ETC2420 Assignment 2

Rounak Agarwal, Jonas Tiong and Daniel Klinger

Contents

1 Question 1

Part A

1.1 Estimating Neonatal Mortality

1.1.1 Introduction

The aim of this paper is to estimate the average neonatal mortality rate (NMR) by a linear regression analysis. This is of interest as we want to analyse the claim that the *composition* of child mortality has changed, where the under 5 mortality rate (U5MR) has declined globally over the decades (since 1950, when our data starts) leading to a higher classification of child mortality as NMR.

Hence, our regression model(s) will look at the effect of NMR on U5MR, as well as other predictors such as by year and by region, if applicable. A rudimentary check of the applicability of such a model including `year`, `region`, `u5mr`, requires us to start by analysing the simple linear model of `scaled_nmr` on `year` and adding more covariates if it improves the fit of our model, as determined by the adjusted R^2. That is, does adding more covariates describe more of our child mortality data?

```
fit1 <- lm(scaled_nmr ~ year, data = dat)
fit2 <- lm(scaled_nmr~ year + region, data = dat)
fit3 <- lm(scaled_nmr ~ year + log_u5mr + region, data = dat)
fit4 <- lm(scaled_nmr ~ year + log_u5mr*region, data = dat)
```

```
model_selec <- tibble("Model" = c(1, 2,3,4),
                      "Adjusted R Squared" = c(summary(fit1)$adj.r.squared,
                                               summary(fit2)$adj.r.squared,
                                               summary(fit3)$adj.r.squared,
                                               summary(fit4)$adj.r.squared))

model_selec%>%
  kable(position = "center")
```

Model	Adjusted R Squared
1	0.04
2	0.43
3	0.55
4	0.61

From the table, it can be seen that our first model is unable to explain most of our data as adduced by its measly ajdusted $R^2 = 0.04$. A visible improvement is made to our model when we include a categorical variable that accounts for region ($R^2 = 0.43$). This is further improved when we add the log of U5MR to our model (we must take the log of `u5mr`, as we are, in fact, modelling a scaled NMR that is more appropriate for regression lines - scaled NMR $= log(\frac{nmr}{u5mr-nmr})$). Finally, we consider the benefits of using an interaction effect between the covariates `log_u5mr` and `region` through an analysis of variance that compares a model with the interaction to one without.

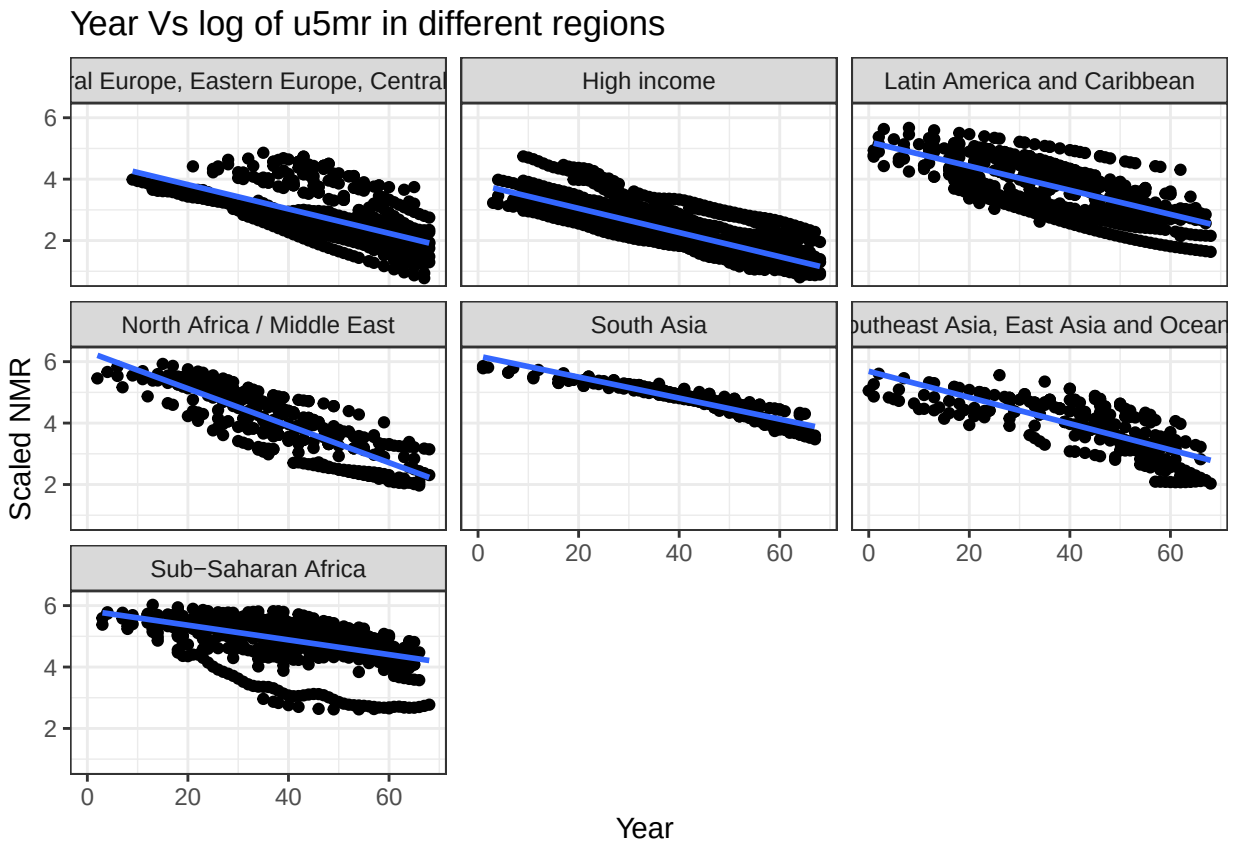
```
anova(fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: scaled_nmr ~ year + log_u5mr + region
## Model 2: scaled_nmr ~ year + log_u5mr * region
##   Res.Df RSS Df Sum of Sq   F Pr(>F)
## 1   4362 854
## 2   4356 740  6       114 112 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

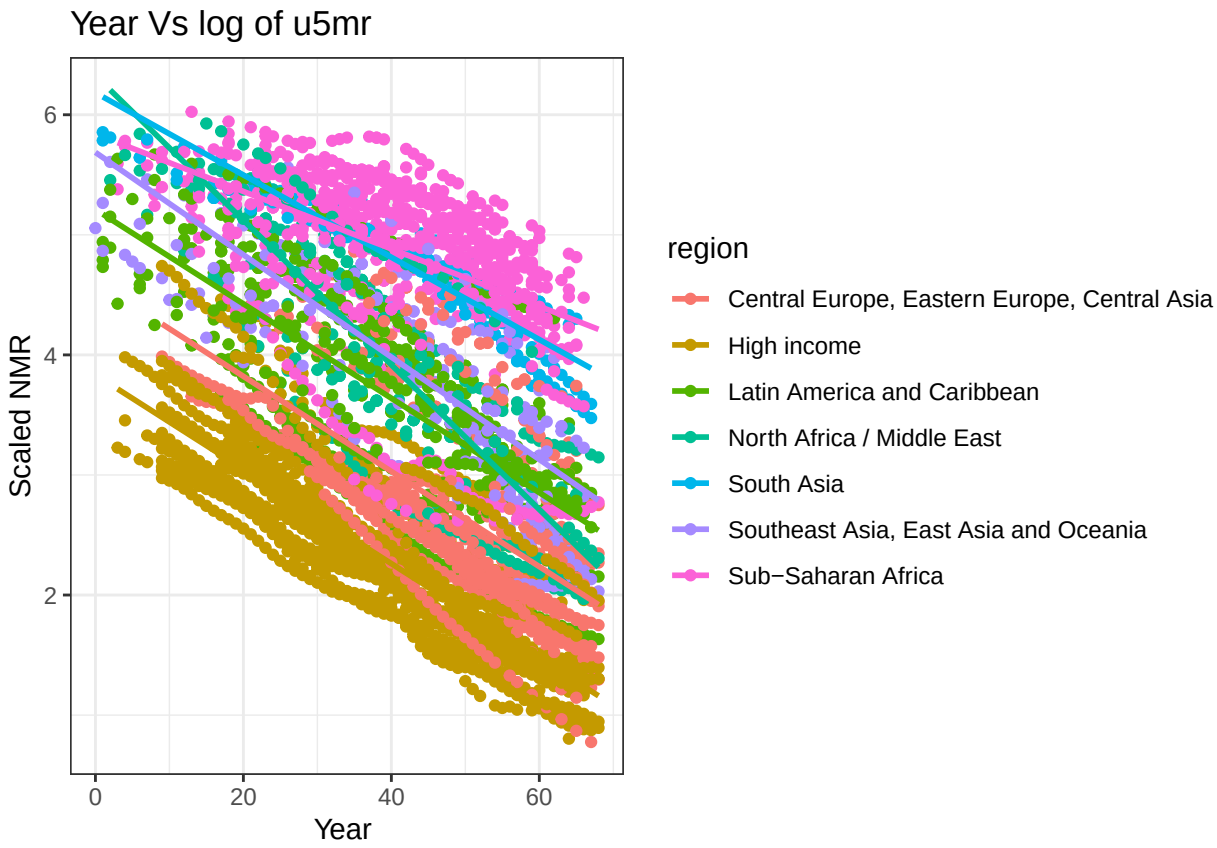
According to the analysis of variance, there is reasonable evidence to conclude at the 0.001 level of significance, that an interaction effect will improve the fit our regression. The residual sum of squares (RSS) also has improved with the interaction effect, which is linked to an improved fit to our data (by the RSE).

Looking at the graphs below, we may observe that every region has its own clear trend, a declining trend, of U5MR. And, this is at paces and level of U5MR at each point of time quite different to every other region. This suggests that it is appropriate to include a region variable and also model the interaction of region with U5MR. The second set of graphs plotting NMR v U5MR describes a downward trend that follows in NMR with the decline in U5MR. Hence, it also gives reason to think we should include U5MR in our regression.

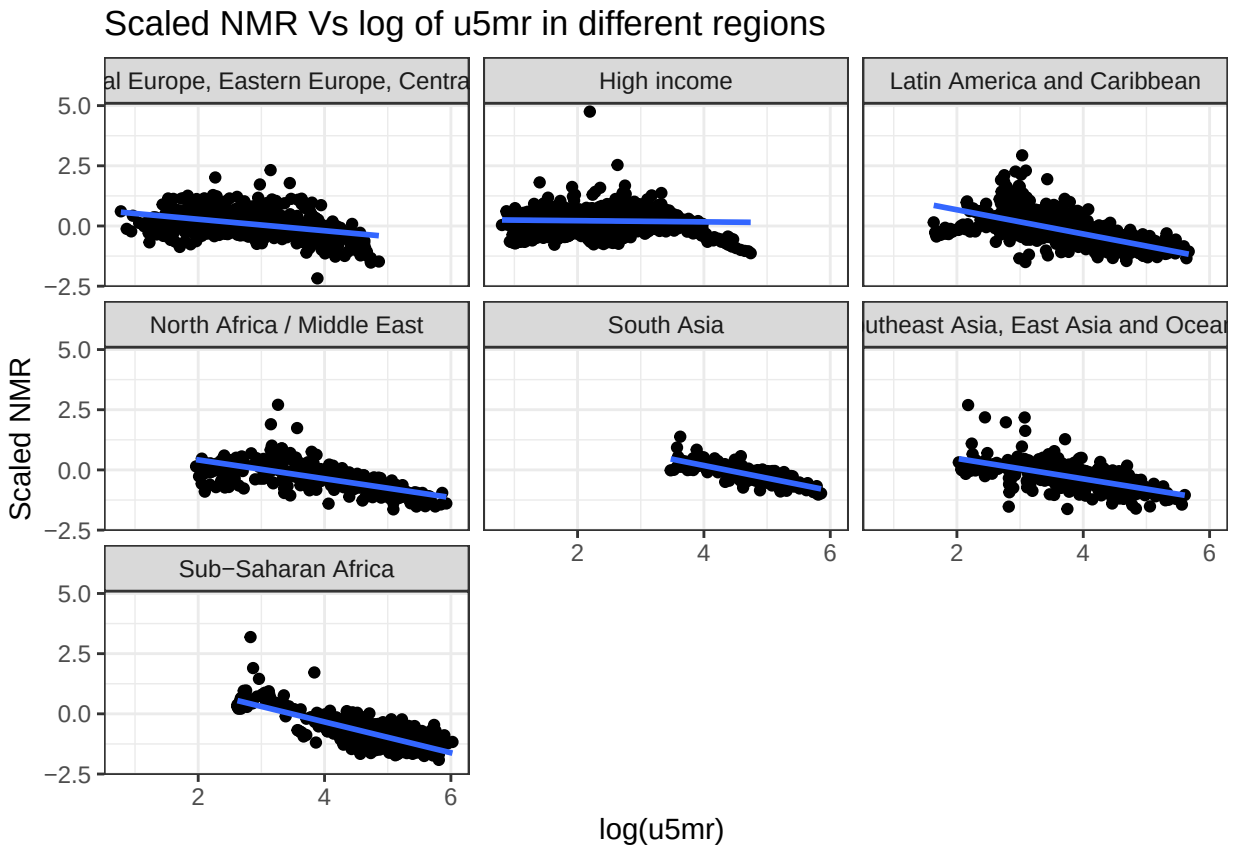
```
dat %>% ggplot(aes(x = year, y = log_u5mr)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE) +
  facet_wrap(~region)+
  labs(x = "Year ",
       y = "Scaled NMR",
       title = "Year Vs log of u5mr in different regions")+theme_bw()
```



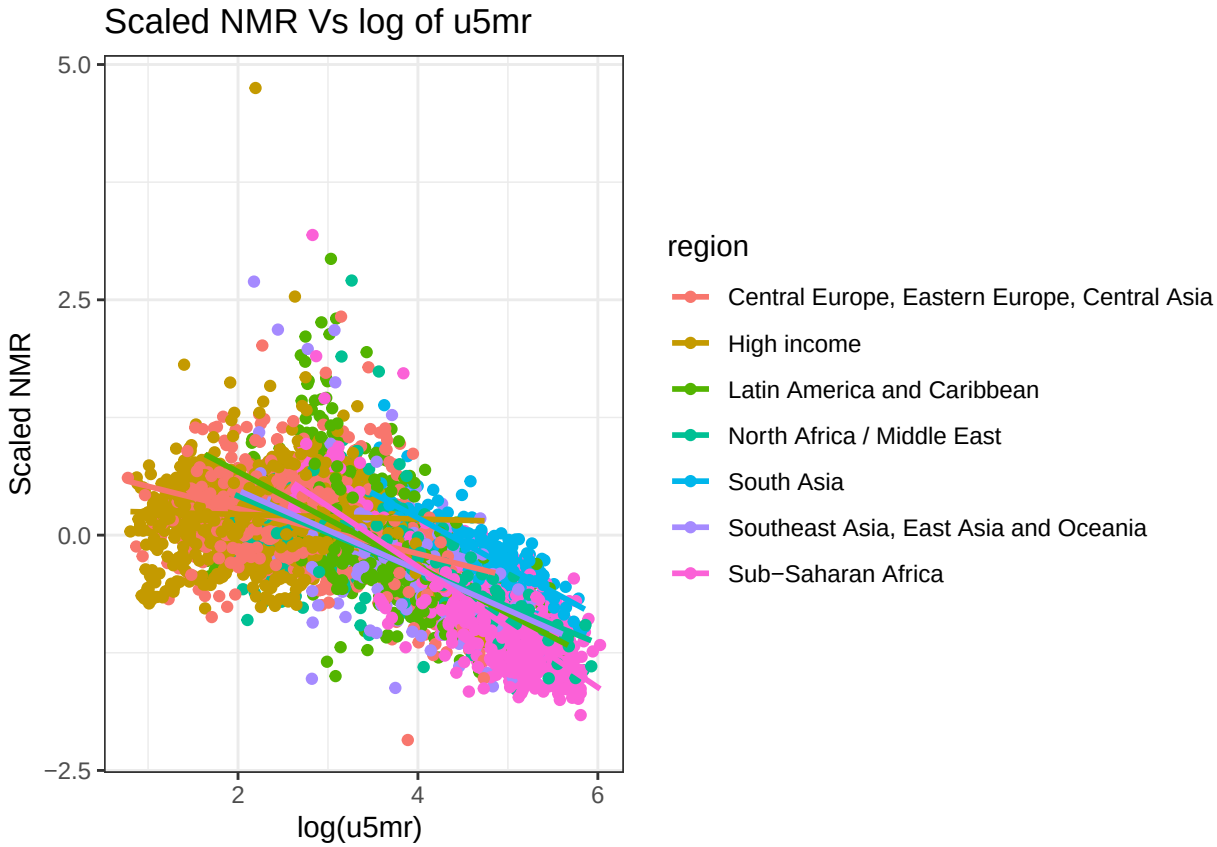

