Chapter 15

Mixed Models

A flexible approach to correlated data.

15.1 Overview

Correlated data arise frequently in statistical analyses. This may be due to grouping of subjects, e.g., students within classrooms, or to repeated measurements on each subject over time or space, or to multiple related outcome measures at one point in time. Mixed model analysis provides a general, flexible approach in these situations, because it allows a wide variety of correlation patterns (or variance-covariance structures) to be explicitly modeled.

As mentioned in chapter 14, multiple measurements per subject generally result in the correlated errors that are explicitly forbidden by the assumptions of standard (between-subjects) AN(C)OVA and regression models. While repeated measures analysis of the type found in SPSS, which I will call "classical repeated measures analysis", can model general (multivariate approach) or spherical (univariate approach) variance-covariance structures, they are not suited for other explicit structures. Even more importantly, these repeated measures approaches discard all results on any subject with even a single missing measurement, while mixed models allow other data on such subjects to be used as long as the missing data meets the so-called missing-at-random definition. Another advantage of mixed models is that they naturally handle uneven spacing of repeated measurements, whether intentional or unintentional. Also important is the fact that mixed model analysis is

often more interpretable than classical repeated measures. Finally, mixed models can also be extended (as generalized mixed models) to non-Normal outcomes.

The term mixed model refers to the use of both fixed and random effects in the same analysis. As explained in section 14.1, fixed effects have levels that are of primary interest and would be used again if the experiment were repeated. Random effects have levels that are not of primary interest, but rather are thought of as a random selection from a much larger set of levels. Subject effects are almost always random effects, while treatment levels are almost always fixed effects. Other examples of random effects include cities in a multi-site trial, batches in a chemical or industrial experiment, and classrooms in an educational setting.

As explained in more detail below, the use of both fixed and random effects in the same model can be thought of hierarchically, and there is a very close relationship between mixed models and the class of models called hierarchical linear models. The hierarchy arises because we can think of one level for subjects and another level for measurements within subjects. In more complicated situations, there can be more than two levels of the hierarchy. The hierarchy also plays out in the different roles of the fixed and random effects parameters. Again, this will be discussed more fully below, but the basic idea is that the fixed effects parameters tell how population means differ between any set of treatments, while the random effect parameters represent the general variability among subjects or other units.

Mixed models use both fixed and random effects. These correspond to a hierarchy of levels with the repeated, correlated measurement occurring among all of the lower level units for each particular upper level unit.

15.2   A video game example

Consider a study of the learning effects of repeated plays of a video game where age is expected to have an effect. The data are in MMvideo.txt. The quantitative outcome is the score on the video game (in thousands of points). The explanatory variables are age group of the subject and "trial" which represents which time the subject played the game (1 to 5). The "id" variable identifies the subjects. Note

the the data are in the tall format with one observation per row, and multiple rows per subject,
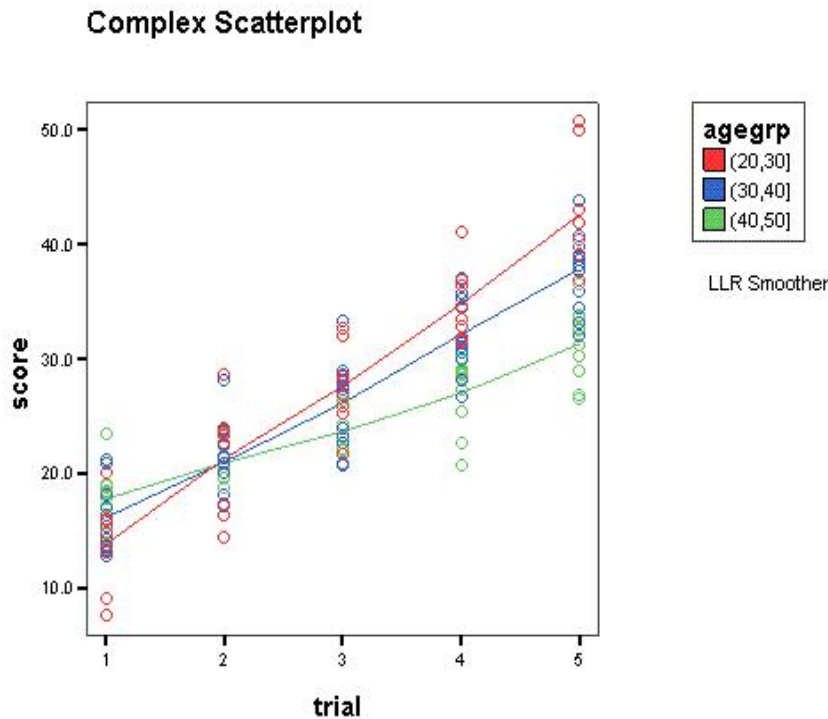
Complex Scatterplot

Figure 15.1: EDA for video game example with smoothed lines for each age group.

Some EDA is shown in figure 15.1. The plot shows all of the data points, with game score plotted against trial number. Smoothed lines are shown for each of the three age groups. The plot shows evidence of learning, with players improving their score for each game over the previous game. The improvement looks fairly linear. The y-intercept (off the graph to the left) appears to be higher for older players. The slope (rate of learning) appears steeper for younger players.

At this point you are most likely thinking that this problem looks like an AN-COVA problem where each age group has a different intercept and slope for the relationship between the quantitative variables trial and score. But ANCOVA assumes that all of the measurements for a given age group category have uncorrelated errors. In the current problem each subject has several measurements and

the errors for those measurements will almost surely be correlated. This shows up as many subjects with most or all of their outcomes on the same side of their group's fitted line.

