THE ANALYSIS OF HIERARCHICAL AND OTHER DEPENDENT DATA

Lecture Notes

Linda Sharples, James Carpenter, Matteo Quartagno

(With thanks to Bianca L. De Stavola)

LSHTM MSc in Medical Statistics 2022-23

Contents

0 Revision of linear regression modelling **6**
0.1 Revision of the linear regression model . . . 6
0.1.1 t-test and ANOVA . . . 6
0.1.2 Simple Linear Regression . . . 8
0.1.3 Adding covariates and interactions . . . 11
0.1.4 Non linear effects . . . 12
0.1.5 Residual diagnostics . . . 12

1 Simple treatment of dependent data **14**
1.1 Dependent data . . . 14
1.2 Origins of dependency . . . 17
1.3 Consequences of dependency . . . 18
1.3.1 Marginal and Conditional Models . . . 19
1.4 Aims of the Module . . . 19
1.5 Notation . . . 20
1.6 Aggregation . . . 20
1.7 Disaggregation . . . 22
1.8 Joint modelling . . . 23
1.9 The fixed effects model . . . 23

2 The random intercept model **29**
2.1 The random intercept model . . . 29
2.2 Estimation . . . 30
2.3 Analysis in Stata . . . 31
2.4 Inference . . . 33

3 The random intercept model with covariates **36**
3.1 Extension of the multivariable regression model . . . 36
3.2 Example: birth weight of siblings . . . 37
3.3 Assigning values to the random components . . . 38
3.4 Diagnostics . . . 41
3.5 Covariates at cluster level . . . 43
3.6 Within and between effects . . . 44
3.7 Fixed or random? . . . 46

4 The random coefficient model **47**
4.1 Example: GCSE scores in schools . . . 47
4.2 Model specification . . . 48
4.3 Example (cont'd) . . . 49
4.4 Inference . . . 51
4.5 On the random effects variances . . . 52
4.6 Predictions . . . 53
4.7 Model assessment . . . 53

5 Longitudinal data I **56**
5.1 Introduction . . . 56
5.2 Fixed occasions . . . 56
5.3 Variable occasions . . . 64
5.4 Example: Infant growth data . . . 64
5.5 Predicting trajectories . . . 66

6 Longitudinal data II **68**
6.1 Marginal structures . . . 68
6.2 Matrix notation . . . 70
6.3 General Formulation of the Mixed Model . . . 71
6.4 Alternative specification . . . 71
6.5 Comments . . . 72
6.6 Unbalanced data . . . 73
6.7 Crossed effects . . . 73

7 Longitudinal data III **74**
7.1 Level 1 heterogeneity . . . 74
7.2 Level 2 heterogeneity . . . 75
7.3 Strategy of analysis . . . 76
7.3.1 Selection steps for a linear mixed model (see Verbeke and Molenberghs, 2000) . . . 76
7.4 More on REML . . . 77

8 Generalized estimating equations **79**
8.1 Introduction . . . 79
8.2 Notation . . . 79
8.3 Links to GLMs . . . 80
8.4 Independent estimating equations (IEEs) . . . 80
8.5 Working correlation matrices . . . 80
8.6 GEEs in Stata: `xtgee` . . . 81
8.7 Example . . . 82
8.8 Comments . . . 83

9 Further issues **85**
9.1 Three or more level data . . . 85
9.2 Binary dependent variables . . . 87
9.3 Designing multilevel/clustered studies . . . 88
9.4 Summary . . . 89

10 Revision **91**

11 Missing data I **92**

12 Missing data II **93**

Outline

Each session in this module starts with a lecture, followed by a practical based on the material presented. There will be a mixture of face-to-face lectures and practical sessions, recorded lectures accessed via moodle and self guided practical sessions with facilitators to answer questions.
There is a pre-course session to revise linear regression.
There is a face-to-face guest lecture in session 13 which accounts for 10% of the assignment mark and practical 14 is an optional Q&A session.

Session	**Date**		**Title**	**Lecturers**
0	**Pre-course**		Revision of linear regression modelling Access lecture and practical via moodle	L Sharples
1	**20 Feb**	am	Introduction and simple treatment of dependent data	L Sharples
2		pm	The random intercept model	J Carpenter
-	**21 Feb**	am	Study time - no lecture	-
3		pm	The random intercept model with covariates	L Sharples
4	**27 Feb**	am	The random coefficients model	J Carpenter
4 cont		pm	Hierarchical models overview	L Sharples
5	**28 Feb**	am	Longitudinal data I	L Sharples
6		pm	Longitudinal data II	J Carpenter
7	**6 Mar**	am	Longitudinal data III	L Sharples
8		pm	Generalized Estimating Equations	J Carpenter
9	**7 Mar**	am	Further issues and summary	L Sharples
10		pm	Revision and Assignment	L Sharples
11	**13 Mar**	am	Missing Data I	M Quartagno
12		pm	Missing Data II	M Quartagno
13	**14 Mar**	am	Guest lecture and discussion	Halima Twabi
14		pm	Optional Q & A session	L Sharples & J Carpenter
	21 Mar	am/pm	Study time - assignment preparation	
	22 Mar	5pm	**Assignment deadline**	

References on which the course is based

Rabe-Hesketh, S. and Skrondal, A. (2012) *Multilevel and Longitudinal Modeling Using Stata, 3rd Edition.* Stata Press.

Snijders, T. and Bosker, R. (1999) *Multilevel Analysis* SAGE Publications Ltd.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* Springer Verlag.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2011) *Applied Longitudinal Analysis. 2nd edition.* John Wiley and Sons, New York.

For a simple introductory text

Kreft, I. and de Leeuw, J. (1998) *Introducing multilevel modelling.* London: SAGE.

Other references

Diggle, P.J.,Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002) *Analysis of Longitudinal Data, Second Edition* Oxford University Press.

Dwyer, J.H., Feinleib, M., Lippert, P. and Hoffmeister, H. eds (1990) *Statistical Methods for Longitudinal Studies of Health.* Oxford University Press.

Goldstein, H. (2011) *Multilevel Statistical Models, Fourth Edition.* Arnold, London.

Jones, B. and Kenward, M.G. (2003) *The Design and Analysis of Cross-Over Trials. Second Edition.* CRC/Chapman & Hall.

Longford, N.T. (1993) *Random Coefficient Models.* Oxford University Press.

Molenberghs, G. and Kenward, M.G. (2007) *Missing data in Clinical Studies.* Wiley.

Chapter 0

Revision of linear regression modelling

In this pre-course session, we will revise linear regression modelling, with emphasis on analysis of variance. In particular, for the module it is important that you have a clear idea of the difference between conditional and marginal models.

0.1 Revision of the linear regression model

A brief revision of the main features and assumptions of the linear regression model is given here using data from a survey of 500 mothers who had singleton births in a large London hospital. The survey included information on the age of the mothers and sex, birth weight and gestational age of their babies. In this session we shall find whether boys and girls differ in terms of mean birth weight and quantify the effect of gestational period on birth weight.

0.1.1 t-test and ANOVA

Let (Y_i, X_i) be the dependent variable and explanatory variable of interest, with Y_i representing birth weight and X_i the sex of baby i, $i = 1, \ldots, N = 500$.

We can formally compare mean birth weight in the two sexes with a t-test (in the dataset `sex` is coded 1 for boys and 2 for girls), assuming that the two groups have the same **population variance** σ^2:

```
. use births,clear
. ttest bweight, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.
---------+--------------------------------------------------
       1 |     264    3229.902    38.99802    633.6428
       2 |     236    3032.831    40.80225     626.816
---------+--------------------------------------------------
combined |     500    3136.884     28.5077    637.4515
---------+--------------------------------------------------
    diff |             197.071    56.47605
------------------------------------------------------------
<OMITTED OUTPUT>
```

The sample mean birth weights in boys and girls, $\hat{\mu_1}$ and $\hat{\mu_2}$ are, respectively, $3230g$ and $3033g$ and

$N = 500$. The t-test of whether the true means μ_1 and μ_2 are the same is

$$t = \frac{\hat{\mu_1} - \hat{\mu_2}}{\hat{SE}(\hat{\mu_1} - \hat{\mu_2})} = \frac{197.071}{56.476} = 3.4895 \tag{1}$$

where SE is the standard error of the difference in estimated means. The relevant part of the Stata output is:

```
diff = mean(1) - mean(2)                          t =   3.4895
Ho: diff = 0                     degrees of freedom =      498
Ha: diff != 0
Pr(|T| > |t|) = 0.0005
```

Under the null hypothesis of no difference between the means, the statistic has a t distribution with $df = N - J = 500 - 2 = 498$ degrees of freedom, where $J = 2$ is the number of estimated means (i.e. the number of groups being compared).

The model underlying this t-test is also called the one-way analysis of variance (ANOVA) model. Analysis of variance involves partitioning the total sum of squares (TSS) (i.e. the sum of squared deviations of the Y_i from their overall mean) into the model sum of squares (MSS) and the residual sum of squares (RSS):

$$\text{TSS} = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 = \text{ MSS} + \text{ RSS}$$

where

$$\begin{aligned} \text{MSS} &= \sum_{i=1}^{N} \left(\hat{Y}_i - \overline{Y}\right)^2 \\ &= \sum_{i \in 1} \left(\overline{Y}_1 - \overline{Y}\right)^2 + \sum_{i \in 2} \left(\overline{Y}_2 - \overline{Y}\right)^2 \\ &= n_1 (\overline{Y}_1 - \overline{Y})^2 + n_2 (\overline{Y}_2 - \overline{Y})^2 \end{aligned}$$

and

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 \\ &= \sum_{i \in 1} \left(Y_i - \overline{Y}_1\right)^2 + \sum_{i \in 2} \left(Y_i - \overline{Y}_2\right)^2 \end{aligned}$$

The model mean square (MMS) and the mean square error (MSE) can be obtained from the corresponding sums of squares by dividing them by the appropriate degrees of freedom as described in the table below:

Source	Sum of Squares	DF	Mean Square
Model	MSS	J-1	MMS
Residual	RSS	N-J	MSE
Total	TSS	N-1	

J = number of groups; N = Total sample size; DF: degrees of freedom

The MSE is the pooled within-group sample variance, and is an estimate of the residual population variance σ.

The F statistic for the null hypothesis that the population means are the same is then

$$F = \frac{\text{MMS}}{\text{MSE}}$$

Under the null hypothesis this statistic has an F distribution with $(J-1, N-J)$ degrees of freedom. When J=2 the F statistic is the square of the t statistic (assuming equal variances).

We can perform an ANOVA of birth weight (measured in kg) in terms of sex using Stata, either with the command `anova` (or equivalently `oneway`) and obtain $F = 12.18 = t^2 = (3.4895)^2$:

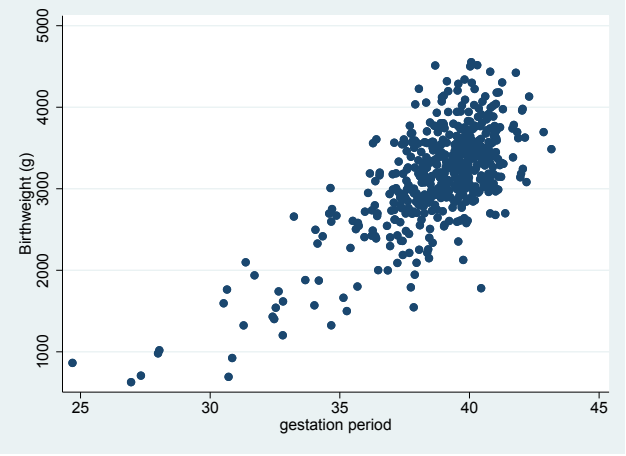
```
. anova bweight sex

          Number of obs =     500     R-squared     =  0.0239
          Root MSE      = 630.431     Adj R-squared =  0.0219

  Source |  Partial SS    df       MS           F     Prob  F
---------+----------------------------------------------------
   Model |  4839398.61     1  4839398.61      12.18     0.0005
     sex |  4839398.61     1  4839398.61      12.18     0.0005
Residual |   197926455   498   397442.68
---------+----------------------------------------------------
   Total |   202765853   499  406344.395
```

The entries for the rows headed `Model` and `sex` are identical because only one explanatory variable is considered here.

0.1.2 Simple Linear Regression

Now we will use linear regression to model birth weight (measured in kg) in terms of a continuous explanatory variable, gestational period (measured in weeks), where Y_i is again birth weight and X_i is gestational period of baby i (Figure 0.1.2).

Figure 1: Birth weight versus gestational period of London babies



A simple linear regression model of Y on X can be written:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $i = 1, \ldots, N$ and the residuals ϵ_i are independent with

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

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

$$Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2).$$

Here β_0 is the intercept and represents the conditional expectation of Y_i when $X_i = 0$

$$E(Y_i \mid X_i = 0) = \beta_0$$

and β_1 is the slope and represents the difference in conditional expectations when X_i increases by one unit, for example from a to $a + 1$:

$$E(Y_i \mid X_i = a + 1) - E(Y_i \mid X_i = a) = (\beta_0 + \beta_1(a + 1)) - (\beta_0 + \beta_1 a) = \beta_1$$

Hence the implicit assumption of this model is that the conditional expectations of every baby fall on a straight line: $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$ (this is the assumption of *linearity*). It implies that residuals have mean zero, conditional on X_i (i.e. $E(\epsilon_i \mid X_i) = 0$) and also that $\text{Cor}(\epsilon_i, X_i) = 0$ (assumption of *'exogeneity'* of the covariate). In addition the model assumes that residuals have constant variance σ (assumption of *'homoscedasticity'*).

As before we can partition the TSS into MSS and RSS, with

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The OLS estimates for β_0 and β_1 are found by minimizing the RSS while again the estimate of σ is found by dividing the RSS by $N - 2$, 2 being the number of estimated parameters.

ML estimation is achieved via the log-likelihood function:

$$\ell(\beta_0, \beta_1, \sigma^2 \mid \mathbf{Y}, \mathbf{X}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_i)^2.$$

where $\mathbf{Y}$ is the $(N \times 1)$ dependent variable vector and $\mathbf{X}$ the $(N \times 1)$ explanatory variable vector for the N babies.

The MLE's of β_0 and β_1 can be obtained from the **score equations**:

$$\begin{aligned} U(\beta_0) &= \ell'(\beta_0) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_i) \\ U(\beta_1) &= \ell'(\beta_1) = \frac{1}{\sigma^2}\sum_{i=1}^{N} X_i(Y_i - \beta_0 - \beta_1 X_i). \end{aligned}$$

Solving these, i.e. setting $U(\hat{\beta}_0) = 0$ and $U(\hat{\beta}_1) = 0$ imply:

$$\begin{aligned} \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} \\ \hat{\beta}_1 &= \frac{\sum X_i Y_i - N\overline{XY}}{\sum X_i^2 - N\overline{X}^2} \end{aligned}$$

They are the same as the OLS estimates. The MLE for σ instead is the RSS divided by N, not $N - 2$.

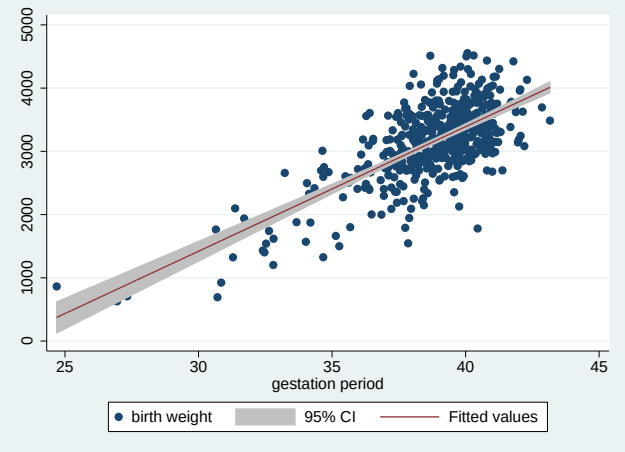
In `Stata` we can fit a linear regression model with:

```
. regress bweight gestwks

  Source |       SS       df       MS          Number of obs =     490
---------+------------------------------       F(  1,   488) =  502.36
   Model |   101603845     1   101603845       Prob > F      =  0.0000
Residual |  98698697.8   488   202251.43       R-squared     =  0.5073
---------+------------------------------       Adj R-squared =  0.5062
   Total |   200302543   489  409616.652       Root MSE      =  449.72

------------------------------------------------------------------------------
bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
--------+------------------------------------------------------------------
gestwks |   196.9726   8.788133    22.41   0.000     179.7054    214.2399
  _cons |   -4489.14   340.8988   -13.17   0.000     -5158.95   -3819.329
------------------------------------------------------------------------------
```

The regression parameter labelled `gstwks` is the estimated slope coefficient, and represents the increase in birth weight predicted by the fitted model for each unit increase in gestational period, in this case one *week*. The other parameter, `_cons`, is the intercept, the average weight at period 0 given by the fitted model. This is an impossible negative value, as it gives the expected birth weight in a baby whose gestation is 0 weeks!

Figure 2: Birth weight and gestational period: fitted regression line with pointwise confidence interval



The ANOVA table, reproduced below for clarity, provides a breakdown of the variability of the data. The Model Mean Square represents the variability in the data that is explained by the fitted line, the Residual Mean Square (also called *Mean Square Error*, MSE) quantifies the remaining variability that is NOT explained by the model. The estimated population residual SD σ is $\sqrt{98698697.8/488} = \sqrt{202251.43} = 449.72$, i.e. around 450g.

Source	Sum of Squares	DF	Mean Square	F
Model	101,603,845	1	101,603,845	502.36
Residual	9,869,869,7.8	488	202,251.43	
Total	200,302,543	489		

The fitted model (a straight line) is plotted together with the pointwise confidence interval in Figure 2 with the command:

```
. twoway (scatter bweight gestwks) (lfitci bweight gestwks)
```

Note how wide the confidence intervals become at the extremes of the range of X values. Indeed at $X = 0$ the confidence interval goes from -5158.95 to -3819.329. Thus, when the data on the explanatory variable do not include the value 0 as in this case, it is best to centre the explanatory variable around its mean value (or some other sensible value) and refit the model:

$$Y_i = \beta_0 + \beta_1(X_i - \hat{\mu}_X) + \epsilon_i$$

In Stata:

```
. su gestwks

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     gestwks |       490    38.72186    2.314167      24.69      43.16

. gen c_gest=gestwks-r(mean)
. regress bweight c_gest
<EDITED OUTPUT>
------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_gest |   196.9726   8.788133    22.41   0.000     179.7054    214.2399
       _cons |   3138.006   20.31645   154.46   0.000     3098.088    3177.925
------------------------------------------------------------------------------
```

The ANOVA table and the estimated slope (and corresponding SE) obtained when fitting this last model have not changed (beside minimal differences in the ANOVA table due to rounding error). The only change is in the estimated intercept that now refers to the expected weight of a baby who was 38.7 weeks at birth and which is now more precisely estimated.

0.1.3 Adding covariates and interactions

Now we consider whether the effect of gestational period on birth weight is in anyway confounded or modified by the sex of the baby. We first add `sex` to the linear model to see whether the coefficient for gestational age `c_gest` changes.

```
. regress bweight c_gest i.sex
<OMITTED OUTPUT>
------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_gest |   196.3436   8.599464    22.83   0.000     179.4469    213.2402
       2.sex |  -189.9113    39.8007    -4.77   0.000    -268.1136    -111.709
       _cons |   3228.698   27.50261   117.40   0.000      3174.66    3282.737
------------------------------------------------------------------------------
```

The estimate of the slope obtained from the simple model indicated a $197g$ increase in birth weight per gestational week. This does not appear to be confounded by `sex` as the sex-adjusted estimate of the slope is $196g$ per week of gestation. The intercepts in boys and girls are however significantly different, with girls being on average $190g$ lighter than boys for any given gestational period.

Now add the interaction between sex and gestational age, then test for modification as follows:

```
. regress bweight i.sex##c.c_gest

------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |  -189.8624   39.82193    -4.77   0.000    -268.1068    -111.618
      c_gest |   203.3581   13.26672    15.33   0.000     177.2909    229.4253
             |
sex#c.c_gest |
          2  |  -12.10667   17.42919    -0.69   0.488    -46.35255     22.1392
             |
       _cons |   3228.461   27.51936   117.32   0.000      3174.39    3282.533
------------------------------------------------------------------------------
```

```
. testparm i.sex#c.cgest
 ( 1)  2.sex#c.cgest = 0
       F(  1,   486) =    0.48
            Prob > F =    0.4876
```

The coefficient `c_gest` is now interpreted as the effect of gestational age on birth weight **for boys**. The interaction term estimates that the girls' growth rate is $12g$ lower than boys, but according to the interaction test, there is no evidence of effect modification (Partial F-test: P=0.49 for the interaction between sex and gestational age). Hence it is likely that the birth weight of boys and girls has the same relation with gestational period.

0.1.4 Non linear effects

It may be unreasonable to assume that birth weight increases linearly with gestational age. To test whether the relationship is quadratic instead of linear we generate a new variable equal to the square of X and add it to the model. Because we are centering gestational age, the model we are considering is:

$$Y_i = \beta_0 + \beta_1(X_i - \hat{\mu}_X) + \beta_2(X_i - \hat{\mu}_X)^2 + \beta_3 \text{ female} + \epsilon_i \tag{2}$$

In Stata:

```
. gen cgestsq=c_gest^2
. regress bweight c_gest cgestsq i.sex

<EDITED OUTPUT>
------------------------------------------------------------
     bweight |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------------
      c_gest |   184.9798   12.20059    15.16   0.000
     cgestsq |  -2.260747   1.723046    -1.31   0.190
       2.sex |  -185.7864   39.89531    -4.66   0.000
       _cons |   3238.811   28.54257   113.47   0.000
------------------------------------------------------------
```

There does not seem to be evidence of departure from linearity of the effect of gestational period.

0.1.5 Residual diagnostics

We can estimate residuals from a fitted model as the differences between observed and predicted values of Y. For the last fitted model these are:

$$\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1(X_i - \hat{\mu}_X) + \hat{\beta}_2(X_i - \hat{\mu}_X)^2)$$

The estimated standardized residuals are defined as

$$\hat{r}_i = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2}}$$

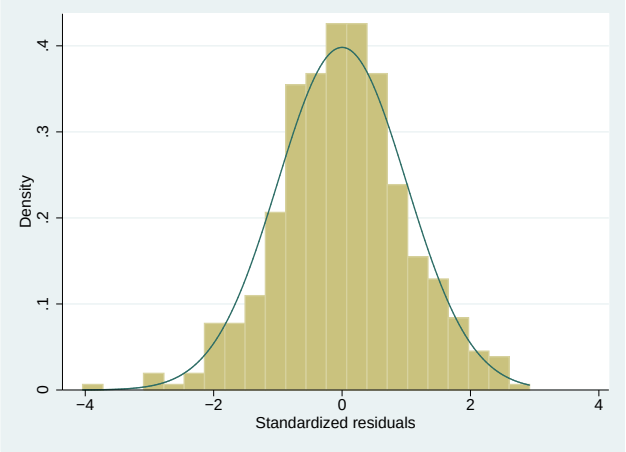
They are often referred to as 'Pearson residuals'.

If the model is correct, these residuals are normally distributed with constant variance and therefore can be used to assess the model's assumptions. In Stata:

```
. predict r, rst
. hist r,normal
. count if r>3 | r<-3
   12
```

The histogram of the standardized residuals show no departure from the normality assumption, nor is there evidence of too many outliers (out of 500):

Figure 3: Birth weight data: histogram of standardized residuals from model (2)



Chapter 1

Simple treatment of dependent data

In this session we introduce situations where measurements are clustered. We discuss the problems that might arise when we reduce the data by creating cluster specific summaries and when we ignore the correlation within clusters. These approaches are compared to standard linear regression models in which cluster differences are modelled by including them as fixed effects.

1.1 Dependent data

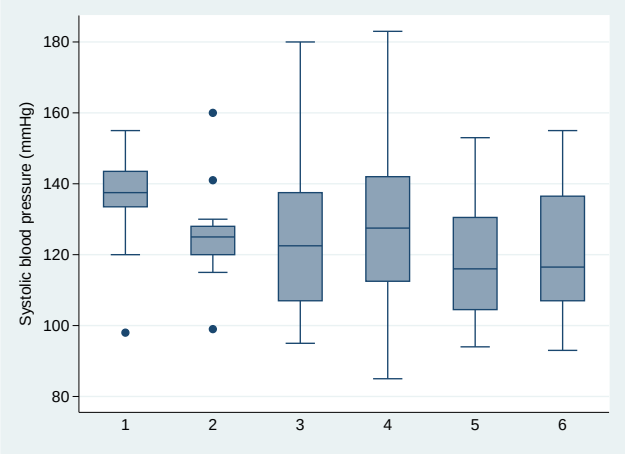
The assumption behind linear and generalized regression modelling is that observations are mutually **independent** and identically distributed, possibly conditional on some covariates. There are many common settings where this assumption does not hold. Here are some examples.

SBP of patients from different hospitals

Consider the data in Figure 1.1 which summarizes the systolic blood pressure measured on 12 patients in each of six hospitals. The hospital-specific means vary from a maximum of 135.7 to a minimum of 117.8, while the overall mean is 125.6. Note how the hospital-specific dispersion also varies.

