# Brief Introduction to Generalized Linear Models

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised August 23, 2024

The purpose of this handout is to briefly show that several seemingly unrelated models are actually all special cases of the generalized linear model. (Indeed, I think most of these techniques were initially developed without people realizing they were interconnected.) We will also briefly introduce the use of factor variables and the `margins` command, both of which will be used heavily during the course.

## The Generalized Linear Model:

$$G(E(Y)) = \alpha + \sum_{k=1}^{K} \beta_k X_{ik}$$

Where G(E(Y)) is some function of the expected value of Y and Y ~ F (i.e. Y has some sort of distribution, e.g. normal, binomial, logistic, etc.) G is referred to as the link function, while F is the distributional family. NOTE: I'm using notation similar to that used by the Stata 13 reference manual when describing the `glm` command; but rather than E(Y), E(Y|X) might be more precise.

## Model 1: OLS regression

$$E(Y) = \alpha + \sum_{k=1}^{K} \beta_k X_{ik}$$

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. regress  income educ jobexp i.black

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  3,   496) =  787.14
       Model |  33206.4588     3  11068.8196           Prob > F      =  0.0000
    Residual |  6974.79047   496  14.0620776           R-squared     =  0.8264
-------------+------------------------------           Adj R-squared =  0.8254
       Total |  40181.2493   499  80.5235456           Root MSE      =  3.7499

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.840407   .0467507    39.37   0.000     1.748553    1.932261
      jobexp |   .6514259   .0350604    18.58   0.000     .5825406    .7203111
     1.black |   -2.55136   .4736266    -5.39   0.000    -3.481921   -1.620798
       _cons |   -4.72676   .9236842    -5.12   0.000    -6.541576   -2.911943
------------------------------------------------------------------------------
```

Note that

- The notation `i.black` tells Stata that black is a categorical variable. In this case, it doesn't affect the results (since black is already coded 0/1) but it would matter if the variable had more than 2 categories. In effect, Stata will create the dummy variables for you. Even more critically, post-estimation commands like `margins` work better when they know which variables are continuous and which are categorical.

- Y has, or can have, a *normal/Gaussian* distribution. Alternatively, you can use regression if Y | X has a normal distribution (or equivalently, if the residuals have a normal distribution and other OLS assumptions are met). That is, the distributional "family" for Y is normal/Gaussian.
- We predict E(Y). E(Y) is in the same units as Y. Alternatively, G(E(Y)) = E(Y). In this case G(E(Y)) is the *identity* link function. Hence, using the `glm` command,

```
. glm income educ jobexp i.black, family(gaussian) link(identity)

Iteration 0:   log likelihood = -1368.3316

Generalized linear models                          No. of obs      =        500
Optimization     : ML                              Residual df     =        496
                                                   Scale parameter =   14.06208
Deviance         =  6974.790467                    (1/df) Deviance =   14.06208
Pearson          =  6974.790467                    (1/df) Pearson  =   14.06208

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]

                                                   AIC             =   5.489327
Log likelihood   = -1368.331633                    BIC             =   3892.345

------------------------------------------------------------------------------
             |                 OIM
      income |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.840407   .0467507    39.37   0.000     1.748777    1.932036
      jobexp |   .6514259   .0350604    18.58   0.000     .5827087    .7201431
     1.black |   -2.55136   .4736266    -5.39   0.000    -3.479651   -1.623069
       _cons |   -4.72676   .9236842    -5.12   0.000    -6.537147   -2.916372
------------------------------------------------------------------------------
```

**MODEL 2: LOGISTIC REGRESSION**. The *logistic regression model (LRM)* (also known as the logit model) can then be written as

$$\ln\frac{P(Y_i=1)}{1-P(Y_i=1)} = \ln\frac{E(Y_i)}{1-E(Y_i)} = \ln(Odds_i) = \alpha + \sum_{k=1}^{K}\beta_k X_{ik} = Z_i$$

The above is referred to as the *log odds* and also as the *logit*. Zi is sometimes used as a convenient shorthand for $\alpha + \Sigma\beta_k X_{ik}$.

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-logit, clear
. logit  grade gpa tuce i.psi, nolog

Logistic regression                               Number of obs   =         32
                                                  LR chi2(3)      =      15.40
                                                  Prob > chi2     =     0.0015
Log likelihood = -12.889633                       Pseudo R2       =     0.3740

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   2.826113   1.262941     2.24   0.025     .3507938    5.301432
        tuce |   .0951577   .1415542     0.67   0.501    -.1822835    .3725988
       1.psi |   2.378688   1.064564     2.23   0.025       .29218    4.465195
       _cons |  -13.02135   4.931325    -2.64   0.008    -22.68657    -3.35613
------------------------------------------------------------------------------
```

