Biostatistics 140.623
Third Term, 2002-2003

Laboratory Exercise 4
Answer Key

The following model explores the relationship between child's age and breastfeeding (1-yes, 0-no) for the 302 mother-child pairs drawn at random from the Nepali class data set:

$$\text{logit Pr(BF} = 1) = \log \ \text{odds}\left(\text{BF} = 1\right) = \beta_0 + \beta_1\left(\text{child's \ age} - 36\right)$$

The following are the results of a logistic regression analysis of breastfeeding on age (in months) using these data in Stata

```
. gen age36=age_chld-36
. logit bf  age36

Iteration 0:   log likelihood = -209.30396
Iteration 1:   log likelihood =  -114.3689
Iteration 2:   log likelihood = -102.25897
Iteration 3:   log likelihood = -100.58092
Iteration 4:   log likelihood = -100.52192
Iteration 5:   log likelihood = -100.52182

Logit estimates                                   Number of obs   =        302
                                                  LR chi2(1)      =     217.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -100.52182                       Pseudo R2       =     0.5197

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age36 |  -.1761668   .0191232    -9.21   0.000    -.2136476   -.1386861
       _cons |  -.6315363   .1908287    -3.31   0.001    -1.005554   -.2575189
```

```
. logistic bf age36

Logit estimates                                   Number of obs   =        302
                                                  LR chi2(1)      =     217.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -100.52182                       Pseudo R2       =     0.5197

------------------------------------------------------------------------------
          bf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age36 |   .8384781   .0160344    -9.21   0.000      .807633    .8705012
------------------------------------------------------------------------------
```

1. From the regression results above, estimate the prevalence of breastfeeding among 36-month old infants.

$$\log \text{ odds } (\text{BF}=1) = b_0 + b_1(\text{child's age} - 36)$$

$$\log\left(\frac{\text{p}}{1-\text{p}}\right) = -0.6315 + (-0.1762(\text{child's age} - 36))$$

$$\log\left(\frac{\text{p}}{1-\text{p}}\right) = -0.6315$$

$$e^{\log\left(\frac{\text{p}}{1-\text{p}}\right)} = e^{-0.6315+0}$$

$$\left(\frac{\text{p}}{1-\text{p}}\right) = 0.5318$$

p = (1-p) 0.5318

1.5318p=0.5318

p = 0.3472 which is the same as $p = \dfrac{e^{-0.6315+0}}{1+e^{-0.6315+0}}$ where $p = \dfrac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$

The following model includes child's gender (0=male; 1=female).

```
. logit bf age36 sex_chld

Iteration 0:   log likelihood = -209.30396
Iteration 1:   log likelihood = -114.05425
Iteration 2:   log likelihood = -101.75438
Iteration 3:   log likelihood = -100.00867
Iteration 4:   log likelihood = -99.943901
Iteration 5:   log likelihood = -99.943775

Logit estimates                                   Number of obs   =        302
                                                  LR chi2(2)      =     218.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -99.943775                       Pseudo R2       =     0.5225

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age36 |  -.1785173   .0194601    -9.17   0.000    -.2166585   -.1403761
    sex_chld |  -.3892598   .3643308    -1.07   0.285    -1.103335    .3248154
       _cons |  -.4514222   .2525563    -1.79   0.074    -.9464234    .0435791
------------------------------------------------------------------------------

. lrtest, saving (0)


.quietly logit bf age36
```

```
. lrtest

Logit:  likelihood-ratio test                              chi2(1)     =       1.16
                                                           Prob > chi2 =     0.2823
```

2. Test whether this additional covariate is needed in the model by

a) Using a z-test (Wald test=estimate/se)

	Coeff	SE	Z	p
sex_chld |	-.3892598	.3643308	-1.07	0.285

b) Comparing the extended and null models using the likelihood-ratio test result.
From above,

```
chi2(1)     =       1.16
Prob > chi2 =     0.2823
```

c) Verifying by hand the result of the likelihood ratio test.

