

# Problem Set 5: Multilevel Models

Alejandro Sarria

## Preliminaries

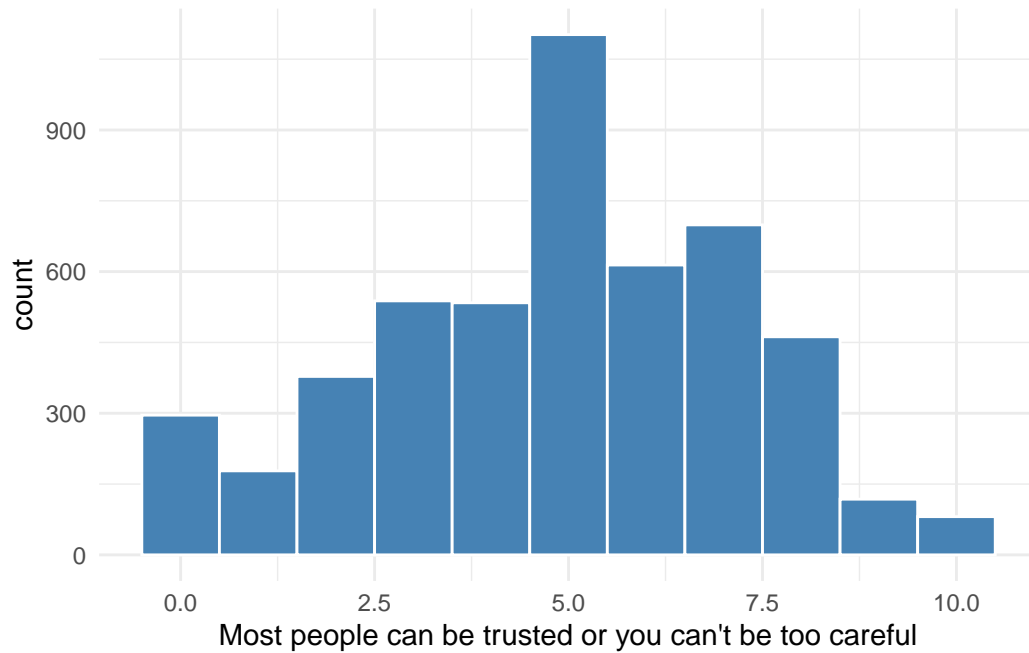
1.

```
psych::describe(d) |>
  filter(vars == 3) |>
  mutate(variance = sd^2) |>
  select(mean, variance)
```

```
      mean variance
ppltrst 4.87      5.55
```

2.

```
d |>
  ggplot(aes(x=ppltrst)) +
  geom_histogram(binwidth = 1, fill="steelblue", color="white") +
  theme_minimal()
```

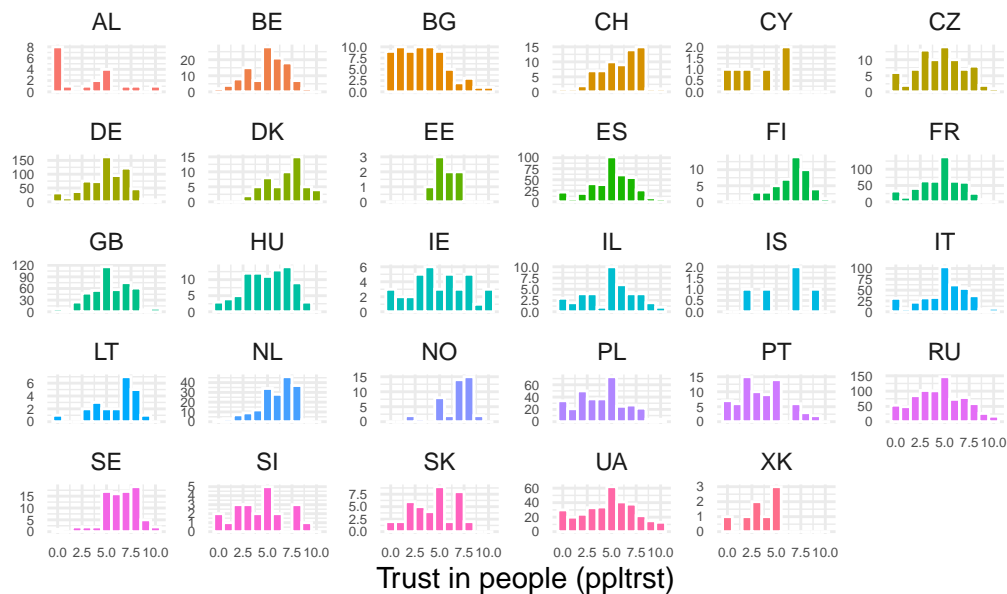


The modal value is 5.

