Generalized Linear Models (2462)

Course Manual — Spring Term 2023

Module Organiser: Chris Frost

Lecturers: Chris Frost and Ruth Keogh

Practical Support: Tess Poole and Amy Macdougall

London School of Hygiene & Tropical Medicine

Generalized Linear Models (2462) 2023

Objectives

The overall objective of this course is to introduce the theory of Generalized Linear Models and illustrate their application to medical and epidemiological data. Specifically, by the end of this study module you should be able to

1. demonstrate an understanding of the theoretical basis of generalized linear models;
2. use generalized linear and other related models for analysis of discrete data;
3. present results in a form suitable for publication in a medical journal, and
4. have an appreciation of different analysis strategies.

Organization

The core of the module is sixteen timetabled lecture/practical sessions. The course manual provides the most detailed description of the material in each of these sessions and you are advised to read ahead if you can.

There will be a pre-recorded lecture for each of these sessions and a live practical. You are strongly advised to listen to the lecture before attending the practical session. These practical sessions will be live face-to-face sessions unless you have been otherwise advised.

Five of the lecture/practical sessions (2, 5, 6, 7 and 10) will be led by Ruth Keogh. The remainder will be given by the module organiser, Chris Frost. Practical support will be provided by Tess Poole and Amy Macdougall.

In addition, five online question and answer sessions will allow you to raise questions related to the material in the lectures and practicals.

See the module timetable for details of all sessions.

Assessment

This module has a written assessment and, for MSc. Medical Statistics students, will also be assessed in the summer exam.

The written assignment will be distributed on Wednesday 25th January. There will be a dedicated session to work on this in class on the afternoon of Tuesday 31st January. The submission deadline is noon on Wednesday 8th February.

Some key references

1. Dobson, A.J and Barnett, A.G. (2008) *An Introduction to Generalized Linear Models, Third Edition.* Chapman & Hall.
2. Collett, D. (2002) *Modelling Binary Data.* Chapman & Hall. [Accessible coverage of logistic regression & other models for binary data].
3. Hosmer, D.W. and Lemeshow, S. (2013) *Applied Logistic Regression.* Wiley Interscience.
4. Hardin, J.W. and Hilbe, J.M. (2012) *Generalized Linear Models and Extensions, Third Edition.* Stata Press. [Thorough coverage of GLMs using Stata].
5. Pawitan, Y. *In All Likelihood: statistical modelling and inference using likelihood.* Oxford University Press. [A very nice book looking at various aspects of likelihood, including GLMs].
6. Harrell, F.E. *Regression Modeling Strategies.* Springer. [Extensive coverage of practical strategies for modelling data].
7. Agresti, A. (1996) *An Introduction to categorical data Analysis.* Wiley. [An excellent reference for categorical data analysis, including logistic regression models].
8. McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models, Second Edition.* Chapman & Hall. [The original reference. Arguably somewhat dated now].

Table of Contents

1 Introduction to Generalized Linear Models and how to fit them in Stata 1
- 1.1 Aims and objectives . . . 1
- 1.2 Linear Regression (revision) . . . 1
  - 1.2.1 Simple linear regression . . . 1
  - 1.2.2 Multiple linear regression . . . 2
  - 1.2.3 Inference for linear regression . . . 4
  - 1.2.4 Lowess smoothing - Stata's `lowess` command . . . 5
- 1.3 The Exponential family of distributions . . . 6
  - 1.3.1 Two other important examples . . . 7
  - 1.3.2 Notes . . . 8
- 1.4 Definition of a Generalized Linear Model (GLM) . . . 8
  - 1.4.1 Example 1: Simple Linear Regression . . . 9
  - 1.4.2 Example 2: Poisson Regression with a single explanatory variable . 9
  - 1.4.3 Example 3: Logistic regression for a Bernouilli outcome with a single explanatory variable . . . 9
  - 1.4.4 Notes . . . 10
- 1.5 Fitting GLMs . . . 10
  - 1.5.1 Stata's `glm` command . . . 10
- 1.6 Practical 1 . . . 12

2 Types of Investigation 15
- 2.1 Aims . . . 15
- 2.2 Specifying research questions . . . 15
- 2.3 Different types of investigation . . . 15
  - 2.3.1 Classification . . . 15
  - 2.3.2 Implications of investigation type . . . 16
  - 2.3.3 The role of study design . . . 17
- 2.4 Properties of different types of investigation . . . 17
  - 2.4.1 Description . . . 17
  - 2.4.2 Prediction . . . 18
  - 2.4.3 Causality and explanation . . . 18
  - 2.4.4 Is there a fourth investigation type? . . . 19
- 2.5 An example . . . 19
- 2.6 Role of explanatory variables in different types of investigation . . . 19
  - 2.6.1 Prediction . . . 21
  - 2.6.2 Causality and explanation . . . 21
- 2.7 Example: the role of regression in different types of investigation . . . 22
  - 2.7.1 Prediction . . . 23

2.7.2 Causality and explanation . . . 23
2.7.3 The "Table 2 Fallacy" . . . 24
2.8 Other considerations . . . 24
2.9 Practical 2 . . . 27

3 Models for Binary Data 29
3.1 Examples of binary data . . . 29
3.2 Grouped and individual binary data . . . 29
3.3 GLM for binary data . . . 29
3.4 Notes . . . 31
3.5 Extended example . . . 32
3.5.1 The data . . . 32
3.5.2 Analysis aims . . . 32
3.5.3 Logistic regression model 1: dilution factor + herd . . . 32
3.5.4 Parameter estimation . . . 32
3.5.5 Logistic regression model 2: + dilution factor*herd . . . 34
3.5.6 Does the effect of dilution factor differ between herds 1 and 2? . . . 35
3.6 Analysis using other link functions . . . 35
3.6.1 The impact of changing a link function on interactions . . . 35
3.7 Models for individual binary data in Stata . . . 36
3.8 Practical 3 . . . 37

4 Model comparison and goodness of fit 41
4.1 Aims and Objectives . . . 41
4.2 Introduction . . . 41
4.3 Comparing nested models . . . 41
4.4 The saturated model, deviance, and goodness of fit . . . 44
4.4.1 The saturated model . . . 44
4.4.2 Deviance . . . 45
4.4.3 Deviance for grouped binary data . . . 46
4.4.4 Pearson's goodness of fit statistic . . . 47
4.5 Assessing goodness of fit for individual binary data . . . 47
4.6 Conclusions . . . 50
4.7 Practical 4 . . . 51

5 Estimating treatment effects using observational data 55
5.1 Aims . . . 55
5.2 Motivating example: treatment for kidney stones . . . 55
5.3 Defining a treatment effect (the 'estimand') . . . 57
5.3.1 Marginal and conditional treatment effects . . . 57
5.3.2 The 'do' notation . . . 57
5.4 Estimating treatment effects . . . 59
5.4.1 Marginal treatment effects . . . 59
5.4.2 Conditional treatment effects . . . 59
5.4.3 Marginal treatment effects revisited: standardization . . . 60
5.5 Estimating the treatment effect using logistic regression . . . 61
5.6 Extensions to continuous and multiple confounders . . . 62
5.7 Extension to a continuous outcome . . . 63
5.8 Estimates of uncertainty . . . 64

5.9 Concluding remarks . . . 64
5.10 Practical 5 . . . 66

6 Collapsibility and non-collapsibility **71**
6.1 Aims . . . 71
6.2 Introduction to collapsibility . . . 71
6.3 Example 1: No confounding by Z . . . 72
6.4 Example 2: With confounding by Z . . . 74
6.5 Using logistic regression . . . 76
6.6 Implications of non-collapsibility for randomized controlled trials . . . 78
6.7 Implications of non-collapsibility for observational studies . . . 80
6.8 Continuous outcomes . . . 80
6.9 Summary . . . 82
6.10 Practical 6 . . . 84

7 Logistic Regression in Cohort and Case-Control Studies **87**
7.1 Aims . . . 87
7.2 Study designs in epidemiology . . . 87
7.3 Studies with a single binary exposure . . . 88
7.3.1 Notation . . . 88
7.3.2 Example data . . . 89
7.3.3 Odds ratios . . . 89
7.3.4 Simple analyses . . . 91
7.4 Logistic regression with a single binary exposure . . . 91
7.4.1 Cohort studies . . . 91
7.4.2 Case-control studies . . . 92
7.4.3 Likelihoods . . . 92
7.4.4 Example using Stata . . . 93
7.5 An alternative logistic regression for case-control studies . . . 94
7.6 Logistic regression with multiple covariates . . . 97
7.6.1 Simple analyses . . . 97
7.6.2 Likelihood for cohort and case-control studies . . . 98
7.6.3 Logistic regression for a general case-control study . . . 98
7.7 Final remarks . . . 99
7.8 Practical 7 . . . 100

8 Count outcomes **103**
8.1 Aims & Objectives . . . 103
8.2 Motivation . . . 103
8.3 Poisson GLM . . . 103
8.4 Example - TRUST asthma trial . . . 104
8.5 Overdispersion . . . 106
8.5.1 Checking for overdispersion . . . 107
8.5.2 Random-effects for overdispersion . . . 108
8.5.3 Estimating equation and quasilikelihood approaches . . . 109
8.6 Summary . . . 110
8.7 Practical 8 . . . 112

9 Models for rates **115**

9.1 Examples of rates . . . 115
9.2 Construction of Poisson frequency records . . . 116
9.3 Poisson process . . . 116
9.4 Models for rates . . . 117
9.5 GLM for rates . . . 117
9.6 Example: British doctors study . . . 118
9.7 Summary . . . 121
9.8 Practical 9 . . . 123

10 Analysis Strategies **125**
10.1 Aims . . . 125
10.2 The need for an analysis strategy . . . 125
10.3 Motivating example: vitamin C study . . . 126
10.4 Model building for causal investigations: popular methods . . . 126
10.4.1 Adjusting for all covariates . . . 127
10.4.2 Stepwise selection methods . . . 127
10.4.3 Change in estimates method . . . 128
10.4.4 Example . . . 129
10.5 Model building for causal investigations: A strategy based on mean squared error . . . 132
10.5.1 Example . . . 133
10.6 Model building in prediction investigations . . . 133
10.6.1 Stepwise selection methods in prediction . . . 134
10.6.2 Example . . . 135
10.7 Including non-linear terms and interactions . . . 135
10.8 Use of analysis plans . . . 136
10.9 Conclusions . . . 137
10.10Practical 10 . . . 138

11 Model checking **141**
11.1 Aims & Objectives . . . 141
11.2 Checking GLMs — how models can be misspecified . . . 141
11.3 Specification of the linear predictor . . . 142
11.3.1 Pearson Residuals . . . 142
11.3.2 Deviance and Anscombe Residuals . . . 143
11.3.3 Residuals for models fitted with `glm` in Stata . . . 144
11.3.4 Use of residuals to examine the adequacy of the linear predictor . . . 144
11.4 Example: alcohol consumption in NHANES . . . 144
11.5 Covariate pattern residuals . . . 147
11.5.1 Example: alcohol consumption in NHANES . . . 148
11.6 Covariate pattern or individual residual plots? . . . 149
11.7 Pearson's goodness of fit test with groups defined by covariate patterns . . . 150
11.8 Link function . . . 150
11.9 Summary . . . 151
11.10Practical 11 . . . 152

12 Assessing model performance **155**
12.1 Aims & Objectives . . . 155
12.2 Calibration . . . 155

12.2.1 Flexible calibration curves . . . 157
12.3 Explained variation and R^2 measures . . . 159
12.4 Discrimination, sensitivity and specificity, and ROC curves . . . 160
12.4.1 Sensitivity and specificity . . . 160
12.4.2 ROC curves . . . 160
12.4.3 Area under the ROC curve . . . 161
12.4.4 Overfitting, cross-validation, and external validation . . . 161
12.5 Predictiveness curves . . . 162
12.6 Summary . . . 163
12.7 Practical 12 . . . 164

13 Matched studies and their analysis **167**
13.1 Aims . . . 167
13.2 Definition and examples . . . 167
13.3 Matching in case-control studies . . . 168
13.4 Rationale for Matching . . . 169
13.5 Contrasting the effects of matching in case-control and other matched designs 170
13.6 'Reversed' analysis of matched case-control studies . . . 172
13.7 Matched studies estimate conditional effects . . . 172
13.8 Matched studies analysed with statistical models with a continuous dependent variable . . . 172
13.8.1 Introduction and notation . . . 172
13.8.2 Simple approaches for the analysis of matched pair studies . . . 172
13.8.3 Analysis using regression models that explicitly acknowledge the matching . . . 173
13.8.4 Analysis using regression models that adjust for the matching variables . . . 173
13.9 Statistical analysis of matched studies with a binary outcome and binary exposure . . . 174
13.9.1 Introduction . . . 174
13.9.2 Tabulation of results . . . 175
13.9.3 McNemar's Test . . . 175
13.9.4 Odds ratios in matched studies with binary outcomes and binary exposures . . . 176
13.9.5 Confidence Intervals for Odds Ratios from matched Studies . . . 177
13.9.6 General analysis of matched binary outcome and binary exposure studies in Stata . . . 179
13.10 Conditional versus marginal odds ratios . . . 180
13.11 Practical 13 . . . 183

14 Conditional Logistic Regression **187**
14.1 Aims . . . 187
14.2 Logistic models for matched studies . . . 187
14.2.1 Preliminaries . . . 187
14.2.2 Matched case-control study . . . 188
14.2.3 Conditional logistic regression for other matched studies . . . 189
14.3 Conditional logistic regression: binary exposure . . . 189
14.3.1 Note on sufficient statistics . . . 190

14.3.2 Development of the conditional logistic regression model 190
14.3.3 The conditional likelihood 191
14.3.4 Extensions 192
14.4 Conditional logistic regression: general situation 193
14.5 Conditional logistic regression: interactions between exposures and matching variables 194
14.6 Conditional logistic regression: implementation 194
14.6.1 Examples 194
14.7 Non-regularity of the matched case-control model 196
14.8 Alternative analyses for matched case-control data 197
14.9 Practical 14 199

15 Multinomial Logistic Regression **203**
15.1 Aims and Objectives 203
15.2 Nominal outcomes and the multinomial distribution 203
15.3 Example - consumption of alcohol 203
15.4 The multinomial logistic regression model 204
15.4.1 Estimation 205
15.5 Gender and alcohol consumption 205
15.6 Age and alcohol consumption 207
15.7 Model comparison and postestimation 209
15.8 Summary 209
15.9 Practical 15 210

16 Ordinal Logistic Regression **213**
16.1 Aims and objectives 213
16.2 Motivation 213
16.3 The ordinal logistic model 213
16.4 Example: gender and alcohol consumption 214
16.5 An alternative formulation of ordinal (and standard) logistic regression . . 215
16.6 Model comparison and goodness of fit 216
16.7 Summary 218
16.8 Practical 16 219

Session 1

Introduction to Generalized Linear Models and how to fit them in Stata

Linear regression is a powerful tool for developing models for continuous outcomes. The family of generalized linear models (GLMs) is a larger class of models which enables us to develop and fit models for a much wider range of outcome types, including binary, count and categorical outcomes.

1.1 Aims and objectives

This session will begin by revising the linear regression model, focussing on the interpretation of regression coefficients (including those from models that include interaction terms). One new technique, the lowess smoother, will also be introduced.

A general family of distributions known as the *exponential family* will then be defined as will *generalized linear models* (GLMs), the class of models than can be used for the analysis of outcome variables from this family.

At the end of this session you should understand that the normal, binomial and Poisson distributions belong to the exponential family, and the key components of a generalized linear model.

You should also have learnt the basic commands to fit GLMs in Stata.

1.2 Linear Regression (revision)

1.2.1 Simple linear regression

Suppose that two measurements (y_i, x_i) are made on each of n individuals. A simple linear regression model of y on x can be written:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where the ϵ_i are independent and identically distributed with

$$\epsilon_i \sim N(0, \sigma^2).$$

Alternatively we can say that the random variable Y_i has the conditional distribution

$$Y_i \mid x_i \sim N(\alpha + \beta x_i, \sigma^2).$$

Example: Morphine and mental activity trial

As an example consider a randomized trial carried out to compare the effect of an injection of morphine against placebo on mental activity. 24 subjects were randomized to each group, and mental activity was assessed by the sum of scores on 7 items in a questionnaire, with a high score indicating high mental activity. The following are the mental activity scores recorded 2 hours after injection of the drug.

Placebo						Morphine					
7	1	10	10	6	2	4	2	14	0	2	0
7	5	6	6	3	8	6	0	1	6	5	3
0	11	10	0	8	6	0	10	0	0	1	2
7	1	8	5	1	5	9	0	2	7	2	12

The following output from Stata was obtained by fitting a model to the data to investigate the effect of treatment (with `treat` taking the value 1 in the placebo group and 2 in the morphine group in the output below) on mental activity.

```
. regress mentact i.treat

  Source |       SS       df       MS                  Number of obs =      48
---------+------------------------------               F(  1,    46) =    2.99
   Model |     42.1875     1     42.1875               Prob > F      =  0.0905
Residual |  649.291667    46  14.1150362               R-squared     =  0.0610
---------+------------------------------               Adj R-squared =  0.0406
   Total |  691.479167    47  14.7123227               Root MSE      =   3.757

------------------------------------------------------------------------------
 mentact |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 2.treat |      -1.875   1.084552     -1.729   0.091       -4.05809    .3080899
   _cons |   5.541667   .7668941      7.226   0.000       3.997989    7.085344
------------------------------------------------------------------------------
```

EXERCISE 1.1 *Simple linear regression*

1. Provide an interpretation of both estimated regression coefficients.
2. What is the estimated residual variance?

1.2.2 Multiple linear regression

Suppose now that we have two predictor variables (z_i as well as x_i). A multiple linear regression model relating y to x and z can be written:

$$y_i = \alpha + \beta x_i + \gamma z_i + \epsilon_i.$$

β here represents the effect of a 1 unit increase in x on y holding z **constant**. The model implies that this effect is the same whatever this constant level of z is. This type of model

estimates the effect of x controlling (or adjusting) for z. In epidemiological applications we might say that we are controlling for the confounding effect of z.

We can extend this model to include an interaction between x and z as follows:

$$y_i = \alpha + \beta x_i + \gamma z_i + \delta x_i z_i + \epsilon_i.$$

It is important to appreciate that adding the interaction term changes the interpretation of β. It is now the effect of a 1 unit increase in x on y holding z **constant at zero**. If z is held constant at some other value, say k, then the effect of a 1 unit increase in x on y is $\beta + k\delta$.

Example: Morphine and mental activity trial (continued)

In the morphine and mental activity trial considered above, mental activity before treatment (`prement` in the following Stata output) was also measured on each patient. Below is the Stata output from a model that includes an interaction term.

```
. regress mentact i.treat##c.prement

  Source |       SS       df       MS                  Number of obs =      48
---------+------------------------------                F(  3,    44) =    8.81
   Model |  259.482345     3   86.494115                Prob > F      =  0.0001
Residual |  431.996822    44  9.81810958                R-squared     =  0.3753
---------+------------------------------                Adj R-squared =  0.3327
   Total |  691.479167    47  14.7123227                Root MSE      =  3.1334

------------------------------------------------------------------------------
        mentact |      Coef.  Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
        2.treat |  -1.211742   1.64558    -0.736   0.465    -4.528191    2.104707
        prement |   .5939394  .1724872     3.443   0.001     .2463142    .9415646
                |
treat#c.prement |
              2 |  -.0895258  .2334819    -0.383   0.703    -.5600776     .381026
                |
          _cons |    1.97803  1.216616     1.626   0.111     -.473898    4.429959
------------------------------------------------------------------------------
```

EXERCISE 1.2 *Multiple linear regression*

1. Provide an interpretation of each estimated regression coefficient.
2. Is R^2 guaranteed to be larger here than in the earlier model?
3. Is the estimated residual variance guaranteed to be smaller here than in the earlier model?

Note that strictly the `'i.'` in the command used above is not needed. It would have been sufficient to use the following command.

```
. regress mentact treat##c.prement
```

Stata's default is to consider variables included in interaction terms as categorical. This is potentially a source of confusion since the `'i.'` is needed to indicate that a variable is categorical when not included as part of an interaction. For this reason `'i.'` is retained when Stata interaction terms involving categorical variables are used in this module.

1.2.3 Inference for linear regression

Inference for regression models can be based on the likelihood function corresponding to the conditional distribution of the outcomes given the covariates. For example, the likelihood function for the simple linear regression model can be written:

$$L(\alpha, \beta, \sigma^2 \mid \mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right\}.$$

It follows that the log-likelihood function is:

$$\ell(\alpha, \beta, \sigma^2 \mid \mathbf{y}, \mathbf{x}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

The MLE's of α and β can be obtained from the **score equations**:

$$U(\alpha) = \ell'(\alpha) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)$$

$$U(\beta) = \ell'(\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i)$$

$$U(\hat{\alpha}) = 0 \Rightarrow \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$U(\hat{\beta}) = 0 \Rightarrow \hat{\beta} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Note that we do not need to know σ^2 to solve these. These estimators are identical to the **least squares** estimators derived in the Regression course by minimising the residual sum of squares. Hence in this case, maximum likelihood is **equivalent** to least squares.

The MLE of σ^2 can also be obtained from the **score equations**:

$$U(\sigma^2) = \ell'(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

$$U(\hat{\sigma^2}) = 0 \Rightarrow \hat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n}$$

This MLE differs from the usual unbiased estimate by the factor $\frac{n-2}{n}$. Hence the MLE is only asymptotically unbiased and so it is customary to use the following to estimate the residual variance in a linear regression model with p parameters:

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_{i1} - \hat{\beta}_2 x_{i2}....)^2}{n - p}$$

A corollary of the need to estimate the residual variance in linear regression is that test statistics follow F distributions, rather than $\chi 2$ distributions.

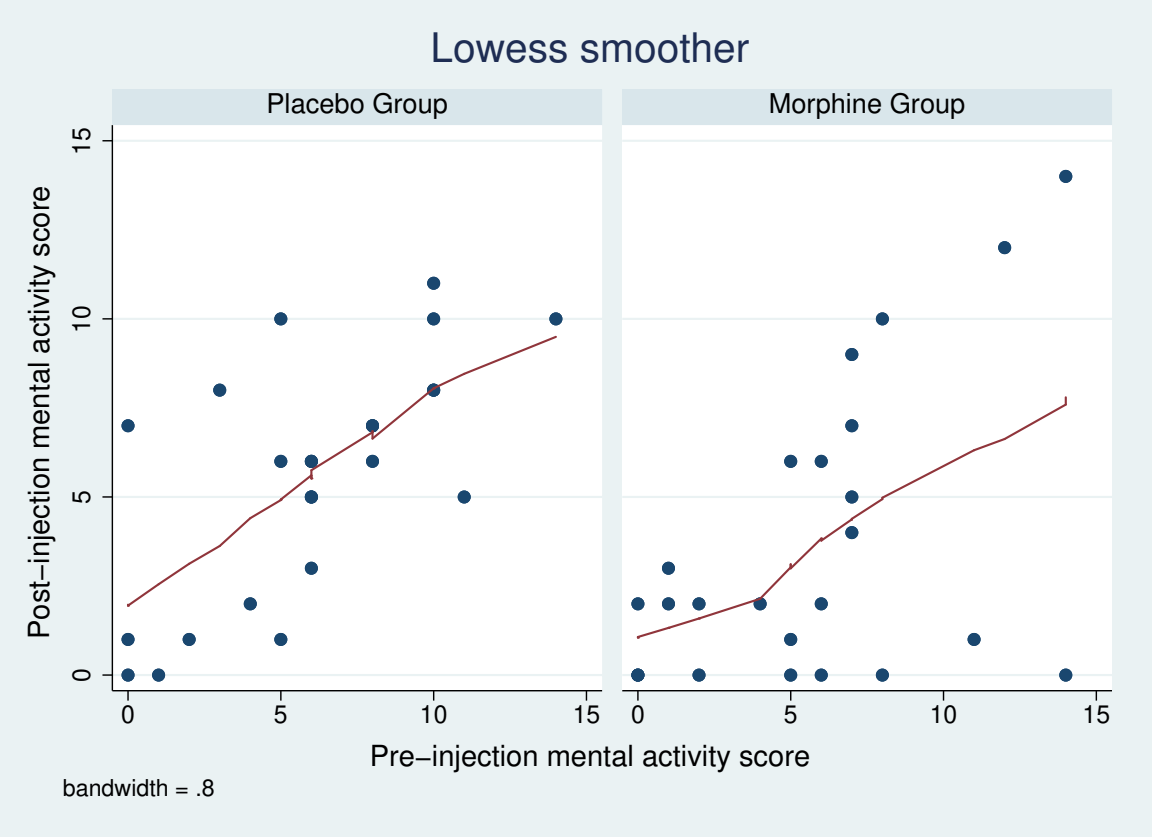


Figure 1.1: Lowess smoothers, with bandwidth set to 0.8, for the mental activity trial data.

1.2.4 Lowess smoothing - Stata's `lowess` *command*

In the linear regression course a number of techniques for assessing whether a particular model fits the data well were introduced. These included methods of identifying outlying and influential points and graphical techniques, such as quantile-quantile plots of residuals and plots of residuals against fitted values.

Visualising the data, and the judging the fits of models from residual plots, can be tricky for some GLMs because of the discrete nature of the outcome variable. One useful technique is to display a so-called lowess smoother along with the data. Here, an introduction to their use is provided by illustrating their application to continuous data.

See the Stata manual for full details, but in brief a lowess smoother provides locally weighted scatterplot smoothing. The smoothed values are obtained by fitting a separate regression model to the data points in the vicinity of each value taken by the predictor variable, with each regression point being weighted such that points with predictor variable values close to the value in question are given more weight than those which are not close. The amount of smoothing is affected by a user-chosen constant termed the *bandwidth*. The bandwidth relates to the proportion of the data that is used in each local regression model: the smaller the bandwidth the more closely the fitted line will follow the data. Stata's default bandwidth is 0.8, but you are encouraged to experiment with different values.

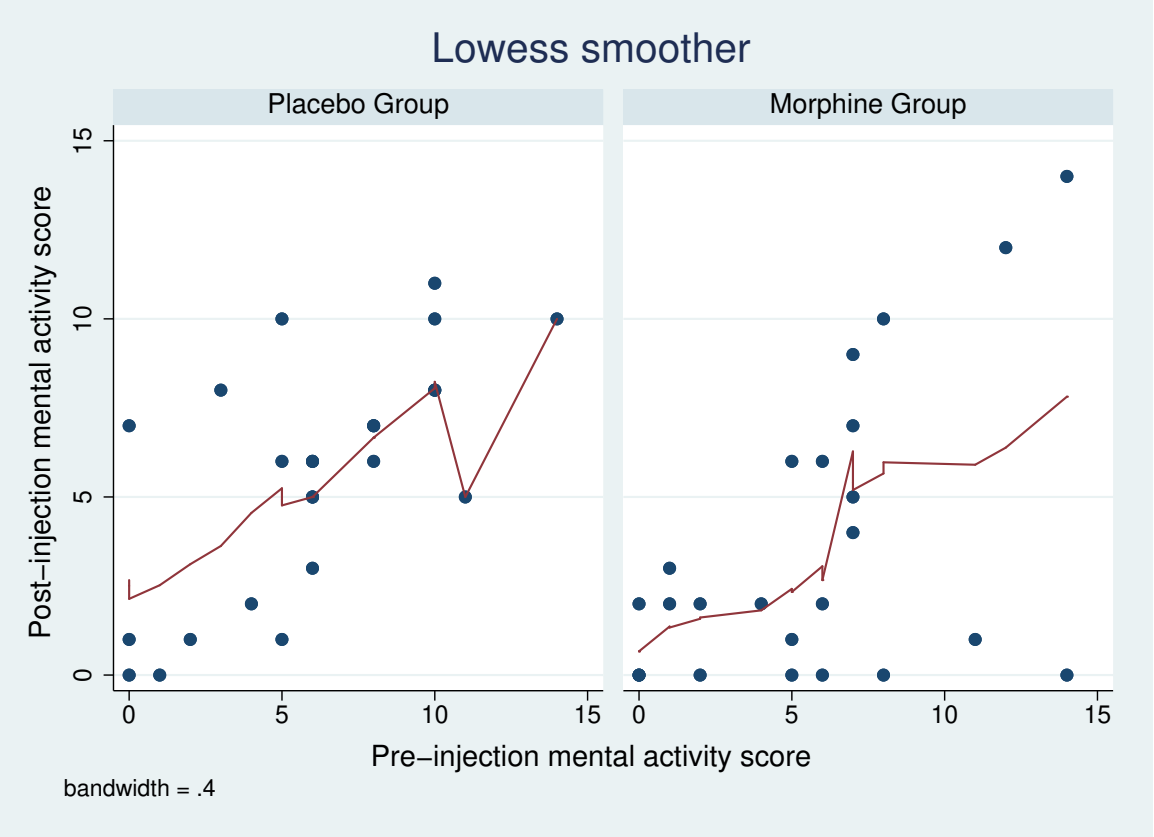


Figure 1.2: Lowess smoothers, with bandwidth set to 0.4, for the mental activity trial data.

Figures 1.1 and 1.2 show scatter plots of the post- and pre-injection mental activity data in the two groups in the mental activity trial along with lowess smoothers for bandwidths of 0.8 and 0.4 respectively. Figure 1.1 suggests that assuming a linear relationship between the two scores is not too unreasonable. In Figure 1.2 it seems that the low bandwidth has resulted in the fitted lines following the observed data rather more closely than is desirable here.

1.3 The Exponential family of distributions

The density of a $N(\mu, \sigma^2)$ random variable Y can be written

$$f(y) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

and the logarithm of this is

$$\ln\{f(y)\} = -\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2).$$

With some reparameterisation and rearranging this takes on a very special form:

$$\ln\{f(y)\} = \frac{y\theta - b(\theta)}{\phi} - c(y,\phi). \tag{1.1}$$

where

$$\begin{aligned}\theta &= \mu \\ \phi &= \sigma^2 \\ b(\theta) &= \frac{\mu^2}{2} \\ c(y,\phi) &= \frac{y^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2).\end{aligned}$$

Distributions that can be written in the form of (1.1) are said to belong to the **exponential family.**

1.3.1 Two other important examples

1. *The Poisson distribution.*

$$\begin{aligned}f(y) = \Pr(Y=y) &= \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0,1,2,\ldots \\ \ln\{f(y)\} &= y\ln(\mu) - \mu - \ln(y!)\end{aligned}$$

$$\boxed{\theta = \ln(\mu), \quad \phi = 1, \quad b(\theta) = \mu, \quad c(y,\phi) = \ln(y!)}$$

2. *The binomial distribution.*

$$\begin{aligned}f(y) = \Pr(Y=y) &= \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0,1,\ldots,n. \\ \ln\{f(y)\} &= y\ln(\pi) + (n-y)\ln(1-\pi) + \ln\left\{\binom{n}{y}\right\} \\ &= y\ln\left(\frac{\pi}{1-\pi}\right) + n\ln(1-\pi) + \ln\left\{\binom{n}{y}\right\}.\end{aligned}$$

$$\boxed{\theta = \ln\left(\tfrac{\pi}{1-\pi}\right), \quad \phi = 1, \quad b(\theta) = -n\ln(1-\pi), \quad c(y,\phi) = -\ln\left\{\binom{n}{y}\right\}}$$

1.3.2 Notes

1. Suppose that we have a random variable Y that follows a distribution in the exponential family, and that we have a single observation from this distribution y. The following, derived from likelihood theory in text books such as that by McCullagh and Nelder, are properties that apply for distributions in the exponential family.

   (a) $E(Y) = b'(\theta)$.

   (b) $\text{Var}(Y) = \phi b''(\theta)$.

   (c) $\hat{\theta}$ solves $b'(\theta) = y$.

   (d) Asymptotically $\text{Var}(\hat{\theta}) = \frac{\phi}{b''(\theta)}$.

2. θ is called the **canonical** or **natural** parameter. The associated function (log for Poisson, logit for binomial) is called the **canonical link function**. We could use other link functions (e.g. probit for binomial data); however sufficient statistics (important for conditional inference) only exist with the canonical link.

3. The parameter ϕ is known as the **scale** or **dispersion** factor. For the Poisson and binomial distributions the scale factor is 1. For the normal distribution the scale parameter is σ^2 and usually needs to be estimated from the data. Whether or not the scale factor has to be estimated from the data has consequences for statistical procedures such as hypothesis tests.

EXERCISE 1.3 *Properties of distributions in the exponential family*

Suppose that we have a random variable Y that follows a distribution in the exponential family, and that we have a single observation from this distribution y. Use the formulae above to derive expressions for each of the following if the distribution is i) normal, ii) Poisson and iii) binomial.

1. $E(Y)$.
2. $\text{Var}(Y)$.
3. The maximum likelihood estimate of θ.
4. The asymptotic variance of $\hat{\theta}$.

1.4 Definition of a Generalized Linear Model (GLM)

Generalized linear modelling is a unified approach to statistical modelling within the exponential family, and so brings a wide range of statistical models 'under one roof'.

A GLM has three components.

1. *Response Distribution:* The response variables $Y_i, i = 1, \ldots, n$ are assumed to be independent, arising from an exponential family distribution, with $\text{E}(Y_i) = \mu_i$.

2. *Linear Predictor:* The explanatory variables $(x_1, \ldots, x_n)$ enter the model in a linear combination with unknown parameters: for the ith response we have the linear predictor:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

3. *Link Function:* The link function relates the linear predictor η_i to the mean μ_i:

$$g(\mu_i) = \eta_i.$$

1.4.1 Example 1: Simple Linear Regression

1. *Response Distribution:* $Y_i, i = 1, \ldots, n$ are assumed to be independent, arising from a normal distribution, with mean μ_i and variance σ^2.
2. *Linear Predictor:* There is only a single explanatory variable x:

$$\eta_i = \beta_0 + \beta_1 x_i.$$

3. *Link Function:* The link function is the **identity** function:

$$g(\mu_i) = \mu_i = \eta_i.$$

1.4.2 Example 2: Poisson Regression with a single explanatory variable

1. *Response Distribution:* $Y_i, i = 1, \ldots, n$ are assumed to be independent, arising from a Poisson distribution, with mean μ_i.
2. *Linear Predictor:* There is only a single explanatory variable x:

$$\eta_i = \beta_0 + \beta_1 x_i.$$

3. *Link Function:* The link function is the **logarithm**:

$$g(\mu_i) = \log(\mu_i) = \eta_i.$$

This model is covered in much more detail in session 8.

1.4.3 Example 3: Logistic regression for a Bernouilli outcome with a single explanatory variable

1. *Response Distribution:* $Y_i, i = 1, \ldots, n$ are assumed to be independent, arising from a Bernouilli distribution, with probability of success (i.e. mean) μ_i.
2. *Linear Predictor:* There is only a single explanatory variable x:

$$\eta_i = \beta_0 + \beta_1 x_i.$$

3. *Link Function:* The link function is the **logit** $(\log(\mu_i/(1-\mu_i)))$ function:

$$g(\mu_i) = \log(\mu_i/(1-\mu_i)) = \eta_i.$$

This model is studied in detail in a number of sessions, starting with session 3. Note that μ_i is customarily replaced by π_i when specifying this model.

1.4.4 Notes

1. The linear predictor is linear in the parameters, not necessarily in the explanatory variables. For example, in polynomial regression, powers of the variables occur:

$$\eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p.$$

   On the other hand

$$\eta_i = \beta_0(1 - e^{\beta_1 x_{i1}})$$

   is **not** a linear predictor.

2. With minor exceptions the underlying theory (parameter estimation, inference, model assessment and comparison) proceeds in the same way for all GLMs; it is just the distribution and link that differ.

3. The link function is chosen to provide a suitable scale for the effects of explanatory variables to operate in a linear manner. Frequently the range of μ will be transformed to the whole real line (-∞ to ∞).

4. The 'log likelihood of the model' is usually shorthand for 'the log likelihood function of the model for the given data evaluated at the MLEs of the parameters,' i.e. the maximum of the log likelihood function.

5. From now on the terms 'model', 'generalized linear model' and 'GLM' will be used interchangeably.

1.5 Fitting GLMs

Throughout this module, we will use the method of maximum likelihood to fit GLMs. This involves finding the values of the model parameters which maximize the likelihood function. Whereas for linear regression models the MLEs can be expressed in a closed form equation, for GLMs more generally the MLEs must be found by iterative methods.

As we shall see, outside of the normal linear regression model, where exact distributional results are available, our inferences (hypothesis tests and confidence intervals) will usually rely on asymptotic results.

1.5.1 Stata's `glm` command

The `glm` command in Stata can be used to fit any generalized linear model. The basic form of the command is

```
glm <response variable> <explanatory variables to form linear
predictor>, family (<name of distribution>) link(<link function>).
```

As an example, consider the interaction model fitted above to the data from the mental activity and morphine trial using the `regress` command. The same model can also be fitted using the `glm` command as follows.

```
. glm mentact i.treat##c.prement, family(gaussian) link(identity)

Iteration 0:   log likelihood = -120.84226

Generalized linear models                         No. of obs      =         48
Optimization     : ML                             Residual df     =         44
                                                  Scale parameter =    9.81811
Deviance         =  431.9968216                   (1/df) Deviance =    9.81811
Pearson          =  431.9968216                   (1/df) Pearson  =    9.81811

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   5.201761
Log likelihood   = -120.8422629                   BIC             =    261.664

------------------------------------------------------------------------------
             |                 OIM
    mentact |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    2.treat |  -1.211742   1.64558    -0.74   0.462    -4.43702    2.013536
    prement |   .5939394   .1724872    3.44   0.001     .2558706    .9320082
            |
treat#c.prement |
          2  |  -.0895258   .2334819   -0.38   0.701    -.5471419    .3680903
            |
      _cons |    1.97803   1.216616    1.63   0.104    -.4064932    4.362554
------------------------------------------------------------------------------
```

Notice that, although parameter estimates and standard errors are identical to those obtained with `regress`, p-values are not. This is because the `glm` command approximates the exact t distribution by the (asymptotically valid) z distribution in calculating the significance of the test statistic. In practice this means that `regress` rather than `glm` should always be used to fit linear models in Stata.

For two other special cases, Stata has specific commands.

- **Logistic regression (binary outcome)**: Distribution binomial, link logit.
  Stata: `logistic` or `logit`.
- **Poisson regression (rates and counts)**: Distribution Poisson, link log.
  Stata: `Poisson`.

These commands will be considered in greater detail in later sessions.

1.6 Practical 1

Dataset required: `mental.dta`

Introduction

The purpose of today's session is to perform a full analysis of the mental activity data trial introduced earlier We will use a series of normal error models, and compare their respective fits using F-tests.

The data are from a three-arm placebo-controlled randomised trial assessing the effects of injection with morphine and heroin on mental activity. Seventy two people were randomized to one of three arms (placebo, morphine, heroin), and mental activity was assessed by the sum of scores on seven items of a questionnaire, with a high score indicating high mental activity.

The dataset contains three variables:

i **treat** - Randomised arm: 1=placebo; 2=morphine; 3=heroin

ii **prement** - Mental activity score before injection

iii **mentact** - Mental activity score 2 hours after injection with drug.

The questions that the analysis should address are as follows.

Aims

1 Do the mean post-injection mental activity scores differ among groups?

2 What effect does adjustment for pre-injection mental activity score have on your conclusions?

3 Are the effects of the drugs different for subjects with different levels of pre-injection mental activity?

Analysis

1 Read the data into Stata and examine. Use tables, histograms and plots to look at the distributions of the pre- and post-injection scores.

Discuss: Discuss with your colleagues (in your Breakout room if online) what can be concluded about the appropriateness of linear regression models for post-injection scores on the basis of these preliminary explorations.

2 Use Stata's `lowess` command to plot the post-injection scores against the pre-injection scores with a lowess smoother, separately in each of the three groups. Explore what happens if you change the bandwidth. The basic form of the command is along the lines of the following:

```
. lowess mentact prement if treat==1, bw(0.8)
```

3 Write down in algebraic form the linear predictors that correspond to each of the models described in Table 1 below. Carefully define each term that you use in the various models.

Note that the distributional assumption in each model is $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ where Y_i is the post-injection mental activity score. In each model the link function is the identity $\eta_i = \mu_i$.

Table 1

#	Terms fitted	Linear predictor
1	Overall mean	
2	Drugs	
3	Pre-inj (Pre)	
4	Drugs + Pre	
5	Drugs + Pre + (Drugs-by-Pre interaction)	

4 Fit each linear regression model in Stata, using the `regress` command, to complete Table 2.

Table 2

#	Terms fitted	RSS	Residual df
1	Overall mean		
2	Drugs		
3	Pre-inj (Pre)		
4	Drugs + Pre		
5	Drugs + Pre + (Drugs-by-Pre interaction)		

5 Use Stata to calculate fitted values for model 5. Display these for each trial arm in a graph. Compute the intercepts and slopes of the three fitted lines from the parameters estimated in the model.

Table 3

Trial arm	Fitted Intercept	Fitted Slope
Placebo		
Morphine		
Heroin		

6 Pen & paper exercise: use the results in Table 2 to carry out formal comparisons of

(a) the fits of models 3 and 4

(b) the fits of models 4 and 5

Note: it is good practice to use a calculator for these calculations, rather than your phone or computer software such as Excel or Stata.

Confirm your results using Stata's `test` command.

Discuss: What do you conclude from these tests? Why can an analogous test not be used to compare the fits of models 2 and 3?

7 Use the `glm` command to fit model 4. In what way do the results differ from those with `regress`?

8 Use results from models 2 and 4 to complete Table 4.

Table 4

	Mean	Mean difference from placebo	SE	Adjusted difference from placebo	SE
Placebo		0	-	0	-
Morphine					
Heroin					

9 Briefly summarise your conclusions regarding the effects of morphine and heroin on mental activity scores in this trial.

Working together with one or more of your colleagues (in your Breakout Room if online), write a paragraph to summarise your conclusions from these analyses. You should try to answer each of the aims listed above. If online, one of you should post your group's paragraph in the Zoom chat.

10 If you have time, explore how to obtain an estimate and confidence interval for the difference in mean post-treatment scores between the morphine and heroin groups using model 4.

There are two ways to do this:

(a) use the post-estimation command `lincom`, or

(b) re-parameterise the model using the "bx." syntax to replace "i.", where x is the number of the group you wish to make the baseline.

Session 2

Types of Investigation

2.1 Aims

At the end of the session you should be able to:

- Describe three different types of investigations that arise in medical statistics and health data science.
- Link a research question to an investigation type and compare the properties of different investigation types.
- Explain how and why explanatory variables are used differently in prediction studies and in causal investigations

2.2 Specifying research questions

Specifying the research question or questions is a crucial starting point for an investigation. In some cases the research question will be highly specific, and in others could be more wide ranging with several components. The research question then informs the subsequent stages of the investigation, ranging from choice of study population; study design; data collection; monitoring and quality control; data analysis; presentation of conclusions; interpretation. Figure 1 illustrates one way of representing the whole process of an investigation.

The statistician/data scientist plays an important role at all stages of an investigation, not just at the data analysis phase. It is perhaps most usual for collaborators who are subject-matter experts (e.g. clinicians) to pose the initial research question. However, the statistician very often plays a key part in refining these initial ideas in order to translate them into something formal and clearly specified.

2.3 Different types of investigation

2.3.1 Classification

The research question informs what type of investigation is required. Investigations can be divided broadly into the following types:

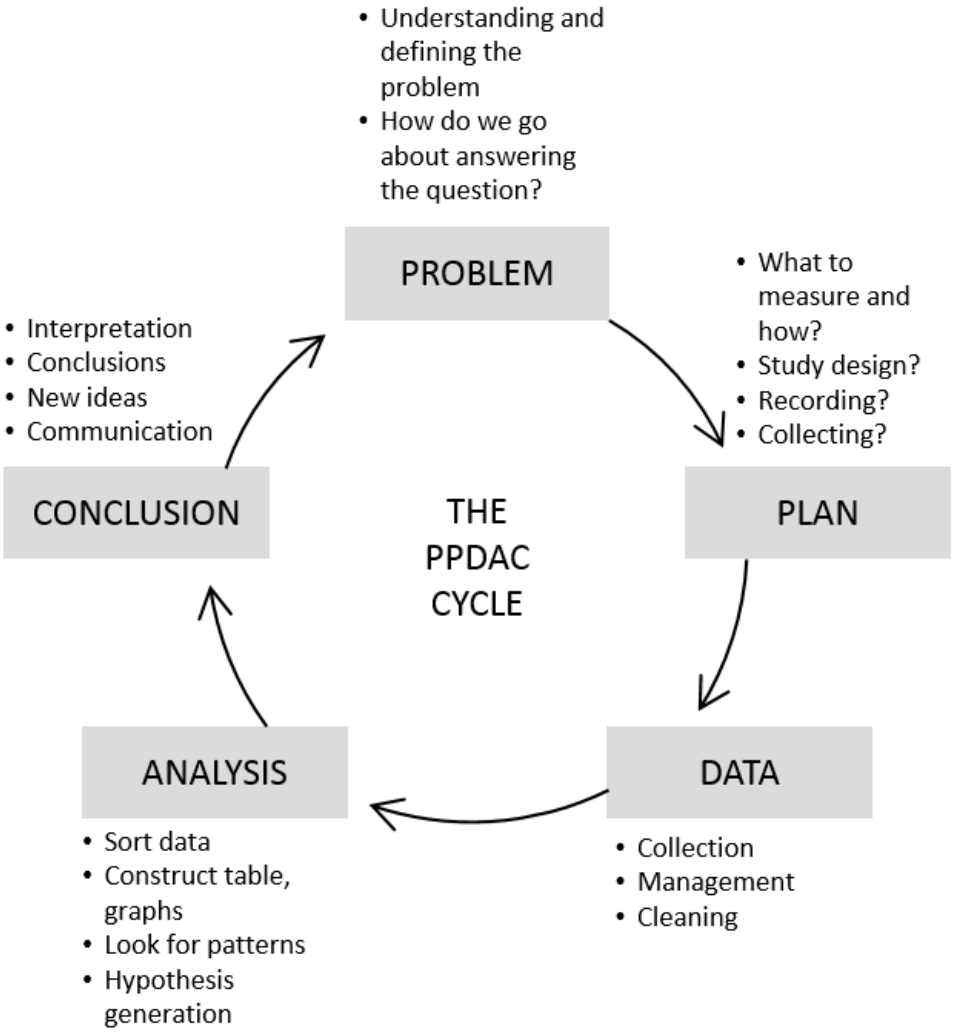
I Description

II Prediction

III Causality and explanation

Hernán, Hsu & Healy (Chance, 2019) set out to classify data science tasks and used three classifications: *Description*, *Prediction*, and *Counterfactual prediction* (meaning causality). Schmueli (Statistical Science, 2010) also described similar classifications: *Descriptive*

Figure 2.1: The PPDAC problem-solving cycle: going from Problem, Plan, Data, Analysis, to Conclusion and communication. Taken from Spiegelhalter 2019.



modelling, *Predictive modelling*, and *Explanatory modelling*. See also Hand (Harvard Data Science Review, 2019) for a nice discussion on this topic.

2.3.2 Implications of investigation type

The distinction between the different types of investigation is crucial because it has a fundamental impact on the steps of the analysis and beyond. For example, the investigation type influences:

- How we decide what variables are to be included in the analysis
- What analysis methods to use
- How we assess the fit/performance of the model or oter analysis approach used
- How we present the results from the analysis
- How the findings might be used in practice
- How we need to work with other experts at different stages

2.3.3 The role of study design

The different types of investigation may be performed using data from studies of different design. Having posed a research question, we can consider (with input from collaborators) what data are required to answer it robustly, including whether new data collection is needed, or whether there are existing data that could be used to address the question. This process needs to take into account considerations of cost, timeliness, feasibility and ethics. For example, for some questions our ideal study could be a randomized controlled trial, but to perform one would require such long follow-up that it would be infeasible and unethical, and so we would turn to observational data to address the research question. There is a major emphasis in the recent biostatistical and epidemiological literature on the use of 'found' data from sources such as electronic health records, which present great opportunities to answer research questions using data on a large number of individuals, but also present challenges for analysis and interpretation. All three types of investigation may make use of observational data. Randomized controlled trials are designed to estimate treatment effects (i.e. for causal investigations), but secondary analyses of trial data can be used in other types of investigation, such as to develop a prediction model.

2.4 Properties of different types of investigation

2.4.1 Description

In a descriptive investigation the data are used to provide a quantitative summary of features of the population of interest, or in other words the data are summarised in a compact way.

Simple descriptive analyses involve calculating proportions of individuals with a particular characteristic (e.g. males and females; smokers and non-smokers), or estimating features of the distribution of continuous variables (e.g. mean and variance of weight or blood pressure). The resulting information is then presented using tables and data visualisation.

Some descriptive analyses may extend to use of more complex methods of analysis. For example, the research question may concern how individuals within a population cluster together in terms of their dietary habits, requiring clustering methods. It may be of interest to estimate the expected survival time post-disease diagnosis in the presence of censored survival times, which would require survival analysis techniques.

All investigations should start with some basic descriptive analysis to gain understanding of the features of the data at hand. It is at this stage that we can uncover challenges such as missing data, gain insights into how certain variables are distributed, and, where relevant, gain understanding of correlations between key variables, including to identify collinearities. Some investigations then go on to the main research question, which goes beyond description, and others may be entirely descriptive and not proceed onto other questions.

Huebner et al. (2019) provide useful guidance on 'initial data analysis'. See also Spiegelhalter (2019) for an accessible discussion of summarising and communicating descriptions of data.

2.4.2 Prediction

Prediction is about using data on some features of individuals to predict other features with the aim of predicting the outcome for new or future observations. More formally, prediction is concerned with mapping data on variables $X_1, X_2, \ldots, X_p$ to an outcome Y. The prediction model could be developed using statistical models such as regression, or approaches that would be described as machine learning algorithms.

Results from prediction investigations are used for a range of purposes: to inform people of their risk or prognosis; to identify people at high risk of an adverse event and hence take action such as more frequent screening (though the investigation will not tell us whether such screening would be effective).

Prediction models are typically developed using observational data. A well known example is the Framingham Risk Score, which provides predictions of a person's 10-year of developing coronary heart disease (D'Agostino et al 2008).

There is a huge literature on prediction in the medical setting. See for example the books by Riley et al. (2019) and Steyerberg (2019).

2.4.3 Causality and explanation

In causal investigations we seek to understand the causal effect of one or more variables on an outcome. Hernán et al. (2019) describe this as "Using data to predict certain features of the world as if the world had been different". For a simple example of a causal investigation, consider a continuous outcome Y (e.g. blood pressure) and a binary treatment variable X, where $X = 1$ denotes treated and $X = 0$ denotes untreated. A causal investigation asks how the mean of Y would be different if all individuals had $X = 1$ compared with if all individuals had $X = 0$. In other words, if we could change X what would be the expected change in Y?

Questions such as this can be arguably simple to answer using a randomized controlled trial, where there is no confounding of the treatment-outcome association. However, issues of drop-out and non-compliance are important to consider. Historically, some have considered answering causal questions to lie only in the domain of randomized experiments. However, randomized experiments are not feasible or ethical to address many important questions. It is now recognised that causality is often the goal of investigations using observational data. See for example the paper of Hernán (2018), who wrote "being explicit about the causal objective of a study reduces ambiguity in the scientific question, errors in the data analysis, and excesses in the interpretation of the results". The field of 'causal inference' has developed in recent decades, with particular advances in recent years, to enable this.

Schmeuli (2010) equates causality with 'explanation', meaning explanation of mechanisms of how one (or more) variable affects another. However, Hernán et al. (2019) make the point that we may be able to say that X causes Y without understanding the underlying mechanism. For example we may find strong evidence from a trial that a drug is effective for a given outcome, but the precise biological mechanisms through which the effect is transmitted are not well understood.

The variable of interest in a causal investigation could be use of a medical treatment (a drug) or application of a procedure. More generally it could be an 'exposure' such as

'smoking' or 'exercising for at least 30 minutes per day'. The 'hypothetical intervention' of interest should be (reasonably) well defined, even if we could never in reality intervene on it in the real world (e.g. it would be impractical, not to say unethical, to intervene on smoking status). See Hernán (2016) for a discussion of related issues.

2.4.4 Is there a fourth investigation type?

There is arguably a fourth investigation type which is concerned with exploring how several explanatory variables $X_1, \ldots, X_p$ are associated with an outcome Y. This might be described as an "exploration of risk factors" investigation. It may involve univariable analyses, looking at the association of each explanatory variable ("risk factor") individually with the outcome, and multivariable analyses which look at association of several variables with the outcome in a single model. These types of analysis are typically carried out using observational data, and many (or perhaps most) epidemiological studies are investigations of this type, at least historically.

These types of investigation can be useful for understanding associations between variables in the population of interest and, as such, some may consider these analyses to be descriptive. However, as we all know, association is not causation! These types of investigation often do not consider the relative temporal ordering of explanatory variables, which means that interpretation of estimated associations as causal effects can be misleading. There is recent emphasis in the epidemiological literature on more principled investigations which are more explicit about the aim of the investigation.

Like in a prediction investigation, the interest is in several explanatory variables. However, unlike in a prediction investigation, the aim is to actually explore quantitatively the unconditional and conditional associations of the explanatory variables with Y, rather than being purely on predicting Y. Unlike in a causal investigation, there is not a particular focus on a single variable. However, there is often an attempt to discuss the associations as though they may be causal even though an explicit causal question has not been posed.

Investigators should be wary of over-interpreting findings from "exploration of risk factors" investigations. And if we are really interested in addressing a causal question we should be explicit about that and carry out our analysis and interpretations accordingly.

2.5 An example

Table 1 provides an example of the features of different investigation types. The overall topic is stroke in women. The table (taken from Hernán et al. 2019) provides an example research question, the features of data that would be required to answer it, and the types of analysis that could be used for investigations of three types: Description, Prediction and Causal inference.

2.6 Role of explanatory variables in different types of investigation

The role of explanatory variables in different types of investigation differs. We focus here on prediction investigations and causal investigations.

Table 2.1: From Hernán, Hsu & Healy 2019. Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

	Description	Prediction	Causal inference
Example of scientific question	How can women aged 60-80 years with stroke history be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	- Eligibility criteria - Features (symptoms, clinical parameters ...)	- Eligibility criteria - Output (diagnosis of stroke over the next year) - Inputs (age, blood pressure, history of stroke, diabetes at baseline)	- Eligibility criteria - Outcome (diagnosis of stroke over the next year) -Treatment (initiation of statins at baseline) - Confounders - Effect modifiers (optional)
Examples of analytics	Cluster analysis	Regression Decision trees Random forests Support vector machines Neural networks	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation

2.6.1 Prediction

In prediction investigations the aim is to use $X_1, \ldots, X_p$ to predict Y. In this setting the $X_1, \ldots, X_p$ are often referred to as the 'predictors' for obvious reasons. For a prediction problem we may well use all of the explanatory variables $X_1, \ldots, X_p$ in the prediction model or algorithm. Crucially, in prediction we are not interested in the inter-relationships between the explanatory variables $X_1, \ldots, X_p$ and their temporal ordering. The only aim is to achieve a good prediction of the outcome Y. It may be desirable to reduce the number of explanatory variables, particularly in settings where the number of potential predictors p is very large. Various principled procedures are available for reducing the number of predictor variables.

2.6.2 Causality and explanation

In investigations of causality, one of the explanatory variables is designated as the treatment or exposure of interest. Let's suppose this is variable X_1 and the research question is about how X_1 affects Y. Or, in other words, if X_1 had been different, how would Y have been different? Let's consider the setting of an RCT and an observational study separately and think of the situation where X_1 is a binary treatment variable.

Randomized controlled trials (RCT)

Suppose individuals are randomized to receive treatment ($X_1 = 1$) or not ($X_1 = 0$), and the outcome Y is observed after some period of follow-up. It is straightforward to estimate the treatment effect in this setting because of the randomization. For a continuous outcome, we would quantify the treatment effect using a difference in the mean outcome in the two treatment groups ($E(Y|X_1 = 1) - E(Y|X_1 = 0)$). For a binary outcome we could quantify the treatment effect in terms of a risk difference ($\Pr(Y = 1|X_1 = 1) - \Pr(Y = 1|X_1 = 0)$), risk ratio ($\Pr(Y = 1|X_1 = 1)/\Pr(Y = 1|X_1 = 0)$) or odds ratio ($\{\Pr(Y = 1|X_1 = 1)/\Pr(Y = 0|X_1 = 1)\}/\{\Pr(Y = 1|X_1 = 0)/\Pr(Y = 0|X_1 = 0)\}$), for example.

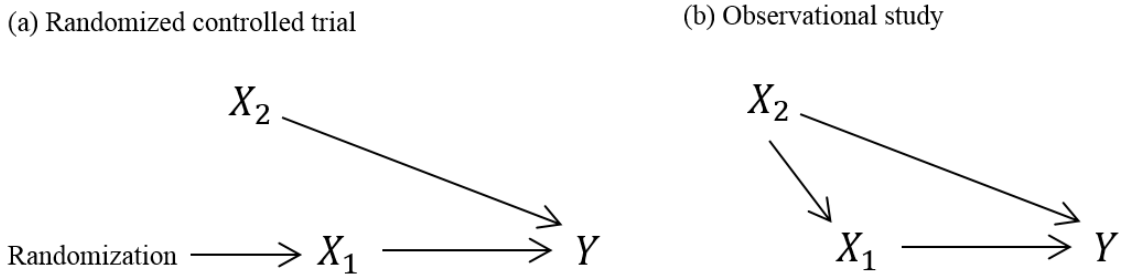
Some of the other explanatory variables $X_2, \ldots, X_p$ are likely to be associated with Y, but we do not need to use them to estimate the treatment effect due to the study design. Sometimes investigators will adjust for baseline variables, measured at the start of the trial prior to treatment. By the study design, baseline variables are not associated with the treatment. There can be advantages of adjusting for baseline variables that are predictors of the outcome. Though there are particular nuances to the interpretation of the resulting estimates depending on the types of outcome (continuous, binary, etc) and on how the treatment effect is quantified.

Of course, there are many important considerations surrounding the validity and interpretation of treatment effects estimated using RCTs, such as whether the effect is a 'per-protocol' or 'intention-to-treat' effect, whether there is drop-out, non-adherence or treatment switching.

