Workshop 9 - Longitudinal Data

JPS

Introduction

We are going to explore one of the longitudinal data modelling approaches presented in the lecture: growth curve models. In essence these are multilevel models applied to longitudinal data. As such, most of the procedures that we will use today were also used last week, which is why for today's practical you are required to take more of a lead, unlike last week, where full guidance was provided.

In today's exercise we will use sentencing data from the Czech Republic. This data is truly unique since it is the first dataset where sentences are linked to the judges who imposed them chronologically. This allow us to explore changes in the sentencing practice across judges as they progress in their career. A research question explored in Drápal & Pina-Sánchez (2022). Here, we will replicate some of the analyses undertook in that paper to model changes in severity (measured as the probability of imposing a custodial sentence) throughout roughly the first 1,000 cases sentenced by different judges from the Czech Republic since they join the judiciary. This analysis will allow us to assess whether judges become harsher or more lenient through their careers, while at the same time allowing us to measure changes in between-judge disparities across time. These between-judge characteristics are problematic as they make the system less transparent and consistent. Using longitudinal data we will be able to see whether sentencing practices of judges converge across time, as it should be expected if they communicate and learn from each other.

Modelling Judicial Trajectories

We start by importing the data, which is saved as a .csv file.

```
judges = read.csv("judgesCZ.csv")
```

The original data could be classified as sensitive so I have trimmed it extensively to ensure that it is fully anonymised. In addition, I have prepared the variables that we are going to use, so no data cleaning is required this time. Still, it is always a good idea to get started with an exploratory analysis, even if brief, to assess what is in the dataset.

```
summary(judges)
```

```
##      range              prevconv             female       judge_sentence
##  Min.   :-1.16358   Min.   :-0.276906   Min.   :0.0000   Min.   :0.0010
##  1st Qu.:-0.49691   1st Qu.:-0.276906   1st Qu.:0.0000   1st Qu.:0.2410
##  Median :-0.49691   Median :-0.176906   Median :0.0000   Median :0.5350
##  Mean   :-0.03034   Mean   : 0.001873   Mean   :0.1542   Mean   :0.6419
##  3rd Qu.: 0.16976   3rd Qu.: 0.123094   3rd Qu.:0.0000   3rd Qu.:0.9550
##  Max.   :10.50309   Max.   : 1.723094   Max.   :1.0000   Max.   :2.0580
##  NA's   :4
##     judge_ID       custody
##  Min.   :4000   Min.   :0.0000
##  1st Qu.:4072   1st Qu.:0.0000
##  Median :4110   Median :0.0000
##  Mean   :4110   Mean   :0.1414
##  3rd Qu.:4163   3rd Qu.:0.0000
```

```
##  Max.   :4191   Max.   :1.0000
##
```

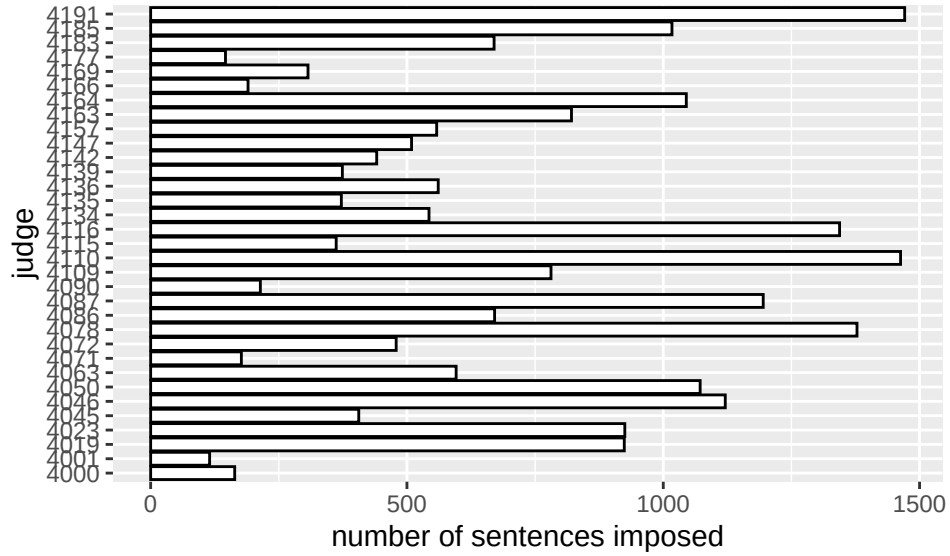
```
length(unique(judges$judge_ID))
```

```
## [1] 33
```

