The reduction childhood mortality rates

Luka.C, Jason.S, Gurtej B

19/10/2021

Q1 : A linear normal regression model

```
df <- read.csv("neonatal_mortality.csv")
```

```
df_scaled <- mutate(df,
                    nmr_log = log(nmr/(u5mr-nmr)),
                    u5mr_log = log(u5mr))
```

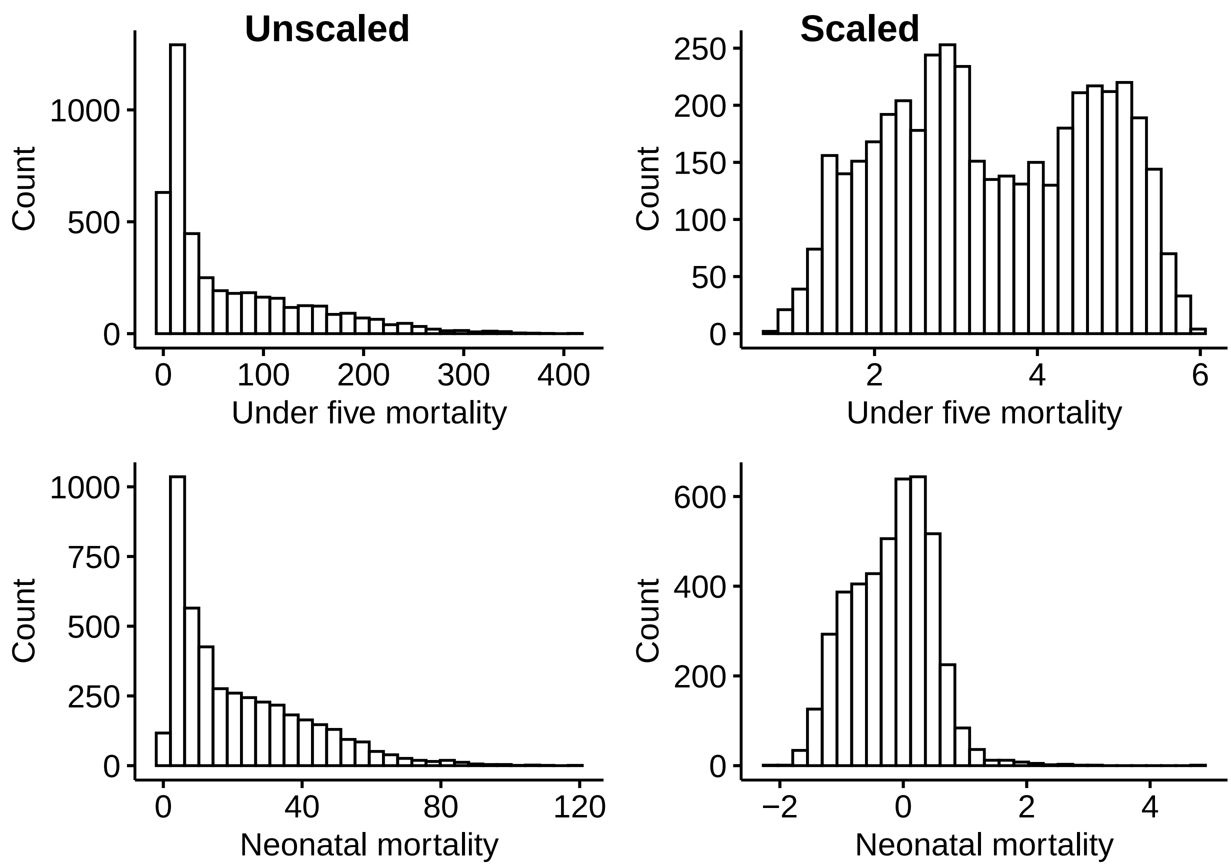
Introduction

One of the century's global goals has been the reduction childhood mortality rates. The purpose of this report is to choose significant variables and from these variable construct models that allow us to estimate neonatal mortality and build prediction intervals for neonatal martality. The variables we will be assessing include neonatal mortality rate per thousand live births, year, under five mortality rate, and region.

```
model_1 <- lm(nmr_log ~ year + region + u5mr_log, data = df_scaled)
model_2 <- lm(nmr_log ~ year + region + u5mr_log + region*u5mr_log, data = df_scaled)
model_selec <- tibble("Model" = c(1, 2),
                      "AdjustedRSquared" = c(summary(model_1)$adj.r.squared, summary(model_2)$adj.r.squa
```

Variable and model Selection

Out of the vairables given to us to build our models, we have chosen to use, time, region, and under five mortality rate to build a model that estimates the average neonatal mortality rate conditional on those variables. Neonatal mortality and under five mortality were transformed using a log transformation as they were both significantly skewed, this transformation turned it into a more normalized datatset suitable for regression to be used on as can be seen by the graphs below.



We have chosen all of the variables to use due to the fact that all but 2 are statistically significant in estimating neonatal mortality. This can be seen from the p values in the table below. We have used Centeral Europe / Eastern Europe / Central Asia as our reference region, this means that North Africa / Middle East and Southeast Asia / East Asia / Oceania are not statistically insignificant from one another at a significance level of 5%.

term	estimate	std.error	statistic	p.value
(Intercept)	15.49	1.35	11.49	0.00
year	-0.01	0.00	-10.77	0.00
regionHigh income	-0.14	0.02	-5.70	0.00
regionLatin America and Caribbean	0.07	0.03	2.52	0.01
regionNorth Africa / Middle East	0.00	0.03	0.03	0.97
regionSouth Asia	0.51	0.04	12.40	0.00
regionSoutheast Asia, East Asia and Oceania	0.03	0.03	0.77	0.44
regionSub-Saharan Africa	-0.15	0.03	-4.53	0.00

term	estimate	std.error	statistic	p.value
u5mr_log	-0.42	0.01	-33.51	0.00

The choice of an interaction variable was considered and we chose to construct two models, one without an interaction vairable called model 1, and with with an interaction variable called model 2. The interaction variable is between region and under five mortality rate. This will allow the effect of under five mortality rate to vary with region on the neonatal mortality rate. Out of the two models we chose to use the model with the interaction effect due to the logic behind allowing under five mortality rate to vary with region. In addition to this we compared the adjusted R Squared as this takes into account the incresae of variables and allows us to make comparison between two models with a different number of dependent variables. As can be seen by the table below the model we have chosen to use is model 2 as it has a higher adjusted r squared of 0.61.

Model	AdjustedRSquared
1	0.55
2	0.61

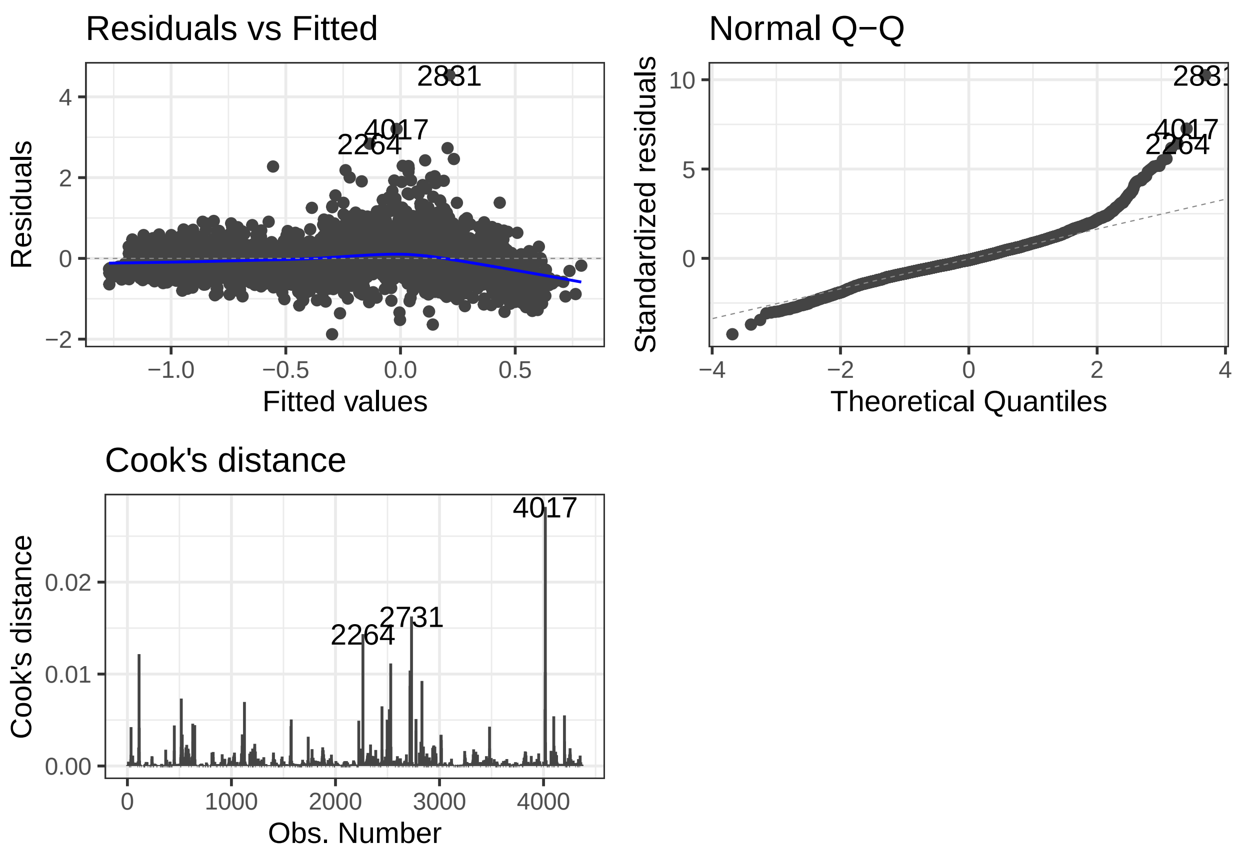
1.1 a) Model Fit Diagnostics for all data simultaneously

An inspection of a residual plot for all the data shows that there doesn't look to be any patterns which is a good sign for the regression model. The only alarming characteristic is the increase in variation from -0.25 to 0.5, This could possibly indicate heteroskedsticity, a variance that isn't constant, this violates one of the characteristics that have to be met for this model to be oujr best linear unbiased estimate od neonatal mortality. The consequences of this is that it affects our ability to perform t tests and F tests on our models regressors. A graph of fitted vs residuals can be seen below.

Another imporant aspect of assessiong the models fit is identifying any high influential points as they could be skewing thew model and effecting the models ability to give accurate estimates. As can be seen in the Cooks Distance graph below there is 3 highly influential points, 2717, 2731, and 4017. This warrants futher investigation into those specific data points as one possible explanation could be that there are errors in how they were recorded, removal of these point could improve the overal accuracy of the models ability to estimate effectivly, however sadly we cannot just remove data because it doesnt suit us as this would present bias in the estimates.