Observational studies

Suppose we have available observational data on the treatment variable X_1 and the outcome Y, for example from electronic health records. In this setting the treatment in non-randomized, and there are very likely to be confounders of the association between the treatment and the outcome. A confounder is a variable that affects both the treatment and

Figure 2.2: Directed acyclic graphs (DAGs) illustrating relationships between a treatment X_1, outcome Y and third variable X_2 in a randomized controlled trial and in an observational study.



the outcome. Confounding variables occur prior in time to both the treatment/exposure and the outcome. See VanderWeele and Schpitser (2013) for a formal statistical discussion of confounding.

To estimate the causal effect of X_1 on Y requires us to control for confounding. Consider a simple setting in which there is only one other variable at play, X_2, which in the observational setting affects whether a person gets the treatment X_1 and also affects their outcome Y. For example, if X_1 is a blood pressure-lowering medication and Y is blood pressure 1 year later, then X_2 could be the person's blood pressure at the time origin. The assumed relationships between the three variables X_1, X_2 and Y are illustrated in Figure 2 using directed acyclic graphs (DAGs), contrasting the relationships in an RCT and in an observational study. DAGs, also called 'causal diagrams', are used to graphically describe mechanistic relationships between variable using uni-directional arrows. An arrow connecting two variables indicates (potential) causation in the direction of the arrow and the absence of an arrow indicates an assumption that there is no direct causal effect of the first variable on the second. See Greenland et al. (Epidemiology, 1999) and Shrier and Platt (2008) for introductions to causal diagrams. Some other useful more recent articles on this are from Etminan et al. (2020) and Tenant et al. (2019). In simple situation such as this example, we don't need a DAG to tell us that we need to account for the confounding by X_2 in our analysis in order to estimate the effect of X_1 on Y. However, when there are lots of variables at play DAGs become very useful, and have formal theory attached.

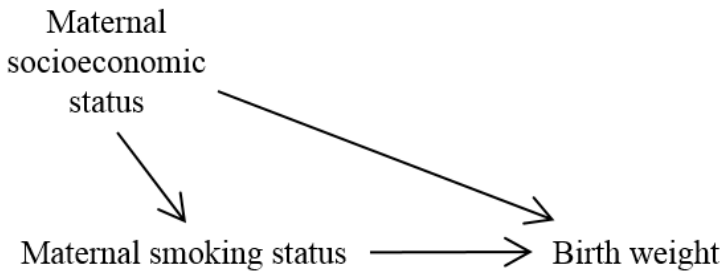
In summary, in a causal investigation the variables on which the research question focuses are X_1 and Y. However, depending on the study design, we may need to account for other variables in the analysis, though those other variables are not our main focus. The concept of confounding is not relevant in prediction investigations.

2.7 Example: the role of regression in different types of investigation

Above, we have contrasted the different types of investigation. Although there are important conceptual differences between the methods, we often use the same statistical tools to address research questions for investigations of different types. Regression is a key tool for analyses in all the types of investigation. In this section we illustrate how the same regression model could be used in prediction investigations and causal investigations, but that the output from the regression should be used and interpreted differently.

We focus on a simple (fictitious) observational study involving three variables: two binary explanatory variables 'maternal smoking status' ($X_1 = 1$: smoker, $X_1 = 0$: non-smoker) and maternal socioeconomic status ($X_2 = 1$: low, $X_2 = 0$: high), and a continuous outcome 'birth weight' (measured in grams). The assumed relationships between the three variables are summarised in the DAG in Figure 3. For the purposes of a simple illustration, we suppose that these are the only three variables at play in this 'system'. In reality of course there are many other maternal and other characteristics that affect a baby's birthweight, such as genetics, maternal diet and alcohol consumption, mother's access to prenatal care, and other features of the environment.

Figure 2.3: Directed acyclic graph (DAG) illustrating relationships between maternal smoking status (X_1), maternal socioeconomic status (X_2) and baby's birth weight (Y)



Consider a linear regression of Y on X_1, X_2, i.e.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{2.1}$$

The estimated regression coefficients and corresponding 95% confidence intervals for this fictitious example are $\hat{\beta}_0 = 3227$ (95% CI $1603, 4851$), $\hat{\beta}_1 = -341$ (95% CI $-513, -169$), $\hat{\beta}_2 = -214$ (95% CI $-18, -410$). We now consider how the output from this regression could be used in different investigation types.

2.7.1 Prediction

If the aim is to predict birth weight based on the two characteristics of the mother, this model allows us to do this. We could obtain the expected value of Y given X_1 and X_2 (in this very simple example there are only 4 possible combinations). In this prediction setting we do not, however, particularly care about the estimates of the regression coefficients. We should instead be concerned with the predictive performance of the model. This could be measured, for example, using R^2. There are many details about how to appropriately assess and quantify the predictive performance of a prediction model which we do not discuss here.

2.7.2 Causality and explanation

Suppose instead that the aim is to assess the causal effect of maternal smoking (X_1) on birth weight (Y) . In the simple setting of Figure 3, maternal socioeconomic status (X_2) is the only confounder of the association between X_1 and Y. The regression model in equation (1) adjusts for X_2 and hence the coefficient for X_1 can be interpreted as the conditional causal effect of X_1 on Y. The 95% confidence interval excludes 0. We can make the interpretation that if all mothers in the study population had smoked, the mean

birthweight would have been 341 grams lower than had all mothers in the study population not smoked. This is referred to as an 'average causal effect'. Here, we have not given any interpretation of the estimate of β_2 because it wasn't relevant for our research question, even though it was important to adjust for X_2 to adjust for confounding. In a more realistic setting, there will be many other variables that confound the association between X_1 and Y and which would need to be accounted for to enable a causal interpretation of β_1.

2.7.3 The "Table 2 Fallacy"

After adjusting for maternal socioeconomic status, maternal smoking was associated with a lowering of 341 grams in mean birthweight. After adjusting for maternal smoking, low maternal socioeconomic status was associated with a lowering of 214 grams in mean birthweight. However, β_1 and β_2 in model (1) do not have the same type of interpretation and this is due to the relationships between the three variables. According to the causal diagram (Figure 3), maternal smoking status is on the causal pathway from socioeconomic status to birth weight. Hence the parameter β_2 in fact represents the effect of socioeconomic status on birth weight that does not go through smoking status - this is a 'direct effect' rather than a 'total effect'. By contrast, β_1 represents the total effect of smoking status on birth weight. We do not go into details about definitions of different types of effect. The aim here is simply to point out that the correct interpretation of the coefficients in the regression model in (1) depends on assumptions about the inter-relationships between the three variables, including how they are ordered in time.

In some (or perhaps many) epidemiological investigations that involve exploration of risk factors, estimates of regression coefficients from multivariable models such as that in (1) (and versions with many more explanatory variables) are presented alongside one another in a table, together with confidence intervals and p-values. They may then be interpreted as though all coefficients had the same meaning, ignoring possible inter-relationships between the variables and temporal ordering. As we have seen from the above example, this could be misleading. This problem has been referred to in the literature the 'Table 2 fallacy', because the estimates of regression coefficients are often presented in 'Table 2' in a paper (where 'Table 1' is usually a table of descriptive statistics). See Westreich and Greenland (2013) for a description of the Table 2 fallacy. Bandoli et al. (2018) provide an example in the context of preeclampsia and preterm birth.

2.8 Other considerations

Above we placed some emphasis on how the investigation type affects what variables should be included in the analysis and on how the results might be interpreted. There are naturally many other things to consider which are beyond the scope of this session. The above example focused on regression. The next few sessions in this module will focus on regression models of different types. They are a fundamental part of the statistician's toolbox and are used in investigations of different types. However, there are many other specialised methods available for specific tasks. For example, in descriptive analyses we may use clustering methods and principal components analysis. In prediction tasks, machine learning methods not based on regression are increasingly used. In studies of causal effects many specialised methods have been developed over recent years. Some of these involve regression and others not.

The type of investigation affects how we should assess the performance and assumptions of a model/analysis. For example, in prediction tasks we should assess how well the prediction model performs in terms of predicting the outcome for a new individual. This requires tools such as cross validation, and measures of predictive performance such as R^2, area under the curve, sensitivity and specificity. In causal analyses we are concerned with whether the assumptions of the models used are valid and whether the model is correctly specified, alongside the validity of untestable assumptions such as whether there are any important confounders that have not been accounted for in the analysis.

This session aimed to provide a broad overview of different types of investigation used in medical statistics/health data science, and which you are likely to encounter in your future careers. This topic has seen some recent emphasis in the literature. The statistical and epidemiological community is increasingly emphasising the need for researchers to ensure they conduct meaningful studies and interpret findings appropriately, particularly relating to the use of observational data. It is a wide topic, and we have only touched on some aspects here.

References

NOTE: You are not expected to read all of these references! It is intended as a list of resources that you may find useful in the future or if you wish to follow-up on some of the topics discussed in more detail.

Bandoli G., Palmsten K., Chambers C.D., et al. Revisiting the Table 2 fallacy: A motivating example examining preeclampsia and preterm birth. Pediatric and Perinatal Epidemiology 2018; 32: 390-397.

D'Agostino R.B., Vasan R.S., Pencina M.J., et al. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. Circulation 2008; 117: 743–753.

Etminan M, Collins GS, Mansournia MA. Using Causal Diagrams to Improve the Design and Interpretation of Medical Research. CHEST 2020; 158: Supplement S21-S28.

Greenland S., Pearl J., Robins J.M. Causal diagrams for epidemiological research. Epidemiology 1999; 10:37–48.

Hand D. What is the Purpose of Statistical Modelling? Harvard Data Science Review 2019 https://doi.org/10.1162/99608f92.4a85af74

Hernán M.A. Does water kill? A call for less casual causal inferences. Annals of Epidemiology 2016; 26: 674-680.

Hernán M.A. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. Am J Public Health. 2018;108: 616–619.

Hernán M.A., Hsu J., Healy B.. A second chance to get causal inference right: a classification of data science tasks. Chance 2019; 32: 42-49.

Huebner M., le Cessie S., Schmidt C., Wach W. A Contemporary Conceptual Framework for Initial Data Analysis. Observational Studies 2019; 4: 171-192.

Riley R.D. et al. Prognosis Research in Healthcare: Concepts, Methods, and Impact. 2019. Oxford University Press.

Schmueli. To explain or to predict? Statistical Science 2010; 25: 289-310.

Schooling CM, Jones H. Clarifying questions about "risk factors": predictors versus explanation. Emerging Themes in Epidemiology 2018; 15: 10.

Schrier I., Platt R.W. Reducing bias through directed acyclic graphs. BMC Medical Research Methodology 2008; 8: 70.

Spiegelhalter D. *The Art of Statistics: Learning from Data.* 2019. Penguin.

Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd Edition. 2019. Springer.

Tennant PWG, Harrison WJ, Murray EJ, et al. Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations. MedRxiv 2019. https://www.medrxiv.org/content/10.1101/2019.12.20.19015511v1

VanderWeele T.J., Shpitser I. On the definition of a confounder. Annals of Statistics 2013; 41: 196-220.

Westreich D., Greenland S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. American Journal of Epidemiology 2013; 177: 292-298.

2.9 Practical 2

In the practical you will divide into groups to discuss example papers from the literature with respect to the issues raised in this session. One week in advance of the practical we will send out information on your groups and we will provide the paper that each group should discuss. Please read the paper assigned for the group discussions in advance.

In your group discussion please consider the following questions:

1. What type of investigation is presented in the paper?
2. Is the investigation type and research question made clear?
3. What data sources were used?
4. What analysis approaches were used?
5. How were the models used assessed?
6. How were the results presented?
7. What conclusions were drawn and did the authors discuss strengths/limitations of their study?
8. Did you discover methods or concepts that are new to you?

In the last part of the practical the whole class will join together and each group is asked to present their findings for 5-7 minutes.

- You do not have to cover all of the above questions in your discussion or presentation – some papers may raise more discussion on certain questions than others.
- You can choose one member of your group to present, or do it as a group or subset thereof.
- Your presentation can be verbal, or you can make slides, or any other suitable format that you can think of. You can choose to present figures or tables extracted from the paper.
- You are not expected to understand everything covered in these papers!
- This is not an assessed presentation!

Session 3

Models for Binary Data

This session will cover:

1 the difference between grouped and individual binary data;

2 classes of GLM for binary data;

3 examples of using logistic regression for binary data.

We begin with a general discussion and then consider an extended example. The practical gives you a chance to work through a similar example.

3.1 Examples of binary data

- Clinical trial of bypass surgery vs angioplasty, outcome: death or myocardial infarction (MI) yes/no;
- mechanical heart valve, outcome: fail/not fail;
- toxicological experiment where mice are exposed to different doses of carbon disulphide, outcome: mouse dead/alive;
- prospective study of effect of general practice lifestyle intervention on risk of death/MI, outcome death yes/no.

3.2 Grouped and individual binary data

Consider the data shown in Figure 3.1. It comes from an experiment in which insects are exposed to different doses of a drug `xi`.

`insect-grouped` contains **grouped** data, in dose group $i, (i = 1, \ldots, 8)$, `yi` deaths are observed out of `ni` individuals.

`insect-individual` contains individual data, for each individual $i, (i = 1, \ldots, 481)$, `yi` is 1 if the individual dies, 0 if the individual survives. Usually, if the data are stored as individual records they are called **Bernoulli** data, while grouped records are called **binomial** data. However, both representations correspond to the same set of binary data.

3.3 GLM for binary data

As usual, the GLM has thee components. For binary data these are:

insect-grouped

xi	yi	ni
49.06	6	59
52.99	13	60
56.91	18	62
60.84	28	56
64.76	52	63
68.69	53	59
72.61	60	62
76.54	59	60

insect-individual

xi	yi
49.06	1
49.06	1
49.06	1
49.06	1
49.06	1
49.06	1
49.06	0
49.06	0
49.06	0
49.06	0
.	.
.	.
.	.

Figure 3.1: Insect data, arranged in groups (left) and as individual observations (right)

1 *Distribution:* independent response variables,

$$\begin{aligned} Y_i &\sim \text{Bin}(n_i, \pi_i), \quad i = 1, \ldots, n. \\ \text{E}(Y_i) &= \mu_i = n_i \pi_i \end{aligned}$$

2 *Linear predictor:* let $x_{i1}, x_{i2}, \ldots, x_{ip}$ be measured values of explanatory variables/covariates for the ith group (or ith individual when $n_i = 1$ for all i). The linear predictor is

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

3 *Link function:* the function that specifies how $\mu_i = n_i \pi_i$ relates to the explanatory variables. With binary data this is usually expressed in terms of π_i rather than μ_i. There are a number of different link functions used in applications including:

(a) **logit:**

$$\eta_i = \text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right).$$

(b) **log:**

$$\eta_i = \log(\pi_i) = \ln(\pi_i).$$

(c) **identity:**

$$\eta_i = \pi_i.$$

(d) **probit:**

$$\eta_i = \text{probit}(\pi_i) = \Phi^{-1}(\pi_i).$$

(e) **complementary log-log:**

$$\eta_i = \text{cloglog}(\pi_i) = \ln\{-\ln(1-\pi_i)\}.$$

3.4 Notes

1 For the probit link, Φ is defined to be the cumulative distribution function of a standard normal distribution, i.e. $\Phi(z) = \Pr(Z \leq z)$ for $Z \sim N(0, 1)$.

2 The natural (canonical) parameter of the binomial distribution is $\theta_i = \text{logit}(\pi_i)$.

3 The logit, probit and complementary log-log link functions each map probability (on the range (0,1)) to the linear predictor on the range $(-\infty, \infty)$. The log link maps probability on the range $(-\infty, 0)$. The finite boundary can lead to non-convergence of the iterative fitting procedures.

4 The identity, logit and probit functions are symmetric about $\pi = 0.5$ but the log and complementary log-log functions are not.

5 The log function has the advantage that parameters correspond to (log) risk ratios. Furthermore, by using a log link for binary data we do not have to worry about the issue of non-collapsibility, a topic which we will return to later. Set against these apparent advantages are that models fitted using this link sometimes do not converge, and that it is possible to have predicted probabilities greater than one!

6 With the identity function the parameters correspond to risk differences. As with the log link there is no issue of non-collapsibility. However, models fitted using this link sometimes do not converge, and it is again possible to have predicted probabilities greater than one.

7 The logit function has the advantage of direct interpretation in terms of log odds and is computationally convenient. For this reason the logit is the most commonly used link for binary data, hence the term **logistic regression**.

8 With individual data, $n_i = 1$, and i indexes each individual. Then Y_i takes the value 0 ('failure') or 1 ('success'). With grouped data, i indexes each group as defined by the levels of the explanatory variables; n_i is the number of individuals in the ith group, and Y_i is the number of 'successes' out of the n_i subjects. This distinction is important in interpreting some of the output from logistic regression.

EXERCISE 3.1 *Link functions*

Derive, in terms of the parameters $\beta_0, \beta_1, \ldots, \beta_p$, expressions for π_i and μ_i for the identity, log, logit and complementary log-log links.

3.5 Extended example

3.5.1 The data

Let y_i denote the number of BSE infected cattle out of n_i, where x_{i1} is the dilution factor of the feed received and x_{i2} is an indicator variable: 0 if the cattle are from herd one and 1 from herd two. The data are as follows.

group	dfactor	cattle	infect
1	1	11	8
1	2	10	7
1	3	12	5
1	4	11	3
1	5	12	2
2	1	10	10
2	2	10	9
2	3	12	8
2	4	11	6
2	5	10	4

(A dilution factor of 2 means a dilution of 10^{-2}.)

3.5.2 Analysis aims

We wish to answer the questions:

1 What is the effect of dilution factor on the chance of infection, allowing for differences between the two herds of cattle?

2 Does the effect of dilution factor differ between herds 1 and 2?

3.5.3 Logistic regression model 1: dilution factor + herd

Assume $Y_i \sim \text{Bin}(n_i, \pi_i)$ independently, and

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \quad \textit{i.e. } \pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}}.$$

3.5.4 Parameter estimation

The log-likelihood $\ell(\boldsymbol{\pi} \mid \mathbf{y})$ can be written as a function of $\boldsymbol{\beta}$: $\ell(\boldsymbol{\beta} \mid \mathbf{y})$. This is maximised with respect to $\boldsymbol{\beta}$ to give the set of MLE's $\hat{\boldsymbol{\beta}}$.

```
. glm infect i.group dfactor, family(binomial cattle) link(logit)

Iteration 0:   log likelihood =  -13.131766
Iteration 1:   log likelihood =  -13.12687
Iteration 2:   log likelihood =  -13.12687

Generalized linear models                         No. of obs      =        10
Optimization     : ML: Newton-Raphson             Residual df     =         7
                                                  Scale parameter =         1
Deviance         =  2.450823011                   (1/df) Deviance =  .3501176
Pearson          =   1.82300189                   (1/df) Pearson  =  .2604288

Variance function: V(u) = u*(1-u/cattle)          [Binomial]
Link function    : g(u) = ln(u/(cattle-u))        [Logit]

                                                  AIC             =  3.225374
Log likelihood   = -13.12686977                   BIC             = -13.66727

------------------------------------------------------------------------------
             |                 OIM
      infect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.group |   1.305903   .4653996     2.81   0.005     .3937366     2.21807
     dfactor |   -.787442   .1813483    -4.34   0.000    -1.142878   -.4320059
       _cons |   2.131041   .6113036     3.49   0.000     .9329085    3.329174
------------------------------------------------------------------------------
```

Note that the fitted values from the model obtained by the Stata command `predict` are the fitted probabilities $\hat{\pi}_i$ where

$$\text{logit}(\hat{\pi}_i) = \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2},$$

which can be re-expressed as

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}}}.$$

Parameter interpretation

- The intercept, $\hat{\beta}_0$ is the estimated log odds of infection given x_1 and x_2 are both 0, i.e. at dilution 0 in herd 1.
- The slope, $\hat{\beta}_1$ is the estimated increase in log odds of infection per unit increase in x_1, the dilution factor, given herd is held constant.
- $\hat{\beta}_2$ is the estimated change in log odds of infection comparing $x_2 = 1$ with $x_2 = 0$, that is, between herds 2 and 1, given dilution is held constant. A difference in log odds is a **log odds-ratio**, and so the estimated odds ratio for comparing herd 2 with herd 1 is

$$\exp(\hat{\beta}_2), \quad \text{with 95\% CI} \quad \exp\{\hat{\beta}_2 \pm 1.96 \times \text{Std. Err. } (\hat{\beta}_2)\}.$$

EXERCISE 3.2 *Interpretation of odds ratios*

What is the estimated odds-ratio and 95% CI per unit increase in dilution factor, holding herd constant (i.e adjusting for herd)?

*3.5.5 Logistic regression model 2: + dilution factor*herd*

Question: Does the effect of dilution factor differ between the two herds? That is, does the odds ratio per unit increase in dilution differ between herds 1 and 2?

Model 2: Define $x_{i3} = x_{i1}x_{i2}$, and extend the linear predictor from model 1:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

As before, we use Stata to obtain the MLE's:

```
. glm infect i.group##c.dfactor, family(binomial cattle) link(logit)

Iteration 0:   log likelihood = -12.975946
Iteration 1:   log likelihood = -12.973836
Iteration 2:   log likelihood = -12.973836

Generalized linear models                         No. of obs      =         10
Optimization     : ML: Newton-Raphson             Residual df     =          6
                                                  Scale parameter =          1
Deviance         =  2.144755929                   (1/df) Deviance =  .3574593
Pearson          =  1.693215575                   (1/df) Pearson  =  .2822026

Variance function: V(u) = u*(1-u/cattle)          [Binomial]
Link function    : g(u) = ln(u/(cattle-u))        [Logit]

                                                  AIC             =  3.394767
Log likelihood   = -12.97383623                   BIC             = -11.67075

------------------------------------------------------------------------------
             |                 OIM
      infect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.group |   1.988669   1.344712     1.48   0.139    -.6469174    4.624255
     dfactor |  -.7050829   .2295473    -3.07   0.002    -1.154987   -.2551785
             |
group#c.dfactor |
           2 |  -.2058331   .3755349    -0.55   0.584     -.941868    .5302018
             |
       _cons |   1.890284   .7358402     2.57   0.010     .4480635    3.332504
------------------------------------------------------------------------------
```

The odds-ratio per unit increase in dilution factor is

$$\exp(-0.705) = 0.49 \text{ in herd 1, and}$$
$$\exp(-0.705 - 0.206) = 0.40 \text{ in herd 2.}$$

3.5.6 Does the effect of dilution factor differ between herds 1 and 2?

Having fitted the model which allows for an interaction between herd and dilution factor, we can now perform a test to answer the second of our questions. The effect of dilution factor is the same in the two herds when $\beta_3 = 0$, and thus we want to test the null hypothesis that $\beta_3 = 0$. Because the null hypothesis concerns only a single parameter, we can perform a (univariate) Wald test.

This is in fact the test that is reported in each row of the output, for each parameter. Thus the p-value 0.584 is the Wald test p-value for testing that $\beta_3 = 0$. There is thus no evidence against the null hypothesis of no interaction, and we might decide that the simpler model (1) is preferable.

Note: if we wanted to perform a multivariate Wald test that more than one coefficients are simultaneously zero, we can use Stata's `test` command.

3.6 Analysis using other link functions

In order to obtain estimates of risk ratios and risk differences the logit link function needs to be replaced by log (`link(log)` in Stata) and identity (`link(id)`) link functions respectively.

3.6.1 The impact of changing a link function on interactions

Statistical interactions are scale dependent, so changing a link function can alter whether or not interaction terms are required in a statistical model.

Suppose that a generalized linear model for some data includes two predictor variables that are both related to an outcome and that in truth, for the specified link function, there is no interaction between them. If the link function is changed in a non-linear fashion then there will almost always be an interaction with this new choice of link function. In the BSE in cattle example considered above if a 1 unit increase in dilution factor has the same effect on the log odds of infection in the two herds then it cannot have the same effect on the log risk of infection in both herds.

As a second illustrative example suppose that two factors (A and B) are related to risk of some outcome as in the second and third columns of the following table.

Table 3.1: An illustration of the fact that interactions are scale dependent

	Risk of outcome		Effect of Factor A		
	Factor A absent	Factor A present	risk difference	risk ratio	odds ratio
Factor B absent	0.1	0.25	0.15	2.5	3.0
Factor B present	0.25	0.5	0.25	2.0	3.0

The table also shows risk differences, risk ratios and odds ratio for the effect of Factor A on the outcome according to the presence or absence of Factor B. The odds ratio relating Factor A to the outcome is 3 whether B is absent ($\frac{0.25/0.75}{0.1/0.9}$) or present ($\frac{0.5/0.5}{0.25/0.75}$). However, the corresponding risk ratios and risk differences are not the same. If we were to model this data using either a log or identity link then we would require an interaction between Factor A and Factor B, but this would not be required with a logit link.

Of course, when we fit statistical models it is quite possible that an interaction term that is not statistically significant can remain so when switching from a logit to a log link function, but in truth it is not possible for an interaction not to be required for at least one of these link functions (other than in the trivial case where one factor is not related to the outcome).

EXERCISE 3.3 *Risk ratios and risk differences*

Suppose the following commands had been used to analyse the data considered above. In both cases what would be the interpretation of each parameter estimate (if the models converge)?

```
. glm infect i.group##c.dfactor, family(binomial cattle) link(log)

. glm infect i.group##c.dfactor, family(binomial cattle) link(id)
```

3.7 Models for individual binary data in Stata

As well as the `glm` command, Stata includes a number of commands for fitting GLMs for individual binary data using particular link functions. The table below shows some of these, along with the corresponding options that achieve the same result but using the `glm` command.

Link function	Stata command	`glm` options
logit	`logit, logistic`	`,fam(binomial) link(logit)`
probit	`probit`	`,fam(binomial) link(probit)`
cloglog	`cloglog`	`,fam(binomial) link(cloglog)`

3.8 Practical 3

Dataset required: `insect.dta`

Introduction

The purpose of this session is to learn how to fit and interpret generalised linear models to binary data.

We will use data from a toxicological experiment. Eight groups of insects were exposed for five hours to gaseous carbon disulphide (CS_2) at different concentrations, with the purpose of investigating how the risk of death depends on the dose received.

The data are in `insect.dta`, with one record per group as defined by CS_2 dose, and three variables, as follows:

dose = CS_2 dose (mg/l);

r = number of insects killed;

n = number of insects in group.

You should have your own Do file and run the commands yourself, but you should discuss the results, interpretations and any queries within your Breakout Room.

Aims

The aims of the analysis are to answer the following questions:

1 Is there an association between dose level and the proportion of insects killed?

2 On what scale is this association best modelled: identity, log, logit?

3 What is the nature of this association: linear or quadratic?

Investigation into effect of CS_2 at different doses on insect survival

1 Open Stata, start a new Do file, and load the data. Explore the dataset. Are there any missing values? How many insects do you have data for?

2 Generate a new variable for the proportion of insects killed at each dose of CS2. List these proportions and plot them against the dose. What do you conclude? Why would simple linear regression not be appropriate for analysing the association between the proportion killed and dose?

3 Generate new variables for the

   (a) log

   (b) log odds

   of being killed at each dose. Plot these values against the dose. What do you conclude?

   Discussion: Looking at these plots, which link function is likely to best fit these data?

4 Write down algebraically an appropriate generalized linear model for these data, which can be used to investigate the association between dose received and risk of death. Your model should specify the distribution, the linear predictor and the link function.

5 Fit the model and obtain MLE's of the parameters using the `glm` command in Stata.

(a) According to the fitted model, what is the estimated probability of death at a dose of 55 mg/l CS_2?

(b) Calculate (using pen, paper and a calculator) the dose that, on the basis of this model, would lead to a 50% death rate (sometimes termed the LD50).

(c) Is there evidence of an increasing risk of death with increasing dose?

(d) Interpret the dose parameter in terms of an odds-ratio, and calculate its 95% CI. Check your calculation by using the eform option of the `glm` command.

6 Look at the fitted values from the model obtained using predict. What variable has been fitted? Plot both the fitted and observed proportions against the dose.

Discussion: Compare the fitted and observed proportions. What do you conclude?

7 Calculate a new variable for the square of the dose. Use this to investigate whether the inclusion of a quadratic dose term improves the fit of the linear dose model.

(a) Test whether there is evidence against linearity by performing a Wald test of the quadratic coefficient. What do you conclude?

(b) Compare the fitted proportions from this model with those you obtained from the previous model and also with the original data, for example by plotting a graph against dose. In what way has the fit been improved?

8 Now consider dose as a categorical variable, fitting generalized linear models with logit, log and identity links.

(a) Fit the models using the following series of commands, interpreting each of the parameter estimates obtained. What do you notice?

```
egen dosecat = rank(dose)

glm r i.dosecat, family(binomial n) link(logit)

glm r i.dosecat, family(binomial n) link(log)

glm r i.dosecat, family(binomial n) link(id)
```

(b) Fit a logistic regression model that includes dose as both categorical and continuous. Test the effect of the categorical dose variable.

```
glm r dose i.dosecat, family(binomial n) link(logit)

test 2.dosecat 3.dosecat 4.dosecat 5.dosecat 6.dosecat 7.dosecat
```

Discussion: Why are there estimates for only six (of eight) dose categories in this model? Discuss the interpretation of each parameter estimate and the test statistic computed by the final command above.

Discussion: Working together with one or more colleagues (in your Breakout Room if online), write a paragraph to answer the aims of the analysis. If online, one of you should post your group's paragraph in the Zoom chat.

Session 4

Model comparison and goodness of fit

4.1 Aims and Objectives

The aim of this session is to introduce the key ideas involved in assessing how to compare the fit of nested GLMs, and also to look at one approach for assessing goodness of fit.

4.2 Introduction

Fitting a GLM to data means defining the log likelihood of the parameters $\boldsymbol{\beta}$ given the data and maximising this to give the MLE's $\hat{\boldsymbol{\beta}}$. In many situations the MLEs cannot be expressed algebraically in 'closed form' and so iterative methods are necessary to find the solution.

The absolute value of the maximised log likelihood will be very dependent on the particular data observed and so is not of interest in itself. However, the difference in this quantity for two models, one nested within the other, does provide a measure of the comparative fit of the two models. The more general model will necessarily have the greater log likelihood, so the question is whether the difference in log likelihoods is large enough to indicate that the more general model provides a 'real' improvement in fit. We consider this in the following sections.

4.3 Comparing nested models

Suppose we have a dataset and we want to compare the fit of two particular GLMs, model 1 and model 2. In the following, we describe how we can assess the relative fit of the two models, provide they are *nested*. Model 1 is *nested* in model 2 if model 1 can be obtained by some restriction (often setting to zero) of parameters in model 2.

EXAMPLE 4.1 *Nested models I*

If model 1 has linear predictor for the ith unit

$$\eta_i = \beta_0 + \beta_1 x_{i1}$$

and model 2, with the same distribution, link function and scale parameter, has linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

then model 1 is nested in model 2 because it can be derived from it by setting $\beta_2 = \beta_3 = 0$.

EXAMPLE 4.2 *Nested models II*

Suppose the variable x_1 is continuous. Suppose, as before, model 1 has linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1}$$

while model 2 has linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2,$$

Model 1 is nested in model 2 because it can be obtained by setting $\beta_2 = 0$.

More generally, assume that model 1 is nested in model 2. Further, suppose that we write the parameters for model 2 as $(\boldsymbol{\psi}, \boldsymbol{\lambda})$, such that the restriction that $\boldsymbol{\psi} = 0$ gives us model 1. To compare the fit of model 1 and model 2, we can test whether the extra complexity of model 2 is necessary. That is, we want to test the null hypothesis that $\boldsymbol{\psi} = 0$.

One approach to testing the null hypothesis that $\boldsymbol{\psi} = 0$ would be to perform a (possibly multivariate) Wald test. However, as outlined in the first session, the Wald test is not invariant to parameter transformation, whereas the log-likelihood ratio test is. Because the null hypothesis only constrains the parameter(s) $\boldsymbol{\psi}$, but not $\boldsymbol{\lambda}$, we can appeal to the profile likelihood results from inference. This states that to test $\boldsymbol{\psi} = 0$ we can calculate the profile log likelihood ratio statistic

$$-2pllr(\boldsymbol{\psi} = 0) = -2\left\{\ell_p(\boldsymbol{\psi} = 0) - \ell_p(\hat{\boldsymbol{\psi}})\right\}$$

where $\hat{\boldsymbol{\psi}}$ denotes the MLE of $\boldsymbol{\psi}$ in model 2. Asymptotically, under the null hypothesis this statistic follows a χ^2 distribution, with degrees of freedom equal to the dimension of $\boldsymbol{\psi}$. The profile log-likelihood value $\ell_p(\boldsymbol{\psi} = 0)$ is the maximum value of the likelihood function of $(\boldsymbol{\psi}, \boldsymbol{\lambda})$ when $\boldsymbol{\psi}$ is constrained to be zero. This is precisely the value of the maximized likelihood function for the simpler model, model 1, which we denote ℓ_1. The profile log-likelihood $\ell_p(\hat{\boldsymbol{\psi}})$ is the maximum value of the ordinary likelihood function when $\boldsymbol{\psi}$ is held fixed at its MLE $\hat{\boldsymbol{\psi}}$. Thus $\ell_p(\hat{\psi})$ is simply the value of the maximized log likelihood function for model 2, which we denote ℓ_2. So, we have

$$-2pllr(\boldsymbol{\psi} = 0) = -2(\ell_1 - \ell_2)$$

In summary, to compare the fit of two nested GLMs, we can compare minus twice the difference in their log-likelihood values to a χ^2 distribution, with degrees of freedom equal to the difference in the number of parameters between the two models. A significant p-value indicates evidence against the null hypothesis that the simpler model (model 1) is correctly specified. Or put another way, significant evidence that model 2 fits the data better than model 1.

Example: BSE infection of cattle

To illustrate, we return to the example of the previous session in which we modelled how the odds of infection of cattle with BSE varied by herd and dilution factor of the feed received. In the previous session we tested whether there was evidence that the effect of dilution factor differed between herds 1 and 2. We tested this hypothesis by performing a Wald test that the coefficient corresponding to the interaction variable was zero. This resulted in a p-value of 0.584.

We now consider testing the same null hypothesis by comparing the fits of model 1 (without the interaction) and model 2 (with the interaction). We can do this by performing a

(profile) log-likelihood ratio test as described above. The log likelihood value for model 1 was -13.1269 and for model 2 it was -12.9738. Therefore we have

$$-2pllr(\beta_3 = 0) = -2(\ell_1 - \ell_2) = -2(-13.1269 + 12.9738) = 0.3062$$

Because the difference in the number of parameters between the models is one, we compare 0.3062 to a χ^2 distribution on one degree of freedom, giving a p-value 0.580. Notice that this very similar to the Wald test p-value obtained previously. The two tests will typically give similar results because they are asymptotically equivalent.

In Stata the `lrtest` command can be used to carry out this test.

```
. glm infect i.group dfactor, family(binomial cattle) link(logit)

Generalized linear models                         Number of obs   =         10
Optimization     : ML                             Residual df     =          7
                                                  Scale parameter =          1
Deviance         =  2.450823011                   (1/df) Deviance =   .3501176
Pearson          =   1.82300189                   (1/df) Pearson  =   .2604288

Variance function: V(u) = u*(1-u/cattle)          [Binomial]
Link function    : g(u) = ln(u/(cattle-u))        [Logit]

                                                  AIC             =   3.225374
Log likelihood   = -13.12686977                   BIC             =  -13.66727

------------------------------------------------------------------------------
             |                 OIM
      infect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.group |   1.305903   .4653996     2.81   0.005     .3937366     2.21807
     dfactor |   -.787442   .1813483    -4.34   0.000    -1.142878   -.4320059
       _cons |   2.131041   .6113036     3.49   0.000     .9329085    3.329174
------------------------------------------------------------------------------

. est store model1

. glm infect i.group##c.dfactor, family(binomial cattle) link(logit)

** output omitted **

. est store model2

. lrtest model1 model2

likelihood-ratio test                           LR chi2(1)  =      0.31
(Assumption: model1 nested in model2)           Prob > chi2 =    0.5801
```

4.4 The saturated model, deviance, and goodness of fit

Recall that for linear regression models, the residual sum of squares gives a measure of how well a model fits the data. The F-test compares the fits of two nested models by formally assessing the reduction in the residual sum of squares achieved by adding extra covariates. However, note that the residual sum of squares cannot be used to assess the fit of an individual model in isolation; this is because in linear regression the residual sum of squares is needed to estimate the unknown residual variance and so cannot additionally be used to assess model fit. Were the residual variance known (as is possible theoretically, but not in practice) then it could be.

In this section we will extend this approach to GLMs, first introducing the notion of a saturated model, and then using this to define the deviance. The deviance is a measure of model fit. The log-likelihood ratio test introduced above is a comparison of deviances. In addition, for certain types of GLM, under certain conditions, deviance can be used to assess the goodness of fit of an individual model in isolation.

4.4.1 The saturated model

The **saturated model** is one with the maximum possible number of parameters, with no redundancies. Usually this implies the same number of parameters as observations, and a model whose fitted values exactly reproduce the observations. For example, if

1 there are n observations $y_i, i = 1, \ldots, n$, each from a normal distribution, with expected value μ_i;

2 the linear predictor is $\eta_i = \beta_1 x_{i1} + \cdots + \beta_n x_{in}$, where x_{ij} is an indicator variable equal to one if $i = j$ and zero otherwise;

3 the link function is the identity, so $\eta_i = \mu_i$;

then the model is saturated and $\hat{\beta}_i = y_i, i = 1, \ldots, n$.

For a given data set, the saturated model provides the best possible fit, and so has a log likelihood that is larger than for any other model. Whether the saturated model can be a reasonable model, in the sense of being useful for answering scientific questions of interest and/or predicting future outcomes, depends on the structure of the covariates. For example, if there are only two binary covariates, the saturated model contains the main effects of these covariates plus their interaction. If the interaction is in truth non-zero, this will be an appropriate model. Conversely, if there is only a single continuous covariate, a separate parameter must be estimated for each realised value of the continuous covariate, and this saturated model will typically be useless in terms of answering our scientific question or being used to predict future outcomes.

Example: BSE infection of cattle

To illustrate, we again return to the example where we explored how the odds of infection of cattle with BSE varied by herd and dilution factor of the feed received. The data is in 10 groups so the saturated model has to have 10 parameters. The easiest way to fit this model is to include an interaction between herd and dilution factor, treating dilution factor as a categorical variable.

```
. glm infect i.group##i.dfactor, family(binomial cattle) link(logit)

Generalized linear models                         Number of obs   =         10
Optimization     : ML                             Residual df     =          0
                                                  Scale parameter =          1
Deviance         =  2.23478e-07                   (1/df) Deviance =          .
Pearson          =  1.43315e-07                   (1/df) Pearson  =          .

Variance function: V(u) = u*(1-u/cattle)          [Binomial]
Link function    : g(u) = ln(u/(cattle-u))        [Logit]

                                                  AIC             =   4.380292
Log likelihood   = -11.90145838                   BIC             =   2.23e-07
------------------------------------------------------------------------------
             |                 OIM
      infect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.group |   17.66102   3532.022     0.01   0.996    -6904.975    6940.297
             |
     dfactor |
           2 |  -.1335284   .9667057    -0.14   0.890    -2.028237     1.76118
           3 |   -1.31729   .8950912    -1.47   0.141    -3.071637    .4370562
           4 |  -1.961637   .9574247    -2.05   0.040    -3.838155   -.0851186
           5 |  -2.590201   1.028741    -2.52   0.012    -4.606498    -.573905
             |
       group#|
     dfactor |
         2 2 |  -16.31134   3532.022    -0.00   0.996    -6938.947    6906.325
         2 3 |   -16.6314   3532.022    -0.00   0.996    -6939.267    6906.004
         2 4 |  -16.49788   3532.022    -0.00   0.996    -6939.134    6906.138
         2 5 |   -16.4571   3532.022    -0.00   0.996    -6939.093    6906.179
             |
       _cons |   .9808183   .6770015     1.45   0.147    -.3460803    2.307717
------------------------------------------------------------------------------
```

4.4.2 Deviance

If L_c is the (maximized) likelihood of some current model of interest and L_s is the likelihood of the saturated model, the likelihood ratio is L_c/L_s and the **scaled deviance** S of the current model is defined as

$$S = -2\ln\left(\frac{L_c}{L_s}\right) = -2(\ell_c - \ell_s)$$

for ℓ_c and ℓ_s the corresponding log likelihoods. The unscaled, or ordinary, **deviance**, D, is defined as ϕS, for ϕ the scale parameter. For the examples which we will use in this course (binomial and Poisson distributions), $\phi = 1$, and so the scaled and unscaled deviances are identical.

Note: in some texts what we have defined as the scaled deviance is defined as the deviance and vice-versa. Our definition matches that used in Stata and in McCullagh and Nelder.

The deviance can be described as measure of 'badness of fit'; a large value of D, i.e. $L_c << L_s$, suggests that the current model does not fit the data well.

The relative fit of two nested models can be compared by comparing their respective scaled deviances. The change in scaled deviance comparing models 1 and 2 is

$$S_1 - S_2 = -2(\ell_1 - \ell_2)$$

This difference in scaled deviances exactly matches the profile log-likelihood ratio statistic we derived earlier. Therefore, comparing the fit of two nested GLMs through comparison of their scaled deviances is in fact simply the profile log-likelihood ratio test which we described previously.

4.4.3 Deviance for grouped binary data

Suppose we have n independent observations from a binomial distribution, with $Y_i \sim Bin(n_i, \pi_i)$, $i = 1, .., n$, i.e. grouped binary data. Ordinarily, the groups are defined by combinations of levels of categorical/factor covariates. The saturated model for this data allows the probability of 'success' to be different in each group, so that $\tilde{\pi}_i = y_i/n_i$, i.e. the fitted probability of success in each group is simply the observed proportion of successes in the data. The log-likelihood of this saturated model is then

$$\ell_s = \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i \log(\tilde{\pi}_i) + (n_i - y_i) \log(1 - \tilde{\pi}_i) \right\}$$

Now suppose we have fitted some (non-saturated) model (e.g. a logistic regression) to the data, giving predicted probabilities $\hat{\pi}_i$, $i = 1, .., n$. Then the log-likelihood value for this model is similarly

$$\ell_c = \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i \log(\hat{\pi}_i) + (n_i - y_i) \log(1 - \hat{\pi}_i) \right\}$$

The deviance is then equal to

$$\begin{aligned} S &= -2(\ell_c - \ell_s) \\ &= 2 \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{\tilde{\pi}_i}{\hat{\pi}_i} \right) + (n_i - y_i) \log \left(\frac{1 - \tilde{\pi}_i}{1 - \hat{\pi}_i} \right) \right\} \end{aligned}$$

Looking at this expression, we can see that the deviance will be small when the predicted probabilities $\hat{\pi}_i$ under our current model are close to the predicted probabilities under the saturated model, $\tilde{\pi}_i$. Large values of the deviance indicate that our model is not fitting the data very well.

In order to assess whether the deviance is larger than we might expect by chance, we need to know the sampling distribution of the deviance, assuming that the null hypothesis is true: that our current model is correctly specified. As we have noted previously, the deviance is the profile log-likelihood ratio test statistic for comparing the saturated model to the current model. Therefore asymptotically it has a χ^2 distribution on $n-p$ degrees of freedom, where p is the number of parameters in our model. Intuitively, this test assesses whether the observed proportions $\tilde{\pi}_i$ differ from the model predicted proportions $\hat{\pi}_i$ by more than would be expected by chance.

Example: BSE infection of cattle

To illustrate, we return to the example where we modelled how the odds of infection of cattle with BSE varied by herd and dilution factor of the feed received.

For the first model fitted to the BSE infection of cattle data (the model with dose considered as continuous and no interaction) the deviance is 2.4508. This is equal to minus twice the difference in log-likelihoods for this model (-13.1269) and the saturated model (-11.9015). Since there are 10 observations, and the model has 3 parameters (intercept, dilution parameter, herd parameter), the residual d.f. is thus $10 - 3 = 7$. If the model is correct, we expect the deviance to be close to 7. In fact, since the deviance is much smaller than 7, the model fits somewhat better than we would expect by chance. Formally the p-value is 0.93 (comparing the deviance to a χ^2 distribution on 7 d.f.), so no evidence of poor fit.

4.4.4 Pearson's goodness of fit statistic

When fitting a GLM in Stata Pearson's goodness of fit statistic is provided along with the Deviance. This is

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\widehat{Var}(Y_i)}$$

where y_i denotes the ith observed outcome, $\hat{\mu}_i$ is its predicted mean, and $\widehat{Var}(Y_i)$ is the estimated variance of Y_i, given the model covariates. For grouped binary data, provided the fitted values are not too small (e.g. less than 5), and the number of groups would not increase with increasing sample size (i.e. increasing the n_i in the groups, but not the number of groups n), Pearson's chi-square statistic approximately follows a χ^2 distribution on $n - p$ degrees of freedom, where n denotes the number of parameters in the saturated model.

The restriction that the fitted values are not too small rules out the use of this for assessing goodness of fit for individual binary data, where necessarily the expected/fitted values are between zero and one.

4.5 Assessing goodness of fit for individual binary data

If we have ungrouped binary data, it turns out that the deviance (like Pearson's goodness of fit statistic) is not useful as a measure of goodness of fit. Intuitively, this is because for individual binary data, whatever our model predicted probability $\hat{\pi}_i$ is, an observed 0/1 value of Y does not, on its own, give us any information about how good the predicted probability is. Put another way, if $\hat{\pi}_i$ is close to zero, but we observed a success for individual i, we have no idea whether this is due to lack of fit, or is just due to chance - even if the true π_i is close to zero, occassionally the outcome will be a success by chance.

Formally the reason that this approach cannot be used with individual binary data is that the number of parameters needed in the saturated model tends to infinity as the sample size increases, invalidating one of the conditions needed for the profile log-likelihood test to be valid.

Individual binary data will arise whenever one or more covariates are continuous, which is a very common situation. The most commonly used approach for assessing goodness

of fit in this setting was proposed by Hosmer and Lemeshow (1980, Communications in Statistics, A10, 1043-1069). The idea is that individuals are grouped according to their model predicted probabilties $\hat{\pi}_i$, such that we return to a situation which is similar to that with grouped binary data, i.e. where the number of groups does not increase with sample size. The mean of the predicted probabilities in each group is then compared to the observed proportion of 'successes' in the group.

More formally, suppose we group into g equal sized groups on the basis of their predicted probabilities. That is, with $g = 10$, the first group contains the individuals with the smallest 10% predicted probabilities, the second group contains the next 10%, etc. In the kth group, we let y_k denote the total number of 'successes', let $\bar{\pi}_k$ denote the average of the predicted probabilities in group k and let n_k denote the number of individuals in the kth group. We then form two $2 \times g$ contingency tables, the first of which contains the observed number of successes and failures by group.

Table 4.1: Table of observed data after grouping by predicted probabilities

	1	2	...	g
$Y = 0$	$n_1 - y_1$	$n_2 - y_2$	...	$n_g - y_g$
$Y = 1$	y_1	y_2	...	y_g

and the second which contains the predicted (expected under the null hypothesis that our model is correct) table:

Table 4.2: Table of expected data after grouping by predicted probabilities

	1	2	...	g
$Y = 0$	$n_1(1 - \bar{\pi}_1)$	$n_2(1 - \bar{\pi}_2)$	...	$n_g(1 - \bar{\pi}_g)$
$Y = 1$	$n_1\bar{\pi}_1$	$n_2\bar{\pi}_2$	...	$n_g\bar{\pi}_g$

We now calculate Pearson's chi-square statistic, which sums over the cells of the table, calculating $(observed - expected)^2/expected$. After a little bit of simplification this gives the following formula.

$$\hat{C} = \sum_{k=1}^{g} \frac{(y_k - n_k\bar{\pi}_k)^2}{n_k\bar{\pi}_k(1 - \bar{\pi}_k)}$$

Through simulation studies, Hosmer and Lemeshow demonstrated that $\hat{C}$ approximately follows a χ^2 distribution on $g - 2$ degrees of freedom under the null hypothesis that the model under consideration is correctly specified. The approximation is expected to perform well provided none of the expected counts is small.

In Stata, the Hosmer-Lemeshow test can be performed after fitting a logistic regression model, but the `logistic` command needs to be used, rather than `glm`. The following shows the test being performed after fitting a logistic regression model to Stata's low birth weight dataset.

```
. webuse lbw, clear

. logistic low age lwt i.race smoke ptl ht ui

** <output omitted> **

. estat gof, group(10) table

Logistic model for low, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0827 |     0 |   1.2 |    19 |  17.8 |    19 |
  |     2 | 0.1276 |     2 |   2.0 |    17 |  17.0 |    19 |
  |     3 | 0.2015 |     6 |   3.2 |    13 |  15.8 |    19 |
  |     4 | 0.2432 |     1 |   4.3 |    18 |  14.7 |    19 |
  |     5 | 0.2792 |     7 |   4.9 |    12 |  14.1 |    19 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.3138 |     7 |   5.6 |    12 |  13.4 |    19 |
  |     7 | 0.3872 |     6 |   6.5 |    13 |  12.5 |    19 |
  |     8 | 0.4828 |     7 |   8.2 |    12 |  10.8 |    19 |
  |     9 | 0.5941 |    10 |  10.3 |     9 |   8.7 |    19 |
  |    10 | 0.8391 |    13 |  12.8 |     5 |   5.2 |    18 |
  +--------------------------------------------------------+

      number of observations =       189
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =         9.65
                Prob > chi2 =         0.2904
```

The logistic regression model relates the probability of a baby being born underweight to various putative risk factors. After fitting the model using `logistic`, we use the `estat gof` command to perform the Hosmer-Lemeshow test. Here we specify 10 groups, and ask Stata to display the observed versus expected table.

The non-significant p-value shows that there is no evidence to reject the null hypothesis that the fitted model is correctly specified. However, we might be somewhat be concerned with the test's validity here, given that some of the expected counts are small. One approach to alleviating this is to perform the test with fewer groups.

The grouping approach taken in the Hosmer-Lemeshow test is often used informally to assess how well a model is calibrated (a topic we shall return to in a later session). Thus in papers reporting a new prognostic model, plots are often presented of observed against expected risk according to (say) deciles of predicted risk. See the following paper for an example: BMJ 2005;331:869.

An issue with using the Hosmer-Lemeshow test in practice is the choice of g, and it is important to be aware that, particularly with smaller datasets, it is possible to obtain quite different results (p-values) through use of different values of g. For this and other reasons

the Hosmer-Lemeshow test has been criticised in more recent literature. For example, Van Calster and colleagues (Calibration: the Achilles heel of predictive analytics. BMC Med 17, 230 (2019). https://doi.org/10.1186/s12916-019-1466-7) say "The commonly used Hosmer–Lemeshow test is often presented as a calibration test, though it has many drawbacks – it is based on artificially grouping patients into risk strata, gives a P value that is uninformative with respect to the type and extent of miscalibration, and suffers from low statistical power". They recommend using flexible calibration curves rather than the Hosmer-Lemeshow test. We will describe these when we return to this topic in later sessions.

4.6 Conclusions

In this session we have looked at how to compare nested GLMs and how to assess goodness of fit. Another route to assessing whether a model is adequate is to compare its fit to a more complex one, which say allows for an interaction between two covariates. If the models do not differ significantly in fit, this does not necessarily imply the simpler model is adequate, but it gives us additional evidence to support this hypothesis.

A drawback of the 'summary' goodness of fit tests we have described is that they give no indication of the cause of bad fit. A number of possibilities exist, including that the linear predictor is mis-specified, that we have used an incorrect link function, or that there is overdispersion, i.e. more variability than one would expect under the assumed model. In a subsequent session we will explore other methods, based on residuals, which can be used to help understand the causes of lack of fit.

4.7 Practical 4

Datasets required: `lbw (webuse) and insect.dta`

Introduction

In the first part of this session we will use a publicly available dataset on birthweights of babies, taken from the book *A*pplied Logistic Regression by David W Hosmer and Stanley Lemeshow.

> "Data were collected as part of a larger study at Baystate Medical Center in Springfield, Massachusetts. This dataset contains information on 189 births to women seen in the obstetrics clinic. Fifty nine of these births were defined as low birth weight [defined as less than 2500 grams]."
>
> ***H**osmer & Lemeshow, Applied Logistic Regression (second edition, 2000), Pub. by Wiley-Interscience*

In the second part of the session we will re-visit the insect dataset we used in Practical 3 and assess the goodness of fit of the logistic regression model we used then.

Aims

The aims of this session are:

1. understand how to compare logistic regression models (in Stata)
2. understand how to assess goodness of fit of logistic regression models (in Stata)

Part A: Investigation into low birthweights in Massachusetts, USA

We will now investigate factors which may be related to the chance of having a low birthweight baby, using the lbw dataset. This dataset is available directly from the Stata website, using the "webuse" command:

```
webuse lbw, clear
```

Explore the dataset, checking for missing values. Be sure to identify the outcome variable, and check how it is coded. Also note that the mother's weight is given in lbs, whilst the baby's weight is given in grams.

1. Fit a logistic model for the low birthweight variable with the mother's weight at last menstrual period (in lbs) as the single covariate. Run the model using the glm, logistic and logit commands, and compare the output.

   Discuss: What are the main differences between the commands and the output they produce?

2. To visually assess the reasonableness of the assumption of a linear lwt effect (on the logit scale), we can compare the fitted values from this model to a lowess smoother plot. First, we calculate the fitted probabilities, then construct the plot using the twoway command:

```
predict fitted_prob , pr

twoway (lowess low lwt) (scatter fitted_prob lwt) (scatter low lwt)
```

Discuss: What do you conclude regarding the appropriateness of the linearity assumption here?

3 Using pen & paper and a calculator, calculate the probability that a women who weighs 120lbs (about 54.4kg) has a baby with birth weight less than 2500g.

4 Use Stata to construct a confidence interval around your estimate from question 3. You will need to use either the lincom command, or re-parameterise the model to centre weight at 120lbs.

Using lincom:

```
lincom 120*lwt + _cons , eform
```

Re-parameterise the model:

```
gen lwt120 = lwt - 120

logit low lwt120, or
```

5 Fit an appropriate model (using logistic or logit), and perform an appropriate test, to examine whether race has an independent association with the probability of having a low birthweight baby, after adjusting for the mother's weight.

Discuss: What do you conclude?

6 When the logistic model is fitted using logistic or logit, a `LR chi2( )` and `Prob > chi2` output is given in the top right hand corner. To understand what this is, perform a profile log-likelihood ratio test comparing the model you fitted in the previous part to the null model. The null model is the model which contains no covariates. What can you deduce about the meaning of the `LR chi2( )` and `Prob > chi2` values in the model you fitted in the previous part? How can you interpret the result in this case?

7 Assess the goodness of fit of the model you fitted in question 5 using the Hosmer-Lemeshow goodness of fit test with ten groups.

```
estat gof, group(10)
```

Repeat the test with five groups, and compare the results. Try using groups of other sizes.

Discuss: What do you conclude about the goodness of fit of this model, and about the Hosmer-Lemeshow test?

Part B: Investigation into effect of CS_2 at different doses on insect survival

8 Reload the insect data used in the last practical, and use the `glm` command to fit the logistic model which assumes (on the logit scale) a linear effect of dose.

Based on the reported deviance, test whether there is any evidence that this model does not fit the data well. You will need to be make use of the `chi2tail` command, or otherwise your Neave tables.

Discuss: Why can the deviance be used to test the fit of this model, but not that of any of the models in Part A?

9 As we did in the last practical, add the square of the dose to the model, and use a profile log-likelihood ratio test to compare the fit of this model to the simpler one which assumes a linear effect. How does the result compare with the corresponding Wald test?

Session 5

Estimating treatment effects using observational data

5.1 Aims

This session is about the type of investigation in which the aim is to estimate causal effects of treatments or exposures on an outcome ('causality and explanation'). The aims of this session are to:

- Define different ways of quantifying a treatment effect on a binary outcome, including marginal effects (i.e. 'population average' effects) and conditional effects.
- Introduce the 'do' notation and its use in defining treatment effects.
- Show how we can estimate marginal and conditional treatment effects using observational data, when the treatment-outcome is confounded.
- Illustrate the use of standardization for estimating marginal treatment effects.

The primary focus is on binary outcomes, and a later section extends the discussion to continuous outcomes. The focus of this session is on observational studies in which the association between treatment and outcome is confounded, though randomized controlled trials are also mentioned.

5.2 Motivating example: treatment for kidney stones

The concepts and methods discussed in the rest of this session focus on the setting of a binary treatment X, binary outcome Y and binary confounder Z, and will be illustrated throughout using an example from an observational study comparing two treatments for kidney stones.

In this example,the two treatments are surgery and lithotripsy, the latter of which is less invasive. We let X denote the treatment variable, with $X = 0$ for surgery and $X = 1$ for lithotripsy. The outcome of interest is binary: whether the treatment was a success ($Y = 0$) or a failure ($Y = 1$). There is a third binary variable indicating the size of the kidney stone: small stone size ($<$ 2cm diameter) (denoted $Z = 0$) or large stone size ($\geq$ 2cm diameter) (denoted $Z = 1$).

In this observational setting subject matter knowledge tells us that stone size will influence the treatment that a doctor recommends for the patient, and also that larger stone size is related to an increased likelihood of a bad outcome. The assumed relationships between the three variables X, Z and Y are illustrated in Figure 5.1 using a *directed acyclic graph* (DAG) (or *causal diagram*). The DAG illustrates that stone size confounds the association between treatment and outcome.

The data from this study (i.e. on X, Z, Y) are summarised in Table 5.1. Some observations we can make from these data are:

- Patients with small stone size were more likely to receive lithotripsy ($X = 1$) than those with large stone size ($\Pr(X = 1|Z = 0) = 0.76$ vs $\Pr(X = 1|Z = 1) = 0.23$).
- The probability of failure ($Y = 1$) is slightly higher in those who received surgery compared to those who received lithotripsy ($\Pr(Y = 1|X = 0) = 0.22$ vs $\Pr(Y = 1|X = 1) = 0.17$). However, this could be due to a mixture of the effects of treatment and of the stone size, because the association between X and Y is confounded by Z.
- In those with small stone size ($Z = 0$), the probability of failure is higher in those who received lithotripsy ($\Pr(Y = 1|X = 1, Z = 0) = 0.13$) compared with surgery ($\Pr(Y = 1|X = 0, Z = 0) = 0.07$). Also, in those with large stone size ($Z = 1$) the probability of failure is higher in those who received lithotripsy ($\Pr(Y = 1|X = 1, Z = 1) = 0.31$) compared with surgery ($\Pr(Y = 1|X = 0, Z = 1) = 0.27$).

