MA50258 Applied Statistics: Coursework 2

Zoë Ganess

1 The Effect of Socioeconomic Factors on Infant Mortality Rate: Accounting for Differences between Countries, Continents, Regions and Time

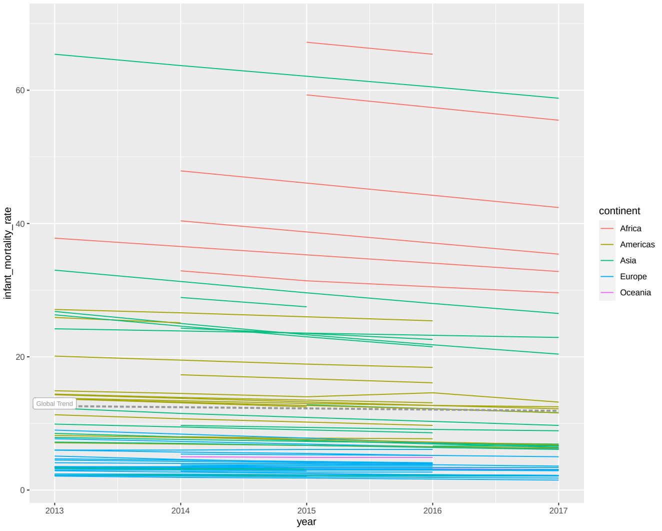
The Sustainable development goals (SDG) data presents itself as a multilevel data problem. Information is reported longitudinally over time, from 2013 to 2017, for each country. Not only does this imply that time itself influences changes any outcome variable of interest, which in this case is infant mortality rate, but countries also have internal trends and characteristics which are influential. There is further complexity considering that countries are nested in continents and regions, such that regional and continental trends may have some additional influence; time-based trends can be different across countries and continents also.

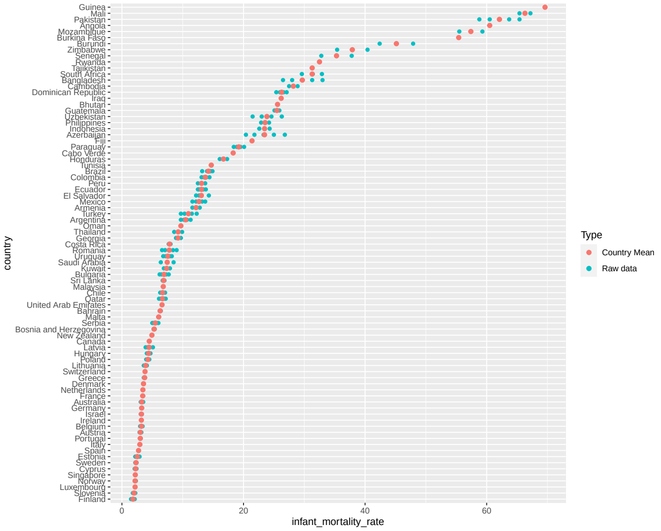
Considering this, in order to determine the effect of socioeconomic factors on infant mortality rate, we firstly simplify things and only look at the the relationship between infant mortality rate and time and how that relationship differs across continent and country. Then, by implementing a mixed effect model, we are able to quantify the variability in infant mortality rate between countries (via the random effects) as well as investigate the relationship between socioeconomic factors and infant mortality rate (via the fixed effects).

NB: where (*) appears, that plot/table/graph etc. was not included due to the page limit but is not central to the main analysis…refer to rmarkdown file for original output.

After adjusting the dataframe so as to omit any missing values and with re-encoded variables (note that year was centered around 2015), in order to visualise the data, we first plot a histogram of the frequency of the respective infant mortality rates, which demonstrates a strong positive skew as seen below (*). Hence, we should anticipate having to either fit a Gamma generalised linear model or a transformed linear model in order to best model the data.

Before implementing any models, it seemed sensible to carry out some basic exploratory analysis. In particular, we first want to determine how infant mortality rate changes over time for each country, at the most baseline level.





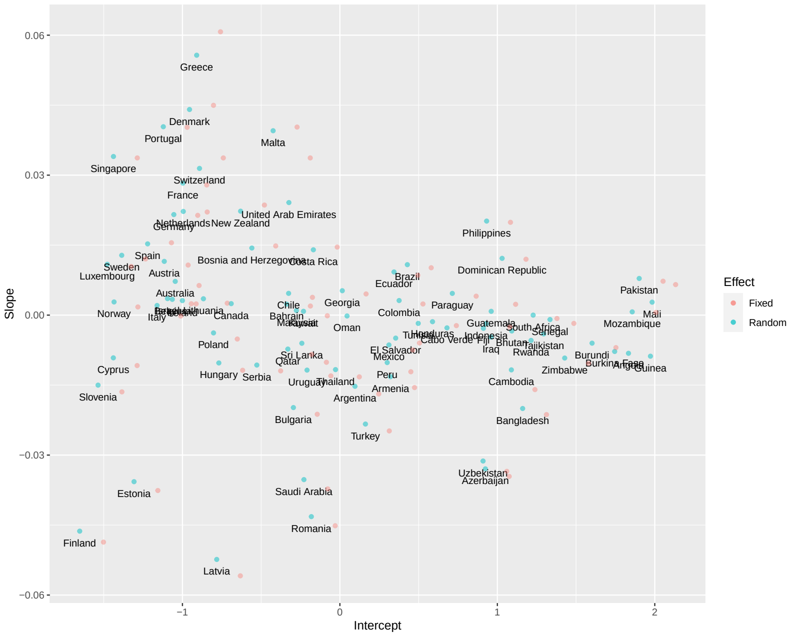
From the plot above on the left, it is clear that there is a general decreasing trend in infant mortality rate as time goes on. Likewise, we can infer that countries in Africa tend to have higher infant mortality rates (with an overall average of 46.3) when compared to countries in Europe (with an overall average of 3.8). Further supporting this notion, from the plot on the right we can see that Guinea has the highest average infant mortality rate of 69.6 whereas Finland has the lowest of 1.8 (specific means derived using basic linear model (*) with country as explanatory variable).

By fitting a model with year as the only fixed effect, we can see that when the year is zero, or in 2013, the average infant mortality rate is 12.25, and it decreases by 0.18 units each year after that. However, under this model, the underlying structure of the data is ignored; none of the country-specific or region-specific trends that might also influence life expectancy are accounted for. This is clear from the prediction plot below (*) which depicts the same overall trend line, rather than one for each specific country. Note that this line corresponds to the global trend line indicated in the first plot.

The model building process began from this initial point. From the plots, it is clear that each country has a different intercept, but it is less clear whether the slopes differ, hence making it sensible to explore the difference between a random intercept model and a random intercept + slope model. Before this, however, I ensured that incorporation of `country` as a random effect was warranted. (*) In particular, under the initial random intercept model (log transformed), there was a country-to-country variability of 1.014278; since this value is more than zero, this inclusion is supported. Interestingly, under the corresponding null model (log transformed), the fixed effects are estimated at 2.01943330 versus 2.22713645 under the initial random intercept model (log transformed), hence implying that the data is imbalanced; in other words, some countries have more data points than others.

```
## Data: sdg
## Models:
## mod_int_ML: log(infant_mortality_rate) ~ 1 + (1 | country) + yearc
## mod_int_slope_ML: log(infant_mortality_rate) ~ yearc + (1 + yearc | country)
##                  npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## mod_int_ML          4 -358.99 -344.76 183.49  -366.99
## mod_int_slope_ML    6 -583.40 -562.06 297.70  -595.40 228.42  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above output, we can see that the p-value corresponding to random intercept + slope model is significantly less than 0.05, hence inferring that this model improves upon the initial random intercept model. (*) This is also supported by the prediction plot below which visually shows an improved fit to the data.



From the plot of the estimated fixed and random effects, there is some inconsistency with shrinkage as some random effect points appear to be closer to (0,0) whereas others do not. Hypothesising why this may be, one plausible reason is that the current random effects are not distributed normally with mean 0 and a constant variance, this of which is an assumption for random effect variables. From the plot, we can also infer that countries to the top of the plot, such as Greece and Denmark, have the quickest growing infant mortality rate whereas those to the bottom, such as Latvia and Finland, have the quickest decreasing infant mortality rate.

Considering that variability not only exists from country to country, but the fact that there is also the possibility of regional and continental differences, I then decided to explore the inclusion of `region` and `continent` as random effects.

