Exercise 18 - Basics of linear mixed models

Zoltan Kekecs

16 May 2020

Table of Contents

Abstract

In this exercise you will learn important concepts related to linear mixed models. The exercise also describes how to formulate linear mixed models. You will also learn how to make a decision about which random effect term to use, and what to report of the results of mixed models.

Data management and descriptive statistics

Load bullying data

In this exercise we will work with simulated data about bullying. In this dataset we are interested in determining the extent to which body size affects vulnerability to bullying among 4th grade primary school children. Vulnerability to bullying in this study is quantified by the number of lunch sandwiches taken from the child during the period of one month based on self report. The predictor of interest in this study is weight. The

researchers hypothesize that the smaller the child, the more sandwitches will be taken from him or her. Participants come from different classes in the same school which is denoted in the variable 'class'.

The datasets can be downloaded from this link:

https://github.com/kekecsz/SIMM32/blob/master/2020/Lab_4/Data_bully_int.sav

https://github.com/kekecsz/SIMM32/blob/master/2020/Lab_4/Data_bully_slope.sav

Check the dataset

First, we are going to work with the data containing a random intercept (see the meaning below), which is called Data_bully_int.sav.

As always, you should start by checking the dataset for coding errors or data that does not make sense, by eyeballing the data through the data view tool, checking descriptive statistics and through data visualization.

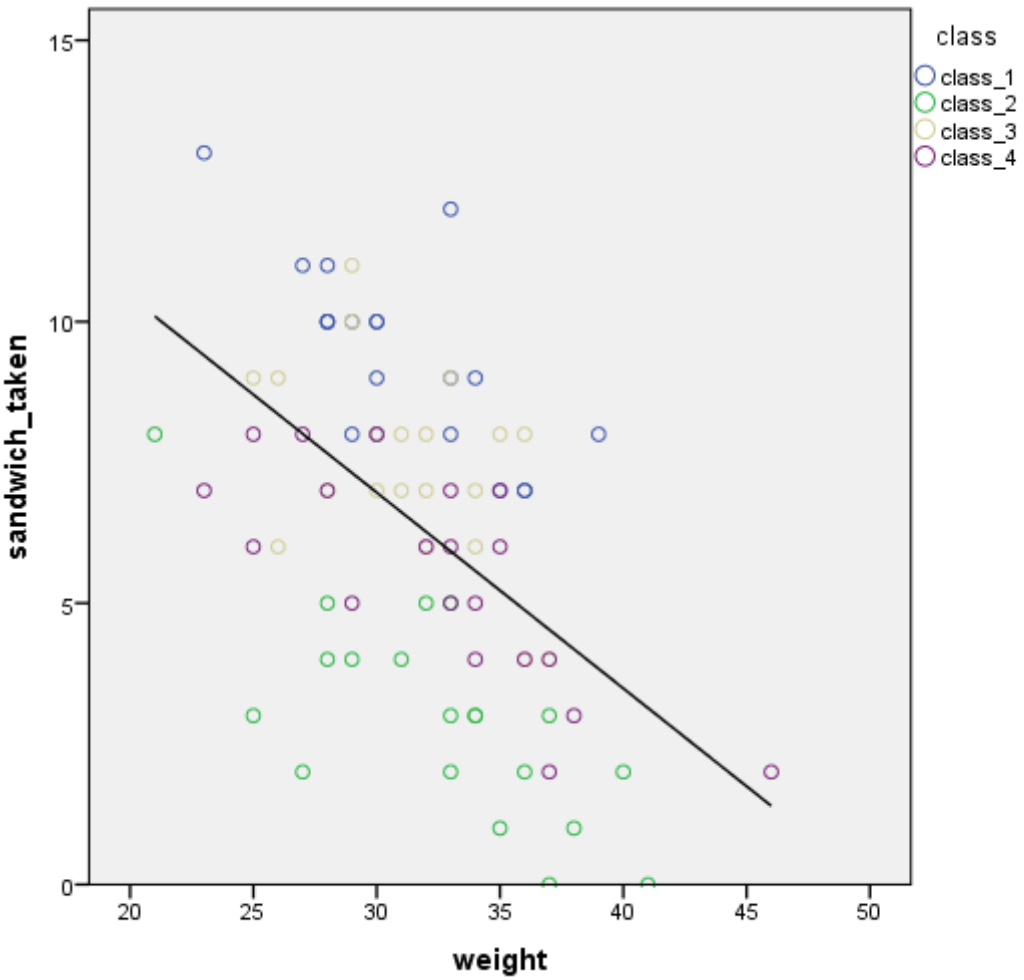
Basics of linear mixed models

Exploring clustering in the data

Let's plot the relationship for the simple regression model of sandwitch_taken predicted by weight on a scatterplot. It seems that there is a clear negative relationship if weight and the number of sandwitches taken from the children, howeverm the variability seems pretty large.