15.3   Mixed model approach

The solution to the problem of correlated within-subject errors in the video game example is to let each subject have his or her own "personal" intercept (and possibly slope) randomly deviating from the mean intercept for each age group. This results in a group of parallel "personal" regression lines (or non-parallel if the slope is also random). Then, it is reasonable (but not certain) that the errors around the personal regression lines will be uncorrelated. One way to do this is to use subject identification as a categorical variable, but this is treating the inherently random subject-to-subject effects as fixed effects, and "wastes" one parameter for each subject in order to estimate his or her personal intercept. A better approach is to just estimate a single variance parameter which represents how spread out the random intercepts are around the common intercept of each group (usually following a Normal distribution). This is the mixed models approach.

From another point of view, in a mixed model we have a hierarchy of levels. At the top level the units are often subjects or classrooms. At the lower level we could have repeated measurements within subjects or students within classrooms. The lower level measurements that are within the same upper level unit are correlated, when all of their measurements are compared to the mean of all measurements for a given treatment, but often uncorrelated when compared to a personal (or class level) mean or regression line. We also expect that there are various measured and unmeasured aspects of the upper level units that affect all of the lower level measurements similarly for a given unit. For example various subject skills and traits may affect all measurements for each subject, and various classroom traits such as teacher characteristics and classroom environment affect all of the students in a classroom similarly. Treatments are usually applied randomly to whole upper-level units. For example, some subjects receive a drug and some receive a placebo, Or some classrooms get an aide and others do not.

In addition to all of these aspects of hierarchical data analysis, there is a variety of possible variance-covariance structures for the relationships among the lower level units. One common structure is called compound symmetry, which indicates the same correlation between all pairs of measurements, as in the sphericity char-

acteristic of chapter 14. This is a natural way to represent the relationship between students within a classroom. If the true correlation structure is compound symmetry, then using a random intercept for each upper level unit will remove the correlation among lower level units. Another commonly used structure is autoregressive, in which measurements are ordered, and adjacent measurements are more highly correlated than distant measurements.

To summarize, in each problem the hierarchy is usually fairly obvious, but the user must think about and specify which fixed effects (explanatory variables, including transformations and interactions) affect the average responses for all subjects. Then the user must specify which of the fixed effect coefficients are sufficient without a corresponding random effect as opposed to those fixed coefficients which only represent an average around which individual units vary randomly. In addition, correlations among measurements that are not fully accounted for by the random intercepts and slopes may be specified. And finally, if there are multiple random effects the correlation of these various effects may need to be specified.

> **To run a mixed model, the user must make many choices including the nature of the hierarchy, the fixed effects and the random effects.**

In almost all situations several related models are considered and some form of model selection must be used to choose among related models.

The interpretation of the statistical output of a mixed model requires an understanding of how to explain the relationships among the fixed and random effects in terms of the levels of the hierarchy.

15.4 Analyzing the video game example

Based on figure 15.1 we should model separate linear relationships between trial number and game score for each age group. Figure 15.2, shows smoothed lines for each subject. From this figure, it looks like we need a separate slope and intercept for each age group. It is also fairly clear that in each group there is random subject-to-subject variation in the intercepts. We should also consider the possibilities that the "learning trajectory" is curved rather than linear, perhaps using the square of the trial number as an additional covariate to create a quadratic curve. We should
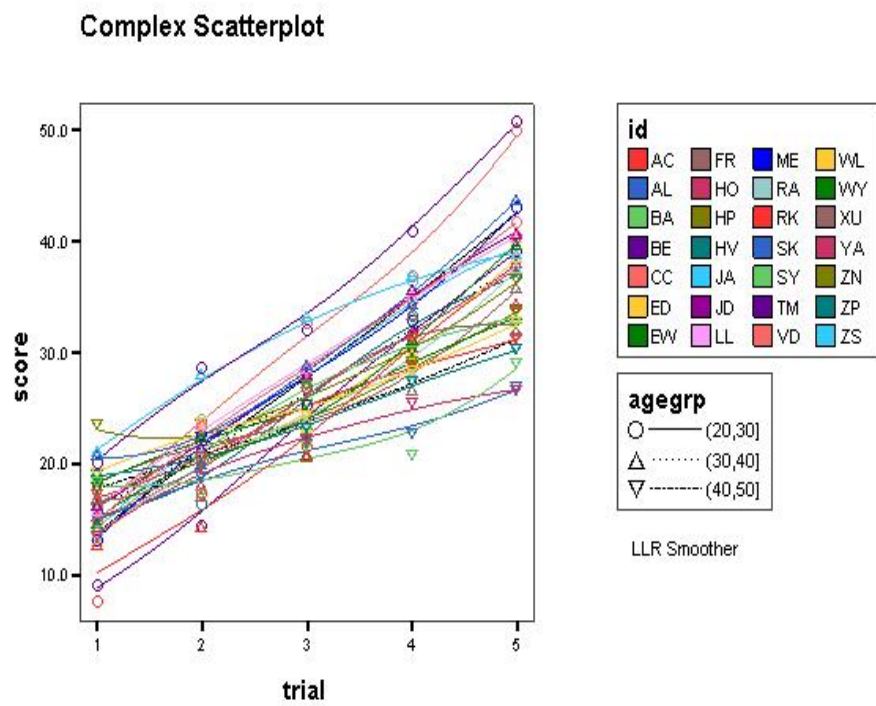
Figure 15.2: EDA for video game example with smoothed lines for each subject.

also check if a random slope is needed. It is also prudent to check if the random intercept is really needed. In addition, we should check if an autoregressive model is needed.