```
## Data: sdg
## Models:
## mod_int_slope_ML: log(infant_mortality_rate) ~ yearc + (1 + yearc | country)
## mod_cont_ML: log(infant_mortality_rate) ~ yearc + (1 + yearc | country) + (1 | region)
## mod_cont_reg_ML: log(infant_mortality_rate) ~ yearc + (1 + yearc | country) + (1 | region) + (1 | continent)
##                  npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## mod_int_slope_ML    6 -583.40 -562.06 297.70  -595.40
## mod_cont_ML         7 -640.42 -615.52 327.21  -654.42 59.013  1  1.567e-14 ***
## mod_cont_reg_ML     8 -648.96 -620.50 332.48  -664.96 10.539  1   0.001169 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above output, the addition of the two random effects prove to be beneficial as the corresponding p-value is less than 0.05.

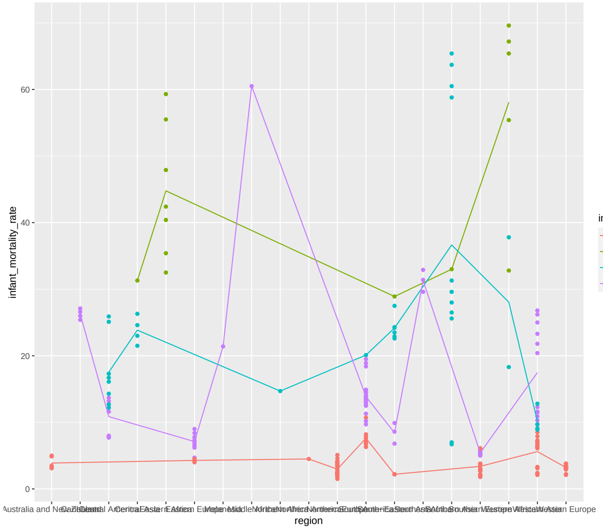
Now, using LOESS smoothing plots as seen below (*), I was able to determine the type of relationship the respective explanatory variables had with infant mortality rate. In particular, log transforming `infant_mortality_rate` proved to improve the fit of the data (hence why this transformation was used in the models thus far); log transforming `gdp` and cubing `life_expectancy` allowed for improved fit whereas `years_education` was retained as a linear term.

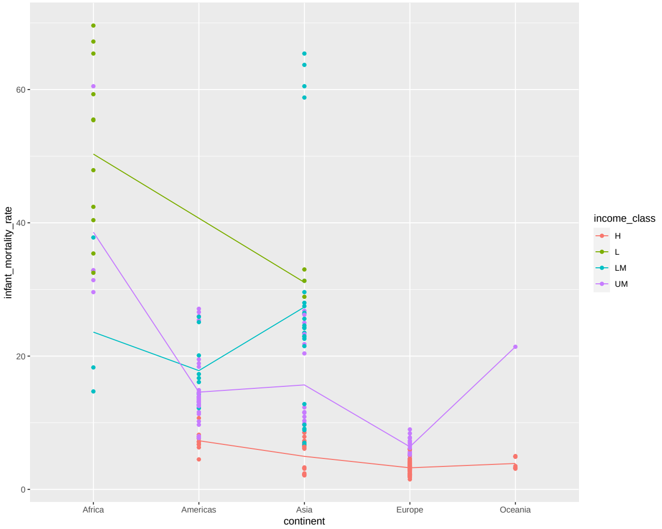
From the above output (*), we can see that removing `years_education` and `log(gdp)` would result in some model improvement. We test this by removing the variables and using ANOVA to compare the models as seen below:

```
## Data: sdg
## Models:
## mod_rm_ML: log(infant_mortality_rate) ~ 1 + poly(life_expectancy, 3) + (1 + yearc | country) + yearc + (1 | cont
## mod_build_ML: log(infant_mortality_rate) ~ 1 + poly(life_expectancy, 3) + (1 + yearc | country) + yearc + log(gd
##               npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## mod_rm_ML       11 -702.10 -662.98 362.05  -724.10
## mod_build_ML    13 -700.95 -654.71 363.48  -726.95 2.8482  2     0.2407
```

We can now confirm that the two variables did not in fact add anything to our model, and so we retain the reduced model going forward.

When considering whether to include interaction terms, it is fairly reasonable to assume that there might be significant differences in infant mortality rate due to income class differences between regions.





The plots above contain substantial information about the data. Firstly, lower class regions tend to have a higher average infant mortality rate, followed by lower middle class regions then upper middle class regions, while upper class regions have the lower rates. Since regions are nested within continents, we can easier visualise the relationship from the second plot. Notably, Africa does not have any high income class regions, Oceania and Europe only have high and upper middle class income regions and the Americas do not have any lower class income regions; Asia is the only continent with regions of all income classes. We also note the higher proportion of high and upper middle income class regions when compared lower middle and low income class regions. Lastly, notice how many regions only consist of one income class rather than a mix.

```
## Data: sdg
## Models:
## mod_re_ML: log(infant_mortality_rate) ~ poly(life_expectancy, 3) + (1 + yearc | country) + yearc + income_class
## mod_inter_ML: log(infant_mortality_rate) ~ income_class + (1 + yearc | country) + yearc + poly(life_expectancy,
##               npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## mod_re_ML       14 -697.42 -647.62 362.71  -725.42
## mod_inter_ML    14 -699.00 -649.21 363.50  -727.00 1.5866  0
```