Is the overall mean useful?
Does the overall mean provide complete information?
Could hospitals differ in any systematic way?

Figure 1.1: Box and whiskers plot of measured SBP in patients from six hospitals

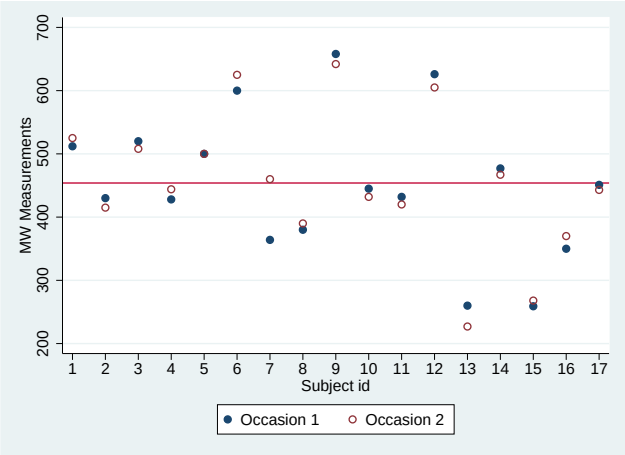


Peak-expiratory-flow rate (PEFR)

The quality of two instruments for the measurement of peak-expiratory-flow rate (PEFR) was measured on 17 people in an experiment reported by Bland and Altman (*Lancet* I, 1986, 307-310). The two instruments were the Standard Wright and the Mini Wright peak flow meter.

Each method was used twice, in a random order to avoid confounding the effect of experience of measurement with that of the method. Figure 1.2 shows the scatter plot of the repeated measures from the Mini Wright measures.

Figure 1.2: Two recordings of PEFR using the Mini Wright meter

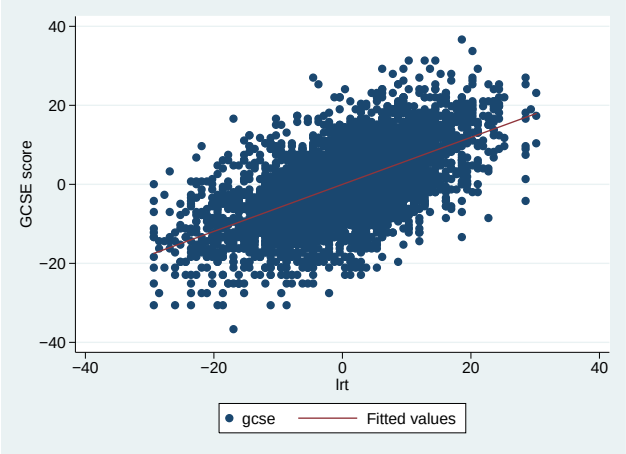


Are repeated observations from the same patient independent?

GCSE results from different schools

There has been increasing interest in the UK in comparing schools performance. Figure 1.3 shows the plot of GCSE scores against reading scores at school entry (denoted LRT) of 4,059 pupils from 65 schools. Both GCSE and LRT scores are centred around their overall averages. In this example we are interested in the relationship between GCSE score at age 16 and reading ability at age 11 when the pupils enter the school. A fitted linear regression line is superimposed onto the data in Figure 1.3.

Figure 1.3: GCSE by LRT in all 65 schools



The schools vary in terms of structure (for example 31% are girls-only schools). They could however vary in other ways, for example in the social background of their pupils or additional educational resources provided by families (*which were not measured*).

Figure 1.4 plots GCSE versus LRT scores for pupils in 4 randomly selected schools. There is variation in both spread of data and the slopes across these 4 schools, which may not be fully explained by known differences between schools, such as single versus mixed-sex intake.

Figure 1.4: GCSE by LRT in four randomly selected schools

Would you assume that all schools have the same relation between GCSE and LRT?
How would you analyse GCSE-LRT relationships across schools?

Birth weight in siblings

Another example where observations are correlated arises when they are from members of the same family. For example a dataset we will consider in this course holds information on 8,604 children born to 3,978 mothers. Most of these mothers (N=3,300) have two children. Figure 1.5 shows the birth weight of their first two children plotted against the mothers identifier. It is quite clear that children born to the same mother are more likely to have similar birth weights.

If a mother is particularly tall, she will on average have heavier children. In the absence of any information on factors that may explain the correlation among children of the same mother, analyses of the data might be confounded by these (unaccounted for) associations.

Would you treat these birth weights as independent?

Children's growth

Consider now a study of how children grow in infancy. We have 198 data records for 68 children who were weighed on up to four occasions, at around 6 weeks of age and then at 8, 12 and 27 months. The observed data are shown in Figure 1.6 and indicate that the relationship between weight in infancy and age is not linear. It also shows a strong degree of 'tracking' (correlation) shown by the fact that the individual growth trajectories are quite well separated, possibly more so for boys.

How would you analyse these data?

If we estimated the regression lines for each child we would produce a collection of poorly estimated curves since we have at most 4 measurements per child. If we ignored the tracking present in the data, we would possibly summarize the data incorrectly.

Figure 1.5: Birth weight of siblings by maternal identifier

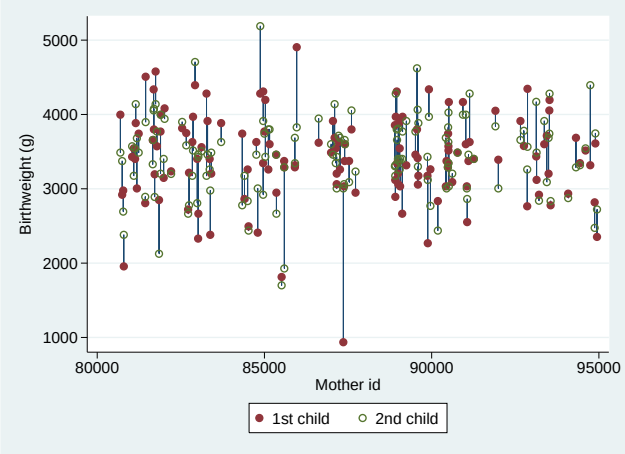
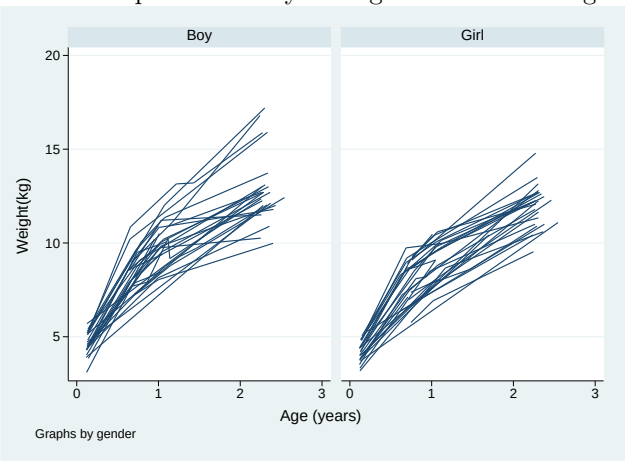


Figure 1.6: Growth profiles of boys and girls in the Infant growth data



1.2 Origins of dependency

In these examples it may not be appropriate to assume that the observation units are independent, even if we condition on known measured variables. This is because:

1. patients from the same hospital may share several unknown features, e.g. disease severity, which will make their SBP more similar;
2. repeated measures of the same underlying feature in one individual will be more similar to each other than measures taken from different individuals;
3. children in the same school may share similar unmeasured home circumstances and therefore obtain similar GCSE scores for a given LRT;
4. siblings share genetic and environmental factors;
5. measures of growth taken on the same individual will be more correlated with each other than with those taken on another individual.

In each of these examples there is a **hierarchy** in the data. There are **elementary units** and

aggregate units (with the latter also called **clusters**). Elementary units nested within aggregate units in a hierarchy:

Level	
Aggregate	**Elementary**
hospitals	patients
individuals	PEFR measures
schools	pupils
mothers	children
children	visits

For this reason data such as those in the examples above are called **hierarchical** or **nested** or **multilevel** or **clustered** data, depending on the context.

In certain settings the clustering is a direct result of the study design (e.g. by *multistage sampling*) which could be carried out for example to reduce costs or improve efficiency. This often arises in sociology or demography. Clinical trials that have a crossover design or where the intervention is applied to a group of individuals (cluster randomised trials) also have clustering as part of the design. In other medical settings however the clustering is an aspect of the population being studied (e.g. children who belong to the same family) or a consequence of what is being studied (e.g. growth).

The growth data example belongs to a particular class of hierarchical data: it is an example of **longitudinal data**, where the strength of dependency within each cluster is influenced by the time interval elapsed between observations.

All the examples described in this chapter have two levels of aggregation *(elementary: 1st level; aggregate: 2nd level)*. However we could easily find other examples with more levels, e.g. pupils clustered within classes which are clustered within schools, or patients, clustered within operating surgeons, who in turn are clustered within hospital.

1.3 Consequences of dependency

If the statistical dependency among observations is ignored, any subsequent inferences are potentially invalid. Both the estimates of a parameter and its confidence interval may be biased. Dependency therefore must be dealt with. We will discuss alternative approaches to achieve this, with the choice driven by the aim of the analysis. We will be fitting models and therefore deal with estimation of parameters of interest and their inference.

Estimation

A simple approach is to ignore the dependence when estimating parameters, but this can (although not necessarily) be an *inefficient* procedure. Examples of this are the use of ordinary least squares when data are dependent, and so-called generalized estimating equations (which we cover later in this module). Alternatively we can fit models that explicitly specify the nature of the dependency, with several modelling options available.

Inference

Even if the dependency in the elementary level observations is ignored when estimating model parameters, the dependence among observations *must* be accommodated when making inferences (i.e. when testing hypotheses). To do this we could either:

- estimate parameters and their confidence intervals directly from a likelihood analysis with a statistical model that incorporates the dependency, or
- after estimating the parameters, estimate their precision in a second stage, using a *robust* approach (as in the Robust Methods Module in Term 1).

1.3.1 Marginal and Conditional Models

The distinction between **marginal** and **conditional** models occurs quite generally, i.e. we don't need dependent data to make it, but it becomes especially important in the dependent data setting.

Suppose that we have a generalised linear regression model for Y, in which we separate out a single covariate (Z say) for special attention, i.e.

$$g\{\mathrm{E}(Y \mid \mathbf{X}, Z)\} = \beta\mathbf{X} + \gamma Z.$$

This model as it stands is a *conditional model*, it describes how the expectation of Y depends on the full set of covariates: the interpretation of each coefficient is in terms of its influence on the mean of Y, *conditional on the values taken by all the other covariates.*

In this sense all the regression models met so far are conditional.

Suppose now that we want to look at the effect of the covariates in $\mathbf{X}$ on the behaviour of the mean of Y *averaged over the levels of* Z.

This leads to the *marginal* model:

$$\mathrm{E}_Z\{\mathrm{E}(Y \mid \mathbf{X}, Z)\} = \mathrm{E}_Z\{g^{-1}(\beta\mathbf{X} + \gamma Z)\}.$$

Assume without loss of generality that $\mathrm{E}(Z) = 0$.

Then if we have a **linear** regression model, g is the identity link i.e.

$$\mathrm{E}(Y \mid \mathbf{X}, Z) = \beta\mathbf{X} + \gamma Z.$$

and marginally β has the same interpretation as in the conditional model:

$$\mathrm{E}_Z\{\mathrm{E}(Y \mid \mathbf{X}, Z)\} = \mathrm{E}_Z(\beta\mathbf{X} + \gamma Z) = \beta\mathbf{X} + \gamma\mathrm{E}(Z) = \beta\mathbf{X}.$$

With a non-linear regression model such as the logistic or Cox proportional hazards model (but excluding the Poisson log-linear model) no such simplification exists in general, and there may be no simple relationship between the parameters in the marginal and the parameters in the conditional model.

This difference is a consequence of *averaging* and *scale*. For example for logistic regression:

$$\mathrm{logit}(E[Y \mid \mathbf{X}, Z]) = \beta\mathbf{X} + \gamma Z$$

Marginalising involves averaging on the scale of the *probabilities:*

$$\mathrm{P}[Y = 1] = \mathrm{E}_Z[\mathrm{expit}(\mathrm{E}[Y \mid \mathbf{X}, Z])]$$

where expit is the inverse of logit.

Why does this matter for dependent data?

For hierarchical data we often fit models that include terms for the aggregate level effects to represent what is common among observations from one "cluster" "or group". For example, we may want to accommodate similarities between multiple measurements made on the same person by including a term in the model for that person. If we then want to make marginal conclusions about the population, we need to average over these individual effects. This only becomes a major issue in non-linear models; the focus in this model is linear models but this provides good groundwork for future modules.

1.4 Aims of the Module

The aim of this module is to introduce methods for the statistical analysis of dependent data when the outcome of interest is continuous and when there are only two levels of aggregation. The analysis of three or more levels and the analysis of binary dependent data are covered in "Advanced Statistical Models" (an optional module in term 3).

The material taught here is a development of ANOVA and the linear regression model and knowledge of these methods is assumed. Because of the special nature of the dependency structure in longitudinal data, we will dedicate separate lectures to their analysis. We will also discuss key concepts regarding the mechanisms leading to missing data and relate them to the implicit assumptions made by the models we are going to use. We will also have a guest lecture on the application of these methods in practice.

1.5 Notation

Let Y_{ij} denote an observation made on the i-th elementary unit of the j-th cluster, where $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, and $N = \sum_{j=1}^{J} n_j$. Note that n_j is the size of cluster j and J is the total number of clusters. If all clusters have the same size n, then $N = Jn$ and the data structure is referred to as **balanced**.

1.6 Aggregation

In the past, a common procedure adopted for clustered data was to aggregate the elementary level data, for example by generating averages for each cluster. This is appropriate as long as one appreciates the following issues.

- The reliability of these averages may vary from cluster to cluster (as this depends on the number of elementary units per cluster).
- The meaning of the variable at the aggregated level is different from that at the original elementary unit.This may lead to invalid inference due to the 'ecological fallacy', as correlations at the macro level (aggregate) cannot be used to make assertions at the micro (individual) level.
- Important information is lost.
- Analyses of aggregate data may have low power, especially if the number of clusters is small.

Artificial example

The data below refer to 5 clusters of 2 observations (Snijders and Bosker, 1999, p 28). Each symbol represents a different cluster.

NB: there is no random variation in the data points as they lie on straight lines, see Figure 1.7.

If we ignore the clustering and fit a regression line we obtain

$$\hat{Y}_{ij} = 5.33 - 0.33X_{ij}$$

This model will be referred to as **Total Regression**. The fitted regression is shown as a dashed line in Figure 1.8, where the individual data points are identified by dots.

Alternatively, if we only use the cluster mean values (which are: (2,6), (3,5), (4,4), (5,3), and (6,2)) we obtain

$$\hat{\overline{Y}}_j = 8.0 - 1.0\overline{X}_j$$

. This model will be referred to as **Between Regression** and is shown as a dotted line in Figure 1.8.

Thus the association between X and Y is steeper across the cluster means, as shown in Figure 1.8 where the cluster-specific means are identified by red crosses.

Which model would you use to summarize the data?

cluster (j)	id (i)	X	Y	$\overline{X}$	$\overline{Y}$
1	1	1	5	2	6
1	2	3	7	2	6
2	1	2	4	3	5
2	2	4	6	3	5
3	1	3	3	4	4
3	2	5	5	4	4
4	1	4	2	5	3
4	2	6	4	5	3
5	1	5	1	6	2
5	2	7	3	6	2

Figure 1.7: Artificial data: scatter of clustered data

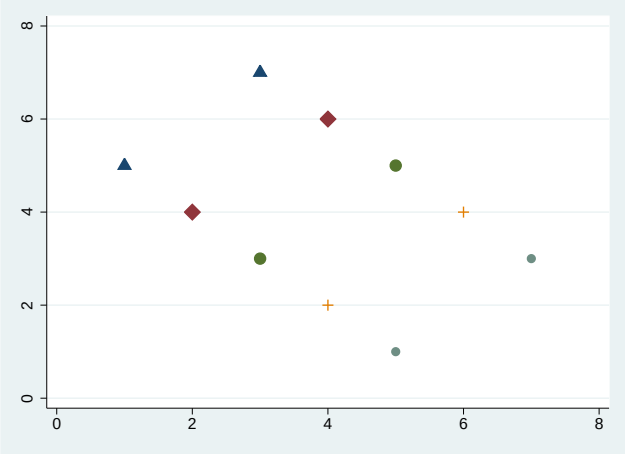
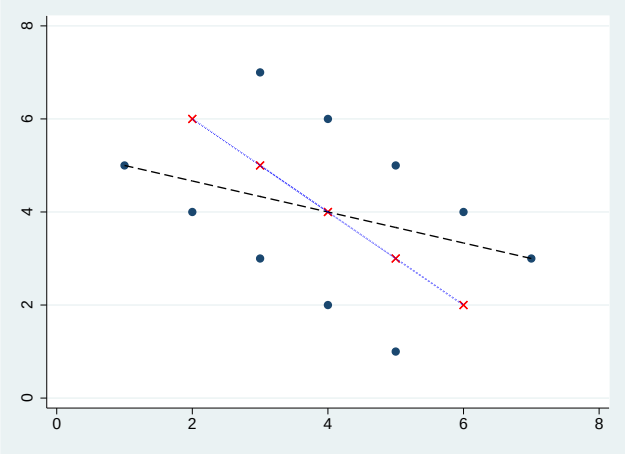


Figure 1.8: Artificial data: total and between clusters regression



Ecological fallacy

The ecological fallacy follows from thinking that relationships observed for groups necessarily hold for individuals: if countries with more salt in their population's diet have higher rates of cardiovascular disease (CVD) incidence, then we might infer that people who eat salty foods must be more likely to suffer from CVD. This may or may not be true at the individual level.

An example of ecological fallacy that is often cited concerns the relationship between country of birth and literacy. For each of the 48 states in the USA of 1930, Robinson (Ecological correlations and the behaviour of individuals (1950). *Am Soc Review*, 15,351-7) computed two numbers from the 1930 Census: the percent of the population who were born outside the USA, and the percent who were literate (in English). The correlation between the 48 pairs of numbers is 0.53. This is an ecological correlation (the unit of analysis is a group of people) which suggests a positive association between birth outside the USA and literacy: that is, individuals born outside the USA are more likely to be literate (in English) than the individuals born inside the USA. In reality, the association is negative: the correlation computed at the individual level is -0.11. The ecological correlation gives the wrong inference: it is positive because individuals born outside the USA tended to live in states where the USA-born individuals were relatively more literate.

The same bias arises in Figure 1.8 where the relation among the cluster means is in the opposite direction of that within the clusters, i.e. at the individual level.

1.7 Disaggregation

One option we have when analysing clustered data is to analyse each cluster separately. In the artificial data example, the data can be re-analysed fitting separate regression lines in each cluster (after centering both variables about their means to facilitate comparisons). Thus we estimate the following equation for each of the five clusters:

$$\hat{Y}_{ij} - \overline{Y}_j = \beta(X_{ij} - \overline{X}_j), \qquad j = 1, \ldots, 5$$

This model will be referred to as **Within Regression**.

The estimated slope is 1 for each regression model because the data were generated without including any random error (Figure 1.9. However, the estimated intercepts differ as each cluster has a distinctive baseline level, with cluster 1 having the highest value and cluster 5 the lowest.

It is only because of the way the data were generated that the estimated cluster specific slopes are all equal. In general the estimates of cluster-specific slopes will not be identical, even if the relation within each cluster is the same.

If we are prepared to assume that the relation within each cluster is the same, we could combine these estimates. Otherwise we are left with many sub-analyses.

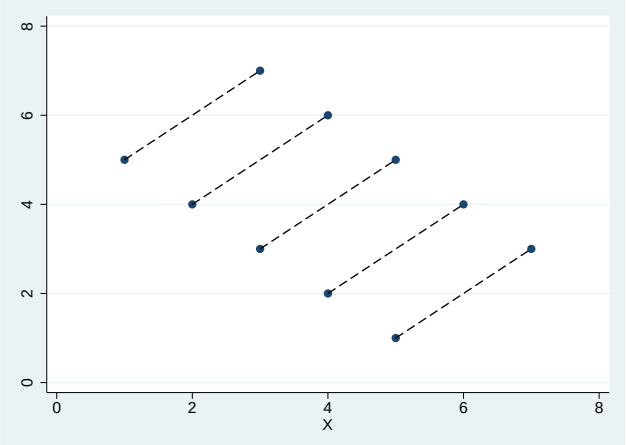
Be aware of data duplication

When disaggregating data one has to be careful not to 'duplicate data'. For example if we are interested in variables that are available only at the cluster level, and we analyse the data at the individual level, we would pretend to have more data than we actually have.

Example: GCSE results from different schools
The GCSE data introduced in Section 1.1 include information on whether schools are single-sex or mixed-sex. The table below gives their distribution using the clustered level information and the elementary level information. Note how even in this simple example the use of the disaggregated data is misleading, with 34% of schools appearing to be girls-only schools if one uses the disaggregated data, while the correct value is 31%:

School type	**Cluster level**		**Elementary level**	
	N	%	N	%
mixed	35	54	2,169	53
boys only	10	15	513	13
girls only	20	31	1,377	34
Total	65	100	4,059	100

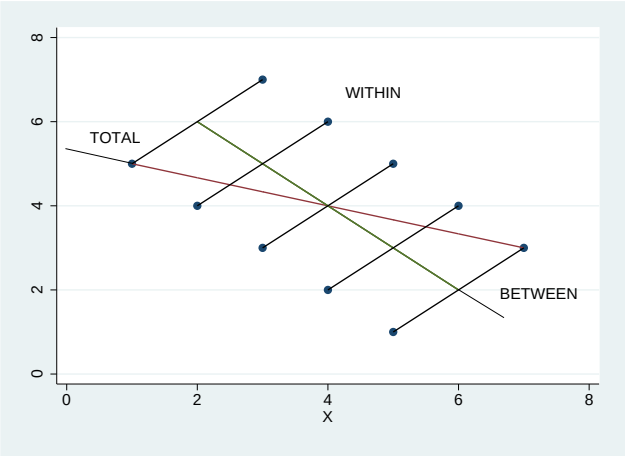
Figure 1.9: Artificial data: within cluster regressions



1.8 Joint modelling

The three models fitted to the artificial data are shown again in Figure 1.10:

Figure 1.10: Artificial data: within, between and total regressions



They are labelled 'TOTAL', 'BETWEEN' and 'WITHIN' as they refer respectively to the first model where the clustering was ignored, the second model that only used the cluster-level means, and the last model that was fitted separately within each cluster.

We have already discussed how the Total model ignores the dependency among the level 1 observations and therefore gives inefficient estimates and biased measures of precision (see Section 1.3). As regards the Within and Between model, instead of choosing between them, we could combine them within a more general model. There are two main ways to specify such a model depending on whether we treat the between cluster variation as random or fixed. In this chapter we will consider the specification where this variation is fixed.

1.9 The fixed effects model

Whenever the clusters have a specific interpretation, e.g. if the clusters in the artificial example refer to five ethnic groups, then we would want to quantify the contribution of ethnicity to the variation in the response. This is achieved by fitting a multivariable regression model with five dummy indicators for the ethnic groups (excluding the intercept from the model). Formally we would fit:

$$Y_{ij} \quad = \quad \alpha_1 I_{i,j=1} + \alpha_2 I_{i,j=2} + \ldots + \alpha_5 I_{i,j=5} + \beta_1 X_{ij} + \epsilon_{ij}$$

where j indicates the cluster, i the individual observation within that cluster, and $I_{i,j=1} = 1$ if the observation belongs to cluster 1 and 0 otherwise, $I_{i,j=2} = 1$ if the observation belongs to cluster 2 and 0 otherwise, etc. Here the error terms ϵ_{ij} are assumed to be independent, identically distributed random variables with mean 0 and variance σ^2. A short-hand for this model is:

$$Y_{ij} = \alpha_j + \beta_1 X_{ij} + \epsilon_{ij} \tag{1.1}$$

where α_1 for example stands for $\alpha_1 I_{i,j=1}$.

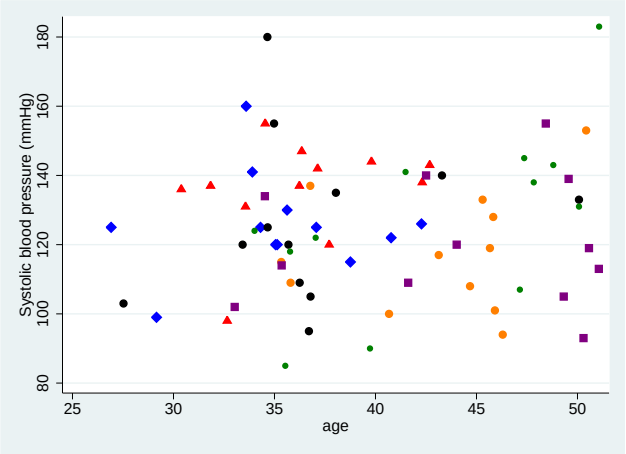
Models such as 1.1 are referred to as a **Fixed effects models** in the context of hierarchical data for reasons that will be clearer later in the course.

