Introduction to Generalized Linear Models: Binary Logistic Regression

Dr Cameron Hurst
cphurst@gmail.com

DAMASAC and CEU, Khon Kaen University

2nd September, 2557

What we will cover....

1. Intro to GLMs
   - General vs. Generalized Linear models
   - Link functions
   - MLE and the exponential family of distributions
   - Classical models $\rightarrow$ GLMs
2. Logistic Regression
   - Kinds of Logistic regression
   - Binary Logistic Regression
   - Revision of the odds ratio
   - Example of Logistic regression analysis

The **general** linear model

Previously, we have considered the **GENERAL** linear model:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i$$

Equivalently(using the Matrix representation):

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Or,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k-1} \\ 1 & x_{2,1} & \cdots & x_{2,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Note: Xs can be continuous variables (covariates) or dummy variables (for factors)

General Linear models and least squares estimation

In the General (normal) linear model we use the Principle of Least Squares to estimate β

- Specifically, the model (i.e. line, plane or hyperplane) is fit such that error sums of squares is minimized:

$$SSE = min\left(\Sigma_{i=1}^{n}\epsilon^2\right)$$

For this reason the estimates of β (obtained by this approach) are called **least squares estimates**.

General → Generalized Linear Models

The General linear model can be further generalized:

$$\mathbf{g}(\mathbf{y}) = \mathbf{X}\beta + \epsilon$$

Or, for the expected (predicted) value of y (dropping ϵ),

$$\mathbf{g}(\mu) = \mathbf{X}\beta$$

where $g() = 1()$ is a function that links $\mathbf{y}$ to $\mathbf{X}$, so

$$\mathbf{1}(\mu) = \mathbf{X}\beta$$

This may seem like an arbitrary (and unnecessary) thing to do, but what about if y (for one reason or another) needed some mathematical transformation to be linearly related to $\mathbf{X}$

General → Generalized Linear Models

For a normally distributed y (given a set level of X), y is often linearly related to the X without needing modification.

What if outcome, y, is not normally distributed (e.g. it might be a binary outcome), the link function $g(y)$ might need to be something other than the identify link, $1(y)$. E.G. we may need to **log** transform y to linearly relate it to X. That is:

$$g(\mu) = \ln(\mu) = \mathbf{X}\beta$$

Or,

$$\mu = g^{-1}(\mathbf{X}\beta) = \exp(\mathbf{X}\beta)$$

In this case a **Log Link** is needed to employ the linear model.

The log link function

The natural log, *ln(y)*, is a very common link function in GLMs; it has some very desirable mathematical and statistical properties.

Canonical link

- For most distributions, a particular link function tends to work best
- This link function is called a **Canonical**, or **Natural** link
- Below are the Distribution-Canonical link pairs
  - Normal (Gaussian) distribution $\leftrightarrow$ Identity link i.e. 1()
  - Binomial distribution (Binary outcomes) $\leftrightarrow$ Logit link
  - Poisson distribution (Counts and rates) $\leftrightarrow$ Log link

The Generalized Linear Model

DEFINITION: A *Generalized Linear Model* is a model that can be represented:

$$g(y) = \mathbf{X}\beta + \epsilon$$

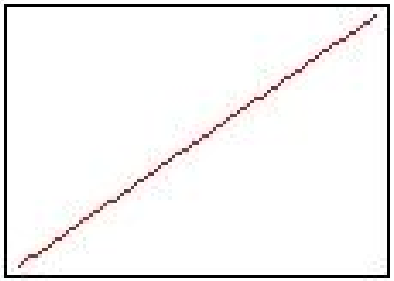
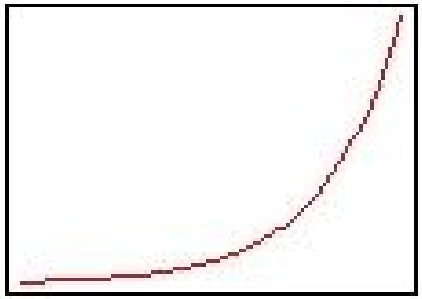
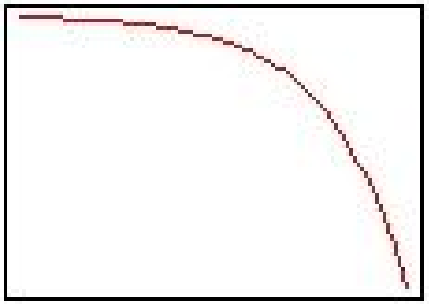
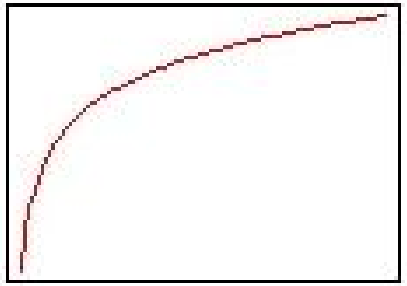
Or equivalently,

