```
#### Biost 518: Applied Biostatistics II
#### Emerson, Winter 2008

#### Homework #2 Key
#### Annotated Stata Log File
#### February 6, 2008

#### The following output was used to generate the numbers that I wanted to present
#### in tables, as well as the plots I wanted to present as figures in the paper.
#### I note that Stata does not present its output in a form suitable for presentation.
#### Numbers need to be rounded to an interesting number of significant digits, and
#### the columns and rows need to facilitate comparison of relevant measures.
#### I used Excel to manipulate this output into the form I wanted, then copied the
#### resulting tables to the MS-Word document.

#### Comments edited into the log file produced by Stata are
#### on the lines that start with the four '#' signs and are
#### printed in italics.

#### The Stata commands are put in bold face.

#### Stata output is displayed in regular typeface in blue.

#### Read in data. I drop the first case, because it was just the column headings.
. infile ptid nadir pretx ps bss grade age obstime str8 inrem using psa.txt
'ptid' cannot be read as a number for ptid[1]
'nadirpsa' cannot be read as a number for nadir[1]
'pretxpsa' cannot be read as a number for pretx[1]
'ps' cannot be read as a number for ps[1]
'bss' cannot be read as a number for bss[1]
'grade' cannot be read as a number for grade[1]
'age' cannot be read as a number for age[1]
'obstime' cannot be read as a number for obstime[1]
'NA' cannot be read as a number for pretx[8]
'NA' cannot be read as a number for pretx[15]
'NA' cannot be read as a number for pretx[18]
'NA' cannot be read as a number for grade[22]
'NA' cannot be read as a number for grade[24]
'NA' cannot be read as a number for grade[26]
'NA' cannot be read as a number for grade[27]
'NA' cannot be read as a number for grade[30]
'NA' cannot be read as a number for grade[33]
'NA' cannot be read as a number for pretx[35]
'NA' cannot be read as a number for grade[35]
```

```
'NA' cannot be read as a number for ps[37]
'NA' cannot be read as a number for bss[37]
'NA' cannot be read as a number for pretx[43]
'NA' cannot be read as a number for ps[43]
'NA' cannot be read as a number for bss[43]
'NA' cannot be read as a number for grade[43]
'NA' cannot be read as a number for pretx[46]
'NA' cannot be read as a number for pretx[51]
'NA' cannot be read as a number for grade[51]
(51 observations read)

. drop in 1
(1 observation deleted)


#### Creating variables to indicate relapse and relapse within 24 months
#### I compare the variables to make sure I did not make a coding mistake

. g relapse= .
(50 missing values generated)
. replace relapse= 1 if inrem=="no"
(36 real changes made)
. replace relapse= 0 if inrem=="yes"
(14 real changes made)

. g relapse24= .
(50 missing values generated)
. replace relapse24=1 if obstime<=24 & inrem=="no"
(22 real changes made)
. replace relapse24=0 if obstime > 24 | (obstime==24 & inrem=="yes")
(28 real changes made)

. table relapse24 relapse
          |  relapse
relapse24 |    0     1
---------+----------
        0 |   14    14
        1 |         22

. bysort relapse24: summ obstime
-> relapse24 = 0
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+-------------------------------------------------------
```

```
     obstime |        28    42.07143    12.05214         24         75

-> relapse24 = 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     obstime |        22    11.13636    6.401603          1         22

##########################################################################################
#### Problem 1
#### Descriptive statistics: Since a major focus is whether patients relapse or not, I choose to
#### provide descriptive statistics within groups defined by relapse within 24 months (I could
#### tell who was in which group, because the earliest censoring time was 24 months)
##########################################################################################

. tabstat age ps bss grade pretx nadir, stat(n mean sd min q max) col(stat) by(relapse24)
Summary for variables: age ps bss grade pretx nadir
     by categories of: relapse24
relapse24 |         N      mean        sd       min       p25       p50       p75       max
        0 |        28  66.71429  5.842736        58        63      65.5      69.5        81
          |        28  83.92857  9.560445        50        80        80        90       100
          |        28  2.321429  .7723735         1         2       2.5         3         3
          |        24  2.083333  .8297022         1         1         2         3         3
          |        23   617.187   1252.08       4.8        45       100       387      4377
          |        28  4.117857  17.27921        .1        .2        .2       .95        92
        1 |        22  68.36364  5.678241        61        64        68        71        86
          |        20      76.5  11.82103        50        70        80        80       100
          |        20       2.8  .4103913         2         3         3         3         3
          |        17  2.235294   .752447         1         2         2         3         3
          |        20    732.35  1357.341        25      69.5       174       530      4797
          |        22  31.94091  52.49686        .5       1.2      10.5        38       183
    Total |        50     67.44  5.771711        58        63        66        70        86
          |        48  80.83333  11.07678        50        80        80        90       100
          |        48  2.520833  .6838434         1         2         3         3         3
          |        41  2.146341  .7924953         1         2         2         3         3
          |        43  670.7512  1287.638       4.8        46       127       429      4797
          |        50     16.36   39.2462        .1        .2       .95        10       183

##########################################################################################
#### Problem 2
#### Comparing means using dichotomized value of bone scan score: bss3
##########################################################################################
```

```
. g bss3= bss
(2 missing values generated)
. recode bss3 1/2=0 3=1
(bss3: 48 changes made)

#### Problem 2a
#### Generate the descriptive statistics: I used tabstat and means to get the descriptives and CI, respectively.
#### However, it should be noted that ttest gives all the descriptive statistics I need for this problem.

. tabstat nadir, stat(n mean sd) col(stat) by(bss3)
Summary for variables: nadir
     by categories of: bss3
    bss3 |         N      mean        sd
       0 |        18  3.544444   8.94856
       1 |        30     24.85  48.66416
   Total |        48  16.86042  39.98558

. bysort bss3: means nadir
-> bss3 = 0
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
       nadir | Arithmetic       18    3.544444       -.9055699   7.994459
             |  Geometric       18    .7572496        .3407211   1.682981
             |   Harmonic       18    .3681841        .2617558   .6204585

-> bss3 = 1
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
       nadir | Arithmetic       30       24.85        6.678504    43.0215
             |  Geometric       30    2.861879        1.191359   6.874798
             |   Harmonic       30    .5643297        .3720422   1.168005

-> bss3 = .
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
       nadir | Arithmetic        2        4.35       -42.02765   50.72765
             |  Geometric        2    2.366432        4.49e-07   1.25e+07
             |   Harmonic        2    1.287356               .          .
Missing values in confidence intervals for harmonic mean indicate
that confidence interval is undefined for corresponding variables.
Consult Reference Manual for details.

#### Problem 2b: t test which presumes equal variances
. ttest nadir, by(bss3)
Two-sample t test with equal variances
```

```
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+-----------------------------------------------------------------
       0 |      18    3.544444    2.109196     8.94856   -.9055699    7.994459
       1 |      30       24.85    8.884819    48.66416    6.678504     43.0215
---------+-----------------------------------------------------------------
combined |      48    16.86042    5.771422    39.98558    5.249814    28.47102
    diff |           -21.30556    11.63363               -44.72283    2.111722
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.8314
Ho: diff = 0                                     degrees of freedom =       46

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0368         Pr(|T| > |t|) = 0.0735          Pr(T > t) = 0.9632
```