In this specification of the model the intercept is α_1 for cluster 1, α_2 for cluster 2, etc., while the slope is assumed to be β_1 for all clusters. Thus, we combine the between and within modelling approaches making certain assumptions for the data generation, namely that the within-cluster slopes are the same.

The blood pressure data

Now let us consider a real dataset. Earlier in Chapter 0 we saw an example where systolic blood pressure (SBP) measurements had been collected on 72 patients from 6 different hospitals. Of interest now is the relation between the patients' measurements of SBP and age. Figure 1.11 shows a scatter plot of the data with different symbols for different hospitals.

Figure 1.11: SBP versus age: different symbols identify the six hospitals



To fit the fixed effects model of equation (1.1) using Stata where X stands for age, and the intercepts refer to predicted SBP at the sample average age, we first generate a new variable (centered age) `c_age` equal to the observed age minus the overall mean age and then the dummy indicators `h_1, h_2`, etc. using `tabulate` with the `generate` option (and `qui` (i.e. `quietly`) simply silences the output for that command). We also use the `noconstant` option to obtain hospital specific intercepts:

```
. use bp_feb,clear
. su age
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        72    39.24429    5.976611   27.56244   52.56657

. gen c_age=age-r(mean)
. qui ta hosp,gen(h_)
. reg bp c_age h_*, noconstant

      Source |       SS           df       MS      Number of obs   =        72
-------------+----------------------------------   F(7, 65)        =    522.27
       Model |  1142756.32         7  163250.902   Prob > F        =    0.0000
    Residual |  20317.6848        65  312.579766   R-squared       =    0.9825
-------------+----------------------------------   Adj R-squared   =    0.9806
       Total |     1163074        72  16153.8056   Root MSE        =     17.68

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       c_age |   1.693115   .4071832     4.16   0.000     .8799139    2.506316
         h_1 |    142.044    5.32922    26.65   0.000     131.4008    152.6872
         h_2 |   130.0188   5.209971    24.96   0.000     119.6137    140.4238
         h_3 |   130.1199    5.17088    25.16   0.000     119.7929    140.4468
         h_4 |   123.8905   5.167309    23.98   0.000     113.5707    134.2103
```

```
        h_5 |   115.0748   5.146691     22.36   0.000     104.7961    125.3534
        h_6 |   112.1854   5.459843     20.55   0.000     101.2814    123.0895
------------------------------------------------------------------------------
```

The estimated parameters tell us that in this population of patients from the six hospitals there is on average a 1.69 mmHg increase in SBP per year increase in age. The six hospitals have different intercepts, with hospital 1 having the largest value and hospital 6 the smallest.

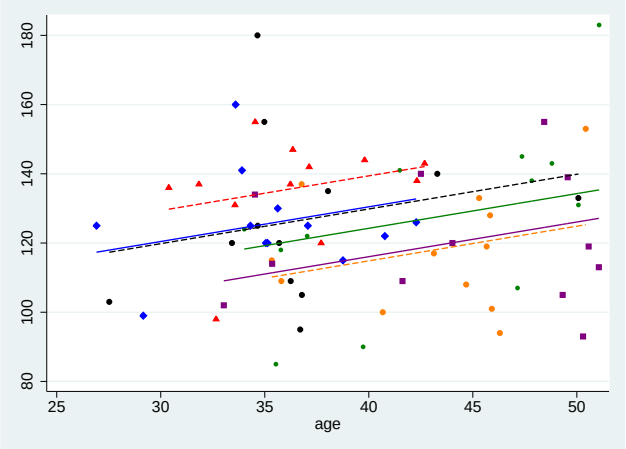
We can test whether these intercepts are significantly different from each other, controlling for differences in age distribution, with

```
. test h_1==h_2==h_3==h_4==h_5==h_6

 ( 1)  h_1 - h_2 = 0
 ( 2)  h_1 - h_3 = 0
 ( 3)  h_1 - h_4 = 0
 ( 4)  h_1 - h_5 = 0
 ( 5)  h_1 - h_6 = 0

       F(  5,    65) =    3.62
            Prob > F =    0.0059
```

which gives a partial F-test $F(5, 65) = 3.62$ with P=0.006. Therefore there is evidence against the null hypothesis that the six hospital have the same intercept, controlling for age.

This model assumes that there is the same SBP-age relation in every hospital and therefore that the regression lines for the six hospitals are parallel (see Figure 1.12). We could assess this by testing the joint significance of the interaction terms between the hospital dummy indicators and `c_age`.

Figure 1.12: SBP and age: fitted regression lines for the six hospitals



The peak-expiratory-flow rate data

We use the peak-expiratory-flow rate (PEFR) data described in Section 1.1 to examine the special setting where there are no explanatory variables in the fixed effects model.

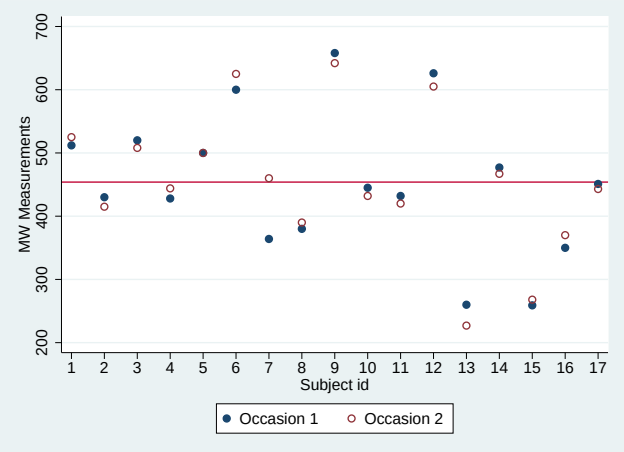
The data concern the results of repeated measurement of PEFR (measured in litres per minute), taken on 17 people with the Mini Wright instrument. The data are listed below (with `id` indicating the subject identifier, `wm1` and `wm2` the two 'Mini Wright' measures next to the subject-specific means `mean_wm`):

id	wm1	wm2	mean_wm
1	512	525	518.5
2	430	415	422.5
3	520	508	514
4	428	444	436
5	500	500	500
6	600	625	612.5
7	364	460	412
8	380	390	385
9	658	642	650
10	445	432	438.5
11	432	420	426
12	626	605	615.5
13	260	227	243.5
14	477	467	472
15	259	268	263.5
16	350	370	360
17	451	443	447

The overall mean is 453.9118 *l/min* and the SD of the 17 cluster specific means is 111.2912.

Figure 1.13 shows how the repeated measures on the same subject are closer to each other than to measures taken on any other subject.

Figure 1.13: PEFR data: simple scatter plot



If we fit a fixed effects regression model to these data with dummy indicators for each subject and no intercept:

$$Y_{ij} = \alpha_j + \epsilon_{ij}$$

we estimate the mean PEFR value for each individual with α_j representing the mean value for individual j. In Stata:

```
. use pefr,clear
. reshape long  wm wp, i(id) j(occasion)
. qui tab id,gen(id_)
. reg wm id_*,nocons
```

```
<EDITED OUTPUT>
------------------------------------------------------
          wm |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------
        id_1 |      518.5   14.07908    36.83   0.000
        id_2 |      422.5   14.07908    30.01   0.000
        id_3 |        514   14.07908    36.51   0.000
        id_4 |        436   14.07908    30.97   0.000
        id_5 |        500   14.07908    35.51   0.000
        id_6 |      612.5   14.07908    43.50   0.000
        id_7 |        412   14.07908    29.26   0.000
        id_8 |        385   14.07908    27.35   0.000
        id_9 |        650   14.07908    46.17   0.000
       id_10 |      438.5   14.07908    31.15   0.000
       id_11 |        426   14.07908    30.26   0.000
       id_12 |      615.5   14.07908    43.72   0.000
       id_13 |      243.5   14.07908    17.30   0.000
       id_14 |        472   14.07908    33.52   0.000
       id_15 |      263.5   14.07908    18.72   0.000
       id_16 |        360   14.07908    25.57   0.000
       id_17 |        447   14.07908    31.75   0.000
------------------------------------------------------
```

Alternatively we could fit a model with an intercept representing the mean value for one subject, e.g. the first one, and 16 additional parameters representing the expected differences in mean PEFR of each individual relative to the first:

```
.  reg wm i.id
<EDITED OUTPUT>
------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          id |
           2 |        -96   19.91083    -4.82   0.000    -138.0082   -53.99182
           3 |       -4.5   19.91083    -0.23   0.824    -46.50818    37.50818
           4 |      -82.5   19.91083    -4.14   0.001    -124.5082   -40.49182
           5 |      -18.5   19.91083    -0.93   0.366    -60.50818    23.50818
           6 |         94   19.91083     4.72   0.000     51.99182    136.0082
           7 |     -106.5   19.91083    -5.35   0.000    -148.5082   -64.49182
           8 |     -133.5   19.91083    -6.70   0.000    -175.5082   -91.49182
           9 |      131.5   19.91083     6.60   0.000     89.49182    173.5082
          10 |        -80   19.91083    -4.02   0.001    -122.0082   -37.99182
          11 |      -92.5   19.91083    -4.65   0.000    -134.5082   -50.49182
          12 |         97   19.91083     4.87   0.000     54.99182    139.0082
          13 |       -275   19.91083   -13.81   0.000    -317.0082   -232.9918
          14 |      -46.5   19.91083    -2.34   0.032    -88.50818   -4.491819
          15 |       -255   19.91083   -12.81   0.000    -297.0082   -212.9918
          16 |     -158.5   19.91083    -7.96   0.000    -200.5082   -116.4918
          17 |      -71.5   19.91083    -3.59   0.002    -113.5082   -29.49182
             |
       _cons |      518.5   14.07908    36.83   0.000     488.7957    548.2043
------------------------------------------------------------------------------
```

In this case the model is:

$$Y_{ij} = \alpha_1 + \delta_j + \epsilon_{ij}$$

where $\delta_j = \alpha_j - \alpha_1$ and $\delta_1 = 0$.

Alternatively still we could fit a model for the differences of each observation with the overall mean,

i.e. for $Y_{ij} - \hat{\mu}$ where $\hat{\mu} = \overline{Y}$. Here the model is:

$$\begin{aligned} Y_{ij} - \mu &= \gamma_j + \epsilon_{ij} \\ Y_{ij} &= \mu + \gamma_j + \epsilon_{ij} \end{aligned} \tag{1.2}$$

where μ is the overall mean and $\sum_{j=1}^{J} \gamma_j = 0$.

In Stata:

```
. su wm
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          wm |        34    453.9118    110.5198        227        658
. gen c_wm=wm-r(mean)

. reg c_wm id_*,nocon
<EDITED OUTPUT>
------------------------------------------------------------
        c_wm |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------------
        id_1 |   64.58823   14.07908     4.59   0.000
        id_2 |  -31.41177   14.07908    -2.23   0.039
        id_3 |   60.08823   14.07908     4.27   0.001
        id_4 |  -17.91176   14.07908    -1.27   0.220
        id_5 |   46.08823   14.07908     3.27   0.004
        id_6 |   158.5882   14.07908    11.26   0.000
        id_7 |  -41.91177   14.07908    -2.98   0.008
        id_8 |  -68.91177   14.07908    -4.89   0.000
        id_9 |   196.0882   14.07908    13.93   0.000
       id_10 |  -15.41176   14.07908    -1.09   0.289
       id_11 |  -27.91177   14.07908    -1.98   0.064
       id_12 |   161.5882   14.07908    11.48   0.000
       id_13 |  -210.4118   14.07908   -14.94   0.000
       id_14 |   18.08824   14.07908     1.28   0.216
       id_15 |  -190.4118   14.07908   -13.52   0.000
       id_16 |  -93.91177   14.07908    -6.67   0.000
       id_17 |  -6.911765   14.07908    -0.49   0.630
------------------------------------------------------------
```

These three specifications of the fixed effects model are equivalent. Thus their Residual Mean Square (`MSE`) is the same. It is equal to 396.44118, leading to an estimated residual SD of 19.91.

In the second specification we find that the difference in mean value of each individual relative to this first (i.e. of the δ_j) ranges from -255 (for `id`=15) to +131.5 (for `id`=9), with mean −64.58824 and SD 111.2912.

Similarly in the third specification we find that the difference in mean value of each individual relative to the overall mean (i.e. the mean of the γ_j) ranges from -210 (for `id`=13) to +191 (for `id`=9), with mean 0 and again SD 111.2912. This SD is also the same as that among the sample cluster-specific means (see page 26).

In each case this estimate of the between-cluster SD is biased because it is affected by noise within each cluster, i.e. each cluster-specific mean is estimated with error. This point will be revised in the next session.

Chapter 2

The random intercept model

In this session we introduce the simplest specification of the *Mixed Effects* model: the random intercept model. We review alternative approaches to estimating this model and drawing inferences. There are different ways of achieving this in Stata and we will review them.

2.1 The random intercept model

Sometimes we are not really interested in the cluster specific intercepts. Also, by fitting separate intercepts for each cluster we ignore the information on the between cluster variation. Because of these considerations we may decide to treat the variation in cluster intercepts not as fixed (and therefore estimated individually) but as random draws from some distribution.

The statistical model then becomes the combination of:

- an overall mean μ plus a cluster level component u_j that represents the departure of cluster j's mean from the overall mean,
- the departure for individual i in cluster j from the cluster mean, e_{ij}

We can express this as:

$$Y_{ij} = \mu + u_j + e_{ij} \tag{2.1}$$

Here u_j is a random variable with mean 0 and variance σ_u^2 (the population between-cluster variance) while the residuals e_{ij} are also random variables with mean 0 and variance σ_e^2 (the population within-cluster variance). u_j and e_{ij} are assumed to be independent, so that $Cov(u_j, e_{ij}) = 0$. The total variance of Y_{ij} is then $\sigma_u^2 + \sigma_e^2$.

For this reason this model is also called **variance-component model** and also **one-way random effects ANOVA model**, because it generalizes the standard one-way ANOVA.

This random effects model differs substantially from the fixed effects model as formulated in equation 1.2 which stated:

$$Y_{ij} = \mu + \gamma_j + \epsilon_{ij}$$

where μ was as here the overall mean. In equation 1.2 however there is the constraint that $\sum_{j=1}^{J} \gamma_j = 0$ while in model 2.1 the random variables u_j have replaced the fixed parameters γ_j and have a distribution with mean zero and variance σ_u^2.

We use the terms **fixed effect(s)** for the linear model parameters (μ in this example) and **random effects** for u_j. A model with both fixed and random (excluding the residual) effects is called a **mixed effects model** or **mixed model** for short.

The independence of the u_j and e_{ij} means that observations from two different clusters are independent, but the presence of u_j in the values from subjects in the same cluster induces a within-cluster (or within class) correlation :

$$\begin{aligned}\mathrm{Cov}(Y_{1j}, Y_{2j}) &= \mathrm{Cov}(u_j, u_j) + \mathrm{Cov}(u_j, e_{2j}) + \mathrm{Cov}(e_{1j}, u_j) + \mathrm{Cov}(e_{1j}, e_{2j}) \\ &= \mathrm{Cov}(u_j, u_j) = \mathrm{V}(u_j, u_j) \\ &= \sigma_u^2.\end{aligned}$$

Given $\mathrm{Var}(Y_{1j}) = \mathrm{Var}(Y_{2j}) = \sigma_u^2 + \sigma_e^2$, the correlation between the pair of observations is then

$$\lambda = \frac{\mathrm{Cov}(Y_{1j}, Y_{2j})}{\mathrm{SD}(Y_{1j})\mathrm{SD}(Y_{2j})} = \frac{\sigma_u^2}{\sigma_e^2 + \sigma_u^2}$$

This is the within-cluster or **intra-class correlation**, that we will denote λ. Note that it is also the proportion of total variance that is accounted for by the cluster.

2.2 Estimation

The random intercept model 2.1 has three parameters: μ, σ_u^2, and σ_e^2. A classical method for estimating parameters in a statistical model is **Maximum Likelihood (ML)**. This requires the assumption that, for example, the u_j and e_{ij} are all normally distributed.

When the data are *balanced* (i.e. have the same number n of units in each cluster j, $j = 1, \ldots, J$), the ML estimators for model 2.1 have simple expressions:

$$\hat{\mu} = \overline{Y}$$

$$\hat{\sigma}_e^2 = \mathrm{MSE}$$

$$\hat{\sigma}_u^2 = \frac{\mathrm{MSS}}{Jn} - \frac{\hat{\sigma_e^2}}{n} \tag{2.2}$$

where MSE is the Mean Squared Error and MSS the Model Sum of Squares, as defined in section 0.1.1.

The first two estimators are unbiased if the model is true, while that for σ_u^2 is downward biased. An alternative unbiased estimator for σ_u^2 is the moment (or ANOVA) estimator, given by

$$\tilde{\sigma}_u^2 = \frac{\mathrm{MSS}}{(J-1)n} - \frac{\hat{\sigma}_e^2}{n} = \frac{(\text{MMS-MSE})}{n} \tag{2.3}$$

That is, the model sum of squares (ie the between clusters SS) is divided by the model degrees of freedom (df) $(J-1)n$ (recall that the ML estimator of the residual variance from ordinary regression has denominator N, the sample size, not the residual df $(N-1)$ which we use in practice. Here $N = nJ$.)

For balanced data the ANOVA (*'Moment'*) estimator is also the **Restricted Maximum Likelihood** or **REML** estimator. Thus, REML generalizes the principle of using the residual df. It is a genuine maximum likelihood procedure, but applied to a linear transformation of the data that "removes" the fixed effect (in this case μ). The variance parameters are estimated from this likelihood and then the fixed effects are estimated in a second step. More details concerning REML are given later in the course when this simple random intercept model is generalised.

Another estimation method is **Generalized Least Squares (GLS)**, commonly used in econometrics. Here the fixed parameters are estimated by a weighted version of OLS, with weights that depend upon the variance components (in the simple balanced case this gives the same as the ANOVA/REML estimates).

For unbalanced data REML and ML estimators implicitly use GLS.

2.3 Analysis in Stata

To fit a random intercept model in Stata the data need to be in long format. The PEFR data, which are in wide format:

```
. use pefr, clear
. l id wm1 wm2 if id==1,noobs
  +------------------+
  | id    wm1   wm2  |
  |------------------|
  |  1    512   525  |
  +------------------+
```

need to be reshaped:

```
. reshape long wp wm, i(id) j(occasion)

.  l id occasion wm if id==1,noobs
  +------------------------+
  | id   occasion     wm   |
  |------------------------|
  |  1          1    512   |
  |  1          2    525   |
  +------------------------+
```

To fit a random intercept model in Stata we have a choice of using either of 2 commands: we can use `mixed` or `gllamm` (as well as two older commands, `xtreg` and `xtmixed`, but we are not going to use them as they are outdated). For the moment we will only consider `mixed`.

The command `mixed` is a general command for fitting random effects models, including the random intercept model. Hence to estimate the parameters in the model for the Mini-Wright measurements using REML we would use:

```
. mixed wm || id:, reml

<OUTPUT OMITTED>

Mixed-effects REML regression                   Number of obs      =        34
Group variable: id                              Number of groups   =        17

                                                Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2
------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9118   26.99208    16.82   0.000     401.0083    506.8153
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                  var(_cons) |   12187.51   4379.546      6026.127    24648.57
-----------------------------+------------------------------------------------
               var(Residual) |    396.441   135.9781      202.4039    776.4942
------------------------------------------------------------------------------
```

The syntax of this command requires some explanation: the command is followed by the name of the dependent variable and then by two vertical bars `||` which indicate the beginning of the specification of the random effects (besides e_{ij}).

Note that we can obtain the standard deviations (instead of the variances) of the error terms using the option `stddev` or just typing a new line:

```
. mixed, stddev

Mixed-effects REML regression                   Number of obs      =        34
Group variable: id                              Number of groups   =        17

                                                Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2
                                                Wald chi2(0)       =         .
Log restricted-likelihood = -180.37921          Prob > chi2        =         .
------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9118   26.99208    16.82   0.000     401.0083    506.8153
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                  sd(_cons) |   110.3971   19.83543      77.62813    156.9986
-----------------------------+------------------------------------------------
               sd(Residual) |   19.91083   3.414678      14.22687    27.86564
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    46.96 Prob >= chibar2 = 0.0000
```

In this simple model we have only the random effect u_j for clusters $j = 1, \ldots, J$ identified by the individual identifier `id`. This is why `id` follows the two vertical bars. It is itself followed by `:` to allow the inclusion of explanatory variables for its random variation (we will revisit this part of the syntax in future sessions). At the moment we do not have any explanatory variables and therefore nothing follows the `:`.

Turning to the interpretation of the parameter estimates we can now compute an unbiased estimate of λ using the output of `mixed`:

$$\hat{\lambda} = \frac{\hat{\sigma}_u^2}{(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)} = \frac{110.40^2}{(110.40^2 + 19.91^2)} = 0.97$$

This within-cluster correlation is a measure of **reliability** of the Mini Wright meter, where σ_u^2 represents the between-individual variance of the true PEFR and σ_e^2 the variance of the measurement error within an individual. A value of 97% is therefore indicating very good reliability. We can interpret this as 97% of the variation being attributable to between-individual differences and 3% to differences between repeated measurements within an individual.

Comments

1. The option `reml` of `mixed` could be replaced by the option `mle` to obtain the ML estimates (ML estimates of σ_u would however be biased).

2. A REML estimate of σ_u can also be obtained with the ANOVA moment estimator. Using the command `loneway` we find $\hat{\lambda}$ reported under 'Intraclass correlation':

   ```
   . loneway wm id
   ```

```
                    One-way Analysis of Variance for wm:
                                                Number of obs =        34
                                                    R-squared =    0.9833
    Source                SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between id             396343.24     16    24771.452     62.48     0.0000
Within id                 6739.5     17    396.44118
------------------------------------------------------------------------
Total                  403082.74     33    12214.628

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ---------------------------------------------
            0.96850     0.01527       0.93856     0.99843
         Estimated SD of id effect                110.397
         Estimated SD within id                  19.91083
<OMITTED OUTPUT>
```

3. In all these fitted models, including the fixed effects model, the estimate of σ_e ('the estimated SD within id' in `oneway`) is 19.91. What varies is the estimate of σ_u and therefore of λ, with $\hat{\sigma}_u$ obtained by ML always smaller than that obtained by REML, and both of these always smaller than the observed variance in cluster-level means, which include the imprecision in the mean estimates.

4. A quick way to obtain summary statistics overall, between and within clusters is to use the command `xtsum`. For example, still using the PEFR data:

```
. xtsum wm,i(id)

Variable    |      Mean   Std. Dev.       Min        Max |    Observations
------------+--------------------------------------------+----------------
wm  overall |  453.9118   110.5198        227        658 |     N =      34
    between |             111.2912      243.5        650 |     n =      17
    within  |             14.29081   405.9118   501.9118 |     T =      2
```

2.4 Inference

Inference for the fixed parameter

When data are balanced, i.e. $n_j = n$, the MLE of μ is the overall mean. The estimated SE is

$$\hat{\text{SE}}(\hat{\mu}) = \sqrt{\frac{n\hat{\sigma}_u^2 + \hat{\sigma}_e^2}{Jn}}$$

Note that for a fixed effects model the ML (and equivalently OLS) estimate of μ is the same but the SE is:

$$\hat{\text{SE}}(\hat{\mu}^{\text{F}}) = \sqrt{\frac{\hat{\sigma}_e^2}{Jn}}$$

Hence $\hat{\text{SE}}(\hat{\mu}^{\text{F}}) < \hat{\text{SE}}(\hat{\mu})$.

In the unbalanced case the MLE of μ obtained from the random intercept model is a weighted mean of the cluster specific means:

$$\hat{\mu} = \frac{\sum_j w_j \overline{Y}_{\cdot j}}{\sum_j w_j}$$

where $w_j = \frac{1}{\sigma_u^2 + \sigma_e^2/n_j}$. So small clusters have the same weight as large clusters if σ_e^2 is small relative to σ_u^2.

The null hypothesis that $\mu = 0$ is tested using the z (or z^2, the Wald test) statistics

$$z = \frac{\hat{\mu}}{\hat{\text{SE}}(\hat{\mu})}$$

The 95% confidence interval for μ is computed with

$$\hat{\mu} \pm z_{0.975}\ \hat{\text{SE}}(\hat{\mu})$$

Note that Stata provides only asymptotic results, unlike SAS.

Inference for the random components

Having fitted a random intercept model we should be interested in checking whether there is evidence for such cluster level variation in the intercept. Thus we wish to test whether $\sigma_u^2 = 0$ against $\sigma_u^2 > 0$. Normally we would use the Likelihood Ratio Test (LRT) (between models with and without σ_u^2), and compare the resulting statistic with a χ^2 distribution with 1 degree of freedom. However the null hypothesis is on the boundary of the parameter space because $\sigma_u^2 \geq 0$. Thus standard statistical test theory is invalid.

