Multivariable Logistic Model

Professor Tuan V. Nguyen
NHMRC Senior Research Fellow, Garvan Institute of Medical Research
University of New South Wales
Sydney – Australia

Contents

- Multiple logistic regression
- Grouped data in multiple linear regression
- Deviances
- Models and submodels

Multiple Logistic Regression

- The logistic regression is easily extended to handle more than one explanatory variable. For k explanatory variables $x_1,\ldots,x_k$, and binary response Y, the model is

$$\pi = \Pr(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}$$

Odds and log-odds form

$$Odds\ Form:$$

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)$$

$$Log\text{-}odds\ \ form:$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Interpretation of coefficients

- As before, a unit increase in x_j multiplies the ***odds*** by $\exp(\beta_j)$

- A unit increase in x_j adds β_j to the ***log-odds***

Grouped and ungrouped data in multiple LR

- To group two individuals in multiple LR, the individuals must have the same values for ***all*** the covariates
- Each distinct set of covariates is called a ***covariate pattern***
- If there are m distinct covariate patterns, we record for each pattern the number of individuals having that pattern (n) and the number of "successes" (r).

Log -likelihood

- **For grouped data, the log-likelihood is**

$$l(\beta_{0,},..,\beta_k)=\sum_{i=1}^{m}\left\{r_i(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik})-n_i\log(1+\exp(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}))\right\}$$

For ungrouped data:

- **The log-likelihood is**

$$l(\beta_{0,},..,\beta_k) = \sum_{i=1}^{N} y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}) - \log(1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))$$

Example: Kyphosis risk factors

- Kyphosis is a curvature of the spine that may be a complication of spinal surgery.
- In a study to determine risk factors for this condition, data were gathered on 83 children following surgery.
- Variables are
  - Kyphosis: (binary, absent=no kyphosis, present=kyphosis)
  - Age: continuous, age in months
  - Start: continuous, vertebrae level of surgery
  - Number: continuous, no of vertebrae involved.

Data

	Kyphosis	Age	Number	Start
1	absent	71	3	5
2	absent	158	3	14
3	present	128	4	5
4	absent	2	5	1
5	absent	1	4	15
6	absent	1	2	16
7	absent	61	2	17
8	absent	37	3	16
9	absent	113	2	16
10	present	59	6	12

