LOGISTIC REGRESSION
Joseph Hilbe

PROJECT: Sample Answer

Data to be used for model: lr_model.

Response is *heartatk*, patient had a heart attack (1/0).

strata: 1-32 levels

race1-3 1:white; 2=black; 3=other

region (*reg1-4*) 1=NE; 2=MidW; 3= S; 4=W

hsizgp (*hsi1-5*) # in household; 5 if >=5

gender is the 1/0 version of sex (1/2), and is in the preferred format.

```
. d

Contains data from c:\data\lr_model.dta
  obs:        10,351
 vars:            29                          25 May 2005 19:40
 size:       517,550 (95.1% of memory free)
------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------
heartatk        byte   %8.0g
strata          byte   %8.0g
region          byte   %8.0g
sex             byte   %8.0g
race            byte   %8.0g
age             byte   %8.0g
height          float  %9.0g
weight          float  %9.0g
bpsystol        int    %8.0g
tcresult        int    %8.0g
```

```
tgresult        int    %8.0g
hdresult        int    %8.0g
hgb             float  %9.0g
diabetes        byte   %8.0g                 1/0
albumin         float  %9.0g
hsizgp          byte   %8.0g
race1           byte   %8.0g                 race== 1.0000
race2           byte   %8.0g                 race== 2.0000
race3           byte   %8.0g                 race== 3.0000
reg1            byte   %8.0g                 region== 1.0000
reg2            byte   %8.0g                 region== 2.0000
reg3            byte   %8.0g                 region== 3.0000
reg4            byte   %8.0g                 region== 4.0000
gender          int    %8.0g                 1/0
hsi1            byte   %8.0g                 hsizgp== 1.0000
hsi2            byte   %8.0g                 hsizgp== 2.0000
hsi3            byte   %8.0g                 hsizgp== 3.0000
hsi4            byte   %8.0g                 hsizgp== 4.0000
hsi5            byte   %8.0g                 hsizgp== 5.0000
------------------------------------------------------------
```

-Perform a univariate logit model on prospective explanatory predictors.

UNIVARIATE LOGIT MODEL ON BINARY AND CONTINUOUS PREDICTORS

Binary : diabetes (1=yes), gender (1-male)

Continuous: other covariates in univariate model below.

```
. glm heartatk age height weight bpsystol tcresult hdresult hgb diabetes albumin
hsizgp sex, fam(bin) nolog

Generalized linear models                         No. of obs      =      8500
Optimization     : ML                             Residual df     =      8488
                                                  Scale parameter =         1
Deviance         =  2511.550617                   (1/df) Deviance =  .2958943
Pearson          =  6701.307213                   (1/df) Pearson  =  .7895037

Variance function: V(u) = u*(1-u/1)               [Binomial]
Link function    : g(u) = ln(u/(1-u))             [Logit]
                                                  AIC             =  .2983001
Log likelihood   = -1255.775308                   BIC             = -74286.36
------------------------------------------------------------------------------
             |                 OIM
    heartatk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0829698   .0066466    12.48   0.000     .0699428    .0959969
      height |  -.0060337   .0093518    -0.65   0.519    -.0243629    .0122956
      weight |   .0015754   .0047375     0.33   0.739      -.00771    .0108608
    bpsystol |  -.0025459   .0025014    -1.02   0.309    -.0074485    .0023567
    tcresult |   .0027107   .0011988     2.26   0.024     .0003611    .0050603
    hdresult |  -.0229772   .0047715    -4.82   0.000    -.0323292   -.0136252
         hgb |   .0252234   .0475914     0.53   0.596    -.0680539    .1185008
    diabetes |   .3244255   .1932492     1.68   0.093     -.054336     .703187
     albumin |  -.2810867   .1977556    -1.42   0.155    -.6686805     .106507
      hsizgp |    .005208   .0563548     0.09   0.926    -.1052454    .1156613
         sex |  -.8643173   .1760941    -4.91   0.000    -1.209455   -.5191791
       _cons |  -3.942782   2.110771    -1.87   0.062    -8.079818    .1942536
------------------------------------------------------------------------------
```

```
. glm heartatk age height weight bpsystol tcresult hdresult hgb diabetes albumin
hsizgp sex, fam(bin) nolog eform

Generalized linear models                         No. of obs      =      8500
Optimization     : ML                             Residual df     =      8488
                                                  Scale parameter =         1
Deviance         =  2511.550617                   (1/df) Deviance =  .2958943
Pearson          =  6701.307213                   (1/df) Pearson  =  .7895037

Variance function: V(u) = u*(1-u/1)               [Binomial]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  .2983001
Log likelihood   = -1255.775308                   BIC             = -74286.36
------------------------------------------------------------------------------
             |                 OIM
    heartatk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.086509   .0072216    12.48   0.000     1.072447    1.100756
      height |   .9939845   .0092956    -0.65   0.519     .9759315    1.012372
      weight |   1.001577    .004745     0.33   0.739     .9923196     1.01092
    bpsystol |   .9974573    .002495    -1.02   0.309     .9925791    1.002359
    tcresult |   1.002714    .001202     2.26   0.024     1.000361    1.005073
    hdresult |   .9772848   .0046631    -4.82   0.000     .9681878    .9864672
         hgb |   1.025544    .048807     0.53   0.596     .9342101    1.125808
    diabetes |   1.383236   .2673092     1.68   0.093     .9471139    2.020181
     albumin |   .7549629   .1492981    -1.42   0.155     .5123842    1.112386
      hsizgp |   1.005222   .0566491     0.09   0.926     .9001036    1.122616
         sex |   .4213391   .0741953    -4.91   0.000     .2983597    .5950088
       _cons |   .0193942   .0409367    -1.87   0.062     .0003097    1.214404
------------------------------------------------------------------------------
```