Problem 2d: t test which allows for unequal variances

```
. ttest nadir, by(bss3) unequal
Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+-----------------------------------------------------------------
       0 |      18    3.544444    2.109196     8.94856   -.9055699    7.994459
       1 |      30       24.85    8.884819    48.66416    6.678504     43.0215
---------+-----------------------------------------------------------------
combined |      48    16.86042    5.771422    39.98558    5.249814    28.47102
    diff |           -21.30556    9.131743               -39.90208   -2.709027
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.3331
Ho: diff = 0                     Satterthwaite's degrees of freedom =  32.1863

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0130         Pr(|T| > |t|) = 0.0260          Pr(T > t) = 0.9870
```

Problem 2e: Classical linear regression on a binary predictor

```
. regress nadir bss3
                                                      Number of obs =      48
      Source |       SS       df       MS              F(  1,    46) =    3.35
-------------+------------------------------           Prob > F      =  0.0735
       Model |  5106.67535     1  5106.67535           R-squared     =  0.0680
    Residual |  70039.1194    46  1522.58955           Adj R-squared =  0.0477
-------------+------------------------------           Root MSE      =   39.02
       Total |  75145.7948    47   1598.8467
------------------------------------------------------------------------------
       nadir |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bss3 |   21.30556   11.63363     1.83   0.074    -2.111722    44.72283
       _cons |   3.544444    9.19719     0.39   0.702    -14.96854    22.05743
------------------------------------------------------------------------------
```

Problem 2i: Linear regression on a binary predictor using robust SE

```
. regress nadir bss3, robust
Linear regression                                      Number of obs =      48
                                                       F(  1,    46) =    5.40
                                                       Prob > F      =  0.0246
```

```
                                                      R-squared     =  0.0680
                                                      Root MSE      =   39.02
------------------------------------------------------------------------------
             |               Robust
       nadir |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bss3 |   21.30556   9.165734     2.32   0.025     2.855889    39.75522
       _cons |   3.544444   2.093856     1.69   0.097    -.6702693    7.759158
------------------------------------------------------------------------------
```

```
##########################################################################################
#### Problem 3
#### Comparing means using bone scan score modeled continuously: bss
##########################################################################################

#### Problem 3a
#### Generate the descriptive statistics: I used tabstat and means to get the descriptives and CI, respectively.

. tabstat nadir, stat(n mean sd) col(stat) by(bss)
Summary for variables: nadir
     by categories of: bss

     bss |         N      mean        sd
---------+----------------------------------
       1 |         5        .2         0
       2 |        13  4.830769  10.34355
       3 |        30     24.85  48.66416
---------+----------------------------------
   Total |        48  16.86042  39.98558

. bysort bss: means nadir
-> bss = 1

    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
       nadir | Arithmetic        5          .2              .2          .2
             |  Geometric        5          .2              .2          .2
             |   Harmonic        5          .2              .2          .2

-> bss = 2
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
       nadir | Arithmetic       13    4.830769       -1.419774    11.08131
             |  Geometric       13    1.263651        .4741617    3.367657
             |   Harmonic       13    .5441928        .3286613    1.580974

-> bss = 3
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+-------------------------------------------------------------
       nadir | Arithmetic       30       24.85        6.678504     43.0215
```

```
             |  Geometric      30    2.861879        1.191359   6.874798
             |   Harmonic      30    .5643297        .3720422   1.168005
-------------+--------------------------------------------------------------

-> bss = .
    Variable |    Type         Obs        Mean       [95% Conf. Interval]
-------------+--------------------------------------------------------------
       nadir | Arithmetic       2        4.35       -42.02765   50.72765
             |  Geometric       2    2.366432        4.49e-07   1.25e+07
             |   Harmonic       2    1.287356               .          .
-------------+--------------------------------------------------------------
Missing values in confidence intervals for harmonic mean indicate
that confidence interval is undefined for corresponding variables.
Consult Reference Manual for details.
```

