SMIII lectures

null

Week 8

Stratification

Common odds ratios for several 2×2 tables.

Sometimes it happens that an experiment or observational study is replicated across many **strata**.

Example - admissions data for UC-Berkeley

- Observational study
- Applications for admission to graduate study
- Classified according to
  - whether the applicant was admitted;
  - the sex of the applicant;
  - the department applied to.

Purpose of analysis

The purpose of the analysis is to identify any systematic effect of sex upon the probability of admission.

- In this context,
- `Sex` is the "treatment'' variable,
- `Admission` is the response variable, and
- `Department` can be thought of as defining **strata**.

Contingency table - overall

The data can be represented as 2×2 contingency table.

```
## Warning in file_diff_dbl(chr): NAs introduced by coercio

## Warning: `columns = vars(...)` has been deprecated in gt
## * please use `columns = c(...)` instead

## Warning: `columns = vars(...)` has been deprecated in gt
## * please use `columns = c(...)` instead
```

	Admitted	Rejected	p
Female	557	1278	0.30
Male	1198	1493	0.45

Department

```
## Warning in file_diff_dbl(chr): NAs introduced by coercion

## Warning: ‘columns = vars(...)‘ has been deprecated in gt 0.3.0:
## * please use ‘columns = c(...)‘ instead

## Warning: ‘columns = vars(...)‘ has been deprecated in gt 0.3.0:
## * please use ‘columns = c(...)‘ instead
```

	Admitted	Rejected	p
A			
Male	512	313	0.62
Female	89	19	0.82
B			
Male	353	207	0.63
Female	17	8	0.68
C			
Male	120	205	0.37
Female	202	391	0.34
D			
Male	138	279	0.33
Female	131	244	0.35
E			
Male	53	138	0.28
Female	94	299	0.24
F			

Odds ratios

Let π_{ij} denote the probability of admission for sex j (F: $j = 1$; M: $j = 2$) in department i .

	Department i		
	Admitted	Rejected	Total
Female	π_{i1}	$1 - \pi_{i1}$	1
Male	π_{i2}	$1 - \pi_{i2}$	1

The log odds-ratio,

$$\beta_i = \log\left(\frac{\pi_{i2}/(1-\pi_{i2})}{\pi_{i1}/(1-\pi_{i1})}\right)$$

can be used as a measure of the association between `Sex` and `Admission`.

In particular

$\beta_i > 0 \Rightarrow$Admission more likely for males;

$\beta_i < 0 \Rightarrow$Admission more likely for females;

$\beta_i = 0 \Rightarrow$Admission equally likely for males and females.

Example - Department A

```r
# Female
p11 <-  89 / (89 + 19)
# Male
p12 <-  512 / (512 + 313)
p11
```

```
## [1] 0.8240741
```

```r
p12
```

```
## [1] 0.6206061
```

```r
b1 <- log((p12 / (1 - p12)) / (p11 / (1 - p11)))
b1
```

```
## [1] -1.052076
```

The common odds ratio model

Consider now the logistic regression model

$$\text{logit}(\pi_{ij}) = \mu + \alpha_i + \beta_j. \quad (\dagger)$$

It follows that

$$\log\left(\frac{\pi_{i2}/(1-\pi_{i2})}{\pi_{i1}/(1-\pi_{i1})}\right) = \text{logit}(\pi_{i2}) - \text{logit}(\pi_{i1}) = \beta_2 - \beta_1.$$

Analysis of the Berkeley Data

In the analysis, the (common) log odds-ratio is estimated to be -0.09987 with a standard error of 0.08085.

This is not significantly different from 0, which implies no strong evidence of discrimination within department in either direction.

Simpson's Paradox

The Department effects are highly significant and cannot be removed, even though they are not of direct interest.

For the model

$$\text{logit}(\pi_{ij}) = \mu + \beta_j.$$

The estimated log odds-ratio is now 0.61035 with a standard error of 0.06389.

