Code

MARP exploration - Team 40b

Tamas Nagy

2021-02-14

```
# Packeges needed
install.packages(c("tidyverse", "psych", "skimr", "ggridges", "tidytext", "lme4",
"performance", "sjPlot", "here", "emo", "broom.mixed"))
```

```
library(tidyverse)
library(psych)
library(skimr)
library(ggridges)
library(tidytext)
library(lme4)
library(performance)
library(sjPlot)
library(broom.mixed)

theme_set(theme_light())
```

Main questions

In this blind analysis, we are going to address 2 research questions:

1. Do religious people report higher well-being?
2. Does the relation between religiosity and well-being depend on how important people consider religion to be in their country (i.e., perceived cultural norms of religion)?

Blinded data

The point of blinded data is to let analysts explore the data without the danger of p-hacking. Therefore, only the relationship of the outcomes and predctor variables are destroyed by shuffling, and the remainder is kept intact. This means that:

- Data reduction techinques (PCA, EFA) will yield valid results.
- Outiers, missing data can be observed and treated.
- Confounders can be explored, meaning that not only univariate distributions but correlations are also kept intact.
- Country level means should remain intact

Analysis issues to resolve

In order to best address the hypotheses, we need to decide a few things.

- Operationalization of variables: How should we conceptualize key variables?
- Outliers how should we handle them?
- Choose statistical model (multilevel?)
    - Confounders (which ones to use?)
    - Moderators
    - Lumping levels for nominal variables

Read and pre-process data

At this point, we only exclude participants who did not pass the attention check.

```
# Read raw data
marp_raw <-
  read_csv(here::here("data/MARP_data.csv"))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   country = col_character(),
##   gender = col_character(),
##   ethnicity = col_character(),
##   denomination = col_character(),
##   sample_type = col_character(),
##   compensation = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
marp_proc <-
  marp_raw %>%
  filter(attention_check == 1) %>%
  mutate( gender = recode(gender,
                          "man" = "Male",
                          "woman" = "Female",
                          "other" = "Other"))
```

Descriptives and potential moderators

```
marp_proc %>%
        select(country,
               gender, ses, education, ethnicity, denomination,
               sample_type, compensation,
               ends_with("_mean")) %>%
        skim()
```

Data summary

Name	Piped data
Number of rows	10195
Number of columns	12
_______________________	
Column type frequency:	
character	6
numeric	6
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
country	0	1.00	2	11	0	24	0
gender	0	1.00	4	6	0	3	0
ethnicity	405	0.96	5	22	0	17	0
denomination	5567	0.45	4	41	0	20	0
sample_type	0	1.00	5	14	0	4	0
compensation	0	1.00	6	31	0	5	0

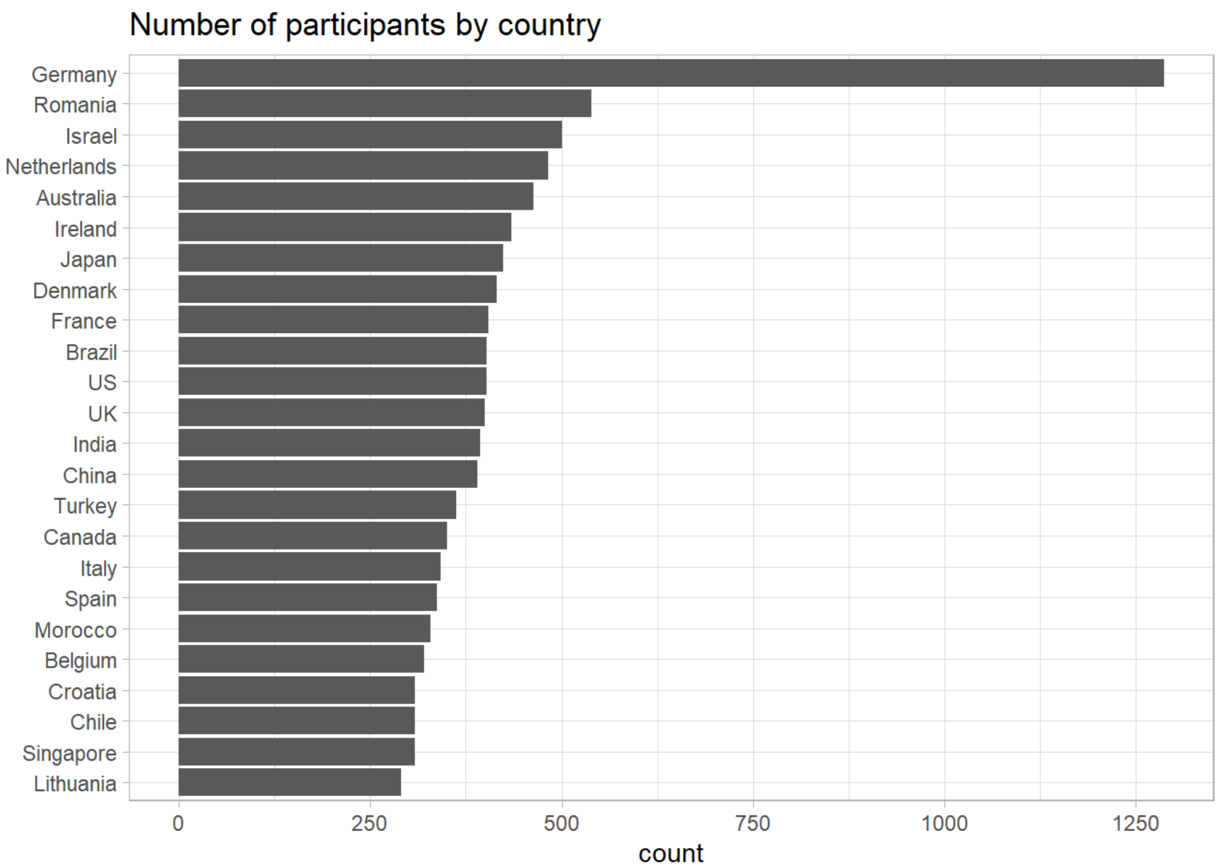
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ses	5	1	6.10	1.77	1.00	5.00	6.00	7.00	10	
education	0	1	4.64	1.26	1.00	4.00	5.00	5.00	7	
wb_overall_mean	0	1	3.67	0.61	1.22	3.28	3.78	4.11	5	
wb_phys_mean	0	1	3.84	0.66	1.00	3.43	4.00	4.29	5	
wb_psych_mean	0	1	3.51	0.72	1.00	3.00	3.67	4.00	5	
wb_soc_mean	0	1	3.56	0.87	1.00	3.00	3.67	4.17	5	

Demographics by country

N of participants | Age | Gender | Education | SES | Denomination | Ethnicity | Importance of religion | GDP per capita | Sample type | Compensation

```
marp_proc %>%
        mutate(country = fct_infreq(country) %>% fct_rev()) %>%
        ggplot() +
        aes(y = country) +
        geom_bar() +
        scale_x_continuous(breaks = seq(0, 1500, 250)) +
        labs(title = "Number of participants by country",
             y = NULL)
```