$$y = g^{-1}(\mathbf{X}\beta + \epsilon)$$

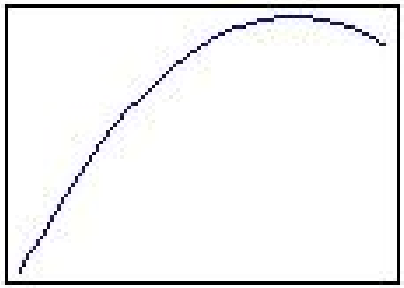
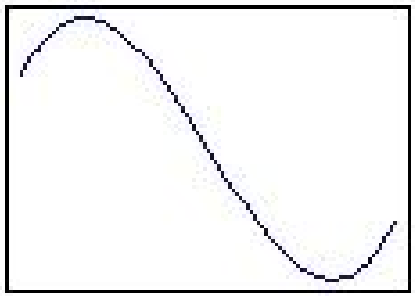
where $g()$ is some link function that is both **monotonic** and **differentiable**

Aside: Monotonic vs Non-monotonic relationships

Monotonic relationships:

Non-monotonic relationships:

Monotonic relationships

Think of monotonic relationships as always walking uphill, or always downhill. The slope doesn't have to be constant, but always going in the same direction (Up or down).

Back to GLMs: Why bother???

That seems like a very nice definition, but why would we bother??

ANS:This flexibility allows us to implement a models for normal outcomes, but also...

- Categorical outcomes:
  - Binary (Binomial distribution)
  - Nominal (Multinomial distribution)
  - Ordinal scale outcomes
  - Count outcomes (e.g. Poisson distribution)
- Other continuous outcomes (not so common in health sciences):
  - Gamma distributed
  - Beta distributed
  - Others

Is there any limit

- Is there a limit to what type of outcomes that GLMs can model??
- **ANS: Yes**. The distribution of the outcome must be from the **Exponential** family of distributions.
- **WHY??????**
- To discuss this we should discuss the differences between the **General** Linear Model and the **Generalized** Linear Model

General Vs Generalized Linear Models

- As we have already discussed **General** Linear models use the **method of Least Squares** to estimate the model parameters
- In contrast, **Generalized** Linear Models use **Maximum Likelihood Estimation** to obtain parameter estimates
- It is the use of Maximum Likelihood estimation, along with the use of link functions that fundamentally differentiates the Generalized Linear Model from the General Linear Model

So what is Maximum Likelihood Estimation

- Simple idea: What is the most likely value of the parameter(s), given the data.
- But can get mathematically messy
- Maximum Likelihood Estimation involves the use of differentiate calculus (with a bit of probability theory) to derive Maximum Likelihood **Estimators**
- It can be shown that for a normally distributed outcome and an identity link, 1(), that the Maximum Likelihood Estimates (**MLEs**) are equivalent to the Least Squares Estimates
- BUT, once we move away from normal outcomes and identity links, the least squares estimates no longer work

Generalized Linear Models

So for a model to be (successfully) implemented as a Generalized Linear Model, Maximum likelihood estimation must be achievable requiring:

Pre-conditions for a Generalized Linear Model (GLM)

- The distribution (of the outcome) MUST come from the **exponential family** of distributions
- The link function must be both **monotonic** and **differentiable**

FORTUNATELY, the derivation of the Maximum Likelihood Estimators (all of the messy bits involving calculus based on random variables) has been done for us. We can simply run our data through software and obtain our **Maximum Likelihood estimates** or **MLEs** of our model βs

Is there any downside to Generalized Linear Models

We now have a family of models that seem much more flexible than General Linear Models, Is there any disadvantage to GLMs??