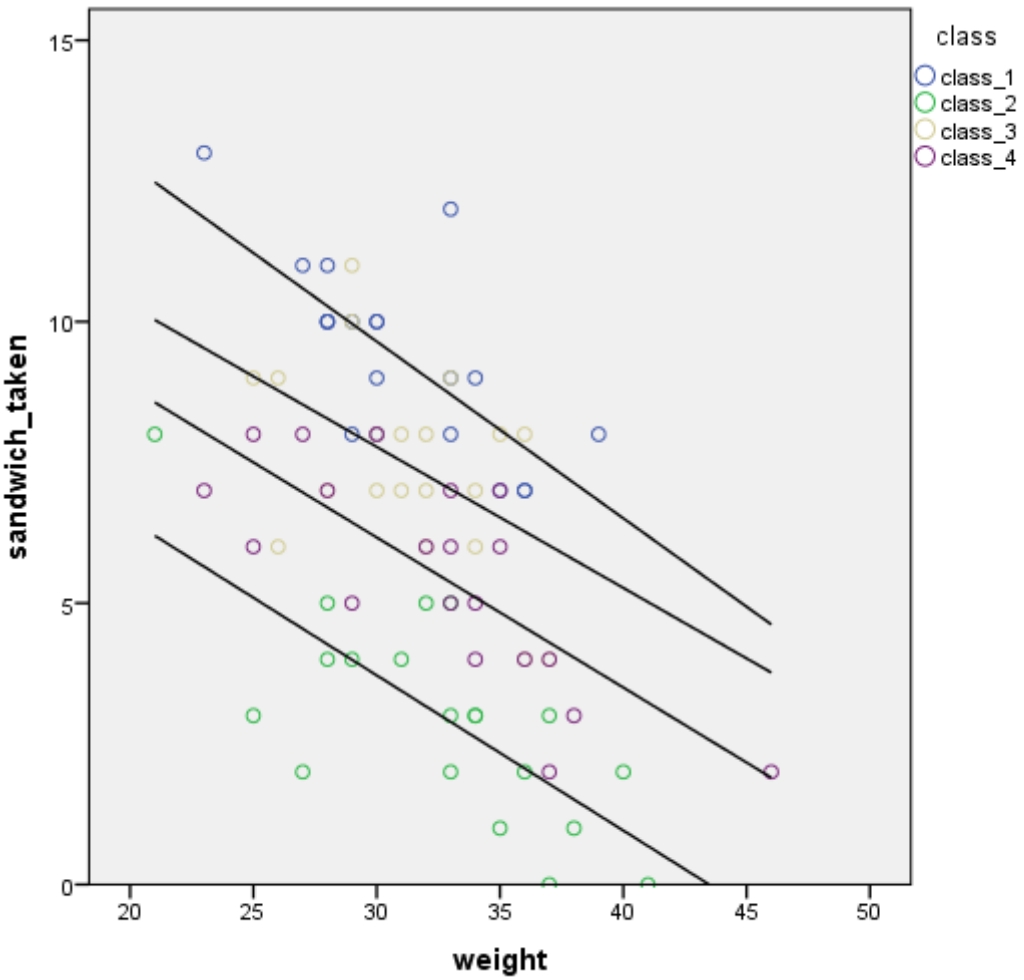
If we look at the color of the dots showing class membership, we may notice that children coming from the same class are grouped together on the scatterplot. This indicates that there is some "clustering" in the data, so observations might not be completely independent from each other.

```
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=weight sandwich_taken
class MISSING=LISTWISE
    REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: weight=col(source(s), name("weight"))
  DATA: sandwich_taken=col(source(s), name("sandwich_taken"))
  DATA: class=col(source(s), name("class"), unit.category())
  GUIDE: axis(dim(1), label("weight"))
  GUIDE: axis(dim(2), label("sandwich_taken"))
  GUIDE: legend(aesthetic(aesthetic.color.exterior), label("class"))
  ELEMENT: point(position(weight*sandwich_taken), color.exterior(class))
  ELEMENT: line(position(smooth.linear(weight*sandwich_taken)))
END GPL
```



Let's see if class membership can explain some of this variability. We can plot the regression lines for each class separately. This seems to be able to explain some of the variability in the data, bringing the regression lines closer to the actual observations. So it seems that it would be worthwile to take into account class membership of the participant when assessing their vulnerability to bullying to get good predictions. (the separate lines only worked for me when I did it through syntax adding the following line to the code:

```
ELEMENT: line(position(smooth.linear(weight*sandwich_taken)), split(class)))
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=weight sandwich_taken
class MISSING=LISTWISE
    REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: weight=col(source(s), name("weight"))
  DATA: sandwich_taken=col(source(s), name("sandwich_taken"))
  DATA: class=col(source(s), name("class"), unit.category())
  GUIDE: axis(dim(1), label("weight"))
  GUIDE: axis(dim(2), label("sandwich_taken"))
  GUIDE: legend(aesthetic(aesthetic.color.exterior), label("class"))
  ELEMENT: point(position(weight*sandwich_taken), color.exterior(class))
  ELEMENT: line(position(smooth.linear(weight*sandwich_taken)), split(class))
END GPL.
```



