UCL Institute of Health Informatics

UCL

# Generalized Linear Models

Session 1 – Introduction

Rachel Pearson

January 2020

## Today's objectives

UCL

This session will introduce you to generalised linear models (GLMs). By the end of the morning you will be able to:

- Define the components of a GLM
- Identify when to use a GLM to model an outcome
- Fit GLMs in Stata for normal, poisson and binomial outcomes
- Interpret output from a GLM in Stata

2

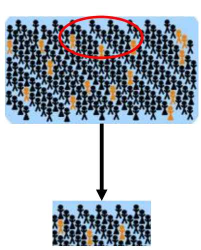
## Statistical inference – linear regression – GLM theory – GLM examples

UCL

MODEL

RECAP

Given a large, random sample of n of a population where a proportion ($\pi$) have some event – the proportion of the sample with the event, K, will follow an Bin(n, $\pi$) distribution.



3

MODEL

PARAMETERS

RECAP

observable

$K \sim bin(n, \pi)$

unknown





MODEL

PARAMETERS

RECAP

ESTIMATORS

Estimator for $\pi$: $\hat{\pi} = \frac{K}{n}$

Estimate for $\pi$: $\frac{k}{n}$, where k is the observed K.





MODEL

PARAMETERS

RECAP

ESTIMATORS

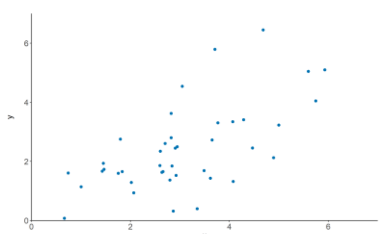
simple hypothesis
($\pi$=0.5 vs $\pi$=0.9)
composite hypothesis
($\pi$ <= 0.5 vs $\pi$ > 0.5)

**Standard Errors:**
- measure of the precision with which the sample statistics approximates the true population
- There are different standard error formulae for different statistical measures (means, percentages, ORs)

**95% Confidence Intervals:**
The interval (sample statistic +/- 1.96 SE) will contain the 'true value' for 95% of random samples



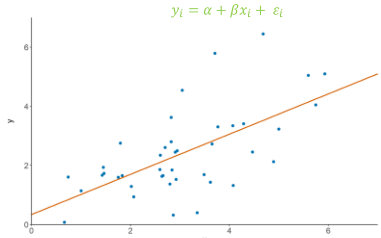


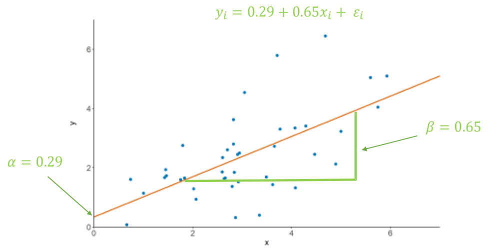






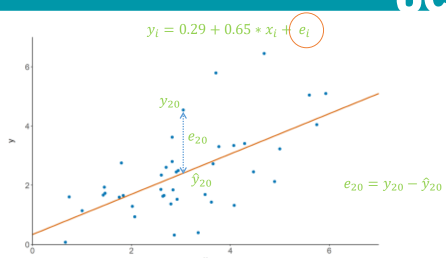

$y_i = \alpha + \beta x_i + \varepsilon_i$






10






11



$$\boldsymbol{y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} + \varepsilon_i}$$

12

$$bmi_i = \alpha + \beta_1 age_i + \beta_2 sex_i + \varepsilon_i$$

```
. reg bmi age 1.sex
```

| Source | SS | df | MS | | | |
|---|---|---|---|---|---|---|
| | | | | Number of obs | = | 2,014 |
| | | | | F(2, 2011) | = | 277.51 |
| Model | 7460.15291 | 2 | 3730.07645 | Prob > F | = | 0.0000 |
| Residual | 27029.9391 | 2,011 | 13.4410438 | R-squared | = | 0.2163 |
| | | | | Adj R-squared | = | 0.2155 |
| Total | 34490.092 | 2,013 | 17.1336771 | Root MSE | = | 3.6662 |

| bmi | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | -.150729 | .0064041 | -23.54 | 0.000 | -.1632882 | -.1381697 |
| 1.sex | -.3156341 | .1637774 | -1.93 | 0.054 | -.6368253 | .0055571 |
| _cons | 28.87251 | .3851415 | 74.97 | 0.000 | 28.11719 | 29.62783 |





$$bmi_i = \alpha + \beta_1 age_i + \beta_2 sex_i + \varepsilon_i$$

```
. reg bmi age 1.sex
```