UNIVARIATE LOGIT MODEL ON CATEGORICAL PREDICTORS

RACE 1=WHITE, 2=BLACK, 3=OTHER

```
.logit heartatk i.race, or nolog

Logistic regression                                     Number of obs   =      10349
                                                        LR chi2(2)      =       2.71
                                                        Prob > chi2     =     0.2583
Log likelihood = -1929.2401                             Pseudo R2       =     0.0007

------------------------------------------------------------------------------
    heartatk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |
      Black  |   .9216795   .1448923    -0.52   0.604     .6772782    1.254275
      Other  |   .5224359   .2380396    -1.42   0.154      .213893    1.276055
             |
       _cons |   .0490798   .0024413   -60.60   0.000     .0445207    .0541057
------------------------------------------------------------------------------
```

REGION 1=NE, 2= MW, 3=S, 4=W

```
. logit heartatk i.region, or nolog

Logistic regression                                     Number of obs   =      10349
                                                        LR chi2(3)      =      10.28
                                                        Prob > chi2     =     0.0164
Log likelihood = -1925.4552                             Pseudo R2       =     0.0027

------------------------------------------------------------------------------
    heartatk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
      region |
          MW |   1.195755   .1778586     1.20   0.229     .8933736    1.600484
           S |   1.261286   .1848699     1.58   0.113     .9463479    1.681034
           W |    1.55282   .2233489     3.06   0.002     1.171356     2.05851
             |
       _cons |   .0381566   .0044305   -28.13   0.000     .0303901    .0479078
------------------------------------------------------------------------------
```

HIS – NUMBER IN HOUSEHOLD

```
. logit heartatk i.hsizgp, or nolog

Logistic regression                               Number of obs   =      10349
                                                  LR chi2(4)      =     130.86
                                                  Prob > chi2     =     0.0000
Log likelihood = -1865.1636                       Pseudo R2       =     0.0339

------------------------------------------------------------------------------
    heartatk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hsizgp |
           2 |   1.859421   .2511039     4.59   0.000     1.427013    2.422855
           3 |   .8652179   .1511619    -0.83   0.407     .6143451    1.218537
           4 |   .4854752   .1032751    -3.40   0.001     .3199565    .7366193
           5 |   .4282782   .0930159    -3.90   0.000     .2798062    .6555332
             |
       _cons |   .0440821   .0052719   -26.10   0.000      .034871    .0557263
------------------------------------------------------------------------------
```