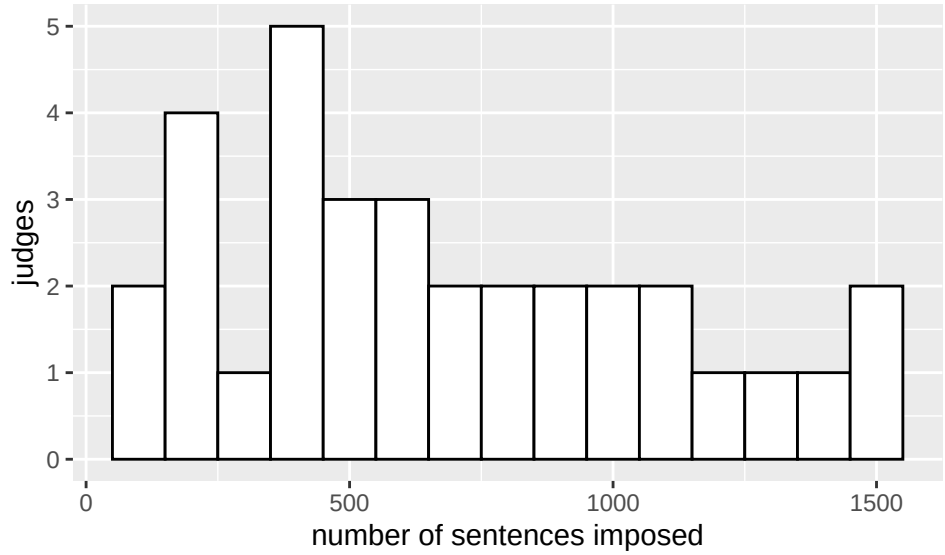
We have eight variables: 'range' captures the recommended sentence length for the offence type according to the Czech criminal code, which we can use as a proxy for offence type, this variable was recorded originally in months, but here it has been divided by ten and centered around the mean (this process helps simplify the computational process); 'prevconv', indicating the number of previous convictions, which has also been centered around the mean; 'female', indicating whether the offender is a woman or a man; 'judge_sentence' indicating the order in which each judge imposed the sentences recorded (this has not been centered around its mean to facilitate its interpretation but it has been divided by 1,000 so the range is comparable with that of other continuous variables used in the model); 'judge_ID' to differentiate by judge presiding over each sentence; and 'custody' indicating whether a custodial sentence was imposed or not.

The longitudinal dimension of the dataset stems from the 22,412 sentences imposed by 33 judges ('judge_ID'). Given the different levels of experience for the judges used in the sample, and as a result of some of the trimming processes that I undertook to prepare the dataset, the distribution of the number of the sentences seen by each judge varies importantly. Growth curve models can handle this issue under the assumption that the missing segments of individual trajectories are missing at random.

```
#The following line creates a dataset with the number of sentences recorded by each
#judge.
count = as.data.frame(table(judges$judge_ID))
names(count) = c("judge_ID", "n")
count
summary(count$n) #The number of cases seen by each judge.
library(ggplot2)
#We can plot this to see the number of cases processed by each judge visually.
ggplot(data=count, aes(x=as.character(judge_ID), y=n)) + geom_bar(stat="identity",
                  color="black", fill="white") + coord_flip() + labs(x="judge",
                  y="number of sentences imposed")
```




```
#We can also plot the distribution of cases seen by judges in our sample.
ggplot(count, aes(x=n)) + geom_histogram(binwidth=100, color="black", fill="white") +
                          labs(y="judges", x="number of sentences imposed")
```



We can see how the mean number of cases processed by judge is around 679, but that distribution is rather right-skewed, with more than 10 judges seeing less than 500 cases, and four judges seeing more than 1250. This uneven distribution is not ideal since those four judges will have a stronger influence in the estimation of trajectories to be explored in our models.

