# Generalized linear models

Christopher F Baum

EC 823: Applied Econometrics

Boston College, Spring 2013

# Introduction to generalized linear models

The generalized linear model (GLM) framework of McCullaugh and Nelder (1989) is common in applied work in biostatistics, but has not been widely applied in econometrics. It offers many advantages, and should be more widely known.

GLM estimators are maximum likelihood estimators that are based on a density in the linear exponential family (LEF). These include the normal (Gaussian) and inverse Gaussian for continuous data, Poisson and negative binomial for count data, Bernoulli for binary data (including logit and probit) and Gamma for duration data.

GLM estimators are essentially generalizations of nonlinear least squares, and as such are optimal for a nonlinear regression model with homoskedastic additive errors. They are also appropriate for other types of data which exhibit intrinsic heteroskedasticity where there is a rationale for modeling the heteroskedasticity.

The GLM estimator $\hat{\theta}$ maximizes the log-likelihood

$$Q(\theta) = \sum_{i=1}^{N} \left[ a\left(m(x_i, \beta)\right) + b(y_i) + c\left(m(x_i, \beta)\right) \right]$$

where $m(x, \beta) = E(y|x)$ is the conditional mean of $y$, $a(\cdot)$ and $c(\cdot)$ correspond to different members of the LEF, and $b(\cdot)$ is a normalizing constant.

For instance, for the Poisson, where the mean equals the variance, $a(\mu) = -\mu$ and $c(\mu) = \log(\mu)$. Given definitions of these two functions, the mean and variance are $E(y) = \mu = -a'(\mu)/c'(\mu)$ and $Var(y) = 1/c'(\mu)$. For the Poisson, $a'(\mu) = 1$, $c'(\mu) = 1/\mu$, so $E(y) = Var(y) = \mu$.

GLM estimators are consistent provided that the conditional mean function is correctly specified: that $E(y_i|x_i) = m(x_i, \beta)$. If the variance function is not correctly specified, a robust estimate of the VCE should be used.

To use the GLM estimator, you must specify two options: the `family()`, which defines the member of the LEF to be employed, and the `link()`, which is the inverse of the conditional mean function. The family option may be chosen as `gaussian, igaussian, binomial, poisson, binomial, gamma`.

The link function essentially expresses the transformation to be applied to the dependent variable. Each family has a canonical link, which is chosen if not specified: for instance, `family(gaussian)` has default `link(identity)`, so that a GLM with those two options would essentially be linear regression via maximum likelihood.

The `binomial` family has a default `link(logit)`, while the `poisson` and `binomial` families share `link(log)`. However, a number of other combinations of `family` and `link` are valid: for instance, `link(power` *n*) is valid for all distributional families.

# Some applications

As an illustration of the GLM methodology, consider a model in which we seek to explain a ratio variable, such as a firm's ratio of R&D expenditures to total assets. In micro data, we find that many firms report a zero value for this ratio. A linear regression model would ignore the zero lower bound, and would not take account of managers' decision not to engage in R&D activity.

Much of the empirical research in this area has made use of a Tobit model, which combines the Probit likelihood that a zero value will be observed with the linear regression likelihood to explain non-zero values, and a Tobit approach certainly improves upon standard linear regression by taking account of the mass point at zero.

However, some researchers (e.g., Papke and Wooldridge, *J. Appl. Econometrics*, 1996) have argued that the Tobit model, a censored regression technique, is not applicable where values beyond the censoring point are infeasible.

The motivation for Tobit is often that of an underlying latent variable, such as consumer utility, which is observed only in a limited range: for instance, those deriving positive expected utility from a purchase are observed spending that amount, while those with negative expected utility do not purchase the item. That latent variable interpretation is difficult to motivate in the R&D expenditure setting.

