In today’s lecture we continue our discussion of the logistic regression model: Topics covered are

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

Reference: Coursebook, section 5.2.3, 5.2.4
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 a binary response \( Y \), the model is

\[
\pi = Pr[Y = 1] = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.
\]
Odds & log-odds form

Odds form:
\[
\frac{\pi}{1 - \pi} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)
\]

Log-odds form:
\[
\log \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.
\]
Interpretation of coefficients

- As before, a unit increase in $x_j$ multiplies the odds by $\exp(\beta_j)$.
- As before, a unit increase in $x_j$ increases the log-odds by $\beta_j$. 
To group 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” ($r$).
For grouped data, the log-likelihood is

\[ \ell(\beta_0, \ldots, \beta_k) = \sum_{i=1}^{m} \left\{ r_i (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) - n_i \log(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})) \right\}. \]

The \( i \)th covariate pattern is \( (x_{i1}, \ldots, x_{ik}) \).

The \( \beta \)'s are chosen to maximise this expression, using an algorithm known as “iteratively weighted least squares” (IRLS).
For ungrouped data:

The log-likelihood is

$$\ell(\beta_0, \ldots, \beta_k) = \sum_{i=1}^{n} \{ y_i (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) - \log(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})) \}.$$  

Again, the $\beta$’s are chosen to maximise this expression. The two forms give the same result.
Example: Kyphosis risk factors

- Kyphosis is a curvature of the spine that may be corrected by spinal surgery. However, not all surgery is successful, and the condition may still be present after the corrective operation.
- In a study to determine risk factors for this, the following variables were measured:
  - **Kyphosis**: (binary, absent=no kyphosis, present=kyphosis)
  - **Age**: continuous, age in months
  - **Start**: continuous, vertebrae level of surgery
  - **Number**: continuous, no of vertebrae involved.
Example: Illustration
Data

Data are in R330 data frame `kyphosis.df`:

<table>
<thead>
<tr>
<th>Kyphosis</th>
<th>Age</th>
<th>Number</th>
<th>Start</th>
</tr>
</thead>
<tbody>
<tr>
<td>absent</td>
<td>71</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>absent</td>
<td>158</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>present</td>
<td>128</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>absent</td>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>absent</td>
<td>1</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>absent</td>
<td>1</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>absent</td>
<td>61</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>absent</td>
<td>37</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>absent</td>
<td>113</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>present</td>
<td>59</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

... 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”:

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

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

Loess smoother added, quadratic effect?
Age around 100 months seems bad.
Fitting (1)

Seems age is important, fit as a quadratic

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

Call:
glm(formula = Kyphosis ~ Age + I(Age^2) + Start + Number,
       family = binomial, data = kyphosis.df)

Deviance Residuals:

     Min  1Q Median  3Q Max
-2.23573 -0.51241 -0.24509 -0.06108 2.35495

Coefficients:

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.3835660  2.0548871 -2.133 0.03291 *
   Age        0.0816412  0.0345292  2.364 0.01806 *
I(Age^2)   -0.0003965  0.0001905 -2.082 0.03737 *
     Start   -0.2038421  0.0706936 -2.883 0.00393 **
    Number    0.4268659  0.2365134  1.805 0.07110 .

(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

Number of Fisher Scoring iterations: 6
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*
Recall our model had two parts:

1. The binomial assumption \((r \text{ is } Bin(n, \pi))\)
2. The logistic assumption (logit of \(\pi\) 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 \(\pi\)'s, one for each covariate pattern:

\[
L(\pi_1, \ldots, \pi_M) = \prod_{i=1}^{M} \binom{n_i}{r_i} \pi_i^{r_i} (1 - \pi_i)^{n_i - r_i}
\]
The log-likelihood is (ignoring bits not depending on the $\pi$’s)

$$\ell(\pi_1, \ldots, \pi_m) = \sum_{i=1}^{m} \{ r_i \log \pi_i + (n_i - r_i) \log(1 - \pi_i) \}.$$ 

The maximum value of this (log)-likelihood is when $\pi_i = r_i/n_i$. If $r_i$ equals 0 or $n_i$ then use $0 \log 0 = 0$ Call this maximum value of $L L_{\text{max}}$. 
$L_{max}$ 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. Recall:
  - $m$: number of covariate patterns
  - $k$: number of covariates
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.
Null deviance