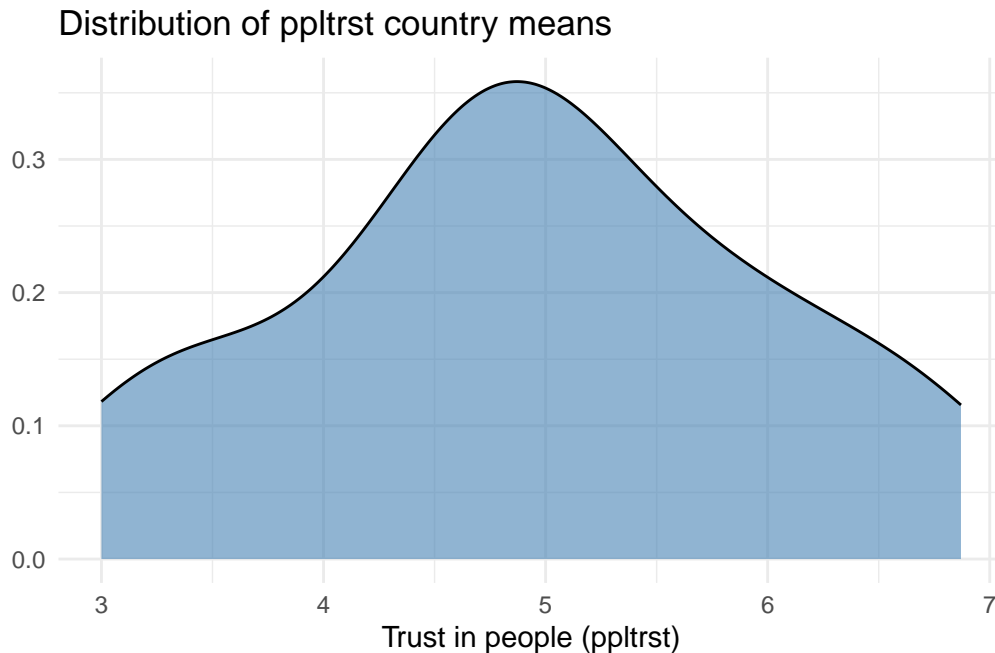
3.

```
d |>
  ggplot(aes(x = ppltrst, fill = cntry)) +
  geom_histogram(binwidth = 1, color = "white") +
  facet_wrap(~cntry, scales = "free_y") +
  theme_minimal() +
  labs(
    title = "Distribution of ppltrst by country",
    x = "Trust in people (ppltrst)",
    y = ""
  ) +
  theme(legend.position = "none",
        axis.text = element_text(size=5))
```

## Distribution of ppltrst by country



```
d |>
  group_by(cntry) |>
  summarise(mean = mean(ppltrst)) |>
  ggplot(aes(x=mean)) +
  geom_density(fill="steelblue", color="black", alpha=0.6) +
  theme_minimal() +
  labs(
    title = "Distribution of ppltrst country means",
    x = "Trust in people (ppltrst)",
    y = ""
  )
```



If we estimate the country means using partial pooling, the estimate of the means of countries with fewer people in the sample will be closer to the overall mean ( $\sim 5$ ). In some countries, this won't have much of an effect, but in cases like Albania, Cyprus and Iceland the estimated mean will be quite different from the sample mean.

## Single Dimension

4.

```
m1 <- lmer(ppltrst ~ 1 + (1 | cntry),
           data = d,
           REML = FALSE)
summary(m1)
```

Linear mixed model fit by maximum likelihood ['lmerMod']  
 Formula: ppltrst ~ 1 + (1 | cntry)  
 Data: d

AIC	BIC	logLik	-2*log(L)	df.resid
22490.9	22510.4	-11242.4	22484.9	4997

Scaled residuals:

	Min	1Q	Median	3Q	Max
	-2.79077	-0.66911	0.00127	0.67580	2.87402

Random effects:

Groups	Name	Variance	Std.Dev.
cntry	(Intercept)	0.8744	0.9351
Residual		5.1737	2.2746

Number of obs: 5000, groups: cntry, 29

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	4.9732	0.1866	26.66

5.

```
icc <- 0.8744 / (0.8744 + 5.1737)
icc
```

```
[1] 0.1445743
```

6.

```
performance::icc(m1)
```

# Intraclass Correlation Coefficient

Adjusted ICC: 0.145  
Unadjusted ICC: 0.145

7.