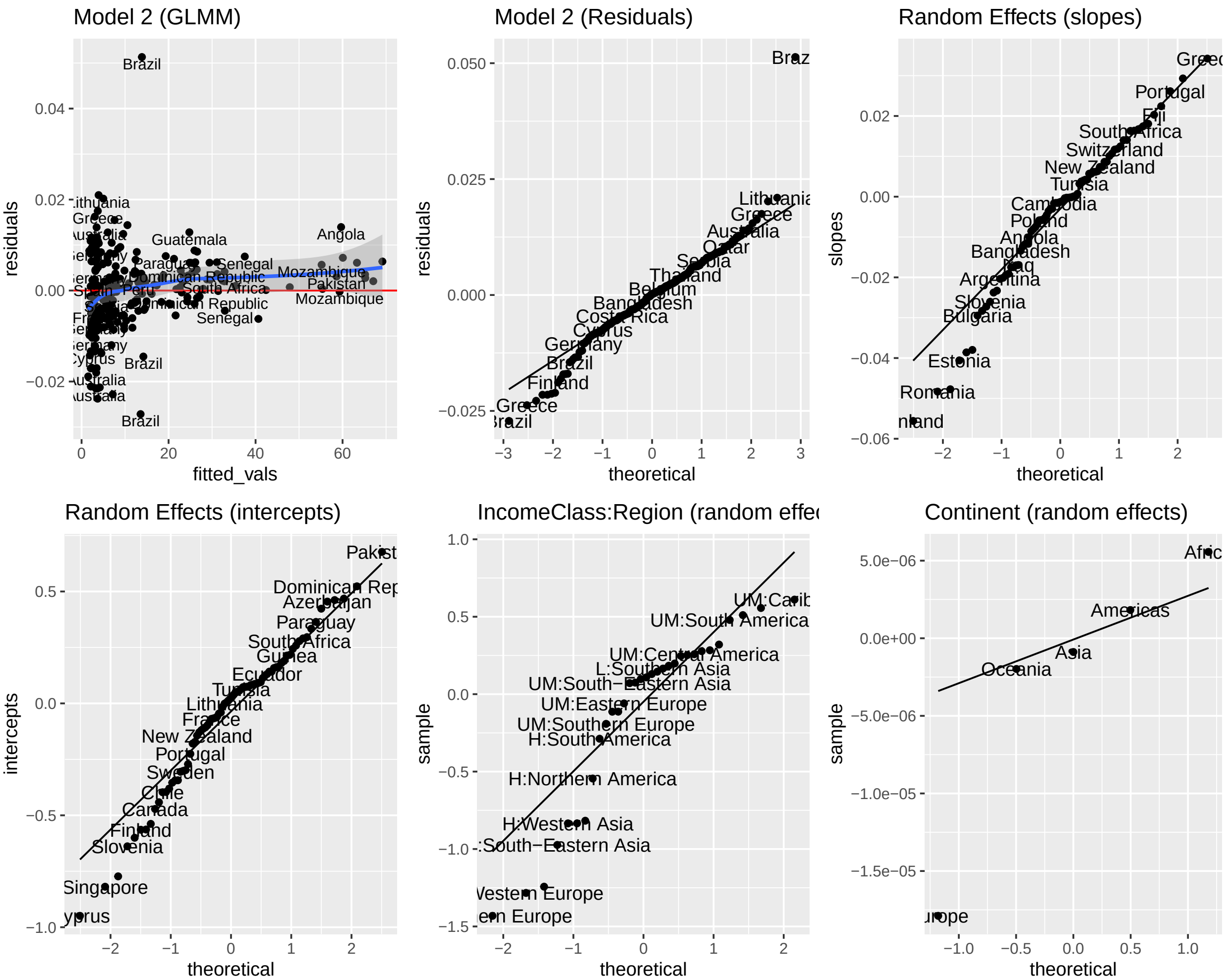
From the above output, we can now retain the composite random effects term as we reject the null hypothesis of it being insignificant in our model at the 95% confidence interval.

From the diagnostic plots above (*), we can see that the normality assumption is in doubt due to the deviations in the qqplots. Notably, most of this is due to data corresponding to Europe and America as inferred from the random effects plot for Continent.

Recalling the strong positive skew in the data, I then implemented a Gamma log-linked GLMM in an attempt to improve the model even further. I created two models and used the same variables as the previous model, the only difference being that for one model I also tried to allow income class random effects by region via `(income_class|region)`.

```
## Data: sdg
## Models:
## mod_glm_inter: (infant_mortality_rate) ~ income_class + (1 + yearc | country) + yearc + poly(life_expectancy, 3)
## mod_glm_reg: (infant_mortality_rate) ~ poly(life_expectancy, 3) + (1 + yearc | country) + yearc + (income_class
##               npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## mod_glm_inter   14 -100.46 -50.663 64.229  -128.46
## mod_glm_reg     23  -85.10  -3.293 65.550  -131.10 2.6413  9     0.9768
```

From the ANOVA output, we fail to reject the null hypothesis and so, we retain the GLMM with the original composite random effects term.

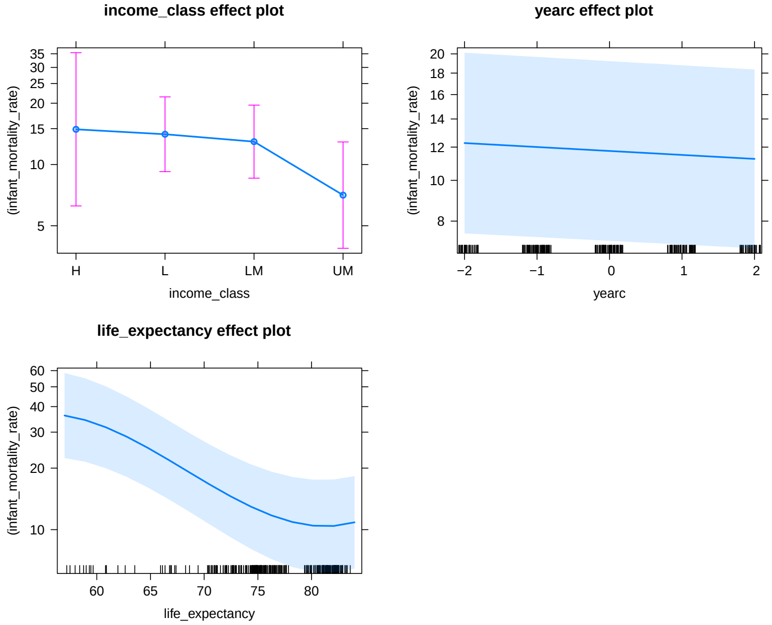


From the diagnostic plots above, there seems to be improvement in terms of the general qqplot such that there are less extreme deviations. However, when examining the qqplot for `incomeclass:region`, we now see a new pattern that appears almost polynomial in nature. The qqplot for continent still shows extreme deviation for Europe. Because of the few amount of points corresponding to the higher fitted values (ie. due to the strong skew), it is a bit difficult to assess the residual plot, however, it does not seem to be of immediate concern if we ignore these extreme values.

```
## [1] 475.5545
## [1] 369.502
```

```
## [1] -100.4586
```

The AIC of the log-linked Gamma GLMM is -100.4586 whereas the AIC for the log transformed model is 369.502. Using this, along with the discussed improvements in the diagnostic plots, it seemed sensible to choose the GLMM for use in interpretation.