Operationalization of variables

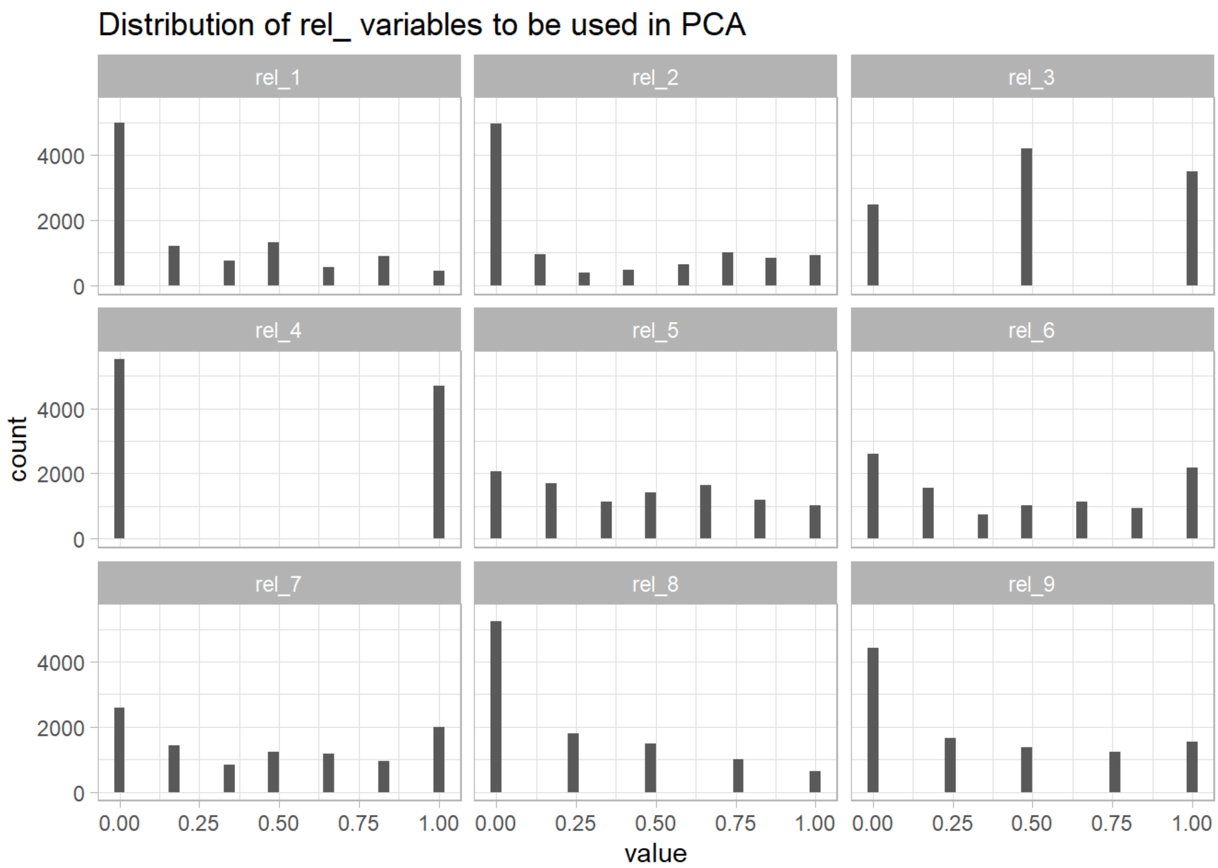
Religiosity and *well-being* has multiple items that we can use. According to previous studies of similar topics, *norms about religion* should be aggregated to the country level.

Religiosity | Well-being | Cultural norms about religiosity into one variable.

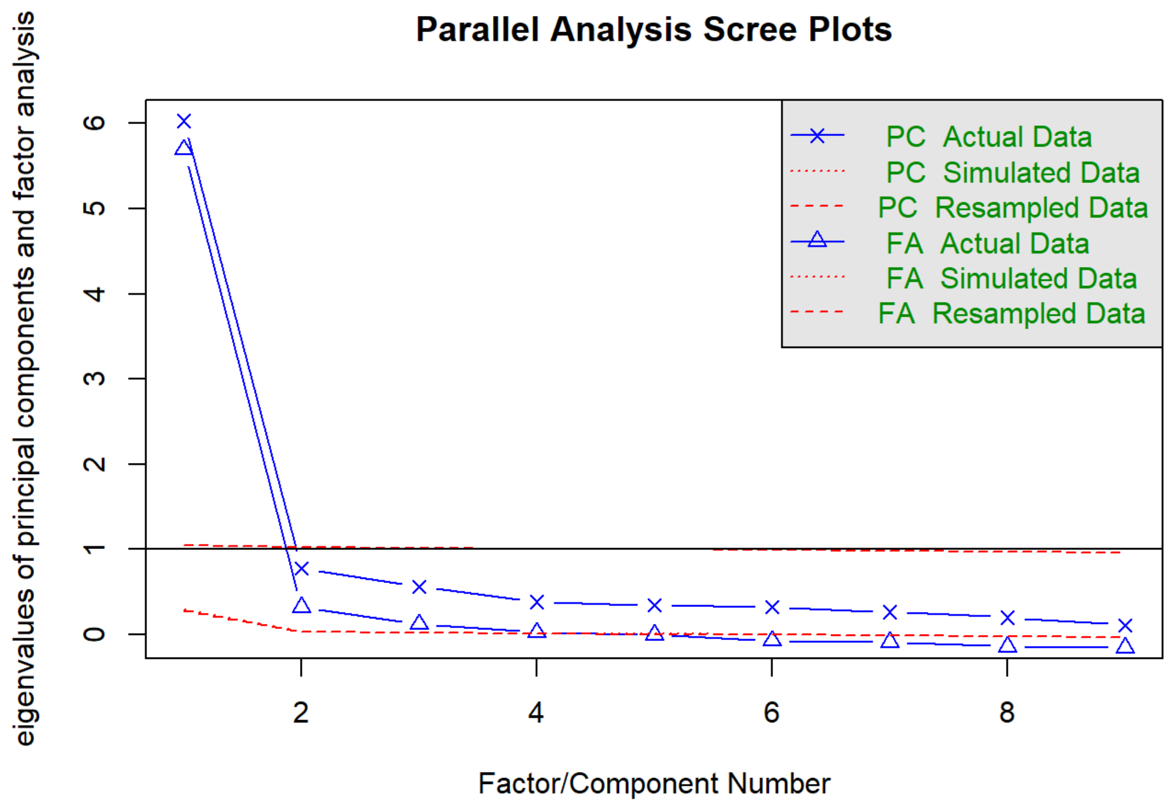
Although religiosity is an elusive concept, and no one-size-fits-all metric is available. We don't feel competent to choose just one question, so We try to use as much information from all available questions as possible. I'm also not feeling confident to relevel specific questions (e.g. rel_3). Therefore, We choose to use PCA to extract an aggregated variable.

```r
marp_proc %>%
  select(starts_with("rel_")) %>%
  pivot_longer(everything()) %>%
  ggplot(aes(value)) +
  geom_histogram() +
  facet_wrap(~name) +
  labs(title = "Distribution of rel_ variables to be used in PCA")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

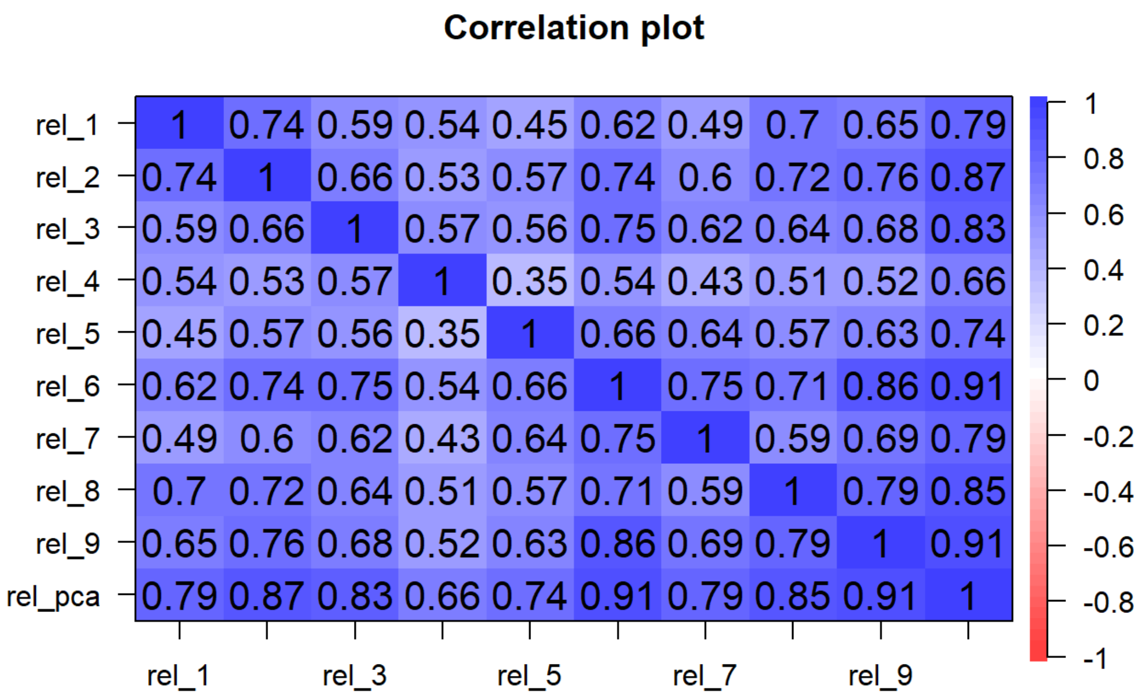



