Exercise 19 - Repeated measures analysis with mixed models

Zoltan Kekecs

May 19 2020

Table of Contents

Abstract

This exercise is focused on the use of linear mixed models in case of repeated measures designs, and how to re-structure data from a 'wide format' to a 'long format' to be useful for linear mixed model analysis

Data management and descriptive statistics

Load wound healing data

In this exercise we will work with simulated data about wound healing over time after a surgical procedure.

The data can be downloaded from:

https://github.com/kekecsz/SIMM32/blob/master/2020/Lab_5/Wound_healing.sav

We know that psychological factors, especially stress, can influence recovery after surgery, and the rate of wound healing. Let's say that we have a theory that it is important for hospitalized patients to have a connection with the outside world. So we may think that patients who have a window close to their hospital bed may have a better mood and thus, would show a faster recovery after surgery. This hypothesis is tested in a simple study looking at whether the distance of the patients' bed from the closest window would predict rate of wound healing. Distance is measured in meters, and wound healing is measured by rating the wound using a standardized wound rating measure taking into account the size of the wound, its inflammation and scarring. A physician rates the wound each day for seven days in the afternoon at the same time of the day. We will use this variable as our outcome measure.

Our hypothesis extends to the role of sunlight in this context, where we suppose that the more sunlight a patient gets the better their recovery would be. To test this hypothesis, our model will take into account whether the bed of the patient is in the north wing or the south wing of the hospital (since the hospital is in the northern hemisphere, we can assume that patients in the south wing would get more sunlight overall during their hospital stay).

To begin with, we will dummy-code the location variable. In this process I created a variable called "south_wing", where south wing is coded as 1 and north wing is coded as 0.

```
RECODE location ('south_wing'=1) ('north_wing'=0) INTO south_wing.

EXECUTE.
```

Lets inspect the layout of this dataset. Notice that each row contains all the data collected from the same participant, and the wound rating data for each day are stored in variables 'day_1', 'day_2', …, 'day_7' respectively.

Check the dataset

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.

Exploring clustering in the data

One exploratory analysis we can perform when dealing with data with repeated measures from the same unit of observation is to look at the correlation between the repeated measurements. This is also called auto-correlation. If the dataset is structured the way it is structured in our example, we can assess this suing a simple correlation procedure in SPSS using **Analyze > Correlate > Bivariate**, and entering the variables day 1 through day 7. Notice that the repeated measures data points are highly correlated. This shows that the different observations of wound rating are definitely not independent from each other. If we know the previous day's would rating, we also get at least some information about the next day's wound rating as well. This is normal, since the wound rating and the initial size of the incision and the wound healing rate depends on the patient. So this is clustered data.

Just like the data in the previous exercise was clustered in school classes, here, the data is clustered within participants!

```
CORRELATIONS
  /VARIABLES=day_1 day_2 day_3 day_4 day_5 day_6 day_7
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlations

		day_1	day_2	day_3	day_4	day_5	day_6	day_7
day_1	Pearson Correlation	1	,851**	,657**	,547**	,465**	,385*	,271
	Sig. (2-tailed)		,000	,000	,002	,010	,036	,148
	N	30	30	30	30	30	30	30
day_2	Pearson Correlation	,851**	1	,836**	,769**	,593**	,468**	,346
	Sig. (2-tailed)	,000		,000	,000	,001	,009	,061
	N	30	30	30	30	30	30	30
day_3	Pearson Correlation	,657**	,836**	1	,862**	,708**	,602**	,441*
	Sig. (2-tailed)	,000	,000		,000	,000	,000	,015
	N	30	30	30	30	30	30	30
day_4	Pearson Correlation	,547**	,769**	,862**	1	,854**	,718**	,602**
	Sig. (2-tailed)	,002	,000	,000		,000	,000	,000
	N	30	30	30	30	30	30	30
day_5	Pearson Correlation	,465**	,593**	,708**	,854**	1	,890**	,798**
	Sig. (2-tailed)	,010	,001	,000	,000		,000	,000
	N	30	30	30	30	30	30	30
day_6	Pearson Correlation	,385*	,468**	,602**	,718**	,890**	1	,904**
	Sig. (2-tailed)	,036	,009	,000	,000	,000		,000
	N	30	30	30	30	30	30	30
day_7	Pearson Correlation	,271	,346	,441*	,602**	,798**	,904**	1
	Sig. (2-tailed)	,148	,061	,015	,000	,000	,000	
	N	30	30	30	30	30	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Repeated measures analysis using linear mixed models

Reshape dataset

Because of this clustering, we need to use a mixed linear model to predict wound rating instead of running a regular fixed-effect-only procedure. However, first we need to re-structure the dataset to a format that will be interpretable to the linear mixed effect regression.

At this point, the dataset contains 7 observations of wound rating from the same participant in one row (one for each day of the week while data was collected). This format is called the **wide format**.

For the mixed model analysis to be able to interpret the data correctly, we will have to restructure the dataset so that each row contains a single observation. This would mean that each participant would have 7 rows instead of 1. Data in the variables ID, distance_window, and south_wing would be duplicated in each of the rows of the same participant, and there would be only a single column for wound rating in this new dataset. This format of the data is usually referred to as the **long format**.

We can do this using the **Data > Restructure** function. You will need to follow the following steps:

- Restructure selected variables into cases

Number of variable Groups:

- Indicate that we want to restructure only one variable group (the wound healing data)
- We can use the variable "ID" for case group identification
- The Target variable should be named wound_healing
- Put variables day_1, day_2, day_3, day_4, day_5, day_6, day_7 into the variables to be transformed box
- Select distance_window and south_wing as fixed variables (you can add location as well if you want, but it is not necessary, we have already re-coded it)

Create index variables

- Indicate that you only want one index variable (time, which will represent which day the data was measured)

Create one index variable

- Sequential numbers is ok

- In the “Edit the index variable and label” box, change the name of the index variable to time

Options

- In the options screen everything can remain as default, you can click Finish

This syntax will produce the same effect:

```
VARSTOCASES
 /MAKE wound_healing FROM day_1 day_2 day_3 day_4 day_5 day_6 day_7
 /INDEX=time(7)
 /KEEP=ID distance_window south_wing
 /NULL=KEEP.
```

Explore how this new data structure looks like. It should have 7 lines for each participant ID, adding up to a total of 210 rows (7*30) in this particular example.

Building the linear mixed model

Now that we have our dataset in an approriate formate, we can build our prediction model. The outcome will be wound rating, and the fixed effect predictors will be day after surgery, distance of the bed from the window, and whether a person was in the south wing or not (these information are stored in the variables time, distance_window, and south_wing).

Since our outcome is clustered within participants, the random effect predictor will be participant ID. As in the previous exercise, we will fit two models, one will be a random intercept model, the other a random + intercept slope model (for simplicity, we will call this the random slope model).

Note that the **random intercept model** means that we suspect that the each participant is different in their overall wound rating (or baseline wound rating, basically the size of the surgical incision wound), but that the effect of the fixed effect predictors (time, distance from window, and location) is the same for each participant. On the other hand, the **random slope model** not only baseline wound rating will be different across participants, but also that the effects of the fixed effect predictors will be different from participant to participant as well.

We can start building these models in Analyze > Mixed models > Linear. For the **random intercept** model we should do the following:

Specify subjects and repeated:

- Put “ID” into the subjects field (you don’t have to put anything in the repeated field)

Linear mixed models:

- Set wound_healing as the dependent
- Distance_window and time should be a covariate, because they are continuous measures
- South_wing should be a Factor
- Fixed menu
    - Put distance_window, time, and south_wing into the model box, since these will be our fixed effect predictors (don’t enter interaction terms, so make sure to select the main effects from the dropdown menu between the boxes)
- Random menu
    - Check the include intercept checkbox, and put ID into the combinations field in the subjects grouping box, this way indicating that the data is clustered within participants
    - For the random intercept model, it does not matter what you choose for covariance type, so it can be left as default.
- Statistics menu
    - Ask for parameter estimates

```
MIXED wound_healing BY south_wing WITH distance_window time

  /CRITERIA=DFMETHOD(SATTERTHWAITE) CIN(95) MXITER(100)
MXSTEP(10) SCORING(1)

    SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE) LCONVERGE(0,
ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)

  /FIXED=time south_wing distance_window | SSTYPE(3)

  /METHOD=REML

  /PRINT=CORB  SOLUTION

  /RANDOM=INTERCEPT | SUBJECT(ID) COVTYPE(VC)