After eliminating predictors that fail to contribute to understanding having a heart attack, the tentative final model appears as:

Age **age in years: 20-74**

tcresult **serum cholesterol (mg/dL)**

hdresult **high density lipids (mg/dL)**

```
. logit heartatk age tcresult hdresult diabetes i.sex i.region , or nolog

Logistic regression                             Number of obs   =       8718
                                                LR chi2(8)      =     542.57
                                                Prob > chi2     =     0.0000
Log likelihood = -1281.5632                     Pseudo R2       =     0.1747

------------------------------------------------------------------------------
    heartatk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.086874   .0059781    15.15   0.000      1.07522    1.098654
    tcresult |   1.002727   .0011498     2.38   0.018     1.000476    1.004983
    hdresult |   .9754325   .0044938    -5.40   0.000     .9666646    .9842801
    diabetes |   1.471177   .2728365     2.08   0.037     1.022835    2.116041
             |
         sex |
      Female |   .4513925   .0549022    -6.54   0.000     .3556509    .5729078
             |
      region |
          MW |   1.455976   .2445512     2.24   0.025     1.047571    2.023603
           S |   1.373518   .2336904     1.87   0.062     .9840358    1.917158
           W |   1.477689    .246818     2.34   0.019     1.065141    2.050025
             |
       _cons |   .0007175   .0003383   -15.35   0.000     .0002847    .0018079
------------------------------------------------------------------------------
```

Employing strata as a cluster effect, the results are:

```
. logit heartatk age tcresult hdresult diabetes i.sex i.region , or nolog
cluster(strata)
```

```
Logistic regression                                     Number of obs   =       8718
                                                        Wald chi2(8)    =     724.29
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1281.5632                       Pseudo R2       =     0.1747

                                (Std. Err. adjusted for 31 clusters in strata)
------------------------------------------------------------------------------
             |               Robust
    heartatk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.086874   .0051563    17.56   0.000     1.076814    1.097027
    tcresult |   1.002727   .0012412     2.20   0.028     1.000297    1.005163
    hdresult |   .9754325    .005993    -4.05   0.000     .9637569    .9872496
    diabetes |   1.471177   .2706587     2.10   0.036     1.025807     2.10991
             |
         sex |
      Female |   .4513925   .0445654    -8.06   0.000     .3719772    .5477626
             |
      region |
          MW |   1.455976   .2071615     2.64   0.008     1.101647    1.924271
           S |   1.373518   .1107238     3.94   0.000     1.172779    1.608616
           W |   1.477689   .1141462     5.05   0.000      1.27008    1.719236
             |
       _cons |   .0007175   .0003324   -15.63   0.000     .0002894    .0017788
------------------------------------------------------------------------------
```

We first determine if the model is specified properly.

```
. linktest

Iteration 0:   log likelihood = -1552.8477
Iteration 1:   log likelihood = -1481.3895
Iteration 2:   log likelihood = -1291.0283
```

```
Iteration 3:   log likelihood = -1278.8476
Iteration 4:   log likelihood = -1278.7873
Iteration 5:   log likelihood = -1278.7873

Logistic regression                                     Number of obs   =       8718
                                                        LR chi2(2)      =     548.12
                                                        Prob > chi2     =     0.0000
Log likelihood = -1278.7873                             Pseudo R2       =     0.1765

------------------------------------------------------------------------------
    heartatk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .4703558   .2406071     1.95   0.051    -.0012254    .9419369
      _hatsq |   -.091053   .0411849    -2.21   0.027     -.171774   -.0103321
       _cons |  -.6633379   .3257353    -2.04   0.042    -1.301767   -.0249084
------------------------------------------------------------------------------
```

The model is now evaluated using Hosmer-Lemeshow GOF test

```
. estat gof, table group(10)

Logistic model for heartatk, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0019 |     0 |   1.1 |   872 | 870.9 |   872 |
  |     2 | 0.0033 |     1 |   2.2 |   871 | 869.8 |   872 |
  |     3 | 0.0055 |     1 |   3.7 |   871 | 868.3 |   872 |
  |     4 | 0.0097 |     6 |   6.4 |   866 | 865.6 |   872 |
  |     5 | 0.0180 |    11 |  11.7 |   860 | 859.3 |   871 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.0323 |    25 |  21.4 |   847 | 850.6 |   872 |
```

