In today’s lecture we finish the course with two final topics. Tomorrow we will summarise the whole course and give some guidance in studying for the final exam.

Topics:

▶ Prospective vs case-control sampling

▶ Regression and classification trees
Prospective sampling

- Suppose we want to estimate the association between lung cancer and smoking, using an odds ratio. We could take a random sample from our target population and cross-classify the resulting data into a contingency table:

<table>
<thead>
<tr>
<th>Lung Cancer</th>
<th>S = Yes</th>
<th>S = No</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC = Yes</td>
<td>$n_1$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>LC = No</td>
<td>$n_3$</td>
<td>$n_4$</td>
</tr>
</tbody>
</table>

- We would estimate the association using the odds ratio

$$\frac{n_1 n_4}{n_2 n_3}.$$ 

- The standard error of the log OR is (recall from STATS 201/8)

$$\sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}.$$
A problem

- The prevalence of lung cancer in the NZ population is about 130 per 100,000 persons. Thus, the chance of a randomly chosen person having lung cancer is about 0.0013.

- Unless the sample is very large, the counts \( n_1 \) and \( n_2 \) will be small, and hence the standard error of the log odds ratio will be large. We won’t get enough lung cancer sufferers in our sample to get an accurate estimate of the odds ratio.

- To get round this problem, we can sample in a different way.
Case-control sampling

- Take a sample of lung cancer cases (chosen for example from a cancer registry) and record their smoking status. These are the cases.

- Take a sample of lung-cancer free subjects (chosen for example from the general population) and record their smoking status. These are the controls.

- Form the contingency table as before and calculate the odds ratio.

- Is this valid????
A remarkable fact

- Not only does this give a valid estimate, but the standard error is also correct.

- Justification: in case-control sampling, we are sampling from the conditional distribution of smoking, given cancer.

- $\frac{n_1}{n_2}$ estimates the conditional odds $P[S = Y|\text{case}]/P[S = N|\text{case}]$

- $\frac{n_3}{n_4}$ estimates the conditional odds $P[S = Y|\text{control}]/P[S = N|\text{control}]$

- Thus, the odds ratio $\frac{n_1/n_2}{n_3/n_4}$ estimates

\[
\frac{P[S = Y|\text{case}]/P[S = N|\text{case}]}{P[S = Y|\text{control}]/P[S = N|\text{control}]}
\]
Some mathematics

\[
\frac{P[S = Y | \text{case}]}{P[S = N | \text{case}]} \cdot \frac{P[S = Y | \text{control}]}{P[S = N | \text{control}]}
\]

\[
= \frac{P[S = Y | \text{case}]P[S = N | \text{control}]}{P[S = Y | \text{control}]P[S = N | \text{case}]}
\]

\[
= \frac{(P[S = Y, \text{case}] / P[\text{case}]) \times (P[S = N, \text{control}] / P[\text{control}])}{(P[S = Y, \text{control}] / P[\text{control}]) \times (P[S = N, \text{case}] / P[\text{case}])}
\]

which is the OR under prospective sampling. Thus the “case-control” OR estimates the correct thing, namely the “prospective OR”. 
Example

In a famous article in the British Medical Journal (1950), Doll and Hill investigated the association between smoking and lung cancer in a case-control study. They obtained the following table:

<table>
<thead>
<tr>
<th>Lung Cancer</th>
<th>Smoker S = Yes</th>
<th>Smoker S = No</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC = Yes</td>
<td>688</td>
<td>21</td>
</tr>
<tr>
<td>LC = No</td>
<td>650</td>
<td>59</td>
</tr>
</tbody>
</table>

Note how many smokers!!
Odds ratios

- Odds of being a case for smokers = \( \frac{688}{650} \)
- Odds of being a case for non-smokers = \( \frac{21}{59} \)
- Odds ratio = \( \frac{688/650}{21/59} = 2.97 \).
- Odds of having lung cancer almost 3 times greater for smokers!!
Confidence interval

- Log odds $= \log(2.97) = 1.08$

- Standard error for log-odds is

$$\sqrt{\frac{1}{688} + \frac{1}{650} + \frac{1}{21} + 59} = 0.260$$

- CI for log-odds: (0.580, 1.599)

- CI for odds (1.787, 4.599)
Consider a model of the form

\[ y = f(x_1, \ldots, x_k) + e \]

where \( f \) is a member of some class of “flexible” functions e.g. linear functions (although linear functions are not very flexible.) By “flexible”, we mean able to capture a wide range of relationships.
An alternative to linear functions: trees

Suppose we have just two variables \((k = 2)\) and consider “step functions” that are constant on rectangular regions of the plane:
Fitting the step function

We choose the first region by splitting by either horizontally or vertically:
Fitting the step function (cont)

And again (but can’t undo earlier splits)
Fitting the step function (cont)

And again

$x_2$

$x_1$
Fitting the step function (cont)

And again, finally getting
Representing splits as a tree

\[ x_1 \leq 3 \quad \text{and} \quad x_1 > 3 \]

\[ x_1 \leq 5 \quad \text{and} \quad x_1 > 5 \]

\[ x_2 \leq 4 \quad \text{and} \quad x_2 > 4 \]

\[ x_2 \leq 2 \quad \text{and} \quad x_2 > 2 \]
Choosing the splits

When choosing how to split a region, possible splits are of the form $X_1 < c$, (a vertical split) or $X_2 < c$ (a horizontal split). Each split divides the data in a region into two groups $G_1$ and $G_2$ on the basis of their $x$-values. Let $\bar{y}_1$ and $\bar{y}_2$ be the mean $y$-values of these two groups, $\bar{y}$ the mean of all the data in the region. We choose the split that maximises the reduction in the residual sum of squares

$$\sum_{i} (y_i - \bar{y})^2 - \sum_{i \text{ in } G_1} (y_i - \bar{y}_1)^2 - \sum_{i \text{ in } G_2} (y_i - \bar{y}_2)^2$$
Fitted values

When the splits have been chosen, the fitted function value is the mean of the y-values of all points falling in the region:
Example: the cherry tree data

RSS is 483.1677 (421.9214 for least squares fit)
Logistic regression

- Can do the same thing for binary responses, except when choosing a split, we minimise $p(1 - p)$ where $p$ is the proportion of 1’s in the left hand group.
- To predict $y$, for a new case, we see which region the $x$’s fall in. Then if the majority of the training set data in that region are 1’s we predict the new case as a 1, otherwise as a zero.
Example: the spam data

The tree is

The error rate is about 5% (for logistic the error rate was about 7%)