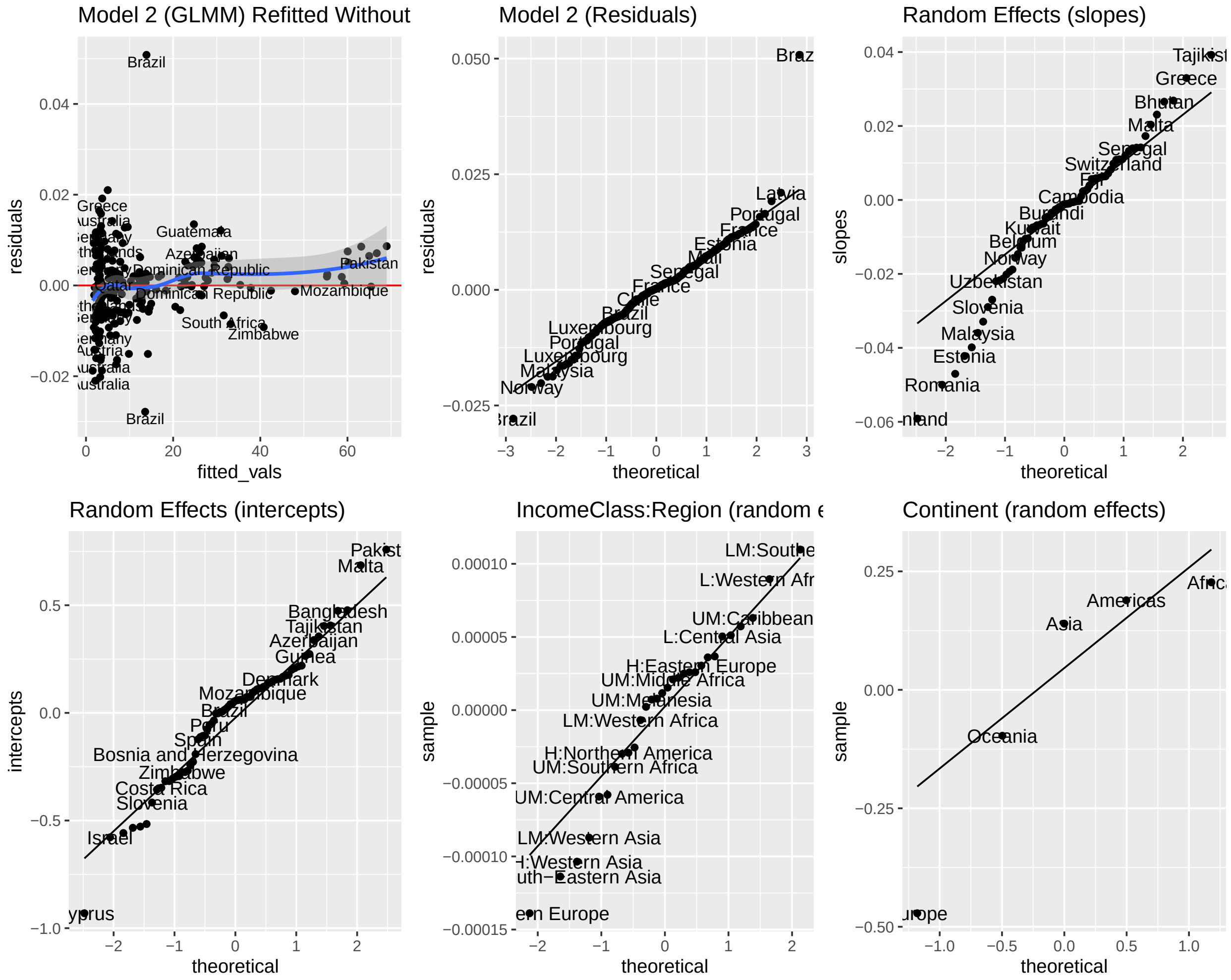


When examining the effect plots, we can see that time and life expectancy both have a negative relationship with infant mortality rate. On the other hand, the predicted relationship between income class and infant mortality rate is particularly interesting as it implies that higher income class regions are estimated to have higher infant mortality rates when compared to regions with the other income classes. This goes against the regular intuition that higher income classes would have lower infant mortality rates. One possible reason why this relationship may have been predicted in this way is due to confounding. More specifically, the data lacks information regarding population size. Presumably, regions with increased population size would have more births and hence, an increased infant mortality rate (due to increased probability of occurrence). Likewise, as we saw before, some regions only have one income class and the proportion of high income class was higher than the other income classes. Hence, it is possible that these regions with high income class are greater in population size, hence resulting in increased infant mortality rate.

Considering this discrepancy in the predictions, as well as the possibility of outliers in the data, I decided to conduct a test for outliers by deriving the Cook's distance. This was done using `HLMdiag` library and hlm_influence() function which takes `lmer` objects to derive influence diagnostics.

```
## # A tibble: 7 x 5
##   country   cooksd mdffits covtrace covratio
##   <fct>      <dbl>   <dbl>    <dbl>    <dbl>
## 1 Argentina 0.168   0.148    0.159      1.16
## 2 Paraguay  0.0932  0.0747   0.275      1.28
## 3 Singapore 0.0918  0.0838   0.136      1.14
## 4 Lithuania 0.0850  0.0773   0.122      1.12
## 5 Hungary   0.0728  0.0669   0.175      1.18
## 6 Bulgaria  0.0520  0.0487   0.0910     1.09
## 7 Sri Lanka 0.0368  0.0352   0.0556     1.06
```

From the plot above (*), we can see that 6 countries exceed the threshold value as indicated by the red line. Using the corresponding dataframe containing the influence data, I arranged the Cook's distance values so as to show the countries with the highest values first as these would correspond to the outliers. Argentina, Paraguay, Singapore, Lithuania, Hungary and Bulgaria all have Cook's distance values of more than 0.05 and so, we remove them from the original dataframe and refit the GLMM model.



From diagnostic plots for the refitted model, we can see that the overall qqplot shows improvement when compared to the previous GLMM as deviations only really exist at the tails. The polynomial-like pattern previously observed in the `incomeclass:region` qqplot is less prominent and Europe is no longer as extremely deviated in terms of the continent qqplot.

```
## [1] -118.0778
```

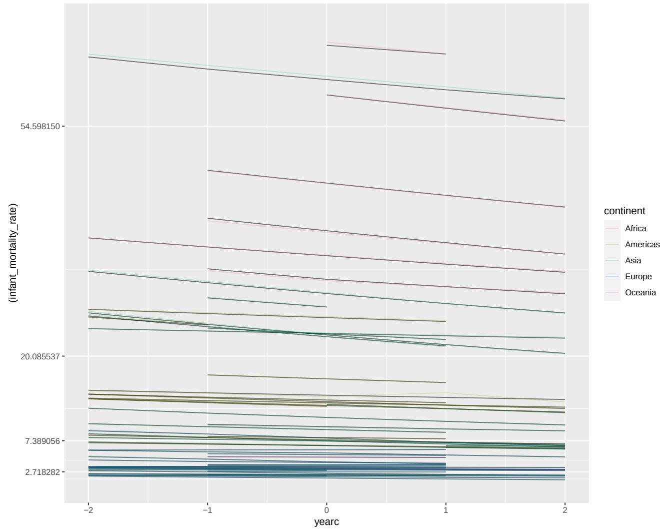
With an AIC of -118.0778, we can now conclude that the refitted GLMM improves upon the previously discussed models. This is further confirmed by overlaying the actual data with the fitted model predictions as seen below:

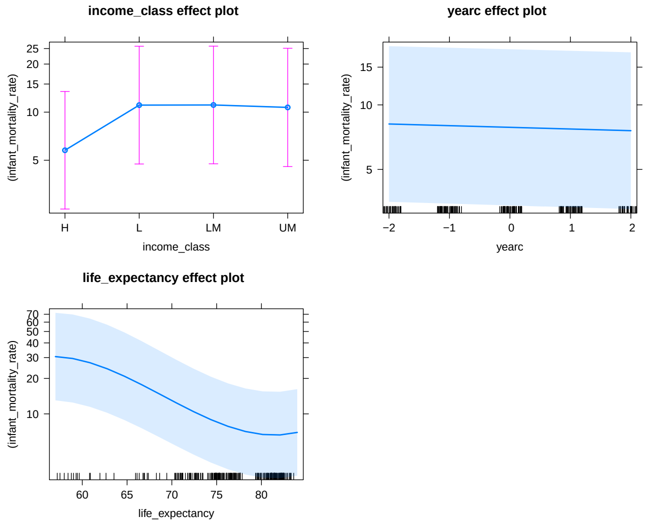
```
##                          Estimate  Std. Error   t value     Pr(>|z|)
## (Intercept)            1.85313123 0.429335513  4.316278 1.586824e-05
## income_classL          0.65082480 0.123948847  5.250753 1.514786e-07
## income_classLM         0.65281099 0.123202782  5.298671 1.166487e-07
## income_classUM         0.61776070 0.127629826  4.840253 1.296737e-06
## yearc                 -0.01795766 0.003632736 -4.943288 7.681604e-07
## poly(life_expectancy, 3)1 -5.64786500 0.738537171 -7.647367 2.051358e-14
## poly(life_expectancy, 3)2  1.11590229 0.197370444  5.653847 1.568959e-08
## poly(life_expectancy, 3)3  0.80786598 0.146849945  5.501303 3.769956e-08
```