Problem 3b: Linear regression on bss modeled continuously and using robust SE

```
. regress nadir bss, robust
Linear regression                                      Number of obs =      48
                                                       F(  1,    46) =    6.04
                                                       Prob > F      =  0.0179
                                                       R-squared     =  0.0632
                                                       Root MSE      =  39.121
------------------------------------------------------------------------------
             |               Robust
       nadir |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bss |   14.69526   5.981668     2.46   0.018     2.654788    26.73573
        _cons |  -20.18389   10.01486    -2.02   0.050    -40.34275   -.0250243
------------------------------------------------------------------------------
```

Problem 3c: Finding the predicted values from the regression model.
Note that Stata would really have done this for us by using the "predict" command

```
. disp  -20.18389 + 14.69526 * 1, -20.18389 + 14.69526 * 2, -20.18389 + 14.69526 * 3
-5.48863 9.20663 23.90189

. predict fitnadir
(option xb assumed; fitted values)
(2 missing values generated)

. tabstat fitnadir, stat(n mean sd) col(stat) by(bss)
Summary for variables: fitnadir
     by categories of: bss

     bss |         N      mean        sd
---------+----------------------------------
       1 |         5 -5.488626         0
       2 |        13  9.206635         0
```

```
       3 |         30   23.9019         0
   Total |         48  16.86042  10.04926


##################################################################################################
#### Problem 4
#### Comparing probability and odds of relapse using the dichotomized bone scan score: bss3
##################################################################################################

#### Problem 4c
#### Generate the descriptive statistics using tabulate.

. tabulate relapse24 bss3, col

| Key                |
|     frequency      |
| column percentage  |

           |         bss3
 relapse24 |         0          1 |     Total
         0 |        14         14 |        28
           |     77.78      46.67 |     58.33
         1 |         4         16 |        20
           |     22.22      53.33 |     41.67
     Total |        18         30 |        48
           |    100.00     100.00 |    100.00

#### Computing the odds
. disp 4/14, 16/14
.28571429 1.1428571

#### Generating CI using ci with the binomial option. These are exact CI.
. bysort bss3: ci relapse24, binomial
-> bss3 = 0
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
   relapse24 |         18    .2222222    .0979908         .064092    .4763728

-> bss3 = 1
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
   relapse24 |         30    .5333333     .091084        .3432552    .7165819
```

```
-> bss3 = .
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |          2           1           0        .1581139           1*
(*) one-sided, 97.5% confidence interval

#### Computing the odds and CI for the odds by transforming the prob and the CI for the prob
. disp .2222222/(1-.2222222), .064092/(1-.064092), .4763728/(1-.4763728)
.28571425 .06848109 .90975564

. disp .5333333/(1-.5333333), .3432552/(1-.3432552), .7165819/(1-.7165819)
1.142857 .52266147 2.5283562

#### Computing the CI for the prob by asymptotic formula (using phat * (1 - phat)
.  disp .2222222 - invnorm(0.975) * sqrt(.2222222*(1-.2222222)/18)
.03016379

.  disp .2222222 + invnorm(0.975) * sqrt(.2222222*(1-.2222222)/18)
.41428061

.  disp .5333333 - invnorm(0.975) * sqrt(.5333333*(1-.5333333)/30)
.35481193

.  disp .5333333 + invnorm(0.975) * sqrt(.5333333*(1-.5333333)/30)
.71185467

#### Computing the CI for the odds by transforming the asymptotic CI for the prob
. disp .03016379/(1-.03016379), .41428061/(1-.41428061)
.03110194 .70730219

. disp .35481193/(1-.35481193), .71185467/(1-.71185467)
.54993566 2.470471

#### Computing the CI for the odds by transforming the asymptotic CI for the log odds

.  disp exp( log(.2222222/(1-.2222222)) - invnorm(0.975) * sqrt(1/.2222222/(1-.2222222)/18) )
.09404722

.  disp exp( log(.2222222/(1-.2222222)) + invnorm(0.975) * sqrt(1/.2222222/(1-.2222222)/18) )
.86799621

.  disp exp( log(.5333333/(1-.5333333)) - invnorm(0.975) * sqrt(1/.5333333/(1-.5333333)/30) )
```

```
.55780708

.  disp exp( log(.5333333/(1-.5333333)) + invnorm(0.975) * sqrt(1/.5333333/(1-.5333333)/30) )
2.3415302

#### Problem 4d: Performing the chi square test

. cs relapse24 bss3
                 | bss3                   |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        16           4  |         20
        Noncases |        14          14  |         28
-----------------+------------------------+------------
           Total |        30          18  |         48
                 |                        |
            Risk |  .5333333    .2222222  |   .4166667
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
-----------------+------------------------+------------------------
Risk difference  |         .3111111       |    .0488969    .5733254
      Risk ratio |             2.4        |    .9499462    6.063501
 Attr. frac. ex. |         .5833333       |   -.0526912    .8350788
 Attr. frac. pop |         .4666667       |
-----------------+------------------------+------------------------
                             chi2(1) =     4.48  Pr>chi2 = 0.0343

#### Problem 4e: Performing the t test which presumes equal variances

. ttest relapse24, by(bss3)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      18    .2222222    .1008317    .4277926     .009486    .4349585
       1 |      30    .5333333    .0926411    .5074163     .343861    .7228057
---------+--------------------------------------------------------------------
combined |      48    .4166667    .0719124    .4982238    .2719976    .5613358
---------+--------------------------------------------------------------------
    diff |           -.3111111    .1429691               -.5988929   -.0233293
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.1761
Ho: diff = 0                                     degrees of freedom =       46

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0174         Pr(|T| > |t|) = 0.0347          Pr(T > t) = 0.9826

#### Problem 4f: Performing the t test which allows unequal variances

. ttest relapse24, by(bss3) unequal
Two-sample t test with unequal variances
```

```
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      18    .2222222    .1008317    .4277926     .009486    .4349585
       1 |      30    .5333333    .0926411    .5074163     .343861    .7228057
---------+--------------------------------------------------------------------
combined |      48    .4166667    .0719124    .4982238    .2719976    .5613358
---------+--------------------------------------------------------------------
    diff |           -.3111111    .1369285               -.5876891   -.0345332
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.2721
Ho: diff = 0                     Satterthwaite's degrees of freedom =    40.78

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0142         Pr(|T| > |t|) = 0.0284          Pr(T > t) = 0.9858
```