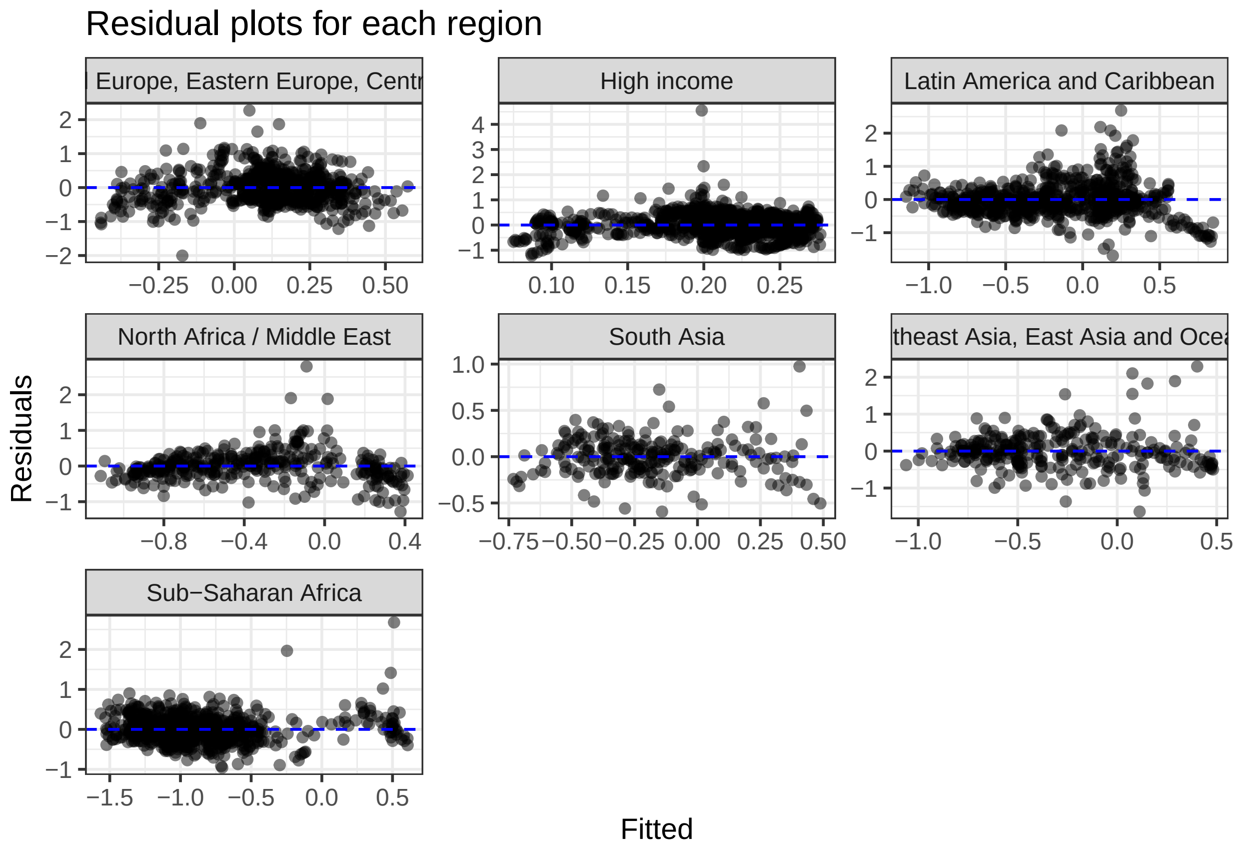
Assessing a QQ plot of the standrdised residuals will give an indication of whether or not the data is normally distributed, as this is one of the assumptions that must hold for this model to be the best linear unbiased estimate. As can be seen by the QQ plot presented below, both tails fall quite far from the line which gives evidence to suggest that the data perhaps comes from a distribution that isn't normal.

```
library(ggfortify)
model_2_aug <- augment(model_2)
model_2_aug <- cbind(model_2_aug, df_scaled[c("country_name")])

autoplot(model_1, which = c(1,2,4)) + theme_bw()
```



Inspecting residual plots for independent regions shows the same as all the data for the most part, mostly random with inconsistnace variances for different fitted values, once again suggesting that the residuals aren't homoskedastic. Sub-Subharan Africa shows two closely clustered groups of residuals, this could possibly indicate that there is a variable that hasn't been included in this model that could explain that. There is also significant differences in the scale of the variation of the residuals from region to region, this is more evidence to suggest that there heteroskedasticity present.

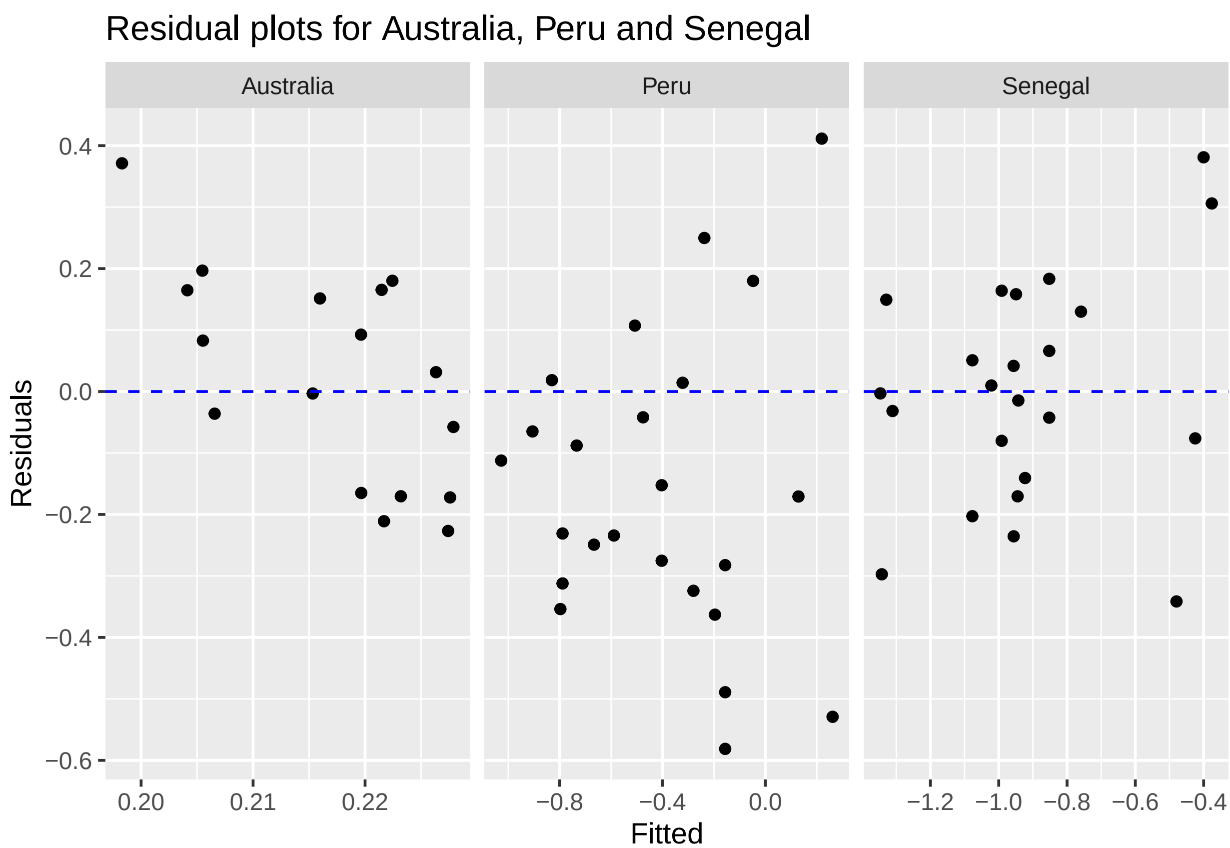
```
model_2_aug %>% ggplot(aes(x = .fitted, y = .resid)) + geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "blue") +
  facet_wrap(~ region, scales = "free") +
  theme_bw() + xlab("Fitted") + ylab("Residuals") +
  ggtitle("Residual plots for each region")
```



An inspection of the residual plots for the three countries selected shows different scales of variance in the two poorer countries, Peru and Senegal, in addition to this the residuals for Peru don't seem to be centered around zero like Senegal and Australia.

Furthermore, lots of points of Peru seem to be negative whereas the Australia's and Senegal's ones are scattered residual plot can be seen below.

```
countries <- c("Senegal", "Peru", "Australia")
country_model <-  model_2_aug %>% filter(country_name == countries)%>% rowid_to_column( "ID")
#residual plot by country
country_model%>%  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "blue") +
  facet_wrap(~ country_name, scales = "free_x") + xlab("Fitted") + ylab("Residuals") +
  ggtitle("Residual plots for Australia, Peru and Senegal")
```



Mean square error and mean absolute error for the model

```
df_scaled_split <- initial_split(df_scaled, strata = region)
df_scaled_test <- testing(df_scaled_split)
```

```
df_scaled_pred <- predict(model_2, df_scaled_test)
df_scaled_pred <- tibble(pred = df_scaled_pred)
df_scaled_test_pred <- cbind(df_scaled_test, df_scaled_pred[c("pred")])
df_scaled_test_pred <- df_scaled_test_pred %>% mutate(error = nmr_log - pred,
                                              error2 = error^2,
                                              abs_error = abs(error))
mse <- sum(df_scaled_test_pred$error2)/nrow(df_scaled_test_pred)
mae <- sum(df_scaled_test_pred$abs_error)/nrow(df_scaled_test_pred)
```

The mean square error and the mean absolute error on a test set are 0.17 and 0.29 respectively.