```r
# Number of components
marp_proc %>%
        select(starts_with("rel_")) %>%
        fa.parallel()
```



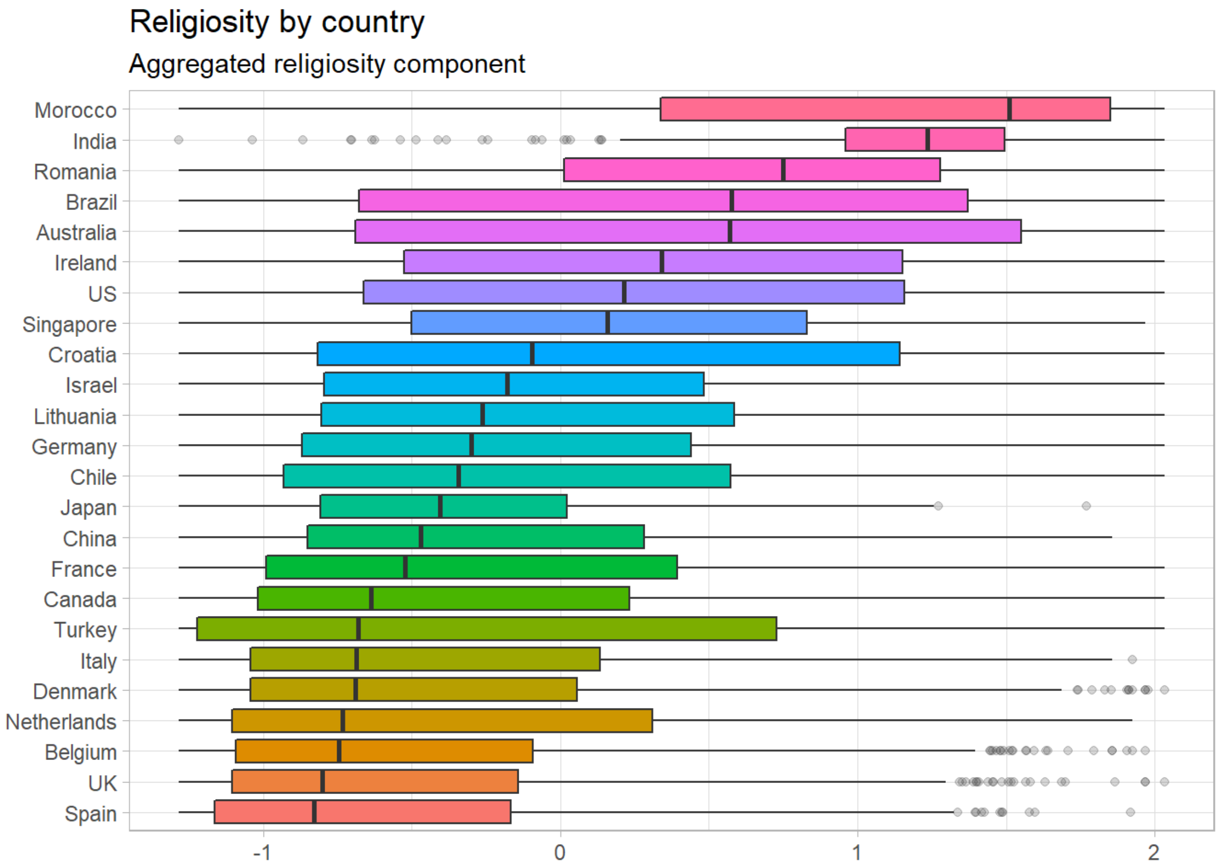

```
## Parallel analysis suggests that the number of factors =  4  and the number of co
mponents =  1
```

```
# Parallel analysis suggests to use a single component.
rel_pca <-
        marp_proc %>%
        select(starts_with("rel_")) %>%
        pca(nfactors = 1)

# Correlation of religiosity items and the PCA component
marp_proc %>%
  select(starts_with("rel")) %>%
  mutate(rel_pca = rel_pca$scores[,1]) %>%
  cor.plot()
```



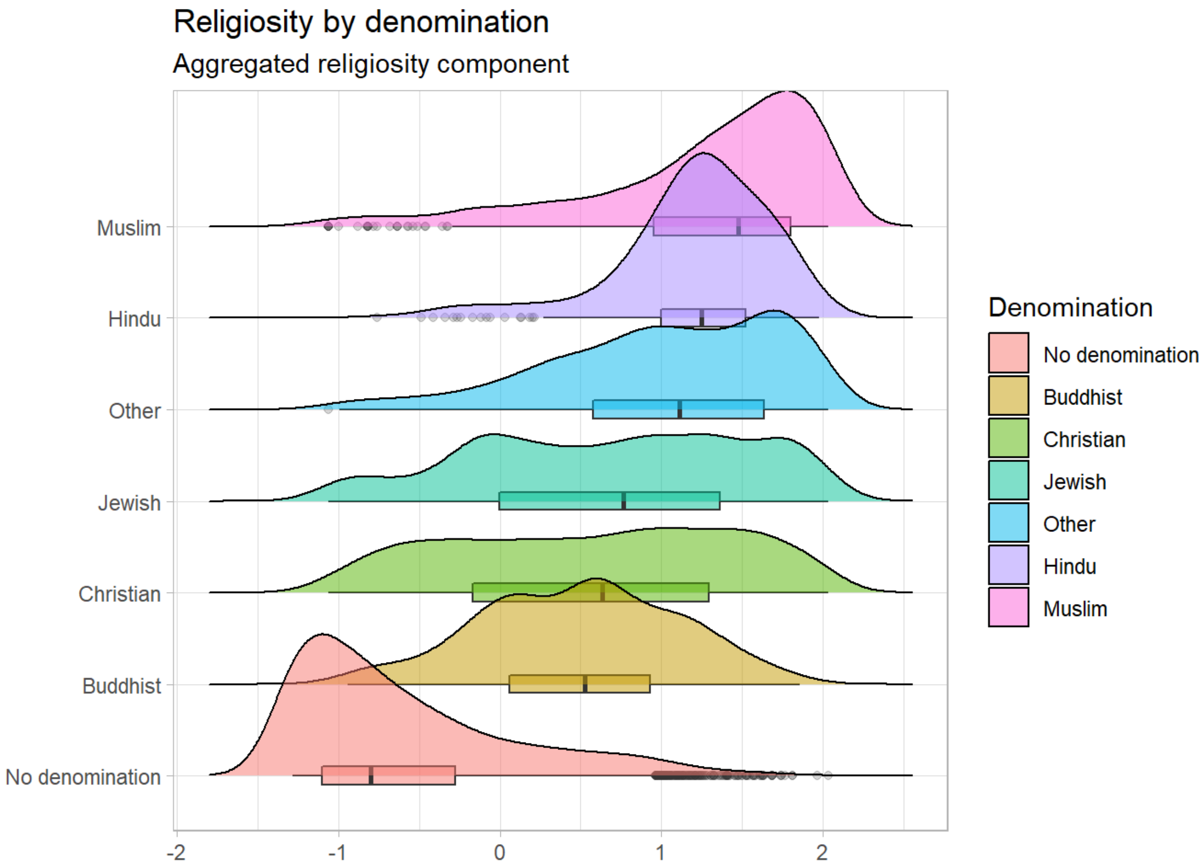
The religiosity values seems to vary considerably by country.

