Multivariate: Linear Mixed Models

Lecture 2: Model Essentials

Tom Booth

Lecture 2

Good morning

Resources

- Reading list is now on LEARN.
  - For reference, there are big lists of additional resources compiled **here, here** and **here**
  - I have not vetted everyting on these lists, and they are not the only sources, but hopefully they are useful.
- I also came across **this** really wonderful visualization that uses the same salary and experience example from lecture 1 (though not same data)
- Lastly, the link to an `lme4()` overview paper by the authors is **here**
  - It gives an excellent brief summary of all key elements of package.

Course Announcements

- Coursework information will be posted on LEARN this week.
- Answersheet to week 1 lab posted.
- Lab 1 content - quick check with class.

Recap

From last week...

- We established discussed how, when data have nested structures, we need a slightly different model to our standard linear model.
  - Namely, the linear mixed model.
- We discussed conceptually how LMM include both fixed and random effects.
  - Fixed effects given our overall average.
  - Random effects capturing the variation around the average via decomposing the model residuals.
- And we then started to look a little at our model equation.

Today

- Intro to the data for our working example
- Intro to `lme4`
- Fixed Effects, random Effects and variance-covariance matrices
- Pooling: None, complete and partial
- Intraclass correlation
- Model specification and interpretation, with examples

Examples

- Example 1: Repeated measures experiment
- Example 2: Two group intervention study
- Example 3: Multiple trial experiment
- Example 4: Cross-sectional multi-level study
- Example 5: Longitudinal growth model
- Example 6: Experience sampling study
- Example 7: Multi-level CFA

Any questions from last week or lab?

Intro to our data

Data: Predicting salary from years of service

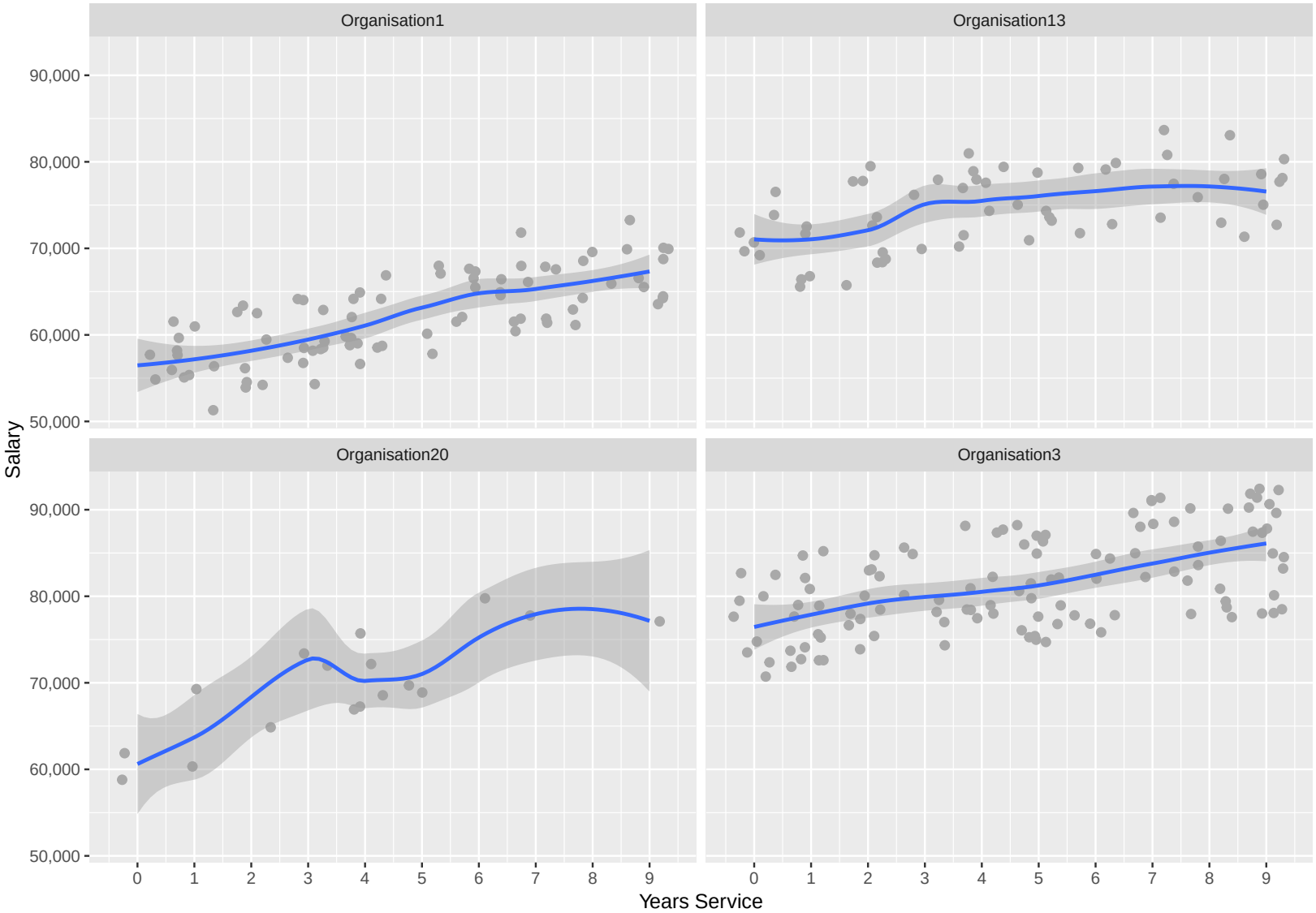
- Last lecture we used salary and years of service as an example.
- Here we have generated some data.
  - Suppose we have 20 organisations
  - Varied number of employees within each organisation
  - We have measured each employees salary
  - And each employees years of service

