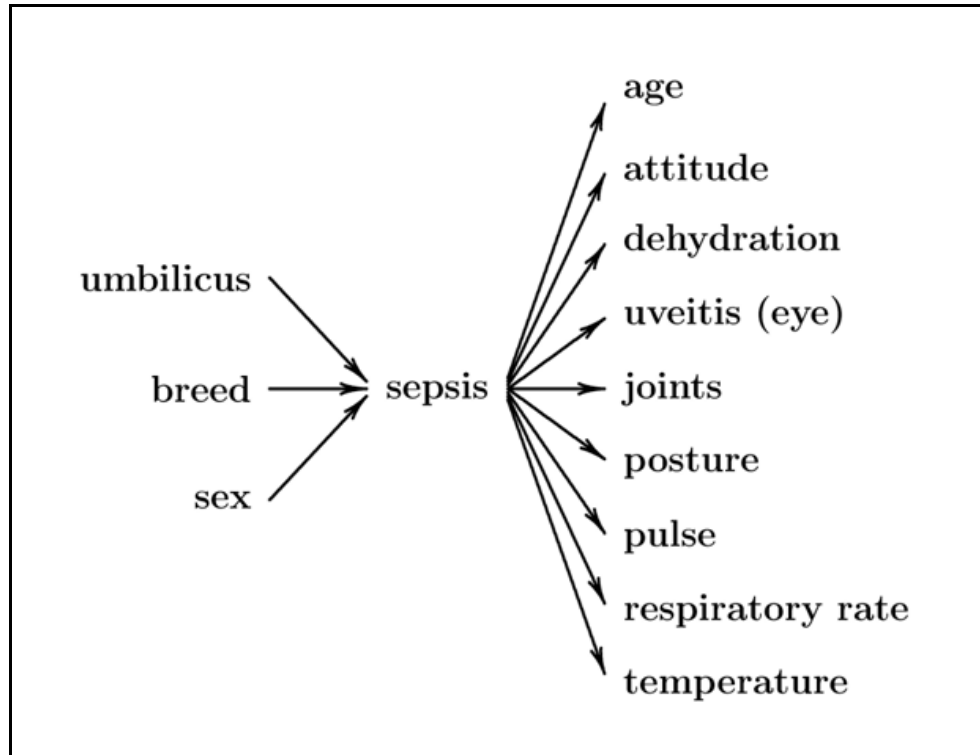


Logistic Regression Model Building Exercise (VER 16.2) Solution

1. Causal structure



This problem represents a very different causal structure than previous exercises and examples we have discussed. In this situation, we want to identify which clinical signs and demographic factors are best able to predict sepsis. Some (umbilicus, breed and sex) are potential “causes” of sepsis. However, most of the clinical signs are, in fact, outcomes of sepsis (e.g., whether or not the calf is septicemic will influence the respiratory rate). (Note: while sepsis can not “cause” age, sepsis might influence the age at which a calf is hospitalised, and this is the age that was recorded in the dataset; if we on the other hand had known the age at which a calf was infected, that would have come before sepsis).

Normally, we do not include consequences of an outcome in a model. But, from a predictive sense we will use all of these factors (signs and demographic characteristics) to try to predict the presence or absence of sepsis. Because the factors may be inter-related we should consider aspects of confounding (e.g. total versus direct effects) in our predictions as we will inevitably find evidence of associations between predictors. We will also need to consider the possibility of interactions if we want to come up with the best predictive model.

2. Linearity of relationships

I will do this for -age- and leave the other variables up to you to do; see the solution do-file for suggested approaches for the other predictors. First, look at the summary statistics for -age- and determine the quartile values.