15.5   Setting up a model in SPSS

The mixed models section of SPSS, accessible from the menu item "Analyze / Mixed Models / Linear", has an initial dialog box ("Specify Subjects and Repeated"), a main dialog box, and the usual subsidiary dialog boxes activated by clicking buttons in the main dialog box. In the initial dialog box (figure 15.3) you will always specify the upper level of the hierarchy by moving the identifier for that level into the "subjects" box. For our video game example this is the subject "id" column. For a classroom example in which we study many students in each classroom, this would be the classroom identifier.
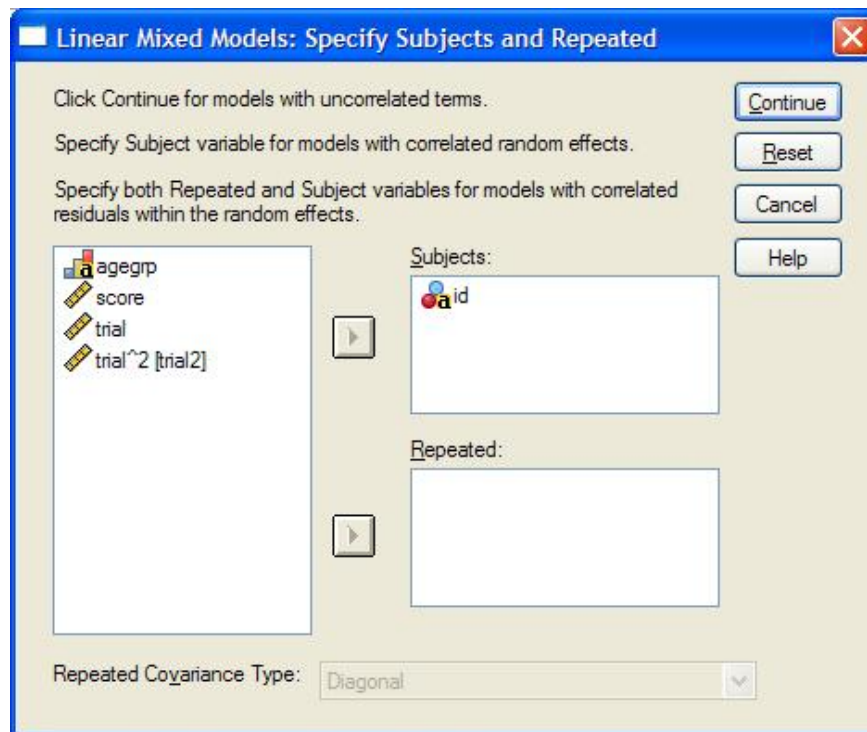

Figure 15.3: Specify Subjects and Repeated Dialog Box.

If we want to model the correlation of the repeated measurements for each subject (other than the correlation induced by random intercepts), then we need to specify the order of the measurements within a subject in the bottom ("repeated") box. For the video game example, the trial number could be appropriate.
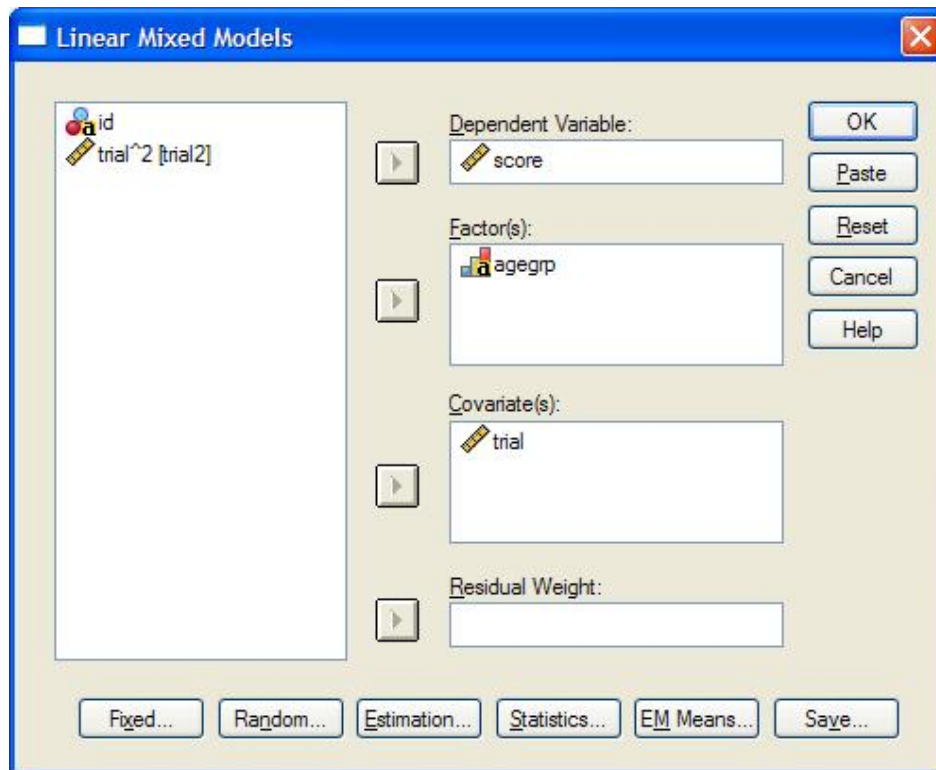

Figure 15.4: Main Linear Mixed Effects Dialog Box.

The main "Linear Mixed Models" dialog box is shown in figure 15.4. (Note that just like in regression analysis use of transformation of the outcome or a quantitative explanatory variable, i.e., a covariate, will allow fitting of curves.) As usual, you must put a quantitative outcome variable in the "Dependent Variable" box. In the "Factor(s)" box you put any categorical explanatory variables (but not the subject variable itself). In the "Covariate(s)" box you put any quantitative explanatory variables. **Important note:** For mixed models, specifying factors and covariates on the main screen does *not* indicate that they will be used in the model, only that they are available for use in a model.

