Logistic Regression
Project 2 Model Answer

Joseph M Hilbe

The project asks to determine the best fitted model for being out of work in Germany during 1988. Potential explanatory predictors are listed below, and are available in the rwm1yr data, which is abstracted from the German Health Reform Registry for the year 1988.

Data: rwm1yr

outwork	1=not working; 0=working	binary	female	1=femaie; 0=male	binary
married	Married=1; Single=0	binary	kids	1=have children; 0=no children	binary
edlevel	Level of education	categorical	docvis	MD visits/year	continuous
hospvis	Days in hospital/year	continuous	age	Ages 25-64	continuous
hhninc	Household income (Marks,OECD wgt)	continuous			

A summary profile of the response, *outwork*, and possible predictors is given as:

```
. su

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     outwork |      4483    .3276824    .4694207          0          1
      female |      4483    .4840509    .4998013          0          1
     married |      4483    .7521749    .4317979          0          1
        kids |      4483    .3794334    .4853001          0          1
     edlevel |      4483    1.491189    .9475775          1          4
-------------+--------------------------------------------------------
      docvis |      4483    2.871961    5.144336          0         90
     hospvis |      4483    .1490074    .8763926          0         35
         age |      4483    43.44011    11.28801         25         64
      hhninc |      4483    3.487401    1.641828          0         20
```

edlevel, a 4 level categorical variable is factored into 4 dummy or indicator variables. We will use the first level, ‘not a high school graduate’, as the reference. It has over three-fourths of the observations.

```
. tab edlevel, gen(educ)

   Level of |
  education |      Freq.     Percent        Cum.
------------+-----------------------------------
Not HS grad |      3,401       75.86       75.86
    HS grad |        294        6.56       82.42
  Coll/Univ |        456       10.17       92.59
Grad School |        332        7.41      100.00
------------+-----------------------------------
      Total |      4,483      100.00
```

A univariable logistic regression is used to determine of any binary or continuous predictor is clearly not associated with the response, *outwork*. A univariabel logistic regression is also provided of outwork on the levels of *edlevel*, with level 1 as the reference. The levels, each binary indicator variables, have been given the names *educ2 – educ4*.

```
. logistic outwork female married kids docvis hospvis age hhninc

Logistic regression                               Number of obs   =       4483
                                                  LR chi2(7)      =    1136.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -2267.3785                       Pseudo R2       =     0.2004

------------------------------------------------------------------------------
     outwork | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   6.033494   .4621664    23.46   0.000     5.192383    7.010855
     married |   1.529322    .151078     4.30   0.000     1.260117    1.856039
        kids |   1.274946   .1132263     2.74   0.006     1.071267     1.51735
      docvis |   1.016134   .0070402     2.31   0.021     1.002429    1.030027
     hospvis |   1.088353   .0509789     1.81   0.071     .9928859       1.193
         age |   1.046549   .0040013    11.90   0.000     1.038736    1.054421
      hhninc |   .6413558   .0192654   -14.79   0.000     .6046863     .680249
       _cons |   .0643173   .0136003   -12.98   0.000     .0424948    .0973464
------------------------------------------------------------------------------
```

```
. glm outwork educ2-educ4, nolog fam(bin) nohead eform

------------------------------------------------------------------------------
             |                 OIM
     outwork | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educ2 |   .9435484   .1215855    -0.45   0.652      .732957    1.214646
       educ3 |   .9342676   .0988733    -0.64   0.521      .759257    1.149619
       educ4 |   .2733037   .0461847    -7.68   0.000     .1962473    .3806162
       _cons |   .5299145   .0190971   -17.62   0.000     .4937763    .5686977
------------------------------------------------------------------------------

. di e(ll)
-2796.9669
```

A model with all predictors except level 2 of *edlevel* appears to be a well fitted model.

```
. glm outwork female-kids docvis hospvis age hhninc educ3 educ4, fam(bin) eform

Iteration 0:   log likelihood = -2269.6826
Iteration 1:   log likelihood = -2246.2986
Iteration 2:   log likelihood = -2246.2469
Iteration 3:   log likelihood = -2246.2469

Generalized linear models                          No. of obs      =      4483
Optimization     : ML                              Residual df     =      4473
                                                   Scale parameter =         1
Deviance         =  4492.493702                    (1/df) Deviance =  1.004358
Pearson          =  5121.343676                    (1/df) Pearson  =  1.144946

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

                                                   AIC             =  1.006579
Log likelihood   = -2246.246851                    BIC             =  -33116.7

------------------------------------------------------------------------------
             |                 OIM
```

```
     outwork | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   6.238166   .4856971    23.51   0.000     5.355293    7.266588
     married |   1.575058   .1593084     4.49   0.000      1.29182    1.920396
        kids |   1.323519   .1184544     3.13   0.002     1.110575    1.577293
      docvis |   1.017057   .0070992     2.42   0.015     1.003237    1.031066
     hospvis |   1.089815   .0514316     1.82   0.068     .9935324    1.195428
         age |     1.0501   .0040986    12.52   0.000     1.042097    1.058164
      hhninc |   .6391308   .0197298   -14.50   0.000     .6016076    .6789944
       educ3 |   2.126172   .2689686     5.96   0.000     1.659275    2.724448
       educ4 |   .6815386   .1282189    -2.04   0.042     .4713599    .9854359
       _cons |   .0498737   .0109012   -13.72   0.000     .0324951    .0765463
------------------------------------------------------------------------------
```