  /SAVE=PRED.
```

Note that here we have 3 different fixed effect predictors, so we could in theory allow for a random slope for any or all of them. The issue is though that each of the random effects is a separate parameter in the model, and as we have seen in the model comparison exercise, overparameterization can lead to overfitting.

So when we ask for a random intercept in our example data, we are estimating 30 different parameters (since we have 30 clusters – the 30 participants). When we also allow for random slope for one of the predictors, we will estimate another 30 parameters, since we estimate the effect of time for each participant separately. So if we allow for the random slope of all of the predictors in our model and the random intercept, we will estimate 120 parameters (30 for the intercept and 3 * 30 for each slope). This is getting very close to the number of observations we have, and increases the possibility of overfitting. You can see from this that it is very easy to reach the number of observations with the number of predictors. Thus, it is very uncommon in papers that you would see multiple random slopes. Usually, the researchers select 1 or 2 variables for which a random slope will be allowed at most, but even more common are simple random intercept models.

Also, in our example the predictor's values are the same for all measurement times, except for "time" itself. So if we were to fit a different slope for each participant, it would be considerably less effective than fitting a slope for all participants combined.

In our particular case, we will allow for a single random slope: the slope of the effect of time.

The random slope model is built the same way as the random intercept model, except that in the Random menu we would move time into the Model field, and choose Unscructured in as covariance type. This way we allow for the effect of time to be different across participants (basically saying that the rate of wound healing can be different from person to person), but restrict the model to predict the same effect for the other two fixed predictors: distance_window and location. By choosing unstructured covariance type, we allow for the random slopes and random intercepts to be correlated. In SPSS 26, you may have to paste the analysis into the syntax and manually edit the COVTYPE() parameter to UN, so that you get the following syntax:

```
MIXED wound_healing BY south_wing WITH distance_window time

 /CRITERIA=DFMETHOD(SATTERTHWAITE) CIN(95) MXITER(100)
MXSTEP(10) SCORING(1)

  SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE) LCONVERGE(0,
ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)

 /FIXED=time south_wing distance_window | SSTYPE(3)

 /METHOD=REML

 /PRINT=CORB  SOLUTION

 /RANDOM=INTERCEPT time | SUBJECT(ID) COVTYPE(UN)

 /SAVE=PRED.
```

Comparing models

Now let's compare the model predictions of the different random effect models to see which one fits the data better.

Visualization is very important when working with mixed effect models. First, let's visualize the predictions. For this we will have to save the predicted values (from the "predicted values and residuals" box) into new variables. I will rename the new variables as pred_intercept and pred_slope so that I will know later which prediction comes from which model.

Then, we should restructure the dataset to put the observed and predicted wound healing values in the same column, and create a new variable which will indicate that the given value is the observed wound heling, the predicted wound healing by the random intercept model, or the predicted wound healing by the random slope model. Also, the participant ID needs to indicate which value belongs to which participant.

IMPORTANT! Before you restructure your dataset like this, save your current dataset so we can revert back to it after the visualization.

We can use the **Data > restructure** command again to do this. The target variable is wound healing, which should contain wound_healing, predicted wound healing by the random intercept model, and the predicted wound healing by the random slope model. The index variable I will call obs_or_pred, which will indicate the source of each value. I will use the previous variable names for the indexing (wound_healing, pred_intercept, pred_slope).

```
VARSTOCASES
 /MAKE wound_healing FROM wound_healing pred_intercept pred_slope
 /INDEX=obs_or_pred(wound_healing)
 /KEEP=ID distance_window south_wing time
 /NULL=KEEP.
```

Now we will split the file and organize it by the ID variable.

```
SORT CASES  BY ID.
SPLIT FILE SEPARATE BY ID.
```

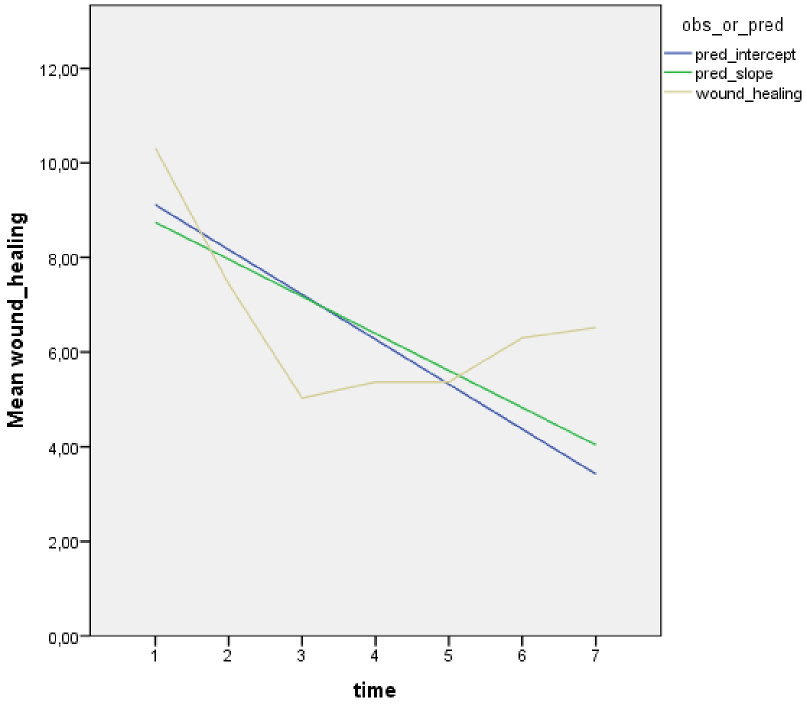
Finally, we can build line plots in the chart builder, including the actual wound healing and predicted wound healing values on the same plot marked by different colors, and we can get this for each participant separately.

By looking at the graphs, we can get a sense of whether the predictions fit the underlying pattern of the data well, and which model is better at this.

Based on the graphs, we can see that the predictions of the random intercept and slope models are almost identical.