Under the null hypothesis that $\sigma_u^2 = 0$ we would expect the correlation among the observations to be positive half of the time and negative the other half. Thus we would expect that the estimate of σ_u^2 would be positive half of the time and zero the other half. The correct sampling distribution of the LRT statistic in this case is a mixture of a χ^2 with 1 df and a χ^2 with 0 df. The correct p-value is obtained by halving the 'naive' p-value. This is what is shown at the bottom of the Stata output and labelled `chibar2(01)`. Revisiting our last estimated model:

```
. mixed wm || id:, reml

<OUTPUT OMITTED>
Mixed-effects REML regression                   Number of obs      =        34
Group variable: id                              Number of groups   =        17

                                                Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2
------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9118   26.99208    16.82   0.000     401.0083    506.8153
------------------------------------------------------------------------------


------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                  var(_cons) |   12187.51   4379.546      6026.127    24648.57
-----------------------------+------------------------------------------------
               var(Residual) |    396.441   135.9781      202.4039    776.4942
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    46.96 Prob >= chibar2 = 0.0000
```

Here σ_u^2 is so significantly different from 0 that halving the p-value does not have an impact on the naive conclusions.

Comments

1. The SE of the random effects SD should not be used to construct CIs. However the estimated confidence intervals for the variance components can be used.

2. When the study size is 'sufficiently large' (see practical) estimating a linear mixed model using ML leads to very similar results to REML. In this case we can use LRT based on ML results to perform hypothesis testing involving the fixed effects.

3. When models to be compared are not nested we use criteria such as AIC or BIC.

Chapter 3

The random intercept model with covariates

In this session we:

- extend the random intercept model to include explanatory variables,
- obtain estimates of the random effects that are useful for model assessment,
- discuss the impact on random effects estimates of including additional explanatory variables and
- consider generalizations of the model to include explanatory variables at both levels (within and between clusters).

3.1 Extension of the multivariable regression model

Consider a two-level linear regression model with two covariates:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \epsilon_{ij}$$

With correlated data it may be unrealistic to assume that the residuals ϵ_{ij} for individuals in the same cluster are uncorrelated. Instead we could assume that ϵ_{ij} has two components:

$$\epsilon_{ij} = u_j + e_{ij}$$

Substituting this expression in the linear regression equation above we see that the model is a generalization of the random intercept model:

$$Y_{ij} = (\beta_0 + u_j) + \beta_1 X_{1ij} + \beta_2 X_{2ij} + e_{ij} \tag{3.1}$$

As discussed before this is an example of a **linear mixed model** where there are both fixed and random effects. Here the fixed effect parameters are β_0, β_1, and β_2, while the random effects are u_j and e_{ij}.

It is generally assumed that $E(u_j|X_1, X_2) = 0$ and $E(e_{ij}|X_1, X_2, u_j) = 0$, from which it follows that $E(e_{ij}|X_1, X_2) = 0$. Thus

$$E(Y_{ij}|X_1, X_2) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij}$$

i.e. the marginal regression (over u_j and e_{ij}) is also linear.

It is also generally assumed that $u_j|X_{1ij}, X_{2ij} \sim N(0, \sigma_u^2)$ and $e_{ij}|X_{1ij}, X_{2ij}, u_j \sim N(0, \sigma_e^2)$.

Notes:

- The intra-class correlation between two observations in the same cluster is now conditional on the explanatory variables included in the model.

- The cluster-specific random intercept u_j represents the effect of cluster level characteristics that have not been included in the model (i.e. heterogeneity due to unmeasured cluster-level variables). Its size will decrease as more explanatory variables for the cluster differences are included in the model.

3.2 Example: birth weight of siblings

We have information on 8,604 children born to 3,978 mothers. In particular we know the babies' birth weights (measured in g), gestational age (measured in *weeks*) and sex. We also know the maternal smoking status and age when becoming pregnant with that baby. These variables are, respectively, `birwt`, `gestat`, `male` (coded 0/1), `smoke` (coded 0/1), and `mage`.

To identify the clusters we will also use the variable `momid`, i.e. the mother identifiers.

We are interested in comparing a simple random intercept model for birth weight (the *Null Model*) with a model that also includes these explanatory variables (the *Full Model*). These variables are available on all 8,604 babies so there are no missing values.

Let's start using REML in `mixed` and compare the results obtained from fitting the *Null Model* with those obtained from the *Full Model*. Starting with the null model:

```
. use siblings,clear

. *NULL MODEL
. mixed birwt  || momid:, reml
Mixed-effects REML regression                   Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

                                                Obs per group: min =         2
                                                               avg =       2.2
                                                               max =         3


                                                Wald chi2(0)       =         .
Log restricted-likelihood = -65472.601          Prob > chi2        =         .
------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   3467.969   7.138554   485.81   0.000     3453.977     3481.96
------------------------------------------------------------------------------


------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                  var(_cons) |     135686   4755.873      126677.6      145335
-----------------------------+------------------------------------------------
               var(Residual) |   142625.4   2965.965      136929.1    148558.7
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  1316.12 Prob >= chibar2 = 0.0000
```

Note that we can obtain the standard deviations (instead of the variances) of the error terms using the option `stddev` or just typing a new line:

```
. mixed, stddev
```

To fit the full model (for easier interpretation) we first centre the two continuous variables, gestational age and maternal age, around some sensible values such as 38 weeks of gestation and 30 years of age:

```
. gen c_gestat=gestat-38
. gen c_mage=mage-30
. *FULL MODEL
. mixed birwt c_gestat male smoke c_mage || momid:, reml
Mixed-effects REML regression                   Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

                                                Obs per group: min =         2
                                                               avg =       2.2
                                                               max =         3


                                                Wald chi2(4)       =   2184.89
Log restricted-likelihood =  -64492.44          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    c_gestat |   85.42416   2.160715    39.54   0.000     81.18924    89.65909
        male |   133.9476   8.869438    15.10   0.000     116.5638    151.3314
       smoke |  -239.9995   15.97937   -15.02   0.000    -271.3185   -208.6805
      c_mage |   13.15788    1.09034    12.07   0.000     11.02085    15.29491
       _cons |   3341.096   8.664204   385.62   0.000     3324.114    3358.077
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                  var(_cons) |   99783.52   3713.955      92763.47    107334.8
-----------------------------+------------------------------------------------
               var(Residual) |   118012.6   2471.723      113266.2    122957.9
------------------------------------------------------------------------------
```

All fixed effects included in the model are significant according to their z-test statistics. The interpretation is as follows. Controlling for the other variables in the model the expected effect of an increase of 1 week in gestational age is to increase the expected birth weight of a baby either from the same or another mother (i.e. in any cluster) by over $85g$, the effect of being a baby boy is to increase the expected birth weight by $134g$, etc.

As expected both cluster-level and elementary-level SDs decrease when we include additional explanatory variables (see columns 2 and 3 of Table 3.1).

Also of interest is whether we get different results from using ML instead of REML with these sample and cluster sizes. The right hand side of Table 3.1 shows the estimated SD of the random components obtained using ML. They are very similar to those obtained with REML.

Note that we cannot calculate a LRT of the joint significance of the fixed effects in the full model using the results obtained via REML because the models are not strictly nested. Recall that REML uses a transformation of the data based on the fixed effects in the model, which will differ for these two models.

3.3 Assigning values to the random components

After fitting statistical models it is advisable to check the model's assumptions before embracing any conclusions. Also, it may be of interest to visualize the predicted individual intercept and the predicted

Table 3.1: Summary of estimates of the variation of the random effects of the null and full model obtained using REML or ML

	REML		ML	
Random part	**Null Model**	**Full Model**	**Null Model**	**Full Model**
$\hat{\sigma}_u$	368.3558	315.8853	368.2864	315.732
$\hat{\sigma}_e$	377.6577	343.5296	377.6579	343.4581

cluster level random effects.

There are two possible approaches to predicting these values: ML and Empirical Bayes (EB).

Simple prediction

As in standard linear regression we can substitute the estimated parameters into the model to derive the predicted residuals. These predicted residuals will be a combination of the two random components e_{ij} and u_j. For a model with one explanatory variable, for example:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{1ij} + u_j + e_{ij} \\ &= \beta_0 + \beta_1 X_{1ij} + \epsilon_{ij} \end{aligned}$$

Then we have that

$$\hat{\epsilon}_{ij} = Y_{ij} - \hat{\beta}_0 + \hat{\beta}_1 X_{1ij}$$

We note that the total residual for an individual's measurement $\hat{\epsilon}_{ij}$, is the sum of

1. the distance between overall and cluster-specific average prediction ($\hat{u}_j$) and
2. the difference between individual measurement and cluster-specific average prediction ($\hat{e}_{ij}$).

A simple estimate of u_j is the average of the $\hat{\epsilon}_{ij}$ in cluster j, since $\mathrm{E}(e_{ij} \mid u_j) = 0$.

The straight line in Figure 3.3 shows the overall predicted birth weight by gestational age $\hat{Y}_{ij}$, and the dotted line shows $\hat{Y}_{ij} + \hat{u}_j$ for mother 14.

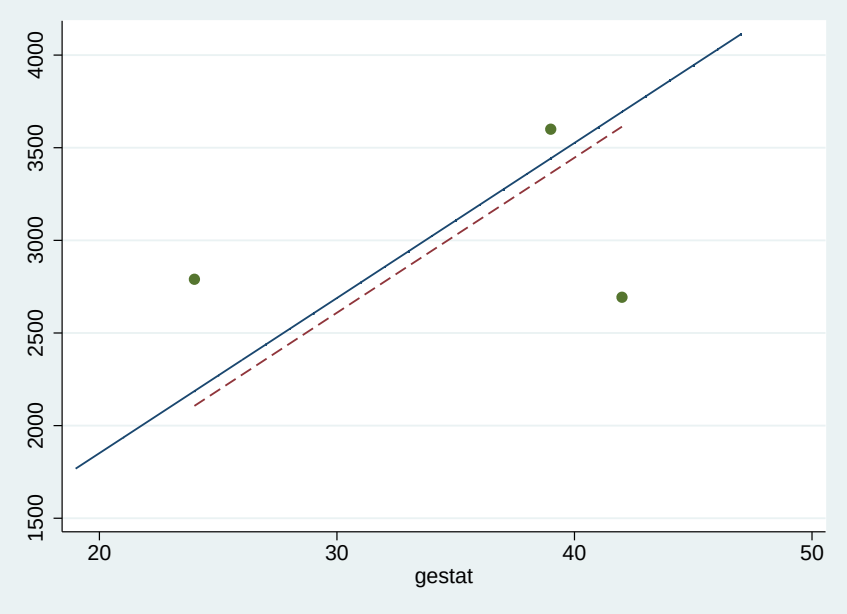
We can obtain these values in Stata with:

```
. qui mixed birwt c_gestat male smoke c_mage || momid:, ml
. predict yhat
. gen res=birwt-yhat
. egen uhat=mean(res),by(momid)
```

where yhat contains the fitted model estimates for each child, res represents the residuals for each child and uhat contains the average residuals for each mother.

Checking the values for mother 14:

```
. sort momid gestat
. l momid gestat birwt yhat res uhat if momid==14,noobs
+-------------------------------------------+
|momid gest bwt   yhat     res         uhat |
|-------------------------------------------|
| 14   24   2790 2186.289  603.7109 -79.67953 |
| 14   39   3600 3442.276  157.7239 -79.67953 |
| 14   42   2693 3693.473 -1000.473 -79.67953 |
+-------------------------------------------+
```



Here we see that mother 14 had 3 children, for one of whom the predicted birth weight is quite low, with $\hat{\epsilon}_{ij} = -1000.473$. On average, the mother's predicted deviation from the overall mean, conditional on the gestational age and sex of the baby and her smoking status and age during the pregnancy, is $\hat{u}_j = 79.7g$.

We can summarise all the mothers' predicted u_js with the following command (note that the function `tag` allows us to pick only one value per mother):

```
. egen pickone=tag(momid)
. summ uhat if pickone
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        uhat |      3978   -.5115937    395.0677  -1386.937   1772.722
```

The mean of the estimated random intercepts is close to zero, as expected, and the SD is nearly $400g$.

The estimates of the random intercepts calculated by Stata in this way are "ML"-based estimates of u_j, which we denote $\hat{u}_j^{\mathrm{ML}}$.

Empirical Bayes prediction

An alternative estimator for random intercepts is the Empirical Bayes (EB) estimator. As before we use the estimates of the model's fixed parameters β to predict the cluster level residuals. In addition however we use the information on their assumed distribution defined by $u_j \sim N(0, \hat{\sigma}_u^2)$.

So we combine this **prior** distribution of u_j with the **likelihood** to obtain their posterior distribution.

In linear random intercept models the ML and EB predictions have a simple relationship:

$$\hat{u}_j^{\mathrm{EB}} = \hat{\mathrm{R}}_j \hat{u}_j^{\mathrm{ML}}$$

where $\hat{\mathrm{R}}_j$ is a measure of **reliability** of $\hat{u}_{ij}^{\mathrm{ML}}$ which is defined as:

$$\hat{\mathrm{R}}_j = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_j} = \hat{\sigma}_u^2 \hat{w}_j$$

where w_j was defined on page 33.

$\hat{R}_j$ is also known as the **shrinkage** factor because it lies between 0 and 1 and therefore shrinks the ML estimator of u_j towards zero. There will be more shrinkage when σ_u is small (clusters are similar), σ_e is large, or n_j is small.

There are good reasons for favouring EB predictions. First, the prediction error, i.e. the difference $\hat{u}_j^{EB} - u_j$ has zero mean over repeated samples of clusters and units within clusters. In addition they are the estimators with the smallest possible variance. In linear mixed models these predictors are therefore also known as **'best linear unbiased predictors' (BLUPs)**. (They are however conditionally biased, see Rabe-Hesketh and Skrondal, page 82.)

Continuing with the same model for the birth weight data, we obtain EB estimates for the cluster level residuals by first calculating the reliability R_j of the ML estimates and then by multiplying reliability by the ML estimated level 2 residuals. Stata does all this for us with the `predict` command with the option `reffects`:

```
. qui mixed birwt c_gestat male smoke c_mage || momid:, ml
. predict uhat_eb2, reffects
```

The `predict` command in Stata can also be used to calculate the standard errors (and therefore confidence intervals) of the random intercepts with the option `reses`:

```
. qui mixed birwt c_gestat male smoke c_mage || momid:, ml
. predict uhat_eb2, reffects reses(uhat_eb2_se)
. gen lower_ci = uhat_eb2 - 1.96*uhat_eb2_se
. gen upper_ci = uhat_eb2 + 1.96*uhat_eb2_se
```

3.4 Diagnostics

In linear mixed models the estimated elementary level (level-1) residuals have normal sampling distributions if the model is true. Thus we can use them to assess the model assumptions. The same is not true for cluster level (level-2) residuals because they are based on all data in the cluster, but it is still worth examining them to identify outliers. Because these estimated residuals are not homoscedastic, we should standardize them first.

Before standardisation, the variance of the EB level 1 predictions is equal to σ_e^2. The standardized level-1 residuals are found after `mixed` with `predict` and the `rstandard` option (as for linear regression). For example (using ML estimation to be consistent with the `gllamm` analysis that is used below):

```
. qui mixed birwt c_gestat male smoke c_mage || momid:, ml
. predict ehat, rstandard
```

However, standardized level-2 residuals are not yet made available by `predict`.

As noted, the marginal variance of the EB level 2 predictions is equal to:

$$R_j \hat{\sigma}_u^2 \tag{3.2}$$

In this simple example (for which $\hat{\sigma}_u^2 = 315.7338^2$ and $\hat{\sigma}_e^2 = 343.4572^2$) we can calculate their variance using equation (3.2) as follows:

```
. predict uhat_eb2, reffects
. sort momid
. qui by momid: gen Nchild=_N
. gen R= 315.7338^2/(315.7338^2+ (343.4572^2)/Nchild)
. gen var_eb=R*(315.7338^2)
. gen uhat_st2=uhat_eb2/sqrt(var_eb)
```

Having calculated the standardized residuals at both levels we can now,

- assess the model assumptions by examining the residuals at level 1, and
- look for outliers by examining the distribution of both types of residuals.

Figure 3.1 and Figure 3.2 show their histograms and Q-Q plots.

Figure 3.1: Histogram and Q-Q plot of cluster (mother) level standardized residuals for the intercept

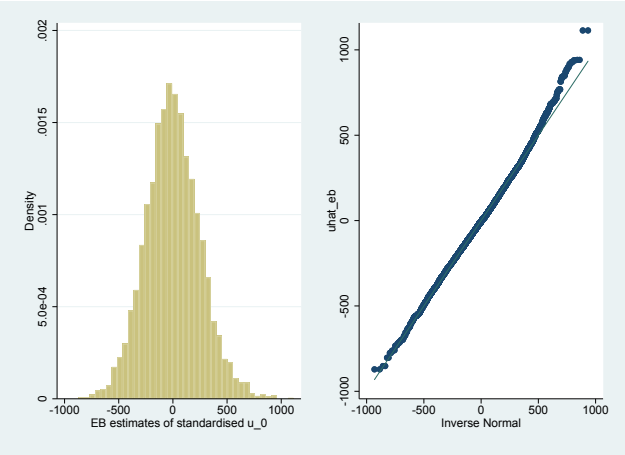
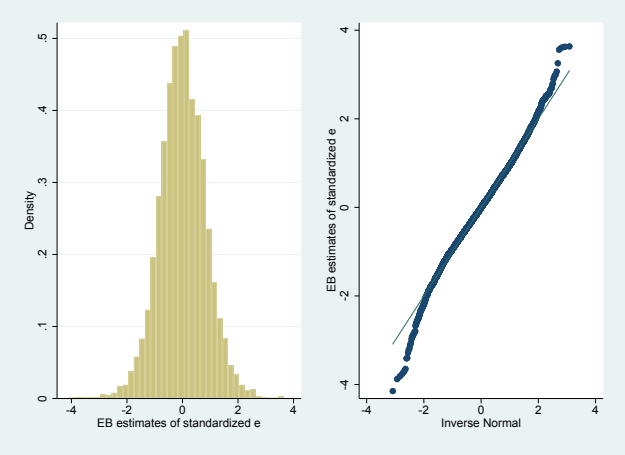


Figure 3.2: Histogram and Q-Q plot of elementary (child) level standardized residuals



gllamm

Direct calculation of the cluster level (i.e. level-2) residuals is available via another set of Stata commands: `gllamm` and `glapred`. These commands are not part of core Stata; they are introduced here because they have features that are not available in `mixed` and because they may be useful beyond this course. First we can use `gllamm` to estimate the random intercept model and then `gllapred` to predict its standardized errors at both levels.

`gllamm` is a very general program that fits a large class of multilevel models including multilevel generalized linear mixed models. It uses a method of numerical integration called adaptive quadrature to evaluate the marginal log likelihood, which is then maximised. Because of its generality it can however be quite slow. Also note that its estimates are ML not REML.

First we check that `gllamm` is installed:

```
. which gllamm
```

If it is not found then install it with:

```
. ssc install gllamm, replace
```

Next we estimate the same random intercept model as that fitted before and then generate the predicted standardized level-1 and level-2 residuals:

```
. gllamm birwt c_gestat male smoke c_mage, i(momid) nip(12) adapt
. gllapred uhat_g,ustd
. gllapred ehat_g,pearson
```

The syntax of these commands is quite straightforward with the exception of two unusual options: `nip` stands for 'number of integration points' and `adapt` for 'adaptive quadrature'; together they control the numerical calculation and maximization of the model log-likelihood function.

Note that the predicted EB cluster-level residuals are held in `uhat_g` and the elementary level ones in `ehat_g`.

3.5 Covariates at cluster level

The dataset holds information on maternal variables that do not change from child to child. These are race (`black` coded 0/1), education (`hsgrad` coded 0/1), and marital status (`married`, coded 0/1). Because they do not vary from child to child but only from mother to mother their effect can only help explain the between cluster variation. We can include them in the model for birth weight data as follows:

```
. mixed birwt c_gestat male smoke c_mage black married hsgrad|| momid:, reml stddev
Mixed-effects REML regression                   Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

                                                Obs per group: min =         2
                                                               avg =       2.2
                                                               max =         3


                                                Wald chi2(7)       =   2285.46
Log restricted-likelihood = -64442.329          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    c_gestat |   84.45556    2.15844    39.13   0.000     80.22509    88.68602
        male |    133.789   8.851405    15.12   0.000     116.4406    151.1375
       smoke |   -227.942   16.33232   -13.96   0.000    -259.9527   -195.9312
      c_mage |    11.0375   1.141231     9.67   0.000      8.80073    13.27427
       black |  -177.8954   25.97089    -6.85   0.000    -228.7974   -126.9934
     married |   61.17158   22.27904     2.75   0.006     17.50546    104.8377
      hsgrad |  -4.211799   13.97923    -0.30   0.763    -31.61059      23.187
       _cons |   3297.461   23.30362   141.50   0.000     3251.787    3343.135
------------------------------------------------------------------------------


------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                   sd(_cons) |   311.2006     5.8627      299.9194     322.906
-----------------------------+------------------------------------------------
                sd(Residual) |   343.5894   3.596225      336.6127    350.7106
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  1073.56 Prob >= chibar2 = 0.0000
```

Now $\hat{\sigma}_u^2$ is 96845.79, a little smaller than before, while $\hat{\sigma}_e^2$ is essentially the same as before.

3.6 Within and between effects

When we fit a model that includes an explanatory variable, say gestational age, that varies among the elementary units, its coefficient represents the effect per unit difference between two children from the same mother or between two children from different mothers. No distinction is made.

If we wish to distinguish between within and between cluster effects we can specify a more general model. Say that X_{ij} varies within and between clusters, e.g. gestational age. It may be that X has a different impact when it varies within the same mother and when it varies across mothers. To study this we separate these two aspects by replacing X_{ij} with its observed mean for mother j, $\overline{X}_{.j}$, and with its difference from the cluster mean for each child in the cluster, $X_{ij} - \overline{X}_{.j}$.

The more general model becomes:

$$Y_{ij} = \beta_0 + \beta_{1B}\overline{X}_{.j} + \beta_{1W}(X_{ij} - \overline{X}_{.j}) + u_j + e_{ij},$$

where β_{1B} and β_{1W} are respectively the between- and within-clusters regression coefficients.

For simplicity consider a model with gestational age only and fit a simple random intercept model:

```
. mixed birwt c_gestat || momid:, reml stddev
Mixed-effects REML regression                   Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

                                                Obs per group: min =         2
                                                               avg =       2.2
                                                               max =         3


                                                Wald chi2(1)       =   1408.15
Log restricted-likelihood = -64819.221          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    c_gestat |    83.7325   2.231357    37.53   0.000     79.35913    88.10588
       _cons |   3358.544     7.1841   467.50   0.000     3344.463    3372.624
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                   sd(_cons) |   336.2624   6.031088       324.647    348.2933
-----------------------------+------------------------------------------------
                sd(Residual) |   352.5369   3.677487      345.4024    359.8188
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  1232.25 Prob >= chibar2 = 0.0000
```

This says that for every additional week of gestation, birth weight is expected to increase by $84g$.

Then we define two new variables to replace `gestat` `avegest`: $\overline{X}_{.j}$ and `difgest`: $(X_{ij} - \overline{X}_{.j})$ with:

```
. egen avegest=mean(gestat),by(momid)
. gen difgest=gestat-avegest
. gen c_avegest=avegest-38
```

where we also centre `avegest` to aid the interpretation of the model's intercept.

The values for mother 14 are:

```
. l momid nchild birwt gestat avegest difgest if momid==14,noobs

  +-------------------------------------------------------+
  | momid   nchild   birwt   gestat   avegest   difgest |
  |-------------------------------------------------------|
  |    14        1    2790       24        35       -11 |
  |    14        2    2693       42        35         7 |
  |    14        3    3600       39        35         4 |
  +-------------------------------------------------------+
```

That is, the average gestation period for mother 14's children was 35 weeks, but the individual children differed from their mother's average gestation by -11, 7 and 4 weeks.

If we fit the extended model using `mixed` we find:

```
. mixed birwt c_avegest difgest || momid:, reml  stddev
Mixed-effects REML regression                   Number of obs      =      8604
Group variable: momid                           Number of groups   =      3978

                                                Obs per group: min =         2
                                                               avg =       2.2
                                                               max =         3


                                                Wald chi2(2)       =   1497.74
Log restricted-likelihood = -64778.628          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   c_avegest |   113.2183   4.029358    28.10   0.000     105.3209    121.1157
     difgest |   70.93503   2.665478    26.61   0.000     65.71079    76.15928
       _cons |   3320.008   8.383309   396.03   0.000     3303.577    3336.439
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                   sd(_cons) |   333.3415   5.972229      321.8393    345.2548
-----------------------------+------------------------------------------------
                sd(Residual) |    351.668   3.655725      344.5754    358.9066
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  1234.20 Prob >= chibar2 = 0.0000
```

There seems to be a stronger effect between ($113g$) than within ($71g$) mothers. The cluster level SD is also smaller in this new model. The interpretation of these results is that,

- for each additional week in average gestation between mothers, the average difference in birth weight between their children increases by $113g$,
- each additional week in gestation between two children from the same mother is associated with a $71g$ increase in birth weight.

To test whether this model is better than the one we started from, we can compute a LRT based however on the results from ML, not those from REML (because otherwise they are not nested). This gives a LRT statistic of 76.22 on 1 d.f., indicating that the new model gives a much better fit to the data. Alternatively we could carry out a Wald test of the equality of β_{1B} and β_{1W} using:

```
. test c_avegest=difgest

 ( 1)  [birwt]c_avegest - [birwt]difgest = 0

           chi2(  1) =   76.60
         Prob > chi2 =    0.0000
```

For more details on how this model relates to within and between regression see Mann et al (2004, *Statist. in Med.* 23:2745-2756).