(a) indicator variables

```
. codebook age
```

```
-----  
age                                     age at admission (in days)  
-----  
  
      type:  numeric (byte)  
      range:  [1,28]  
unique values: 27                               units: 1  
                                                    missing .: 1/254  
  
      mean:   9.36364  
      std. dev: 6.19477  
  
percentiles:      10%      25%      50%      75%      90%  
                  2        5        8        13       18
```

Note one missing value for age. In the absence of any biologically relevant way of categorizing age, I will divide the data at the quartiles, and then check the variable to make sure it has worked.

```
. egen age_c4=cut(age), at(0,5,8,13,30)  
(1 missing value generated)
```

```
. tab age_c4
```

age_c4	Freq.	Percent	Cum.
0	56	22.13	22.13
5	65	25.69	47.83
8	65	25.69	73.52
13	67	26.48	100.00
Total	253	100.00	

Fit a logistic model with -age_c4- as the only predictor.

```
. logit sepsis i.age_c4
```

```
...  
Iteration 0:  log likelihood = -152.01177  
Iteration 1:  log likelihood = -146.23504  
Iteration 2:  log likelihood = -146.15399  
Iteration 3:  log likelihood = -146.15399
```

```
Logistic regression                               Number of obs   =       253  
LR chi2(3)                                       =        11.72  
Prob > chi2                                       =       0.0084  
Log likelihood = -146.15399                       Pseudo R2      =       0.0385
```

```
-----  
sepsis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]  
-----+-----  
age_c4 |  
  5 |   -.7411016   .3823161    -1.94   0.053    -1.490427    .0082242  
  8 |   -1.149667   .4035236    -2.85   0.004    -1.940559   -.3587757  
 13 |   -1.188134   .4026057    -2.95   0.003    -1.977226   -.399041  
     |  
_cons |   -.1431008   .2679457    -0.53   0.593    -.6682647    .382063  
-----
```

It appears as if there is a decrease in risk as age increases, but the rate of decrease drops off as you move up through the categories.

(b) plot of log odds vs predictor

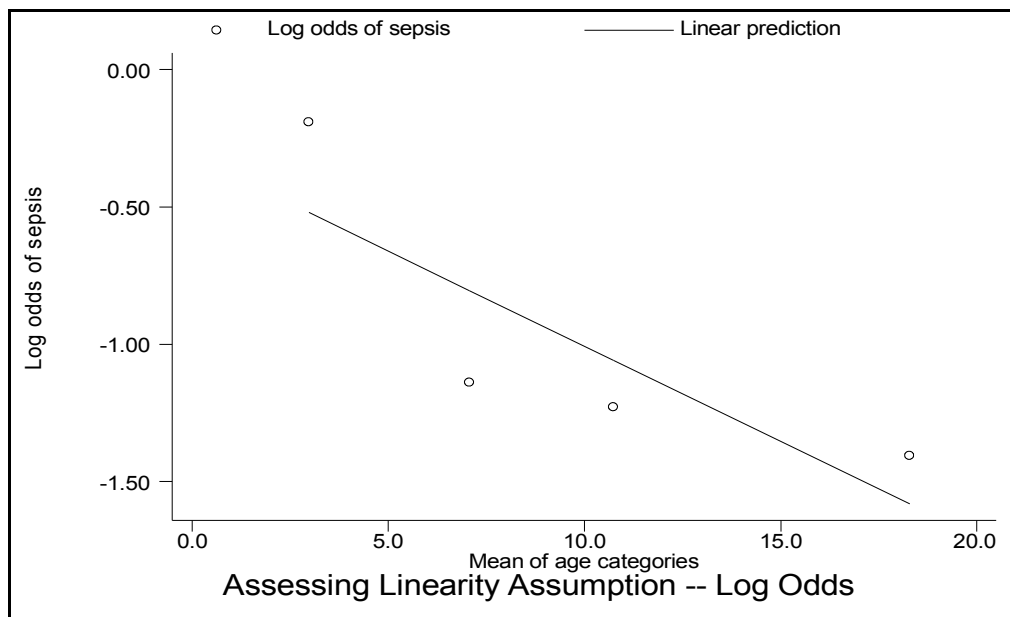
For this we use the lintrend command, which is a Stata add-on command that you need to download (use the findit command to start the process).

```
. lintrend sepsis age, g(4) plot(log) xlab ylab
```