```
GGRAPH

  /GRAPHDATASET NAME="graphdataset" VARIABLES=time
MEAN(wound_healing)[name="MEAN_wound_healing"]

    obs_or_pred MISSING=LISTWISE REPORTMISSING=NO

  /GRAPHSPEC SOURCE=INLINE.

BEGIN GPL

  SOURCE: s=userSource(id("graphdataset"))

  DATA: time=col(source(s), name("time"), unit.category())

  DATA: MEAN_wound_healing=col(source(s), name("MEAN_wound_healing"))

  DATA: obs_or_pred=col(source(s), name("obs_or_pred"), unit.category())

  GUIDE: axis(dim(1), label("time"))

  GUIDE: axis(dim(2), label("Mean wound_healing"))

  GUIDE: legend(aesthetic(aesthetic.color.interior), label("obs_or_pred"))

  SCALE: linear(dim(2), include(0))

  ELEMENT: line(position(time*MEAN_wound_healing),
color.interior(obs_or_pred), missing.wings())

END GPL.
```

The plots should look like this (I only copied one of the 30 plots here), each plot representing data corresponding to a different participant:



The difference between the predictions of the two models is unremarkable.

Furthermore, we can compare the model fit indices (such as AIC) of the two models to get further information about the model fit of the two models in comparison to each other.

Intercept model:

Information Criteria[a]

-2 Restricted Log Likelihood	772,937
Akaike's Information Criterion (AIC)	776,937
Hurvich and Tsai's Criterion (AICC)	776,996
Bozdogan's Criterion (CAIC)	785,592
Schwarz's Bayesian Criterion (BIC)	783,592

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: wound_healing.

Slope model:

Information Criteria[a]

-2 Restricted Log Likelihood	770,279

Akaike's Information Criterion (AIC)	778,279
Hurvich and Tsai's Criterion (AICC)	778,478
Bozdogan's Criterion (CAIC)	795,591
Schwarz's Bayesian Criterion (BIC)	791,591

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: wound_healing.

None of these methods indicate a significant difference between the prediction efficiency of the models. So in this particular sample there is not too much benefit for assuming a different slope of time for each participant. But this does not necesseraly mean that there is no point of using it in another sample. Previous studies and theory needs to be evaluated as well.

For now, without any prior knowledge from the literature, we can continue using the random intercept model.

Adding the quadratic term of time to the model

While exploring the plots we might notice that there is a non-linear relationship between time and wound rating. It seems that wounds improve fast in the first few days, and the healing is slower in the days after that.

Let's add the quadratic term of time to the random intercept model to account for this non-linear relationship. As you could see in exercise 14 – model diagnostics, adding higher order terms as predictors can create structural multicollinearity, so to avoid that we should center our variable: time before we compute its squared value and save it as a new variable (see all this in exercise 14).

```
DESCRIPTIVES VARIABLES=time

  /STATISTICS=MEAN STDDEV MIN MAX.


COMPUTE time_centered=time-4.

EXECUTE.


COMPUTE time_centered_sq=time_centered*time_centered.

EXECUTE.
```

So now that we have computed the centered value of time and its squared value, we can build a new mixed model, with wound_healing as a dependent, and south_wing, distance_window, time_centered, and time_centered_sq as the fixed effect predictors. We only allow for random intercept here, as we have seen above that a random slope does not make a big difference. Again, we save the predicted values from the "predicted values and residuals" box for visualization purposes.