The next step is to specify the fixed effects components of the model, using

the Fixed button which brings up the "Fixed Effects" dialog box, as shown in figure 15.5. Here you will specify the structural model for the "typical" subject, which is just like what we did in ANCOVA models. Each explanatory variable or interaction that you specify will have a corresponding parameter estimated, and that estimate will represent the relationship between that explanatory variable and the outcome if there is no corresponding random effect, and it will represent the mean relationship if there is a corresponding random effect.
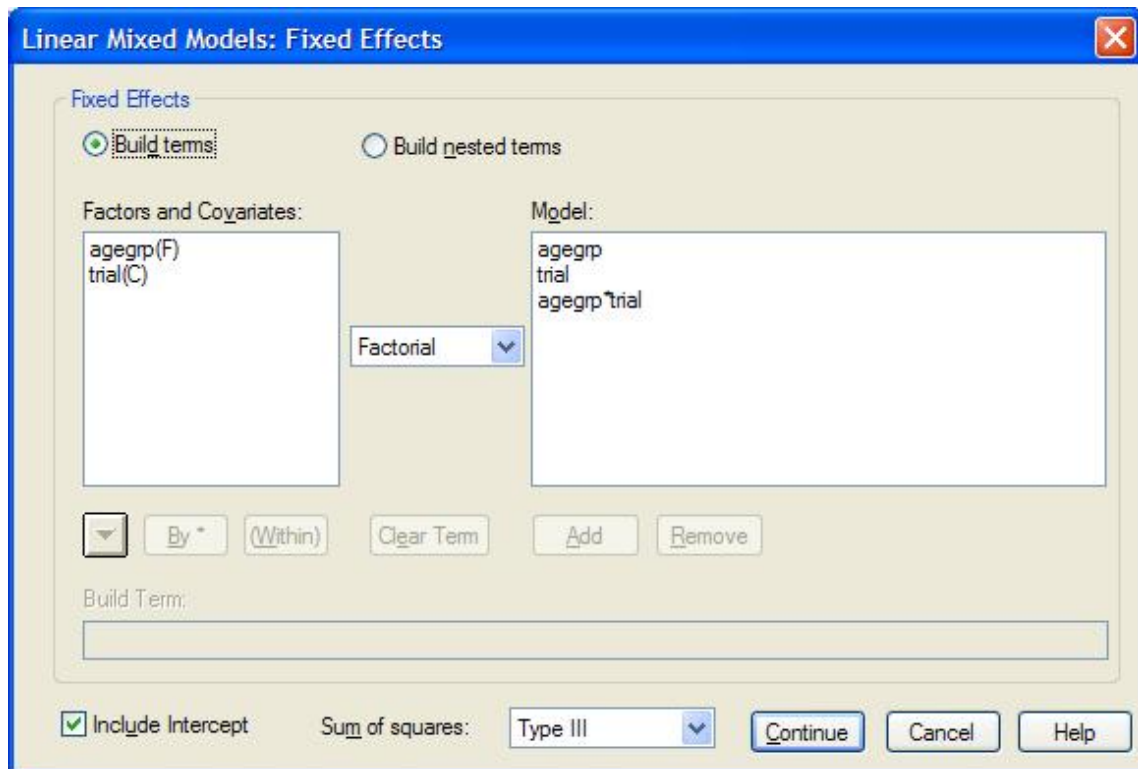

Figure 15.5: Fixed Effects Dialog Box.

For the video example, I specified main effects for age group and trial plus their interaction. (You will always want to include the main effects for any interaction you specify.) Just like in ANCOVA, this model allows a different intercept and slope for each age group. The fixed intercept (included unless the "Include intercept" check box is unchecked) represents the (mean) intercept for the baseline age group, and the $k - 1$ coefficients for the age group factor (with $k = 3$ levels) represent differences in (mean) intercept for the other age groups. The trial co-

efficient represents the (mean) slope for the baseline group, while the interaction coefficients represent the differences in (mean) slope for the other groups relative to the baseline group. (As in other "model" dialog boxes, the actual model depends only on what is in the "Model box", not how you got it there.)

In the "Random Effects" dialog box (figure 15.6), you will specify which parameters of the fixed effects model are only means around which individual subjects vary randomly, which we think of as having their own personal values. Mathematically these personal values, e.g., a personal intercept for a given subject, are equal to the fixed effect plus a random deviation from that fixed effect, which is zero on average, but which has a magnitude that is controlled by the size of the random effect, which is a variance.
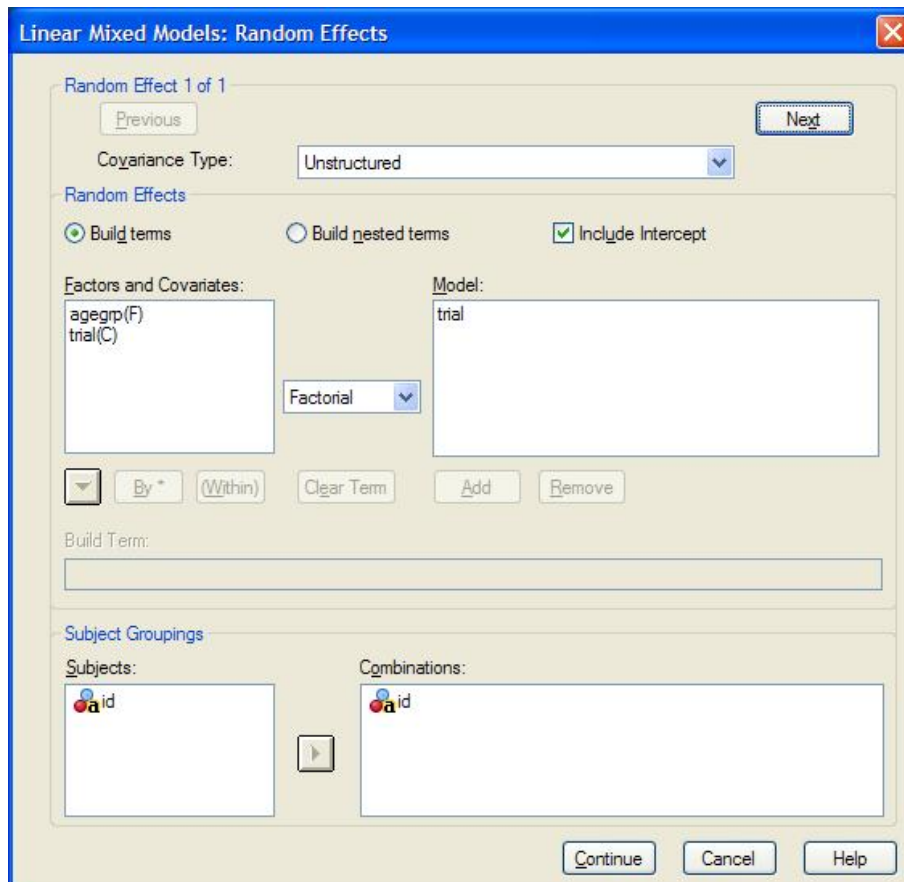