```
dat %>% ggplot(aes(x = year, y = log_u5mr,
                   colour = region)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE)+
  labs(x = "Year ",
       y = "Scaled NMR",
       title = "Year Vs log of u5mr")+theme_bw()
```




```
dat %>% ggplot(aes(x = log_u5mr, y = scaled_nmr)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE) +
  facet_wrap(~region)+
  labs(x = "log(u5mr) ",
       y = "Scaled NMR",
       title = "Scaled NMR Vs log of u5mr in different regions")+theme_bw()
```




```
dat %>% ggplot(aes(x = log_u5mr, y = scaled_nmr,
                   colour = region)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE)+
  labs(x = "log(u5mr) ",
       y = "Scaled NMR",
       title = "Scaled NMR Vs log of u5mr")+theme_bw()
```



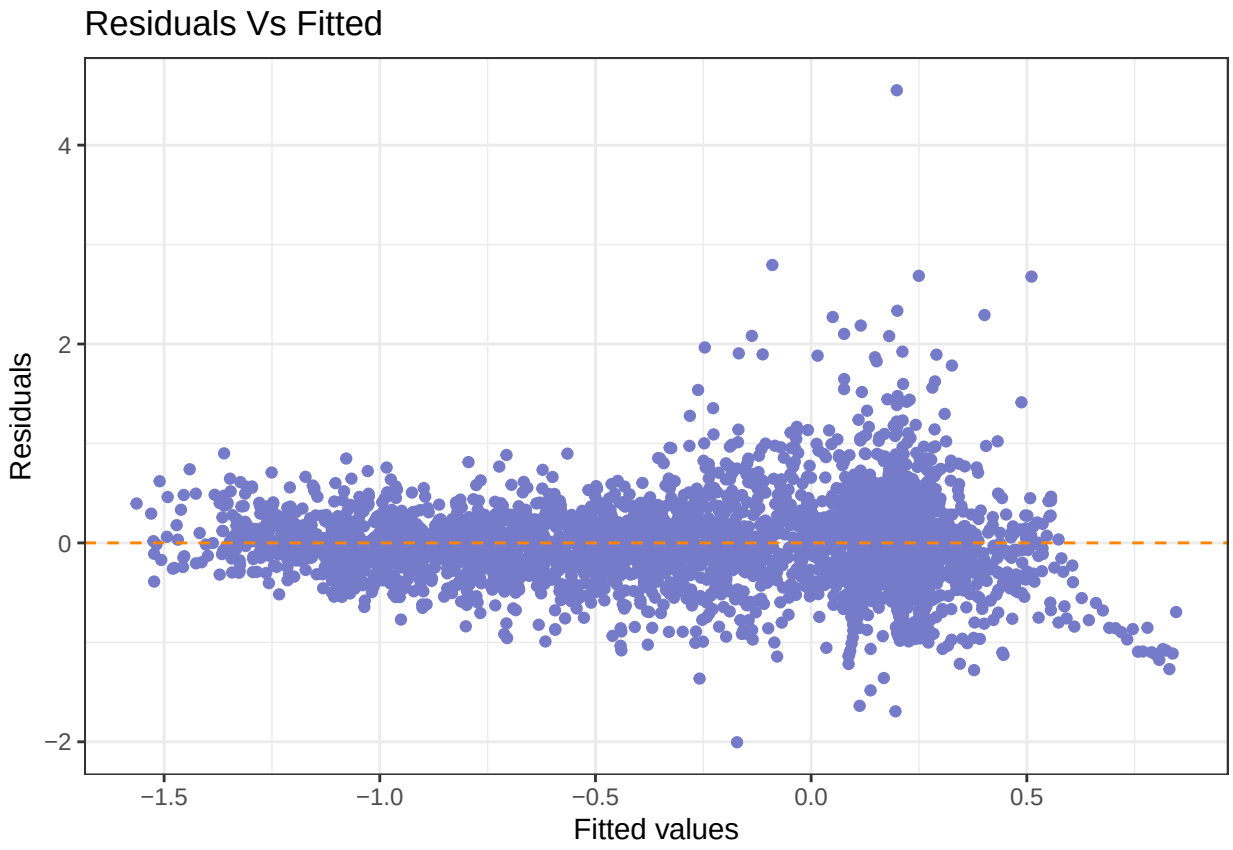
We conclude that our chosen linear regression model is `scaled_nmr` on `year`, the log of `u5mr`, and `region`, where we will model, in addition to the effects of condition on `u5mr` and `region` on `scaled_nmr`, a separate slope for `u5mr` for every `region` there is in the data (the interaction effect).

1.2 Adding the model to our workflow

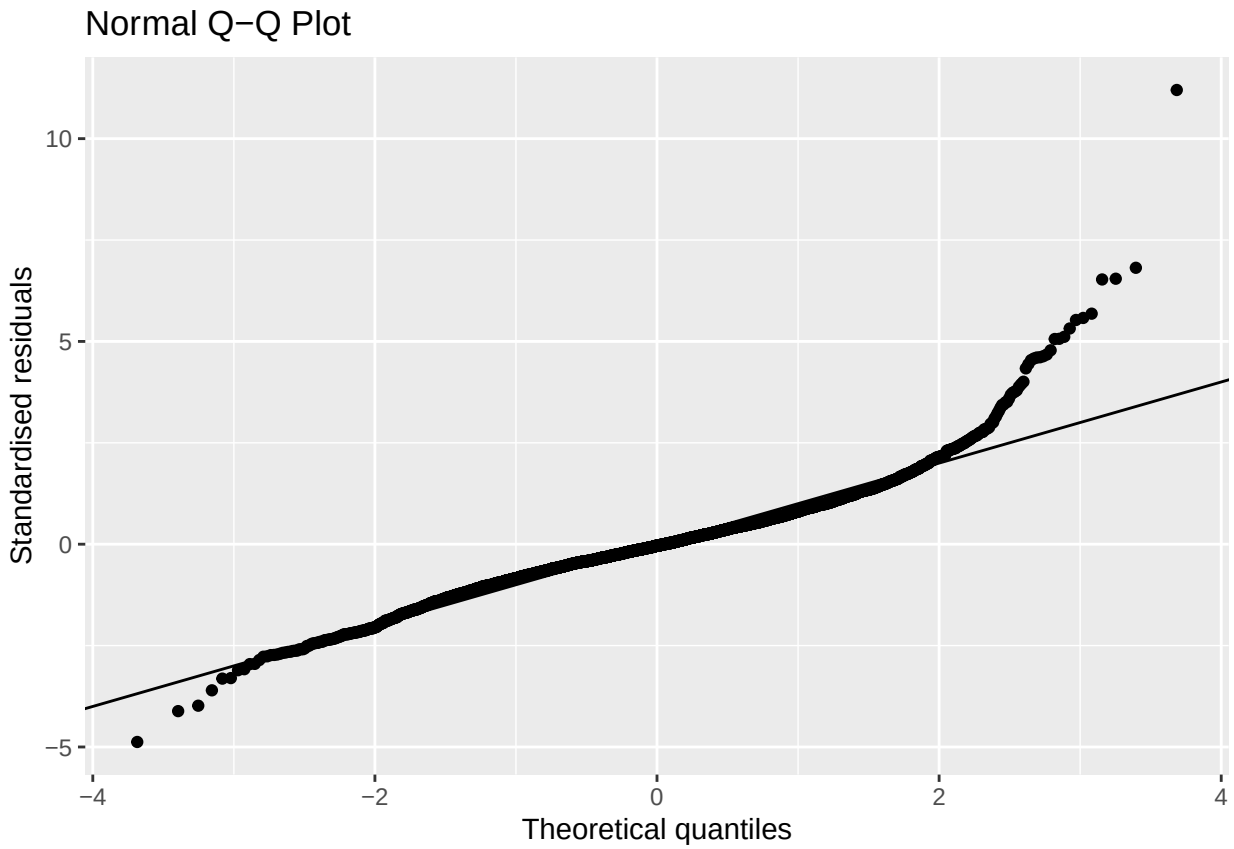
```
dat_rec <- recipe(scaled_nmr ~ year + u5mr + region, data = dat) %>%
  step_log(u5mr) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ u5mr*starts_with("region"))
#Step_interactallows for interaction between the covariates.
#Needed to convert all non-numeric covariates to dummy variables

# Model
lm_spec <- linear_reg() %>%
  set_engine("lm")

# Workflow linking the recipe & model
dat_wf <- workflow() %>%
  add_recipe(dat_rec) %>%
  add_model(lm_spec)

# Our regression line
fit_chosen <- dat_wf %>% fit(data = dat) %>% extract_fit_parsnip()
dat_chosen <- fit_chosen$fit %>% augment()
dat_chosen <- cbind(dat %>% select(-year, -u5mr), dat_chosen)
```

1.3 Assessing fit of all data

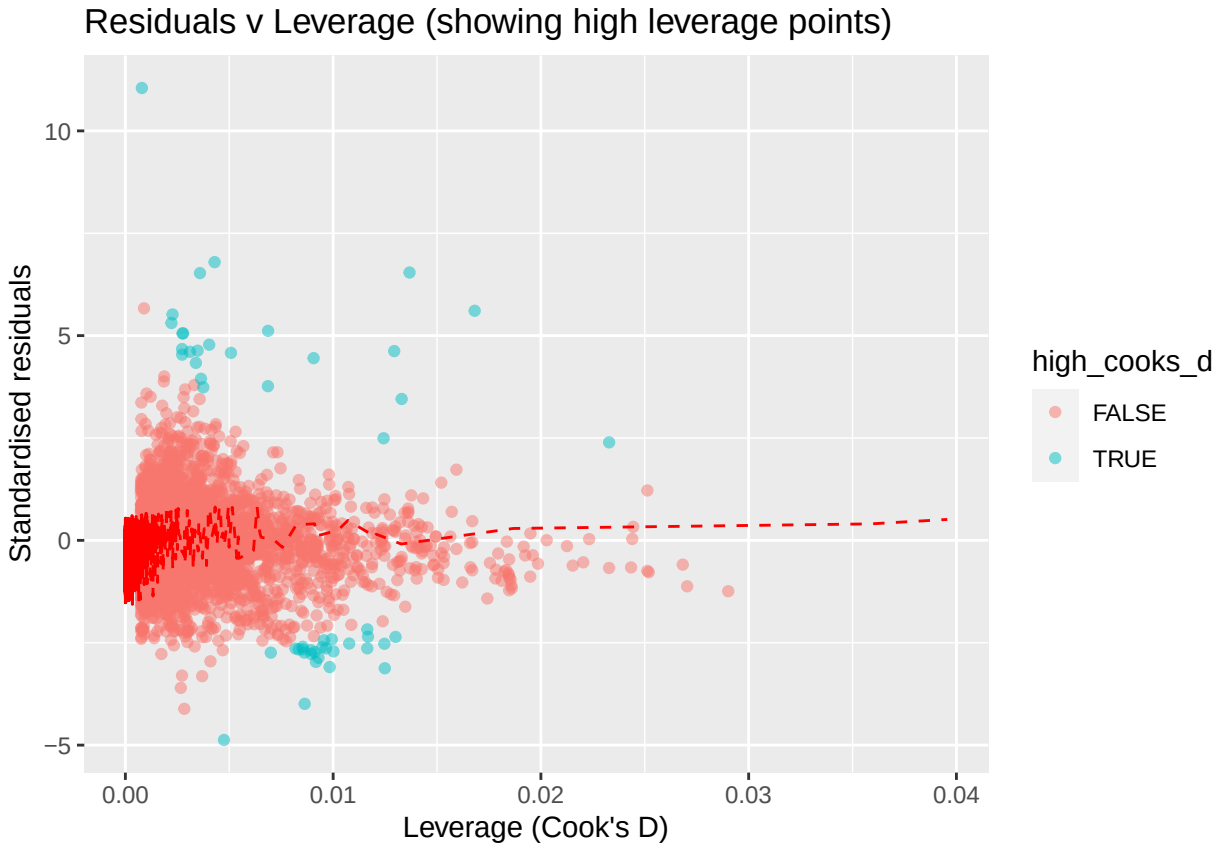
```
dat_chosen %>% ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(color = "#757bc8") +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "#ff8400") +
  labs(x = "Fitted values",
       y = "Residuals",
       title = "Residuals Vs Fitted")+theme_bw()
```




```
dat_chosen %>% ggplot(aes(sample = .resid / .sigma)) +
  geom_qq() +
  geom_abline(intercept = 0, slope = 1) +
  labs(x = "Theoretical quantiles",
       y = "Standardised residuals",
       title = "Normal Q-Q Plot")
```