An ICC of 0.145 can be interpreted as saying that 14.5% percent of the variance in ppltrst is explained by variation between countries, meaning that the resulting 76.5% of variance is explained by within country variation.

8.

```
m2 <- lm(ppltrst ~ 1,
        data = d)
summary(m2)
```

Call:

```
lm(formula = ppltrst ~ 1, data = d)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.8678	-1.8678	0.1322	2.1322	5.1322

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.86780	0.03331	146.1	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.355 on 4999 degrees of freedom

The mean estimate for the partially pooled model (4.9732) is not significantly different from the completely pooled model (4.8678). Suggesting that the complexity of random intercepts is not necessary in this case.

9.

```
global_mean <- fixef(m1)["(Intercept)"]

eb <- ranef(m1)$cntry |>
  rownames_to_column(var="cntry") |>
  rename(u_j = `(Intercept)`) |>
  mutate(partial_mean = global_mean + u_j)

#highest mean
eb |>
  arrange(desc(partial_mean)) |>
  slice(1)
```

```

      cntry      u_j partial_mean
1      DK 1.709806      6.683016

```

```

#lowest mean
eb |>
  arrange(partial_mean) |>
  slice(1)

```

```

      cntry      u_j partial_mean
1      BG -1.510355      3.462854

```

```

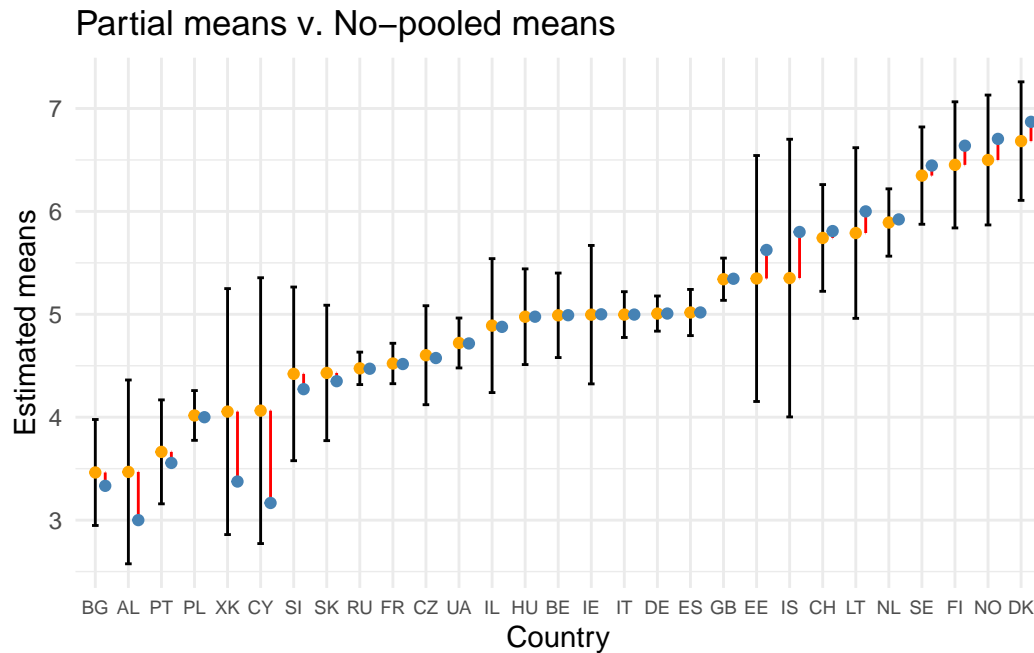
se <- attr(ranef(m1, condVar=TRUE)[[1]], "postVar")
eb$se <- sqrt(se[1,1,])

```