Papke and Wooldridge suggest that a GLM with a binomial distribution and a logit link function, which they term the 'fractional logit' model, may be appropriate even in the case where the observed variable is continuous. To model the ratio $y$ as a function of covariates $x$, we may write

$$g\{E(y)\} = \mathbf{x}\beta,\ y \sim F$$

where $g(\cdot)$ is the link function and $F$ is the distributional family. In our case, this becomes

$$\mathrm{logit}\{E(y)\} = \mathbf{x}\beta,\ y \sim \mathrm{Bernoulli}$$

which should be estimated with a robust VCE.

We illustrate with proportions data in which both 0 and 1 are observed, first fitting with a Tobit specification:

```
 . use http://stata-press.com/data/hh3/warsaw, clear
. g proportion = menarche/total
. tobit proportion age, ll(0) ul(1) vsquish
Tobit regression                                  Number of obs   =         25
                                                  LR chi2(1)      =      81.83
                                                  Prob > chi2     =     0.0000
Log likelihood =  23.393423                       Pseudo R2       =     2.3352

------------------------------------------------------------------------------
  proportion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2336978   .0108854    21.47   0.000     .2112314    .2561642
       _cons |  -2.554451   .1454744   -17.56   0.000    -2.854696   -2.254207
-------------+----------------------------------------------------------------
      /sigma |   .0780817   .0119052                      .0535105    .1026528
------------------------------------------------------------------------------
  Obs. summary:          3  left-censored observations at proportion<=0
                        21     uncensored observations
                         1 right-censored observation  at proportion>=1
```

As Papke and Wooldridge's critique centers on the interpretation of the dependent variable, we might want to make use of Stata's `linktest`, a specification test that considers whether the 'link' is appropriate. In the link test, we regress the dependent variable on the predicted values and their squares. If the model is specified correctly, the squares of the predicted values will have no power.

```
. linktest, ll(0) ul(1) vsquish
Tobit regression                                  Number of obs   =         25
                                                  LR chi2(2)      =      90.81
                                                  Prob > chi2     =     0.0000
Log likelihood =  27.886535                       Pseudo R2       =     2.5917

------------------------------------------------------------------------------
  proportion |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   1.452772   .1440383    10.09   0.000     1.154806    1.750738
      _hatsq |  -.4089519    .123241    -3.32   0.003    -.6638952   -.1540085
       _cons |  -.0729681   .0351176    -2.08   0.049    -.1456144   -.0003218
-------------+----------------------------------------------------------------
      /sigma |   .0640866   .0098612                      .0436872    .0844859
------------------------------------------------------------------------------
  Obs. summary:          3  left-censored observations at proportion<=0
                        21     uncensored observations
                         1 right-censored observation  at proportion>=1
```

As is evident, the link test rejects its null, and casts doubt on the Tobit specification.

# Let us reestimate the model with a fractional logit GLM:

```
 . glm proportion age, family(binomial) link(logit) robust nolog
note: proportion has noninteger values
Generalized linear models                          No. of obs      =         25
Optimization     : ML                              Residual df     =         23
                                                   Scale parameter =          1
Deviance         =       .221432                   (1/df) Deviance =   .0096275
Pearson          =   .1874651097                   (1/df) Pearson  =   .0081507
Variance function: V(u) = u*(1-u/1)                [Binomial]
Link function    : g(u) = ln(u/(1-u))              [Logit]
                                                   AIC             =   .5990425
Log pseudolikelihood = -5.488031244                BIC             =  -73.81271

------------------------------------------------------------------------------
             |               Robust
  proportion |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.608169   .0541201    29.71   0.000     1.502095    1.714242
       _cons |  -20.91168   .7047346   -29.67   0.000    -22.29294   -19.53043
------------------------------------------------------------------------------

. qui margins, at(age=(10(1)18))
. marginsplot, addplot(scatter proportion age, msize(small) ylab(,angle(0))) //
> /
> ti("Proportion reaching menarche") legend(off)
  Variables that uniquely identify margins: age
```