Holding sex constant, a unit increase in age (i.e. a year) is associated with a 0.15 $kg/m^2$ decrease in BMI (95% CI: -0.16 to -0.14).

| bmi | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | -.150729 | .0064041 | -23.54 | 0.000 | -.1632882 | -.1381697 |
| 1.sex | -.3156341 | .1637774 | -1.93 | 0.054 | -.6368253 | .0055571 |
| _cons | 28.87251 | .3851415 | 74.97 | 0.000 | 28.11719 | 29.62783 |





$$bmi_i = \alpha + \beta_1 age_i + \beta_2 sex_i + \varepsilon_i$$

```
. reg bmi age 1.sex
```

Holding age constant, there is weak evidence that being female is associated with a 0.32 $kg/m^2$ lower BMI (95% CI: -0.64 to 0.01).

| bmi | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | -.150729 | .0064041 | -23.54 | 0.000 | -.1632882 | -.1381697 |
| 1.sex | -.3156341 | .1637774 | -1.93 | 0.054 | -.6368253 | .0055571 |
| _cons | 28.87251 | .3851415 | 74.97 | 0.000 | 28.11719 | 29.62783 |

UCL

$\boldsymbol{bmi_i = \alpha + \beta_1 age_i + \beta_2 sex_i + \varepsilon_i}$

```
. reg bmi age i.sex

      Source |       SS           df       MS      Number of obs   =     2,014
-------------+----------------------------------   F(2, 2011)      =    277.51
       Model |  7460.15291         2  3730.07645   Prob > F        =    0.0000
    Residual |  27029.9391     2,011  13.4410438   R-squared       =    0.2163
-------------+----------------------------------   Adj R-squared   =    0.2155
       Total |   34490.092     2,013  17.1336771   Root MSE        =    3.6662
```

Age and Sex explain an estimated 22% of the variability in BMI

16



UCL

Assumptions of linear regression:

- $Y|x \sim N(\alpha + \beta x, \sigma^2)$
- The relationship between Y and the X's is linear
- The residuals (error terms) are normally distributed i.e. $e_i \sim N(0, \sigma^2)$
- The residuals are independent one of another
- The variance of the residuals is constant, independent of x's

- For many health-related outcomes, these assumptions will not be appropriate
- For example, number of days spent in hospital or whether a patient has diabetes or not

17



UCL

Generalized linear models *generalize* linear modelling and allow us to model a wider variety of outcome types.

The **Exponential Family** of distributions

$$ln\{f(y)\} = \frac{y\theta - b(\theta)}{\varphi} - c(y, \varphi)$$

18

$$ln\{f(y)\} = \frac{y\theta - b(\theta)}{\varphi} - c(y,\varphi)$$

$$f(y) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

$$ln(f(y)) = ln\left(\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}\right)$$





$$ln\{f(y)\} = \frac{y\theta - b(\theta)}{\varphi} - c(y,\varphi)$$

$$ln(f(y)) = ln\left(\sqrt{\frac{1}{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y-\mu)^2$$

$$ln(f(y)) = -\frac{1}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)$$

$$ln(f(y)) = \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \left(\frac{y^2}{2\sigma^2} + \frac{1}{2}ln(2\pi\sigma^2)\right)$$





$$ln\{f(y)\} = \frac{y\theta - b(\theta)}{\varphi} - c(y,\varphi)$$

Distributions that can be written in this form belong to the exponential distribution of families

$$\theta = \mu, \quad b(\theta) = \frac{\mu^2}{2}, \quad \varphi = \sigma^2,$$

$$c(y,\varphi) = \left(\frac{y^2}{2\sigma^2} + \frac{1}{2}ln(2\pi\sigma^2)\right)$$

$ln\{f(y)\} = \frac{y\theta - b(\theta)}{\varphi} - c(y,\varphi)$

The Poisson distribution:

$$f(y) = \Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, y = 0,1,2,3 \ldots$$

$$\ln(f(y)) = y\ln(\mu) - \mu - \ln(y!)$$

$$\theta = \ln(\mu), \quad b(\theta) = \mu, \quad \varphi = 1, \quad c(y,\varphi) = \ln(y!)$$

