Statistical Thinking Assignment 2

Xinyi Cui, Janice Hsin Hsu, Pranali Angne, Sahinya Akila

10/17/2021

```r
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
```

```r
# Loading Libraries

library(printr)
library(tidyverse)
library(tidymodels)
library(broom)
library(splines)
library(dagitty)
library(ggdag)
library(knitr)
library(gtsummary)
library(kableExtra)
```

Task 1: Estimating Neonatal Mortality

Introduction

One of this century's global goals has been the reduction of childhood mortality across all countries. There has been enormous effort put into this goal at all levels from the united nations down to local interventions. The aim of this report is to produce a linear regression model to estimate the average neonatal mortality rate (NMR).

Data

The source of the child mortality data is from the UN Inter-agency Group for Child Mortality Estimation.

It contains the following columns:

- `country_name`: Name of the country
- `year`: The year the data was measured
- `region`: The name of the continent the country is from
- `nmr`: The observed number of neonatal deaths per thousand live births (the neonatal mortality rate). This is measured either using a country's vital registration system (births and deaths register) or using some sort of high-quality survey.
- `u5mr`: The estimated under-five mortality rate
- `nmr_transformed`: log of the number of neonatal deaths per 1000 live births divided by the number of non-neonantal deaths per 1000 live births.

$$log(\frac{nmr}{u5mr - nmr})$$

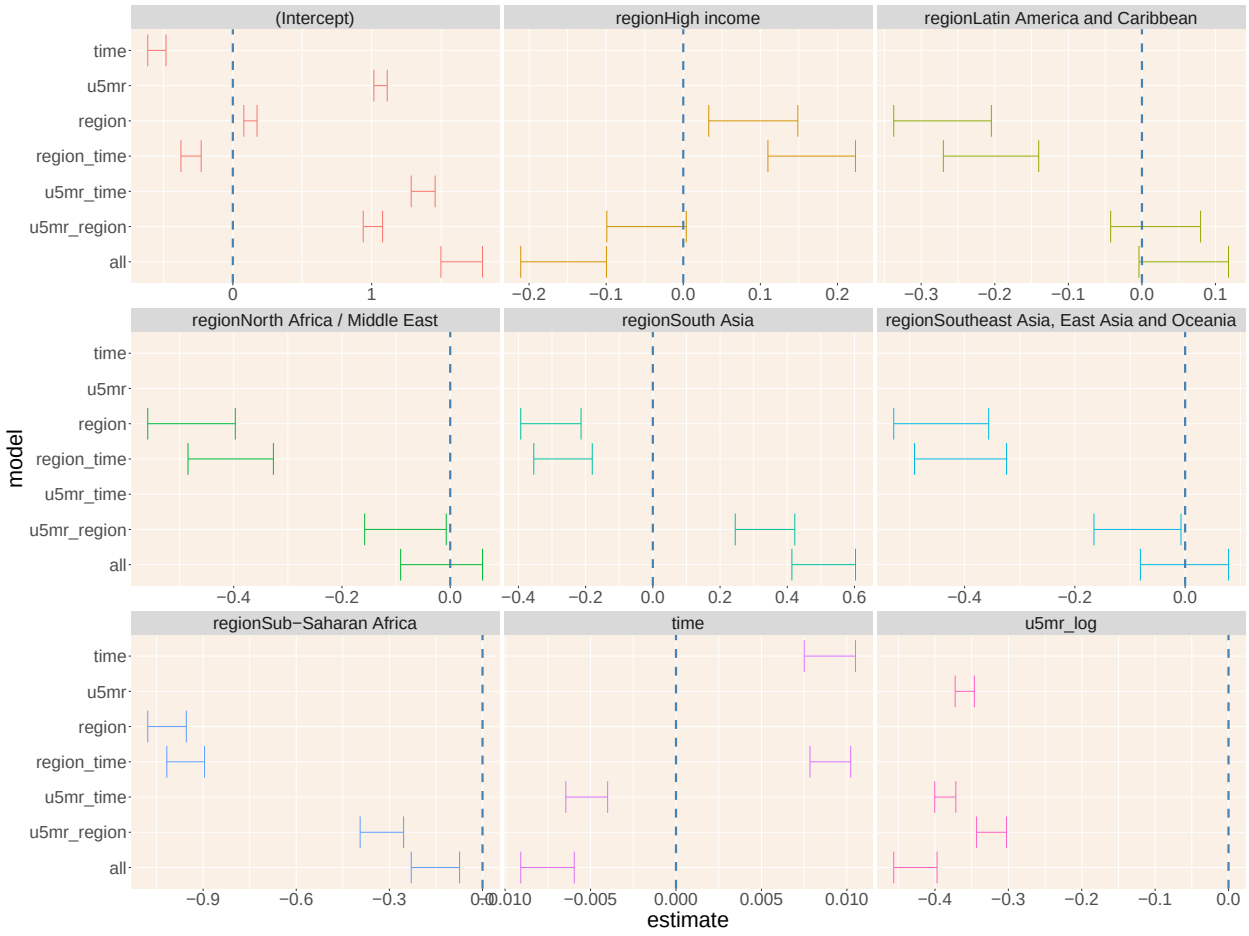
```r
# Reading the data
neonatal_mortality <- read_csv("neonatal_mortality.csv")

# Adding log ratio between neonatal mortality and non-neonatal mortality rate and time
nmr_data <- neonatal_mortality %>%
  mutate(u5mr_log = log(u5mr),
         transformed_nmr = log(nmr/(u5mr-nmr)),
         time = year - min(year))
```

Task 1.1: Linear Regression

Task 1.1.1: Explain the choice of variables in your model (you should not use country_name, if you use U5MR, you should use it on the log-scale!). In particular you should consider whether an interaction effect should be used.

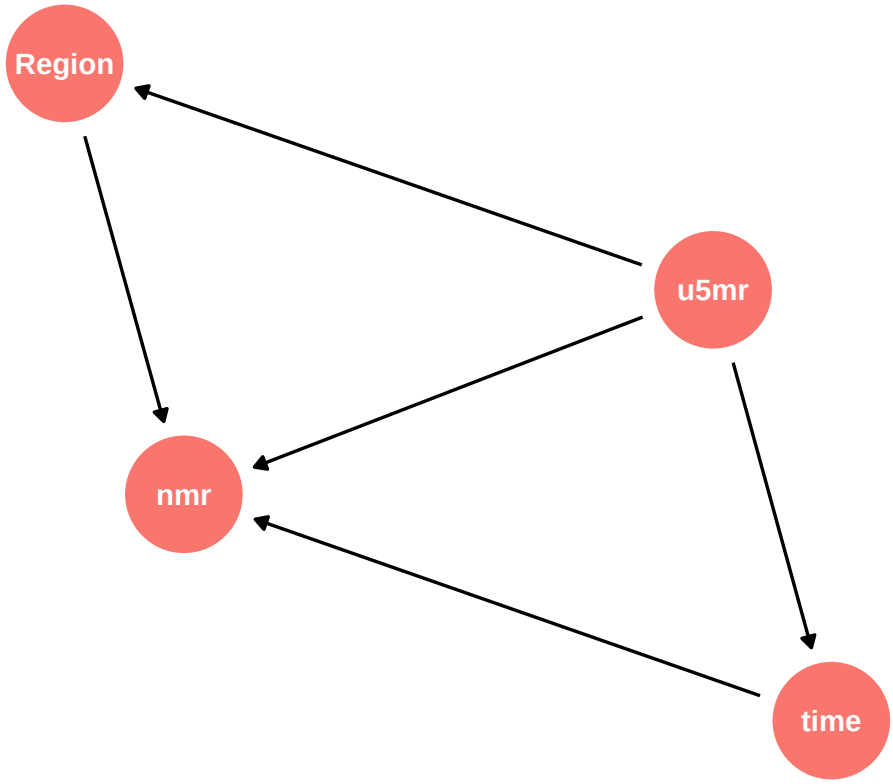
```
# Splitting the data
nmr_split <- initial_split(nmr_data, strata = region)
nmr_train <- training(nmr_split)
nmr_test <- testing(nmr_split)

# Choosing the variables by trying different combinations of variables
fit1 <- lm(transformed_nmr ~ u5mr_log + region + time, nmr_train) %>%
  tidy() %>%
  mutate(model = "all")

fit2 <- lm(transformed_nmr ~ u5mr_log + region, nmr_train) %>%
  tidy() %>%
  mutate(model = "u5mr_region")

fit3 <- lm(transformed_nmr ~ u5mr_log + time, nmr_train) %>%
  tidy() %>%
  mutate(model = "u5mr_time")

fit4 <- lm(transformed_nmr ~ region + time, nmr_train) %>%
  tidy() %>%
  mutate(model = "region_time")

fit5 <- lm(transformed_nmr ~ region, nmr_train) %>%
  tidy() %>%
  mutate(model = "region")

fit6 <- lm(transformed_nmr ~ u5mr_log, nmr_train) %>%
  tidy() %>%
  mutate(model = "u5mr")

fit7 <- lm(transformed_nmr ~ time, nmr_train) %>%
  tidy() %>%
  mutate(model = "time")

# joining all the data frames and calculating the upper and lower values
full_model <- list(fit1, fit2, fit3, fit4, fit5, fit6, fit7) %>%
  reduce(full_join) %>%
  mutate(upper = estimate + 1.96 * std.error,
         lower = estimate - 1.96 * std.error)

full_model %>% ggplot(aes(estimate, y = model)) +
  geom_errorbar(aes(xmin = lower, xmax = upper, color = term)) +
  geom_vline(aes(xintercept = 0),
             linetype = "dashed",
             size = 1.2,
             color = "steel blue") +
  facet_wrap(.~term, scale = "free_x") +
```

```
  scale_y_discrete(limits = c("all", "u5mr_region", "u5mr_time", "region_time", "region", "u5mr", "time
  theme(panel.background = element_rect(fill = "linen"), legend.position = "none", text = element_text(s
```