Figure 15.6: Random Effects Dialog Box.

In the random effects dialog box, you will usually want to check "Include Intercept", to allow a separate intercept (or subject mean if no covariate is used) for each subject (or each level of some other upper level variable). If you specify any random effects, then you must indicate that there is a separate "personal" value of, say, the intercept, for each subject by placing the subject identifier in the "Combinations" box. (This step is very easy to forget, so get in the habit of doing this every time.)

To model a random slope, move the covariate that defines that slope into the "Model" box. In this example, moving trial into the Model box could be used to model a random slope for the score by trial relationship. It does not make sense to include a random effect for any variable unless there is also a fixed effect for that variable, because the fixed effect represents the average value around which the random effect varies. If you have more than one random effect, e.g., a random intercept and a random slope, then you need to specify any correlation between these using the "Covariance Type" drop-down box. For a single random effect, use "identity". Otherwise, "unstructured" is usually most appropriate because it allows correlation among the random effects (see next paragraph). Another choice is "diagonal" which assumes no correlation between the random effects.

What does it mean for two random effects to be correlated? I will illustrate this with the example of a random intercept and a random slope for the trial vs. game score relationship. In this example, there are different intercepts and slopes for each age group, so we need to focus on any one age group for this discussion. The fixed effects define a mean intercept and mean slope for that age group, and of course this defines a mean fitted regression line for the group. The idea of a random intercept and a random slope indicate that any given subject will "wiggle" a bit around this mean regression line both up or down (random intercept) and clockwise or counterclockwise (random slope). The variances (and therefore standard deviations) of the random effects determine the sizes of typical deviations from the mean intercept and slope. But in many situations like this video game example subjects with a higher than average intercept tend to have a lower than average slope, so there is a negative correlation between the random intercept effect and the random slope effect. We can look at it like this: the next subject is represented by a random draw of an intercept deviation and a slope deviation from a distribution with mean zero for both, but with a negative correlation between these two random deviations. Then the personal intercept and slope are constructed by adding these random deviations to the fixed effect coefficients.

Some other buttons in the main mixed models dialog box are useful. I recommend that you always click the Statistics button, then check both "Parameter estimates" and "Tests for covariance parameters". The parameter estimates are needed for interpretation of the results, similar to what we did for ANCOVA (see chapter 10). The tests for covariance parameters aid in determining which random effects are needed in a given situation. The "EM Means" button allows generation of "expected marginal means" which average over all subjects and other treatment variables. In the current video game example, marginal means for the three video groups is not very useful because this averages over the trials and the score varies dramatically over the trials. Also, in the face of an interaction between age group and trial number, averages for each level of age group are really meaningless.

As you can see there are many choices to be made when creating a mixed model. In fact there are many more choices possible than described here. This flexibility makes mixed models an important general purpose tool for statistical analysis, but suggests that it should be used with caution by inexperienced analysts.

Specifying a mixed model requires many steps, each of which requires an informed choice. This is both a weakness and a strength of mixed model analysis.

15.6   Interpreting the results for the video game example

Here is some of the SPSS output for the video game example. We start with the model for a linear relationship between trial and score with separate intercepts and slopes for each age group, and including a random per-subject intercept. Table 15.1 is called "Model Dimension". Focus on the "number of parameters" column. The total is a measure of overall complexity of the model and plays a role in model selection (see next section). For quantitative explanatory variables, there is only one parameter. For categorical variables, this column tells how many parameters are being estimated in the model. The "number of levels" column tells how many lines are devoted to an explanatory variable in the Fixed Effects table (see below), but lines beyond the number of estimated parameters are essentially blank (with

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed	Intercept	1		1	
Effects	agegrp	3		2	
	trial	1		1	
	agegrp * trial	3		2	
Random Effects	Intercept	1	Identity	1	id
Residual				1	
Total		9		8	

Table 15.1: Model dimension for the video game example.

parameters labeled as redundant and a period in the rest of the columns). We can see that we have a single random effect, which is an intercept for each level of id (each subject). The Model Dimension table is a good quick check that the computer is fitting the model that you intended to fit.

The next table in the output is labeled "Information Criteria" and contains many different measures of how well the model fits the data. I recommend that you only pay attention to the last one, "Schwartz's Bayesian Criterion (BIC)", also called Bayesian Information Criterion. In this model, the value is 718.4. See the section on model comparison for more about information criteria.

