Logistic Regression - Exercise 3
Solutions

1. Assessment of overall fit

First, fit the model

```
. logit sepsis i.post umb c.age_ct##c.age_ct

Iteration 0:   log likelihood = -146.60743
Iteration 1:   log likelihood = -124.07141
Iteration 2:   log likelihood = -123.17905
Iteration 3:   log likelihood = -123.17402
Iteration 4:   log likelihood = -123.17402

Logistic regression                                     Number of obs   =        240
                                                        LR chi2(5)      =      46.87
                                                        Prob > chi2     =     0.0000
Log likelihood = -123.17402                             Pseudo R2       =     0.1598

------------------------------------------------------------------------------
       sepsis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         post |
      sternal |   1.594338   .4175655     3.82   0.000     .7759249    2.412752
      lateral |   1.838978   .4345141     4.23   0.000     .9873464    2.690611
              |
          umb |   1.078453   .3577357     3.01   0.003     .3773043    1.779602
       age_ct |  -.0823109   .0289085    -2.85   0.004    -.1389705   -.0256513
              |
c.age_ct#c.age_ct |   .0086873   .0033643     2.58   0.010     .0020934    .0152812
              |
        _cons |  -2.768343   .4018875    -6.89   0.000    -3.556028   -1.980658
------------------------------------------------------------------------------
```

The Pearson's chi-square goodness of fit test is highly significant which might appear to say that the model does not fit the data very well at all. This is not surprising since there are on average 3 observations per covariate pattern as there are 95 covariate patterns created by this model with 240 cases---this is too small for the chi-square based tests to be valid as the data are essentially binary. This suggests that there might be many covariate patterns with only one record and they would have a large influence on the Chi2 statistic.

```
. estat gof /*Pearson X2 statistic*/

Logistic model for sepsis, goodness-of-fit test

       number of observations =       240
 number of covariate patterns =        95
             Pearson chi2(89) =       114.86
                  Prob > chi2 =         0.0339
```

Also you could carry out a Hosmer-Lemeshow goodness-of-fit test with 10 groups in order to increase the number of observation per group.

```
 estat gof, g(10) table /*Hosmer-Lemeshow*/

Logistic model for sepsis, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0580 |     2 |   1.2 |    22 |  22.8 |    24 |
  |     2 | 0.0946 |     1 |   1.9 |    23 |  22.1 |    24 |
  |     3 | 0.1745 |     4 |   4.2 |    25 |  24.8 |    29 |
  |     4 | 0.2101 |     3 |   3.8 |    16 |  15.2 |    19 |
  |     5 | 0.2604 |     8 |   5.7 |    16 |  18.3 |    24 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.3202 |     6 |   7.0 |    18 |  17.0 |    24 |
  |     7 | 0.3670 |     8 |   9.0 |    18 |  17.0 |    26 |
  |     8 | 0.4761 |     8 |  10.2 |    15 |  12.8 |    23 |
  |     9 | 0.5708 |    17 |  14.5 |    10 |  12.5 |    27 |
  |    10 | 0.8311 |    15 |  14.5 |     5 |   5.5 |    20 |
  +--------------------------------------------------------+

      number of observations =       240
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =         4.63
               Prob > chi2 =         0.7962
```

Since the observed and expected values are close, the model appears to fit well over the range of predicted probability values. Overall, there is no evidence of "lack of fit" because the value is much larger than 0.05.