```
marp_proc %>%
mutate(rel_pca = rel_pca$scores[,1]) %>%
mutate(country = fct_reorder(country, rel_pca, median)) %>%
ggplot() +
aes(y = country, x = rel_pca, fill = country) +
geom_boxplot(outlier.alpha = .2, show.legend = FALSE) +
labs(title = "Religiosity by country",
     subtitle = "Aggregated religiosity component",
     x = NULL,
     y = NULL)
```




```
marp_proc %>%
  bind_cols(rel_pca = rel_pca$scores[,1]) %>%
  left_join(lumped_denom, by = "subject") %>%
  mutate(denom_lump = fct_reorder(denom_lump, rel_pca, median)) %>%
  ggplot() +
  aes(y = denom_lump, x = rel_pca, fill = denom_lump) +
  geom_boxplot(alpha = .5,
               width = .2,
               outlier.alpha = .2, show.legend = FALSE) +
  geom_density_ridges(alpha = .5) +
  labs(title = "Religiosity by denomination",
       subtitle = "Aggregated religiosity component",
       fill = "Denomination",
       x = NULL,
       y = NULL)
```

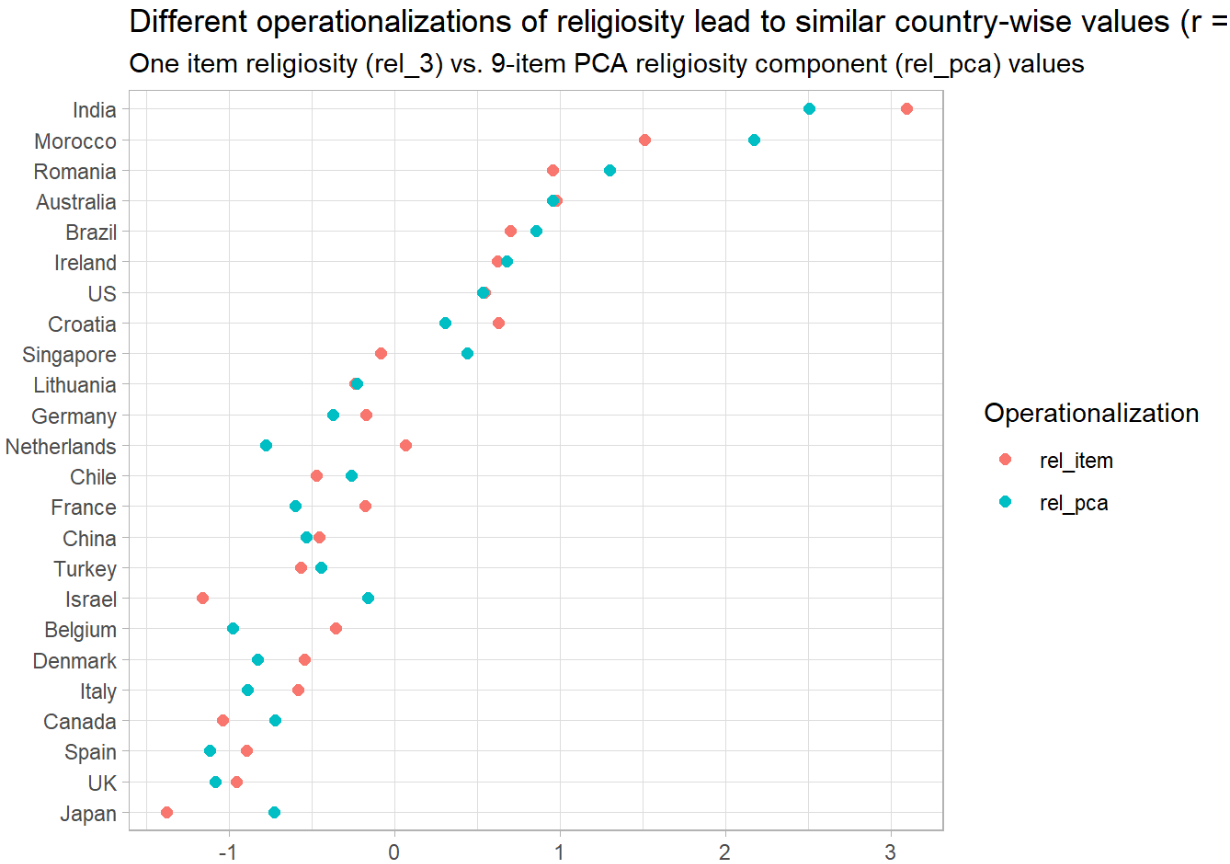
```
## Picking joint bandwidth of 0.173
```



We compared the PCA operationalization with the self-admitted single item religiosity. The difference on the country level seems quite subtle.

```
marp_proc %>%
  mutate(rel_pca = rel_pca$scores[,1]) %>%
  group_by(country) %>%
  summarise(rel_item = mean(rel_3 == 1),
                rel_pca = mean(rel_pca),
                n = n()) %>%
  mutate(across(starts_with("rel_"), ~scale(.x) %>% as.numeric())) %>%
  pivot_longer(cols = c("rel_item", "rel_pca")) %>%
  mutate(country = fct_reorder(country, value)) %>%
  ggplot() +
  aes(x = value, y = country, color = name) +
  geom_point(size = 2) +
  labs(title = "Different operationalizations of religiosity lead to similar countr
y-wise values (r = .91)",
       subtitle = "One item religiosity (rel_3) vs. 9-item PCA religiosity componen
t (rel_pca) values",
       y = NULL, x = NULL, color = "Operationalization")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Construct final dataset

Using all information from the exploratory analysis, we create a dataset for modeling. This dataset still doesn't contain potential problems that may emerge during model diagnostics.

We add the religiosity component, the country-wise norms, the lumped denomination data, and set baselines for categorical variables. We also drop participants with missing values in varibles that we want to use in the statistical models, as those can cause difficulties when comparing models. This means dropping 25 participants.

```
marp_nodiag <-
  marp_proc %>%
  # Add religiosity scores from PCA
  bind_cols(religiosity = rel_pca$scores[,1]) %>%
  # Add country level norms
  left_join(country_norms, by = "country") %>%
  # Merge different branches of the same religion, lump levels < 1%
  left_join(lumped_denom, by = "subject") %>%
  # Set baselines
  mutate(sample_type = fct_relevel(sample_type, "general public"),
         denom_lump = fct_relevel(denom_lump, "No denomination"),
         gender = fct_relevel(gender, "Female")) %>%
  # Drop participants with missing variables
  drop_na(age,  gender, ses, education, denom_lump, sample_type)
```

Investigating model assumptions

Before creating the final dataset and models, we investigate if there is anything strange in the model diagnostics that would necessitate further changes in the dataset. Therefore we create a model that contains all the terms that we want to include in the analysis, and we check all assumptions.

```
model_diag <-
  lmer(wb_overall_mean ~ religiosity * cnorm_mean +
         # personal level confounders
                         age + gender + ses + education + denom_lump +
         # country and sample level confounders
                         gdp_scaled + sample_type +
         # random intercept and slope model
                         (religiosity|country), data = marp_nodiag)

check_model(model_diag)
```

