Today’s agenda

- In today’s lecture we continue our discussion of logistic regression, with a few R tricks thrown in.

- Topics
  - R tricks
  - Binary anova
  - Variable selection
  - Under/over dispersion

Coursebook reference: Section 5.2.7
> A=matrix(1:4,2,2)
> A

[,1] [,2]
[1,] 1 3
[2,] 2 4
> apply(A,1,sum) # sum rows
[1] 4 6
> apply(A,2,max) # find max of columns
[1] 2 4
R Tricks: functions

# functions
moments = function(x,r=2){
  xbar = mean(x)
  mean(((x-xbar)^r)
}
> x=rnorm(10000)
> moments(x,3)
[1] -0.06573193
R Tricks: apply with functions

```r
> X = matrix(rnorm(15), 5,3)
> X[5,1]=NA
> X

[,1]      [,2]      [,3]
[1,] 1.09434119 0.3995507 0.5876889
[2,] 0.37134630 -0.3494054 1.9481628
[3,] 0.05114511 0.3633751 -0.6316826
[4,] 0.82542167 -1.0183255 -1.5921251
[5,]    NA  0.2775317 0.1498247

# write function to check for NA's
# expects a vector x
any.na = function(x){
  y = is.na(x)
  any(y)
}

# more compact version
any.na = function(x)any(is.na(x))
```
R Tricks: removing rows

```r
> any.na(X[,1])
[1] TRUE

> na.rows = apply(X, 1, any.na)
> na.rows
[1] FALSE FALSE FALSE FALSE TRUE

> X[!na.rows,]

[,1]       [,2]       [,3]
[1,] 1.09434119 0.3995507 0.5876889
[2,] 0.37134630 -0.3494054 1.9481628
[3,] 0.05114511 0.3633751 -0.6316826
[4,] 0.82542167 -1.0183255 -1.5921251
```
R Tricks: using a built-in function

```r
> no.na = complete.cases(X)
> no.na
[1] TRUE  TRUE  TRUE  TRUE  FALSE
> X[no.na,]
     [,1]       [,2]       [,3]
[1,] 1.094341  0.3995507  0.5876889
[2,] 0.3713463 -0.3494054  1.9481628
[3,] 0.05114511 0.3633751 -0.6316826
[4,] 0.82542167 -1.0183255 -1.5921251
```
R Tricks: Vectors and matrices

> # vectors
> x=c(2,4,2,7)
> x
[1] 2 4 2 7
> x[3]
[1] 2
> length(x)
[1] 4
> range(x)
[1] 2 7
> # matrices
> A[1,2]
[1] 3
[1] 5
> sum(A)
[1] 10
> # lists
> my.list = list(numbers=c(3,4,7,2),aMatrix=A)
> my.list$aMatrix
>  [,1]  [,2]
> [1,] 1  2
> [2,] 3  4
> my.list[[2]]
>  [,1]  [,2]
> [1,] 1  2
> [2,] 3  4
> names(my.list)
> [1] "numbers" "aMatrix"
> y=rnorm(10); x=rnorm(10); lm.obj = lm(y~x)
> names(lm.obj)
> [1] "coefficients" "residuals" "effects" "rank"
> [5] "fitted.values" "assign" "qr" "df.residual"
> [9] "xlevels" "call" "terms" "model"
> lm.obj$df.residual
> [1] 8
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.

This is the model containing all possible interactions i.e. if there are 4 factors A, B, C, D the model is A*B*C*D.
Binary Anova

- The plum tree data: see the coursebook, p 139
- 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 are in R330 data frame plum.df.
<table>
<thead>
<tr>
<th>length</th>
<th>time</th>
<th>r</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>long</td>
<td>autumn</td>
<td>156</td>
<td>240</td>
</tr>
<tr>
<td>long</td>
<td>spring</td>
<td>84</td>
<td>240</td>
</tr>
<tr>
<td>short</td>
<td>autumn</td>
<td>107</td>
<td>240</td>
</tr>
<tr>
<td>short</td>
<td>spring</td>
<td>31</td>
<td>240</td>
</tr>
</tbody>
</table>
> 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 $r/n$
The fitted logit is

\[
\log \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right)
\]

For a saturated (maximal model) that puts no restrictions on the probabilities, \( \hat{\pi} = r/n \). Thus, the fitted logit is

\[
\log \left( \frac{r/n}{1 - r/n} \right) = \log \left( \frac{r}{n - r} \right).
\]

If \( r = 0 \) or \( r = n \) we set \( r = 0.5 \) or \( r = n - 0.5 \).
Fitted logits: hand calculation

The baselines are for time: Autumn, for length: Long. The coefficients are

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.6190</td>
</tr>
<tr>
<td>lengthshort</td>
<td>-0.8366</td>
</tr>
<tr>
<td>timespring</td>
<td>-1.2381</td>
</tr>
<tr>
<td>lengthshort:timespring</td>
<td>-0.4527</td>
</tr>
</tbody>
</table>