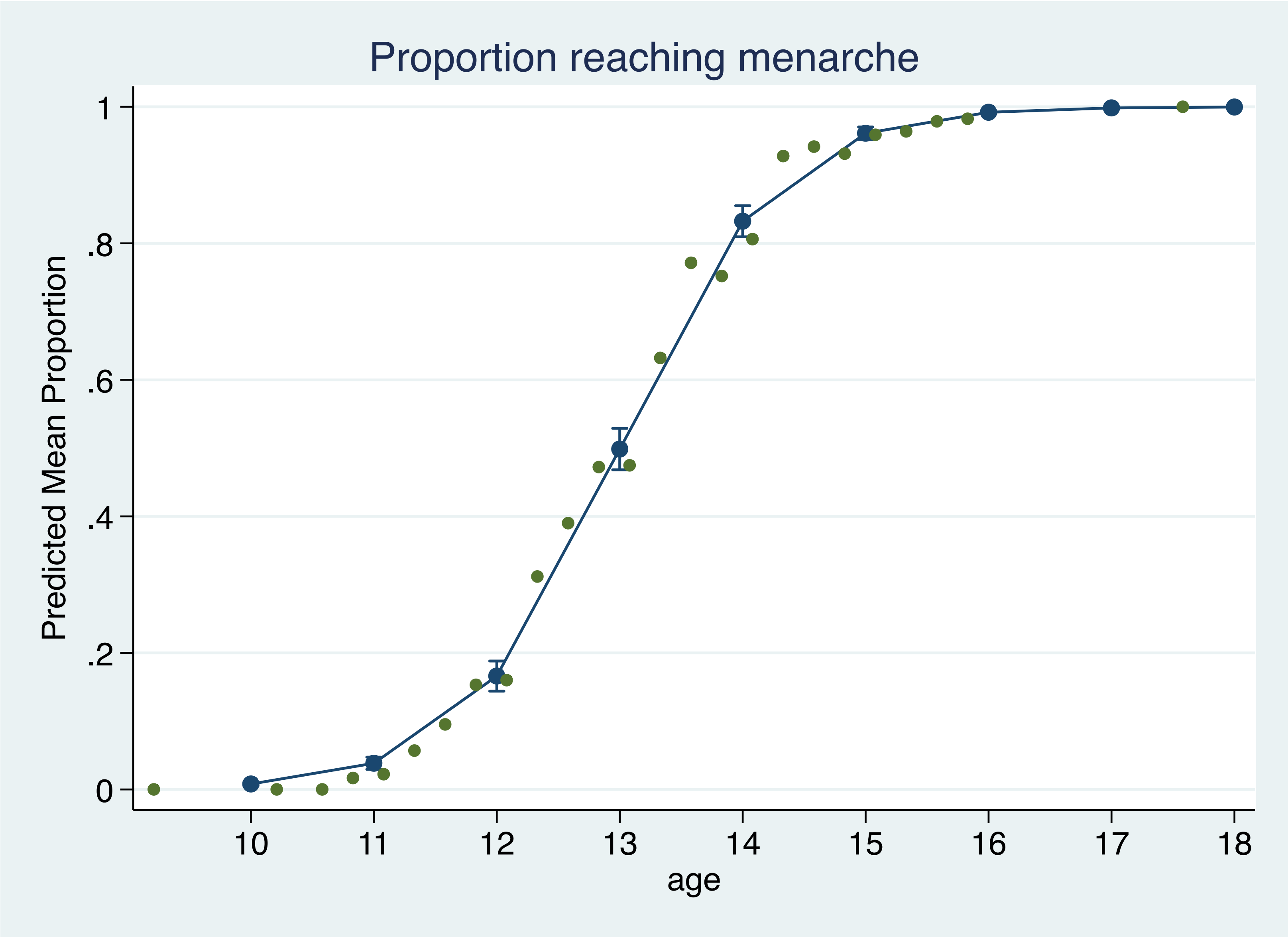
## The link function now is satisfied with the specification:

```
. linktest, robust vsquish
Iteration 0:   log pseudolikelihood =  17.299744
Generalized linear models                          No. of obs      =        25
Optimization     : ML                              Residual df     =        22
                                                   Scale parameter =   .016672
Deviance         =  .3667845044                    (1/df) Deviance =   .016672
Pearson          =  .3667845044                    (1/df) Pearson  =   .016672
Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
                                                   AIC             =  -1.14398
Log pseudolikelihood =  17.29974429                BIC             = -70.44848

                              Robust
  proportion       Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

        _hat    .1173394   .0114055    10.29   0.000     .0949851    .1396938
      _hatsq   -.0030241   .0036441    -0.83   0.407    -.0101665    .0041182
       _cons     .524775   .0337826    15.53   0.000     .4585623    .5909878
```