Prediction intervals for neonatal mortality rate

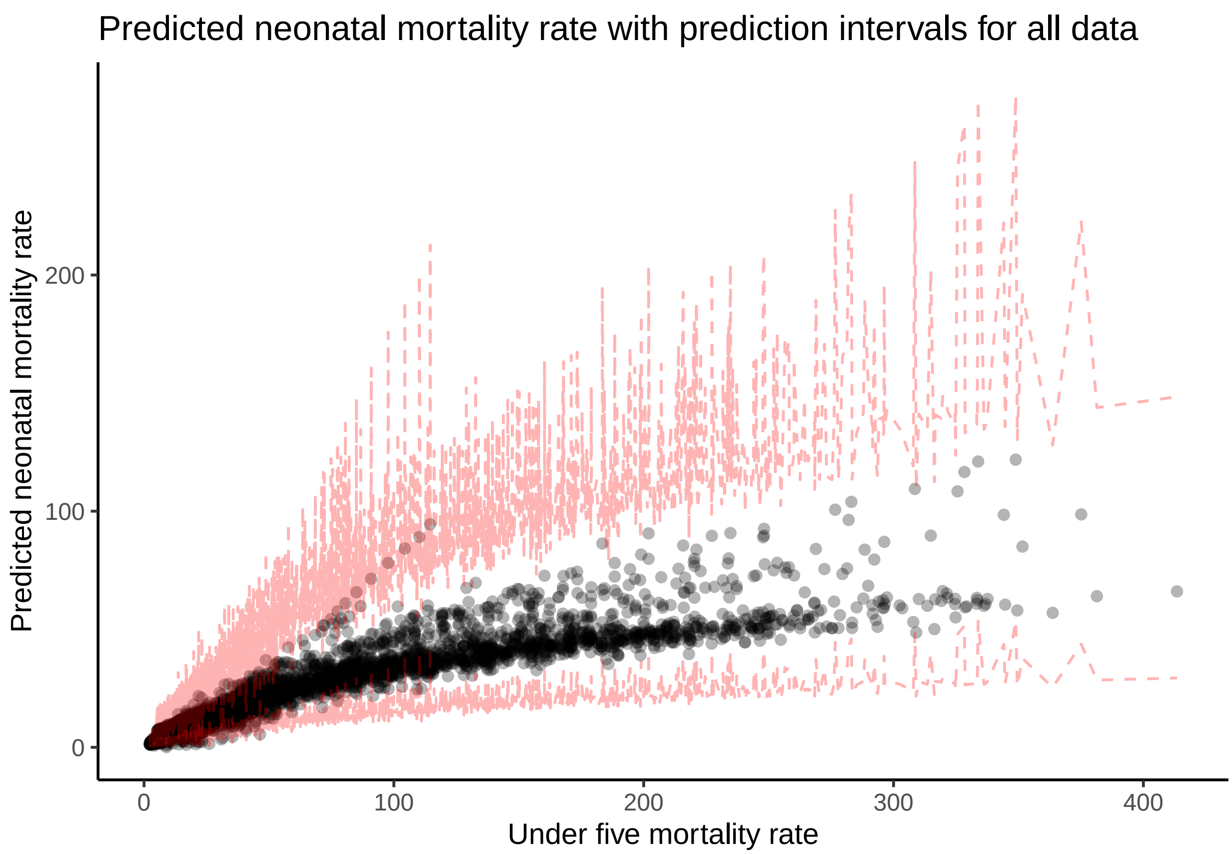
```
pred_int <- predict(model_2, df_scaled, interval = "prediction")
pred_int_tib <- tibble(pred = pred_int[,1],
                       lower = pred_int[,2],
```

```
                         upper = pred_int[,3])
pred_int_df <- cbind(df_scaled, pred_int_tib[c("pred", "lower", "upper")])
pred_int_df <- cbind(pred_int_df, model_2_aug[c("country_name")])
```

```
pred_graph <- tibble(nmr = df_scaled$nmr,
                     u5mr = df_scaled$u5mr,
                     nmr_pred_log = pred_int_df$pred,
                     nmr_pred_lower_log = pred_int_df$lower,
                     nmr_pred_upper_log = pred_int_df$upper,
                     nmr_pred = (exp(nmr_pred_log))*(u5mr - nmr),
                     nmr_pred_lower = (exp(nmr_pred_lower_log))*(u5mr - nmr),
                     nmr_pred_upper = (exp(nmr_pred_upper_log))*(u5mr - nmr),
                     region = df_scaled$region,
                     country_name = df_scaled$country_name)
```

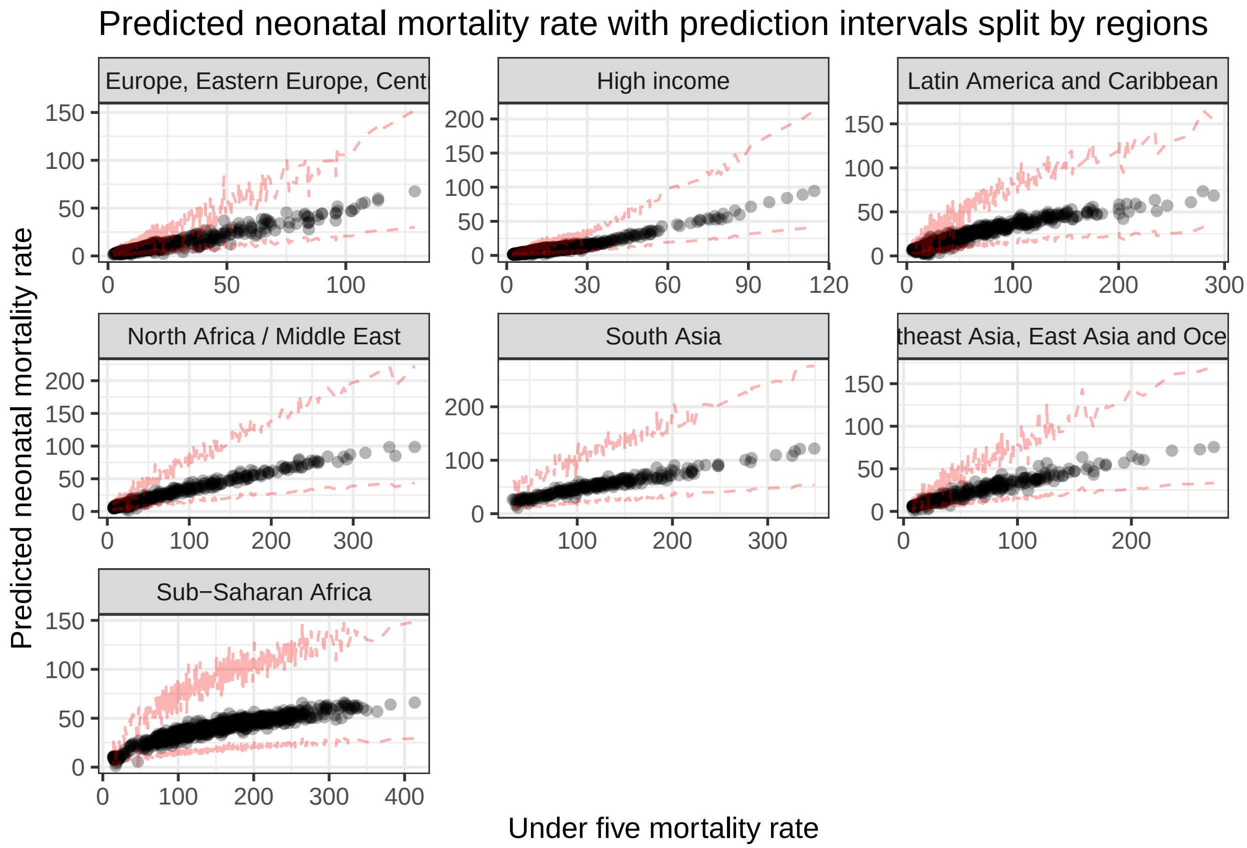
Below is the graph of the predicted values and the prediction intervals for neonatal mortality rate for all the data. The predicted values are the points and the upper and lower prediction interval is given by the red lines.

```
pred_graph %>% ggplot(aes(x = u5mr, y = nmr_pred)) +
  geom_point(alpha = 0.3) +
  geom_line(aes(y = nmr_pred_upper),
            linetype = "dashed",
            alpha = 0.3,
            colour = "red") +
  geom_line(aes(y = nmr_pred_lower),
            linetype = "dashed",
            alpha = 0.3,
            colour = "red") + xlab("Under five mortality rate") + ylab("Predicted neonatal mortality ra
  ggtitle("Predicted neonatal mortality rate with prediction intervals for all data")+theme_classic()
```



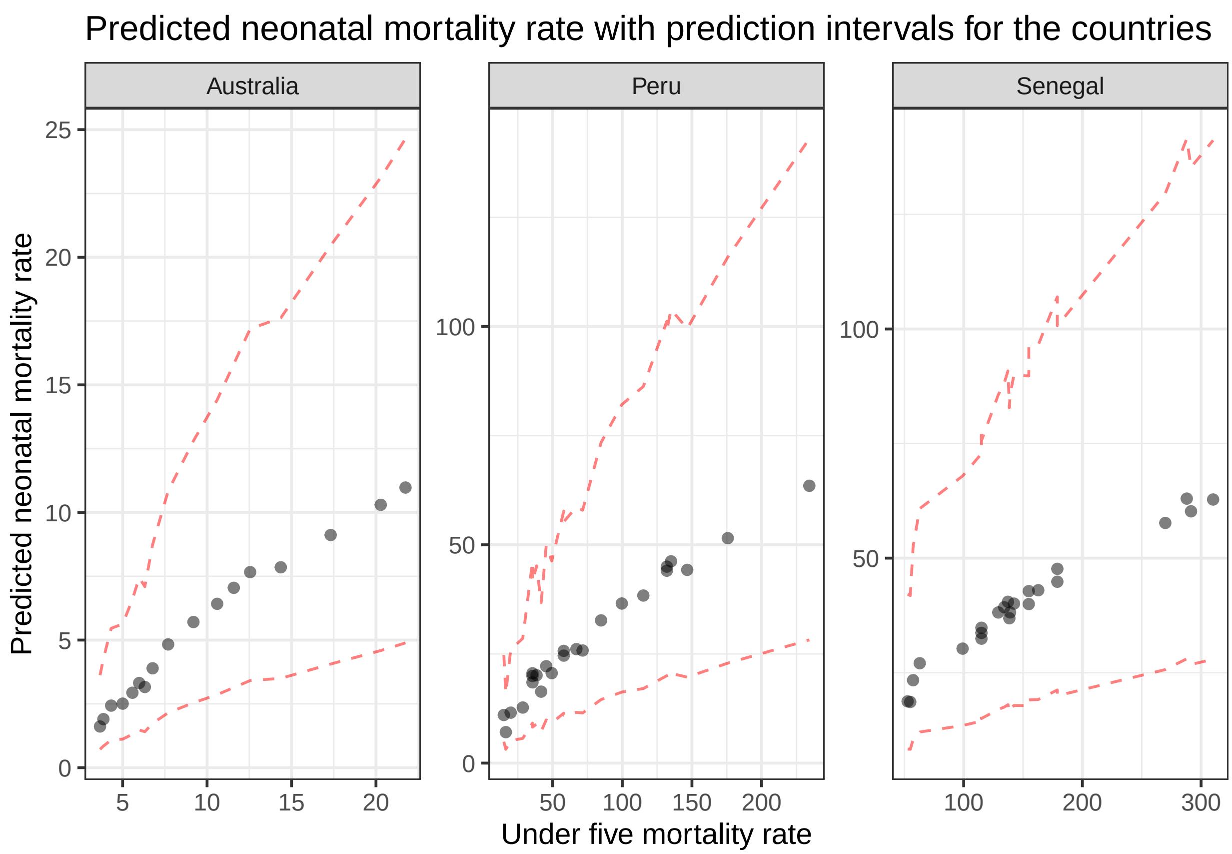
Below is the graph of the predicted values and the prediction intervals for neonatal mortality rate split by region. The predicted values are the points and the upper and lower prediction interval is given by the red lines.

```
pred_graph %>% ggplot(aes(x = u5mr, y = nmr_pred)) + geom_point(alpha = 0.3) +
  geom_line(aes(y = nmr_pred_upper),
            linetype = "dashed",
            alpha = 0.3,
            colour = "red") + facet_wrap(~ region, scales = "free") +
  geom_line(aes(y = nmr_pred_lower),
            linetype = "dashed",
            alpha = 0.3,
            colour = "red") +
  xlab("Under five mortality rate") +
  ylab("Predicted neonatal mortality rate") +
  ggtitle("Predicted neonatal mortality rate with prediction intervals split by regions")+theme_bw()
```



