Stata Lecture Notes
Class 3

As a review, there are customized commands in Stata for performing regression analyses. In addition the `glm` command with appropriate family and link designations can perform the same regression analyses.

The following examples use the `nepalibf.dta` data set where
bf= breastfeeding = 1 if yes, 0 if no;
sex_chld=0 if male, 1 if female;
age_mom is age in years;
parity is number of previous live births.

1. The customized commands in Stata for logistic regression analysis are `logit` and `logistic`:

The `logit` command provides the estimated regression coefficients from the model log(odds) = $\beta_0 + \beta_1 X$

```
. logit bf sex_chld

Iteration 0:   log likelihood = -320.00632
Iteration 1:   log likelihood = -319.98468
Iteration 2:   log likelihood = -319.98468

Logistic regression                               Number of obs     =        472
                                                  LR chi2(1)        =       0.04
                                                  Prob > chi2       =     0.8352
Log likelihood = -319.98468                       Pseudo R2         =     0.0001

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sex_chld |   .0389756   .1873558     0.21   0.835    -.3282351    .4061863
       _cons |  -.3692173   .1281411    -2.88   0.004    -.6203693   -.1180653
------------------------------------------------------------------------------
```

The `logistic` command provides the exponentiated estimated regression coefficients, exp(b_0) and exp(b_1) from the model log(odds) = $\beta_0 + \beta_1 X$

```
. logistic bf sex_chld

Logistic regression                               Number of obs     =        472
                                                  LR chi2(1)        =       0.04
                                                  Prob > chi2       =     0.8352
Log likelihood = -319.98468                       Pseudo R2         =     0.0001

------------------------------------------------------------------------------
          bf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sex_chld |   1.039745   .1948023     0.21   0.835     .7201937    1.501082
       _cons |   .6912752   .0885808    -2.88   0.004     .5377458     .888638
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

2. The `glm` command with family(binomial) and link(logit) provides the estimated regression coefficients (same as the `logit` command).

```
. glm bf sex_chld, family(binomial) link(logit)

Iteration 0:   log likelihood = -320.27607
Iteration 1:   log likelihood =  -319.9847
Iteration 2:   log likelihood = -319.98468

Generalized linear models                         No. of obs      =        472
Optimization     : ML                             Residual df     =        470
                                                  Scale parameter =          1
Deviance         =   639.969364                   (1/df) Deviance =   1.361637
Pearson          =  471.9999981                   (1/df) Pearson  =   1.004255

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

                                                  AIC             =   1.364342
Log likelihood   =  -319.984682                   BIC             =  -2253.811

------------------------------------------------------------------------------
             |                 OIM
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sex_chld |   .0389756   .1873558     0.21   0.835    -.3282351    .4061863
       _cons |  -.3692173   .1281411    -2.88   0.004    -.6203693   -.1180653
------------------------------------------------------------------------------
```

3. However, the `glm` command with family(binomial) and link(identity) provides the estimated regression coefficients provides the estimated regression coefficients, exp(b_0) and exp(b_1) from the model $Pr(Y=1)= \beta_0 + \beta_1 X$. The estimate b_1 is a risk difference.

```
. glm bf sex_chld, family(binomial) link(identity)

Iteration 0:   log likelihood = -319.98468
Iteration 1:   log likelihood = -319.98468

Generalized linear models                         No. of obs      =        472
Optimization     : ML                             Residual df     =        470
                                                  Scale parameter =          1
Deviance         =   639.969364                   (1/df) Deviance =   1.361637
Pearson          =          472                   (1/df) Pearson  =   1.004255

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

                                                  AIC             =   1.364342
Log likelihood   =  -319.984682                   BIC             =  -2253.811
------------------------------------------------------------------------------
             |                 OIM
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sex_chld |   .0094517   .0454416     0.21   0.835    -.0796123    .0985156
       _cons |   .4087302   .0309678    13.20   0.000     .3480343     .469426
------------------------------------------------------------------------------
Coefficients are the risk differences.
```

4. The customized command in Stata for linear regression analysis is `regression`:

The `regress` command provides the estimated regression coefficients from the model $E(Y) = \beta_0 + \beta_1 X$

```
. regress age_mom parity

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(1, 498)       =    983.55
       Model |  13635.4276         1  13635.4276   Prob > F        =    0.0000
    Residual |  6904.02039       498  13.8634947   R-squared       =    0.6639
-------------+----------------------------------   Adj R-squared   =    0.6632
       Total |   20539.448       499  41.1612184   Root MSE        =    3.7234

------------------------------------------------------------------------------
     age_mom |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      parity |   2.031015   .0647612    31.36   0.000     1.903776    2.158253
       _cons |   19.62211   .3326105    58.99   0.000     18.96862     20.2756
------------------------------------------------------------------------------
```

5. The `glm` command with family(normal) and link(identify) provides the estimated regression coefficients (same as the `regress` command).

```
. glm age_mom parity, family(normal) link(identity)

Iteration 0:   log likelihood =  -1365.782

Generalized linear models                         No. of obs      =        500
Optimization     : ML                             Residual df     =        498
                                                  Scale parameter =   13.86349
Deviance         =  6904.020385                   (1/df) Deviance =   13.86349
Pearson          =  6904.020385                   (1/df) Pearson  =   13.86349

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

                                                  AIC             =   5.471128
Log likelihood   = -1365.782038                   BIC             =   3809.146

------------------------------------------------------------------------------
             |                 OIM
     age_mom |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      parity |   2.031015   .0647612    31.36   0.000     1.904085    2.157944
       _cons |   19.62211   .3326105    58.99   0.000      18.9702    20.27401
------------------------------------------------------------------------------
```

6. Please note that family(normal) and link(identity) are the defaults for the `glm` command if no family and no link are specified.

```
. glm age_mom parity

Iteration 0:   log likelihood =  -1365.782

Generalized linear models                         No. of obs      =        500
Optimization     : ML                             Residual df     =        498
                                                  Scale parameter =   13.86349
Deviance         =  6904.020385                   (1/df) Deviance =   13.86349
Pearson          =  6904.020385                   (1/df) Pearson  =   13.86349

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

                                                  AIC             =   5.471128
Log likelihood   = -1365.782038                   BIC             =   3809.146

------------------------------------------------------------------------------
             |                 OIM
     age_mom |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      parity |   2.031015   .0647612    31.36   0.000     1.904085    2.157944
       _cons |   19.62211   .3326105    58.99   0.000      18.9702    20.27401
------------------------------------------------------------------------------
```

7. A lowess smoother of the relationship between mother's age and parity suggests that the relationship between them appears to be approximately linear until parity exceeds 9 or so. The regression analysis could be repeated on the subset of mothers with parity < 10.

```
. lowess age_mom parity
```