```
p <- 14
n <- length(dat$nmr)
threshold <- (p + 1) / n

dat_chosen %>%
  mutate(high_cooks_d = .cooksd > threshold) %>%
  ggplot(aes(x = .hat, y = .std.resid,
             colour = high_cooks_d)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(x = .cooksd, y = .fitted), col = "red", lty = "dashed") +
  labs(x = "Leverage (Cook's D)",
       y = "Standardised residuals",
       title = "Residuals v Leverage (showing high leverage points)")
```



We take the residuals vs fitted; normal q-q plot; and residuals vs leverage as diagnostics for the fit of our data.

Residuals vs Fitted A linear model is appropriate when there is a constant variance against fitted values (of our regression). From our graph, there is a large change in variance after -0.5, where there are some large outliers compared to the previously constant variance data point before. After 0.5, the residuals sit well below the zero line as well. Hence, this is a signs of heretoskedasticity, meaning our regression line is suffering from the effects of high leverage or outlier data points and that our plot is not perfect. It does not, however, seem like that there is clear pattern emerging from this plot, thus it may stand that the residuals are independent which would be a good sign for this regression.

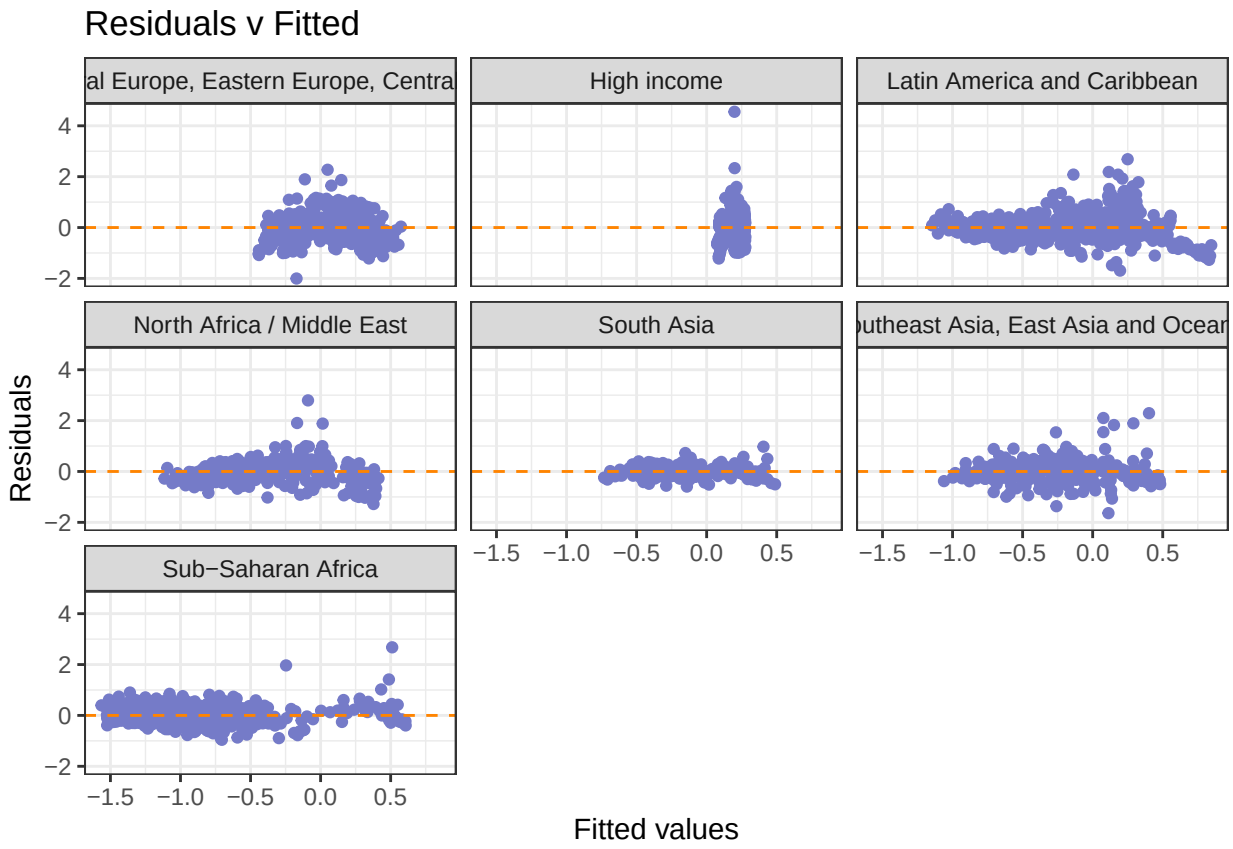
Normal Q-Q Plot A linear model may still be useful for prediction even if the assumptions that residuals are i.i.d normal do not hold (these are the assumptions that residuals should have constant variances and by independent in the residuals vs fitted plot). What has to hold is that they look approximately normal, as per the q-q plot. Hence, our regression model suffers in that the standardised residuals stray at the ends and beginning of the q-q plot, accounting for the large uncertainty (or noise) in our data in modelling the `scaled_nmr`.

Residuals vs Leverage This plot is useful for identifying those data points that greatly influence our regression. If there is high residual (poorly described by the model) and high leverage for a point (high influence on the fit), then such a point is considered an outlier. Outliers mean our regression will suffer in fitting our data well. Looking at the blue points, there are many outliers in our data, hence our regression model is not fitting it very well.

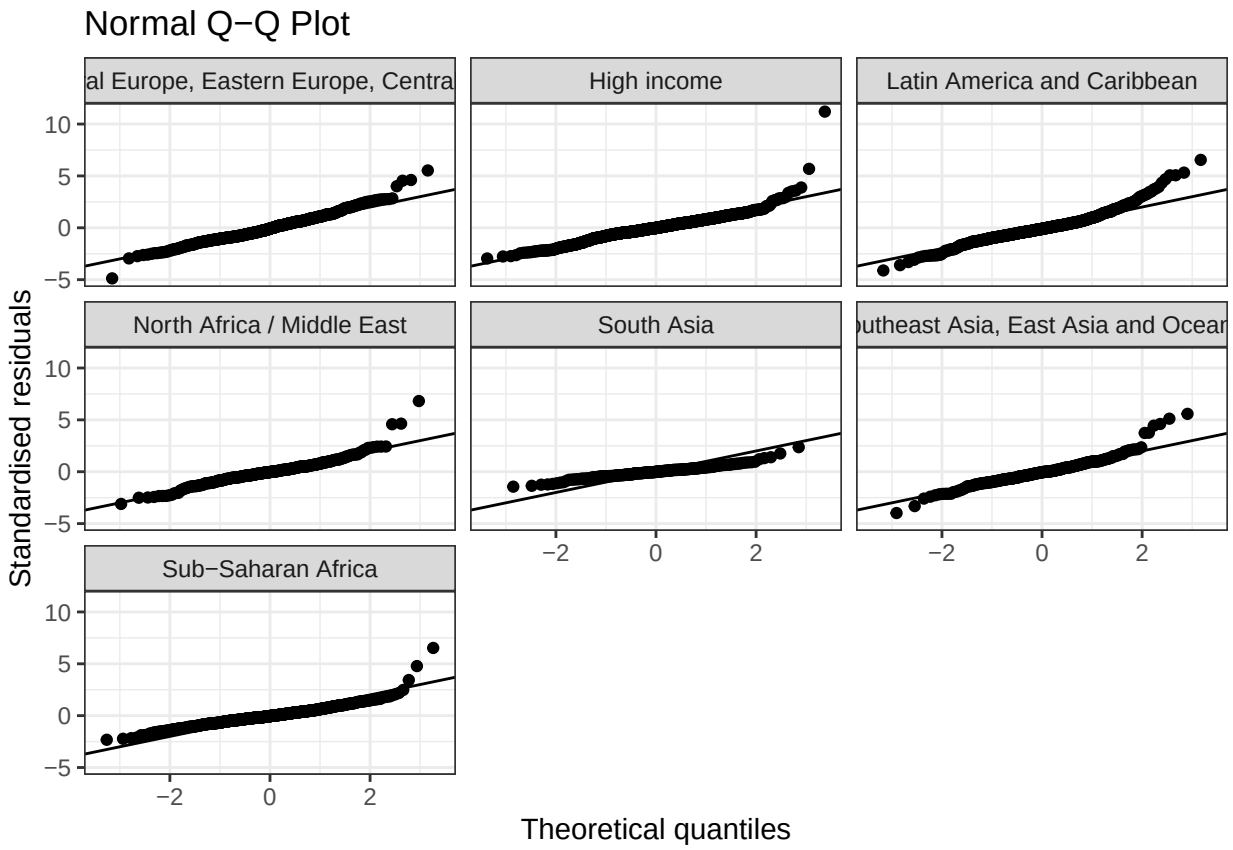
Overall, the fit of model is not perfect, however, its applicability for prediction is not solely determined by these diagnostic plots. Moreover, despite the inadequacies, the fit is not terrible as there is no clear pattern in the fitted vs residuals plot for instance (suggesting we are using the wrong type of model).

1.4 Fitting model by region

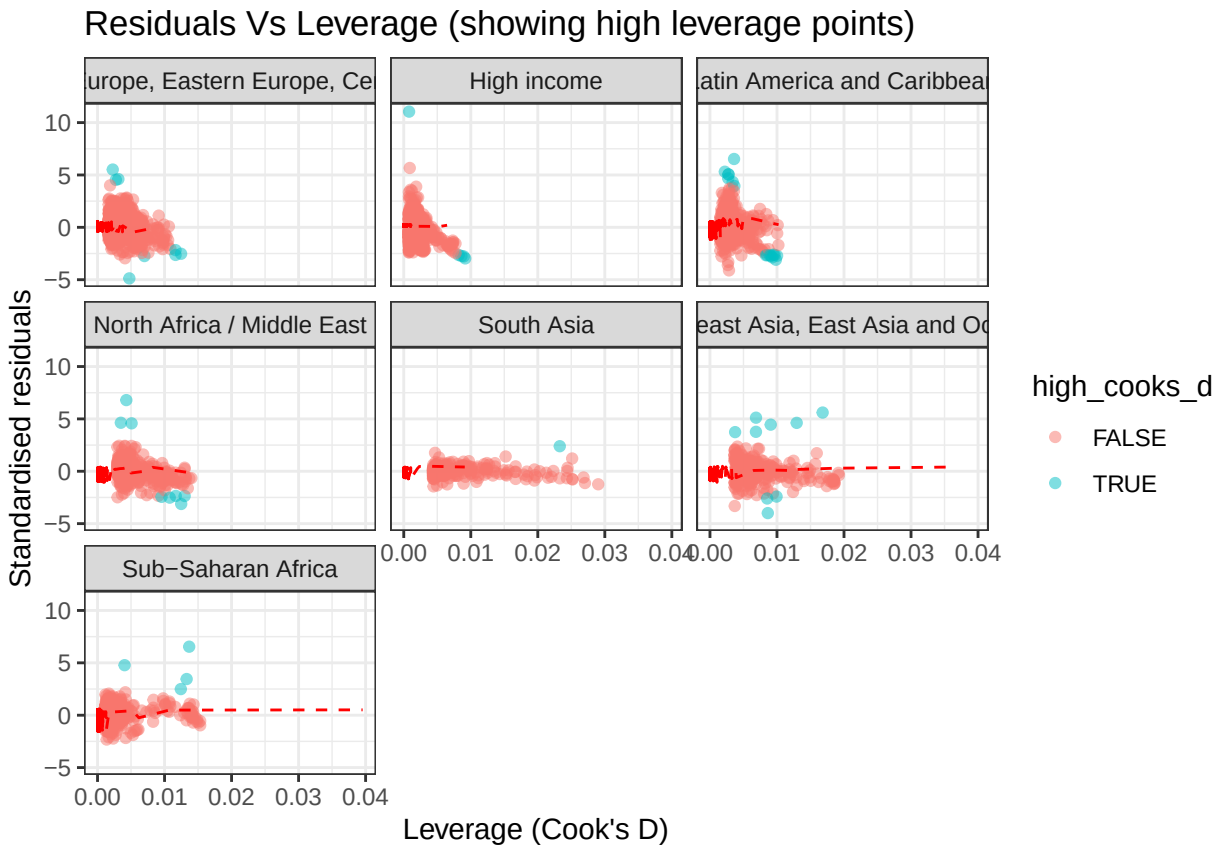
```
dat_chosen %>% ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(color = "#757bc8") +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "#ff8400") +
  facet_wrap(~region) +
  labs(x = "Fitted values",
       y = "Residuals",
       title = "Residuals v Fitted")+theme_bw()
```




```
dat_chosen %>% ggplot(aes(sample = .resid / .sigma)) +
  geom_qq() +
  geom_abline(intercept = 0, slope = 1) +
  facet_wrap(~region) +
  labs(x = "Theoretical quantiles",
       y = "Standardised residuals",
       title = "Normal Q-Q Plot")+theme_bw()
```




```
p <- 14
n <- length(dat$nmr)
threshold <- (p + 1) / n

dat_chosen %>%
  mutate(high_cooks_d = .cooksd > threshold) %>%
  ggplot(aes(x = .hat, y = .std.resid,
             colour = high_cooks_d)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(x = .cooksd, y = .fitted), col = "red", lty = "dashed") +
  facet_wrap(~region) +
  labs(x = "Leverage (Cook's D)",
       y = "Standardised residuals",
       title = "Residuals Vs Leverage (showing high leverage points)")+theme_bw()
```

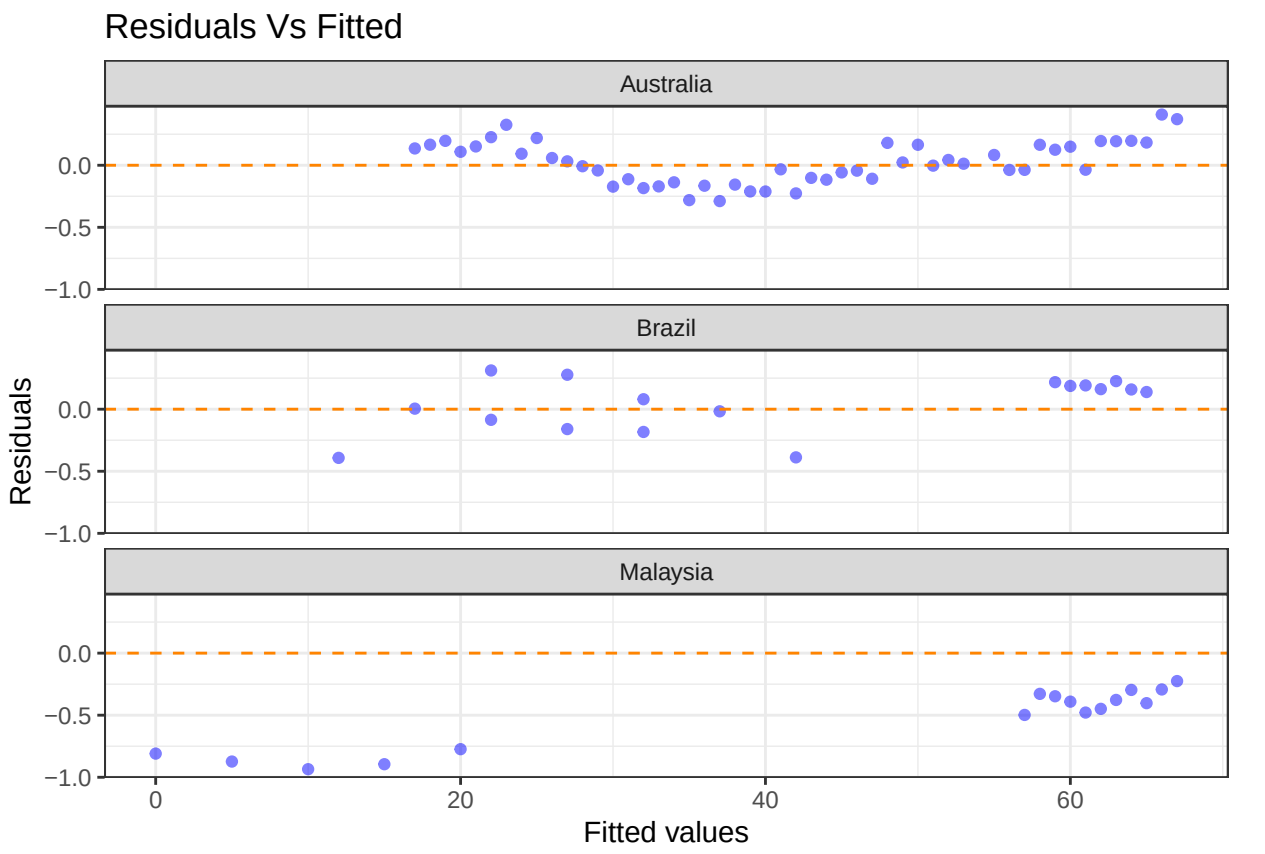


The regression fares very well for South Asia and not so well for the other regions. By region, the residuals seem very high at some fitted values and for most part are constant in variance as it was for the whole data set. Visible differences are for the region Eurasia (Central Europe, Eastern Europe, Central Asia) and the high income countries. This will require further exploration, however, we think that in the case of Eurasia, this may be caused by the demographic isolation and make-up of some settlements being far away from each other and the central city (giving rise to the poor fit). Such factors would lead to increased risk of death of children, and particularly for young babies who need intensive care that may not be able to provided in such places. In the case of high income countries, fitted values tend to center, we believe, because they generally have a minimum understanding of where they want there child mortality rates to be at, and its correlated to the high expenditure and wealth of these countries to support such rates. Moreover, the high residuals may be caused from the fact some of these countries have very strong policies and institutions that prevent neonatal deaths. What is clear is that these regions, and with the issue of heteroskedaticity for most of the other regions, make our fit to the data limited by our model. Such issues can be seen by the normalised q-q plot, where residuals stray at the tails and the high number of outliers (save South Asia). It will also be interesting to explore why South Asia is fitting our regression model very well. This may, however, be largely due to the lack of outliers in the data.

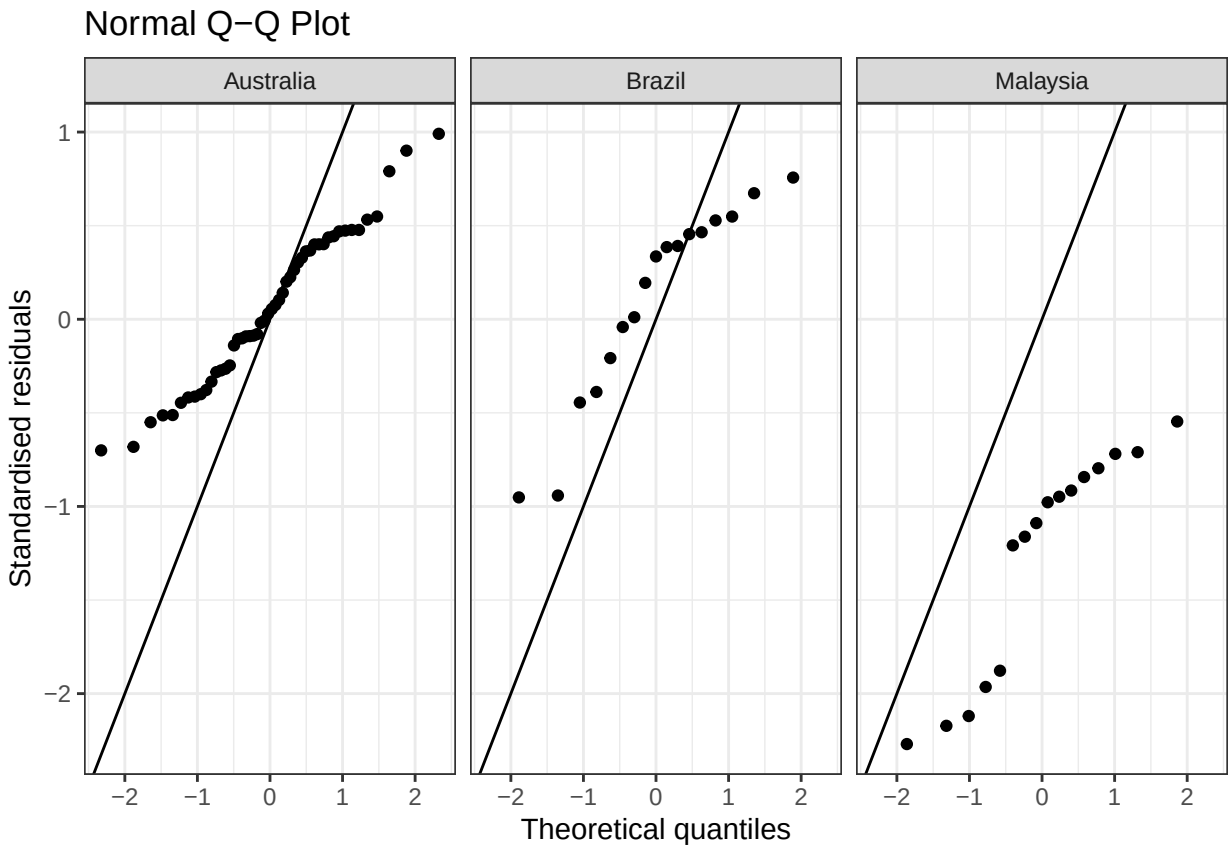
1.5 Fitting model by the three selected countries