We may also plot the predictions of the GLM model against the actual proportions data:



# Log-gamma model

Consider a situation where a GLM approach might be useful in simplifying the interpretation of an estimated model. Say that an outcome variable is strictly positive, and we want to model it in a nonlinear form. A common approach would be to transform the outcome variable with logarithms.

This raises the issue that the predictions of the model in levels are biased, even when adjustments are made for the ‘retransformation bias’ (see `sec describe levpredict`).

Alternatively, we can address this problem by using a log-gamma GLM, with the family chosen as gamma and the link function specified as the log. The predictions, residuals and other regression diagnostics of the model are then kept in the natural units of measurement, which may make estimation of the model in this context more attractive than estimating the log-linear regression model.

```
. sysuse cancer
(Patient Survival in Drug Trial)
. glm studytime age i.drug, family(gamma) link(log) nolog vsquish
Generalized linear models                          No. of obs      =         48
Optimization     : ML                              Residual df     =         44
                                                   Scale parameter =   .3180529
Deviance         =  16.17463553                    (1/df) Deviance =   .3676054
Pearson          =  13.99432897                    (1/df) Pearson  =   .3180529
Variance function: V(u) = u^2                      [Gamma]
Link function    : g(u) = ln(u)                    [Log]
                                                   AIC             =   7.403608
Log likelihood   = -173.6866032                    BIC             =  -154.1582
------------------------------------------------------------------------------
             |                 OIM
   studytime |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0447789    .015112    -2.96   0.003    -.0743979     -.01516
        drug |
          2  |   .5743689   .1986342     2.89   0.004      .185053    .9636847
          3  |     1.0521   .1965822     5.35   0.000     .6668056    1.437394
       _cons |   4.646108   .8440093     5.50   0.000      2.99188    6.300336
------------------------------------------------------------------------------
```

```
. predict stimehat
(option mu assumed; predicted mean studytime)

. su studytime stimehat

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   studytime |        48        15.5    10.25629          1         39
    stimehat |        48    15.73706    8.412216   5.185771   34.77219

. corr studytime stimehat
(obs=48)

             | studyt~e stimehat
-------------+------------------
   studytime |   1.0000
    stimehat |   0.6820   1.0000

. di _n "R^2: `=r(rho)^2´"

R^2: .4650907146848232
```

# Poisson on panel data

GLM estimators can be applied to panel or repeated-measures data. In the following example from McCullagh and Nelder, we have data on ships' accidents, with records of the periods the ships were in service, the periods in which they were constructed, and a measure of exposure: how many months they were in service.

As these are discrete (count) data, we model them with a Poisson distribution and a log link. First we consider a pooled estimator with a cluster-robust covariance matrix.

```
. webuse ships, clear
. // cluster by repeated observations on ship type
. glm accident op_75_79 co_65_69 co_70_74 co_75_79, family(poisson) ///
> link(log) vce(cluster ship) exposure(service) nolog vsquish
Generalized linear models                        No. of obs      =        34
Optimization     : ML                            Residual df     =        30
                                                 Scale parameter =         1
Deviance         =  62.36534078                  (1/df) Deviance =  2.078845
Pearson          =  82.73714004                  (1/df) Pearson  =  2.757905
Variance function: V(u) = u                      [Poisson]
Link function    : g(u) = ln(u)                  [Log]
                                                 AIC             =  4.947995
Log pseudolikelihood = -80.11591605              BIC             = -43.42547
                                   (Std. Err. adjusted for 5 clusters in ship)
------------------------------------------------------------------------------
             |               Robust
    accident |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    op_75_79 |   .3874638   .0873609     4.44   0.000     .2162395    .5586881
    co_65_69 |   .7542017    .134085     5.62   0.000        .4914    1.017003
    co_70_74 |    1.05087    .217247     4.84   0.000     .6250737    1.476666
    co_75_79 |   .7040507   .2109515     3.34   0.001     .2905933    1.117508
       _cons |   -6.94765   .0288689  -240.66   0.000    -7.004232   -6.891068
 ln(service) |          1  (exposure)
------------------------------------------------------------------------------
```

```
. margins, by(ship) vsquish
Predictive margins                                Number of obs   =         34
Model VCE    : Robust
Expression   : Predicted mean accident, predict()
over         : ship
```