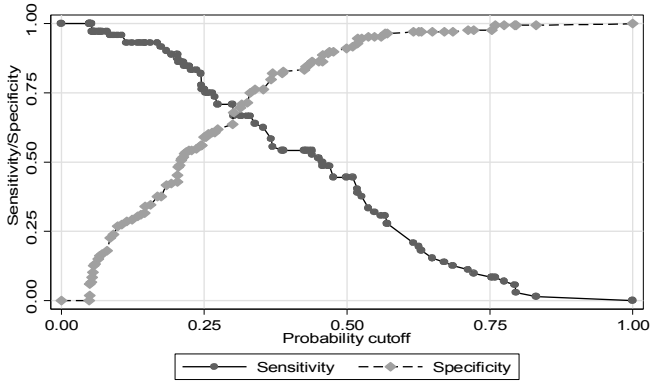
Next we will evaluate the predictive ability (sensitivity and specificity) of the model.

```
. estat class

Logistic model for sepsis

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        32             14  |         46
     -     |        40            154  |        194
-----------+--------------------------+-----------
   Total   |        72            168  |        240

Classified + if predicted Pr(D) >= .5
True D defined as sepsis != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   44.44%
Specificity                     Pr( -|~D)   91.67%
Positive predictive value       Pr( D| +)   69.57%
Negative predictive value       Pr(~D| -)   79.38%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    8.33%
False - rate for true D         Pr( -| D)   55.56%
```

```
False + rate for classified +   Pr(~D| +)   30.43%
False - rate for classified -   Pr( D| -)   20.62%
--------------------------------------------------
Correctly classified                        77.50%
--------------------------------------------------
```

The sensitivity of the model is quite low (44%) while the specificity is quite high (92%). However, both of these values depend on the cutpoint chosen to classify calves as septicemic or not. The values shown are based on the default value of 0.5, and as we shall see later (Q3) this can be manipulated depending on the prevalence of the outcome in the study group and the seriousness of making a false prediction.

Next we will generate a graph of the sensitivity and specificity against possible cutpoints. The sensitivity and specificity are equal (both approximately 70%) at a cutpoint of about 0.3. This suggests that the model has moderate predictive ability. However, if you were using this model in a clinical setting, your choice of cutpoint would depend on how serious you considered false positive results to be compared to false negative ones (see discussion in Q3).

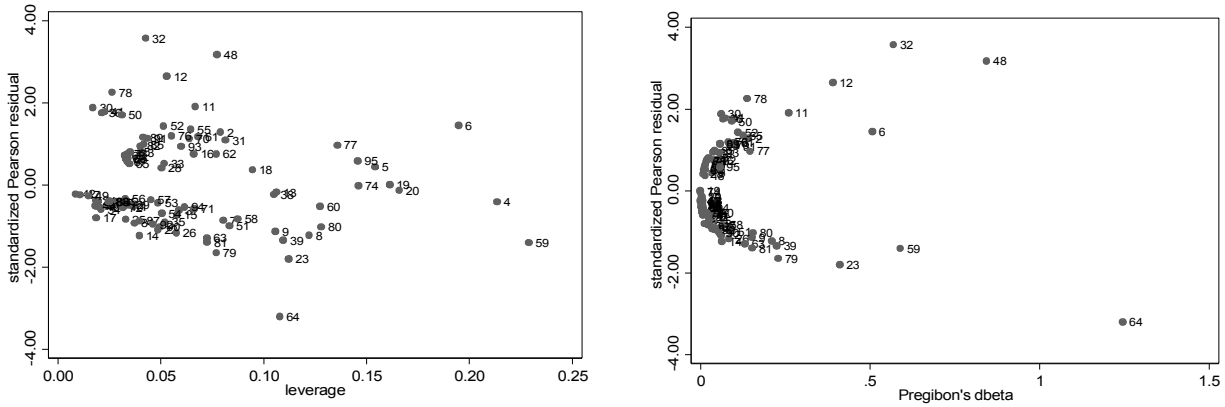


2. Outliers and influential observations

First we will compute the various diagnostic values (residuals etc.) and generate some summary statistics and a couple of graphs to identify outliers and influential observations. To do this, we will reduce the dataset to 1 record per covariate pattern.