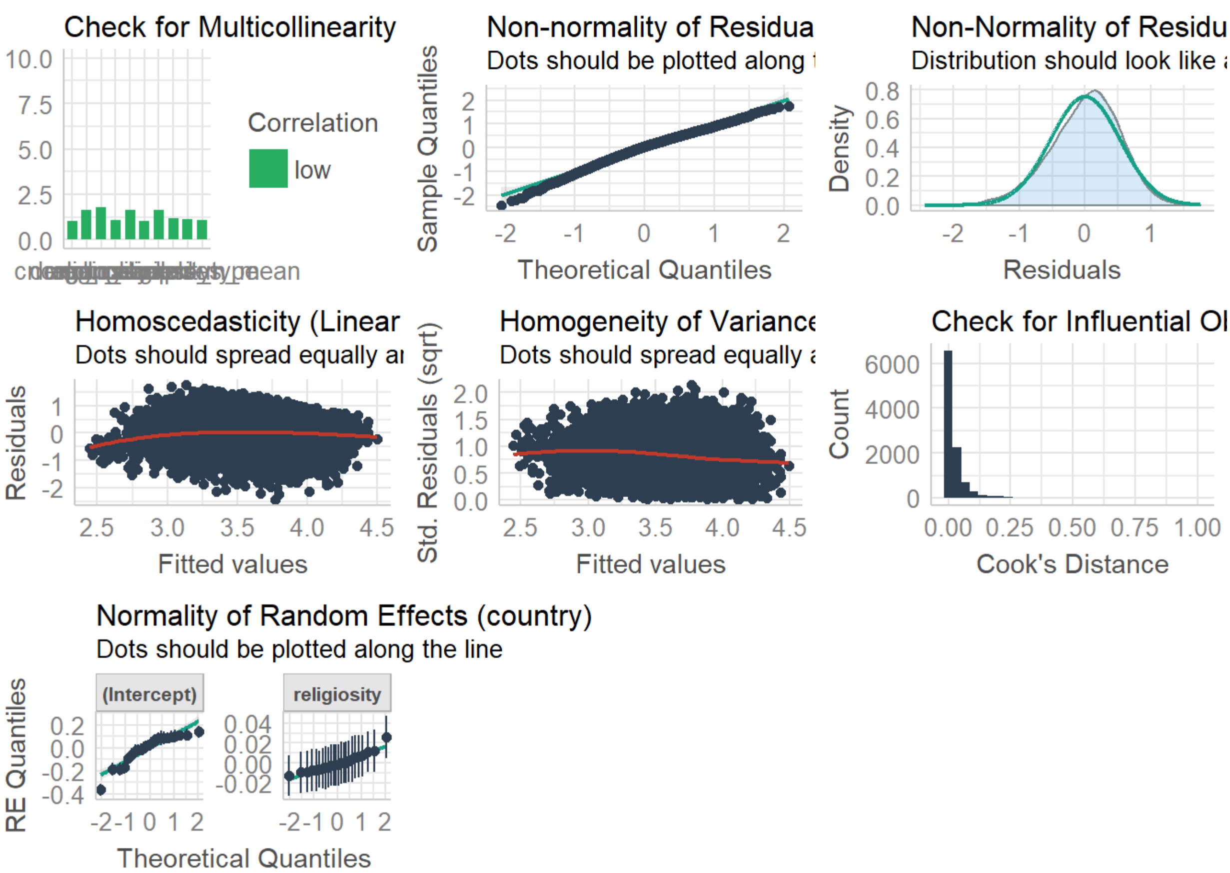
```
## Loading required namespace: qqplotr
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10170 rows containing missing values (geom_text_repel).
```

```
## `geom_smooth()` using formula 'y ~ x'
```

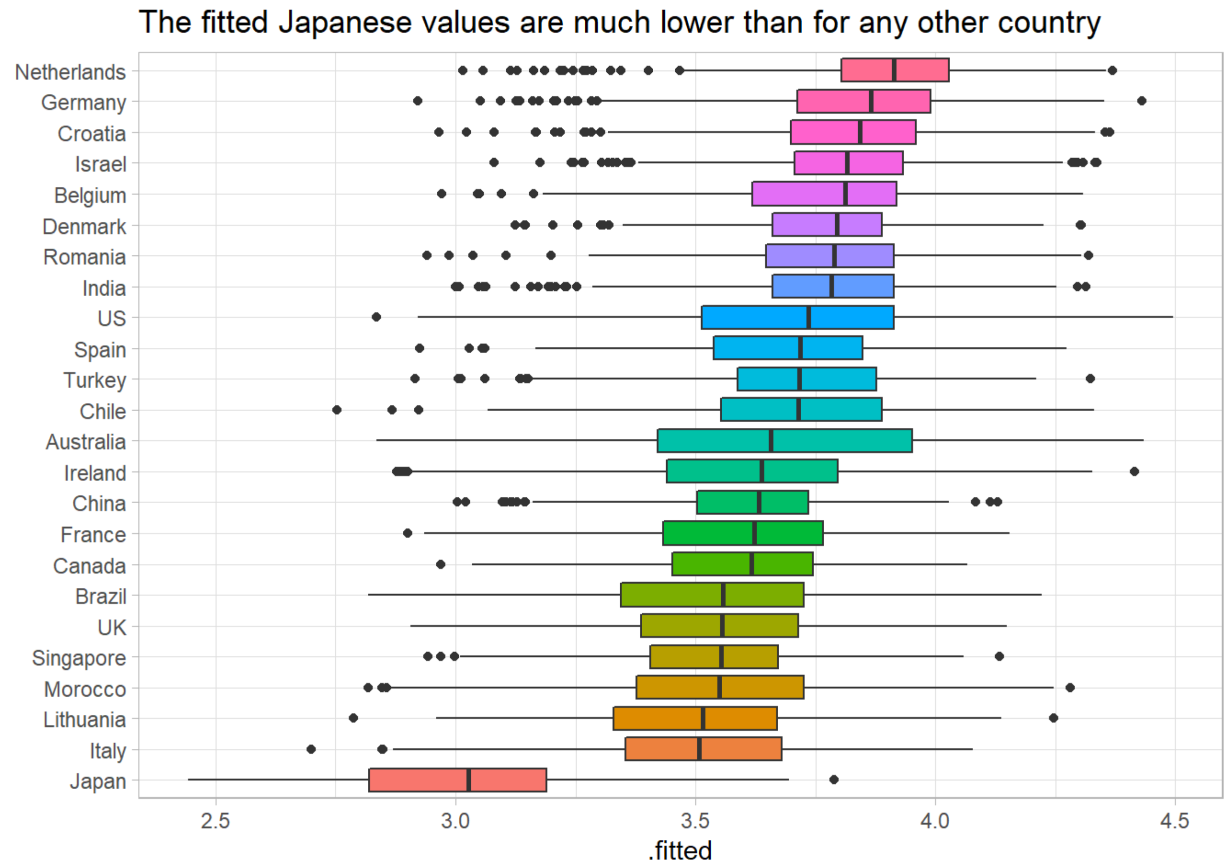


Model diagnostics show:

- ☑ No multicollinearity,
- ☑ Normally distributed residuals
- ☑ No influential cases
- ☑ Normally distributed random effects
- ✗ Homoskedasticity

Apart from heteroscedasticity, it seems like there is a strange separation in the fitted values. All residuals on the left hand side come from the Japanese sample. As the separation is complete and the difference is huge, we should handle the Japanese data with extra care. Further, there is very small variability in the Japanese fitted values.
Taken together, we decided to remove the Japanese data.

```
augment(model_diag) %>%
  mutate(country = fct_reorder(country, .fitted)) %>%
  ggplot() +
  aes(x = .fitted, y = country, fill = country) +
  geom_boxplot(show.legend = FALSE) +
  labs(title = "The fitted Japanese values are much lower than for any other countr
y",
       y = NULL)
```



Correcting the final dataset

In the final dataset we remove the Japanese answers.

```
marp <-
  marp_nodiag %>%
  filter(country != "Japan") %>%
  force()
```

Building models

1) Do religious people report higher well-being?