An example in social sciences surrounds social economic position (SEP) in different countries: there could be an effect within each country that is different from that across countries. In social sciences this approach takes the name of **contextual analysis**.

3.7 Fixed or random?

The models described in equation 1.2 and equation 2.1 show that hierarchical data can be modelled with either fixed or random effects. Which is most appropriate depends on the aims of the analysis, the nature of the clusters, the size of the clusters (up to a point), and the population distribution involved. More specifically:

1. if the groups/clusters are regarded as unique entities (like types of housing) and we want to draw conclusions on each of them, then a fixed effects model is appropriate;
2. if the groups/clusters can be regarded as samples from a real or hypothetical population and we want to draw conclusions on this population, then a random effects model is appropriate;
3. if there are few groups/clusters and we use a random effects model, the random components will be poorly estimated; however if interest is in the fixed parameters of the model a small number of groups/clusters is sufficient;
4. if a fixed effects approach is adopted the size of groups/clusters should be large; in contrast, this is not required in a random effects model (if its assumptions are correct!);
5. the random effects model estimates fewer parameters but relies on distributional assumptions that may not be appropriate; in this case results may be unreliable.

For more discussion see Snijders and Bosker (1999), pp 43-45, and Rabe-Hesketh and Skrondal (2008), pp 61-62.

Chapter 4

The random coefficient model

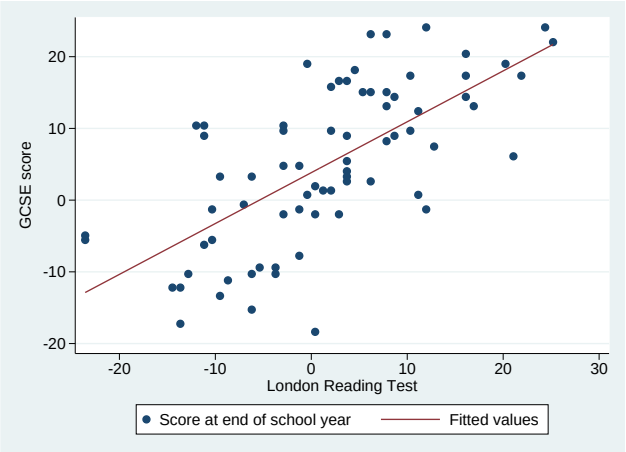
In this session we extend the random intercept model by introducing a new random component for the slope. This model, called the random coefficient model, is also referred to as the random slope model. The assumptions of this more general model are discussed. An example is used to illustrate the new model's specification and interpretation, and to derive predictions.

4.1 Example: GCSE scores in schools

The school data were introduced in Chapter 1. They hold information on the Graduate Certificate of Secondary Education (GCSE) score at age 16 and reading level before entering the school (measured by the London Reading Test (LRT) score) at age 11, of each pupil attending 65 schools, as well as other pupil and school level characteristics. Note that both GCSE and LRT scores are centred about their mean to allow consistent interpretation of the intercepts.

If we regress the GCSE score against the LRT score for pupils in the first school (see Figure 4.1) we find:

Figure 4.1: GCSE versus LRT scores in school 1

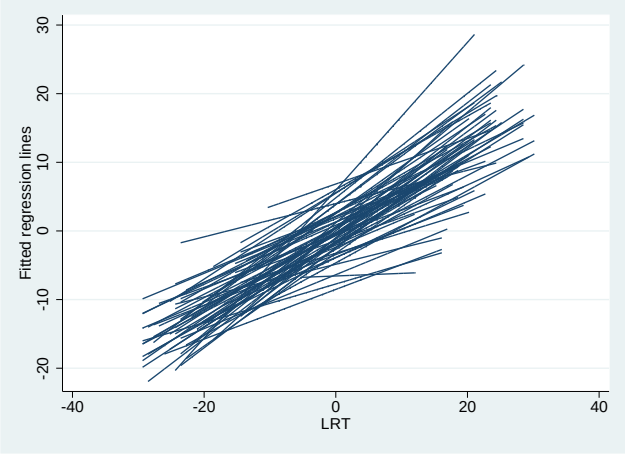



```
. use gcse, clear
. reg gcse lrt if school==1
Source |       SS       df       MS              Number of obs =      73
 ------+------------------------------           F(  1,    71) =   59.44
 Model |  4084.89189     1  4084.89189            Prob > F      =  0.0000
Resid  |  4879.35759    71  68.7233463            R-squared     =  0.4557
```

```
-------+------------------------------           Adj R-squared =  0.4480
 Total |  8964.24948    72  124.503465           Root MSE      =    8.29

------------------------------------------------------------------------------
  gcse |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
   lrt |   .7093406   .0920061     7.71   0.000     .5258856    .8927955
 _cons |   3.833302   .9822377     3.90   0.000     1.874776    5.791828
------------------------------------------------------------------------------
```

If we repeat this for all schools separately, excluding school 48 because it only reports 2 pupils, and save the predicted regression lines for each of them, we find:

Figure 4.2: Predicted regression lines of GCSE versus LRT scores: separate estimates from each school



These school-specific models have varying intercept and slope as seen in Figure 4.2. Indeed their intercepts have mean -0.18 (SD=3.29), with range from -8.52 to 6.84, while their slopes have mean 0.54 (SD=0.18), with range from 0.04 to 1.08.

Figure 4.3 shows the scatter of the school specific intercepts and slopes (with their respective means identified by the lines). Intercepts and slopes are positively correlated (the correlation is 0.36): schools whose children had higher GCSE scores for an average entry LRT score are also those with the greatest improvements in GCSE scores with increasing LRT scores.

We come to this conclusion after fitting 64 regression models, each with 3 regression parameters (intercept, slope and residual variance). A more parsimonious modelling approach is to assume that the school specific intercepts and slopes are distributed randomly around some population means, where, in this case, 'population' stands for the population of all schools in the UK.

4.2 Model specification

In the random intercept model the intercepts vary around a mean value but the coefficient for LRT (i.e. the effect of an explanatory variable) does not vary between schools. We now consider the extension in which the coefficient (slope) also varies randomly across the clusters.

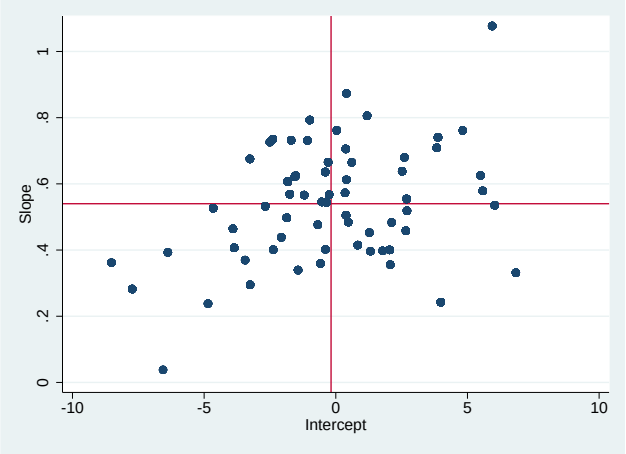
Consider the model:

$$Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})X_{1ij} + e_{ij} \tag{4.1}$$

As before we make a number of assumptions:

- the model is correctly specified (*i.e.* the relationship between X with Y is linear)
- X_{ij} is exogenous (i.e. not correlated with random error), so that

Figure 4.3: School specific slopes and intercepts



- $E(u_{0j}|X_{ij}) = 0$, $E(u_{1j}|X_{ij}) = 0$, $E(e_{ij}|X_{ij}, u_{0j}, u_{1j}) = 0$, and
- $u_0, u_1 \perp X$, $u_0, u_1 \perp e$.

The u's are uncorrelated with e_{ij} and all three are uncorrelated with X_{ij}. Here u_{0j} represents the deviation of the intercept for cluster j from the mean intercept β_0, and u_{1j} represents the deviation of the slope for cluster j from the mean slope β_1.

The main difference between this model and the random intercept model of the previous chapter is that the latter leads to parallel regression lines for all schools.

To estimate this model using ML or REML we need to specify the distribution of all random terms. We usually assume that, given the explanatory variables, the two level-2 terms, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{\Sigma}_u)$ where $\mathbf{\Sigma}_u$ is a 2×2 matrix with a non-zero covariance between u_{0j} and u_{1j} with elements $\sigma^2_{u00}, \sigma_{u01}, \sigma^2_{u11}$.

4.3 Example (cont'd)

Let us start by fitting a fixed effects model, that is by assuming a common slope for all schools but allowing different (but fixed) intercepts. For consistency with our earlier analyses we drop the data from school 48 (which had only 2 pupils):

```
. use gcse, clear
. drop if school==48
. qui ta school,gen(sch_)
. reg gcse lrt sch_*, nocon

  Source |       SS       df       MS              Number of obs =    4057
---------+------------------------------           F( 65,  3992) =   48.69
   Model |  179037.853    65  2754.42851           Prob > F      =  0.0000
Residual |  225851.871  3992    56.57612           R-squared     =  0.4422
---------+------------------------------           Adj R-squared =  0.4331
   Total |  404889.724  4057  99.8002771           Root MSE      =  7.5217

------------------------------------------------------------------------------
       gcse |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
        lrt |   .5594835   .0125347    44.63   0.000     .5349085    .5840585
      sch_1 |   4.082324   .8805959     4.64   0.000     2.355864    5.808784
      sch_2 |   5.620176   1.015436     5.53   0.000     3.629355    7.610997
```

```
.....
<OUTPUT OMITTED>
.....
  sch_64 |  -1.770126     .84147    -2.10   0.035    -3.419877   -.1203747
------------------------------------------------------------------------------
```

The model gives an estimated coefficient (slope) for lrt of 0.56 (0.01) plus 64 intercepts, one per school included in the analysis. The intercepts vary from -9.63 to 7.91; their mean is -0.03 and the SD=3.38. The estimated residual SD is 7.52 (see `Root MSE`).

If we instead fit a random intercept model we find (using REML):

```
. mixed gcse lrt || school: , reml stddev
Mixed-effects REML regression                   Number of obs      =      4057
Group variable: school                          Number of groups   =        64

                                                Obs per group: min =         8
                                                               avg =      63.4
                                                               max =       198

                                                Wald chi2(1)       =   2040.11
Log restricted-likelihood = -14022.025          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
        gcse |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lrt |   .5632679   .0124706    45.17   0.000      .538826    .5877099
       _cons |   .0310058   .4052647     0.08   0.939    -.7632984      .82531
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school: Identity             |
                  sd(_cons) |   3.070397    .311598       2.51658    3.746092
-----------------------------+------------------------------------------------
               sd(Residual) |   7.523608   .0842294      7.360319    7.690519
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   405.60 Prob >= chibar2 = 0.0000
```

The estimated slope is again 0.56 (0.01). The estimated mean intercept is 0.03 (not significantly different from 0 according to the z test); the estimated SD of its (normal) distribution across schools is 3.07. The estimated residual elementary level SD is again 7.52 (because it is estimated by the MSE).

All of these estimates are very close (or equal) to the corresponding ones from the fixed effects model. Those from the linear mixed model were obtained estimating fewer parameters (4 instead of 66) but making additional assumptions on the structure of the (random effects) error terms.

To fit a random intercept and slope model with the command `mixed` we add the name of the explanatory variable for which the regression (slope) parameter is random (in this simple case we add `lrt`) after `school:` in the Stata command. To fit the model as specified on page 49 we also need to use the option `covariance(unstructured)` to specify that u_{0j} and u_{1j} are allowed to be correlated (the default is for their correlation to be equal to zero) and that we are not putting any constraints on the values of $\sigma^2_{u00}, \sigma^2_{u11}$ and σ_{01}. We find:

```
. mixed gcse lrt || school: lrt, reml cov(unstructured) stddev
Mixed-effects REML regression                   Number of obs      =      4057
Group variable: school                          Number of groups   =        64

                                                Obs per group: min =         8
                                                               avg =      63.4
```

```
                                                              max =       198
                                                Wald chi2(1)      =    765.71
Log restricted-likelihood = -14001.482          Prob > chi2       =    0.0000
------------------------------------------------------------------------------
        gcse |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lrt |   .5566002   .0201146    27.67   0.000     .5171762    .5960241
       _cons |  -.1092061   .4026458    -0.27   0.786    -.8983773    .6799651
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
school: Unstructured         |
                     sd(lrt) |   .1223192   .0191969      .0899303    .1663732
                   sd(_cons) |   3.041435   .3105418      2.489818    3.715261
              corr(lrt,_cons) |   .4937123    .149395      .1525706    .7297272
-----------------------------+------------------------------------------------
                sd(Residual) |    7.44194   .0839802      7.279149    7.608372
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =   446.69   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

This specification of the model allows not only the intercept but also the regression coefficient for LRT to vary across schools. This has an impact on the estimated mean intercept which is now -0.11 (although still not significantly different from zero) with a slightly reduced SD (decreased from 3.07 to 3.04). The mean slope is 0.56 but is now allowed to vary across schools with an estimated SD of 0.12. The estimated correlation between slopes and intercepts is 0.49, similar to our naive estimate of 0.36, found when examining the distribution of the school specific regression lines (which was affected by greater noise). The estimated residual (individual level) error SD is 7.44, slightly smaller than in the other models. A summary of these results is given in Table 4.1.

Table 4.1: Comparison of fixed, random intercept and random coefficient models: school data

	Model					
Parameter	**Fixed effect**		**Random intercept**		**Random coeff.**	
	Est	**SE**	**Est**	**SE**	**Est**	**SE**
Fixed part						
β_0	$(-0.03)^a$	-	0.031	0.405	-0.109	0.403
β_1	0.560	0.013	0.563	0.013	0.557	0.020
Random part						
σ_{u00}	[3.38]	-	3.070	0.312	3.041	0.311
σ_{u11}	-	-	-	-	0.122	0.019
corr(0,1)	-	-	-	-	0.494	0.149
σ_e	7.522	-	7.524	0.084	7.442	0.084

corr(0,1)=$\frac{\sigma_{u01}}{\sigma_{u00}\sigma_{u11}}$; (a) sample mean

4.4 Inference

We should assess whether the random coefficient model fits the data better than the random intercept one, i.e. whether both the variance of u_{1j} and the covariance between u_{0j} and u_{1j} are zero.

As noted before when discussing tests for the random intercept model, tests of these hypotheses involve parameter values at the boundary of the parameters space. Hence the LRT of the random coefficient versus the random intercept model does not have chi-square distribution under the null hypothesis (and the degrees of freedom reported by Stata are incorrect). See the `Note` in the Stata output.

Refitting the random intercept and random intercept and slope models to save the values of the maximum likelihood and then computing the LRT:

```
. qui mixed gcse lrt || school: , reml
. estimates store ri
. qui mixed gcse lrt || school: lrt, reml cov(unstructured)
. estimates store rc
. lrtest rc ri

Likelihood-ratio test                                 LR chi2(2)  =     41.09
(Assumption: ri nested in rc)                         Prob > chi2 =    0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the
      boundary of the parameter space.  If this is not true, then the reported
      test is conservative.
Note: LR tests based on REML are valid only when the fixed-effects specification
      is identical for both models.
```

There is strong evidence against the null hypothesis that the regression slopes are all identical (i.e. have zero variance).

Note that the correct sampling distribution of the LRT statistic in this case is a mixture of a χ^2 with 2 df and a χ^2 with 1 df.

But how do we interpret the results of the random coefficient model? The model says that the population mean intercept is -0.11 (which is not significantly different from 0) and the population mean slope is 0.56 (95% CI:0.52, 0.60). Here by 'population' we mean the population of all schools in London, or the UK, or whatever the population this study's data represents. The school level intercepts and slopes vary around these means with SD of respectively 3.04 and 0.12 and a positive correlation of nearly 0.50. The residual (level 1) SD is much larger at 7.44.

These results can be used to construct **ranges** within which 95% of all school intercepts and slopes are expected to fall. These are calculated as: $-0.11 \pm 1.96 \times 3.04$ and $0.56 \pm 1.96 \times 0.12$, that is 95% of the cluster (school)-specific intercepts are expected to range between -6.07 and 5.85, while their slopes lie between 0.33 and 0.80. Comparing these values with those found by separate regression analyses on each school ($-0.18 \pm 1.96 \times 3.29$, i.e. -6.63 to 6.27, and $0.54 \pm 1.96 \times 0.18$, i.e. 0.19 to 0.89) we see that the random coefficient model gives us a tighter description of the data (see also the last section of this chapter).

4.5 On the random effects variances

We should take care when interpreting the random part of the model. First we should keep in mind that the units of the random intercepts and random slopes variances are not the same.

The units of the variance of the intercept are the square of the units of Y (GCSE score) whilst the units of the variance of the slope are the square of the ratio of the units of Y over the units of X_1 (GCSE score per LRT score).

Further we should appreciate that the residual variance, i.e. the variance of Y_{ij} given X_{1ij}, is not constant for all values of X_1.

We see this once we re-write the random coefficient model, separating the fixed from the random part:

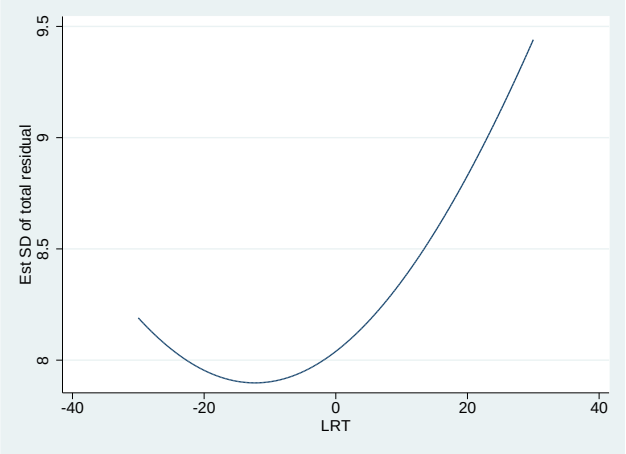
$$\begin{aligned} Y_{ij} &= (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})X_{1ij} + e_{ij} \\ &= (\beta_0 + \beta_1 X_{1ij}) + (u_{0j} + u_{1j}X_{1ij} + e_{ij}) \\ &= (\beta_0 + \beta_1 X_{1ij}) + \epsilon_{ij} \end{aligned}$$

Thus the total residual variance depends on X_1 (see page 49 to revisit the notation):

$$\begin{aligned} \text{Var}(Y_{ij}|X_{1ij}) &= \text{Var}(u_{0j} + u_{1j}X_{1ij} + e_{ij}) && (4.2) \\ &= \sigma^2_{u00} + X^2_{1ij}\sigma^2_{u11} + 2X_{1ij}\sigma_{u01} + \sigma^2_e && (4.3) \end{aligned}$$

For the schools data, the random coefficient model has total residual SD varying with LRT as shown in Figure 4.4.

Figure 4.4: SD of the total residual error ϵ_{ij}



Also note that intra-class correlation is conditional on X_1, i.e. the correlation between two level 1 individuals belonging to the same cluster, depends on the values of X_1 taken by these two individuals.

4.6 Predictions

A useful representation of the results is to plot them using the post estimation command `predict` after fitting the model with `mixed`.

We can then plot them against the explanatory variable with:

```
. qui mixed gcse lrt || school: lrt, reml cov(unstructured)
. predict yhat_re, fitted
```

The results are in Figure 4.5. As expected these estimates are less spread out than those obtained by separate linear regression models.

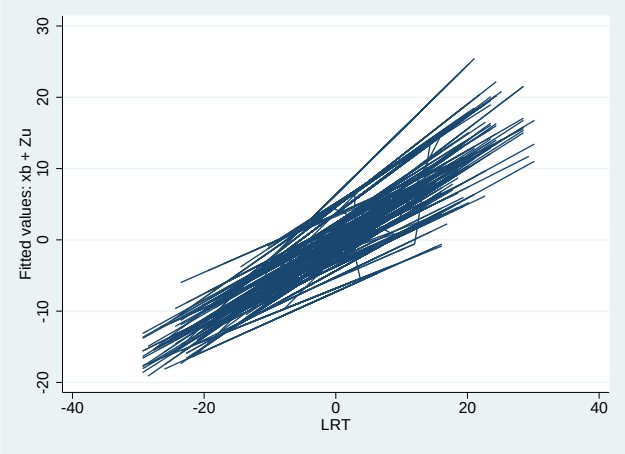
4.7 Model assessment

We could also examine the distribution of the level 1 and level 2 residuals to assess the appropriateness of the model's assumptions. As already discussed, the post estimation command `predict` used after fitting the model with `mixed` does not give standardized level 2 residuals, only standardized level 1 residuals. If the clusters are large (e.g.school classes), then approximate level 2 standardized residuals can be obtained by dividing the estimated random effects by their standard errors using the option `reses`.

We can gather some understanding of the model's appropriateness by examining the original (non-standardized) cluster level residuals. After creating an indicator `pickone` to pick only one student per school to represent that school we can do the following:

```
. egen pickone = tag(school)
. predict ebslope ebinterc, reffects reses(ebslope_se, ebinterc_se)
. predict rst1,rstandard
```

Figure 4.5: EB predictions of school-specific regression lines: random intercept and slope model




