# Hierarchical models

Silje Synnøve Lyder Hermansen

2023-04-27

# Recap: our course

# Recap: our course

**We are entering the last part of this course**

1. R-skills (week 1-3)
2. Limited and categorical outcome variables (GLMs) (week 4-10)
3. **Data structures** (week 11-14)

# The purpose of this course

# The purpose of this course

$\Rightarrow$ *The purpose of this course is to find solutions when the assumptions of the linear model are not satisfied*

# Two assumptions in ordinary regression

# Two assumptions in ordinary regression

**Linear models (OLS) rely on two assumptions that are often violated**

1. **Assumption 1:** outcomes are continuous and unbounded (week 4-10)
2. **Assumption 2:** observations are independent and identically distributed (iid) (week 11-14)
   - independent: probabiliy of observing one unit is not dependent on observing another
   - identically distributed: observations come from the same data generating process/probability distribution

$\Rightarrow$ *strategies for when these are not satisfied*

# Solutions to violations of those assumptions

1. **Assumption 1:** Limited and categorical outcome variables (GLMs): - recode the dependent variable and describe the data generating process w/probability distribution - choice of model depends on the data generating process - e.g. logit, multinomial, ordinal, poisson, neg.bin, zero-inflated, coxph…
2. **Assumption 2:** Observations are not iid: - hierarchical/nested data - missing data

$\Rightarrow$ *what do we do when observations are not iid?*

# Today (week 11 and 12)

# Today (week 11 and 12)

**Phenomena are sometimes observed within a common context**

- we suspect that there are unobserved covariates that influence
  - the outcome and our predictors → *spurious relationships/confounders*
  - our standard error → *observations are too similar/too many*
- examples:
  - geographic context:
    - patients in hospitals: same administrative procedures
    - unemployed in municipalities: same job market/economy
    - conflicts in countries: same competition for resources/power
  - time:
    - patients/unemployed/conflicts: years
  - time and space:
    - time-series cross-sectional/panel data
    - e.g. MEPs in years from countries

# Data contains variation

**Analysis is about strategically leveraging variation**

- information
- noise:
  - bias : confounders
  - random noise: lack of precision

$\Rightarrow$ *hierarchical models are very explicit about this*

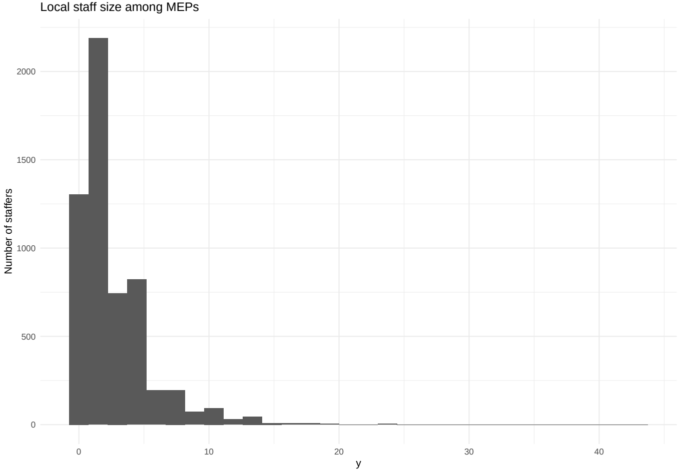
# Our example: MEPs and their local investments

**All Members of the European Parliament have the same budget for local staff**

- time-series cross-section data with three groups:
  - MEPs are observed every 6 months (`MEP`)
  - there is variation in nationality (`Nationality`)
  - there is variation over time (`Period`)
- covariates at the group-level:
  - MEP: gender, nationality
  - Nationality: electoral system
  - Period: election, reform
- covariates across groups:
  - MEP/time: age
  - Nationality/time: labor cost