Null model:	age	Log likelihood = -100.52182
Extended model:	age, gender	Log likelihood = -99.943775

LRT = -2 (difference in log-likelihoods)
= -2(log-likelihood of null model – log-likelihood of extended model)
= -2(-100.52 –(-99.94)) = 1.16 with 1 degree of freedom

d) Does inclusion of the additional covariate improve the fit of the model?

No, there is no statistically significant contribution of gender from the results of either the Wald test or the Likelihood Ratio Test.

3. Interpret the estimated logistic regression coefficients for age and gender.

b_1 = the difference in the log odds of breastfeeding for an infant of age x+1 months and x months, controlling for gender

b_2 = the difference in the log odds of breastfeeding for males and females, controlling for age

4. Estimate the prevalence of breastfeeding for a 36-month old female child versus that for a 36-month old male child.

$$p = \frac{e^{b_0+b_1x_1+b_2x_2}}{1+e^{b_0+b_1x_1+b_2x_2}} = p = \frac{e^{-0.4514+(-0.1785)x_1+(-0.3893)x_2}}{1+e^{-0.4514+(-0.1785)x_1+(-0.3893)x_2}}$$

For females (X_2=1):

$$p = \frac{e^{-0.4514+(-0.1785)x_1+(-0.3893)x_2}}{1+e^{-0.4514+(-0.1785)x_1+(-0.3893)x_2}} = 0.3014$$

For males (X_2=0):

$$p = \frac{e^{-0.4514+(-0.1785)x_1+(-0.3893)x_2}}{1+e^{-0.4514+(-0.1785)x_1+(-0.3893)x_2}} = 0.3890$$

5. Estimate the prevalence of breastfeeding for a 12-month old male child.

$$p = \frac{e^{-0.4514+(-0.1785)(12-36)+(-0.3893)0}}{1+e^{-0.4514+(-0.1785)(12-36)+(-0.3893)0}} = 0.98$$

6. The following is a Hosmer-Lemeshow goodness-of-fit test for the model that includes child's age and gender. Interpret the result of this test.

```
. quietly logit bf age36 sex_chld

. lfit, group(5)

Logistic model for bf, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

       number of observations =       302
             number of groups =         5
      Hosmer-Lemeshow chi2(3) =         2.13
                  Prob > chi2 =        0.5468
```

Based on p=0.5468, this model is a good fit by the Hosmer-Lemeshow criteria.

7. The following model includes only child's gender (0=male; 1=female). Compare these results to the previous logistic regression results.

```
. logit bf sex_chld

Iteration 0:   log likelihood = -209.30396
Iteration 1:   log likelihood = -209.13468
Iteration 2:   log likelihood = -209.13468

Logit estimates                                   Number of obs   =        302
                                                  LR chi2(1)      =       0.34
                                                  Prob > chi2     =     0.5607
Log likelihood = -209.13468                       Pseudo R2       =     0.0008

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sex_chld |   .1340247   .2304102     0.58   0.561     -.317571    .5856203
       _cons |  -.0387145    .160674    -0.24   0.810    -.3536297    .2762007
------------------------------------------------------------------------------
```

8. Which model do you prefer and why? Justify your choice and summarize the findings of your analysis in a sentence or two.

We can compare the model with gender only and the model with age and gender:

Null model:	gender	Log likelihood = -209.13468
Extended model:	age, gender	Log likelihood = -99.943775

LRT = -2 (difference in log-likelihoods)
= -2(log-likelihood of null model – log-likelihood of extended model)
= -2(-209.13 –(-99.94)) = 218.4 with 1 degree of freedom

which indicates a significant addition when age is added to the model. The coefficient for age does not substantially change when gender is added to the model. Although gender is not statistically associated with the odds of breastfeeding, one might keep gender in the model as a controlling variable.

Our results indicate that, after controlling for gender, the odds of breastfeeding are substantially reduced with increased age of the child (OR=0.84, 95% CI: 0.81, 0.87). After controlling for age, the odds of breastfeeding are reduced in females as compared to males (OR=0.68, 95% CI: 0.33,1.38).