... 81 cases in all

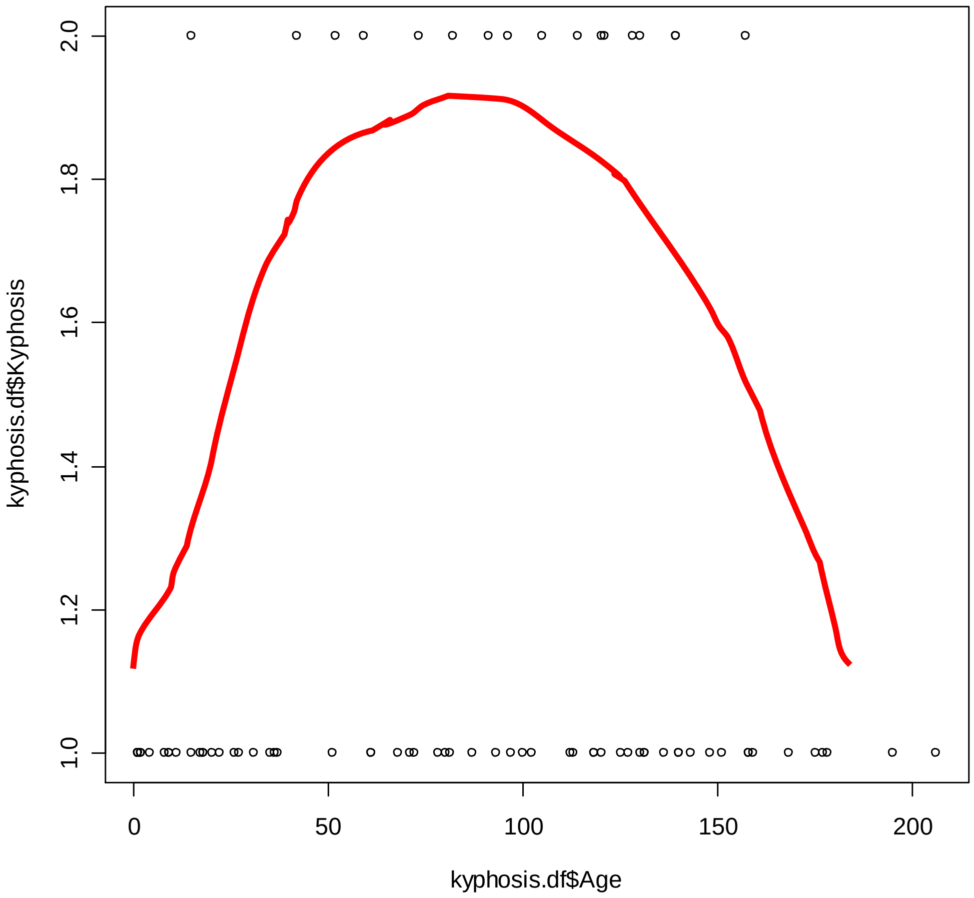
Caution

- **In this data set Kyphosis is not a binary variable with values 0 and 1 but rather a factor with 2 levels “absent” and “present”:**

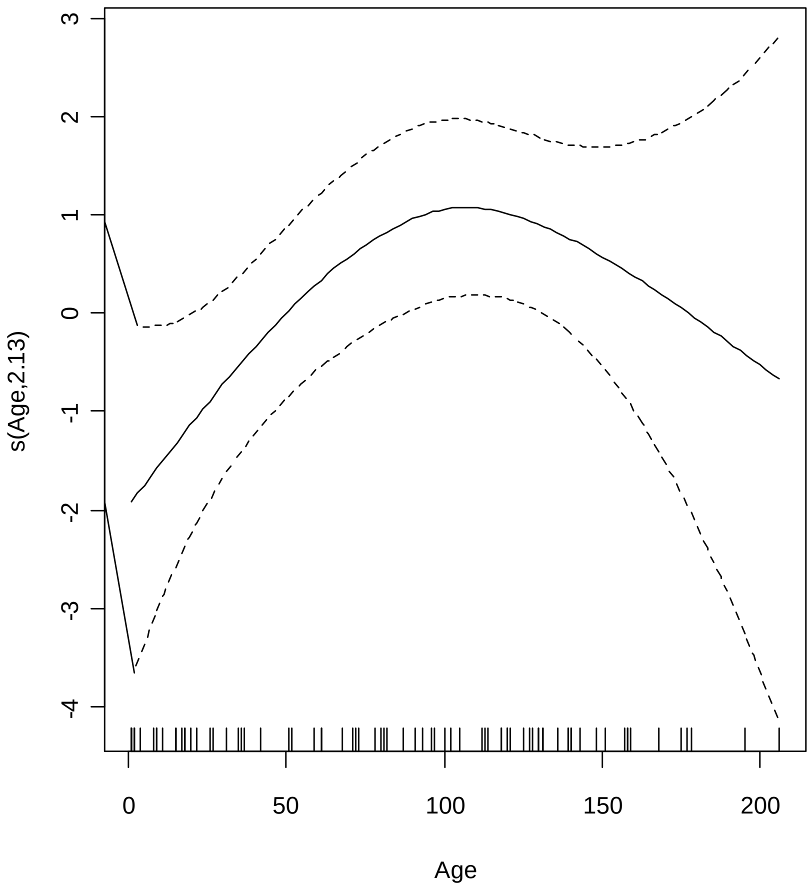
```
levels(kyphosis.df$Kyphosis)
[1] "absent"  "present"
```

NB: if we fit a regression with Kyphosis as the response we are modelling the prob that Kyphosis is “present”: In general, R picks up the first level of the factor to mean “failure (ie in this case “absent” or Y=0) and combines all the other levels into “success” (in this case “present” or Y=1).

plot(kyphosis.df$age,kyphosis.df$Kyphosis)



```
plot(gam(Kyphosis~s(Age) + Number + Start,
family=binomial, data=kyphosis.df))
```



Fitting (i)

```
> kyphosis.glm<-glm(Kyphosis~
   Age + I(Age^2) + Start + Number,
   family=binomial, data=kyphosis.df)
> summary(kyphosis.glm)
```

Fitting (ii)

```
Call:
glm(formula = Kyphosis ~ Age + I(Age^2) + Start + Number,
family = binomial, data = kyphosis.df)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.23572  -0.51241  -0.24509  -0.06109   2.35494
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.3834531  2.0478366  -2.141   0.0323 *
Age          0.0816390  0.0343840   2.374   0.0176 *
I(Age^2)    -0.0003965  0.0001897  -2.090   0.0366 *
Start       -0.2038411  0.0706232  -2.886   0.0039 **
Number       0.4268603  0.2361167   1.808   0.0706 .
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 83.234  on 80  degrees of freedom
Residual deviance: 54.428  on 76  degrees of freedom
AIC: 64.428
```