Below is the graph of the predicted values and the prediction intervals for neonatal mortality rate split by selected countries. The predicted values are the points and the upper and lower prediction interval is given by the red lines.

```
country_predict <-  pred_graph %>% filter(country_name == countries)%>% rowid_to_column( "ID")
country_predict %>% ggplot(aes(x = u5mr, y = nmr_pred)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(y = nmr_pred_upper),
            linetype = "dashed",
            alpha = 0.5,
            colour = "red") + facet_wrap(~ country_name, scales = "free") +
  geom_line(aes(y = nmr_pred_lower),
            linetype = "dashed",
            alpha = 0.5,
            colour = "red") + xlab("Under five mortality rate") + ylab("Predicted neonatal mortality ra
  ggtitle("Predicted neonatal mortality rate with prediction intervals for the countries")+theme_bw()
```



==

Q1 : A linear regression model with incorporating an appropriate non-linear effect

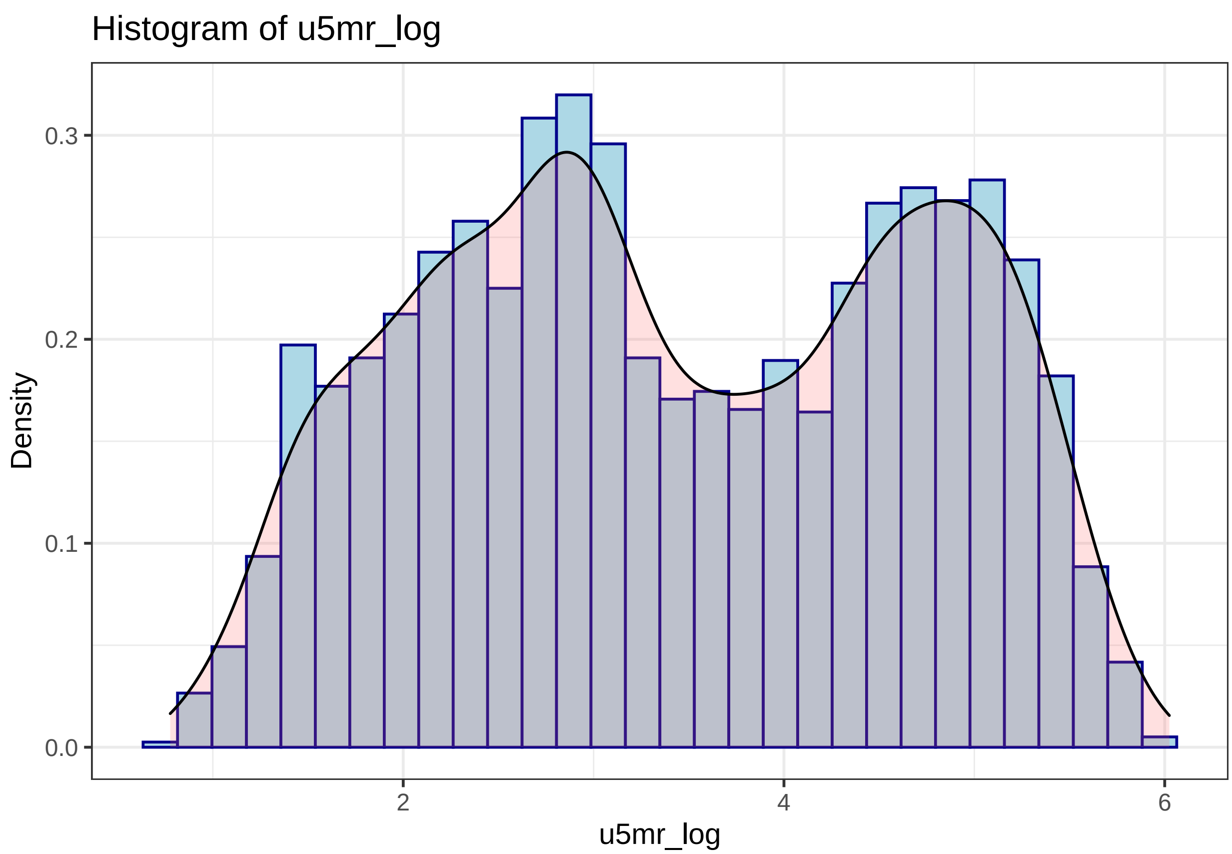
Q1.1) Explain the choice of model using appropriate visualisations.

To explain the choice of model, there are 3 questions to be answered:

Q1: Why are we still using regression?

Due to the limitation of the dataset in which we are predicting a numeric variable, it is impossible to use other supervised learning models like clustering. Using regression is a simple method to tackle this problem where the independent and dependent variables are clearly defined.

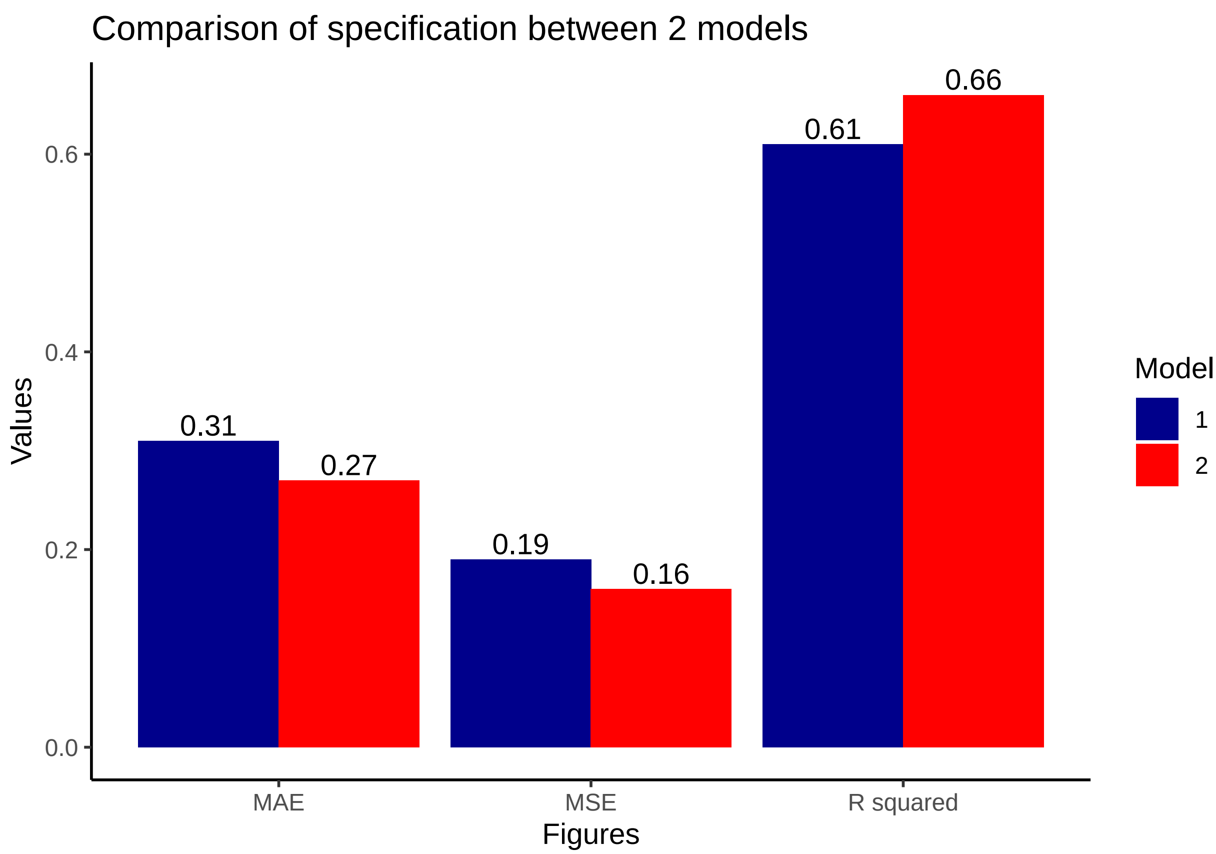
```
ggplot(df_scaled, aes(x=u5mr_log))+
  geom_histogram(aes(y=..density..), colour="darkblue", fill="lightblue")+
 geom_density(alpha=.2, fill="#FF6666") +
  theme_bw()+
  ggtitle("Histogram of u5mr_log")+ylab("Density")
```



Q2: If using regression, why did we put bs()?

The purpose of this linear model is to incorporate an appropriate non-linear effect. And based on the histogram above, we can show that the variable – u5mr_log – is of no linearity. Therefore using bs() for u5mr_log allows us to smooth it out by considering its nonlinear effect. bs() is a better method than polynomial regression for this problem.

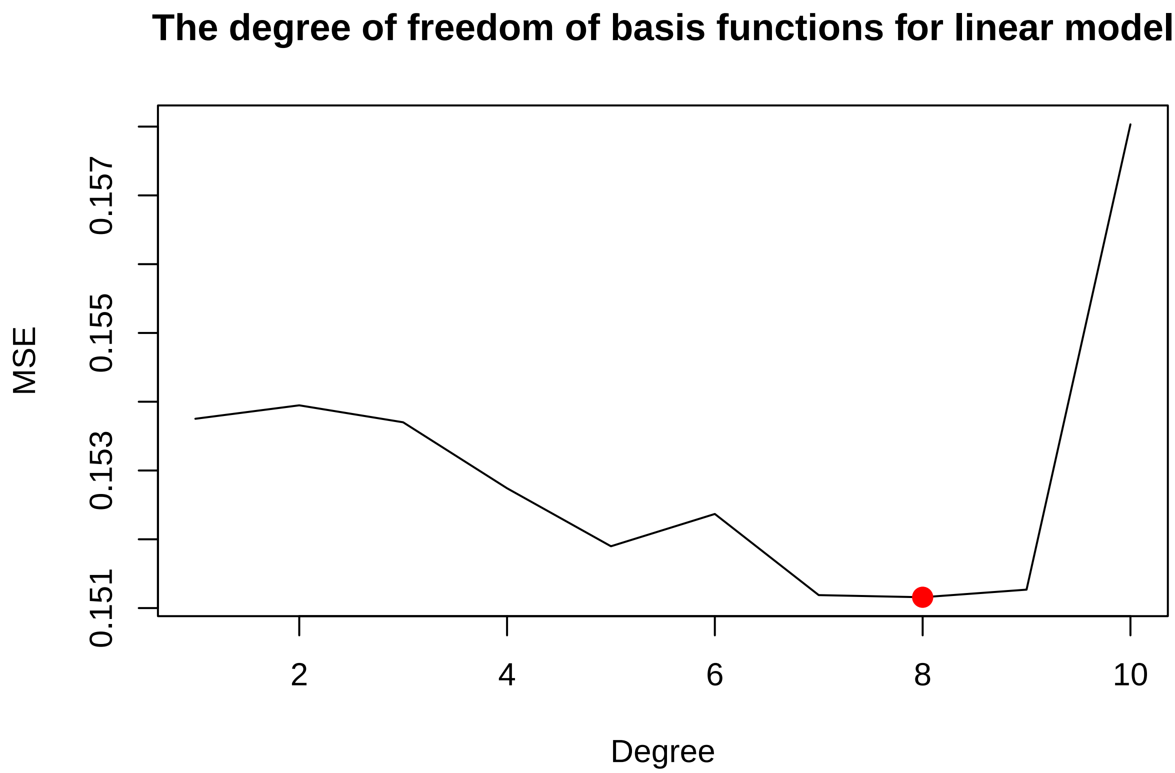
```
compare <- data.frame(Model = c(1,2,1,2,1,2),
                  Values= c(0.61,0.66,0.19,0.16,0.31,0.27),
                  Figures =c("R squared","R squared","MSE","MSE","MAE","MAE"))

compare[,1] <-as.factor(compare[,1])


ggplot(compare,
       aes(x=Figures,
                y=Values,
                fill=Model )) +
    geom_col(position = "dodge")+
  scale_fill_manual(values=c("Dark Blue", "Red"))+
  geom_text(aes(label=Values), position=position_dodge(width=0.9), vjust=-0.25) +
  ggtitle("Comparison of specification between 2 models")+theme_classic()
```



Q3: Why do we prefer model 2?