The probabilities discussed above actually estimates, and so strictly we should denote them as write them using 'hats', e.g. $\widehat{\Pr}(Y = 1|X = 0) = 0.22$. Because the probabilities are estimates they are subject to sampling error. For the purposes of explaining the concepts, and simplifying the notation, in much of this session we ignore sampling error.

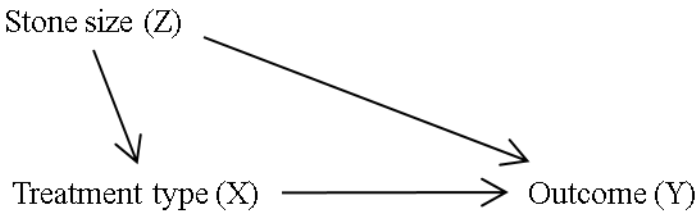


Figure 5.1: Directed acyclic graph (DAG) illustrating assumed relationships between stone size, treatment type and outcome.

Table 5.1: Data on type of treatment received (X), kidney stone size (Z) and binary outcome (Y).

	$(X = 0)$	$(X = 1)$	Total
Success $(Y = 0)$	273	289	562
Failure$(Y = 1)$	77	61	138
Total	350	350	700

	Small stone size $(Z = 0)$		Large stone size $(Z = 1)$	
	Surgery $(X = 0)$	Lithotripsy $(X = 1)$	Surgery $(X = 0)$	Lithotripsy $(X = 1)$
Success $(Y = 0)$	81	234	192	55
Failure$(Y = 1)$	6	36	71	25
Total	87	270	263	80

5.3 Defining a treatment effect (the 'estimand')

5.3.1 Marginal and conditional treatment effects

In the kidney stones example the effect of treatment type on the outcome is confounded by stone size, because smaller stone size is associated both with an increased likelihood of getting lithotripsy rather than surgery, and with greater chance of a good outcome. In this situation we know that to obtain an estimate of the causal effect of treatment type on the outcome we have to control for stone size.

But what exactly do we mean by the 'treatment effect'? There are different ways of defining or measuring the treatment effect. The quantity that we use to measure the treatment effect is called the 'estimand' - meaning the thing we aim to estimate. In the context of a binary outcome the estimand may be a risk difference, risk ratio or odds ratio. We also need to specify whether we are interested in a *marginal* treatment effect or a *conditional* treatment effect. In this section we define what we mean by these terms.

Questions about causal effects of treatments (or, more generally, exposures) are 'what if?' questions. In the kidney stones example, the causal effect of lithotripsy on the outcome could be thought of in terms of contrasting what the distribution of the outcome Y would have been under two hypothetical scenarios:

- *What if* all individuals had received lithotripsy? i.e. What would the distribution of Y have been if all individuals had had $X = 1$. vs
- *What if* all individuals had received surgery? i.e. What would the distribution of Y have been if all individuals had had $X = 0$.

This contrast is expressed in terms of 'what if everyone in this study had received $X = 1$ vs what if everyone had received $X = 0$'. Because the question is about 'everyone' this contrast is about a 'population average' treatment effect. This is called a *marginal treatment effect*.

We may also be interested in the effects of treatment in subsets of the population. For example, consider these hypothetical scenarios:

- *What if* all individuals with large stone size had received lithotripsy? i.e. In those with $Z = 1$, what would the distribution of Y have been if they had all had $X = 1$. vs
- *What if* all individuals with large stone size had received surgery? i.e. In those with $Z = 1$, what would the distribution of Y have been if they had all had $X = 0$.

This contrast is expressed in terms of 'what if people in this study with $Z = 1$ had received $X = 1$ vs what if people in this study with $Z = 1$ had received $X = 0$'. Because the question is about people with $Z = 1$ this is called a *conditional treatment effect'* (it is conditional on a particular value of Z). We could repeat this comparison for people with small stone size. Here we use the term *conditional treatment effect'* to refer to the treatment effect in a subset of study participants.

5.3.2 The 'do' notation

Above we described what we mean by a marginal or conditional treatment effect in words. We need some notation that also allows us to define the treatment effects using statistical

expressions. Consider the marginal treatment effect discussed above, which is defined by considering 'what if everyone in this study had received $X = 1$ vs what if everyone had received $X = 0$'. In real life we cannot observe every individual under both possible treatments ($X = 0, 1$). In the kidney stones example, a comparison of estimates of $\Pr(Y = 1|X = 1)$ and $\Pr(Y = 1|X = 0)$ does not provide an estimate of the treatment effect. This is because of the confounding by Z, which means that the groups of people with $X = 1$ and $X = 0$ have different distributions of Z, which also affects the outcome. In other words, differences that we see in the outcome between those who received lithotripsy and those who received surgery may not just be due to the different treatments they received, but also reflect differences in the distribution of kidney stone sizes in the two treatment groups. Suppose that we had instead performed a randomized controlled trial to investigate the treatment effect. We would then intervene on the treatment to be assigned (through the randomization), meaning that people in the $X = 0$ and $X = 1$ groups would have a similar distribution of Z. So in a randomized trial a comparison of estimates of $\Pr(Y = 1|X = 1)$ and $\Pr(Y = 1|X = 0)$ provides an estimate of the treatment effect.

So a comparison of estimates of $\Pr(Y = 1|X = 1)$ and $\Pr(Y = 1|X = 0)$ provides an estimate of the treatment effect if X was randomized, but not in an observational study in which the association between X and Y is confounded. But the definition of the treatment effect we wish to estimate should not depend on what type of study we are going to use to estimate it, or how we are going to estimate it in terms of the statistical analysis applied to the data we have. This is why it is useful (or essential, many would say) to use some new notation to define what we mean by a treatment effect.

The 'do' notation was introduced by Pearl (1995). Writing $do(X = 1)$ means that we imagine intervening on X to set it to 1. In the context of our kidney stones example, $\Pr(Y = 1|do(X = 1))$ is the probability of failure if everyone in the study population had been given lithotripsy, and $\Pr(Y = 1|do(X = 0))$ is the probability of failure if everyone in the study population had been given surgery. Pearl defines confounding as anything that makes $\Pr(Y = 1|do(X = x))$ differ from $\Pr(Y = 1|X = x)$. Typically when we write $\Pr(Y = 1|X = x)$ we mean $\Pr(Y = 1|see(X = x))$ rather than $\Pr(Y = 1|do(X = x))$, where by $see(X = x)$ we mean observing $X = x$ rather than intervening ('doing') to set $X = x$. Accessible introductions to the 'do' notation (or *do-operator*) and related concepts are given in the books by Pearl, Glymour and Jewell (2016) and Pearl & Mackenzie (2018).

Now that we have this notation, we can define the treatment effect precisely. There are different ways of comparing $\Pr(Y = 1|do(X = 1))$ and $\Pr(Y = 1|do(X = 0))$. The treatment effect can be defined in terms of the difference in the probability of the outcome (*risk difference*) under the two hypothetical interventions

$$\Pr(Y = 1|do(X = 1)) - \Pr(Y = 1|do(X = 0)). \tag{5.1}$$

This is sometimes referred to as the Average Causal Effect (ACE). Alternatively we could measure the treatment effect using a ratio of probabilities (*risk ratio*)

$$\frac{\Pr(Y = 1|do(X = 1))}{\Pr(Y = 1|do(X = 0))} \tag{5.2}$$

or using an *odds ratio*

$$\frac{\Pr(Y = 1|do(X = 1))\Pr(Y = 0|do(X = 0))}{\Pr(Y = 1|do(X = 0))\Pr(Y = 0|do(X = 1))}. \tag{5.3}$$

The treatment effect estimands in (5.1), (5.2) and (5.3) are 'marginal' treatment effects because they involve 'marginalized' or 'population averaged' probabilities, where the averaging is over the distribution of Z in the study population.

The question arises as to what population the marginal treatment effect refers to. It is the 'population average treatment effect', where the population is the study population. Therefore, in general, the marginal treatment effect depends on the distribution of Z in the study population because it is defined as averaged over the distribution of Z in the study population. In a different study with a different distribution of Z the marginal treatment effect would be different.

Conditional treatment effect estimands can also be expressed using the do notation:

$$\text{Conditional risk difference} \quad \Pr(Y=1|do(X=1),Z=z) - \Pr(Y=1|do(X=0),Z=z) \tag{5.4}$$

$$\text{Conditional risk ratio} \quad \frac{\Pr(Y=1|do(X=1),Z=z)}{\Pr(Y=1|do(X=0),Z=z)} \tag{5.5}$$

$$\text{Conditional odds ratio} \quad \frac{\Pr(Y=1|do(X=1),Z=z)/\Pr(Y=0|do(X=1),Z=z)}{\Pr(Y=1|do(X=0),Z=z)/\Pr(Y=0|do(X=0),Z=z)} \tag{5.6}$$

In this section we have focused on defining treatment effects. In the next section we move on to discussing how we can estimate the treatment effects.

5.4 Estimating treatment effects

5.4.1 Marginal treatment effects

In reality each individual can only receive one treatment at a given time, so we can never observe the outcome Y under the hypothetical situations in which all individuals in the population of interest received treatment $X = 0$ (e.g. lithotripsy) and in which all individuals received treatment $X = 1$ (e.g. surgery). As noted above, a randomized trial mimics this scenario through randomization, which makes the groups of individuals in the two treatment groups comparable in terms of their other characteristics. In a randomized trial we 'do' X on two comparable groups of individuals and so the probabilities $\Pr(Y = 1|do(X = x))$ $(x = 0, 1)$ can be estimated directly from the data on Y and X because $\Pr(Y = 1|do(X = x)) = \Pr(Y = 1|X = x)$ $(x = 0, 1)$. In our observational study $\Pr(Y = 1|do(X = x)) \neq \Pr(Y = 1|X = x)$ due to the confounding by Z. In section 5.4.3 we will explain how *marginal* treatment effects can be estimated using observational data when there is confounding. First, however, it is helpful to consider how we can estimate *conditional* treatment effects using observational data when there is confounding.

5.4.2 Conditional treatment effects

In the kidney stones example, Z is the only confounder, and therefore after conditioning on Z we have $\Pr(Y = 1|do(X = x), Z = z) = \Pr(Y = 1|X = x, Z = z)$. That is, within the $Z = 1$ (large stone size) group we can estimate what the probability of failure would have been if everyone had received lithotripsy $(X = 1)$ and what the probability of failure would have been if everyone had received surgery $(X = 0)$. We can do the same in the small stone size group. We say the treatment is conditionally randomized given Z. This means that conditional treatment effects can be estimated from the observational data using estimates of the probabilities $\Pr(Y = 1|X = 1, Z = z)$.

	Estimand	Estimate
Conditional risk differences	$\Pr(Y=1\|do(X=1),Z=0)-\Pr(Y=1\|do(X=0),Z=0)$ $\Pr(Y=1\|do(X=1),Z=1)-\Pr(Y=1\|do(X=0),Z=1)$	0.13-0.07=0.06 0.31-0.27=0.03
Conditional risk ratios	$\Pr(Y=1\|do(X=1),Z=0)/\Pr(Y=1\|do(X=0),Z=0)$ $\Pr(Y=1\|do(X=1),Z=1)/\Pr(Y=1\|do(X=0),Z=1)$	0.13/0.07=1.85 0.31/0.27=1.15
Conditional odds ratios	$\frac{\Pr(Y=1\|do(X=1),Z=0)/\Pr(Y=0\|do(X=1),Z=0)}{\Pr(Y=1\|do(X=0),Z=0)/\Pr(Y=0\|do(X=0),Z=0)}$ $\frac{\Pr(Y=1\|do(X=1),Z=1)/\Pr(Y=0\|do(X=1),Z=1)}{\Pr(Y=1\|do(X=0),Z=1)/\Pr(Y=0\|do(X=0),Z=1)}$	$\frac{0.13/(1-0.13)}{0.07/(1-0.07)}=1.99$ $\frac{0.31/(1-0.31)}{0.27/(1-0.27)}=1.21$

Table 5.2: Effect of lithotripsy on outcome conditional on stone size: Estimates of the conditional risk difference, conditional risk ratio, and conditional odds ratio obtained using the kidney stones data

From Table 5.1 we have the estimates $\widehat{\Pr}(Y=1|X=1,Z=0)=0.13$, $\widehat{\Pr}(Y=1|X=0,Z=0)=0.07$, $\widehat{\Pr}(Y=1|X=1,Z=1)=0.31$, $\widehat{\Pr}(Y=1|X=0,Z=1)=0.27$. Estimates of the conditional treatment effects are given in Table 5.2. The treatment effect estimates suggest a higher risk (or odds) of failure under lithotripsy compared with surgery for individuals with both small and large stone sizes. Estimates of all three estimands (the conditional risk difference, risk ratio and odds ratio) are all higher for individuals with small stone compared to those with large stone size. This could be because stone size modifies the treatment effect, or it could be due to random variation. Note that the results in Table 5.2 are just the point estimates. They should be accompanied by estimates of their uncertainty (e.g. 95% confidence intervals) - see Section 5.8. We could also perform tests of the hypothesis that the treatment effect is the same (on some scale) in those with small and large stone size.

5.4.3 Marginal treatment effects revisited: standardization

We have seen above how conditional treatment effects can be estimated from the observational data where there is confounding (and when we have measured the confounding variable(s)). Now let's consider how we could estimate a marginal treatment effect, as defined in equations (5.1), (5.2) and (5.3). As noted above, the probabilities $\Pr(Y=1|do(X=x))$ do not correspond to the probabilities $\Pr(Y=1|X=x)$ that refer to the observational data, due to the confounding by Z.

Using the law of total probability, we can write $\Pr(Y=1|do(X=x))$ as

$$\Pr(Y=1|do(X=x))=\sum_{z=0,1}\Pr(Y=1|do(X=x),Z=z)\Pr(Z=z|do(X=x)). \quad (5.7)$$

In the previous section we argued that the conditional probabilities $\Pr(Y=1|do(X=x),Z=z)$ can be estimated easily from the observational data because $\Pr(Y=1|do(X=x),Z=z)=\Pr(Y=1|X=x,Z=z)$, under the assumption that Z is the only confounder. Now let's think about the second term in the sum on the right-hand-side of (5.7), $\Pr(Z=z|do(X=x))$. It helps to look at the DAG in Figure 5.1 for this part. If we intervene on X, as implied by $do(X=x)$, then this has no impact on Z because Z comes before X, i.e. it is not downstream from X. It follows that $\Pr(Z=z|do(X=x))=\Pr(Z=z)$. Therefore, under the assumption that Z is the only confounder of the

association between X and Y, we can write (5.7) as

$$\Pr(Y = 1|do(X = x)) = \sum_{z=0,1} \Pr(Y = 1|X = x, Z = z) \Pr(Z = z). \tag{5.8}$$

The probabilities $\Pr(Y = 1|X = x, Z = z)$ and $\Pr(Z = z)$ can be estimated from the observational data and hence we can estimate the marginal treatment effects. This technique of expressing $\Pr(Y = 1|do(X = x))$ in terms of the conditional probabilities $\Pr(Y = 1|X = x, Z = z)$ and $\Pr(Z = z)$ is called 'standardization'. It is an example of the use of weighted averaging. Standardization is also referred to as 'marginalizing' or 'averaging' over Z and it has a long history of use in epidemiology (e.g. see Lash, Vanderweele, Haneuse, Rothman 2021).

The estimate of $\Pr(Y = 1|do(X = x))$ obtained using standardization is *adjusted for* Z, but it is not *conditional on* Z. When we say that an effect of X is adjusted for Z (or standardized for Z) there is no assumption that this effect applies at an particular value for Z. In contrast, conditional effects apply at particular values of Z, or at all values of Z if it is assumed that the effect of X is not modified by Z (treatment effect homogeneity). Sometimes the terms 'adjusted for' and 'conditional on' are used interchangeably, but this is imprecise. We should take care to be clear about which we mean, i.e whether we are discussing a marginal effect (where the estimation has involved adjusting for Z) or an effect that is conditional on Z.

Let's apply this approach to the kidney stones data. We replace $\Pr(Y = 1|X = x, Z = z)$ and $\Pr(Z = z)$ in (5.8) by their estimates from the data. This gives

$$\begin{aligned}\widehat{\Pr}(Y = 1|do(X = 1)) &= \sum_{z=0,1} \widehat{\Pr}(Y = 1|X = 1, Z = z)\widehat{\Pr}(Z = z) \\ &= \frac{36}{270} \times \frac{270 + 87}{700} + \frac{25}{80} \times \frac{263 + 80}{700} = 0.2211\end{aligned} \tag{5.9}$$

and

$$\begin{aligned}\widehat{\Pr}(Y = 1|do(X = 0)) &= \sum_{z=0,1} \widehat{\Pr}(Y = 1|X = 0, Z = z)\widehat{\Pr}(Z = z) \\ &= \frac{6}{87} \times \frac{270 + 87}{700} + \frac{71}{263} \times \frac{263 + 80}{700} = 0.1675\end{aligned} \tag{5.10}$$

The resulting estimates of the marginal risk difference, risk ratio and odds ratio are given in Table 5.3. The marginal treatment effect estimates are all in the direction of a higher risk of failure if a patient in this population receives lithotripsy, compared with if they receive surgery. As before, these are the point estimates and they should be accompanied by information about their uncertainty.

5.5 Estimating the treatment effect using logistic regression

So far we have not used any regression models to estimate treatment effects, as they were not required for our simple setting. When we move to more complex settings, such as when there are several confounders to control for, a regression modelling approach is typically needed. For our setting of a binary outcome, the probabilities involved in estimating treatment effects can be estimated using logistic regression.

	Estimand	Estimate
Marginal risk difference	$\Pr(Y=1\|do(X=1)) - \Pr(Y=1\|do(X=0))$	0.0.2211-0.1675=0.05
Marginal risk ratio	$\Pr(Y=1\|do(X=1))/\Pr(Y=1\|do(X=0))$	0.2211/0.1675=1.32
Marginal odds ratio	$\frac{\Pr(Y=1\|do(X=1))/\Pr(Y=0\|do(X=1))}{\Pr(Y=1\|do(X=0))/\Pr(Y=0\|do(X=0))}$	$\frac{0.2211/(1-0.2211)}{0.1675/(1-0.1675)} = 1.41$

Table 5.3: Effect of lithotripsy on outcome: Estimates of the marginal risk difference, marginal risk ratio, and marginal odds ratio obtained using the kidney stones data

The conditional probabilities $\Pr(Y=1|X=x,Z=z)$ can be estimated using the logistic regression model

$$\Pr(Y=1|X=x,Z=z) = \frac{\exp(\beta_0+\beta_X X+\beta_Z Z+\beta_{XZ}XZ)}{1+\exp(\beta_0+\beta_X X+\beta_Z Z+\beta_{XZ}XZ)}. \tag{5.11}$$

The conditional odds ratios are given by combinations of e^{β_X} and β_{XZ}. When Z is binary then the conditional odds ratio given $Z=0$ is e^{β_X} and the conditional odds ratio given $Z=1$ is $e^{\beta_X+\beta_{XZ}}$. In Stata, after fitting a logistic regression model using 'glm' one can obtain estimates of the probabilities $\Pr(Y=1|X=x,Z=z)$ using 'predict'. Conditional risk differences and risk ratios can then be estimated using the resulting conditional probabilities. To obtain marginal treatment effect estimates we could use the formula in (5.8). The 'margins' command in Stata can be used to obtain marginal treatment effect estimates. We will try this out in the practical.

The above logistic regression model includes an interaction term between X and Z. A model without the interaction term could be used, which would make the assumption that the conditional odds ratio for X given Z is not modified by Z. In our example, this would be the assumption that the odds ratio for lithotripsy conditional on stone size is the same in patients with large stone size and small stone size. The model would then be of the form:

$$\Pr(Y=1|X=x,Z=z) = \frac{\exp(\beta_0+\beta_X X+\beta_Z Z)}{1+\exp(\beta_0+\beta_X X+\beta_Z Z)}. \tag{5.12}$$

If our treatment effect estimand of interest is a conditional odds ratio, and we assume no $X \times Z$ interaction, then the conditional odds ratio is e^{β_X}.

When are our estimands of interest are conditional odds ratios, the estimate of our estimand of interest corresponds directly parameters in our logistic regression model (β_X or $e^{\beta_X+\beta_{XZ}}$). However, this is not true in general. If our estimand of interest is marginal, then further steps are needed after fitting the logistic regression model to get to the estimate of interest. Also, if our estimand is a conditional risk difference or risk ratio then then further steps are needed after fitting the logistic regression model to get to the estimate of interest.

5.6 Extensions to continuous and multiple confounders

Up to now in this session we have focused on a single binary confounder Z. In most observational settings there are several potential confounders that we wish to control for. The methods described above extend to settings with multiple confounders, and with continuous confounders.

Standardization extends to more than one variable that we wish to average over. For example, with two binary confounders Z_1 and Z_2 the standardization formula in 5.8

becomes

$$\Pr(Y=1|do(X=x)) = \sum_{z_1=0,1}\sum_{z_2=0,1} \Pr(Y=1|X=x,Z_1=z_1,Z_2=z_2)\Pr(Z_1=z_1,Z_2=z_2) \tag{5.13}$$

Suppose that Z were continuous instead of binary, then the standardization formula requires an integral rather than a sum:

$$\Pr(Y=1|do(X=x)) = \int \Pr(Y=1|X=x,Z=z)f(z)dz \tag{5.14}$$

where $f(z)$ denotes the probability density function for Z. If Z is continuous then a regression model will typically be required to estimate the conditional expectations $\Pr(Y=1|X=x,Z=z)$. To perform the standardization requires an assumption about the distribution of Z. For example it might be assumed that Z is normally distributed. If there are multiple Z variables to control for then the standardization requires specifying their joint distribution. In general, it would be difficult to specify this distribution, and mis-specifying it would result in a biased estimate of the marginal treatment effect. An alternative approach, which avoids having to assume a particular form for the joint distribution of the confounders, is to use *empirical standardization*, which we now describe.

In empirical standardization the sum or integral in the standardization formula is replaced by a sum over individuals. Suppose our study has n individuals. Using the empirical standardization formula, the marginal treatment effect can be estimated using

$$\widehat{\Pr}(Y=1|do(X=x)) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\Pr}(Y=1|X=x,Z=z_i). \tag{5.15}$$

Consider $X=1$. Using a logistic regression model (or otherwise) we can estimate conditional probability $\Pr(Y=1|X=1,Z=z_i)$ for each individual $i=1,\ldots,n$ in the study population using their value of Z, z_i, and setting their treatment to 1 (regardless of their observed treatment). We then take the average of these estimates across the n individuals - this is $\frac{1}{n}\sum_{i=1}^{n}\widehat{\Pr}(Y=1|X=x,Z=z_i)$. We can do the same for $X=0$. This empirical standardization approach extends to the situation in which Z is a vector of confounders to be controlled for.

5.7 Extension to a continuous outcome

Above we focused on a binary outcome Y. In this section we outline the extension of the above concepts and methods to the setting of a continuous outcome, again denoted Y. Consider a binary treatment X and single binary confounder Z, and assume the relationships as depicted in the DAG in Figure 5.1. In the context of the kidney stones example, we can imagine that the binary outcome (success/failure) is replaced by a continuous outcome such as as measure of kidney function.

When the outcome is continuous, the treatment effect is measured using a difference in means. Using the 'do' notation as above, the marginal treatment effect is defined as:

$$E(Y|do(X=1)) - E(Y|do(X=0)). \tag{5.16}$$

The conditional treatment effect (conditional on Z) is defined as

$$E(Y|do(X=1), Z=z) - E(Y|do(X=0), Z=z). \tag{5.17}$$

The expectations conditional on Z can, as above, be estimated from the observational data using the result that $E(Y|do(X=x), Z=z) = E(Y|X=x, Z=z)$ if Z is the only confounder. The marginal probabilities can again be estimated using standardization:

$$E(Y|do(X=x)) = \sum_{z=0,1} E(Y|X=x, Z=z) \Pr(Z=z). \tag{5.18}$$

In simple situations such as when Z is a single binary confounder, the conditional expectations $E(Y|X=x, Z=z)$ can be estimated from the data without the need for any regression model. In general, however, the conditional expectations $E(Y|X=x, Z=z)$ will need to be estimated using a linear regression model. The empirical standardization method outlined in section 5.6 also extends directly to this setting of a continuous outcome.

Consider the following linear regression model for continuous outcome Y:

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \epsilon. \tag{5.19}$$

The parameter β_X corresponds to the conditional mean difference $E(Y|X=1, Z=z) - E(Y|X=0, Z=z)$. If there is no interaction between X and Z in the linear regression model then the conditional mean difference $E(Y|X=1, Z=z) - E(Y|X=0, Z=z)$ is equal to the marginal mean difference $E(Y|do(X=1)) - E(Y|do(X=0))$. This is assuming that Z is the only confounder. This is a special feature of linear models, and does not extend to logistic regression models.

In the practical you will try out these methods for a continuous outcome.

5.8 Estimates of uncertainty

When we applied the methods above to the kidney stones data, we focused on point estimates of treatment effects. As mentioned above, these should be accompanied by estimates of the degree of uncertainty in the estimate, e.g. using a 95% confidence interval.

Confidence intervals for treatment effect estimates can be estimated analytically in some cases, for example when the conditional treatment effect corresponds to a parameter in a linear or logistic regression model. In general, however, an alternative approach is needed. To obtain estimated 95% CIs for the marginal risk difference, for example, the variance transformation method could be used - this is covered in Analytical Techniques in Term 1 (also sometimes called the 'delta method'). An alternative is to use bootstrapping (introduced in Term 1 in Robust Methods), which is easier and does not require an approximation. The 'margins' command in Stata provides estimates of 95% CIs for risk differences (but not other estimands, to our knowledge), which it obtains using bootstrapping.

5.9 Concluding remarks

This session has focused on a binary treatment or exposure. The concepts and methods extend to other types of exposure, for a example a continuous exposure (e.g. dose). In

that case the treatment effect is defined in terms of a contrast in the outcome between two levels of X. We focused on a medical treatment in the example, but we could use the same methods to investigate effects of 'lifestyle' exposures, for example relating to dietary intake or physical activity level. Exposures considered in causal investigations should be well defined. There has been much debate over whether it makes sense to estimate causal effects of such things as sex or ethnicity, since these are things that cannot be different for an individual (depending on how they are defined). See Hernán (2016) for a discussion on this topic.

Assumptions about the inter-relationships between all of the variables at play in a given study are key for informing which variables should be adjusted for in an analysis when the aim is to estimate the effect of a particular treatment or exposure. DAGs are helpful in setting out assumptions about relationships between variables, and in particular their temporal ordering, and can be used to establish which variables need to be controlled for to estimate certain effects. Tennant et al. (2021) give recommendations for use of DAGs in applied health research.

We have focused on simple settings in this session to illustrate the main points. The book by Pearl, Glymour and Jewell (2016) is an excellent source of additional detail and extensions to more complex settings.

References

Hernán M.A. Does water kill? A call for less casual causal inferences. Annals of Epidemiology 2016; 26: 674-680.

Lash T.L., Vanderweele T.J., Haneuse S., Rothman K., Greenland S., *M*odern Epidemiology. 3rd Edition. 2008. Lippincott Williams & Wilkins.

Pearl J. Causal diagrams for empirical research. Biometrika 1995; 82:669–710.

Pearl J, Glymour M., Jewell N.P. *Causal Inference in Statistics: A Primer*. 2016. Wiley.

Pearl J., Mackenzie D. *The Book of Why*. 2018. Penguin.

Tennant P.W.G., Murray E.J., Arnold K.F., et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. International Journal of Epidemiology 2021; 50: 620–632.

5.10 Practical 5

Datasets required
kidney_example.csv
kidney_example_continuousY.csv

Introduction

This practical is in two parts. Part A uses the kidney stones example data from the Notes, in which the outcome Y is binary. Part B considers data on a similar study but with a continuous outcome Y. The other variables are the same: X denotes a binary treatment and Z a binary confounder. Both data sets include 700 individuals.

Variable	Description
`y`	Binary: 0 = success; 1 = failure
	Continuous: a score on a quality of life questionnaire
`x`	Treatment. 0 = Surgery; 1 = Lithotripsy
`z`	Size of kidney stone. 0 = small; 1 = large

Aims

1 To estimate conditional treatment effects for both binary and continuous outcomes.

2 To estimate marginal treatment effects for both binary and continuous outcomes using standardization.

3 To interpret estimates of marginal and conditional treatment effects.

Part A: Binary outcome

The data for this section are in the file "kidney_example.csv". You can load the dataset into Stata using: `import delimited kidney_example.csv , clear`

The data are summarised in the table below.

	$(X=0)$	$(X=1)$	Total
Success $(Y=0)$	273	289	562
Failure$(Y=1)$	77	61	138
Total	350	350	700

	Small stone size $(Z=0)$		Large stone size $(Z=1)$	
	Surgery $(X=0)$	Lithotripsy $(X=1)$	Surgery $(X=0)$	Lithotripsy $(X=1)$
Success $(Y=0)$	81	234	192	55
Failure$(Y=1)$	6	36	71	25
Total	87	270	263	80

1 In this question we will use logistic regression to obtain an estimate of the conditional odds ratio

$$\frac{\Pr(Y=1|do(X=1),Z=z)/\Pr(Y=0|do(X=1),Z=z)}{\Pr(Y=1|do(X=0),Z=z)/\Pr(Y=0|do(X=0),Z=z)}$$

for $Z = 0, 1$.

i) First using a logistic regression model for Y with X and Z as explanatory variables (not including their interaction).

ii) Using a logistic regression model for Y with X, Z and their interaction $X \times Z$ as explanatory variables. You might find the `lincom` command useful for this.

2 Using your model from Question 1(ii) obtain estimates of:

i) The conditional probabilities $\Pr(Y = 1|do(X = 1), Z = z)$ and $\Pr(Y = 1|do(X = 0), Z = z)$ for $Z = 0, 1$. After fitting the logistic regression model the 'predict' command can be used to obtain estimates of the $\Pr(Y = 1|X = x, Z = z)$ components.

```
predict ypred
```

The estimates of $\Pr(Y = 1|X = 0, Z = 0)$ (for example) can be seen using

```
list ypred if x==0 & z==0
```

ii) The conditional risk differences

$$\Pr(Y = 1|do(X = 1), Z = z) - \Pr(Y = 1|do(X = 0), Z = z), \qquad Z = 0, 1.$$

iii) The conditional risk ratios

$$\frac{\Pr(Y = 1|do(X = 1), Z = z)}{\Pr(Y = 1|do(X = 0), Z = z)}, \qquad Z = 0, 1.$$

Discuss: Compare your results from the logistic regression models with those given in Table 5.2 in the Notes (which were not obtained using regression). Interpret your conditional treatment effect estimates obtained in questions 1 and 2. What assumption do we make to interpret these estimates 'causally'.

3 We will next estimate marginal treatment effects.

i) Using the results from your model in Question 1(ii), and information about $\Pr(Z = z)$, use the standardization formula in (5.8) to estimate $\Pr(Y = 1|do(X = 1))$ and $\Pr(Y = 1|do(X = 0))$. The standardization formula is:

$$\Pr(Y = 1|do(X = x)) = \sum_{z=0,1} \Pr(Y = 1|X = x, Z = z) \Pr(Z = z)$$

Note that the probabilities of having small and large kidney stone size are $\Pr(Z = 0) = 0.51$ and $\Pr(Z = 1) = 0.49$ (check you agree!).

You may find it useful to use the table below to keep track of your estimates.

ii) Using your results from (i) obtain estimates of the marginal risk difference, marginal risk ratio, and marginal odds ratio, as defined in equations (5.1)-(5.3) in the Notes.

$\Pr(Y=1\|do(X=1))$	$\Pr(Y=1\|X=1,Z=0)$	
	$\Pr(Y=1\|X=1,Z=1)$	
	$\Pr(Y=1\|X=1,Z=0)\times\Pr(Z=0)$	
	$\Pr(Y=1\|X=1,Z=1)\times\Pr(Z=1)$	
	$\Pr(Y=1\|X=1,Z=0)\times\Pr(Z=0)$ $+\Pr(Y=1\|X=1,Z=1)\times\Pr(Z=1)$	
$\Pr(Y=1\|do(X=0))$	$\Pr(Y=1\|X=0,Z=0)$	
	$\Pr(Y=1\|X=0,Z=1)$	
	$\Pr(Y=1\|X=0,Z=0)\times\Pr(Z=0)$	
	$\Pr(Y=1\|X=0,Z=1)\times\Pr(Z=1)$	
	$\Pr(Y=1\|X=0,Z=0)\times\Pr(Z=0)$ $+\Pr(Y=1\|X=0,Z=1)\times\Pr(Z=1)$	

Discuss: What are the interpretations of the causal marginal treatment effects you have estimated? Compare these with the conditional treatment effects estimated in questions 1 and 2.

4 We can also use the 'margins' postestimation command in Stata to get to the treatment effect estimates more directly.

i) After running the logistic regression model of Y on $X, Z, X \times Z$ run the following command
`margins x`

This provides estimates of $\Pr(Y=1|do(X=1))$ and $\Pr(Y=1|do(X=0))$ using the same standardization procedure that you used above.

ii) Estimates of the marginal risk difference can be obtained using

```
margins x, dydx(x)
```

iii) The results from the 'margins' command provide an estimate of the 95% confidence interval for the marginal risk difference. Interpret the point estimate and its 95% confidence interval.

5 Lastly, we will use the empirical standardization method to estimate the marginal probabilities $\Pr(Y=1|do(X=1))$ and $\Pr(Y=1|do(X=0))$. This was outlined in section 5.6 of the Notes (see equation (5.15)). The Stata code for this is a bit tricky so please follow the steps outlined in the accompanying Stata Do file, and make sure you understand what is being done in each step. Check that you get the same estimates as found in Questions 3 and 4.

EXTRA: Extra Stata code is provided to show how we can obtain bootstrap estimates of 95% confidence intervals for treatment effect estimates. Those who have done the Robust Methods module may wish to look through this as an extra exercise.

Part B: Continuous outcome

In the second part of this practical we will use a modified version of the kidney study data with a continuous outcome Y instead of a binary outcome. For example, Y could be a score on a quality of life questionnaire. The data are available in the file kidney_example_continuousY.csv. Load the data.

6 What are the means of Y in each of the four groups defined by X and Z? That is, what are $E(Y|X=x,Z=z)$ for $x,z=0,1$?

7 Fit a linear regression of Y on X, Z and their interaction $X \times Z$. How do the coefficients relate to the means of each group?

8 Use your results from question 4 and 5 to estimate the conditional effect of X on Y given Z for $Z = 0, 1$, using a conditional mean difference:

$E(Y|do(X = 1), Z = 0) - E(Y|do(X = 0), Z = 0)$
$E(Y|do(X = 1), Z = 1) - E(Y|do(X = 0), Z = 1)$

Discuss: Interpret the estimates. What assumption do you make to interpret the conditional mean differences as causal treatment effect estimates?

9 We now estimate the causal marginal treatment effect.

i) By hand, use standardization to complete the table below, and hence to estimate the causal marginal treatment effect:

$$E(Y|do(X = 1)) - E(Y|do(X = 0))$$

Reminder, by standardization $E(Y|do(X = x))$ is given by

$$E\left(Y|X = x, Z = 0\right) \times Pr\left(Z = 0\right) + E\left(Y|X = x, Z = 1\right) \times Pr\left(Z = 1\right)$$

$E(Y\|do(X = 1))$	$E(Y\|X = 1, Z = 0)$	
	$\Pr(Z = 0)$	0.51
	$E(Y\|X = 1, Z = 1)$	
	$\Pr(Z = 1)$	0.49
$E(Y\|do(X = 1))$		
$E(Y\|do(X = 0))$	$E(Y\|X = 0, Z = 0)$	
	$\Pr(Z = 0)$	0.51
	$E(Y\|X = 0, Z = 1)$	
	$\Pr(Z = 1)$	0.49
$E(Y\|do(X = 0))$		
$E(Y\|do(X = 1)) - E(Y\|do(X = 0))$		

ii) You can also obtain the same estimate using the 'margins' command after fitting your linear regression model as follows:

```
regress y i.x##i.z

margins x, dydx(x)
```

10 Use the empirical standardization approach described in the notes (section 5.6, equation 5.15), and used above in Question 5, to estimate the marginal treatment effect. You can do this by modifying the Stata code used for Question 5.

Discuss: Interpret the marginal treatment effect estimate.

11 Fit a linear regression of Y on X and Z, without an interaction term between X and Z. What is the conditional mean difference $E(Y|do(X = 1), Z = z) - E(Y|do(X = 0), Z = z)$ for $Z = 0, 1$? Compare this with the result from using

```
regress y i.x i.z
margins x, dydx(x)
```

Session 6

Collapsibility and non-collapsibility

6.1 Aims

As in session 5, the focus of this session is on estimating treatment effects, i.e. on 'causality'. This session concentrates on a particular property of effect estimands called 'collapsibility'. The aims are to:

- Introduce the concept of collapsibility and non-collapsibility.
- Discuss measures of association that are collapsible (e.g. mean differences, risk ratios) and measures of association that are non-collapsible (e.g. odds ratios).
- Discuss the implications of non-collapsibility for interpretation of regression coefficients in linear and logistic regression analyses.

6.2 Introduction to collapsibility

Collapsibility is about the relation between marginal and conditional effects. We are therefore going to consider the marginal and conditional treatment effects as defined in session 5. As in session 5 we focus on a binary treatment X, binary outcome Y and binary covariate Z. In section 6.8 we will consider continuous outcomes.

For a binary outcome the treatment effect can be quantified using a risk difference, a risk ratio, or an odds ratio. Treatment effects can be expressed formally using the 'do' notation. As a reminder, the marginal risk difference (RD), risk ratio (RR), and odds ratio (OR) are defined respectively as

$$\begin{aligned} RD &= \Pr(Y=1|do(X=1)) - \Pr(Y=1|do(X=0)) && (6.1)\\ RR &= \Pr(Y=1|do(X=1))/\Pr(Y=1|do(X=0)) && (6.2)\\ OR &= \frac{\Pr(Y=1|do(X=1))/\Pr(Y=0|do(X=1))}{\Pr(Y=1|do(X=0))/\Pr(Y=0|do(X=0))} && (6.3) \end{aligned}$$

The conditional risk difference, risk ratio, and odds ratio (conditional on $Z=z$) are defined respectively as

$$\begin{aligned} RD_z &= \Pr(Y=1|do(X=1),Z=z) - \Pr(Y=1|do(X=0),Z=z) && (6.4)\\ RR_z &= \Pr(Y=1|do(X=1),Z=z)/\Pr(Y=1|do(X=0),Z=z) && (6.5)\\ OR_z &= \frac{\Pr(Y=1|do(X=1),Z=z)/\Pr(Y=0|do(X=1),Z=z)}{\Pr(Y=1|do(X=0),Z=z)/\Pr(Y=0|do(X=0),Z=z)} && (6.6) \end{aligned}$$

Different authors have defined collapsibility in slightly different ways. We focus on defining collapsibility in terms of the relation between conditional and marginal treatment effect

measures (e.g. as in Huitfeldt et al. 2019). An effect measure is described as *collapsible* if the marginal treatment effect can be expressed as a weighted average of conditional treatment effects. Risk differences and risk ratios are *collapsible*. But odds ratios are *non-collapsible*.

The relation between the conditional risk differences and the marginal risk difference is

$$\begin{aligned}&\Pr(Y=1|do(X=1)) - \Pr(Y=1|do(X=0)) = \\ &\sum_{z=0,1} \{\Pr(Y=1|do(X=1), Z=z) - \Pr(Y=1|do(X=0), Z=z)\} \Pr(Z=z).\end{aligned} \tag{6.7}$$

That is, the marginal risk difference is a weighted average of the conditional risk differences given $Z = 0, 1$, with weights $\Pr(Z = 0)$ and $\Pr(Z = 1)$.

The relation between the conditional risk ratios and the marginal risk ratio is

$$\frac{\Pr(Y=1|do(X=1))}{\Pr(Y=1|do(X=0))} = \sum_{z=0,1} \frac{\Pr(Y=1|do(X=1), Z=z) w_z}{\Pr(Y=1|do(X=0), Z=z)} \tag{6.8}$$

where $w_z = \Pr(Y = 1|do(X = 0), Z = z) \Pr(Z = z) / \Pr(Y = 1|do(X = 0))$. That is, the marginal risk ratio is a weighted average of the conditional risk ratios given $Z = 0, 1$, with weights $w_0 = \Pr(Y = 1|do(X = 0), Z = 0) \Pr(Z = 0) / \Pr(Y = 1|do(X = 0))$ and $w_1 = \Pr(Y = 1|do(X = 0), Z = 1) \Pr(Z = 1) / \Pr(Y = 1|do(X = 0))$.

As we will shortly see, there is no way of writing the marginal odds ratio as a weighted average of the conditional odds ratios.

Below we will give some examples, and discuss why the property of collapsibility is important.

6.3 Example 1: No confounding by Z

Suppose we have some data in which the underlying relationships between X, Y, Z are as depicted in the DAG in Figure 6.1, where X represents a treatment (or, more generally, an exposure), Y represents an outcome of interest, and Z is a covariate. Compared with the DAG that we considered in Chapter 5 (Figure 5.1) the arrow from Z to X has been removed, indicating an assumption that Z does not affect X. The covariate Z still affects the outcome Y. In other words, in Figure 6.1 Z does not confound the association between X and Y. A situation such as this would arise if X is a randomized treatment in a trial, for example.

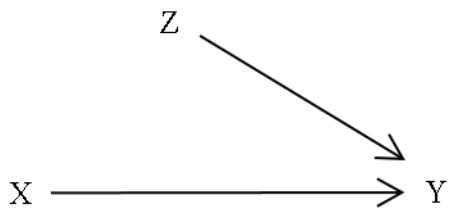


Figure 6.1: Example 1: Directed acyclic graph (DAG) illustrating relationships between three variables.

In this example we consider the data given in Table 6.1, on Z, X, Y (all binary) for 20 individuals. The data were generated according to the relationships depicted in the DAG in Figure 6.1. From the table we have $\Pr(X = 1|Z = 1) = 0.5$ and $\Pr(X = 1|Z = 0) = 0.5$, meaning that X and Z are (marginally) independent, as in the DAG. For the purposes of explaining the concepts we shall ignore sampling variation for now, and omit the 'hats' from estimated probabilities.

Table 6.1: Example 1: Data on covariate Z, treatment X and outcome Y for 20 individuals.

	$Z=0$		$Z=1$		Total
	$X=0$	$X=1$	$X=0$	$X=1$	
$Y=0$	3	1	4	2	10
$Y=1$	2	4	1	3	10
Total	5	5	5	5	20

Let's estimate the marginal and conditional treatment effects using the data in the table. Under the assumption that there is no confounding of the association between X and Y, we have $\Pr(Y = 1|do(X = x)) = \Pr(Y = 1|X = x)$ and $\Pr(Y = 1|do(X = x), Z = z) = \Pr(Y = 1|X = x, Z = z)$. Hence we can estimate the marginal and conditional treatment effects easily using data in which the underlying relationships are generated under the DAG in Figure 6.1. Unlike in Session 5, where Z was a confounder, we do not need to use standardization to estimate the marginal treatment effects.

Table 6.2 shows the marginal and conditional RDs, RRs, and ORs estimated using the data in Table 6.1. These make use of the following probabilities:

$$\Pr(Y = 1|X = 0) = \frac{1+2}{5+5} = 0.3, \quad \Pr(Y = 1|X = 1) = \frac{3+4}{5+5} = 0.7$$

$$\Pr(Y = 1|X = 0, Z = 0) = 0.4, \quad \Pr(Y = 1|X = 1, Z = 0) = 0.8$$

$$\Pr(Y = 1|X = 0, Z = 1) = 0.2, \quad \Pr(Y = 1|X = 1, Z = 1) = 0.6$$

First consider the risk differences (RD). The conditional risk differences given $Z = 0$ and $Z = 1$ are both equal to 0.4. The marginal risk difference is also equal to 0.4. Risk differences are 'collapsible'. Note that we are in a simple situation in which Z does not modify the effect of X on Y, because the conditional RDs are the same for $Z = 0$ and $Z = 1$. In this case collapsibility means that the marginal RD is equal to the two conditional RDs.

Next consider the risk ratios (RR). The conditional RRs are different for $Z = 0$ and $Z = 1$ ($0.8/0.4 = 2$ and $0.6/0.2 = 3$ respectively), and the marginal RR is $0.7/0.3 = 2.33$. Due to the effect modification we do not expect the marginal RR to be equal to either of the conditional RRs. However, the marginal RR is in between the two conditional RRs - the marginal RR of 2.33 is a weighted average of the conditional RRs 2 and 3. The weights are given by w_0 and w_1 as defined in section 6.2. Here we have

$$w_0 = \frac{\Pr(Y = 1|do(X = 0), Z = 0)\Pr(Z = 0)}{\Pr(Y = 1|do(X = 0))} = \frac{\Pr(Y = 1|X = 0, Z = 0)\Pr(Z = 0)}{\Pr(Y = 1|X = 0)} = \frac{0.4 \times 0.5}{0.3} = 0.67,$$

$$w_1 = \frac{\Pr(Y = 1|do(X = 0), Z = 1)\Pr(Z = 1)}{\Pr(Y = 1|do(X = 0))} = \frac{\Pr(Y = 1|X = 0, Z = 1)\Pr(Z = 1)}{\Pr(Y = 1|X = 0)} = \frac{0.2 \times 0.5}{0.3} = 0.33,$$

Table 6.2: Treatment effect estimates in Example 1, obtained using the data from Table 6.1

	Estimand	Estimate
Risk differences		
Marginal	$\Pr(Y=1\|do(X=1)) - \Pr(Y=1\|do(X=0))$	$0.7 - 0.3 = 0.4$
Conditional on $Z=0$	$\Pr(Y=1\|do(X=1), Z=0) - \Pr(Y=1\|do(X=0), Z=0)$	$0.8 - 0.4 = 0.4$
Conditional on $Z=1$	$\Pr(Y=1\|do(X=1), Z=1) - \Pr(Y=1\|do(X=0), Z=1)$	$0.6 - 0.2 = 0.4$
Risk ratios		
Marginal	$\frac{\Pr(Y=1\|do(X=1))}{\Pr(Y=1\|do(X=0))}$	$\frac{0.7}{0.3} = 2.33$
Conditional on $Z=0$	$\frac{\Pr(Y=1\|do(X=1),Z=0)}{\Pr(Y=1\|do(X=0),Z=0)}$	$\frac{0.8}{0.4} = 2$
Conditional on $Z=1$	$\frac{\Pr(Y=1\|do(X=1),Z=1)}{\Pr(Y=1\|do(X=0),Z=1)}$	$\frac{0.6}{0.2} = 3$
Odds ratios		
Marginal	$\frac{\Pr(Y=1\|do(X=1))/\Pr(Y=0\|do(X=1))}{\Pr(Y=1\|do(X=0))/\Pr(Y=0\|do(X=0))}$	$\frac{0.7/0.3}{0.3/0.7} = 5.44$
Conditional on $Z=0$	$\frac{\Pr(Y=1\|do(X=1),Z=0)/\Pr(Y=0\|do(X=1),Z=0)}{\Pr(Y=1\|do(X=0),Z=0)/\Pr(Y=0\|do(X=0),Z=0)}$	$\frac{0.8/0.2}{0.4/0.6} = 6$
Conditional on $Z=1$	$\frac{\Pr(Y=1\|do(X=1),Z=1)/\Pr(Y=0\|do(X=1),Z=1)}{\Pr(Y=1\|do(X=0),Z=1)/\Pr(Y=0\|do(X=0),Z=1)}$	$\frac{0.6/0.4}{0.2/0.8} = 6$

where the replacing of $do(X=x)$ with $X=x$ is possible because there is no confounding of the X-Y association in this example. The weighted average of the two conditional RRs is $2 \times 0.67 + 3 \times 0.33 = 2.33$, which is equal to our marginal RR.

If we had a different example in which the conditional RRs given $Z=0$ and $Z=1$ were equal, then these would equal the marginal RR.

Finally, consider the odds ratios (OR). The conditional ORs given $Z=0$ and $Z=1$ are both equal to 6. However, the marginal OR is equal to 5.44. The marginal OR of 5.44 cannot be expressed as a weighted average of the two conditional ORs. Odds ratios are 'non-collapsible'. The marginal and conditional OR estimates are both 'valid', but they have different interpretations.

In the above example, the association between X and Y is modified by Z when we measure the association between X and Y using a risk ratio, but not when we measure the association using a risk difference or odds ratio. This is a reminder that interactions are scale-dependent. Usually, if there is no interaction between X and Z when the X-Y association is measured on one scale, then there will be an interaction between X and Z when the X-Y association is measure on another scale. Also note that no interaction on the OR scale always implies an interaction on the RR scale (unless Z is not related to Y), and vice versa. The example here is atypical in that there is no interaction between X and Z on both the RD and OR scales.

6.4 Example 2: With confounding by Z

In Example 1, there was no confounding of the association between X and Y. This meant that the marginal causal treatment effect of X on Y could be estimated using the 'crude' marginal association between X and Y. That is, we could write (using the RD as an

example):

$$RD = \Pr(Y=1|do(X=1)) - \Pr(Y=1|do(X=0)) = \Pr(Y=1|X=1) - \Pr(Y=1|X=0) \quad (6.9)$$

In this second example we consider a situation in which the association between X and Y is confounded by Z, as depicted in the DAG in Figure 6.2. Table 6.3 shows data on 20 individuals that were generated according to the relationships depicted in the DAG in Figure 6.2. The covariate Z affects both X and Y, i.e. it is a confounder of the X-Y association.

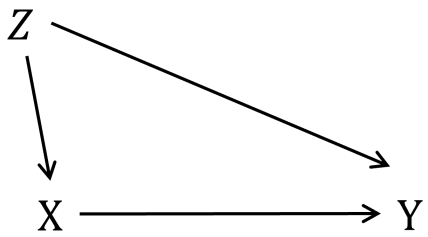


Figure 6.2: Example 2: Directed acyclic graph (DAG) illustrating relationships between three variables.

Table 6.3: Example 2: Data on covariate Z, treatment X and outcome Y for 20 individuals.

	$Z=0$		$Z=1$		Total
	$X=0$	$X=1$	$X=0$	$X=1$	
$Y=0$	2	5	1	2	10
$Y=1$	2	1	5	2	10
Total	4	6	6	4	20

Table 6.4 shows the marginal and conditional RDs, RRs, and ORs estimated using the data in Table 6.3. To estimate the marginal treatment effects we need to use standardization, due to the confounding by Z. Using standardization we have:

$$\Pr(Y=1|do(X=0)) = \frac{2}{4} \times \frac{4+6}{20} + \frac{5}{6} \times \frac{6+4}{20} = 0.5 \times 0.5 + 0.83 \times 0.5 = 0.67$$

$$\Pr(Y=1|do(X=1)) = \frac{1}{6} \times \frac{4+6}{20} + \frac{2}{4} \times \frac{6+4}{20} = 0.17 \times 0.5 + 0.5 \times 0.5 = 0.34$$

Because Z is the only confounder we can write $\Pr(Y=1|do(X=x), Z=z) = \Pr(Y=1|X=x, Z=z)$. From Table 6.3:

$$\Pr(Y=1|X=0, Z=0) = 0.5, \quad \Pr(Y=1|X=1, Z=0) = 1/6 = 0.17$$

$$\Pr(Y=1|X=0, Z=1) = 5/6 = 0.83, \quad \Pr(Y=1|X=1, Z=1) = 0.5$$

In this example, the treatment X has a protective effect on the outcome, because the RDs are less than 0,and the RRs and ORs are less than 1. In example 1 the treatment effect was in the other direction.

Table 6.4: Treatment effect estimates in Example 2, obtained using the data from Table 6.3

	Estimand	Estimate
Risk differences		
Marginal	$\Pr(Y=1\|do(X=1)) - \Pr(Y=1\|do(X=0))$	$0.34 - 0.67 = -0.33$
Conditional on $Z=0$	$\Pr(Y=1\|do(X=1), Z=0) - \Pr(Y=1\|do(X=0), Z=0)$	$0.17 - 0.5 = -0.33$
Conditional on $Z=1$	$\Pr(Y=1\|do(X=1), Z=1) - \Pr(Y=1\|do(X=0), Z=1)$	$0.5 - 0.83 = -0.33$
Risk ratios		
Marginal	$\frac{\Pr(Y=1\|do(X=1))}{\Pr(Y=1\|do(X=0))}$	$\frac{0.34}{0.67} = 0.51$
Conditional on $Z=0$	$\frac{\Pr(Y=1\|do(X=1),Z=0)}{\Pr(Y=1\|do(X=0),Z=0)}$	$\frac{0.17}{0.5} = 0.34$
Conditional on $Z=1$	$\frac{\Pr(Y=1\|do(X=1),Z=1)}{\Pr(Y=1\|do(X=0),Z=1)}$	$\frac{0.5}{0.83} = 0.60$
Odds ratios		
Marginal	$\frac{\Pr(Y=1\|do(X=1))/\Pr(Y=0\|do(X=1))}{\Pr(Y=1\|do(X=0))/\Pr(Y=0\|do(X=0))}$	$\frac{0.34/0.66}{0.67/0.33} = 0.25$
Conditional on $Z=0$	$\frac{\Pr(Y=1\|do(X=1),Z=0)/\Pr(Y=0\|do(X=1),Z=0)}{\Pr(Y=1\|do(X=0),Z=0)/\Pr(Y=0\|do(X=0),Z=0)}$	$\frac{0.17/0.83}{0.5/0.5} = 0.20$
Conditional on $Z=1$	$\frac{\Pr(Y=1\|do(X=1),Z=1)/\Pr(Y=0\|do(X=1),Z=1)}{\Pr(Y=1\|do(X=0),Z=1)/\Pr(Y=0\|do(X=0),Z=1)}$	$\frac{0.5/0.5}{0.83/0.17} = 0.20$

6.5 Using logistic regression

Above we focused on simple settings, in which the estimands of interest could be estimated without the need to fit a regression model. More generally, we need to use regression for our analysis. In this section we repeat Example 2 using regression, focusing primarily on the odds ratios. The conditional treatment effects (measured using the odds ratio) can be estimated using the logistic regression model

$$\Pr(Y=1|X=x,Z=z) = \frac{e^{\beta_0+\beta_X x+\beta_Z z+\beta_{XZ}xz}}{1+e^{\beta_0+\beta_X x+\beta_Z z+\beta_{XZ}xz}}. \tag{6.10}$$

We have included the interaction between X and Z in this model, meaning that it is a saturated model. The results from fitting this model in Stata are shown below. We see that the estimate of e^{β_X} is 0.2, the estimate of e^{β_Z} is 5, and the estimate of $e^{\beta_{XZ}}$ is 1 ($\beta_{XZ}=0$). There is no interaction between X and Z in these data, so we expect to see the estimated interaction term is equal to 1. The estimated odds ratio for the conditional association between X and Y is therefore 0.2 for both $Z=0$ and $Z=1$. That is, our estimate of $\frac{\Pr(Y=1|do(X=1),Z=1)/\Pr(Y=0|do(X=1),Z=1)}{\Pr(Y=1|do(X=0),Z=1)/\Pr(Y=0|do(X=0),Z=1)}$ is 0.2 for both values of Z. Our focus here is on the effect of X on Y and so we are not really interested in e^{β_Z} - we are including Z because we wish to control for it as a confounder, but we are not interested in the association between Z and Y.

```
. glm y i.x##i.z, family(binomial) eform

** output omitted **
------------------------------------------------------------------------------
             |                 OIM
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         1.x |         .2   .2966479    -1.09   0.278     .0109268    3.660715
         1.z |          5   7.416198     1.09   0.278     .2731707    91.51787
             |
         x#z |
       1 1   |          1   2.097618     0.00   1.000     .0163872    61.02326
             |
       _cons |          1          1    -0.00   1.000     .1408635    7.099071
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

The causal marginal odds ratio $\frac{\Pr(Y=1|do(X=1))/\Pr(Y=0|do(X=1))}{\Pr(Y=1|do(X=0))/\Pr(Y=0|do(X=0))}$ can be estimated using standardization, using the conditional estimates. Using the `margins` command in Stata gives the results shown below. This tells us that the estimates of $\Pr(Y=1|do(X=0))$ and $\Pr(Y=1|do(X=1))$ are respectively 2/3 and 1/3. The marginal odds ratio can then be estimated by plugging in these estimates, giving 0.25, as we found in the 'by hand' calculations.

```
. margins x

Predictive margins                              Number of obs     =         20
Model VCE    : OIM

Expression   : Predicted mean y, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |
           0 |   .6666667   .1463285     4.56   0.000      .379868     .9534653
           1 |   .3333333   .1463285     2.28   0.023     .0465347      .620132
------------------------------------------------------------------------------
```

Suppose we had instead fitted a logistic regression of Y on X without conditioning on Z:

$$\Pr(Y=1|X=x) = \frac{e^{\beta_0+\beta_X x}}{1+e^{\beta_0+\beta_X x}}. \tag{6.11}$$

Fitting this crude marginal model in Stata gives the results shown below. The estimate of the crude marginal odds ratio, $\frac{\Pr(Y=1|X=1)/\Pr(Y=0|X=1)}{\Pr(Y=1|X=0)/\Pr(Y=0|X=0)}$, is 0.18. This differs from the causal marginal odds ratio estimate of 0.25 because it fails to account for the confounding by Z. That is, in this example the crude marginal odds ratio gives a biased estimate of the marginal treatment effect (measured using the causal marginal odds ratio) due to confounding.

