Today’s agenda

- R tricks of the day: paste, loops
- Introduction to categorical and count responses
  - For the next few lectures, we concentrate on the first case, where the response is binary, or a proportion
  - Then we will move on to the analysis of count data, including the analysis of contingency tables
- Logistic regression: first ideas
Today’s R tricks

paste

> paste("John", "Smith")
[1] "John Smith"
> paste("Graph", 1:3)
[1] "Graph 1" "Graph 2" "Graph 3"
> paste(c("y ~ A","B","A:B"),collapse="+")
[1] "y ~ A+B+A:B"
Today’s R tricks (2)

Looping, par

data(cherry.df)
varnames = names(cherry.df)
graph.titles = paste("Histogram of ", varnames)
par(mfrow=c(1,3))
for(j in 1: length(varnames)){
  hist(cherry.df[,j], xlab = varnames[j],
       main = graph.titles[j])
}
Histogram of diameter

Histogram of height

Histogram of volume
The prevalence of coronary heart disease (CHD) depends very much on age: the probability that a person randomly chosen from a population is suffering from CHD depends on the age of the person (and on lots of other factors as well, such as smoking history, diet, exercise and so on).

CHD data: There are 2 variables

  AGE: age in years (continuous variable)
  CHD: 0=no CHD, 1=CHD (binary variable)
## CHD data (cont)

<table>
<thead>
<tr>
<th>age</th>
<th>chd</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>39</td>
<td>1</td>
</tr>
<tr>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>44</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>54</td>
<td>1</td>
</tr>
<tr>
<td>56</td>
<td>1</td>
</tr>
<tr>
<td>57</td>
<td>1</td>
</tr>
<tr>
<td>....</td>
<td></td>
</tr>
</tbody>
</table>

NB: 0=no chd, 1=chd
library(R330)
data(chd.df)
plot(chd ~ age, data = chd.df, pch = 19)
Ungrouped and grouped data

- In the data set on the previous slide, every line of the data file corresponds to an individual in the sample. This is called **ungrouped data**
- If there are many identical ages, a more compact way of representing the data is as **grouped data**
CHD data (cont)

```
attach(chd.df)
sorted.chd.df=chd.df[order(age),]
sorted.chd.df
```

<table>
<thead>
<tr>
<th>age</th>
<th>chd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>30</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>31</td>
</tr>
</tbody>
</table>

Alternative way of entering data: record

- $n=$ Number of times each age occurs (30 occurs 5 times)
- $r=$ Number of CHD cases out of 5 (0 in this case)

Hence $r = 5$ and $n = 5$. 
CHD data: alternate form

```
    group.age r n
   1     20 0 1
   2     23 0 1
   3     24 0 1
   4     25 1 2
   5     26 0 2
   6     28 0 2
   7     29 0 1
   8     30 0 5
   9     31 1 1 ...
```

The number of lines in the data frame is now the number of distinct ages, not the number of individuals. This is grouped data
# ungrouped to grouped

```r
grouped.chd.df <- data.frame(
group.age = sort(unique(chd.df$age)),
r = tapply(chd.df$chd, chd.df$age, sum),
n = tapply(chd.df$chd, chd.df$age, length))
plot(I(r/n) ~ group.age, xlab = "age",
ylab = "r/n", data = grouped.chd.df)
```
Grouped data plot
For grouped data, the response variable is of the form “\( r \) successes out of \( n \) trials” (sounds familiar??)

The natural distribution for the response is binomial: the distribution of the number of “successes” (CHD!!!!) out of \( n \) trials

The probability of success, \( \pi \) say, depends on the age in some way

For each age, the probability \( \pi \) is estimated by \( r/n \)
Suppose there are $n$ people in the sample having a particular age (age 30 say). The probability of a 30 year-old-person in the target population having CHD is $\pi$, say. What is the probability $r$ out of the $n$ in the sample have CHD?

Use the binomial distribution:

$$\text{Prob}[r \text{ have CHD}] = \binom{n}{r} \pi^r (1 - \pi)^{n-r}$$
Probability of \( r \) out of \( n \) “successes” e.g. for \( n = 10, \ p = 0.25 \): 

\[
\begin{align*}
&> r = 0:10 \\
&> n=10 \\
&> p=0.25 \\
&> \text{probs} = \text{dbinom}(r,n, p) \\
&> \text{names(probs)} = r \\
&> \text{probs} \\
0 & 1 & 2 \\
5.631351e-02 & 1.877117e-01 & 2.815676e-01 \\
3 & 4 & 5 \\
2.502823e-01 & 1.459980e-01 & 5.839920e-02 \\
6 & 7 & 8 \\
1.622200e-02 & 3.089905e-03 & 3.862381e-04 \\
9 & 10 \\
2.861023e-05 & 9.536743e-07 \\
\end{align*}
\]

\[
> \text{barplot(probs, col="red")}
\]
Relationship between $\pi$ and age

Could try $\pi = a + b \times \text{AGE}$:
Better: Logistic function

\[ \pi = \frac{\exp(a + b \times \text{AGE})}{1 + \exp(a + b \times \text{AGE})} \]

\(a\) and \(b\) are constants controlling shape of curve
Logistic regression

- **To sum up:**
  - In a random sample of \( n \) persons aged \( x \), the number that have CHD has a binomial distribution \( Bin(n, \pi) \).
  - The probability \( \pi \) is related to age by the logistic function
    \[
    \pi = \frac{\exp(a + b \times AGE)}{1 + \exp(a + b \times AGE)}
    \]