```
countries <- c("Malaysia", "Brazil", "Australia")
dat_chosen_three <- dat_chosen %>%
  filter(country_name %in% countries) #IS THERE A BETTER CHOICE OF COUNTRIES?

dat_chosen_three %>%
  ggplot(aes(x = year, y = .resid)) +
  geom_point(alpha = 0.5,color = "blue") +
  geom_hline(yintercept = 0,
```

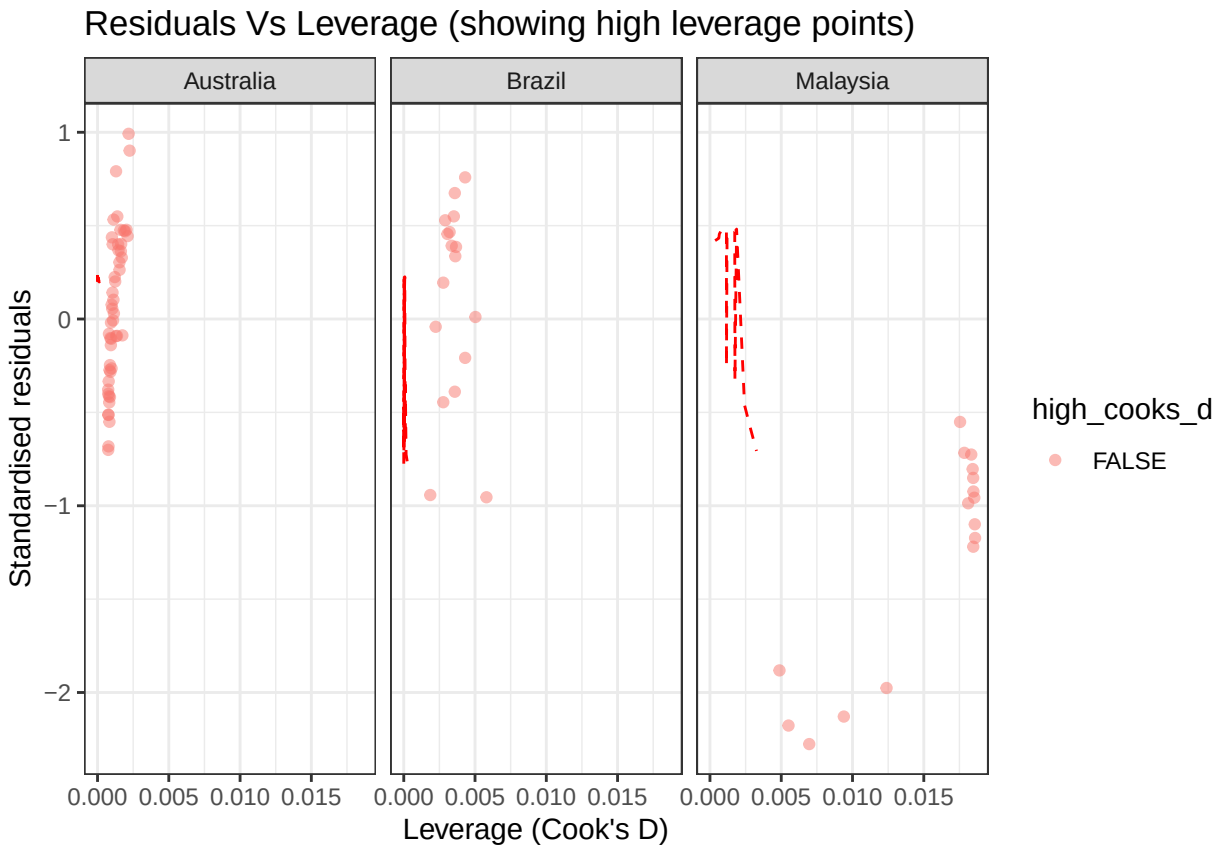
```
            linetype = "dashed",
            colour = "#ff8400")+
  facet_wrap(~country_name, nrow = 3) +
  labs(x = "Fitted values",
       y = "Residuals",
       title = "Residuals Vs Fitted")+theme_bw()
```




```
dat_chosen_three %>% ggplot(aes(sample = .resid / .sigma)) +
  geom_qq() +
  geom_abline(intercept = 0, slope = 1) +
  facet_wrap(~country_name) +
  labs(x = "Theoretical quantiles",
       y = "Standardised residuals",
       title = "Normal Q-Q Plot")+theme_bw()
```




```
dat_chosen_three %>%
  mutate(high_cooks_d = .cooksd > threshold) %>%
  ggplot(aes(x = .hat, y = .std.resid,
             colour = high_cooks_d)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(x = .cooksd, y = .fitted), col = "red", lty = "dashed") +
  facet_wrap(~country_name) +
  labs(x = "Leverage (Cook's D)",
       y = "Standardised residuals",
       title = "Residuals Vs Leverage (showing high leverage points)") +
  theme_bw()
```



We have chosen Australia, Brazil, and Malaysia. The reason being is that they are part of three different continents and are in very different stages of their growth. Australia is a high income country, Brazil and Malaysia are not. There is a pattern for all these countries in their residuals, which means the regression does not do well to model these countries' NMR. This is, however, not the biggest of concern, for each country will likely have huge variance and its own distinct pattern for NMR over time compared to global averages and regional averages. The residuals are clearly not normal as per the q-q plot. None of these countries have outliers or even high leverage points. Yet, there standardised residuals is quite high.

1.6 Validation Set Approach to RMSE & MAE

We are concerned here to estimate the test root mean square error (RMSE) and test mean absolute error (MAE), through the validation set approach. A low estimate for these metrics would be a favourable sign to the fit of our model.

```
dat_split <- initial_split(data = dat, strata = region)
dat_train <- training(dat_split)
dat_valid <- testing(dat_split)

fit_chosen_train <- dat_wf %>% fit(dat_train)
dat_chosen_valid <- dat_valid %>%
  bind_cols(predict(fit_chosen_train, dat_valid))

dat_chosen_valid %>% metrics(truth = scaled_nmr, estimate = .pred) %>%
  rename("Estimator" = ".metric", "Estimate"= ".estimate")%>%
  select(-`.estimator`)%>%
  kable()
```

Estimator	Estimate
rmse	0.4
rsq	0.6
mae	0.3

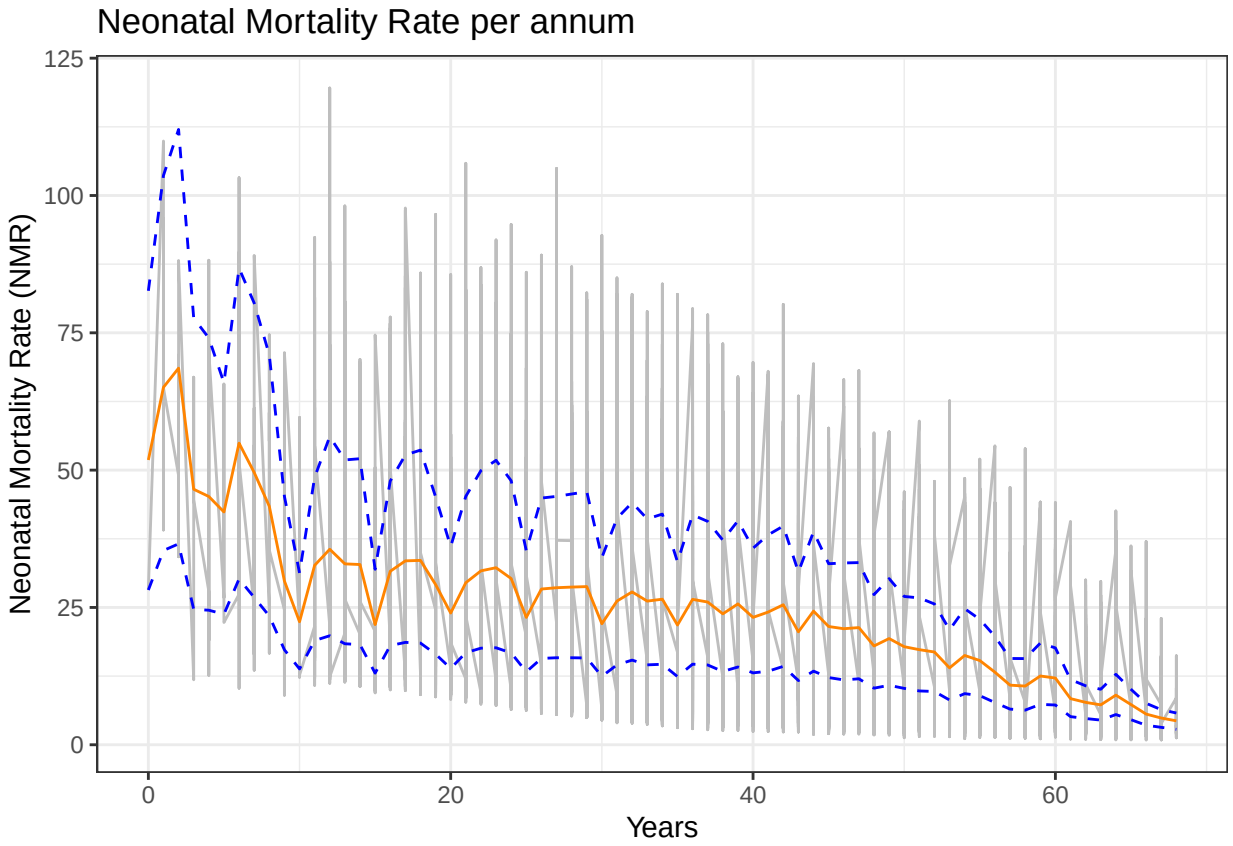
We opted for the default 75 per cent split of our data to training and validation sets. On this, our model (on the training set) has an estimated test RMSE of 0.42 and an estimated test MAE of 0.30. Such values suggest that our model is not doing too bad in fitting our data, with its estimated test MAE of 0.30 being favourable sign.

1.7 Prediction CLT intervals

We test here how well our model can predicate the data despite the inadequancies in its fit as determined and discussed before.

1.8 Prediction CLT interval for whole data