We have fitted seven models in total and the error bar graph represents the significance of each variable. The time variable is considered to be significant among all models under the 95% confidence interval as both the maximum and the minimum are different from 0. Similarly, variables such as region, and u5mr can also be considered as significant. However, while looking at the region_time model which represents the interaction effect between region and time, we can see that the shifts from the region model are only significant for High income and Latin American region; and therefore, the interaction effect should not be used here. Simultaneously, the u5mr_time model and the u5mr_region model both show insignificance as they are not too different from 0. To sum up, u5mr affects region while no effect for the opposite, and u5mr is significant among all the models and its' significance levels are not changed if adding the time interaction on.

```
# understanding relationship between the variables
dagify(nmr ~ u5mr,
       nmr ~ time,
       nmr ~ Region,
       time ~ u5mr,
       Region ~ u5mr
       ) %>%
  ggdag_adjust(node_size = 20) +
  theme_dag(legend.position = "none")
```




```
# Fitting the model using the training data (with the interactive effect)
fit <- lm(transformed_nmr ~ u5mr_log + u5mr_log*region + u5mr_log*time, nmr_train)

summary(fit)
```

```
##
## Call:
## lm(formula = transformed_nmr ~ u5mr_log + u5mr_log * region +
##     u5mr_log * time, data = nmr_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0002 -0.2371 -0.0158  0.2106  4.5385
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                     1.6849250  0.1365730
## u5mr_log                                       -0.4406838  0.0379608
## regionHigh income                              -0.4997255  0.0812943
## regionLatin America and Caribbean               0.7827098  0.1114625
## regionNorth Africa / Middle East                0.2410142  0.1314313
## regionSouth Asia                                1.4655496  0.3155026
```

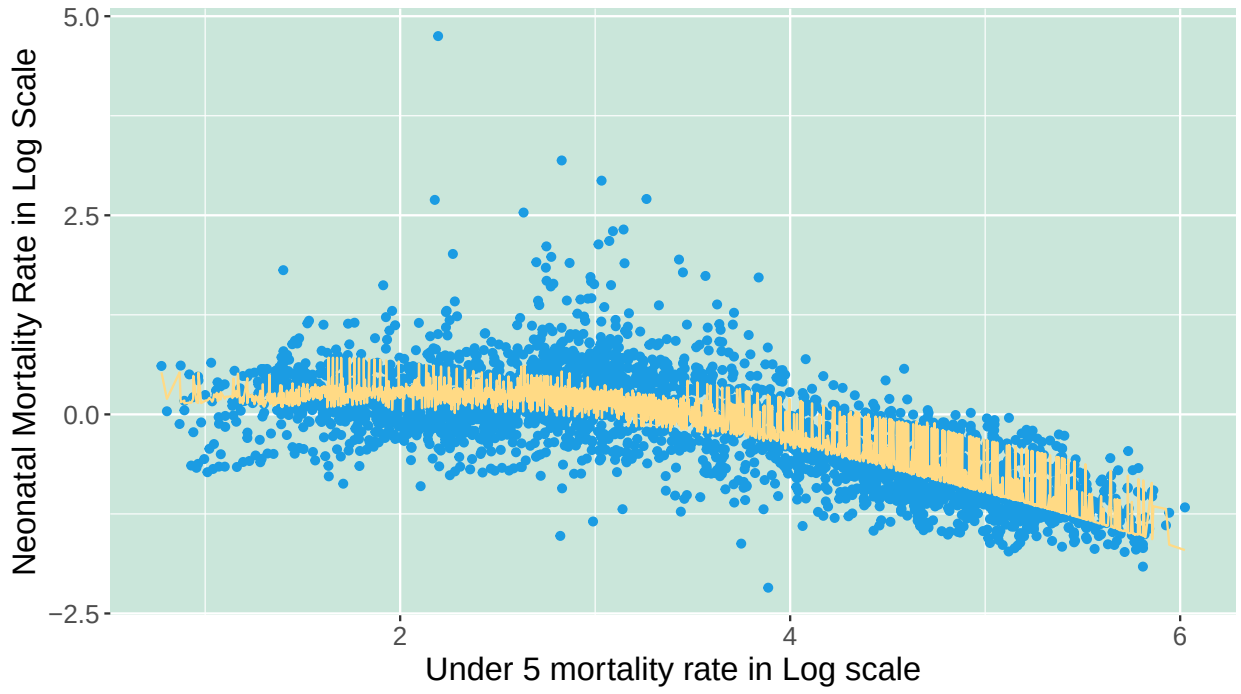
```
## regionSoutheast Asia, East Asia and Oceania           0.5390637  0.1523187
## regionSub-Saharan Africa                              1.3472651  0.1520669
## time                                                 -0.0141736  0.0019642
## u5mr_log:regionHigh income                            0.1713538  0.0307126
## u5mr_log:regionLatin America and Caribbean           -0.2197577  0.0340746
## u5mr_log:regionNorth Africa / Middle East            -0.0876061  0.0371048
## u5mr_log:regionSouth Asia                            -0.2443877  0.0695899
## u5mr_log:regionSoutheast Asia, East Asia and Oceania -0.1665250  0.0423067
## u5mr_log:regionSub-Saharan Africa                    -0.3485632  0.0373663
## u5mr_log:time                                         0.0025870  0.0004901
##                                                      t value Pr(>|t|)
## (Intercept)                                           12.337  < 2e-16 ***
## u5mr_log                                             -11.609  < 2e-16 ***
## regionHigh income                                     -6.147 8.85e-10 ***
## regionLatin America and Caribbean                      7.022 2.65e-12 ***
## regionNorth Africa / Middle East                       1.834 0.066780 .
## regionSouth Asia                                       4.645 3.53e-06 ***
## regionSoutheast Asia, East Asia and Oceania            3.539 0.000407 ***
## regionSub-Saharan Africa                               8.860  < 2e-16 ***
## time                                                  -7.216 6.63e-13 ***
## u5mr_log:regionHigh income                             5.579 2.61e-08 ***
## u5mr_log:regionLatin America and Caribbean            -6.449 1.29e-10 ***
## u5mr_log:regionNorth Africa / Middle East             -2.361 0.018282 *
## u5mr_log:regionSouth Asia                             -3.512 0.000451 ***
## u5mr_log:regionSoutheast Asia, East Asia and Oceania  -3.936 8.45e-05 ***
## u5mr_log:regionSub-Saharan Africa                     -9.328  < 2e-16 ***
## u5mr_log:time                                          5.278 1.39e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4248 on 3261 degrees of freedom
## Multiple R-squared:  0.5956, Adjusted R-squared:  0.5937
## F-statistic: 320.2 on 15 and 3261 DF,  p-value: < 2.2e-16
```

nmr is the dependent variable. From the dag graph we can see that both time and region is piped. We do not want to add them, cause they will end up worth lessed effect which removes the correlation between u5m4 and nmr because we are adding bias through time and region. u5mr is a fork therefore we consider to include it. Moreover, from the error bar we see the indirect affect to u5mr and time, therefore, to consider the interaction effect, we include u5mr*region and u5mr*time.