- This is called the **logistic regression model**.
Interpretation of $a$ and $b$

- If $b > 0$, $\pi$ increases with increasing age
- If $b < 0$, $\pi$ decreases with increasing age
- Slope of curve when $\pi = 0.5$ is $b/4$
- If $a$ is large and positive, the probabilities $\pi$ for any reasonable age are high
- If $a$ is large and negative, the probabilities $\pi$ for any reasonable age are low
Estimation of $a$ and $b$

- To estimate $a$ and $b$, we use the method of maximum likelihood

- Basic idea:
  - Using the binomial distribution, we can work out the probability of getting any particular set of responses.
  - In particular, we can work out the probability of getting the data we actually observed. This will depend on $a$ and $b$.
  - Choose $a$ and $b$ to maximise this probability
Is this reasonable?

- Here is a “thought experiment” to convince you that it is...

- Suppose that
  - the value of the parameter $a$ is known to be 0.
  - the true value of the parameter $b$ is either 1 or 2, but we don’t know which.
Decisions, decisions...

- You observe some data, say the actual $r/n$ values.

- Suppose you can calculate the probability of observing the data, provided you know the value of $a$ and $b$.

- You will win $1000 if you correctly guess which is the true value of $b$ (i.e. either 1 or 2.)

- How do you decide?
Decisions, decisions...

- You calculate
  - if $b = 1$, prob of observing data is $10^{-2}$
  - if $b = 2$, prob of observing data is $10^{-20}$

- Which value do you pick, 1 or 2????

- There are 2 possibilities
  - The true value of $b$ is 1, or
  - The true value of $b$ is 2 and a really rare event occurred.

- A no brainer, you pick $b = 1$. 
Likelihood

- The probability of getting the data we actually observed depends on the unknown parameters $a$ and $b$
- As a function of those parameters, it is called the likelihood
- We choose $a$ and $b$ to maximise the likelihood
- These estimates are maximum likelihood estimates
Log likelihood

- The values of $a$ and $b$ that maximise the likelihood are the same as those that maximize the log of the likelihood. This is because the log function is an increasing function.

- The log of the likelihood is called the log-likelihood and is denoted by $l$ (letter ell).

- For our logistic regression model, the log likelihood can be written down. The form depends on whether the data is grouped or ungrouped.
For grouped data:

For $m$ groups:

- There are $m$ distinct ages, $x_1, \ldots, x_m$.
- The repeat counts are $n_1, \ldots, n_m$.
- The CHD counts are $r_1, \ldots, r_m$.
- The log-likelihood is

$$l(a, b) = \sum_{i=1}^{m} \left\{ r_i(a + bx_i) - n_i \log(1 + \exp(a + bx_i)) \right\}.$$
For ungrouped data:

- There are $n$ individuals, with ages $x_1, \ldots, x_n$. (there could be repeats)
- The CHD indicators are $y_1, \ldots, y_n$, taking values 0 (=no CHD) or 1(=CHD).
- The log-likelihood is

$$l(a, b) = \sum_{i=1}^{n} \{y_i(a + bx_i) - \log(1 + \exp(a + bx_i))\}.$$
Maximizing the log-likelihood

- The log-likelihood on the previous 2 slides is maximized using a numerical algorithm called iteratively reweighted least squares, or IRLS.
- This is implemented in R by the \texttt{glm} function
- GLM = generalised linear model, a class of models including logistic regression
Doing it in R

For ungrouped data (in data frame chd.df)

```r
> chd.glm <- glm(chd ~ age, family = binomial, data = chd.df)
> summary(chd.glm)
Call:
glm(formula = chd ~ age, family = binomial, data = chd.df)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.9686  -0.8480  -0.4607   0.8262   2.2794

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.2784    1.1296  -4.673  2.97e-06 ***
age            0.1103    0.0240   4.596  4.30e-06 ***
---
```

Use binomial to compute log-likelihood

Estimate of $\alpha$

Estimate of $\beta$
Doing it in R

For grouped data (in data frame grouped.chd.df)

```r
> g.chd.glm<-glm(cbind(r,n-r) ~ age, family = binomial, data=grouped.chd.df)
> summary(g.chd.glm)
Call:
glm(formula = cbind(r, n - r) ~ age, family = binomial, data = grouped.chd.df)
Deviance Residuals:
          Min          1Q       Median          3Q         Max
 -2.50854   -0.61906    0.05055     0.59488    2.00167

Coefficients:  
               Estimate Std. Error z value  Pr(>|z|)  
(Intercept)  -5.27834    1.12690   -4.684 2.81e-06 ***
age           0.11032    0.02395    4.607 4.09e-06 ***
---
```

Points with fitted logistic curve

Probability of CHD at different ages
R code for picture

```r
> coef(chd.glm)
(Intercept)  age
  -5.2784444  0.1103208
> plot(I(r/n)~group.age, xlab= "age", ylab= "r/n", data= grouped.chd.df, cex=1.4, pch = 19, col="blue")
> ages=20:70
> lp = -5.2784444 + 0.1103208*ages
> probs = exp(lp)/(1+exp(lp))
> lines(ages, probs, col="red", lwd=2)
> title(main = "Probability of CHD at different ages")
```