```
pred_nmr <- predict(fit_chosen$fit, dat_chosen, interval = "prediction") %>%
  as.data.frame()

pred_nmr <- bind_cols(dat, pred_nmr) %>% mutate(
  fitted = exp(fit) * u5mr / (1 + exp(fit)),
  upr_ci = exp(upr) * u5mr / (1 + exp(upr)),
  lwr_ci = exp(lwr) * u5mr / (1 + exp(lwr))
  ) %>% select(-fit, -upr, -lwr)

pred_nmr_all <- pred_nmr %>% group_by(year) %>%
  summarise(fit = mean(fitted),
            lwr = mean(lwr_ci),
            upr = mean(upr_ci),
            nmr = nmr)

pred_nmr_all %>% ggplot(aes(x = year)) +
  geom_line(aes(y = nmr), colour = "grey") +
  geom_line(aes(y = fit), colour = "#ff8400") +
  geom_line(aes(y = lwr), colour = "blue", linetype = "dashed") +
  geom_line(aes(y = upr), colour = "blue", linetype = "dashed") +
    labs(
      x = "Years",
      y = "Neonatal Mortality Rate (NMR)",
      title = "Neonatal Mortality Rate per annum")+theme_bw()
```

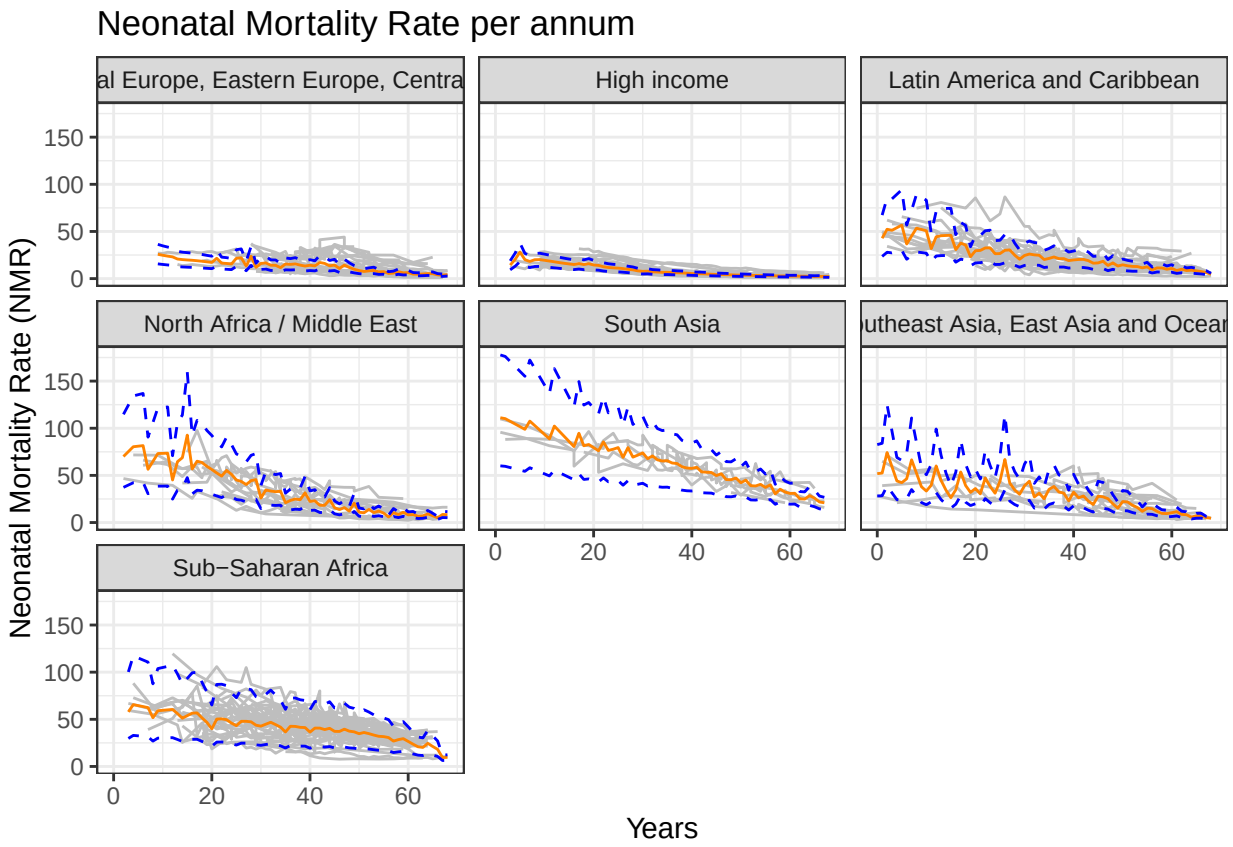


The grey lines are the plot of the points from the data, and the yellow line is the average NMR as predicted by our model. The blue lines that encloses this is the 95 per cent prediction CLT interval. From the graph, we notice that again the downward trend in NMR over the years. We also see that our model predicts an average NMR with its prediction intervals that sits within the true values (the grey lines). This suggests that the fit of model is good on the scale of global averages.

1.9 Prediction CLT interval by region

```
pred_nmr_region <- pred_nmr %>% group_by(year, region) %>%
  summarise(fit = mean(fitted),
            lwr = mean(lwr_ci),
            upr = mean(upr_ci))

pred_nmr %>% ggplot(aes(x = year)) +
  geom_line(aes(y = nmr, group = country_name),
            colour = "grey") +
  geom_line(data = pred_nmr_region, aes(y = fit),
            colour = "#ff8400") +
  geom_line(data = pred_nmr_region, aes(y = lwr),
            colour = "blue",
            linetype = "dashed") +
  geom_line(data = pred_nmr_region, aes(y = upr),
            colour = "blue",
            linetype = "dashed") +
  facet_wrap(~region, nrow = 3) +
  labs(
      x = "Years",
```

```
      y = "Neonatal Mortality Rate (NMR)",
      title = "Neonatal Mortality Rate per annum")+theme_bw()
```

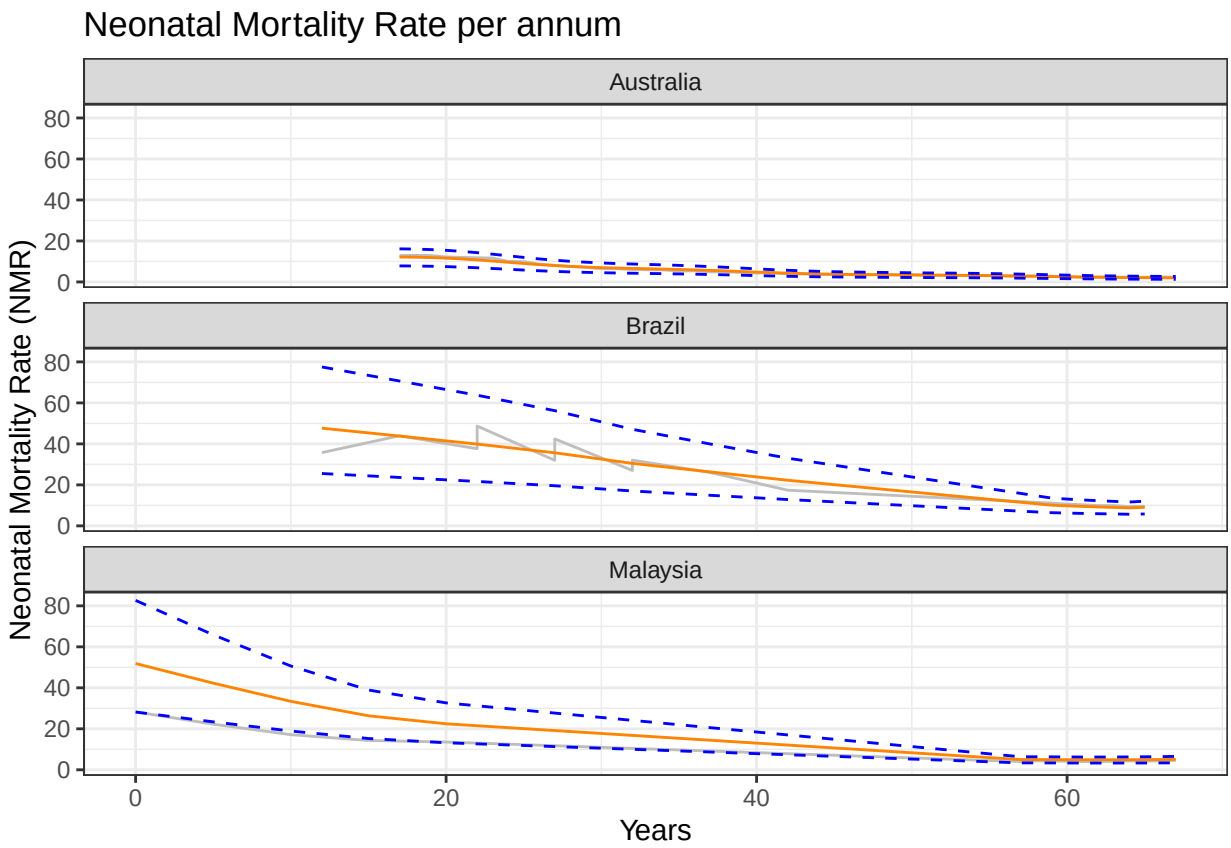


By region, the breakdown isn't as favourable. Our model predicts large uncertainty in its value for NMR for South Asia, which comes from uncertainty in other parts of the whole data set. This is not bad per se, as the future trend of NMR is reliant on the observations of the trajectory of NMR in the past for other countries. For North Africa / the Middle East and Sub-Saharan Africa, the uncertainty is also quite big in the predicted values, which is a reflection of the underlying uncertainty of estimating NMR for them. What is surprising is that for high income countries and Eurasia, this uncertainty is quite low. Perhaps this is because for our q-q plots in §2.2, the residuals only strayed very far off in the tails, which meant that our model was still good for predicting the average NMR.

1.10 Prediction CLT intervals of three countries

```
pred_nmr_three <- pred_nmr %>% filter(country_name %in% countries)


pred_nmr_three <- pred_nmr_three %>% group_by(year, country_name) %>%
  summarise(fit = mean(fitted),
            lwr = mean(lwr_ci),
            upr = mean(upr_ci),
            nmr = nmr)

pred_nmr_three %>% ggplot(aes(x = year)) +
  geom_line(aes(y = nmr), colour = "grey") +
  geom_line(aes(y = fit), colour = "#ff8400") +
  geom_line(aes(y = lwr), colour = "blue", linetype = "dashed") +
```

```
geom_line(aes(y = upr), colour = "blue", linetype = "dashed") +
facet_wrap(~country_name, nrow = 3) +
labs(
    x = "Years",
    y = "Neonatal Mortality Rate (NMR)",
    title = "Neonatal Mortality Rate per annum")+theme_bw()
```



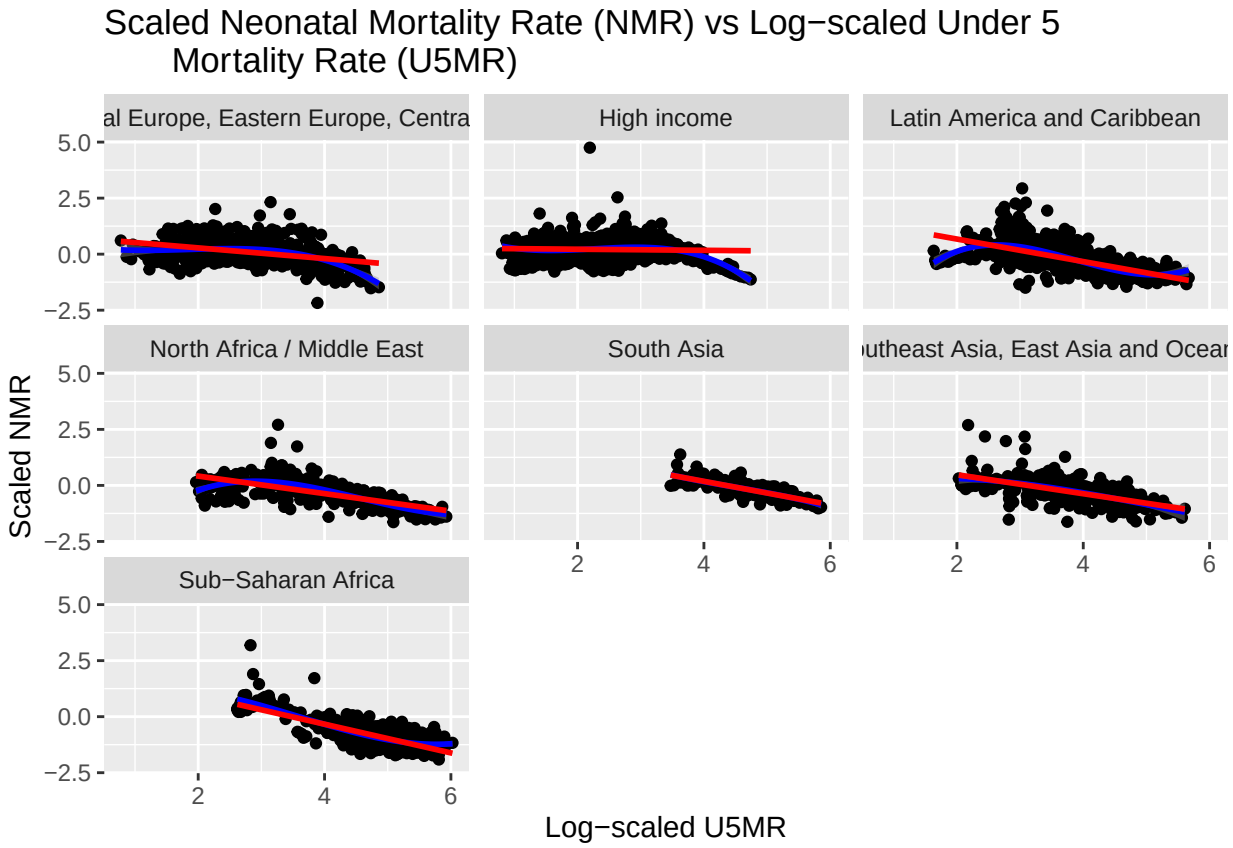
For these three countries, the prediction intervals has large uncertainty that is bigger than the true values. This again is expected, we cannot expect that our model will fare just as well as it did on the whole and for each region as it would for just one country.

Part B

1.11 Non-linearity of U5MR vs NMR

A reason that our fit for our linear regression model was not very good could have possible due to not capturing the relationship between U5MR and NMR in the best way. We consider the linear regression against a non-linear regression model using a B-Spline for the `log_u5mr` covariate.

```
dat  %>%
  ggplot(aes(x = log_u5mr, y = scaled_nmr)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ bs(x), colour = "blue") +
  geom_smooth(method = "lm", colour = "red") +
  facet_wrap(~region, nrow = 3) +
  labs(x = "Log-scaled U5MR",
       y = "Scaled NMR",
       title = "Scaled Neonatal Mortality Rate (NMR) vs Log-scaled Under 5
       Mortality Rate (U5MR)")
```



This model has visibile improvements over its linear counterparts for some regions. Latin America and Caribbean; North Africa / Middle East; Eurasia (Central Europe, Eastern Europe, Central Asia); High income are regions where the B-spline modelling has reduced the distance of the residuals from the fitted values (the blue line). Other countries have had muted benefits to the linear model (the red line), where in South Asia there is no improvements. As noted in §2.2, the model has plotted the relationship very well for South Asia and South Asia benefits from little outliers. Hence, it makes sense that the B-Spline has opted to leave it as a linear model for that region. We thus conclude that it will be of interest to model and compare a regression model using B-spline for the log of `u5mr` to estimate `scaled_nmr`.

In this second model, we consider the non linear affect that U5MR has on NMR

1.12 Calculating DF parameter for B-Spline Function

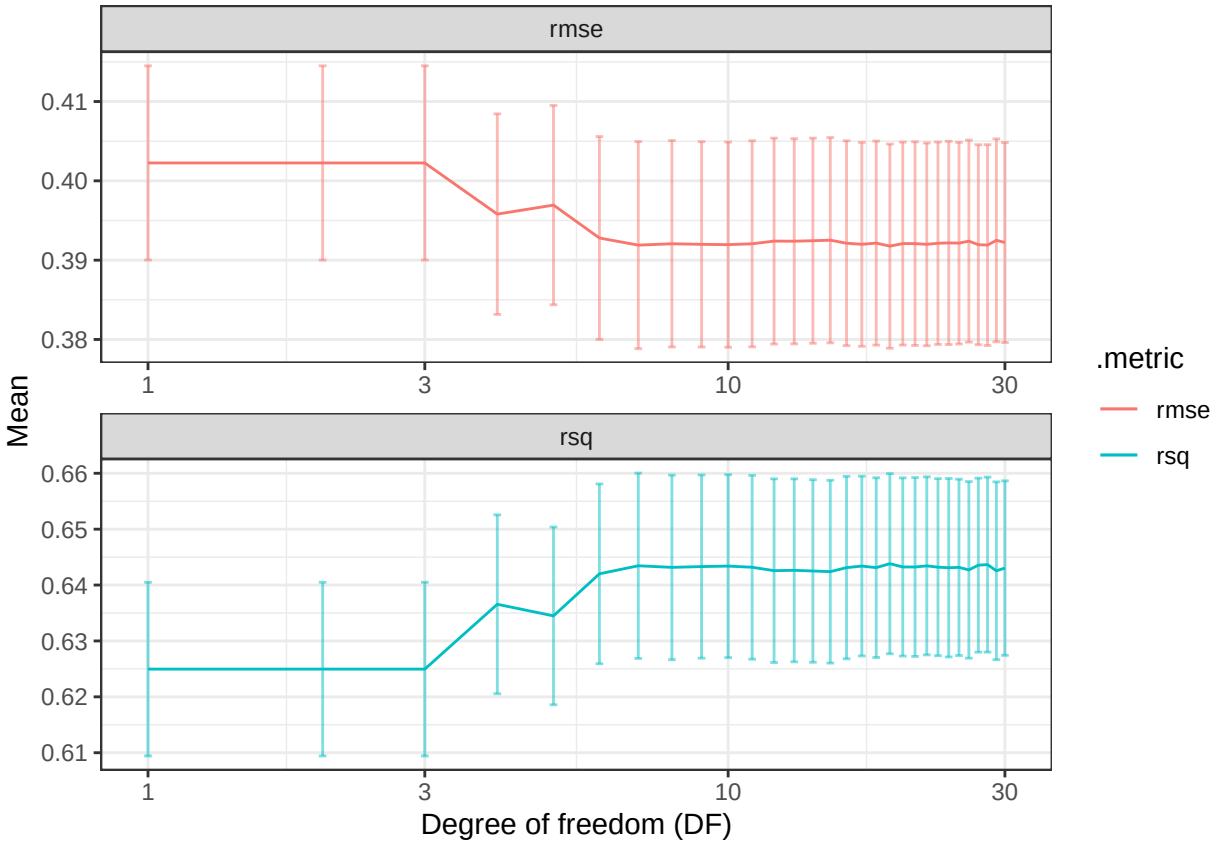
```
dat_rec <- recipe(scaled_nmr ~ year + u5mr + region, data = dat_train) %>%
  step_log(u5mr) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ u5mr*starts_with("region")) %>%
  step_bs(u5mr, deg_free = tune()) # using B-spline for the (log of) u5mr

lm_spec <- linear_reg() %>%
  set_engine("lm")

dat_wf <- workflow() %>%
  add_recipe(dat_rec) %>%
  add_model(lm_spec)