The differences in values of the standard errors when applying a robust variance estimator gives evidence that there is excess correlation in the data. We should use robust SEs then for our model. The *p*-values appear better then with model-based SEs.

```
. glm outwork female-kids docvis hospvis age hhninc educ3 educ4, fam(bin) eform robust

Iteration 0:   log pseudolikelihood = -2269.6826
Iteration 1:   log pseudolikelihood = -2246.2986
Iteration 2:   log pseudolikelihood = -2246.2469
Iteration 3:   log pseudolikelihood = -2246.2469

Generalized linear models                          No. of obs      =      4483
Optimization     : ML                              Residual df     =      4473
                                                   Scale parameter =         1
Deviance         =  4492.493702                    (1/df) Deviance =  1.004358
Pearson          =  5121.343676                    (1/df) Pearson  =  1.144946

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

                                                   AIC             =  1.006579
Log pseudolikelihood = -2246.246851                BIC             =  -33116.7

------------------------------------------------------------------------------
```

```
             |               Robust
     outwork | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   6.238166   .4838883    23.60   0.000     5.358337     7.26246
     married |   1.575058   .1726084     4.15   0.000     1.270616    1.952444
        kids |   1.323519   .1169449     3.17   0.002     1.113061    1.573771
      docvis |   1.017057    .007751     2.22   0.026     1.001978    1.032362
     hospvis |   1.089815   .0449568     2.08   0.037     1.005169    1.181589
         age |     1.0501   .0043535    11.79   0.000     1.041602    1.058667
      hhninc |   .6391308   .0248904   -11.49   0.000     .5921618    .6898253
       educ3 |   2.126172    .294894     5.44   0.000      1.62009    2.790343
       educ4 |   .6815386    .127023    -2.06   0.040     .4729837    .9820526
       _cons |   .0498737   .0126814   -11.79   0.000     .0302995    .0820931
------------------------------------------------------------------------------
```

The AIC and BIC values are lower with this model than any other.

```
. abic
AIC Statistic   =   1.006579           AIC*n      = 4512.4937
BIC Statistic   =    1.01239           BIC(Stata) = 4576.5742
```

The fit of the model is accessed by using various post-estimation tests. We use a HosmerLemeshow Goodness-of-fit test and a Tukey-Pregibon linktest.

```
. qui logistic outwork female married kids docvis hospvis age hhninc educ3 educ4
. estat gof, table group(10)

Logistic model for outwork, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0640 |    40 |  17.4 |   409 | 431.6 |   449 |
  |     2 | 0.1032 |    34 |  37.9 |   414 | 410.1 |   448 |
  |     3 | 0.1460 |    47 |  55.7 |   401 | 392.3 |   448 |
  |     4 | 0.2015 |    45 |  77.0 |   404 | 372.0 |   449 |
  |     5 | 0.2715 |    99 | 104.8 |   349 | 343.2 |   448 |
  |-------+--------+-------+-------+-------+-------+-------|
```

```
  |     6 | 0.3671 |   137 | 141.6 |   311 | 306.4 |   448 |
  |     7 | 0.4619 |   226 | 186.6 |   223 | 262.4 |   449 |
  |     8 | 0.5554 |   234 | 228.2 |   214 | 219.8 |   448 |
  |     9 | 0.6816 |   252 | 276.2 |   196 | 171.8 |   448 |
  |    10 | 0.9837 |   355 | 343.5 |    93 | 104.5 |   448 |
  +-------------------------------------------------------+

       number of observations =      4483
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        70.92
                  Prob > chi2 =         0.0000

. linktest

Iteration 0:   log likelihood = -2835.6103
Iteration 1:   log likelihood = -2253.8784
Iteration 2:   log likelihood = -2239.1808
Iteration 3:   log likelihood = -2238.8469
Iteration 4:   log likelihood = -2238.8453
Iteration 5:   log likelihood = -2238.8453

Logistic regression                               Number of obs   =       4483
                                                  LR chi2(2)      =    1193.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -2238.8453                       Pseudo R2       =     0.2105

------------------------------------------------------------------------------
     outwork |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   1.130544   .0487535    23.19   0.000     1.034989    1.226099
      _hatsq |   .0882216   .0207308     4.26   0.000       .04759    .1288533
       _cons |  -.0554539   .0433141    -1.28   0.200     -.140348    .0294401
------------------------------------------------------------------------------
```