This model also gives us more sensible estimates:

- Average infant mortality rates were approximately 100 x (exp(-0.01795766)-1) = 1.8% less for each year passed (between 2013 and 2017)
- Around log(2)/-0.01795766 = 39 years ago, average infant mortality rates would have been doubled
- For two regions within the same year and with the same life expectancy,if one of those regions is made up of persons with high class income, then we expect to see an average increase of exp(0.65082480)-1 = 0.917 infant mortality rate units of the other region if it is made up of persons with low class income; if the other region is made up of persons with lower middle class income then we expect to see an average increase of exp(0.65281099)-1 = 0.921 infant mortality rate units and if it is a upper middle class income region then we expect to see an average increase of exp(0.61776070)-1 = 0.855 infant mortality rate units.

This can be easier visualised by the plots below:





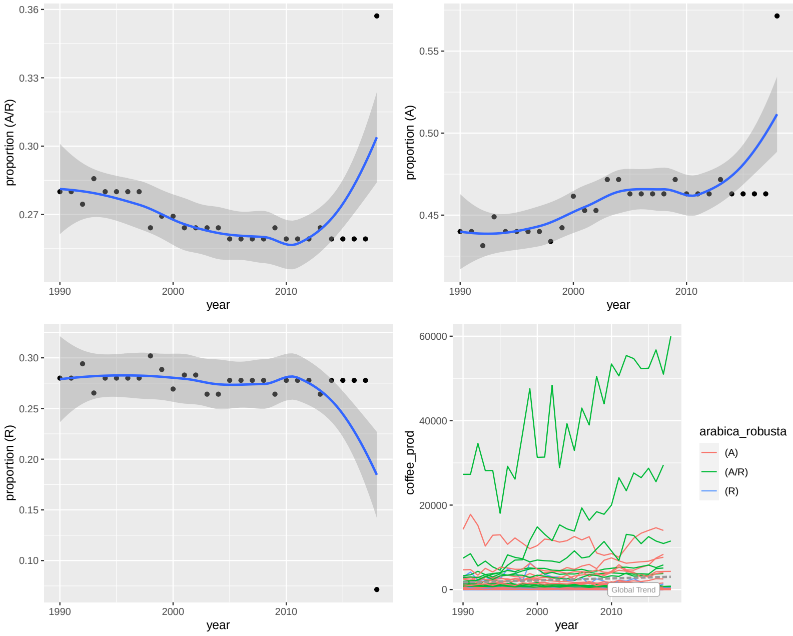
2 An Analysis of Coffee Production Over Time

In this dataset, there are two sources of variability which may result in differences in coffee production: variability due to the type of coffee and variability due to country. The data is also longitudinal as the entries are recorded over time from 1990 to 2018. Here, we aim to quantify the trend of coffee production over time, by also explaining differences between countries and coffee type.

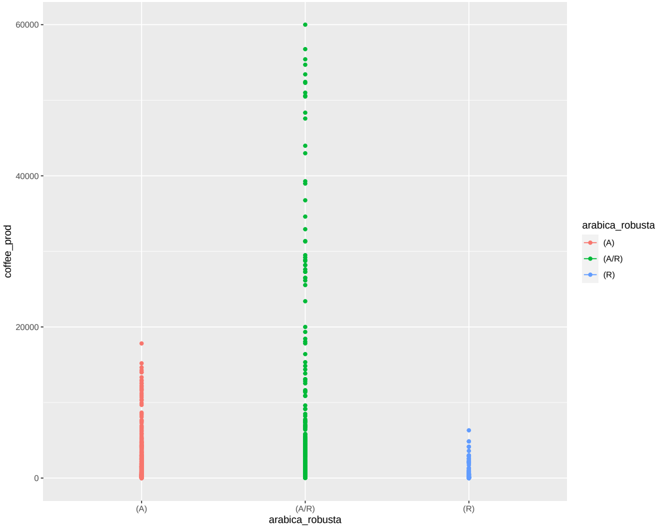
The initial dataframe was melted to make subsequent analysis easier; missing values were omitted and the `arabica_robusta` variable was adjusted so that there were three combinations, (A/R), (A) and (R), corresponding to the two types of coffee. This was done because there were entries initially corresponding to '(A/R)' and '(R/A)' rather than just one factor, this of which would make more sense since they represent the same combination.

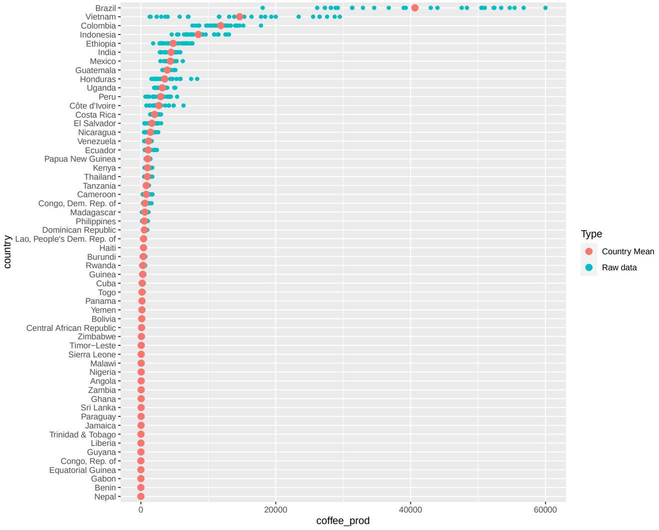
Plotting a barplot of the frequency of the respective coffee production values, there is a strong positive skew of 5.547558 as seen below (*). Hence, we should anticipate having to either fit a Gamma generalised linear model or a transformed linear model in order to best model the data.

By creating binary factors corresponding to whether a country used arabica, robusta or a combination of both, I then plotted graphs to visualise the relative usage (proportion) of these coffee types over time.



From the plot above, we can see that combination of arabica and robusta usage generally decreased from 1990 till around 2012, after which there is some increase. Arabica usage generally increases throughout the entire recorded period whereas robusta remains fairly constant despite the sudden decrease in 2018. Considering the fact that arabica and combination usage both see a drastic increase in 2018, this may account for this sudden decrease for robusta usage. When considering the relative proportion, arabica is used the most whereas robusta is used the least. It is hard to deduce the overall trend from the bottom right plot, however at a higher level, there appears to be an increasing trend for the visible lines; notably, these lines also correspond to combination usage.