Fitted logit for length = long, time = autumn is **0.6190**
Fitted logit for length = long, time = spring is

\[0.6190 - 1.2381 = -0.6191\]

Fitted logit for length = short, time = autumn is

\[0.6190 - 0.8366 = -0.2176\]

Fitted logit for length = long, time = spring is

\[0.6190 - 1.2381 - 0.8366 - 0.4527 = -1.9084\]

In R:

```r
> predict(plum.glm)
[1] 0.6190392 -0.6190392 -0.2175203 -1.9083470
```
interaction.plot(plum.df$length, plum.df$time, log(plum.df$s/(plum.df$n-plum.df$s)), cex.lab=1.5, cex.axis=1.5, cex=1.5, ylab = "average logit", xlab = "length", lwd=2)
> 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 | 151.019 | 3 | 151.019 |<|  |
| length | 45.837 | 2 | 105.182 | 1.285e-11 |
| time | 102.889 | 1 | 2.294 | 3.545e-24 |
| length:time | 2.294 | 0 | 7.727e-14 | 0.130 |

> 1-pchisq(2.294,1)
[1] 0.1298748

Interaction not significant!
Additive model

> 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 is negative)
Prob of survival less for spring planting (coeff is positive)

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 that no-interaction model fits well.
Fitted probabilities

<table>
<thead>
<tr>
<th></th>
<th>length=long</th>
<th>length=short</th>
</tr>
</thead>
<tbody>
<tr>
<td>time =autumn</td>
<td>0.6712</td>
<td>0.4246</td>
</tr>
<tr>
<td>time =spring</td>
<td>0.3288</td>
<td>0.1504</td>
</tr>
</tbody>
</table>

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

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

\[
AIC = \text{Deviance} + 2 \times (\text{number of parameters})
\]

- Pick model with smallest AIC
- Calculate with AIC function
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 a *grahami*. 
### The data

```r
> lizard.df

<table>
<thead>
<tr>
<th>length</th>
<th>height</th>
<th>time</th>
<th>r</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>short</td>
<td>low</td>
<td>early</td>
<td>54</td>
<td>67</td>
</tr>
<tr>
<td>short</td>
<td>high</td>
<td>early</td>
<td>44</td>
<td>49</td>
</tr>
<tr>
<td>long</td>
<td>low</td>
<td>early</td>
<td>25</td>
<td>43</td>
</tr>
<tr>
<td>long</td>
<td>high</td>
<td>early</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>short</td>
<td>low</td>
<td>mid</td>
<td>77</td>
<td>98</td>
</tr>
<tr>
<td>short</td>
<td>high</td>
<td>mid</td>
<td>63</td>
<td>67</td>
</tr>
<tr>
<td>long</td>
<td>low</td>
<td>mid</td>
<td>64</td>
<td>97</td>
</tr>
<tr>
<td>long</td>
<td>high</td>
<td>mid</td>
<td>21</td>
<td>26</td>
</tr>
<tr>
<td>short</td>
<td>low</td>
<td>late</td>
<td>22</td>
<td>36</td>
</tr>
<tr>
<td>short</td>
<td>high</td>
<td>late</td>
<td>25</td>
<td>38</td>
</tr>
<tr>
<td>long</td>
<td>low</td>
<td>late</td>
<td>13</td>
<td>24</td>
</tr>
<tr>
<td>long</td>
<td>high</td>
<td>late</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>
```
Proportion of grahami lizards higher when perches are short and high, and in the earlier part of the day.
Model selection

- Full model is

\[
\text{cbind}(r, n - r) \sim \text{time} \ast \text{length} \ast \text{height}
\]

so fit this first.

- Then use anova and stepwise to select a simpler model if appropriate.
\[
\text{lizard.glm} \leftarrow \text{glm(cbind(r, n-r)~time*length*height, + family=binomial, data=lizard.df)}
\]

\[
\text{anova(lizard.glm, test="Chisq")}
\]

| 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 |

Suggests model
\[
\text{cbind(r, n - r) } \sim \text{ time + length + height}
\]
```r
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(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
AIC

```r
> AIC(glm(cbind(r,n-r)~time*length*height, + family=binomial,data=lizard.df))
[1] 68.21162
> AIC(glm(cbind(r,n-r)~time*length*height - + time:length:height, family=binomial, data=lizard.df))
[1] 67.90428
> AIC(glm(cbind(r,n-r)~time+length+height + + time:length+time:length:height, family=binomial, data=lizard.df))
[1] 65.90559
> AIC(glm(cbind(r,n-r)~time+length+height, + family=binomial,data=lizard.df))
[1] 64.09314
```

Suggests model

\[ \text{cbind}(r, n - r) \sim \text{time} + \text{length} + \text{height} \]
par(mfrow=c(2,2))
plot(model2)
Conclusions

- Strong 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 (cont)

- 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 $\psi$ 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 *   |

Standard errors bigger with quasibinomial.