Let's move now to the modelling phase. We want to explore whether judges become harsher or more lenient through their careers. Additionally, we will try to assess whether differences between judges disparities become more or less pronounced across time. I would divide this into three different steps, in order of complexity:

1. First, we could specify a standard logistic model to determine whether the number of cases imposed has an influence on sentence severity (the probability of imposing custodial sentences).
2. Then, we can use a random intercepts model to assess whether there are meaningful between-judge disparities in the use of such penalty, i.e. whether there are judges systematically harsher or more lenient than others. This involves thinking of longitudinal data as hierarchical data. In this case, sentences will be the level-1 unit and judges the level-2, i.e. sentences are clustered within judges.
3. Lastly, we can specify a random slopes model for the variable 'judge_sentence' to explore whether those judge disparities shrink or expand as judges become more experienced, i.e. as they sentence more cases.

Question: Do judges sentence more harshly or more leniently as they acquire more experience? Hint1: Estimate a logit model for 'custody' using all the other variables in the 'judges' dataset as explanatory variables. Hint2: you can use *glm* with *family=binomial* to do so.

```
logit = glm(custody ~ range + prevconv + female + judge_sentence,
            family="binomial", data=judges)
summary(logit)
```

```
##
## Call:
## glm(formula = custody ~ range + prevconv + female + judge_sentence,
##     family = "binomial", data = judges)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.06970    0.03887 -53.253  < 2e-16 ***
## range           0.58021    0.01842  31.493  < 2e-16 ***
## prevconv        2.72608    0.05482  49.724  < 2e-16 ***
```

```
## female         -0.29047    0.07280  -3.990 6.61e-05 ***
## judge_sentence -0.15631    0.04622  -3.382 0.000721 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18268  on 22407  degrees of freedom
## Residual deviance: 14266  on 22403  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 14276
##
## Number of Fisher Scoring iterations: 5
```

So, as could be expected 'range' and 'prevconv', indicating more serious offences and more persistent offending, are associated with a higher probability of receiving a custodial sentence; similarly female offenders tend to commit less serious crimes, which is reflected by a negative coefficient. The most interesting part is that our main variable of interest, 'judge_sentence' shows a negative and statistically significant effect. This means that judges tend to become more lenient with experience. Specifically, after 1,000 cases processed, compared to the moment when a judge imposes their first sentence, the odds ratio of imprisonment is...

```
exp(summary(logit)$coefficients[5,1])
```

This seems pretty substantial; using the reference category we can also calculate the probabilities of receiving a custodial sentence at those two time points in the judicial career. To define this reference category we can use the average offence type (range = 0) committed by a male offender (female=0) with an average number of previous convictions (prevconv = 0). That is, we just need to consider the regression coefficients for the intercept and 'judge_sentence' as all others are set to zero. Lastly, remember that to go from an odds ratio to a probability we use the following equation, Odds/(1+Odds).

```
#The probability of imposing a custodial sentence for a judge with no experience is
#calculated using the intercept, which means experience is set at 0.
exp(summary(logit)$coefficients[1,1]) / (1+exp(summary(logit)$coefficients[1,1]))
```

```
## [1] 0.1120772
```

```
#The probability of imposing a custodial sentence for a judge that has imposed 1000
#sentences. Remember that we divided the number of cases by 1000, so an increment of
#1 represents 1000 cases.
exp(summary(logit)$coefficients[1,1]+summary(logit)$coefficients[5,1]) /
(1+exp(summary(logit)$coefficients[1,1]+summary(logit)$coefficients[5,1]))
```

```
## [1] 0.09743935
```

Using these comparison of reference categories we can put effect sizes in context a little more clearly, and see how the leniency effect attributed to experience is not really that dramatic. Still, it is interesting to see how the sentencing process is not entirely deterministic, there are discrepancies that can be attributed to non-legal factors such as the stage of the career of the judge imposing the sentence. To assess the extent of between-judge disparities, i.e. differences in severity between judges, we can extend this model by including a random intercepts term. This model can take some time to compute as the addition of random effects makes it considerably more complex.