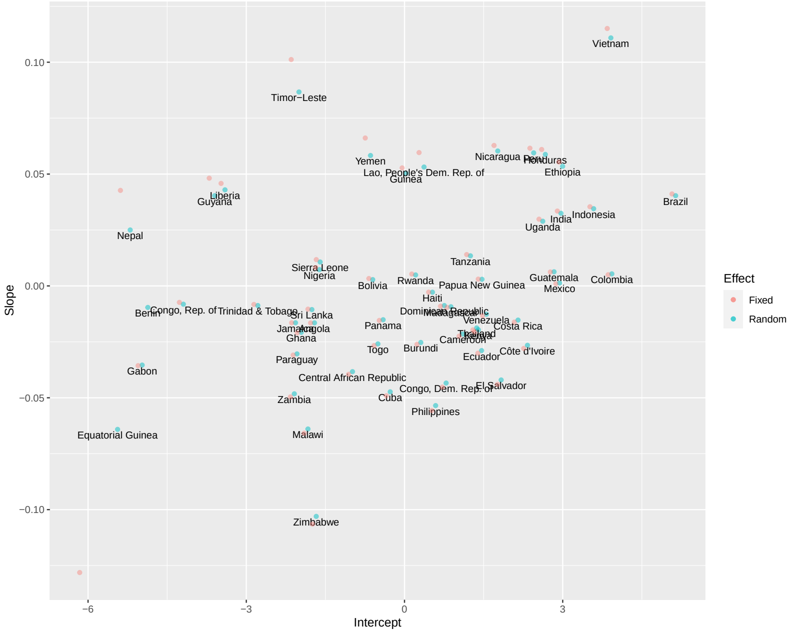


Further expanding on this, from the plot on the left we can see that combination usage resulted in more coffee production with a total annual average of 5851 thousand 60kg bags, followed by arabica only usage (1540 thousand 60kg bags) and lastly robusta only usage (295 thousand 60kg bags) for the 28 year period. From the plot on the right, Brazil has the highest total annual average of 40636 thousand 60kg bags whereas Nepal has the lowest with only 1.5 thousand 60kg bags, for the same period. Notably, there is a substantial amount of variability in coffee production within Brazil itself.

Considering this variability between countries, it seemed sensible to include `country` as a random effect. This was supported as under the initial random intercept model (log transformed), there was a country-to-country variability of 2.48281; since this value is more than zero, this inclusion is supported. Under the corresponding null model (log transformed), the fixed effects are estimated at 5.70480835 versus 5.45965491 under the initial random intercept model (log transformed), hence implying that the data is imbalanced; in other words, some countries have more data points than others.

```
## Data: coffee
## Models:
## mod_int_ML: log(coffee_prod) ~ 1 + (1 | country) + yearc
## mod_int_slope_ML: log(coffee_prod) ~ yearc + (1 + yearc | country)
##                  npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## mod_int_ML          4 2419.8 2441.0 -1205.89   2411.8
## mod_int_slope_ML    6 1735.7 1767.5  -861.85   1723.7 688.08  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above output, we can see that the p-value corresponding to random intercept + slope model is significantly less than 0.05, hence inferring that this model improves upon the initial random intercept model. This is also supported by the prediction plot below which visually shows an improved fit to the data (*).

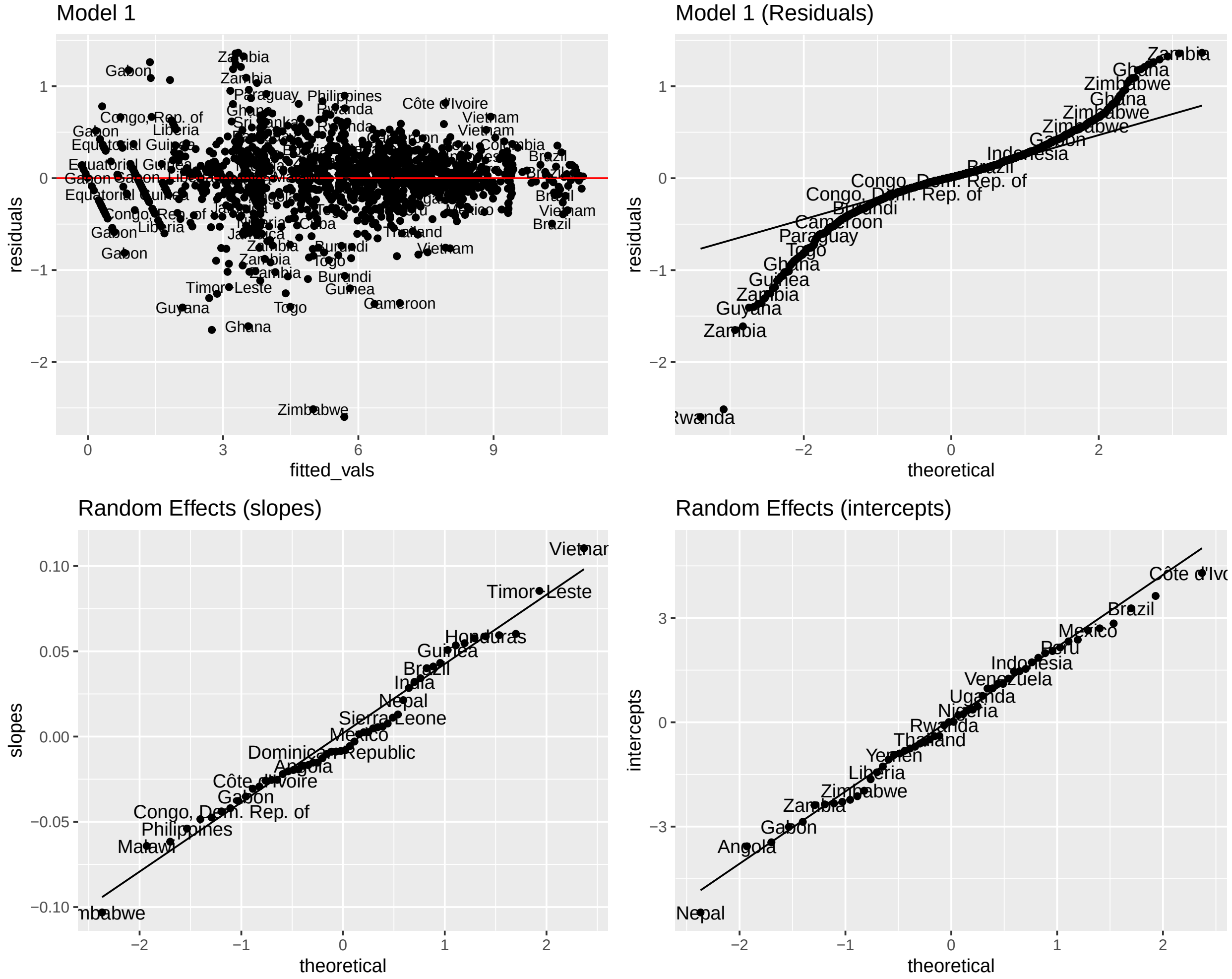


With this data there also appears to be some inconsistency with shrinkage as some random effect points appear to be closer to (0,0) whereas others do not. It is probable that this can also be explained using the same reasoning as hypothesised with the SDG dataset. From the plot, we can also infer that countries to the top of the plot, such as Timor-Leste and Vietnam, have the quickest growing total annual coffee production whereas those to the bottom, such as Zimbabwe and Malawi, have the quickest decreasing total annual coffee production.

Considering, the substantial differences between coffee type in terms of coffee production/yield, this of which may be systematic, we now incorporate the `arabica_robusta` term into our model. In particular, I tried this term as a fixed effect alone, as an interaction term with `yearc` and lastly as a composite random effects term with `country`.

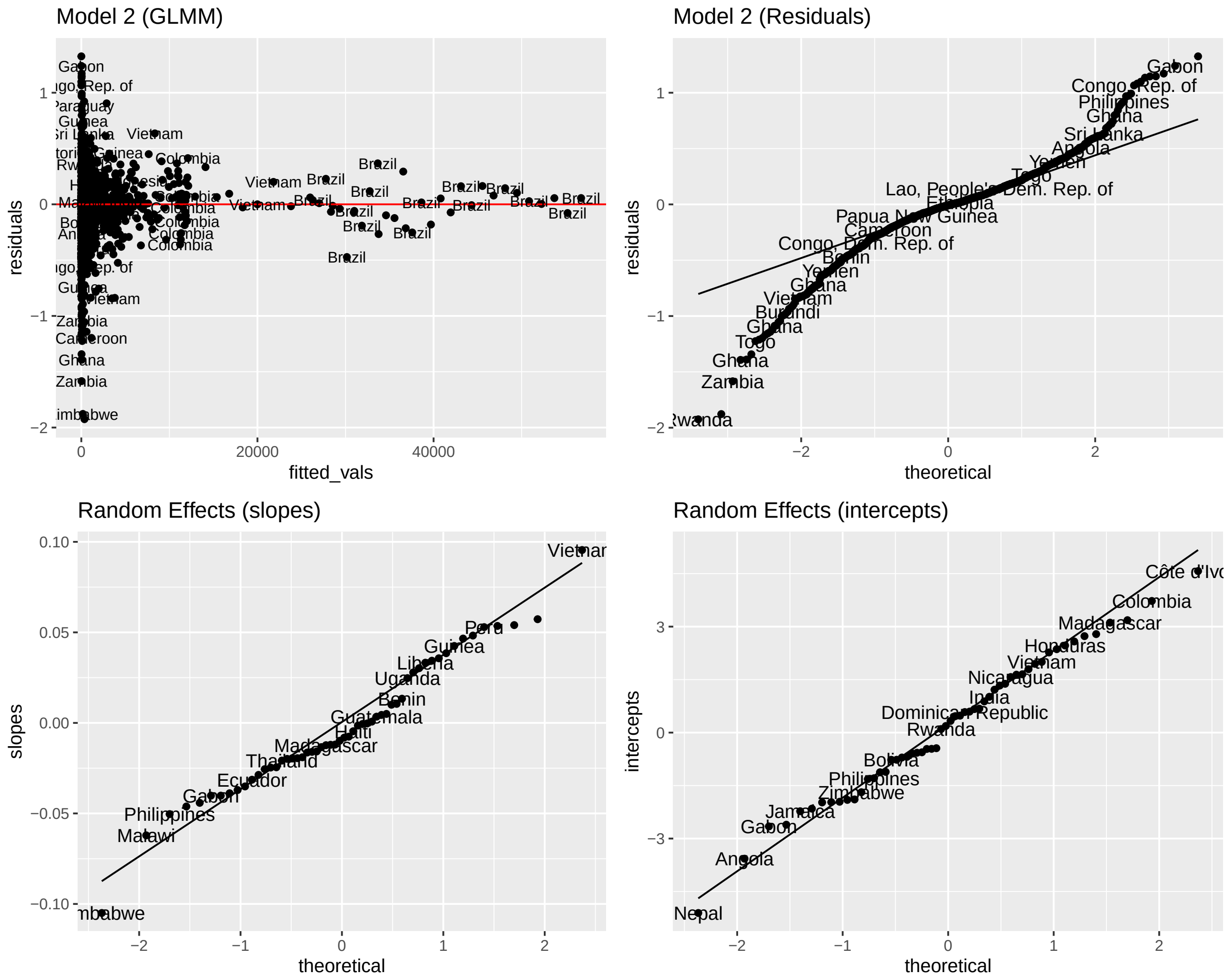
```
## Data: coffee
## Models:
## mod_int_slope_ML: log(coffee_prod) ~ yearc + (1 + yearc | country)
## mod_type_ML: log(coffee_prod) ~ yearc + arabica_robusta + (1 + yearc | country)
## mod_re_inter_ML: log(coffee_prod) ~ yearc + arabica_robusta + (1 + yearc | country) + (1 | arabica_robusta:count
## mod_type_inter_ML: log(coffee_prod) ~ yearc * arabica_robusta + (1 + yearc | country)
##                   npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
## mod_int_slope_ML     6 1735.7 1767.5 -861.85   1723.7
## mod_type_ML          8 1714.6 1757.0 -849.28   1698.6 25.1366  2  3.481e-06 ***
## mod_re_inter_ML      9 1716.6 1764.3 -849.28   1698.6  0.0000  1     0.9995
## mod_type_inter_ML   10 1718.2 1771.2 -849.09   1698.2  0.3869  1     0.5339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA output tells us to retain `arabica_robusta` as a fixed effect over the other tested terms.



