Multivariable Logistic Model: Interaction

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

Contents

Topics

- Binary anova
- Variable selection
- Under/over dispersion

Binary Anova

- Categorical variables are included in logistic regressions in just the same way as in linear regression.
- Done by means of “dummy variables”.
- Interpretation is similar, but in terms of log-odds rather than means.
- A model which fits a separate probability to every possible combination of factor levels is a maximal model, with zero deviance

Example

- The plum tree data: see the coursebook,
- For another example, see Tutorial 8)
- Data concerns survival of plum tree cuttings. Two categorical explanatory variables, each at 2 levels: *planting time* (spring, autumn) and *cutting length* (long, short). For each of these 4 combinations 240 cuttings were planted, and the number surviving recorded.

Data

	length	time	r	n
1	long	autumn	156	240
2	long	spring	84	240
3	short	autumn	107	240
4	short	spring	31	240

Fitting

```
> plum.glm<-glm(cbind(r,n-r)~length*time, family=binomial,
data=plum.df)
> summary(plum.glm)
Call:
glm(formula = cbind(r, n - r) ~ length * time, family =
binomial, data = plum.df)
Deviance Residuals:
[1]  0  0  0  0
Coefficients:                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.6190     0.1353   4.574 4.78e-06 ***
lengthshort                   -0.8366     0.1876  -4.460 8.19e-06 ***
timespring                    -1.2381     0.1914  -6.469 9.87e-11 ***
lengthshort:timespring        -0.4527     0.3009  -1.505    0.132
Null deviance: 1.5102e+02  on 3  degrees of freedom
Residual deviance: 1.7683e-14  on 0  degrees of freedom
AIC: 30.742
```

Zero residuals!

Zero deviance and df!

Points to note

- The model length*time fits a separate probability to each of the 4 covariate patterns
- Thus, it is fitting the maximal model, which has zero deviance by definition
- This causes all the deviance residuals to be zero
- The fitted probabilities are just the ratios r/n

Fitted logits

Logits	Length=long	Length=short
Time=autumn	.6190	.6190 -.8366 = -.2176
Time=spring	.6190 - 1.281= -.6191	.6190 -1.2381 -.8366-.4527 = -1.9084