The proportion and log odds of sepsis by categories of age

(Note: 4 age categories of equal sample size;
Uses mean age value for each category)

age	min	max	d	total	sepsis	logodds
3.0	1	5	33	73	0.45	-0.19
7.1	6	8	16	66	0.24	-1.14
10.7	9	13	12	53	0.23	-1.23
18.3	14	28	12	61	0.20	-1.41



The table (above) gives the proportion and log odds of sepsis by categories of age. The graph plots the log odds of sepsis against the midpoint of the categories of age, in order to make it easier to see the shape of the relationship. As with the indicator variables, it seems like there is a substantial drop in the log odds of sepsis if the calf is under 7 days of age, but then it levels off.

(c) quadratic term

A natural next step would be to create a quadratic term and use that to determine if the relationship was non-linear.

```
.generate age_ct=age-10
(1 missing value generated)
.generate age_ctsq=age_ct^2
(1 missing value generated)
. logit sepsis age_ct age_ctsq
```

```
Iteration 0: log likelihood = -152.01177
Iteration 1: log likelihood = -146.10453
Iteration 2: log likelihood = -146.05013
Iteration 3: log likelihood = -146.05013
```

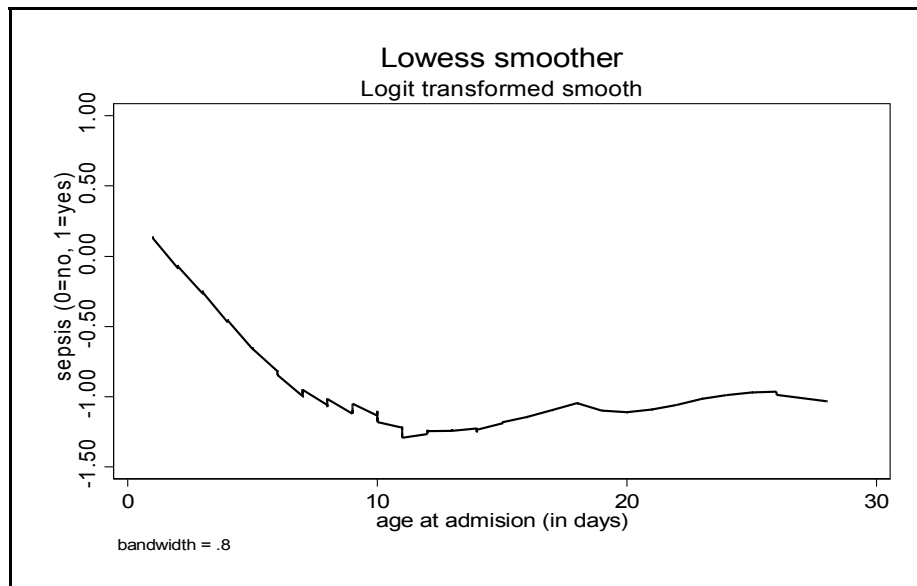
```
Logistic regression          Number of obs   =       253
                             LR chi2(2)            =       11.92
                             Prob > chi2           =       0.0026
                             Pseudo R2             =       0.0392
```

```
Log likelihood = -146.05013
```

sepsis	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age_ct	-.0841157	.0259882	-3.24	0.001	-.1350516	-.0331798
age_ctsq	.007484	.0029367	2.55	0.011	.0017281	.0132399
_cons	-1.289177	.2016961	-6.39	0.000	-1.684494	-.8938603

The quadratic term is significant, suggesting that the relationship is indeed curved (i.e., non-linear). An alternative approach to assessing the shape of this relationship is to fit some sort of smoothed curve through a scatterplot of the -age- vs -sepsis- relationship on the logit scale, although the exercise did not ask you to do this.

```
. format sepsis %6.2f /* to avoid strange-looking and wrong y-axis */
. lowess sepsis age, logit scheme(simono)
```



This graph suggests that either a quadratic relationship or one consisting of two linear splines might be appropriate.