Mixed models

Right now we are only interested in whether or not weight influences bullying vulnerability, and if so, to what extent. Class membership is secondary to our interests, and even if we would be able to establish the particular effect of any of the classes in this school, in other schools this information would be useless, since there would be different classes.

Using linear mixed models we can take into account that data is clustered according to class membership, without entering class as a classical predictor (fixed effect predictor) in our model. Instead, we can enter it into the model as a random effect.

In this context, we call the effect of predictors on the outcome 'fixed effects'. The effect of clustering variables is called a 'random effect'. We use random effect for predictors for which we have many levels in real life, but we only have information about a small subset of those levels in our dataset, and knowing the effects of these particular levels, like the effect of class 1 in school X, would not be very useful when we use the regression equation on new data, because the new data will most likely not come from class 1 in school X. So

this is more of a nuisance variable for us, we can't do much with knowing its effect, but still we need to take into account that data is clustered according to these levels.

Models that can take into account both fixed and random effects are called mixed effect models.

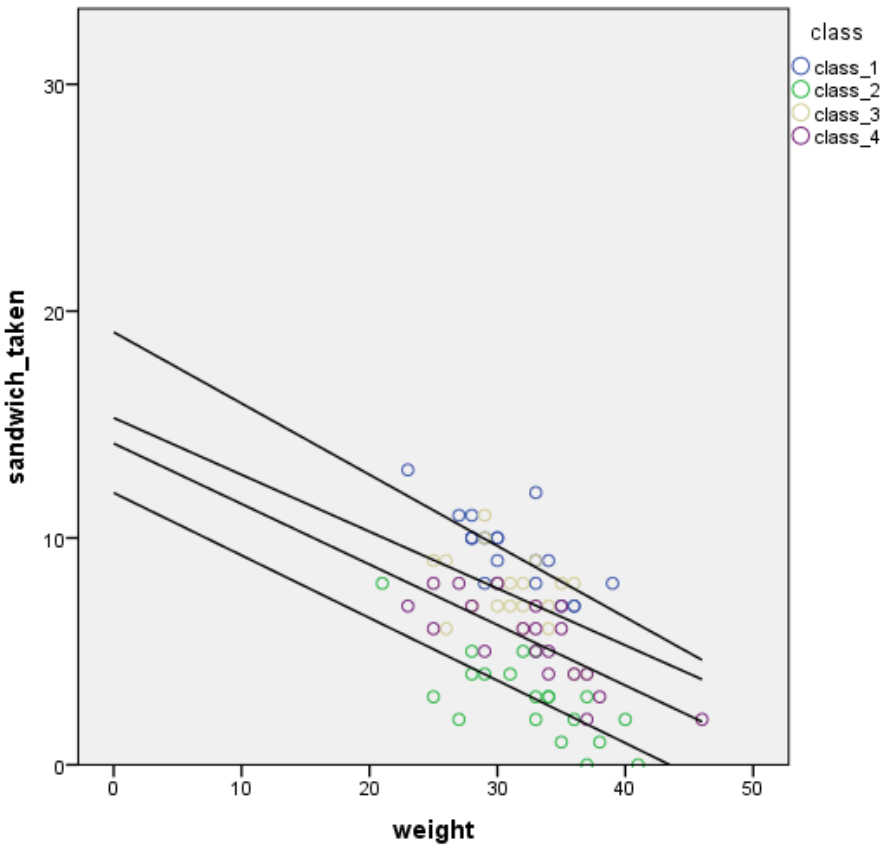
Two types of random effects

There is generally two ways in which clustering (or the random effect term) can influence the outcome variable:

random intercept, but no random slope: it is possible that clusters are only different in how high or low they are on the value of the outcome variable on average, but the effect of the fixed effect predictors are the same in all of the clusters. This is what we see in the dataset named data_bully_int. You can observe that weight affects number of sanwiches taken by rougly the same amount in all classes, but some classes just have fewer sandwiches taken from them in general than others.

On the scatterplot this manifests as the regression lines crossing the y axis at different places (different intercepts), but being almost parallel to each other (similar slopes), indicating that the effect of weight is similar in the classes.