Points arising

- Start and Age clearly significant
- Need age as quadratic
- What is deviance?
- How do we judge goodness of fit? Is there an analogue of R^2?
- What is a dispersion parameter?
- What is Fisher Scoring?
- To answer these, we first need to explain ***deviance***

Deviance

Recall that our model had 2 parts

- **The binomial assumption (r is Bin (n,π))**
- **The logistic assumption (logit of π is linear)**

If we only assume the first part, we have the most general model possible, since we put no restriction on the probabilities. Our likelihood L is a function of the π's, on

$$L(\pi_1,\ldots,\pi_M) = \prod_{i=1}^{M} \binom{n_i}{r_i} \pi_i^{r_i} (1-\pi_i)^{n_i - r_i}$$

Deviance (cont)

$$l(\pi_1,\ldots,\pi_M) = \sum_{i=1}^{M} \left\{ r_i \log(\pi_i) + (n_i - r_i)\log(1-\pi_i) \right\}$$

Deviance (cont)

L_{max} represents the biggest possible value of the likelihood for the most general model.

Now consider the logistic model, where the form of the probabilities is specified by the logistic function. Let L_{Mod} be the maximum value of the likelihood for this model.

The deviance for the logistic model is defined as

$$\text{Deviance} = 2(\log L_{max} - \log L_{Mod})$$

Deviance (cont)

- Intuitively, the better the logistic model, the closer L_{mod} is to L_{max}, and the smaller the deviance should be
- How small is small?
- If m is small and the n_i's are large, then ***when the logistic model is true***, the deviance has approximately a chi-squared distribution with m-k-1 degrees of freedom
  - m: number of covariate patterns
  - k: number of covariates

Deviance (cont)

- **Thus, if the deviance is less than the upper 95% percentage point of the appropriate chi-square distribution, the logistic model fits well**
- **In this sense, the deviance is the analogue of R^2**
- ***NB Only applies to grouped data, when m is small and the n's are large.***
- **Other names for deviance: *model deviance, residual deviance (R)***

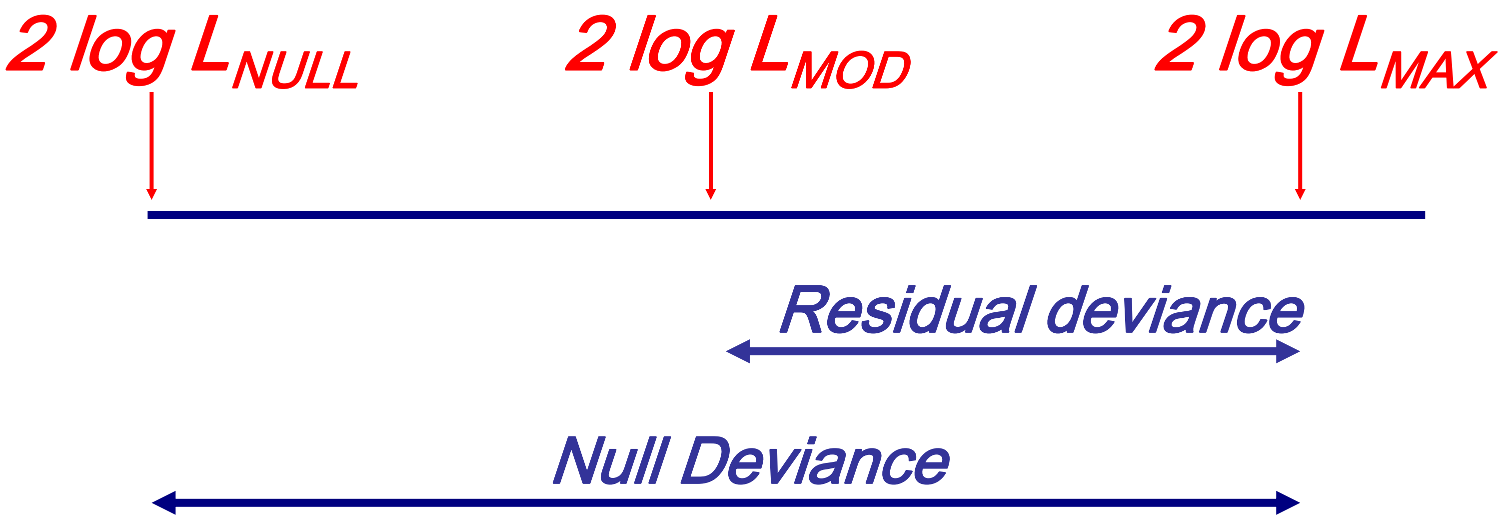
Null deviance

- At the other extreme, the most restrictive model is one where all the probabilities π_i are the same (ie don't depend on the covariates). The deviance for this model is called the ***null deviance***
- Intuitively, if none of the covariates is related to the binary response, the model deviance won't be much smaller then the null deviance