- Unfortunately, simple measures of explanatory power (such as R-Squared) now go out the window
- Some model-specific *pseudo* R-squared values have been developed, such as Nagelkirke's R-squared used in Logistic Regression (see SPSS)
- However, in general we are left to use less friendly measures of model fit and adequacy such as Deviance and AIC (more later)

Classical models → GLMs

It is quite possible that you have used a GLM without knowing it. This is because many of the standard models used in biostatistics pre-date the idea of the Generalized Linear Model.

- For example, Logistic Regression (which pre-dates GLMs by about 50 years) can be shown to be a GLM
- As we have already seen, Linear Regression (or more generally the General Linear Model) can be implemented as a GLMs
- Poisson regression and Log-Linear analysis are also examples of GLMs

Classical Models, distributions and Link functions

There are a number of classical models that fit into the GLM framework, here are the ones most commonly encountered in Biostatistics

Outcome	Distn.	Link function	Old fashioned name
Binary	Binomial	Logit	Logistic regression
Count (or rate)	Poisson	Log	Log-linear analysis and Poisson regression
Continuous	Normal	Identity	General linear model, ANOVA, Linear regression
Many more			

Cox regression for survival analysis: Exception to the rule

- Cox (proportional hazards) regression is also often identified as a GLM.
- Strictly speaking Cox regression is not a GLM as it is not fully parametric (it is semi-parametric)
- However, as cox regression looks and feels like a GLM, it can be practically (if not theoretically) be considered a GLM.

An example of a GLM

Today we will focus on a particular Generalized Linear Model:

Binary Logistic Regression

This method is widely used in the health and biomedical sciences for modelling binary outcomes

Different types of Logistic Regression

BINARY LOGISTIC REGRESSION (what we cover)

- dichotomous or binary outcomes (2)
- yes/no
- live/dead

POLYTOMOUS LOGISTIC REGRESSION (3+ classes)

NOMINAL (aka Multinomial) LOGISTIC REGRESSION

- nominal with no basis for ordering
  - factory worker, office worker, outdoor worker
  - type of cancer

ORDINAL LOGISTIC REGRESSION

- ordinal with natural basis for ordering
  - low, medium, high
  - strongly disagree, disagree, don't know, agree, strongly agree

Logistic regression in a GLM framework

- Classical model was developed for a binary outcome
- Involves modelling the OR (odds ratio)
- Uses the **logit** link, which is simply the **log of the odds ratio**. That is:

$$\ln(OR) = \mathbf{X}\beta + \epsilon$$

Or,

$$\ln\left(\frac{Odds_{exposed}}{Odds_{unexposed}}\right) = \mathbf{X}\beta + \epsilon$$

Alternatively...

$$\ln(OR) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i$$

By exponentiating both sides (and a little algebraic manipulation)...

$$Prob(Case) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i)}}$$

Purpose of logistic regression:

The above representation not that helpful for standard logistic regression analysis (hypothesis testing about associations), but helpful when using logistic regression for diagnostic testing.

Logistic regression: What will discuss

1. To revise the concept of an odds ratio and how it reflects the relationship between a dichotomous outcome and continuous and/or categorical explanatory variables
2. To interpret logistic regression model output generated from statistical software (R)
3. To consider how we might write-up logistic regression methods and results suitable for thesis or publication

Revision: Old concepts and new approaches

- Odds ratios: What are they, and how do we interpret them?
- Confounding: To adjust or to not adjust (that is the question)
- Effect modification: Interaction terms in Logistic Regression models

REVISION - Odds Ratio

If we represent the association of our binary outcome with a 2-level risk factor (e.g. Exposed / non-exposed)

	Outcome	
Risk Factor	+	-
+	a	b
-	c	d

Odds Ratio:Prevalence studies

Then we can calculate the odds ratio (for prevalence study)

- Odds for outcome(exposed): $\frac{a}{b}$
- Odds for outcome(unexposed): $\frac{c}{d}$
- Odds ratio: $\frac{(\frac{a}{b})}{(\frac{c}{d})} = \frac{ad}{bc}$

Odds Ratio:Case-control studies

Similarly, for case control data (where prevalence is artificial) we can calculate the odds of EXPOSURE for cases and control:

- Odds for cases: $\frac{a}{c}$
- Odds for control: $\frac{b}{d}$
- Odds ratio: $\frac{(\frac{a}{c})}{(\frac{b}{d})}$

This OR represents the *Odds of exposure* in the case group, relative to the odds of exposure in the control group.