From the diagnostic plots above, we can see that the normality assumption is in doubt due to the deviations in the tails of the overall qqplot. The residual plot demonstrates an interesting pattern to the left, such that there are a few parallel/diagonal arrangements. This is most probably due to the reported values for those countries being very close to zero. In fact, upon closer inspection, some of these points correspond to Nepal which had the lowest coffee production.

Recalling the strong positive skew in the data, I then implemented a Gamma log-linked GLMM in an attempt to improve the model even further.



From the diagnostic plots above, there is still doubt regarding the normality assumptions. In particular, the residual plot is extremely deviated to the left of the plot, with a few values being found at the upper most range of the fitted values. However, when comparing the AIC values, the GLMM has a lower AIC and thus, improves upon the previous model.

Considering the severe skew of the data, I decided to implement a transformation using the `transformTukey` function. From this (*), we get a lambda of 0.05; since this value is greater than 0, the corresponding transformation is recommended to be x ^ lambda where x is the response.

```
## [1] 18363.62
```

```
## [1] 18517.87
```

```
## [1] 18650
```

From the AIC values, we can see that the Tukey transformed model fits the data better than the corresponding log transformed model but still does not improve upon the GLMM.

Considering the GLMM diagnostics, I thought it would be sensible to conduct a test for outliers by deriving the Cook's distance.