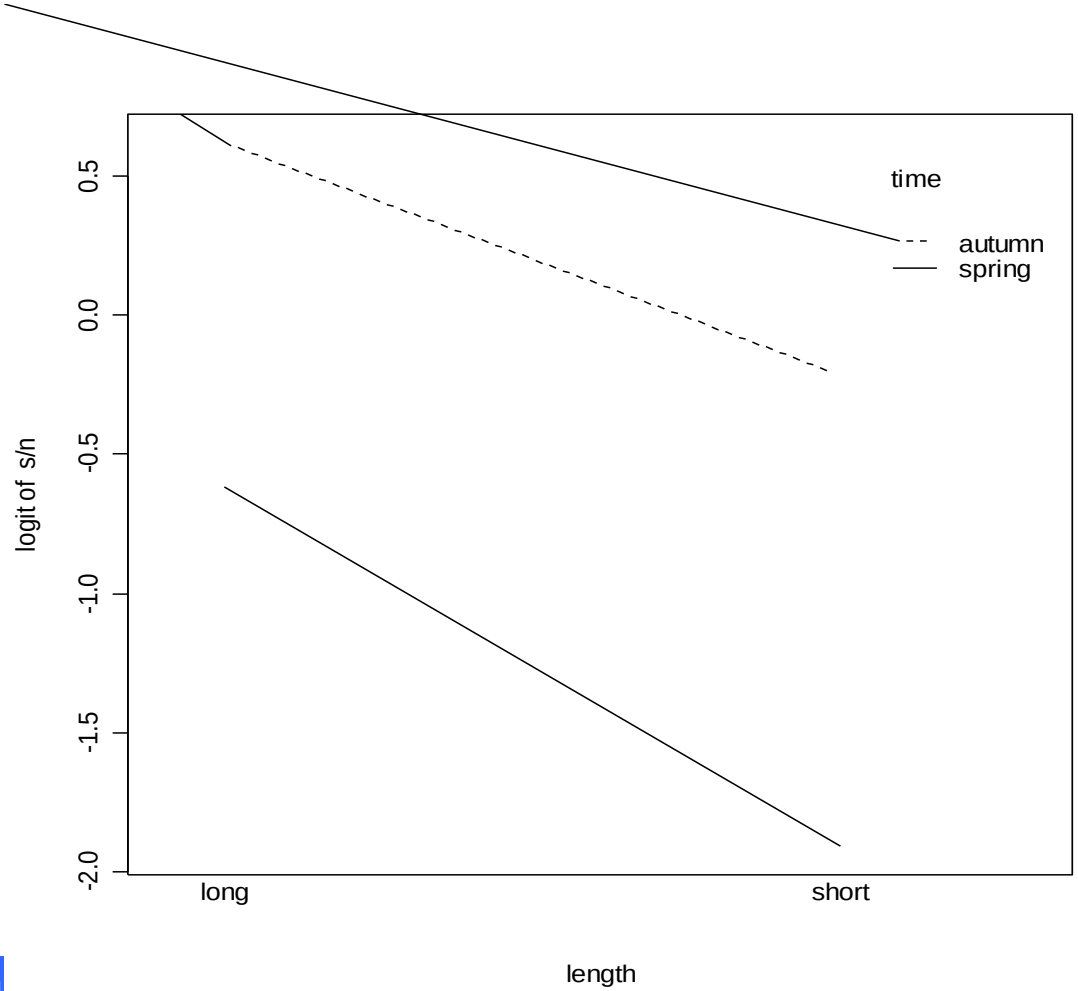
```
Coefficients:                Estimate
(Intercept)                   0.6190
lengthshort                  -0.8366
timespring                   -1.2381
lengthshort:timespring       -0.4527
```

```
> predict(plum.glm)
[1]  0.6190392 -0.6190392 -0.2175203 -1.9083470
```

Interaction plot

```
attach(plum.df)
interaction.plot(length,time,log((r+0.5)/(n-r+0.5)))
```

Lines almost parallel, indicating no interaction on the log-odds scale



Anova

```
> anova(plum.glm,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: cbind(r, n - r)
Terms added sequentially (first to last)

            Df Deviance  Resid.Df  Resid.Dev P(>|Chi|)
NULL                            3    151.019
length       1   45.837          2    105.182 1.285e-11
time         1  102.889          1      2.294 3.545e-24
length:time  1    2.294          0  7.727e-14     0.130
> 1-pchisq(2.294,1)
[1] 0.1298748
```

Interaction not significant

Final model: interpretation and fitted probabilities

```
> plum2.glm<-glm(cbind(r,n-r)~length + time,
family=binomial, data=plum.df)
> summary(plum2.glm)
Call:
glm(formula = cbind(r, n - r) ~ length + time, family =
binomial, data = plum.df)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.7138     0.1217   5.867 4.45e-09 ***
lengthshort   -1.0177     0.1455  -6.995 2.64e-12 ***
timespring    -1.4275     0.1465  -9.747  < 2e-16 ***
Null deviance: 151.0193  on 3  degrees of freedom
Residual deviance:   2.2938  on 1  degrees of freedom
AIC: 31.036
> 1-pchisq(2.2938,1)
[1] 0.1298916
```

Final model: interpretation and fitted probabilities

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.7138     0.1217   5.867 4.45e-09 ***
lengthshort   -1.0177     0.1455  -6.995 2.64e-12 ***
timespring    -1.4275     0.1465  -9.747  < 2e-16 ***
```

Prob of survival less for short cuttings (coeff<0)
Prob of survival less for spring planting (coeff<0)

