Week 10: Temporal data

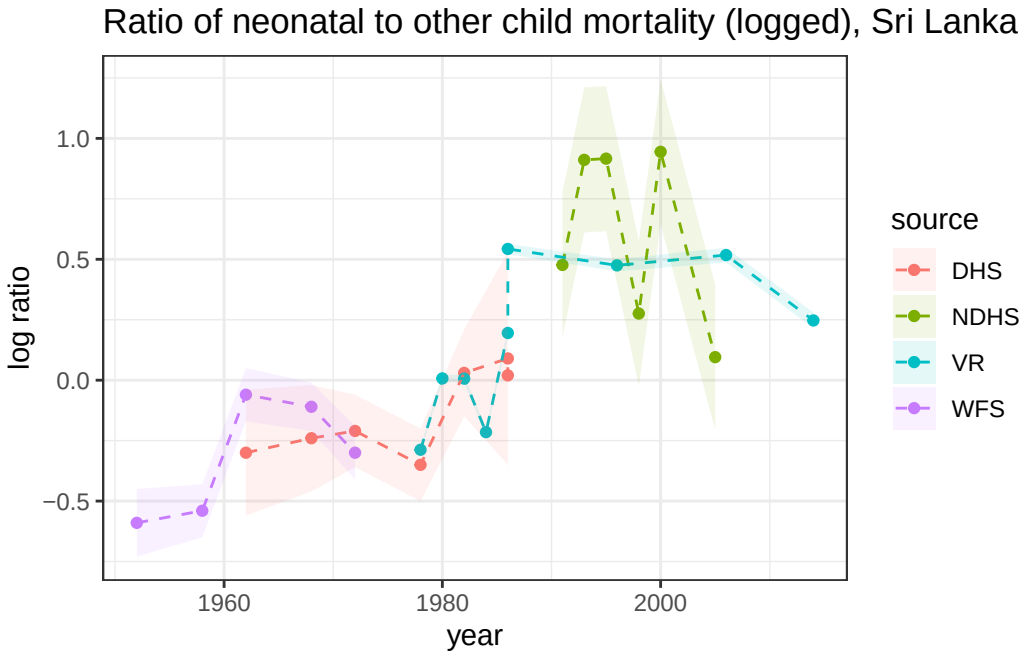
24/03/24

Child mortality in Sri Lanka

In this lab you will be fitting a couple of different models to the data about child mortality in Sri Lanka, which was used in the lecture. Here's the data and the plot from the lecture:

```
library(tidyverse)
library(here)
library(rstan)
library(tidybayes)

lka <- read_csv(here("/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka", y = "log
```




```
head(lka)
```

```
# A tibble: 6 x 7
  country_name country_code source  year logit_ratio    se ratio
  <chr>        <chr>        <chr>  <dbl>       <dbl> <dbl> <dbl>
1 Sri Lanka    LKA          WFS     1972        -0.3   0.11 0.426
2 Sri Lanka    LKA          WFS     1968        -0.11  0.1  0.473
3 Sri Lanka    LKA          WFS     1962        -0.06  0.11 0.485
4 Sri Lanka    LKA          WFS     1958        -0.54  0.11 0.368
5 Sri Lanka    LKA          WFS     1952        -0.59  0.14 0.357
6 Sri Lanka    LKA          DHS     1986         0.09  0.44 0.522
```

```
max(lka$year)
```

```
[1] 2014
```

```
2022-2014
```

```
[1] 8
```

```r
View(lka)
```

Fitting a linear model

Let's firstly fit a linear model in time to these data. Here's the code to do this:

```r
observed_years <- lka$year
years <- min(observed_years):max(observed_years)
nyears <- length(years)

stan_data <- list(y = lka$logit_ratio, year_i = observed_years - years[1]+1,
                  T = nyears, years = years, N = length(observed_years),
                  mid_year = mean(years), se = lka$se)

#mod <- stan(data = stan_data,
#             file = here("/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Stati
#save(mod, file = "/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics/la
```

```r
load("/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics/labs/model_obje
```