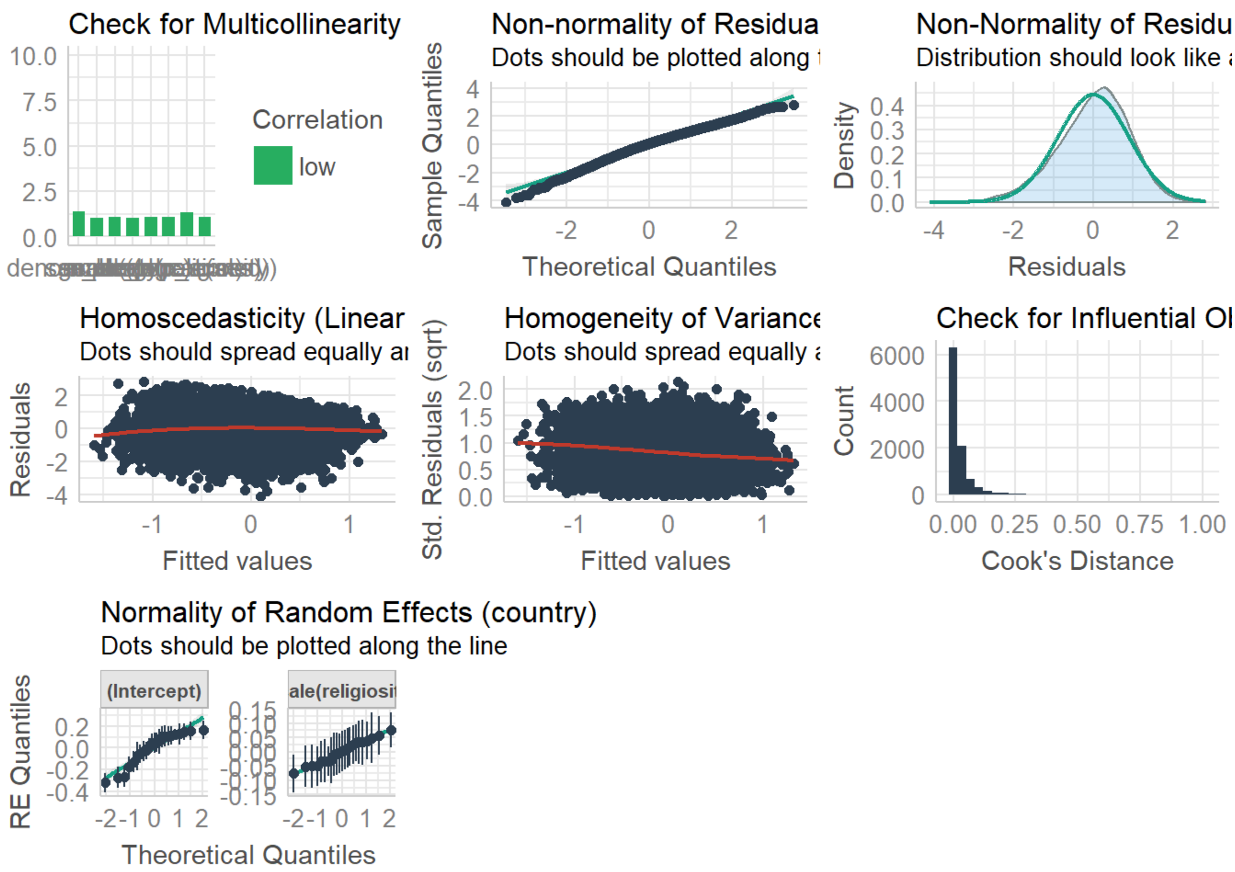
```
h1 <-
  lmer(scale(wb_overall_mean) ~ scale(religiosity) +
         # personal level confounders
         scale(age) + gender + scale(ses) + scale(education) + denom_lump +
         # country and sample level confounders
         scale(gdp_scaled) + sample_type +
         # random intercept and slope model
         (scale(religiosity)|country),
       data = marp)

# Create a null model for comparisons that does not contain the main predictor
h0 <- update(h1, . ~ . -scale(religiosity))

check_model(h1)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 9747 rows containing missing values (geom_text_repel).
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# summary(h1)
```

We can handle heteroscedasticity by using cluster robust standard errors (CR2), using the `clubSandwich` package. https://strengejacke.github.io/sjPlot/articles/tab_model_robust.html (https://strengejacke.github.io/sjPlot/articles/tab_model_robust.html)

```
tab_model(h1,
          show.aic = TRUE,
          show.reflvl = TRUE,
          string.ci = "95% CI",
        # file = "docs/h1_model.html",
          vcov.fun = "CR",
          vcov.type = "CR2",
          vcov.args = list(cluster = h1@frame$country)
          )
```

	scale(wb_overall_mean)		
Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	0.00	-0.12 – 0.13	0.946
Female	*Reference*		
Male	0.07	0.02 – 0.13	**0.011**
Other	-0.48	-0.74 – -0.22	**<0.001**
No denomination	*Reference*		
Buddhist	0.04	-0.09 – 0.16	0.567

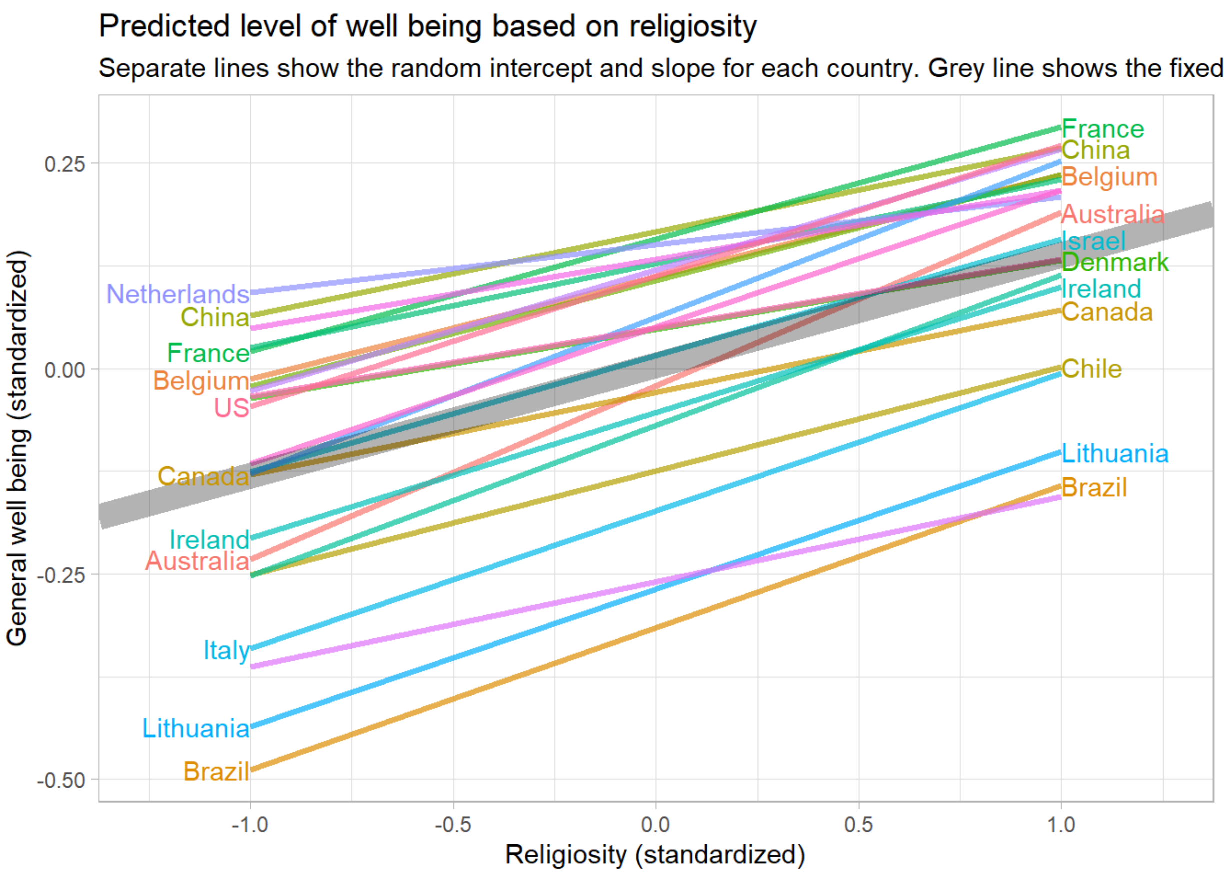
Christian	-0.04	-0.13 – 0.05	0.422
Hindu	-0.14	-0.27 – -0.02	**0.027**
Jewish	-0.13	-0.26 – -0.00	**0.049**
Muslim	-0.15	-0.29 – -0.01	**0.035**
Other	-0.15	-0.31 – 0.01	0.072
general public	*Reference*		
mixed	0.03	-0.13 – 0.19	0.675
online panel	-0.12	-0.18 – -0.06	**<0.001**
students	0.12	-0.28 – 0.52	0.564
scale(age)	0.04	-0.00 – 0.08	0.076
scale(education)	0.08	0.05 – 0.10	**<0.001**
scale(gdp_scaled)	0.04	-0.02 – 0.11	0.204
scale(religiosity)	0.13	0.09 – 0.17	**<0.001**
scale(ses)	0.35	0.30 – 0.39	**<0.001**
Random Effects			
σ^2	0.81		
$\tau_{00\ country}$	0.03		
$\tau_{11\ country.scale(religiosity)}$	0.00		
$\rho_{01\ country}$	-0.27		
ICC	0.03		
$N_{country}$	23		
Observations	9747		
Marginal R^2 / Conditional R^2	0.171 / 0.200		
AIC	25778.669		

