A tutorial on Generalised Additive Mixed Effects Models for bilingualism research

Motivation

...

Generalised additive (mixed) models

...

Pre-requisites

This tutorial assumes readers are already familiar with R and have at least some experience fitting linear models, including models with random effects (variably known as mixed, hierarchical, nested, multi-level models. The following packages need to be installed:

- tidyverse
- mgcv
- tidygam

You can find the tutorial code and data here: XXX

Case study 1: U-shaped learning

U-shaped learning

[Lit review]

The data

We have simulated data of learning scores from 200 subjects, taken at 10 time points. A proficiency score was also included for each subject at each time point. The data is intended to simulate a study in which participants perform a learning task and take a proficiency test at 10 consecutive time points. This is what the data looks like.

```
dat1
```

```
# A tibble: 2,200 x 4
   score proficiency subj   time
   <dbl>       <dbl> <fct> <int>
 1  26.7      -1.37  s1        0
 2  26.4      -1.29  s1        1
 3  26.0      -1.21  s1        2
 4  25.7      -1.14  s1        3
 5  25.5      -1.06  s1        4
 6  25.2      -0.983 s1        5
 7  25.0      -0.906 s1        6
 8  24.9      -0.830 s1        7
 9  24.6      -0.753 s1        8
10  24.5      -0.677 s1        9
# i 2,190 more rows
```

The `score` column contains the learning scores, while the `proficiency` column the proficiency scores. The subject ID is given in `subj`. time point 0 to 9 is in `time`. Figure 1 shows the relationship between proficiency and learning scores for individual subjects. From the figure, it is clear that such relationship has a U-shape, by which learning scores intially decrease, plateau, then increase again with higher proficiency.

In the following sections we will analyse this data using GAMMs. For pedagogical purposes, we will first focus on the effect of proficiency scores on learning scores at a single time point, time point 5. Subsequently, we will analyse the entire data, to illustrate how to conduct a time-series analysis.

Modelling a non-linear effect

Let's first focus on how to model a non-linear effect: here, we can look at the effect of the proficiency score on the learning score. As we have seen in Figure 1, the effect is U-shaped, i.e. it is not linear. To simplify things, let's look only at data from time point. We will extend the analysis to also include time point as a predictor later. We can model a non-linear effect of proficiency on the learning score with the following code.

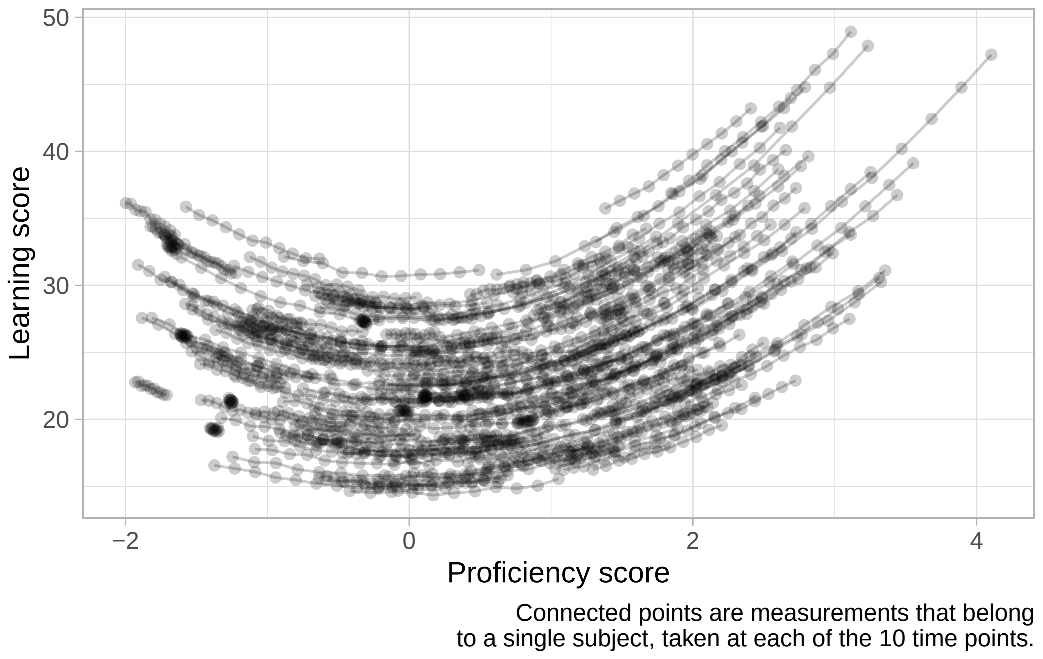

Figure 1: **?(caption)**

```r
# attach the mgcv package
library(mgcv)

# fit the model
gam_1 <- gam(
  score ~ s(proficiency),
  data = dat1
)
```

The formula states that `score`, the outcome variable (also known as the dependent variable) should be modelled as a function of `proficiency`, but we use the `s()` to indicate that we want to estimate a (potentially) non-linear effect. The name of the function, `s`, stands for "smooth term". Smooth terms (aka smoothers) are mathematical objects that allow GAMs to fit non-linear effects. A detailed treatment of smoothers is beyond the scope of this tutorial, so we refer the readers to XXX.