9. Below find two 2x2 tables that show the number of Nepali children breastfeeding by age (< 36 months, 36-60 months) for boys versus girls.

```
-> sex_chld = 0 (Males)

           |         ageb
breast fed | < 36 mont  36+ month |     Total
-----------+----------------------+----------
         0 |        12         67 |        79
         1 |        65         11 |        76
-----------+----------------------+----------
     Total |        77         78 |       155

-> sex_chld = 1 (Females)

           |         ageb
breast fed | < 36 mont  36+ month |     Total
-----------+----------------------+----------
         0 |        17         53 |        70
         1 |        72          5 |        77
-----------+----------------------+----------
     Total |        89         58 |       147
```

Pool the data above to obtain a single 2x2 table that ignores the gender of the child.

```
            |         ageb
breast fed | < 36 mont  36+ month |     Total
-----------+----------------------+----------
         0 |     29      120      |    149
         1 |    137       16      |    153
-----------+----------------------+----------
     Total |    166      136      |     302
```

10. Calculate the log odds ratio and standard error and confidence interval for each of the three tables above:

Group	**OR estimate**	**Log OR**	**se(Log OR)**	**95% CI for log odds ratio**	
Pooled	.0282238	-3.567588	.3355821	-4.225329	-2.909847
Boys	.03031	-3.496278	.4522747	-4.382737	-2.60982
Girls	.0222746	-3.804307	.5399818	-4.862671	-2.745942

Group	**OR estimate**	**95% CI for OR**	
Pooled	.0282238	.0146205	.0544841
Boys	.03031	.0124911	.0735478
Girls	.0222746	.0077298	.0641878

Compare to the `Stata` results on the next page.

```
. cs ageb bf, or

                 | breast fed             |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        16         120  |        136
        Noncases |       137          29  |        166
-----------------+------------------------+----------
           Total |       153         149  |        302
                 |                        |
            Risk |  .1045752    .8053691  |  .4503311
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         -.700794       | -.7807459    -.620842
      Risk ratio |         .1298475       |  .0811279    .2078247
 Prev. frac. ex. |         .8701525       |  .7921753    .9188721
 Prev. frac. pop |         .4408389       |
      Odds ratio |         .0282238       |  .0146843    .0542816  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =   149.77  Pr>chi2 = 0.0000
```

```
. cs ageb bf, or by(sex_chld)

 gender: M=0 F=1 |       OR       [95% Conf. Interval]    M-H Weight
-----------------+---------------------------------------------------
               0 |     .03031      .0125813   .0730224      28.09677 (Cornfield)
               1 |   .0222746        .00797   .0628017      25.95918 (Cornfield)
-----------------+---------------------------------------------------
           Crude |   .0282238      .0146843   .0542816
    M-H combined |   .0264512      .0134179   .0521443
---------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    0.192  Pr>chi2 = 0.6613

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =    149.18
                                                Pr>chi2 =    0.0000
```

11. Let ageb =0 if age < 36 months, 1if age 36+ months. Fit the following logistic regression models:

    Model A: $\text{logit}\Pr(\text{BF}=1)=\beta_0+\beta_1\text{ageb}$

    Model B: $\text{logit}\Pr(\text{BF}=1)=\beta_0+\beta_1\text{ageb}+\beta_2(\text{gender})$

    Model C: $\text{logit}\Pr(\text{BF}=1)=\beta_0+\beta_1\text{ageb}+\beta_2(\text{gender})+\beta_3(\text{ageb}*\text{gender})$

    Match the logistic regression coefficients above to the results of the log odds ratios in question 10.

In **Model A**, logit $\Pr(\text{BF}=1) = \beta_0 + \beta_1\text{ageb}$

When gender=0 (male) and ageb=0 (younger), then
β_0 = the log odds of breastfeeding in younger children (< 36 months).

β_1 = the difference in the log odds of breastfeeding in older children (36+ months) and younger children (< 36 months).

In **Model B**, logit $\Pr(\text{BF}=1) = \beta_0 + \beta_1\text{ageb} + \beta_2(\text{gender})$

β_0 = the log odds of breastfeeding in younger male children (< 36 months).