Question: Are there between-judge disparities in the use of custodial sentences? Hint1: You can add random intercepts to your previous logit model using the command *glmer* (from library *lme4*) and specifying that 'judge_ID' is the variable that captures the level-2 clusters, as follows, *+ (1|judge_ID)*. Hint2: To answer the question you can compare the standard deviation of the random intercept (within the 'Random effects' part of the model's output) to the size of the intercept, or you could also take a look at the random intercept estimated for each judge, which can be obtained using *coef('model_name')*, as we did last week.

```
library(lme4)
RI = glmer(custody ~ range + prevconv + female + judge_sentence +
          (1|judge_ID), family="binomial", data=judges)
summary(RI)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: custody ~ range + prevconv + female + judge_sentence + (1 | judge_ID)
##    Data: judges
##
##      AIC      BIC   logLik deviance df.resid
##  13851.1  13899.2  -6919.6  13839.1    22402
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.6710 -0.3385 -0.2272 -0.1594  7.0107
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  judge_ID (Intercept) 0.2509   0.5009
## Number of obs: 22408, groups:  judge_ID, 33
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.09979    0.09627 -21.811  < 2e-16 ***
## range           0.59535    0.01899  31.347  < 2e-16 ***
## prevconv        2.77357    0.05656  49.038  < 2e-16 ***
## female         -0.35872    0.07422  -4.833 1.34e-06 ***
## judge_sentence -0.26703    0.05626  -4.746 2.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) range  prvcnv female
## range       -0.056
## prevconv    -0.119  0.200
## female      -0.084  0.003  0.126
## judge_sntnc -0.264 -0.111 -0.073 -0.024
```

Notice how now all the standard errors are larger than in the previous model. This is a result of having adjusted for the within cluster correlation (sentences within judges), which were left unadjusted in our previous model (we assumed sentences are independent). We can also see that some regression coefficients have also changed by a few decimal points, due to the different estimation method undertaken in this model, but in general all regression coefficients related to the fixed part of the model are quite similar to what we observed before.

The most interesting part of this model is its random part. We can see that the standard deviation of the random intercepts term is 0.50; but what does that mean? Widespread between-judge disparities, or relatively negligible differences? As we learnt in the previous workshop, we can get the intercept for each judge using *coef*, and the specific random effect for each judge using *ranef*.