```
MIXED wound_healing BY south_wing WITH time_centered time_centered_sq
distance_window

  /CRITERIA=DFMETHOD(SATTERTHWAITE) CIN(95) MXITER(100)
MXSTEP(10) SCORING(1)

    SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE) LCONVERGE(0,
ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)

  /FIXED=south_wing time_centered time_centered_sq distance_window |
SSTYPE(3)

  /METHOD=REML

  /PRINT=CORB  SOLUTION

  /RANDOM=INTERCEPT | SUBJECT(ID) COVTYPE(VC)

  /SAVE=PRED.
```

Now let's look at the AIC, and also, plot the prediction of this model against the actual observations by participant just like we did previously, to compare the effectiveness of this model to the effectiveness of the random intercept model which only included the first order term of time.

Both the AIC and the graphs indicate that the model containing the quadratic term of time is a better fit to the data than the one only containing the first order term of time.

Information Criteria[a]

-2 Restricted Log Likelihood	625,677
Akaike's Information Criterion (AIC)	629,677
Hurvich and Tsai's Criterion (AICC)	629,737
Bozdogan's Criterion (CAIC)	638,323
Schwarz's Bayesian Criterion (BIC)	636,323

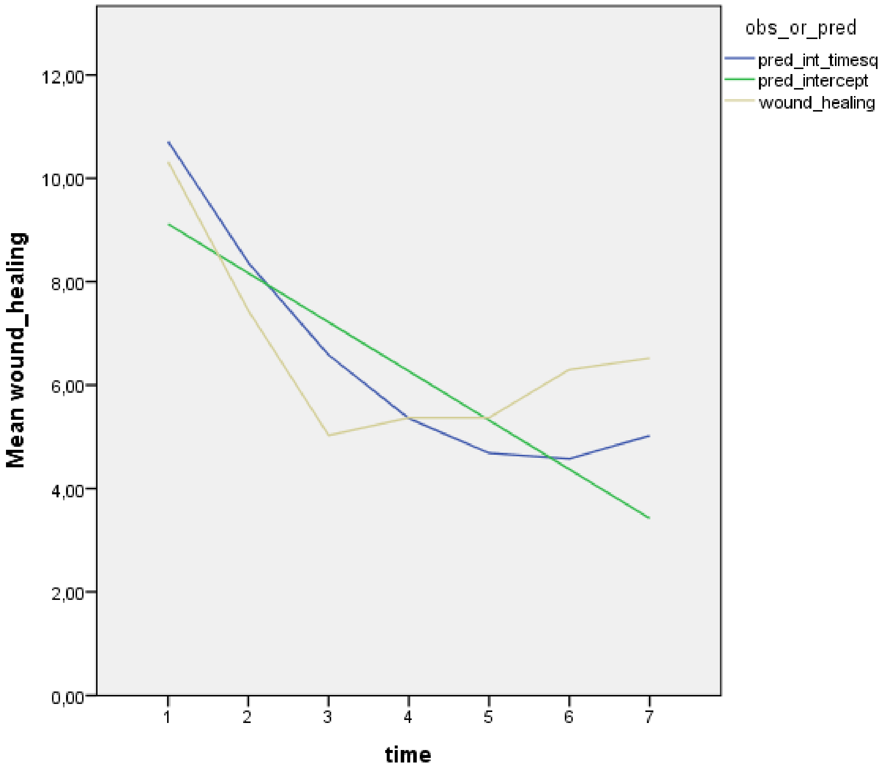
The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: wound_healing.

```
VARSTOCASES
 /MAKE wound_healing FROM wound_healing pred_intercept pred_int_timesq
 /INDEX=obs_or_pred(wound_healing)
 /KEEP=ID distance_window south_wing time time_centered time_centered_sq
 /NULL=KEEP.
```

A

```
SORT CASES  BY ID.

SPLIT FILE SEPARATE BY ID.


GGRAPH

  /GRAPHDATASET NAME="graphdataset" VARIABLES=time
MEAN(wound_healing)[name="MEAN_wound_healing"]

    obs_or_pred MISSING=LISTWISE REPORTMISSING=NO

  /GRAPHSPEC SOURCE=INLINE.

BEGIN GPL

  SOURCE: s=userSource(id("graphdataset"))

  DATA: time=col(source(s), name("time"), unit.category())

  DATA: MEAN_wound_healing=col(source(s), name("MEAN_wound_healing"))

  DATA: obs_or_pred=col(source(s), name("obs_or_pred"), unit.category())

  GUIDE: axis(dim(1), label("time"))

  GUIDE: axis(dim(2), label("Mean wound_healing"))

  GUIDE: legend(aesthetic(aesthetic.color.interior), label("obs_or_pred"))

  SCALE: linear(dim(2), include(0))

  ELEMENT: line(position(time*MEAN_wound_healing),
color.interior(obs_or_pred), missing.wings())

END GPL.
```



The results indicate that a model taking into account the nonlinear relationship of time and wound rating produces a significantly better fit to the observed data than a model only allowing for a linear trend of time and wound healing.

The fit seems reasonable, so we stop here and decide that this will be our final model.

As always, you will need to run model diagnostics before reporting your final results. The next exercis will conver this topic.