When gender =0 (male), then log(odds)= $\beta_0 + \beta_1$age
If we look at age=1 (older), then log(odds of breastfeeding in older males)= $\beta_0 + \beta_1$
If we look at age=0 (younger), then log(odds of breastfeeding in younger males)= β_0
Subtracting these, we get
β_1 =log(odds of breastfeeding in older males) - log(odds of breastfeeding in younger males)

When gender =1 (female), then log(odds)= $\beta_0 + \beta_1$age $+\beta_2$
If we look at age=1 (older), then log(odds of breastfeeding in older females)= $\beta_0 + \beta_1 + \beta_2$
If we look at age=0 (younger), then log(odds of breastfeeding in younger females)= $\beta_0 + \beta_2$
Subtracting these, we get
β_1=log(odds of breastfeeding in older females)- log(odds of breastfeeding in younger females)

Thus, β_1 = the difference in the log odds of breastfeeding between older children (36+ months) and younger children (< 36 months) after controlling for gender.

Similarly, β_2 = the difference in the log odds of breastfeeding in females and males after controlling for age.

12. Interpret the coefficients in Models B and C using the terms "effect modifer" and "confounder" as if for a public health journal.

There is no evidence of confounding since the magnitude of the estimated regression coefficient for age remains similar in both Models A and B.

We see no evidence that the odds ratio is different for girls or boys. Thus, it appears that gender does not modify the relationship between age and breastfeeding. (Gender is not an effect - modifier.) There is not a significant interaction effect of age and gender on breastfeeding.

Note: In the last part of this lab exercise, we have lost information by dichotomizing age. It may be best to keep age as a continuous covariate and explore the possibility of using linear spline terms.

Model A

```
. logit bf ageb

Iteration 0:   log likelihood = -209.30396
Iteration 1:   log likelihood = -128.83764
Iteration 2:   log likelihood = -126.20336
Iteration 3:   log likelihood = -126.16167
Iteration 4:   log likelihood = -126.16164

Logit estimates                                   Number of obs   =        302
                                                  LR chi2(1)      =     166.28
                                                  Prob > chi2     =     0.0000
Log likelihood = -126.16164                       Pseudo R2       =     0.3972

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ageb |  -3.567588    .335582   -10.63   0.000    -4.225317   -2.909859
       _cons |   1.552685   .2044065     7.60   0.000     1.152056    1.953315
------------------------------------------------------------------------------
```

Model B

```
. logit bf ageb sex_chld

Iteration 0:   log likelihood = -209.30396
Iteration 1:   log likelihood =  -128.4437
Iteration 2:   log likelihood = -125.63848
Iteration 3:   log likelihood = -125.58479
Iteration 4:   log likelihood = -125.58474

Logit estimates                                   Number of obs   =        302
                                                  LR chi2(2)      =     167.44
                                                  Prob > chi2     =     0.0000
Log likelihood = -125.58474                       Pseudo R2       =     0.4000

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ageb |  -3.628328   .3449674   -10.52   0.000    -4.304451   -2.952204
    sex_chld |  -.3548912   .3333839    -1.06   0.287    -1.008312    .2985293
       _cons |   1.753121   .2847674     6.16   0.000     1.194987    2.311255
------------------------------------------------------------------------------
```

Model C

```
. gen interact=ageb*sex_chld

. logit bf ageb sex_chld interact

Iteration 0:   log likelihood = -209.30396
Iteration 1:   log likelihood = -128.41831
Iteration 2:   log likelihood = -125.55842
Iteration 3:   log likelihood = -125.48828
Iteration 4:   log likelihood = -125.48808
Iteration 5:   log likelihood = -125.48808

Logit estimates                                   Number of obs   =        302
                                                  LR chi2(3)      =     167.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -125.48808                       Pseudo R2       =     0.4005

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ageb | -3.496278   .4522747    -7.73   0.000     -4.38272   -2.609836
    sex_chld | -.2460278   .4140415    -0.59   0.552    -1.057534    .5654786
    interact | -.3080288   .7043669    -0.44   0.662    -1.688563    1.072505
       _cons |  1.689481   .3141941     5.38   0.000     1.073671     2.30529
------------------------------------------------------------------------------
```