Describe our data

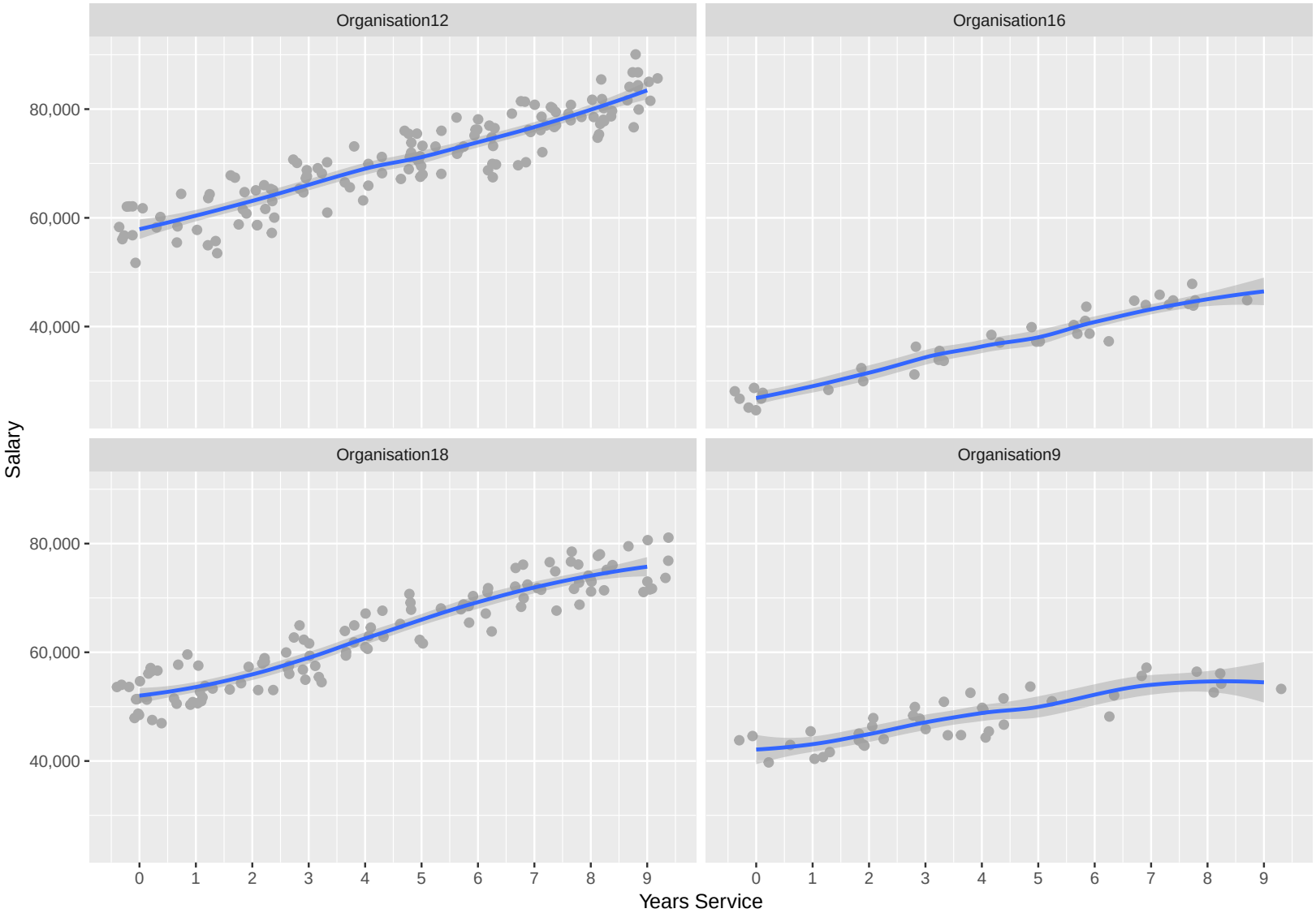
Table 1: Descriptive Statistics by Organisation

