

## Brief Introduction to Generalized Linear Models

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.)

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  $Y \sim F$  (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 the same notation that the Stata 8 reference manual does 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 http://www.nd.edu/~rwilliam/xsoc73994/statafiles/glm-reg, clear
. regress income educ jobexp black
```

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

  

|        | 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  |
| 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

- $Y$  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.

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

```
Iteration 0: log likelihood = -1368.3316
```

```
Generalized linear models           No. of obs   =           500
Optimization       : ML: Newton-Raphson   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]
Standard errors   : OIM

Log likelihood    = -1368.331633         AIC            = 5.489327
BIC              = 3892.34485
```

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

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

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

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

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/glm-logit, clear
```

```
. logit grade gpa tuce psi, nolog
```

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

|                      | grade    | gpa      | tuce      | psi      | _cons     |
|----------------------|----------|----------|-----------|----------|-----------|
| Coef.                | 2.826113 | 2.826113 | .0951577  | 2.378688 | -13.02135 |
| Std. Err.            | 1.262941 | 1.262941 | .1415542  | 1.064564 | 4.931325  |
| z                    | 2.24     | 2.24     | 0.67      | 2.23     | -2.64     |
| P> z                 | 0.025    | 0.025    | 0.501     | 0.025    | 0.008     |
| [95% Conf. Interval] | .3507938 | .3507938 | -.1822835 | .29218   | -22.68657 |
|                      | 5.301432 | 5.301432 | .3725988  | 4.465195 | -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 psi, family(binomial 1) link(logit) nolog

Generalized linear models                No. of obs    =        32
Optimization      : ML: Newton-Raphson    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]
Standard errors   : OIM

Log likelihood    = -12.88963347          AIC              = 1.055602
BIC              = -71.26133835
```

```
-----+-----
```

| 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  |
| 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 |

```
-----+-----
```

**Model 3: Cross-classified data (loglinear model; in this specific case, model of independence).** Consider a simple 2-way cross-classification of data.

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/glm-cat, clear
(Categorical Case II - Tests of Association)
```

```
. tab female dem [fw=freq], chi2 lrchi2 expected
```

```
+-----+-----+
| Key          |
|-----|
| frequency    |
| expected frequency |
+-----+-----+
```

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

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

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 female 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: Newton-Raphson  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]
Standard errors    : OIM
```

```
Log likelihood    = -14.12422743      AIC            =  8.562114
BIC               =  4.00122841
```

| freq   | Coef.     | Std. Err. | z     | P> z  | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| female | -.4054651 | .1443376  | -2.81 | 0.005 | -.6883615 - .1225687 |
| 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

$$P(\text{Male}) * P(\text{Republican}) * N = 95/200 * 120/200 * 200 = 57.$$

Expressing things in terms of the glm,

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

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.

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!

Stata's `glm` program can estimate many of the models we will talk about – 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.