```
. label var ebinter "EB intercept"
. label var ebslope "EB slope"
. label var rst1 "EB level 1 residual"
. qnorm ebinterc if pickone==1
. qnorm ebslope if pickone==1
. hist rst1 , normal freq
. qnorm rst1
```

The plots for unstandardized level 2 residuals are shown in Figure 4.6 and for standardized level 1 residuals in Figure 4.7. There seems to be a large outlier among the level 2 random slopes, which should be investigated further.

Figure 4.6: Q-Q plot of unstandardized EB predictions of the random intercept and slope model

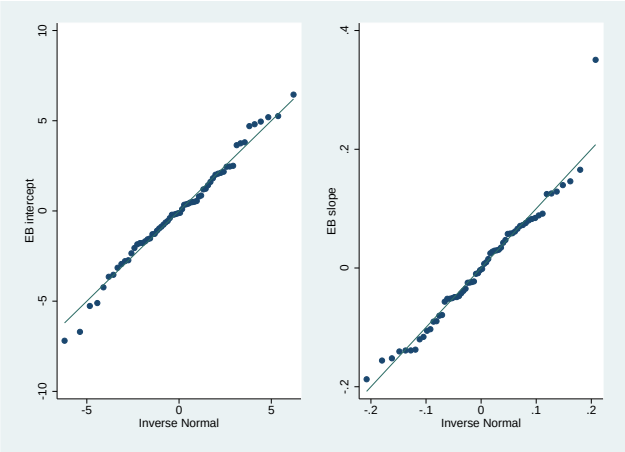
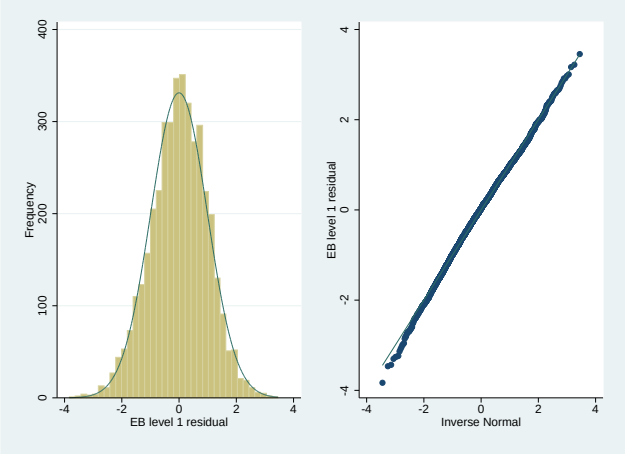


Figure 4.6 describes the unstandardized school-level residuals. Using approximate standard errors for the random effects we can produce the approximate standardized ones as follows.

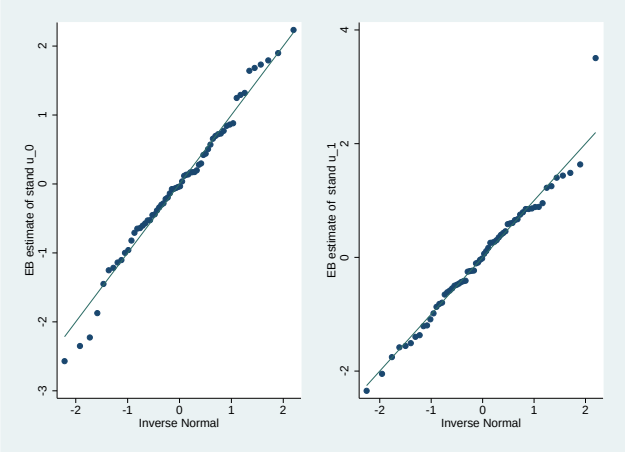
```
. gen ebinterc_std = ebinterc/ebinterc_se
. gen ebslope_std = ebinterc/slope_se
. qnorm ebinterc_std if pickone==1
. qnorm ebslope_std if pickone==1
```

Figure 4.7: Histograms and Q-Q plot of EB predictions of standardized level 1 residuals



These are plotted in (Figure 4.8).

Figure 4.8: Q-Q plots of EB predictions of the standardized level 2 residuals



Chapter 5

Longitudinal data I

In this session we revisit random intercept and random intercept and slope models in the context of longitudinal data using Stata.

In session 6 we will embed this in a general class of Linear Mixed Models for dependent data, while in session 7 we will deal with other issues arising when modelling longitudinal data.

5.1 Introduction

Longitudinal data are prospective data in which a sample is followed over time with information collected at several time points. These time points are not always the same for all individuals, they could vary from individual to individual. In clinical trials they are usually the same (they are collected at *'fixed occasions'*) but in epidemiology the observation times are more likely to vary (*'variable occasions'*). Data such as these are also called **repeated measures data** and in econometrics **panel data** or **cross sectional time series data**. Some Stata commands define the data using the `xt` command and this formulation arose from time series methods.

In longitudinal data the clusters are the individuals who are followed over time. Thus the same family of models as those discussed in earlier sessions are applicable to longitudinal data. Longitudinal data do, however, have a feature that is not usually present with other clustered data: the correlation structure of longitudinal data is governed by time and therefore the longitudinal observations within a cluster have an order. This is unlike the case of standard aggregated data where elements within each cluster can be *exchanged* with each other (i.e. pupils in schools). In other words it is very unlikely that the observations in the same clusters have the same within-cluster correlation, as defined for example on page 29.

There are different possible approaches to deal with such structures depending on whether the observations over time are at fixed occasions, or not.

5.2 Fixed occasions

"Fixed occasions" relates to longitudinal data that are collected at fixed times t_i, $i = 1, \ldots, n$ on all study participants. Note that because occasions are fixed and in equal numbers for all, the data are balanced, i.e. have equal cluster sizes, that is $n_j = n, \forall j$.

As with all types of data analysis one should always start with some simple descriptive summaries of the data. With longitudinal balanced data this is quite easy, unless there are several missing values. For example mean values could be calculated at each time point to examine 'average profiles'. Alternatively, simple regression models could be fitted for each individual over time and the distribution of the subject-specific intercepts and slopes examined.

Missing values

Great care is required when analysing data that are affected by missing values, even when calculating simple summary statistics. If some individuals have missing data at some time points, the overall average profile mentioned above would not represent the true average because different people contribute at different times.

For the results to be unbiased the individuals with missing values should not be in any way selected because of their value of the response Y. If some selection does occur corrections can be made **if the mechanism driving the selection depends on measured data**. With random effects models such corrections are 'automatic'.

Random intercept model

The classical model for repeated measures is called the **compound symmetry model**. It is a slight generalization of the random intercept model as introduced so far.

When there are no explanatory variables it is defined as:

$$Y_{ij} = \mu_i + u_{0j} + e_{ij} \tag{5.1}$$

where μ_i is the mean intercept at occasion i (and is a fixed effect).

To fit this model we create a series of dummy indicators to identify the occasions:

$$Y_{ij} = \sum_{h=1}^{n} \beta_{0h} I_{i=h,j} + u_{0j} + e_{ij}$$

where $I_{i=h,j}$ are dummy indicators that identify whether the i-th observation for subject (i.e. cluster) j was taken at occasion h.

The model implies that for observations within the same cluster:

$$\begin{aligned} \mathrm{Cov}(Y_{1j}, Y_{2j}) &= \mathrm{Cov}(u_{0j} + e_{1j}, u_{0j} + e_{2j}) \\ &= \sigma^2_{u00} \end{aligned}$$

that is, they all have the same covariance, i.e. they are exchangeable, exactly as for the standard random intercept model.

Observations from different clusters are also uncorrelated, as before:

$$\begin{aligned} \mathrm{Cov}(Y_{1j}, Y_{2j^*}) &= \mathrm{Cov}(u_{0j} + e_{1j}, u_{0j^*} + e_{2j^*}) \\ &= 0 \end{aligned}$$

When there are no missing values, the fixed occasion data for each cluster j can be viewed as a vector of $\{Y_{ij}\}$, each of size n, and for which the covariance matrix (of size $n \times n$) is:

$$\Omega_y = \begin{pmatrix} \sigma^2_{u00} + \sigma^2_e & \sigma^2_{u00} & \cdots & \sigma^2_{u00} \\ \sigma^2_{u00} & \sigma^2_{u00} + \sigma^2_e & \cdots & \sigma^2_{u00} \\ \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \\ \sigma^2_{u00} & \sigma^2_{u00} & \cdots & \sigma^2_{u00} + \sigma^2_e \end{pmatrix}$$

It is because of the form of the covariance matrix, that the name of **compound symmetry model** is used to describe this model.

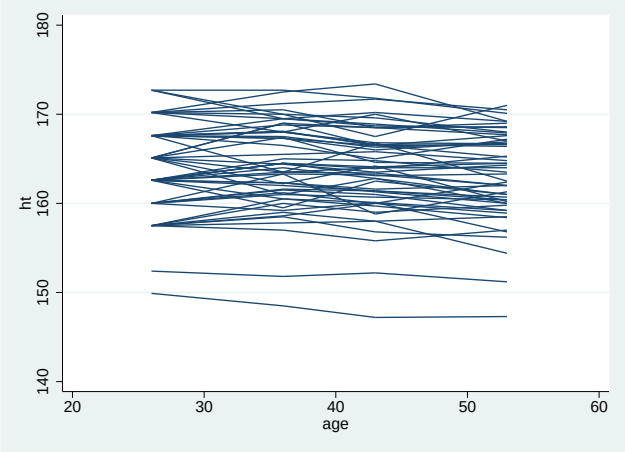
Example: adult height measures

We have height measurements taken in adulthood on 1980 women. Ignoring here any possible differences in measurement quality, we wish to study whether the data show any evidence of decreasing height with age.

The table below gives a summary of the heights measured at ages 26, 36, 43,and 53 and Figure 5.1 shows a sample of their profiles:

Age	N	Mean	SD
26	1758	162.33	6.36
36	1610	162.26	6.05
43	1567	162.27	5.96
53	1462	161.56	5.96

Figure 5.1: Height profiles of some randomly selected women



Note that, in principle, all women should be measured at all times. Missing values could have occurred for many reasons but for the sake of this example we will assume that missingness is completely at random (i.e. is not related to height or age).

Among those with complete data the sample variance-covariance matrix is

```
. corr ht*,cov
(obs=1173)

             |     ht26     ht36     ht43     ht53
-------------+------------------------------------
        ht26 |  39.8134
        ht36 |  34.7585  34.4551
        ht43 |  34.4790  33.3604  34.3315
        ht53 |  34.1282  33.0867  32.9489  34.2152
```

To fit the compound symmetry model using Stata we first generate the dummy variables and then fit a random intercept model without the 'constant', i.e. without a baseline intercept (note that the `noconstant` option goes before the specification of the random part of the model):

```
. qui tab age, gen(age_)
. mixed ht age_*, noconst || id:, reml stddev
Mixed-effects REML regression                   Number of obs      =      6397
Group variable: id                              Number of groups   =      1980
                                                Obs per group: min =         1
                                                               avg =       3.2
                                                               max =         4
                                                Wald chi2(4)       =  1.42e+06
Log restricted-likelihood = -15237.576          Prob > chi2        =    0.0000
```

```
------------------------------------------------------------------------------
          ht |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_1 |   162.3414   .1390258  1167.71   0.000     162.0689    162.6139
       age_2 |   162.3174    .139767  1161.34   0.000     162.0434    162.5913
       age_3 |   162.1943   .1399716  1158.77   0.000       161.92    162.4687
       age_4 |   161.4532   .1404093  1149.88   0.000      161.178    161.7284
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |    5.99213   .0973276      5.804376    6.185957
-----------------------------+------------------------------------------------
                sd(Residual) |   1.409331   .0150026      1.380231    1.439044
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) = 10810.51 Prob >= chibar2 = 0.0000
```

The test for the difference in means across the four occasions is highly significant (using the Wald test after fitting the models using REML chi2(3) = 374.56; using the ML after using ML chi2(3) = 359.60).

The estimated covariance matrix is of size 4×4 and is equal to (because $\sigma^2_{u00} = 5.992^2 = 35.91$ and $\sigma^2_e = 1.409^2 = 1.99$):

$$\hat{\mathbf{\Omega}}_y = \begin{pmatrix} 37.90 & 35.91 & 35.91 & 35.91 \\ 35.91 & 37.90 & 35.91 & 35.91 \\ 35.91 & 35.91 & 37.90 & 35.91 \\ 35.91 & 35.91 & 35.91 & 37.90 \end{pmatrix}$$

Examining the cluster level and elementary level residuals (see Figure 5.2 and Figure 5.3) we see that the latter residuals appear to have too many extreme values.

The approximate, standardized, level 2 residuals were calculated as follows:

```
. sort id age
. egen pickone=tag(id)

. mixed ht age_*, noconst || id:, reml
. predict ehat_st, rstandard
. predict uhat_eb, reffects reses(uhat_eb_se)
. gen R= 5.99213^2/(5.99213^2+ (1.409331^2)/3.23)
. gen uhat_eb_se2 = R*5.99213^2
. gen uhat_st=uhat_eb/uhat_eb_se2

. hist uhat_st if pickone
. qnorm uhat_st  if pickone
. hist ehat_st
. qnorm ehat_st
```

In the above, 3.23 is the average number of measurements per person.

Alternatively we could have treated age as an explanatory variable and fitted a random intercept model with a linear effect of age (after centering age at 26 years to improve the precision of the estimates):

```
. mixed ht age || id:, reml
Log restricted-likelihood = -15285.955
------------------------------------------------------------------------------
          ht |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0314815   .0019214   -16.38   0.000    -.0352474   -.0277157
```

Figure 5.2: Standardized cluster level residuals (intercept) from the compound symmetry model

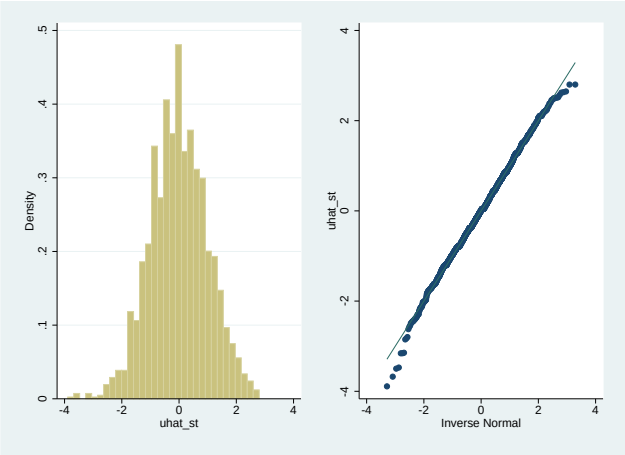
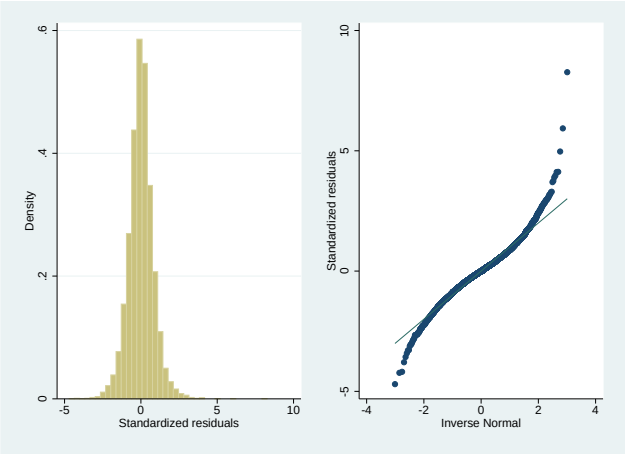


Figure 5.3: Standardized elementary level residuals from the compound symmetry model




```
       _cons |   162.4907   .1379564  1177.84   0.000     162.2203    162.7611
------------------------------------------------------------------------------


------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   5.990112   .0973412      5.802333    6.183968
-----------------------------+------------------------------------------------
                sd(Residual) |   1.424956   .0151655       1.39554    1.454992
------------------------------------------------------------------------------
```

We can see that this last specification leads to a larger residual variance than the compound symmetry, although the cluster level variances are similar..

We cannot use the restricted LRT to compare these two model fits because they have different specifications of the fixed effect part (we could use the LRT based on ML fits however (and this rejects the null hypothesis that the random intercept model with linear effect of age is as good as the model with 4 random intercepts) .

The compound symmetry model is very restrictive. It assumes that within subjects all variances are

equal and all covariances are equal. However when time is involved it is more likely that observations closer in time are more correlated than observations which are further away in time.

Random intercept and slope model

There are many ways to relax the assumption of compound symmetry of the covariance matrix. One way is to include random slopes in the model.

With longitudinal data, a simple random intercept and slope model is:

$$Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})t_i + e_{ij}$$

where, in this section, time t_i does not vary between clusters.

The inclusion of random slopes implies that the variance varies with time too. Collecting together the random terms for the ith observation in cluster j:

$$\begin{aligned} \mathrm{Var}(Y_{ij} \mid t_i) &= \mathrm{Cov}(u_{0j} + u_{1j}t_i + e_{ij}, u_{0j} + u_{1j}t_i + e_{ij}) \\ &= \sigma^2_{u00} + \sigma^2_{u11}t_i^2 + 2t_i\sigma_{u01} + \sigma^2_e \end{aligned}$$

It also implies that for different observations within the same cluster:

$$\begin{aligned} \mathrm{Cov}(Y_{1j}, Y_{2j} \mid t_1, t_2) &= \mathrm{Cov}(u_{0j} + u_{1j}t_1 + e_{1j}, u_{0j} + u_{1j}t_2 + e_{2j}) \\ &= \sigma^2_{u00} + \sigma^2_{u11}t_1t_2 + \sigma_{u01}(t_1 + t_2) \end{aligned}$$

that is, their covariance depends on the times of measurement. In contrast, as before, observations from different clusters are still uncorrelated:

$$\begin{aligned} \mathrm{Cov}(Y_{1j}, Y_{2j^*} \mid t_1, t_2) &= \mathrm{Cov}(u_{0j} + u_{1j}t_1 + e_{1j}, u_{0j^*} + u_{1j^*}t_2 + e_{2j^*}) \\ &= 0 \end{aligned}$$

Example: adult height measures (cont'd)

We now fit a model with just one random intercept but we include the effect of age allowing the slope to vary from woman to woman. We also centre age around its mean to improve the numerical stability of the estimates:

```
 . replace age=age-26
. mixed ht age || id: age, reml cov(unstructured) stddev
Mixed-effects REML regression                   Number of obs      =      6397
Group variable: id                              Number of groups   =      1980
                                                Obs per group: min =         1
                                                               avg =       3.2
                                                               max =         4
                                                Wald chi2(1)       =    192.92
Log restricted-likelihood = -15185.213          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
          ht |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0315799   .0022737   -13.89   0.000    -.0360362   -.0271236
       _cons |   162.4962   .1411298  1151.39   0.000     162.2196    162.7728
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                    sd(age) |   .0599296   .0026642      .0549289    .0653856
```

```
                  sd(_cons) |   6.158847   .1018531      5.962419    6.361746
              corr(age,_cons) |   -.280517   .0359417      -.349351   -.2086715
------------------------------+------------------------------------------------
                 sd(Residual) |    1.25921   .0167512      1.226803    1.292474
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) = 10924.00   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

We can test whether the model with random slopes gives a better fit than a model with just a random intercept and a constant slope using restricted LRT (and indeed it does) but we cannot test this model using a restricted LRT against the compound symmetry one because they have different specifications of the fixed effects part (we could use the LRT based on ML however).

Both models indicate that there is evidence of decreasing height with age. The random intercept and random slope model however has the advantage of allowing a non constant covariance between observations in the same cluster, because the covariance depends on the time gap between occasions.

We can calculate the covariance matrix implied by the random intercept and random slope model using the results of `mixed` with the default option `variance` (it is more precise this way):

```
. mixed, variance
<OUTPUT OMITTED>
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                     var(age) |   .0035916   .0003193      .0030172    .0042753
                   var(_cons) |    37.9314   1.254596      35.55045    40.47182
                cov(age,_cons) |  -.1035381   .0149202     -.1327811   -.0742951
-----------------------------+------------------------------------------------
                  var(Residual) |   1.585611   .0421865      1.505045    1.670488
------------------------------------------------------------------------------
```

The estimated covariance matrix for this model is:

$$\begin{aligned}\hat{\text{Cov}}(Y_{1j}, Y_{2j} \mid t_1, t_2) &= \sigma^2_{u00} + \sigma^2_{u11} t_1 t_2 + \sigma_{u01}(t_1 + t_2) \\ &= 37.93 \; + \; 0.004 \times t_1 \times t_2 \; - \; 0.104 \times (t_1 + t_2)\end{aligned}$$

and

$$\begin{aligned}\hat{\text{Var}}(Y_{1j} \mid t_1) &= \sigma^2_{u00} + \sigma^2_{u11} t_1^2 + 2\sigma_{u01}(t_1) + \sigma^2_e \\ &= 37.93 \; + \; 0.004 \times t_1^2 \; - \; 0.104 \times 2 \times t_1 + 1.59\end{aligned}$$

The covariance matrix from this last fit then is (where $t_1 = 0$,$t_2 = 10$,$t_3 = 17$, $t_4 = 27$):

$$\hat{\boldsymbol{\Sigma}}_u = \begin{pmatrix} 39.52 & 36.90 & 36.17 & 35.14 \\ 36.90 & 37.81 & 35.75 & 35.07 \\ 36.17 & 35.75 & 37.03 & 35.03 \\ 35.14 & 35.07 & 35.03 & 36.54 \end{pmatrix}$$

Thus the covariance between two measures decreases with increasing time interval between them, as one would expect in this context (note that the observations are not equidistant).

Checking now the residuals we see that the new model's elementary level standardized residuals are far less skewed (Figure 5.4 and Figure 5.5).

Figure 5.4: Standardized cluster level residuals (intercept and slope) from the random intercept and slope model

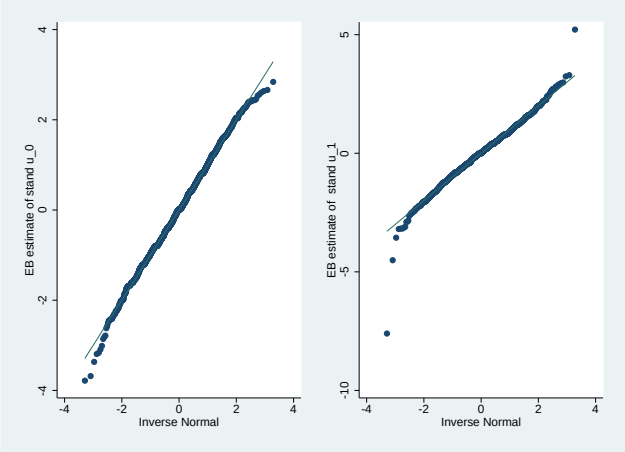
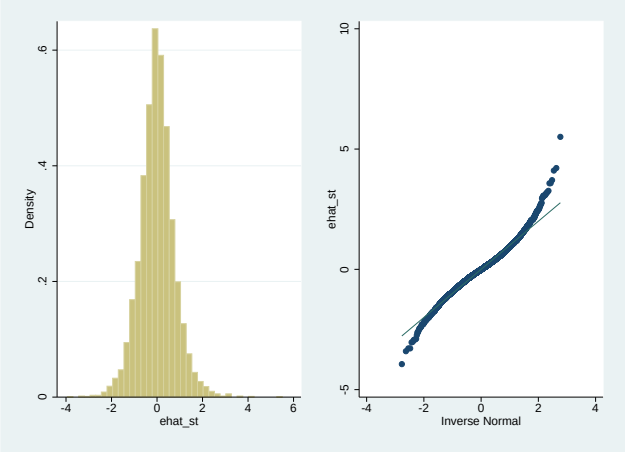


Figure 5.5: Standardized elementary level residuals from the random intercept and slope model



5.3 Variable occasions

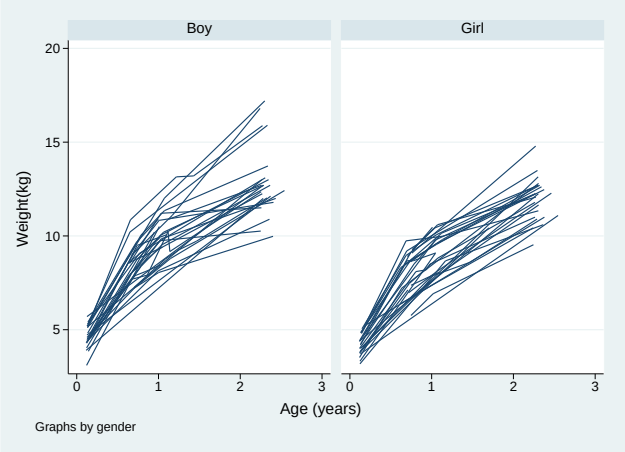
When the repeated observations are not taken at regular times, they are **unbalanced** and therefore we cannot express the data structure in terms of covariance matrices as above, nor produce simple mean profiles.

Unbalanced data can be thought of as balanced data affected by missingness. Thus great care is required when fitting linear mixed models to these data, because we are implicitly assuming that missingness is unrelated to the unmeasured Y.

5.4 Example: Infant growth data

We consider now a dataset which holds information on infants who were weighed on up to four occasions, at around 6 weeks of age and then at 8, 12 and 27 months. We are interested in modelling their growth trajectories. The observed data are shown in Figure 5.6 and indicate that boys may grow faster than girls.

Figure 5.6: Growth profiles of boys and girls in the infant growth data



The shape of the relationship between weight and age is not linear. Further, the observations were not all taken at the same time points.

Random intercept model

We could fit a model with a random intercept and fixed linear and quadratic effects for age, plus a term for being a girl:

$$Y_{ij} = (\beta_0 + u_{0j}) + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 \text{ girl}_j + e_{ij}$$

When we fit this in Stata we find that all the fixed effects are highly significant:

```
. use infant, clear
. gen age2=age^2
. gen girl=gender-1

. mixed weight age age2 girl || id:, reml stddev
Mixed-effects REML regression                   Number of obs      =       198
Group variable: id                              Number of groups   =        68
                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5
```

```
                                                Wald chi2(3)       =   2606.95
Log restricted-likelihood = -276.96721          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.817395    .290517    26.91   0.000     7.247992    8.386798
        age2 |  -1.705479   .1089108   -15.66   0.000     -1.91894   -1.492017
        girl |  -.7341374   .2359099    -3.11   0.002    -1.196512   -.2717625
       _cons |   3.799253   .2121041    17.91   0.000     3.383537     4.21497
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   .8594507   .0950292      .6919965    1.067427
-----------------------------+------------------------------------------------
                sd(Residual) |   .7394062   .0457837      .6549032    .8348128
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    67.58 Prob >= chibar2 = 0.0000
```

In this example, based on a relatively small dataset, the estimates of the cluster level random effects obtained using ML instead of REML are slightly different:

```
. mixed weight age age2 girl|| id:, ml stddev
<OMITTED OUTPUT>
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   .8443384   .0923907      .6813573    1.046305
-----------------------------+------------------------------------------------
                sd(Residual) |   .7339017   .0451127      .6506012    .8278678
------------------------------------------------------------------------------
```

Random intercept and slope model

We extend the model to include a random slope for the linear term for age:

```
. mixed weight age age2 girl|| id: age, reml cov(unstructured) stddev
Mixed-effects REML regression                   Number of obs      =       198
Group variable: id                              Number of groups   =        68
                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5
                                                Wald chi2(3)       =   1935.62
Log restricted-likelihood = -258.97501          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.698437   .2398533    32.10   0.000     7.228333     8.16854
        age2 |  -1.657734   .0885945   -18.71   0.000    -1.831376   -1.484092
        girl |  -.5983843    .199975    -2.99   0.003    -.9903281   -.2064406
       _cons |   3.795497   .1681453    22.57   0.000     3.465939    4.125056
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
```

```
------------------------------+------------------------------------------------
id: Unstructured              |
                      sd(age) |   .5177674   .0884366      .3704643    .7236408
                    sd(_cons) |   .6148786   .1302125      .4060045    .9312104
               corr(age,_cons) |   .1354213   .3143137     -.4552427    .6433159
------------------------------+------------------------------------------------
                 sd(Residual) |   .5741197   .0499052      .4841858    .6807581
------------------------------------------------------------------------------
```

There is a clear decrease in residual SD and SD of the intercept when the slopes are allowed to vary across children (conditional on sex). Further the cluster and elementary level standardized residuals are well behaved (Figure 5.7 and Figure 5.8).