Before looking at the summary of the `gam_1` model, let's plot the predicted effect of proficiency. We will use the tidygam package, which provides users with utility functions that make extracting and plotting predictions from GAMs easier.

```
# attach tidygam
library(tidygam)

# extract model predictions
gam_1_preds <- predict_gam(gam_1)

# plot predictions
plot(gam_1_preds, series = "proficiency")
```

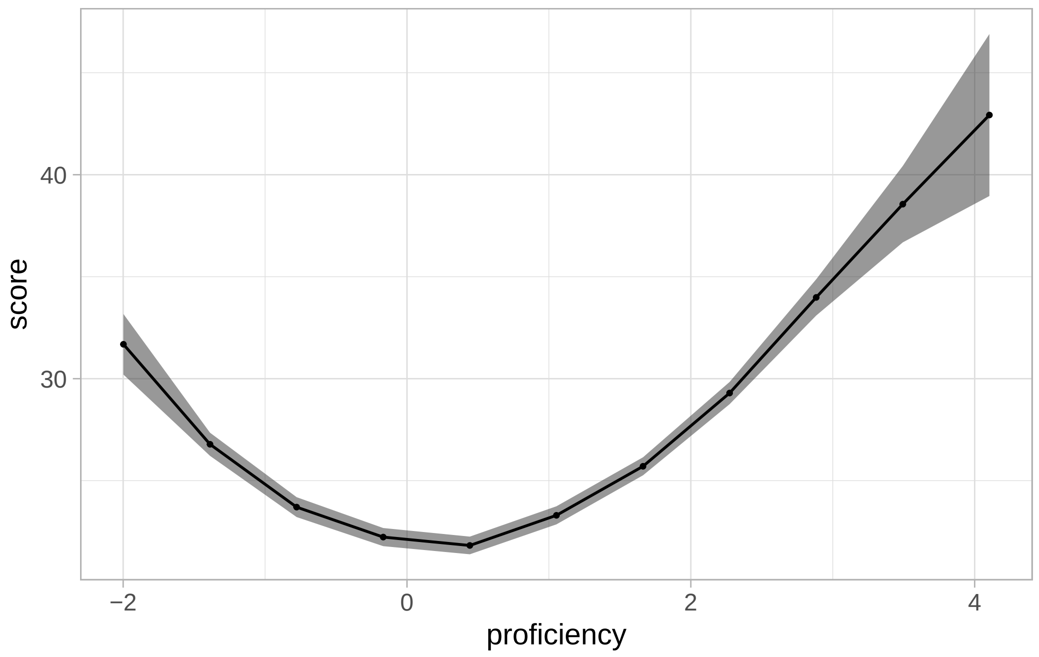

Figure 2: **?(caption)**

Figure 2 is a plot of the predicted effect of proficiency on learning score, based on the `gam_1` model from above. We will look into the details of `predict_gam()`. For now, just notice that the learning score initially decreases with increasing proficiency until about 0.5 and then starts increasing, thus producing the typical U-shaped curve. Let's inspect the model summary now.

```
summary(gam_1)
```

```
Family: gaussian
Link function: identity
```

```
Formula:
score ~ s(proficiency)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.73817    0.09987   247.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df     F p-value
s(proficiency) 6.37  7.551 128.3  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.306   Deviance explained = 30.8%
GCV = 22.018  Scale est. = 21.944    n = 2200
```

The relevant parts of the summary for the time being are the `Parametric coefficients:` table and the `Approximate significance of smooth terms:` table. The first table contains the estimate of the intercept. This intercept is the same intercept you would get in a linear model: here, the estimate of the intercept is the predicted learning score when proficiency is 0. According to the summary, when proficiency is 0, the learning score is about 25. But what we are really interested in is the effect of proficiency on learning score, rather than the intercept per se.

Information on the effect of proficiency is found in the second table, which contains estimates of the smooth terms. Alas, these estimates don't say much about the effect per se. Rather, they just indicate if the effect is linear or not. More specifically, the `edf` estimate, which stands for Estimated Degrees of Freedom, would be 1 for perfectly linear effects and greater than 1 for non-linear effects. This number tells us nothing about the shape of the effect. The only way to assess this is to plot the model predictions, as we have done above.

In our `gam_1` model, the `edf` for the smooth term over proficiency is 6.37. This is significantly different from 1 ($p < 2e\text{-}16$), thus suggesting that the effect of proficiency is not linear (the `Ref.df`, reference degrees of freedom, and `F` values have the only function of being needed to get a *p*-value).