22



$ln\{f(y)\} = \frac{y\theta - b(\theta)}{\varphi} - c(y,\varphi)$

The Binomial distribution:

$$f(y) = \Pr(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, y = 0,1,\ldots,n$$

$$\ln(f(y)) = y\ln\left(\frac{\pi}{1-\pi}\right) + n\ln(1-\pi) + ln\left\{\binom{n}{y}\right\}$$

$$\theta = \ln\left(\frac{\pi}{1-\pi}\right), \quad b(\theta) = -n\ln(1-\pi), \quad \varphi = 1,$$

$$c(y,\varphi) = -ln\left\{\binom{n}{y}\right\}$$

23



Three GLM components:

- Response distribution
  The $Y_i's$ $i = 1,\ldots,n$ are assumed to be independent and arising from an exponential family. $E(Y_i) = \mu_i$
- Linear predictor
  $$n_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}$$
- Link function
  $$g(\mu_i) = n_i$$

24

Statistical inference – linear regression – GLM theory – GLM examples

| Exponential family | Canonical (natural) link | Response variable(s) | examples |
|---|---|---|---|
| Normal (gaussian) | y | Continuous numbers $(-\infty, \infty)$ | Heights, weights, blood pressure, other lab values |
| Poisson | log(y) = log($\mu$) | Counts (0,1,2,3,4….) Rates $[0,\infty)$ | Counts, rates (e.g. average inpatient length of stay, number of emergency admissions…) |
| Binomial | logit(y) = $\log(\frac{\pi}{1-\pi})$ | Binary [0,1] – logistic regression | i.e. “case” vs “non-case” (e.g. death, disease, receiving an intervention…) |

25

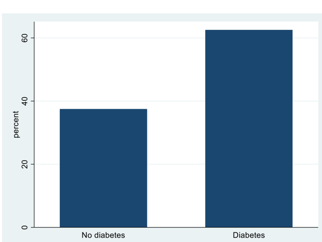
Statistical inference – linear regression – GLM theory – GLM examples

```
. reg bmi age i.sex

      Source |       SS           df       MS
-------------+----------------------------------
       Model |  7460.15291         2  3730.07645
    Residual |  27029.9391     2,011  13.4410438
-------------+----------------------------------
       Total |   34490.092     2,013  17.1336771

         bmi |      Coef.   Std. Err.      t    P
-------------+------------------------------------
         age |   -.150729   .0064041   -23.54   0
       1.sex |  -.3156341   .1637774    -1.93   0
       _cons |   28.87251   .3851415    74.97   0
```

```
. glm bmi age i.sex, family(gaussian) link(identity)

Iteration 0:   log likelihood = -5472.7423

Generalized linear models                         No. of obs      =      2,014
Optimization     : ML                             Residual df     =      2,011
                                                  Scale parameter =   13.44104
Deviance         =  27029.93914                   (1/df) Deviance =   13.44104
Pearson          =  27029.93914                   (1/df) Pearson  =   13.44104

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

                                                  AIC             =   5.437679
Log likelihood   = -5472.742277                   BIC             =    11730.5

------------------------------------------------------------------------------
             |                 OIM
         bmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.150729   .0064041   -23.54   0.000    -.1632807   -.1381772
       1.sex |  -.3156341   .1637774    -1.93   0.054     -.636632    .0053637
       _cons |   28.87251   .3851415    74.97   0.000     28.11765    29.62738
------------------------------------------------------------------------------
```

26

Statistical inference – linear regression – GLM theory – GLM examples




27

Statistical inference – linear regression – GLM theory – GLM examples

UCL

```
. glm diabetes age sex, family(binomial) link(logit)

Iteration 0:   log likelihood = -3302
Iteration 1:   log likelihood = -3297
Iteration 2:   log likelihood = -3297
Iteration 3:   log likelihood = -3297

Generalized linear models
Optimization     : ML

Deviance         =  6594.233323
Pearson          =  4999.872144

Variance function: V(u) = u*(1-u)
Link function    : g(u) = ln(u/(1-u))

Log likelihood   = -3297.116661

                              OIM
    diabetes      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

         age    .0411039   .0093775    4.38   0.000     .0227245    .0594834
         sex   -.0154755   .0592913   -0.26   0.794    -.1316844    .1007334
       _cons   -2.356755   .6561699   -3.59   0.000    -3.642824   -1.070686
```

Odds Ratio = exp(0.0411039) = 1.042

95% C.I. 1.023 to 1.061

Holding sex constant, there is evidence that, for a 1 year increase in age, the odds of developing diabetes increases by 4% (95% C.I. of 2 to 6%).