Note that

- When y is a dichotomy, it does not have a normal distribution; rather it has a *binomial* distribution (family binomial)
- The left hand side is not E(Y), nor is the left-hand side in the same units as Y. The left hand side is expressed in log odds. We predict G(E(Y)), where G is the *logit* link function. Hence, expressing this as a GLM,

```
. glm grade gpa tuce i.psi, family(binomial) link(logit) nolog

Generalized linear models                          No. of obs      =         32
Optimization     : ML                              Residual df     =         28
                                                   Scale parameter =          1
Deviance         =  25.77926693                    (1/df) Deviance =   .9206881
Pearson          =  27.25711646                    (1/df) Pearson  =   .9734684

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

                                                   AIC             =   1.055602
Log likelihood   = -12.88963347                    BIC             =  -71.26134

------------------------------------------------------------------------------
             |                 OIM
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   2.826113   1.262941     2.24   0.025     .3507937    5.301432
        tuce |   .0951577   .1415542     0.67   0.501    -.1822835    .3725988
       1.psi |   2.378688   1.064564     2.23   0.025       .29218    4.465195
       _cons |  -13.02135   4.931324    -2.64   0.008    -22.68657   -3.356129
------------------------------------------------------------------------------
```

See the Appendix for a few additional examples of GLMs. In particular, the Appendix shows that even a simple crosstab is an example of a Generalized Linear Model! Other GLMs will be discussed during the semester.

Stata's `glm` program can estimate many models – OLS regression, logit, loglinear and count. It can't do ordinal regression or multinomial logistic regression, but I think that is mostly just a limitation of the program, as these are considered GLMS too. Part of this gap is filled by my `oglm` program (ordinal generalized linear models). All in all, `glm` can estimate about 25 different combinations of link functions and families (many of which I have no idea why you would want to use them!) In most cases you don't want to use `glm` because there are specialized routines which work more efficiently and which add other bells and whistles. But, this does serve to illustrate how several seemingly unrelated models are all actually special cases of a more general model.

## Appendix: Other GLM Examples

**MODEL 3: CROSS-CLASSIFIED DATA** (loglinear model; in this specific case, model of independence). Consider a simple 2-way cross-classification of data.

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-cat, clear
(Categorical Case II - Tests of Association)

. tab female dem [fw=freq], chi2 lrchi2 expected
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
+--------------------+

           |          dem
    female |     0 Rep      1 Dem |     Total
-----------+----------------------+----------
    0 Male |        65         55 |       120 
           |      57.0       63.0 |     120.0 
-----------+----------------------+----------
  1 Female |        30         50 |        80 
           |      38.0       42.0 |      80.0 
-----------+----------------------+----------
     Total |        95        105 |       200 
           |      95.0      105.0 |     200.0 

          Pearson chi2(1) =   5.3467   Pr = 0.021
 likelihood-ratio chi2(1) =   5.3875   Pr = 0.020
```

As the chi-square statistics indicate, gender and party affiliation are not independent of each other; females are more likely to be Democrats than are men.

This is probably one of the first things you learned in introductory stats. What you may not have learned is that this can also be written as a loglinear model:

$$\ln(Expected\_Cell\_Frequency) = \alpha + \sum_{k=1}^{K} \beta_k X_{ik}$$

In this model,

- The cell frequencies have a *Poisson* distribution, i.e. family Poisson
- The left hand side is not the expected cell frequency; rather it is the log of the expected cell frequency. Hence, expressing this as a GLM

```
. glm freq i.female i.dem, family(poisson) link(log)

Iteration 0:   log likelihood =  -14.13805
Iteration 1:   log likelihood = -14.124228
Iteration 2:   log likelihood = -14.124227

Generalized linear models                          No. of obs      =         4
Optimization     : ML                              Residual df     =         1
                                                   Scale parameter =         1
Deviance         =  5.387522771                    (1/df) Deviance =  5.387523
Pearson          =  5.346700063                    (1/df) Pearson  =    5.3467

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  8.562114
Log likelihood   = -14.12422743                    BIC             =  4.001228