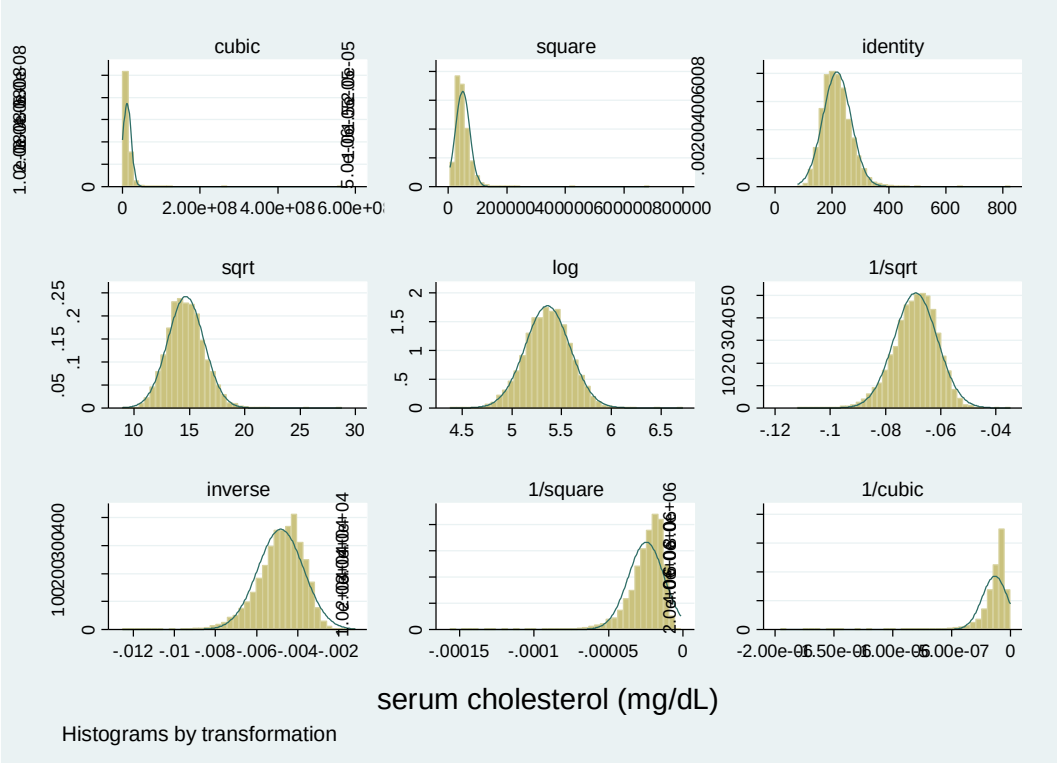
```
|      7 | 0.0514 |     44 |  36.0 |    828 | 836.0 |    872 |
|      8 | 0.0789 |     57 |  55.8 |    815 | 816.2 |    872 |
|      9 | 0.1233 |     85 |  85.1 |    787 | 786.9 |    872 |
|     10 | 0.4705 |    147 | 153.5 |    724 | 717.5 |    871 |
+-------------------------------------------------------------+

      number of observations =      8718
            number of groups =        10
     Hosmer-Lemeshow chi2(8) =      6.69
                Prob > chi2 =      0.5704
```

The model as presented appears to be well fitted.