Confounding

Are the crude and adjusted ORs similar??

- Old methods: Crude OR vs Mantel-Haenzel adjusted OR
- New Logistic Regression approach:
  - Is the OR from a Bivariate (Simple) Logistic Regression (i.e. crude OR) same as the adjusted OR we get from a 'Multi-variable' Logistic regression
  - We can adjust for multiple potential confounders
  - Confounders can be continuous and/or categorical
  - We can test the significance (and report ORs) of the confounders as potential (independent) risk factors in their own right

Confounding: Example

Study to explore whether high cholesterol (+/-) represents a risk factor for CHD(+/-):

- **Crude OR**: We find that the OR = 5.9
  - Individuals with high cholesterol have 5.9 time the odds of CHD (relative to those without high cholesterol)

When we adjust for demographics(Sex, Age, SES):

- **Adjusted OR**: OR = 2
  - After adjusting for demographics, individuals with high cholesterol have 2 time the odds of CHD (relative to those without high cholesterol)

We find that demographics has substantial confounding effect(e.g.$\triangle$OR $\gg$ 10%)

In terms of Logistic regression models

"Simple" (Bivariate) Logistic Regression model

$$\ln(OR) = \beta_0 + \beta_{Chol} Chol_i$$

gives crude $OR = e^{\beta_{Chol}} = 5.9$
Multi-variable Logistic Regression model....

$$\ln(OR) = \beta_0 + \beta_{Chol} Chol_i + \beta_{Fem} Fem_i + \beta_{Age} Age_i + \beta_{Mid} Mid_i + \beta_{Up} Up_i \quad (1)$$

gives $e^{\beta_{chol}} = 2.0$

Note:

Chol, *Fem*, *Mid* and *Up* are dummy variables (0/1) indicating class membership: As in GENERAL linear model lecture

Effect modification

Effect modification is where the effect of a risk factor varies with the **level** of another predictor. That is, if we ran a model with interaction term(s) (to determine whether effect modification was occurring) and observed a significant interaction(s). For example:

$$\ln(OR) = \beta_0 + \beta_{Chol} Chol_i + \beta_{Fem} Fem_i + \beta_{Chol \times Fem}(Chol \times Fem)$$

and we found we reject H_0: $\beta_{Chol \times Fem} = 0$ then we would conclude that Sex modifies the effect of high cholesterol on the Odds (and risk) of CHD.

Effect modification

After identifying effect modification (A significant interaction), the simplest way of gauging the modification effect of Sex on the risk factor (cholesterol) is to run an individual Logistic Regression for each gender. That is,
For males (only), run:

$$\ln(OR) = \beta_0 + \beta_{Chol} Chol_i$$

and for females (only), run:

$$\ln(OR) = \beta_0 + \beta_{Chol} Chol_i$$

We might find that for males: $OR_{Chol} = 10$ ($p < 0.05$), and for females $OR_{Chol} = 1.2$ ($p \not< 0.05 \Rightarrow$ns) suggesting that cholesterol only represents a significant risk factor (for CHD) for males (i.e males with high cholesterol have 10 times the odds of CHD, relative to males without high cholesterol).

Effect modification vs. confounding

Many people get confused about the difference between **Confounding** and **Effect modification**.

Take home message: Confounding vs. effect modification

Confounding is where the effect of a risk factor (or treatment) changes with the **PRESENCE** of another (confounding) risk factor in the model. **We need to adjust**. **Effect modification** is where the effect of a risk factor (or treatment) changes with the **LEVEL** of another risk factor in the model

- We should also note that effect modification 'trumps' confounding. Since interpretation of the main effects is inappropriate when there is effect modification, the notion of confounding is no longer relevant.