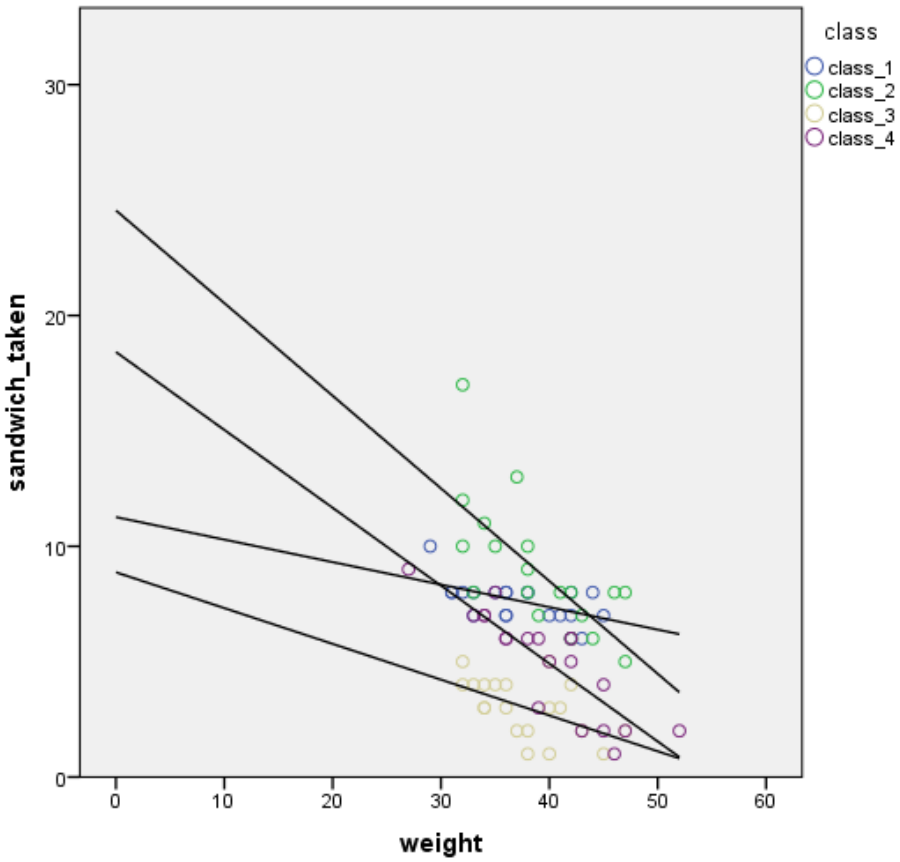
```
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=weight sandwich_taken
class MISSING=LISTWISE
    REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: weight=col(source(s), name("weight"))
  DATA: sandwich_taken=col(source(s), name("sandwich_taken"))
  DATA: class=col(source(s), name("class"), unit.category())
  GUIDE: axis(dim(1), label("weight"))
  GUIDE: axis(dim(2), label("sandwich_taken"))
  GUIDE: legend(aesthetic(aesthetic.color.exterior), label("class"))
  SCALE: linear(dim(1), min(0)  )
  SCALE: linear(dim(2), max(30)  )
  ELEMENT: point(position(weight*sandwich_taken), color.exterior(class))
  ELEMENT: line(position(smooth.linear(weight*sandwich_taken)), split(class))
END GPL.
```



random intercept, AND random slope: Now let's look at the other dataset called Data_bully_slope.sav. This dataset is also simulated, and it comes from the same scenario as the Data_bully_int dataset, but the data looks a bit different. When we plot the regression lines for the classes separately, we find that not only the mean value of the stolen sandwitches is different across classes, but the effect of weight also seems to be different in the different classes.

For example in Class 1, weigth seems to almost make no difference, everyone seems to lose roughly the same number of sanwiches each month regardless of weight. On the other hand in classes 2 and 4, smaller children seem to lose much more sandwitches than their heavy weight classmates.

On the scatterplot, you can easily spot this by noticing that both the intercepts and the slope of the regression lines are different across classes.



Let's see how we can use this knowledge about clustering in the dataset to get more accurate picture about the influence of the predictive value of the predictors on the outcome, and to ultimately make better predictions. We will use the data_bully_slope dataset here.

Now we are going to fit three different models, a simple fixed effect model with no random effects, a **random intercept model**, and a **random slope model**. Remember that based on the above plots, we suspect that the random effect predictor, class, has an effect on both the mean of the outcome (intercept), and the effect of the fixed effect predictor (slope) in the data_bully_slope dataset. So in reality we would probably only fit the random slope model. The other models are just here to show you how they differ in prediction effectiveness and how to formulate them.