TEST FOR NORMALITY OF CONTINUOUS PREDICTORS

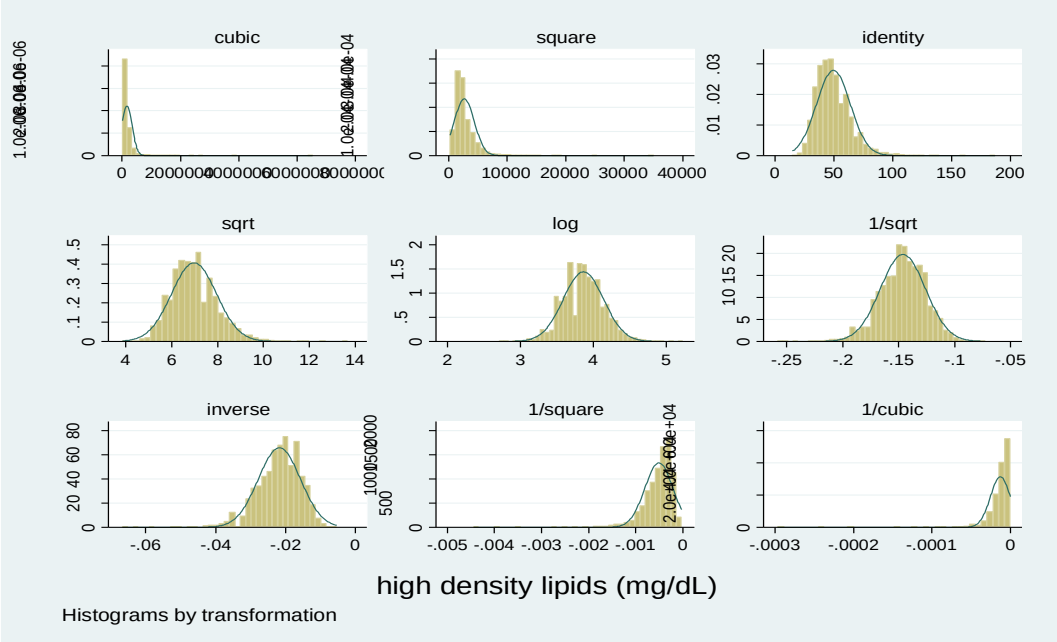
TCRESULT

`gladder tcresult`



HDRESULT

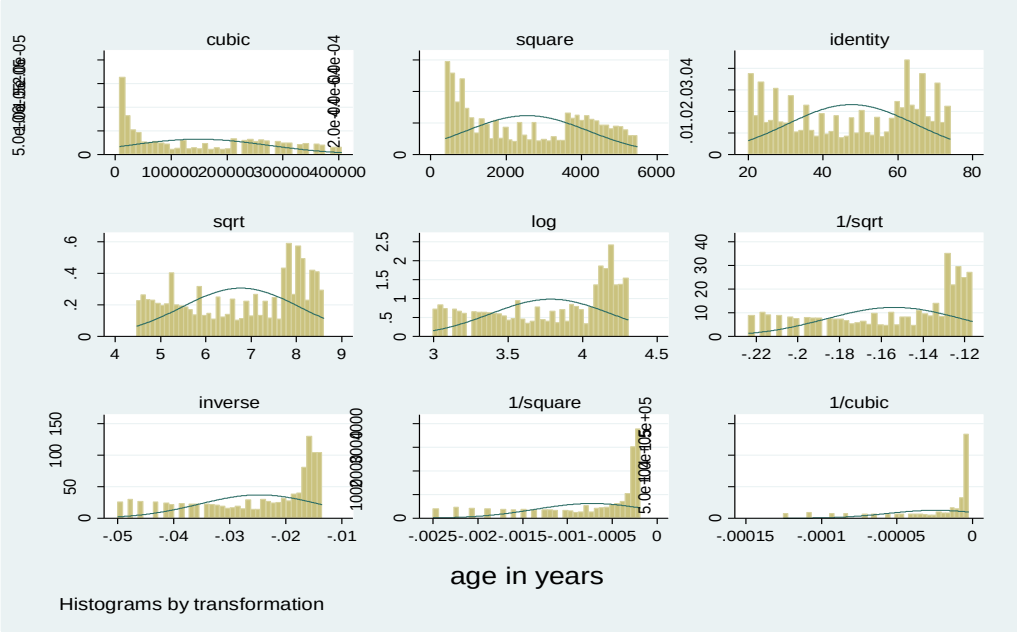
`gladder hdresult`



Both tcresult and hdresult are approximately normal, which assists in model interpretation.