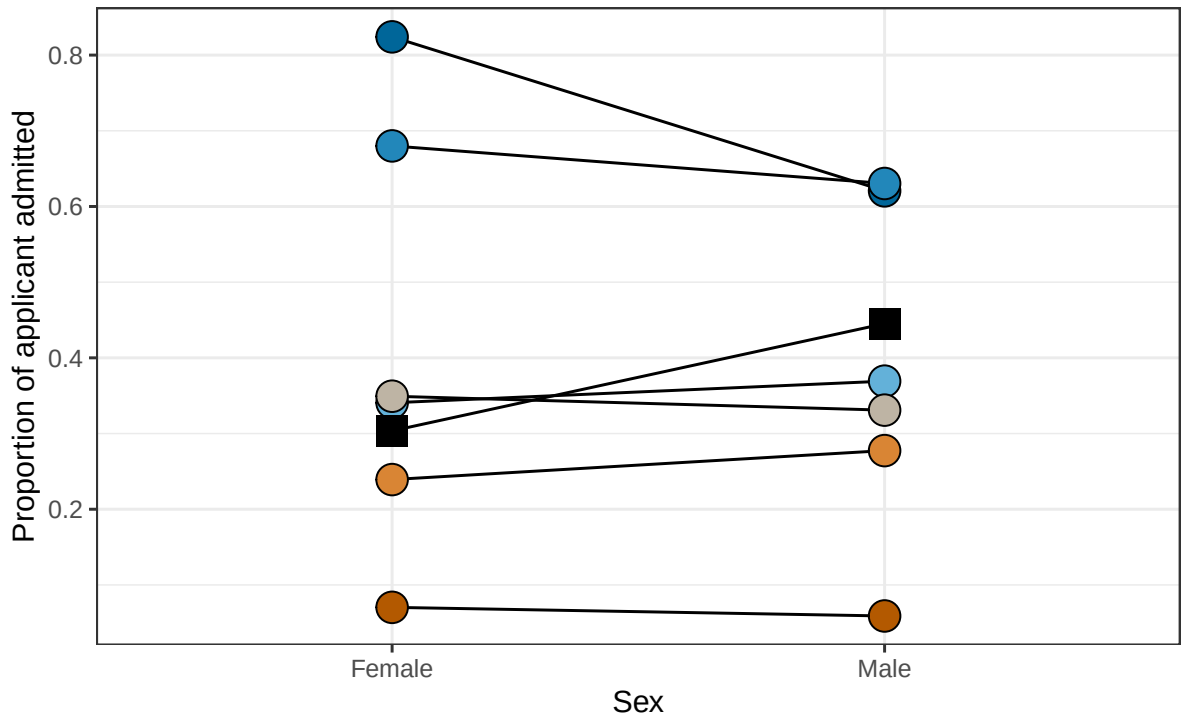
This is indicates a strong apparent bias towards male applicants.

This phenomenon is an example of **Simpson's paradox**.

There is no evidence of bias within any of the 6 departments, but if the data from all of the departments are combined, there is an apparent bias toward male applicants.

Proportion plot

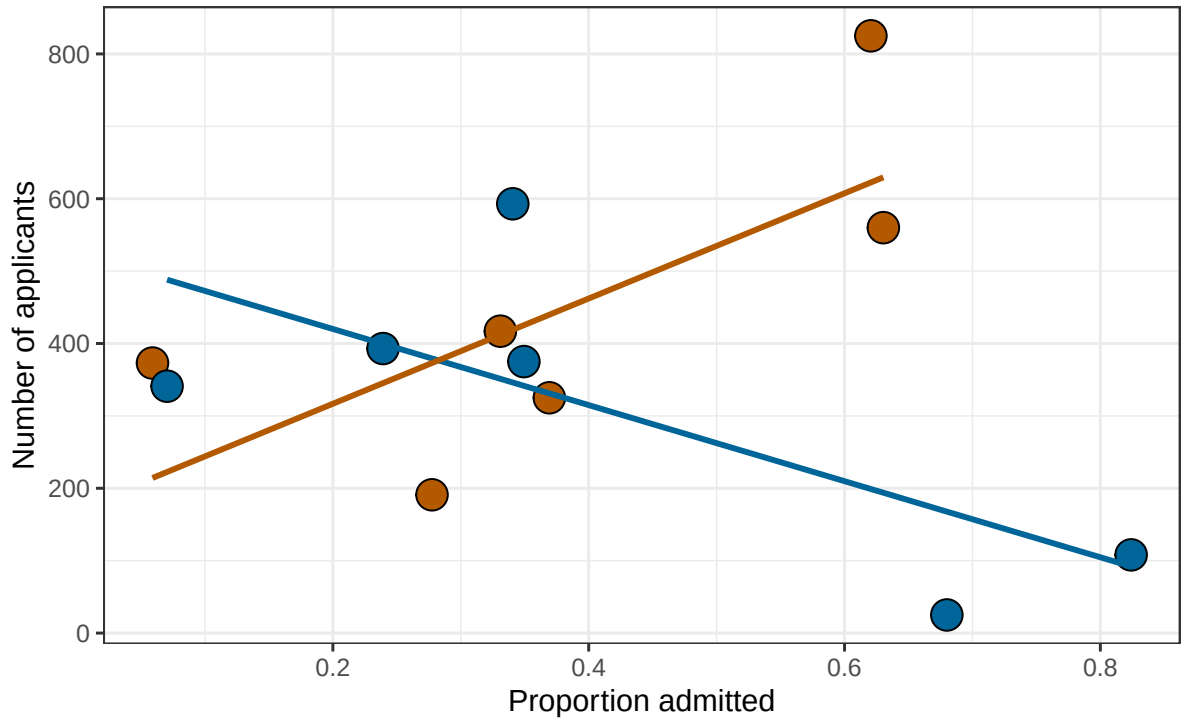
```
## Warning in file_diff_dbl(chr): NAs introduced by coercio
```