3. Model building

The basis for the model building should be findings from the unconditional analyses; some of these were summarised in the solution for the previous exercise (VER 16.1). It is often useful to make sure the unconditional analyses include (logistic regression) models with each predictor separately, in order to get a feeling for total relationship of each predictor to the outcome variable on the scale used for multivariable modelling.

Two specific findings from the unconditional analysis are mentioned here because they lead to construction of new variables. First, -age- was centred and a squared term was generated. This was the only variable for which there was strong enough evidence of a non-linear relationship to warrant this treatment. (Note it is not necessary to centre -age- before squaring it if you just want to evaluate the significance of the quadratic term. However, if the quadratic term is retained in the model, interpretation is usually facilitated by using a centred version of the predictor). Second, it was noted that -jnts- had a too sparse category: there was only one calf with 4 swollen joints, and it is natural to put this observation together with the 5 observations with 2 swollen joints (the new category becoming thus 2 or more swollen joints).

The manual model building (part (a)) will be left for you to do. We show results for an automated variable selection based on backward selection.

```
. generate jnts3=jnts
(11 missing values generated)
. replace jnts3=2 if jnts==4
(1 real change made)

. xi:stepwise, pr(0.1) pe(0.099): logit sepsis (age_ct age_ctsq) dehy resp temp (i.attd) eye (i.jnts3)
(i.post) umb
i.attd      _Iattd_0-2      (naturally coded; _Iattd_0 omitted)
i.jnts3     _Ijnts3_0-2     (naturally coded; _Ijnts3_0 omitted)
i.post      _Ipost_0-2     (naturally coded; _Ipost_0 omitted)
begin with full model
p = 0.7477 >= 0.1000 removing dehy
p = 0.5156 >= 0.1000 removing _Ijnts3_1 _Ijnts3_2
p = 0.3080 >= 0.1000 removing temp
p = 0.2299 >= 0.1000 removing resp
p = 0.1292 >= 0.1000 removing eye
p = 0.1063 >= 0.1000 removing _Iattd_1 _Iattd_2
```

```
Logistic regression                                Number of obs =          215
                                                    LR chi2(5)          =          36.96
                                                    Prob > chi2         =          0.0000
Log likelihood = -110.66878                        Pseudo R2           =          0.1431
```

	sepsis	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
	age_ct	-.0790008	.0301248	-2.62	0.009	-.1380443	-.0199573
	age_ctsq	.0086542	.0034741	2.49	0.013	.0018451	.0154633
	umb	.8591326	.3755848	2.29	0.022	.1229999	1.595265
	_Ipost_1	1.527862	.4364585	3.50	0.000	.6724193	2.383305
	_Ipost_2	1.786502	.4589452	3.89	0.000	.8869858	2.686018
	_cons	-2.706408	.4181162	-6.47	0.000	-3.525901	-1.886915

```
. * final model refitted with all observations included
. logit sepsis i.post umb age_ct age_ctsq
```

```
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
	age_ctsq	.0086873	.0033643	2.58	0.010	.0020934	.0152812
	_cons	-2.768343	.4018875	-6.89	0.000	-3.556028	-1.980658

```
. estimates store final
. estimates esample: post sepsis umb age_ct, replace /* to keep the same dataset below */
```

It is seen that -age- (in a quadratic form), -umb- and -post- remain as significant predictors of sepsis.

4. Confounding and interaction

(a) interaction and/or confounding between -umb- and -post-

First a model which included interaction terms between -umb- and -post- was fit, and the significance of the interaction was assessed with a multiple Wald test.