It has higher R Squared, lower MAE and MSE. The details will be explained in the last part of this analysis.

```
set.seed(100)
deltas <- rep(NA, 10)
for (i in 1:10) {
  ## glm specified as Gaussian is same as lm.
  fit <- glm(nmr_log ~ bs(u5mr_log, i)*region + year, data = df_scaled, family = gaussian)
    deltas[i] <- cv.glm(df_scaled, fit, K = 10)$delta[1]
}
plot(1:10, deltas, xlab = "Degree", ylab = "MSE", type = "l",
     main = "The degree of freedom of basis functions for linear model")
d.min <- which.min(deltas)
points(which.min(deltas), deltas[which.min(deltas)], col = "red", cex = 2, pch = 20)
```



After using cross-validation with 10 folds for linear model, the optimal degree of the piecewise polynomial of bs() is at 8, with the minimum of MSE.

```
final_model <- lm(nmr_log ~ bs(u5mr_log, which.min(deltas))*region + year, data = df_scaled)
model_selec <- tibble("Model" = c("Normal lm", "Final lm"),
                      "Adjusted R Squared" = c(summary(model_2)$adj.r.squared, summary(final_model)$adj
kable(model_selec)
```

Model	Adjusted R Squared
Normal lm	0.61
Final lm	0.66

In terms of R-square,roughly 66% of the variation in the nmr_log can be explained by year, region, and u5mr_log. Which has been roughly improved by 5%.

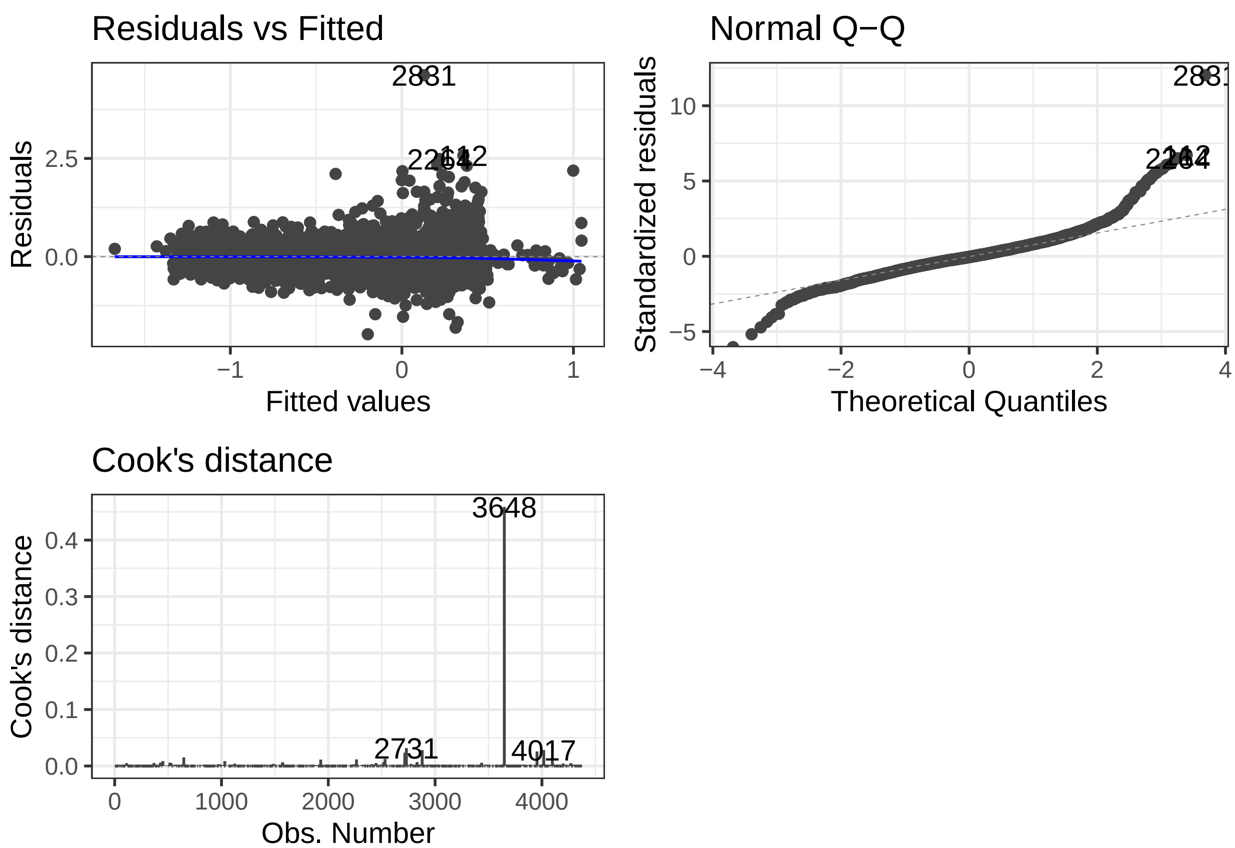
1.2 a) Model Fit Diagnostics for all data simultaneously

Inspecting of a residual plot for all the data, there is no pattern shown in residual plot based on the plot displayed, so we can infer that linear model is good enough and independent. The only alarming characteristic is the increase in variation from -0.25 to 0.5, This could possibly indicate heteroskedsticity, a variance that isn't constant, this violates one of the characteristics that have to be met for this model to be oujr best linear unbiased estimate od neonatal mortality. The consequences of this is that it affects our ability to perform t tests and F tests on our models regressors. A graph of fitted vs residuals can be seen below.

Assessing a QQ plot of the standrdised residuals will give an indication of whether or not the data is normally distributed, as this is one of the assumptions that must hold for this model to be the best linear unbiased estimate. As can be seen by the QQ plot presented below, both tails are deviating the from the diagonal line which gives evidence to suggest that the data perhaps comes from a distribution that isn't normal.

Another imporant aspect of assessiong the models fit is identifying any outliers as they could be skewing thew model and effecting the models ability to give accurate estimates. So, A good linear model should avoid as much outliers as possible. As can be seen in the Cooks Distance graph, therefore, if eliminating the 2731st, 3648th, and 4017th observation, the model will be better. This warrants futher investigation into those specific data points as one possible explanation could be that there are errors in how they were recorded, removal of these point could improve the overal accuracy of the models ability to estimate effectivly, however sadly we cannot just remove data because it doesnt suit us as this would present bias in the estimates.

```
final_model_aug <- augment(final_model)
final_model_aug <- cbind(final_model_aug, df_scaled[c("country_name")]) %>% as.tibble()
final_model_aug <- final_model_aug %>% rename(.resid= .std.resid)
autoplot(final_model, which = c(1,2,4)) + theme_bw()
```

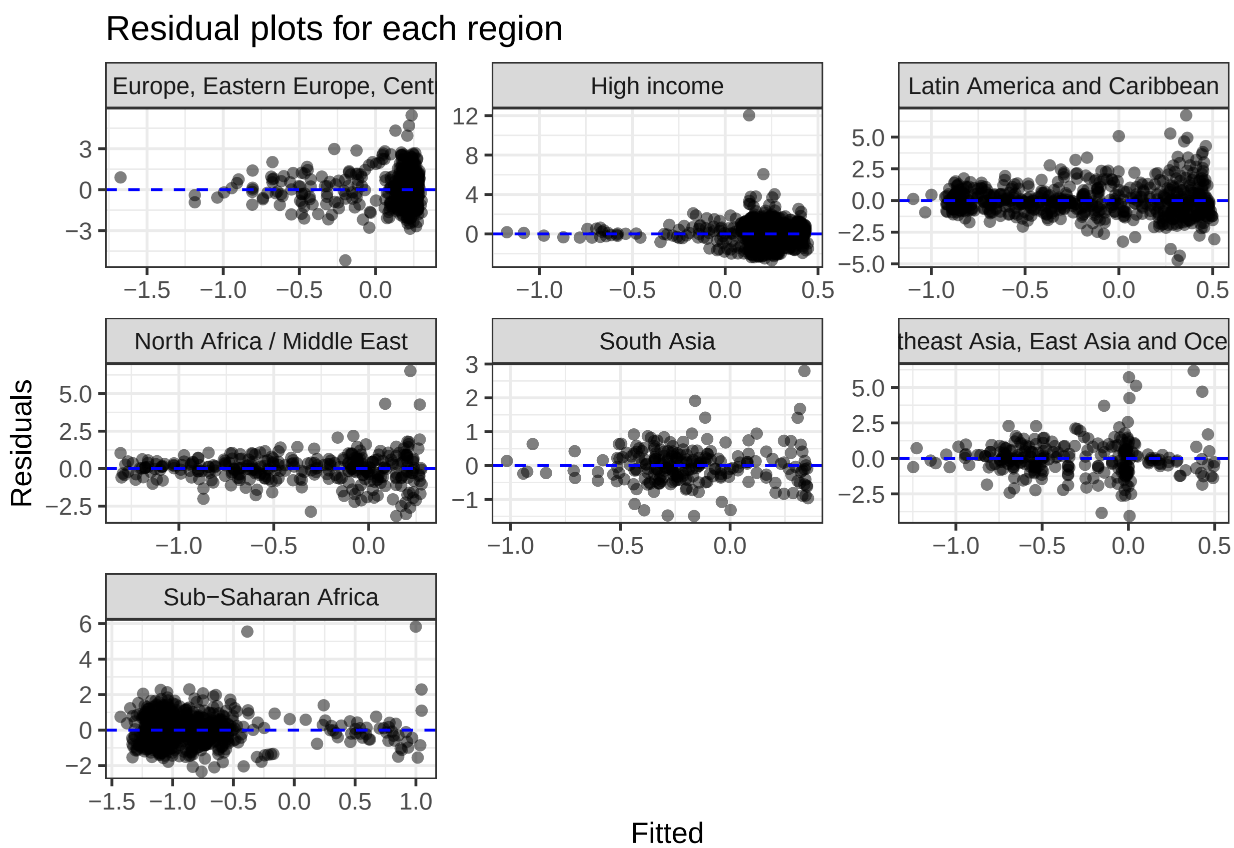


1.2 b) Model Fit Diagnostics for data in each region

Inspecting residual plots for independent regions, the region of North Africa / Mid East shows a rough pattern, with a consistent variances, suggesting that the residuals aren't homoskedastic. Conversely, apart from the region of North Africa / Mid East, the rest of regions mostly random with inconsistent variances for different fitted values, once again suggesting that the residuals aren't homoskedastic.