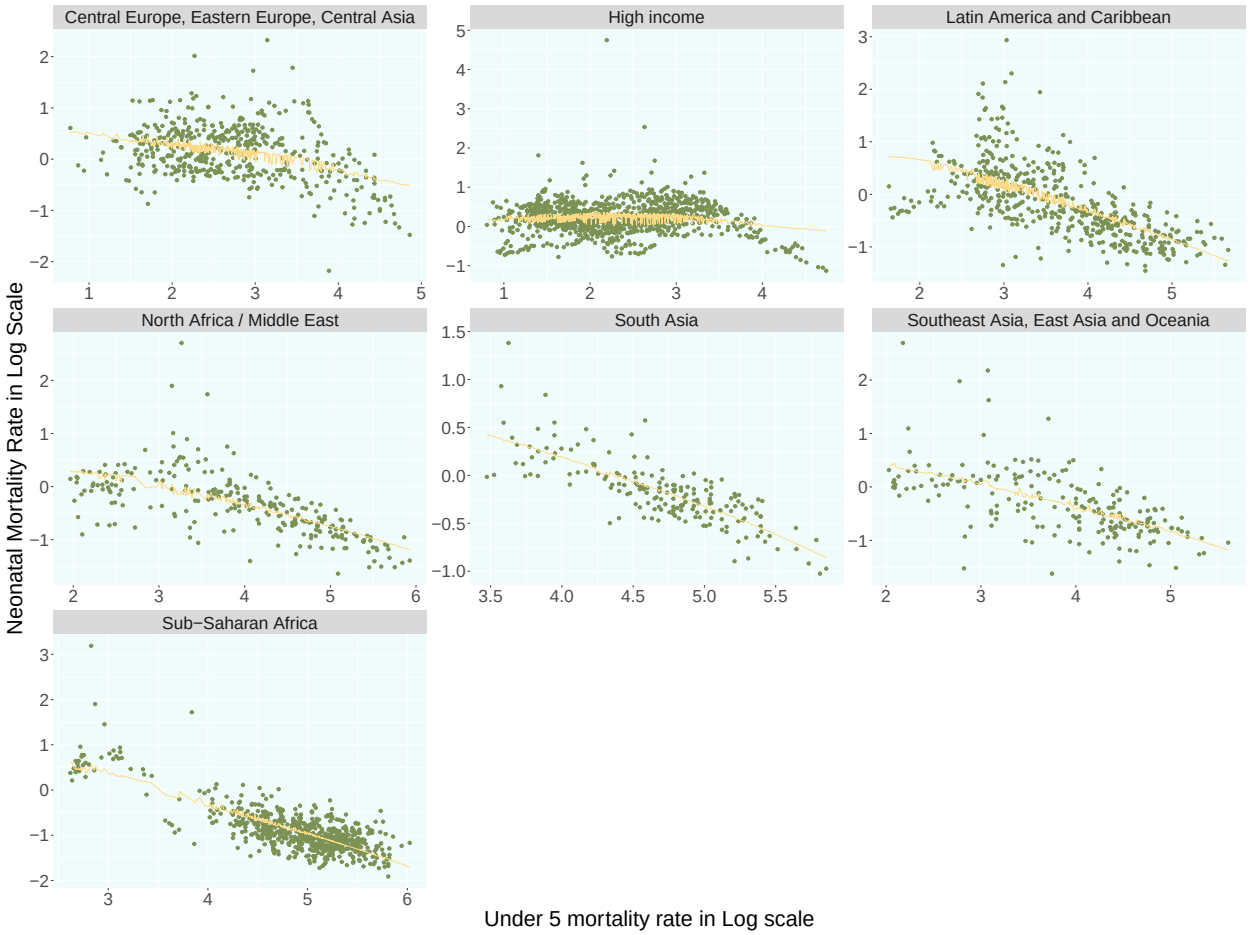
Task 1.1.2: Assess your linear model and comment on its fit. This should be done a) for all data simultaneously; b)for data in each region; and c) for data in a maximum of 3 countries that should be chosen to highlight different aspects of the fit diagnostics.

```r
augment_fit <- fit %>%
  augment(data = nmr_train)
# part a
augment_fit %>%
  ggplot() +
  geom_point(aes(x = u5mr_log, y= transformed_nmr), color = "#1b9ce3")+
  geom_line(aes(x = u5mr_log, y = .fitted), color = "#ffda85") +
  theme(panel.background = element_rect(fill = "#cae6da"), text = element_text(size=15)) +
```

```
  xlab("Under 5 mortality rate in Log scale") +
  ylab("Neonatal Mortality Rate in Log Scale")
```

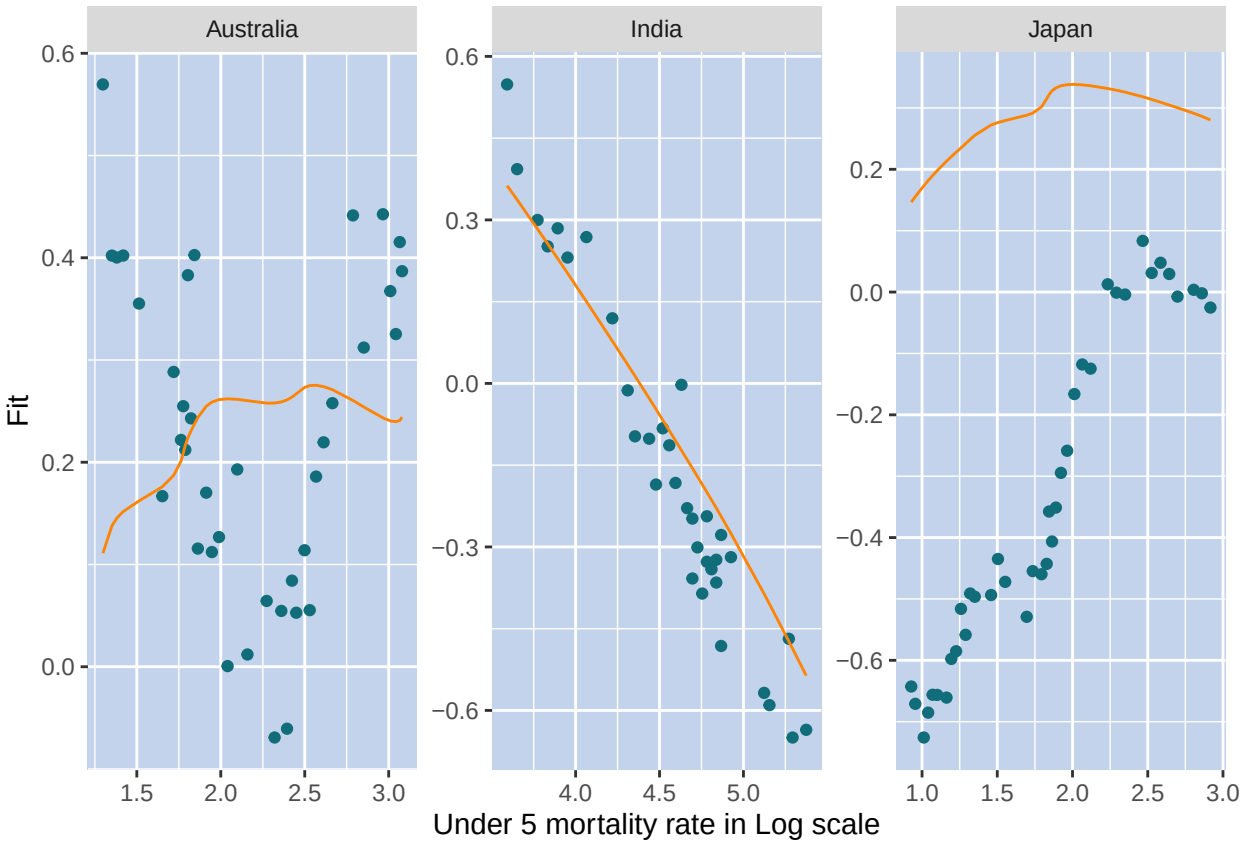



```
# part b
augment_fit %>%
  ggplot() +
  geom_point(aes(x = u5mr_log, y= transformed_nmr), color = "#7c9154")+
  geom_line(aes(x = u5mr_log, y = .fitted), color = "#ffda85")+
  facet_wrap(~region, scale = "free") +
  theme(panel.background = element_rect(fill = "#f0fcfc"), text = element_text(size=25)) +
  xlab("Under 5 mortality rate in Log scale") +
  ylab("Neonatal Mortality Rate in Log Scale")
```




```
# part c
augment_fit %>%
  filter(country_name %in% c("India", "Japan", "Australia")) %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr)) +
  geom_point(color = "#126d7a") +
  geom_line(aes(x = u5mr_log, y = .fitted), color = "#ff8400") +
  facet_wrap(~country_name, scales = "free") +
  theme(panel.background = element_rect(fill = "#c3d3eb")) +
  xlab("Under 5 mortality rate in Log scale") +
  ylab("Fit")
```

.metric	.estimator	.estimate
rmse	standard	0.3672529
rsq	standard	0.6561715
mae	standard	0.2759881