Figure 5.7: Standardized cluster level residuals (intercept and slope) from the random intercept and slope model

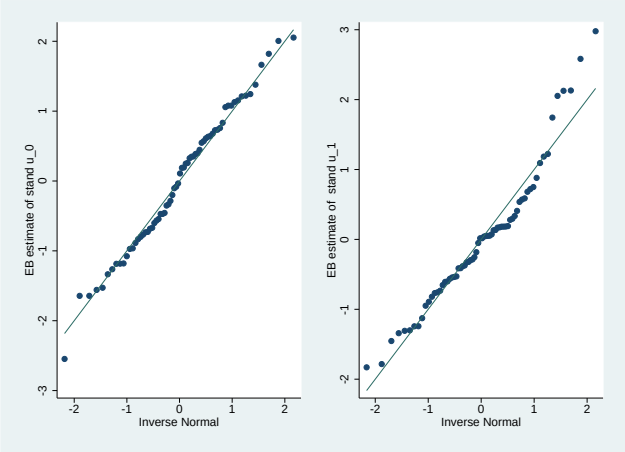
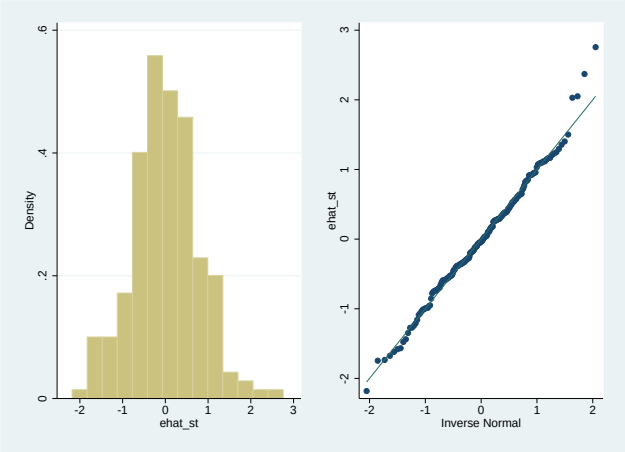


Figure 5.8: Standardized elementary level residuals from the random intercept and slope model



5.5 Predicting trajectories

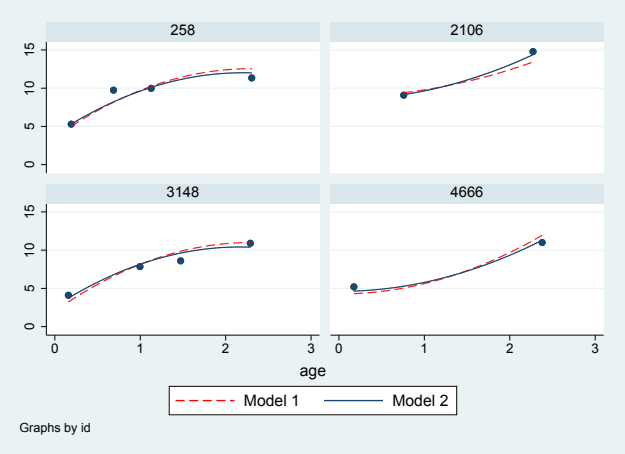
We compare the trajectories predicted by the random intercept and by the random intercept and random slope model. The predictions are derived from the estimates of the fixed effects plus the EB 'estimates'

of the cluster level errors. We do this in Stata for the first model with the command `predict` after fitting the model:

```
. qui mixed weight age age2 girl|| id: , reml cov(unstructured)
. predict traj1,fitted
```

Using the random intercept or the random intercept and slope model leads to similar predictions, see for example the results for four girls in Figure 5.9.

Figure 5.9: Observed weight and predicted growth profiles of four girls in the infant growth data



Chapter 6

Longitudinal data II

In this session we consider alternative ways of formulating the same hierarchical models and introduce a general formulation of the mixed model.

6.1 Marginal structures

The mixed models considered so far are sometimes referred to as conditional models, because the dependent variable is modelled conditional on the cluster level residuals. The different conditional models imply different marginal relations.

Random intercept model

With longitudinal data, where the data may be balanced or unbalanced, a simple random intercept model is specified as:

$$Y_{ij} = (\beta_0 + u_{0j}) + \beta_1 t_{ij} + e_{ij}$$

This implies that the conditional relation is

$$Y_{ij}|t_{ij}, u_{0j} \sim N(\beta_0 + \beta_1 t_{ij} + u_{0j}, \sigma_e^2)$$

with $u_{0j}|t_{ij} \sim N(0, \sigma_u^2)$. Also note that $\mathrm{Var}(Y_{ij}|t_{ij}, u_{0j}) = \sigma_e^2$, i.e. the conditional variance of Y_{ij}, conditional on t and u_{0j} depends only on σ_e^2, and that the conditional covariance between different units within the same cluster, conditional on u_j, is zero: $\mathrm{Cov}(Y_{ij}, Y_{i^*j}|t_{ij}, t_{i^*j}, u_j) = 0$.

These **conditional** statements differ from the marginal ones. The **marginal** (with respect to u_j) distribution of Y_{ij} is:

$$E(Y_{ij}|t_{ij}) = \beta_0 + \beta_1 t_{ij}$$

with $\mathrm{Var}(Y_{ij}|t_{ij}) = \sigma_u^2 + \sigma_e^2$ and $\mathrm{Cov}(Y_{ij}, Y_{i^*j}|t_{ij}, t_{i^*j}) = \sigma_u^2$.

For example the two elementary units within cluster j, Y_{1j} and Y_{2j}, conditional on t_{1j} and t_{2j}, follow a marginal bivariate normal distribution with mean

$$\begin{bmatrix} \beta_0 + \beta_1 t_{1j} \\ \beta_0 + \beta_1 t_{2j} \end{bmatrix}$$

and (compound symmetry) covariance matrix:

$$\begin{bmatrix} \sigma_u^2 + \sigma_e^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 \end{bmatrix}$$

Random intercept and random slope model

Recalling the model:

$$Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})t_{ij} + e_{ij},$$

which can also be written as

$$Y_{ij} = (\beta_0 + \beta_1 t_{ij}) + (u_{0j} + u_{1j}t_{ij}) + e_{ij},$$

the **conditional** relations are:

$$Y_{ij}|t_{ij}, u_{0j}, u_{1j} \sim N(\beta_0 + \beta_1 t_{ij} + u_{0j} + t_{ij}u_{1j}, \sigma_e^2)$$

with $\mathbf{u}_j|t_{ij} \sim N(0, \mathbf{\Sigma}_u)$ where

$$\mathbf{\Sigma}_u = \begin{bmatrix} \sigma_{u00}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u11}^2 \end{bmatrix}$$

and
$\mathrm{Cov}(Y_{ij}, Y_{i^*j}|t_{ij}, t_{i^*j}, u_{0j}, u_{1j}) = 0.$

The **marginal** (with respect to u_j) distribution of Y_{ij} is instead specified as:

$$E(Y_{ij}|t_{ij}) = \beta_0 + \beta_1 t_{ij}$$

with

$$\begin{aligned}
\mathrm{Var}(Y_{ij} \mid t_{ij}) &= \sigma_{u00}^2 + 2\sigma_{u01}t_{ij} + \sigma_{u11}^2 t_{ij}^2 + \sigma_e^2 \\
\mathrm{Cov}(Y_{ij}, Y_{i^*j} \mid t_{ij}, t_{i^*j}) &= \mathrm{Cov}(u_{0j} + u_{1j}t_{ij} + e_{ij}, u_{0j} + u_{1j}t_{i^*j} + e_{i^*j}) \\
&= \sigma_{u00}^2 + \sigma_{u01}(t_{ij} + t_{i^*j}) + \sigma_{u11}^2 t_{ij}t_{i^*j} \quad (for\ i \neq i^*) \\
\mathrm{Cov}(Y_{ij}, Y_{i^*j^*} \mid t_{ij}, t_{i^*j^*}) &= \mathrm{Cov}(u_{0j} + u_{1j}t_{ij} + e_{ij}, u_{0j^*} + u_{1j^*}t_{i^*j^*} + e_{i^*j^*}) \\
&= 0 \quad (for\ j \neq j^*)
\end{aligned}$$

Covariance and correlation between Y_{ij} from the same level 2 unit are non-zero and depend on t_{ij}.

Random coefficient models and unstructured variance models

Let $i = 1, \ldots, I$ index individuals and $j = 1, \ldots, 3$ index three scheduled observation times on each individual, at times t_1, t_2, t_3. We have a dependent variable Y_{ij}.

A random intercept and slope model is:

$$\begin{aligned}
Y_{i,1} &= (\beta_0 + u_{0,i}) + (\beta_1 + u_{1,i})t_1 + \epsilon_{i,1}, \\
Y_{i,2} &= (\beta_0 + u_{0,i}) + (\beta_1 + u_{1,i})t_2 + \epsilon_{i,2}, \\
Y_{i,3} &= (\beta_0 + u_{0,i}) + (\beta_1 + u_{1,i})t_3 + \epsilon_{i,3},
\end{aligned}$$

At level 2, the random intercepts and slopes have a joint Normal distribution,

$$\begin{pmatrix} u_{0,i} \\ u_{1,i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{u00}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u11}^2 \end{pmatrix} \right]$$

At level 1, conditional on cluster mean and fixed effects, the three observations for individual i are independent and identically distributed,

$$\begin{pmatrix} e_{i,1} \\ e_{i,2} \\ e_{i,3} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_e^2 & 0 \\ 0 & 0 & \sigma_e^2 \end{pmatrix} \right]$$

Remember, Σ is a 2-by-2 symmetric matrix with three terms — hence it is an unstructured matrix *for the random effects at level 2.*

All three observations on each individual have a 3-by-3 covariance matrix, which has six terms (three variances and three covariances). In this random intercepts and slopes model these six terms are functions of the four variance terms in the above model — i.e., $\sigma^2_{u_{00}}, \sigma_{u_{01}}, \sigma^2_{u_{11}}, \sigma^2_e$, and the times, t_i. In the lectures we will show how to calculate these functions ; but this detail is not important here. The point is that this is a structured model for the variance, because rather than estimating all six terms, we define the variance model as a function (structure) of four variance terms and time.

For an unstructured, marginal, covariance model, we simply remove the random effects, and allow a unstructured covariance matrix for the level-1 residuals relative to the mean structure:

$$\begin{aligned} Y_{i,1} &= \beta_0 + \beta_1 t_1 + \epsilon_{i,1}, \\ Y_{i,2} &= \beta_0 + \beta_1 t_2 + \epsilon_{i,2}, \\ Y_{i,3} &= \beta_0 + \beta_1 t_3 + \epsilon_{i,3}, \end{aligned}$$

This unstructured, marginal, covariance model has the form,

$$\begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \epsilon_{i,3} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\epsilon_1} & \sigma_{\epsilon_{12}} & \sigma_{\epsilon_{13}} \\ \sigma_{\epsilon_{12}} & \sigma^2_{\epsilon_2} & \sigma_{\epsilon_{13}} \\ \sigma_{\epsilon_{13}} & \sigma_{\epsilon_{23}} & \sigma^2_{\epsilon_3} \end{pmatrix} \right]$$

Again notice that this has six distinct variance terms and there are no constraints on the nature of those terms.

6.2 Matrix notation

If the data are **balanced**, we can write hierarchical models in matrix notation as follows, for every cluster j. Define the vectors $\mathbf{Y}_j$ and $\mathbf{e}_j$ as

$$\mathbf{Y}_j = \begin{bmatrix} Y_{1j} \\ Y_{2j} \\ \cdots \\ \cdots \\ Y_{nj} \end{bmatrix}$$

$$\mathbf{e}_j = \begin{bmatrix} e_{1j} \\ e_{2j} \\ \cdots \\ \cdots \\ e_{nj} \end{bmatrix}$$

Considering only 3 observation times for simplicity, t_1, t_2, t_3, let the matrix $\mathbf{T}$ be

$$\mathbf{T} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \end{bmatrix}$$

Also let

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

and

$$\mathbf{u}_j = \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}$$

Then the random intercept and random slope model can be written as, dropping the subfix j for simplicity, with $\mathbf{Y}$ representing the vector of observations for cluster j:

$$\mathbf{Y} = \mathbf{T}\beta + \mathbf{T}\mathbf{u} + \mathbf{e}$$

where $\mathbf{u} \sim N(\mathbf{0}, \mathbf{\Sigma}_u)$ and $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$ This model then implies that, for every j, Var($\mathbf{Y}$) is:

$$\text{Var}(\mathbf{Y}) \quad = \quad \mathbf{T}\mathbf{\Sigma}_u\mathbf{T}^T + \sigma_e^2\mathbf{I}$$

6.3 General Formulation of the Mixed Model

A general specification of these models (which includes random intercept and random intercept and slope models among others) is then:

$$\mathbf{Y} \quad = \quad \mathbf{T}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{Z}$ is a submatrix of $\mathbf{T}$ specifying random coefficient terms. Its variance is:

$$\text{Var}(\mathbf{Y}) \quad = \quad \mathbf{Z}\mathbf{\Sigma}_u\mathbf{Z}^T + \mathbf{\Sigma}_e$$

Under normality,

$$\mathbf{Y} \sim N(\mathbf{T}\beta, \mathbf{Z}\mathbf{\Sigma}_u\mathbf{Z}^T + \mathbf{\Sigma}_e).$$

So this is simply a **multivariate** (greater than 1 response) linear model, with a very particular structure for the covariance matrix.

Very commonly (but not necessarily) it is assumed that

$$\mathbf{\Sigma}_e = \sigma_e^2\mathbf{I}.$$

6.4 Alternative specification

Instead of specifying these models using the random effects notation and language we could therefore simply specify a model for the marginal expectations and covariance matrices. Models in which a structure (pattern) for the marginal covariance is assumed are referred to as **covariance pattern models**.

The marginal covariance structures that we have met so far are:

- **Compound symmetric structure**
  With clusters of size three:

$$\begin{bmatrix} \sigma_u^2 + \sigma_e^2 & & \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 \end{bmatrix}$$

  We have met this before; it also goes with the name of **exchangeable structure**.

- **Random coefficient (RC) structure**
  Again we have seen this before; with 3 time points at 0, 1 and 2 this is:

$$\begin{bmatrix} \sigma_{u00}^2 + \sigma_e^2 & & \\ \sigma_{u00}^2 + \sigma_{u01} & \sigma_{u00}^2 + 2\sigma_{u01} + \sigma_{u11}^2 + \sigma_e^2 & \\ \sigma_{u00}^2 + 2\sigma_{u01} & \sigma_{u00}^2 + 3\sigma_{u01} + 2\sigma_{u11}^2 & \sigma_{u00}^2 + 4\sigma_{u01} + 4\sigma_{u11}^2 + \sigma_e^2 \end{bmatrix}$$

In addition we could have other covariance structures, for example:

- **Autoregressive (AR) structure**

$$\frac{\sigma_\eta^2}{1-\alpha^2}\begin{bmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{bmatrix}$$

- **Unstructured**

$$\begin{bmatrix} \sigma_{11} & & \\ \sigma_{21} & \sigma_{22} & \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

An autoregressive structure could arise for example when the residual terms depend on each other as in:

$$e_{ij} = \alpha e_{i\ j-1} + \eta_{ij}$$

with $\eta_{ij} \sim N(0, \sigma^2_\eta)$. This is suitable only for balanced data with equidistant observations.

Autoregressive models of order 1 can be fitted in Stata using the `mixed` command:

```
. mixed depvar indepvars ...|| levelvar:, residuals(ar 1, t(time)) ...
```

More flexible structures are available in `gllamm` and specialized multivariate modelling software such as MPlus.

Finally note that an **independent structure**, such as this,

$$\sigma^2 \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0 & 1 \end{bmatrix}$$

would imply independent observations (i.e. would not require any complex modelling or correction for dependency).

6.5 Comments

1. *Random coefficient models versus covariance pattern models*

   - Random coefficient models are appealing in that they build up the variability using models for individual behaviour: unfortunately they can be a bad fit if the behaviour is not amenable to simple modelling.
   - Pattern-based models (e.g. AR(1)) may also be used in combination with random-coefficient models. However they cannot be used when times of measurement are not common.

2. *Formal comparison of fit of covariance structures*

   - *The likelihood ratio test (LRT)*
     Provided that the mean structure (fixed effects) does not change, likelihood ratio tests can be used to compare the fit of **nested** structures. As the unstructured is the most general model, all structures are nested within it. Comparison with this provides a **goodness of fit** test. This may have low power though.
   - *Other measures of fit: information criteria*
     When covariance structures are not nested we need another way of assessing fit.

     We could instead use a general measure like Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). In different ways, these penalize the fit, as measured by the likelihood, with the number of parameters required for the fit, for example AIC is -2(logl - no. parameters in the model). BIC has a greater penalty on the number of parameters and will result in a more parsimonious model. In practice, whichever measure we decide to use, better fitting models will have smaller value. There is no guarantee however that these criteria will agree, they may well not; nor is it necessary that they will lead to an "appropriate" structure.

6.6 Unbalanced data

- If the data are unbalanced in the sense that there are no common times of measurement, there is no unstructured covariance matrix for Y, and hence patterned matrices can't be used.
- Random effects and random coefficient models can be used with unbalanced data and therefore even when data are unbalanced we know what is the implied covariance structure.
- Thus direct assessment of fit through LRTs of nested covariance structure models are not feasible, while comparison of fit of alternative models become important tools for assessing fit.

6.7 Crossed effects

So far we have treated all elementary units as nested within the cluster units, e.g. children clustered in schools. However there may be some factor that makes some of the elementary units also affected equally by a common factor, e.g. GCSE results marked by less strict examiners. Then there is something affecting the elementary units that works *across* cluster levels. Another example not covered in the course comes from cross-over clinical trials where **period** is such a crossing factor.

We can treat this additional clustering factor as a fixed effect and include it in the specification of the fixed part of the mixed effects model. Alternatively they can be treated as random (but this requires introducing another level of hierarchy).

Chapter 7

Longitudinal data III

In this session we discuss some extensions of the hierarchical models we have encountered so far and revisit REML.

7.1 Level 1 heterogeneity

So far we have assumed that the level 1 and level 2 error variances are constant. However we could allow the level 1 variances to vary according to an explanatory variable and therefore we may be able to get a closer representation of the data. This takes the name of **complex level 1 variation**.

Consider again the infant growth data in which weight is measured from birth to around 3 years. The level 1 variance (as well as the mean) may be different for the two sexes. In other words: $\sigma_e = f($ gender$)$, where the function f is a log function to guarantee that the SD is not negative. For example:

$$log(\sigma_e) = \delta_1 I_{\text{gender=boy}} + \delta_2 I_{\text{gender=girl}}$$

.

This model can be fitted in Stata with:

```
. mixed weight age age2 girl|| id: age , reml cov(unstructured) stddev ///
       residuals(independent, by(girl))
Mixed-effects REML regression                   Number of obs      =       198
Group variable: id                              Number of groups   =        68
                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5
                                                Wald chi2(3)       =   2050.35
Log restricted-likelihood = -257.53016          Prob > chi2        =    0.0000
------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.629728   .2343754    32.55   0.000     7.170361    8.089096
        age2 |  -1.635069   .0865352   -18.89   0.000    -1.804675   -1.465463
        girl |  -.6043802   .2045169    -2.96   0.003    -1.005226   -.2035343
       _cons |   3.829412   .1755813    21.81   0.000     3.485279    4.173545
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                   var(age) |   .2434063   .0890538      .1188248    .4986051
                 var(_cons) |   .4159272   .1545263      .2008061     .861505
```

```
              cov(age,_cons) |   .0410678   .0855177     -.1265439    .2086794
-----------------------------+------------------------------------------------
Residual: Independent,       |
    by girl                  |
                  0: var(e) |   .4124363   .0906558      .2680761    .6345351
                  1: var(e) |   .2458869   .0579841      .1548846    .3903574
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(4) =   106.45   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

These are REML estimates and we can see that the variance in weights is greater for boys (0.41) than for girls (0.25). We use a LRT to compare the new sex specific variance estimates with the estimated σ_e^2 found when assuming there was no heterogeneity among the sexes (which was equal to 0.33). Comparing the model that assumes heterogeneity in variance between sexes with the one that assumes homogeneity does not indicate that the more complex model gives a better fit (P=0.09).

7.2 Level 2 heterogeneity

We can also allow for differential level 2 structures. In order to investigate this in Stata, we have to 'pretend' that we have a 3 level model, where level 2 is split between boys and girls, each with its own random intercept and random slope:

```
.  generate  boy=(girl==0)
.  generate  age_girl  =  age*girl
.  generate  age_boy  =  age*boy

. mixed weight age age2 girl ///
|| id: girl age_girl , nocons cov(unstructured) ///
|| id: boy age_boy ,nocons cov(unstructured)   reml stddev

Mixed-effects REML regression                   Number of obs      =       198
Group variable: id                              Number of groups   =        68

                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5


                                                Wald chi2(3)       =   2353.87
Log restricted-likelihood = -254.96233          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.614975   .2351766    32.38   0.000     7.154037    8.075912
        age2 |  -1.646424   .0874516   -18.83   0.000    -1.817826   -1.475022
        girl |  -.6088238   .2031867    -3.00   0.003    -1.007062   -.2105852
       _cons |   3.820436   .1601151    23.86   0.000     3.506617    4.134256
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
```

```
                    sd(girl) |   .7171004    .164404      .4575426    1.123902
                 sd(age_girl) |   .2305558   .1426821      .0685489    .7754467
          corr(girl,age_girl) |   .3348284   .7263757     -.8497283    .9604504
------------------------------+------------------------------------------------
id: Unstructured              |
                     sd(boy) |    .556704   .1821813      .2931358    1.057255
                  sd(age_boy) |    .694514   .1359894      .4731637    1.019414
            corr(boy,age_boy) |   .0367294   .3886928     -.6206779    .6638125
------------------------------+------------------------------------------------
                sd(Residual) |   .5694751   .0493961      .4804428    .6750062
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(6) =   111.59   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

The results show some differences by sex, with girls being generally more variable at the intercept but having more similar slopes. Testing this model against the one that assumes a common level 2 structure we find borderline evidence that the more complex model gives a better fit:

```
. est store m2
. qui mixed weight age age2 girl|| id: age , reml cov(unstructured) stddev
. est store m1

. lrtest m2 m1

Likelihood-ratio test                                 LR chi2(3)  =      8.03
(Assumption: m1 nested in m2)                         Prob > chi2 =    0.0455

Note: The reported degrees of freedom assumes the null hypothesis
       is not on the boundary of the parameter space.
      If this is not true, then the reported test is conservative.
Note: LR tests based on REML are valid only when the fixed-effects
      specification is identical for both models.
```

7.3 Strategy of analysis

Clarify the questions that you are trying to address in your analysis. Is the focus:

(a) the conditional (within cluster) effect of a covariate?

(b) the variability between and within the clusters?

(c) the marginal effect of a covariate?

If (a) & (b), consider mixed effects models. If (a) and the covariate is not defined at the cluster level you may want to use fixed effects models. If (c) you may consider using a marginal model and fitting it using GEEs.

7.3.1 Selection steps for a linear mixed model (see Verbeke and Molenberghs, 2000)

Fitting a linear mixed model implies that an appropriate mean structure as well as a covariance structure is specified. The covariance structure explains the random variation in the data that is not explained by the (assumed) mean structure. Hence the two parts of the model (mean and covariance structures) are highly dependent. Moreover, an appropriate covariance model is essential to obtain valid inferences for the parameters in the mean structure.

Step 1

Since the covariance structure models all the variability in the data which cannot be explained by the fixed effects, we start by first removing all systematic trends. Starting from an over-elaborated model for the mean structure we make sure that the specification of the fixed effects (*the part of the model that refers to the population mean*) is sufficiently general to guarantee that the random effects component is not affected by mis-specification of the fixed effects.

This can be achieved by `'Saturating'` the fixed effects component of the model (i.e. by including all potential explanatory variables, plus quadratic and other non-linear terms if continuous, plus their interactions) before comparing alternative specifications of the random components.

This will lead to over-parametrization of the mean structure of the model in order to get consistent estimators of the covariance structure in the following steps.

Among these alternative structures we might also consider (if suitable) structures that allow heterogeneities at each level (as discussed in the first part of this session). Note that one should not include a polynomial random effect (e.g. a random effect for the quadratic effect of time) unless all hierarchical inferior terms are also included (e.g. a random effect for the linear effect of time).

We can then examine residuals, outliers, predictions versus observed values, etc. derived from each model that has a 'saturated' mean structure and alternative random effects structure.

Step 2

Once the random effects component of the model has been selected:

- Simplify the fixed effect component of the model using Wald tests (if models fitted using REML) or LRTs (if fitted using ML). Start by examining interactions and highest order polynomial terms first and then work down through less complex terms. As for the random effects, one should only include a polynomial effect (e.g. a quadratic effect of time) if all hierarchical inferior terms are also included (e.g. a linear effect of time), similarly for interaction terms.
- Examine residuals, outliers, predictions vs observed values, etc.
- Check the predicted trajectories versus the observed ones
- Interpret the results!

Comment on baseline covariates In some contexts the Y variable is measured at baseline, say in a clinical trial setting before any treatments have been given. In this case there are two main option, the baseline measurement can be taken as an explanatory variable or it can be taken as the first dependent measurement. The choice will depend on the specific context and the aim of the analysis.

7.4 More on REML

REML is a genuine maximum likelihood procedure, but applied to a linear transformation of the data that removes the fixed effects (β_0, β_1, etc).

The steps are:

- Find a linear combination of the Y, e.g. $\mathbf{u}'\mathbf{Y}$ such that $E(\mathbf{u}'\mathbf{Y}) = 0$.

  One such transformation is $\mathbf{u} = \mathbf{I} - \mathbf{P}$ where $\mathbf{P}$ is the 'projection' matrix met in linear regression[1].
- The variance parameters are estimated from the likelihood for these new observations $\mathbf{u}'\mathbf{Y}$. Note that
  - this likelihood is a function of the variance components *only*.
  - *Any* choice of $\mathbf{u}$ satisfying this produces the same REML likelihood function.
- The fixed effects are estimated in a second step via GLS (with weights given by estimated variance parameters).