Furthermore, Sub-Subharan Africa and High income regions show that there are two closely clustered groups of residuals. This could possibly indicate that there is a variable that hasn't been included in this model that could explain that. There is also significant differences in the scale of the variation of the residuals from region to region, this is more evidence to suggest that there heteroskedasticity present.

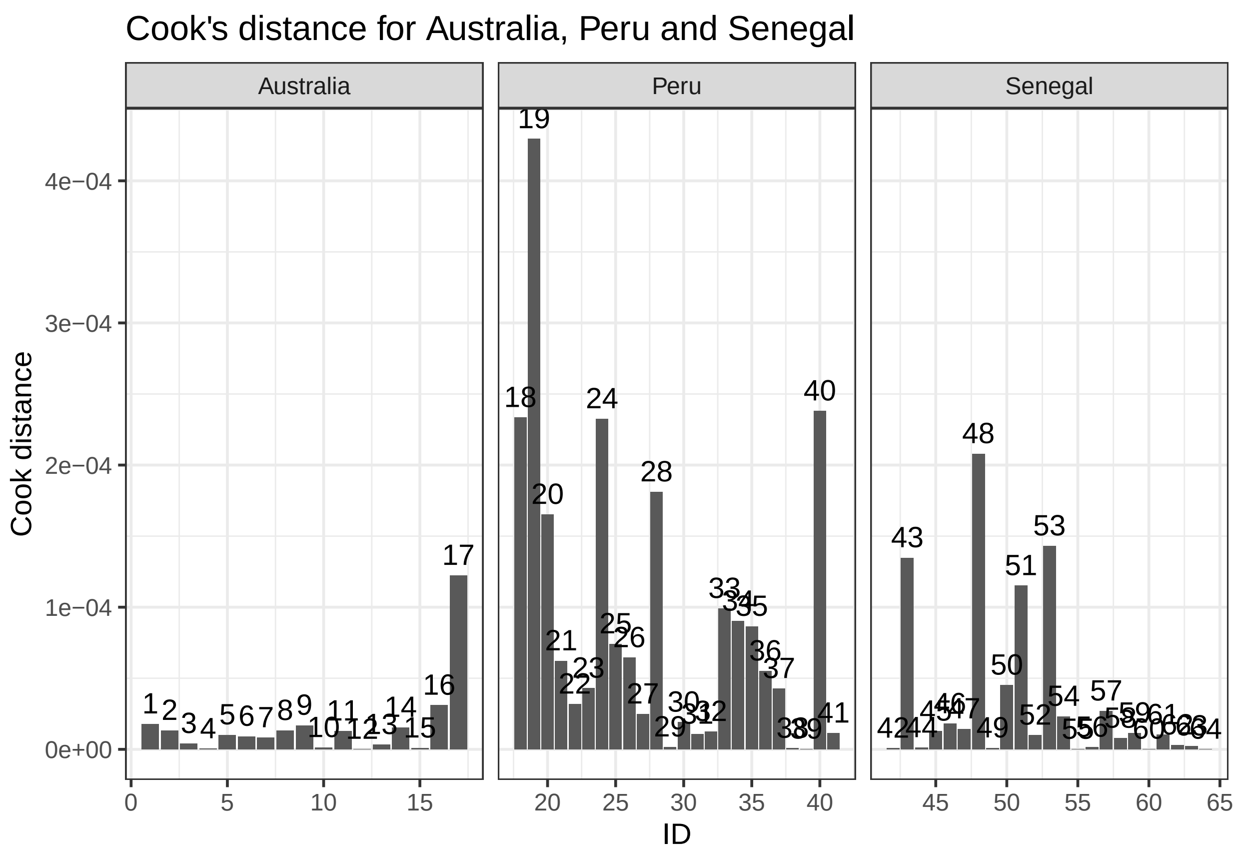
```
final_model_aug %>% ggplot(aes(x = .fitted, y = .resid)) + geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "blue") +
  facet_wrap(~ region, scales = "free") +
  theme_bw() +
  xlab("Fitted") +
  ylab("Residuals") +
  ggtitle("Residual plots for each region")
```



Inspecting the Cook's distance by country, both developing countries (Peru and Senegal) are of outliers.The Peru's outliers (the 19th, 37th, 18th, and 20th observations) as well as the Senegal's ones (the 48th, 53th, and 51th observations) are skewing thew model and effecting the models ability to give accurate estimates.

A good linear model should avoid as much outliers as possible. This warrants further investigation into those specific data points as one possible explanation could be that there are errors in how they were recorded, removal of these point could improve the overall accuracy of the models ability to estimate effectivly, however sadly we cannot just remove data because it doesn't suit us as this would present bias in the estimates.

```
#Cook distance plot by country
ggplot(data = country_model, aes(x = ID, y = .cooksd, label=ID))+geom_col()+
  facet_wrap(~ country_name, scales = "free_x") +
  geom_text(position = position_dodge(width = 0.9), vjust = -0.5) +
  theme_bw() +
  ylab("Cook distance")+
  ggtitle("Cook's distance for Australia, Peru and Senegal")
```

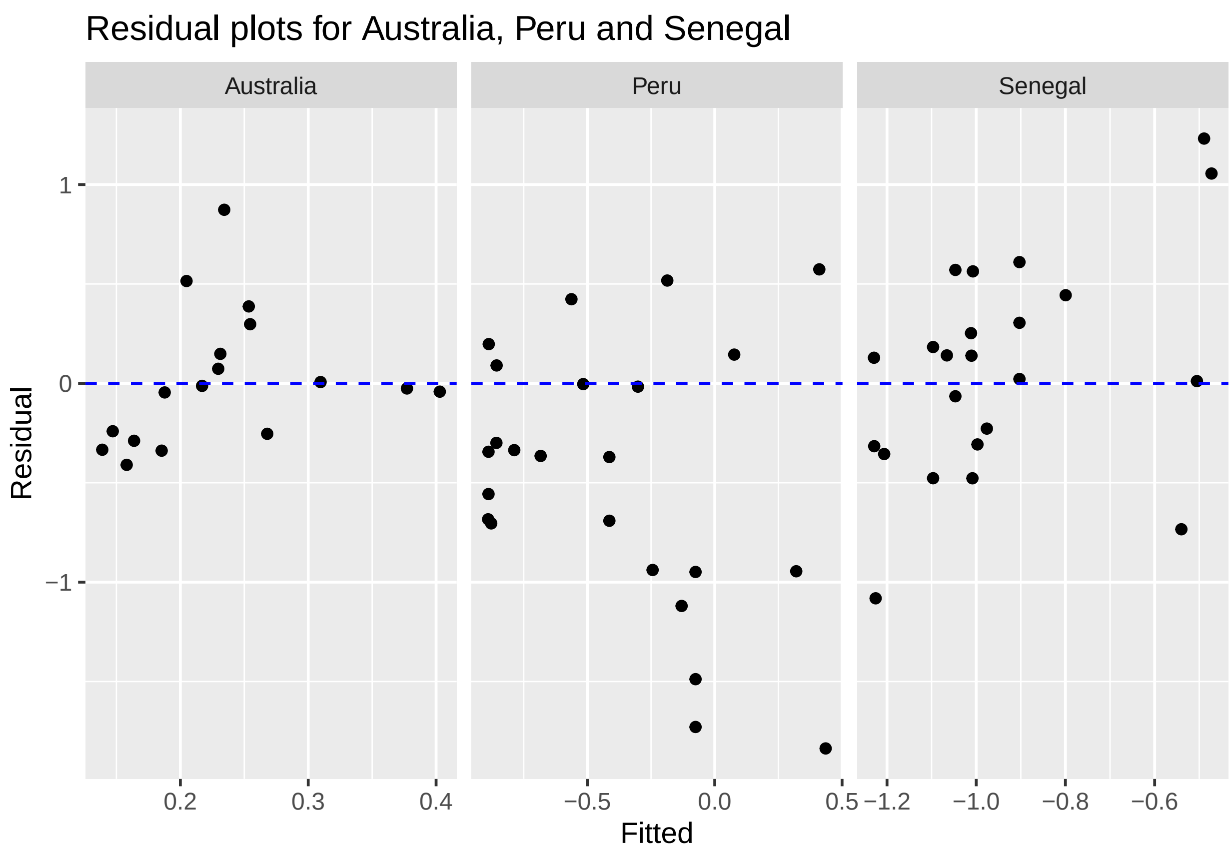


1.2 c) Model Fit Diagnostics for data in a maximum of 3 countries that should be chosen

An inspection of the residual plots for the three countries selected shows different scales of variance in the two poorer countries, Peru and Senegal.Also, the residuals for Peru don't seem to be centered around zero like Australia.

Furthermore, points of Peru seem to be mostly negative whereas the Senegal's ones are prone to be positive. The residual plot can be seen below.

```
countries <- c("Senegal", "Peru", "Australia")
country_model <-  final_model_aug %>% filter(country_name == countries)%>% rowid_to_column( "ID")
#residual plot by country
country_model%>%  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0,
             linetype = "dashed",
             colour = "blue") +
  facet_wrap(~ country_name, scales = "free_x") +
  ggtitle("Residual plots for Australia, Peru and Senegal")+
  xlab("Fitted")+
  ylab("Residual")
```




```r
df_scaled_split <- initial_split(df_scaled, strata = region)
df_scaled_test <- testing(df_scaled_split)
```

```r
df_scaled_pred <- predict(final_model, df_scaled_test)
df_scaled_pred <- tibble(pred = df_scaled_pred)
df_scaled_test_pred <- cbind(df_scaled_test, df_scaled_pred[c("pred")])
df_scaled_test_pred <- df_scaled_test_pred %>% mutate(error = nmr_log - pred,
                                                     error2 = error^2,
                                                     abs_error = abs(error))
mse <- sum(df_scaled_test_pred$error2)/nrow(df_scaled_test_pred)
mae <- sum(df_scaled_test_pred$abs_error)/nrow(df_scaled_test_pred)
```