```
Null deviance: 151.0193  on 3  degrees of freedom
Residual deviance:   2.2938  on 1  degrees of freedom
AIC: 31.036
```

Deviance of 2.2938 on 1 df: pvalue is 0.1299
evidence is that no-interaction model fits well.

Fitted Probabilities

	Length= long	Length= short
Time = autumn	0.6712	0.4246
Time = spring	0.3288	0.1504

```
> predict(plum2.glm,type="response")
[1] 0.6712339 0.3287661 0.4245994 0.1504006
```

Variable selection

- Variable selection proceeds as in ordinary regression
- Use anova and stepwise
- AIC also defined for logistic regression

  AIC = Deviance + 2 ´(number of parameters)

- Pick model with smallest AIC

Example: lizard data

- Site preferences of 2 species of lizard, *grahami* and *opalinus*
- Want to investigate the effect of
  - Perch height
  - Perch diameter
  - Time of day

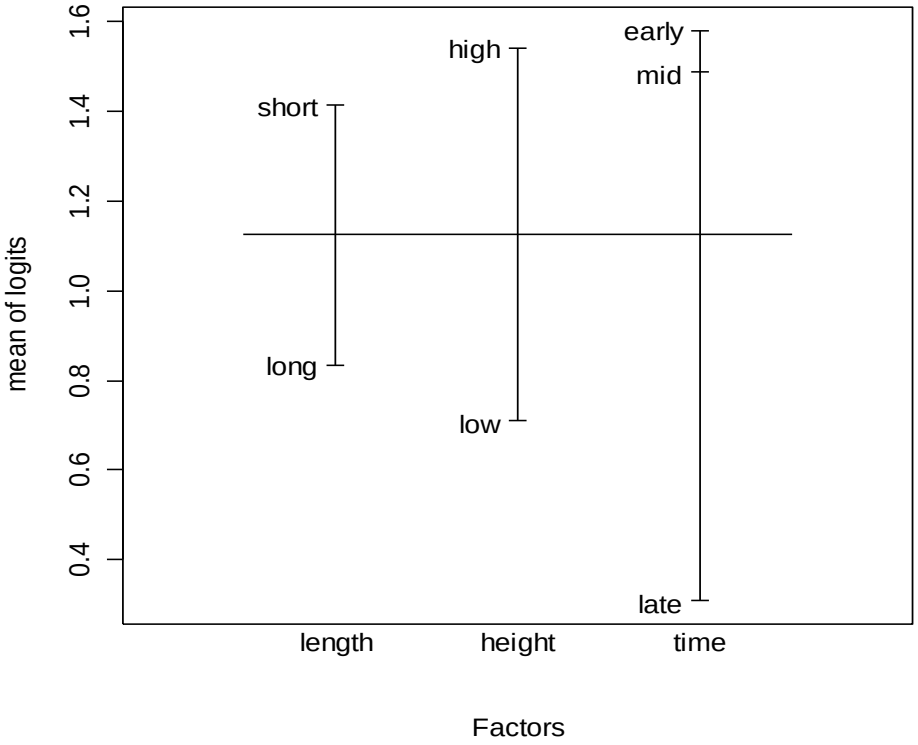
  on the probability that a lizard caught at a site will be grahami

Data

	length	height	time	r	n
1	short	low	early	54	67
2	short	high	early	44	49
3	long	low	early	25	43
4	long	high	early	18	19
5	short	low	mid	77	98
6	short	high	mid	63	67
7	long	low	mid	64	97
8	long	high	mid	21	26
9	short	low	late	22	36
10	short	high	late	25	38
11	long	low	late	13	24
12	long	high	late	5	10

Eyeball analysis