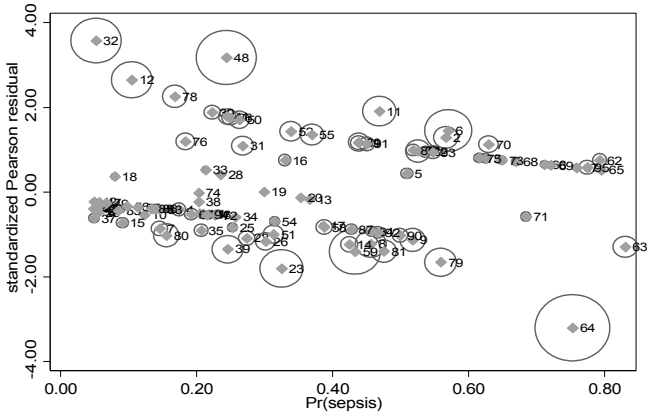
```
. sum num_obs pv rst lev db if withincov==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     num_obs |        95    2.526316    2.082562          1         10
          pv |        95    .3374078    .2166876    .049139   .8311161
         rst |        95       .04099    1.143371  -3.208725   3.571799
         lev |        95    .0631579    .0459353   .0086876   .2290239
          db |        95    .0951538    .1839791   1.22e-06   1.245695
```

There were 95 covariate patterns with and average of 3 calves per pattern (range was 1 to 10). There were some standardised residuals outside of the range of -3 to +3, and we should investigate these for errors or other explanations for their poor fit.



The graph of standardised residuals vs leverage identifies 2 covariate patterns (32 and 48) with large positive residuals and one (64) with a large negative residual. However, none of these had particularly high leverage values. On the other hand, 3 covariate patterns (4, 6 and 59) stand out as having high leverage.

The graph of standardised residuals vs delta-betas identifies 2 covariate patterns (48 and 64) as being quite influential, and they also had large residuals (despite their influence, the model still did not fit those points very well).



This graphs shows the cov. patterns 12, 32, 48 y 64 with high standardized Pearson residuals and high delta beta (size of the bubble). Remember that we are expecting to see large delta beta values at predicted probabilities (Pr(sepsis) in the range of 0.1-0.3 and 0.7-0.9 even if they are not influential. So , cov. pattern 48 is the most influential, but these cov. patterns are worth exploring. Lets look at what types of calves were in each of those covariate patterns.

	cov	num_obs	age	post	umb	avg_se~s	pv	rst	lev	dx2	db
13.	59	2	26	sternal	no	0.00	0.43	-1.41	0.23	1.986182	.5900095

100.	4	8	2	standing	no	0.13	0.17	-0.42	0.21	.1732255	.0470648
219.	6	6	2	lateral	no	0.83	0.57	1.45	0.19	2.096629	.5074989
228.	78	1	9	standing	yes	1.00	0.17	2.25	0.03	5.08206	.1381744
229.	12	4	5	standing	no	0.50	0.11	2.64	0.05	6.98416	.3911124
233.	**48**	**3**	**20**	**sternal**	**no**	**1.00**	**0.24**	**3.17**	**0.08**	**10.05072**	**.8441582**
239.	**32**	**5**	**12**	**standing**	**no**	**0.40**	**0.05**	**3.57**	**0.04**	**12.75775**	**.5694717**

Covariate patterns 32, 48 and 59 stand out as having had a large influence on the model. Neither of these patterns contained a very large number of calves (ranged from 2 to 5), so their influence has derived from the fact that they have high residuals, not because they represented a large number of cases. Neither group had particularly high leverage, or their influence would have been greater. We will now refit the model, leaving each of these covariate patterns out (sequentially), to see what effect they have on the parameter estimates.