Graphical interpretation

$$L_{NULL} \leq L_{MOD} \leq L_{MAX}$$

$$\therefore 2\log L_{NULL} \leq 2\log L_{MOD} \leq 2\log L_{MAX}$$



Example: budworm data

- Batches of 20 moths subjected to increasing doses of a poison, “success”=death
- Data is grouped: for each of 6 doses (1.0, 2.0, 4.0, 8.0, 16.0, 32.0 mg) and each of male and female, we have 20 moths.
- m=12 covariate patterns

Example: budworm data

	sex	dose	r	n
1	0	1	1	20
2	0	2	4	20
3	0	4	9	20
4	0	8	13	20
5	0	16	18	20
6	0	32	20	20
7	1	1	0	20
8	1	2	2	20
9	1	4	6	20
10	1	8	10	20
11	1	16	12	20

Sex:

0=male

1=female

3 models

- Null model: probabilities π_i are constant, equal to π say. Estimate of this common value is total deaths/total moths = sum(r)/sum(n)   =111/240 = 0.4625

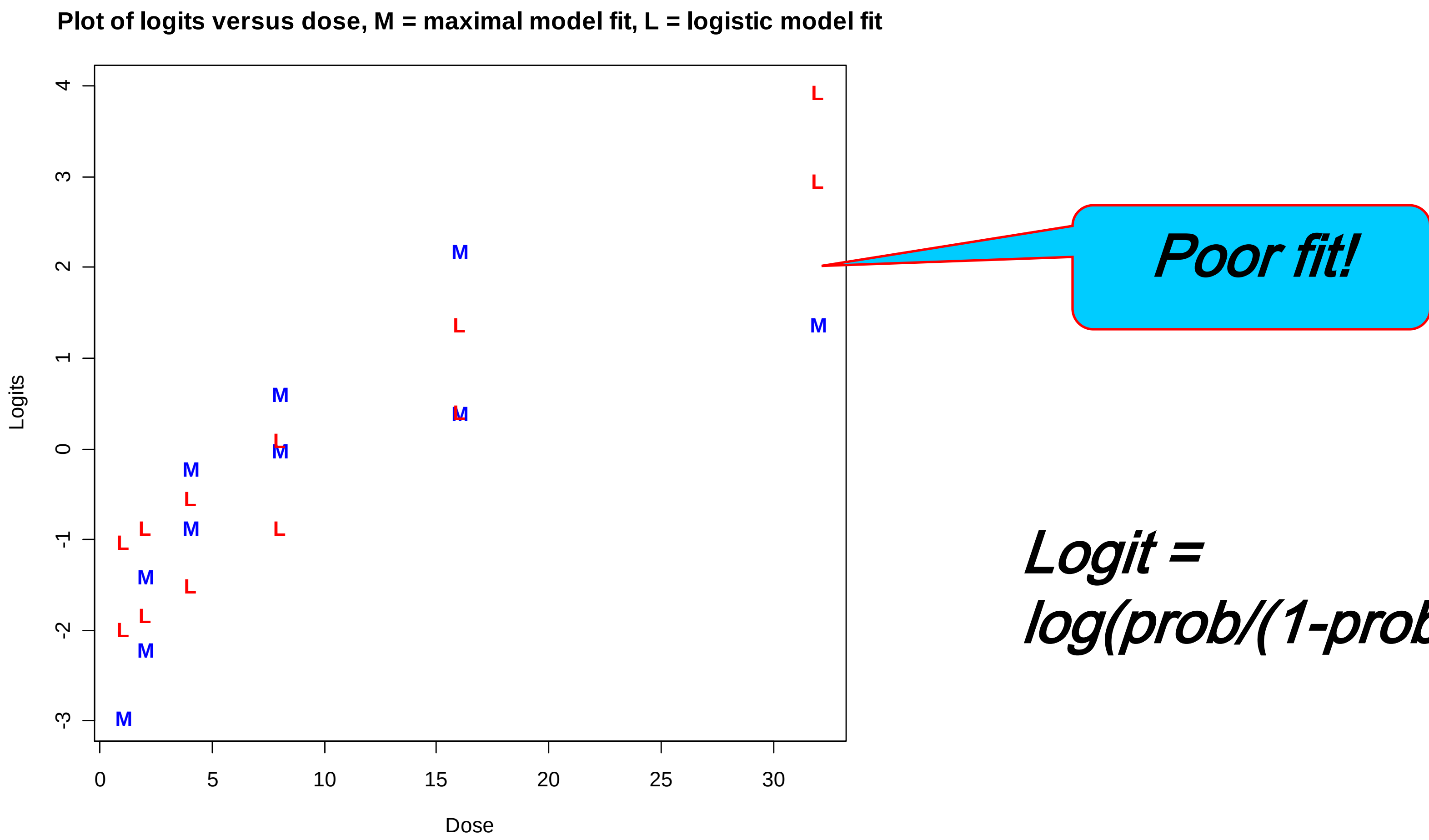
- Logistic model : probabilities estimated using fitted logistic model