```

d |>
  group_by(cntry) |>
  summarise(no_pool_mean = mean(ppltrst)) |>
  inner_join(eb, by=join_by(cntry)) |>
  ggplot(aes(x=reorder(cntry, partial_mean))) +
  geom_errorbar(aes(ymin = partial_mean - 1.96*se,
                    ymax = partial_mean + 1.96*se),
                width = 0.2) +
  geom_point(aes(y=partial_mean),
             color="orange") +
  geom_errorbar(aes(ymin = partial_mean,
                    ymax = no_pool_mean),
                width = 0.01,
                position = position_nudge(0.3),
                color="red") +
  geom_point(aes(y=no_pool_mean),
             color="steelblue",
             position=position_nudge(0.3)) +
  theme_minimal() +
  labs(
    title = "Partial means v. No-pooled means",
    x = "Country",
    y = "Estimated means"
  ) +
  theme(axis.text.x = element_text(size=7))

```



It was the case that the partial means of countries with a low  $n$  differed the most from the their no pooled estimates as they were estimated to be closer to the overall mean.

## Two Dimensions

10.

```
d <- d |>
  mutate(eduyrs_scaled = rescale(eduyrs))

m3 <- lmer(ppltrst ~ eduyrs_scaled + (1 | cntry),
          data = d,
          REML = FALSE)

summary(m3)
```

Linear mixed model fit by maximum likelihood ['lmerMod']  
 Formula: ppltrst ~ eduyrs\_scaled + (1 | cntry)  
 Data: d

AIC	BIC	logLik	-2*log(L)	df.resid
-----	-----	--------	-----------	----------



22409.6    22435.6   -11200.8    22401.6        4996

Scaled residuals:

	Min	1Q	Median	3Q	Max
	-2.80179	-0.67975	0.07135	0.68638	3.01157

Random effects:

Groups	Name	Variance	Std.Dev.
cntry	(Intercept)	0.8144	0.9024
Residual		5.0896	2.2560

Number of obs: 5000, groups: cntry, 29

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	3.9368	0.2131	18.471
eduyrs_scaled	1.6403	0.1789	9.167

Correlation of Fixed Effects:

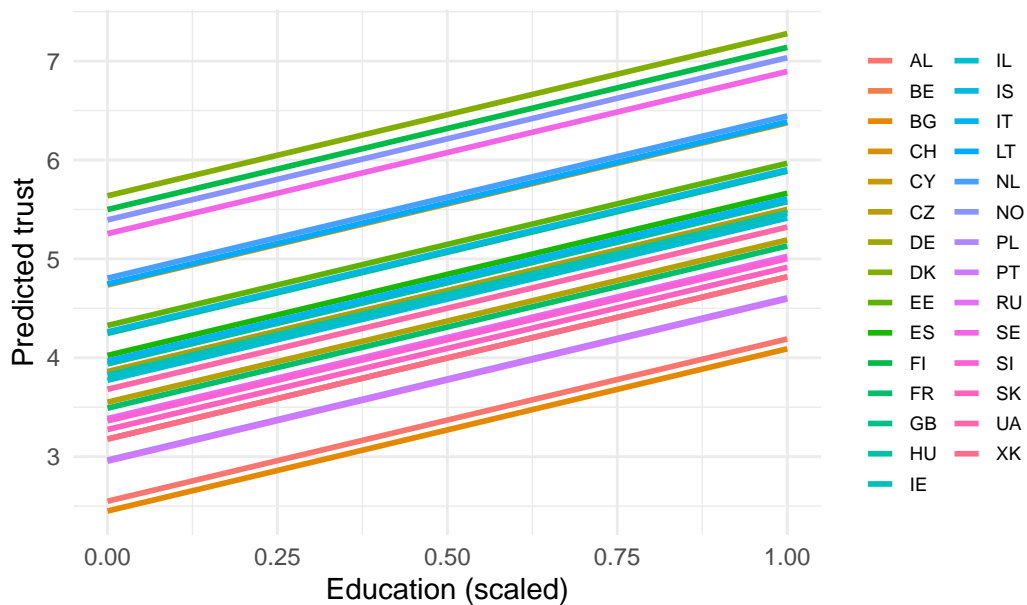
	(Intr)
eduyrs_scld	-0.531

The maximum effect for education on ppltrst is 1.6403. That means that, for the average country, the most that education can increase trust is 1.6403.

## 11.