First, let's fit the simple linear regression model as we have learned in the previous exercises. Notice that this model only contains a single *fixed effect* predictor, weight, so we will save this regression model to a model called mod_fixed.

simple regression (only fixed effect term, no random effect):

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT sandwich_taken
  /METHOD=ENTER weight.
```

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	129,114	1	129,114	17,315	,000[b]
	Residual	581,636	78	7,457		
	Total	710,750	79			

a. Dependent Variable: sandwich_taken

b. Predictors: (Constant), weight

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	15,887	2,366		6,715	,000	11,177	20,597
	weight	-,254	,061	-,426	-4,161	,000	-,375	-,132

a. Dependent Variable: sandwich_taken

random intercept model:

We have to use a different analysis type to fit a mixed effect model (including both fixed and random effects). This is in **Analyze > Mixed models > Linear**. On the first dialog box, we would enter 'class' in the subjects box, click next, and in the next dialog box, we would specify the model: sandwiches_taken will be the dependent variable, and weight will be a covariate. You will also want to specify some other things in the menus:

Fixed…: Put the predictor weight in the Model box

Random…: In the top of the random effects box, “check include intercept”; under covariance type you should enter: Variance components (this is the default); and in the Subject groupings enter class into the Combinations box.

Statistics…: Check Parameter estimation

If all this is set, we can run the model.

This means that we allow the model to fit a separate regression line to each of the clustering levels (in this case, classes), but we restrict it so that all of these regression lines have to have the same slope.

We would do this if we suspected that the clustering variable (random effect predictor) would have no influence on the effect of the fixed effect predictor. So based on what we saw on the plots above, this would be a good fit for the data_bully_int dataset. But here we fit this to the data_bully_slope dataset, so we can compare the effectiveness of random slope and random intercept models.

The “Estimates of Fixed Effects” table in the output provides us with the model coefficients and the confidence intervals.

```
MIXED sandwich_taken WITH weight
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(10) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0,
   ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED=weight | SSTYPE(3)
  /METHOD=REML
  /PRINT=SOLUTION
  /RANDOM=INTERCEPT | SUBJECT(class) COVTYPE(VC).
```

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	16,351544	1,844368	10,810	8,866	,000	12,283383	20,419705
weight	-,266056	,033328	75,105	-7,983	,000	-,332448	-,199665

a. Dependent Variable: sandwich_taken.

You can see the AIC in the Information Criteria table in the output.

random slope model (allowing BOTH random intercept and random slope):

The model can be fit just like in the case of the random intercept model, but in the Random… menu you would put weight in the Model box.

Fixed…: Put the predictor weight in the Model box

Random…: In the top of the random effects box, "check include intercept"; under covariance type you should enter: Unstructured; in the Model box include weight; and in the Subject groupings enter class into the Combinations box.

Statistics…: Check Parameter estimation

If all this is set, we can run the model.

This means that we allow the model to fit a separate regression line to each of the clustering levels (in this case, classes), and we do not restrict the slope or the intercept of this regression line (other than the line needs to be linear).

We do this because we suspect that the clustering variable (random effect predictor) influences both the mean value of the outcome and the effect of the fixed effect predictors.