AGE

gladder age



Age is clearly not normal, but remember, it does not need to as long as it is linear in the logit. However, because of the somewhat U shape it might be wise to categorize age into levels.

```
. gen byte agegrp =1 if age<40
(6409 missing values generated)

. replace agegrp = 2 if age>=40 & age<60
(2563 real changes made)

. replace agegrp = 3 if age>=60 & age !=.
(3846 real changes made)

. tab agegrp, gen(agegrp)

     agegrp |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      3,942       38.08       38.08
          2 |      2,563       24.76       62.84
          3 |      3,846       37.16      100.00
------------+-----------------------------------
      Total |     10,351      100.00
```

MODEL WITH AGE CATEGORIZED INTO 3 LEVELS

```
. logit heartatk agegrp2 agegrp3  tcresult hdresult diabetes i.sex i.region , or nolog
cluster(strata)

Logistic regression                                     Number of obs   =       8718
                                                        Wald chi2(9)    =     417.34
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -1282.5092                       Pseudo R2       =     0.1741

                                (Std. Err. adjusted for 31 clusters in strata)
```

```
------------------------------------------------------------------------------
             |               Robust
    heartatk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     agegrp2 |   20.76892    11.7144     5.38   0.000     6.875532    62.73666
     agegrp3 |   66.48464   34.51542     8.08   0.000     24.03373    183.9169
    tcresult |   1.002269   .0012284     1.85   0.064     .9998641    1.004679
    hdresult |   .9756797   .0061217    -3.92   0.000     .9637549    .9877521
    diabetes |    1.55518    .279145     2.46   0.014     1.093944    2.210884
             |
         sex |
      Female |   .4646493   .0448208    -7.95   0.000     .3846069    .5613496
             |
      region |
          MW |   1.461241   .2027189     2.73   0.006     1.113357    1.917828
           S |   1.366195   .1048544     4.07   0.000     1.175395    1.587968
           W |    1.53215   .1177232     5.55   0.000      1.31795    1.781163
             |
       _cons |   .0029942   .0018272    -9.52   0.000     .0009054    .0099017
------------------------------------------------------------------------------
```

The odds ratios of the age group levels are extraordinarily high, with a corresponding high standard error. This indicates that there is a problem. It is likely that there are too few values in one or more of the cells relating the response to levels of agegrp. This can be determined by tabulating the two variables.

```
. tab heartatk agegrp

     heart |
   attack, |
    1=yes, |              agegrp
      0=no |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |     3,936      2,473      3,464 |     9,873
         1 |         5         89        382 |       476
```

```
-----------+---------------------------------+----------
     Total |     3,941      2,562      3,846 |    10,349
```

It is obvious what the problem is – only 5 successes out of a near 4,000 patients in the reference level. The percentages can be determined by:

```
. tab heartatk agegrp, col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

     heart |
   attack, |
    1=yes, |              agegrp
      0=no |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |     3,936      2,473      3,464 |     9,873
           |     99.87      96.53      90.07 |     95.40
-----------+---------------------------------+----------
         1 |         5         89        382 |       476
           |      0.13       3.47       9.93 |      4.60
-----------+---------------------------------+----------
     Total |     3,941      2,562      3,846 |    10,349
           |    100.00     100.00     100.00 |    100.00
```

The 5 1's in the agegrp1 level represent only .13% of the cases in agegrp1. Even agegrp2 has only 89, of 3.5% 1's. Since age was, in conjunction with the other predictors, linear in the logit, and when included in the model resulted in a well fitted model, we keep it as it is given in the data.

INTERPRETATION

The model shown above appears to be a well fitted model. Displayed again, with adjustment for collection over strata, we have:

heartatk	Odds Ratio	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.086874	.0051563	17.56	0.000	1.076814	1.097027
tcresult	1.002727	.0012412	2.20	0.028	1.000297	1.005163
hdresult	.9754325	.005993	-4.05	0.000	.9637569	.9872496
diabetes	1.471177	.2706585	2.10	0.036	1.025807	2.10991
gender	2.215367	.21872	8.06	0.000	1.825609	2.688336
reg2	1.455976	.2071612	2.64	0.008	1.101647	1.924271
reg3	1.373518	.1107236	3.94	0.000	1.172779	1.608616
reg4	1.477689	.114146	5.05	0.000	1.27008	1.719235

AGE (OR=1.0868)

For a one unit increase in age there is a near 9% increase in the odds of a patient in the study having a heart attack, the other predictors held constant.