```
. glm y i.x, family(binomial) eform

** output omitted **
------------------------------------------------------------------------------
             |                 OIM
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         1.x |   .1836735    .179247    -1.74   0.082     .0271243    1.243753
       _cons |   2.333333   1.610153     1.23   0.220     .6033814    9.023222
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

6.6 Implications of non-collapsibility for randomized controlled trials

The above discussion has implications for randomized trials with binary outcomes when the treatment effect is quantified using an odds ratio. Randomization ensures that the treatment effect in a randomized trial is not subject to confounding. Nonetheless, it can still be appropriate to adjust for baseline covariates measured pre-randomization. Although some clinical trial practitioners favour doing unadjusted analyses in the interests of simplicity, many prefer to adjust for a set of pre-specified baseline variables thought to be predictive of the primary outcome. Usually this will include any variables used for stratification and minimisation.

Baseline adjustment can substantially increase statistical efficiency when an outcome is continuous and linear regression is used for the analysis. If (in each randomized group) the Pearson correlation between a baseline measure and a follow-up measure is r then adjusting for that baseline measure (using ANCOVA) will, in expectation, multiply the standard error of the treatment effect by $(1 - r^2)^{0.5}$. When an outcome is binary, and logistic regression is used for the analysis, the effects of adjusting for a pre-randomization baseline measure that is associated with the outcome variable, but not with the predictor variable of primary interest, are more subtle. These effects are illustrated by returning to Example 1 above. Suppose the data in Table 6.1 arose from a randomized trial, where $X = 1$ denotes an active treatment and $X = 0$ denotes placebo, and Z is a baseline variable that is predictive of the outcome.

The box overleaf shows the results from fitting two logistic regression models in Stata. The first is a logistic regression of Y on X, and the second is a logistic regression of Y on X and Z. The regression of Y on X gives an estimate of the causal marginal odds ratio $\frac{\Pr(Y=1|do(X=1))/\Pr(Y=0|do(X=1))}{\Pr(Y=1|do(X=0))/\Pr(Y=0|do(X=0))}$ of 5.44 ($e^{1.6946}$). Because there is no confounding here, this is an unbiased estimate of causal marginal odds ratio. The regression of Y on X and Z gives an estimate of the causal conditional odds ratio $\frac{\Pr(Y=1|do(X=1),Z=1)/\Pr(Y=0|do(X=1),Z=1)}{\Pr(Y=1|do(X=0),Z=1)/\Pr(Y=0|do(X=0),Z=1)}$ of 6 ($e^{1.7918}$). These are the same estimates as shown in Table 6.2. We see the following impacts of conditioning on Z, which are general results:

- Conditioning on baseline predictors of the outcome increases the estimated log odds ratio due to non-collapsibility. However, the marginal and conditional odds ratios have a different interpretation.
- The conditional log odds ratio estimate has a larger standard error than the marginal log odds ratio estimate. This is in contrast to the analogous situation in linear regression.

- The test statistic for the conditional log odds ratio estimate (1.76 in our example) is larger than the test statistic for the marginal log odds ratio estimate (1.74 in our example). This is because the relative increase in the standard error is not as big as that in the log odds ratio. The difference between the test statistics is not substantial.

The behaviour exhibited here is true generally in logistic regression, as shown by Robinson and Jewell in a paper entitled 'Some surprising results about covariate adjustment in logistic regression' (1991).

The conditional odds ratio (and 95% CI) could be used to obtain an estimate of the marginal OR using standardization. The estimate would be equal to 5.44, and its 95% CI would be slightly narrower than that obtained from the regression of Y on X (ignoring Z).

Both marginal and conditional odds ratio estimates are perfectly valid due to the lack of confounding - they just estimate different quantities. That is, by adjusting for baseline covariates in a randomized trial, we change the quantity that we are estimating - from a marginal to a conditional odds ratio. This is often not appreciated. The marginal treatment effect is (arguably) the relevant quantity for making policy decisions, while conditional effects are more relevant for answering how effective a treatment will be for a particular individual (on the basis of the values of their covariates). Conditional effects are of particular interest if there is effect modification - that is, if the treatment effect differs according to the characteristics of the patient. Marginal quantities refer to a specific population and care should be taken to consider whether marginal estimates from a trial are transportable to a wider population in which a treatment could be used.

```
. glm y i.x, family(binomial)

** output omitted **
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         1.x |   1.694596   .9759001     1.74   0.082    -.2181333    3.607325
       _cons |  -.8472979   .6900656    -1.23   0.220    -2.199802    .5052058
------------------------------------------------------------------------------

. glm y i.x i.z, family(binomial)

** output omitted **
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         1.x |   1.791759   1.020621     1.76   0.079    -.2086204    3.792139
         1.z |  -.9808293   1.020621    -0.96   0.337    -2.981209    1.019551
       _cons |  -.4054651   .8164966    -0.50   0.619    -2.005769    1.194839
------------------------------------------------------------------------------
```

Here we have considered conditioning on a single baseline covariate Z. In general, several baseline covariates could be conditioned on. The non-collapsibility of odds ratios means that the treatment effect we estimate will differ depending on what baseline covariates are conditioned on. This means that different trials may be estimating different quantities

due to differences in the covariates which are being conditioned on.

Partly due to non-collapsibility, and partly due to the fact that risk ratios are regarded as being more easily interpretable than odds ratios, randomized controlled trials with a binary outcome are sometimes analysed with a log (rather than a logistic) link function. A drawback to such models is that they sometimes do not converge, particularly when many covariates are included.

6.7 Implications of non-collapsibility for observational studies

The non-collapsiblity of odds ratios also has important implications for the analysis of data from observational studies. Suppose we wish to estimate the causal effect of a treatment or exposure X on an outcome Y using observational data, in which there are likely to be confounders of the X-Y association. Typically we don't know exactly which covariates are confounders. One approach that is sometimes advocated for deciding whether or not a variable is a confounder is to compare the association between X and Y before and after conditioning on the potential confounder (Z say). Intuitively, we expect the estimated association between X and Y to change when we adjust for an important confounder. Conversely, when we adjust for a variable that is not a confounder, intuitively we do not expect the estimated treatment effect to change. However, because of non-collapsibility, it turns out that this intuition is not correct for odds ratios. Any difference we see between the unconditional odds ratio $\frac{\Pr(Y=1|X=1)/\Pr(Y=0|X=1)}{\Pr(Y=1|X=0)/\Pr(Y=0|X=0)}$ and the conditional odds ratio $\frac{\Pr(Y=1|X=1,Z=z)/\Pr(Y=0|X=1,Z=z)}{\Pr(Y=1|X=0,Z=z)/\Pr(Y=0|X=0,Z=z)}$ could be due to confounding by Z, but it will at least in part be due to the fact that odds ratios are non-collapsible. There will also be some difference due to random variation. The only circumstances in which the true underlying unconditional and conditional odds ratios would be equal are if (a) Z is associated with X but not Y; (b) Z is not associated with either X or Y.

When using odds ratios, we should be aware that a change in the odds ratio estimate could be attributable purely to the non-collapsibility of odds ratios, and is not necessarily due to confounding. In practice, non-collapsibility may not have a big impact on estimates. This is true when a binary outcome is rare, because then the odds ratio approximates a risk ratio, which is collapsible.

We will discuss these issues again in session 10 ('Analysis Strategies')

6.8 Continuous outcomes

We have focused on binary outcomes in this session up to now, because of the special issues that effect measures for binary outcomes raise due to non-collapsibility. This section briefly discusses collapsibility in the context of continuous outcomes.

As covered in session 5, effects of treatments on a continuous outcome are typically quantified using a mean difference. We may be interested in a marginal mean difference

$$E(Y|do(X=1)) - E(Y|do(X=0)), \tag{6.12}$$

or a conditional mean difference (conditional on Z)

$$E(Y|do(X=1), Z=z) - E(Y|do(X=0), Z=z). \tag{6.13}$$

Mean differences are collapsible. This means that a marginal mean difference can be expressed as a weighted average of conditional mean differences. For binary Z we have the following relation between the marginal and conditional estimands

$$E(Y|do(X=1)) - E(Y|do(X=0)) = \sum_{z=0,1} \{E(Y|do(X=1), Z=z) - E(Y|do(X=0), Z=z)\} \Pr(Z=z). \quad (6.14)$$

We will use a simulated data example for illustration. We consider a situation as depicted in Figure 6.1, where Z does not confound the association between X and Y. Data were generated on Y, X and Z for 4000 individuals. 1000 individuals were in each of the four groups $(X=0, Z=0)$, $(X=0, Z=1)$, $(X=1, Z=0)$, $(X=1, Z=1)$. The outcome Y was generated using the linear model

$$Y = 10 + 2X + Z + \epsilon$$

where the residuals ϵ follow a normal distribution with mean 0 and variance 1. The data conform to the assumptions in Figure 6.1: there is no (marginal) association between X and Z, but both X and Z affect Y.

The conditional expectations $E(Y|do(X=x), Z=z)$ can be estimated using $E(Y|do(X=x), Z) = E(Y|X=x, Z=z)$. Here there is no confounding by Z and so the marginal expectation can be written as $E(Y|do(X=x)) = E(Y|X=x)$. That is, in this situation it is legitimate to estimate the marginal treatment effect without having to use standardization (and using standardization will give the same estimate).

Results from two linear regression models are shown below. The regression of Y on X alone provides an estimate of the marginal treatment effect $E(Y|do(X=1)) - E(Y|do(X=0))$ of 1.99. The regression of Y on X and Z provides an estimate of the conditional treatment effect $E(Y|do(X=1), Z=z) - E(Y|do(X=0), Z=z)$ of 1.99, which is assumed by the model to be the same for $Z=0,1$ because we do not include an interaction (and which we know is true because of how the data were simulated). The marginal and conditional treatment effect estimates are identical, which we expect because mean differences are collapsible, and because there is no X-by-Z interaction. The implication of this is that if we adjust for a variable Z and see that the coefficient for X does not change, then this implies that Z does not confound the association between X and Y. In practice, the marginal and conditional estimates will often be slightly different due to random variation - small differences indicate either no confounding or negligible confounding by Z

Note the standard errors of the marginal and conditional effect estimates are different. The standard error of the estimated coefficient from the regression of Y on X is 0.660, and the standard error of the estimated coefficient from the regression of Y on X and Z is 0.643. Corresponding to this, the test statistic is higher in the conditional model vs the marginal model (31.04 vs 30.23). This is what we expect - conditioning on a baseline predictor of the outcome which is not a confounder gives a more precise estimate. This is why it can be advantageous to use baseline adjustment in randomized trials when the outcome is continuous.

```
. regress y x

** output omitted **
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.994514   .0659763    30.23   0.000     1.865163    2.123864
       _cons |   10.48038   .0466523   224.65   0.000     10.38891    10.57184
------------------------------------------------------------------------------

. regress y x z

** output omitted **
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.994514   .0642502    31.04   0.000     1.868548     2.12048
           z |    .950175   .0642502    14.79   0.000     .8242088    1.076141
       _cons |   10.00529   .0556423   179.81   0.000     9.896198    10.11438
------------------------------------------------------------------------------
```

We can show the non-collapsibility result for mean differences algebraically. Consider the linear regression model

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \epsilon \tag{6.15}$$

Under this model, the expectation of Y given X is

$$E(Y|X) = \beta_0 + \beta_X X + \beta_Z E(Z|X) \tag{6.16}$$

If X and Z are marginally independent, then $E(Z|X) = E(Z)$, and we use the notation $E(Z) = \mu_Z$. Then we have

$$E(Y|X) = \beta_0 + \beta_Z \mu_Z + \beta_X X \tag{6.17}$$

Therefore, if we fit the regression model for Y with X as the only covariate, the coefficient for X is identical to the coefficient for X in the model which adjusts for Z (i.e. β_X), **if** X and Z are marginally independent. Note that the intercept changes from β_0 to $\beta_0 + \beta_Z \mu_Z$.

6.9 Summary

This session has introduced the concepts of collapsibility and non-collapsibility, which are properties of treatment effect estimands. More generally, they are features of measures of association between an exposure and outcome (even if we do not aim to attach a causal interpretation to the association measure).

For binary outcomes, risk differences and risk ratios are collapsible, whereas odds ratios are non-collapsible. Odds ratios are widely used in medical statistics and epidemiology because they have other convenient properties, but it is important to be aware of the implications of non-collapsibility for our analyses involving estimation of odds ratios. The non-collapsibility of odds ratios has important implications for analyses of both randomized trials and observational studies with binary outcomes.

For continuous outcomes, treatment and exposure effects are quantified using mean differences, which are collapsible.

We have focused on simple situations to introduce the key concepts. The references given below provide more detailed and advanced discussions about non-collapsibility and related issues.

References

Daniel R., Zhang J., Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal* 2021; 63: 528-557. https://doi.org/10.1002/bimj.201900297

Greenland S., Robins J.M., Pearl J. Confounding and Collapsibility in Causal Inference. *Statistical Science* 1999; 14: 29-46.

Hernán M.A., Robins J.M. *Causal Inference: What If.* 2020. Boca Raton: Chapman Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Huitfeldt A., Stensrud M.J., Suzuki E. On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging Themes in Epidemiology* 2019; 16: 1. doi: 10.1186/s12982-018-0083-9.

Robinson L.D., Jewell N.P. Some surprising results about covariate adjustment in logistic regression *International Statistical Review* 1991; 59 :227-240.

6.10 Practical 6

Datasets required
kidney_example.csv
simdata_binary.dta
simdata_binary2.dta

Introduction

This practical is in three parts. In Part A we return the the kidney stones example used in session 5. Parts B and C use simulated data.

Aims

1 To explore relationships between marginal and conditional treatment effect estimates, considering both collapsible and non-collapsible quantities.

2 To gain further practice at interpreting marginal and conditional treatment effect estimates.

3 To extend the scenarios considered in the notes to include a continuous treatment-outcome confounder.

Part A: Kidney stones example

In session 5 and practical 5 we estimated the marginal and conditional treatment effects (of X on Y) using risk differences, risk ratios and odds ratios. See Tables 5.2 and 5.3.

1 In the notes we saw that the marginal risk difference is equal to a weighted average of the conditional risk differences (conditional on stone size Z). What are the weights? Show that the marginal risk difference is equal to a weighted average of the conditional risk differences.

2 In the notes we saw that the marginal risk ratio is equal to a weighted average of the conditional risk ratios (Conditional on stone size Z). What are the weights? Show that the marginal risk ratio is equal to a weighted average of the conditional risk ratios.

3 Can the marginal odds ratio be expressed as a weighted average of the conditional odds ratios? Explain your answer.

Discuss: What are the interpretations of the marginal and conditional risk differences, risk ratios and odds ratios.

Discuss: For a patient with a large kidney stone, which treatment effect estimate(s) are most relevant to inform their treatment?

Part B: A simulated example with binary outcome

In this part we some simulated data (simdata_binary.dta) on a binary treatment X, binary outcome Y and binary covariate Z. The covariate Z is known to be measured temporally prior to X, meaning that we can be sure it does not lie on the causal pathway between X and Y.

4 Quantify the relationship between X and Z.

5 Using logistic regression (using a saturated model) estimate the conditional probabilities $\Pr(Y = 1|X = x, Z = z)$ for $x = 0, 1$ and $z = 0, 1$, and hence the conditional risk differences

$$\Pr(Y = 1|X = 1, Z = z) - \Pr(Y = 1|X = 0, Z = z), \qquad z = 0, 1$$

6 Without performing any further calculations, what do you expect the marginal risk difference to be?

7 From your logistic regression in question 5, what are the estimates of the conditional odds ratios

$$\frac{\Pr(Y = 1|do(X = 1), Z = z)/\Pr(Y = 0|do(X = 1), Z = z)}{\Pr(Y = 1|do(X = 0), Z = z)/\Pr(Y = 0|do(X = 0), Z = z)}, \qquad z = 0, 1$$

8 Fit another logistic regression to obtain an estimate of the marginal odds ratio

$$\frac{\Pr(Y = 1|do(X = 1))/\Pr(Y = 0|do(X = 1))}{\Pr(Y = 1|do(X = 0))/\Pr(Y = 0|do(X = 0))}$$

What assumption do you make? Explain the difference between this estimate and your estimates in question 7.

Part C: A simulated example with binary outcome and continuous confounder

In this part we some simulated data (simdata_binary2.dta) on a binary treatment X, binary outcome Y and continuous variable Z, where Z confounds the association between X and Y.

9 Fit a logistic regression of Y on X and Z and their interaction. What is the conditional odds ratio for an individual with (a) Z equal to its median value in the data, (b) Z equal to its 10th percentile in the data, (c) Z equal to its 90th percentile in the data.

10 Using the empirical standardization method introduced in session 5, obtain an estimate of the effect of X on Y using a marginal odds ratio.

11 Compare this with the odds ratio from a regression of Y on X alone. What is the reason for the difference between this estimate and that in question 10?

Discuss: What are the interpretations of the marginal and conditional odds ratios?

Session 7

Logistic Regression in Cohort and Case-Control Studies

7.1 Aims

The aims of this lecture and practical are for you to be able to do the following:

- Outline two epidemiological study designs: cohort studies and case-control studies.
- Use simple analyses for cohort studies and case-control studies using two-by-two tables and Mantel-Haenszel methods.
- Understand the need for logistic regression in epidemiological studies.
- Outline the justification for the use of logistic regression in case-control studies, for both binary exposures and more complex situations.
- Analyse data from cohort studies and case-control studies in Stata.

7.2 Study designs in epidemiology

In epidemiology we are interested in studying associations between exposures, treatments, or characteristics of an individual (we will use the term 'exposures') and an outcome. The outcome is often whether or not an individual dies or is diagnosed with a particular disease in some fixed period. In the context that we focus on here the outcome is univariate, but there are several exposures or covariates of interest.

The two main study designs used in epidemiology are cohort (or prospective) studies and case-control (or retrospective) studies. In both situations we start by defining some population we wish to study, for example men and women in a given age range and resident in a particular geographical area during a particular period in time. We refer to this as the underlying population of interest.

In a cohort study a suitable sample of individuals is chosen to represent the population of interest. Values of the exposures, denoted collectively by X, are determined and the individuals are followed through time (e.g. for 1 year, 5 years, 20 years) until the outcome of interest is observed ($Y = 1$), or not ($Y = 0$). A bit more generally, if the main exposure is binary or categorical then individuals may be sampled form the underlying population within strata defined by the exposure. This would be to over-represent exposure groups in the cohort study which may be rare in the underlying population.

By contrast, in a case-control study we start with individuals observed to have a specific outcome, say $Y = 1$, whom we call cases. The cases are chosen to represent those occurring in the population of interest. We then choose a suitable number of controls. The controls are chosen to represent the part of the population of interest with $Y = 0$. The exposures

are then determined retrospectively on the chosen individuals. It is important to specify the underlying population of interest when designing a case-control study. This ensures that it is clear to what population the results from a case-control study refer.

It is important to emphasise that the primary aims of the two types of study is the same, that being to investigate the association between exposures and the outcome. Referring to the types of investigation discussed in session 2, both case-control and cohort studies are used to study causal effects or to explore risk factors for a binary outcome. Unlike cohort studies, case-control studies are not typically used for prediction - this is because case-control studies provide estimates of relative associations but not of absolute risks and hence without the incorporation of additional external information they do not provide predictions of absolute risk, which are the focus of prediction studies.

Study designs in epidemiology are discussed in detail in the epidemiology modules. A nice overview of this topic is given in Chapters 6-8 of Rothman, Greenland & Lash (2008). The two classic books of Breslow and Day (1980, 1987) on case-control studies and cohort studies remain highly relevant and readable. In this session and in the practical we use example data sets from Brelow and Day 1980)

7.3 Studies with a single binary exposure

7.3.1 Notation

We begin by focusing on the simplest situation where there is a single binary outcome Y and a single binary exposure X. The aim of a study of this type is to relate a binary exposure X_i (taking values 0 if unexposed and 1 if exposed) to a binary outcome Y_i (also taking values 0 and 1). Both are observed for n individuals, $i = 1, \ldots, n$. Individuals in the underlying population therefore fall into one of four groups: $(X = 1, Y = 1)$, $(X = 1, Y = 0)$, $(X = 0, Y = 1)$, $(X = 0, Y = 0)$. The probabilities in the underlying population of belonging to these groups are respectively

$$\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}, \tag{7.1}$$

where

$$\pi_{xy} = \Pr(X = x, Y = y), \tag{7.2}$$

and

$$\pi_{11} + \pi_{10} + \pi_{01} + \pi_{00} = 1. \tag{7.3}$$

See Table 7.1(a).

In a cohort study we take a sample from the underlying population, observe the exposure X and later observe the corresponding values of Y. More generally we may take a sample from the subpopulations with $X = 0$ and $X = 1$. From this information we may estimate and compare the probabilities

$$\Pr(Y = y \mid X = x) = \frac{\pi_{xy}}{\pi_{x0} + \pi_{x1}}. \tag{7.4}$$

See Table 7.1(b).

The motivation for a case-control study comes primarily from situations in which $Y = 1$ is a rare event, so that very large sample sizes become necessary if prospective studies are to include even a modest number of such individuals. This situation motivates sampling independently the two sub-populations defined by the outcome status, $Y = 0$ (controls) and $Y = 1$ (cases). Exposure status is then observed for the sampled individuals. A case-control sample therefore leads us to consider the conditional probabilities of the exposure X given the outcome Y, namely

$$\Pr(X = x \mid Y = y) = \frac{\pi_{xy}}{\pi_{0y} + \pi_{1y}}. \tag{7.5}$$

See Table 7.1(c).

Table 7.1: Probabilities associated with binary explanatory and binary response variables in the underlying population, in a cohort study and in a case-control study.

(a) Underlying population structure

	X	
	0	1
Y		
0	π_{00}	π_{10}
1	π_{01}	π_{11}

(b) Cohort study

	X	
	0	1
Y		
0	π_{00}	π_{10}
1	π_{01}	π_{11}
$\Pr(Y = y\|X = x)$	$\pi_{01}/(\pi_{01} + \pi_{00})$	$\pi_{11}/(\pi_{10} + \pi_{11})$

(c) Case-control study: separate samples from subpopulations $Y = 0, 1$ with relevant conditional probabilities

	X		$\Pr(X = x\|Y = y)$
	0	1	
Y			
0	π_{00}	π_{10}	$\pi_{10}/(\pi_{10} + \pi_{00})$
1	π_{01}	π_{11}	$\pi_{11}/(\pi_{11} + \pi_{01})$

7.3.2 Example data

Table 7.2 shows the two-by-two tables summarising data from a cohort study and a case-control study within the cohort. In the cohort study there are 10,000 individuals, 70% of whom are unexposed ($X = 0$) and 30% of whom are exposed ($X = 1$). The case-control study uses all cases from the cohort ($Y = 1$) and a 10% sample of the controls. The case-control study involves a total of 1343 individuals - 382 cases and 961 controls.

7.3.3 Odds ratios

In a study of the association between the exposure X and the outcome Y, as defined above, we are usually interested in how the prevalence of the outcomes $Y = 0$ and $Y = 1$

Table 7.2: Example data from a cohort study and a corresponding case-control study

(a) Cohort study

	Unexposed $X = 0$	Exposed $X = 1$	Total
Disease-free ($Y = 0$)	6793	2825	9618
Disease ($Y = 1$)	207	175	382
Total	7000	3000	10,000

(b) Case-control study

	Unexposed $X = 0$	Exposed $X = 1$	Total
Disease-free ($Y = 0$)	679	282	961
Disease ($Y = 1$)	207	125	382

differs among individuals with different exposures, X. For example, in a study of the association between smoking and lung cancer we are interested in whether the proportion of lung cancer cases which arises in a group of smokers differs from the proportion of lung cancer cases arising in a group of non-smokers.

Therefore, we would like to compare the prospective conditional probabilities $\Pr(Y = 1 \mid X = 1)$ and $\Pr(Y = 1 \mid X = 0)$. One way of comparing these two probabilities is using an *odds ratio*. Firstly we define the odds. The odds of the outcome $Y = 1$ given $X = x$ $(x = 0, 1)$ are

$$\frac{\Pr(Y = 1|X = x)}{1 - \Pr(Y = 1|X = x)} = \frac{\pi_{x1}/(\pi_{10} + \pi_{11})}{1 - \pi_{x1}/(\pi_{10} + \pi_{11})} = \frac{\pi_{x1}}{\pi_{x0}}. \tag{7.6}$$

The ratio of the odds of $Y = 1$ given $X = 1$ versus $X = 0$ is

$$\frac{\Pr(Y = 1|X = 1)/\left(1 - \Pr(Y = 1|X = 1)\right)}{\Pr(Y = 1|X = 0)/\left(1 - \Pr(Y = 1|X = 0)\right)} = \frac{\pi_{11}/\pi_{10}}{\pi_{01}/\pi_{00}} = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}} \tag{7.7}$$

From Table 7.1(b) it is evident that these odds can be estimated from a cohort study. However, we can see from Table 7.1(c) that we cannot estimate the conditional probability $\pi_{11}/(\pi_{10} + \pi_{11})$ directly from a case-control study. We can instead estimate the odds of $X = 1$ conditional on $Y = y$ $(y = 0, 1)$ from a case-control study

$$\frac{\Pr(X = 1|Y = y)}{1 - \Pr(X = 1|Y = y)} = \frac{\pi_{1y}/(\pi_{1y} + \pi_{0y})}{1 - \pi_{1y}/(\pi_{1y} + \pi_{0y})} = \frac{\pi_{1y}}{\pi_{0y}}. \tag{7.8}$$

The ratio of the odds of $X = 1$ given $Y = 1$ versus $Y = 0$ is

$$\frac{\Pr(X = 1|Y = 1)/\left(1 - \Pr(X = 1|Y = 1)\right)}{\Pr(X = 1|Y = 0)/\left(1 - \Pr(X = 1|Y = 0)\right)} = \frac{\pi_{11}/\pi_{01}}{\pi_{10}/\pi_{00}} = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}} \tag{7.9}$$

The cornerstone of the analysis of case-control studies is that the *ratio* of odds of $Y = 1$ given $X = 1$ to that given $X = 0$ calculated from the cohort study is the same as the corresponding ratio of odds in the case-control study of $X = 1$ given $Y = 1$ to that given $Y = 0$:

$$\begin{aligned}&\frac{\Pr(Y = 1 \mid X = 1)/\Pr(Y = 0 \mid X = 1)}{\Pr(Y = 1 \mid X = 0)/\Pr(Y = 0 \mid X = 0)} = \\ &\quad \frac{\Pr(X = 1 \mid Y = 1)/\Pr(X = 0 \mid Y = 1)}{\Pr(X = 1 \mid Y = 0)/\Pr(X = 0 \mid Y = 0)}.\end{aligned} \tag{7.10}$$

In this session our focus is on odds ratios. Other ways of measuring associations include using a risk difference or risk ratio, or a rate ratio, which incorporates person-time. These other quantities can be estimated from cohort studies and using certain case-control sampling designs. We do not consider these measures further in this session.

7.3.4 Simple analyses

In the situation of a binary exposure, simple analyses can be used to estimate the odds ratio from a cohort or case-control study. Suppose that the data from a given study of either type is as in Table 7.2. Using the cohort study data the odds of $Y = 1$ in the $X = 1$ group is $(175/3000)/(1 - 175/3000) = 0.062$ and the odds of $Y = 1$ in the $X = 0$ group is $(207/7000)/(1 - 207/7000) = 0.030$. The odds ratio is therefore $0.062/0.074 = 2.03$. If the observed data are as in Table 7.3 then the odds ratio is given by ad/bc. Woolf's formula for the variance of a log odds ratio from a 2-by-2 table is

$$var \log OR = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{7.11}$$

The variance of the log OR from the cohort study is therefore estimated to be $1/6793 + 1/2825 + 1/207 + 1/175 = 0.011$. An estimate 95% confidence interval for the log OR is given by $\log 2.03 \pm 1.96 \times \sqrt{(0.011)} = (0.503, 0.915)$. Exponentiating this gives the 95% CI for the OR as $(1.65, 2.50)$.

Table 7.3: Example data from a cohort study and a corresponding case-control study

	Unexposed $X = 0$	Exposed $X = 1$
Disease-free ($Y = 0$)	a	b
Disease ($Y = 1$)	c	d

Exercise: use the same methods to obtain an estimate of the odds ratio from the case-control data and a corresponding 95% confidence interval.

7.4 Logistic regression with a single binary exposure

When we have multiple exposure variables we typically need analyses based on regression modelling. To estimate odds ratios the appropriate regression model is a logistic regression model. In this section we outline the use of logistic regression in cohor and case-control studies the simple situation of a single binary exposure. This paves the way for the more general situation considered in the next section. The logistic model has of course been introduced and used in previous sessions, including sessions 1, 3, 4, 5 and 6, where there was an implicit emphasis on cohort studies.

7.4.1 Cohort studies

In a cohort study the outcome Y is a random variable, which for a given individual follows a Bernoulli distribution with probability of $Y = 1$ depending on the exposure:

$$Y_i | X_i = x_i \sim \text{Binomial}(1, \pi_i) \tag{7.12}$$

where

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \lambda + \beta x_i \tag{7.13}$$

The association between Y and X is therefore expressed as a logistic model. We can write this alternatively using

$$\Pr(Y=1|X=1) = \frac{e^{\lambda+\beta}}{1+e^{\lambda+\beta}}, \quad \Pr(Y=1|X=0) = \frac{e^{\lambda}}{1+e^{\lambda}}. \tag{7.14}$$

The intercept parameter λ can be written in terms of the parameters from Table 7.1 as $\lambda = \log \pi_{01}/\pi_{00}$ and β is the log odds ratio, i.e. $\beta = \log \pi_{11}\pi_{00} - \log \pi_{10}\pi_{01}$.

7.4.2 Case-control studies

In a case-control study the outcome Y has been fixed by the study design. It is now the binary exposure X which is a random variable. For individual i we have

$$X_i|Y_i = y_i \sim \text{Binomial}(1, \pi_i^*) \tag{7.15}$$

where

$$\text{logit}(\pi_i^*) = \log\left(\frac{\pi_i^*}{1-\pi_i^*}\right) = \lambda^* + \beta y_i \tag{7.16}$$

The association between X and Y is therefore also expressed as a logistic model. We can write this alternatively using

$$\Pr(X=1|Y=1) = \frac{e^{\lambda^*+\beta}}{1+e^{\lambda^*+\beta}}, \quad \Pr(X=1|Y=0) = \frac{e^{\lambda^*}}{1+e^{\lambda^*}}. \tag{7.17}$$

The parameter β has exactly the same interpretation as under the corresponding cohort study, i.e. $\beta = \log \pi_{11}\pi_{00} - \log \pi_{10}\pi_{01}$. The intercept parameter differs though, and can be written in terms of the parameters from Table 7.1 as $\lambda^* = \log(\pi_{10}/\pi_{00})$. In the cohort formulation, λ is the log odds of disease in the unexposed. In the case-control formulation, λ^* is the log odds of exposure in the disease-free (the controls). These are two quite different things.

7.4.3 Likelihoods

The likelihood for the data from a cohort study is

$$\begin{aligned} L_{\text{cohort}} &= \prod_{i=1}^{n} \Pr(Y=y_i|X=x_i) \\ &= \prod_{i=1}^{n} \Pr(Y=1|X=x_i)^{y_i} \Pr(Y=0|X=x_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\lambda+\beta x_i}}{1+e^{\lambda+\beta x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\lambda+\beta x_i}}\right)^{1-y_i}. \end{aligned} \tag{7.18}$$

The likelihood for the data from a case-control study is

$$\begin{aligned} L_{\text{case-control}} &= \prod_{i=1}^{n} \Pr(X = x_i | Y = y_i) \\ &= \prod_{i=1}^{n} \Pr(X = 1 | Y = y_i)^{x_i} \Pr(X = 0 | Y = y_i)^{1-x_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\lambda^* + \beta y_i}}{1 + e^{\lambda^* + \beta y_i}} \right)^{x_i} \left(\frac{1}{1 + e^{\lambda^* + \beta y_i}} \right)^{1-x_i}. \end{aligned} \tag{7.19}$$

7.4.4 Example using Stata

The box below shows the results from a logistic regression of Y on X using the cohort study data from Table 7.2(a). The estimate of the log OR β is 0.71 β. The estimate of the intercept parameter λ is -3.49 and this is the log odds of $Y = 1$ in the unexposed individuals ($X = 0$). Exponentiating these values gives $e^{\hat{\beta}} = 2.03$ and $e^{\hat{\lambda}} = 0.03$. So the estimated odds of the disease $Y = 1$ in the unexposed is 0.03, and this is estimated to increase by a factor of 2.03 in the exposed. In other words the odds of the disease is 103% higher in the exposed compared to the unexposed. Therefore the estimated odds of the disease in the exposed is 0.062. The 95% CI for e^{β} is $(1.65, 2.50)$ and the p-value is < 0.001. There is strong evidence of an association between the exposure and the disease in the cohort study.

```
. glm y x, family(binomial)
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .7094515   .1051018     6.75   0.000     .5034557    .9154473
       _cons |  -3.490929   .0705559   -49.48   0.000    -3.629216   -3.352642
------------------------------------------------------------------------------

. glm y x, family(binomial) eform
------------------------------------------------------------------------------
             |                 OIM
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   2.032876    .213659     6.75   0.000     1.654429    2.497892
       _cons |   .0304725     .00215   -49.48   0.000      .026537    .0349918
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Corresponding results from the case-control study are shown below. Notice that we are now doing a logistic regression of X on Y (the X and Y are reversed in the GLM formula). The estimate of β here is 0.71, giving $e^{\hat{\beta}} = 2.04$. The estimate of λ^* is -0.88, giving $e^{\hat{\lambda}^*} = 0.42$. Therefore, the estimated odds of being exposed in the disease-free is 0.42. The odds of being exposed in the diseased is estimated to be increased by a factor of 2.04. In other words the ratio of the odds of being exposed in the diseased to the odds of being exposed in the disease-free is 2.04. Using the results from earlier, we know that this can

also be interpreted as the ratio of the odds of being diseased in the exposed to the odds of being disease in the unexposed.

As we expect, the odds ratio estimates from the cohort and case-control study data are almost identical. The intercepts are very different because the represent different quantities.

```
. glm x y, family(binomial)
------------------------------------------------------------------------------
             |                 OIM
           x |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           y |   .7107812    .124756     5.70   0.000      .466264    .9552985
       _cons |  -.8787141   .0708439   -12.40   0.000    -1.017565   -.7398626
------------------------------------------------------------------------------

. glm x y, family(binomial) eform
------------------------------------------------------------------------------
             |                 OIM
           x | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           y |   2.035581   .2539509     5.70   0.000     1.594028    2.599446
       _cons |   .4153166   .0294226   -12.40   0.000     .3614739    .4771795
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

7.5 An alternative logistic regression for case-control studies

In the case-control analysis above we performed a logistic regression of X on Y, and hopefully it is clear to see why that is the 'obvious' thing to do in the simple setting of a single binary exposure. However, the more conventional analysis of case-control data uses a logistic regression of Y on X. In this section we explain why that is justified, and in the next section we explain why it is especially needed when we are in a more complex setting of a continuous exposure and/or multiple exposures. First, to motivate the discussion in this section, we show the results from a logistic regression of Y on X using the case-control data:

```
. glm y x, family(binomial)
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .7107812    .124756     5.70   0.000       .466264    .9552985
       _cons |  -1.187902   .0793957   -14.96   0.000    -1.343515    -1.03229
------------------------------------------------------------------------------

. glm y x, family(binomial) eform
------------------------------------------------------------------------------
             |                 OIM
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   2.035581   .2539509     5.70   0.000     1.594028    2.599446
       _cons |   .3048601   .0242046   -14.96   0.000     .2609269    .3561905
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

The estimate of the coefficient for X in this logistic regression of Y on X is identical to the log odds ratio estimated from the earlier regression of X on Y - this is because of the 'reversibility' of odds ratios as shown in (7.10). However, the intercept is different. In the regression of X on Y the estimated intercept was -0.88, whereas in the regression of Y on X the intercept is -1.19. Next we explain why it is justified to estimate odds ratios from a case-control study using a logistic regression with Y (disease status) as the outcome variable.

In a case-control study the probabilities we are able to estimate are conditional on being in the case-control sample. We let $S = 1$ denote the indicator of being in the case-control sample. So everyone in the case-control sample has $S = 1$, while individuals in the underlying cohort who were not selected for the case-control sample have $S = 0$. Based on case-control data alone we observe $\Pr(Y = 1|S = 1)$, the probability of being a case conditional on being in the case-control sample. We can't estimate $\Pr(Y = 1)$ from a case-control study alone - this is the probability of being a case in the underlying population. In the example data in Table 7.2 the cohort study provides an estimate of $\Pr(Y = 1)$ of $382/10000 = 0.0382$, and from the case-control study we have $\Pr(Y = 1|S = 1) = 382/1343 = 0.28$. So the distribution of cases is quite different in the underlying population and in the case-control sample. Similarly the distribution of the exposure is quite different in the underlying cohort ($\Pr(X = x)$) and in the case-control sample ($\Pr(X = x|S = 1)$). However, conditional on Y, the exposure distribution in the case-control sample is representative of that in the underlying cohort, i.e. $\Pr(X = x|Y = y, S = 1) = \Pr(X = x|Y = y)$. It turns out that this result is key to showing that a case-control study can be analysed using a logistic regression of Y on X.

Using Bayes theorem we can write

$$\Pr(X = 1|Y = 1, S = 1) = \frac{\Pr(Y = 1|X = 1, S = 1)\Pr(X = 1|S = 1)}{\Pr(Y = 1|S = 1)} \tag{7.20}$$

The first term $\Pr(Y = 1|X = 1, S = 1)$ can be written as follows, where we make use of the fact that sampling to the case-control study is done of the basis of Y only, and hence

$\Pr(S = 1|Y = y, X = x) = \Pr(S = 1|Y = y)$:

$$\begin{aligned}
\Pr(Y = 1|X = 1, S = 1) &= \frac{\Pr(S = 1|Y = 1, X = 1)\Pr(Y = 1|X = 1)}{\Pr(S = 1|X = 1)} \\
&= \frac{\Pr(S = 1|Y = 1, X = 1)\Pr(Y = 1|X = 1)}{\Pr(S = 1|X = 1, Y = 1)\Pr(Y = 1|X = 1) + \Pr(S = 1|X = 1, Y = 0)\Pr(Y = 0|X = 1)} \\
&= \frac{\Pr(S = 1|Y = 1)\Pr(Y = 1|X = 1)}{\Pr(S = 1|Y = 1)\Pr(Y = 1|X = 1) + \Pr(S = 1|Y = 0)\Pr(Y = 0|X = 1)} \\
&= \frac{\Pr(Y = 1|S = 1)\Pr(Y = 0)\Pr(Y = 1|X = 1)}{\Pr(Y = 1|S = 1)\Pr(Y = 0)\Pr(Y = 1|X = 1) + \Pr(Y = 0|S = 1)\Pr(Y = 1)\Pr(Y = 0|X = 1)}
\end{aligned} \tag{7.21}$$

In this expression the probabilities $\Pr(Y = y|X = x)$ and $\Pr(Y = y)$ refer to the underlying population, while the probabilities $\Pr(Y = y|S = 1)$ refer to the case-control sample. Now we make use of the logistic model for $\Pr(Y = y|X = x)$. These probabilities were given in (7.14). We plug $\Pr(Y = 1|X = 1) = \frac{e^{\lambda+\beta}}{1+e^{\lambda+\beta}}$ and $\Pr(Y = 1|X = 0) = \frac{e^{\lambda}}{1+e^{\lambda}}$ into the above expression. After a bit of rearrangement it can be shown that

$$\Pr(Y = 1|X = 1, S = 1) = \frac{e^{\alpha^*+\beta}}{1 + e^{\alpha*+\beta}} \tag{7.22}$$

where

$$\alpha^* = \lambda + \log\left(\frac{\Pr(Y = 1|S = 1)}{\Pr(Y = 0|S = 1)}\right) - \log\left(\frac{\Pr(Y = 1)}{\Pr(Y = 0)}\right) \tag{7.23}$$

Using similar steps it can be shown that

$$\Pr(Y = 1|X = 0, S = 1) = \frac{e^{\alpha^*}}{1 + e^{\alpha*}}. \tag{7.24}$$

Putting the results in (7.20), (7.24), and (7.22) together and putting them in the case-control likelihood in (7.19) we see that the likelihood can be written as

$$\begin{aligned}
L_{\text{case-control}} &= \prod_{i=1}^{n} \Pr(X = x_i|Y = y_i) \\
&= \prod_{i=1}^{n} \Pr(X = x_i|Y = y_i, S = 1) \\
&= \prod_{i=1}^{n} \frac{\Pr(Y = y_i|X = x_i, S = 1)\Pr(X = x_i|S = 1)}{\Pr(Y = y_i|S = 1)} \\
&= \prod_{i=1}^{n} \left(\frac{e^{\alpha^*+\beta x_i}}{1 + e^{\alpha*+\beta x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\alpha*+\beta x_i}}\right)^{1-y_i} \Pr(X = x_i|S = 1)/\Pr(Y = y_i|S = 1)
\end{aligned} \tag{7.25}$$

The term $\Pr(Y = y_i|S = 1)$ is just a constant. The term $\Pr(X = x_i|S = 1)$ does not provide any information about the parameters α^* or β. Hence we have

$$L_{\text{case-control}} \propto \prod_{i=1}^{n} \left(\frac{e^{\alpha^*+\beta x_i}}{1 + e^{\alpha*+\beta x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\alpha*+\beta x_i}}\right)^{1-y_i} \tag{7.26}$$

The case-control likelihood is proportional to a likelihood corresponding to a logistic regression of Y on X. It follows that we can analyse case-control study using a logistic regression of Y on X. The coefficient for X, β, is the same log odds ratio as would be estimated had the data arisen from a cohort study. The intercept parameter α^*does not have a useful interpretation - it is a function of α (the intercept in the logistic model for the cohort study), the proportions of cases and controls in the case-control study ($\Pr(Y = 1|S = 1)$ and $\Pr(Y = 0|S = 1)$), and the proportion of cases and controls in the underlying population ($\Pr(Y = 1)$ and $\Pr(Y = 0)$), which are not estimable from the case-control data. Looking at the Stata output given at the start of this section, the estimate of α^* is -1.19.

7.6 Logistic regression with multiple covariates

So far in this session the focus has been on a single binary exposure. Most often, however, there are multiple explanatory variables to take into account in an analysis. We now consider p explanatory variables or exposures, $X_1, \ldots, X_p$. These may include both categorical and continuous variables.

7.6.1 Simple analyses

Suppose we have just two exposures to consider, X_1 and X_2, and suppose that X_1 is a binary variable and X_2 is a categorical variable with C groups. The data from a cohort or case-control could be displayed by making C tables such as Table 7.3 for the association between X_1 and Y, one for each category of X_2. The is illustrated in Table 7.4 for binary X_1 and X_2.

Table 7.4: Example data for a binary outcome Y and two binary exposures, X_1 and X_2.

	$X_2 = 0$		$X_2 = 1$	
	$X_1 = 0$	$X_1 = 1$	$X_1 = 0$	$X_1 = 1$
$Y = 0$	a_1	b_1	a_2	b_2
$Y = 1$	c_1	d_1	c_2	d_2

The odds ratios for the association between Y and X_1 conditional on X_2 is

$$\frac{\Pr(Y = 1 \mid X_1 = 1, X_2)/\Pr(Y = 0 \mid X_1 = 1, X_2)}{\Pr(Y = 1 \mid X_1 = 0, X_2)/\Pr(Y = 0 \mid X_1 = 0, X_2)} = \frac{\Pr(X_1 = 1 \mid Y = 1, X_2)/\Pr(X_1 = 0 \mid Y = 1, X_2)}{\Pr(X_1 = 1 \mid Y = 0, X_2)/\Pr(X_1 = 0 \mid Y = 0, X_2)}. \quad (7.27)$$

One approach to estimating the conditional OR this is the Mantel-Haenszel method. The Mantel-Haenszel estimator for the conditonal odds ratio is

$$OR_{MH} = \frac{\sum_{i=1}^{C} a_i d_i / n_i}{\sum_{i=1}^{C} b_i c_i / n_i} \quad (7.28)$$

where the sums are over the levels of X_2 and $n_i = a_i + b_i + c_i + d_i$ denotes the number of individuals in the ith category of X_2.

This procedure is OK when we just have two categorical variables to consider, however it becomes cumbersome as the number of categories increases and does not accommodate continuous variables. This motivates the need for a regression-based approach.

7.6.2 Likelihood for cohort and case-control studies

In the setting of a cohort study the logistic model for the dependence of Y on $X_1, \ldots, X_p$ is formulated as:

$$Y_i = 1 | X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip} \sim \text{Binomial}(1, \pi_i) \tag{7.29}$$

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \lambda + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \tag{7.30}$$

We can alternatively write this in the form

$$\Pr(Y_i = 1 | X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip}) = \frac{e^{\lambda+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1 + e^{\lambda+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}. \tag{7.31}$$

In this model, λ is the log odds of developing the disease when all the x's take the value zero, and β_k is the log odds-ratio of developing the disease for a 1 unit increase in x_k, with the other x's held fixed, that is, adjusting for (conditional on) the other covariates.

The likelihood for the cohort study is

$$L_{\text{cohort}} = \prod_{i=1}^{n} \Pr(Y_i = y_i | X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip}) \tag{7.32}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\lambda+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1 + e^{\lambda+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}\right)^{y_i} \left(\frac{1}{1 + e^{\lambda+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}\right)^{1-y_i} \tag{7.33}$$

What is the likelihood for the case-control study? When we had a binary exposure we used the probabilities $\Pr(X = x | Y = y)$ to write the likelihood for the case-control study in 7.19. In this more general situation the likelihood is

$$L_{\text{case-control}} = \prod_{i=1}^{n} P(X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip} | Y_i = y_i) \tag{7.34}$$

where the $P(.)$ now represents the joint distribution of the predictor variables conditional on the outcome. This may be a complex distribution, especially if the predictors include both categorical and continuous variables. Now it is more difficult to see how we might use a logistic regression model to estimate the log odds ratios of interest, here $\beta_1, \ldots, \beta_p$, from a case-control study.

7.6.3 Logistic regression for a general case-control study

The arguments from section 8.5, leading to the case-control likelihood in 7.26, can be extended directly to the more general setting with multiple covariates. Specifically it can be shown that the case-control likelihood is

$$L_{\text{case-control}} \propto \prod_{i=1}^{n} \left(\frac{e^{\alpha^*+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1 + e^{\alpha^*+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}\right)^{y_i} \left(\frac{1}{1 + e^{\alpha^*+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}\right)^{1-y_i} \tag{7.35}$$

where α^* is as defined in 7.23.

It can also be shown that the variances of the log odds ratio parameters $\beta_1, \ldots, \beta_p$ can be estimated in the usual way, by using the inverse of the observed information matrix. We do not give the details of this here.

7.7 Final remarks

Logistic regression is widely used in the study on binary outcomes in both cohort studies and case-control studies. In this session we have explained the justification for using a logistic regression of case/control status (Y) on the exposures ($X_1, \ldots, X_p$) to estimate odds ratios from a case–control study, where these odds ratios are the same could be estimated from a corresponding cohort study. In logistic analyses of cohort and case-control studies the intercepts differ and those from a case-control study do not have a useful interpretation are typically ignored.

In the practical you will analyse data from a case-control study with multiple covariates.

References

Breslow, N.E. and Day, N.E. *Statistical Methods in Cancer Research: Volume I. The Analysis of Case-Control Studies*. 1980. Lyon: International Agency for Research on Cancer. Available online at publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications.

Breslow, N.E. and Day, N.E. *Statistical Methods in Cancer Research: Volume II. The Design and Analysis of Cohort Studies*. 1987. Lyon: International Agency for Research on Cancer. Available online at publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications.

Keogh, R.H. and Cox, D.R. *Case-control studies*. 2014. Cambridge: Cambridge University Press.

Prentice, R.L. and Pyke, R. 1979. Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

Rothman K., Greenland S., Lash T. *M*odern Epidemiology. 3rd Edition. 2008. Lippincott Williams & Wilkins.

7.8 Practical 7

Datasets required: `oesophageal_data-1.dta` and `oesophageal_data-2.dta`

Introduction

We will use data from a case-control study of oesophageal cancer. Cases are 200 men diagnosed with oesophageal cancer in the Ille-et-Vilaine area of France between January 1972 and April 1974. Controls are 776 men who were sampled from an electoral register. The data are used extensively in the book Breslow, N.E. and Day, N.E. 1980. Statistical Methods in Cancer Research: Volume 1 The Analysis of Case-Control Studies. This book is freely available online at

`https://publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications/Statistical-Methods-In-Cancer-Research-Volume-I-The-Analysis-Of-Case-Control-Studies-1980`

The first dataset is called oesophageal_data1.dta and consists of one row per participant, and four variables as below:

Variable	Description
`case`	0=control, 1=case
`age_group`	age group 1: 25-34, 2: 35-44, 3: 45-54, 4: 55-64, 5: 65-74, 6: 75+
`tobacco_group`	tobacco intake group (in grams per day) 0: None, 1: 1-9, 2: 10-19, 3: 20-29, 4: 30+
`alcohol_grp`	alcohol intake group (in grams per day) 0: 0-39, 1: 40-79, 2: 80-119, 3: 120+

Aims

We will examine the association between oesophageal cancer as the outcome and a dichotomous smoking status variable (smoker or non-smoker) as an exposure, and attempt to answer the following questions:

1 Is smoking associated with the risk ofoesophageal cancer?

2 Does alcohol intake or age confound this relationship?

3 What is the best way to adjust for these potential confounders: as categorical or continuous variables?

We will also swap the exposure and outcome variables to explore what happens to the estimate of the odds ratio (and its confidence interval).

Analysis

1 Generate a new binary variable for tobacco intake which takes value 0 if intake is 0 and value 1 otherwise. Call this tobacco_2. Explore the associations between:

(a) binary tobacco status and alcohol group

(b) binary tobacco status and age

(c) alcohol and age

Discuss: Are there any associations which may affect your analysis?

2 (a) Use the two-by-two table for the binary tobacco variable and the case-control status to estimate the odds ratio. Use Woolf's method to obtain a 95% confidence interval for this estimate.

(b) Interpret your results.

(c) Compare your estimates to those obtained from the `mhodds` command.

3 Write down the algebraic form of the following logistic regression models, and then use the `glm` command in Stata to fit the models:

(a) with oesophageal cancer as the outcome and tobacco intake as the explanatory variable

(b) with tobacco intake as the outcome and oesophageal cancer as the explanatory variable

Discuss: Compare and contrast the results from the two analyses

4 We will now control for alcohol intake. Using the Mantel-Haenszel method `mhodds`, find the conditional odds ratio:

(a) with oesophageal cancer as the outcome and tobacco intake as the explanatory variable

(b) with tobacco intake as the outcome and oesophageal cancer as the explanatory variable

5 Repeat Q4 a) and b) using the `glm` command in Stata.

Discuss: Compare and contrast the results from the two models, and the two mhodds commands. What do the intercept parameters represent in each logistic model?

6 We will now also condition on age. Add the age group variable into the model with case-control status as the outcome and interpret the results.

Discuss: What impact does adding age to the model have on the odds ratio for tobacco status? What are two possible reasons for any impact that you see?

7 Age and alcohol intake are in fact available as continuous variables. These can be found in the `oesophageal_data-2.dta` dataset.

Fit a logistic model for oesophageal cancer with three explanatory variables: binary tobacco intake, continuous alcohol intake and continuous age.

Discuss: Compare the results from this model to your results from Q6. What are the advantages and disadvantages of the two approaches?

8 Present your results in a table or tables suitable for use in a paper or report.

Session 8

Count outcomes

8.1 Aims & Objectives

In this session we will consider models for outcomes (dependent variables) which are counts. By the end of the session you should

- understand the Poisson regression model for outcomes
- understand why overdispersion can occur for count outcomes
- be familiar with different approaches for accommodating overdispersion

8.2 Motivation

As usual, we let Y denote the outcome, which in this session denotes some sort of count. Examples which arise in clinical and epidemiological studies include:

- number of asthma attacks occurring in the month following baseline for each patient in a clinical trial,
- number of epileptic seizures occurring in a year for patients suffering from epilepsy,
- number of tumours found in an magnetic resonance brain scan of patients.

An often cited early example of use of the Poisson model was in modelling the number of soliders killed by being kicked by horse kicks in the Prussian cavalary.

8.3 Poisson GLM

A random variable Y which is a count can take non-negative integer values $\{0, ..\}$. As described in the Probability sessions last term, Poisson random variables can be derived as arising as the number of events which occur over an interval of time, when the events occur independently of each other and at a constant rate. This suggests that a Poisson model may be a reasonable (starting) model for count outcomes such as the number of asthma attacks occurring over a given time.

Recall that if $Y \sim Po(\mu)$ then

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

The Poisson GLM assumes that Y follows a Poisson distribution conditional on covariates $x_1, .., x_p$. The canonical link function is $\theta = \log(\mu)$. Thus the Poisson GLM assumes that $Y_i \sim Po(\mu_i)$ where

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

The ratio of the means for one subject with covariate vector $\mathbf{x}_1 = (x_{11}, .., x_{1p})$ and another with covariate vector $\mathbf{x}_0 = (x_{01}, .., x_{0p})$ is then equal to

$$\frac{\exp(\beta_0 + \beta_1 x_{11} + ... + \beta_p x_{1p})}{\exp(\beta_0 + \beta_1 x_{01} + ... + \beta_p x_{0p})} = \exp(\beta_1(x_{11} - x_{01}) + ... + \beta_p(x_{1p} - x_{0p}))$$

The coefficients β_0 corresponds to the log of the mean of Y for a subject with all covariates equal to zero. The coefficient β_1 represents the increase in the log of the mean for a one unit increase in the covariate x_1. The exponentiated coefficients are usually referred to as rate-ratios, since this is the interpretation in the common situation when the outcome Y arises as the number of events over a particular period.

Poisson GLMs can be fitted in Stata either using the `glm` command, or (slightly more succinctly) with the `poisson` command.

8.4 Example - TRUST asthma trial

To illustrate Poisson regression we consider data from the TRUST randomised trial carried out in 983 asthma patients. Patients were randomised to receive inhaled steroids every day (the `regular` group) or placebo every day with inhaled steroids only used according to perceived patient need (the `on-demand group`). The outcome considered here is the number of asthma exacerbations each patient experienced over the course of the one-year follow-up period.

The distribution of the numbers of exacerbations experienced in each group was as follows.

Number of exacerbations (`totex`)	On-demand Group	Regular Group
0	263	283
1	109	115
2	62	47
3	24	23
4	20	17
5	2	5
6	2	1
7	2	4
8	1	0
9	1	1
10	0	0
11	0	1
Total	486	497

Table 8.1: Exacerbation data from the TRUST randomised controlled trial in asthma.

To facilitate our understanding of the issues to come, we shall focus for now on a very simple Poisson model in which the count of the number of exacerbations is related just to the treatment group. This can be fitted in Stata using either of the following commands.

```
. poisson totex i.treat

Iteration 0:   log likelihood = -1410.5258
Iteration 1:   log likelihood = -1410.5258

Poisson regression                                Number of obs   =        983
                                                  LR chi2(1)      =       0.59
                                                  Prob > chi2     =     0.4405
Log likelihood    = -1410.5258                    Pseudo R2       =     0.0002

------------------------------------------------------------------------------
       totex |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |
    Regular  |  -.0525111   .0680885    -0.77   0.441    -.1859622    .0809399
       _cons |  -.1039897   .0477818    -2.18   0.030    -.1976404    -.010339
------------------------------------------------------------------------------

. glm totex i.treat, family(Poisson)

Iteration 0:   log likelihood = -1435.6091
Iteration 1:   log likelihood = -1410.5267
Iteration 2:   log likelihood = -1410.5258
Iteration 3:   log likelihood = -1410.5258

Generalized linear models                         No. of obs      =        983
Optimization     : ML                             Residual df     =        981
                                                  Scale parameter =          1
Deviance         =  1752.229977                   (1/df) Deviance =   1.786167
Pearson          =   2101.04419                   (1/df) Pearson  =   2.141737

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   2.873908
Log likelihood   = -1410.525818                   BIC             =  -5007.458

------------------------------------------------------------------------------
             |                 OIM
       totex |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |
    Regular  |  -.0525111   .0680885    -0.77   0.441    -.1859622    .0809399
       _cons |  -.1039897   .0477818    -2.18   0.030    -.1976404    -.010339
------------------------------------------------------------------------------
```

This provides little evidence that the mean number of exacerbations differs between treatment groups. The ratio of the estimated mean for those receiving regular inhaled steroids to those not is $\exp(-0.053) = 0.949$, so that the estimated mean is 5.1% lower. The estimated mean number in those not receiving regular inhaled steroids is $\exp(-0.104) = 0.901$, while for those receiving regular inhaled steroids it is $\exp(-0.104 - 0.053) = 0.855$. Note that Stata can provide the exponentiated parameter estimates by adding `eform` as an option in the `glm` command.

```
. glm totex i.treat, family(Poisson) eform

** <output omitted> **

------------------------------------------------------------------------------
             |                 OIM
      totex |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      treat |
    Regular  |   .9488437   .0646054    -0.77   0.441      .830305    1.084306
      _cons |   .9012346   .0430627    -2.18   0.030     .8206649    .9897143
------------------------------------------------------------------------------
```

For the Poisson model 'null model' which contains no covariates, the MLE of the mean μ is simply the sample mean $(1/n)\sum_{i=1}^{n} Y_i$. Consequently, since our Poisson regression only contains categorical covariates, the MLE of the means in the two groups are identical to the sample means of the variable in these two groups.