```
 DATASET ACTIVATE DataSet2.
MIXED sandwich_taken WITH weight
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(10) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0,
    ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001,
ABSOLUTE)
  /FIXED=weight | SSTYPE(3)
  /METHOD=REML
  /PRINT=SOLUTION
  /RANDOM=INTERCEPT weight | SUBJECT(class) COVTYPE(UN).
```

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	15,888633	3,558453	2,946	4,465	,022	4,444602	27,332664
weight	-,251541	,072336	2,968	-3,477	,041	-,483139	-,019942

a. Dependent Variable: sandwich_taken.

If we would compare the prediction error of the 3 models (if we would ask for the residuals to be saved in the Save menu), we would see that the RSS is largest for the fixed effect model and the smallest for the random slope model.

But this is no surprise, since we allow for increasingly more flexibility in the model to fit our data in the random intercept, and the random slope models. So relying on RSS alone in our original dataset when comparing prediction efficiency would be misleading (we could do that using a test-set like we saw it in exercise 15 – model comparison).

Instead, we can use model fit indices that are sensitive to the number of predictors, such as AIC.

Note that in the case of the random effect models, it is more appropriate to use a model fit index which is corrected for overparametrized models. SPSS provides Schwartz's Bayesian Information Criterion (BIC) Bozdogan's Consistent AIC (CAIC) which can be used for model selection purpose. Both of these have been shown to perform well when the random effects are uncorrelated. However, their performance is poorer and lead to bad decisions in model comparison when the random effects are correlated. So when the correlation of random effects is suspected, the use of regular AIC is advised in SPSS (in other software it is better to use conditional AIC (cAIC, not to be confused with consistent AIC, here refered to as CAIC). More details about the comparative effectiveness of different model fit indices and formulas to compute them can be found in the following publication: (Vallejo, Tuero-Herrero, Núñez & Rosário, 2014). You can get information on the correlation of the random effects by asking for Correlation of parameter estimates in the Statistics menu of the linear mixed model builder dialog box. In the output, you can look at the "Correlation matrix of the estimates of the fixed effects".

Reference: Vallejo, G., Tuero-Herrero, E., Núñez, J. C., & Rosário, P. (2014). Performance evaluation of recent information criteria for selecting multilevel models in behavioral and social sciences. *International Journal of Clinical and Health Psychology*, *14*(1), 48-57.

Determining significance of the model as a whole

To see if the models are better than the null model without any fixed effect predictors (only the random effect of intercept), you can build the null model the same way as a random intercept model, except in the Fixed…: menu you would not enter anything into the Model box (remove weight from there if it is in there). Compare the AIC (or BIC, CAIC as appropriate) of the null model with the model with the fixed effect predictors. Use the regular criteria: smaller AIC is better, and if the difference between the AIC is greater than 2, there is a significant difference between the predictive power of the models.

Deciding whether to use random intercept or slope model

As always, when selecting model components, you should make decisions based on theoretical considerations and prior research results. So in this case, whether it makes sense theoretically for class membership to influence the slope of the effect of weight as well or not, or whether previous exploratory studies have shown that modeling a random slope produces much better model fit than a simple random intercept model.

In some case however, we are exploring the topic without strong prior research or theoretical cues. In these cases it may be appropriate to select between random slope and random intercept models based on model fit. But this decision needs to be clearly documented in the publication. The best if such decisions are pre-registered before data collection. Using such model selection, you could compare model fit indices such as regular AIC in SPSS (or cAIC in R) of the models with different random effect terms as seen before.

