not a serious problem)
- minimum and maximum - 1 and 28, both values seem reasonable
- distribution of values - mean = 9.4 days, median = 8 days, percentiles suggest a reasonable (in fact approximately normal) distribution

A good way to summarize quantitative and dichotomous variables is to request just key summary statistics as shown below:

```
. tabstat age sex dehy eye pulse-sepsis, statistics(n mean min max sd)
```

stats	age	sex	dehy	eye	pulse	resp	temp	umb
N	253	252	240	236	245	235	247	243
mean	9.363636	.5198413	7.033333	.0423729	115.8939	38.79149	38.05745	.2263374
min	1	0	0	0	30	12	32	0
max	28	1	15	1	200	120	42	1
sd	6.19477	.5006004	3.683466	.2018665	29.53279	19.72057	1.716473	.4193237

  

stats	sepsis
N	254
mean	.2874016
min	0
max	1
sd	.4534439

These summaries remind us that at least for some variables there are missing values so we need to be careful of this building multivariable models.

### 3. Unconditional associations

As above, I will select a couple of predictors (-attd- and -age-) to evaluate. Evaluating the rest of the predictors will be left up to you (note: conditions for chi2 test should be checked)

```
. tabulate attd sepsis, chi2 row expected
attitude |
(0=bright, |
1=depr., |
2=comatose | sepsis (0=no, 1=yes)
) | no yes | Total
-----+-----+-----
bright | 27 1 | 28
| 19.9 8.1 | 28.0
| 96.43 3.57 | 100.00
-----+-----+-----
depressed | 125 49 | 174
| 123.5 50.5 | 174.0
| 71.84 28.16 | 100.00
-----+-----+-----
```

comatose		24	22		46
		32.6	13.4		46.0
		52.17	47.83		100.00
-----+-----+-----					
Total		176	72		248
		176.0	72.0		248.0
		70.97	29.03		100.00

Pearson chi2(2) = 16.7596 Pr = 0.000

-attd- has a statistically significant association with sepsis ( $P < 0.001$ ) and there is clear evidence of a trend. As attitude deteriorates (bright  $\rightarrow$  depressed  $\rightarrow$  comatose), the probability of sepsis goes up.

. ttest age, by(sepsis) unequal // (note: the default for two-sample t-tests should be unequal variances which is not the case in Stata, you would need to type "unequal")

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
no	180	9.938889	.4499234	6.036355	9.051053	10.82673
yes	73	7.945205	.7481307	6.392032	6.453834	9.436577
-----+-----						
combined	253	9.363636	.3894619	6.19477	8.596621	10.13065
-----+-----						
diff		1.993683	.8730009		.2661495	3.721217
-----+-----						
diff = mean(no) - mean(yes)				t =	2.2837	
Ho: diff = 0				Satterthwaite's degrees of freedom = 126.827		
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) = 0.9880		Pr( T  >  t ) = 0.0241		Pr(T > t) = 0.0120		

Septic calves were generally younger (7.9 days) than non-septic calves (9.9 days) and this difference is moderately statistically significant ( $p=0.02$ ).

### 3. Identify all variables with a significant ( $p \leq 0.1$ ) association with sepsis.

The unconditional association p-values (and potential functional forms for relationships between continuous predictors and sepsis - determined in Exercise #2, question 2) are as follows:

Categorical		Continuous		
Variable	p-value	Variable	p-value	shape?
breed	0.36	age	0.02	quad?
sex	0.69	dehy	0.07	linear
attd	<0.001	pulse	0.98	NA
eye	0.004	resp	0.004	quad?
jnts	0.01	temp	0.04	quad?
post	<0.001			
umb	0.001			

#### 4. Logistic model with -post- and -umb-

Before fitting the model, we will check if the baseline value of -post- that will be used as a default (ie. the lowest level, post = 0) is a reasonable choice.

```
-----
post                                     posture (0=standing, 1=sternal, =lateral)
-----
      type: numeric (byte)
      label: postlbl

      range: [0,2]                               units: 1
unique values: 3                               missing .. 6/254

      tabulation:  Freq.   Numeric   Label
                   93         0   standing
                   86         1   sternal
                   69         2   lateral
                   6         .
```

There are 93 observations in which post=0 and no biologic reason why we should choose a different level, so we will leave it as it is.

```
. logit sepsis i.post umb
```

```
Logistic regression                               Number of obs =          241
                                                LR chi2(3)         =          36.44
                                                Prob > chi2        =          0.0000
Log likelihood = -128.7436                       Pseudo R2         =          0.1240
```

```
-----
      sepsis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      post |
sternal |  1.627315   .4115099     3.95  0.000     .8207709     2.43386
lateral |  1.889788   .4282958     4.41  0.000     1.050344     2.729232
      |
      umb |  1.090339   .3442872     3.17  0.002     .4155487     1.76513
      _cons | -2.383711   .3639175    -6.55  0.000    -3.096976    -1.670446
-----
```

It appears that both posture and swollen umbilicus have significant associations with being septic. We will formally test the significance of -post- using a multiple Wald test (with 2 df).

```
. testparm i.post
```

```
( 1) [sepsis]1.post = 0
( 2) [sepsis]2.post = 0

      chi2( 2) =    21.29
      Prob > chi2 =    0.0000
```

-post- is a highly significant predictor of -sepsis-. We could also have used a likelihood ratio test, but since there is no doubt about the statistical significance of -post-, the Wald test is adequate.

**(a) explain the relationship between having a swollen umbilicus and the risk of being septic.**

First we note that having a swollen umbilicus increases the risk (to be technically correct we should say that it increases the odds) that a calf is septic by  $e^{1.09} \approx 3$  times.

**(b) explain the relationship between posture and the risk of being septic.**

Being in sternal recumbency increases the risk (odds) of sepsis by  $e^{1.63} \approx 5.1$  times and being in lateral recumbency increases it by  $e^{1.89} \approx 6.6$  times.

**(c) predicted probability of sepsis change as posture changes**

We could calculate the probabilities associated with the postural positions (as shown below in 5, but instead we will get the computer to calculate these for us (also shown in 5). Since -umb- also has an effect we will give the probabilities separately for those with and without a swollen umbilicus.

```
. table umb post, contents(mean pred_prob) /* table of predicted prob's */
-----
swollen |
umbilicus |      posture (0=standing,
(0=no,    |      1=sternal, 2=lateral)
1=yes)    | standing  sternal  lateral
-----+-----
      no | .0844233  .3194294  .3789698
      yes | .2152826  .5827186  .6448359
-----
```

Note that the effect of -post- is constant on the odds scale, but not on the probability scale. For example, going from standing to sternal increases the odds of sepsis by 5.1 times (as noted above). In a calf without a swollen umbilicus, being in sternal recumbency increases the probability of infection  $0.32/0.08 = 4$  times. In a calf with a swollen umbilicus, the effect of sternal recumbency is  $0.58/0.22 = 2.6$  fold increase in probability.

**5. What is the probability of sepsis in calf #1294 ?**

calf 1294 had post=0 and umb = 0 so:

$$\ln \frac{p}{1-p} = \beta_0 = -2.38; \frac{p}{1-p} = e^{-2.38} \rightarrow p = 0.08$$

Of course, you can get the computer to compute the predicted probability for you:

```
. list case sepsis post umb pred_p~b
-----+-----
13. | 1294      no  standing   no  .0844233 |
-----+-----
```