# Our dependent variable: Local staff size

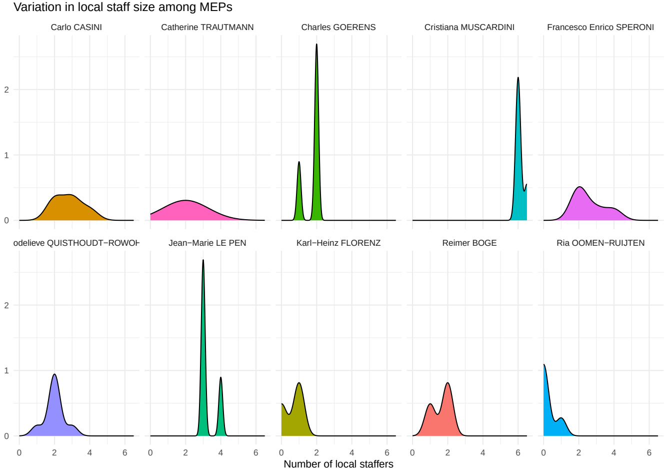
There is variation in the size of MEPs' local staff. What part of this variation am I interested in?



# Groups of observations

# Groups of observations

Let's consider the distribution of local staff *within* and *between* each MEP.



# Variation and group averages

Let's consider the distribution of local staff in light of one of the groupings (individual)

```
## # A tibble: 1,161 x 6
##        ID   y_j   sd_y   n_j  sd_a y_all
##     <int> <dbl>  <dbl> <int> <dbl> <dbl>
##  1    840 1.75   0.463     8  2.96  2.71
##  2    988 2.8    0.837     5  2.96  2.71
##  3    997 2.6    0.894     5  2.96  2.71
##  4   1023 3.25   0.463     8  2.96  2.71
##  5   1037 1.62   0.518     8  2.96  2.71
##  6   1038 0.625  0.518     8  2.96  2.71
##  7   1055 2      0.535     8  2.96  2.71
##  8   1059 2     NA         1  2.96  2.71
##  9   1073 6.1    0.224     5  2.96  2.71
## 10   1122 0.2    0.447     5  2.96  2.71
## # ... with 1,151 more rows
```

**each individual has**

- a mean staff size
- a group size

***within*-individual variation**

- a standard deviation for each distribution

***between*-individual variation**

- the standard deviation of the group means

$\rightarrow$ *we group and label the variation*

$\Rightarrow$ *Which of the variations do I want to leverage?*

## Which of the variations do I leverage?

# Which of the variations do I leverage?

- within-individual variation
    - calculate group means
    - regress residuals on individual/time predictors

→ *individual fixed effects*

- between-individual variation
    - calculate group means
    - regress them on individual predictors (e.g. gender)

→ *an aggregated data frame*

- both
    - ordinary OLS (pooled model)
    - hierarchical models

→ *random effects with predictors on both levels*

# Let's take it step-by-step

**we can separate out group averages**

- fixed effects
  - leverage within-group variation

  $\rightarrow$ a form of varying-intercept model with no pooling
  - fixed effects for between-group regression (a warm-up to level-two variables)
- random intercepts
  - random-intercept only models to cluster errors
  - random intercepts and predictors
    - at either/both levels
- varying intercepts + varying slopes
  - with fixed effects (a warm-up)
  - with random effects