In the case of the Data_bully_slope, the AIC of the random slope model is smaller by more than 2, indicating that the random slope model seems to be a slightly better fit.

Visualization also plays an important role in assessing the model fit of mixed effect models.

We can plot the regression lines of the two models in the following way:

- saving the predictions of the models into new variable (be sure to check the predicted values checkbox in the "Predicted values and residuals" box instead of the "predicted values" box, so that you will use the information from the groups (clusters) as well, instead of just using the predictions based on the fixed effects coefficients).

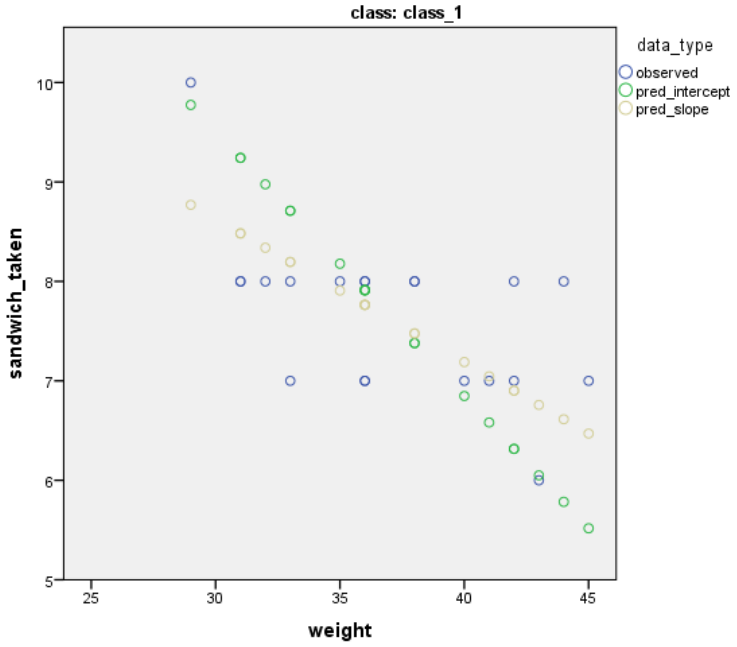
- Restructuring the dataset so that all of the predictions and the actual observed values are in the same column one after the other, and that the corresponding class, and weight information is copied over as well. (You can do this manually or using the Data > Restructure function.
  During this process you will have to create an index variable (for example called "data_type"), which will tell you which lines contain the actual observed sandwiches taken vales, and which contain the predictions from which models. Even if you use a numerical variable to make this distinction, it is best if you define the vale labels in the variable view so that it is obvious which data point corresponds to which prediction or the observed data.

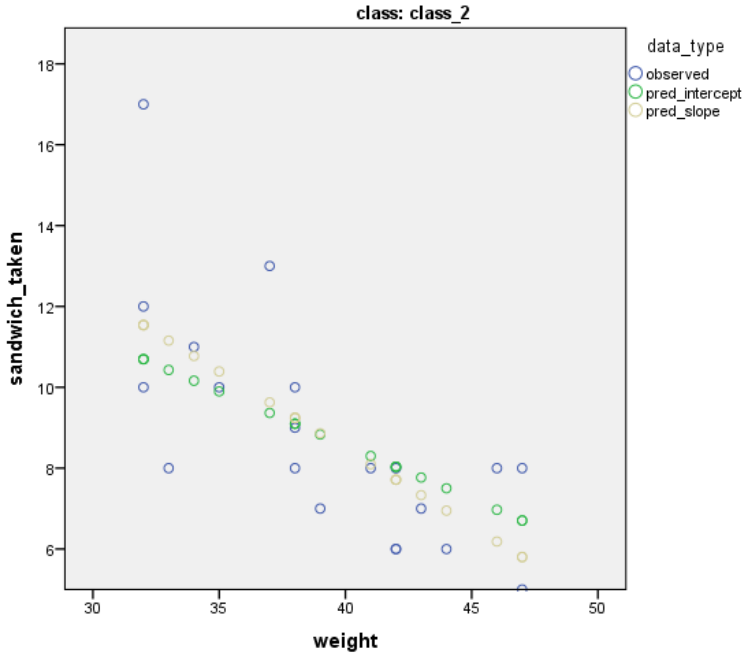
- Split data by Class using the Data > Split data function (you will have to allow SPSS to sort the cases by this variable)

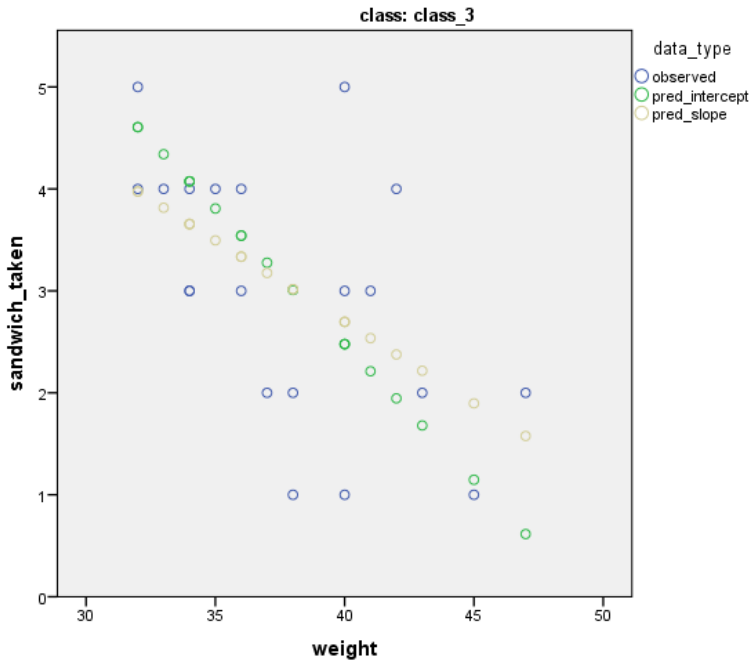
- Use the chart builder to build either a grouped scatterplot or a grouped line plot, where the y axis contains the sandwiches taken values (predicted and observed), the x axis contains weight, and the color will be defined by data_type (the index variable).

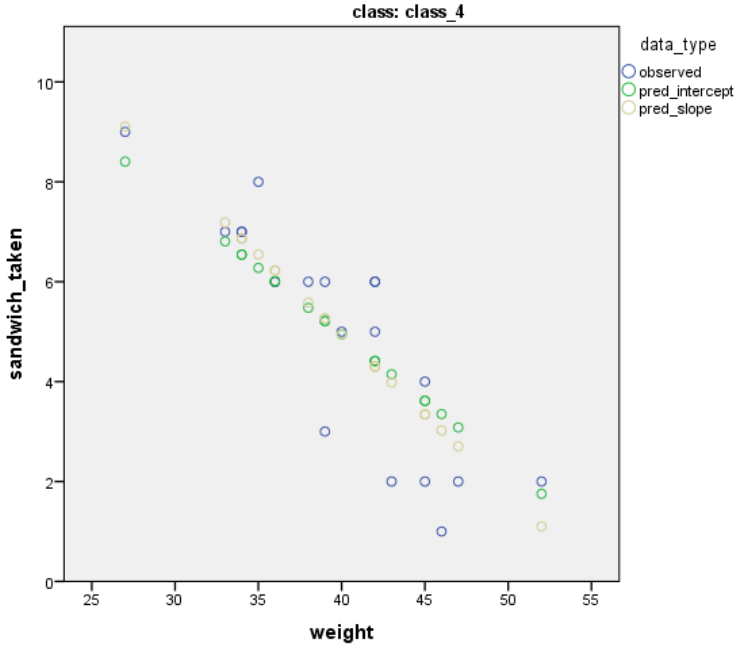
This way, for each class, you will get a plot depicting the actual observed values, and the predicted values by the random intercept and the random slope model. Now you can look at these plots to see which model's predictions match the observed data better. The random slope model will always be somewhat better, but if the difference is not

substantial, you might decide to use the random intercept model, to restrict the flexibility of the model and avoid overfitting.






class: class_3
data_type
observed
pred_intercept
pred_slope
sandwich_taken
weight
0
1
2
3
4
5
30
35
40
45
50


class: class_4
data_type
observed
pred_intercept
pred_slope
sandwich_taken
weight
0
2
4
6
8
10
25
30
35
40
45
50
55

When comparing the plots, we can see a slight improvement in the model fit in the random slope model compared to the random intercept model, but the improvement is not too impressive. We should choose the model which makes more sense theoretically.

What to report

We have to report the same information about linear mixed models as for fixed-effect-only linear models seen in the previous exercises.

You could describe the following in your methods section about the model formulation:

"In order to determine the influence of weight on vulnerability to bullying, we used a linear mixed model. In the model the outcome variable was the number of sandwitches taken, and we used a single fixed effect predictor, weight. Because data was clustered in school classes, we included the random effect of class in the model. No prior data or theory was available about how the random effect of class might manifest, so we built two separate models. In one model we only allowed for a random intercept of classes, while in the other model we allowed for both random intercept and the random slope of the effect of weight across different classes. As pre-registered in our experimental protocol, we compared the model fit of the random intercept and slope models using the AIC model fit index, and we made a choice between these two models based on this index."