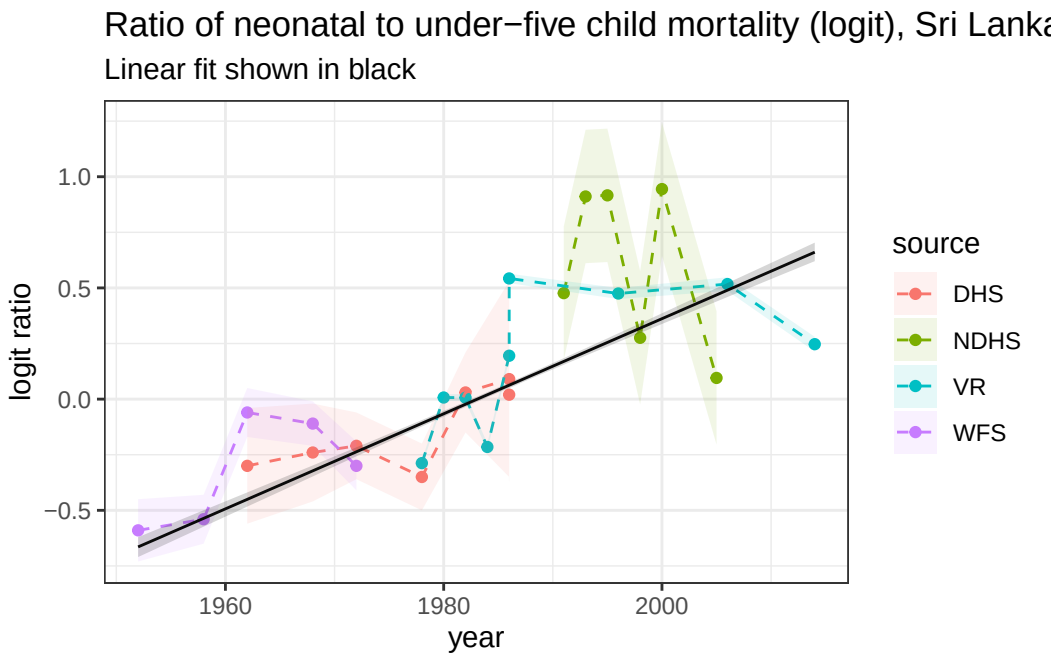
Extract the results:

```r
res <- mod %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])
```

Plot the results:

```r
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  geom_line(data = res, aes(year, .value)) +
  geom_ribbon(data = res, aes(y = .value, ymin = .lower, ymax = .upper), alpha = 0.2)+
  theme_bw()+
```

```
labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
     y = "logit ratio", subtitle = "Linear fit shown in black")
```



Question 1

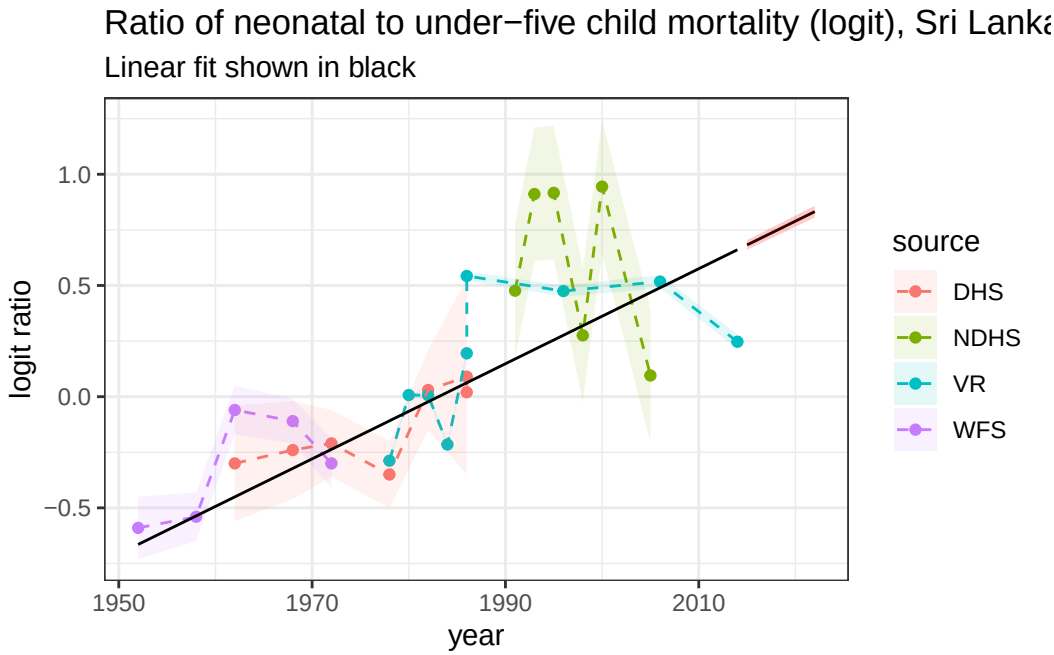
Project the linear model above out to 2022 by adding a `generated quantities` block in Stan (do the projections based on the expected value μ). Plot the resulting projections on a graph similar to that above.

```
#mod2 <- stan(data = stan_data,
#              file = here("/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Stati
#save(mod2, file = "/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics/l
```

```
load("/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics/labs/model_obje
mu_p <- extract(mod2)$mu_p
mu_p_median <- apply(mu_p, 2, median)
proj_mu<-apply(mu_p, 2, median)
proj_se<-apply(mu_p,2,sd)
yearp=seq(2015,2022,1)
proj=data.frame(year=yearp,logit_ratio=proj_mu,se=proj_se)
```

```r
lka$logit_ratio<- as.numeric(lka$logit_ratio)
lka$se<- as.numeric(lka$se)
lka$year<- as.numeric(lka$year)
```

```r
ggplot() +
  geom_point(data = lka, aes(x = year, y = logit_ratio, color = source)) +
  geom_line(data = lka, aes(x = year, y = logit_ratio, color = source), linetype = 2) +
  geom_ribbon(data = lka, aes(x = year, ymin = logit_ratio - se, ymax = logit_ratio + se,
  geom_line(data = proj, aes(x = year, y = logit_ratio), color = "black") + # Projected lo
  geom_line(data = res, aes(year, .value))+
  geom_ribbon(data = proj, aes(x = year, ymin = logit_ratio - se, ymax = logit_ratio + se)
  theme_bw() +
  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
       y = "logit ratio", subtitle = "Linear fit shown in black")
```