# grid for which we will test out by CV which flexibility (1=<DF=<30) is best
df_grid <- grid_regular(deg_free(range = c(1,30)), levels = 30)

# 10-fold CV
folds <- vfold_cv(dat, v = 10, strata = region)

# tuning our model by 10-fold CV going through the grid
fit_tuned <- dat_wf %>%
  tune_grid(resamples = folds,
            grid = df_grid)

# choosing the best DF by selecting the lowest RMSE
lowest_rmse <- fit_tuned %>% select_best("rmse")

# finalizing the workflow on this chosen DF
final <- finalize_workflow(dat_wf, lowest_rmse)

fit_bs <- final %>% fit(dat) %>% extract_fit_parsnip() # our regression
dat_bs <- fit_bs$fit %>% augment()
dat_bs <- cbind(dat %>% select(-year),
                dat_bs) # binding back parts of the data that was removed
```

1.13 Choosing the DF for our B-spline for log_u5mr covariate.

We have chosen the appropriate number of basis functions for our B-spline regression model through tuning our model through a grid of values (for the degrees of freedom it can test which has the lowest RMSE. The best one will minimise the RMSE and R^2.

```
fit_tuned %>% collect_metrics() %>%
  ggplot(aes(x = deg_free, y = mean,
             colour = .metric)) +
  geom_errorbar(aes(ymin = mean - std_err,
                    ymax = mean + std_err),
                alpha = 0.5) +
  geom_line() +
  facet_wrap(~.metric, scales = "free", nrow = 2) +
  scale_x_log10() +
  theme(legend.position = "none") +
  labs(x = "Degree of freedom (DF)",
```

```
       y = "Mean")+theme_bw()
```



1.14 Fitting the Models considering the non-linear effect and comparing them

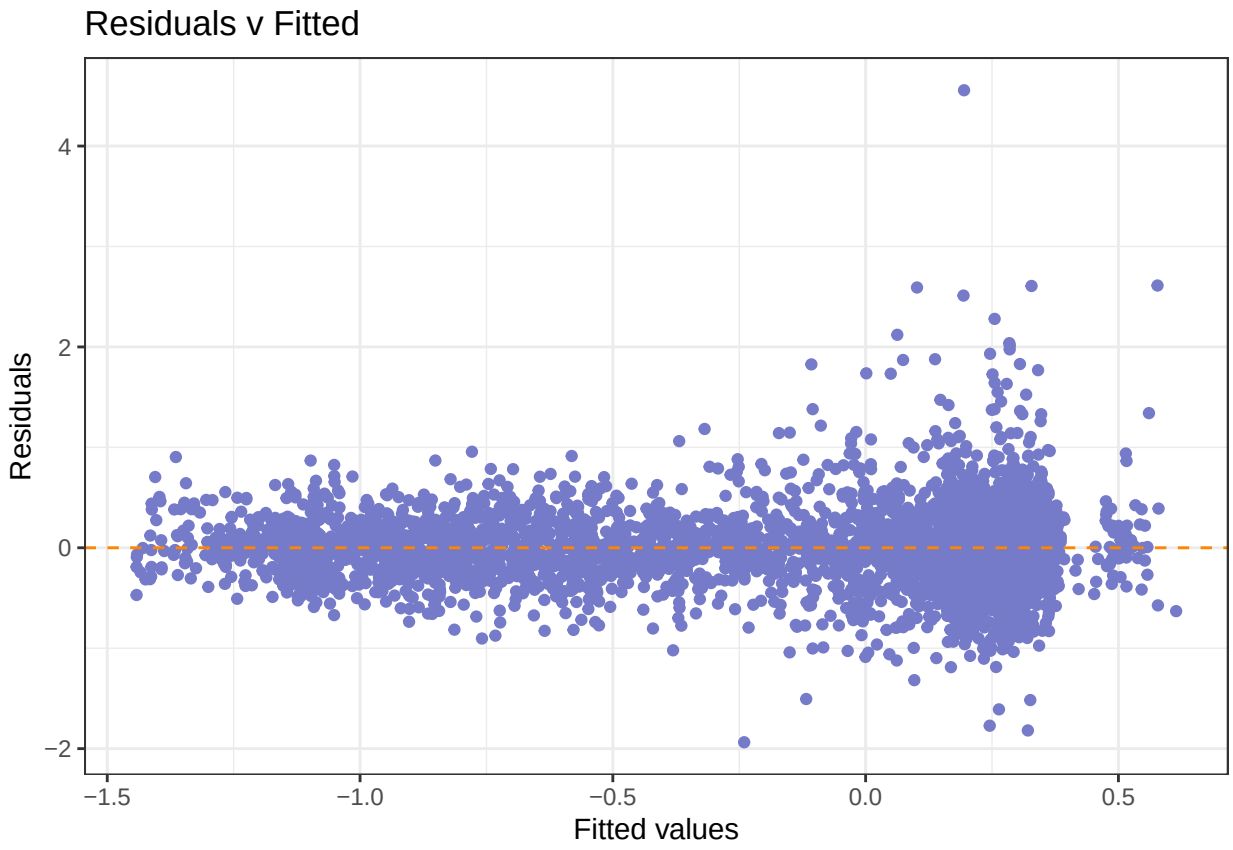
```
fit4 <- lm(scaled_nmr ~ year + log_u5mr*region,
           data = dat)
fit5 <- lm(scaled_nmr ~ year + bs(log_u5mr, df = lowest_rmse$deg_free)*region,
           data = dat)
anova(fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: scaled_nmr ~ year + log_u5mr * region
## Model 2: scaled_nmr ~ year + bs(log_u5mr, df = lowest_rmse$deg_free) *
##     region
##   Res.Df RSS  Df Sum of Sq    F Pr(>F)
## 1   4356 740
## 2   4254 626 102       114 7.58 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

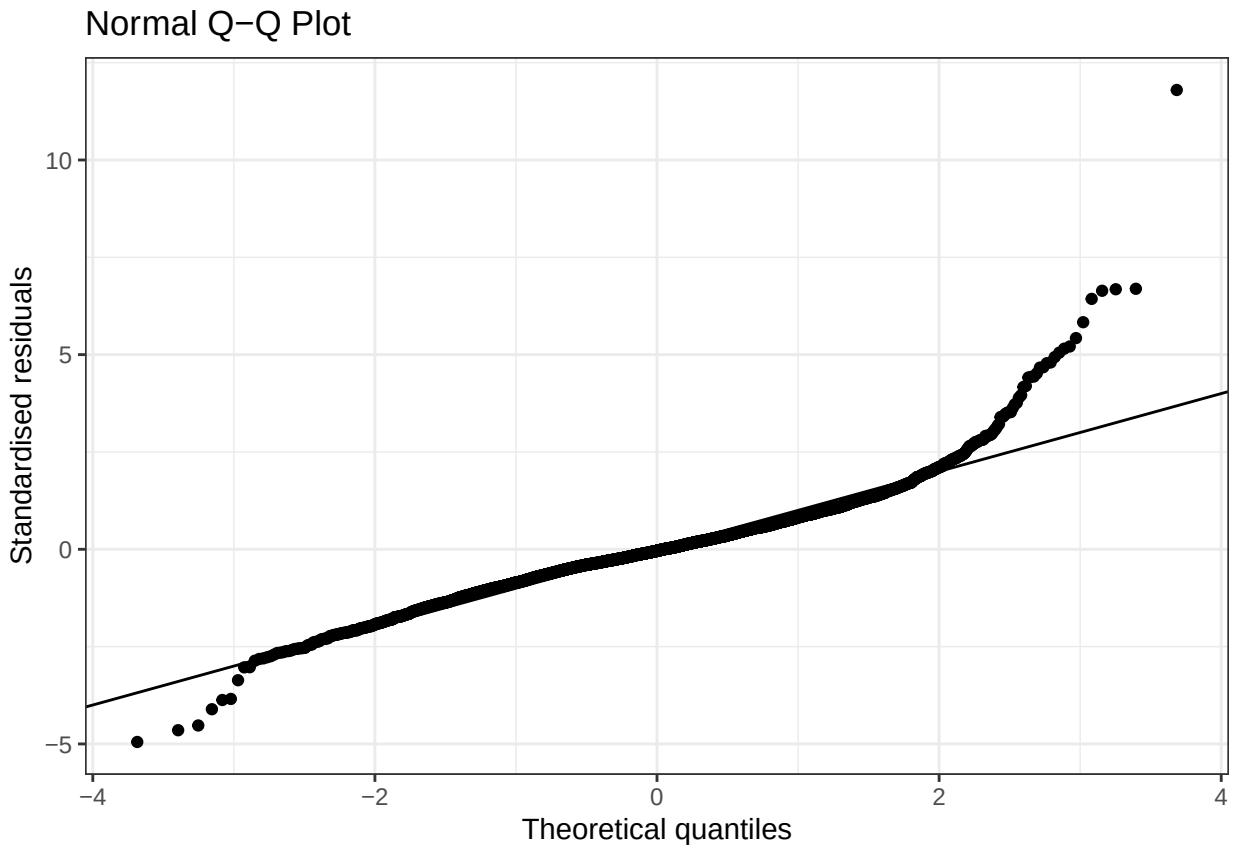
Doing an analysis of variance, it can be seen that there is reasonable evidence to conclude that there is an improvement using the B-spline for the log `u5mr` in our model, at the 0.001 level of significance.

1.15 New fit for all data

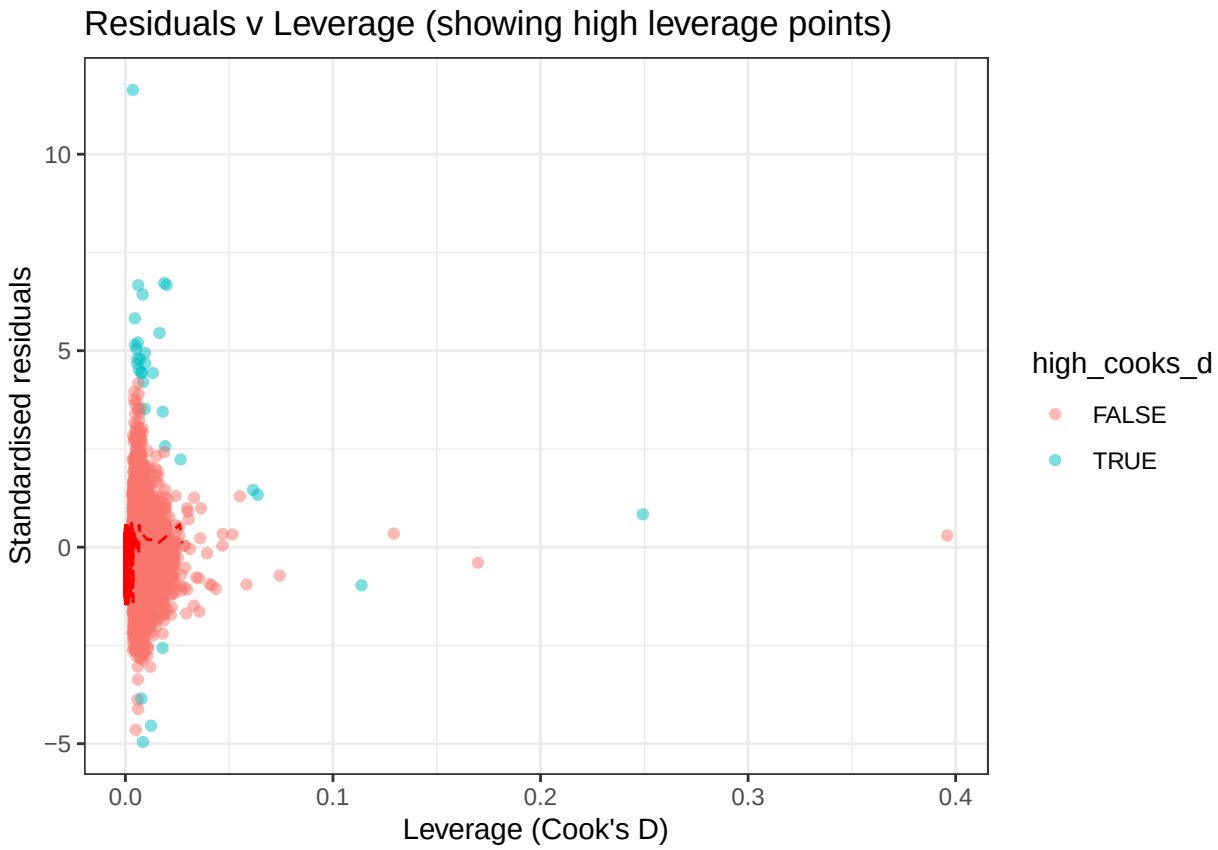
```
dat_bs %>% ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(color = "#757bc8") +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "#ff8400") +
  labs(x = "Fitted values",
       y = "Residuals",
       title = "Residuals v Fitted")+theme_bw()
```




```
dat_bs %>% ggplot(aes(sample = .resid / .sigma)) +
  geom_qq() +
  geom_abline(intercept = 0, slope = 1) +
  labs(x = "Theoretical quantiles",
       y = "Standardised residuals",
       title = "Normal Q-Q Plot")+theme_bw()
```