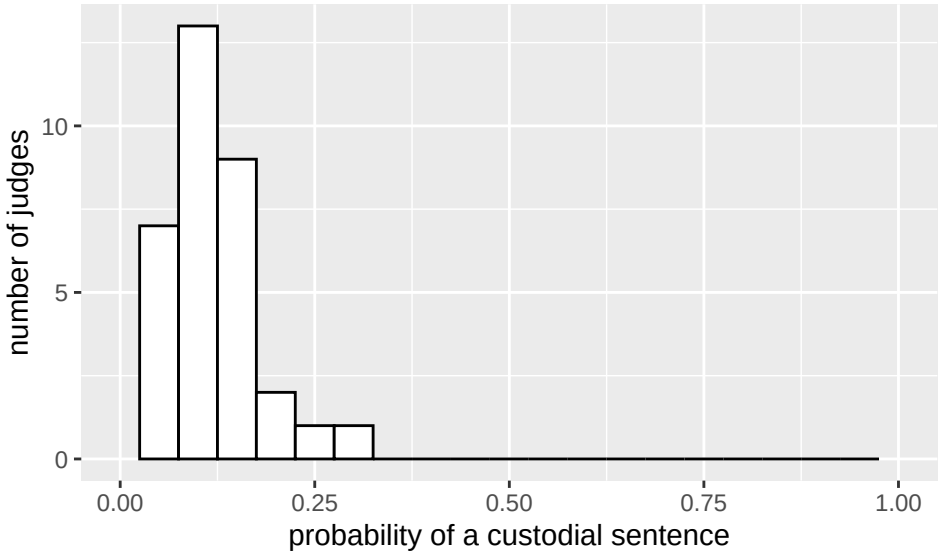
```
coef(RI)
ranef(RI)
```

This between-judge variability looks rather substantive. For example, notice how the second and third to last judges have got intercepts (-2.82 and -0.86) roughly 50% smaller and larger than the model's intercept (-2.10). To obtain a clearer view of the magnitude of these between-judge disparities we can estimate the probability of imposing a custodial sentence for each judge using just the intercept, that is, when 'judge_sentence' is equal to zero, and we are considering the average offence type, number of previous convictions and a male offender.

```
#Estimating an average probability of imposing a custodial sentence for each judge
#based on their respective random intercepts.
prob = exp(coef(RI)$judge_ID[,1]) / (1 + exp(coef(RI)$judge_ID[,1]))
summary(prob)  #The between-judge disparities in the use of custodial sentences
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04918 0.07972 0.10637 0.11843 0.14892 0.29661
```

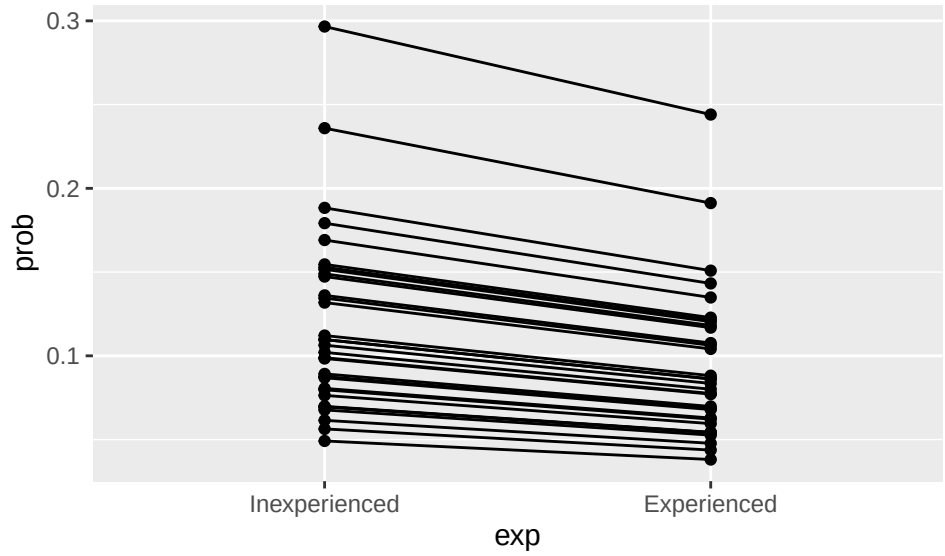
```
#range from 0.05 to 0.30, that is remarkable, probably higher than desirable.
prob = as.data.frame(prob) #I turn this into a dataset so it can be plotted with ggplot.
names(prob) = "intercept" #Prove more meaningful labels.
#Plotting the between-judge disparities.
ggplot(prob, aes(x=intercept)) + geom_histogram(binwidth=0.05, color="black", fill="white") +
       xlim(0,1) + labs(x = "probability of a custodial sentence", y = "number of judges")
```