Model R squared:

We need to use a special type of the R squared called the marginal R squared in the case of mixed effects models that shows the proportion of variance explained by the fixed factor(s) alone, when not taking into account the random effect terms.

Reference: Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 4(2), 133-142. and Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. Methods in Ecology and Evolution, 5(9), 944-946.

Marginal R squared (variance explained by the fixed effects in the model) can be manually computed in SPSS using the formulae in Nakagawa & Schielzeth (2013). When using a Gaussian model without random slopes (just random intercept), the formula is: R^2m = Vf / (Vf + Vr + Ve).

Here, Vf is the fixed effects variance, which can be calculated by saving the fixed effect predicted values based on the model (ask for this in the Save menu, make sure to ask for the predicted values in the "Fixed predicted values" box instead of the "Predicted values and residuals" box!), and calculating their variance (for example by using the Analyze > Descriptive Statistics > Descriptives command, and in the options asking for variance.

Vr is the sum of the random effects variances ,and Ve is the residual variance. This can be found in the Estimates of Covariance Parameters table in the output of the linear mixed model.

Estimates of Covariance Parameters[a]

Parameter	Estimate	Std. Error

Residual		1,491078	,155517
Intercept [subject = hospital]	Variance	,163902	,114583

a. Dependent Variable: pain.

Conditional R^2 (variance explained by both the fixed and random effects in the model) can be computed with this formula:

R^2c = (Vf + Vr) / (Vf + Vr + Ve)

Source:

https://ecologyforacrowdedplanet.wordpress.com/2013/08/27/r-squared-in-mixed-models-the-easy-way/

I don't know the formula to compute marginal R squared for random slope models, so in the case of a random slope model, R squared information can be computed in another statistical software, or left unreported.

In the results section, you should report the linear mixed models analysis as follows:

"The random slope model produced a better model fit according to the AIC (AIC intercept = 308.72, AIC slope = 305.44). Which slightly favors the random slope model, but the visual inspection did not reveal substantial benefit for applying the random slope model, so in order to restrict the flexibility of the model to avoid overfitting, we chose the random intercept model. Thus, we present the results of the random intercept model in the following. (The results of the random slope model are listed in the supplement.)

The random intercept linear mixed model was significantly better than the null model AIC (AIC intercept = 308.72, AIC null = 350.47), where the fixed effect predictor, weight, explained 16.58% of the variance of sandwiches taken (marginal R^2 = 0.17)."

You will also have to report statistics related to the predictors, this is usually done in a table format, because most often we have multiple predictors (even though in this example we only have one). You can get the information about the important results related to the predictors by asking for Parameter estimates in the Statistics menu. The model coefficients and confidence intervals then can be found in the estimated fixed effects table in the output.

Estimates of Fixed Effects[a]

						95% Confidence Interval	
Parameter	Estimate	Std. Error	df	t	Sig.	Lower Bound	Upper Bound
Intercept	16,351544	1,844368	10,810	8,866	,000	12,283383	20,419705
weight	-,266056	,033328	75,105	-7,983	,000	-,332448	-,199665

a. Dependent Variable: sandwich_taken.

Note that the use and interpretation of p-values for linear mixed models is controversial at the moment, so observe the trend in your particular sub-field and decide whether you want to use them or not. Confidence intervals give you information about statistical significance, so it is not necessary to provide p-values.

Standardized betas cannot be extracted in SPSS for mixed linear models, so these can be omitted from the report (or can be computed in another statistical software).