TCRESULT (OR=1.0027)

For a one unit increase in serum cholesterol per mg in the blood of a patient in the study, there is a .3% increase in the odds of having a heart attack, with the other predictors held constant.

HDRESULT (OR=0.9754)

For a one unit increase in high density lipids per mg in the blood of a patient in the study, there is a 2.5% decrease in the odds of having a heart attack, with the other predictors held constant.

DIABETES (OR=1.4712)

There is a 47% increase in the odds of having a heart attack for those patients in the study who have diabetes compared to those not having diabetes.

GENDER (OR=2.2154)

There is a two-and-a-fifth times greater odds of a male having a heart attack among those in the study compared to a female. That is, men have over twice the odds of having a heart attack than women in the study.

REG2 [MidWest] (OR=1.4560)

Those patients in the study living in the Mid West have a some 46% greater odds of having a heart attack than patients living in the North East.

REG3 [South] (OR=1.3735)

Those patients in the study living in the South have a some 37% greater odds of having a heart attack than patients living in the North East.

REG4 [West] (OR=1.4777)

Those patients in the study living in the West have a some 48% greater odds of having a heart attack than patients living in the North East.

PREDICTION

To determine the pattern of model predictors that yield the greatest probability of a patient in the study having a heart attack, we first must predict the probability of having a heart attack for all patients.

```
. predict mu

(option pr assumed; Pr(heartatk))

(1633 missing values generated)
```

To determine the high, low, and mean/median values of mu, the predicted probability of having a heart attack, summarize the data using the detail option.

```
. su mu, detail

                        Pr(heartatk)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0004254       .0000372
 5%     .0006197       .0001473
10%      .000793       .0001825       Obs                8718
25%     .0016821       .0001999       Sum of Wgt.        8718

50%     .0230036                      Mean           .0432439
                        Largest       Std. Dev.      .0523726
75%     .0654579       .2841254
90%     .1269105       .2982205       Variance       .0027429
95%     .1576736       .3215641       Skewness       1.451294
99%     .2015226        .376534       Kurtosis       4.760525
```

LEAST LIKELY TO HAVE A HEART ATTACK

```
. l age tcresult hdresult diabetes sex region if mu<= 0.0002027

      +-----------------------------------------------------+
      | age   tcresult   hdresult   diabetes      sex   region |
      |-----------------------------------------------------|
 6540. |  21        201        101          0   Female       NE |
 6941. |  37        266        187          0   Female       MW |
 7890. |  29        223        112          0   Female        S |
 9069. |  22        331        139          0   Female        W |
      +-----------------------------------------------------+
```

List the values of the predictors for just a bit higher than the lowest value of *mu*.

```
. l age tcresult hdresult diabetes sex region if mu<= 0.0002027

      +-------------------------------------------------------+
      | age   tcresult   hdresult   diabetes      sex   region |
      |-------------------------------------------------------|
 6540. |  21        201        101          0   Female       NE |
 6941. |  37        266        187          0   Female       MW |
 7890. |  29        223        112          0   Female        S |
 9069. |  22        331        139          0   Female        W |
      +-------------------------------------------------------+
```

A 37 year old non-diabetic female patient in the study living in the Mid West, with a serum cholesterol count of 266 and high density lipid count of 187 has the least likelihood of having a heart attack.

MOST LIKELY TO HAVE A HEART ATTACK

List the values of the predictors for just a bit lower than the highest value of *mu*. Since Stata records missing values as high, it is necessary to explicitly exclude them from the listing.

```
. l age tcresult hdresult diabetes gender region mu if mu>= 0.47 & mu!=.

      +-------------------------------------------------------------------+
      | age   tcresult   hdresult   diabetes   gender   region         mu |
      |-------------------------------------------------------------------|
 2545. |  71        331         16          1     Male        3   .4704973 |
      +-------------------------------------------------------------------+
```

A 71 year old diabetic male patient in the study living in the South, with a serum cholesterol count of 331 and high density lipid count of 16 has the highest likelihood of having a heart attack.