Effect modification vs. confounding

For example:

- If GENDER confounds with our treatment effect we need to include it in our model (i.e. adjust for it)
- If GENDER is an effect modifier (of our treatment), then the efficacy our treatment is different for men (e.g. more effective) than women (e.g. less effective)

Motivating example: Risk factors for fertility

- Study examining effects of various risk factors on fertility in women (*Trichopoulos et al.*, 1976).
- In our analysis, we will focus on the occurrence of (at least one) previous spontaneous abortion as a potential risk for infertility.

Risk factors for fertility: Eyeball data

	Infertility	
Spontaneous abort.	Yes	No
Previous	55	52
No previous	28	113

Just a quick perusal of the cross-tabulation **suggests** that previous spontaneous abortion is associated with an increased risk of infertility

Bivariate analysis: 'Simple' Logistic Regression

Now fitting the model:

$$\ln(OR) = \beta_0 + \beta_{Spon} Spon_i + \epsilon_i$$

where $Spon_i = 1$ where the participant has had a previous abortion (exposed) and 0 otherwise (unexposed)

Logistic Regression in R

(As promised) running Logistic regression (and any GLM) in R is a simple extension to the General Linear Model we saw last week.

General Linear Model

```
my.lin.reg <-lm(my.y~x1+x2, data=mydata.df)
```

To run a logistic regression model:

Logistic regression (and any other Generalized Linear Model)

```
my.log.reg <-glm(my.y~x1+x2, data=mydata.df, family=binomial())
```

Note: For the GENERAL linear model, *my.y* is assumed to be continuous; and for Logistic regression, *my.y* must be binary (0,1)

Output (abridged):

```
#Crude estimate of spontaneous effect
mod.biv <- glm(infert ~ spon.bin, data=infert.df,family=binomial())
summary(mod.biv)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.3952     0.2111  -6.609 3.87e-11 ***
spon.binPrev   1.4513     0.2863   5.069 4.00e-07 ***
---
Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1
AIC: 292.81
                       OR    OR.L95    OR.U95
(Intercept)   0.2477876 0.1638273 0.3747769
spon.binPrev 4.2685440 2.4353608 7.4816296
```

Note: (1) $OR = e^{\beta}$; (2) The ORs (and their CIs) are not automatically output, I have written (and provided you) an R function for this purpose.

Interpretation: write-up

The Simple logistic regression analysis indicates that previous spontaneous abortion is associated with an increased risk in infertility.

The odds of infertility in women who have had at least one previous spontaneous abortion have 4.27 the odds of infertility compared to women who have not had a previous spontaneous abortion (OR=4.268, $p<0.05$, CI: 2.43, 7.48)

We should note that the above represents a **crude** estimate of the association between infertility and previous spontaneous abortion. It is quite likely that there are confounders (that need to be adjusted for)

Multi-variable logistic regression: Obtaining adjusted estimates

Let's see if age represents a confounding effect. Let's fit the model:

$$\ln(OR) = \beta_0 + \beta_{Spon} Spon_i + \beta_{Age} Age_i + \epsilon_i$$

where:
$Spon_i$ is defined as before; and
Age_i represents the age of the i^{th}participant.

Output (abridged):

```
#Adjusted estimate of spontaneous effect
mod.mv<-glm(infert~spon.bin+age,data=infert.df,family=binomial())
summary(mod.mv)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.17146    0.92885  -2.338   0.0194 *
spon.binPrev  1.49274    0.29180   5.116 3.13e-07 ***
age           0.02399    0.02777   0.864   0.3877
---
Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1
AIC: 294.06
```

```
                OR.mv  OR.mv.L95 OR.mv.U95
(Intercept)  0.1140111 0.01846273 0.7040411
spon.binPrev 4.4492701 2.51130981 7.8827410
age          1.0242754 0.97001840 1.0815673
```

Note: $OR = e^{\beta}$

Interpretation: write-up

We see that the addition of *Age* to the model only had a marginal effect. Age neither represents a significant risk factor(in it's own right), nor does it substantially alter the OR for previous spontaneous abortion ($\triangle OR = -4.2\%$). In other words, *Age* does not represent a confounding effect.