# Fixed effects

# Separate out group-level variation

# Separate out group-level variation

- we can calculate the same individual averages in an **OLS with fixed effects**
- . . . but we're not interested in statistical significance (se $\neq$ sd)

```
##
## Call:
## lm(formula = y ~ -1 + as.factor(ID), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.125  -0.333   0.000   0.375  36.188
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## as.factor(ID)840    1.750e+00  4.868e-01   3.595 0.000328 ***
## as.factor(ID)988    2.800e+00  6.158e-01   4.547 5.58e-06 ***
## as.factor(ID)997    2.600e+00  6.158e-01   4.222 2.46e-05 ***
## as.factor(ID)1023   3.250e+00  4.868e-01   6.676 2.75e-11 ***
## as.factor(ID)1037   1.625e+00  4.868e-01   3.338 0.000850 ***
## as.factor(ID)1038   6.250e-01  4.868e-01   1.284 0.199250
## as.factor(ID)1055   2.000e+00  4.868e-01   4.108 4.05e-05 ***
## as.factor(ID)1059   2.000e+00  1.377e+00   1.453 0.146420
## as.factor(ID)1073   6.100e+00  6.158e-01   9.906  < 2e-16 ***
## as.factor(ID)1122   2.000e-01  6.158e-01   0.325 0.745349
## as.factor(ID)1129   2.000e+00  6.158e-01   3.248 0.001171 **
## as.factor(ID)1164   2.625e+00  4.868e-01   5.392 7.31e-08 ***
## as.factor(ID)1179   0.000e+00  7.950e-01   0.000 1.000000
## as.factor(ID)1183   1.800e+00  6.158e-01   2.923 0.003482 **
## as.factor(ID)1185   0.000e+00  7.950e-01   0.000 1.000000
## as.factor(ID)1186   1.000e+00  6.158e-01   1.624 0.104447
## as.factor(ID)1191   3.600e+00  6.158e-01   5.846 5.37e-09 ***
## as.factor(ID)1204   1.500e+00  4.868e-01   3.081 0.002073 **
## as.factor(ID)1253   1.000e+00  6.158e-01   1.624 0.104447
## as.factor(ID)1263   4.000e+00  4.868e-01   8.217 2.70e-16 ***
## as.factor(ID)1309   3.000e+00  6.158e-01   4.872 1.14e-06 ***
## as.factor(ID)1351   0.000e+00  1.377e+00   0.000 1.000000
```

# The limits/strengths of fixed effects

**The individual fixed effects in a model without intercept report average staff size per member**

- the fixed effects control away the between-group variation
  - e.g. gender can no longer be estimated (no variation)
- ... to only keep within-group variation
  - e.g. effect of electoral cycle, party size (vary over time)

$\Rightarrow$ *the panel data approach*

# Within-group variation

# Within-group variation

We want to compare the effect of changes in party-funding while holding individual (and thus national) traits constant

- fixed-effects are strictly within individuals
- ... but is the between-individual variation in party-funding really undesirable?

Table 1:

| | *Dependent variable:* | |
|---|---|---|
| | y | |
| | Pooled OLS | OLS w/fixed-effects |
| | (1) | (2) |
| SeatsNatPal.prop | $-0.677^{***}$ | $-1.884^{***}$ |
| | (0.218) | (0.515) |
| Constant | $2.760^{***}$ | $2.111^{***}$ |
| | (0.067) | (0.492) |
| Observations | 5,577 | 5,577 |
| $R^2$ | 0.002 | 0.825 |
| Adjusted $R^2$ | 0.002 | 0.780 |
| Residual Std. Error | 2.902 (df = 5575) | 1.363 (df = 4435) |
| F Statistic | $9.658^{***}$ (df = 1; 5575) | $18.318^{***}$ (df = 1141; 4435) |

*Note:* $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

# Between-group variation

# Between-group variation

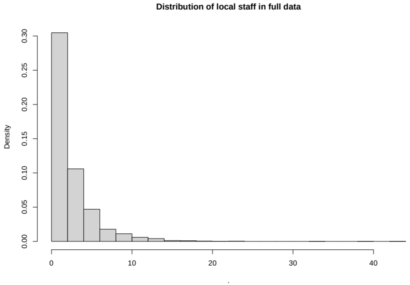
**In fact, most of the variation is *between* individuals**

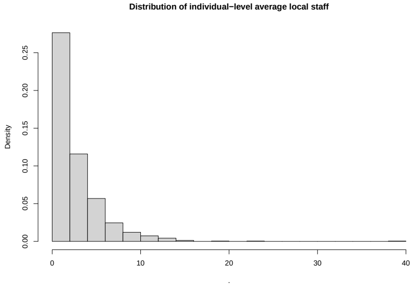
- the variation within individuals may not be representative

```
##      sd_all         sd_group
##  Min.   :0.1785   Min.   :0.00000
##  1st Qu.:0.1785   1st Qu.:0.00000
##  Median :0.1785   Median :0.00466
##  Mean   :0.1785   Mean   :0.02074
##  3rd Qu.:0.1785   3rd Qu.:0.03423
##  Max.   :0.1785   Max.   :0.25664
##                   NA's   :136
```

# A new level of analysis

**The fixed effects without other predictors give us a "new" aggregated data frame with observations at the individual level**



