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

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**:

```
. 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

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

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^2.

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

$$F(J-1, N-J) = \frac{\text{MMS}}{\text{MSE}}$$

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 $F = 12.18 = t^2 = (3.4895)^2$:

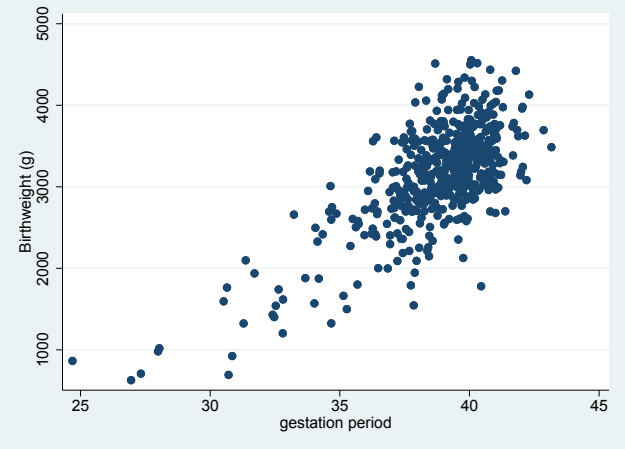
```
. anova bweight sex

          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

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

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

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

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^2(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^2 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:

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

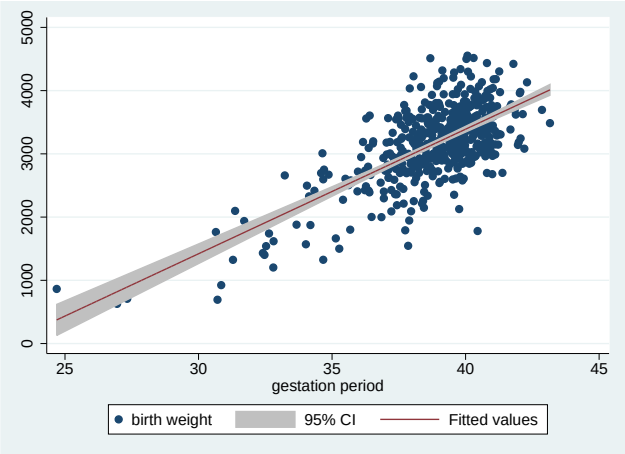
They are the same as the OLS estimates. The MLE for σ^2 instead is the RSS divided by N, not $N-2$.

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:

$$Y_i = \beta_0 + \beta_1(X_i - \hat{\mu}_X) + \beta_2(X_i - \hat{\mu}_X)^2 + \beta_3 \text{ female} + \epsilon_i \quad (1)$$

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 (1)