organisation	N.Employee	Av.Salary	Sd.Salary
Organisation1	82	62119	4803
Organisation10	6	63762	6116
Organisation11	117	64465	3989
Organisation12	137	71003	8382
Organisation13	59	74520	4467
Organisation14	23	53446	7352
Organisation15	26	33954	4866
Organisation16	37	36620	6974
Organisation17	61	45203	7964
Organisation18	114	63344	9148

Visualize our data



Visualize our data



Our data

- So from this we can see we have quite different intercepts across the organisation, but that on the whole, the relationships look positive in all cases.
    - Which I suspect is a relief if you work in those organisations.
- This points to our optimal model as most likely having a random intercept, and possibly a random slope for the effect of `experience` on `salary`.

Model for data: Random intercept, random slope

$$Salary_{ij} = \beta_{0j} + \beta_{1j} Service_{ij} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \upsilon_{0j}$$

$$\beta_{1j} = \gamma_{10} + \upsilon_{1j}$$

- Presented as a single equation:

$$Salary_{ij} = \gamma_{00} + \gamma_{10} Service_{ij} + \upsilon_{0j} + \upsilon_{1j} + \epsilon_{ij}$$

Introduction to `lme4` and `lmer()`

Packages and functions

- You have already been introduced to `tidyverse` packages.
- Our main analytic package for this course is `lme4`.
- In particular the `lmer()` function.
  - `lmer()` fits generalized **l**inear **m**ixed-**e**ffect models in **R**
- The major alternative for this type of model is `nlme()`.
  - Preferable (perhaps) if fitting models with very complex random effect covariance structures (more on this later)

lmer()

```
lmer(
  formula = , # specify the model (focus today)
  data = ,    # as with everything, provide data
  REML =      # Choose estimator (more next week)
)
```

- As always there are many more arguments, but these are the main ones for us to focus on to begin with.

`lmer()` formula

- The formula in `lmer()` is very similar to the formula structure in `lm()`
  - The outcome is placed on the left of $\sim$

```
lmer(Y ~ )
```

- The predictors are separated by $+$;

```
lmer(Y ~ X1 + X2)
```

- So what is different?

`lmer()` formula: Random effects

- We need to split our formula statement to explicitly tell R what are fixed and random effects.
- Random effects appear in parentheses after the fixed effects.
- Random effects have the following structure:
  - `(lowest level | grouping variable)`
- The vertical bar | can be read as *by* or *given*
  - The variable given to the right of the bar is treated as a grouping variable that the variable to the left is within.

lmer() formula: Random effects

- So for example:

```
lmer(Y ~              # outcome
       X1 + X2 +      # fixed effects of 2 predictors
       (1 | X2)       # random intercept by X2
       )
```

lmer() formula: Random effects

- We can add more variables to the left of the | if we wish:

```
lmer(Y ~                    # outcome
       X1 + X2 +            # fixed effects of 2 predictors
       (1 + X1 | X2)        # random intercept & slope X1 by X2
       )
```

- This specification allows the random effects for the intercept and slope of x1 to covary.

lmer() formula: Random effects

- We can force the random effects to not covary by specifying the random effects separately.

```
lmer(Y ~                    # outcome
       X1 + X2 +            # fixed effects of 2 predictors
       (1 | X2) +           # random intercept by X2
       (0 + X1 | X2)        # random slope X1 by X2
       )
```

Fixed Effects, Random Effects & Variance-covariance matrices

Random intercept model

$$Salary_{ij} = \beta_{0j} + \beta_1 Service_i + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \upsilon_{0j}$$

$$Salary_{ij} = \gamma_{00} + \beta_1 Service_i + \upsilon_{0j} + \epsilon_{ij}$$

```
m1 <-
  lmer(salary ~
       1 + service + # fixed intercept & fixed slope
       (1 | organisation), # random intercept
       data = pay
     )
```

A little more `lmer()`

- As with most R model objects, the results from `lmer()` provide a lot of information.
  - `summary()`
  - `fixef()`
  - `ranef()`
  - `coef()`
  - `fitted()`
- The `arm()` package is also used in labs to get standard errors for the different coefficients.