Proportion plot

```
## Warning in file_diff_dbl(chr): NAs introduced by coercio

## `geom_smooth()` using formula 'y ~ x'
```



Prospective and retrospective studies

Prospective and retrospective studies

One of the key assumptions for linear regression models is that each observation of the response variable, y, can be thought of as an observation from the corresponding conditional distribution of $Y|\boldsymbol{x}$.

- For example, in the beetle experiment, it is natural to think of the number of beetles that died at a particular concentration as a binomial response.
- Experiments or observational studies of this type are often called **prospective studies**:
  - In the beetle example, we expose a number of beetles to a certain concentration of carbon disulphide and observe the numbers that die.
  - In an epidemiological study, we might identify individuals with given levels of a certain risk factor (e.g. smoking status) and then observe the number that develop a given disease (e.g. emphysema) within a give time.

Retrospective studies

An alternative design that is frequently used in epidemiological studies is called a **retrospective** experiment.

- In this case, individuals are selected according to their response status and then the exposure to the risk factors of interest is determined.
- For example, a retrospective study to determine the effect of smoking on emphysema could:
  - Select a group of emphysema patients (sometimes called cases);
  - Select a comparable group of subjects without emphysema (sometimes called controls);
  - Determine the relevant smoking history for all of the subjects.

Analysis for retrospective studies

At first sight, it would appear that retrospective studies require a very different analysis to ordinary prospective studies.

While this is generally the case, for certain retrospective studies, the logistic regression model can be validly applied as if it were a prospective study.

Logistic regression and retrospective studies

Consider a population in which the binary response Y satisfies the logistic regression model,

$$P(Y = 1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})}.$$

Now let S be a sampling indicator defined by

$$S = \begin{cases} 1 & \text{if the subject is sampled} \\ 0 & \text{if the subject is not sampled} \end{cases}$$

Logistic regression and retrospective studies

The assumptions for retrospective sampling can be stated as

$$P(S = 1|Y = 1, \boldsymbol{x}) = \rho_1 \text{ and } P(S = 1|Y = 0, \boldsymbol{x}) = \rho_0.$$

- This assumption implies that cases are sampled at rate ρ_1 and controls at rate ρ_0.
- There is no requirement that $\rho_1 = \rho_0$ but
  - it is critical that the probability of being sampled does not depend on the predictor $\boldsymbol{x}$.

Logistic regression and retrospective studies

Now consider the success probability

$$P(Y = 1|S = 1, \boldsymbol{x})$$

for a subject sampled retrospectively.

- To justify the claim that logistic regression can be used, we must show that

$$P(Y = 1|S = 1, \boldsymbol{x})$$

satisfies a logistic regression model with the same parameters.

Logistic regression and retrospective studies

Hence, the probability of the positive outcome $Y = 1$ in a retrospective sample satisfies the same logistic regression model as for a prospective sample

- except that the intercept parameter β_0 is not estimable.

Remarks

- It should be intuitively plausible the intercept term cannot be estimated in a retrospective design. If it could, we would be able to estimate the probability of the positive response $Y = 1$ which is clearly not possible with retrospective sampling.
- Retrospective designs are especially useful for rare diseases, such as cancer, where the rate may be as low as $1/1000$ and where the incubation period may be very long. If a prospective design was used in this situation, we would need to recruit and follow roughly 20,000 subjects in order to observe 20 cases of the disease.