Question 2

The projections above are for the logit of the ratio of neonatal to under-five child mortality. You can download estimates of the under-five child mortality from 1951 to 2022 here: https://childmortality.org/all-cause-mortality/data/estimates?refArea=LKA. Use these data to get estimates and projections of neonatal mortality for Sri Lanka, and plot the results.

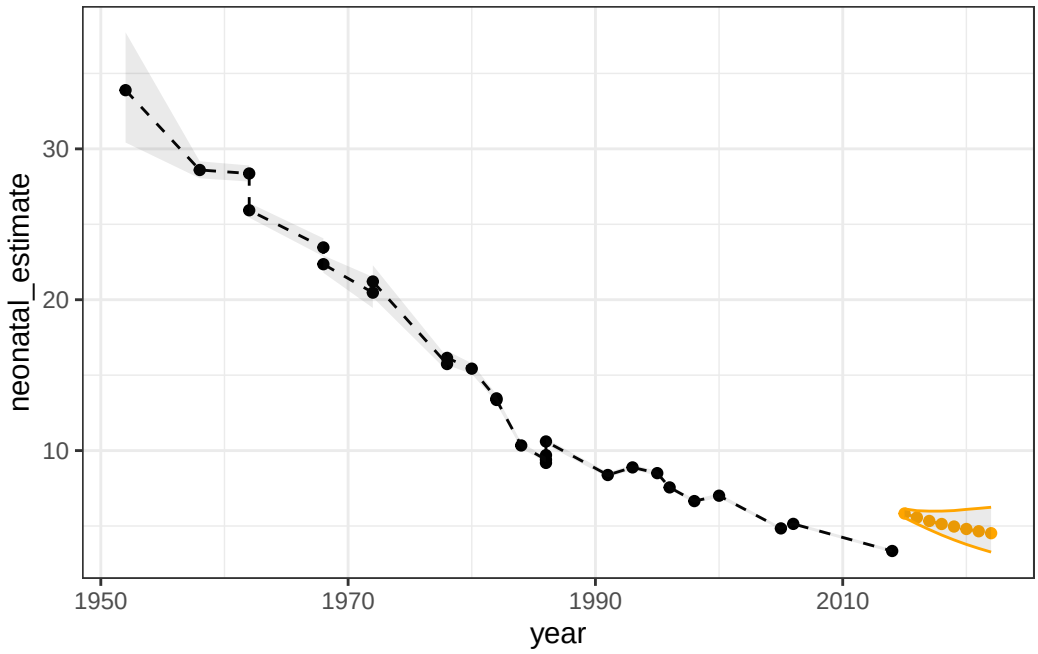
The neonatal mortality for Siri Lanka is calculated by the total mortality estimate multiplied by the rate of neonatal mortality, where the rate of neonatal morality (p) is computed from the logit_ratio log(p/1-p). The projections of neonatal mortality is calculated by first getting the projected rate of the neonatal mortality(from the median of projections from the question1), then multiply the total mortality rate.

```
lka2 <- read_csv(here("/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistic
lka2<-lka2|> rename(year=Year,estimate=Estimate)
lka3<-inner_join(lka,lka2,by="year")
lka3$country_name<-NULL
lka4<-lka3|>mutate(rate=ratio/(1+ratio),neonatal_estimate=estimate*rate,lower=`Lower bound
View(lka4)
```

```
proj2<-inner_join(proj,lka2[1:8,]|>arrange(year),by="year")

proj2<-proj2|>mutate(rate = exp(logit_ratio)/(1+exp(logit_ratio))
                     ,lower=`Lower bound`*rate,upper=`Upper bound`*rate,neonatal_estimate=r

ggplot() +
  geom_point( data = lka4,aes(x = year, y = neonatal_estimate))+
  geom_line(data = lka4, aes(x = year, y = neonatal_estimate), linetype = 2)+theme_bw()+
  geom_ribbon(data = lka4, aes(x = year, ymin = lower, ymax = upper), alpha = 0.1) +
  geom_point( data = proj2,aes(x = year, y = neonatal_estimate),color="orange")+
  geom_line(data = proj2, aes(x = year, y = neonatal_estimate), linetype = 2,color="orange
  geom_ribbon(data = proj2, aes(x = year, ymin = lower, ymax = upper), alpha = 0.1,color="
```