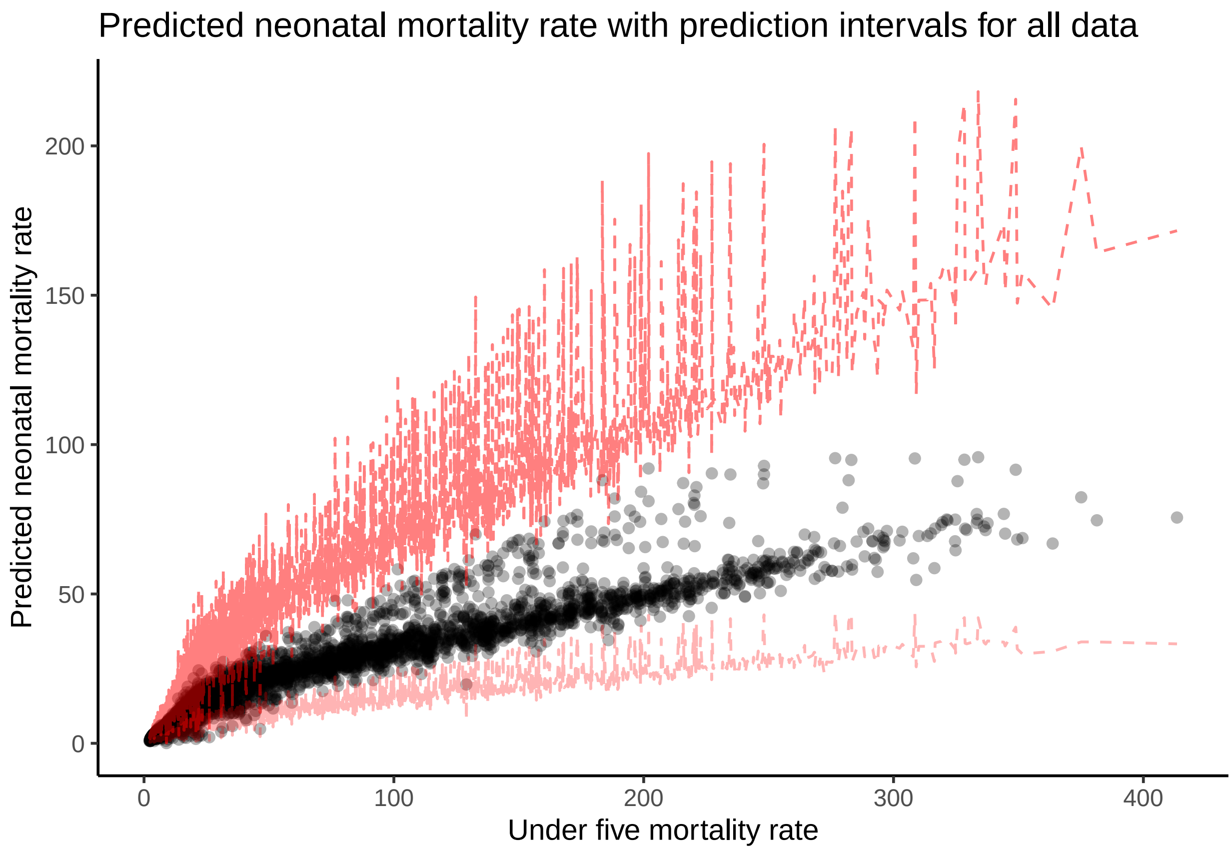
1.3) The root mean square error and the mean absolute error on a test set are 0.16 and 0.27 respectively.

```r
pred_int <- predict(final_model, df_scaled, interval = "prediction")
pred_int_tib <- tibble(pred = pred_int[,1],
                       lower = pred_int[,2],
                       upper = pred_int[,3])
pred_int_df <- cbind(df_scaled, pred_int_tib[c("pred", "lower", "upper")])
pred_int_df <- cbind(pred_int_df, final_model_aug[c("country_name")])
```

```r
pred_graph <- tibble(nmr = df_scaled$nmr,
                     u5mr = df_scaled$u5mr,
                     nmr_pred_log = pred_int_df$pred,
```

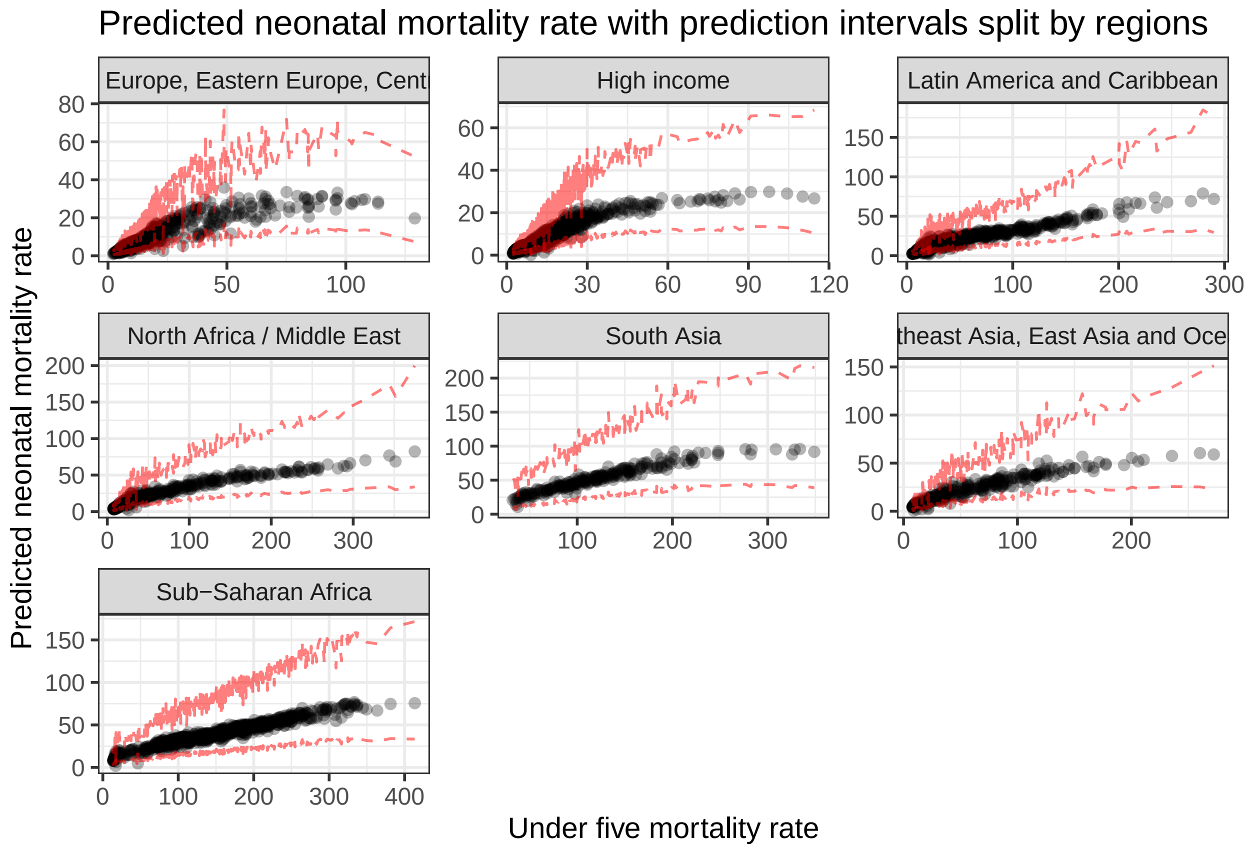
```
                     nmr_pred_lower_log = pred_int_df$lower,
                     nmr_pred_upper_log = pred_int_df$upper,
                     nmr_pred = (exp(nmr_pred_log))*(u5mr - nmr),
                     nmr_pred_lower = (exp(nmr_pred_lower_log))*(u5mr - nmr),
                     nmr_pred_upper = (exp(nmr_pred_upper_log))*(u5mr - nmr),

                     region = df_scaled$region,
                     country_name = df_scaled$country_name)
```

```
pred_graph %>% ggplot(aes(x = u5mr, y = nmr_pred)) +
  geom_point(alpha = 0.3) +
  geom_line(aes(y = nmr_pred_upper),
            linetype = "dashed",
            alpha = 0.5,
            colour = "red") +
  geom_line(aes(y = nmr_pred_lower),
            linetype = "dashed",
            alpha = 0.3,
            colour = "red") +theme_classic()+xlab("Under five mortality rate") + ylab("Predicted neonata
  ggtitle("Predicted neonatal mortality rate with prediction intervals for all data")
```

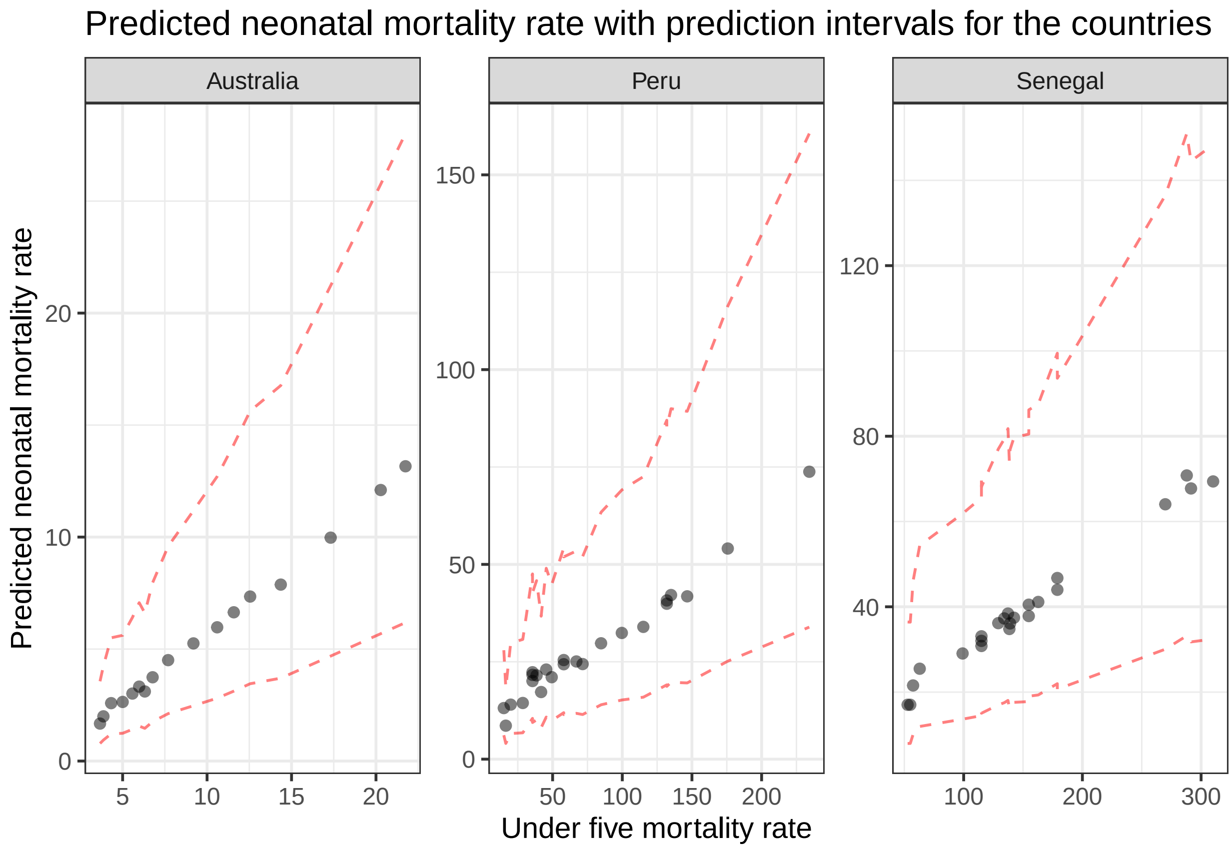



```
pred_graph %>% ggplot(aes(x = u5mr, y = nmr_pred)) + geom_point(alpha = 0.3) +
  geom_line(aes(y = nmr_pred_upper),
            linetype = "dashed",
            alpha = 0.5,
```

```
            colour = "red") +
    geom_line(aes(y = nmr_pred_lower),
            linetype = "dashed",
            alpha = 0.5,
            colour = "red") +
  facet_wrap(~ region, scales = "free")+theme_bw() + xlab("Under five mortality rate") + ylab("Predicted
  ggtitle("Predicted neonatal mortality rate with prediction intervals split by regions")
```