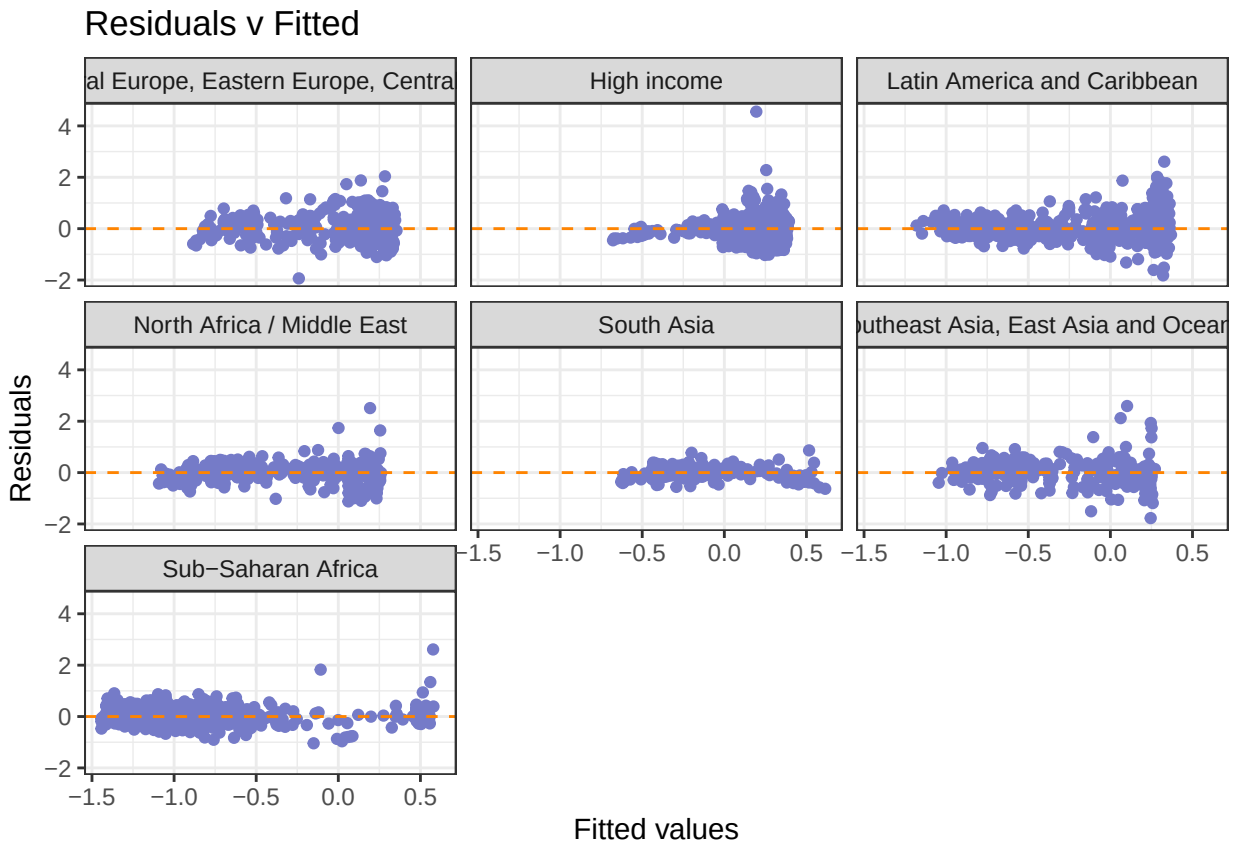

```
p <- 14
n <- length(dat$nmr)
threshold <- (p + 1) / n

dat_bs %>%
  mutate(high_cooks_d = .cooksd > threshold) %>%
  ggplot(aes(x = .hat, y = .std.resid,
             colour = high_cooks_d)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(x = .cooksd, y = .fitted), col = "red", lty = "dashed") +
  labs(x = "Leverage (Cook's D)",
       y = "Standardised residuals",
       title = "Residuals v Leverage (showing high leverage points)")+theme_bw()
```



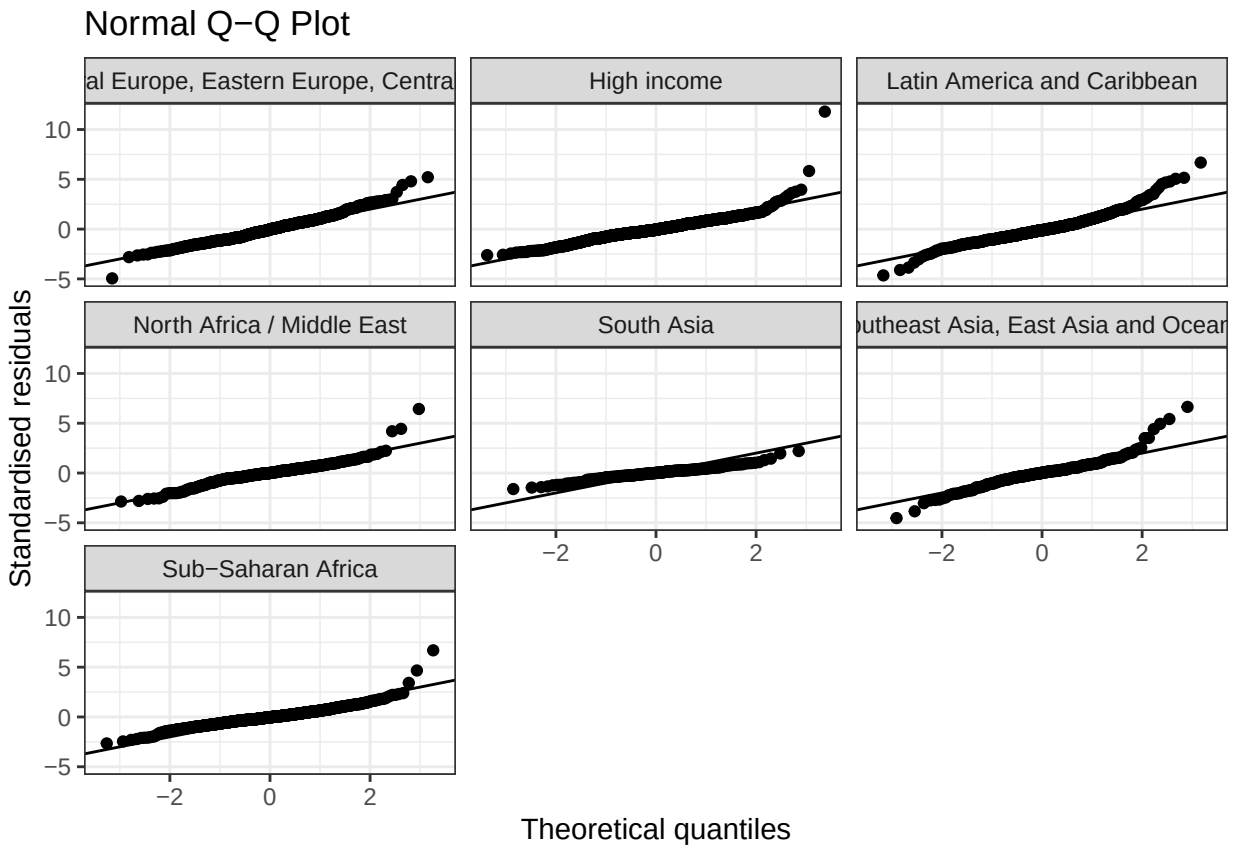
In terms of the fitted vs residual plot, there is prima facie not many signs that there has been a significant improvement. However, near the 0.5 value, the residuals are now more spread out in each side of the zero line, something that was noted to be lacking in the linear model. This is a favourable sign of improvement in the fit of our model. Moreover, looking carefully, more points lie in both sides of the zero line in equal amounts for the entirely of this plot than before. Hence, this suggests that the variances is more constant and independent for this model that the old one. The normal q-q plot, however, does appear to be the same. So, this suggests that the estimating accuracy of this model is not that different to the old linear model. There is also less outlier points under this model. Hence, this model is not severely effected as before by different data points, improving the fit to the data.

1.16 New fit for all data by region

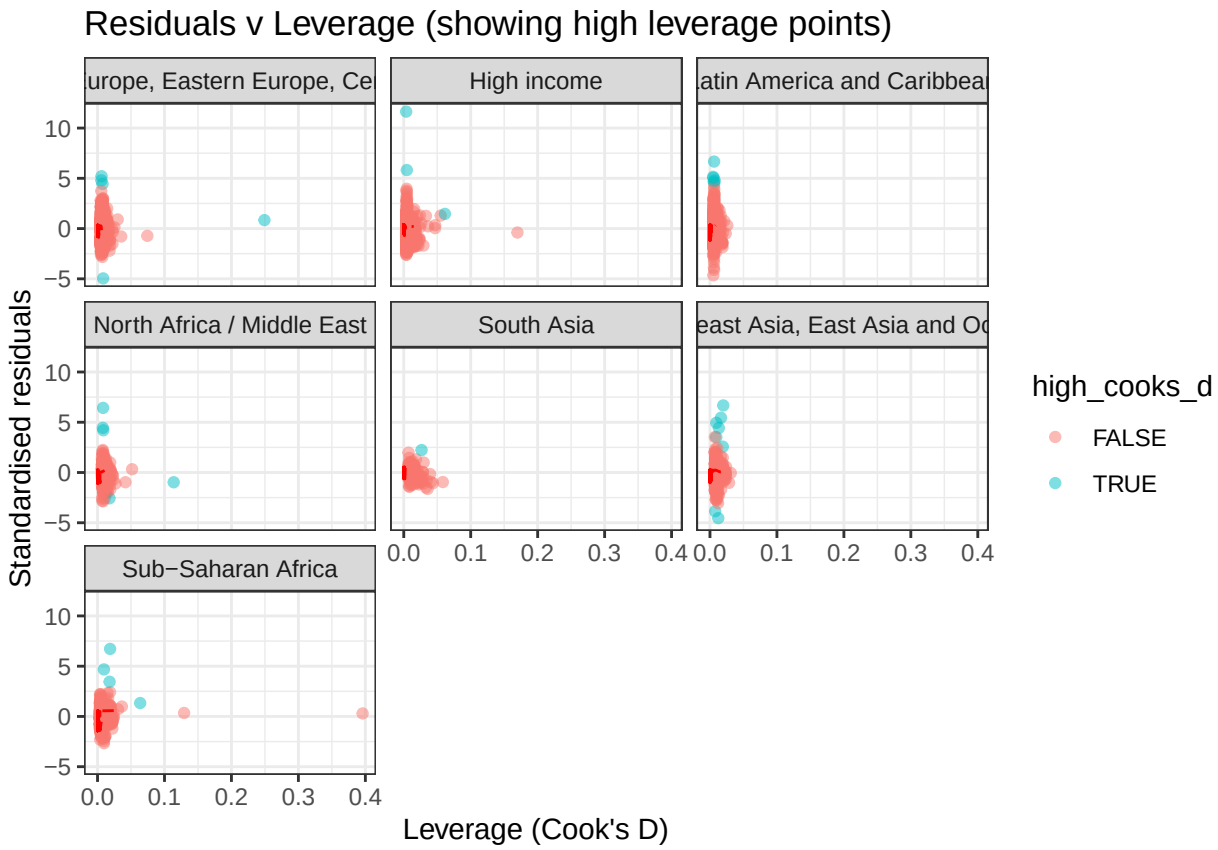
```
dat_bs %>% ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(color = "#757bc8") +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "#ff8400") +
  facet_wrap(~region) +
  labs(x = "Fitted values",
       y = "Residuals",
       title = "Residuals v Fitted")+theme_bw()
```




```
dat_bs %>% ggplot(aes(sample = .resid / .sigma)) +
  geom_qq() +
  geom_abline(intercept = 0, slope = 1) +
  facet_wrap(~region) +
  labs(x = "Theoretical quantiles",
       y = "Standardised residuals",
       title = "Normal Q-Q Plot")+theme_bw()
```




```
dat_bs %>%
  mutate(high_cooks_d = .cooksd > threshold) %>%
  ggplot(aes(x = .hat, y = .std.resid,
             colour = high_cooks_d)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(x = .cooksd, y = .fitted), col = "red", lty = "dashed") +
  facet_wrap(~region) +
  labs(x = "Leverage (Cook's D)",
       y = "Standardised residuals",
       title = "Residuals v Leverage (showing high leverage points)")+theme_bw()
```

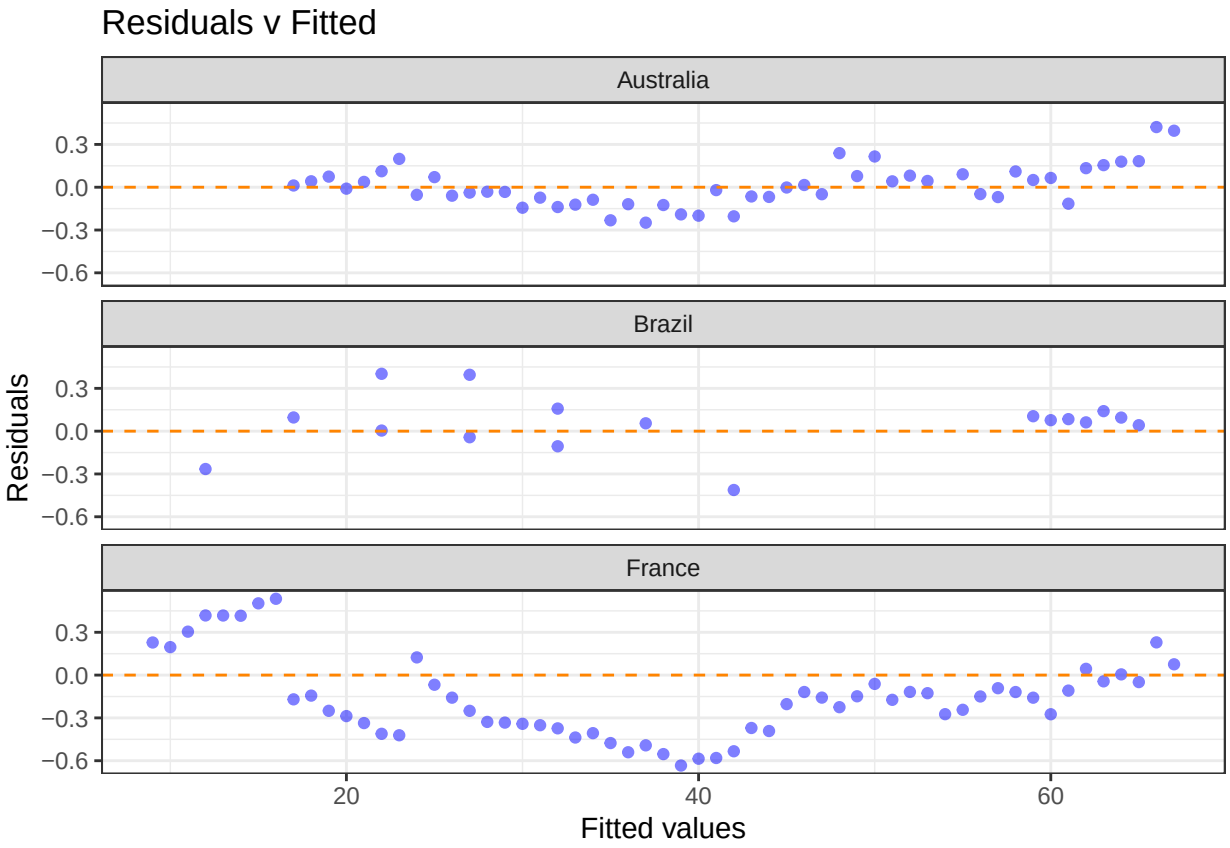


By region, we can see again that the fit has improved. For high income countries and Euarasia, the data is very more spread out across the zero line than it was under the linear model. Again, the q-q plot suggests that the predicting accuracy will not be significantly different. For the number of outliers, this has reduced for some regions, but across the board, the diagnostic plot shows that it has stayed relatively the same. All in all, these plots suggests that there is indeed benefit to using a non-linear model for estiamting the average NMR.

1.17 New fit for the data by the selected countries

```
dat_bs_three <- dat_bs %>%
  filter(country_name == "Australia" |
           country_name == "Brazil" |
           country_name == "France")

dat_bs_three %>%
  ggplot(aes(x = year, y = .resid)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "#Ff8400") +
  facet_wrap(~country_name, nrow = 3) +
  labs(x = "Fitted values",
       y = "Residuals",
       title = "Residuals v Fitted")+theme_bw()
```

.metric	.estimate
rmse	0.43
rsq	0.60
mae	0.29



Plotting for our chosen three countries, there is still a pattern for the residuals vs fitted, however, the data is now more spread across the zero line. As discussed before, this pattern is expected for each country will be different to its regional and the global average in child mortality data. So, we sacrifice the ability to plot any abitrary country for the ability to predict on the regional and global scale NMR. Likewise, this note rings true to the q-q plot which shows that at the tails the data points differing the theoretical normal quantiles (suggesting a lack of fit). A concern for using B-spline is that, however contrary to the data, the predicted leverage (as per Cook's D) is vastly different for these countries. However, such concerns are slightly alleviated by the fit by region, for the plot fares well in those.