Random intercept model: Fixed effects

$$Salary_{ij} = \gamma_{00} + \beta_1 Service_i + \upsilon_{0j} + \epsilon_{ij}$$

```
fixef(m1)
```

```
## (Intercept)     service
##    55873.65     1631.44
```

- γ_{00}: Overall intercept $= 55873.65$
  - Average salary $=$ £55,873.65
- β_1: average slope $= 1631.44$
  - For every year of service, salary increases by £1631.44

Random intercept model: Variance-covariance

$$Salary_{ij} = \gamma_{00} + \beta_1 Service_i + \upsilon_{0j} + \epsilon_{ij}$$

```
as.data.frame(VarCorr(m1))
```

```
##            grp        var1 var2      vcov     sdcor
## 1 organisation (Intercept) <NA> 221880749 14895.662
## 2     Residual        <NA> <NA>  18288174  4276.467
```

- υ_{0j} and ϵ_{ij} are the model residual terms from which we can calculate the variance associated with the random effects.
  - These are shown as variances (in `vcov`) and in standard deviation units (in `sdcor`)

Random intercept model: Random effect estimates

- We can access the individual random effects estimates for the groups using:

```
as.data.frame(ranef(m1))[1:3, 2:4]
```

```
##          term            grp    condval
## 1 (Intercept)  Organisation1 -1492.4466
## 2 (Intercept) Organisation10   539.3024
## 3 (Intercept) Organisation11   949.7160
```

- `term` shows that it relates to the intercept.
- `grp` shows the level of the grouping factor
- `condval` is the residual estimate. Here the difference in intercept for each organisation from the overall average.

Models for individual groups

- Say we wanted to look at the predicted values for employees in organisation 1.
- What would our equation look like?

Relation between `fixef()`, `ranef()` & `coef()`

- Random effects are the differences between the individual coefficients and the fixed effects. So:
  - `ranef()` = `coef()` - `fixef()`
- Individual coefficients are the sum of the fixed and random effects
  - `coef()` = `fixef()` + `ranef()`

Model for organisation 1

- Slope = fixed slope

```
fixef(m1)[2]
```

```
## service
## 1631.44
```

$$\hat{Salary}_{i1} = (\gamma_{00} + \upsilon_{01}) + 1631.44 Service_i$$

Model for organisation 1

- Intercept = fixed slope + the randon effect estimate for group

```
fixef(m1)[1]
```

```
## (Intercept)
##    55873.65
```

```
ranef(m1)$organisation[1,]
```

```
## [1] -1492.447
```