- At the other extreme, the most restrictive model is one where all the probabilities $\pi_i$ are the same (i.e. they 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 greater than the null deviance.
Graphical interpretation

\[ L_{Null} \leq L_{Mod} \leq L_{Max} \]

so

\[ 2 \log L_{Null} \leq 2 \log L_{Mod} \leq 2 \log L_{Max} \]
Example: the 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
- Data are in the R330 data frame `budworm.df`
The data

> budworm.df

    sex dose s 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
  12   1  32 16 20
## Probabilities under the 3 models

```r
> data(budworm.df)
> max.mod.probs <- budworm.df$s / budworm.df$n
> budworm.glm <- glm( cbind(s, n-s) ~ sex + dose,
                  family=binomial, data = budworm.df)
> logist.mod.probs <- predict(budworm.glm, type="response")
> null.mod.probs <- sum(budworm.df$s) / sum(budworm.df$n)
> cbind(max.mod.probs, logist.mod.probs, null.mod.probs)

<table>
<thead>
<tr>
<th></th>
<th>max.mod.probs</th>
<th>logist.mod.probs</th>
<th>null.mod.probs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.2677414</td>
<td>0.4625</td>
</tr>
<tr>
<td>2</td>
<td>0.20</td>
<td>0.3002398</td>
<td>0.4625</td>
</tr>
<tr>
<td>3</td>
<td>0.45</td>
<td>0.3713931</td>
<td>0.4625</td>
</tr>
<tr>
<td>4</td>
<td>0.65</td>
<td>0.5283639</td>
<td>0.4625</td>
</tr>
<tr>
<td>5</td>
<td>0.90</td>
<td>0.8011063</td>
<td>0.4625</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>0.9811556</td>
<td>0.4625</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>0.1218892</td>
<td>0.4625</td>
</tr>
<tr>
<td>8</td>
<td>0.10</td>
<td>0.1400705</td>
<td>0.4625</td>
</tr>
<tr>
<td>9</td>
<td>0.30</td>
<td>0.1832034</td>
<td>0.4625</td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.2983912</td>
<td>0.4625</td>
</tr>
<tr>
<td>11</td>
<td>0.60</td>
<td>0.6046013</td>
<td>0.4625</td>
</tr>
<tr>
<td>12</td>
<td>0.80</td>
<td>0.9518445</td>
<td>0.4625</td>
</tr>
</tbody>
</table>
```
Logits under the 3 models

```r
> max.logit = log((budworm.df$s+0.5)/(budworm.df$n -budworm.df$s+0.5))
> model.logit = predict(budworm.glm)
> cbind(max.logit,model.logit)

       max.logit model.logit
     1  -2.5649494  -1.0061121
     2  -1.2992830  -0.8461564
     3  -0.1910552  -0.5262451
     4   0.5877867   0.1135776
     5   2.0014800   1.3932230
     6   3.7135721   3.9525137
     7  -3.7135721  -1.9746604
     8  -2.0014800  -1.8147047
     9  -0.8023465  -1.4947934
    10   0.0000000  -0.8549707
    11   0.3856625   0.4246747
    12   1.2992830   2.9839654
```
Plotting logits

Poor fit: maximal logits are not linear
Calculating the likelihoods

Likelihood is

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

\[
L_{\text{Max}} = 2.8947 \times 10^{-7}, \quad 2 \log L_{\text{Max}} = -30.1104
\]
\[
L_{\text{Mod}} = 2.4459 \times 10^{-13}, \quad 2 \log L_{\text{Mod}} = -58.0783
\]
\[
L_{\text{Null}} = 2.2142 \times 10^{-34}, \quad 2 \log L_{\text{Null}} = -154.9860
\]
Calculating the deviances

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

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

Coefficients:

|              | Estimate | Std. Error | z value | Pr(>|z|)     |
|--------------|----------|------------|---------|-------------|
| (Intercept)  | -1.16607 | 0.26155    | -4.458  | 8.26e-06    *** |
| sex          | -0.96855 | 0.32954    | -2.939  | 0.00329    **  |
| dose         | 0.15996  | 0.02341    | 6.832   | 8.39e-12   *** |

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 (20), $m$ small (12)
- Can interpret residual deviance as a measure of fit
  
  $1 - \text{pchisq}(27.968, 9)$
  
  [1] 0.0009656815
- Not a good fit!! (as we suspected from the plot) In actual fact log(dose) works better
> logdose.glm<-glm( cbind(r, n-r) ~ sex + log(dose),
family=binomial, data = budworm.df)
> summary(logdose.glm)

Call:
  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