```
> plot.design(lizard.df, y=log((lizard.df$r+0.5)
/(lizard.df$n-lizard.df$r+0.5)), ylab="mean of logits")
```



Proportion of grahami lizards higher when perches are short and high, and in the earlier part of the day

Model selection

- Full model is

  cbind(r,n-r)~time*length*height

  so fit this first.

- Then use anova and stepwise to select a simpler model if appropriate

anova

```
> lizard.glm<-glm(cbind(r,n-r)~time*length*height,
+ family=binomial,data=lizard.df)
> anova(lizard.glm, test="Chisq")

 Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                             11     54.043
time                 2   14.711          9     39.332       0.001
length               1   15.680          8     23.652  7.503e-05
height               1   13.771          7      9.882  2.065e-04
time:length          2    1.170          5      8.711      0.557
time:height          2    5.017          3      3.694      0.081
length:height        1    0.001          2      3.693      0.971
time:length:height   2    3.693          0 -1.354e-14      0.158
```

Both approaches suggest model

cbind(s,n-s) ~ time + length + height

stepwise

```
>null.model<-glm(cbind(r,n-r)~1, family=binomial,
data=lizard.df)
> step(null.model, formula(lizard.glm), direction="both")

Call:  glm(formula = cbind(r, n - r) ~ height + time + length,
family = binomial,       data = lizard.df)

Coefficients:
(Intercept)    heightlow     timelate      timemid  lengthshort
    1.49466     -0.83011     -1.05278      0.04003      0.67630

Degrees of Freedom: 11 Total (i.e. Null);  7 Residual
Null Deviance:      54.04
Residual Deviance: 9.882        AIC: 64.09
```

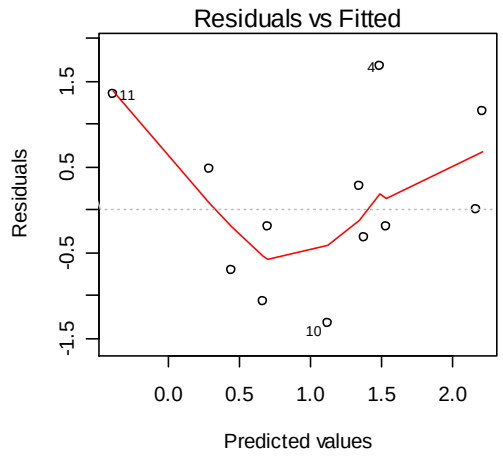
Summary

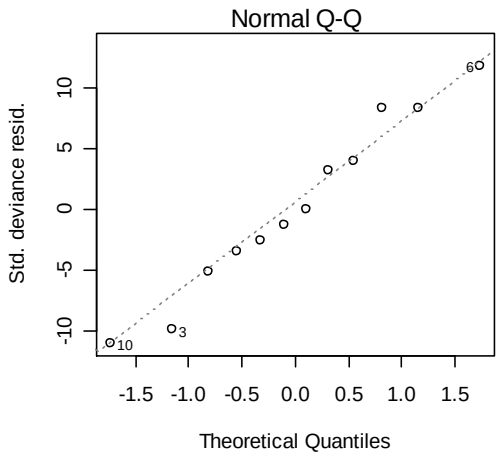
```
> summary(model2)
Call:
glm(formula = cbind(r, n - r) ~ time + length + height,
family = binomial, data = lizard.df)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.49466    0.28809   5.188 2.12e-07 ***
timelate    -1.05278    0.28026  -3.756 0.000172 ***
timemid      0.04003    0.23971   0.167 0.867384
lengthshort  0.67630    0.20588   3.285 0.001020 **
heightlow   -0.83011    0.23204  -3.578 0.000347 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.0430  on 11  degrees of freedom
Residual deviance:  9.8815  on  7  degrees of freedom
```

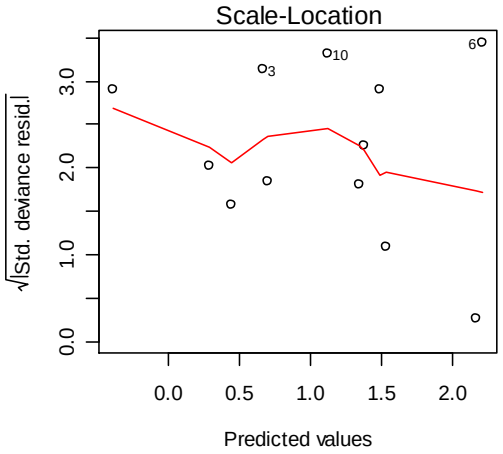
Diagnostics

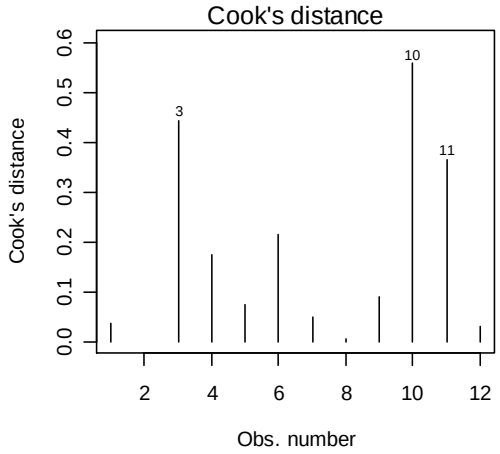
```
> par(mfrow=c(2,2))
> plot(model2,
    which=1:4)
```

No major problems









Conclusions

- Weak suggestion that *Grahami* relatively more numerous in mornings/midday
- Strong suggestion *Grahami* relatively more numerous on short perches
- Strong suggestion *Grahami* relatively more numerous on high perches

Over/under dispersion

- The variance of the binomial B(n,p) distribution is np(1-p), which is always less than the mean np.
- Sometimes the individuals having the same covariate pattern in a logistic regression may be correlated.
- This will result in the variance being ***greater*** than np(1-p) (if the correlation is +ve) or ***less than*** np(1-p) (if the correlation is - ve)

Over/under-dispersion

- If this is the case, we say the data are ***over-dispersed*** (if the variance is greater) or ***under-dispersed*** (if the variance is less)
- Consequence: standard errors will be wrong.
- Quick and dirty remedy: analyse as a binomial, but allow the "scale factor" to be arbitrary: this models the variance as

  $\psi np(1-p)$ where y is the "scale factor"

(for the binomial, the scale factor is always 1)

Over-dispersed model