```
. by treat: summ totex

------------------------------------------------------------------------------
-> treat = Ondemand

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       totex |        486    .9012346    1.334135          0          9

------------------------------------------------------------------------------
-> treat = Regular

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       totex |        497    .8551308     1.40389          0         11
```

8.5 Overdispersion

The Poisson model assumes that the (conditional on covariate values) variance of Y is a particular function of the (conditional) mean - in fact that they are identical. This is a strict assumption, and in practice, this assumption might not hold.

To see how the mean-variance assumption made by the Poission model may be violated, consider our model for Y with covariates $X_1, .., X_p$. This model assumes that two individuals who share the same values of the covariates (or indeed different covariate values but such that their linear predictor values are identical) have outcomes which follow identical Poisson distributions with a particular mean. In reality, it is likely there are some other factors which independently (of the covariates we have included) affect Y, such that the two individuals outcomes have different means.

To make this concept more formal, consider the population of asthma patients in the TRUST trial randomised to receive regular inhaled steroids, and their number of exacerbations. The Poisson regression model fitted previously assumes that each patient's Y follows a Poisson distribution with common mean μ. Now suppose that in truth, there

exist other factors (as there surely do), such that some patients have outcomes which follow a Poisson distribution with higher mean than others. Assume that for patient i, Y_i follows a Poisson distribution with mean μ_i. We assume that, across the population of patients, the μ_i have overall mean μ and variance σ^2. Note that this is an example of a random-effects model (covered in more detail in modules such as that on hierarchical data), where the random-effect μ_i represents the effects of omitted covariates.

Then across the population of patients Y will have (marginal, as opposed to conditional on μ_i) mean

$$E(Y_i) = E(E(Y_i|\mu_i)) = E(\mu_i) = \mu$$

and marginal variance (using the law of total variance)

$$\begin{aligned} Var(Y_i) &= E(Var(Y_i|\mu_i)) + Var(E(Y_i|\mu_i)) \\ &= E(\mu_i) + Var(\mu_i) \\ &= \mu + \sigma^2. \end{aligned}$$

Thus the variability of the counts across the patients randomised to receive regular inhaled steroids will show greater variability than would be expected under the Poisson model. In particular, the variance will be larger than the mean, as opposed to the two being equal, as implied by the standard Poisson distribution.

8.5.1 Checking for overdispersion

If our Poisson GLM has only categorical covariates we may informally examine whether there is overdispersion by comparing the mean and variances of Y according to categories formed by the covariates. In the earlier example of the exacarbation count variable in TRUST, we can see that the means in the two groups (0.901 and 0.855) are markedly smaller than their respective variances ($1.334^2 = 1.780$ and $1.404^2 = 1.971$). There is thus much more variability in the counts than would be expected under a Poisson model.

If using `glm` to fit a Poisson model Stata reports the deviance and Pearson statistics. If these statistics, divided by the residual degrees of freedom, are substantially larger than one, this is suggestive of overdispersion. For the Poisson model for the TRUST data considered above, the deviance and goodness of fit statistics, divided by the residual degrees of freedom, are 1.79 and 2.14, again suggesting that we have overdispersion.

If using `poisson` the deviance and Pearson's statistics are not reported. However the `estat gof` command after `poisson` performs both Pearson's goodness of fit test and the deviance test. This command formally compares each of these statistics to a χ^2 distribution with degrees of freedom equal to the residual degrees of freedom, giving p-values which can be interpreted as relating to a test of overdispersion. However, the asymptotic arguments for comparison to a χ^2 distribution do not hold here. The reason is that (unlike for grouped binary data) as the sample size increases the number of parameters in the saturated model also increases and this violates the necessary assumptions for the tests to be valid. Pearson's goodness of fit test typically performs better than the deviance (type 1 error rate closer to nominal 5%). In the textbook by Pawitan it is stated that the performance of the deviance test improves as the expectations of the counts increase. However, it is recommended that both tests are used with care, and not to rely too closely on either p-value.

8.5.2 Random-effects for overdispersion

An better approach for assessing overdispersion is to compare the fit of the Poisson model with a model which allows for the possibility of 'extra-Poisson' variability. As we saw previously, overdispersion can arise when there are omitted factors which influence Y, conditional on the included covariates $x_1, .., x_p$. These omitted factors can be represented by including in the model a subject-specific random-effect a_i.

The most popular random-effects model for count data is the negative binomial model. This assumes that for each subject there exists a random-effect a_i, and that given a_i and the observed covariates $\mathbf{x}_i = (x_{i1}, .., x_{ip})$, Y_i follows a Poisson distribution with mean $\exp(\beta^T \mathbf{x}_i + a_i)$. In the negative binomial model, the exponentiated subject-specific random-effects $\exp(a_i)$ are assumed to follow a gamma distribution with mean 1 and variance α. This has the consequence that when the expectation of the outcome is μ then its variance is $\mu(1 + \alpha\mu)$. The larger the parameter α, the greater the overdispersion. An alternative name for the model is (for obvious reasons) the gamma-Poisson mixture model.

In Stata negative binomial models can be fitted using `nbreg`. Fitting a negative binomial model to the TRUST data we obtain the following.

```
. nbreg totex i.treat

** <output omitted> **

Negative binomial regression                    Number of obs     =        983
                                                LR chi2(1)        =       0.28
Dispersion        = mean                        Prob > chi2       =     0.5966
Log likelihood    = -1273.1428                  Pseudo R2         =     0.0001

------------------------------------------------------------------------------
       totex |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |
    Regular  |  -.0525111    .099203    -0.53   0.597    -.2469454    .1419231
       _cons |  -.1039899   .0701058    -1.48   0.138    -.2413947    .0334149
-------------+----------------------------------------------------------------
    /lnalpha |   .2460895     .10964                      .0311991    .4609799
-------------+----------------------------------------------------------------
       alpha |   1.279014   .1402311                      1.031691    1.585627
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 274.77               Prob >= chibar2 = 0.000
```

First, we notice that the coefficient estimates for the intercept and treatment effect are identical to those from `poisson` (and `glm`): this occurs here because the model only contains a single binary covariate: more generally the parameter estimates will differ between the two models.

Second, we notice that the standard errors are considerably larger than those from `poisson`. The standard errors from `nbreg` are nearly 50% larger. Correspondingly the 95% confidence interval around the treatment effect is also nearly 50% wider.

When overdispersion is present, the model based standard errors calculated are underestimates, in the sense that they underestimate the variance of the MLE in repeated samples.

Consequently, our hypothesis tests will have type 1 error rates which exceed the nominal 5% level, leading to an increased probability of rejecting the null hypotheses when it is true.

When $\alpha = 0$, the negative binomial model reduces to the standard Poisson model. Thus we can test whether there is evidence of overdispersion by comparing the relative fit of the two models. This is what is presented at the bottom of the output from `nbreg`: there is highly sigificant evidence that the negative binomial model fits the data better. This is unsurprising here, since the number of asthma exacerbations is likely to vary considerably between patients, even in the same treatment group.

In addition to the estimated coefficients for the covariates, we have estimates of α and $\log(\alpha)$, along with 95% Wald based confidence intervals. The magnitude of α gives an indication of how much variation there is in Y, after removing the systematic effects of the covariate(s).

8.5.3 Estimating equation and quasilikelihood approaches

Usually, the regression models we fit in medical statistics are full probability models, in the sense that they fully specify the conditional distribution of Y given the covariates $x_1, .., x_p$. Sometimes we are primarily interested in a less ambitious task - to model how the conditional mean of the outcome Y depends on the covariates.

Suppose that we believe we have correctly modelled $E(Y_i|x_{i1}, .., x_{ip})$, that is, how the expectation of the outcome depends on the covariates, but that our parametric model might not be correctly specified in other respects. The Poisson model for the TRUST data with randomised treatment group as the sole covariate is an example. With only a single binary covariate, there is no possibility of mis-specifying how the mean of Y depends on x. However, we saw that the counts are overdispersed relative to what would be expected with a Poisson model.

A quasilikelihood approach involves estimating the parameters of interest based on the likelihood of a particular parametric model, even though we do not believe that model is correct (in every respect) for the data at hand. Suppose that the GLM we have fitted assumes

$$\begin{aligned} E(Y_i) &= \mu(\beta, \mathbf{x}_i) \\ Var(Y_i) &= \phi v(\beta, \mathbf{x}_i) \end{aligned}$$

for known functions $\mu()$ and $v()$. Then the likelihood score equation for β which we solve to obtain the MLE of β can be shown to be

$$\sum_{i=1}^{n} \frac{\partial \mu(\beta, \mathbf{x}_i)}{\partial \beta} v^{-1}(\beta, \mathbf{x}_i)(Y_i - \mu(\beta, \mathbf{x}_i)) = 0$$

The theory of *estimating equations* says that estimators which are obtained as solutions to estimating equations are consistent if the estimating equations are unbiased: that is they have mean zero when expectations are taken at the true value of the parameters.

Provided we have correctly specified the mean function, i.e. that $E(Y|\mathbf{x}_i) = \mu(\beta, \mathbf{x}_i)$, then $E(Y_i - \mu(\beta, \mathbf{x}_i)) = 0$, and it follows that the GLM estimating equations are unbiased. As a consequence, it follows that, provided we have the mean function is correct, the MLE

of β, based on whatever parametric model we have assumed, is consistent, even if the parametric model is not correct in other respects.

However, the standard errors reported by Stata (generally) rely on the parametric model being correct in every respect - they are based on the observed information matrix. To obtain valid standard errors we can use the robust, or sandwich estimator of variance. In Stata we can do this by adding the `robust` option. Returning to the TRUST Poisson model that we fitted to the alcohol data, we obtain the following.

```
. glm totex i.treat, family(Poisson) robust

Iteration 0:   log pseudolikelihood = -1435.6091
Iteration 1:   log pseudolikelihood = -1410.5267
Iteration 2:   log pseudolikelihood = -1410.5258
Iteration 3:   log pseudolikelihood = -1410.5258

Generalized linear models                         No. of obs      =        983
Optimization     : ML                             Residual df     =        981
                                                  Scale parameter =          1
Deviance         =  1752.229977                   (1/df) Deviance =   1.786167
Pearson          =   2101.04419                   (1/df) Pearson  =   2.141737

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   2.873908
Log pseudolikelihood = -1410.525818               BIC             =  -5007.458

------------------------------------------------------------------------------
             |               Robust
       totex |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |
    Regular  |  -.0525111   .0996094    -0.53   0.598    -.2477421    .1427198
       _cons |  -.1039897   .0671147    -1.55   0.121    -.2355321    .0275526
------------------------------------------------------------------------------
```

Again the standard errors are much larger than those obtained without the `robust` option, which rely on the model being correct in every respect, and are here very close to those obtained from negative binomial regression.

For more details on estimating equation and quasilikelihood approaches, see Chapter 14 of Pawitan's book.

8.6 Summary

In this session we have explored Poisson regression for modelling count outcomes. Overdispersion, due to omitted explanatory variables, is a common issue when modelling count data. Formally modelling such omitted effects, for example through negative binomial regression, is one approach to the problem. An alternative, when the overdispersion is considered merely a nuisance factor, is to use robust sandwich variance estimates. A drawback of the latter is that because we no longer have a full model for the data, we cannot use the model to give (valid) predictions.

We conclude by noting that there are a number of other issues that can arise in modelling count data which we have not covered. One which occurs fairly commonly is to have an excess of zeroes relative to what would be expected under a Poisson (or even negative binomial) model. One approach to modelling such phenomena is to use a zero-inflated Poisson model. For further information on this, see Chapter 15 of the Stata press book 'Generalized Linear Models and Extensions', by Hardin and Hilbe.

8.7 Practical 8

Dataset required: `medpar.dta` (from Stata website)

Introduction

In this practical we will explore models for a count variable and look at the two approaches described in the lecture for handling overdispersion. We will use publicly available data on length of hospital stay from Arizona in the US.

In our analyses we will focus on the following variables.

Variable	Description
`los`	Length of hospital stay, in days
`age`	Age group (factor variable)
`type1`	Binary variable indicating elective admission
`type2`	Binary variable indicating urgent admission
`type3`	Binary variable indicating emergency admission

Note that admission type is here described using three binary indicator (or dummy) variables, rather than as a 3-level categorical variable.

Aims

1 Understand how to fit models to count data using the glm command (in Stata).

2 Understand how to compare such models (in Stata).

3 Understand the concept of overdispersion, and how to identify it.

Analysis

The dataset is available from the Stata Press website; to load the data type:

```
use http://www.stata-press.com/data/hh3/medpar, clear
```

1 Explore the length of stay variable. How is it distributed? What are the minimum and maximum values in this dataset?

   Construct a 95% confidence interval for the mean length of stay using the standard approach taught in the Analytical Techniques course (a Wald 95% confidence interval).

2 Use the `glm` command to fit a Poisson model for length of stay, with no covariates.

   What does the constant coefficient in this model represent? Construct a 95% confidence interval for the mean length of stay.

3 Repeat question 2 using the `robust` option with the `glm` command.

Discuss: Compare and contrast the 95% confidence intervals in questions 1, 2 and 3. Which is most appropriate? Which is least appropriate?

4 We will now explore which (if any) factors are related to length of hospital stay. First fit a Poisson model (without robust standard errors) to assess whether length of hospital stay is related to the type of admission. Interpret each of the parameter estimates in your model.

5 Add age as factor variable to your model in question 4 and perform a likelihood ratio test of whether age (treated as categorical) is a predictor of length of hospital stay, adjusting for type of admission. Is this test an appropriate one?

6 The likelihood ratio test cannot be used with robust standard errors. However, we can still use a (multivariate) Wald test, which is based on the robust variance-covariance matrix, to assess if age group has an effect on length of stay, adjusting for admission type. Refit the model in question 5 with robust standard errors, and then perform this Wald test using the following command.

```
testparm i.age
```

Discuss: Compare and contrast the results of the tests in questions 5 and 6. Which is the appropriate test?

Working together with one or more colleagues (in your Breakout Room if online), write a short paragraph (suitable for a medical journal) to summarise your findings concerning the effects of the type of admission on length of stay in hospital for this model. If online, one of you should post your group's paragraph in the Zoom chat.

7 Use `nbreg` to fit a negative binomial regression model to the length of stay variable, with no covariates. Interpret the likelihood ratio test reported at the bottom of the output, and relate it to your findings from earlier questions.

8 Fit a negative binomial regression model with indicator variables for type of admission as the only predictor variables. Use the estimated parameters to calculate the predicted mean and variance of lengths of stay for each of the three types of admission. To do this you will to use the negative binomial regression model result that when the expectation of the outcome is μ then its variance is $\mu(1+\alpha\mu)$ (see section 8.5.2 of the notes). Compare the predicted variances with the observed variances for each admission type.

Discuss: What do you conclude about the estimates and variances predicted from the negative binomial model? What are your conclusions about the best way to model these data?

Session 9

Models for rates

The aim of this session is to introduce you to Poisson models for rates.

At the end of this session, you should understand the key concepts of Poisson rate modelling, and be able to fit such models in Stata.

9.1 Examples of rates

Rates occur in many medical problems. Examples include:

- Rate of exacerbations/complications in clinical trials: e.g. asthma exacerbations.
- Mortality or incidence rates, e.g.
  - coronary heart disease among doctors
  - Mortality rate of men working in the rubber-manufacturing industry.

Example 9.1 *British doctors study*

agegrp	smokes	deaths	pyrs
1	1	32	52407
2	1	104	43248
3	1	206	28612
4	1	186	12663
5	1	102	5317
1	0	2	18790
2	0	12	10673
3	0	28	5710
4	0	28	2585
5	0	31	1462

Table 9.1: Data from study of effect of smoking on coronary heart disease in British male doctors. Details in the text.

Table 9.1 shows data from a famous cohort study that was used to investigate the effect of smoking on coronary heart disease (CHD) among male British doctors. The variable `smokes` is 1 for smoker and 0 for non-smoker. The variable `agegrp` has categories:

1: 35–44 years

2: 45–54 years

3: 55–64 years

4: 65–74 years

5: 75+ years

9.2 Construction of Poisson frequency records

The data in table 9.1 are stored in Poisson frequency records, showing the number of deaths and the number of person-years of follow-up in different categories of age and smoking. However, the data collected would have been for each individual, and records may have originally looked like this:

id	agein	smokes	doe	dox	CHD
101	44	1	1/2/54	3/8/71	1
102	51	1	3/9/58	5/10/69	0

At first sight, we might consider modelling such data using logistic regression with `CHD` as the outcome variable. However, such an approach cannot be used because each subject has been followed up for a different length of time, and of course the probability of experiencing an event depends strongly on this.

The crude rates can be calculated by assuming that the true rate does not vary during follow-up.

Exercise 9.1 *Calculation of Poisson rates*

Complete Table 9.2 by calculating the crude rate of CHD in smokers and non-smokers for the British doctors study (ignoring age group).

Group	Person-years of follow-up	CHD deaths	Death rate per 1000 person-years	Rate ratio
Non-smokers				1
Smokers				

Table 9.2: Death rates due to CHD in smokers and non-smokers, collapsed over age group.

9.3 Poisson process

Suppose that Y is a random variable (r.v.) representing the number of events (e.g. deaths) during a specified time period of length t. Assume that

- events occur independently, so that events during non-overlapping time periods are independent r.v.'s;
- the probability of an event occurring in a very short time interval δt is $\lambda \times \delta t$ in the limit as δt tends to 0.

under these assumptions it can be shown (see Probability sessions) that the number of events in time t, denoted by the r.v. Y, has a Poisson distribution

$$Y \sim Po(\mu)$$

where $\mu = \lambda t$ and λ is the **rate** (sometimes called the 'force of mortality').

9.4 Models for rates

Consider events which occur independently in periods of time t_i with rates λ_i. Then the r.v.'s Y_i which represent the numbers of events in periods of time t_i have Poisson distributions, with means $\mu_i = \lambda_i t_i$.

The Poisson distribution is a member of the exponential family, so the mean μ_i can be modelled through a generalized linear model using a linear predictor of p explanatory variables $x_{i1}, \ldots, x_{ip}$ via a suitable link function.

The log function is nearly always used with the Poisson distribution:

- it maps positive values of μ to the whole real line for the linear predictor;
- parameters are easily interpreted in terms of **multiplicative** effects on the scale of the rates;
- it is the natural (or canonical) parameterization for the Poisson distribution.

The model we are interested in is one for the **rates** λ_i, and takes the form:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

However, for the generalized linear model we need to express the linear predictor in terms of the mean $\mu_i = \lambda_i t_i$. Using $\lambda_i = \mu_i / t_i$ we have

$$\log(\mu_i) - \log(t_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

or

$$\log(\mu_i) = \log(t_i) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

Hence, in the model for μ_i the term involving the log-person years is in the linear predictor. This term is called an **offset**: it is a predictor variable whose regression coefficient is forced to be equal to one. By using an offset in this way we are able to use a Poisson GLM to fit models for rates.

For the `glm` command in Stata, the variable containing $\log(t_i)$ is passed using the `offset` option.

9.5 GLM for rates

We have now arrived at the GLM for rates. This has the following three components:

1 *Distribution:* independent response variables

$$Y_i \sim Po(\mu_i)$$

where Y_i has expected value $\mu_i = \lambda_i t_i$, for t_i the observed person-time in group i.

2 *Linear Predictor:* explanatory variables whose measured values for group i are $x_{i1}, \ldots, x_{ip}$, together with an 'offset' $\ln(t_i)$ with fixed coefficient 1, form the linear predictor

$$\eta_i = \log(t_i) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

3 *Link function:* the log function is used, giving

$$\log(\mu_i) = \eta_i.$$

As for grouped binomial data, the model deviance can be used to assess goodness of fit for models for Poisson frequency records. Under the null hypothesis that the current model is correctly specified, the deviance will follow a χ^2 distribution on $n-p$ d.f., where n is the number of observations (and therefore number of parameters in the saturated model) and p is the number of parameters in the current model. In this setting, Pearson's statistic can alternatively be used, again compared to a χ^2 distribution on $n-p$ d.f..

9.6 Example: British doctors study

We arrange the data as in Table 9.1. The number of deaths in group i is given by y_i, $(i = 1, \ldots, 10)$, the person-years by t_i. We let x_{i1} denote an indicator variable taking the value 1 if group i consists of smokers and 0 otherwise. We let x_{i2}, x_{i3}, x_{i4} and x_{i5} denote indicators taking the value 1 if group i is the 2nd, 3rd, 4th and 5th age categories respectively.

The analysis aims are:

1 to investigate the effect of smoking on CHD rate, without adjustment for age;

2 to investigate the effect of smoking on CHD rate, after adjustment for age, and

3 to investigate an interaction between the effects of smoking and age on CHD rate.

Model 1: smoking
We start with a model relating CHD rate to smoking alone. This model has the following linear predictor.

$$\log(\mu_i) = \log(t_i) + \beta_0 + \beta_1 x_{i1}$$

Fitting this model in Stata gives the following.

```
. glm deaths i.smokes, family(poisson) link(log) offs(log_pyrs)

Generalized linear models                         No. of obs      =         10
Optimization     : ML                             Residual df     =          8
                                                  Scale parameter =          1
Deviance         =   905.976178                   (1/df) Deviance =    113.247
Pearson          =  1155.096377                   (1/df) Pearson  =    144.387

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   96.50441
Log likelihood   = -480.5220592                   BIC             =   887.5555

------------------------------------------------------------------------------
             |                 OIM
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.smokes |   .5422209   .1071834     5.06   0.000     .3321452    .7522966
       _cons |  -5.961821   .0995037   -59.92   0.000    -6.156845   -5.766798
    log_pyrs |          1  (offset)
------------------------------------------------------------------------------
```

The interpretation of the parameter estimates is as follows:

- $\hat{\beta}_0 = -5.96$: the estimated log rate for non-smokers.
- $\hat{\beta}_1 = 0.547$: the estimated difference in log rate between non-smokers and smokers.

Looking at the deviance statistic, we see that it is dramatically larger than the residual d.f., indicating that our current model is a poor fit to the data. Here this can be attributed to the fact that we have omitted an important covariate (age), which strongly affects rates. Due to the fact we have omitted age (so far), the observed counts vary around that predicted on the basis of smoking status alone by far more than would be expected under the Poisson assumption.

Although unlikely, note that if we were only concerned with modelling the outcome as a function of smoking, we could collapse the data into two observations, by pooling the number of events and person years across the five age categories. We could then fit the same Poisson model, which for this collapsed data is the saturated model. This gives identical inferences to the analysis shown in the Stata output above, but has different values for the deviance and degrees of freedom.

Model 2: smokes + age group
This model has linear predictor:

$$\log(\mu_i) = \ln(t_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}$$

Fitting this model in Stata gives the following.

```
. glm deaths i.smokes i.agecat, family(poisson) link(log) offs(log_pyrs)

** <output omitted> **

Generalized linear models                         No. of obs      =         10
Optimization     : ML                             Residual df     =          4
                                                  Scale parameter =          1
Deviance         =   12.1323693                   (1/df) Deviance =   3.033092
Pearson          =  11.15533736                   (1/df) Pearson  =   2.788834

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   7.920031
Log likelihood   = -33.60015489                   BIC             =   2.922029

------------------------------------------------------------------------------
             |                 OIM
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.smokes |   .3545354   .1073741     3.30   0.001      .144086    .5649848
             |
      agecat |
      45-54  |   1.484007   .1951034     7.61   0.000     1.101611    1.866402
      55-64  |   2.627505   .1837273    14.30   0.000     2.267406    2.987604
      65-74  |   3.350493   .1847992    18.13   0.000     2.988293    3.712692
        75+  |   3.700096   .1922195    19.25   0.000     3.323353     4.07684
             |
       _cons |  -7.919326   .1917618   -41.30   0.000    -8.295172   -7.543479
    log_pyrs |          1  (offset)
------------------------------------------------------------------------------
```

The age-adjusted rate ratio, comparing smokers with non-smokers, is

$$e^{0.3545} = 1.43 \text{ with } 95\% \text{ CI } \left(e^{0.3545-1.96\times0.1074}, e^{0.3545+1.96\times0.1074}\right) = (1.16,\ 1.76).$$

From the Wald test for the smoking covariate, we have strong evidence of an effect of smoking on rates, after adjusting for age group. However, comparing the deviance (12.132) to χ^2 on 4 d.f., we obtain $p = 0.016$: there is evidence that the model is not correctly specified. Recall that this test is the comparison of the fit of the current model with the saturated model. Here the saturated model is the model which includes an interaction between smoking and (categorical) age. We thus have evidence that an interaction is needed.

Note that an alternative modelling approach for this data would be to model the age effect as a continuous covariate. This would have the advantage (should the model be appropriate for the data) of giving a simpler description of how rates vary with age and smoking.

Also note that we can also fit Poisson rate models in Stata using the `poisson` command. In this case we do not need to specify distribution or link. Neither do we need to calculate the log of the person years at risk. We simply put in the person years at risk as the *exposure* using the `e()` option. Check the `help` for details.

Model 3: smokes + age group + their interaction

Fitting this model in Stata gives the following.

```
. glm deaths i.smokes##i.agecat, family(poisson) link(log) offs(log_pyrs)

Generalized linear models                         No. of obs      =         10
Optimization     : ML                             Residual df     =          0
                                                  Scale parameter =          1
Deviance         =  3.03513e-13                   (1/df) Deviance =          .
Pearson          =  3.29314e-13                   (1/df) Pearson  =          .

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

AIC              =  7.506794
Log likelihood   = -27.53397024                   BIC             =  3.04e-13

------------------------------------------------------------------------------
             |                 OIM
      deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.smokes |   1.746872   .7288688     2.40   0.017     .3183158    3.175429
             |
      agecat |
      45-54  |   2.357366   .7637625     3.09   0.002     .8604192    3.854313
      55-64  |   3.830163   .7319249     5.23   0.000     2.395616    5.264709
      65-74  |   4.622656   .7319249     6.32   0.000     3.188109    6.057202
        75+  |   5.294359     .72956     7.26   0.000     3.864447     6.72427
             |
smokes#agecat |
    1#45-54  |  -.9866221   .7900623    -1.25   0.212    -2.535116    .5618717
    1#55-64  |  -1.362808   .7561868    -1.80   0.072    -2.844907    .1192905
    1#65-74  |  -1.442289   .7565318    -1.91   0.057    -2.925064    .0404861
    1#75+    |  -1.846991   .7571736    -2.44   0.015    -3.331024   -.3629583
             |
       _cons |  -9.147932   .7071066   -12.94   0.000    -10.53384   -7.762029
    log_pyrs |          1  (offset)
------------------------------------------------------------------------------
```

Note that the deviance reported here is zero, because this model is saturated.

EXERCISE 9.2 *Interpretation of saturated model*

Describe how the effect of smoking on CHD rate varies with age.

9.7 Summary

We have seen how a Poisson GLM can be used to model rate data, which occurs when we count the number of events over follow-up periods which vary between subjects. In this setting use of logistic regression is inappropriate because such an analysis would not account for how long each subject has been followed up for.

In this session we have focused on grouped Poisson data, which arise when we fit regression models for such data with only categorical covariates (or continuous covariates which take

only a small number of values). Often we may have individual level covariates which are continuous. This leads us to a Poisson GLM where each subject has their own offset value, and the dependent variable is the binary variable indicating whether they experienced the event or not by the end of follow-up. This model is equivalent to the *exponential* survival model: named because under the Poisson assumption the time between events (or from time zero to an event) follows an exponential distribution (see the Survival module for more details).

9.8 Practical 9

Dataset required: `rubber.dta`

Introduction

The purpose of this exercise is to fit and interpret some Poisson regression models for data from a cohort study of men from the rubber manufacturing industry.

In one rubber factory isolated conditions were maintained for the manufacturing process, while in a second workers were exposed to more dirt and fumes. The aim is to investigate differences in death rates between men who worked in the two factories, making appropriate allowances for age.

The dataset (rubber.dta) consists of four variables for groups defined by factory and age category:

Variable	Description
`agegrp`	Age group: 1=50-59, 2=60-69, 3=70-79, 4=80-89
`factory`	Factory (1 or 2)
`deaths`	Number of deaths
`pyrs`	Number of man-years of exposure

Aims

The aim of this session is to address the following questions about these data:

1 Is there a difference between deaths rates in the two factories?

2 Is death rate associated with age?

3 What is the best way to model age (categorical or continuous)?

Analysis

1 Calculate the total number of deaths and total person years of observation. What is the overall rate of death amongst these men? What is the log death rate?

2 Calculate the death rate and log death rate in each factory.

3 In Stata, generate a variable for the log death rates in each row, and plot these against age group, labelled according to factory (hint: use the option `mlabel([varname])` with the `twoway` command). What does this plot suggest?

Discuss: What are your initial conclusions from these descriptions of the data?

4 Write down algebraically a model that can be used to investigate the relationship between death rate and age group (treated as a categorical variable, with age group 50-59 taken as the baseline category).

Discuss: Take a moment to check with your colleagues that you have specified the models in the same way.

5 Using Stata:

(a) obtain the maximum likelihood estimates of the parameters in your model. How strong is the evidence for an age effect?

(b) calculate the estimated rate ratios comparing age groups:

i. 60-69 with 50-59

ii. 80-89 with 70-79

Calculate 95% CI's for these estimates, using `lincom` (or by re-parameterising the model) for ii.

6 Now include both age group and factory effects in an additive linear predictor. What is the evidence for a difference in death rates between the factories (adjusted for age group)? Calculate the estimate (and 95% CI) of the rate ratio comparing the two factories, adjusted for age group.

7 Now fit a model that includes the interaction between age group and factory.

(a) What term can be used to describe this model (Hint: look at its deviance)?

(b) What is the evidence for an interaction? How do you interpret this for these data?

(c) Write down algebraically the linear predictor for this model, and use this to help you calculate the estimated rates of death in

i. factory 1, age 70-79

ii. factory 2, age 50-59

iii. factory 2, age 60-69

(d) Check your answers against your original data

Discuss: Take a moment to check with your colleagues that you have specified the models in the same way.

8 Now consider the age group variable as continuous. Fit a model with age group and factory as covariates and consider its deviance.

Discuss: What do you conclude about the overall fit of the model with continuous age and factory as additive effects? How might the fit be improved?

9 Now consider your analysis as a whole.

Discuss: What are your epidemiological conclusions from this analysis? Working together with one or more colleagues (in your Breakout Room if online), write a short paragraph to summarise your findings concerning the comparison of the two factories. If online, one of you should post your group's paragraph in the Zoom chat.

Session 10

Analysis Strategies

10.1 Aims

The aim of this session is to discuss strategies for building statistical models for data. The objectives are for you to be able to do the following:

- Appreciate that different investigation types can have different aims and understand why these require different analysis strategies.
- Describe considerations for variable selection in investigations of causal effects.
- Describe two variable selection procedures for investigations of causal effects which can be used when there are many possible adjustment variables.
- Describe variable selection procedures for investigations where the aim is prediction of an outcome
- Understand the limitations of variable selection methods

10.2 The need for an analysis strategy

In session 2 we discussed different types of investigation. In this session we focus on two types of investigation:

- Investigations where the aim is to estimate the causal effect of a treatment or exposure on an outcome
- Investigations where the aim is to build a predictive model for an outcome

In studies of causal effects the aim is to estimate the effect of a single variable (in our context) on the outcome. We are therefore concerned about confounding and we want to adjust for variables that confound the association of interest. The focus is on the parameter that quantifies the causal effect - we want to estimate this parameter with little or no bias, but to balance this with obtaining a precise estimate.

In prediction studies the focus is on how well a set of covariates predicts the outcome, rather than on any particular model parameter (or function thereof) such as a mean difference or an odds ratio. We do not need to consider confounding in prediction investigations.

In both types of investigation there are often many potential variables at play and it is necessary to have strategies which help us to decide, in a structured way, which variables are to included in the final model. Strategies typically involve comparison of several different models (including different sets of covariates) based on criteria that are focused on the aims of the investigation. The process used arrive at a final model for a given investigation is called 'variable selection' or 'model building', with the latter being a more

general term. It is important to emphasise that the aims of the two investigation types mean that the considerations for building model are different.

The task of model building is a complex one, and there are many different possible strategies. Typically the process involves a degree of subjectivity, and moreover different epidemiologists or statisticians may have different views on what an appropriate process is in any given situation. In this session we cover some simple possible approaches. However, this should not be taken to mean that other approaches are (necessarily) wrong or ill advised. There are limitations to the approaches we cover, which are commented on throughout.

10.3 Motivating example: vitamin C study

In this session we will use data from a study of serum vitamin C (the **vitc.dta** dataset analysed in the Regression 5 practical last term). The data are from an observational study of 92 individuals and we have the following variables: **seruvitc** (serum vitamin C level (micromol/l)), **age** (age of subject in years), **height** (height in metres), **cigs** (0=non-smoker; 1=smoker), **weight** (weight in kg), **sex** (0=male; 1=female) and **ctakers** (0 = Vitamin C supplement non-user; 0 = Vitamin C supplement user).

We will consider two types of investigations using these data. In the first type of investigation our interest is in the causal effect of vitamin C supplement use on serum vitamin C, with the other covariates considered as potential confounders. In the second type of investigation our aim is to predict serum vitamin C using some or all of the other variables.

The dataset is available on the U Drive.

10.4 Model building for causal investigations: popular methods

We begin by focusing on model building for causal investigations. The discussion here is informed by the papers of Greenland and Pearce (2015), Greenland et al. (2016) and Vansteelandt et al. (2010).

The first step is to identify the pool of variables to be considered for inclusion in the model, in addition to the main exposure (vitamin C supplement use in the example). Use of causal diagrams (directed acyclic graphs - DAGs) is strongly recommended to help with this process. DAGs were discussed in Sessions 2 and 5. There will be certain variables that we definitely want to include due to subject matter knowledge. For example, these may include sex and age. There may be some variables that we know we definitely don't want to include because they are on the causal pathway between the exposure and outcome, i.e. they are mediators. We are often left, however, with a set of variables ('potential confounders') that we are are unsure whether we wish to include in the model or not. The decision to include covariates in a regression model is ideally based on the strength of evidence for these covariates confounding the association between the main exposure of interest and outcome.

We let Y denote the outcome and X_1 denote the main exposure, whose effect on Y we wish to estimate. The other covariates are denoted $X_2, \ldots, X_p$. We assume that variables we definitely don't want to include (e.g mediators) are excluded from this set.

Commonly used variable selection strategies used in this types of investigations are:

1 Adjust for all variables identified as potential confounders - we call this the 'full model'.

2 Use stepwise selection methods: Selecting covariates on the basis of some measure of their ability to predict outcome or exposure (or both) given other covariates in the model.

3 The 'change in estimates' approach: select covariates that change the exposure effect estimate when included in the model.

Let's discuss these in turn.

10.4.1 Adjusting for all covariates

Adjusting for all variables identified as potential confounders is a viable approach when the sample size is large and/or the pool of covariates is fairly small. This is quite often a sensible strategy. It is a 'safe' method because it means that we adjust for all *potential* confounders that we have measured. We should always be aware that there may be other variables that confound the exposure-outcome association but we which we not know about or have not measured/are not recorded in our data. A large sample size brings the luxury of being able to include many covariates in our regression model without too much loss of precision (for estimating the causal effect of interest).

In some settings the pool of potential covariates that we may need to adjust for can be very large, relative to the sample size. The 'rule of thumb' of requiring 10 subjects per variable included in the model (or 10 cases in the situation of a binary outcome) has often been quoted, but among experts in variable selection this rule is considered crude as it does not take into account many important features of the data.

10.4.2 Stepwise selection methods

Stepwise selection strategies are automatic approaches which select a set of predictive covariates from a large group of candidate variables. They come in different forms, including 'backwards elimination' and 'forwards selection'. The backwards elimination procedure starts with a full model (i.e. including all covariates), and then eliminates the least significant variables one at a time until all those remaining have a $p-$value greater than the a pre-set '$p-$value to retain'. In this setting of a causal investigation, the main exposure X_1 is forced to be kept in the model. The procedure for backwards elimination is as follows:

1 Regress the outcome Y on all explanatory variables together, $X_1, X_2, \ldots, X_p$.

2 Eliminate the least significant covariate out of $X_2, \ldots, X_p$ (i.e. the one with the largest p-value) if its $p-$value is greater than the pre-selected constant (e.g. $p = 0.2$).

3 Continue eliminating one variable at a time until all variables remaining in the model have $p-$values smaller than the pre-selected constant.

The object of stepwise regression is to obtain a parsimonious model by searching for a model that explains the most variation in the outcome with the fewest variables. At each step of the backwards selection procedure a variable may be dropped from the model if its removal does not increase the residual variation (according to the pre-defined p-value). Therefore, stepwise methods are focused on overall model fit rather than on identifying and retaining confounders. They may therefore result in important confounders being omitted

from the model. Stepwise methods can result in retaining non-confounders or weak confounders preferentially over strong confounders, particularly where the non-confounders or weak confounders are more strongly predictive of the outcome.

Stepwise methods are arguably more suitable for use prediction investigations, though have also come in for serious criticism in that context too for reasons which we will discuss further below.

10.4.3 Change in estimates method

The stepwise methods do not target the aim of the investigation, which is to remove (or reduce) confounding. The change in estimates approach is a preferable alternative that does target this aim. The basic idea is that if a variable is a confounder then the coefficient for the main exposure will be different in models including and excluding the confounding. The change in variables approach assesses covariates in terms of whether their inclusion changes the coefficient for the main exposure by a degree judged to be 'important'. Often it is investigated whether the coefficient for the main exposure changes by 10% or more.

A bit like in the stepwise regression approaches, the change in variable approach can be implemented in different ways, including using 'backwards' and 'forwards' procedures. In the backwards selection approach the steps are:

1 Fit the full model including the main exposure X_1 plus all potential confounders $X_2, \ldots, X_p$.

2 Remove each covariate $(X_2, \ldots, X_p)$ individually from the model and record how much the coefficient for the main exposure changes. If there are any covariates whose removal results in a change in the coefficient for the main exposure of less than 10%, remove the covariate that changes the coefficient by the smallest amount.

3 Starting with the model from step two, remove each remaining covariate in turn and exclude the covariate whose removal that has the least impact on the main exposure coefficient (provided it is less than 10%).

4 Repeat until there are no more covariates whose removal changes the coefficient by less than 10%.

In the forwards selection approach the steps are:

1 Fit a model including just the main exposure.

2 Add each covariate individually into the model and record how much the coefficient for the main exposure changes. Keep the covariate that changes the coefficient for the main exposure the most (provided it is greater than some preset amount, typically 10%).

3 Starting with the model from step two, which now includes two explanatory variables (including the main exposure) include each remaining covariate in turn and record how much the coefficient for the main exposure changes. Add the covariate that changes the coefficient the most into the model (provided it is greater than 10%).

4 Repeat until there are no more covariates to be added that change the coefficient for the main exposure by 10%.

More complicated stepwise procedures could be used in which variables are added and removed. For example in the forwards selection approach, certain variables may become unimportant once other variables are added, so additional steps are included in which covariates are removed as well as added (forwards stepwise selection). Different procedures can result in different sets of covariates being included in the final model.

The change in estimates approach has some attraction and has become popular. However, it also has some limitations:

- One criticism is that it focuses on change in the point estimate for the main exposure and does not consider uncertainty in the estimate (though other versions have considered changes in confidence limits instead of change in estimates).

- Considering a change of 10%, or some other value, is arbitrary, and the decision as to what constitutes a meaningful change should ideally be based on the context.

- Sometimes it may be sets of covariates that are important rather than individual covariates.

- The analyst should take care to notice whether including certain variables results in a large increase in the standard error for the main effect.

- A major drawback of this approach is that it does not account for non-collapsibility. The concept of non-collapsibility was introduced in session 6. Including or excluding covariates in logistic regression will result in a change in the log odds ratio estimate for the main exposure even if the covariate is not a confounder, because odds ratios are non-collapsible. As noted in the earlier session this is not a major issue in certain settings, including when the outcome is rare. Hence the change in variables approach may still be approximately appropriate for use in logistic regression in some settings. In other settings with a binary outcome it would be preferable to consider changes in marginal effect estimates (obtained through standardization) instead of conditional estimates, and perform the change in variables strategy on the basis of changes in the marginal effect estimate.

10.4.4 Example

We apply the methods outlined above to the vitamin C data. In this example,the aim is to estimate the causal effect of vitamin C supplement use on serum vitamin C, measured using a mean difference. The boxes below show the results from a linear regression of serum vitamin C on supplement use only, and a linear regression of serum vitamin C on supplement use plus all the other covariates. In the univariable model, the mean difference in serum vitamin C in the supplement users compared with the non-users is 22.09, and there is strong evidence against the null hypothesis of no association (95% CI (10.73,33.45)). After adjusting for all of the other variables the effect estimates reduces to 19.86 (95% CI (8.83,30.89)). This suggests there may be some (perhaps weak) confounding of the association between supplement use and serum vitamin C by the other covariates.

```
. regress seruvitc i.ctakers

      Source |       SS           df       MS      Number of obs   =        92
-------------+----------------------------------   F(1, 90)        =     14.94
       Model |  7358.65215         1  7358.65215   Prob > F        =    0.0002
    Residual |  44334.4239        90   492.60471   R-squared       =    0.1424
-------------+----------------------------------   Adj R-squared   =    0.1328
       Total |  51693.0761        91  568.055781   Root MSE        =    22.195

------------------------------------------------------------------------------
    seruvitc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ctakers |
        yes  |   22.09301   5.716168     3.87   0.000     10.73684    33.44917
       _cons |   48.64384   2.597693    18.73   0.000     43.48306    53.80461
------------------------------------------------------------------------------
```

```
. regress seruvitc i.ctakers age height i.cigs weight i.sex

      Source |       SS           df       MS      Number of obs   =        91
-------------+----------------------------------   F(6, 84)        =      4.25
       Model |  11547.5668         6  1924.59447   Prob > F        =    0.0009
    Residual |  38079.4222        84  453.326455   R-squared       =    0.2327
-------------+----------------------------------   Adj R-squared   =    0.1779
       Total |   49626.989        90  551.410989   Root MSE        =    21.291

------------------------------------------------------------------------------
    seruvitc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ctakers |
        yes  |   19.85927   5.547339     3.58   0.001     8.827777    30.89076
         age |  -.3533428   .8397259    -0.42   0.675     -2.02323    1.316544
      height |   -1.57296   42.25903    -0.04   0.970    -85.60968    82.46376
             |
        cigs |
        yes  |  -11.79955   6.757432    -1.75   0.084    -25.23745    1.638341
      weight |   .1072991   .2230579     0.48   0.632     -.336276    .5508742
       1.sex |   10.36657   7.284816     1.42   0.158    -4.120082    24.85323
       _cons |    65.3516   89.48158     0.73   0.467    -112.5923    243.2955
------------------------------------------------------------------------------
```

The stepwise methods can be performed in an automated way in Stata. The results from backwards elimination and forwards selection procedures are shown below. For this we are using a p-value of 0.2 to retain or add a given covariate. Note that the 'lockterm1' option is used to force the main exposure (ctakers) into the model. In this example the backwards and forwards procedures result in the same model, including sex and smoking status. However, they are not guaranteed to result in the same model. Changing the p-value also impacts on which variables are in the final model of course. The coefficient for supplement use in the model chosen in the stepwise selection procedures in 19.83, which is very close to that from the full model. The 95% CI is slightly narrower.

```
. xi: stepwise, pr(0.2) lockterm1: regress seruvitc i.ctakers age height i.cigs
> weight i.sex
i.ctakers         _Ictakers_0-1       (naturally coded; _Ictakers_0 omitted)
i.cigs            _Icigs_0-1          (naturally coded; _Icigs_0 omitted)
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)
                      begin with full model
p = 0.9704 >= 0.2000  removing height
p = 0.6715 >= 0.2000  removing age
p = 0.4486 >= 0.2000  removing weight

      Source |       SS           df       MS      Number of obs   =        91
-------------+----------------------------------   F(3, 87)        =      8.46
       Model |  11208.6598         3  3736.21995   Prob > F        =    0.0001
    Residual |  38418.3292        87   441.58999   R-squared       =    0.2259
-------------+----------------------------------   Adj R-squared   =    0.1992
       Total |   49626.989        90  551.410989   Root MSE        =    21.014

------------------------------------------------------------------------------
    seruvitc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Ictakers_1 |   19.83167   5.452864     3.64   0.000     8.993509    30.66982
    _Icigs_1 |    -12.255   6.588755    -1.86   0.066    -25.35086    .8408641
     _Isex_1 |   9.764167   4.487498     2.18   0.032      .844779    18.68355
       _cons |   46.02833   3.551566    12.96   0.000     38.96921    53.08746
------------------------------------------------------------------------------
```

```
. xi: stepwise, forward pe(0.2) lockterm1: regress seruvitc i.ctakers age height
>  i.cigs weight i.sex
i.ctakers         _Ictakers_0-1       (naturally coded; _Ictakers_0 omitted)
i.cigs            _Icigs_0-1          (naturally coded; _Icigs_0 omitted)
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)
                      begin with term 1 model
p = 0.0165 <  0.2000  adding  _Isex_1
p = 0.0663 <  0.2000  adding  _Icigs_1

      Source |       SS           df       MS      Number of obs   =        91
-------------+----------------------------------   F(3, 87)        =      8.46
       Model |  11208.6598         3  3736.21995   Prob > F        =    0.0001
    Residual |  38418.3292        87   441.58999   R-squared       =    0.2259
-------------+----------------------------------   Adj R-squared   =    0.1992
       Total |   49626.989        90  551.410989   Root MSE        =    21.014

------------------------------------------------------------------------------
    seruvitc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Ictakers_1 |   19.83167   5.452864     3.64   0.000     8.993509    30.66982
     _Isex_1 |   9.764167   4.487498     2.18   0.032      .844779    18.68355
    _Icigs_1 |    -12.255   6.588755    -1.86   0.066    -25.35086    .8408641
       _cons |   46.02833   3.551566    12.96   0.000     38.96921    53.08746
------------------------------------------------------------------------------
```

There is not (to my knowledge) an automated way of performing the change in estimates approach. However, with a small number of variables in this example it is easy enough to perform ourselves. A backwards change in estimates procedure was performed. This starts with the full model, as shown above. Each covariate is then eliminated one by one and the coefficient for supplement use recorded. The effect estimate obtained from the full model is 19.86. After excluding each variable in turn the coefficient for supplement use was:

- Excluding age: 19.89
- Excluding height: 19.84
- Excluding smoking status: 20.08
- Excluding weight: 19.71
- Excluding sex: 20.86

The variable whose removal changes the coefficient the least is height. The change is less than 10% (comparing 19.84 with 19.86). So we exclude height and go though the elimination procedure again. At the next step age is removed, followed by smoking status, then weight, then sex. In this case all covariates are eventually eliminated using this procedure and we are left with the univariable model.

10.5 Model building for causal investigations: A strategy based on mean squared error

Greenland et al. (2016) described an approach to confounder selection based on the mean squared error (MSE). Consider a parameter representing a causal effect β^* - for example, this may represent a mean difference, a risk difference, a log risk ratio, or a log odds ratio. And suppose that we can estimate that effect by fitting a regression model - we denote the estimator β. The MSE is the expected squared difference between the true value β^* and β:

$$MSE = E\left\{(\beta - \beta^*)^2\right\} \tag{10.1}$$

It can be shown that the MSE can also be expressed as

$$\begin{aligned} MSE =& E\left\{(\beta - \beta^*\right\}^2 + E\left\{(\beta^* - E(\beta)^2\right\} \\ =& \text{Bias}(\beta)^2 + SE(\beta)^2 \end{aligned} \tag{10.2}$$

where $\text{Bias}(\beta)$ denotes the bias in the estimator β and $SE(\beta)$ denotes its standard error. Thus, the MSE is a function of both bias and variance. A low MSE is a desirable feature of an estimator. This model building strategy focuses on reducing the standard error of an estimator, but not at the expense of serious bias.

Now suppose that we have a full model that adjusts for all covariates. The coefficient from this model that denotes the causal effect estimator is denoted β_{full}, with standard error $SE(\beta_{\text{full}})$. The full model is assumed to include all confounders and hence the estimator β_{full} is assumed unbiased: $\text{Bias}(\beta_{\text{full}}) = 0$. The MSE for the full model is therefore $MSE_{\text{full}} = SE(\beta_{\text{full}})^2$. The MSE for the full model could be large therefore if the effect is estimated imprecisely, i.e. with large standard error.

Now consider a model with a reduced set of covariates being adjusted for. The coefficient from this model that denotes the causal effect estimator is denoted β_{red}, with standard

error $SE(\beta_{\text{red}})$. The MSE for the reduced model is $MSE_{\text{red}} = \text{Bias}(\beta_{\text{red}})^2 + SE(\beta_{\text{red}})^2$. The bias can be estimated using $\text{Bias}(\beta_{\text{red}}) = \hat{\beta}_{full} - \hat{\beta}_{red}$, where $\hat{\beta}_{full}$ and $\hat{\beta}_{red}$ are the estimates from the full model and the reduced model respectively.

A backwards elimination algorithm based on the MSE is as follows:

1 Fit the full model and obtain the MSE, MSE_{full}

2 Fit a series a reduced models, omitting each covariate in turn, and calculate the MSE for each reduced model, MSE_{red}. Assess whether any model reduces the MSE compared to the full model. If so, omit the covariate associated with the biggest reduction in the MSE.

3 Starting with the reduced model from step 2, again omit each remaining covariate in turn and assess whether any of the further reduced models gives a further reduction in the MSE.

4 Repeat until there are no further reductions in the MSE

There is also a forwards selection equivalent. Other variations on this approach have also been described. Vansteelandt et al. (2010) described a more sophisticated but more difficult and computationally intensive model building approach based on the MSE.

10.5.1 Example

There is no automated way of performing this procedure in Stata or other software packages, as far as we are aware. We performed the backwards MSE approach on the vitamin C data. Stata code for this example is provided on Moodle.

The MSE for the full model is 30.773. After excluding each of the following variables individuals, the MSE in the reduced model was: age (30.471), height (30.169), smoking status (31.548), weight (30.424), sex (31.65). So omitting age, height and weight all resulted in a small reduction in the MSE, with the removal of height giving the biggest reduction. So we exclude height from the model and run the procedure again. At the next step we exclude age, followed by weight. We are left with a model including smoking status and sex, in addition to the main exposure of course. Removing smoking status or sex form the model results in an increase in the MSE, and so this is our final model.

10.6 Model building in prediction investigations

There is a large literature on model building in prediction and we can only touch on a few points in this session. The book of Steyerberg (2019) provides a comprehensive account of modern methods in prediction investigations.

The first step is to identify the pool of variables that could be included in your prediction model, for example through discussion with subject matter experts and from a review of the literature. If this yields only a relatively small number of predictor variables then a full model including all variables may be fitted. However, more often perhaps, there is a large pool of potential predictors and it is necessary or desirable to reduce the set of variables to be included in the model. Some variables may be very weak predictors or not predictive of the outcome at all, and so it is appropriate to exclude them from the model. When considering the pool of potential predictors we should bear in mind what variables

will be easily available for individuals on whom the model is to be used in practice - that is, on future individuals not used in the prediction model development. We might sometimes also make trade-offs between a model with a large number of predictors and high predictive performance and a model with a smaller number of predictors but only slightly reduced predictive performance. Decisions about this need to take into account how the model is going to be used in practice.

10.6.1 Stepwise selection methods in prediction

Stepwise selection methods, which we introduced in the context of causal investigations above, are widely used in the development of prediction models. As stated earlier, they are automatic approaches that select a set of predictive covariates from a large group of candidate variables. Stepwise methods are arguably more appropriate in the prediction investigation setting than in the causal investigation setting, though many also consider them unsuitable for use in prediction as well.

Above we outlined backwards and forwards approaches. Steyerberg recommends that the backwards approach is preferable because all variables are considered in combination with others. In the earlier description the main covariate of interest was always retained in the model, whereas in the prediction context the stepwise selection is performed for the full set of covariates. Also in the earlier description the decision as to whether a variable was retained in the model was based on a p-value. In prediction the p-value used is typically smaller than when the stepwise procedure is used in confounder selection (e.g. 0.05). In prediction, other measures such as Akaike's Information Criterion (AIC) can be used to make the decisions instead of p-values. This quantity (the AIC) is not discussed in any detail in this module, but is just mentioned here to make you aware there are other approaches.

Advantages of stepwise selection in prediction investigations include the following:

- It is straightforward to perform in any statistical package.
- Stepwise methods are objective in the sense that different analysts should arrive at the same model given the same decision rules are used in the stepwise procedure.
- It usually succeeds in the goal of making a smaller model.

As already mentioned, stepwise methods are no panacea. Some of the problems are as follows.

1 Multiple testing: in repeated sampling from a population, both the effect estimates and the standard error of variables retained in stepwise procedures will be biased (hence the p-value will be biased). This is because covariates only make it into the model in the subset of data sets where they are statistically significantly associated with the outcome.

2 The selection process is driven by the statistical significance of the association between each covariate and the outcome. The size of estimated regression coefficients and their standard errors are not separately considered.

3 If used simplistically background knowledge is ignored. For example for the vitamin C example considered above some background reading would inform the analyst

that supplement taking and gender are related to serum vitamin C levels, so in a realistic analysis both of these should have been forced into the model.

4 Results from forwards and backwards selection may differ - this is most likely when variables are highly collinear.

5 Results will vary depending on choice of '$p-$value to enter' and '$p-$value to retain'.

6 When dummy indicators for a single categorical variable are included they should all be in or all be out; this is a common pitfall, and can be avoided in Stata by grouping variables in parentheses (check the help for examples).

7 Missing data in covariates are problematic. A simple approach when there are missing data is to apply stepwise methods using the subset of the data with no missing values on covariates. However, more principled approaches to handling missing data in variable selection procedures have been developed.

These and other criticisms are elaborated on by a number of authors, for example the *Regression Modeling Strategies* book by Harrell. Harrell's criticisms are also given on a Stata help page entitled 'Problems with stepwise regression'. It should be pointed out that some of the criticisms made of stepwise methods also apply to other methods considered here. In particular concerns over bias arising from multiple testing apply to every method considered so far.

Following the development of a prediction model additional steps are needed to assess the predictive performance of the model. Some of the basic techniques for doing this are covered in the session on Assessing Model Performance.

10.6.2 Example

One measure of predictive performance of a model is the R^2. In the vitamin C example, the R^2 from the model including only vitamin C supplement use was 0.142, and this increased in 0.232 in the full model (see earlier Stata output). We also applied backwards elimination using a p-value of 0.2 to retain variables in the model. Note that our application of the backwards elimination procedure is slightly different from when we used it in the setting of a causal investigation, because here we do not force the supplement use variable into the model. The reduced model after backwards elimination here is actually identical to that found when supplement use was forced in the model (see earlier results) and includes smoking status and sex. The R^2 from the reduced model is 0.226, which is very close to that from the full model.

10.7 Including non-linear terms and interactions

In a single lecture it is not possible to do more than introduce some of the key concepts underlying analysis strategies. One issue not dealt with concerns the form of the assumed relationship between continuous covariates and the outcome variable. By default we include continuous predictors linearly in regression models. Note that this does not necessarily mean assuming the predictor has a linear effect on the outcome. For example in logistic regression we model the log odds of a positive outcome, and so including a predictor linearly corresponds to assuming the log odds, after adjusting for other covariates, depends linearly on the continuous predictor. While some predictors may have an

approximate linear effect (on the appropriate scale), in general there is no reason why this should be the case. Thus if our dataset is 'sufficiently' large, it is a good idea to explore non-linear effects. We may also consider transforming continuous predictors before proceeding, particularly if they exhibit a highly skewed distribution.

The easiest approach to allow for a non-linear effect of a covariate X is to add X^2 as an additional covariate, which means assuming the transformed mean of the outcome is a quadratic function of the covariate. One approach which permits more flexibility than quadratic effects (or more generally polynomials) is to use *splines*. Splines are functions which consist of piecewise polynomials, i.e. different polynomial functions stuck together across the range of the predictor. To do this we split the range of the predictor into intervals at user-defined knots. So-called *fractional polynomials* are another flexible approach to including non-linear terms.

Possible interactions are another consideration. In a causal investigation, interactions are important to consider because there may be factors that modify the effect we are interested in. If there are a number of potential adjustment variables that we are considering then this results in many possible interactions. For this reason we usually restrict our attention to variables identified as potential effect-modifiers before we have started the analysis. Thus, in bio-medical applications we usually focus on factors for which an interaction makes biological sense. Typical examples are `age`, `sex` and variables relating to co-morbidities. If interactions are important then the the model building techniques described above can be adapted to start with a model that includes the important interactions.

In the prediction setting, including variable interactions may improve the predictive performance of a model. If we have a very large data set we may be willing to include many interaction terms in our prediction model, because we are not too concerned about interpretation of the final model, except in terms of its predictive performance. The stepwise methods can be extended to incorporate interactions.

10.8 Use of analysis plans

It is a good idea to agree an analysis plan with the people you are working with, which may include clinicians, epidemiologists and other statisticians. In some settings, in particular in clinical trials, a formal detailed analysis plan is required by regulatory bodies before any analysis takes place. Even away from this kind of setting, e.g. in epidemiological studies, there are good reasons for preparing an analysis plan, including the following.

- An analysis plan is sometimes required as part of the process of obtaining ethical approval for the study.
- It helps you, the statistician, to think carefully about how to best answer the questions you are working on.
- It helps your collaborators, who are likely to be posing the clinical or epidemiological questions, to focus on what the most important questions are and to understand from discussion with you how these questions will be addressed and any problems that could arise, e.g. due to missing data or errors in measurements.
- Journal reviewers like to see that the hypotheses being investigated in a paper were pre-defined. This provides reassurance that the associations being reported are not

the result of so-called 'data dredging'.

- Sometimes quite some time may pass between discussing the analyses and actually receiving the data and performing the analyses. An analysis plan serves as a record of what has been discussed and agreed.
- The analysis plan provides a ready-made outline for the methods section of a paper.

Of course it is not always possible to set out in detail every single analysis that you will perform on a given set of data to answer some pre-specified questions. Upon receiving the data some issues may arise which require further discussion and some modification or extension of the plan.

10.9 Conclusions

The appropriate analysis strategy depends on the aim of the analysis. Thus what constitutes a reasonable strategy in one study may be inappropriate in another. We also reiterate that there is no single correct strategy, even for a given aim and setting, and that the model building process typically requires judgements and decisions which to some extent will always be subjective. This is not necessarily a bad thing: it is part of the statistician's role to blend the subject matter knowledge from collaborators with statistical expertise in order to devise a suitable strategy and execute it.

A number of more advanced methods have been developed that rectify some of the limitation of the methods discussed in this session. Well known methods include *ridge regression* and the *LASSO*. Advanced methods for variable selection in causal investigations are also emerging. Machine learning methods can be appropriate in some settings, particularly in high dimensional settings with hundreds or perhaps thousands of variables (e.g. genomic data). Machine learning methods have been extensively used and developed for prediction problems, but are now also being used in causal investigations. For more details on these methods, the book *An Introduction to Statistical Learning* by James et al. is recommended.

References

Greenland S, Daniel R, Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *International Journal of Epidemiology* 2016; 45: 565-575.

Greenland S, Pearce N. Statistical foundations for model-based adjustments. *Annual Review of Public Health* 2015; 36: 89-208.

Harrell F. *Regression Modelling Strategies.* Springer. 2001.

James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. Springer. 2013. Freely available online: `http://faculty.marshall.usc.edu/gareth-james/ISL/`

Steyerberg EW. *Clinical Prediction Models: A practical approach to development, validation, and updating.* Second Edition. Springer. 2019.

Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* 2010; 21: 7-30.

10.10 Practical 10

Dataset required: `lbw.dta` (Stata webuse)

Introduction

In this practical we will use a number of different analysis strategies to build models with birth weight as the outcome using the same dataset introduced in practical 4. There are two parts to this session.

- In Part A we will consider ***causal*** investigation
- In Part B we will conduct a ***prediction*** investigation

We will use the same dataset in both parts. This The variables we will consider are described in the table below.

Variable	Description
`bwt`	Baby's birth weight (grams)
`age`	Mother's age (years)
`lwt`	Mother's weight at last menstrual period (in lbs)
`race`	Mother's race (1=white; 2=black; 3=other)
`smoke`	Mother's smoking status during pregnancy
`ptl`	Number of previous premature labours
`ht`	Mother's history of hypertension (0=None, 1=previous hypertension)
`ui`	Presence of uterine irritability in mother (1=No, 1=Yes)
`ftv`	Number of visits to physician during first trimester

Aims

The aims of this session are to:

1 Understand why different modelling approaches are needed according to the aims of the research question

2 Understand how to select variables for possible inclusion in a causal investigation

3 Understand how to select variables for possible inclusion in a prediction investigation

4 Be able to use stepwise, change in estimate, and MSE approaches to modelling

Part A. Causal investigation

First, imagine that our aim is to estimate the effect of maternal hypertension on birthweight adjusting for all relevant confounders. So we have a single exposure of interest and a number of other variables which could potentially confound the relationship between hypertension and birthweight. Clinical input suggests that presence of uterine irritability (`ui`) and number of visits to the physician (`ftv`) could be on the causal pathway between maternal hypertension and birthweight, so we will not consider these variables as potential confounders.

1 The dataset is available directly from Stata:

```
webuse lbw, clear
```

(a) Explore the dataset. Are there any variables with missing values?

One issue to look for is where there are very few observations in a particular level of a categorical variable. If there are, we can merge categories to avoid running into problems during regression analysis. Are there any variables you would consider re-categorising in this dataset?

(b) Investigate the associations of the potential confounders with the exposure and outcome variables, using tables and plots as you see fit.

(c) Compare the mean birthweight in the two groups of women defined by hypertension status.

2 (a) Fit a simple linear regression model with birthweight as the outcome and history of hypertension as the only explanatory variable.

(b) Next fit the full model including all covariates (re-categorised if necessary), except the two deemed inappropriate to include.

Discuss: Compare the coefficient for history of hypertension in the two models. What conclusions do you draw?

3 (a) Perform backwards selection, forcing history of hypertension to be included in the model and using a threshold of 0.2 for exclusion from the model.

Notice that Stata treats categorical variables as a series of dummy variables and considers them separately when deciding whether or not to include them in the model. This means that, to take the race variable as an example, that white women could be included in the model and black women excluded. This is clearly undesirable. To force Stata to include or exclude all categories together you should enclose the variable in parentheses in the command statement.

(b) Repeat the above using forwards selection, with a threshold of 0.2 for inclusion in the model.

Discuss: Compare the results from the forwards and backwards methods. What impact does changing the p-value criteria for retaining and adding variables make to the final selection?

4 (a) Use the change in estimates method (section 10.4.3 of the Notes) to perform the model building, using the backwards approach and with a threshold of a 10% change in the value of the coefficient for hypertension.

(b) Investigate the effect of changing the threshold. Do you always select the same variables for your final model? If you're working in a group, each person in the group could investigate a different threshold.

Discuss: Compare the results from the different variable selection methods. Working with a few colleagues decide which model you prefer. Then present the results from the unadjusted model and your preferred adjusted model in an appropriate table. Write a paragraph summarising your conclusions concerning the effect of maternal hypertension on birthweight, including noting any assumptions or caveats. If online post the paragraph in the Zoom chat.

Part B. Prediction investigation

Now suppose that we instead want to build a predictive model for birthweight.

5 In the investigation into the causal effect of maternal hypertension on birthweight, we excluded two variables because it was thought that they could be on the causal pathway between the exposure the outcome. What will affect your choice of the set of variables you will consider for a prediction investigation?

6 (a) We will include all of the variables listed in the table above. Run a model with all of the variables included.