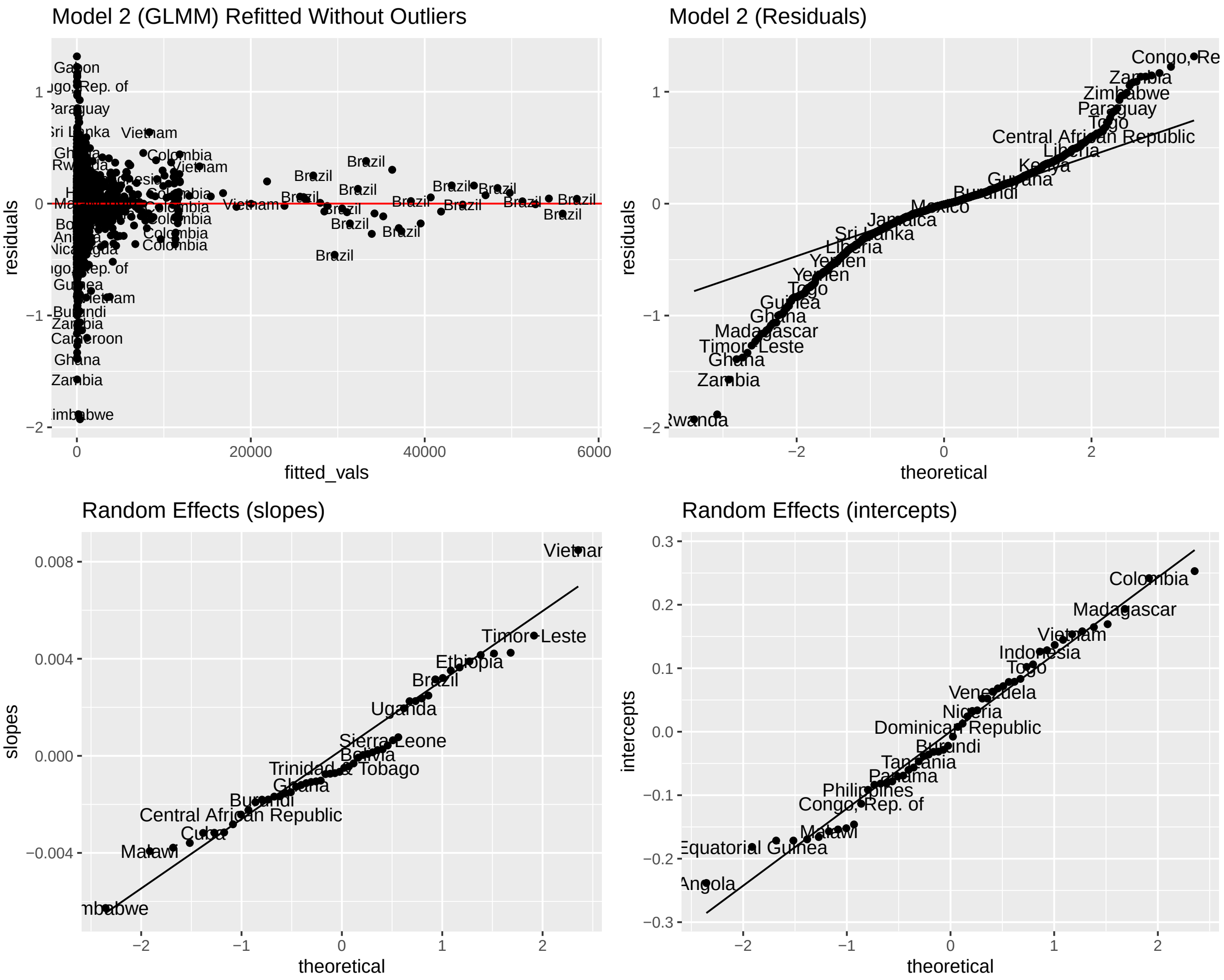
```
## # A tibble: 3 x 6
##   country        cooksd mdffits covtrace covratio leverage.overall
##   <fct>           <dbl>   <dbl>    <dbl>    <dbl>            <dbl>
## 1 Côte d'Ivoire 0.107   0.101    0.0816     1.08            0.0700
## 2 Nepal         0.0823  0.0792   0.0558     1.06            0.0678
## 3 Angola        0.0650  0.0604   0.0958     1.10            0.118
```

From the plot above (*), we can see that 2 countries exceed the threshold value as indicated by the red line. Using the corresponding dataframe containing the influence data, I arranged the Cook's distance values so as to show the countries with the highest values first as these would correspond to the outliers. Côte d'Ivoire and Nepal all have Cook's distance values of more than 0.08 and so, we remove them from the original dataframe and refit both the GLMM and Tukey transformed models.

```
## [1] 17997.68
```

```
## [1] 17848.69
```

After refitting the respective models to the adjusted data, we can see that the AIC values corresponding to each of respective models have decreased; the GLMM still has the lower AIC so we retain it as our final model.



Visually, there is not a noticable difference in the diagnostic plots, however, it should be noted that the random effect plots (both for slopes and intercepts) now fall at a more 45 degree angle when compared to the previous model.

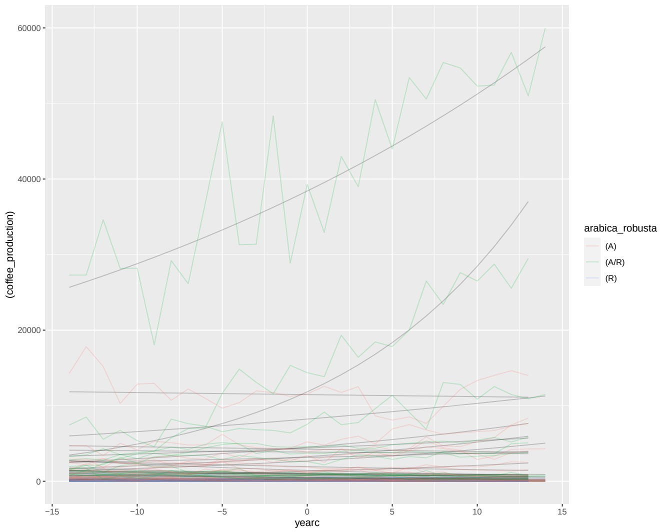
```
##                          Estimate  Std. Error   t value     Pr(>|z|)
## (Intercept)           5.911028854 0.389567494 15.173311 5.312755e-52
## yearc                -0.009958371 0.007283693 -1.367215 1.715581e-01
## arabica_robusta(A/R)  1.427126221 0.624290887  2.285996 2.225452e-02
## arabica_robusta(R)   -2.997384295 0.574100536 -5.221009 1.779505e-07
```

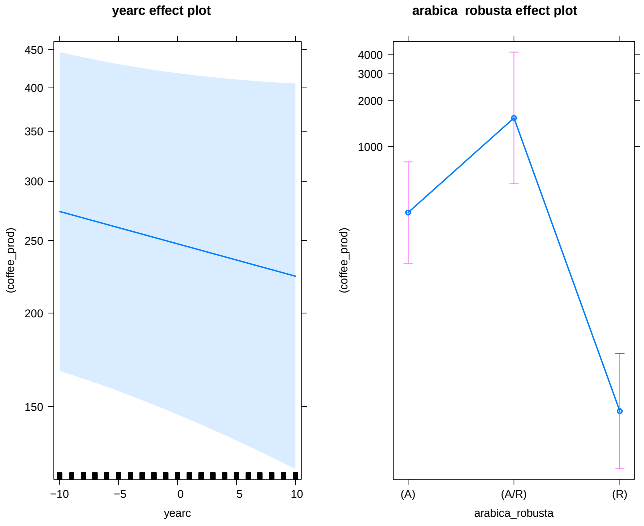
The estimates under the final model are as follows:

- Total annual coffee production (in thousand 60kg bags) was approximately 100 x (exp(-0.009958371)-1) = 99% less for each year passed (between 1990 and 2018)
- In the same year, if arabica coffee was used then we expect to see an average increase of exp(1.427126221)-1 = 3.2 thousand 60kg bags in total annual coffee production if a combination of arabica and robusta was used instead.
- In the same year, if arabica coffee was used then we expect to see an average decrease of exp(-2.997384295)-1 = 0.95 thousand 60kg bags in total annual coffee production if a combination of arabica and robusta was used instead.

This can be easier visualised by the plots below:

- Note that there is a significant difference between coffee production with robusta type versus arabica as well as combination (as the respective pink bars do not overlap). From this, it may be sensible to think that countries using a combination of coffee types may be using more arabica type than robusta type. This also follows from the proportion plots which were first discussed in this section.


(coffee_production)
yearc
arabica_robusta
(A)
(A/R)
(R)


yearc effect plot
(coffee_prod)
yearc
arabica_robusta effect plot
(coffee_prod)
(A)
(A/R)
(R)
arabica_robusta