$$\hat{Salary}_{i1} = (55873.65 - 1492.447) + 1631.44 Service_i$$

Model for organisation 1

- or

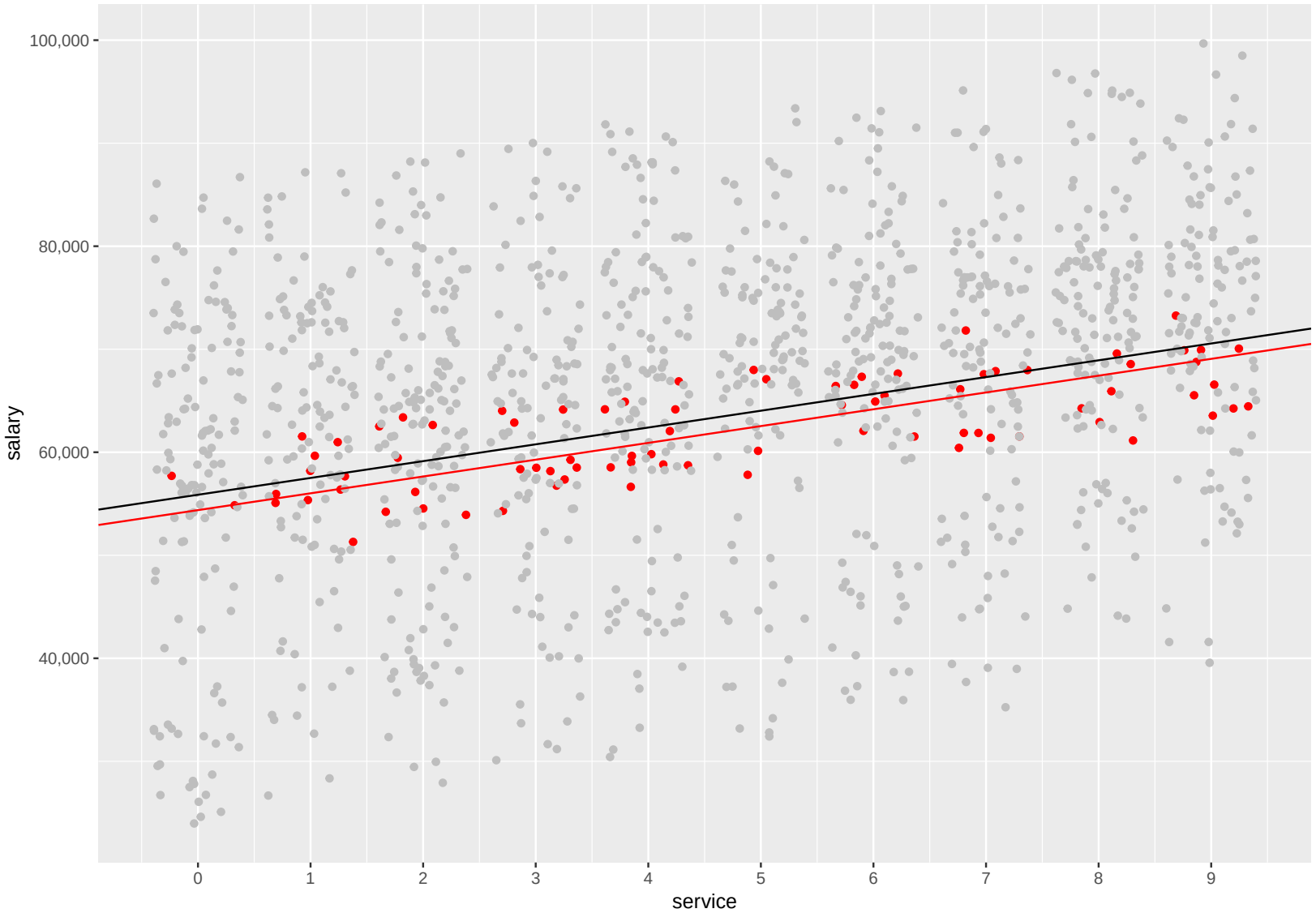
```
coef(m1)$organisation[1,]
```

```
##                 (Intercept) service
## Organisation1       54381.2 1631.44
```

- Prediction equation for an individual in organisation 1:

$$\hat{Salary}_{i1} = 54381.2 + 1631.44 Service_i$$

Model for organisation 1



Random intercept-random slope

$$Salary_{ij} = \beta_{0j} + \beta_{1j} Service_i + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \upsilon_{0j}$$

$$\beta_{1j} = \gamma_{01} + \upsilon_{1j}$$

$$Salary_{ij} = \gamma_{00} + \gamma_{01} + \upsilon_{0j} + \upsilon_{1j} + \epsilon_{ij}$$

```
m2 <-
lmer(salary ~
   1 + service + # fixed intercept & slope
   (1 + service | organisation), # random intercept & slope
   data = pay
 )
```

Random intercept-random slope: Fixed effects

$$Salary_{ij} = \gamma_{00} + \gamma_{01} + \upsilon_{0j} + \upsilon_{1j} + \epsilon_{ij}$$

```
fixef(m2)
```

```
## (Intercept)     service
##    55643.11     1693.03
```

- γ_{00}: Overall intercept = 55643.11 (m1 γ_{00} = 55873.65)
  - Average salary = £55,643.11
- β_1: Overall slope = 1693.03 (m1 β_1 = 1631.44)
  - For every year of service, average salary increases by £1693.03