# Between-group regression

# Between-group regression

I can keep the between-group variation by regressing my fixed effects on national party size (i.e. funding).

```
##
## Call:
## lm(formula = y_a ~ SeatsNatPal.prop, data = df1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.749 -1.748 -0.717  0.969 36.814
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.7490     0.1454  18.906   <2e-16 ***
## SeatsNatPal.prop  -0.3079     0.4867  -0.633    0.527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.885 on 1139 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.0003513,  Adjusted R-squared:  -0.0005264
## F-statistic: 0.4002 on 1 and 1139 DF,  p-value: 0.5271
```

- a new data frame with group-averages (one per MEP)
- regress on party size

# Trade-offs

- I don't control for all the individual-level confounders
- I put too much weight to MEPs that are observed only a few times

MEPs from majority parties stay longer in office; there are too many small parties in the sample

```
## # A tibble: 2 x 2
##   Majority Periods
##      <dbl>   <dbl>
## 1        0    4.84
## 2        1    5.66
```

# Random intercepts

# Random intercepts

The **hierarchical model** allows me to manage my variation better.

- consider the random-intercept only model:

$$y_i \sim \alpha_j$$

- group intercepts are defined by both types of information
- the weight of each depends on:
    - size of the groups
    - within-group variation
    - between-group variation

# Random intercepts

```
## # A tibble: 1,161 x 6
##       ID y_j   sd_y   n_j  sd_a y_all
##    <int> <dbl> <dbl> <int> <dbl> <dbl>
##  1   840 1.75   0.463     8  2.96  2.71
##  2   988 2.8    0.837     5  2.96  2.71
##  3   997 2.6    0.894     5  2.96  2.71
##  4  1023 3.25   0.463     8  2.96  2.71
##  5  1037 1.62   0.518     8  2.96  2.71
##  6  1038 0.625  0.518     8  2.96  2.71
##  7  1055 2      0.535     8  2.96  2.71
##  8  1059 2     NA         1  2.96  2.71
##  9  1073 6.1    0.224     5  2.96  2.71
## 10  1122 0.2    0.447     5  2.96  2.71
## # ... with 1,151 more rows
```

$$\alpha_j \sim \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}$$

- $n_j$: number of observations of the MEP (size of group)
- $\sigma_y^2$: variance within the MEP (within-group variation)
- $\bar{y}_j$: group estimate (group means)
- $\sigma_\alpha^2$: variance between MEPs (between-group variation)
- $\bar{y}_{all}$: overall mean (mean of means)

# Fit a random-intercept model

We can fit a random intercept model with an intercept for each MEP

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + (1 | ID)
##    Data: df
##
## REML criterion at convergence: 23402.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.8823 -0.2281 -0.0566  0.2532 26.3067
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 7.914    2.813
##  Residual             1.905    1.380
## Number of obs: 5729, groups:  ID, 1161
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.69962    0.08505   31.74
```

- $\hat{\sigma}_\alpha$ (between-group variation): 2.8131432
- $\hat{\sigma}_y$ (within-group variation): 1.3801693

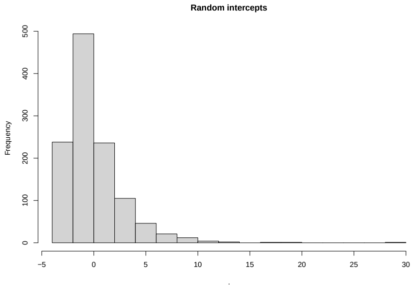
**intra-class correlation**

- $\hat{\sigma}^2_\alpha/(\hat{\sigma}^2_\alpha + \hat{\sigma}^2_y) = 0.67$
- 0 (grouping contributes with no info) to 1 (groups are homogenous)

# The random intercepts

## The random intercepts can also be reported separately