```
> model3<-glm(cbind(r,n-r)~time+length+height,
      family=quasibinomial,data=lizard.df)
> summary(model3)
Call:
glm(formula = cbind(r, n - r) ~ time + length + height,
family = quasibinomial, data = lizard.df)
> Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.49466    0.33128   4.512  0.00276 **
timelate    -1.05278    0.32228  -3.267  0.01374 *
timemid      0.04003    0.27565   0.145  0.88864
lengthshort  0.67630    0.23675   2.857  0.02446 *
heightlow   -0.83011    0.26683  -3.111  0.01706 *
---
(Dispersion parameter for quasibinomial family taken to be
1.322352)
Null deviance: 54.0430  on 11  degrees of freedom
Residual deviance:  9.8815  on  7  degrees of freedom
```

Comparison

Binomial

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.49466	0.28809	5.188	2.12e-07	***
timelate	-1.05278	0.28026	-3.756	0.000172	***
timemid	0.04003	0.23971	0.167	0.867384	
lengthshort	0.67630	0.20588	3.285	0.001020	**
heightlow	-0.83011	0.23204	-3.578	0.000347	***

Quasibinomial

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.49466	0.33128	4.512	0.00276	**
timelate	-1.05278	0.32228	-3.267	0.01374	*
timemid	0.04003	0.27565	0.145	0.88864	
lengthshort	0.67630	0.23675	2.857	0.02446	*
heightlow	-0.83011	0.26683	-3.111	0.01706	*