Plots

```
# Prepare predictions for plotting
h1_lines <-
  coef(h1)$country %>%
  rownames_to_column("country") %>%
  as_tibble() %>%
  transmute(country,
            intercept = `(Intercept)`,
            slope = `scale(religiosity)`,
            fix_int = fixef(h1)["(Intercept)"],
            fix_slo = fixef(h1)["scale(religiosity)"],
            x = -1,
            xend = 1,
            yend = intercept + slope,
            y = intercept - slope)
```

Predictions for religiosity by country

```
h1_lines %>%
  ggplot() +
  aes(x = x, xend = xend, y = y, yend = yend, color = country) +
  geom_segment(show.legend = FALSE, size = 1.2, alpha = .7) +
    geom_abline(aes(intercept = fix_int,
                    slope = fix_slo),
                color = "black", size = 5, alpha = .3) +
    geom_text(aes(label = country),
              show.legend = FALSE, hjust = 1, check_overlap = TRUE) +
    geom_text(aes(x = xend, y = yend, label = country),
              show.legend = FALSE, hjust = 0, check_overlap = TRUE) +
    xlim(-1.25,1.25) +
    labs(title = "Predicted level of well being based on religiosity",
         subtitle = "Separate lines show the random intercept and slope for each co
untry. Grey line shows the fixed effect",
         y = "General well being (standardized)",
         x = "Religiosity (standardized)")
```



2) Does the relation between religiosity and well-being depend on how important people consider religion to be in their country (i.e., perceived cultural norms of religion)?

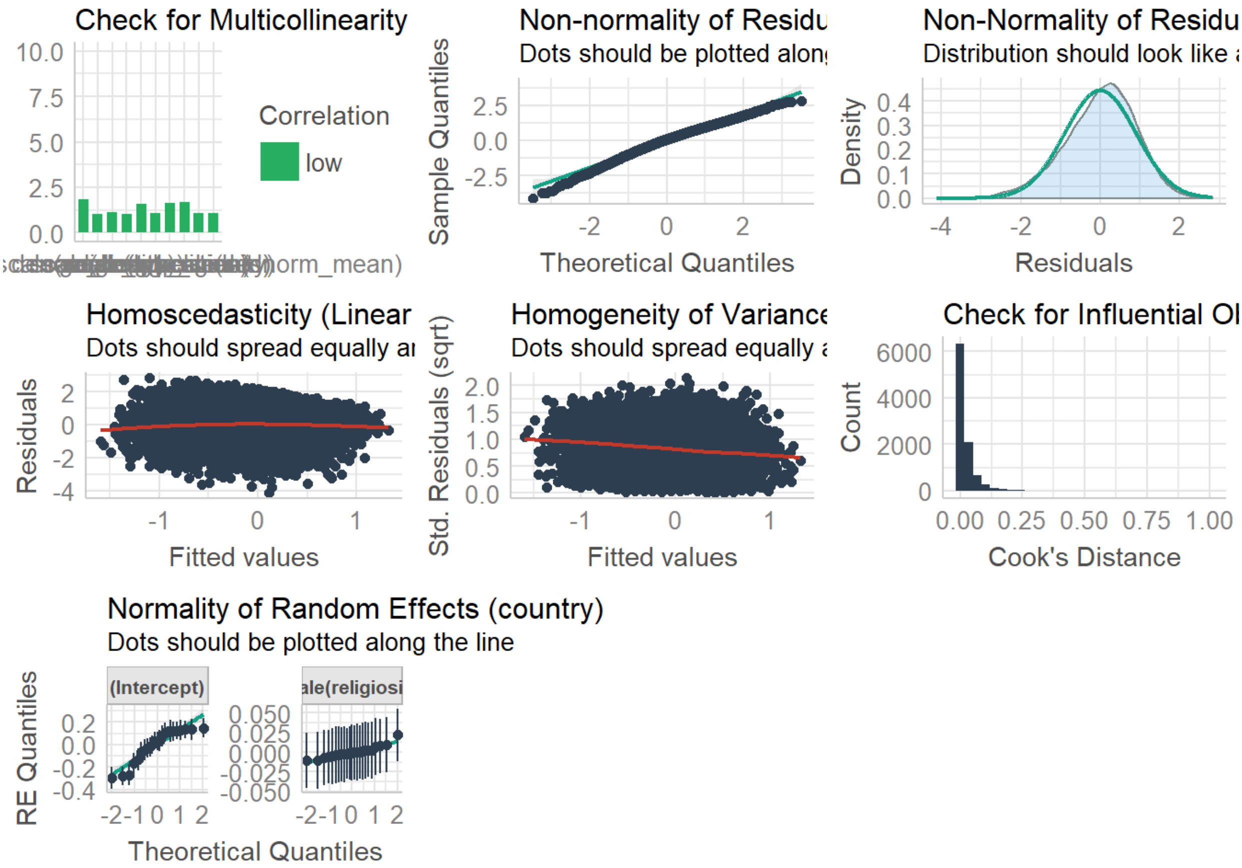
```
h2 <-
  lmer(scale(wb_overall_mean) ~ scale(religiosity) * scale(cnorm_mean) +
         # personal level confounders
                         scale(age) + gender + scale(ses) + scale(education) + deno
m_lump +
         # country level confounders
                         scale(gdp_scaled) + sample_type +
         # random intercept and slope model
                         (scale(religiosity)|country),
       data = marp)

check_model(h2)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 9747 rows containing missing values (geom_text_repel).
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# summary(h2)
```

Model diagnostics show heteroscedasticity, therefore cluster robust standard errors are calculated.

```
tab_model(h2,
          show.aic = TRUE,
          show.reflvl = TRUE,
          string.ci = "95% CI",
          # file = "docs/h2_model.html",
          vcov.fun = "CR",
          vcov.type = "CR2",
          vcov.args = list(cluster = h1@frame$country)
          )
```

	scale(wb_overall_mean)		
Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	-0.00	-0.13 – 0.12	0.968
Female	*Reference*		
Male	0.07	0.02 – 0.13	**0.012**
Other	-0.49	-0.75 – -0.23	**<0.001**
No denomination	*Reference*		

Buddhist	0.04	-0.08 – 0.16	0.465
Christian	-0.03	-0.12 – 0.06	0.483
Hindu	-0.15	-0.29 – -0.01	**0.030**
Jewish	-0.13	-0.26 – -0.01	**0.039**
Muslim	-0.18	-0.29 – -0.06	**0.002**
Other	-0.16	-0.32 – 0.01	0.063
general public	*Reference*		
mixed	0.03	-0.14 – 0.20	0.713
online panel	-0.13	-0.18 – -0.07	**<0.001**
students	0.13	-0.26 – 0.52	0.509
scale(age)	0.04	-0.00 – 0.08	0.067
scale(cnorm_mean)	-0.05	-0.12 – 0.02	0.138
scale(education)	0.08	0.05 – 0.10	**<0.001**
scale(gdp_scaled)	0.03	-0.06 – 0.11	0.526
scale(religiosity)	0.13	0.10 – 0.16	**<0.001**
scale(religiosity):scale(cnorm_mean)	0.05	0.04 – 0.07	**<0.001**
scale(ses)	0.35	0.30 – 0.40	**<0.001**
Random Effects			
σ^2	0.81		
$\tau_{00\ country}$	0.03		
$\tau_{11\ country.scale(religiosity)}$	0.00		
$\rho_{01\ country}$	0.09		
ICC	0.03		
$N_{country}$	23		
Observations	9747		
Marginal R^2 / Conditional R^2	0.173 / 0.200		
AIC	25775.529		