Random intercept-random slope: Variance-covariance

$$Salary_{ij} = \gamma_{00} + \gamma_{01} + \upsilon_{0j} + \upsilon_{1j} + \epsilon_{ij}$$

```
out <- as.data.frame(VarCorr(m2))
kable(out) # for slide formatting only
```

grp	var1	var2	vcov	sdcor
organisation	(Intercept)	NA	276523735.9	16629.0028525
organisation	service	NA	582158.3	762.9929890
organisation	(Intercept)	service	-7538486.4	-0.5941518
Residual	NA	NA	12645384.3	3556.0349157

Random intercept-random slope: Variance-covariance

grp	var1	var2	vcov	sdcor
organisation	(Intercept)	NA	276523735.9	16629.0028525
organisation	service	NA	582158.3	762.9929890
organisation	(Intercept)	service	-7538486.4	-0.5941518
Residual	NA	NA	12645384.3	3556.0349157

- Note here we have a correlation between random effects of intercept and slope (r = -0.59)
  - Organisations with higher starting salary (intercept), have shallower slopes (pay increase per year services)
- Why do we have this estimate?
  - *HINT: Think back to the lmer() code intro*
- Why did we not have this parameter in the random intercept model?

Random effects structure

- The lowest level residuals (ϵ_{ij}) are assumed to have a mean of 0 and a variance to be estimated.
- The group level residuals (here υ_{0j} and υ_{1j}) are also assumed to have a mean of 0 and a variance to be estimated.
- ϵ_{ij} are assumed to be uncorrelated with all group level residual terms.
- But υ_{0j} and υ_{1j} are usually assume to have non-zero covariances.

$$\begin{matrix} \sigsq_{\upsilon_{01}} & \\ \sigsq_{\upsilon_{12}} & \sigsq_{\upsilon_{02}} \end{matrix}$$

- These variances are what we extract from `VarCorr()`

Random effects structure

- A big topic in LMM is the particular structure that this matrix takes.
- By structure we mean:
  - What covariances are estimated vs fixed to 0.
  - What pattern of constraints in placed on the matrix
- Much of this is determined by study designs, and we will discuss this more in the context of some specific models over the next 3 weeks.

Random intercept-random slope: Random effects

- Again we can access the individual random effect estimates for groups using `ranef()`.
- Note that here we have terms relating to both the intercept and the slope.
  - Here we have 40 estimates.
  - The difference in intercept and slope for each of the 20 organisations.

```
# select rows to show org 1 and 10 intercept & slope values
as.data.frame(ranef(m2))[c(1:2, 21:22), 2:4]
```

```
##           term            grp    condval
## 1  (Intercept)  Organisation1  178.38351
## 2  (Intercept) Organisation10  823.09931
## 21     service  Organisation1 -364.69092
## 22     service Organisation10  -71.51081
```

Model for organisation 1

```
fixef(m2)
```

```
## (Intercept)     service
##    55643.11     1693.03
```