| | Margin | Delta-method Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ship | | | | | | |
| 1 | 4.271097 | .6324781 | 6.75 | 0.000 | 3.031463 | 5.510731 |
| 2 | 40.00104 | 3.886872 | 10.29 | 0.000 | 32.38291 | 47.61916 |
| 3 | 2.338215 | .3196475 | 7.31 | 0.000 | 1.711718 | 2.964713 |
| 4 | 1.896671 | .2694686 | 7.04 | 0.000 | 1.368522 | 2.42482 |
| 5 | 2.741811 | .4428016 | 6.19 | 0.000 | 1.873936 | 3.609686 |

We may also fit an unconditional fixed-effects estimator, appropriate for the case where there are a finite number of panels in the population. A conditional fixed-effects model can be fit with Stata's `xtpoisson` command, as may random-effects alternatives.

```
. // unconditional fixed effects for ship type
. glm accident op_75_79 co_65_69 co_70_74 co_75_79 i.ship, family(poisson) ///
> link(log) exposure(service) nolog vsquish
Generalized linear models                          No. of obs      =        34
Optimization     : ML                              Residual df     =        25
                                                   Scale parameter =         1
Deviance         =  38.69505154                    (1/df) Deviance =  1.547802
Pearson          =  42.27525312                    (1/df) Pearson  =   1.69101
Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
                                                   AIC             =  4.545928
Log likelihood   = -68.28077143                    BIC             = -49.46396

------------------------------------------------------------------------------
             |                 OIM
    accident |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    op_75_79 |    .384467   .1182722     3.25   0.001     .1526578    .6162761
    co_65_69 |   .6971404   .1496414     4.66   0.000     .4038487    .9904322
    co_70_74 |   .8184266   .1697736     4.82   0.000     .4856763    1.151177
    co_75_79 |   .4534266   .2331705     1.94   0.052    -.0035791    .9104324
        ship |
          2  |  -.5433443   .1775899    -3.06   0.002    -.8914141   -.1952745
          3  |  -.6874016   .3290472    -2.09   0.037    -1.332322    -.042481
          4  |  -.0759614   .2905787    -0.26   0.794    -.6454851    .4935623
          5  |   .3255795   .2358794     1.38   0.168    -.1367357    .7878946
       _cons |  -6.405902   .2174441   -29.46   0.000    -6.832084   -5.979719
 ln(service) |          1  (exposure)
------------------------------------------------------------------------------
```

```
. margins, by(ship) vsquish
Predictive margins                                Number of obs   =         34
Model VCE    : OIM
Expression   : Predicted mean accident, predict()
over         : ship

             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ship |
          1  |          6   .9258201     6.48   0.000     4.185426    7.814574
          2  |   36.14286   2.272282    15.91   0.000     31.68927    40.59645
          3  |   1.714286   .4948717     3.46   0.001     .7443551    2.684216
          4  |   2.428571   .5890151     4.12   0.000     1.274123     3.58302
          5  |   5.333333    .942809     5.66   0.000     3.485462    7.181205
------------------------------------------------------------------------------
```

For more information, see *Generalized Linear Models and Extensions, 3d ed.*, JW Hardin and JM Hilbe, Stata Press, 2012.