Variable	final	no48	no32	no59
post				
sternal	1.5943382	1.4733121	1.838132	1.6534685
	.4175655	.42444626	.45133212	.42025725
lateral	1.8389785	1.841746	2.0903666	1.8597241
	.43451415	.43797718	.46810051	.43629705
umb	1.0784534	1.1658411	1.1709217	1.057926
	.35773568	.36228683	.36938763	.36004641
age_ct	-.08231092	-.09908383	-.09132431	-.07794532
	.02890848	.03008591	.02981698	.02889291
c.age_ct#				
c.age_ct	.0086873	.00842812	.00986948	.0104358
	.00336432	.0034537	.00348273	.00357614
_cons	-2.7683427	-2.8140512	-3.0959294	-2.8344969
	.4018875	.40678358	.44786843	.40997449
N	240	237	235	238

legend: b/se

The biggest effect of leaving covariate pattern #48 out is that the linear effect of -age- is increased. This is not surprising since this group of calves were old and and hence expected to have a low probability of sepsis, but in fact all 3 were septicemic. This has reduced the effect of age in the full model.

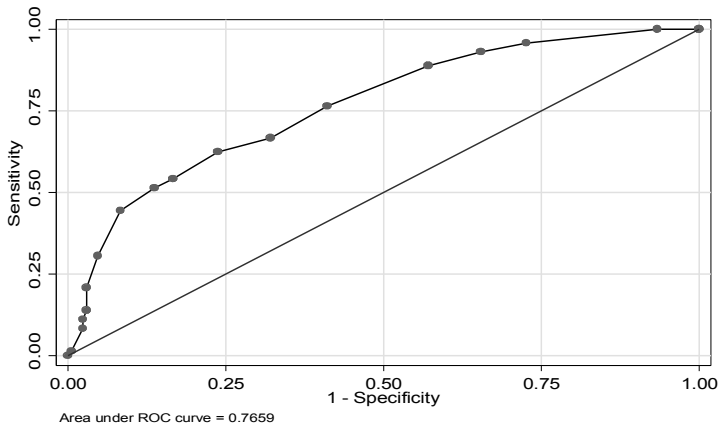
Leaving out covariate pattern #32 results in a larger estimated effect for -umb- and -post- also a small larger linear effect of -age- was observed. These were calves just above average age (average 9 days) which were standing and did not have swollen umbilicus so were expected to be septicemic free, but two of them were.

Leaving out covariate pattern #59 results in a larger estimated effect for -post- and a smaller linear effect of -age-. These were older calves which were expected were expected to be septicemic free, but with sternal posture and did not have swollen umbilicus and were septicemic free.

For these covariate patterns, removal of that individual group of calves resulted in more extreme estimates for one or more parameters. Consequently, the full model may be underestimating these effects, but there is no valid reason for not using the full model.

3. Predicting sepsis

As we saw above, the sensitivity and specificity are approximately equal (about 70%) when the threshold is 0.3. Since this is the approximately the prevalence of sepsis in the study population you might use that as a cutpoint. However, where you set the threshold will depend on whether it is more important to maximize sensitivity or specificity and this will depend on what decisions you are going to make as a result of your predictions. To evaluate the overall predictive ability of the model, you can generate an ROC curve using various cutpoints of predicted values. To do this, we will first categorize the predicted values at increments of 0.05 units.



Detailed report of sensitivity and specificity

Cutpoint	Sensitivity	Specificity	Correctly Classified	LR+	LR-
(>= 0)	100.00%	0.00%	30.00%	1.0000	
(>= .05)	100.00%	6.55%	34.58%	1.0701	0.0000
(>= .1)	95.83%	27.38%	47.92%	1.3197	0.1522
(>= .15)	93.06%	34.52%	52.08%	1.4212	0.2011
(>= .2)	88.89%	42.86%	56.67%	1.5556	0.2593
(>= .25)	76.39%	58.93%	64.17%	1.8599	0.4007
(>= .3)	66.67%	67.86%	67.50%	2.0741	0.4912
(>= .35)	62.50%	76.19%	72.08%	2.6250	0.4922
(>= .4)	54.17%	83.33%	74.58%	3.2500	0.5500
(>= .45)	51.39%	86.31%	75.83%	3.7536	0.5632
(>= .5)	44.44%	91.67%	77.50%	5.3333	0.6061
(>= .55)	30.56%	95.24%	75.83%	6.4167	0.7292
(>= .6)	20.83%	97.02%	74.17%	7.0000	0.8160

```
( >= .65 )         13.89%        97.02%       72.08%       4.6667      0.8875
( >= .7 )          11.11%        97.62%       71.67%       4.6667      0.9106
( >= .75 )          8.33%        97.62%       70.83%       3.5000      0.9390
( >= .8 )           1.39%        99.40%       70.00%       2.3333      0.9920
( >  .8 )           0.00%       100.00%       70.00%       1.0000
```

```
ROC                    -Asymptotic Normal--
Obs       Area     Std. Err.      [95% Conf. Interval]
-------------------------------------------------------
240     0.7659       0.0330        0.70119     0.83064
```