As anticipated, the between-judge disparities are quite substantial, with some judges being more punitive than others. To assess how those disparities look like after judges become more experienced we can compare the probabilities that we have estimated to a new set of probabilities by judge after they have sentenced their first 1,000 cases.

```
#Calculating judges' average use of custody after having sentenced 1000 cases.
#We are only expanding what we did above by incorporating the coeffficient for
##judge_sentence' from the 'RI' model.
prob2 = exp(coef(RI)$judge_ID[,1] + coef(RI)$judge_ID[,5]) /
        (1 + exp(coef(RI)$judge_ID[,1] + coef(RI)$judge_ID[,5]))
#As before we turn these probabilities into a dataset so we can plot them with ggplot.
prob2 = as.data.frame(prob2)
names(prob2) = "intercept2"
#We create a dataset composed of a variable capturing the two set of probabilities
#estimated, another variable indicating whether the probability is from the judge
```

```
#when she is unexperienced or experienced, and another variable listing the 33 judges.
lineplot = data.frame(exp=factor(rep(c("Inexperienced", "Experienced"), each=33),
                      levels=c("Inexperienced", "Experienced")),
                      judge=rep(1:33, 2), prob=c(prob$intercept, prob2$intercept2))
#We change the type of ggplot to a lineplot so we can compare the probabilities at
#judge_sentence=0 and at judge_sentence=1.
ggplot(data=lineplot, aes(x=exp, y=prob, group=judge)) + geom_line() + geom_point()
```



We see that the probability of imposing custodial sentences is lower for all judges after they become more experienced, but that is because we are assuming that such effect is constant across judges. To test whether that is the case we need to estimate a random slopes model. It also appears that the range of between judge disparities is narrower when judges become more experienced, but that is just the result of pushing the average custody for each judge closer to 0, i.e. when modelling probabilities there is a floor effect at 0 (and a ceiling effect at 1) that makes it harder to observe differences the closer we are to those values compared to probabilities lying closer to the middle point, 0.5. Again, remember that we are still assuming that the effect of 'judge_sentence' is constant across judges.

Question: Do between-judge disparities shrink as judges become more experienced? Is this effect statistically significant and meaningful? Hint1: You can explore this question by expanding our previous random intercepts model including a random slopes term, you can do that changing the random part of the model, so, in our case: *(1 + 'variable_name'|judge_ID)*. Hint2: To test whether the random slopes term is statistically significant you can run a likelihood ratio test. Remember from last week that you will need first to load the library *lmtest*, then you will need the command, *lrtest(model1, model2)*. Hint3: To interpret whether the random slopes term is meaningful you can take a look at the standard deviation of the random intercept compare it to that of the standard deviation of the random slope (does it seem remarkable?), and the correlation between these two terms (is it positive, negative, large or small?). The answer to our research question (do between-judge disparities shrink or grow as judges become more experienced?) can be derived from that correlation term. Lastly, see if you can plot the between-judge disparities for judges when they start their careers and when they have processed one thousand cases using the above lineplot as a template. Specifically, you will need to estimate 'prob' and 'prob2' as we did before, with the only difference that now these need to be derived from your new random slopes model, as we did last week. Once you have those, you can simply put them together in the same dataset as we did above (we called it 'lineplot'), and run the same *ggplot* code.

```
RS = glmer(custody ~ range + prevconv + female + judge_sentence +
                     (1 + judge_sentence|judge_ID), family="binomial", data=judges)
summary(RS)
```