Problem 4h: Performing classical logistic regression with a binary predictor

```
. logit relapse24 bss3
Iteration 0:   log likelihood = -32.601277
Iteration 1:   log likelihood = -30.284272
Iteration 2:   log likelihood =  -30.26243
Iteration 3:   log likelihood = -30.262411

Logistic regression                               Number of obs   =         48
                                                  LR chi2(1)      =       4.68
                                                  Prob > chi2     =     0.0306
Log likelihood = -30.262411                       Pseudo R2       =     0.0717
------------------------------------------------------------------------------
   relapse24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bss3 |   1.386294   .6748011     2.05   0.040     .0637085     2.70888
       _cons |  -1.252763   .5669462    -2.21   0.027    -2.363957   -.1415689
------------------------------------------------------------------------------
```

Computing the odds of relapse for each group and the OR across groups

```
. disp exp(-1.252763), exp(-1.252763+ 1.386294),exp( 1.386294)
.28571428 1.1428567 3.9999986
```

Computing the CI for odds of relapse in the group with bss3==0

```
. disp exp(-2.363957), exp(-.1415689)
.09404734 .86799537
```

Computing the prob of relapse for each group

```
. disp .28571428 / (1+.28571428), 1.1428567/(1+1.1428567)
.22222222 .53333324
```

```
. logistic relapse24 bss3
Logistic regression                               Number of obs   =         48
                                                  LR chi2(1)      =       4.68
```

```
                                                  Prob > chi2     =     0.0306
Log likelihood = -30.262411                       Pseudo R2       =     0.0717
------------------------------------------------------------------------------
   relapse24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bss3 |          4   2.699204     2.05   0.040     1.065782    15.01246
------------------------------------------------------------------------------
```

Problem 4l: Performing logistic regression with a binary predictor and using robust SE

```
. logit relapse24 bss3, robust

Iteration 0:   log pseudolikelihood = -32.601277
Iteration 1:   log pseudolikelihood = -30.284272
Iteration 2:   log pseudolikelihood =  -30.26243
Iteration 3:   log pseudolikelihood = -30.262411

Logistic regression                               Number of obs   =         48
                                                  Wald chi2(1)    =       4.13
                                                  Prob > chi2     =     0.0421
Log pseudolikelihood = -30.262411                 Pseudo R2       =     0.0717
------------------------------------------------------------------------------
             |               Robust
   relapse24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bss3 |   1.386294   .6819416     2.03   0.042     .0497134    2.722875
       _cons |  -1.252763   .5729452    -2.19   0.029    -2.375715   -.1298109
------------------------------------------------------------------------------

. logistic relapse24 bss3, robust

Logistic regression                               Number of obs   =         48
                                                  Wald chi2(1)    =       4.13
                                                  Prob > chi2     =     0.0421
Log pseudolikelihood = -30.262411                 Pseudo R2       =     0.0717
------------------------------------------------------------------------------
             |               Robust
   relapse24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bss3 |          4   2.727766     2.03   0.042      1.05097    15.22403
------------------------------------------------------------------------------
```

Computing the CI for odds of relapse in the group with bss3==0

```
. disp exp(-2.375715), exp(-.1298109)
.09294801 .87826149
```

##
Problem 5
Comparing probability and odds of relapse modeling bone scan score continuously: bss
##

Problem 5a

```
#### Generate the descriptive statistics using ci.

. bysort bss: ci relapse24, binomial
-> bss = 1
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |          5           0           0               0    .5218238*
(*) one-sided, 97.5% confidence interval

-> bss = 2
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |         13    .3076923    .1280077        .0909204    .6142617

-> bss = 3
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |         30    .5333333     .091084        .3432552    .7165819

-> bss = .
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |          2           1           0        .1581139           1*
(*) one-sided, 97.5% confidence interval

#### Transform proportions to odds.
. disp .5218238/(1-.5218238)
1.0912793

. disp .3076923/(1-.3076923), .0909204/(1-.0909204), .6142617/(1-.6142617)
.44444443 .10001368 1.5924312

. disp .5333333/(1-.5333333), .3432552/(1-.3432552), .7165819/(1-.7165819)
1.142857 .52266147 2.5283562

#### Problem 5b: Classical logistic regression

. logit relapse24 bss

Iteration 0:   log likelihood = -32.601277
Iteration 1:   log likelihood = -29.442314
Iteration 2:   log likelihood = -29.313767
```

```
Iteration 3:   log likelihood = -29.312348
Iteration 4:   log likelihood = -29.312348

Logistic regression                               Number of obs   =         48
                                                  LR chi2(1)      =       6.58
                                                  Prob > chi2     =     0.0103
Log likelihood = -29.312348                       Pseudo R2       =     0.1009
------------------------------------------------------------------------------
   relapse24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bss |   1.310016   .5851009     2.24   0.025      .163239    2.456792
        _cons |  -3.742337   1.605597    -2.33   0.020    -6.889249   -.5954246
------------------------------------------------------------------------------
```