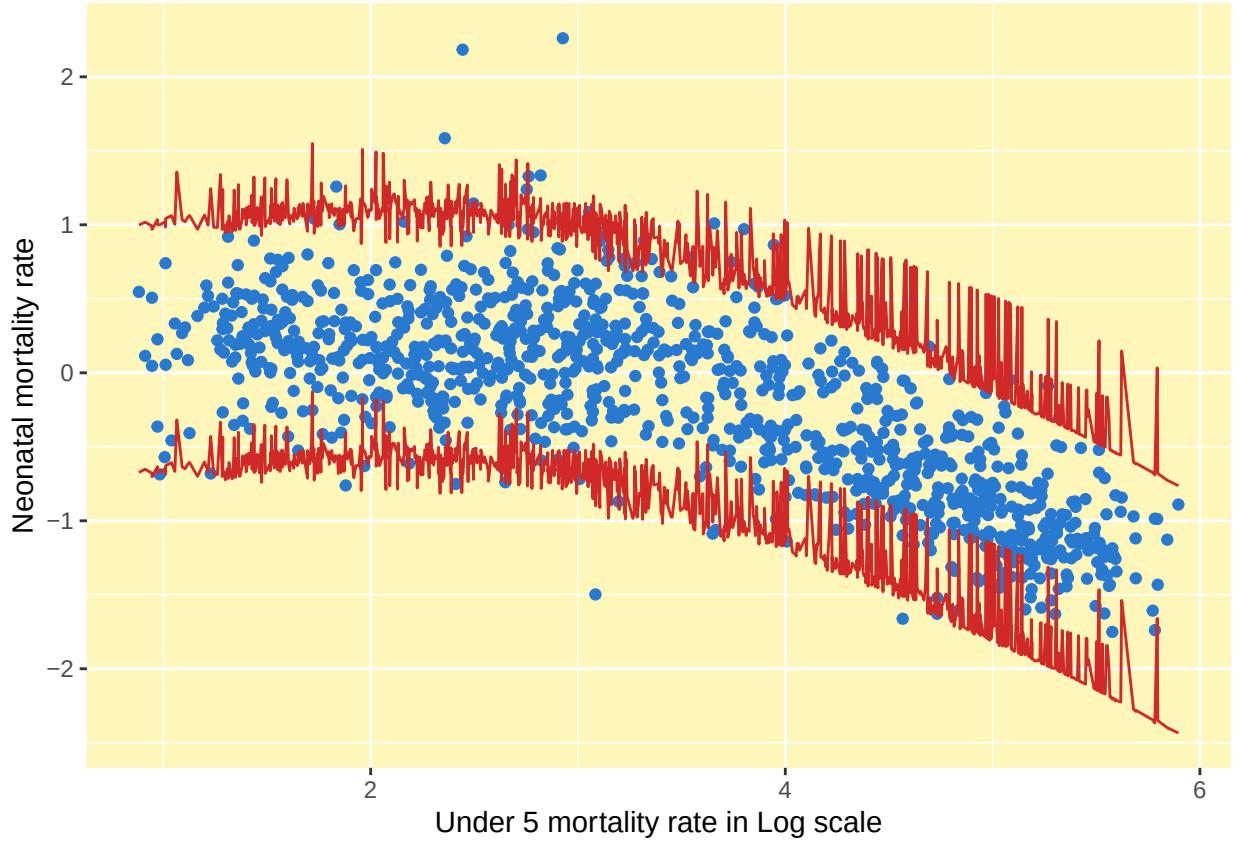
We choose India, Australia and Japan to highlight different aspects of the fit. Australia and Japan both are developed countries, and India is a developing country. For Australia and Japan, they fit terribly for the model, especially Japan, non of the points are on the model. On contrast, India fits better than the other two countries.

Task 1.1.3: Estimate the root mean square error and the mean absolute error on a test set. The test set should be produced using the argument strata = region.

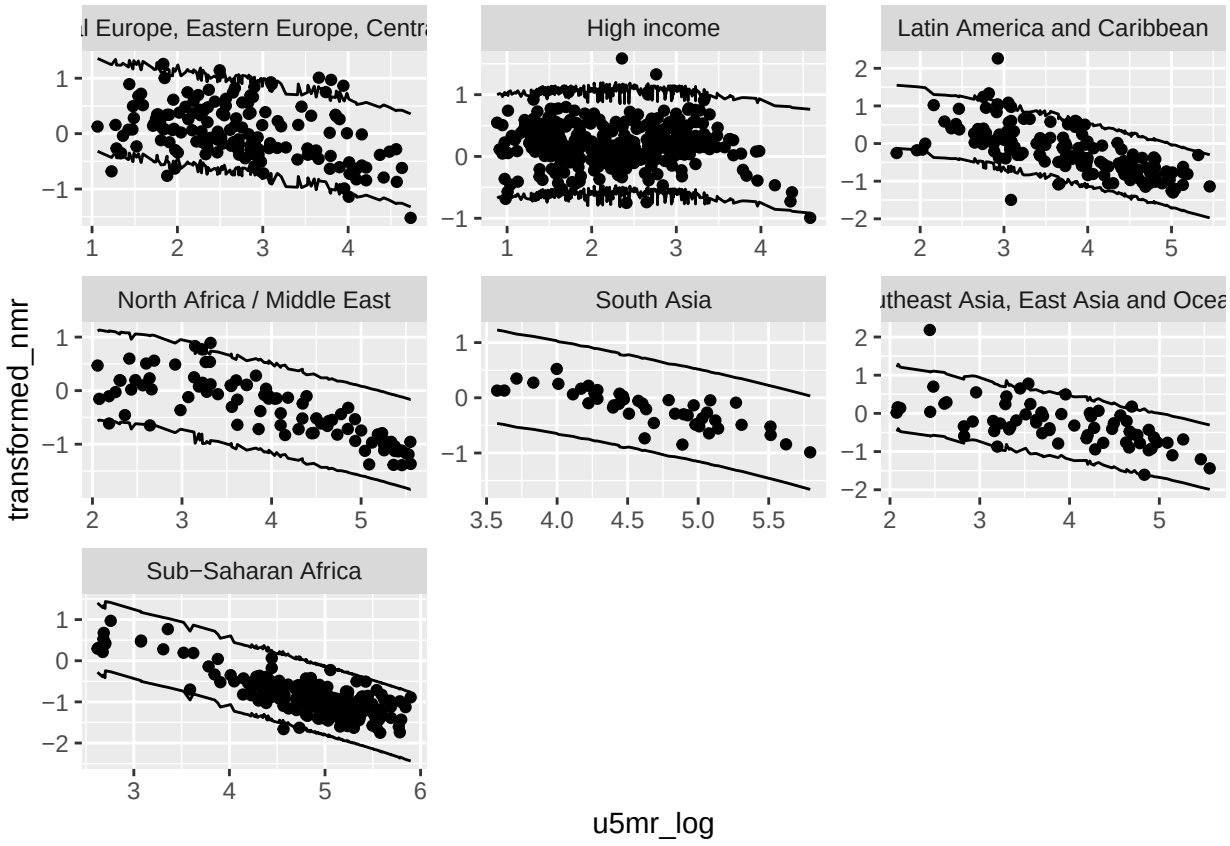
```
# Predict the test dataset
lm_pred <- tibble(pred = predict(fit, nmr_test))
lm_pred <- bind_cols(nmr_test, lm_pred)
lm_pred %>%
  metrics(truth = transformed_nmr,
          estimate = pred) %>%
  kbl(booktabs = T) %>%
  kable_styling(position = "center")
```

Task 1.1.4: Produce a prediction, with prediction intervals, of the NMR on its natural scale (aka not on the log-scale) and plot these a) for all data simultaneously; b) for data in each region; and c) for data in a maximum of 3 countries that show different aspects of the fit.

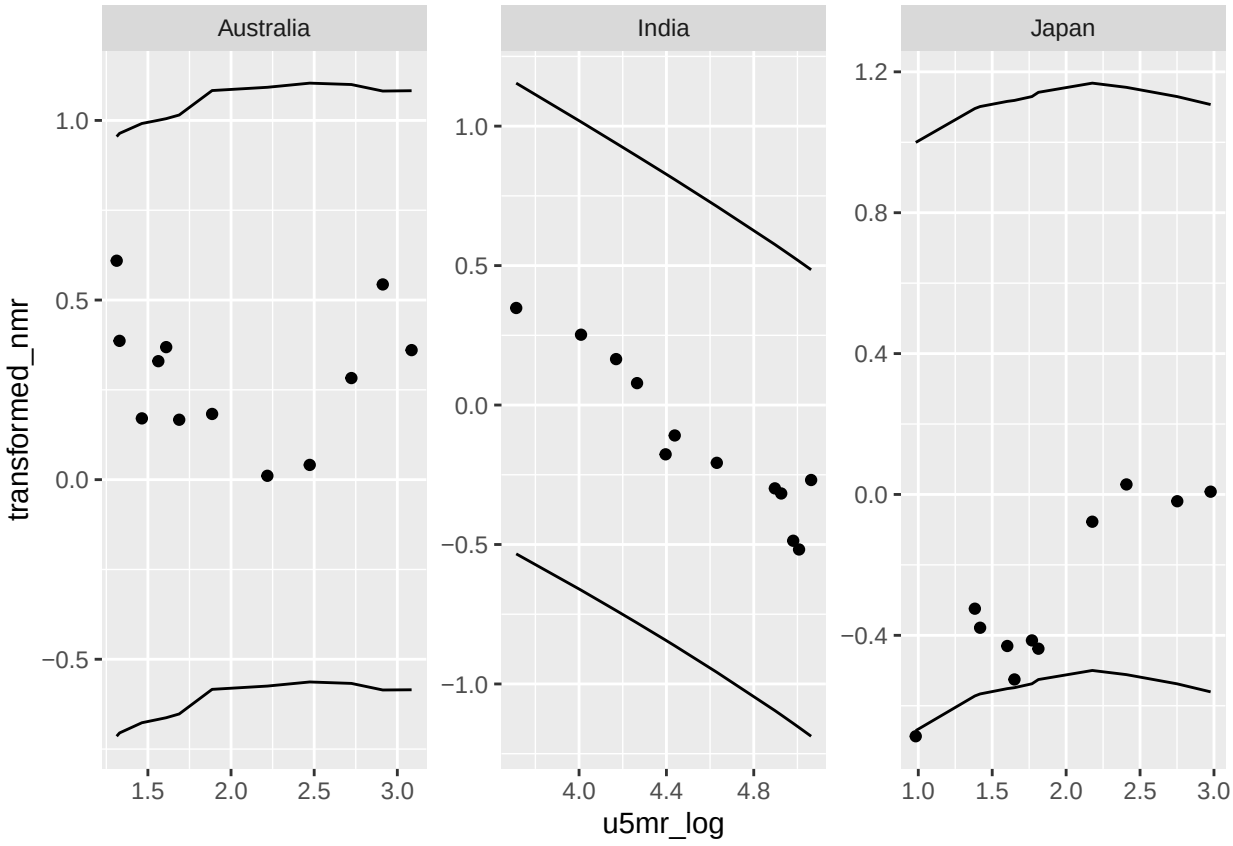
```
# Prediction for all data simultaneously
linear_prediction  <- as_tibble(predict(fit, nmr_test, interval = "prediction")) %>%
  bind_cols(lm_pred)

linear_prediction  %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr))+
  geom_point(color = "#287ad1")+
  geom_line(aes(y = lwr), color = "#d12828")+
  geom_line(aes(y = upr), color = "#d12828") +
  theme(panel.background = element_rect(fill = "#fff6ba")) +
  xlab("Under 5 mortality rate in Log scale") +
  ylab("Neonatal mortality rate")
```




```
#  Prediction for region
linear_prediction %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr))+
  geom_point()+
  geom_line(aes(y = lwr)) +
  geom_line(aes(y = upr))+
  facet_wrap(~ region, scale = "free")
```