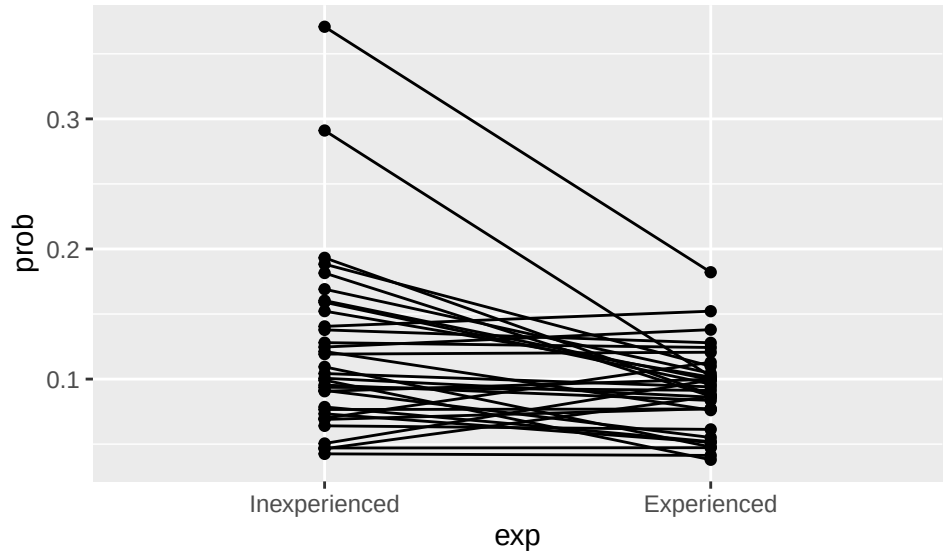
```
coef(RS)
```

Looking at the fixed part of the model the first thing to notice is that 'judge_sentence' is not significant anymore, this is likely due to the within cluster correlation not being perfectly accounted for with the simpler random intercepts specification. On the random part of the model we can also see that the standard deviation for the random intercepts term remains quite substantial (even bigger than before), and so is the standard deviation for the random slopes. If we want to make sure that this random slopes term is significant we can run a likelihood ratio test

```
library(lmtest)
lrtest(RI, RS)
```

And of course, it is significant. Those random slopes disparities appear to be massive, indicating that although the average effect of 'judge_sentence' is not statistically significant (i.e. the average effect of experience is not significantly different from zero), there is evidence pointing at very different trajectories followed by individual judges. In considering whether those different trajectories will converge (reducing the judge disparities that we observed in our previous model) or diverge (increasing those judge disparities), we can get an important clue from the correlation between the random intercepts and random slopes terms. This is negative, meaning that judges with a higher than average intercept (those who initially were harsher) will be associated with a negative experience effect (will become more lenient as they become more experienced), and the other way around, those judges with a lower than average intercept (those that were initially more lenient) will be associated with a positive experience effect (will become harsher as they become more experienced). We can see this visually as follows:

```
prob = exp(coef(RS)$judge_ID[,1] ) / (1 + exp(coef(RS)$judge_ID[,1]))
summary(prob)
prob = as.data.frame(prob)
names(prob) = "intercept"
prob2 = exp(coef(RS)$judge_ID[,1] + coef(RS)$judge_ID[,5]) /
        (1 + exp(coef(RS)$judge_ID[,1] + coef(RS)$judge_ID[,5]))
prob2 = as.data.frame(prob2)
names(prob2) = "intercept2"
lineplot = data.frame(exp=factor(rep(c("Inexperienced", "Experienced"), each=33),
                      levels=c("Inexperienced", "Experienced")),
                      judge=rep(1:33, 2), prob=c(prob$intercept, prob2$intercept2))
ggplot(data=lineplot, aes(x=exp, y=prob, group=judge)) + geom_line() + geom_point()
```



This is an important finding. Even though experience does not seem to have a fixed effect on sentence severity, it does produce a really interesting effect in diminishing between court disparities. It seems that judges, as they sentenced more and more cases, become more in tune with each other. When I think of this it doesn't surprise me, when marking student essays as part of a module teaching team it is absolute essential to discuss the marking procedure before and after seeing one or a few essays, otherwise I tend to note between-marker disparities. Going back to judicial decision-making, this finding has got important implications for sentencing policy since it seems there are procedures that could be taken into consideration to enhance consistency in sentencing without having to rely on prescriptive sentencing guidelines, which are often considered too intrusive and could undermine the principle of individualisation.

Now, putting things into perspective, our sample is huge but we have only looked at 33 judges. In addition, it is possible that our results are also affected by selection bias if those judges for whom we can see their trajectories for a much longer timespan are different from judges for which we can only see how they operate for a short timespan. Plenty of additional work needs to be done on those two fronts, but this is an interesting start.