```
ggpredict(m3, terms=c("eduyrs_scaled", "cntry"), type="random") |>
  ggplot(aes(x = x , y = predicted , group = group, color = group )) +
  geom_smooth(formula= y ~ x,
              method = "lm",
              alpha = .5, se=FALSE) +
  theme_minimal() +
  labs(title = "Predicted trust by education by country - no pooling",
       y = "Predicted trust",
       x = "Education (scaled)",
       color = NULL) +
  theme(
    legend.position = "right",
    legend.text = element_text(size = 7),
    legend.key.size = unit(0.4, "cm")
  )
```

Predicted trust by education by country – no pooling

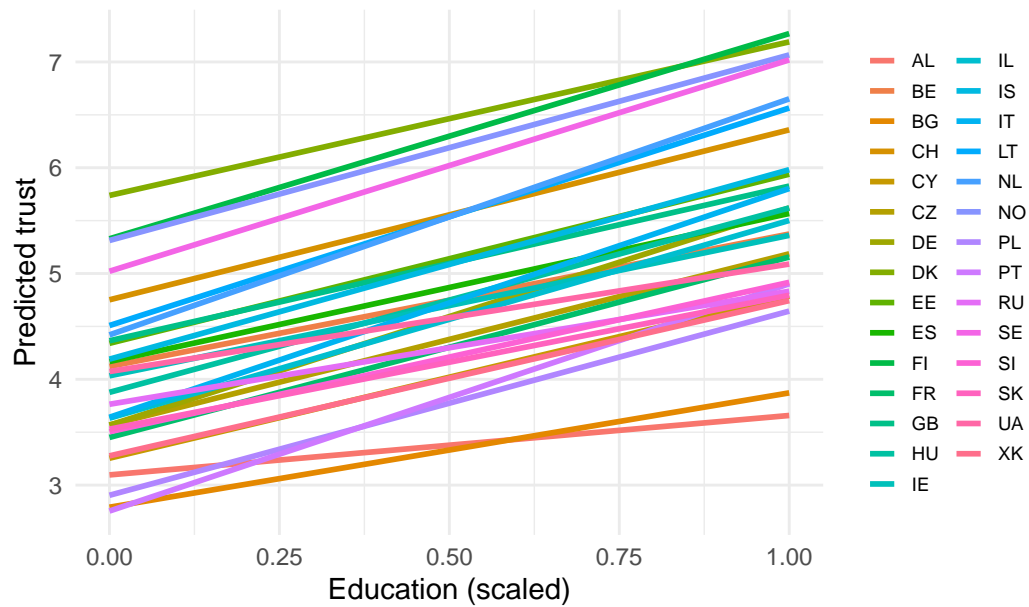


12

```
m4 <- lmer(ppltrst ~ eduyrs_scaled + (eduyrs_scaled | cntry),
  data = d,
  REML = FALSE)

ggpredict(m4, terms=c("eduyrs_scaled", "cntry"), type="random") |>
  ggplot(aes(x = x , y = predicted , group = group, color = group )) +
  geom_smooth(formula= y ~ x,
    method = "lm",
    alpha = .5, se=FALSE) +
  theme_minimal() +
  labs(title = "Predicted trust by education by country - random slopes",
    y = "Predicted trust",
    x = "Education (scaled)",
    color = NULL) +
  theme(
    legend.position = "right",
    legend.text = element_text(size = 7),
    legend.key.size = unit(0.4, "cm")
  )
```

Predicted trust by education by country – random slopes



13

```
compare_performance(m3, m4, metrics=c("BIC", "RMSE", "R2"))
```

Some of the nested models seem to be identical and probably only vary in their random effects.

# Comparison of Model Performance Indices

Name	Model	BIC (weights)	R2 (cond.)	R2 (marg.)	RMSE
m3	lmerMod	22435.6 (0.999)	0.151	0.015	2.250
m4	lmerMod	22450.6 (<.001)	0.155	0.014	2.246

The model without random slopes performs better. This means that the effect of education on trust does not vary significantly across countries.

```
d <- d |>
  mutate(agea_scaled = rescale(agea),
         attend_scaled = rescale(attend),
         health_scaled = rescale(health))

m5 <- lmer(ppltrst ~ eduyrs_scaled + I(agea_scaled^2) + attend_scaled +
         health_scaled + female + (1 | cntry),
         data = d,
         REML = FALSE)

ggeffect(m5, terms="agea_scaled") |>
  plot() +
  labs(title = "Effect of age on trust",
       y = "Predicted trust",
       x = "Age (scaled)")
```