```
regress bwt i.ht age lwt i.race i.smoke i.premature i.ui i.visits
```

(b) Investigate stepwise methods for building your prediction model, as above, investigating the impact of different thresholds and the difference between forwards and backwards approaches.

Discuss: With your colleagues discuss the models and agree on a preferred model from the stepwise investigations. Contrast the results obtained from this model with those obtained from the alternative strategy of including all predictor variables. Which model do you prefer?

Bonus question

7 If you have time, return to the casual investigation and use the MSE method (section 10.5 of the Notes) to perform the model building. Code is provided in the solutions to this Practical.

Session 11

Model checking

11.1 Aims & Objectives

The aim of this session is to introduce some further concepts and approaches for assessing whether a model is correctly specified.

11.2 Checking GLMs — how models can be misspecified

Before reporting the results based on a fitted model, we should check, as far as possible, that the model is correctly specified. If it is not, our inferences may well be invalid, and we may draw incorrect conclusions regarding our research question(s) of interest.

There are three obvious ways in which a GLM may be inadequate:

1 *Mis-specified distribution:* the assumption that the observed responses are independent realizations of a given distribution may not be valid. For example, count data may be overdispersed for a Poisson distribution due to an omitted covariate.

2 *Mis-specified linear predictor:* the linear predictor omits interactions which should be included, or models the effects of continuous covariates inappropriately (i.e. using the wrong transformation).

3 *Mis-specified link function:* for the specified linear predictor, the link function may not be correct. There may also exist an alternative link function which would allow a simpler linear predictor to be used (e.g. switching from a logit link to a log link, or vice-versa, may remove the need for interactions).

There exists an enormous range of methods for examining the adequacy of a fitted model, known collectively as **diagnostics**. Some methods involve formal statistical tests, but most involve inspection of plots. A comprehensive survey of methods for binomial (and binary) data is given in Chapters 5 and 6 of Collett. Many of these can be adapted to other distributions. See also McCullagh and Nelder, Chapter 12. Many of these are based on residuals.

In a previous session we explored the deviance and Pearson goodness of fit tests, and the Hosmer-Lemeshow test for individual binary data. It is important to remember that failure to reject the null hypotheses of correct specification from such global goodness of fit tests does not prove a model is correctly specified. In particular, the tests may not have sufficient power to detect poor fit. In this session we will explore some of the methods developed which aim to detect specific types of mis-specification. Since we have explored issues surrounding overdispersion for counts in an earlier session, we will focus on the second and third aspects, that of specification of the linear predictor and the link function.

11.3 Specification of the linear predictor

Our primary concern when considering specification of the linear predictor is whether we have omitted any important interactions and/or whether we have included continuous covariates in the appropriate way. That is, for example if we have modelled log odds as a linear function of age, is this appropriate?

In the case of omitted interactions, we can re-fit the model including the interaction, and assess through significance testing and consideration of the magnitude of the interaction parameter(s) whether it should be included. Concerning modelling a continuous covariate's effect, we briefly described in the session on analysis strategies a number of approaches for allowing for non-linear effects.

To assess whether a covariate which is currently not included in the model should be included, we can of course just add the covariate to the model and, through significance testing and/or consideration of the magnitude of the corresponding parameter estimate judge whether the covariate should be retained in the model. However, we must be wary of the fact that if we compare many different models, the validity of the final model inferences may be impaired due to the issue of multiple testing.

Next, we describe a number of alternative definitions of residuals for GLMs, which extend the familiar concept of residuals in linear regression. We then go on to describe how they can be used to assess whether the linear predictor is correctly specified.

11.3.1 Pearson Residuals

Measures of 'distance' or 'agreement' between observations and fitted values are known as **residuals**. We will give both the generic definitions, and their specific values in the case of grouped binomial data, in which $Y_i \sim Bin(n_i, \pi_i)$, and where $\text{logit}(\pi_i) = \eta_i$.

The ith **raw residual** is the difference

$$r_i = y_i - \hat{\mu}_i$$

i.e. the difference between y_i and its fitted mean. In the binomial data case, this is equal to

$$r_i = y_i - n_i\hat{\pi}_i$$

Since the variability of observations is not in general constant (it varies depending on covariates), larger residuals may be due simply to $Var(Y_i)$ being larger. In the binomial case, we have that $Var(Y_i) = n_i\pi_i(1-\pi_i)$. To address this, we can scale the raw residual by the (estimated) standard deviation of Y_i, to give the **Pearson residual**:

$$p_i = \frac{r_i}{\sqrt{\hat{Var}(Y_i)}}$$

In the binomial case we have

$$p_i = \frac{r_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

Note that the sum of squared Pearson residuals is equal to the Pearson's chi-squared statistic (hence the name):

$$\sum_{i=1}^{n} p_i^2 = \text{Pearson's } \chi^2 \text{ statistic}$$

Although we have divided by the (estimated) variance of Y_i, additional variability is introduced into the residuals due to variability in $\hat{\mu}_i$. This variability is due to the uncertainty in the regression coefficient estimates. The *standardized Pearson residuals* allow for this, and are defined so that they should have variance one. The expression for the standardization is somewhat complicated – for further details see the Collett book.

Although standardised Pearson residuals have unit standard deviation they are typically not normally distributed. For example, consider individual binary data where just 10% of data points have $Y = 1$. All of these points will have positive Pearson residuals (and therefore positive standardised Pearson residuals) whilst the other 90% of data points will have negative Pearson residuals. So the distribution of the Pearson residuals cannot be Gaussian. What we would expect is that 10% of points will have large positive residuals with the other 90% having small negative residuals.

11.3.2 Deviance and Anscombe Residuals

By analogy to the construction of Pearson residuals we might consider decomposing in the same way the deviance statistic. The **deviance residual** is chosen so that it has the same sign as r_i and $\sum_{i=1}^{n} d_i^2 = D$ where D denotes the deviance statistic. Recall that for the binomial distribution we have that the deviance is

$$D = \sum_{i=1}^{n} 2\left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right\}.$$

It follows that

$$d_i = \text{sign}(r_i)\sqrt{2\left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right\}}.$$

The deviance residuals can also be standardized to have (approximately) variance one.

There is a third class of residual, called the **normalised** or **Anscombe** residual:

$$a_i = \frac{f(y_i) - f(\hat{\mu}_i)}{\sqrt{\widehat{Var}\{(f(y_i) - f(\hat{\mu}_i))\}}}$$

The function $f(\cdot)$ is defined in such a way that the Anscombe residuals have, as best as possible, a normal distribution, provided the model is correctly specified. We will not go into further details regarding this function here - for more details see the Stata press book 'Generalized Linear Models and Extensions' by Hardin and Hilbe.

11.3.3 Residuals for models fitted with `glm` *in Stata*

If the model is fitted using `glm`, the following residuals can be generated using `predict`:

1 `predict <newvar>, pearson` computes the Pearson residuals;

2 `predict <newvar>, deviance` computes the deviance residuals;

3 `predict <newvar>, anscombe` computes the Anscombe residuals.

The option `standardized` can be used if the standardized version is desired. See `help glm` for the full range of options for `predict`.

If the data is stored as grouped binary data then these commands will give one residual per group: if the data is individual binary data then there will be one residual per individual.

11.3.4 Use of residuals to examine the adequacy of the linear predictor

This topic has already been covered in detail in the Regression Course (term 1). The main use of residuals is to produce graphical displays rather than make formal tests (although these do exist). Here is a summary of some possible plots and their uses.

1 Plot of residuals against observation number (**index plot**) – to detect observations with unusually large residuals, and other effects associated with the order of the data.

2 Plot of residuals against the values of the linear predictor – to detect outliers, and to detect patterns of fit associated with the mean.

3 Plot of residuals against a particular explanatory covariate which is already included in the model. If there is some (remaining) relationship it means the covariate's effect is not being correctly modelled.

4 Plot of residuals against a new covariate, or a higher power of an existing covariate ('added variable plot') – to assess whether the covariate needs to be included in the linear predictor.

These plots can be performed using the `lowess` command, in order to produce a smoothed nonparametric estimate of the (if any) association between a residual and linear predictor or covariate.

11.4 Example: alcohol consumption in NHANES

To illustrate these concepts we consider data from the 2003-2004 US National Health and Nutrition Examinations Survey (NHANES). The analyses shown in this and subsequent sessions are intended to be purely illustrative. Data are available on 2,548 men and women for the following variables.

Here we will consider logistic regression models for a dichotomised version of the alcohol variable, which takes the value one if the participant reported consuming > 5 drinks on days they did consume alcohol and zero otherwise. We start with a model which includes gender and age as covariates.

Variable	Description
`gender`	1=male, 2=female
`ageyrs`	Age in years at survey
`bmi`	Body mass index (kg/m^2)
`sbp`	Systolic blood pressure (mmHg)
`ALQ130`	Reported average number of drinks per day, on days when alcohol consumed

```
. glm heavydrinker i.gender ageyrs, family(binomial)

** output omitted **

------------------------------------------------------------------------------
             |                 OIM
heavydrinker |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
     Female  |  -1.824947   .1736042   -10.51   0.000    -2.165205   -1.484689
      ageyrs |  -.0304506   .0039882    -7.64   0.000    -.0382674   -.0226338
       _cons |  -.1838589   .1781737    -1.03   0.302     -.533073    .1653552
------------------------------------------------------------------------------
```

There is strong evidence that the probability of > 5 drinks per day depends on both gender and age, with women and older participants less likely to report drinking > 5 drinks.

An obvious concern is whether we have correctly modelled the effect of age on logit($P(Y = 1)$) correctly. One approach is to include a suitable non-linear transformation of age as an additional covariate, and see if it improves fit, or we could use a more flexible spline type approach. Doing the former, we obtain a p-value for age squared of $p = 0.073$, which particularly in light of the large sample size, indicates that a linear effect (on the log odds) is not unreasonable.

As a further check, we can plot the (individual level) residuals from the model with gender and linear age against age. Figure 11.1 shows a plot of the standardized Pearson residuals against age from this model, constructed using the following Stata commands.

```
. predict p, pearson standardized

. lowess p ageyrs, yline(0, lcolor(black)) msize(tiny)
```

The patterns seen here are a result of the discreteness of the binary outcome. For each age and gender, every subject with $Y_i = 1$ has the same residual value; and each subject with $Y_i = 0$ has the same residual value. So for each age there are four possible residual values, according to gender and outcome. The two higher 'lines' correspond to subjects with $Y_i = 1$. One 'line' is for males and the other for females. The two lower 'lines' correspond to subjects with $Y_i = 0$. Again one 'line' is for males and the other for females. Increasing age is associated with decreased odds of heavy drinking, so for those with $Y_i = 1$, the residual is larger for larger ages.

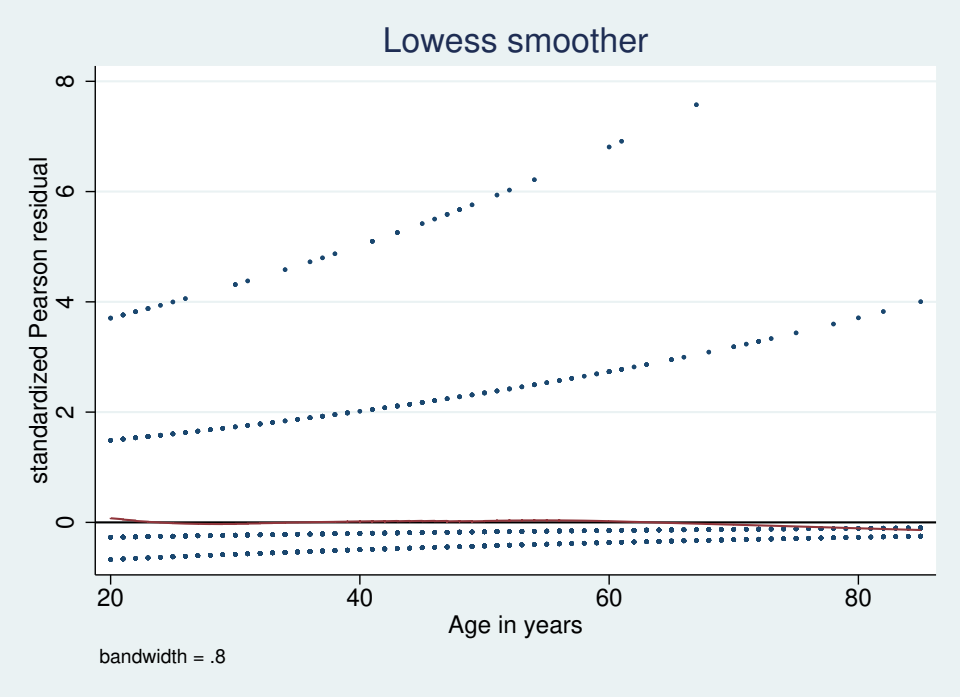


Figure 11.1: Standardized Pearson residuals (for individuals) against age, in logistic model with gender and linear age as covariates

The solid line is the lowess smoother. There is some slight departure from the $Y = 0$ line (useful to show this line on the plot) but this is not dramatic suggesting that including age linearly in the logistic regression model is not unreasonable. Notice how much more informative the lowess smoother is here than it was when we introduced this technique in the context of linear regression.

As a warning against entirely relying on the strategy of first fitting a model that assumes a linear relationship and then, if this is statistically significant, adding the square of a continuous covariate to check if the functional form is correct, we now consider whether BMI should be added to the logistic regression model which has gender and age as covariates. To do this, we add BMI as a covariate and see if it improves fit as follows.

```
. glm heavydrinker i.gender ageyrs bmi, family(binomial)

** output omitted **

------------------------------------------------------------------------------
             |                 OIM
heavydrinker |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
     Female  |  -1.829978   .1738184   -10.53   0.000    -2.170656   -1.489301
      ageyrs |   -.030688   .0040105    -7.65   0.000    -.0385485   -.0228274
         bmi |    .008346   .0117336     0.71   0.477    -.0146515    .0313435
       _cons |  -.4048605   .3588733    -1.13   0.259    -1.108239    .2985182
------------------------------------------------------------------------------
```

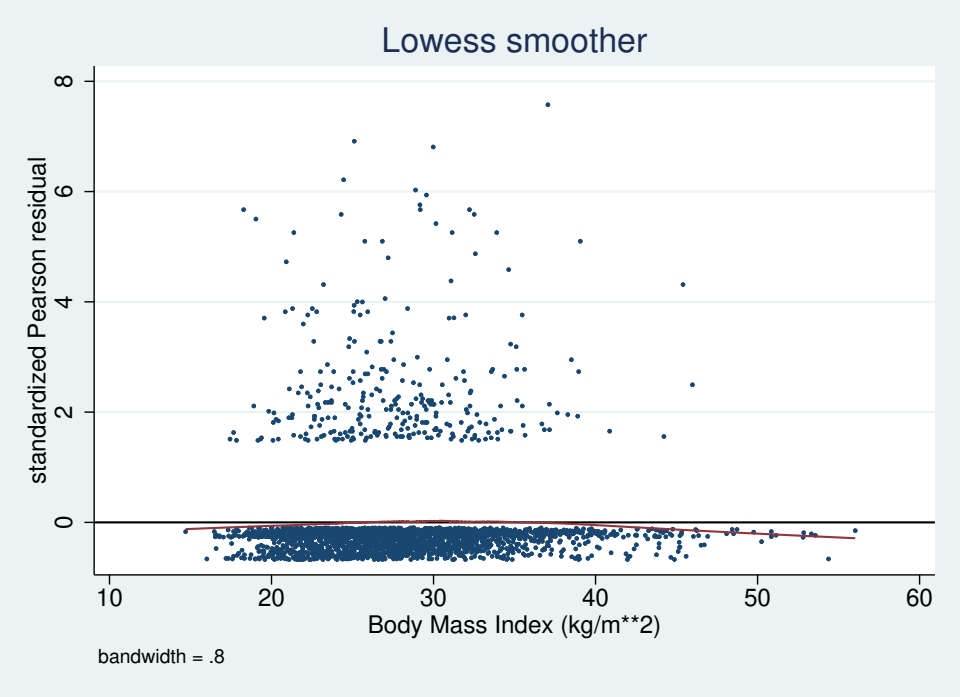


Figure 11.2: Standardized Pearson residuals against BMI, in logistic model with gender and linear age as covariates

The Wald p-value for BMI indicates there is no evidence that the outcome is related to BMI, conditional on gender and age. However, if we plot the standardized Pearson residuals from the model with just gender and age against BMI (Figure 11.2), we see a suggestion of a non-linear, perhaps quadratic, relationship.

Adding BMI and BMI squared, and comparing to the model with gender and age only with a log-likelihood ratio test, we obtain $p = 0.02$, indicating some evidence that BMI does indeed have an independent association with the outcome. In this case we were unlucky in the sense that the association with BMI is such that assuming a linear effect (on the log odds) gives a slope of close to zero, wrongly leading us to conclude that the outcome is unrelated to BMI (after adjustment for gender and age). Having said this, the lowess association is not particularly large in magnitude, and obtaining $p = 0.02$ in such a large dataset also suggests that the association is small in magnitude: hence one might choose to still not include it in the model.

Exercise 11.1 *Residual plot appearance*

Explain why there are no small positive residuals, and no large negative residuals, in Figure 11.2.

11.5 Covariate pattern residuals

So far we have defined residuals at the individual observation level. If a model for individual binary data is fitted using `logit` or `logistic`, the values are calculated by 'covariate' pattern. Given a particular fitted model, with a particular specification of covariates, a covariate pattern is a particular combination of values of the included covariates which occur in the data. To explain, consider a model in which gender and ethnicity (3 levels)

are included as covariates (without an interaction). Then there would be 6 covariate patterns in the data (assuming each ethnic group has men and women in the dataset), corresponding to each combination of gender and ethnicity.

If instead our model included a continuous covariate and if each individual in the dataset had a different value of the covariate, the number of covariate patterns would equal the sample size, n. More commonly, variables which we refer to as continuous may take on fewer values due to rounding (e.g. age to nearest year), such that even with a 'continuous' covariate, the number of covariate patterns m is less than the sample size n. As soon as our model has at least one continuous covariate, the number of covariate patterns m will typically equal, or be close the sample size n. This is because each individual in the dataset has a unique combination of the covariates which are included in the model.

To define covariate pattern residuals, suppose there are m covariate patterns, and consider the jth covariate pattern, which contains n_j individuals. Since each individual outcome is Bernoulli 0/1, we can think of the data as really being grouped binary data. We let y_j denote the total number of successes in the jth covariate pattern, and let x_j denote the values of the (multiple) covariates in the jth pattern. The covariate pattern residual is then

$$\tilde{y}_j = y_j - n_j\hat{\pi}_j$$

where $\hat{\pi}_j$ denotes the fitted probability for the jth covariate pattern, and the corresponding Pearson residual is defined as

$$\tilde{p}_j = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

The motivation for calculating residuals by covariate pattern is as follows. With individual binary data, a single residual cannot on its own give information about model fit: if an individual has predicted probability of success 0.1 but has a 1 (success) outcome, we cannot distinguish between whether this is due to poor fit or chance: if the model is correctly specified we expect 10% of individuals with a predicted success probability of 0.1 to have successes. In contrast, if we calculate residuals by covariate patterns, and for a given pattern the predicted success probability is 0.1, we expect the observed proportion to be close (aside from sampling variability) to 0.1.

11.5.1 Example: alcohol consumption in NHANES

To illustrate covariate pattern residuals we return to our Stata analysis relating alcohol consumption to age and sex in the NHANES data, but now use `logit` and the `rstandard` option with `predict` to get standardised Pearson residuals by covariate pattern. The relevant commands are as follows.

```
logit heavydrinker i.gender ageyrs
predict p2, rstandard
egen pickone=tag(gender ageyrs)
lowess p2 ageyrs if pickone==1, yline(0, lcolor(black))
```

With covariate pattern residuals we need to be a little careful in how we use them: each individual with the same covariate pattern will be assigned the same covariate pattern

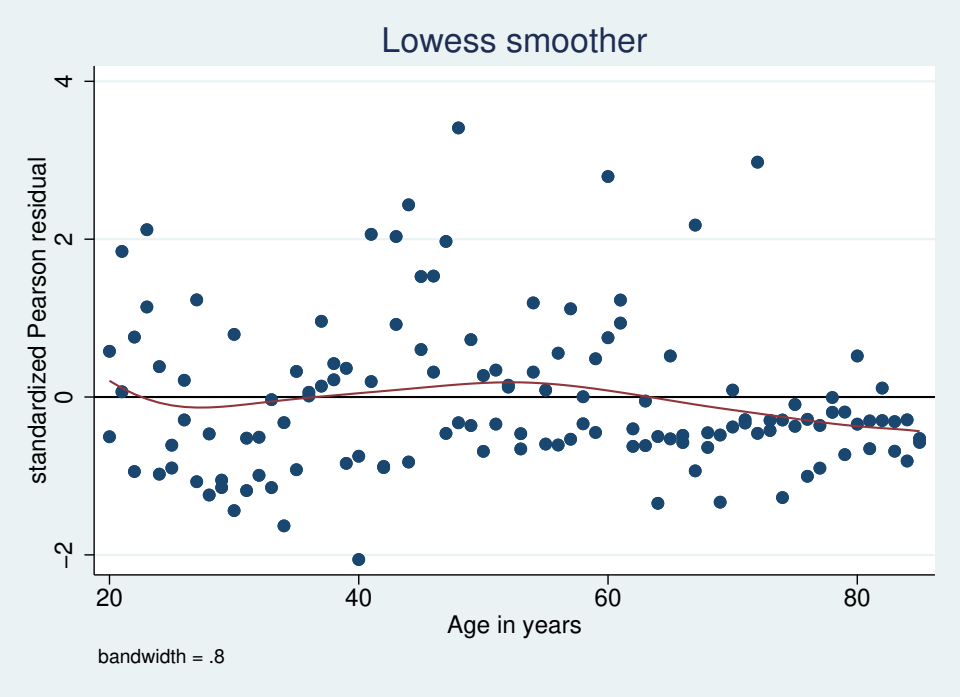


Figure 11.3: Standardized Pearson residuals (by covariate pattern) against age, in logistic model with gender and linear age as covariates

residual. This is usually not an issue in generating plots, but if we want to use lowess smoothers on such plots, we should treat each covariate pattern as a single (grouped binary) observation. The use of the `pickone` variable in the above series of commands achieves this aim.

The plot produced by these commands is shown in Figure 11.3. There are now just two residuals shown at each age, one for males and one for females. Again the fitted lowess line exhibits some slight departure from the $Y = 0$ line but this is not dramatic suggesting that including age linearly in the logistic regression model is not unreasonable.

11.6 Covariate pattern or individual residual plots?

When the number of covariate patterns, m, is less than n, we have a choice about which residuals to use, and consequently which residuals to use in diagnostic plots. There is no consensus on which is preferable, and indeed software packages other than Stata (e.g. SAS) typically calculate individual level residuals after fitting logistic models.

Hosmer and Lemeshow (Applied Logistic Regression) advocate the use of covariate pattern residuals, essentially because this gets one back to a grouped binary data setting, such that various asymptotic properties of the residuals become closer to holding. However, as previously noted, as soon as our model starts to have continuous covariates the number of covariate patterns typically approaches or equals n, and we are back to the individual level residuals.

One drawback to covariate pattern residuals is that if we have calculated covariate pattern residuals for a fitted model and want to check whether a further variable needs to be added to the model, we cannot really plot the covariate pattern residuals against the new

variable, since individuals who share the same covariate values of the fitted model likely have different values of the new covariate. All we can do in this situation is to add the variable into the model and then check the appearance of residual plots.

11.7 Pearson's goodness of fit test with groups defined by covariate patterns

In session 4 we described the Hosmer-Lemeshow goodness of fit test, which can be performed using a user-specified number of groups. For example `estat gof, group(10)` carries out the test with 10 groups. It is also possible to perform this test with groups defined by covariate pattern. In Stata this is achieved simply by omitting a specification of the number of groups after `estat`.

For the earlier model fitted to the NHANES data with heavy drinking as the outcome and just gender and age as covariates we get the following.

```
. estat gof

Logistic model for heavydrinker, goodness-of-fit test

       number of observations =      2548
 number of covariate patterns =       132
            Pearson chi2(129) =       120.33
                  Prob > chi2 =         0.6951
```

The degrees of freedom are 129, which is the number of covariate patterns minus the 3 model parameters. This is the Pearson goodness of fit test for the collapsed grouped data for the 132 combinations of age and gender observed in this data.

For individual binary data, this test is valid provided that the number of covariate patterns is not 'too large' relative to n. Here we had $n = 2548$ and $m = 132$, so it is probably reasonable to use the test. But when m is of a similar magnitude to n, e.g. when we have multiple continuous covariates, the test should not be used. Instead we can use the Hosmer-Lemeshow approach, specifying `group()`.

The Pearson goodness of fit test for the model with gender and age gave $p = 0.695$, indicating no evidence of poor fit. But we also saw earlier that if we added BMI and BMI squared this improved model fit significantly (note that this does not necessarily imply a practically important improvement in the predictive, i.e. discriminatory power, of the model: see session 12).

It is important to realise a non-significant Pearson goodness of fit test does not mean there are not other variables which might help predict the outcome. It merely means that there is no evidence indicating that the model you have specified for outcome given those covariates currently included is not correctly specified.

11.8 Link function

Provided the link function has an inverse which maps to valid values of the mean (or probability in the case of binary data), typically we tend to worry less about choice of link function. Indeed, for binary data, the logit and probit link functions are in fact very similar to each other, and tend to give very similar conclusions.

One approach is to test whether the link function is correct, given the current specification of the linear predictor. The test involves calculating the linear predictor from the fitted model, then re-fitting the model with the original linear predictor, plus an additional covariate equal to the original linear predictor value squared. If the original model is correctly specified, the squared linear predictor covariate will not significantly improve model fit. If it does, this is taken as evidence that the link function is wrong. In this case, we would try an alternative link function, and perform the same test.

In Stata, this test can be performed using the `linktest` command. After using `logistic` simply type `linktest` without any options, but for the command to work properly after using `glm` the distribution and link function need to be re-specified, for example after fitting a logistic regression model using `glm` use `linktest, family(binomial) link(logit)`.

11.9 Summary

We have outlined the three overall routes through which a GLM can be mis-specified. We have described how Pearson, deviance and Anscombe residuals can be derived for GLMs, and how they can be used to assess adequacy of the linear predictor and the link function.

One topic we have not considered in detail is the use of diagnostics for assessing the influence of individual observations and for the identification of outliers. For the latter, we can examine whether any observations have standardized residuals which are very large (after standardization). For this purpose, particularly in small samples, the Anscombe residuals are preferable.

11.10 Practical 11

Dataset required: `nhanesglm.dta`

Introduction

In this practical we will explore model development and model checking, making use of the residuals described in the lecture. We will continue analysing the NHANES alcohol data introduced in the lecture. In this practical we will develop a model for whether a participant in the study has hypertension, defined as having a systolic blood pressure of 140 mmHg or above.

In this session we give you most of the Stata code you will need; we ask that you spend time on interpreting the output from each command, and in discussing conclusions and model choices with your colleagues.

Variable	Description
`gender`	1 = male, 2 = female
`ageyrs`	Age in (whole number of) years
`bmi`	Body Mass Index (kg/m^2)
`sbp`	Systolic blood pressure (mmHg)
`ALQ130`	Average number of alcoholic drinks per day in past year
`alccat`	Average number of alcoholic drinks per day in past year: 1 = 1 drink per day 2 = 2-5 drinks per day 3 = 6+ drinks per day

Aims

The aims of this session are to:

- Understand how to assess model fit for logistic regression models
- Interpret residual plots to assess if explanatory variables are modelled correctly

Analysis

Load the data and explore the variables.

Before we start, we need to generate a new binary variable to indicate whether each observation has hypertension (1) or not (0).

```
gen ht = (sbp>=140)
```

Univariable model

1 Fit a logistic regression model relating the log odds of hypertension to age using the glm command.

```
glm hypertension ageyrs, family(binomial)
```

We will explore various options with the predict command to investigate different ways to check if the model provides a good fit to the data.

First, we will obtain the predicted probability of hypertension for each person in the study. This uses the option mu, the expected value of Y.

```
predict pr1, mu
```

Examine these probabilities. We can also look at the average predicted probability among people without hypertension and with hypertension. We would hope that the mean will be close to zero in the first group and close to one in the second group.

```
bysort ht:  sum pr1
```

2 To obtain the *i*ndividual Pearson standardized residuals we use the following:

```
predict sp1, pearson standardized
```

Examine the distribution of these residuals. What is their mean and variance? Are they normally distributed?

(a) The first plot we will look at is an index plot. This can be obtained using

```
gen id = _n
scatter sp1 id
```

Discuss: What do you notice from this plot? Are there any outliers you would want to investigate? Can you explain why there are no residuals with a value between 0 and about 0.8?

(b) To plot the residuals against the linear predictor we can use the following: predict xb, xb

```
lowess sp1 xb, yline(0)
```

(c) We can plot the residuals against age using:

```
lowess sp1 ageyrs, yline(0)
```

Discuss: Why does the plot appear to only show a maximum of two residuals at each age?

In seeking to understand this you may find it helpful to 'jitter' the points a little.

```
lowess sp1 ageyrs, yline(0) jitter(3)
```

3 Since this is a relatively large dataset and age is only recorded to the nearest year we can consider a grouped approach.

(a) There are a number of ways of converting the individual patient data to grouped data. One way is as follows.

```
egen n=count(ht), by(ageyrs)
egen r=sum(ht), by(ageyrs)
egen pickone=tag(ageyrs)
```

The first two lines generate the denominator and numerator, respectively, for each group by age. The third line selects one observation per group, which is marked by the variable pickone taking the value 1; for all other rows pickone takes the value of 0. Browse the data to make sure that you understand the new variables.

(b) Now refit the model to the data in the grouped form; when we do this we must restrict the dataset to just one row per group using "`if pickone==1`". Then

obtain the grouped standardised Pearson residuals from this model.

```
glm r ageyrs if pickone==1, family(binomial n)
predict grp1, pearson standardized
```

Plot these against age.

```
lowess grp1 ageyrs if pickone==1, yline(0)
```

4 If the model is fitted in Stata using the `logit` or `logistic` command then covariate pattern residuals can be obtained using the option "`rstandard`":

```
logit ht ageyrs
predict cp1, rstandard
lowess cp1 ageyrs if pickone==1, yline(0)
```

Note here that although we must fit the data on all observations, when we plot the residuals we must again restrict it to one per group to allow the lowess function to appropriately smooth the curve.

Discuss: Compare the covariate pattern residuals (and plots) to those obtained from the grouped analysis (in Q3) and the individual residuals in Q2. What do you notice?

Discuss: From looking at all of the plots, what do you conclude about the functional form in which age should be included into the logistic regression? In which plot do you find it easiest to discern patterns?

Multivariable models

5 Add the square of the age variable to the linear predictor for the logistic regression model. Use residual plots analogous to those above to explore the fit of this model.

Discuss: What do you conclude about the most appropriate way to include age in this model?

6 Use individual-level residuals from the model in 5 to examine whether BMI ought to be included in the model, and depending on what you find, continue with your previous model or add BMI. In the latter case, generate new residuals and assess if you have included BMI using the most appropriate functional form.

7 (optional) Explore whether gender should be included in the model, including whether or not the other covariates already included interact with gender in their effects on the log odds of hypertension.

Discuss: What is your final model? Have you included the same variables in the same way as your colleagues?

8 (optional) Based on your final model, use margins to calculate fitted probabilities for an individual aged 60 years, at BMI values from 20 to 40 in increments of 5, separately for men and women, and plot the resulting values.

Session 12

Assessing model performance

12.1 Aims & Objectives

The aim of this session is to consider the assessment of the performance of a model for individual binary data, through introducing the concepts of calibration, explained variation, and discrimination.

As usual, we assume we have a binary outcome Y and covariates $X_1, .., X_p$, and that we have fitted an appropriate GLM for modelling how $E(Y|x_1, .., x_p)$ depends on the covariates. Usually we will choose the canonical logit link function, such that the model is a logistic regression. Without loss of generality, in the following we will assume that we have used logistic regression, although the definitions and concepts which follow would also apply if a different link were used.

After fitting the model to our dataset, we can calculate the predicted probability of 'success' for each subject, which with the logit link is given by:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ...\hat{\beta}_1 x_{ip})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ...\hat{\beta}_1 x_{ip})}$$

12.2 Calibration

The model is said to be well calibrated if the predicted probabilities match (subject to sampling variability) the true expected proportions of successes. If the model is to be used for prediction, and these predictions are to be used to aid clinical decision making, it is important for the model to be well calibrated. For example, if amongst those subjects assigned a predicted risk of 5%, 20% experience the event, we may not give an intervention which would have been given had the predicted risk been 20%.

If the model only contains categorical covariates, then we can perform this comparison by covariate pattern – for each pattern, we compare the observed proportion of subjects who experienced a success with the predicted probability of success for that pattern (remembering that all subjects in the same covariate pattern have an identical fitted probability of success).

To illustrate, we again consider models for the `heavydrinker` variable in the NHANES data. Recall from the previous session that this binary variable takes the value 1 if an individual reported drinking more than 5 alcoholic drinks on average when they drink alcohol, and 0 otherwise. First, we fit a very simple model which has only gender as a covariate.

```
. logistic heavydrinker i.gender

** output omitted **

------------------------------------------------------------------------------
heavydrinker | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
     Female  |   .1727592   .0297217   -10.21   0.000     .1233104    .2420376
       _cons |   .2180385   .0152498   -21.78   0.000     .1901076    .2500731
------------------------------------------------------------------------------
```

After doing so, we can use the estat gof command to give us a table of expected and observed counts by covariate pattern, by specifying the `table` option.

```
. estat gof, table

Logistic model for heavydrinker, goodness-of-fit test

  +-------------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 |  Exp_0 | Total |
  |-------+--------+-------+-------+-------+--------+-------|
  |     1 | 0.0363 |    42 |  42.0 |  1115 | 1115.0 |  1157 |
  |     2 | 0.1790 |   249 | 249.0 |  1142 | 1142.0 |  1391 |
  +-------------------------------------------------------------+

  +-------------------------+
  | Group |   Prob | gender |
  |-------+--------+--------|
  |     1 | 0.0363 | Female |
  |     2 | 0.1790 |   Male |
  +-------------------------+

      number of observations =      2548
 number of covariate patterns =         2
             Pearson chi2(0) =         0.00
                 Prob > chi2 =            .
```

For the current model, there are of course two covariate patterns, one corresponding to the males and the other corresponding to the females. Since the number of parameters in the model equals the number of covariate patterns, the model predicted numbers of events matches the observed in each group - we have perfect calibration, and this is a consequence of the model having the same number of parameters as covariate patterns.

Now we add the `highbmi` variable as an additional covariate, which define as 1 if $BMI >$ 25, and 0 otherwise. Fitting the model and using `estat gof, table`, we obtain the following.

```
. estat gof, table

Logistic model for heavydrinker, goodness-of-fit test

  +-------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0340 |    13 |  14.5 |   413 | 411.5 |   426 |
  |     2 | 0.0376 |    29 |  27.5 |   702 | 703.5 |   731 |
  |     3 | 0.1686 |    74 |  72.5 |   356 | 357.5 |   430 |
  |     4 | 0.1837 |   175 | 176.5 |   786 | 784.5 |   961 |
  +-------------------------------------------------------+

  +-----------------------------------+
  | Group |   Prob | gender | highbmi |
  |-------+--------+--------+---------|
  |     1 | 0.0340 | Female |       0 |
  |     2 | 0.0376 | Female |       1 |
  |     3 | 0.1686 |   Male |       0 |
  |     4 | 0.1837 |   Male |       1 |
  +-----------------------------------+

      number of observations =      2548
 number of covariate patterns =         4
             Pearson chi2(1) =         0.30
                 Prob > chi2 =         0.5848
```

Now there are 4 covariate patterns, corresponding to combinations of gender and `highbmi`. Now, since there are more covariate patterns (4) than model parameters (3), the observed and expected counts are not identical. As we described in session 11, when the number of covariate patterns is small (i.e. when we have only categorical covariates in our model), the Pearson test statistic reported by `estat gof` is a test of the null hypothesis that our model is correctly specified (which imples it is well calibrated). Here, we have no evidence ($p = 0.58$) against this null, and hence no evidence of poor calibration.

12.2.1 Flexible calibration curves

Once we have continuous covariates in our model, as discussed in session 11, the comparison of observed and expected becomes difficult because the number of covariate patterns is usually large, and close to the sample size n. This leads to the numbers in each covariate pattern being small or even one. In session 4 we saw how the Hosmer-Lemeshow test attempts to overcome this issue by grouping subjects according to their predicted probabilities, and comparing observed with expected in these groups. In Stata this test (and table) can be requested using a command such as `estat gof, table group(10)`. The number of groups must be specified: here we have opted for ten.

However, as explained in session 4, there are drawbacks to the Hosmer-Lemeshow test, with alternative approaches nowadays recommended. See Van Calster and colleagues (Calibration: the Achilles heel of predictive analytics. BMC Med 17, 230 (2019). https://doi.org/10.1186/s12916-019-1466-7) for an overview of alternatives. Austin and Steyer-

berg (Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33(3):517-535. doi:10.1002/sim.5941) suggest graphical approaches. At its simplest this involves relating the binary outcome to the predicted probability of the outcome using a locally weighted scatter plot smoother, such as a lowess smoother. The following shows how this can be done in Stata with the low birth weight data used to introduce the Hosmer-Lemeshow test in session 4.

```
. webuse lbw, clear

. logistic low age lwt i.race smoke ptl ht ui

** <output omitted> **

. predict fitval

. twoway (lowess low fitval) (scatteri 0 0 1 1, recast(line) ///
lpattern(dash)) (scatter low fitval, msymbol(lgx)), ///
title(Calibration curve) ytitle(Underweight baby (1 = yes, 0 =  no)) ///
xtitle(Fitted probability) legend(off)
```

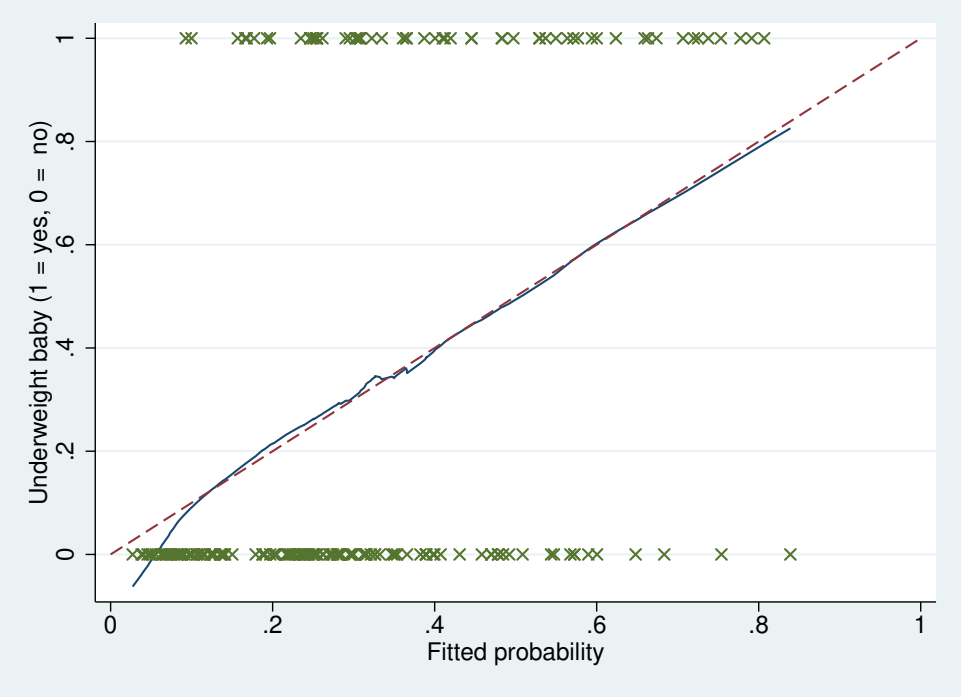


Figure 12.1: Flexible calibration curve for a logistic regression model predicting the probability of a baby being born underweight from multiple covariates. Fitted line from a lowess smoother.

Here the lowess smoother lies close to the line of equality throughout its length, so we have no evidence of poor calibration. See the literature referenced above for sophistications of this approach, including adding confidence intervals to the lowess smoother.

12.3 Explained variation and R^2 measures

Ensuring that our model is well calibrated is important. However, it in no way implies that our model accurately predicts individual risk. This is clear from the fact that the model for heavy drinking with only gender as a covariate has perfect calibration, yet we know that there are other variables in the dataset which significantly improve the model's fit: i.e. other variables which give us more accurate predictions of the risk of heavy drinking. The problem with the model for heavy drinking which includes only gender as a covariate is that it explains very little of the variation in risk between subjects. In particular, it only explains variation in risk between individuals which can be attributed to differences in gender. This leads us to consider whether we can construct a measure which quantifies how much variation in risk our model explains.

For linear regression models, R^2 measures how much of the variation in Y is explained by the covariates $x_1, .., x_p$. It is usually defined as

$$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Generalizing R^2 to logistic regression models has proved somewhat difficult. There have been many different proposed measures which aim to be the logistic regression analogue of R^2 for linear regression – see Mittlbock and Schemper (1996), 'Explained variation for logistic regression', Statistics in Medicine, 15; 1987-1997.

Here, we shall content ourselves to examination of the measure that is reported by Stata's `logistic` and `logit` commands as 'pseudo-R2'. The measure is often referred to as McFadden's likelihood ratio index, and is given by

$$R^2_{\text{McFadden}} = 1 - \frac{\ell_c}{\ell_{null}}.$$

where ℓ_c and ℓ_{null} are the maximized log-likelihood values of the current model and the null model respectively. To try and see if this definition makes intuitive sense, suppose our current model fits no better than the null model (as measured by the maximized log-likelihood). Then $R^2_{\text{McFadden}} = 0$. Conversely, if our model fits perfectly, in the sense that $\hat{\pi}_i \approx 1$ for those subjects with $y_i = 1$ and $\hat{\pi}_i \approx 0$ for those subjects with $y_i = 0$ then

$$\ell_c = \sum_{i=1}^{n} y_i \log(\hat{\pi}_i) + (1 - y_i) \log(1 - \hat{\pi}_i) \approx 0$$

and so $R^2_{\text{McFadden}} \approx 1$.

One criticism of McFadden's index is that when applied to a linear regression model its value is not the same as R^2, which lead some to question its appropriateness as a measure of explained variation for logistic regression. Also, whether one believes it is possible to construct a model which predicts $\hat{\pi}_i \approx 1$ for those subjects with $y_i = 1$ and $\hat{\pi}_i \approx 0$ for those subjects with $y_i = 0$ depends on one's philosophical views regarding nature. If we believe the world is deterministic, and our model is capturing the uncertainty due to imperfect knowledge, then McFadden's measure seems appropriate. In contrast, suppose we believe that there is intrinsic randomness in our outcome, such that even if we had measured all of the explanatory factors relevant to our outcome, our predicted probabilities would

still not be 0 or 1. Then arguably McFadden's measure is less useful, since it would be impossible to attain a value of $R^2_{\text{McFadden}} = 1$.

Our earlier model for heavy drinking with gender and `highbmi` has a pseudo R2 value of just 0.079, suggesting that the two covariates only explain a relatively small amount of the variation in the outcome.

Despite the fact that McFadden's index (like others) has certain drawbacks, in particular that its interpretation is somewhat unintuitive, it is useful to examine its value from fitted models, to give some sense of how much variation in the outcome is being explained by the predictors.

12.4 Discrimination, sensitivity and specificity, and ROC curves

12.4.1 Sensitivity and specificity

The final approach to assessing model performance that we shall examine is that based on the model's ability to discriminate between those subjects with $y_i = 1$ and those with $y_i = 0$. To do this, suppose we classify subjects as 'positive' if their predicted risk lies above some value p, and negative otherwise. The sensitivity and specificity of this classification rule is then given by the proportion of cases who are classified positive, and the proportion of non-cases who are classified negative:

$$\begin{aligned} \text{sensitivity} &= P(+|y_i = 1) = \frac{\sum_{i=1}^{n} 1(\hat{\pi}_i > p \ \& \ y_i = 1)}{\sum_{i=1}^{n} 1(y_i = 1)} \\ \text{and specificity} &= P(-|y_i = 0) = \frac{\sum_{i=1}^{n} 1(\hat{\pi}_i \leq p \ \& \ y_i = 0)}{\sum_{i=1}^{n} 1(y_i = 0)}. \end{aligned}$$

The values of sensitivity and specificity will clearly depend on the cut-off p that we use to classify subjects. If we use a high value of p to classify as positive, we will have lower sensitivity but higher specificity compared with what would be obtained with a lower value of p. After fitting a model using `logistic` we can visualize this using the `lsens` command. Figure 12.2 shows this plot for a logistic model for `heavydrinker` with gender, age and BMI (continuous) as covariates. Ideally there would exist values of p such that sensitivity and specificity are both high. However, whether this is possible will depend on how well risk can be predicted. In Figure 12.2, we see that exists no p which would give sensitivity and specificity both above (say) 80%.

12.4.2 ROC curves

An alternative popular plot of the sensitivity and specificity is the so called 'receiver operating characteristic' (ROC) curve. The plot has its origins in the analysis of radar signals in World War II. The plot is of sensitivity against $1 -$ specificity as the value of p is varied. After `logistic`, the ROC curve can be plotted using `lroc`. Figure 12.3 shows the ROC curve for the preceding model for heavy drinking with gender, age and BMI as covariates. If the current model has no predictive power, the line in the ROC curve is the line of equality. As the model's predictive ability increases, the curve moves towards the top left corner.

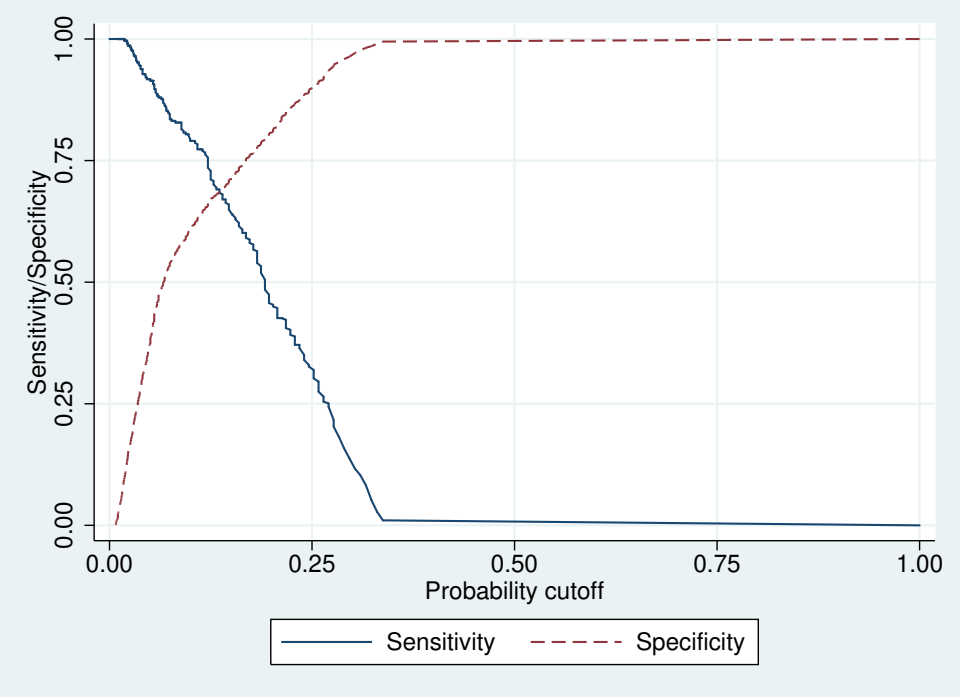


Figure 12.2: Sensitivity and specificity vs predicted probability cut-off. Model for heavy drinking with gender, age and BMI (continuous) as covariates.

12.4.3 Area under the ROC curve

Since a model with perfect discrimination has an ROC curve which passes through the top left hand corner, the area under the ROC curve (AUC) is commonly reported as a summary measure of a model's predictive ability. When the AUC is close to one, this means there exists a value p such that sensitivity and specificity are both close to 1. Conversely, a model with no predictive ability, whose ROC line is the line of equality, has an AUC of 0.5.

It turns out that the AUC has another helpful interpretation. Consider two randomly selected subjects, one with $y_i = 1$ and the other $y_j = 0$. Then one can prove (see Practical) that $AUC = P(\pi_i > \pi_j | y_i = 1 \ \& \ y_j = 0)$. That is, for a randomly selected pair of subjects, one with a success and one with a failure, the AUC is the probability that the model correctly ranks them, by giving the subject with an observed success a higher probability of success.

12.4.4 Overfitting, cross-validation, and external validation

There are issues with sensitivity, specificity and ROC curves which fall beyond the scope of this module but which are important, particularly if a model is to be used to aid clinical decision making. The first is that of overfitting, which can be a particular problem in small datasets. If we build a model on one data sample then the model will be optimised to maximise discrimination for that sample and hence its discriminatory performance is likely to be better in that sample than on another, even if the samples are drawn from the same population. One way of avoiding this problem is to divide the sample in two, and to use the first half of the data to build a model, and the second half to assess its performance. However this approach is inefficient and so other approaches tend to be

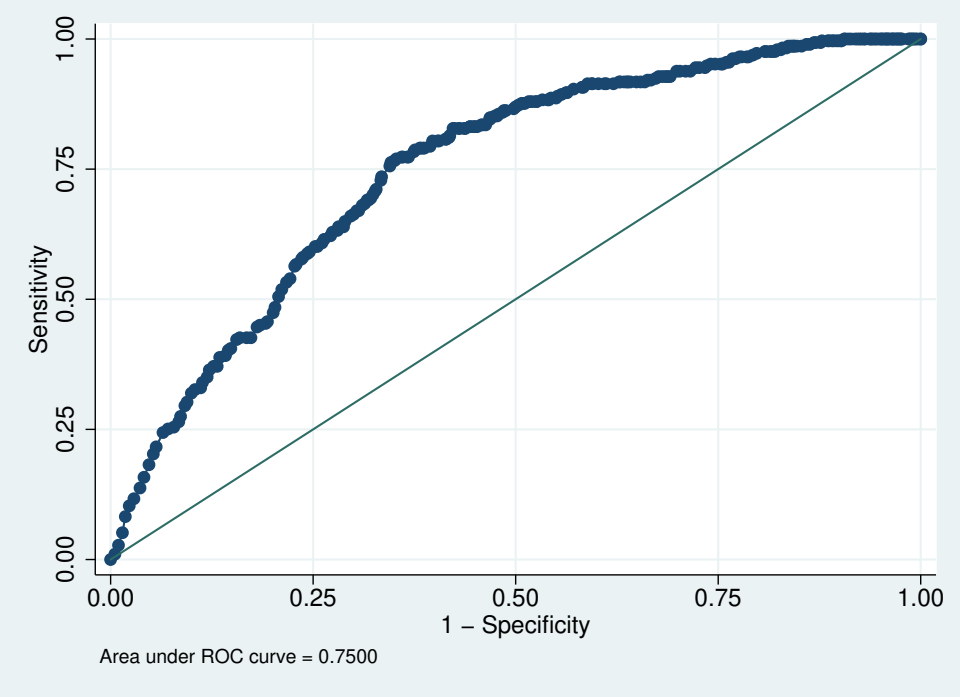


Figure 12.3: Receiver operating curve for model for heavy drinking with gender, age and BMI (continuous) as covariates.

used in practice. These include 'leave one out', 'k-fold cross-validation' and bootstrap approaches (*e.g.* the '0.632 bootstrap').

The second issue is external validation, and more generally the applicability of models developed using data from one population to individuals from a different population.

More details on overfitting, cross-validation, and external validation can be found in the book Clinical Prediction Models by Steyerberg (Springer, 2009). There are also published guidelines (the TRIPOD guidelines) on reporting standards for multivariable prediction models (Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement, published in 11 journals including Annals of Internal Medicine 2015;162(1):55-63).

12.5 Predictiveness curves

The ROC curve has been criticised in some quarters (see Cook, Circulation 2007;115:928–935) because it does not readily allow clinically useful quantities to be derived. In particular, if we are interested in prognostic modelling, we are arguably interested in the distribution of Y given the covariates. In contrast, the ROC curve is a plot of sensitivity and (one minus) specificity, quantities which concern the distribution of predicted risk *conditional* on outcome Y. A similar criticism can be levelled at the AUC, since it concerns a comparison of predicted risks conditional on outcome – it is looking backwards in time, whereas in predictive model building our outcome usually follows (in time) our covariates (and hence predicted risk).

Here we mention one alternative proposal, that of predictiveness curves (see Pepe *et. al.*,

American Journal of Epidemiology 2008;167:362–368). The predictiveness curve is simply a plot of each predicted risk $\hat{\pi}_i$ against that risks' percentile in the distribution. That is, if the 50% centile of the predicted risks is 0.21, we plot 0.21 against 50. The plot enables one to immediately assess how widely dispersed the predicted risks are among the subjects. A model which assigns most subjects similar values of $\hat{\pi}_i$ cannot have good discrimination, whereas a model which assigns risk with more dispersion has improved discrimination (assuming the model is well calibrated). The plot also allows us to read off what proportion of subjects have risk above or below a certain level.

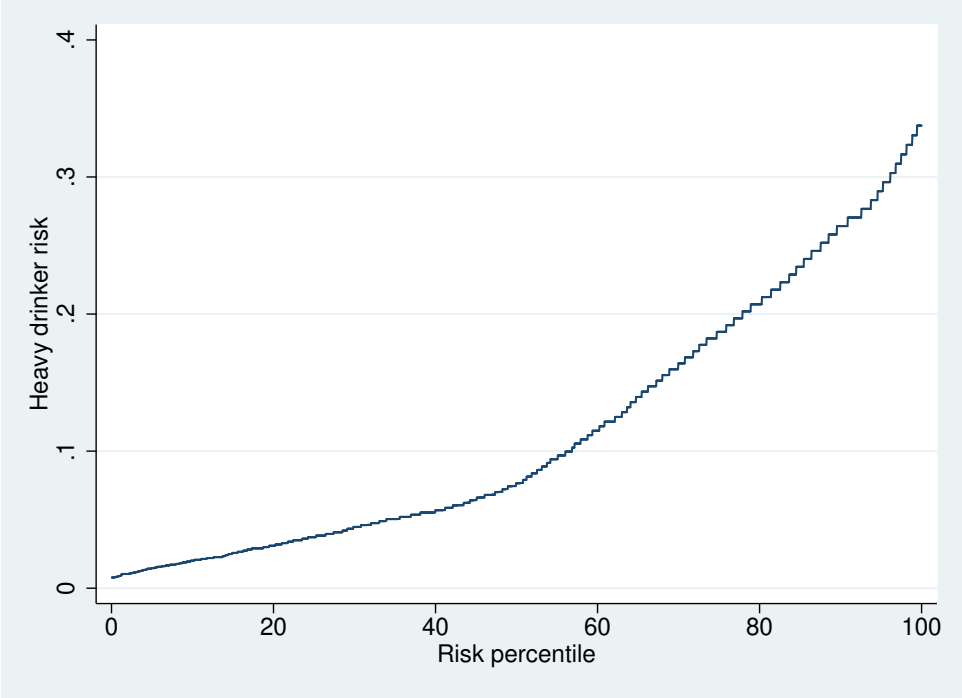


Figure 12.4: Predictiveness curve for model for heavy drinking with gender, age and BMI (continuous) as covariates.

Figure 12.4 shows the predictiveness curve for our model for heavy drinking with gender, age and BMI as covariates. This shows that the predicted risks are fairly well dispersed, although no subjects have a predicted risk above around 35%. We can also (for example) read off that approximately 20% of the subjects have a predicted risk above 0.2.

Lastly, one possible addition to the predictiveness curve is add the observed proportions of events in groups defined by (say) deciles of risk. These are of course the same quantities that the Hosmer-Lemeshow test uses to assess calibration. If the model is well calibrated, the observed proportions should be reasonably close to the predicted risks.

12.6 Summary

In this session we have explored some approaches for assessing the performance of logistic regression models. We have seen that a model being well calibrated does not imply that it necessarily discriminates well between subjects, in that predicted risks may be quite different from a subject's 'true' risk. How best to quantify and summarise a logistic model's discrimination ability is still an open area of research and debate, but the approaches we have covered here are those which are typically used in current practice.

12.7 Practical 12

Dataset required: `nhanesglm.dta`

Introduction

There are three sections to this practical session.

The first section is a data analysis, looking at the predictive power of logistic regression models for hypertension.

The second section uses simulated data to explore the relationship between odds ratios, pseudo R^2 and the area under the ROC curve.

In the final optional section we ask you to prove a result which was stated in the notes, relating to the value of the AUC.

Aims

Understand how to assess the performance of a logistic regression model using

- model predictions
- pseudo R^2
- area under the ROC curve

Part A. Analysis of hypertension

1 Load the nhanesglm.dta dataset. As in the previous practical we will analyse hypertension as a binary variable, so first create a new variable for hypertension, defined as having systolic blood pressure of 140 mmHg or above. Tabulate the new variable.

2 Fit a logistic regression model for hypertension with BMI entered as a single linear covariate.

(a) What is the reported pseudo R^2?

(b) Use the predict command to generate a new variable containing the fitted probabilities from this model.

(c) Use the `estat gof` command to obtain observed and predicted values, according to deciles of predicted risk.

Discuss: How would you summarise the association between BMI and hypertension? What do you conclude about the model's performance from the predicted values and the pseudo R^2?

3 Add gender and age to your model, and again generate a variable containing the predicted probabilities from this model.

Plot the two predicted probabilities against each other.

4 Essentially the same comparison of the ranges of the fitted probabilities from the two models can be made by constructing predictiveness curves (section 12.5 in notes) for each of the two models, displaying both on the same graph. First create the new variables:

```
egen rank1 = rank(pr)
egen rank2 = rank(pr2)
egen num = max(rank1)
gen cent1=100*(rank1-0.5)/num
gen cent2=100*(rank2-0.5)/num
```

And then create the plot:

```
#delimit ;
twoway (line pr cent1, sort) (line pr2 cent2, sort lpattern(dash))
,
xtitle("Risk percentile") ytitle("Hypertension risk")
legend(order(1 "prediction using BMI alone"
2 "prediction using BMI, age and gender") rows(2))
;
#delimit cr
```

(The "`#delimit ;`" tells Stata that the end of each command is marked by a semi-colon. The "`#delimit cr`" tells Stata to again accept a carriage return as the end of line marker. This allows us to split a long command over several lines, to make it easier to read).

Discuss: What do you conclude about the two models from these plots?

5 Plot the ROC curve for the model including age and gender as well as BMI.

Discuss: What do you conclude regarding the model's discrimination ability from the ROC curve?

6 In the session 11 practical we investigated more complex models for the risk of hypertension. Return to the final model from that practical and use the techniques explored above to investigate the improvements in predictiveness and discrimination from using this model.

Part B. Simulation

To further explore the concept of explained variation and discrimination in binary regression models we will simulate some data, so that we know what the true conditional distribution of the outcome given the covariates is.

7 Use the following code to clear Stata's memory, set the number of observations (to 10,000), set Stata's random number seed mechanism (so we all obtain the same answers), and generate a covariate x from a Bernoulli distribution with probability of 1 equal to 0.5:

```
clear
set obs 10000
set seed 91413
gen x=(runiform()<0.5)
```

8 Next, add to your do file code to randomly generate Y , such that

$$P(Y = 1|x = 0) = 0.3$$

and

$$P(Y = 1|x = 1) = 0.7$$

Then fit the logistic regression model for y on x, and note the pseudo R^2 and AUC values.

9 Re-run your code from the previous question, but with

$$P(Y = 1|x = 0) = 0.1$$

and

$$P(Y = 1|x = 1) = 0.9$$

Again, note the pseudo R^2 and AUC values.

10 Re-run one more time with

$$P(Y = 1|x = 0) = 0.01$$

and

$$P(Y = 1|x = 1) = 0.99$$

Discuss: From these results, what do you conclude about explained variation in logistic regression models, and the pseudo R2 and AUC measures in particular, as the strength of association increases?

Part C. Theory (optional)

Prove the result stated in the lecture notes in section 12.4.3 regarding the area under the ROC curve, AUC:

$$AUC = P(\pi_i > \pi_j | y_i = 1 \ \& \ y_j = 0)$$

for two randomly selected subjects, one with $y_i = 1$ and the other with $y_j = 0$.

Session 13

Matched studies and their analysis

13.1 Aims

The aim of this lecture is to introduce you to the concepts and basic analysis tools for matched data. It will also lay the groundwork for the next session, which is on the use of conditional logistic regression.

Following this lecture and practical you should be able to do the following:

- Give examples of the use of matching in randomized trials and observational studies.
- Outline how matching can be used to increase precision of estimates.
- Outline how matching can be used in cohort studies and case-control studies to control confounding.
- Use statistical models to analyse matched studies where the dependent variable is (appropriately) continuous.
- Use McNemar's test and odds ratios to analyse matched studies with binary outcomes and exposures.

13.2 Definition and examples

A *matched* comparison study is one in which the design of the study is such that observations made under one condition are specifically, or intrinsically, linked to one or more observations made under one or more other conditions. In such a matched comparison study, the set of matched observations is sometimes referred to as a *block*.

Example 13.1 *Matched cross-over study with continuous outcome*

Fifty subjects are given an anti-hypertensive drug designed to lower their diastolic blood pressure. Diastolic blood pressure is measured before (Y_{i1}) and after (Y_{i2}) administration of the drug. Here the outcome is continuous and the blocks are the subjects.

Example 13.2 *Matched intervention study with binary outcome*

Seventy seven patients with diabetic retinopathy have one eye chosen at random to receive a new laser treatment while the other eye receives the standard treatment. After five years the outcome (blind or not) is recorded for both eyes. This outcome variable is binary and the blocks are the patients.

Example 13.3 *Matched cross-sectional study with continuous outcome*

Twenty patients with Alzheimer's disease and their unaffected spouses each carry out a memory test. The test yields an integer valued score in the range 0–30. This outcome variable is (commonly considered) continuous. The blocks are the patient-spouse pairs.

EXAMPLE 13.4 *Matched cohort study with binary outcome*

One hundred individuals taking statins are each matched to an individual not taking statins and with the same hypercholesterolaemia status (yes/no). The pairs arre followed up for 3 years during which it is recorded whether or not they are diagnosed with cardiovascular disease. The blocks are the matched pairs.

EXAMPLE 13.5 *Matched case-control study I*

Two hundred lung cancer patients are each matched to a person without cancer of the same age and sex. All subjects are asked whether or not they have ever smoked cigarettes. The blocks here are the matched pairs. As explained in session 7, in a case control study we sample by outcome status, and then measure other variables which (in reality) usually occurred earlier. It is natural to consider these other factors as exposures (and covariates) and case-control status as the outcome, but we will sometimes fit models in which one such exposure is the dependent variable and case-control status a predictor.

13.3 Matching in case-control studies

The use of matching in case-control studies is extremely common. In a matched case-control study, as noted above, a case is matched to one or more controls. The exposures of interest could be of any type (binary categorical, continuous) and there could be multiple exposures and other adjustment variables to consider. In a cohort study the use of matching is really restricted to binary exposures. So in a case-control study the matching concerns a feature which is inherent to all case-control studies - the use of cases and controls. Whereas in a cohort study the use of matching is restricted to a specialised scenario. Matched cohort studies are not particularly common.

In a case-control study it is common to match on well known confounders which are not themselves of primary interest. Matching has also sometimes been used just as a convenient method for selecting a control group. In implementing matching there may be some choice between insisting on exact match with respect to one or a small number of features, and forcing approximate match with respect to a larger number of properties. The pool of available controls may not always be large enough to insist on exact matching on several variables, which would result in a loss of some of the cases from the study if a match cannot be found. To avoid the loss of many cases from the study, the matching criteria may therefore sometimes need to be relaxed. For example, if age to the nearest year is used as a matching variable, then for some cases it may not be possible to find a control of the same age. Hence one may have to relax the matching criteria and select controls whose age is within, say, three years of that of the case. Where the matching is quite crude fitting an appropriate model may be needed to account for residual dependency.

It is also common in a case-control study to match cases to controls who have 'survived' at least as long as the case (according to the relevant time scale, which may be age). This is further discussed in the survival module.

EXAMPLE 13.6 *Matched case-control study II*

A case-control sample within the EPIC-Norfolk cohort was used to investigate the association between dietary fibre intake, measured using 7-day diet diaries, and risk of colorectal cancer (Dahm et al, Journal of the National Cancer Institute, Volume 102, Issue 9, Pages 614–626, https://doi.org/10.1093/jnci/djq092). Cases were all men and women diagnosed with colorectal cancer from entry to the underlying prospective cohort study, between 1993 and 1998, until the end of 2006. The diet diaries were completed at entry to the cohort study and, to avoid the possibility of bias from individuals changing their diet because of undiagnosed colorectal cancer, cases diagnosed within one year of completing the diary were excluded. Cases with a prior cancer were also excluded. Each case was matched to four controls on sex, age at entry to the study (within 3 years), date of diary completion (within 3 months). The study comprised 318 colorectal cancer cases and 1272 controls.

13.4 Rationale for Matching

There are a number of different reasons for using matching. Here we focus on two, i) controlling for confounding and ii) improving precision of estimates.

Individuals, or units of observation (e.g. eyes), within a block are expected to have more similar outcomes (or exposures in a matched case-control study) than individuals not eligible to be in the same block.

- In Example 13.1, certain individuals are much more likely to have high blood pressure than others for reasons aside from the treatment under study and therefore an individual serves as his or her own 'perfect' control.
- In Example 13.4, individuals with hypercholesterolaemia are more likely than those without to develop cardiovascular disease regardless of whether they are taking statins. Hypercholesterolaemia may also be associated with whether one is prescribed statins in the first place.
- In Example 13.5, age and sex may be associated with whether a person gets lung cancer (certainly age) and these factors are also likely to be associated with smoking status, so it is desirable to look at the association between smoking and lung cancer in people who are the same with respect to the other factors.

Matching controls for confounding at the design stage, and also allows differences between conditions to be more precisely estimated by removing 'between block variability' from estimates. This applies for both continuous and binary outcomes, although there are particular complexities for binary outcomes.

It is important to appreciate that a matched study must always be accompanied by an appropriate analysis that acknowledges the matching in order for valid results to be obtained. In later sections of this session, and in session 14, appropriate methods will be described, but first an important distinction must be made between the role of matching in case-control studies and its role in all of the other matched designs described above (cross-over studies and matched intervention, cross-sectional and cohort studies).

13.5 Contrasting the effects of matching in case-control and other matched designs

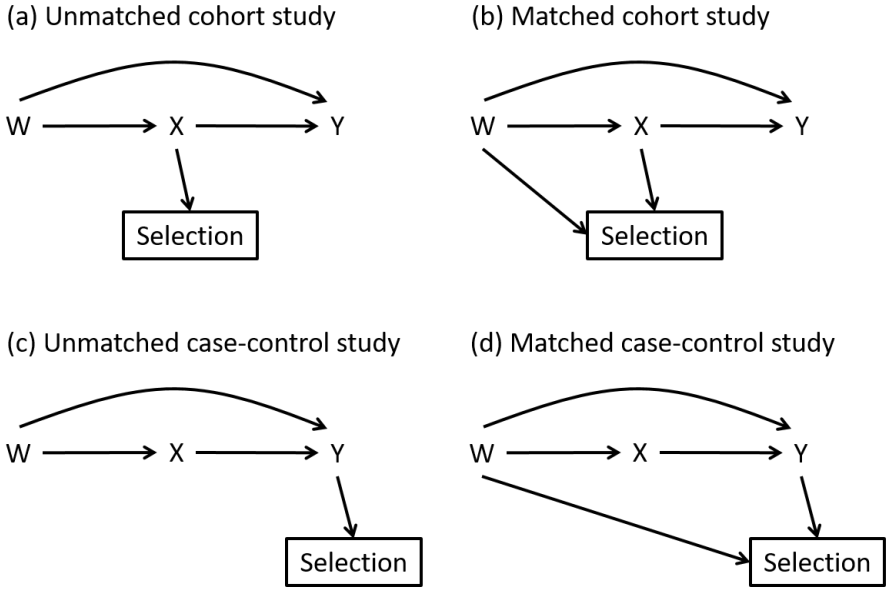
In all of the matched designs considered above, with the exception of the matched case-control design, matching eliminates confounding even if the analysis ignores the matching. However, in a matched-case control study, this is not the case. To appreciate the reason for this we return to the design of each of these types of study. To simplify the language used we focus on the contrast between matched case-control studies and matched cohort studies, although it should be appreciated that the essential comparison is between matched case-control designs and all of the other matched designs considered above.

In a **cohort study** exposed individuals can be matched to unexposed individuals based on one or more characteristics of an individual. These may be variables such as age and sex. In Example 13.4, individuals taking statins were matched to individuals not taking statins but with the same hyprcholesterolaemia status.

In a **case-control study** diseased individuals (cases) can be matched to non-diseased individuals (controls) based on one or more characteristics. As above, these may be variables such as age and sex, or some aspect of medical history. It is also common to match cases to controls based on something such as 'neighbourhood of residence' or 'GP practice'. In other examples, controls may be chosen from the same family as the case so that there is matching on 'genetic background'.

Figure 13.1 uses path diagrams to illustrate relationships between variables in matched and unmatched cohort and case-control studies, where the matching variables are confounders. Here, as in earlier sessions, X is the exposure of main interest, whilst W is a set of confounders.

Figure 13.1: Path diagrams showing relationships between variables in unmatched and matched cohort and case-control studies.



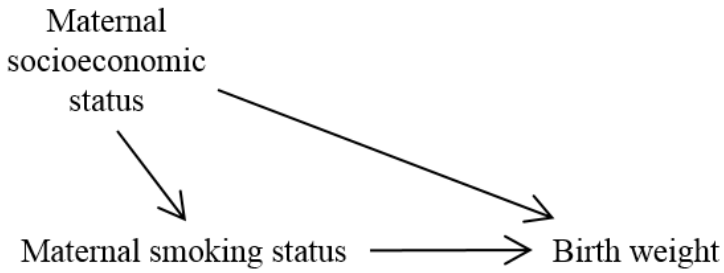
In a **cohort study** matching breaks the association between the matching variables and the exposure, that is, in the study sample there is no association between the exposure and the variables used in the matching. The association referred to here is the unconditional

association between the matching variables and the exposure, *i.e.* the association ignoring the outcome. Breaking this is sufficient to eliminate confounding by W.

In contrast in a **case-control study**, matching breaks the unconditional (*i.e.* ignoring the exposure) association between the outcome and the variables used in the matching. This does not guarantee that the matching variables and the outcome remain independent conditional on exposure, and it is this conditional independence that that would need to be eliminated in order to prevent confounding.

To illustrate this let us return to the simple (fictitious) observational situation considered in session 2. The DAG in Figure 13.2 shows three variables: two binary explanatory variables 'maternal smoking status' ($X = 1$: smoker, $X = 0$: non-smoker) and maternal socioeconomic status ($W = 1$: low, $W = 0$: high), and a continuous outcome 'birth weight' (measured in grams).

Figure 13.2: Directed acyclic graph (DAG) illustrating relationships between maternal smoking status (X), maternal socioeconomic status (W) and baby's birth weight (Y)



The arrow linking maternal socioeconomic status to maternal smoking status represents the unconditional association between these two variables. If we were to carry out a matched cohort study (by matching each smoker in our cohort to a non-smoker of the same socioeconomic status), we would break the unconditional association between maternal socioeconomic status and smoking status and so eliminate confounding.

However, the arrow linking maternal socioeconomic status to birth weight represents a conditional association. For the linear regression model $E(Y) = \beta_0 + \beta_1 X + \beta_2 W$ this arrow is represented by β_2, it is the effect of a change in maternal socioeconomic status on birth weight **holding smoking status constant**. If we were to carry out a matched case-control study (by choosing a cut-off and dichotomising birth weight, then matching each 'low-birth weight baby' to a 'healthy-birth weight baby' with the same maternal socioeconomic status) we would break the unconditional association between birth weight and maternal association, **not the conditional association**, so confounding would not be eliminated.

What this illustrates is that in the analysis of matched case-control studies allowance for the matching variables should always be made.

13.6 'Reversed' analysis of matched case-control studies

Occasionally, particularly when an exposure is continuous, matched case-control studies are analysed in a 'reverse' fashion with a difference in exposure levels between matched cases and controls reported. For example, Wald and colleagues (British Journal of Cancer (1989), **59**, 936-938) carried out a matched case-control study in which people diagnosed with cancer were matched to people without cancer on a number of potential confounding variables. The case-control study was nested within a cohort study with all participants having had cholesterol levels measured at the start of follow-up (and those with cancer at the start of follow-up omitted from this nested case-control study). Rather than reporting the odds ratio for a unit increase in cholesterol (which might have a causal interpretation), they chose to report the mean difference in cholesterol levels between the matched cases and controls. This is a simple and legitimate comparison, but obviously not one that quantifies a causal effect (because the cholesterol is measured prior to disease onset).

13.7 Matched studies estimate conditional effects

In the previous sections we have emphasised that matched studies must be analysed in a way that takes account of the matching. This means that all parameters estimated are conditional on the matching variables. For example, in the matched case-control study where patients with lung cancer are each matched to a person without lung cancer of the same age and sex, an estimated odds ratio for the association between smoking and lung cancer is conditional on age and sex.

13.8 Matched studies analysed with statistical models with a continuous dependent variable

13.8.1 Introduction and notation

We begin by considering matched studies where the dependent variable in the analysis model is continuous. The study in question may be a randomized trial or a cohort study with a continuous outcome. Or it could be a case-control study where we (artificially) consider a continuous exposure as a dependent variable and case-control status as a predictor. To avoid confusion with the approach used throughout the rest of the module (where Y denotes case-control status) we will use Z to denote the dependent variable, V to denote the predictor of primary interest and W to denote other covariates in this section.

13.8.2 Simple approaches for the analysis of matched pair studies

We focus on matched studies in which the block is a pair. Suppose that we have continuous measures of a dependent variable from n matched pairs, denoted by Z_{i1} and Z_{i2} ($i = 1, \ldots, n$). The two members of the pair have been subject to a different treatment or exposure (denoted by the subscript 1 or 2), or for a matched case-control study they are controls and cases respectively. The basic analysis strategy is to calculate the differences for each block:

$$Z_{i2} - Z_{i1}, \quad (i = 1, \ldots, n) \tag{13.1}$$

and then to use these differences as observations in appropriate one-sample statistical procedures. These involve testing the null hypothesis that the mean difference is zero, and constructing confidence intervals for the difference.

Three commonly used simple analyses for matched studies are

1 A paired t-test for comparing two means: this is the one sample t-test applied to differences with a null hypothesis that the true mean difference is 0.

2 Wilcoxon matched pairs test: this is the Wilcoxon signed rank test applied to differences with a null hypothesis that the true median difference is 0.

3 Sign test: the sign test can also be used to test a null hypothesis that a true median difference is 0.

13.8.3 Analysis using regression models that explicitly acknowledge the matching

Matched studies with continuous dependent variables can also be analysed using regression models for the individual responses. Such regression models can be used when each block is a pair, but they also extend to situations where some or all blocks include three or more individuals (for example, a matched case-control study with two matched controls for each case). To allow for this we extend the notation used in the previous section such that Z_{ij} denotes the dependent variable for individual j in the ith matched set or block and V_{ij} is an indicator (0/1) variable denoting the exposure or treatment of interest (or case-control status). For a continuous dependent variable we may use the following model for Z_{ij}:

$$Z_{ij} = \beta_0 + \beta_1 V_{ij} + \gamma_i + \epsilon_{ij} \tag{13.2}$$

where γ_i is a fixed effect for the ith matched set and ϵ_{ij} is residual error. This model can be fitted simply by adjusting for the categorical variable representing the matched sets in the regression of Z on V. If the study is a matched pair study fitting the model in (13.2) is equivalent to performing a paired t-test. However, as stated above, the approach can also be used where blocks are not pairs. Further, the model in (13.2) can also be extended to incorporate covariates that vary within blocks. For example in 13.3 we may wish to adjust for age and sex when comparing patients with Alzheimer's disease with their spouses.

Including fixed effects for each matched set is only appropriate in linear regression models. As we will see later, in non-linear models such as logistic regression models it is not appropriate to adjust for the categorical variable representing the matched sets using such fixed effects. As an alternative to using a fixed effect for the jth matched set, a so-called 'random effect' can be used, and this extends to non-linear models. The resulting models are referred to as hierarchical models, mixed effects models, or random effects models. These are covered in the module on hierarchical data and we do not give further details here.

13.8.4 Analysis using regression models that adjust for the matching variables

It can often be useful to classify matched studies into one of two broad types.

1 The matching is based on something which is unique to each matched set. The most obvious example of this is when the blocks are the subjects: *e.g.* an individual provides his or her own control (Example 13.1), or the left eye provides a control for the right eye (Example 13.2). Another example is when an individual is matched to their spouse, so that each block is a married couple (Example 13.3).

2 The matching is based on characteristics which are not unique to the block, for example when cases are matched to controls based on sex and age group. In this situation there are in principle several potential matches in the underlying population for a given individual.

The approaches described in the previous section are applicable for both types of matching. However, for the second type there is an alternative, which is to incorporate the matching variables (rather than indicators for matched sets) into the analysis. To describe this model we can simplify our notation. Let Z_k denote the dependent variable for individual k and V_k the main exposure (or treatment or case-control status) and W_k the vector of matching variables. The appropriate model is as follows.

$$Z_k = \beta_0 + \beta_1 V_k + \beta_2^T W_k + \delta_k \tag{13.3}$$

It might be thought that the W_k could be omitted from the model. This will indeed give a similar (identical if the matching is exact) estimate of β_1. This is because the matching means that the exposure (V_k) and the matching variables (W_k) are uncorrelated in the study sample, even if they are correlated in the underlying population (which may have been the whole reason for the matching in the first place if the matching was used to control confounding). However, omitting the W_k will render the standard estimate of the standard error of β_1 incorrect, because residuals from observations in the same matched set will not be independent.

An analysis which adjusts for the matching variables may be more efficient than one which includes an indicator variable for each matched set, but can involve additional modelling choices (for example, is it appropriate to assume that the effects of the matching variables are linear?). We return to this topic, in the context of binary outcomes, in the next session.

13.9 Statistical analysis of matched studies with a binary outcome and binary exposure

13.9.1 Introduction

We now consider matched studies with a binary outcome and a binary exposure. Unlike for models with continuous dependent variables (where in some settings we can get unbiased effect estimates if we ignore the matching), it is always necessary to allow for the matching variables in this setting.

The study in question may be a randomized trial or a cohort study with a binary response. Here we assume that we have n matched binary outcomes denoted by Z_{i1} and Z_{i2}, ($i = 1, \ldots, n$). The corresponding treatments or exposures are denoted V_{i1} and V_{i2}.

The study in question could also be a matched case-control study with a binary exposure. In this setting we will use Z_{i1} and Z_{i2}, ($i = 1, \ldots, n$) to denote the n matched binary exposures with case-control status denoted by V_{i1} and V_{i2}.

In this session we focus on binary variables. In the next session we will consider methods of analysis where categorical, continuous and multiple exposures are related to binary outcomes, methods which are especially relevant for matched case-control studies.

We will consider two closely connected ways of studying whether there is an association between a binary exposure and a binary outcome in a matched study:

1 Using McNemar's test.

2 Using inferences concerning odds ratios.

13.9.2 Tabulation of results

There are two ways of tabulating results from matched studies with a binary outcome and a binary exposure. Consider the diabetic retinopathy example (Example 13.2). We can have an *unmatched presentation*, as shown in Table 13.1(a). This tabulation is not completely informative in that it does not link results from the two eyes. By contrast, look at the matched presentation in Table 13.1(b). This is clearly preferable.

Table 13.1: Presentation of data from the diabetic retinopathy study

(a) Unmatched presentation

	Standard treatment ($V = 0$)	New treatment ($V = 1$)
Not blind ($Z = 0$)	43	67
Blind ($Z = 1$)	34	10
Total	77	77

(b) Matched presentation

		Standard treatment ($V = 0$)		
		Not blind ($Z = 0$)	Blind ($Z = 1$)	
New Treatment	Not blind ($Z = 0$)	39	28	67
($V = 1$)	Blind ($Z = 1$)	4	6	10
		43	34	77

13.9.3 McNemar's Test

Table 13.2 shows a general arrangement of data from a matched study, which is a useful arrangement for McNemar's test.

Table 13.2: General arrangement of data for McNemar's Test

		$V = 0$		
		$Z = 0$	$Z = 1$	
$V = 1$	$Z = 0$	q	r	$q + r$
	$Z = 1$	s	t	$s + t$
		$q + s$	$r + t$	n

When $V = 0$ the proportion of individuals with $Z = 1$ is $p_1 = (r + t)/n$. When $V = 1$ the proportion of individuals with $Z = 1$ is $p_2 = (s + t)/n$. We are interested in whether the population proportions that these two proportions estimate are different. The observed difference in proportions is

$$p_2 - p_1 = (s - r)/n. \tag{13.4}$$

The null hypothesis in McNemar's test is that this difference between the proportions under two conditions (or two exposure or treatment groups) is 0. We can see from the above formula that this is can be achieved by consideration only of the counts r and s. If the null hypothesis is true then r and s should, in expectation, be equal. In other words if the null hypothesis is true we have

$$r \sim \text{Binomial}(r + s, 0.5). \tag{13.5}$$

This is the basis of McNemar's test. We can use either an exact binomial test or a Normal approximation to the binomial to test the null hypothesis.

Example 13.2 continued: Diabetic retinopathy data Consider the data in Table 13.1(b). In total there are 32 (28+4) discordant pairs. The McNemar test p-value is the probability of observing 28 or more 'successes' (or 4 or fewer) out of 32 when the true probability of 'success' is 0.5. Stata can calculate this exactly as follows:

```
. bitesti 32 28 0.5

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
       32          28           16       0.50000      0.87500

  Pr(k >= 28)           = 0.000010  (one-sided test)
  Pr(k <= 28)           = 0.999999  (one-sided test)
  Pr(k <= 4 or k >= 28) = 0.000019  (two-sided test)
```

Here there is strong evidence against the null hypothesis that the two treatments are equally effective.

13.9.4 Odds ratios in matched studies with binary outcomes and binary exposures

McNemar's test does not quantify the association between an exposure or treatment and the outcome. As in unmatched studies, the association between two binary variables can be measured using an odds ratio. In the next session we will show how to do this using a variant of logistic regression (conditional logistic regression). Here, we introduce an alternative approach, which builds on a method for the estimation of pooled odds ratios in stratified studies relating a binary factor (V) to a binary dependent variable (Z), the Mantel-Haenszel method. This gives an odds ratio which is conditional on the variable(s) which define the strata.

For the ith strata, denote the number of $V = 0$ individuals having $Z = 0$ by a_i, the number of $V = 1$ individuals having $Z = 0$ by b_i, the number of $Z = 0$ individuals having $Z = 1$ by c_i and the number of $Z = 1$ individuals having $Z = 1$ by d_i (Table 13.3).

Table 13.3: Stratified data in stratum i: Numbers of individuals in stratum i with each combination of $V = 0, 1$ and $Z = 0, 1$.

	$V = 0$	$V = 1$
$Z = 0$	a_i	b_i
$Z = 1$	c_i	d_i

The Mantel-Haenszel odds ratio ($V = 1$ vs $V = 0$) is

$$\Psi_{MH} = \frac{\sum_i (a_i d_i / n_i)}{\sum_i (b_i c_i / n_i)} \tag{13.6}$$

where n_i is the number of individuals in each strata.

In the matched pair setting we can use the same approach, but now the strata are the matched sets and each strata contains only 2 units of observation (in fact the approach does generalise to settings where the block size is greater than two).

In the matched pair setting stratum i is a pair of individuals, so the numbers a_i, b_i, c_i, d_i are all either 0 or 1. Further, because of the matched design, in each stratum we will have one individual with $V = 0$ and one with $V = 1$, so we have $a_i + c_i = 1$ and $b_i + d_i = 1$. The only strata that will contribute to the numerator in 13.6 are those where $a_i = 1$ and $d_i = 1$, whilst the only strata that will contribute to the denominator are those where $b_i = 1$ and $c_i = 1$. In the notation of Table 13.2 there are s strata where $a_i = 1$ and $d_i = 1$ and r strata where $b_i = 1$ and $c_i = 1$. It follows that here the Mantel-Haenszel odds ratio is as follows.

$$\begin{aligned} \Psi_{MH} &= \frac{\sum_i (a_i d_i / n_i)}{\sum_i (b_i c_i / n_i)} \\ &= \frac{s}{r} \text{ (in notation of Table 13.2)} \end{aligned}$$

where $n_i = 2$ is the number of individuals in each table.

Notice that the odds ratio estimate only depends on the discordant pairs, the numbers of which are given by r and s. In a matched randomized trial or matched cohort study the matched pairs in which both individuals have the same outcome are uninformative, and in a matched case-control study the matched pairs in which both individuals (the case and the control) have the same exposure are uninformative.

13.9.5 Confidence Intervals for Odds Ratios from matched Studies

Confidence intervals for odds ratios can be calculated using the standard method for Mantel-Haenszel estimates (not described here). A better alternative, though, is to construct an exact confidence interval as follows:

1 Denote the expected *proportion* of discordant pairs in which $V = 0$ results in $Z = 0$ and $V = 1$ results in $Z = 1$ by π. The parameter π is estimated by $s/(r + s)$. Also denote the true ratio of discordant pairs (estimated by s/r) by Ψ. The relationship between Ψ and π is:

$$\Psi = \frac{\pi}{1 - \pi}.$$

2 s (and r) follows a binomial distribution

$$s \sim \text{Binomial}(r + s, \pi)$$

Using our knowledge of the Binomial distribution we can calculate lower and upper confidence limits π_L and π_U for π, which could be exact or from a Normal approximation. Note the connection between this step and the basis for McNemar's test in equation (13.5).

3 Finally, we can back transform the confidence limits for π to give a confidence interval for Ψ as:

$$\left(\frac{\pi_L}{1-\pi_L}, \frac{\pi_U}{1-\pi_U}\right).$$

Example 13.5 continued: Estimating odds ratios for the lung cancer data

Two hundred lung cancer patients are each matched to a person without cancer of the same age and sex. All subjects are asked whether or not they have ever smoked. Table 13.4 shows the data.

Table 13.4: Matched case control data to investigate the association between smoking and lung cancer.

		Controls ($V = 0$)		
		Never smoked ($Z = 0$)	Ever smoked ($Z = 1$)	
Cancer cases ($V = 1$)	Never smoked ($Z = 0$)	20	10	30
	Ever smoked ($Z = 1$)	70	100	170
		90	110	200

The ratio of the odds of being a smoker if you are a case to the odds of being a smoker if you are a control is $70/10 = 7$. Note that this is the same as the ratio of the odds of being a case if you are a smoker to the odds of being a case if you are a non-smoker.

To obtain a 95% confidence interval for this odds ration first get a 95% confidence interval for the proportion of discordant pairs in which the case is the smoker as follows.

```
. cii proportions 80 70
                                                    -- Binomial Exact --
Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
         |      80        .875    .0369755        .7821109    .9383979
```

Then the 95% confidence interval of the odds ratio relating case-control status to smoking (the ratio of discordant pairs) is:

$$\left(\frac{0.7821}{1-0.7821}, \frac{0.9384}{(1-0.9384}\right)$$
$$= (3.59, 15.23)$$

13.9.6 General analysis of matched binary outcome and binary exposure studies in Stata

All matched binary outcome and binary exposure studies can be analysed using the `mcc` command in Stata. Since this is designed for matched case-control (mcc) studies, care must be taken in interpretation if it used with other designs. Below we show its use in the lung cancer data (Example 13.5 - a matched case-control study) and the diabetic retinopathy data (Example 13.2 - a matched trial). Note that in the diabetic retinopathy application of `mcc`, 'cases' equate to 'new (treatment)', 'controls' to 'standard (treatment)' and 'exposed' to 'blind'.

Example 13.5 continued: Analysis of the lung cancer data using `mcc`

```
. list
            id       case    control
  1.         1          1          1 -- case and control smoked
  2.         2          0          0 -- case and control never smoked
  3.         3          1          0 -- case smoked, control never smoked
   .         .          .          .
 200.      200          1          1

.  mcc case control

. * or

.  mcci 100 70 10 20

                | Controls               |
Cases           |   Exposed   Unexposed  |      Total
----------------+------------------------+-----------
        Exposed |       100          70  |        170
      Unexposed |        10          20  |         30
----------------+------------------------+-----------
          Total |       110          90  |        200

McNemar's chi2(1) =     45.00    Prob > chi2 = 0.0000
Exact McNemar significance probability       = 0.0000

Proportion with factor
        Cases            .85
        Controls         .55     [95% Conf. Interval]
                    ---------    --------------------
        difference        .3      .2178361   .3821639
        ratio       1.545455      1.359508   1.756834
        rel. diff.  .6666667      .5542088   .7791246

        odds ratio         7       3.58949   15.23321   (exact)
```

Example 13.2 continued: Analysis of the diabetic retinopathy data using `mcc`

```
.   list id new standard
            id        new   standard     |eye given new treatment
  1.         1          0          1  -- |not blind, standardly
  2.         2          0          1     |treated eye blind
  3.         3          1          0
   .         .          .          .
   .         .          .          .
 76.        76          1          1  -- |both eyes blind
 77.        77          0          0  -- |neither eye blind

 . mcc new standard
                  | Controls               |
Cases             |   Exposed   Unexposed  |     Total
------------------+------------------------+----------
         Exposed |         6           4  |        10
       Unexposed |        28          39  |        67
-----------------+------------------------+----------
           Total |        34          43  |        77

McNemar's chi2(1) =     18.00    Prob > chi2 = 0.0000
Exact McNemar significance probability       = 0.0000

Proportion with factor
        Cases       .1298701
        Controls    .4415584     [95% Conf. Interval]
                   ---------     --------------------
        difference -.3116883     -.4507167    -.17266
        ratio       .2941176      .1612072   .5366088
        rel. diff. -.5581395     -.8799927  -.2362864

        odds ratio  .1428571      .0364098   .4083484   (exact)
```

13.10 Conditional versus marginal odds ratios

Odds ratios from matched analyses are *conditional odds ratios*. Their interpretation is *conditional* on the blocks. Conditional odds ratios can be contrasted with *marginal odds ratios* or *population odds ratios* - these are odds ratios within a certain population with its given distribution of other features. We discussed this issue in a previous session, where we saw that in some settings the marginal and conditional odds ratios could both be of interest.

Conditional odds ratios can be converted to marginal odds ratios if we know the distribution of risks in unexposed/untreated people in that population. It is possible to estimate this distribution in a cohort study or intervention study. But this distribution cannot be inferred from the data in a case-control study. In a 1-1 matched case-control study the prevalence of the outcome is 50% and this renders the distribution of other features atypical of the population. Even the distribution of other features in controls alone will be atypical. So it is never appropriate to try and estimate a marginal odds ratio in a matched case-control study using just the case-control study data.

Example 13.2 continued: Marginal and conditional odds ratios in the diabetic retinopathy study

For the diabetic retinopathy study, the estimated conditional odds ratio relating type of treatment to odds of blindness is $s/r = 0.143$. The blocks are individual patients with diabetic retinopathy. The interpretation is that, *for eyes from the same individual*, the odds of an eye given the new treatment going blind are 0.143 times that of an eye given the standard treatment. In equations:

$$\text{Conditional OR} = \frac{\Pr(\text{Blind}|\text{new, individual } i)/\Pr(\text{Not blind}|\text{new, individual } i)}{\Pr(\text{Blind}|\text{standard, individual } i)/\Pr(\text{Not blind}|\text{standard, individual } i)}$$

This is assumed to be the same for all i.

The conditional odds ratios is *not equivalent to saying* that across the population the odds of eyes given the new treatment going blind are 0.143 times that of eyes given the standard treatment. That would be given by the marginal odds ratio:

$$\text{Marginal OR} = \frac{\Pr(\text{Blind}|\text{new})/\Pr(\text{Not blind}|\text{new})}{\Pr(\text{Blind}|\text{standard})/\Pr(\text{Not blind}|\text{standard})}$$

To estimate the marginal odds ratio we would need to know something about the individuals in the population, in particular concerning their risk of blindness. Suppose that we have the following two pieces of information:

- 50% of patients can be classified as 'high risk' (HR) and the other 50% as 'low risk' (LR).
- Using the standard treatment, the 'high risk' patients have a 90% chance of going blind and the 'low risk' patients have a 10% chance of going blind.

This information gives us:

$$\begin{aligned}\Pr(\text{Blind}|\text{standard}) &= \Pr(\text{Blind}|\text{standard, HR})Pr(\text{HR}) + \Pr(\text{Blind}|\text{standard, LR})\Pr(\text{LR}) \\ &= 0.9 \times 0.5 + 0.1 \times 0.5 = 0.5.\end{aligned}$$

We also need the probability Pr(Blind|new). To find this we use the conditional OR together with the above information, which gives us:

$$\begin{aligned}\frac{\Pr(\text{Blind}|\text{new, HR})}{\Pr(\text{Not blind}|\text{new, HR})} &= 0.143 \times \frac{\Pr(\text{Blind}|\text{standard, HR})}{\Pr(\text{Not blind}|\text{standard, HR})} = 0.143 * \frac{0.9}{0.1} = 1.287 \\ \frac{\Pr(\text{Blind}|\text{new, LR})}{\Pr(\text{Not blind}|\text{new, LR})} &= 0.143 \times \frac{\Pr(\text{Blind}|\text{standard, LR})}{\Pr(\text{Not blind}|\text{standard, LR})} = 0.143 * \frac{0.1}{0.9} = 0.016.\end{aligned}$$

Rearrangement gives:

$$\begin{aligned}\Pr(\text{Blind}|\text{new, HR}) &= 1.287/(1 + 1.287) = 0.563 \\ \Pr(\text{Blind}|\text{new, LR}) &= 0.016/(1 + 0.016) = 0.016.\end{aligned}$$

Hence we can find

$$\begin{aligned}\Pr(\text{Blind}|\text{new}) &= \Pr(\text{Blind}|\text{new, HR})\Pr(\text{HR}) + \Pr(\text{Blind}|\text{new, LR})\Pr(\text{LR}) \\ &= 0.563 \times 0.5 + 0.016 \times 0.5 = 0.290.\end{aligned}$$

The marginal OR is therefore

$$\text{Marginal OR} = \frac{\Pr(\text{Blind}|\text{new})/\Pr(\text{Not blind}|\text{new})}{\Pr(\text{Blind}|\text{standard})/\Pr(\text{Not blind}|\text{standard})} = \frac{0.5/(1-0.5)}{0.290/(1-0.290)} = 0.408.$$

Here the marginal OR is quite markedly closer to 1 than the conditional OR.

13.11 Practical 13

Dataset required: `vit_E.dta`

Introduction

The aim of this practical is to consolidate the ideas of matching discussed in the lecture. We will use data from a matched case-control study of the association between vitamin E measured in the blood and occurrence of cancer. The dataset is called vit_E.dta

High levels of vitamin E are thought by some to be protective against cancer. This hypothesis was investigated by measuring Vitamin E in stored blood samples from 271 men in a large cohort study initially aged 35–64 years who subsequently developed cancer. These values were compared with those from control men who had not, at that time, developed cancer. One control for each case was selected randomly from within the cohort study, subject to matching for age (within 5 years), duration of storage of the blood sample (within 3 months) and smoking status.

Note that this dataset does not include the matching variables. It does include 'observation time', which is the time from blood collection to diagnosis of cancer in the cases. In this session we will analyse the data assuming no confounding.

Variable	Description
`set`	Case-control unique identifier taking values from 1 to 271
`case`	0: control 1: case
`vitE`	Vitamin E (mg/dl)
`time`	Observation time: 1: up to 1 year 2: 1-3 years 3: over 3 years

Aims

- Understand how to appropriately conduct hypothesis tests on data from a matched case-control study.
- Use appropriate statistical tests to analyse matched studies with continuous dependent variables.

Analysis

1 Start by looking at the form of the data. It will help to sort the data so that the data from the same set (block) are shown on consecutive rows. Use the code below to list the first five sets, with the control appearing first within each pair.

```
sort set case
list in 1/10, sep(2)
```

2 (a) Summarize the vitamin E measurements for cases and controls. What is the difference between the means in the two groups?

Hint: use the `bysort` prefix to `summarize` to get the summaries within cases and controls automatically.

(b) Use suitable plots to display vitamin E levels in the cases and controls.

For questions 3-5 we will analyse these data treating vitamin E as the dependent variable, and cancer as the independent variable. In reality the vitamin E reading is best thought of as an exposure since blood was taken and stored before the onset of cancer.

3 We will first analyse these data using linear regression.

(a) We first demonstrate an invalid approach, to demonstrate the importance of allowing for block in any analysis. Fit a linear regression model with just the case-control variable as an explanatory variable.

```
regress vitE i.case
```

What is the coefficient for the case-control variable? What is the reported standard error?

(b) Now add the set variable, as a categorical predictor in the regression model.

```
regress vitE i.case i.set
```

What is the coefficient for the case-control variable? What is the reported standard error?

Discuss: Compare the estimates of the case-control coefficient and its standard error. Is there evidence that people who developed cancer had different vitamin E levels than people who did not develop cancer?

4 In order to make the data suitable for analysis with a paired t-test (and other analyses), we need to reconfigure it so that the data from each pair is contained in one row. This can be done using the reshape command in Stata, which is an extremely useful command, but can be fiddly to use in practice.

Look at your data in the browser, then use the following command to reshape it.

```
reshape wide vitE, i(set) j(case)
```

Now look at the data again in the browser to see how it is now arranged.

5 With the data in wide format, carry out a paired t-test of the null hypothesis that the mean blood vitamin E levels are the same in cases and controls.

Discuss: How do the estimated mean difference and its standard error compare to those from the linear regression models above?

For the remainder of this session we will dichotomise vitamin E into a binary variable and perform analysis to estimate the odds ratio of association. Here we can use the reversible nature of an odds ratios so that we can interpret vitamin E as the exposure and cancer as the outcome or vise-versa.

6 (a) Define a "high" vitamin E status as a level above 12mg/dl and complete the table below. The table records counts of sets, that is pairs of people, one of whom developed cancer and the other who did not.

		Controls		
	Vitamin E status	Low	High	Total
Cases	Low			
	High			
	Total			

[Hint: create two new variables for vitamin E status in cases and controls respectively, which take the value 1 for individuals with high vitamin E and the value 0 for individuals with low vitamin E. In the Stata do file these variables are h_vitE_control and h_vitE_case.]

(b) Using the completed table, compute the test statistic for McNemar's test. What is the null hypothesis?

(c) Use the `bitesti` command to perform a test of this null hypothesis.

(d) Again using the completed table, calculate the estimated odds ratio by hand.

(e) Use the `cii` command to calculate an exact 95% confidence interval for this estimate.

7 Use the `mcc` command to analyse this matched case-control study using dichotomised vitamin E as the exposure.

Discuss: What are your epidemiological conclusions from the analysis where vitamin E status is dichotomised? Working together with one or more colleagues (in your Breakout Room if online), write a short paragraph to summarise your findings. If online, one of you should post your group's paragraph in the Zoom chat.

Session 14

Conditional Logistic Regression

14.1 Aims

The aim of this lecture is to build on what we learnt in previous sessions about matched studies and about logistic regression in cohort and case-control studies, by introducing conditional logistic regression for matched studies. Following this lecture and practical you should be able to do the following:

- Write down a logistic regression model suitable for matched studies.
- Outline the steps leading to the use of conditional logistic regression for a matched case-control study for a single binary exposure and in the more general situation of continuous or multiple exposures.
- Describe why an unconditional logistic regression analysis would be inappropriate for matched studies when using a logistic model which includes a separate parameter for each matched set, and summarize how such an analysis would affect the odds ratio estimates.
- Outline how an unconditional logistic regression analysis could be used for matched case-control data for some types of matching variables.

14.2 Logistic models for matched studies

14.2.1 Preliminaries

In this section we focus on matched studies where the blocks or matched sets are *pairs* and where there is a single binary exposure of interest. Suppose that our matched data is made up of n matched pairs. We begin by defining the exposure X_i and outcome Y_i which refer to members of the sub-population from which the ith matched pair originates ($i = 1, \ldots, n$). The sub-population here is that in which all individuals have the same features as those two individuals in pair i. In some special circumstances that sub-population may only contain those two individuals (e.g. where the pair is the left and right eye of an individual).

Note that here (in contrast to the second half of session 13) we are reverting to the usual notation for matched case-control studies where X denotes exposure and Y the outcome. We do this because the most common use of conditional logistic regression is in the context of the analysis of matched case-control studies.

Generalizing the probabilities π_{xy} introduced previously, we define

$$\pi_{i;xy} = \Pr(X_i = x, Y_i = y). \tag{14.1}$$

The odds ratio relating exposure to disease for pair i is then

$$(\pi_{i;11}\pi_{i;00})/(\pi_{i;10}\pi_{i;01}). \tag{14.2}$$

Only one matched pair gives information about this odds ratio. Clearly there is far too little information in the data to proceed without making further assumptions. The most direct approach, in the absence of additional information, is to assume that the true log odds ratio relating exposure to disease is the same for all pairs, that is that

$$(\pi_{i;11}\pi_{i;00})/(\pi_{i;10}\pi_{i;01}) = e^{\beta} \tag{14.3}$$

for all $i = 1, \ldots, n$. This means that although the pairs have different features by definition, the association between the exposure and the outcome is assumed to be the same in the different sub-populations from which the pairs originate.

We showed in the previous lecture that a way of estimating the odds ratio in the pair-matched setting is to use the Mantel-Haenszel method. Here we instead pursue a logistic-regression-based approach. We start by considering a matched case-control study setting.

14.2.2 Matched case-control study

In the session on logistic regression in cohort and case-control studies we defined the following logistic model for an *unmatched* case-control study with a binary exposure

$$\Pr(X_k = 1|Y_k = y_k) = \frac{e^{\lambda^* + \beta y_k}}{1 + e^{\lambda^* + \beta y_k}} \tag{14.4}$$

where k refers to an individual.

Here we extend this to the matched pair setting. Each pair contains one case and one control. We let X_{i0} denote the value of the binary explanatory variable X for the control in the ith matched pair, and X_{i1} denote the value of X for the case in the ith matched pair. A logistic model which incorporates the matched pair information is

$$\Pr(X_{i0} = 1) = \frac{e^{\lambda_i^*}}{1 + e^{\lambda_i^*}}, \qquad \Pr(X_{i1} = 1) = \frac{e^{\lambda_i^* + \beta}}{1 + e^{\lambda_i^* + \beta}} \tag{14.5}$$

Another way of writing this is

$$\Pr(X_{iy} = 1) = \frac{e^{\lambda_i^* + \beta y}}{1 + e^{\lambda_i^* + \beta y}}, \qquad y = 0, 1. \tag{14.6}$$

The parameter β is the log odds ratio in pair i, which is assumed to be the same in all pairs $i = 1, \ldots, n$. The parameter λ_i^* is specific to the ith matched pair. It is the log odds of exposure ($X = 1$) for the control in the ith pair:

$$\lambda_i^* = \log\left(\frac{\pi_{i;10}}{\pi_{i;00}}\right). \tag{14.7}$$

There are four possible outcomes (x_{i1}, x_{i0}) for pair i, namely (0, 0), (1, 0), (0, 1), (1, 1). The likelihood from pair i is

$$\frac{e^{\lambda_i^* x_{i0}}}{(1 + e^{\lambda_i^*})} \frac{e^{(\lambda_i^* + \beta)x_{i1}}}{(1 + e^{(\lambda_i^* + \beta)})}. \tag{14.8}$$

The full likelihood is formed from this by multiplying over i and therefore is

$$\frac{\exp(\Sigma \lambda_i^* x_{i0})}{\Pi(1 + e^{\lambda_i^*})} \frac{\exp\{\Sigma(\lambda_i^* + \beta)x_{i1}\}}{\Pi(1 + e^{(\lambda_i^* + \beta)})}. \tag{14.9}$$

where the sums and products are over the matched sets i.The contribution to the likelihood in (14.8) is the only part of the total likelihood involving λ_i^*; that is, the only information we have about λ_i^* is that which comes from the case and the control in the ith matched pair. Because of the very small amount of information available about the parameters λ_i^* it would be a mistake to estimate β by maximising the above full likelihood. The consequences of doing this are discussed in a later section.

A different form of analysis is therefore required.

14.2.3 Conditional logistic regression for other matched studies

We comment briefly on the corresponding arguments for matched cohort studies (and other matched studies such as matched cross-over studies and matched intervention and cross-sectional studies). In a matched cohort study an exposed individual is matched to an unexposed individual, which is in contrast to the matched case-control study where a case is matched to a control. We use notation analogous to that used above and let Y_{i1} denote the outcome (0 or 1) for the exposed individual in the ith matched pair and Y_{i0} denote the outcome (0 or 1) for the unexposed individual in the ith matched pair.

A logistic model for the outcome Y given the exposure X which incorporates the matched pair information is

$$\Pr(Y_{i0}=1)=\frac{e^{\lambda_i}}{1+e^{\lambda_i}}, \qquad \Pr(Y_{i1}=1)=\frac{e^{\lambda_i+\beta}}{1+e^{\lambda_i+\beta}} \tag{14.10}$$

As in the unmatched situation discussed in the session on logistic regression in cohort and case-control studies, λ_i and λ_i^* are different, but the models for the matched cohort study and matched case-control study have the same log odds ratio parameter β. This is assuming we are comparing equivalent cohort and case-control studies which are based on the same exposure and outcome and using the same matching variables.

In the matched cohort setting there are four possible outcomes (y_{i1}, y_{i0}) for pair i and the the likelihood from pair i is

$$\frac{e^{\lambda_i y_{i0}}}{(1+e^{\lambda_i})}\frac{e^{(\lambda_i+\beta)y_{i1}}}{(1+e^{\lambda_i+\beta})}. \tag{14.11}$$

The full likelihood is formed from this by multiplying over i. Again, the contribution to the full likelihood from the ith matched pair is the only part of the full likelihood involving the parameter λ_i. Because of this, again it would be a mistake to estimate β by maximising the full likelihood and an alternative approach to estimating β is needed.

14.3 Conditional logistic regression: binary exposure

In this section we return to focusing on a matched case-control study with one control per case, i.e. pair matching, in the simplest situation of a single binary exposure. Our aim is to find an analysis which allows us to somehow manage to estimate the log odds ratio parameter β without having to also estimate the nuisance parameters λ_i^* $(i=1,\ldots,n)$.

14.3.1 Note on sufficient statistics

The key to estimating parameters of interest without having to estimate nuisance parameters is to make use of the *sufficient statistics* for the nuisance parameters, which here are the parameters λ_i^* $(i = 1, \ldots, n)$. The formal definition of a sufficient statistic is as follows. Suppose we have a probability function or density for the random variable $\mathbf{y}$ which depends on the parameters $\theta_1, \ldots, \theta_p$:

$$f(\mathbf{y} \mid \theta_1, \ldots, \theta_p).$$

The statistic T_k, which is some function of the data, is said to be 'sufficient' for the parameter θ_k if the conditional distribution of $\mathbf{y}$ given T_k does not depend on θ_k. More informally, T_k is a sufficient statistic for θ_k if T_k gives us all the information we would need to estimate θ_k, i.e. if we had the entire sample data we would do not better at estimating θ_k than if we have T_k, which is some function of the data.

For example, suppose we have some sample data $y_1, y_2, \ldots, y_n$ that is assumed to come from a normal distribution and we wish to estimate the mean of that distribution, denoted μ. It is assumed that the variance of the distribution, σ^2 say, is known. The likelihood is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\}. \tag{14.12}$$

To find the maximum likelihood estimate for μ all we need is $\sum_i y_i$ (or equivalently $\bar{y}$). We can not do better in estimating μ by somehow bringing in the whole sample data. Therefore $\sum_i y_i$ is the sufficient statistic for μ.

14.3.2 Development of the conditional logistic regression model

In our context of a pair-matched case-control study it can be shown that the set of exposures (x_{i0}, x_{i1}) is sufficient for the nuisance parameter λ_i^* in the ith matched pair. The set of exposures (x_{i0}, x_{i1}) does not include the information on which individual in the pair the values (x_{i0}, x_{i1}) came from, i.e. it does not tell us which value belongs to the case and which belongs to the control. For a binary exposure, knowing the set of exposures (x_{i0}, x_{i1}) is equivalent to knowing the total number of subjects who are exposed in each case-control set. The set of exposures (x_{i0}, x_{i1}) can therefore be replaced by $T_i = x_{i0} + x_{i1}$. There are three possible values for T_i:

0: case and control both exposed.

1: either the case or the control is exposed and the other is unexposed.

2: both the case and the control are exposed.

Using the information given above about sufficient statistics, we condition on $T_i = x_{i0} + x_{i1}$ when finding the contribution to the likelihood for the ith matched set. With this conditioning, the only random phenomenon left to assess is how the fixed pattern of exposures is distributed to the case and control. That is, we calculate the probability that the case (and the control) had its observed exposure conditional on the fact that we know that these exposure patterns belong to this particular matched pair. In equations, we are interested in the conditional probabilities:

$$\Pr(X_{i0} = x_{i0}, X_{i1} = x_{i1} | T_i = x_{i0} + x_{i1}) \tag{14.13}$$

If $x_{i0} = x_{i1}$ (either case and control both exposed or neither exposed) the conditional probability that the case (and the control) is assigned its observed exposure is 1, i.e. the probability that the case is exposed and the control is exposed given that both the case and the control are exposed is 1. In equations:

$$\Pr(X_{i0} = 0, X_{i1} = 0|T_i = 0) = 1, \quad \Pr(X_{i0} = 1, X_{i1} = 1|T_i = 2) = 1. \tag{14.14}$$

Now suppose x_{i0} and x_{i1} are different (T_i =1). There are only two possibilities, either the case is exposed and the control not, or vice-versa:

$$\Pr(X_{i0} = 1, X_{i1} = 0|T_i = 1), \quad \Pr(X_{i0} = 0, X_{i1} = 1|T_i = 1). \tag{14.15}$$

These probabilities can be written as

$$\Pr(X_{i0} = 1, X_{i1} = 0|T_i = 1) = \frac{\Pr(X_{i0} = 1)\Pr(X_{i1} = 0)}{\Pr(X_{i0} = 1)\Pr(X_{i1} = 0) + \Pr(X_{i0} = 0)\Pr(X_{i1} = 1)} \tag{14.16}$$

$$\Pr(X_{i0} = 0, X_{i1} = 1|T_i = 1) = \frac{\Pr(X_{i0} = 0)\Pr(X_{i1} = 1)}{\Pr(X_{i0} = 1)\Pr(X_{i1} = 0) + \Pr(X_{i0} = 0)\Pr(X_{i1} = 1)}. \tag{14.17}$$

Using the logistic model introduced in the previous section we can see that

$$\begin{aligned}\Pr(X_{i0} = 1, X_{i1} = 0|T_i = 1) &= \frac{e^{\lambda_i^*}/(1 + e^{\lambda_i^*+\beta})(1 + e^{\lambda_i^*})}{e^{\lambda_i^*}/(1 + e^{\lambda_i^*+\beta})(1 + e^{\lambda_i^*}) + e^{\lambda_i^*+\beta}/(1 + e^{\lambda_i^*+\beta})(1 + e^{\lambda_i^*})} \\ &= \frac{1}{1 + e^{\beta}}.\end{aligned} \tag{14.18}$$

Similarly we can find that

$$\Pr(X_{i0} = 0, X_{i1} = 1|T_i = 1) = \frac{e^{\beta}}{1 + e^{\beta}}. \tag{14.19}$$

14.3.3 The conditional likelihood

The data from a pair-matched case-control study with a binary exposure can be tabulated in the matched format shown in Table 14.1. This table is similar to that used in the previous session in the discussion of McNemar's Test. Here we use a different notation for the numbers in the table.

Table 14.1: Data from a matched case control study with a single binary exposure

		Y=1	
		$X = 0$	$X = 1$
$Y = 0$	$X = 0$	n_{00}	n_{10}
	$X = 1$	n_{01}	n_{11}

Using the results in the preceding section and the data in Table 14.1, the conditional likelihood for a matched case-control study with 1 control per case and a single binary exposure is

$$L = \left(\frac{e^{\beta}}{1 + e^{\beta}}\right)^{n_{10}} \left(\frac{1}{1 + e^{\beta}}\right)^{n_{01}}. \tag{14.20}$$

We can find the maximum likelihood estimate for β in the usual way. The log likelihood is:

$$l = n_{10}\beta - (n_{10} + n_{01}) \log(1 + e^{\beta}) \tag{14.21}$$

with first derivative

$$\frac{\mathrm{d}l}{\mathrm{d}\beta} = n_{10} - (n_{10} + n_{01}) \frac{e^{\beta}}{1 + e^{\beta}}. \tag{14.22}$$

This gives us the maximum likelihood estimate for β:

$$\hat{\beta} = \log \frac{n_{10}}{n_{01}}. \tag{14.23}$$

Notice how, as with the application of the Mantel-Haenszel odds ratio to matched case-control data, sets where both case and control are exposed, or both are unexposed, give no information. Conditional logistic regression for matched case control data with binary exposure gives the same odds ratio as Mantel-Haenszel. The inference is asymptotic.

14.3.4 Extensions

Above we focus on a pair-matched case-control study and single binary exposure. The conditional argument used in this section to obtain the conditional logistic regression model can be extended as follows:

- The argument extends easily to matched cohort studies or matched randomized trials, by replacing X_{iy} $(y = 0, 1)$ by Y_{ix} $(x = 0, 1)$ in the above workings. The resulting conditional likelihood is the same as given above except that n_{10} becomes the number of pairs in which the exposed individual becomes a case and the unexposed individual becomes a control, and vice versa for n_{01}.
- Matched case-control studies with more than one control per case. It is common to match up to about 5 controls per case.
- Matched case-control studies with multiple exposures, which may combine both categorical and continuous exposures. We outline this more general situation in the next section.

Exercise: Derive the likelihood for a conditional logistic regression of a matched case-control study with 2 controls per case, still assuming a binary exposure. Hint: You may wish to use the following notation:

X_{i1}: exposure for the case in the ith matched set

X_{i0_1}: exposure for the first control in the ith matched set

X_{i0_2}: exposure for the second control in the ith matched set

Consider the sufficient statistic $T_i = X_{i1} + X_{i0_1} + X_{i0_2}$.

14.4 Conditional logistic regression: general situation

We now consider a general predictor X which may be binary, categorical or continuous, and which may be a vector of predictors of various types. We are concerned with the conditional distribution of the predictors given case or control status, denoted $P(X_{i1} = x)$ and $P(X_{i0} = x)$. Suppose that in a particular matched set the control is observed to have predictors x_0 and the case is observed to have predictors x_1. The conditional probability of interest is now the probability that the individual with predictors x_0 is in fact the control and the individual with predictors x_1 is the case, given that the predictors for the matched pair are (x_0, x_1). The sufficient statistic is the set of predictors (x_0, x_1) and the joint conditional distribution of interest is

$$\begin{aligned} &\mathrm{P}(X_{i0} = x_0, X_{i1} = x_1 | T_i = (x_0, x_1)) \\ &= \frac{\mathrm{P}(X_{i0} = x_0)\mathrm{P}(X_{i1} = x_1)}{\mathrm{P}(X_{i0} = x_0)\mathrm{P}(X_{i1} = x_1) + \mathrm{P}(X_{i0} = x_1)\mathrm{P}(X_{i1} = x_0)} \end{aligned} \tag{14.24}$$

It is only for single binary X that the distributions $P(X_{i1} = x)$ can be formulated using logistic models so here we need to do something different. The solution is to proceed as we did in the session on logistic regression in cohort and case-control studies in the setting of an unmatched case-control study. We start by letting Y_{ix} denote the case or control status for an individual in the ith matched set with predictor x. A logistic regression model for Y_{ix} is

$$\Pr(Y_{ix} = 1) = \frac{e^{\lambda_i + \beta^T x}}{1 + e^{\lambda_i + \beta^T x}}. \tag{14.25}$$

This model refers to the underlying population. By Bayes Theorem we have

$$P(X_{i1} = x) = \Pr(Y_{ix} = 1) \times P(X_{i.} = x) / \Pr(Y_{i.} = 1) \tag{14.26}$$

where $P(X_{i.} = x)$ refers to the unconditional (marginal) distribution of X in the sub-population which generates the ith matched set and $\Pr(Y_{i.} = 1)$ is the unconditional probability of being a case in that sub-population. Plugging this into the above conditional joint distribution gives

$$\begin{aligned} &\mathrm{P}(X_{i0} = x_0, X_{i1} = x_1 | T_i = (x_0, x_1)) \\ &= \frac{\Pr(Y_{ix_0} = 0)\Pr(Y_{ix_1} = 1)}{\Pr(Y_{ix_0} = 0)\Pr(Y_{ix_1} = 1) + \Pr(Y_{ix_0} = 1)\Pr(Y_{ix_1} = 0)}. \end{aligned} \tag{14.27}$$

Using the above logistic regression model this simplifies to

$$\mathrm{P}(X_{i0} = x_0, X_{i1} = x_1 | T_i = (x_0, x_1)) = \frac{e^{\beta^T x_1}}{e^{\beta^T x_1} + e^{\beta^T x_0}}. \tag{14.28}$$

This is the contribution to the conditional likelihood for the ith matched pair. The full likelihood is the product over such terms for all pairs, giving:

$$L_{\text{matched}} = \prod_i \frac{\exp\{\beta^T x_{i1}\}}{\exp\{\beta^T x_{i1}\} + \exp\{\beta^T x_{i0}\}}. \tag{14.29}$$

We can extend this to a matched case-control study with c controls per case, i.e. each matched set comprises 1 case and c controls. The likelihood for a conditional logistic regression is

$$L_{\text{matched}} = \prod_i \frac{\exp\{\beta^T x_{i1}\}}{\exp\{\beta^T x_{i1}\} + \sum_{k=1}^{c} \exp\{\beta^T x_{i0k}\}} \tag{14.30}$$

where x_{i0k} denotes the exposure for the kth control in the ith matched set.

14.5 Conditional logistic regression: interactions between exposures and matching variables

In a matched study our design precludes us from estimating main effects of the matching variables. For example, in the smoking and lung cancer matched case-control study we cannot estimate the effect of sex (one of the matching variables) because there is no variability in gender within matched pairs. However, it is possible to estimate interactions between matching variables and exposures, for example we can use the smoking and lung cancer data to explore whether the odds ratio relating smoking to lung cancer differs between males and females. This is one of those relatively rare situations in statistics where it makes sense to include a two-way interaction in a statistical model without including the main effect of one of the variables involved in the interaction.

14.6 Conditional logistic regression: implementation

It can be seen that the conditional probabilities are in the form of a logistic regression, and so the conditional MLE's can be obtained by setting up a logistic regression in which the units for analysis are the matched pairs with either just the case or the control exposed.

In practice it can be tedious, and a potential source of error, to set up a new form of the data set for the conditional logistic regression as described above. Fortunately, statistical packages now have conditional logistic regression built in, which removes this extra step. For example, in Stata we use the `clogit` command, specifying the variable containing the matched (case-control) set indicator in a `group(varname)` option after the comma. There is a similar command in `R` (`clogit` in the `survival` library).

14.6.1 Examples

We first show how to analyse the lung cancer and smoking case-control study using `clogit`. To use this command it is first necessary to reshape the data so that each line holds data for a single individual. The case-control status also needs to be coded as 0 and 1.

Example 13.5 continued: Analysis of the smoking and lung cancer case-control study using conditional logistic regression