1.18 Validation set approach to estimating the test RMSE & MAE

```
fit_bs_train <- final %>% fit(dat_train)
dat_bs_valid <- dat_valid %>%
  bind_cols(predict(fit_bs_train, dat_valid))

dat_bs_valid %>% metrics(truth = scaled_nmr, estimate = .pred) %>%
  select(-`.estimator`)%>%
  kbl(booktabs = T) %>%
  kable_styling(position = "center")
```

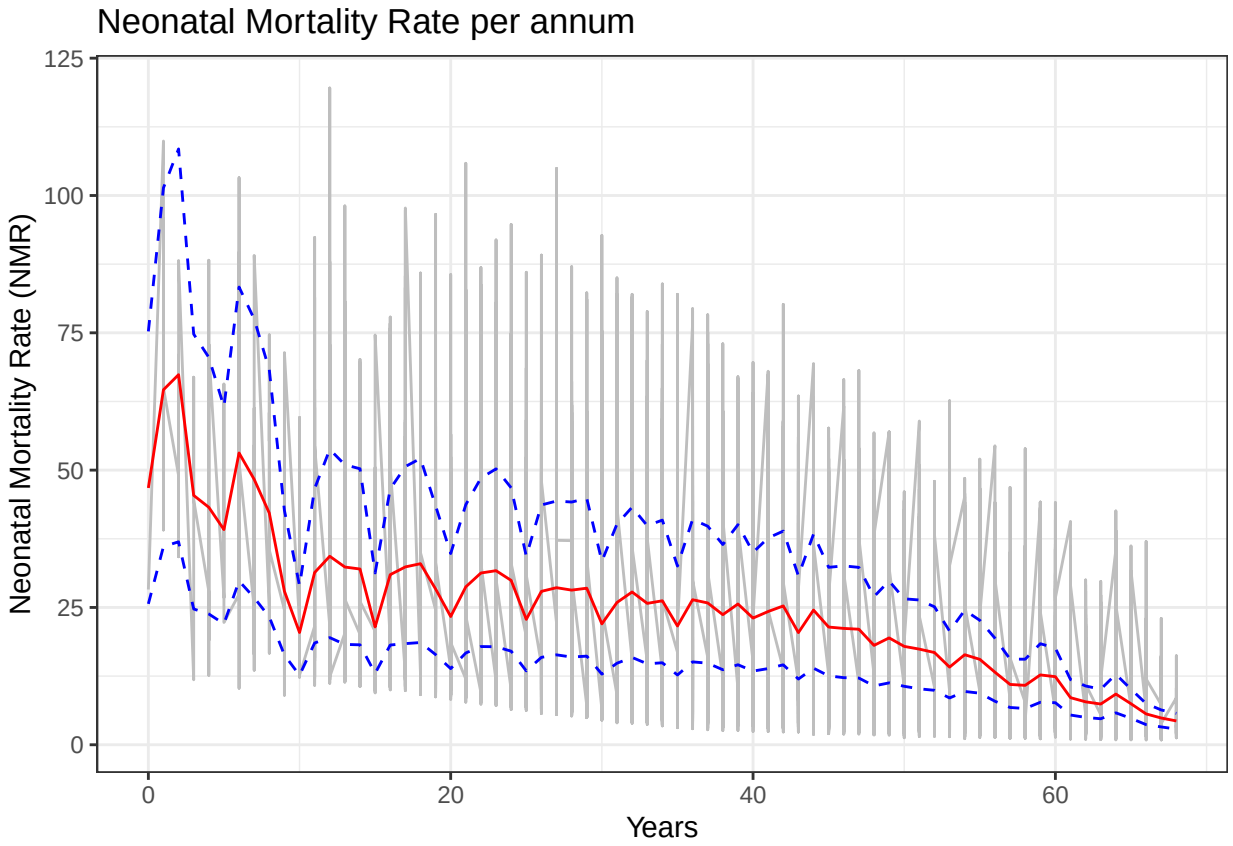
Using the validation set approach, splitting our data again in 75-25 per cent (by region) as per §2.4, and running the model (fitted to the training set) on the validation set, we see that the estimated test RMSE and MAE is approximately the same as under the linear model. This resulted is not suprising for the q-q plots aforementioned suggested that for predicting, our model will not be significally different.

1.19 Prediction CLT intervals

We test here how well our model can predict the data compared to the old model.

1.20 Prediction CLT intervals for whole data

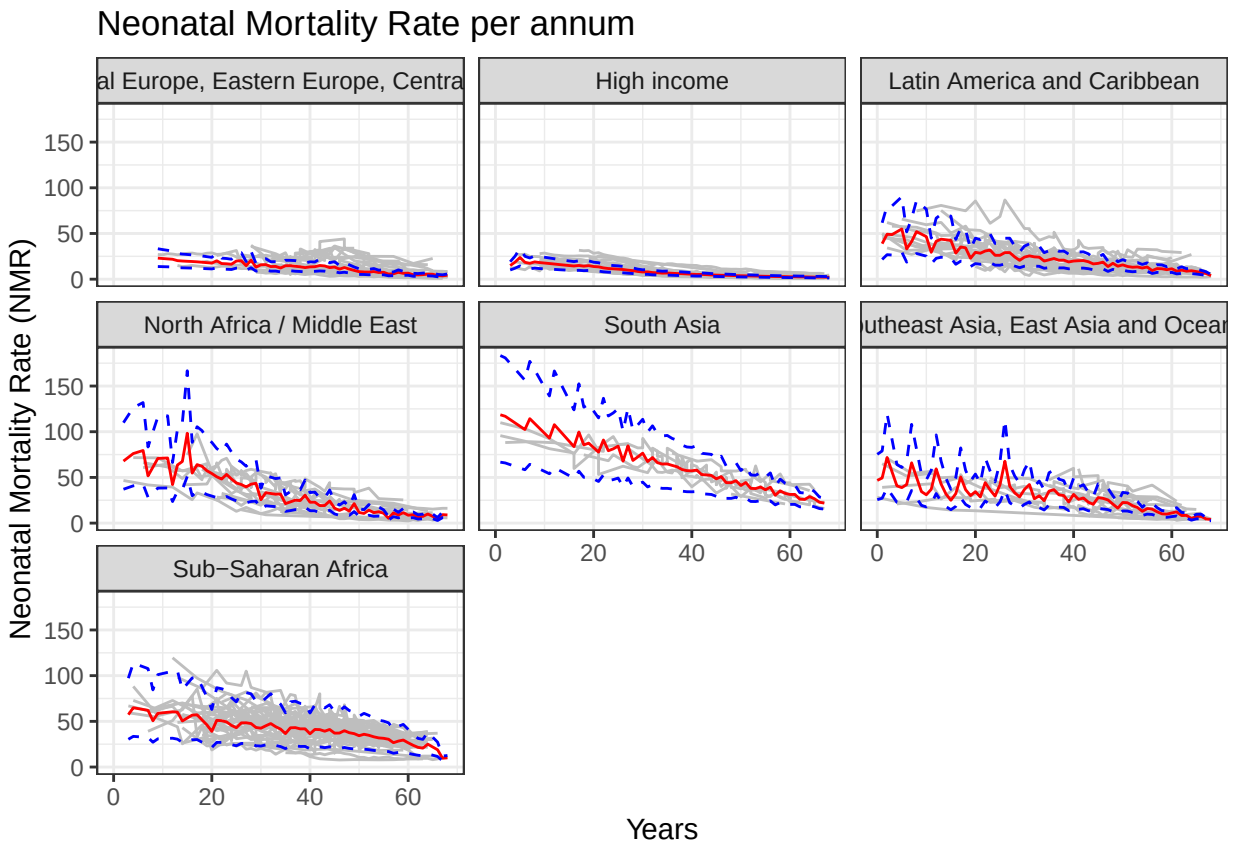
```
pred_nmr <- predict(fit_bs$fit, dat_bs, interval = "prediction") %>%
  as.data.frame()

pred_nmr <- bind_cols(dat, pred_nmr) %>% mutate(
  fitted = exp(fit) * u5mr / (1 + exp(fit)),
  upr_ci = exp(upr) * u5mr / (1 + exp(upr)),
  lwr_ci = exp(lwr) * u5mr / (1 + exp(lwr))
  ) %>% select(-fit, -upr, -lwr)

pred_nmr_all <- pred_nmr %>% group_by(year) %>%
  summarise(fit = mean(fitted),
            lwr = mean(lwr_ci),
            upr = mean(upr_ci),
            nmr = nmr)

pred_nmr_all %>% ggplot(aes(x = year)) +
  geom_line(aes(y = nmr), colour = "grey") +
  geom_line(aes(y = fit), colour = "red") +
  geom_line(aes(y = lwr), colour = "blue", linetype = "dashed") +
  geom_line(aes(y = upr), colour = "blue", linetype = "dashed") +
    labs(
      x = "Years",
      y = "Neonatal Mortality Rate (NMR)",
      title = "Neonatal Mortality Rate per annum")+theme_bw()
```



According to the plot below, our predicted values and the 95 per cent prediction CLT interval is very similar to the one produced by the linear model. This is again suggested by the q-q plots before. Hence, for estimating our data, if by global average for NMR, then either models will do.

1.21 Prediction CLT intervals by region

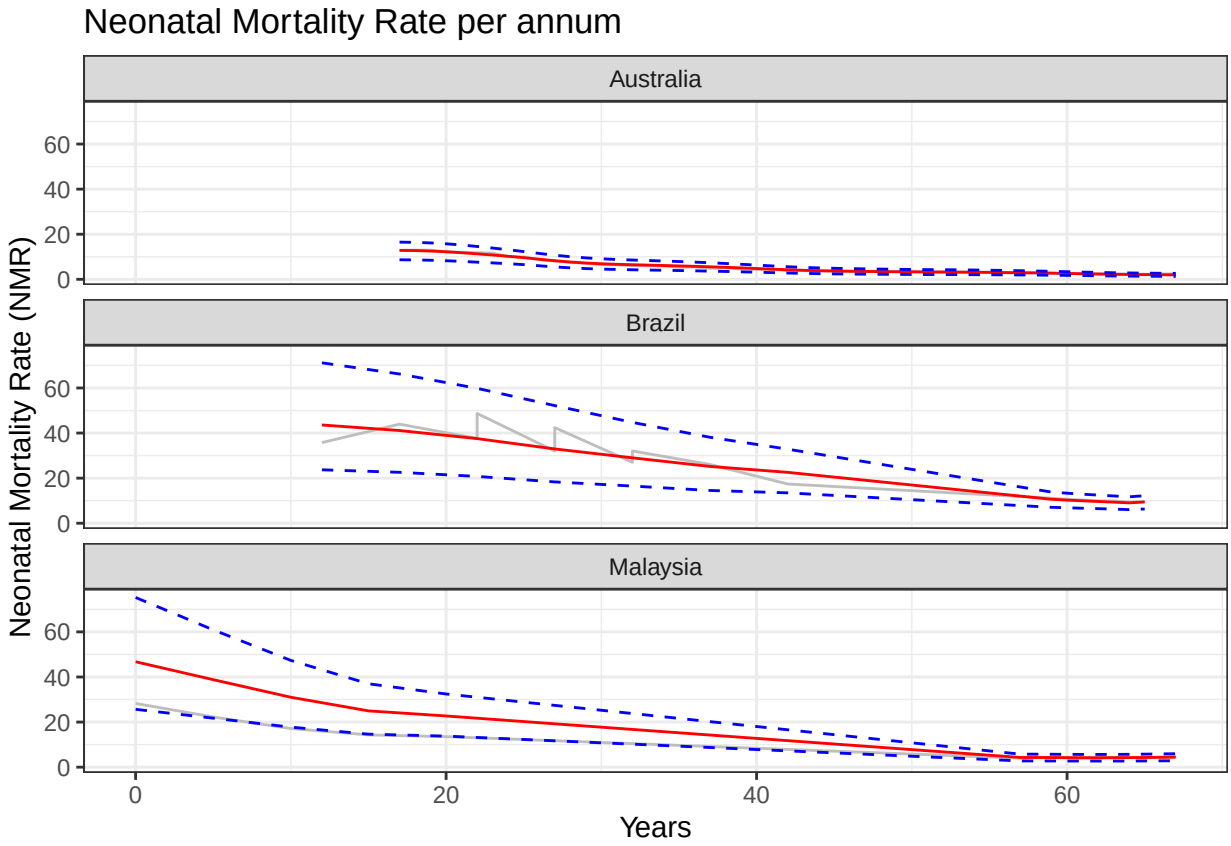
```
pred_nmr_region <- pred_nmr %>% group_by(year, region) %>%
  summarise(fit = mean(fitted),
            lwr = mean(lwr_ci),
            upr = mean(upr_ci))

pred_nmr %>% ggplot(aes(x = year)) +
  geom_line(aes(y = nmr, group = country_name),
            colour = "grey") +
  geom_line(data = pred_nmr_region, aes(y = fit),
            colour = "red") +
  geom_line(data = pred_nmr_region, aes(y = lwr),
            colour = "blue",
            linetype = "dashed") +
  geom_line(data = pred_nmr_region, aes(y = upr),
            colour = "blue",
            linetype = "dashed") +
  facet_wrap(~region, nrow = 3) +
  labs(
      x = "Years",
      y = "Neonatal Mortality Rate (NMR)",
      title = "Neonatal Mortality Rate per annum")+theme_bw()
```



By region, this stays the same, that the prediction value and the uncertainty in that (by the blue prediction interval) is very similar to the linear model one.

1.22 Prediction CLT intervals by three countries

```
pred_nmr_three <- pred_nmr %>% filter(country_name %in% countries)


pred_nmr_three <- pred_nmr_three %>% group_by(year, country_name) %>%
  summarise(fit = mean(fitted),
            lwr = mean(lwr_ci),
            upr = mean(upr_ci),
            nmr = nmr)

pred_nmr_three %>% ggplot(aes(x = year)) +
  geom_line(aes(y = nmr), colour = "grey") +
  geom_line(aes(y = fit), colour = "red") +
  geom_line(aes(y = lwr), colour = "blue", linetype = "dashed") +
  geom_line(aes(y = upr), colour = "blue", linetype = "dashed") +
  facet_wrap(~country_name, nrow = 3) +
  labs(
      x = "Years",
      y = "Neonatal Mortality Rate (NMR)",
      title = "Neonatal Mortality Rate per annum")+theme_bw()
```



Even by the three countries, this looks the same still. Hence, we conclude that for estimating, either model will do, though the B-spline model leads to lesser residuals for some regions and thus is still favourable in that regard.

Summary of Findings & Conclusion

We have modelled the covariates of `region`, the log of `u5mr`, and `year` to the `scaled_nmr`, with an interaction effect for `u5mr` at every `region`. Such a model we found to be the best. It was also found that, in terms of outputted residuals, that having a B-Spline for `u5mr` was favourable for there was some non-linearity in this variable against `u5mr`. However, overall, the linear model and the non-linear one was very similar to each other in predicting the average NMR. We also saw that there was a downward trend for all countries in their NMR, and this was evidenced to be due to the changing composition of mortality that we wanted to test. Further exploration of some of this is, however, needed for each region, as some regions were different in their pace and value of NMR to others. Our models suffered in the lack of fit across differing regions, notably those in high income countries and in Eurasia. In conclusion, for estimating the average NMR, although the B-Spline model showed clear improvement in its analysis of a variance and evidenced in its fitted vs residual plot, there is little other evidence to suggest that we should favour it for linear model in prediction.