```
## $ID
##          (Intercept)
## 840    -0.9218818205
## 988     0.0957703069
## 997    -0.0950437968
## 1023    0.5343046665
## 1037   -1.0432306944
## 1038   -2.0140216857
## 1055   -0.6791840727
## 1059   -0.5638894996
## 1073    3.2442030192
## 1122   -2.3848130421
## 1129   -0.6674861081
## 1164   -0.0724397031
## 1179   -2.4991054029
## 1183   -0.8583002119
## 1185   -2.4991054029
## 1186   -1.6215566270
## 1191    0.8590267220
## 1204   -1.1645795683
## 1253   -1.6215566270
## 1263    1.2623979100
## 1309    0.2865844107
## 1351   -2.1758791452
## 1394    3.7212382786
## 1403   -0.6791840727
## 1405   -0.0950437968
## 1407   -0.6674861081
## 1419    0.2865844107
```

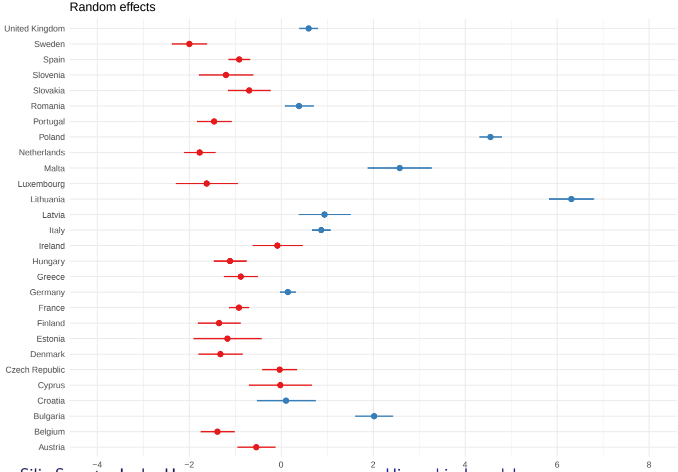



- we have an intercept per MEP
- they are centered around zero/a grand mean/intercept
- we read them in relation to the mean/intercept

# Reporting

**It is common to illustrate them in a coefplot if there are reasonably few**

- here, there are a bit too many, so I illustrate with nationality



# Pooling and smoothing

- the **group intercept** weighs *more* when:
  - group size is consequential ($n_j$)
  - between-group variation is large/groups are distinguishable ($\sigma^2_\alpha$)
- the **group intercept** weighs *less* when:
  - group size is small
  - within-group variation is large/group is "mushy" $\sigma^2_y$
- the **grand mean** (mean of means) steps in to compensate:
  - when a group is small or imprecise
  - when groups are indistinguishable

$$\alpha_j \sim \frac{\frac{n_j}{\sigma^2_y}\bar{y}_j + \frac{1}{\sigma^2_\alpha}\bar{y}_{all}}{\frac{n_j}{\sigma^2_y} + \frac{1}{\sigma^2_\alpha}}$$

# Varying intercepts with predictors

# Varying intercepts with predictors

**We are not generally interested in the intercepts**

- they are a way to cluster the errors
- give correct standard errors for level-two variables

# Fit a varying intercept model with a fixed predictor

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ SeatsNatPal.prop + (1 | ID)
##    Data: df
##
## REML criterion at convergence: 22650.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.9897 -0.2201 -0.0519  0.2502 26.4800
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 7.464    2.732
##  Residual             1.868    1.367
## Number of obs: 5577, groups:  ID, 1141
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)        2.9114     0.1189  24.492
## SeatsNatPal.prop  -1.0529     0.3496  -3.011
##
## Correlation of Fixed Effects:
##             (Intr)
## StsNtPl.prp -0.712
```

- 

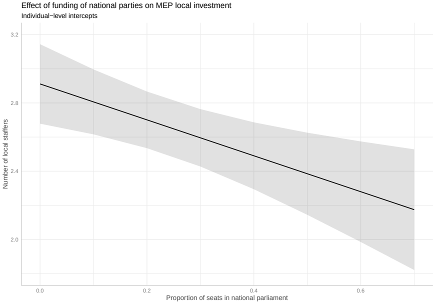
# Results from four models

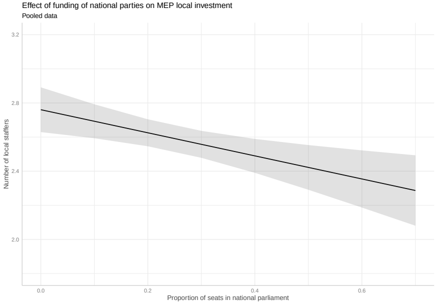
# Regression table

Table 2:

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | y | | y_a | y |
| | *OLS* | | *OLS* | *linear mixed-effe* |
| | Pooled | Fixed effects | Grouped | Random eff |
| | (1) | (2) | (3) | (4) |
| SeatsNatPal.prop | −0.677*** | −1.884*** | −0.308 | −1.053** |
| | (0.218) | (0.515) | (0.487) | (0.350) |
| Constant | 2.760*** | 2.111*** | 2.749*** | 2.911*** |
| | (0.067) | (0.492) | (0.145) | (0.119) |
| Observations | 5,577 | 5,577 | 1,141 | 5,577 |
| $R^2$ | 0.002 | 0.825 | 0.0004 | |
| Adjusted $R^2$ | 0.002 | 0.780 | −0.001 | |
| Log Likelihood | | | | −11,325.4 |
| Akaike Inf. Crit. | | | | 22,658.82 |
| Bayesian Inf. Crit. | | | | 22,685.33 |
| Residual Std. Error | 2.902 (df = 5575) | 1.363 (df = 4435) | 2.885 (df = 1139) | |
| F Statistic | 9.658*** (df = 1; 5575) | 18.318*** (df = 1141; 4435) | 0.400 (df = 1; 1139) | |
| *Note:* | | | | *p<0.1; **p<0.05; ***p< |

# Effect plot


Effect of funding of national parties on MEP local investment
Individual−level intercepts
Number of local staffers
3.2
2.8
2.4
2.0
0.0
0.2
0.4
0.6
Proportion of seats in national parliament


Effect of funding of national parties on MEP local investment
Pooled data
Number of local staffers
3.2
2.8
2.4
2.0
0.0
0.2
0.4
0.6
Proportion of seats in national parliament

# Varying slopes

# Varying slopes

**We sometimes want to know if the slope is similar across groups**

- we do this through interactions
- let's check if women have as many staffers as men

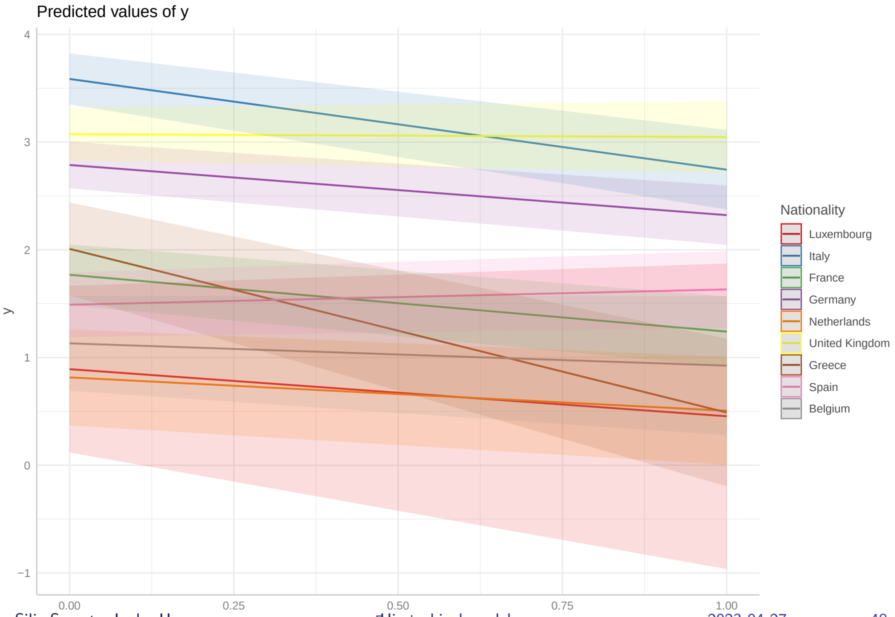
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ Female + (1 | ID) + (1 | Nationality)
##    Data: df
##
## REML criterion at convergence: 22905.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.6986 -0.2442 -0.0351  0.2683 26.0773
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  ID          (Intercept) 4.532    2.129
##  Nationality (Intercept) 4.135    2.033
##  Residual                1.910    1.382
## Number of obs: 5729, groups:  ID, 1161; Nationality, 28
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.6918     0.3975   6.772
## Female       -0.4299     0.1393  -3.086
##
## Correlation of Fixed Effects:
##        (Intr)
## Female -0.129
```

# Fixed effects with interaction

## The brutal way of estimating intercepts and slopes is with an interaction