Next comes the Fixed Effects tables (tables 15.2 and 15.3). The tests of fixed effects has an ANOVA-style test for each fixed effect in the model. This is nice because it gives a single overall test of the usefulness of a given explanatory variable, without focusing on individual levels. Generally, you will want to remove explanatory variables that do not have a significant fixed effect in this table, and then rerun the mixed effect analysis with the simpler model. In this example, all effects are significant (less than the standard alpha of 0.05). Note that I converted the SPSS p-values from 0.000 to the correct form.

The Estimates of Fixed Effects table does not appear by default; it is produced by choosing "parameter estimates" under Statistics. We can see that age group 40-50 is the "baseline" (because SPSS chooses the last category). Therefore the (fixed) intercept value of 14.02 represents the mean game score (in thousands of points) for 40 to 50 year olds for trial zero. Because trials start at one, the intercepts are not meaningful in themselves for this problem, although they are needed for calculating and drawing the best fit lines for each age group.

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	57.8	266.0	<0.0005
agegrp	2	80.1	10.8	<0.0005
trial	1	118.9	1767.0	<0.0005
agegrp * trial	2	118.9	70.8	<0.0005

Table 15.2: Tests of Fixed Effects for the video game example.

						95% Conf. Int.	
		Std.				Lower	Upper
Parameter	Estimate	Error	df	t	Sig.	Bound	Bound
Intercept	14.02	1.11	55.4	12.64	<0.0005	11.80	16.24
agegrp=(20,30)	-7.26	1.57	73.0	-4.62	<0.0005	-10.39	-4.13
agegrp=(30,40)	-3.49	1.45	64.2	-2.40	0.019	-6.39	-0.59
agegrp=(40,50)	0	0	.	.	.	.	.
trial	3.32	0.22	118.9	15.40	<0.0005	2.89	3.74
(20,30)*trial	3.80	0.32	118.9	11.77	<0.0005	3.16	4.44
(30,40)*trial	2.14	0.29	118.9	7.35	<0.0005	1.57	2.72
(40,50)*trial	0	0	.	.	.	.	.

Table 15.3: Estimates of Fixed Effects for the video game example.

As in ANCOVA, writing out the full regression model then simplifying tells us that the intercept for 20 to 30 year olds is 14.02-7.26=6.76 and this is significantly lower than for 40 to 50 year olds (t=-4.62, p<0.0005, 95% CI for the difference is 4.13 to 10.39 thousand points lower). Similarly we know that the 30 to 40 years olds have a lower intercept than the 40 to 50 year olds. Again these intercepts themselves are not directly interpretable because they represent trial zero. (It would be worthwhile to recode the trial numbers as zero to four, then rerun the analysis, because then the intercepts would represent game scores the first time someone plays the game.)

The trial coefficient of 3.32 represents that average gain in game score (in thousands of points) for each subsequent trial *for the baseline 40 to 50 year old age group*. The interaction estimates tell the *difference* in slope for other age groups compared to the 40 to 50 year olds. Here both the 20 to 30 year olds and the 30 to 40 year olds learn quicker than the 40 to 50 year olds, as shown by the significant interaction p-values and the positive sign on the estimates. For example, we are 95% confident that the trial to trial "learning" gain is 3.16 to 4.44 thousand points *higher* for the youngest age group compared to the oldest age group.

Interpret the fixed effects for a mixed model in the same way as an ANOVA, regression, or ANCOVA depending on the nature of the explanatory variables(s), but realize that any of the coefficients that have a corresponding random effect represent the mean over all subjects, and each individual subject has their own "personal" value for that coefficient.

The next table is called "Estimates of Covariance Parameters" (table 15.4). It is very important to realize that while the parameter estimates given in the Fixed Effects table are estimates of mean parameters, the parameter estimates in this table are estimates of variance parameters. The intercept variance is estimated as 6.46, so the estimate of the standard deviation is 2.54. This tells us that for any given age group, e.g., the oldest group with mean intercept of 14.02, the individual subjects will have "personal" intercepts that are up to 2.54 higher or lower than the group average about 68% of the time, and up to 5.08 higher or lower about 95% of the time. The null hypothesis for this parameter is a variance of zero, which would indicate that a random effect is not needed. The test statistic is called a Wald Z statistic. Here we reject the null hypothesis (Wald Z=3.15, p=0.002)

					95% Conf. Int.	
		Std.	Wald		Lower	Upper
Parameter	Estimate	Error	Z	Sig.	Bound	Bound
Residual	4.63	0.60	7.71	<0.0005	3.59	5.97
Intercept(Subject=id) Variance	6.46	2.05	3.15	0.002	3.47	12.02

Table 15.4: Estimates of Covariance Parameters for the video game example.

and conclude that we do need a random intercept. This suggests that there are important unmeasured explanatory variables for each subject that raise or lower their performance in a way that appears random because we do not know the value(s) of the missing explanatory variable(s).

The estimate of the residual variance, with standard deviation equal to 2.15 (square root of 4.63), represents the variability of individual trial's game scores around the individual regression lines for each subjects. We are assuming that once a personal best-fit line is drawn for each subject, their actual measurements will randomly vary around this line with about 95% of the values falling within 4.30 of the line. (This is an estimate of the same σ^2 as in a regression or ANCOVA problem.) The p-value for the residual is not very meaningful.

Random effects estimates are variances. Interpret a random effect parameter estimate as the magnitude of the variability of "personal" coefficients from the mean fixed effects coefficient.

All of these interpretations are contingent on choosing the right model. The next section discusses model selection.

15.7   Model selection for the video game example

Because there are many choices among models to fit to a given data set in the mixed model setting, we need an approach to choosing among the models. Even then, we must always remember that all models are wrong (because they are idealized simplifications of Nature), but some are useful. Sometimes a single best model

is chosen. Sometimes subject matter knowledge is used to choose the most useful models (for prediction or for interpretation). And sometimes several models, which differ but appear roughly equivalent in terms of fit to the data, are presented as the final summary for a data analysis problem.