```
. list
            id       case    control
  1.         1          1          1 -- case and control smoked
  2.         2          0          0 -- case and control never smoked
  3.         3          1          0 -- case smoked, control never smoked
   .         .          .          .
 200.      200          1          1

. rename case smoke2
. rename control smoke1
. reshape long smoke, i(id) j(casecon)
. replace casecon=casecon-1
. list
            id    casecon      smoke

  1.         1          0          1 -- control smoked
  2.         1          1          1 -- case smoked
  3.         2          0          0 -- control never smoked
  4.         2          1          0 -- case never smoked
  5.         3          0          0 -- control never smoked
  6.         3          1          1 -- case smoked
   .         .          .          .
399.       200          0          1
400.       200          1          1

. clogit casecon i.smoke, group(id) or

Iteration 0:   log likelihood = -114.14167
Iteration 1:   log likelihood = -113.31955
Iteration 2:   log likelihood = -113.31927
Iteration 3:   log likelihood = -113.31927

Conditional (fixed-effects) logistic regression

                                                Number of obs     =        400
                                                LR chi2(1)        =      50.62
                                                Prob > chi2       =     0.0000
Log likelihood = -113.31927                     Pseudo R2         =     0.1826

------------------------------------------------------------------------------
     casecon | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.smoke |          7   2.366432     5.76   0.000     3.608605    13.57865
------------------------------------------------------------------------------
```

Notice that the estimated odds ratio is the same as when the study was analysed using the `mcc` command in the previous session. However, the 95% CI is somewhat different.

We now turn to the analysis of the diabetic retinopathy trial. When conditional logistic regression is used for a matched case-control study the dependent variable should be the case-control status, whereas in matched cohort and intervention studies the dependent variable should be the outcome (here blindness). In a simple setting with a single binary exposure and a single binary outcome exchanging the outcome and exposure will not alter

the results, but in more complex situations with multiple covariates it will.

Example 13.2 continued: Analysis of the diabetic retinopathy data using conditional logistic regression

```
. list
            id      treat      blind

  1.         1          1          1
  2.         1          2          0
  3.         2          1          1
  4.         2          2          0
  5.         3          1          0
  6.         3          2          1
   .
154.        77          1          0
154.        77          2          0

. clogit blind i.treat, group(id) or
note: multiple positive outcomes within groups encountered.
note: 45 groups (90 obs) dropped because of all positive or
      all negative outcomes.

Iteration 5:   log likelihood = -12.056645

Conditional (fixed-effects) logistic regression

                                                Number of obs     =         64
                                                LR chi2(1)        =      20.25
                                                Prob > chi2       =     0.0000
Log likelihood = -12.056645                     Pseudo R2         =     0.4564

------------------------------------------------------------------------------
       blind | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.treat |   .1428571   .0763604    -3.64   0.000      .050109    .4072755
------------------------------------------------------------------------------
```

As with the smoking and lung cancer example the estimated odds ratio is identical to that obtained using **mcc**, but the 95% CI is somewhat different.

14.7 Non-regularity of the matched case-control model

There may be a temptation to deal with the matched pair analysis by a direct use of maximum likelihood, obtaining an estimate by maximizing the full unconditional likelihood (14.9) with respect to $(\beta, \lambda_1^*, \ldots, \lambda_n^*)$. To preserve the matching in the analysis, a separate parameter is needed for each case-control set. As the sample size (the number of sets in this situation) increases, the number of parameters in the model increases at the same rate. This is a 'non-regular' problem, and the usual properties of MLE's do not apply, in particular the parameter estimates will not be consistent, that is, they will remain biased even as the sample size tends to ∞.

Let's investigate this in more detail in the simplest situation of a matched study with one control per case and a single binary exposure.

The concordant case-control pairs make no contribution to the estimate of β from the unconditional likelihood, and it can be shown that for one-to-one matching the profile unconditional likelihood for β is

$$\begin{aligned} & n_{10}\left\{\frac{\beta}{2}-\log\left(1+e^{-\frac{\beta}{2}}\right)-\log\left(1+e^{\frac{\beta}{2}}\right)\right\} \\ + \quad & n_{01}\left\{-\frac{\beta}{2}-\log\left(1+e^{-\frac{\beta}{2}}\right)-\log\left(1+e^{\frac{\beta}{2}}\right)\right\}. \end{aligned} \tag{14.31}$$

Differentiation with respect to β gives the estimated log odds ratio to be

$$2\log\left(\frac{n_{10}}{n_{01}}\right), \tag{14.32}$$

which is twice the consistent log odds ratio estimate from the conditional analysis. It can also be shown that the asymptotic variance of the unconditional log odds ratio estimate is twice that of the conditional log odds ratio estimate.

Although the unconditional analysis using the full likelihood in (14.9) might seem a good way of analysing the matched case-control data, it results in an inconsistent estimate of the odds ratio, with the relative magnitude of the inconsistency becoming more severe as the true odds ratio increases. This is because maximum likelihood theory assumes that the parameter space is fixed while the number of observations, matched sets in this case, tends to infinity. In the unconditional analysis, however, the dimension of the parameter space increases as the number of observations increases.

14.8 Alternative analyses for matched case-control data

We return for a moment to the logistic model for matched studies, from which the conditional likelihood for matched case-control studies is derived:

$$\Pr(Y_{ix}=y)=\frac{e^{\lambda_i+\beta^T x_{iy}}}{1+e^{\lambda_i+\beta^T x_{iy}}}, \quad y=0,1. \tag{14.33}$$

The regularity problem was discussed in Section 14.7, and the conditioning approach is one solution to that problem. In this section we discuss briefly two alternative approaches:

1 Assume that the matched set parameters are themselves a sample from some distribution - this leads to the so-called **generalized linear mixed model**, which is introduced in Advanced Statistical Modelling.

2 A different approach is to represent the variation between matched sets by including appropriate regression terms in the logistic model instead of having a separate parameter λ_i for each set. This is followed by an unconditional logistic regression analysis.

We focus on the second possibility. Its use depends on the matching criteria which were used. There are some types of matching criteria that are easily summarised in one or a small number of variables, for example if case and controls are matched on age and sex. In this situation instead of using the model in (14.33) we could use the model

$$\Pr(Y_{ix}=y)=\frac{e^{\lambda+\beta^T x_{iy}+\gamma\text{age}_i+\delta\text{sex}_i}}{1+e^{\lambda_i+\beta^T x_{iy}+\gamma\text{age}_i+\delta\text{sex}_i}}, \quad y=0,1. \tag{14.34}$$

where age_i denotes the age of the case and controls in the uth matched set and sex_i denotes their sex. These two variables are the same for all individuals in the set.

We make the following comments on this approach, as contrasted with the conditional approach:

- If we applied the conditioning approach to this model the age and sex terms would be eliminated and we would end up with the same conditional logistic regression model as before.
- It might be thought that no adjustment for the matching variables is necessary. However, this is not true because although matching results in unconditional independence of the matching variables and the outcome this does not guarantee that the matching variables and the outcome remain independent conditional on exposure. An adjustment should always be made.
- It may appear that an advantage of this approach is that we can estimate the effects of the matching variables if they are of interest. However, the parameters associated with age and sex are not the same parameters as would be the case if we had not matched on age and sex - because these variables were used in the matching.
- A drawback of this approach is that we have to specify a form for the effect of the matching variables on the outcome. In the above model the effect of age on the log odds is assumed to be linear. However, in reality the association may be non-linear. By using the conditional approach we avoid having to specify a form for the relationship between matching variables and the outcome.
- In the conditional approach, although we cannot estimate the main effects of matching variables we can still estimate interactions between matching variables and the main exposures. The same goes for the above unconditional regression approach - but we can still estimate any interaction terms without bias from the case-control study.
- There are some types of matching criteria which are not easily summarized in one or more regression variables. For example, if cases are matched on 'neighbourhood', then it may be difficult to capture everything about this in regression variables, though attempts could be made to at least partially capture the matching variable, e.g. using available measures of social deprivation within geographical areas. This could result in residual confounding, however, if not all the important features are captured.
- If we are interested in the main effect of a particular variable on the outcome then that variables should not be used as part of the matching criteria.
- If matching variables are not related to the outcome, then there may be some loss of precision.

14.9 Practical 14

Datasets required: `vitE.dta` and `infertility.dta`

Introduction

There are two parts to this session.

In the first part we re-visit the vitamin E dataset from the previous practical and apply the conditional logistic regression methods we learned in lecture 13.

In the second part we use a new dataset and investigate the association between infertility in women and number of previous spontaneous and induced abortions.

Aims

- Learn how to fit a conditional logistic regression model to matched data (in Stata).
- Learn how to fit a standard logistic regression model to matched data by including the matching variables (in Stata).
- Understand the different assumptions underpinning the two models described above

Part A. Vitamin E dataset

Reload the vitamin E dataset and generate a binary variable to indicate high vitamin E levels (above 12mg/dl).

As a reminder, in the previous session we found that the estimated conditional odds ratio for the association between the binary vitamin E variable and a person's cancer status (case or control) was 0.76, indicative of a protective effect of high vitamin E levels.

1 We will first demonstrate an invalid method of analysis which might seem superficially attractive.

Fit a logistic regression model with case as the dependent variable, and the binary vitamin E indicator variable and the set variable as categorical explanatory variables.

```
logistic case h_vitE i.set
```

Confirm that this does not give the same (correct) estimate of the odds ratio as we calculated in the previous session (0.76).

This situation contrasts with the linear regression we used in the previous practical session, where adjusting for the matched set produced a valid estimate of the mean difference in vitamin E levels between cases and controls. This type of analysis is not valid when the dependent variable is binary.

2 Fit a conditional logistic regression analysis as below.

```
clogit case h_vitE, group(set)
```

3 We could also use the original, continuous vitamin E variable as an explanatory variable. However, as the dependent variable is still binary, we must again use conditional logistic regression.

```
clogit case vitE, group(set)
```

4 Suppose now that we wish to explore whether the effect of dichotomised vitamin E on the conditional odds of cancer varies with observation time (time from blood collection to diagnosis of cancer in the case). Carry out an appropriate analysis to explore this.

Discuss: What do you conclude from this analysis?

Part B. Infertility dataset

The data for Part B come from a matched case-control study of the association between infertility in women and previous spontaneous or induced abortion. This is an example of a matched case-control study with two controls per case. The dataset is called infertility_data.dta. The main research question of interest is whether previous abortions (induced or spontaneous) affect risk of infertility.

Further details of the study can be found in the original paper by Tzonou et al, which is available online: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1059707`

Each case (a woman diagnosed as infertile) was matched to two controls on age, parity and education level. The variables shown in the data are as follows:

Variable	Description
`case`	case/control indicator (0=control, 1=case)
`stratum`	index for matched set, numbered 1–271
`spontaneous`	Number of prior spontaneous abortions [0 (zero), 1 (one), 2 (2 or more)]
`induced`	Number of prior induced abortions [0 (zero), 1 (one), 2 (2 or more)]
`age`	age in years
`parity`	number of children
`education`	years of education [1 (0-5 years), 2 (6-11 years), 3 (12+ years)]

5 Open the dataset and familiarise yourself with the data and its structure.

(a) Do rows contain information on individual people or matched sets?

(b) How many cases, controls, and matched sets are included in the dataset?

(c) Are all of the sets the same size?

6 We will examine the effect of spontaneous abortion on infertility using conditional logistic regression. Use `clogit` to perform a matched analysis using the variable `spontaneous` as a categorical explanatory variable.

What do you conclude?

7 As discussed in the lecture, an alternative analysis for a matched case-control study with 'well-defined' matching variables is to perform an unmatched analysis with regression adjustment for the matching variables. Fit the following three models using `glm` or `logit`.

(a) A model that only includes the matching variables (age, parity and education level),

(b) A model that only includes the exposures of interest, the categorised count of spontaneous abortions, as a categorical explanatory variable,

(c) A model that includes the exposure of interest and the matching variables.

Discuss: Compare the estimated coefficients for the matching variables in the models in parts (a) and (c). What do you conclude?

Also compare the estimated coefficients relating to the categorised count of spontaneous abortions in (b) and (c) with those from the matched analysis using conditional logistic regression. Which approaches give valid estimates of the odds ratios?

Session 15

Multinomial Logistic Regression

15.1 Aims and Objectives

The aim of this session is to introduce models appropriate for nominal outcomes. That is, categorical outcomes with no ordering.

By the end of this session you should understand how multinomial logistic regression can be used to model the dependency of a nominal outcome on covariates.

15.2 Nominal outcomes and the multinomial distribution

In this session we consider data where the outcome Y takes values from a finite set $\{1, .., J\}$. We are interested in relating how the probability of Y taking each of the J values varies according to a collection of covariates $x_1, ., .x_p$.

A random variable which takes values from a finite set (e.g. Y) is said to follow a *multinomial* distribution (with $n = 1$). The multinomial can be seen as an extension of the binomial distribution to the setting where there are more than two possible outcomes. Unfortunately the multinomial does not belong to the exponential family. When the covariates are all categorical, it is in fact possible to set up an appropriate GLM using the Poisson distribution, which is based on modelling the number of individuals who fall into different cells of a contingency table. For more on this, see Chapter 9 in the book by Dobson, for example.

In practice, the covariates often include continuous variables, precluding the use of a GLM. However, we will see that the logistic regression model can be readily extended to accommodate the outcome Y taking more than two values, and as with a GLM, the parameters of this model can be estimated via maximum likelihood.

15.3 Example - consumption of alcohol

To illustrate multinomial logistic regression, we again consider data from the 2003-2004 US National Health and Nutrition Examinations Survey (NHANES). As before the analyses shown in this and subsequent sessions are intended to be purely illustrative. In this session we will examine how a categorised version of the alcohol variable is related to the other variables. Specifically, we will 'cut' the alcohol variable according to whether individuals reported one drink per day (1), between 2 and 5 (inclusive) drinks per day (2), and more than 5 drinks per day (3).

Of course this categorised alcohol variable has an ordering. In the next session we will examine how models can be fitted to the same variable which attempt to exploit this ordering, but in this session we treat the variable as unordered, or nominal. The distribution of this categorised version of the variable is as follows.

```
. tab alccat

            alccat |      Freq.     Percent        Cum.
-------------------+-----------------------------------
   1 drink per day |        898       35.24       35.24
2-5 drinks per day |      1,359       53.34       88.58
 >5 drinks per day |        291       11.42      100.00
-------------------+-----------------------------------
             Total |      2,548      100.00
```

15.4 The multinomial logistic regression model

To define the multinomial logistic regression model we must pick a reference level of the outcome Y. For the purposes of this section, we assume Y takes values from 1 to J, and we shall choose 1 as the reference level. The multinomial logistic regression assumes that, for $j = 2, .., J$

$$\log\left(\frac{P(Y=j|x_1,..,x_p)}{P(Y=1|x_1,..,x_p)}\right) = \beta_{j0} + \beta_{j1}x_1 + ... + \beta_{jp}x_p$$

Thus we are modelling the (log of the) ratio of the probability of Y taking value j to the probability that it takes the reference level. The intercept for the jth level corresponds to the log odds of outcome j, given that the outcome is either j or 1. Similarly, the coefficients of the covariates for the jth level can be interpreted as in a standard logistic regression, but conditional on the outcome taking value either j or 1.

Note that the $J-1$ equations determine the log odds for any pair of outcome values. For example, for $1 < a, b \leq J$, we have

$$\begin{aligned}
\log\left(\frac{P(Y=b|x_1,..,x_p)}{P(Y=a|x_1,..,x_p)}\right) &= \log\left(\frac{P(Y=b|x_1,..,x_p)/P(Y=1|x_1,..,x_p)}{P(Y=a|x_1,..,x_p)/P(Y=1|x_1,..,x_p)}\right) \\
&= \log\left(\frac{P(Y=b|x_1,..,x_p)}{P(Y=1|x_1,..,x_p)}\right) - \log\left(\frac{P(Y=a|x_1,..,x_p)}{P(Y=1|x_1,..,x_p)}\right) \\
&= \beta_{b0} + \beta_{b1}x_1 + ... + \beta_{bp}x_p - (\beta_{a0} + \beta_{a1}x_1 + ... + \beta_{ap}x_p) \\
&= \beta_{b0} - \beta_{a0} + (\beta_{b1} - \beta_{a1})x_1 + ... + (\beta_{bp} - \beta_{ap})x_p \qquad (15.1)
\end{aligned}$$

Using the fact that $P(Y = 1|x_1, .., x_p) = 1 - \sum_{j=2}^{J} P(Y = j|x_1, .., x_p)$, we can derive the following expressions for the probabilities that Y takes each value, for given values of the covariates.

$$\begin{aligned} P(Y = 1|x_1, .., x_p) &= \frac{1}{1 + \sum_{j=2}^{J} e^{\beta_{j0} + \beta_{j1}x_1 + ... + \beta_{jp}x_p}} \\ P(Y = 2|x_1, .., x_p) &= \frac{e^{\beta_{20} + \beta_{21}x_1 + ... + \beta_{2p}x_p}}{1 + \sum_{j=2}^{J} e^{\beta_{j0} + \beta_{j1}x_1 + ... + \beta_{jp}x_p}} \\ &... \\ P(Y = J|x_1, .., x_p) &= \frac{e^{\beta_{J0} + \beta_{J1}x_1 + ... + \beta_{Jp}x_p}}{1 + \sum_{j=2}^{J} e^{\beta_{j0} + \beta_{j1}x_1 + ... + \beta_{jp}x_p}} \end{aligned}$$

15.4.1 Estimation

To estimate the parameters of the multinomial logistic a first thought is to fit $J-1$ separate logistic regressions. While this is possible, it is actually (generally) more efficient to estimate the parameters simultaneously. To help understand the reason for this, suppose we have a 3-level outcome Y, and we were to fit logistic regressions for level 2 vs level 1, and for level 3 vs level 1. Using the equations (15.1), the estimates from these two models would give us estimates for the coefficients for level 3 vs level 2. However, these estimates in general will not coincide exactly with the estimates we would obtain from a logistic model for level 3 vs level 2. Fitting the models simultaneously exploits the fact that the estimates should respect the relationship shown in equation (15.1), and this improves efficiency.

The model parameters are estimated by maximum likelihood. The `mlogit` command in Stata can be used to do this. The syntax is largely the same as for `logistic`. An important difference is the `baseoutcome` option, which enables specification of the reference level. The default is to choose the most frequent outcome level as the baseline.

15.5 Gender and alcohol consumption

To illustrate, we fit a multinomial logistic model to the categorised alcohol variable in the NHANES datasets, and include gender as covariate (see overleaf).

We first note that because the 2-5 drinks per day category occurs most frequently, Stata has chosen this level as the reference level. The estimated constant for the 1 drink per day category corresponds to the log odds for having 1 drink per day vs 2-5 for men.

The coefficent for females of 0.599 is the estimated difference in these log odds between women and men. Thus, women are more likely to have 1 drink per day as opposed to 2-5 drinks per day than men. Specifically, the odds ratio for females versus males for having 1 drink per day as opposed to 2-5 (amongst those who drink 5 or less drinks per day) is $\exp(0.599) = 1.82$. This effect is highly statistically significant.

```
. mlogit alccat i.gender

Multinomial logistic regression                     Number of obs   =       2548
                                                    LR chi2(2)      =     189.97
                                                    Prob > chi2     =     0.0000
Log likelihood = -2327.1432                         Pseudo R2       =     0.0392

------------------------------------------------------------------------------
             alccat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------
1_drink_per_day     |
             gender |
             Female |   .5992096   .0870421     6.88   0.000     .4286103    .769809
              _cons |  -.7195339   .0630546   -11.41   0.000    -.8431186   -.5959492
--------------------+---------------------------------------------------------
2_5_drinks_per_day  |  (base outcome)
--------------------+---------------------------------------------------------
_5_drinks_per_day   |
             gender |
             Female |  -1.517809   .1755555    -8.65   0.000    -1.861892   -1.173727
              _cons |  -1.126337   .0729256   -15.45   0.000    -1.269268   -.9834052
------------------------------------------------------------------------------
```

Turning to the 5 drinks per day estimates, we see that being female is associated with a reduced odds of having 5 drinks per day (as opposed to 2-5 drinks per day) compared to men. Together, these results suggest that women are less likely than men to be drink heavily rather than moderately, and more likely than men to drink lightly rather than moderately.

Adding the `rrr` option gives us exponentiated coefficients. The rrr stands for relative-risk ratio. As explained in the Stata manual, this is because the exponentiated coefficients correspond to ratios of relative risks. Stata's manual states that the coefficients are not (log) odds ratios. This is indeed correct if one does not add the 'conditional on the outcome being either j or 1' when giving interpretation. However, in keeping with textbooks, we shall refer to the exponentiated coefficients corresponding to covariates as odds ratios, always bearing in mind that they are defined conditional on the outcome being either j or 1.

To re-assure ourselves regarding this, we fit a logistic regression model, where the outcome is 1 for those individuals who drink on average one drink per day, 0 for those who drink 2-5 drinks per day, and missing for those who drink more than 5 drinks per day, such that they are excluded from the analysis.

```
. gen oneperday=1
. replace oneperday=0 if alccat==2
(1359 real changes made)
. replace oneperday=. if alccat==3
(291 real changes made, 291 to missing)
```

```
. logit oneperday i.gender

Logistic regression                                     Number of obs   =       2257
                                                        LR chi2(1)      =      47.97
                                                        Prob > chi2     =     0.0000
Log likelihood = -1493.0353                             Pseudo R2       =     0.0158

------------------------------------------------------------------------------
   oneperday |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
     Female  |   .5992096   .0870421     6.88   0.000     .4286103     .769809
       _cons |  -.7195339   .0630546   -11.41   0.000    -.8431186   -.5959492
------------------------------------------------------------------------------
```

As expected, fitting the logistic model to those individuals who drink either one or 2-5 drinks per day, we recover the same parameter estimates (for 1 drink vs 2-5 drinks) as using `mlogit`. This occurs here because we only have a single binary covariate. More generally we would obtain slightly different parameter estimates.

With a single covariate which is binary, arguably the use of multinomial logistic regression is unnecessary. Instead we could have presented a cross-tabulation of the two variables, and used appropriate estimates and tests for the parameters of interest. However, more generally such an approach will not be feasible. In particular, this will be the case when we have one or more continuous covariates.

15.6 Age and alcohol consumption

We now fit a model for the categorised alcohol variable with age entering linearly.

```
. mlogit alccat ageyrs
Multinomial logistic regression                         Number of obs   =       2548
                                                        LR chi2(2)      =     264.22
                                                        Prob > chi2     =     0.0000
Log likelihood = -2290.0175                             Pseudo R2       =     0.0545

------------------------------------------------------------------------------
             alccat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------
1_drink_per_day     |
             ageyrs |   .0347866   .0025138    13.84   0.000     .0298596    .0397137
              _cons |  -2.103926   .1320333   -15.93   0.000    -2.362707   -1.845146
--------------------+---------------------------------------------------------
2_5_drinks_per_day  |  (base outcome)
--------------------+---------------------------------------------------------
_5_drinks_per_day   |
             ageyrs |  -.0143282    .004093    -3.50   0.000    -.0223502   -.0063061
              _cons |  -.9521576   .1754993    -5.43   0.000     -1.29613   -.6081853
------------------------------------------------------------------------------
```

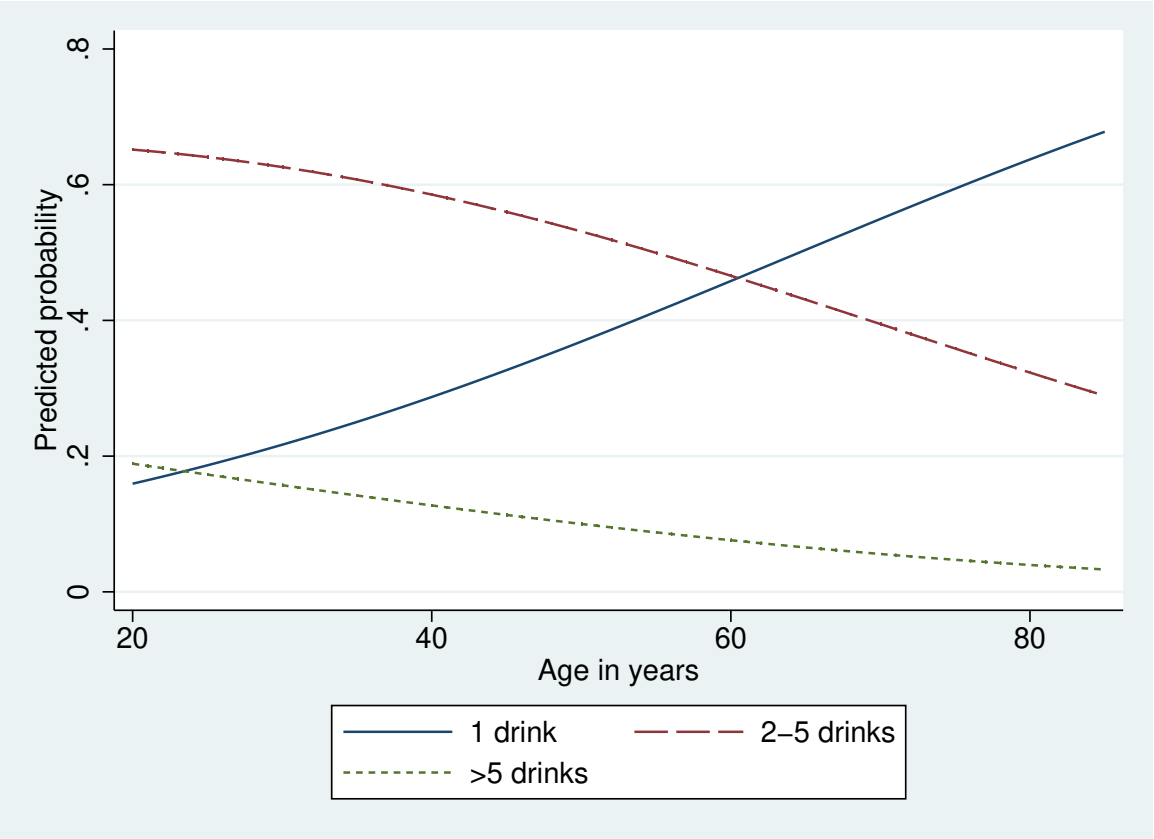


Figure 15.1: Predicted probabilities of alcohol category by age

As with gender, there is strong evidence that age is associated with the probabilities of falling into the three different categories of the alcohol variable. We see that increasing age is associated with increased odds of drinking one drink per day as opposed to 2-5 drinks per day. And second, that increasing age is associated with decreased odds of drinking more than 5 drinks per day as opposed to 2-5 drinks per day. Together these results suggest that the number of drinks being consumed per day (as reported) is highest among the young, and decreases with age.

To visualise the fitted model, we can use the predict command to calculate the predicted probability of being in each of the three categories for each individual and then plot these (15.1). This nicely shows that while the proportion of individuals drinking more than 5 drinks per day (on those days they drink) is low, the proportion decreases steadily with age. The proportion drinking one drink strongly increases with age, and lastly (as must be the case since the probabilities sum to one), the proportion drinking 2-5 drinks decreases strongly with age.

Of course, this interpretation all assumes that entering age linearly (on the model scale) is appropriate. If it is not, these predicted probabilities may systematically differ from the true values (i.e. be biased). One approach to exploring this would be to add age squared as a covariate to the model, and see if it materially improves the fit of the model.

15.7 Model comparison and postestimation

Nested models can, as usual, be compared using (profile) log-likelihood ratio tests. Note that since a covariate is either not included in a model or is included in all $J-1$ equations, to examine whether there is evidence that a covariate is needed in the model, either a joint Wald test (using `test`) or the profile log-likelihood ratio test (of the model including and excluding the covariate in question) should be used. These test the null hypothesis that the coefficients of the covariate in all $J-1$ equations are zero.

After `mlogit`, the predict command can be used to generate variables which contain the predicted probability of each individual falling into each of the J levels of Y, based on their covariate values. The syntax in Stata is as follows.

```
predict pr0 pr1 pr2, pr
```

where as many new variable names need to be specified as there are levels in the outcome.

15.8 Summary

We have seen how logistic regression can be relatively easily extended to model categorical outcomes with more than two levels. This brings with it an increase in the complexity when interpreting model parameters: parameters for a given level of the outcome measure how the probability of that level occuring relative to the reference level change with the covariates.

To read about work on goodness of fit for multinomial logistic regression, see Fagerland, Hosmer and Bofin, 'Multinomial goodness-of-fit tests for logistic regression models', Statistics in Medicine, 2008; 27:4238-4253 and Fagerland and Hosmer 'A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models', The Stata Journal, 2012; 12:447-453.

15.9 Practical 15

Dataset required: `nhanesglm.dta`

Introduction

In this practical we will use multinomial logistic regression techniques to analyse the NHANES alcohol data introduced in the lecture. The outcome variable is a categorical measure of daily alcohol consumption, in three levels.

The dataset is called `nhanesglm.dta` and contains six variables as below.

Variable	Description
`gender`	1 = male, 2 = female
`ageyrs`	Age in (whole number of) years
`bmi`	Body Mass Index (kg/m^2)
`sbp`	Systolic blood pressure (mmHg)
`ALQ130`	Average number of alcoholic drinks consumed on days when alcohol is consumed
`alccat`	Categorised `ALQ130`: 1 = 1 drink per day 2 = 2-5 drinks per day 3 = 6+ drinks per day

Aims

- Compare the estimates from multinomial logistic regression to those from standard logistic regression.
- Understand how to calculate estimated probabilities from multinomial regression estimates.
- Learn how to use the margins command in Stata to obtain the estimated probabilities.

Analysis

1 Tabulate the categorised alcohol consumption variable (`alccat`) by gender (as in the lecture slides).

2 Fit a (standard) logistic regression model with categorised alcohol consumption as the dependent variable and gender as the only covariate, **omitting the highest drinking category**.

Do this by first using the following code to create a binary outcome that is 1 when `alccat` is equal to 1, 0 when `alccat` is equal to 2 and missing when `alccat` is equal to 3 and then fitting the model.

```
gen alc12 = alccat if alccat < 2.5

replace alc12 = 0 if alc12 == 2

logit alc12 i.gender, nolog
```

Also fit a second analogous logistic regression model, this time omitting the lowest drinking category.

3 Fit a multinomial logistic model to the categorised alcohol consumption variable with gender as the only covariate.

Discuss: Compare the parameter estimates from the logistic regression models with those from the multinomial logistic regression model and check that you understand the interpretation of each of the estimated coefficients for the multinomial logistic regression model.

4 Fit a multinomial logistic model to the categorised alcohol consumption variable with gender and age as covariates. Give an exact interpretation of each estimated coefficient.

Discuss: Compare your interpretations with one or more of your colleagues (in your Breakout Room if online).

5 Use the estimated coefficients from the model to predict the probabilities that a 50 year old female is in each of the three alcohol consumption categories.

Use the `predict` command to generate the fitted probabilities of being in each alcohol consumption category for each participant. You will need to name three variables to be the predicted probabilities.

```
predict pr1 pr2 pr3
```

Check that your predictions for a 50 year old female match those predicted here.

Plot line graphs of these fitted probabilities against age, separately in males and females. Hint: in order to see the lines sort the data by age before plotting the probabilities.

6 The `margins` command can also be used to display fitted probabilities. For example, use the following commands to produce and display the fitted probabilities for the 1 drink per day category by gender, at ages 20, 30, 40, ... and 80 after refitting the multinomial logistic regression model with age and gender as covariates.

```
margins, at(ageyrs=(20(10)80)) over(gender) predict(outcome(1))

marginsplot
```

Discuss: With one or more colleagues discuss the patterns exhibited by the various plots. Write a few sentences that describe how the distribution of the categorised alcohol consumption variable varies by age and gender. Post these in the zoom chat if online.

7 Add an interaction between age and gender to the model in part 4 and use an appropriate test to see whether there is evidence that this improves the fit of the model.

8 (optional) Calculate fitted probabilities for the model in part 7 and use plots analogous to those used earlier to explore whether the predictions are materially altered by the inclusion of the interaction.

Session 16

Ordinal Logistic Regression

16.1 Aims and objectives

The aims of this session are to introduce the ordinal logistic regression model for modelling ordinal outcomes. By the end of the session you should

- understand the formulation of the ordinal logistic model and interpretation of its parameters,
- appreciate the model's alternative derivation based on a latent continuous variable, and
- be familiar with different approaches to assessing the proportional odds assumption.

16.2 Motivation

In the previous session we saw how logistic regression could be extended to model a dependent/outcome variable which takes more than two values. Often such categorical outcomes have an ordering. The categorised alcohol variable we analysed in the previous session for example, has a natural ordering, since it is derived from the number of alcoholic drinks reported by study participants. When the outcome variable has an ordering, often this can be used to specify a model with fewer parameters, and easier interpretation, than the model which does not utilise the ordering.

16.3 The ordinal logistic model

Recall that if Y is a binary 1/0 variable, a logistic regression model relates

$$\text{logit}(P(Y=1)) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right)$$

to a vector of covariates $x_1, .., x_p$.

Now consider an ordinal outcome variable Y, which takes values $1, .., J$. Different modelling approaches for ordinal outcomes are based on different comparisons of the probabilities that Y takes the different possible values. Following the terminology used by Stata, in an ordinal logistic regression model, we consider, for $j = 2, .., J$

$$\begin{aligned}\text{logit}(P(Y \geq j)) &= \log\left(\frac{P(Y \geq j)}{1 - P(Y \geq j)}\right) \\ &= \log\left(\frac{P(Y \geq j)}{P(Y < j)}\right)\end{aligned}$$

There is of course no need to consider $P(Y \geq 1)$, since this equals one. For a given value of j, a model for $\text{logit}(P(Y \geq j))$ can be thought of as a model for the collapsed binary variable which takes value one if $Y \geq j$ and value zero if $Y < j$. We could, for example, for each $j = 2, .., J$ assume that

$$\text{logit}(P(Y \geq j)) = \beta_{j0} + \beta_{j1}x_1 + ... + \beta_{jp}x_p$$

and estimate these parameters by fitting $J - 1$ separate logistic regression models. However, it may be possible to reduce the number of parameters used in this set of logistic regression models by fully exploiting the fact that the levels of Y are ordered.

For increasing values of j, $P(Y \geq j)$ will decrease, and so we must allow a separate intercept parameter β_{j0} for each value of j. However, given that the outcome Y is ordered, we might assume that the covariate effects are homogeneous, in some sense.

In (what we are calling) ordinal logistic regression, the covariate effect parameters $\beta_{j1}, .., \beta_{jp}$ are assumed to be equal for all j. That is, the model assumes

$$\text{logit}(P(Y \geq j)) = \beta_{j0} + \beta_1 x_1 + ... + \beta_p x_p \tag{16.1}$$

where only the intercept is allowed to vary across different values of j.

Consider two individuals, one with covariate vector $\mathbf{x}_1 = (x_{11}, .., x_{1p})$ and another with covariate vector $\mathbf{x}_0 = (x_{01}, .., x_{0p})$. Then the log odds ratio of $Y \geq j$ comparing the first to the second individual is equal to

$$\begin{aligned} \log\left(\frac{P(Y \geq j|\mathbf{x}_1)/P(Y < j|\mathbf{x}_1)}{P(Y \geq j|\mathbf{x}_0)/P(Y < j|\mathbf{x}_0)}\right) &= \text{logit}(P(Y \geq j|\mathbf{x}_1)) - \text{logit}(P(Y \geq j|\mathbf{x}_0)) \\ &= \beta_1(x_{11} - x_{01}) + ... + \beta_p(x_{1p} - x_{0p}) \end{aligned}$$

Thus we see that the odds ratio comparing these two individuals is the same irrespective of the value of j. Because of this, ordinal logistic regression is often referred to as the *proportional odds model*.

For a model with p covariates, recall that the multinomial logistic model contains $(J - 1) \times (p + 1)$ parameters. In contrast, the ordinal logistic model contains only $J - 1$ intercept parameters plus p coefficients corresponding to the covariates. Consequently, ordinal logistic regression is a more parsimonious modelling approach when the outcome is ordered, provided that the model assumptions hold. In particular, interpretation of covariate effects is considerably simpler for the ordinal logistic model, since there is only a single coefficient for each covariate. However, the proportional odds assumption is a strong one and we should be careful to check that there is no evidence that it doesn't hold in any given setting.

16.4 Example: gender and alcohol consumption

Returning to the NHANES data, we fit an ordinal logistic regression model for the categorised alcohol variable, with gender as covariate:

```
. ologit alccat i.gender

Ordered logistic regression                       Number of obs   =       2548
                                                  LR chi2(1)      =     156.12
                                                  Prob > chi2     =     0.0000
Log likelihood = -2344.0638                       Pseudo R2       =     0.0322

------------------------------------------------------------------------------
      alccat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
     Female  |   -.979379   .0797411   -12.28   0.000    -1.135669   -.8230893
-------------+----------------------------------------------------------------
       /cut1 |  -1.101782   .0597772                     -1.218943   -.9846213
       /cut2 |   1.679587   .0681337                      1.546047    1.813126
------------------------------------------------------------------------------
```

The estimated coefficient for gender is the log odds ratio, i.e. the log of the ratio of odds for women compared to men. Remembering that the ordinal logistic regression estimates $\text{logit}(P(Y \geq j))$, this means that the log odds for $Y \geq 2$ to $Y = 1$ is 0.979 lower for women compared to men. That is, women are at lower odds of being in the higher alcohol categories 2 and 3 compared to 1. Under the ordinal logistic model, this coefficient is also the log odds ratio for being in category 3 versus categories 1 or 2.

From the output we also see estimates for `/cut1` and `/cut2`. These are the estimates of $-\beta_{20}$ and $-\beta_{30}$ (notice the minus signs here), the intercept parameters in the equations for $\text{logit}(P(Y \geq 2))$ and $\text{logit}(P(Y \geq 3))$. Thus, for a male, the estimated log odds of $Y \geq 2$ are 1.102, and the estimated log odds of $Y \geq 3$ are -1.680. Thus we have that, for a male, $P(Y \geq 2) = \exp(1.102)/(1 + \exp(1.102)) = 0.75$.

16.5 An alternative formulation of ordinal (and standard) logistic regression

It is not immediately obvious why in Stata's output from `ologit` we receive estimates of minus the intercept parameters, and indeed why these are labelled 'cut' points. The explanation is due to the fact that Stata's notation is based on an alternative, equivalent, expression of the ordinal logistic regression model, which in some cases may be useful conceptually.

It is often argued (e.g. for psychometric measures) that an ordinal variable arises as a categorised and crude measure of some underlying continuous latent quantity. Assume that, for a subject with covariate values $x_1, .., x_p$, this latent quantity is equal to

$$\tilde{Y} = \beta_1 x_1 + ... + \beta_p x_p + U$$

where U is a random variable which follows the *logistic* distribution with mean zero and scale parameter one. The logistic distribution is a real valued continuous probability distribution. For our purposes it suffices to know that for U following the logistic distribution with mean zero and scale parameter one

$$P(U \leq u) = \frac{\exp(u)}{1 + \exp(u)}$$

We then assume that Y takes value j when $\kappa_{j-1} < \tilde{Y} \leq \kappa_j$, with $\kappa_0 = -\infty$ and $\kappa_J = +\infty$. The κ parameters are thus the 'cut-points' used to categorise the continuous latent $\tilde{Y}$ to form the observed ordinal Y. Thus we have

$$P(Y = j) = P(\kappa_{j-1} < \tilde{Y} \leq \kappa_j) = P(\kappa_{j-1} < \beta_1 x_1 + ... + \beta_p x_p + U \leq \kappa_j)$$

Some algebra (see Practical) shows that this model implies an ordinal logistic regression model as given in equation (16.1), with the log odds ratio parameters $\beta_1, .., \beta_p$ identical, and $\beta_{j0} = -\kappa_{j-1}$.

It is important to note that although the ordinal logistic regression can be derived in this way (assuming the existence of the latent quantity U), this does not mean the model is only appropriate when we think such a latent continuous quantity exists: it merely shows that ordinal logistic regression would be the correct model in settings where the ordinal outcome arises as a categorised version of the latent quantity $\tilde{Y}$ as described. Indeed, for the categorised alcohol variable, such an interpretation would be inappropriate because we know how the ordinal variable arises; we have categorised it from a count variable.

Lastly, we note that the preceding derivation of the ordinal logistic model carries through in the case that $J = 2$, meaning that standard logistic regression can be considered as arising from the categorising of a latent continuous quantity.

16.6 Model comparison and goodness of fit

Nested ordinal logistic regression models can be compared, as usual, by performing likelihood ratio tests. Similarly, we can include interaction terms and non-linear covariate terms to assess whether the independent variables have been included appropriately in the linear predictor.

For a categorical outcome, the ordinal logistic model contains fewer parameters than the corresponding multinomial logistic model. This is achieved by assuming the covariate effects on the log odds of $Y \geq j$ are the same irrespective of the value of j. This is often known as the proportional odds assumption. If it does not hold, but we use the ordinal logistic model, we will obtain invalid inferences.

Unfortunately the ordinal logistic is not nested in the multinomial logistic, precluding testing the assumption by formally comparing the fits of the two models. However, there are a number of alternative less formal approaches that we can use.

The multinomial model contains $p(J-2)$ more parameters than the ordinal model, and so an informal assessment of the proportional odds assumption can be obtained by computing twice the difference in log-likelihoods. If this is large, it suggests that the proportional odds assumption may be violated. One could also compare the value to a χ^2 on $p(J-2)$ degrees of freedom, the asymptotic distribution of this test statistic were the two models in fact nested, although bearing in mind that this test is not formally justified.

The output from `glm` (which we are not using here!) reports AIC and BIC values. These are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC): measures developed to quantify how well a model fits the data at hand, but with a penalty made for the complexity of the model. For a model with k parameters, the AIC is defined as $2k - 2\ell$ where ℓ denotes the maximized log likelihood of the model. Models

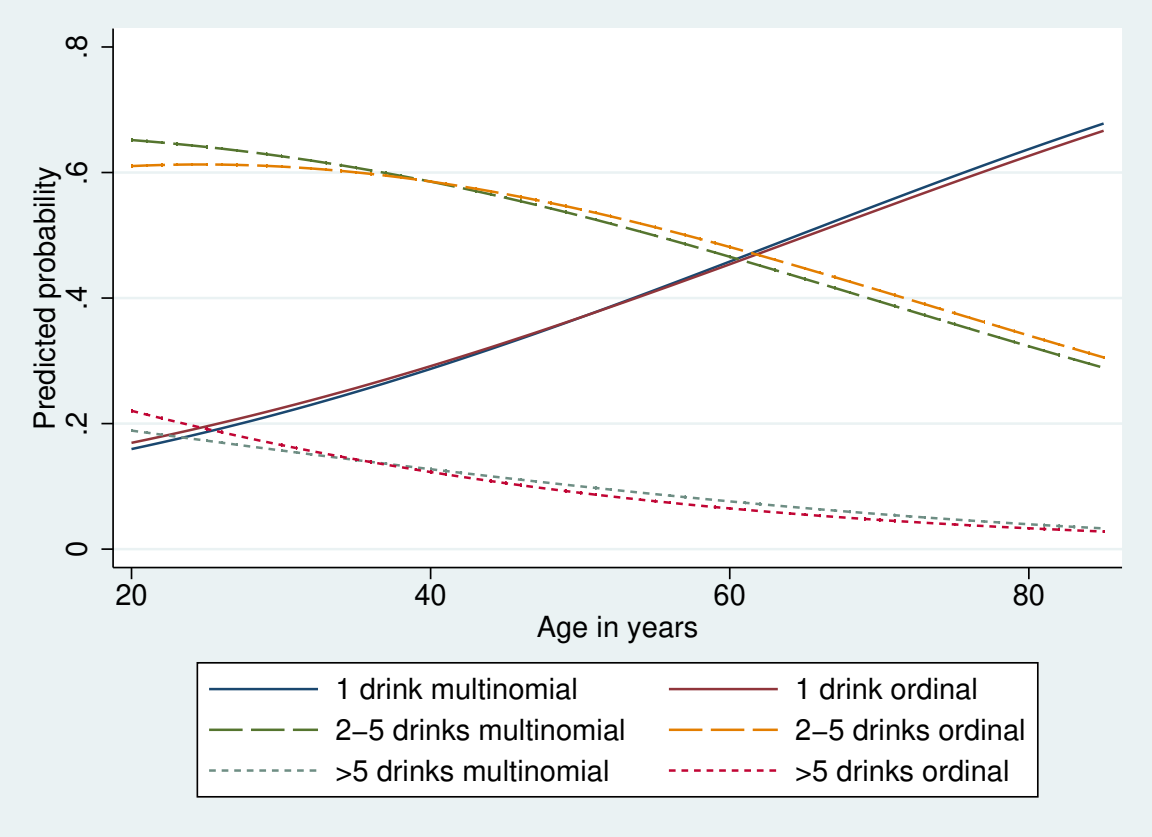


Figure 16.1: Predicted probabilities of alcohol category by age, from both ordinal and multinomial logistic regression models.

are preferred which have lower AIC. An advantage of using AIC for model comparison is that it allows non-nested models to be compared, although the comparison does not produce a statistical test (and hence no p-value).

After fitting a model using `mlogit` or `ologit` the `estat ic` command can be used to report the AIC and BIC. For the model fitted earlier with gender as covariate, the AIC is 4694.128 for the `ologit` and 4662.286 for the `mlogit`, indicating that the multinomial logistic is preferable.

Another informal approach to assessing the proportional odds assumption is to fit the separate logistic regression models corresponding to 'cutting' Y at the different possible values. If the covariate effect estimates are similar across the logistic fits, this suggests that the proportional odds assumption is not unreasonable.

Lastly, one can compare predicted values under two competing models (e.g. multinomial versus logistic). If the predicted values (probabilities in the case of categorical outcomes) are similar, then we would ordinarily prefer the simpler model. Figure 16.1 shows the predicted probabilities of the three alcohol categories in the NHANES data, estimated using both ordinal and multinomial logistic regression models, with age as the only covariate. The closeness of the predictions from the two models suggests that the additional complexity of the multinomial model does not here materially improve fit, since predictions from both are very similar.

As has been discussed previously, model selection and assessment of goodness of fit are complex, and arguably subjective, processes. An important consideration when these are conducted are the issues of sample size and power. With large datasets the power to detect differences in fit will be large, even when the differences in fit are negligible from a substantive perspective. Conversely, with small sample sizes models might fit poorly but formal significance tests of goodness of fit might not show this, due to lack of power.

16.7 Summary

We have seen how the ordinal logistic regression model allows the dependence of ordinal outcome variables on covariates to be modelled. In settings where the proportional odds assumption holds, the model gives a far simpler characterisation of the covariates' effects on the outcome, since there is only one parameter per covariate. However, we must be careful to ensure that, at least approximately, such a simplifying assumption holds in the data at hand.

The ordinal logistic (proportional odds) model we have introduced here is by no means the only model which takes ordering of the dependent variable into account. Alternatives are the adjacent-category logit and continuation-ratio logit models: for more information see the book by Agresti (1996).

16.8 Practical 16

Dataset required: `nhanesglm.dta`

Introduction

In this practical we will again use the NHANES data introduced in previous sessions, but here we will analyse a categorised version of systolic blood pressure (SBP) as the outcome variable, using an ordinal logistic regression model.

For the purposes of this session, we will ignore diastolic blood pressure and categorise blood pressure according to SBP as follows:

Category	SBP
Normal	< 120
Prehypertension	$120 - 139$
High (stage 1)	$140 - 159$
High (stage 2)	≥ 160

Given the above classification system, it is of interest to model how the probability of falling into the four categories depends on age and gender.

Aims

- Interpret coefficients from ordinal logistic regression
- Understand the differences between multinomial and ordinal regression
- Know how to assess the fit of an ordinal regression model

Analysis

1 Use the `egen` command (or otherwise) to create a new variable for SBP, categorised as above. Tabulate the new variable to see its distribution.

What is the mean age in each blood pressure group?

2 We will fit an ordinal logistic regression model to the categorised SBP variable with age as the only covariate. But to help the interpretation of the intercept estimate, we will first centre the age variable at 50 years.

```
gen cen_age = ageyrs - 50

ologit sbpcat cen_age
```

Discuss: Carefully describe the interpretation of each of the estimated parameters.

3 (a) From your fitted model, calculate the estimated probability of having SBP in the highest (High (stage 2)) category for an individual aged 70 years.

(b) From your fitted model, calculate the estimated probability of having SBP in the second highest category (High (stage 1)) for an individual aged 70 years.

(c) Use the `margins` command in Stata to obtain these estimates.

4 Use the `predict` command to generate predicted probabilities of subjects being in each of the four SBP categories, and plot the probabilities against age. Hint: in order to see the lines sort the data by age before plotting the probabilities.

5 To assess whether the proportional odds assumption is reasonable here, fit the corresponding `mlogit` model. Compare the log-likelihoods, AIC values, and fitted probabilities from the two models.

6 As a final check of the model fit, we can compare the coefficients to those from a series of logistic regression models with different cut-offs. First create three new variables, to dichotomise the SBP variable in three different ways:

(a) Normal vs All other categories

(b) Normal or Prehypertension vs Hypertension (Stage 1 or 2)

(c) All other categories vs Stage 2 Hypertension

Fit a standard logistic regression model for each of these new variables as the outcome, and centred age as the only covariate. Compare the age coefficient from each model to the equivalent from the ordinal logistic regression model.

Discuss: Considering all of the above, which model do you prefer for these data?

7 Fit an ordinal logistic regression model and perform an appropriate test to assess if there is evidence that gender is associated with SBP category, after adjusting for age. What is the estimated effect of gender?

8 Use one or more of the approaches in parts 5 and 6 to assess whether the ordinal logistic regression model with age and gender provides a good fit to these data.

Discuss: Which model do you now prefer for these data?

9 (optional) Prove that the latent variable model defined in Section 16.5 of the notes implies the ordinal logistic regression model given by equation (16.1).