------------------------------------------------------------------------------
             |                 OIM
        freq |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.female |  -.4054651   .1443376    -2.81   0.005    -.6883615   -.1225687
       1.dem |   .1000835   .1415985     0.71   0.480    -.1774444    .3776114
       _cons |   4.043051    .117727    34.34   0.000     3.812311    4.273792
------------------------------------------------------------------------------
```

Note that the chi-square statistics in the original crosstab correspond to the Deviance and Pearson statistics presented in the GLM. Further, as the crosstab shows, under the model of independence the expected number of male Republicans is 57. To confirm, the formula for computing the expected cell frequency is

P(Male) * P(Republican) * N = 95/200 * 120/200 * 200 = 57.

Expressing things in terms of the glm,

$$\begin{aligned}\ln(\text{Expected\_Male\_Republicans}) &= \alpha + \sum_{k=1}^{K} \beta_k X_{ik} = 4.043051 - .4054651 * \mathit{female} + .1000835 * \mathit{dem} \\ &= 4.043051 - .4054651 * 0 + .1000835 * 0 \\ &= 4.043051\end{aligned}$$

Since the log of the expected cell frequency for male Republicans is 4.043051, this means that the expected cell frequency for male Republicans is exp(4.043051), which equals 57.

Using the `margins` command (more on it later) we can easily reproduce all the expected frequencies under the model of independence:

```
. margins female#dem

Adjusted predictions                              Number of obs   =          4
Model VCE    : OIM

Expression   : Predicted mean freq, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  female#dem |
        0 0  |         57    6.71044     8.49   0.000     43.84778    70.15222
        0 1  |         63   7.143529     8.82   0.000     48.99894    77.00106
        1 0  |         38    5.10196     7.45   0.000     28.00034    47.99966
        1 1  |         42   5.479964     7.66   0.000     31.25947    52.74053
------------------------------------------------------------------------------
```

So in other words, you could say that a generalized linear model with link log and family poisson produces a significant likelihood ratio chi-square statistic of 5.3875 with 1 d.f. – and many people would never guess that all you had done was run a simple crosstab!

**MODEL 4: PROBIT MODEL**. The probit model is a popular alternative to logit, generally producing very similar predictions. The probit model can be written as

$$y^* = \alpha + \sum X\beta + \varepsilon, \varepsilon \sim N(0,1)$$

If y* >= 0, y = 1
If y* < 0, y = 0

The logit model can actually be written the same way, except the error term has a logistic distribution rather than Normal(0, 1). The parameter estimates in a logistic regression tend to be 1.6 to 1.8 times higher than they are in a corresponding probit model. The predicted values in a probit model are like Z-scores. Somebody who has a predicted score of 0 has a 50% chance of success. Somebody with a score of 1 has about an 84% chance of success.

We proceed as we did with logistic regression, except we use the `probit` command instead of `logit`, and with `glm` we specify `link(probit)` rather than `link(logit)`.

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-logit, clear
. probit grade gpa tuce i.psi, nolog

Probit regression                               Number of obs     =         32
                                                LR chi2(3)        =      15.55
                                                Prob > chi2       =     0.0014
Log likelihood = -12.818803                     Pseudo R2         =     0.3775

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |    1.62581   .6938825     2.34   0.019     .2658255    2.985795
        tuce |   .0517289   .0838903     0.62   0.537    -.1126929    .2161508
       1.psi |   1.426332   .5950379     2.40   0.017     .2600795    2.592585
       _cons |   -7.45232   2.542472    -2.93   0.003    -12.43547   -2.469166
------------------------------------------------------------------------------

. glm grade gpa tuce i.psi, family(binomial) link(probit) nolog

Generalized linear models                         No. of obs      =         32
Optimization     : ML                             Residual df     =         28
                                                  Scale parameter =          1
Deviance         =  25.63760665                   (1/df) Deviance =   .9156288
Pearson          =  26.25160404                   (1/df) Pearson  =   .9375573

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = invnorm(u)              [Probit]

                                                  AIC             =   1.051175
Log likelihood   = -12.81880332                   BIC             =    -71.403

------------------------------------------------------------------------------
             |                 OIM
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |    1.62581   .6938825     2.34   0.019     .2658255    2.985795
        tuce |   .0517289   .0838903     0.62   0.537    -.1126929    .2161508
       1.psi |   1.426332   .5950379     2.40   0.017     .2600795    2.592585
       _cons |   -7.45232   2.542472    -2.93   0.003    -12.43547   -2.469166
------------------------------------------------------------------------------
```