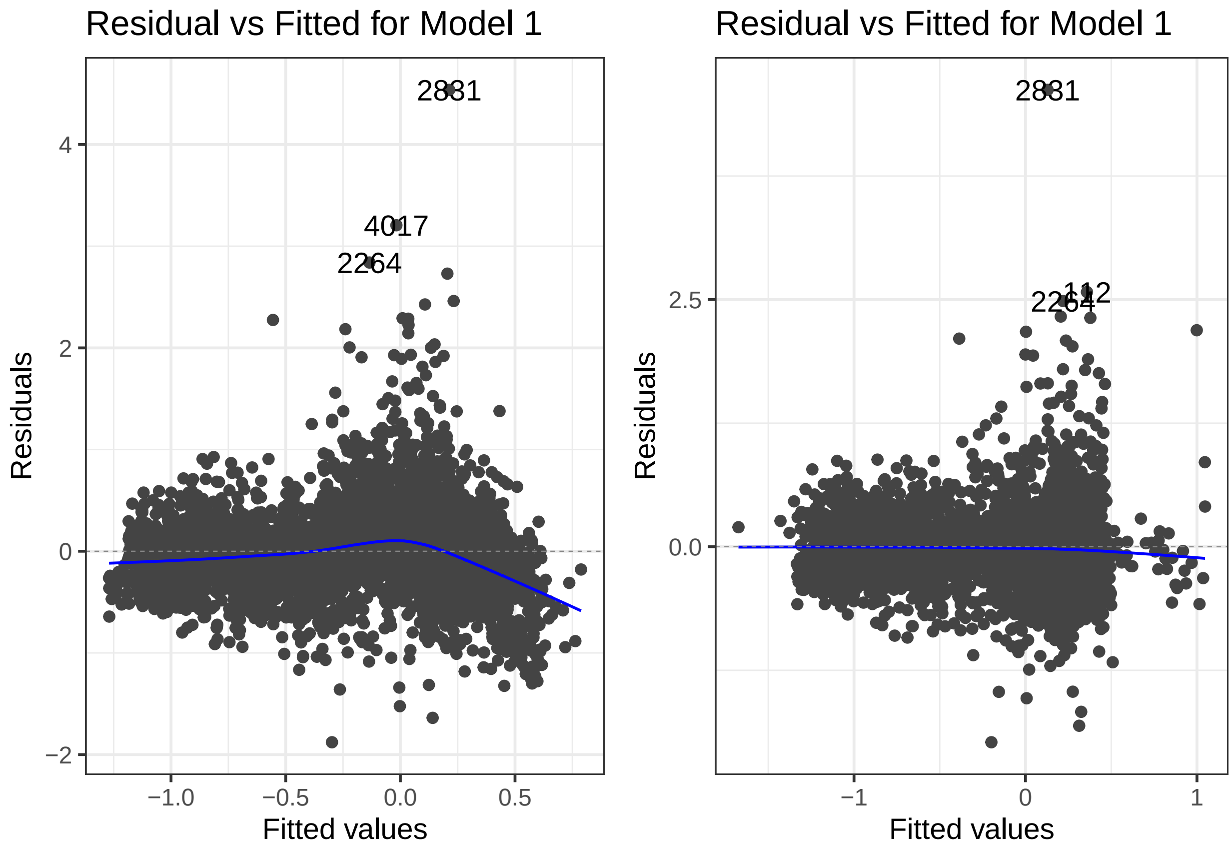



```
country_predict <-  pred_graph %>% filter(country_name == countries)%>% rowid_to_column( "ID")
country_predict %>% ggplot(aes(x = u5mr, y = nmr_pred)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(y = nmr_pred_upper),
            linetype = "dashed",
            alpha = 0.5,
            colour = "red")  +
    geom_line(aes(y = nmr_pred_lower),
            linetype = "dashed",
            alpha = 0.5,
            colour = "red") + facet_wrap(~ country_name, scales = "free")+theme_bw()+ xlab("Under five m
  ylab("Predicted neonatal mortality rate") +
  ggtitle("Predicted neonatal mortality rate with prediction intervals for the countries")
```

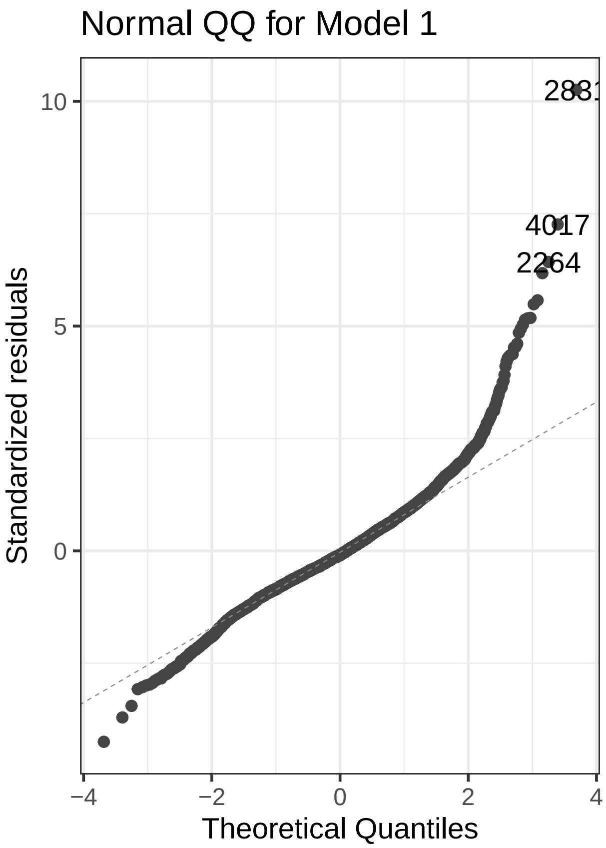


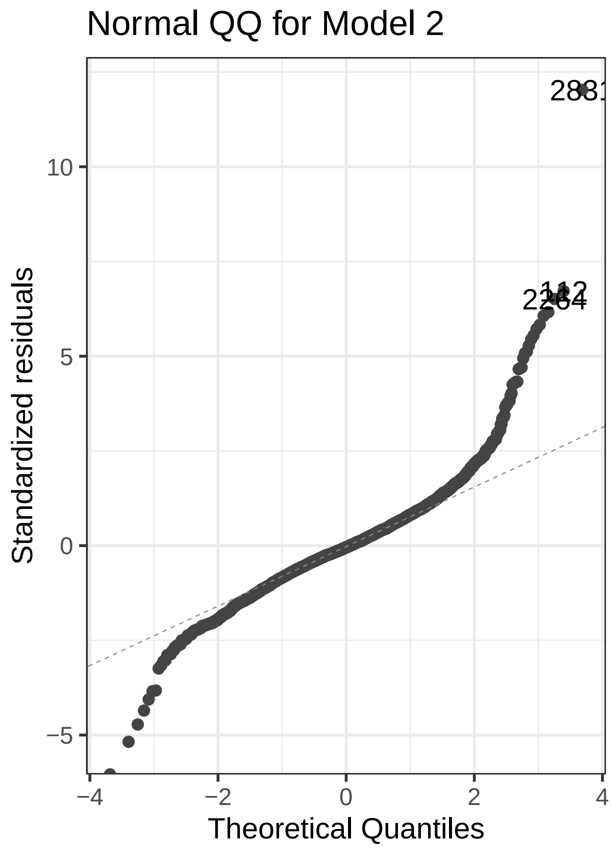
Write a paragraph or two describing the differences between the two models and explaining which you think is a more appropriate model of the data.

```
#model 1 residual plot
model1_resi <- autoplot(model_1, which = 1) +
  theme_bw()+
  ggtitle("Residual vs Fitted for Model 1")
#model 1 residual plot
model2_resi <- autoplot(final_model, which = 1) +
  theme_bw()+
  ggtitle("Residual vs Fitted for Model 1")

#model 1 qq plot
model1_qq <- autoplot(model_1, which = 2) +
  theme_bw() +
  ggtitle("Normal QQ for Model 1")
#model 2 qq plot
model2_qq <- autoplot(final_model, which = 2) +
  theme_bw()+
  ggtitle("Normal QQ for Model 2")

#combine residual plots
model1_resi+model2_resi
```




```
model1_qq + model2_qq
```





In Model 2(modified linear model), the residual plot is relatively evenly distributed as there are minimal peaks present which represent a greater difference between the actual and predicted values. The blue line on the residual plot shows the relative trend between the fitted values and their residual values.Model 1(regular linear model) has a generally uneven slope as it increases as the fitted values approach 0, and decreases after Fitted Values = 0 is reached.On the contrary Model 2(modified linear model) has a relatively flat line throughout the plot, displaying the spread of distribution being relatively even as compared to Model 1, hence a relatively less difference between the actual and predicted values in Model 2 as compared to Model 1.

When observing the Residual Standard Error(RSE) for Model 2 (RSE = 0.41) is reduced by 0.03 when shifted from Model 1(RSE = 0.44).Also supported by the bar graphs provided, the Mean Square Error(MSE) reduces by 0.02 when Model 2(MSE = 0.16) is used and the Mean Absolute Error(MAE) is also reduced by 0.03(Model 1 MAE: 0.32,Mode 2 MAE: 0.29).the r^2 value increases by 0.06 (6%), meaning the variation in mortality rates explained by the variation in year, under 5 mortality rate and region is 6% higher in Model 2(r^2 =0.61) than Model 1Model 1(r^2 = 0.55).

Finally, upon inspecting the QQ plots,Model 1(regular linear model) has a slight bend on its curve as compared to Model 2(modified linear model), therefore Model 1 is not as accurate in fitting the model with the testing data as compared to Model 2.Ideally the QQ plot should be closely fit towards the diagonal line which represents the respective regression model.Although both plots share common outliers (observations 2381 and 2264), the differing outliers(Model 1: observation 4017, Model 2: 112) highlight the variation in predicting certain observations of the two models, but overall the fitting of the modified linear model using bsplines showcases a closer fit to the regression line than the regular linear model.

From the evidence found from analysing the 3 visualisations, there is sufficient evidence suggesting that the modified linear model produces higher accuracy and therefore fits and predicts the average mortality rates better than the traditional linear model.