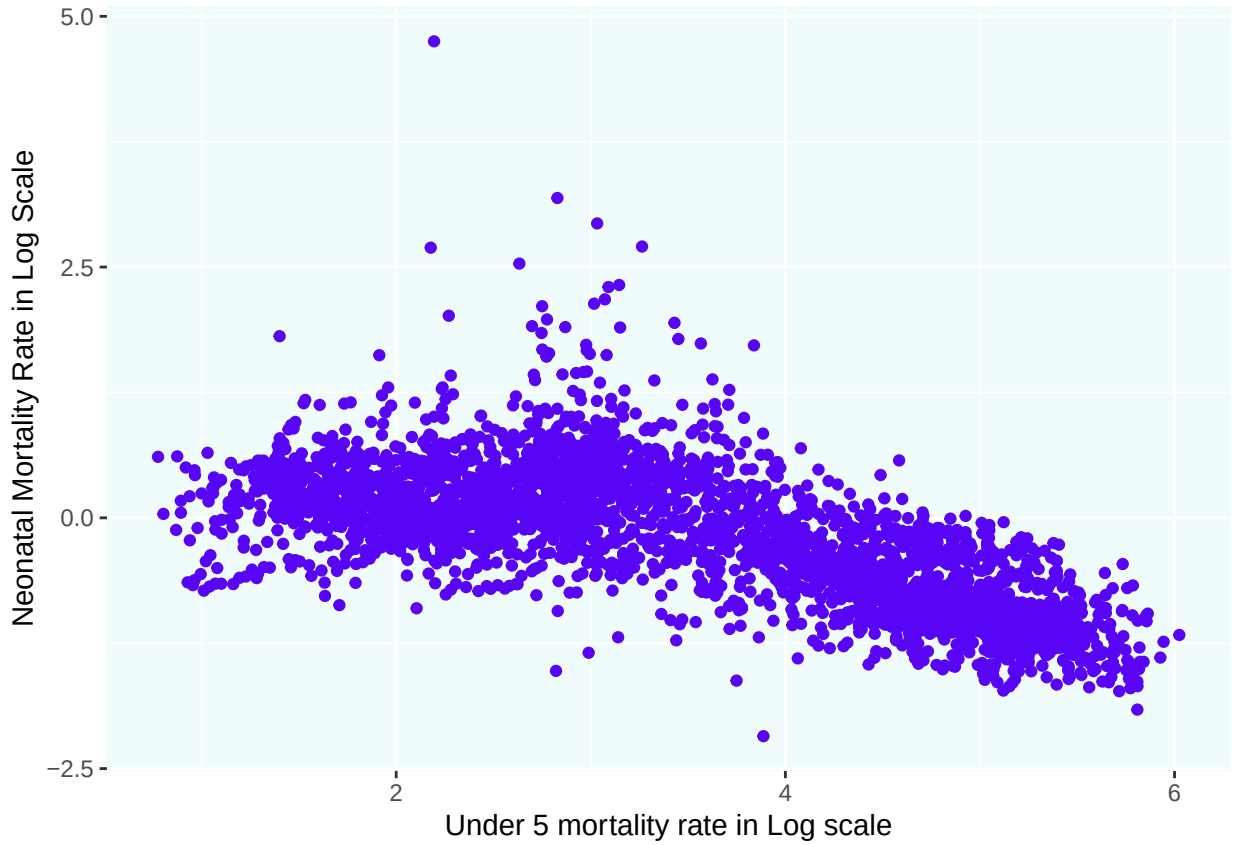

```
#  Prediction for 3 Countires
 linear_prediction %>%
  filter(country_name %in% c("India", "Japan", "Australia")) %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr))+
  geom_point()+
  geom_line(aes(y = lwr)) +
  geom_line(aes(y = upr))+
  facet_wrap(~ country_name, scale = "free")
```



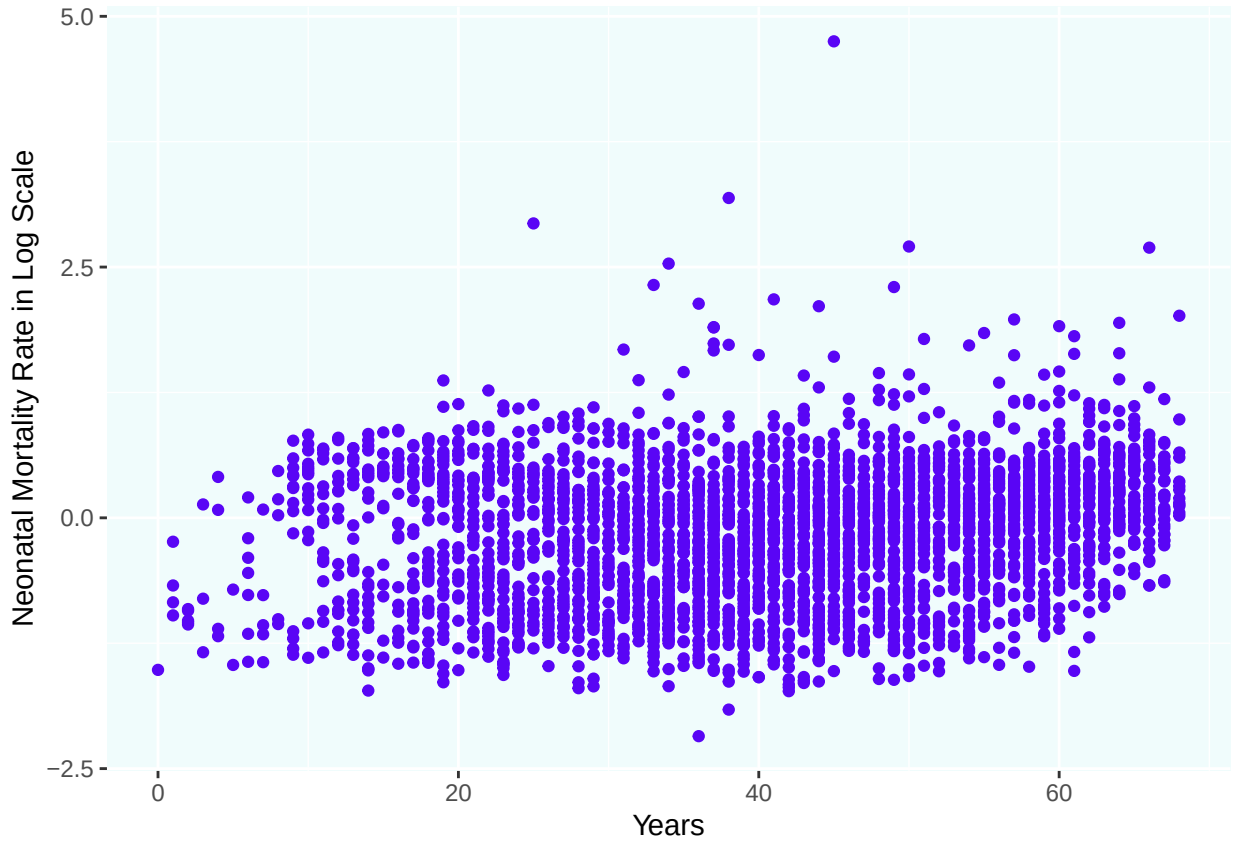
Task 1.2: Non-linear Regression

Task 1.2.1: Explain your choice of model, using appropriate visualisations to support your choice.