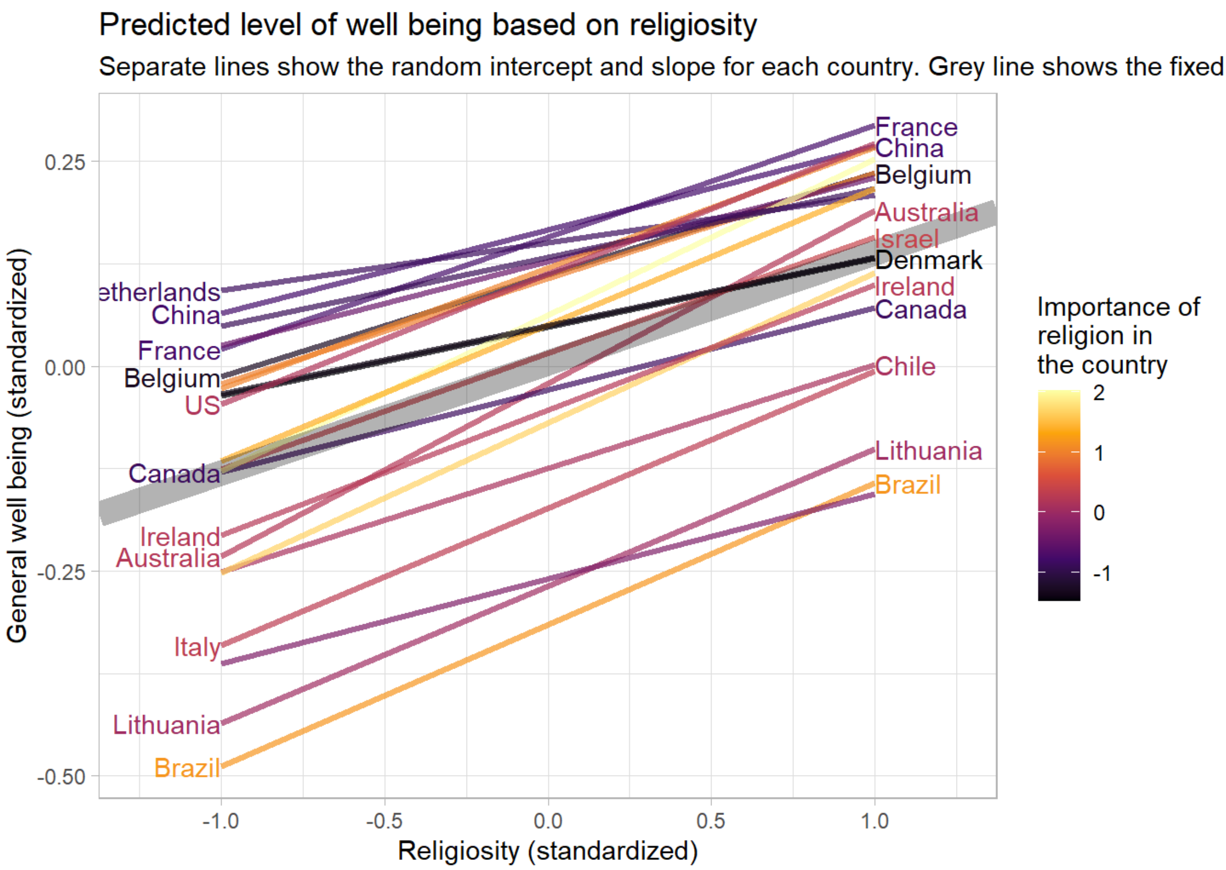
The relationship between religiosity and well-being is moderated by country norms about religion. In countries where religion is more important , religion has a stronger association with well-being.

Plots

Show random intercept and slope by country norms | On different facets

Show only the slopes

```
h1_lines %>%
        left_join(country_norms, by = "country") %>%
        ggplot() +
        aes(x = x, xend = xend, y = y, yend = yend, color = cnorm_mean) +
        geom_segment(size = 1.2, alpha = .7) +
        geom_abline(aes(intercept = fix_int,
                        slope = fix_slo),
                    color = "black", size = 5, alpha = .3) +
        scale_color_viridis_c(option = "inferno") +
        geom_text(aes(label = country),
                  show.legend = FALSE, hjust = 1, check_overlap = TRUE) +
        geom_text(aes(x = xend, y = yend, label = country),
                  show.legend = FALSE, hjust = 0, check_overlap = TRUE) +
        xlim(-1.25,1.25) +
        labs(title = "Predicted level of well being based on religiosity",
             subtitle = "Separate lines show the random intercept and slope for eac
h country. Grey line shows the fixed effect",
             y = "General well being (standardized)",
             x = "Religiosity (standardized)",
             color = "Importance of\nreligion in\nthe country")
```



Model comparisons and Bayes Factors

```
anova(h0, h1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: marp
## Models:
## h0: scale(wb_overall_mean) ~ scale(age) + gender + scale(ses) + scale(education)
+
## h0:     denom_lump + scale(gdp_scaled) + sample_type + (scale(religiosity) |
## h0:     country)
## h1: scale(wb_overall_mean) ~ scale(religiosity) + scale(age) + gender +
## h1:     scale(ses) + scale(education) + denom_lump + scale(gdp_scaled) +
## h1:     sample_type + (scale(religiosity) | country)
##    npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## h0   20 25730 25873 -12845    25690
## h1   21 25699 25850 -12829    25657 32.335  1  1.297e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(h1, h2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: marp
## Models:
## h1: scale(wb_overall_mean) ~ scale(religiosity) + scale(age) + gender +
## h1:     scale(ses) + scale(education) + denom_lump + scale(gdp_scaled) +
## h1:     sample_type + (scale(religiosity) | country)
## h2: scale(wb_overall_mean) ~ scale(religiosity) * scale(cnorm_mean) +
## h2:     scale(age) + gender + scale(ses) + scale(education) + denom_lump +
## h2:     scale(gdp_scaled) + sample_type + (scale(religiosity) | country)
##    npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## h1   21 25699 25850 -12829    25657
## h2   23 25683 25849 -12819    25637 19.971  2  4.606e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(h0, h2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: marp
## Models:
## h0: scale(wb_overall_mean) ~ scale(age) + gender + scale(ses) + scale(education)
+
## h0:     denom_lump + scale(gdp_scaled) + sample_type + (scale(religiosity) |
## h0:     country)
## h2: scale(wb_overall_mean) ~ scale(religiosity) * scale(cnorm_mean) +
## h2:     scale(age) + gender + scale(ses) + scale(education) + denom_lump +
## h2:     scale(gdp_scaled) + sample_type + (scale(religiosity) | country)
##    npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## h0   20 25730 25873 -12845    25690
## h2   23 25683 25849 -12819    25637 52.306  3  2.577e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Calculate BIC based Bayes factors for
# H0 vs H1
exp((BIC(h0) - BIC(h1))/2)
```

```
## [1] 3640.829
```

```
# H1 vs H2
exp((BIC(h1) - BIC(h2))/2)
```

```
## [1] 0.003643478
```

```
# H0 vs H2
exp((BIC(h0) - BIC(h2))/2)
```

```
## [1] 13.26528
```

```
# Get std. beta and conf ints for both models
h1_coef <-
  tidy(h1, conf.int = TRUE) %>%
  filter(term == "scale(religiosity)") %>%
  select(estimate, conf.low, conf.high) %>%
  mutate(across(everything(), round, 2)) %>%
  summarise(str_glue("std. beta = {.$estimate} 95% CI[{.$conf.low}, {.$conf.high}]"
)) %>%
  pull()

h2_coef <-
  tidy(h2, conf.int = TRUE) %>%
  filter(term == "scale(religiosity):scale(cnorm_mean)") %>%
  select(estimate, conf.low, conf.high) %>%
  mutate(across(everything(), round, 2)) %>%
  summarise(str_glue("std. beta = {.$estimate} 95% CI[{.$conf.low}, {.$conf.high}]"
)) %>%
  pull()
```

Conclusion

The BF for the first research question indicates that the data are 3640.83 more likely under the alternative hypothesis than the null. Therefore, religiosity seems to have an effect of std. beta = 0.13 95% CI[0.1, 0.17] on general well being while controlling for gender, age, denomination, ses, education, sample type, and country gdp.

The BF for the second research question indicates that the data are 13.27 more likely under the alternative hypothesis than the null. Therefore, country norms about religiosity seem to moderate (std. beta = 0.05 95% CI[0.03, 0.08]) the effect of religiosity on general well being while controlling for gender, age, denomination, ses, education, sample type, and country gdp.

Based on the collected evidence, the answer to both research questions is ‘yes’.