- Maximal model: probabilities estimated by r_i/n_i

Probabilities under the 3 models

```
> max.mod.probs<-budworm.df$r/budworm.df$n
> budworm.glm<-glm( cbind(r, n-r) ~ sex + dose, family=binomial,
data = budworm.df)
> logist.mod.probs<-predict(budworm.glm, type="response")
> null.mod.probs<-sum(budworm.df$r)/sum(budworm.df$n)
> cbind(max.mod.probs,logist.mod.probs,null.mod.probs)
   max.mod.probs logist.mod.probs null.mod.probs
1           0.05        0.2677414         0.4625
2           0.20        0.3002398         0.4625
3           0.45        0.3713931         0.4625
4           0.65        0.5283639         0.4625
5           0.90        0.8011063         0.4625
6           1.00        0.9811556         0.4625
7           0.00        0.1218892         0.4625
8           0.10        0.1400705         0.4625
9           0.30        0.1832034         0.4625
10          0.50        0.2983912         0.4625
11          0.60        0.6046013         0.4625
12          0.80        0.9518445         0.4625
```

Plotting logits



Logit = log(prob/(1-prob))

Calculating the likelihoods

Likelihood is

$$L(\pi_1,\ldots,\pi_M) = \prod_{i=1}^{12} \binom{n_i}{r_i} \pi_i^{r_i} (1-\pi_i)^{n_i - r_i}$$

$L_{MAX} = 2.8947 \times 10^{-7}$, $2 \log L_{MAX} = -30.1104$

$L_{MOD} = 2.4459 \times 10^{-13}$, $2 \log L_{MOD} = -58.0783$

$L_{NULL} = 2.2142 \times 10^{-34}$, $2 \log L_{NULL} = -154.9860$

Calculating the deviances

Residual deviance = -30.1104 – (-58.0783) = 27.9679

Null deviance = -30.1104 – (-154.9860) = 124.8756

```
summary(budworm.glm)
Call:
glm(formula = cbind(r, n - r) ~ sex + dose, family =
binomial, data = budworm.df)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1661     0.2615  -4.459 8.24e-06 ***
sex          -0.9686     0.3295  -2.939  0.00329 **
dose          0.1600     0.0234   6.835 8.19e-12 ***
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:  27.968  on  9  degrees of freedom
AIC: 64.078
```

Goodness of fit

- N's reasonably large, m small
- Can interpret residual deviance as a measure of fit

```
> 1-pchisq(27.968,9)
[1] 0.0009656815
```

- Not a good fit!! (as we suspected from the plot)
- In actual fact log(dose) works better

Improvement!

```
> logdose.glm<-glm( cbind(r, n-r) ~ sex + log(dose),
family=binomial, data = budworm.df)
> summary(logdose.glm)
glm(formula = cbind(r, n - r) ~ sex + log(dose), family
= binomial, data = budworm.df)Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3724     0.3854  -6.156 7.46e-10 ***
sex          -1.1007     0.3557  -3.094  0.00197 **
log(dose)     1.5353     0.1890   8.123 4.54e-16 ***
Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:    6.757  on  9  degrees of freedom
AIC: 42.867
> 1-pchisq( 6.757 ,9)
[1] 0.6624024
>
```

Big reduction in deviance, was 27.968

P-value now large