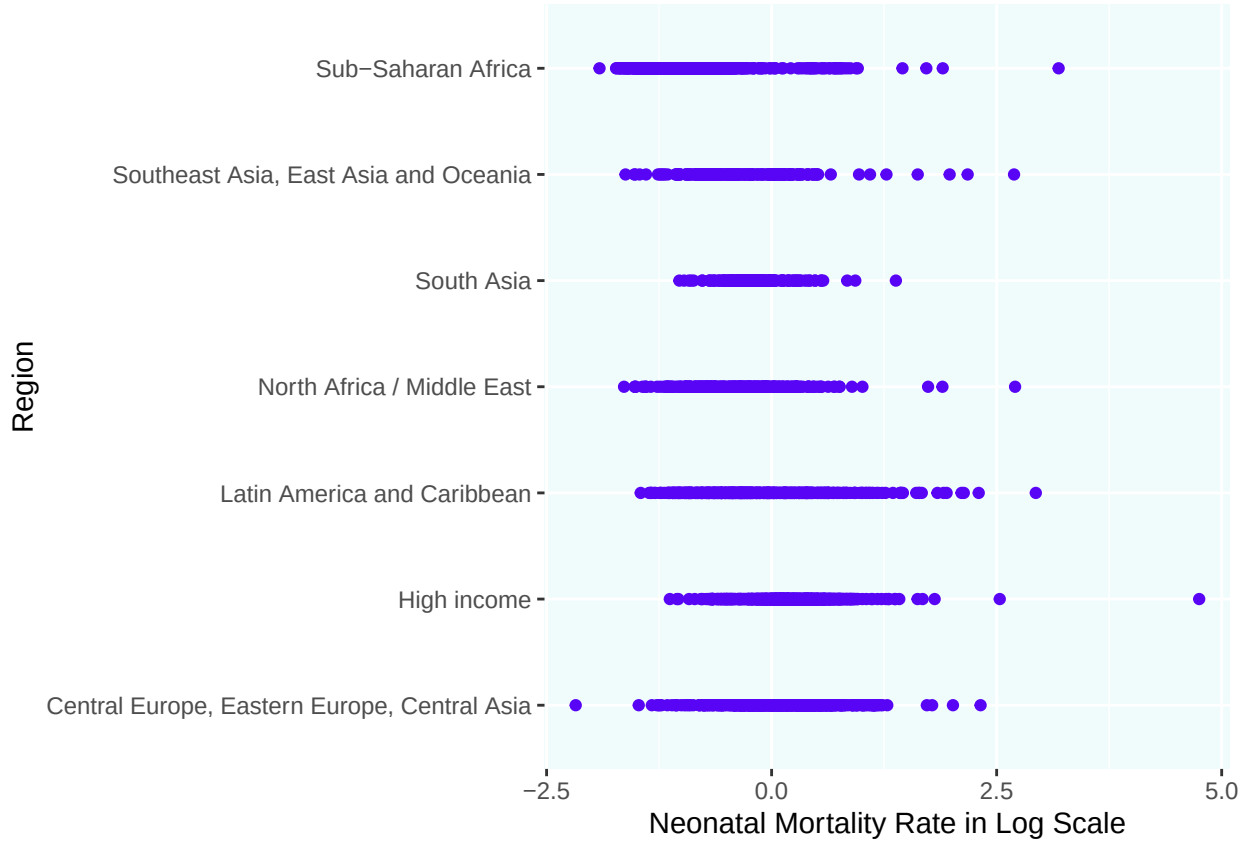
```
#non-linear variable

#non-linear variable part1 choose variable


# find which variables have non-linear relationship
#logu5 and logr non linear relationship
ggplot(nmr_train)+
  geom_point(aes(x = u5mr_log, y = transformed_nmr), color = "#5905f5")+
  theme(panel.background = element_rect(fill = "#f0fcfc")) +
  xlab("Under 5 mortality rate in Log scale") +
  ylab("Neonatal Mortality Rate in Log Scale")
```




```
#time and logr, no non-linear relationship
ggplot(nmr_train)+
  geom_point(aes(x = time, y = transformed_nmr), color = "#5905f5")+
  theme(panel.background = element_rect(fill = "#f0fcfc")) +
  xlab("Years") +
  ylab("Neonatal Mortality Rate in Log Scale")
```




```
#region and logr, no non-linear
ggplot(nmr_train)+
  geom_point(aes(y = region, x = transformed_nmr), color = "#5905f5")+
  theme(panel.background = element_rect(fill = "#f0fcfc")) +
  ylab("Region") +
  xlab("Neonatal Mortality Rate in Log Scale")
```



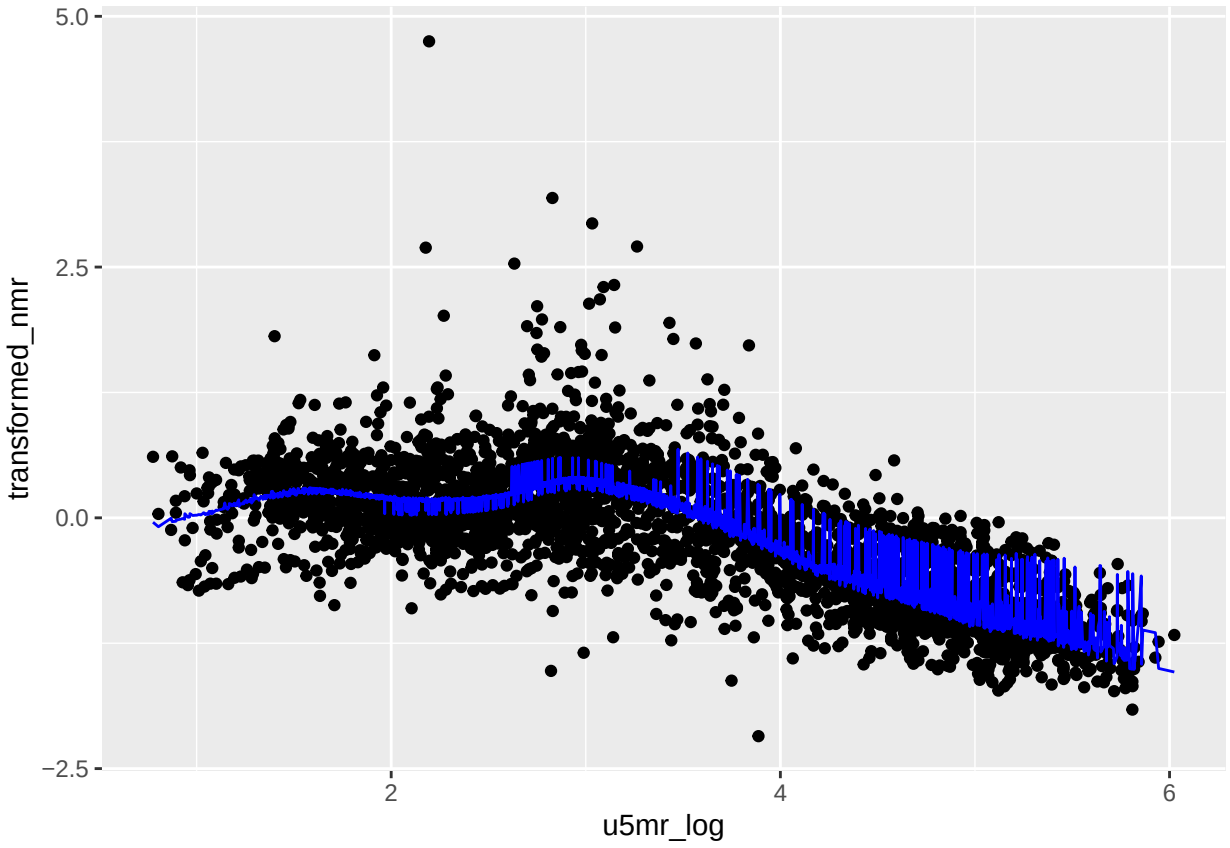

```
# non-linear model
fit_non_linear <- lm(transformed_nmr ~ bs(u5mr_log, df= 15) + u5mr_log*region + u5mr_log*time, data = nm
```

Task 1.2.2: Use cross-validation to select an appropriate number of basis functions for bs()