The odds of infertility in women who have had at least one previous spontaneous abortion have 4.27 the odds of infertility compared to women who have not had a previous spontaneous abortion (OR_{crude}=4.27, $p<0.05$, CI: 2.43, 7.48). It was suspected age might represent a confounding effect, so a multivariable model including age was run. Their was only a marginal difference in the age adjusted estimate of the spontaneous abortion effect, so only interpretation of the crude estimate is provided. In addition, age could not be

Motivating example: Effect modification

We suspect that age may be an effect modifier. To make interpretation easier we will collapse age into two categories ("Younger than 34", "34 years and older"). This results in the model:

$$\ln(OR) = \beta_0 + \beta_{Spon} Spon_i + \beta_{AgeGr} AgeGr_i + \beta_{Spon \times AgeGr}(Spon \times AgeGr) + \epsilon_i \quad (2)$$

where:
$Spon_i$ is defined as before; and
$AgeGr_i$ represents a dummy variable which equals 1 for women 34+ years (0 otherwise)

Output (abridged):

```
#With spon.bin x age.bin interaction
mod.mv2<- glm(infert~spon.bin+age.bin+spon.bin:age.bin,data=infert.df,family=binomial())
summary(model.mv2)
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.2939     0.2738  -4.726 2.29e-06 ***
spon.binPrev                1.1528     0.3627   3.178  0.00148 **
age.bin>=34                -0.2400     0.4307  -0.557  0.57732
spon.binPrev:age.bin>=34    0.8331     0.5991   1.390  0.16439
---
Signif. codes:  *** 0.001 ** 0.01 * 0.05 . 0.1

AIC: 294.43
                                OR.mv OR.mv.L95 OR.mv.U95
(Intercept)                 0.2741935 0.1603303 0.4689201
spon.binPrev                3.1671827 1.5556545 6.4481194
age.bin>=34                 0.7866205 0.3382040 1.8295818
spon.binPrev:age.bin>=34    2.3003770 0.7108843 7.4438762
```

Note: $OR = e^{\beta}$

Interpretation

- From the analysis *agegr* is not a significant effect modifier ($OR_{AgeGr \times spon}$=2.3, CI: 0.71 7.44)
- If it was, there are two ways of determining the Age group specific ORs
  1. More difficult: Get estimated marginal ORs (using above model). Involves lots of ln()s and exp()s
  2. Easier: Run analysis separately on each age group

That is,the OR_{spon} for each stratum can be obtained,
Younger:

```
glm(infert
~spon.bin,data=infert.df[infert.df$age<34,],family=binomial())
```

Older:

```
glm(infert
~spon.bin,data=infert.df[infert.df$age≥34,],family=binomial())
```

Performing this analysis....

```
######Younger group######
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.2939     0.2738  -4.726 2.29e-06 ***
spon.binPrev   1.1528     0.3627   3.178  0.00148 **

                 OR.mv OR.mv.L95 OR.mv.U95
(Intercept)  0.2741935 0.1603303 0.4689201
spon.binPrev 3.1671827 1.5556545 6.4481194

######Older group######
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.5339     0.3324  -4.614 3.95e-06 ***
spon.binPrev   1.9859     0.4769   4.165 3.12e-05 ***

                 OR.mv OR.mv.L95  OR.mv.U95
(Intercept)  0.2156863 0.1124199  0.4138109
spon.binPrev 7.2857143 2.8612553 18.5518684
```

Interpretation

Although the age effect modification was not statistically significant ($\beta_{sponxagegr} = 0.8331$, $p = 0.164$), there still seems to be a substantial difference between the effect of previous spontaneous abortion for younger ($OR_{<34}$=3.17, p<0.05, CI: 1.56, 6.45) and older ($OR_{\geq 34}$=7.28, p<0.05, CI: 2.86, 18.55) women.

Note: I have used the β rather than the OR to talk about the significance of interaction effect (difficult to interpret ORs for an interaction terms).

Concluding remarks

- GLMs represent a very versatile, and perhaps the most useful, family of models in biostatistics
- Primarily, they can model most of the outcome types (distributions) that arise in health studies
- Mainly differ from General Linear Models (linear regression etc) in two ways:
  1. they formally relate the outcome variable to a linear model through a **link function**
  2. they use **maximum likelihood estimation** (rather than least squares estimation)