In both circumstances, the test fail to confirm that the model properly fits the data. After attempting a variety of models, the following model apparently fits. The model below produces the near same results whether a robust variance estimator is used for standard errors.

```
. logistic outwork female hospvis educ4

Logistic regression                               Number of obs   =       4483
                                                  LR chi2(3)      =     696.20
                                                  Prob > chi2     =     0.0000
Log likelihood = -2487.5081                       Pseudo R2       =     0.1228

------------------------------------------------------------------------------
     outwork | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   5.348065   .3837571    23.37   0.000     4.646412    6.155676
     hospvis |   1.091037   .0489109     1.94   0.052     .9992638    1.191238
       educ4 |   .3431769   .0603354    -6.08   0.000     .2431446    .4843636
       _cons |   .1997761   .0116325   -27.66   0.000     .1782297    .2239273
------------------------------------------------------------------------------
```

We first apply a Hosmer-Lemeshow goodness-of-fit test, collapsing on three levels. A TukeyPregibon *linktest* follows, indicating that the model fits as a logistic regression.

```
. estat gof, table group(3)

Logistic model for outwork, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 |  Exp_1 | Obs_0 |  Exp_0 | Total |
  |-------+--------+-------+--------+-------+--------+-------|
  |     1 | 0.1665 |   315 |  333.2 |  1828 | 1809.8 |  2143 |
  |     2 | 0.5165 |  1017 | 1008.2 |  1091 | 1099.8 |  2108 |
  |     3 | 0.9575 |   137 |  127.6 |    95 |  104.4 |   232 |
  +--------------------------------------------------------+

       number of observations =      4483
             number of groups =         3
      Hosmer-Lemeshow chi2(1) =         2.87
                  Prob > chi2 =         0.0905
```

```
. linktest

Iteration 0:   log likelihood = -2835.6103
Iteration 1:   log likelihood = -2496.8493
Iteration 2:   log likelihood = -2487.5291
Iteration 3:   log likelihood = -2487.4605
Iteration 4:   log likelihood = -2487.4605

Logistic regression                               Number of obs   =       4483
                                                  LR chi2(2)      =     696.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -2487.4605                       Pseudo R2       =     0.1228

------------------------------------------------------------------------------
     outwork |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   1.042692   .1442169     7.23   0.000     .7600317    1.325352
      _hatsq |   .0258423   .0834595     0.31   0.757    -.1377353      .18942
       _cons |  -.0023463   .0423691    -0.06   0.956    -.0853881    .0806956
------------------------------------------------------------------------------

. abic
AIC Statistic   =   1.111536          AIC*n      = 4983.0161
BIC Statistic   =   1.112225          BIC(Stata) = 5008.6484
```

The failure of the square of the hat statistic to be significant indicates that the model fits as a logistic model; i.e. that the link has been correctly specified. The AIC and BIC values are also low. Note that the left hand side AIC and BIC are consistent, which indicates fit.

A classification test given a cutoff point at the mean of the model fitted values shows only a 69% correct classification rate. This is not particularly good, but the model does apparently fit as a logistic regression model.

```
. predict mu
(option pr assumed; Pr(outwork))
```

```
. su mu

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          mu |      4483    .3276823    .1811158   .0641598   .9575343

. estat class, cutoff(.3276823)

Logistic model for outwork

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      1079           994  |       2073
     -     |       390          2020  |       2410
-----------+--------------------------+-----------
   Total   |      1469          3014  |       4483

Classified + if predicted Pr(D) >= .3276823
True D defined as outwork != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   73.45%
Specificity                     Pr( -|~D)   67.02%
Positive predictive value       Pr( D| +)   52.05%
Negative predictive value       Pr(~D| -)   83.82%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   32.98%
False - rate for true D         Pr( -| D)   26.55%
False + rate for classified +   Pr(~D| +)   47.95%
False - rate for classified -   Pr( D| -)   16.18%
--------------------------------------------------
Correctly classified                        69.13%
--------------------------------------------------
```

To determine the covariate profile that best predicts being out-of-work, we have

```
. sort mu

. su mu


    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          mu |      4483    .3276823    .1811158   .0641598   .9575343


. l female hospvis educ4 if mu>.9575


      +--------------------------+
      | female   hospvis   educ4 |
      |--------------------------|
4483. |      1        35       0 |
      +--------------------------+
```

Therefore, a female patient without graduate level education who has been in the hospital 35 days during 1988 had a 96% chance of being unemployed during that year. The aim of the model is to extrapolate to future years. We can expect that females without post collegiate education who spend more than a month in the hospital during a calendar year will be unemployed.