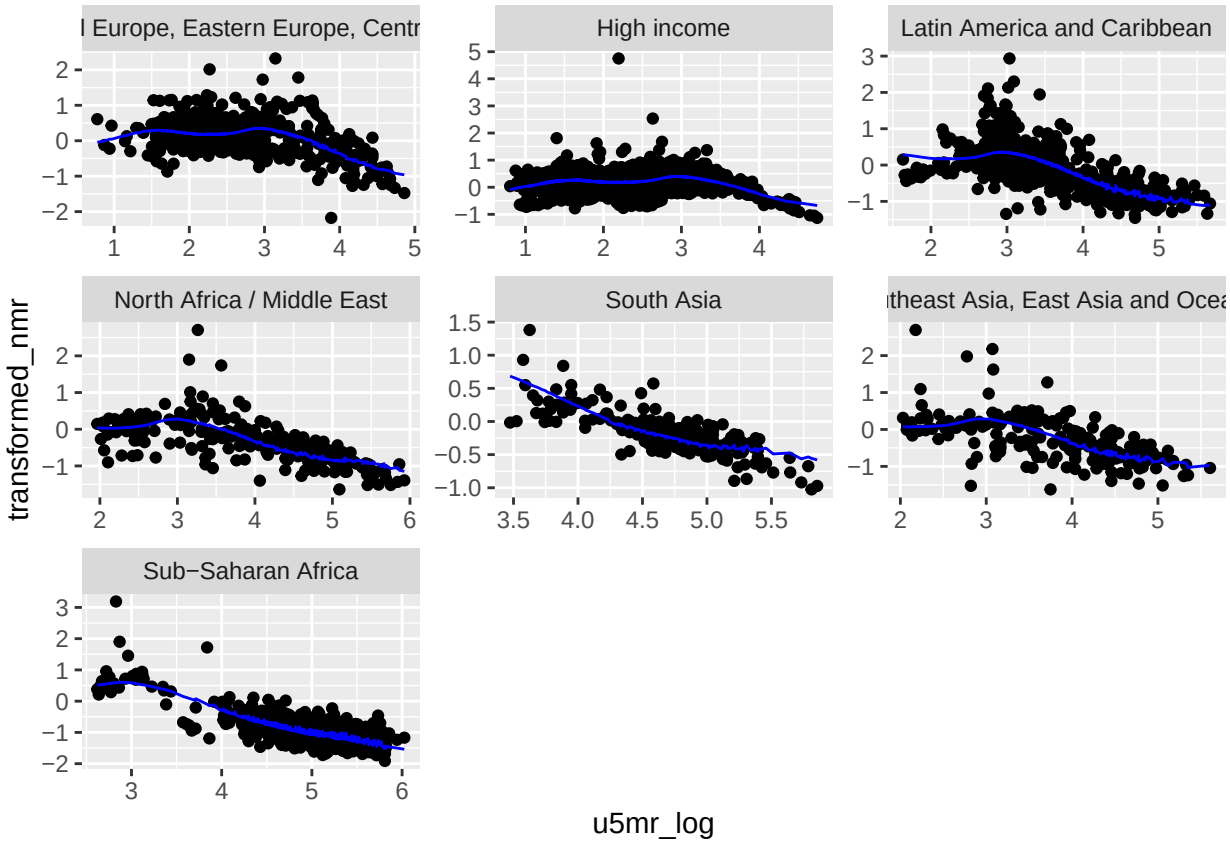
```
#NOT DONE
```

Task 1.2.3: For your final model, assess your linear model and comment on its fit. This should be done a) for all data simultaneously; b) for data in each region; and c) for data in a maximum of 3 countries that should be chosen to highlight different aspects of the fit diagnostics.

```
# non-linear all
augment_non_linear <- fit_non_linear %>%
  augment(data = nmr_train)


augment_non_linear %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr)) +
  geom_point()+
  geom_line(aes(x = u5mr_log, y = .fitted), color = "blue")
```

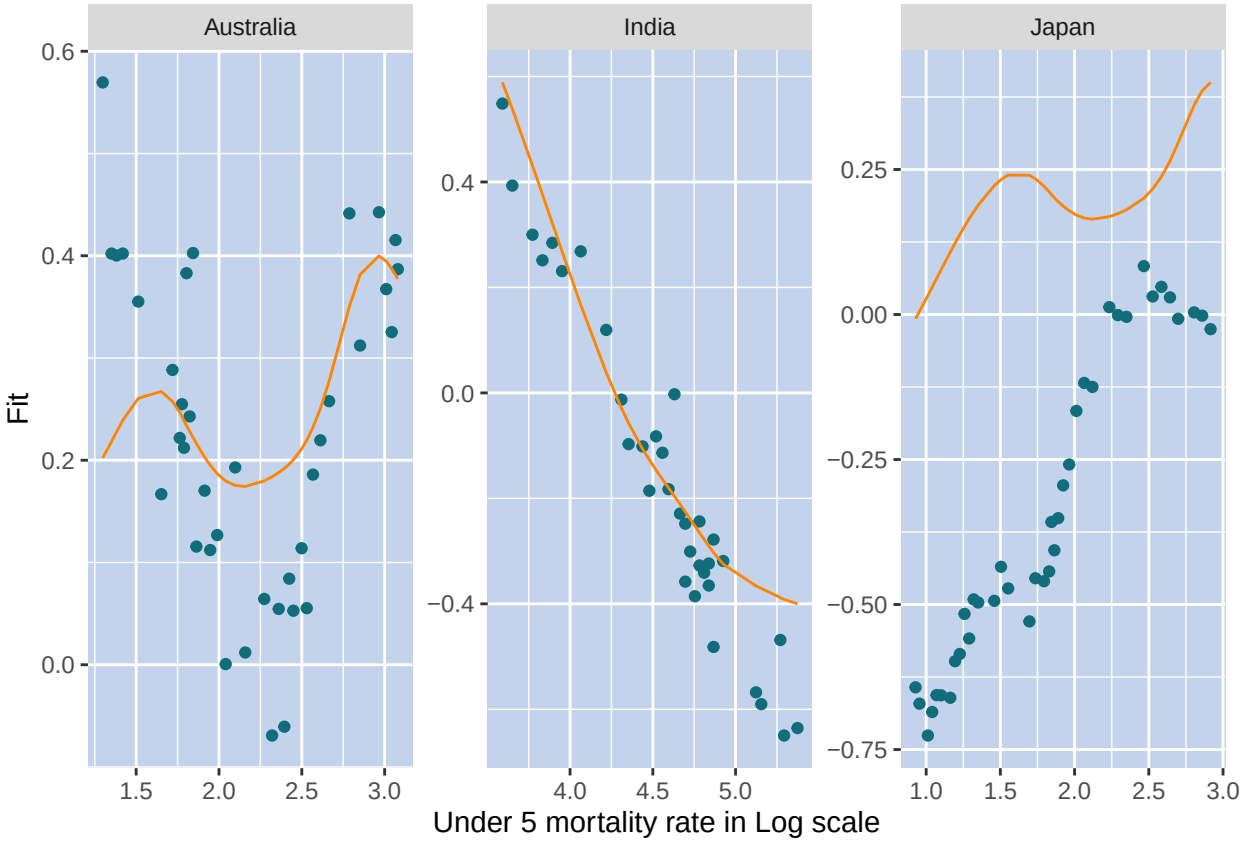



```
#non- linear region
augment_non_linear %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr)) +
  geom_point()+
  geom_line(aes(x = u5mr_log, y = .fitted), color = "blue")+
  facet_wrap(~ region, scale = "free")
```




```
# 3 countries
augment_non_linear %>%
  filter(country_name %in% c("India", "Japan", "Australia")) %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr)) +
  geom_point(color = "#126d7a") +
  geom_line(aes(x = u5mr_log, y = .fitted), color = "#ff8400") +
  facet_wrap(~country_name, scales = "free") +
  theme(panel.background = element_rect(fill = "#c3d3eb")) +
  xlab("Under 5 mortality rate in Log scale") +
  ylab("Fit")
```

.metric	.estimator	.estimate
rmse	standard	0.3575515
rsq	standard	0.6750786
mae	standard	0.2686656

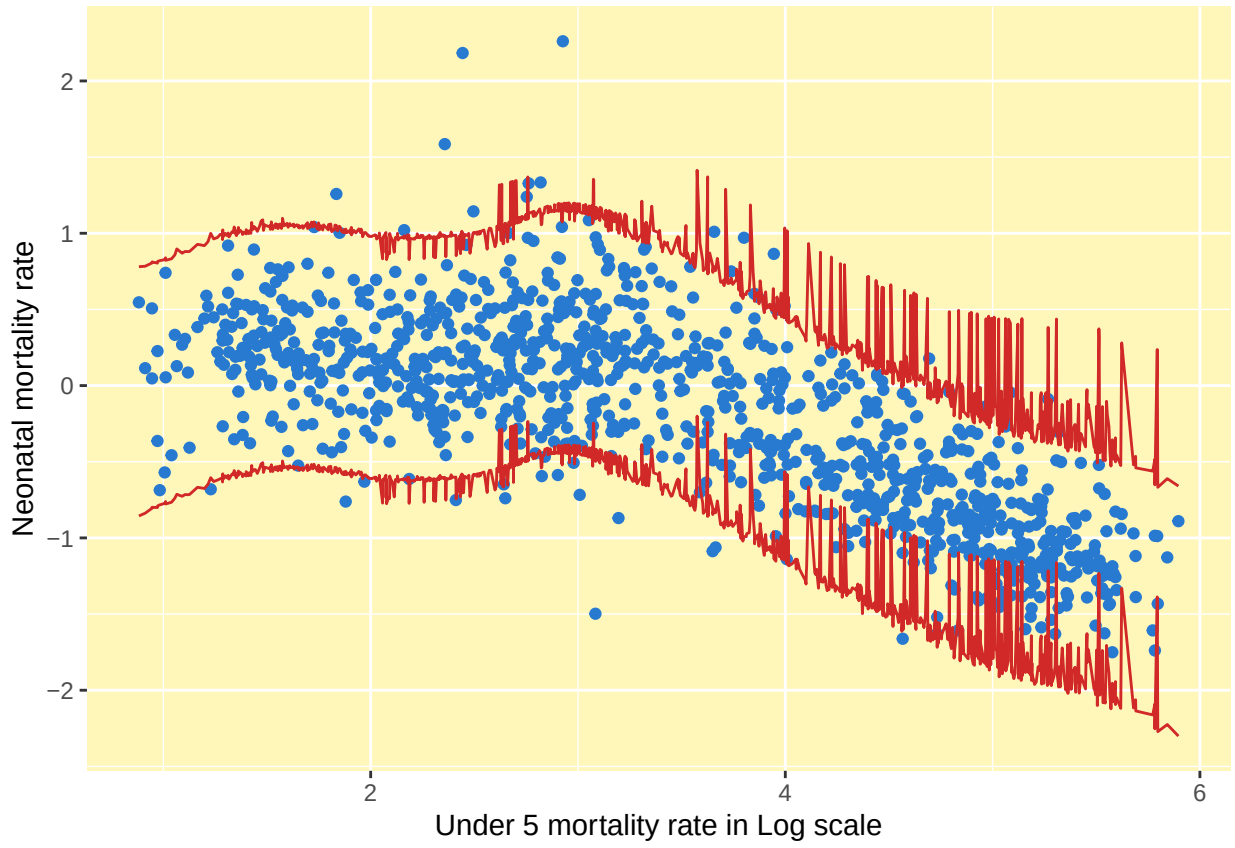


Task 1.2.4: Estimate the root mean square error and the mean absolute error using the same test set as before.