Log Odds Ratio

28

Statistical inference – linear regression – GLM theory – GLM examples

UCL

```
. glm diabetes age sex, family(binomial) link(logit)

Iteration 0:   log likelihood = -3302
Iteration 1:   log likelihood = -3297
Iteration 2:   log likelihood = -3297
Iteration 3:   log likelihood = -3297

Generalized linear models
Optimization     : ML

Deviance         =  6594.233323
Pearson          =  4999.872144

Variance function: V(u) = u*(1-u)
Link function    : g(u) = ln(u/(1-u))

Log likelihood   = -3297.116661

                              OIM
    diabetes      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

         age    .0411039   .0093775    4.38   0.000     .0227245    .0594834
         sex   -.0154755   .0592913   -0.26   0.794    -.1316844    .1007334
       _cons   -2.356755   .6561699   -3.59   0.000    -3.642824   -1.070686
```

Odds Ratio = exp(-0.0154755) = 0.985

95% C.I. 0.877 to 1.106

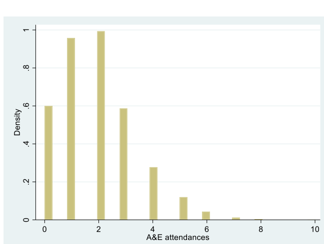
Holding age constant, being female is associated with a 1% decrease in the odds of developing diabetes (95% C.I. of -12% to +11%).

Log Odds Ratio

29

Statistical inference – linear regression – GLM theory – GLM examples

UCL




30

## Statistical inference – linear regression – GLM theory – GLM examples

```
. glm ae_atd age i.sex, family(poisson) link(log)

Iteration 0:   log likelihood = -8522.4235
Iteration 1:   log likelihood
Iteration 2:   log likelihood
Iteration 3:   log likelihood

Generalized linear models
Optimization     : ML

Deviance         =  6127.4037
Pearson          =  5375.7933

Variance function: V(u) = u
Link function    : g(u) = ln(

Log likelihood   = -8499.5734
```

| ae_atd | Coef. | OIM Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | .0101182 | .0032787 | 3.09 | 0.002 | .003692 | .0165444 |
| 1.sex | .0095216 | .0207984 | 0.46 | 0.647 | -.0312425 | .0502857 |
| _cons | -.0742624 | .2300892 | -0.32 | 0.747 | -.525229 | .3767041 |

Log Risk Ratio

Risk Ratio = exp(0.0101182) = 1.01

95% C.I. 1.00 to 1.02

Holding sex constant, there is evidence that, for a 1 year increase in age, the risk of A&E attendances increases by 1% (95% C.I. of 0 to 2%).

31

## Statistical inference – linear regression – GLM theory – GLM examples

```
. glm ae_atd age i.sex, family(poisson) link(log)

Iteration 0:   log likelihood = -8522.4235
Iteration 1:   log likelihood = -8499.5987
Iteration 2:   log likelihood
Iteration 3:   log likelihood

Generalized linear models
Optimization     : ML

Deviance         =  6127.4037
Pearson          =  5375.7933

Variance function: V(u) = u
Link function    : g(u) = ln(

Log likelihood   = -8499.5734
```

| ae_atd | Coef. | OIM Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | .0101182 | .0032787 | 3.09 | 0.002 | .003692 | .0165444 |
| 1.sex | .0095216 | .0207984 | 0.46 | 0.647 | -.0312425 | .0502857 |
| _cons | -.0742624 | .2300892 | -0.32 | 0.747 | -.525229 | .3767041 |

Log Risk Ratio

Risk Ratio = exp(0.0095216) = 1.01

95% C.I. 0.97 to 1.05

Holding age constant, being female is associated with a 1% increase in risk of A&E attendances (95% C.I. of -3% to +5%).

32

## Summary

- We have recapped some of the key definitions underpinning statistical inference
- We have recapped linear regression modelling
- We have learnt about the exponential family of distributions and its properties
- We have learnt about GLMS and its three key components
- We have learnt how to use GLMs to model binary and count outcomes
- We have learnt how to interpret parameter estimates in GLMs

33

## References


- Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General),135*(3), 370-384. doi:10.2307/2344614
- Dobson AJ. *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC: Boca Raton, FL, 1990.


34

## Generalized Linear Models

**Fitting GLMs in Stata:**

glm <response variable> <explanatory variables to form linear predictor>, family (<name of distribution>) link(<link function>).

35