REML is typically the approach of choice when fitting these models.

[1] $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (it is used to find $\hat{\mathbf{Y}}$: $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$)

- For example, the usual regression estimator of the residual variance σ^2 is actually a REML, not an ML, estimator.
- REML usually produces estimates of the variance components with smaller bias than ML. The difference diminishes as the number of subjects increases.
- There is no general rule guaranteeing the superiority of REML however.
- Important point: when the fixed effects part of the model changes, the "data" underlying the REML likelihood function also changes, hence we should not use LRT from REML.
- Remember that the LRT compares the probability of observing the data for two different hypothesized models. If the data change (because we are using REML) then we are fitting the two models to different data.

Chapter 8

Generalized estimating equations

In this session we consider marginal models that require making only "working assumptions" about the joint distribution of the outcomes. By "working assumptions" we mean assumptions that are unlikely to be true but that, under certain conditions, even if wrong do not invalidate the inferences we draw. These models are useful when the error structure is not of interest itself, but we want to take account of correlation between observations in the same cluster when estimating fixed model parameters.

8.1 Introduction

Liang and Zeger (Biometrika, 1986) proposed an approach that extends generalized linear models (GLMs) to a setting with correlated observations within clusters. It is called **generalized estimating equations (GEEs)**. This approach is particularly useful in the analysis of non-Gaussian outcomes, for which estimation of mixed effects models can be difficult. We will only introduce GEEs here: more will be discussed in *Advanced Statistical Methods* in Term 3.

GEEs characterize the marginal expectation of the outcome as a function of covariates using a generalized linear model for which a (possibly incorrect) correlation matrix is assumed. The advantage of this approach is that the estimating equations give consistent estimates of the regression parameters under weak assumptions about the joint distribution. Consistent estimates of their variances are achieved using a robust variance estimator. [Aside: Recall that a consistent estimator is one that converges (in probability) to the true value of the parameter as the number of observations increases]

8.2 Notation

For simplicity consider a balanced dataset (*i.e.* with the same number n of observations per cluster) and let $\boldsymbol{Y}_j = (Y_{1j}, ..., Y_{nj})^T$ be the $(n \times 1)$ vector of outcome values for cluster j $(j = 1, 2, \ldots, J)$ and $\boldsymbol{X}_j$ be the $(n \times p)$ matrix of covariate values for the j-th cluster,

$$\begin{pmatrix} X_{11j} & X_{21j} & \cdot & \cdot & X_{p1j} \\ X_{12j} & X_{22j} & \cdot & \cdot & X_{p2j} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{1nj} & X_{2nj} & \cdot & \cdot & X_{pnj} \end{pmatrix}.$$

Note that this is a generalization of the notation used in section 6.3 where the explanatory variables were the intercept (equivalent to $X_{1ij} = 1$ for all i and j), and time ($X_{1ij} = t_i$ for all j).

8.3 Links to GLMs

In a generalized linear model (GLM) for $\boldsymbol{Y}_j$ expressed as a function of $\boldsymbol{X}_j$, with identity link and Gaussian distribution, the expectation of $\boldsymbol{Y}_j$ is an $(n \times 1)$ vector

$$\boldsymbol{\mu}_j = \boldsymbol{X}_j \boldsymbol{\beta},$$

with $\boldsymbol{\beta}$ a $(p \times 1)$ vector of regression coefficients. Assume the true variance-covariance of $\boldsymbol{Y}_j$, the $(n \times n)$ matrix $\boldsymbol{\Omega}_j$, is equal to $\boldsymbol{V}_j$. Then the score equation for $\boldsymbol{\beta}$ is:

$$\boldsymbol{U}(\boldsymbol{\beta}) \quad = \quad \sum_{j=1}^{J} \boldsymbol{D}_j^T \boldsymbol{V}_j^{-1} \{\boldsymbol{Y}_j - \boldsymbol{\mu}_j(\boldsymbol{\beta})\}$$

where $\boldsymbol{D}_j$ is an $(n \times p)$ matrix with elements $\{\frac{\partial \mu_j}{\partial \beta_k}\}$. For a Gaussian model with identity link $\boldsymbol{D}_j = \boldsymbol{X}_j$ and hence

$$\boldsymbol{U}(\boldsymbol{\beta}) \quad = \quad \sum_{j=1}^{J} \boldsymbol{X}_j^T \boldsymbol{V}_j^{-1} \{\boldsymbol{Y}_j - \boldsymbol{\mu}_j(\boldsymbol{\beta})\}$$

Solution of these score equations leads to ML estimation of $\boldsymbol{\beta}$. However in a GEEs setting we are not assuming that $\boldsymbol{Y}_j$ are jointly normally distributed but only that the means and variances are characterized as in this GLM.

8.4 Independent estimating equations (IEEs)

In settings where the outcomes, conditional on the explanatory variables, are independent, the true variance-covariances matrix $\boldsymbol{\Omega}_j$ is a diagonal matrix $\boldsymbol{V}_j$ with diagonal elements v_{ij}, possibly identical across clusters: $v_{ij} = v_i$.

If we stack $\boldsymbol{X}_j$, $\boldsymbol{Y}_j$ and $\boldsymbol{\mu}$ and make a block diagonal matrix $\boldsymbol{V}$ of these assumed $\boldsymbol{V}_j$ matrices, we can write the score function as:

$$\boldsymbol{U}(\boldsymbol{\beta}) \quad = \quad \boldsymbol{X}^T \boldsymbol{V}^{-1} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\} \tag{8.1}$$

This expression has to be solved iteratively (*e.g.* by iterative weighted least squares (IWLS)). It leads to the vector estimator $\hat{\boldsymbol{\beta}}$ which is consistent and asymptotically has a normal distribution with covariance matrix $\mathbf{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1}$, if the assumption regarding $\boldsymbol{\Omega}_j = \boldsymbol{V}_j$ is correct. Note that even if the data are correlated the estimator of $\boldsymbol{\beta}$ derived from the score equations above (which assume $\boldsymbol{V}_j$ is a diagonal matrix) is still unbiased (though not efficient).

[Aside: Recall that a more efficient estimator is one that arrives at the true estimate of a set of parameters with a smaller number of observations].

8.5 Working correlation matrices

Assuming the true cluster variance-covariance matrix $\boldsymbol{\Omega}_j$ is a diagonal matrix could be a useful option, given that, although unlikely to be met, estimation of $\boldsymbol{\beta}$ via GEEs is unbiased.

However, Liang and Zeger showed that modelling the correlations, as opposed to assuming that they are zero, may boost efficiency of $\hat{\boldsymbol{\beta}}_j$. We could therefore consider other specifications of the variance-covariance matrix, because whatever we choose might be incorrect but the estimator of $\boldsymbol{\beta}$ would still be unbiased (Liang and Zeger, 1986).

An alternative specification for $\boldsymbol{V}_j$ is:

$$\boldsymbol{V}_j(\boldsymbol{\alpha}) = \boldsymbol{A}_j^{\frac{1}{2}} \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{A}_j^{\frac{1}{2}}$$

where $\boldsymbol{A}_j$ is a diagonal matrix with elements v_{ij} and $\boldsymbol{R}(\boldsymbol{\alpha})$ is a correlation matrix of our choice, a so-called "working matrix", which is indexed by a vector of parameters $\boldsymbol{\alpha}$.

Several structures may be appropriate for modelling this matrix.

- The independent structure. This is defined as

$$\begin{aligned} R_{ii'}(\alpha) &= 1 \quad \text{if} \quad i = i' \\ &= 0 \text{ otherwise} \end{aligned}$$

(Note: no new parameters)

- The exchangeable structure. This is defined as

$$\begin{aligned} R_{ii'}(\alpha) &= 1 \quad \text{if} \quad i = i' \\ &= \alpha \text{ otherwise} \end{aligned}$$

(Note: 1 new parameter)

- The unstructured matrix. This is a symmetric matrix defined as

$$\begin{aligned} R_{ii'}(\alpha) &= 1 \quad \text{if} \quad i = i' \\ &= \alpha_{ii'} \text{ with } \alpha_{ii'} = \alpha_{i'i} \quad \text{otherwise} \end{aligned}$$

(Note: $n \times (n-1)/2$ new parameters)

Once a working matrix is selected, estimation is achieved using the estimating equations (8.1) but with $\boldsymbol{V}(\boldsymbol{\alpha})_j$ instead of $\boldsymbol{V}_j$.

If the mean model is correct, the vector $\hat{\boldsymbol{\beta}}$ is consistent even if the correlation structure is misspecified. The variance of $\hat{\boldsymbol{\beta}}$ however is only consistently estimated if the assumed variance-covariance matrix is correctly specified. Robust estimation of the variance of $\hat{\boldsymbol{\beta}}$ is therefore preferable since it is consistent when only the mean is correctly specified.

8.6 GEEs in Stata: `xtgee`

The syntax for `xtgee`:

```
xtgee depvar varlist, family(family) link(link) corr(correlation)
           i(cluster) t(occasion) robust
```

- `Family`: binomial, gaussian, gamma, etc.
- `Link`: identity, cloglog, log, etc.
- `Correlation`:

OPTION	WORKING MATRIX
`exchangeable`	exchangeable
`independent`	independent
`unstructured`	unstructured
`fixed` *matname*	user-specified

- `i()`: defines the cluster identifier
- `t()`: defines the within-cluster counter, *e.g.* the timing of the observations

Note that one could set the `i()` and `t()` variables in advance with

```
.  xtset idvar timevar
```

This is particularly useful if fitting alternative models.

8.7 Example

Let us revisit the infant growth data. In order to treat the growth observations as having been collected at fixed occasions (which they nearly were) we use rounded age in months as the within-cluster counter: this variable is called `age_round` (in the dataset called `growth_gee.dta`).

In chapter 5 we found that weight increased non-linearly with age in these children. So we allow for a quadratic term in `age_round` here too.

Estimating the marginal effects of age and gender assuming an independent working matrix and using robust estimation of the standard errors we find:

```
. xtset id age_round
. xtgee weight age_round age_r2 girl,  cor(indep) vce(robust)

GEE population-averaged model                   Number of obs      =       197
Group variable:                         id      Number of groups   =        68
Link:                             identity      Obs per group: min =         1
Family:                           Gaussian                     avg =       2.9
Correlation:                   independent                     max =         5
                                               Wald chi2(3)       =   1477.79
Scale parameter:                  1.311073     Prob > chi2        =    0.0000

Pearson chi2(197):                  258.28     Deviance           =    258.28
Dispersion (Pearson):             1.311073     Dispersion         =  1.311073

                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   age_round |   .6105166   .0287151    21.26   0.000     .5542359    .6667973
      age_r2 |  -.0110303   .0009153   -12.05   0.000    -.0128243   -.0092364
        girl |  -.7270275   .2555978    -2.84   0.004     -1.22799    -.226065
       _cons |   4.288499   .1658479    25.86   0.000     3.963443    4.613554
------------------------------------------------------------------------------
```

Note that the `robust` option leads to valid standard errors even if the correlations within group are not as hypothesized by the specified correlation structure (as long as the model mean is correctly specified).

Using an exchangeable working matrix instead of an independent working matrix slightly changes the parameter estimates:

```
. xtgee weight age_round age_r2 girl, cor(exch) vce(robust)
GEE population-averaged model                   Number of obs      =       197
Group variable:                         id      Number of groups   =        68
Link:                             identity      Obs per group: min =         1
Family:                           Gaussian                     avg =       2.9
Correlation:                 exchangeable                     max =         5
                                               Wald chi2(3)       =   1897.86
Scale parameter:                  1.316525     Prob > chi2        =    0.0000

                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   age_round |   .6392608   .0223122    28.65   0.000     .5955298    .6829919
      age_r2 |  -.0118552   .0007118   -16.66   0.000    -.0132502   -.0104602
```

```
       girl |  -.7313195   .2356071    -3.10   0.002    -1.193101    -.269538
      _cons |   4.137133   .1446893    28.59   0.000     3.853547    4.420719
------------------------------------------------------------------------------
```

The exchangeable working matrix is also the structure implied by a random intercepts model. Fitting it using MLE we find very similar parameter estimates:

```
. mixed weight age_round age_r2 girl || id:, ml

Mixed-effects ML regression                     Number of obs      =       197
Group variable: id                              Number of groups   =        68

                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5


                                                Wald chi2(3)       =   2622.36
Log likelihood = -272.11841                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   age_round |   .6374622   .0233183    27.34   0.000     .5917591    .6831653
      age_r2 |  -.0118043   .0007514   -15.71   0.000     -.013277   -.0103316
        girl |  -.7302047   .2344282    -3.11   0.002    -1.189676   -.2707339
       _cons |   4.145148   .2037314    20.35   0.000     3.745842    4.544455
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |    .853041   .0930156       .688898    1.056294
-----------------------------+------------------------------------------------
                sd(Residual) |   .7370556   .0454524      .6531437    .8317479
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    68.18 Prob >= chibar2 = 0.0000
```

8.8 Comments

- GEE models are expressed in terms of marginal parameters for the explanatory variables, that is they measure the change in Y per average change in X_k across all clusters; mixed effects models are expressed in terms of conditional parameters for the explanatory variables, that is they measure the change in Y per change in X_k in a given cluster. In general these are different. Only for models with identity or logarithmic link are they the same.
- For data arising from a normal distribution the first two moments fully determine the likelihood and for this reason, with completely balanced data, if the variance-covariance matrix assumed when implementing GEEs is the same as the one implied by a mixed effects models the two methods lead to the same $\hat{\boldsymbol{\beta}}$. However note that this is not true in general.
- The consistency property discussed here is a large sample result: hence there must a sufficiently large number of clusters for the results to apply.
- The working correlation matrix needs to be chosen carefully:

  - If observations are clustered without any ordering then exchangeable may be appropriate.
  - If the number of clusters is relatively small (but not too small) and data are completely balanced, an unstructured matrix may be appropriate.

- Standard GEE models assume that missing observations are **Missing Completely at Random (MCAR)**; that is missing data are independent of both the value of the missing response and the observed data. If they are missing by other mechanisms estimates are biased (see the practical). Recall that conditional models are valid under the weaker assumption of **Missing at Random (MAR)**.

Chapter 9

Further issues

9.1 Three or more level data

We can generalize all we have learnt so far using further levels of dependence. Consider again the PEFR data and the results from the two meters together, the Mini Wright and the Standard Wright. This time we will analyse them together. First we treat their difference in performance as a fixed effect. In order to do so we need to reshape the data so that all four measures are stacked up for each individual and we have a marker for the Mini Wright measurement method:

```
. use pefr,clear
. reshape long  wm wp, i(id) j(occasion)
. gen i=_n
. reshape long w, i(i) j(meth) string
. gen mini=meth=="m"
```

The 2 level random intercept model with meter as a fixed effect is then fitted with:

```
. mixed w mini || id:, reml stddev
------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mini |   6.029412   5.776497     1.04   0.297    -5.292314    17.35114
       _cons |   447.8824   27.60971    16.22   0.000     393.7683    501.9964
------------------------------------------------------------------------------


------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   112.5851    20.1255      79.30889    159.8233
-----------------------------+------------------------------------------------
                sd(Residual) |   23.81711    2.38171      19.57802    28.97404
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 133.92        Prob >= chibar2 = 0.0000
```

This tells us that the Mini Wright meter overestimates the overall mean by 6.03 units, on average. Given its SE however this difference is not statistically significant. On the other hand such a shift may not be the same for all individuals.

We could model this between-methods heterogeneity (across subjects) by including another random intercept for each combination of method and subject. The three level model can be written as:

$$Y_{ikj} = \beta_0 + u^{(2)}_{kj} + u^{(3)}_{j} + e_{ikj}$$

where $u^{(2)}_{kj}$ is the component of the random intercept for method k on subject j and $u^{(3)}_{j}$ of subject j (and influences all measures taken on j). Here method is nested within subject.

The distributional assumptions are : $u^{(3)}_{j} \sim N(0, \sigma^2_{3u0})$, $u^{(2)}_{kj}|u^{(3)}_{k} \sim N(0, \sigma^2_{2u0})$, and $e_{ikj}|u^{(2)}_{kj}, u^{(3)}_{j} \sim N(0, \sigma^2_{e})$.

This model can be fitted in `mixed` using an additional layer of double vertical bars as follows, with `mini` corresponding to the k index in the notation above:

```
. mixed w || id: || mini:, reml stddev

------------------------------------------------------------
                |     No. of       Observations per Group
Group Variable  |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
             id |         17          4        4.0          4
           mini |         34          2        2.0          2
------------------------------------------------------------

------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   450.8971   27.45823    16.42   0.000     397.0799    504.7142
------------------------------------------------------------------------------

------------------------------------------------------------------------------
Random-effects Parameters    |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                  sd(_cons) |   112.0211   20.23062      78.62772    159.5966
-----------------------------+------------------------------------------------
mini: Identity               |
                  sd(_cons) |   19.47623   4.829488      11.97937    31.66474
-----------------------------+------------------------------------------------
               sd(Residual) |   17.75859   2.153545      14.00184    22.52329
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 145.19                Prob > chi2 = 0.0000
```

This model can also include fixed terms, as in

$$Y_{ikj} = \beta_0 + \beta_1 X_j + u^{(2)}_{kj} + u^{(3)}_{j} + e_{ikj}$$

Note that X_j may include mini,

```
. mixed w mini || id: || mini:, reml stddev

------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------
             id |         17          4        4.0          4
           mini |         34          2        2.0          2
------------------------------------------------------------

                                              Wald chi2(1)      =       0.56
```

```
Log restricted-likelihood = -337.79137          Prob > chi2       =     0.4540

------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mini |   6.029412   8.053187     0.75   0.454    -9.754544    21.81337
       _cons |   447.8824    27.7519    16.14   0.000     393.4896    502.2751
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   111.9893   20.23686       78.5889    159.5848
-----------------------------+------------------------------------------------
mini: Identity               |
                   sd(_cons) |   19.83869   5.005767       12.0986    32.53053
-----------------------------+------------------------------------------------
                sd(Residual) |   17.75859   2.153545      14.00184    22.52329
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 143.25                Prob > chi2 = 0.0000
```

9.2 Binary dependent variables

When the outcome is binary- as opposed to continuous, similar concerns arise regarding the need to account for dependencies among the observations. There are however additional complications regarding the interpretation of the parameters used to specify the models.

We will start with a simple example and only consider a random intercept model for a binary outcome. The topic is covered at greater length in *Advanced Statistical Models* (optional module in term 3).

The logistic regression model for a binary outcome Y_j observed on subject j, assumed to be distributed according to a Bernoulli (i.e. binomial) distribution, can be defined in terms of the logit function:

$$\text{logit } (\Pr(Y_j = 1|X_j)) = ln\{\frac{\Pr(Y_j = 1|X_j)}{1 - \Pr(Y_j = 1|X_j)}\}$$

and this is expressed in terms of a constant plus a linear function of X_j:

$$\text{logit } (\Pr(Y_j = 1|X_j)) = \beta_0 + \beta_1 X_j$$

In the presence of dependency among the observations belonging to the same cluster, as for example in repeated measures of a binary indicator of disease status (e.g. infection), we generalize this model as follows:

$$\text{logit } (\Pr(Y_{ij} = 1|X_{ij}, u_j)) = \beta_0 + \beta_1 X_{ij} + u_j$$

where u_j is assumed to be normally distributed with mean 0 and variance σ_u^2, and $Y_{ij} \sim$ Binomial$(1, \pi_{ij})$.

This is a random intercept logistic regression model, an example of **generalized linear mixed models**.

The main Stata commands that can be used to fit such models are `melogit`, `xtlogit` and `gllamm`. The first two are equivalent to `mixed` for linear random effects regression. All three require numerical integration and therefore are relatively slow.

As an example consider the Siblings data, with the binary response defined by low birth weight of less than $2500g$. We fit a mixed effects model using the `melogit` command.

```
. gen lowbw=birwt<2500
```

```
. melogit lowbw male || momid:,
Mixed-effects logistic regression
Group variable:           momid

Integration method: mvaghermite
--------------------------------------
lowbw        |      Coef.   Std. Err.
-------------+------------------------
male         |  -.3985269    .142015
_cons        |  -4.494467   .2187952
-------------+------------------------
momid        |
var(_cons)   |   3.332344   .5916439
--------------------------------------
```

Note that β_{male} is a conditional (on u_j) effect. That is, the odds ratio of low birth weight for a boy is 0.671, relative to a girl with the same mother.

To estimate the marginal effect over the population of a child being male we can use the post-estimation command `margins` (Note that this returns probabilities of low birth weight for girls and boys in the population). We can calculated the marginal effect (over all babies) by hand.

```
. margins, by(male)
--------------------------------------
             |     Margin   Std. Err.
-------------+------------------------
Female       |   .0401334   .0031566
Male         |   .0291492   .0026683
--------------------------------------

. scalar or_f = 0.0401334/(1 - 0.0401334)
. scalar or_m = 0.0291492/(1 - 0.0291492)
. disp or_m/or_f       0.7180903
```

The marginal effect (over all u_j) is 0.718. That is, the odds ratio of low birth weight for a boy is 0.734, relative to a girl from the same population. Hence, care should be taken when interpreting the results from fitting such models.

9.3 Designing multilevel/clustered studies

To design a good (simple or multilevel) study it is essential to be clear about its primary objectives. This is usually expressed in terms of one parameter of interest. Therefore the problem becomes one of determining the sample size that will give a small enough SE for that parameter.

It is well known that in simple random samples the SE of the mean is related to the sample size by the relation $SE = SD/\sqrt{N}$ where N is the sample size. When using 2 stage-sampling the clustering needs to be taken into account and therefore the SE of the mean, for example, needs updating. If all clusters have the same size, n, the total sample size is $N = nJ$, with J being the total number of clusters.

The **design effect** is the amount that the total study size N needs to be inflated in order to give the same SE of the estimator that would be obtained from a random sample of a given size. A simple expression for the design effect when the cluster sizes are equal is:

$$\text{design effect} = 1 + (n-1)\lambda$$

where λ is the intraclass correlation and n the common cluster size. Hence the more homogeneous the clusters are, or the greater the cluster size, the greater the design effect is and the greater the total required sample size.

Note that this simple expression will underestimate the required sample size if there are few clusters available or if the size of the clusters is likely to vary substantially. Both these issues are common in practice and alternative sample size adjustments have been derived to accommodate them.

A complementary measure to the design effect is the **effective study size**. This is equal to the total number of elementary units divided by the the design effect:

$$N_{\text{effective}} = \frac{nJ}{\text{design effect}}$$

.

The required steps when designing a study are therefore:

1. Work out the required sample size for a single stage design
2. Make an informed guess (or a series of guesses) of the size of λ (some databases containing example ICCs have been compiled and made freely available)
3. Calculate the design effect
4. Inflate the required sample size of step 1 accounting for the design effect.

The design effect above is defined for the estimation of the overall mean. When the focus of the study is the effect of an explanatory variable that varies between clusters or that varies within cluster we should consider the formula for the SE of their fixed effect regression coefficients. Details are given in Rabe-Hesketh and Skrondal (2008), section 3.10.3.

For a more general discussion of study design, including those aimed at the variance structure, see Chapter 10 of Snijder and Bosker (1999).

9.4 Summary

The aim of this module was to introduce methods for the statistical analysis of dependent data when the outcome of interest is continuous and when there are two levels of aggregation.

The main points covered were:

- *Consequences of dependency*
  If the statistical dependency among observations is ignored, any subsequent inferences are potentially invalid. Dependency therefore must be dealt with.
- *Estimation*
  In many settings dependence can be ignored when estimating parameters, but this can (although not necessarily) be an *inefficient* procedure. Alternatively we fit models that explicitly specify the nature of the dependency, with several modelling options available.
- *Inference*
  Even if the dependency in the elementary level observations is ignored, when making inferences (for example using hypothesis tests) the dependence among observations *must* be accommodated. This may come directly from a likelihood analysis with a model that incorporates the dependency, or after estimation of the parameters, i.e. in a second stage, using a *robust* approach to the estimation of precision.

We have started with development of ANOVA and the linear regression model and then developed random intercept, random intercept and slope models and then the more general marginal covariance structures leading to generalized linear mixed models. We have examined special issues arising with longitudinal data. We have used the Stata software package in lectures and practicals. You should therefore be quite adept at dealing with relatively complex data structures.

In the next session we will release this year's assignment and provide a supervised practical session. There will be an opportunity to ask questions about the assignment and anything else covered in the course.

The subsequent two lectures are dedicated to treatment of missing data. Failure to obtain the full dataset is common in medical and public health research, particularly where measurements are taken over a long period.

The module will finish with a guest lecture from a researcher from Malawi and will cover application of three methods to a real public health problem. **Material in this lecture will form part of the assessment and is worth 10% of the marks, so it is very important that you attend**.

Chapter 10

Revision

The material for this session will be developed in response to requests from the students and made available close to the time of delivery

Chapter 11

Missing data I

The material for this session is held in a separate section in Moodle

Chapter 12

Missing data II

The material for this session will be distributed at a later date

Bibliography