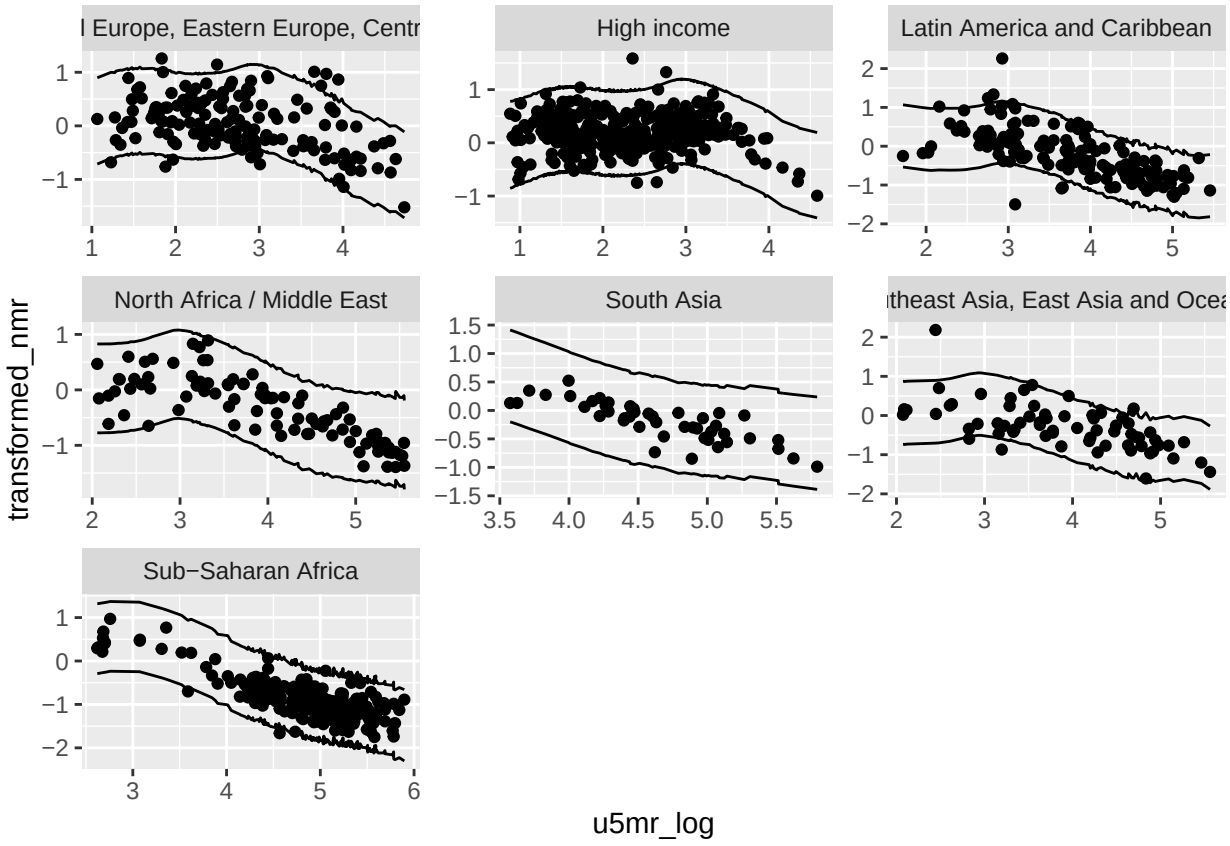
```
# Predict the test dataset
lm_pred <- tibble(pred = predict(fit_non_linear, nmr_test))
lm_pred <- bind_cols(nmr_test, lm_pred)
lm_pred %>%
  metrics(truth = transformed_nmr,
          estimate = pred) %>%
  kbl(booktabs = T) %>%
  kable_styling(position = "center")
```

Task 1.2.5: Produce a prediction, with prediction intervals8, of the NMR on its natural scale (aka not on the log-scale) and plot these a) for all data simultaneously; b) for data in each region; and c) for data in a maximum of 3 countries that show different aspects of the fit.

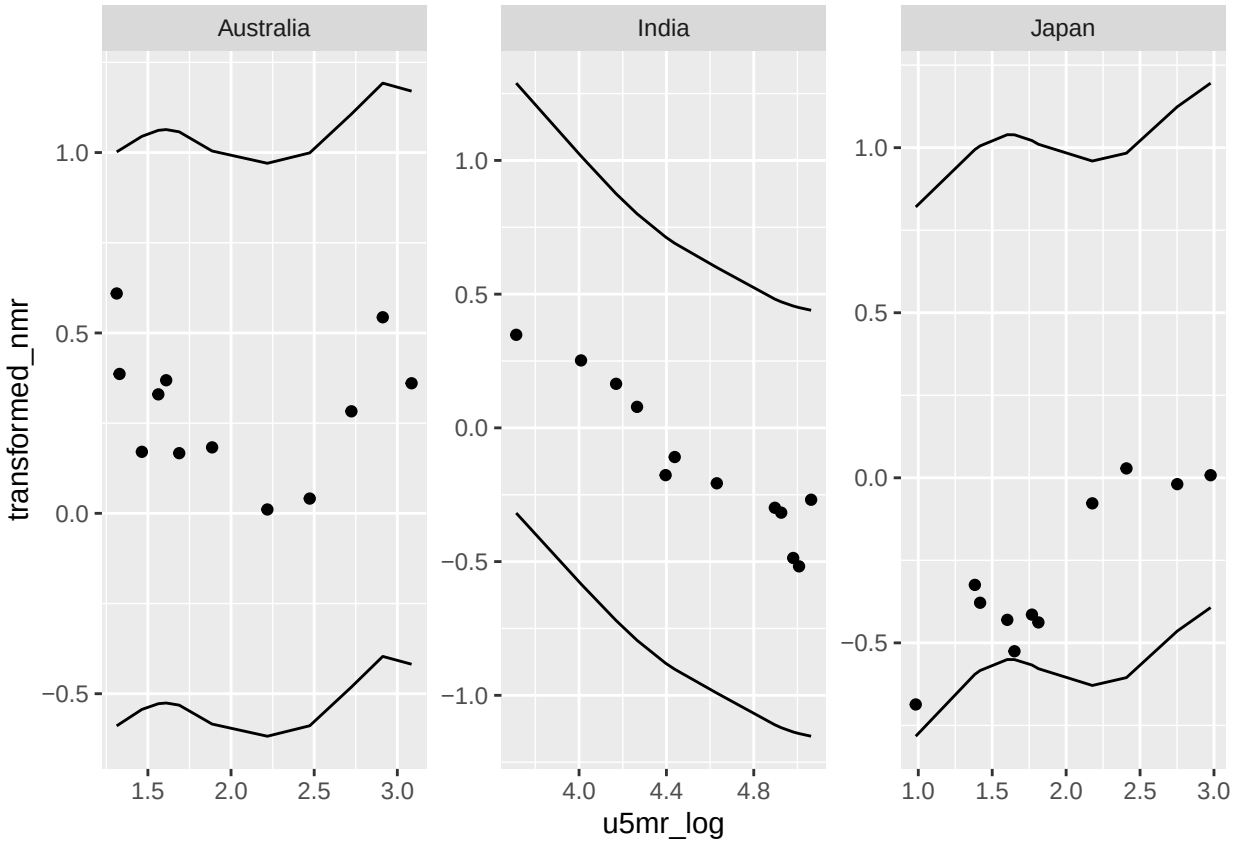
```
# Prediction for all data simultaneously
non_linear_prediction <- as_tibble(predict(fit_non_linear, nmr_test, interval = "prediction")) %>%
  bind_cols(lm_pred)

non_linear_prediction %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr))+
  geom_point(color = "#287ad1")+
  geom_line(aes(y = lwr), color = "#d12828")+
  geom_line(aes(y = upr), color = "#d12828") +
  theme(panel.background = element_rect(fill = "#fff6ba")) +
  xlab("Under 5 mortality rate in Log scale") +
  ylab("Neonatal mortality rate")
```




```
#  Prediction for region
non_linear_prediction %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr))+
  geom_point()+
  geom_line(aes(y = lwr)) +
  geom_line(aes(y = upr))+
  facet_wrap(~ region, scale = "free")
```




```
#  Prediction for 3 Countires
 non_linear_prediction %>%
  filter(country_name %in% c("India", "Japan", "Australia")) %>%
  ggplot(aes(x = u5mr_log, y = transformed_nmr))+
  geom_point()+
  geom_line(aes(y = lwr)) +
  geom_line(aes(y = upr))+
  facet_wrap(~ country_name, scale = "free")
```


Australia
India
Japan
transformed_nmr
u5mr_log
1.0
0.5
0.0
−0.5
−1.0
1.5
2.0
2.5
3.0
4.0
4.4
4.8