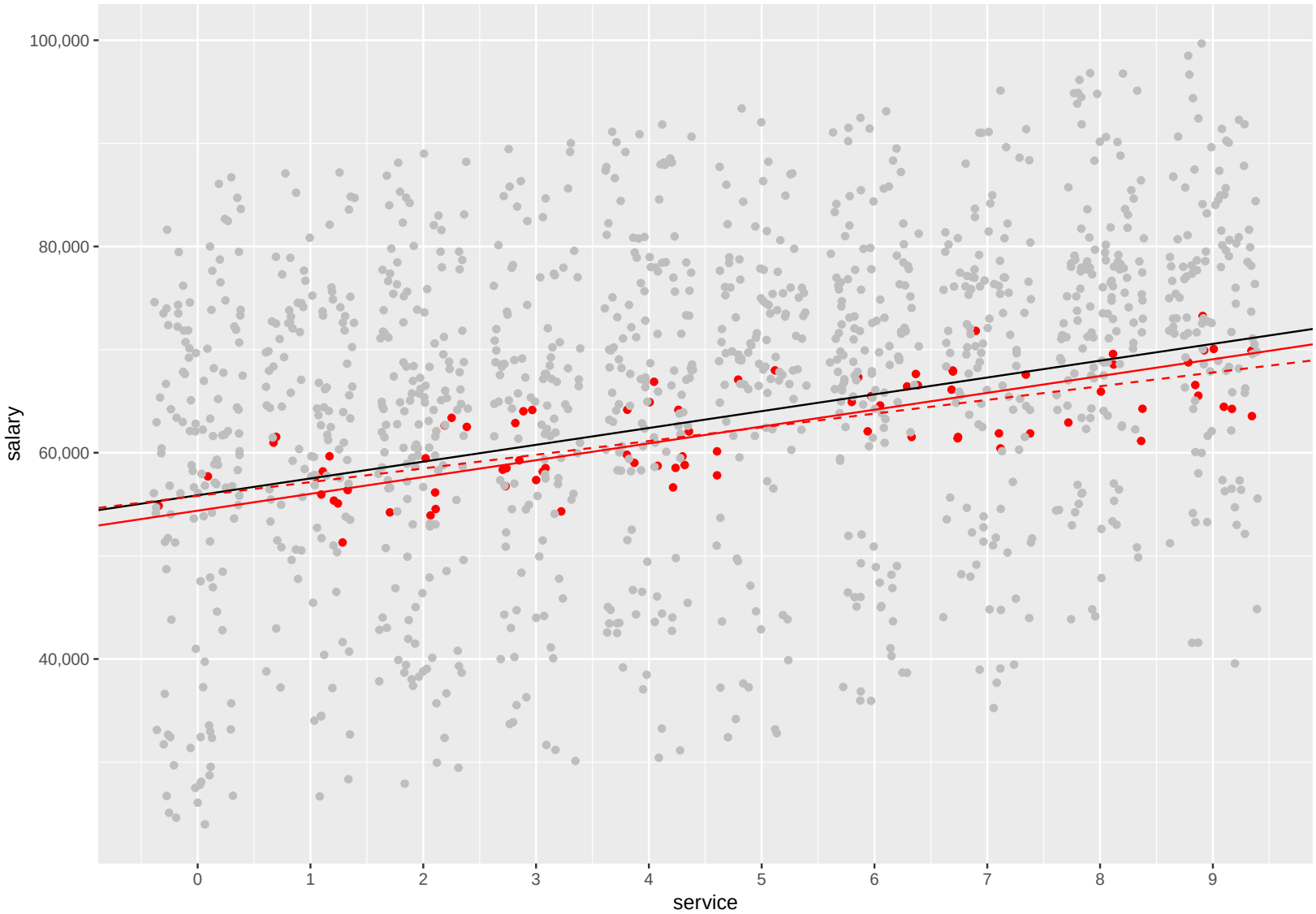
```
coef(m2)$organisation[1,]
```

```
##                (Intercept) service
## Organisation1    55821.49 1328.34
```

- Prediction equation for an individual (i) in organisation 1:

$$\hat{Salary}_{i1} = 55821.49 + 1328.34 Service_i$$

Model for organisation 1



Model for individual groups: Organisation 1

- Note that the inclusion of the additional random effect for slope has changed the estimate of the intercept.
- The intercept is now closer to the overall average, and the difference for organisation 1 is captured in the shallower slope.
- ***Why note download the notebook and code, and plot some lines for different organisations***

Pooling

Broad idea

- Last week I introduced some approaches to clustering and introduced the term pooling.
  - It is an important topic so we reutn to it more fully now.
- Recall our three key phrases were:
  - Complete pooling
  - No pooling
  - Partial pooling
- We can explore the effects using our salary example.
  - Structure here follows Gellman & Hill (2007)

A model with no predictors

- Suppose we were interested in estimating the distribution of salary in each organisation.
- We could approach this in different ways.
  - Complete pooling: Use the average and standard deviation across all data points as the estimate for each orgaisation.
  - No pooling: Estimate the mean and standard deviation independently in every organisation.
  - Partial pooling: Use LMM (!)

Estimates

Table 2: Pooled Estimates

Organisation	Complete_Pool	Partial_Pool	No_Pool
Organisation1	65724.17	62120.90	62119
Organisation10	65724.17	63737.14	63762
Organisation11	65724.17	64462.90	64465
Organisation12	65724.17	70992.64	71003
Organisation13	65724.17	74484.07	74520

Comments on Estimates

- Our complete pooling model ignores variation across organisation.
- The no pooling model overstates the differences across organisation.
- Partial pooling is a compromise.
    - The differences between organisations are smaller for the LMM estimates than the no pooling estimates (although marginally in this example).
    - What is going on?

Shrinkage

- This effect is what is referred to as shrinkage.
- In a LMM, in groups were estimates are less accruate, we see more shrinkage
- Accuracy is largely driven by:
  1. the N of the group (lower N = less information = less accurate)
  2. the distance of the group estimate from the overall average (further away = less reliable)
- The logic here is that when there are conditions under which we might make a bad guess at our coefficients of interest, using information from across the model is useful.
  - i.e. pulling estimates (or shrinking them) back towards the overall average effect.