Problem 5c: Generate the predicted values for log odds, then odds, then proportions

```
. disp -3.742337 + 1 * 1.310016, -3.742337 + 2 * 1.310016, -3.742337 + 3 * 1.310016
-2.432321 -1.122305 .187711

. disp exp(-2.432321), exp( -1.122305), exp( .187711)
.08783274 .32552859 1.2064848

. disp .08783274/(1+.08783274), .32552859/(1+.32552859), 1.2064848/(1+1.2064848)
.08074103 .24558398 .54679044

. logistic relapse24 bss

Logistic regression                               Number of obs   =         48
                                                  LR chi2(1)      =       6.58
                                                  Prob > chi2     =     0.0103
Log likelihood = -29.312348                       Pseudo R2       =     0.1009
------------------------------------------------------------------------------
   relapse24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bss |   3.706232   2.168519     2.24   0.025     1.177318    11.66733
------------------------------------------------------------------------------
```

Problem 5d: Logistic regression using robust SE

```
. logit relapse24 bss,  robust

Iteration 0:   log pseudolikelihood = -32.601277
Iteration 1:   log pseudolikelihood = -29.442314
Iteration 2:   log pseudolikelihood = -29.313767
Iteration 3:   log pseudolikelihood = -29.312348
Iteration 4:   log pseudolikelihood = -29.312348

Logistic regression                               Number of obs   =         48
```

```
                                                  Wald chi2(1)     =       6.66
                                                  Prob > chi2      =     0.0099
Log pseudolikelihood = -29.312348                 Pseudo R2        =     0.1009
------------------------------------------------------------------------------
             |               Robust
   relapse24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bss |   1.310016    .507584     2.58   0.010     .3151693    2.304862
       _cons |  -3.742337    1.38282    -2.71   0.007    -6.452615   -1.032059
------------------------------------------------------------------------------
```

. logistic relapse24 bss, robust

```
Logistic regression                               Number of obs   =         48
                                                  Wald chi2(1)    =       6.66
                                                  Prob > chi2     =     0.0099
Log pseudolikelihood = -29.312348                 Pseudo R2       =     0.1009
------------------------------------------------------------------------------
             |               Robust
   relapse24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bss |   3.706232   1.881224     2.58   0.010     1.370491    10.02279
------------------------------------------------------------------------------
```

```
##########################################################################################
#### Problem 6
#### Comparing mean nadir PSA across groups defined by relapse within 24 months
##########################################################################################

#### Problem 6a
#### Generate the descriptive statistics and perform inference using t test that allows unequal variance.
```

. tabstat nadir, stat(n mean sd min q max) col(stat) by(relapse24)

```
Summary for variables: nadir
     by categories of: relapse24

relapse24 |         N      mean        sd       min       p25       p50       p75       max
----------+----------------------------------------------------------------------------------
        0 |        28  4.117857  17.27921        .1        .2        .2       .95        92
        1 |        22  31.94091  52.49686        .5       1.2      10.5        38       183
    Total |        50     16.36   39.2462        .1        .2       .95        10       183
```

. ttest nadir, by(relapse24) unequal

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      28    4.117857    3.265464    17.27921   -2.582322    10.81804
       1 |      22    31.94091    11.19237    52.49686    8.665104    55.21671
combined |      50       16.36    5.550251     39.2462    5.206354    27.51365
    diff |           -27.82305      11.659               -51.8556   -3.790503
------------------------------------------------------------------------------
```

```
    diff = mean(0) - mean(1)                                      t =  -2.3864
Ho: diff = 0                     Satterthwaite's degrees of freedom =  24.5887

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0125         Pr(|T| > |t|) = 0.0250          Pr(T > t) = 0.9875

##########################################################################################
#### Problem 7
#### Comparing geometric mean nadir PSA across groups defined by relapse within 24 months
##########################################################################################

#### Problem 7a
#### Generate a variable measuring log transform of nadir PSA: lnadir

. g lnadir= log(nadir)

#### Generate the descriptive statistics and perform inference using t test that allows unequal variance.

. ttest lnadir, by(relapse24) unequal

Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
       0 |      28   -.6548003    .2775986    1.468914   -1.224386    -.085215
       1 |      22    2.118018    .3992543    1.872669    1.287724    2.948313
combined |      50    .5652399    .3041384    2.150583    -.045949    1.176429
    diff |           -2.772819    .4862767               -3.756323   -1.789315
    diff = mean(0) - mean(1)                                      t =  -5.7021
Ho: diff = 0                     Satterthwaite's degrees of freedom =   39.104

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000


##########################################################################################
#### Problem 8
#### Comparing odds of relapse within 24 months across groups defined by nadir PSA
##########################################################################################

#### Problem 8: Descriptive statistics (Not requested of you, but I did it anyway)
#### Even though I will model nadir PSA continuously (either untransformed or log transformed), it
#### is useful to provide descriptive statistics of the prob and odds of relapse across
#### categories of nadir PSA (scatterplots just don't cut it with binary data)