```
##
## Call:
## lm(formula = y ~ Female * Nationality, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.360 -1.300 -0.322  0.841 33.640
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.297619   0.261949   8.771  < 2e-16 ***
## Female                           -1.017619   0.428829  -2.373 0.017676 *
## NationalityBelgium               -1.166040   0.345220  -3.378 0.000736 ***
## NationalityBulgaria               2.361722   0.363257   6.502 8.63e-11 ***
## NationalityCroatia               -0.964286   0.531124  -1.816 0.069491 .
## NationalityCyprus                -0.081403   0.473705  -0.172 0.863567
## NationalityCzech Republic         0.274962   0.339263   0.810 0.417706
## NationalityDenmark               -1.130952   0.411992  -2.745 0.006069 **
## NationalityEstonia               -0.922619   0.654872  -1.409 0.158933
## NationalityFinland               -1.297619   0.453708  -2.860 0.004252 **
## NationalityFrance                -0.529633   0.298915  -1.772 0.076473 .
## NationalityGermany                0.489749   0.284168   1.723 0.084862 .
## NationalityGreece                -0.289216   0.342130  -0.845 0.397957
## NationalityHungary               -0.785619   0.338715  -2.319 0.020408 *
## NationalityIreland                0.270563   0.446781   0.606 0.544816
## NationalityItaly                  1.288675   0.288524   4.466 8.11e-06 ***
## NationalityLatvia                 0.748893   0.450178   1.664 0.096258 .
## NationalityLithuania              7.062381   0.381403  18.517  < 2e-16 ***
## NationalityLuxembourg            -1.405727   0.473705  -2.968 0.003015 **
```

# Fixed effects with interaction: illustrated



# Random effects with interaction

# Random effects with interaction

**The smoother way is to make the interaction with random effects**

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ (Female | Nationality)
##    Data: df
##
## REML criterion at convergence: 26454.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4160 -0.5565 -0.1400  0.3590 14.0749
##
## Random effects:
##  Groups      Name        Variance Std.Dev. Corr
##  Nationality (Intercept) 4.4156   2.1013
##              Female      0.6026   0.7763   -0.70
##  Residual                5.7660   2.4012
## Number of obs: 5729, groups: Nationality, 28
```

# Separate estimates: in numbers

We get separate estimates for each gender/nationality pair with an intercept and a slope

```
## $Nationality
##                      Female (Intercept)
## Austria        -0.682786687   2.1922004
## Belgium        -0.078665680   1.1056759
## Bulgaria       -0.519431708   4.6477406
## Croatia         1.129558801   1.9099753
## Cyprus          0.266615697   2.3407704
## Czech Republic -0.441600150   2.5342893
## Denmark         0.011033704   1.1378970
## Estonia         0.004313265   1.2878466
## Finland         0.171665067   1.0067988
## France         -0.457629252   1.7436918
## Germany        -0.442287304   2.7785125
## Greece         -1.001039566   1.8913030
## Hungary        -0.352722849   1.4645368
## Ireland        -0.294476308   2.5050055
## Italy          -0.782810147   3.5681890
## Latvia          0.344490203   3.2251358
## Lithuania      -2.039448127   9.2027101
## Luxembourg      0.034769460   0.8381631
## Malta          -0.741676434   5.2866482
## Netherlands    -0.166663834   0.7709452
## Poland         -1.643994123   7.3758183
## Portugal       -0.020780981   1.0175385
## Romania        -0.809898298   3.1306021
## Slovakia       -0.597770740   1.9982510
## Slovenia       -0.011463886   1.2668577
## Spain           0.136254830   1.4934447
## Sweden         -0.074087062   0.5093264
```

# Separate estimates: in images