The area under the curve being less than 0.8 suggests only a very good level of predictability by the model. The estimates of sensitivity and specificity at various cutpoints show how the sensitivity decreases as the cutpoint is raised, while the specificity goes up. A calf with a model-based probability of sepsis of at least 0.35 is 2.6 times more likely to be septic than a calf with a model based probability below this value.

4. Choosing an appropriate cutpoint

As we saw above, the sensitivity and specificity are approximately equal (about 70%) when the threshold is 0.3. Since this is the approximately the prevalence of sepsis in the study population you might use that as a cutpoint. However, where you set the threshold will depend on whether it is more important to maximize sensitivity or specificity and this will depend on what decisions you are going to make as a result of your predictions.

If you wanted to be certain of detecting a high proportion of septicemic calves in order to ensure that they received aggressive therapy for the condition, then you would choose a low cutpoint (eg. 0.2).

```
. estat class, cut(0.2)

Logistic model for sepsis

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        64            96  |        160
     -     |         8            72  |         80
-----------+--------------------------+-----------
   Total   |        72           168  |        240

Classified + if predicted Pr(D) >= .2
True D defined as sepsis != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   88.89%
Specificity                     Pr( -|~D)   42.86%
Positive predictive value       Pr( D| +)   40.00%
Negative predictive value       Pr(~D| -)   90.00%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   57.14%
```

```
False - rate for true D         Pr( -| D)   11.11%
False + rate for classified +   Pr(~D| +)   60.00%
False - rate for classified -   Pr( D| -)   10.00%
--------------------------------------------------
Correctly classified                        56.67%
--------------------------------------------------
```

This gets the sensitivity up to 89%, but the positive predictive value (assuming the prevalence in the study population is similar to the prevalence in your clinical population) is only 40%. This means that 60% of the calves that are treated much more aggressively, would, in fact, not have needed the more extensive treatment.

On the other hand, if you were going to euthanise all calves that were classified as septicemic (on the grounds that the prognosis was very poor), you might want to maximise specificity (ie minimise false positives). A cutpoint of 0.6 provides a specificity of 97%.

```
. estat class , cut(0.6)

Logistic model for sepsis

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        15              5  |         20
     -     |        57            163  |        220
-----------+--------------------------+-----------
   Total   |        72            168  |        240

Classified + if predicted Pr(D) >= .6
True D defined as sepsis != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   20.83%
Specificity                     Pr( -|~D)   97.02%
Positive predictive value       Pr( D| +)   75.00%
Negative predictive value       Pr(~D| -)   74.09%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    2.98%
False - rate for true D         Pr( -| D)   79.17%
False + rate for classified +   Pr(~D| +)   25.00%
False - rate for classified -   Pr( D| -)   25.91%
--------------------------------------------------
Correctly classified                        74.17%
--------------------------------------------------
```

The positive predictive value is now 75% so you are reasonably sure that calves that are euthanised were, in fact, septicemic. However, the negative predictive value is down to 74%, which means that 26% of calves that were assumed to be non-septicemic are, in fact, septicemic. If they only receive conventional therapy for diarrhea, their prognosis is very poor.