Two of the most commonly used methods for **model selection** are **penalized likelihood** and testing of individual coefficient or variance estimate p-values. Other more sophisticated methods include model averaging and cross-validation, but they will not be covered in this text.

15.7.1   Penalized likelihood methods for model selection

Penalized likelihood methods calculate the likelihood of the observed data using a particular model (see chapter 3). But because it is a fact that the likelihood always goes up when a model gets more complicated, whether or not the additional complication is "justified", a model complexity penalty is used. Several different penalized likelihoods are available in SPSS, but I recommend using the **BIC (Bayesian information criterion)**. AIC (Akaike information criterion) is another commonly used measure of model adequacy. The BIC number penalizes the likelihood based on both the total number of parameters in a model and the number of subjects studied. The formula varies between different programs based on whether or not a factor of two is used and whether or not the sign is changed. In SPSS, just remember that "smaller is better".

The absolute value of the BIC has no interpretation. Instead the BIC values can be computed for two (or more) models, and the values compared. A smaller BIC indicates a better model. A difference of under 2 is "small" so you might use other considerations to choose between models that differ in their BIC values by less than 2. If one model has a BIC more than 2 lower than another, that is good evidence that the model with the lower BIC is a better balance between complexity and good fit (and hopefully is closer to the true model of Nature).

In our video game problem, several different models were fit and their BIC values are shown in table 15.5. Based on the "smaller is better" interpretation, the (fixed) interaction between trial and age group is clearly needed in the model, as is the random intercept. The additional complexity of a random slope is clearly not justified. The use of quadratic curves (from inclusion of a trial^2 term) is essentially no better than excluding it, so I would not include it on grounds of parsimony.

Interaction	random intercept	random slope	quadratic curve	BIC
yes	yes	no	no	718.4
yes	no	no	no	783.8
yes	yes	no	yes	718.3
yes	yes	yes	no	727.1
no	yes	no	no	811.8

Table 15.5: BIC for model selection for the video game example.

The BIC approach to model selection is a good one, although there are some technical difficulties. Briefly, there is some controversy about the appropriate penalty for mixed models, and it is probably better to change the estimation method from the default "restricted maximum likelihood" to "maximum likelihood" when comparing models that differ only in fixed effects. Of course you never know if the best model is one you have not checked because you didn't think of it. Ideally the penalized likelihood approach is best done by running all reasonable models and listing them in BIC order. If one model is clearly better than the rest, use that model, otherwise consider whether there are important differing implications among any group of similar low BIC models.

15.7.2   Comparing models with individual p-values

Another approach to model selection is to move incrementally to one-step more or less complex models, and use the corresponding p-values to choose between them. This method has some deficiencies, chief of which is that different "best" models can result just from using different starting places. Nevertheless, this method, usually called **stepwise model selection** , is commonly used.

Variants of step-wise selection include forward and backward forms. Forward selection starts at a simple model, then considers all of the reasonable one-step-more-complicated models and chooses the one with the smallest p-value for the new parameter. This continues until no addition parameters have a significant p-value. Backward selection starts at a complicated model and removes the term with the largest p-value, as long as that p-value is larger than 0.05. There is no guarantee that any kind of "best model" will be reached by stepwise methods, but in many cases a good model is reached.

15.8 Classroom example

The (fake) data in schools.txt represent a randomized experiment of two different reading methods which were randomly assigned to third or fifth grade classrooms, one per school, for 20 different schools. The experiment lasted 4 months. The outcome is the after minus before difference for a test of reading given to each student. The average sixth grade reading score for each school on a different statewide standardized test (stdTest) is used as an explanatory variable for each school (classroom).

It seems likely that students within a classroom will be more similar to each other than to students in other classrooms due to whatever school level characteristics are measured by the standardized test. Additional unmeasured characteristics including teacher characteristics, will likely also raise or lower the outcome for a given classroom.

Cross-tabulation shows that each classroom has either grade 3 or 5 and either placebo or control. The classroom sizes are 20 to 30 students. EDA, in the form of a scatterplot of standardized test scores vs. experimental test score difference are shown in figure 15.7. Grade differences are represented in color and treatment differences by symbol type. There is a clear positive correlation of standardized test score and the outcome (reading score difference), indicating that the standardized test score was a good choice of a control variable. The clustering of students within schools is clear once it is realized that each different standardized test score value represents a different school. It appears that fifth graders tend to have a larger rise than third graders. The plot does not show any obvious effect of treatment.

A mixed model was fit with classroom as the upper level ("subjects" in SPSS mixed models) and with students at the lower level. There are main effects for stdTest, grade level, and treatment group. There is a random effect (intercept) to account for school to school differences that induces correlation among scores for students within a school. Model selection included checking for interactions among the fixed effects, and checking the necessity of including the random intercept. The only change suggested is to drop the treatment effect. It was elected to keep the non-significant treatment in the model to allow calculation of a confidence interval for its effect.