#### First find a suitable categorization. I will look for categories that are in a sense multiplicative:
```

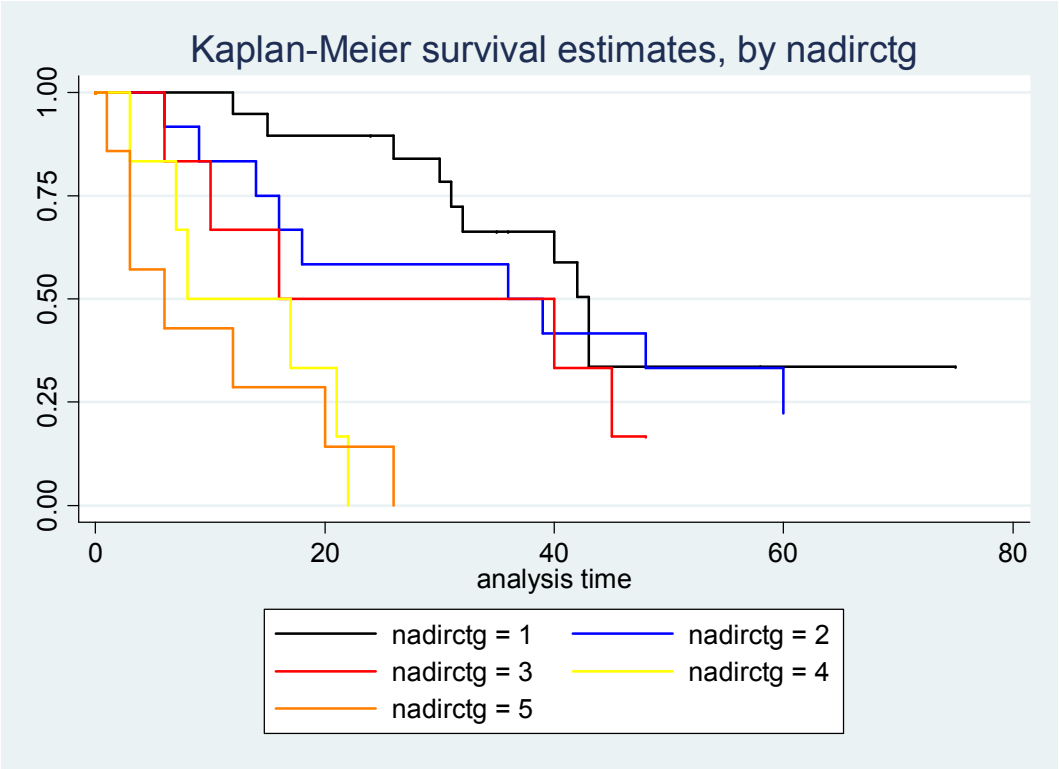
```
#### A constant ratio between the endpoints (more or less). It takes a couple tries to find the
#### categories that have some scientific interpretation, but still have sufficient numbers of
#### subjects to be able to get meaningful statistics. I eventually choose to have categories that
#### each range over an approximate four fold difference: 0.1-0.5 0.5-2.0 2.0-8.0 8.0-32.0 32.0-183.0

. tabstat nadir, stat(n mean sd min q max)

    variable |         N      mean        sd       min       p25       p50       p75       max
-------------+----------------------------------------------------------------------------------
       nadir |        50     16.36   39.2462        .1        .2       .95        10       183
----------------------------------------------------------------------------------------------

. g nadirctg= nadir
. recode nadirctg 0/1=1 1/2=2 2/4=4 4/max=8
(nadirctg: 48 changes made)

. tabstat nadir, stat(n mean sd min q max) by(nadirctg) col(stat)
Summary for variables: nadir
     by categories of: nadirctg
nadirctg |         N      mean        sd       min       p25       p50       p75       max
---------+----------------------------------------------------------------------------------
       1 |        26        .4  .2814249        .1        .2        .2        .7         1
       2 |         5      1.46  .2880972       1.1       1.2       1.6       1.7       1.7
       4 |         2      2.55  .4949748       2.2       2.2      2.55       2.9       2.9
       8 |        17  46.77647   56.8047       5.2        10        16        52       183
   Total |        50     16.36   39.2462        .1        .2       .95        10       183

. drop nadirctg
. g nadirctg= nadir
. recode nadirctg min/0.5=1 0.5/2=2 2/8=3 8/32=4 32/max=5
(nadirctg: 50 changes made)

. tabstat nadir, stat(n mean sd min q max) by(nadirctg) col(stat)
Summary for variables: nadir
     by categories of: nadirctg
nadirctg |         N      mean        sd       min       p25       p50       p75       max
---------+----------------------------------------------------------------------------------
       1 |        19  .2473684  .1172292        .1        .2        .2        .2        .5
       2 |        12  1.083333  .3857303        .7       .75       .95       1.4       1.7
       3 |         6  5.216667  2.280716       2.2       2.9       5.6         7         8
       4 |         6  15.16667  6.177918        10        11      13.5        16        27
       5 |         7  96.85714  59.73393        38        40        92       169       183
   Total |        50     16.36   39.2462        .1        .2       .95        10       183
```

```
#### Generate prob and odds for each category
. bysort nadirctg: ci relapse24, binomial
-> nadirctg = 1
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |         19    .1052632    .0704059        .0130122    .3313767

-> nadirctg = 2
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |         12    .4166667    .1423188        .1516522    .7233303

-> nadirctg = 3
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |          6          .5    .2041241        .1181172    .8818828

-> nadirctg = 4
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |          6           1           0        .5407419           1*
 (*) one-sided, 97.5% confidence interval

-> nadirctg = 5
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
   relapse24 |          7    .8571429      .13226        .4212768    .9963897

. disp .1052632/(1-.1052632), .0130122/(1-.0130122), .3313767/(1-.3313767)
.11764711 .01318375 .49561046

. disp  .4166667/(1- .4166667), .1516522/(1-.1516522), .7233303/(1-.7233303)
.71428581 .17876182 2.6144182

. disp .5/(1-.5), .1181172/(1-.1181172), .8818828/(1-.8818828)
1 .13393753 7.4661675

. disp .5407419/(1-.5407419)
1.1774249

. disp .8571429/(1-.8571429), .4212768/(1-.4212768), .9963897/(1-.9963897)
6.0000021 .72794179 275.98529
```