```
. logit sepsis post##umb age_ct age_ctsq

Iteration 0:  log likelihood = -146.60743
Iteration 1:  log likelihood = -124.24964
Iteration 2:  log likelihood = -123.10101
Iteration 3:  log likelihood = -123.08409
Iteration 4:  log likelihood = -123.08408
Iteration 5:  log likelihood = -123.08408

Logistic regression              Number of obs   =       240
                                LR chi2(7)        =       47.05
                                Prob > chi2         =       0.0000
Log likelihood = -123.08408      Pseudo R2       =       0.1605

-----+-----
      sepsis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      post   |
      sternal|  1.705165   .513114     3.32  0.001   .6994803   2.71085
      lateral|  1.890318   .5280981    3.58  0.000   .8552651   2.925372
      umb    |
      yes    |  1.273292   .7244878    1.76  0.079  -.1466783   2.693261
      post#umb |
      sternal#yes | -.3574362   .9007454   -0.40  0.691  -2.122865   1.407992
      lateral#yes | -.1143824   .9642375   -0.12  0.906  -2.004253   1.775488
      age_ct | -.0830928   .0290381   -2.86  0.004  -.1400064  -.0261791
      age_ctsq | .0088271    .0033888    2.60  0.009   .0021851   .015469
      _cons  | -2.843414   .4770802   -5.96  0.000  -3.778474  -1.908354
-----+-----

. testparm post#umb

( 1) [sepsis]1.post#1.umb = 0
( 2) [sepsis]2.post#1.umb = 0

      chi2( 2) =    0.18
      Prob > chi2 =  0.9137
```

Therefore, there is no indication or evidence of interaction whatsoever. To check for confounding, we explore the association between -umb- and -post-, as well as the impact of omitting one of the predictors on the coefficient for the other one. These analyses are not based on the causal structure, but as discussed above the causal structure is unusual here and should probably not be used for making specific choices about how to assess confounding.

```
. tabulate umb post, chi2

swollen |
umbilicus | posture (0=standing, 1=sternal,
(0=no,    |          2=lateral)
1=yes) | standing  sternal  lateral | Total
-----+-----
      no |      70      65      51 |    186
      yes |      19      21      15 |     55
-----+-----
      Total |      89      86      66 |    241

      Pearson chi2(2) =    0.2345   Pr =  0.889
```

```
. logit sepsis umb age_ct age_ctsq if e(sample)
(results not shown)
. estimates store no_post
. logit sepsis i.post age_ct age_ctsq if e(sample)
(results not shown)
. estimates store no_umb
. estimates table no_umb no_post final , stat(N)
```

Variable	no_umb	no_post	final
post			
sternal	1.591951		1.5943382
lateral	1.7999995		1.8389785
age_ct	-.08346151	-.08749481	-.08231092
age_ctsq	.0084945	.00877887	.0086873
umb		1.0131637	1.0784534
_cons	-2.4797113	-1.5521923	-2.7683427
N	240	240	240

There was no indication or evidence of an association between -post- and -umb-, thus precluding any of those variables to play a confounding role. Also the changes in the coefficients when the other variable was omitted were relatively small (much less than 20%). We conclude that no interaction or confounding exists between -post and -umb-.

(b) -age- as a confounder

As above, a model will be fit leaving -age- (both linear and quadratic terms) out to see if the coefficients for -post- and -umb- change.

```
. logit sepsis i.post umb if e(sample)
(results not shown)
. estimates table . final, stat(N)
```

Variable	active	final
post		
sternal	1.6430163	1.5943382
lateral	1.8882768	1.8389785
umb	1.0801901	1.0784534
age_ct		-.08231092
age_ctsq		.0086873
_cons	-2.37994	-2.7683427
N	240	240

The coefficients for both -post- and -umb- were only little affected by omitting the age terms from the model. So in this case, there is really no need to continue with associations between age and the predictors. We conclude that no confounding is exerted by -age- on the other predictors in the model.