Here are some results:

We note that non-graphical EDA (ignoring the explanatory variables) showed that individual students test score differences varied between a drop of 14 and a
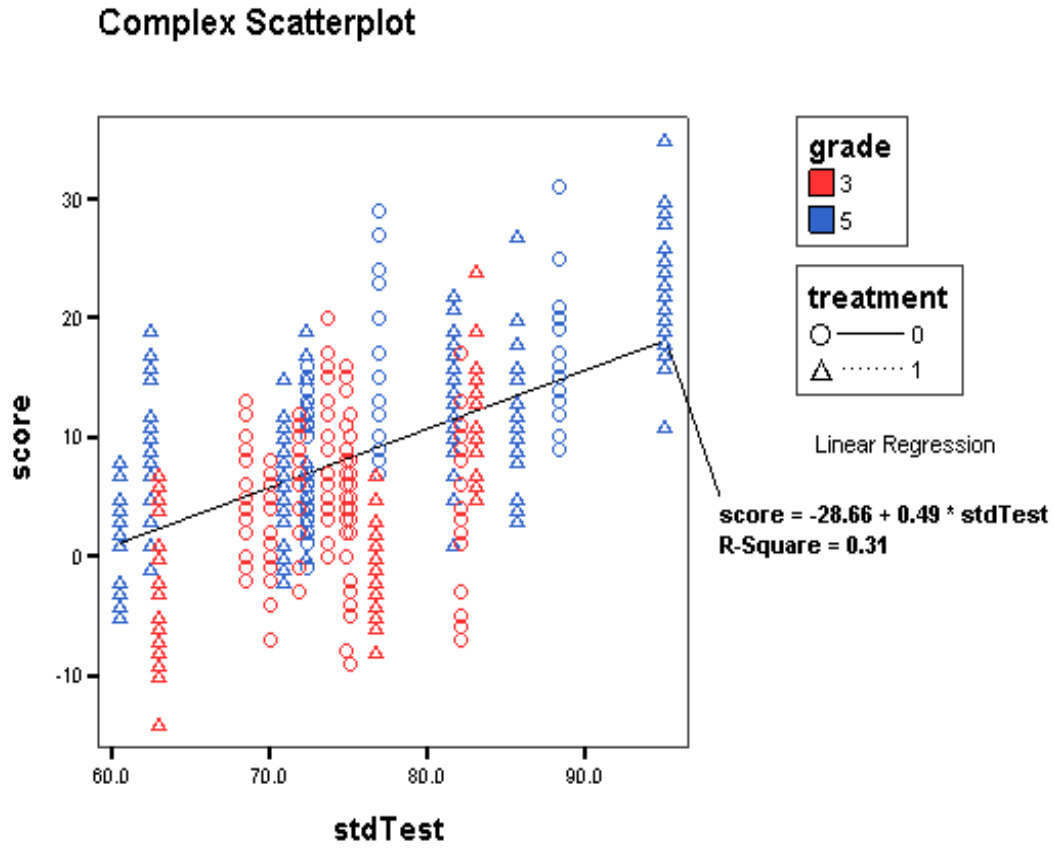
Figure 15.7: EDA for school example

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	15.9	14.3	0.002
grade	1	16.1	12.9	0.002
treatment	1	16.1	1.2	0.289
stdTest	1	15.9	25.6	<0.0005

Table 15.6: Tests of Fixed Effects for the school example.

Parameter	Estimate	Std. Error	df	t	Sig.	95% Conf. Int. Lower Bound	95% Conf. Int. Upper Bound
Intercept	-23.09	6.80	15.9	-3.40	0.004	-37.52	-8.67
grade=3	-5.94	1.65	16.1	-3.59	0.002	-9.45	-2.43
grade=5	0	0	.	.	.	.	.
treatment=0	1.79	1.63	16.1	1.10	0.289	-1.67	5.26
treatment=1	0	0	.	.	.	.	.
stdTest	0.44	0.09	15.9	5.05	<0.0005	0.26	0.63

Table 15.7: Estimates of Fixed Effects for the school example.

Parameter	Estimate	Std. Error	Wald Z	Sig.	95% Conf. Int. Lower Bound	95% Conf. Int. Upper Bound
Residual	25.87	1.69	15.33	<0.0005	22.76	29.40
Intercept(Subject=sc.) Variance	10.05	3.94	2.55	0.011	4.67	21.65

Table 15.8: Estimates of Covariance Parameters for the school example.

rise of 35 points.

The "Tests of Fixed Effects" table, Table 15.6, shows that grade (F=12.9, p=0.002) and stdTest (F=25.6, p<0.0005) each have a significant effect on a student's reading score difference, but treatment (F=1.2, p=0.289) does not.

The "Estimates of Fixed Effects" table, Table 15.7, gives the same p-values plus estimates of the effect sizes and 95% confidence intervals for those estimates. For example, we are 95% confident that the improvement seen by fifth graders is 2.43 to 9.45 *more* than for third graders. We are particularly interested in the conclusion that we are 95% confident that treatment method 0 (control) has an effect on the outcome that is between 5.26 points more and 1.67 points less than treatment 1 (new, active treatment).

We assume that students within a classroom perform similarly due to school and/or classroom characteristics. Some of the effects of the student and school characteristics are represented by the standardized test which has a standard deviation of 8.8 (not shown), and Table 15.7 shows that each one unit rise in standardized test score is associated with a 0.44 unit rise in outcome on average. Consider the comparison of schools at the mean vs. one s.d. above the mean of standardized test score. These values correspond to $\mu_{stdTest}$ and $\mu_{stdTest} + 8.8$. This corresponds to a 0.44*8.8=3.9 point change in average reading scores for a classroom. In addition, other unmeasured characteristics must be in play because Table 15.8 shows that the random classroom-to-classroom variance is 10.05 (s.d.= 3.2 points). Individual student-to-student, differences with a variance 23.1 (s.d. = 4.8 points), have a somewhat large effect that either school differences (as measured by the standardized test) or the random classroom-to-classroom differences.

In summary, we find that students typically have a rise in test score over the four month period. (It would be good to center the stdTest values by subtracting their mean, then rerun the mixed model analysis; this would allow the Intercept to represent the average gain for a fifth grader with active treatment, i.e., the baseline group). Sixth graders improve on average by 5.9 more than third graders. Being in a school with a higher standardized test score tends to raise the reading score gain. Finally there is no evidence that the treatment worked better than the placebo.

> **In a nutshell: Mixed effects models flexibly give correct estimates of treatment and other fixed effects in the presence of the correlated errors that arise from a data hierarchy.**