```
#### Problem 8a: Logistic regression modeling nadir continuously and untransformed
#### I will use logistic regression with robust SE to answer the question, but I also
#### use classical logistic regression just to illustrate the differences

. logit relapse24 nadir, robust
Iteration 0:   log pseudolikelihood =  -34.29649
Iteration 1:   log pseudolikelihood = -30.762407
Iteration 2:   log pseudolikelihood =  -30.19455
Iteration 3:   log pseudolikelihood = -30.060232
Iteration 4:   log pseudolikelihood = -30.051269
Iteration 5:   log pseudolikelihood = -30.051225

Logistic regression                               Number of obs   =         50
                                                  Wald chi2(1)    =       0.73
                                                  Prob > chi2     =     0.3914
Log pseudolikelihood = -30.051225                 Pseudo R2       =     0.1238
------------------------------------------------------------------------------
             |               Robust
   relapse24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |   .0407059     .04749     0.86   0.391    -.0523728    .1337845
       _cons |  -.6762589   .3681406    -1.84   0.066    -1.397801    .0452834
------------------------------------------------------------------------------

. logistic relapse24 nadir, robust
Logistic regression                               Number of obs   =         50
                                                  Wald chi2(1)    =       0.73
                                                  Prob > chi2     =     0.3914
Log pseudolikelihood = -30.051225                 Pseudo R2       =     0.1238
------------------------------------------------------------------------------
             |               Robust
   relapse24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |   1.041546    .049463     0.86   0.391     .9489751    1.143146
------------------------------------------------------------------------------

. logistic relapse24 nadir
Logistic regression                               Number of obs   =         50
                                                  LR chi2(1)      =       8.49
                                                  Prob > chi2     =     0.0036
Log likelihood = -30.051225                       Pseudo R2       =     0.1238
------------------------------------------------------------------------------
   relapse24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |   1.041546   .0244315     1.74   0.083     .9947449    1.090548
------------------------------------------------------------------------------


#### Problem 8b: Logistic regression modeling nadir continuously after log transformation
#### I will use logistic regression with robust SE to answer the question, but I also
```

```
#### use classical logistic regression just to illustrate the differences

. logit relapse24 lnadir, robust
Iteration 0:   log pseudolikelihood =  -34.29649
Iteration 1:   log pseudolikelihood =  -22.83943
Iteration 2:   log pseudolikelihood = -22.065678
Iteration 3:   log pseudolikelihood = -22.031786
Iteration 4:   log pseudolikelihood = -22.031677

Logistic regression                               Number of obs   =         50
                                                  Wald chi2(1)    =      10.55
                                                  Prob > chi2     =     0.0012
Log pseudolikelihood = -22.031677                 Pseudo R2       =     0.3576
------------------------------------------------------------------------------
             |               Robust
   relapse24 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnadir |   .8913371   .2744225     3.25   0.001     .3534789    1.429195
       _cons |  -.7109372   .3677853    -1.93   0.053    -1.431783    .0099087
------------------------------------------------------------------------------

#### Computing the odds ratio for a doubling of nadir PSA
. disp 2^.8913371, 2^.3534789 , 2^1.429195
1.8548945 1.2776378 2.6929641

. logistic relapse24 lnadir, robust
Logistic regression                               Number of obs   =         50
                                                  Wald chi2(1)    =      10.55
                                                  Prob > chi2     =     0.0012
Log pseudolikelihood = -22.031677                 Pseudo R2       =     0.3576
------------------------------------------------------------------------------
             |               Robust
   relapse24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnadir |   2.438388   .6691485     3.25   0.001     1.424013    4.175338
------------------------------------------------------------------------------

#### Computing the odds ratio for a doubling of nadir PSA (agrees with above from logit)
. disp 2.438388^log(2),  1.424013^log(2),    4.175338^log(2)
1.8548945 1.2776378 2.6929647

. logistic relapse24 lnadir
Logistic regression                               Number of obs   =         50
                                                  LR chi2(1)      =      24.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -22.031677                       Pseudo R2       =     0.3576
------------------------------------------------------------------------------
   relapse24 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnadir |   2.438388   .5879702     3.70   0.000     1.520029    3.911593
------------------------------------------------------------------------------
```

```
##########################################################################################
#### Problem 9
#### Comparing hazard of relapse across groups defined by nadir PSA
##########################################################################################

#### Problem 9a: Descriptive statistics

. stset obstime relapse
     failure event:  relapse != 0 & relapse < .
obs. time interval:  (0, obstime]
 exit on or before:  failure
       50  total obs.
        0  exclusions
       50  obs. remaining, representing
       36  failures in single record/single failure data
     1423  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        75

. sts graph, by(nadirctg) col(black blue red yellow orange)
         failure _d:  relapse
   analysis time _t:  obstime
```




```
. sts list, by(nadirctg) at(12 24 36 48)
         failure _d:  relapse
   analysis time _t:  obstime

              Beg.                      Survivor      Std.
    Time      Total      Fail            Function     Error      [95% Conf. Int.]
-------------------------------------------------------------------------------
nadirctg=1
      12         19         1             0.9474    0.0512     0.6812    0.9924
      24         17         1             0.8947    0.0704     0.6408    0.9726
      36         10         4             0.6624    0.1130     0.3955    0.8329
      48          3         3             0.3365    0.1626     0.0744    0.6335
nadirctg=2
      12         11         2             0.8333    0.1076     0.4817    0.9555
      24          8         3             0.5833    0.1423     0.2701    0.8009
      36          7         1             0.5000    0.1443     0.2085    0.7361
```

```
      48         5        2              0.3333    0.1361     0.1027    0.5884
nadirctg=3
      12         5        2              0.6667    0.1925     0.1946    0.9044
      24         4        1              0.5000    0.2041     0.1109    0.8037
      36         4        0              0.5000    0.2041     0.1109    0.8037
      48         1        2              0.1667    0.1521     0.0077    0.5168
nadirctg=4
      12         4        3              0.5000    0.2041     0.1109    0.8037
      24         1        3                   .         .          .         .
      36         1        0                   .         .          .         .
      48         1        0                   .         .          .         .
nadirctg=5
      12         3        5              0.2857    0.1707     0.0411    0.6115
      24         2        1              0.1429    0.1323     0.0071    0.4649
      36         1        1                   .         .          .         .
      48         1        0                   .         .          .         .
Note:  Survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.