Pooling with predictors

- Our example discussed shrinkage with respect to the means of the groups.
- But the same principle applies to estimates of other coefficients (e.g. slopes) when allowed to vary across groups.

Intraclass Correlation

How big of an influence is grouping structure?

- How much variance in our data comes from each of the different sources in our data?
    - It is a good question.
- This can be investigated based on the intraclass correlation coefficient (ICC).
- The ICC is calculated from an *intercept only model*

Intercept-only model

- The intercept-only model is a model with no predictor variables (x's) and a random intercept.

$$Y_{ij} = \beta_{0j} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \upsilon_{0j}$$

- As a single equation:

$$Y_{ij} = \gamma_{00} + \upsilon_{0j} + \epsilon_{ij}$$

Intercept-only model

$$Y_{ij} = \gamma_{00} + \upsilon_{0j} + \epsilon_{ij}$$

- This model does not explain variance in the outcome (Y_{ij}).
  - γ_{00} is the average of Y_{ij} across all observations.
  - It does decompose the variance in Y_{ij}
- The variance has two componenets:
  - σ^2_ϵ which is the variance in the lowest level residual ϵ_{ij}
  - $\sigma^2_{\upsilon_0}$ which is the variance in the lowest level residual υ_{ij}

ICC

- From this it is possible to calculate the amount of the total variance ($\sigma^2_\epsilon + \sigma^2_{\upsilon_0}$) that is associated with the highest level errors (or the grouping/nesting/clustering variable)

$$ICC = \frac{\sigma^2_{\upsilon_0}}{\sigma^2_\epsilon + \sigma^2_{\upsilon_0}}$$

- The ICC is the proportion variance explained by the grouping structure.

ICC in practice

- ICC's range from 0 to 1.
- The magnitude of the ICC will vary a lot dependent on the study type, thus grouping structure.
    - ICC's when nesting is within participants can be quite high.
    - ICC's when nesting is within a social group (e.g. companies) is often much lower.
- The presence of ICC of any non-zero ICC may suggest grouping structure would impact the estimates of the model parameters.
    - Thus an LMM approach may be sensible.

Calculating ICC

```
# Run the model
i.only <- lmer(salary ~ 1 + (1 | organisation), data = pay)

# Save the summary of results
ICC_res <- summary(i.only)
```

Calculating ICC

- We need to pull the variance of the random effects.
- This has a few steps as the info we would like is a little hidden.

```
variances <- as.data.frame(ICC_res$varcor)
# save the variances from within varcor
variances
```

```
##            grp        var1 var2      vcov     sdcor
## 1 organisation (Intercept) <NA> 219641629 14820.311
## 2     Residual        <NA> <NA>  40343767  6351.674
```

```
ICC <- variances[1,4]/sum(variances[,4])
ICC
```

```
## [1] 0.8448229
```

Interpret the ICC

- An ICC of 0.85 is pretty high.
- It suggests 85% of the variation in the data is at the level of organisation, and 15% at the level of the individual.
- So with respect to the example and salary, it could be argued that it is organisational factors, not individual that drive the differences in salary.

Model Specification & Interpretation

Example 1: Intervention study

- A research team conducts an intervention study on exercise. They want to know if total hours exercise increases whether someone uses an in gym personal trainer, or has one-to-one sessions from home.
- They randomly assign 100 people to each condition.
- They measure number of hours exercised one week pre, and two weeks post, a 3 week training intervention period.
- They also measure a set of demographics.

Fill in the blanks...

```
lmer(ExHrs ~          # outcome
       +              # fixed effects
     ( | )            # random effects
     )
```

Example 2: Longitudinal study

- A research team is interested in change in aggressive behaviour across adolescents.
- They measure aggression using a questionnaire measure every year from age 7 to age 17.
- The children in the study come from different areas of the same city.

Fill in the blanks...

```
lmer(Aggression ~          # outcome
       +                   # fixed effects
     ( | )                 # random effects
     )
```

Example 3: Experimental study

- A research team is interested in the non-word reaction time in a lexical decision task.
- They are interested in whether real word neighbourhood density is predictive of RT.
- Conduct an experiment with 30 trials per participant.

Fill in the blanks...

```
lmer(RT ~          # outcome
       +           # fixed effects
     ( | )         # random effects
     )
```

That's all for today