. stci , by(nadirctg) p(75)
         failure _d:  relapse
   analysis time _t:  obstime

             |    no. of
nadirctg     |  subjects         75%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
           1 |        19           .            .           42           .
           2 |        12          60            .           36           .
           3 |         6          45     2.102052           10           .
           4 |         6          21     1.675193            7           .
           5 |         7          20     3.394819            3           .
       total |        50          48     9.337316           40           .

. stci , by(nadirctg) p(50)
         failure _d:  relapse
   analysis time _t:  obstime

             |    no. of
nadirctg     |  subjects         50%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
           1 |        19          43      .8370397           31           .
           2 |        12          36      8.958518            9          60
           3 |         6          16      6.029328            6           .
           4 |         6           8      1.629969            3           .
```

```
           5 |         7           6      1.030499            1          20
       total |        50          30      6.644128           17          40

. stci , by(nadirctg) p(25)
         failure _d:  relapse
   analysis time _t:  obstime

             |    no. of
nadirctg     |  subjects         25%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
           1 |        19          31      2.419997           12          42
           2 |        12          14      2.666917            6          36
           3 |         6          10      1.405999            6          40
           4 |         6           7      1.244302            3          17
           5 |         7           3      .2685666            1           6
       total |        50          12      3.721935            6          18


#### Problem 9b: Prop Hzd regression modeling nadir continuously and untransformed
#### I will use PH regression with robust SE to answer the question, but I also
#### use classical PH regression just to illustrate the differences

. stcox nadir
         failure _d:  relapse
   analysis time _t:  obstime

Iteration 0:   log likelihood = -118.96593
Iteration 1:   log likelihood = -113.88229
Iteration 2:   log likelihood = -113.29294
Iteration 3:   log likelihood = -113.28918
Iteration 4:   log likelihood = -113.28918
Refining estimates:
Iteration 0:   log likelihood = -113.28918

Cox regression -- Breslow method for ties

No. of subjects =           50                     Number of obs   =         50
No. of failures =           36
Time at risk    =         1423
                                                   LR chi2(1)      =      11.35
Log likelihood  =   -113.28918                     Prob > chi2     =     0.0008
------------------------------------------------------------------------------
           _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |    1.01551   .0038122     4.10   0.000     1.008066     1.02301
```

```
. stcox nadir, robust
         failure _d:  relapse
   analysis time _t:  obstime

Iteration 0:   log pseudolikelihood = -118.96593
Iteration 1:   log pseudolikelihood = -113.88229
Iteration 2:   log pseudolikelihood = -113.29294
Iteration 3:   log pseudolikelihood = -113.28918
Iteration 4:   log pseudolikelihood = -113.28918
Refining estimates:
Iteration 0:   log pseudolikelihood = -113.28918

Cox regression -- Breslow method for ties

No. of subjects      =           50                 Number of obs   =        50
No. of failures      =           36
Time at risk         =         1423
                                                    Wald chi2(1)    =     16.79
Log pseudolikelihood =   -113.28918                 Prob > chi2     =    0.0000
             |                Robust
           t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
       nadir |    1.01551   .0038139     4.10   0.000     1.008063    1.023013
```

Problem 9c: Prop Hzd regression modeling nadir continuously after log transformation
I will use PH regression with robust SE to answer the question, but I also
use classical PH regression just to illustrate the differences

```
. stcox lnadir
         failure _d:  relapse
   analysis time _t:  obstime

Iteration 0:   log likelihood = -118.96593
Iteration 1:   log likelihood = -108.18837
Iteration 2:   log likelihood = -107.32052
Iteration 3:   log likelihood = -107.31899
Refining estimates:
Iteration 0:   log likelihood = -107.31899

Cox regression -- Breslow method for ties
```

```
No. of subjects =           50                     Number of obs   =        50
No. of failures =           36
Time at risk    =         1423
                                                   LR chi2(1)      =     23.29
Log likelihood  =   -107.31899                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnadir |   1.535399   .1339713     4.91   0.000     1.294046    1.821767
------------------------------------------------------------------------------
```

```
. stcox lnadir, robust
         failure _d:  relapse
   analysis time _t:  obstime

Iteration 0:   log pseudolikelihood = -118.96593
Iteration 1:   log pseudolikelihood = -108.18837
Iteration 2:   log pseudolikelihood = -107.32052
Iteration 3:   log pseudolikelihood = -107.31899
Refining estimates:
Iteration 0:   log pseudolikelihood = -107.31899

Cox regression -- Breslow method for ties

No. of subjects      =           50                Number of obs   =        50
No. of failures      =           36
Time at risk         =         1423
                                                   Wald chi2(1)    =     34.04
Log pseudolikelihood =   -107.31899                Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnadir |   1.535399   .1128444     5.83   0.000     1.329419    1.773293
------------------------------------------------------------------------------
```