Random walks

Question 3

Code up and estimate a first order random walk model to fit to the Siri Lankan data, taking into account measurement error, and project out to 2022.

The figure below shows the estimate of first order random walk. The redline is the projected line, whereas the black line is the fitted RW1, and the blue shadows represents the measurement errors.

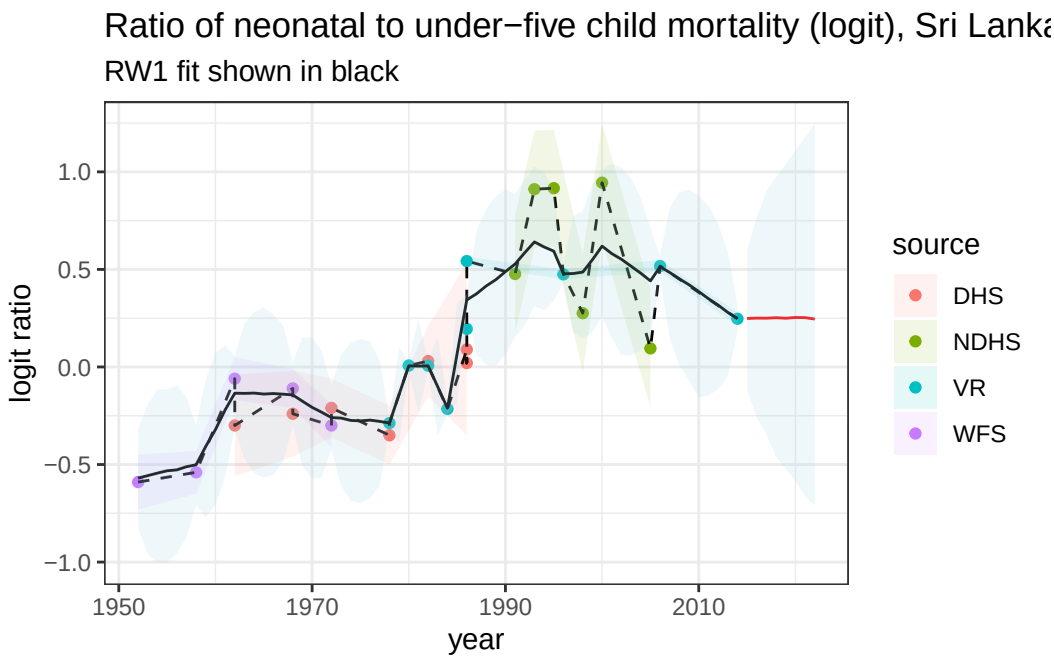
```
stan_data2 <- list(y = lka$logit_ratio, year_i = observed_years - years[1]+1,
                   T = nyears, years = years, N = length(observed_years),
                   mid_year = mean(years), se = lka$se,P=8)

#mod3<-stan(data=stan_data2,
#           seed=100,
#           file ="~/Documents/2024/STA2201-Methods-of-Applied-Statistics/labs/lab10_mode2
#save(mod3, file = "/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics/l
load("/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics/labs/model_obje
res2 <- mod3 |>
  gather_draws(mu[t]) |>
```

```
  median_qi() |>
  mutate(year = years[t])

proj3 <- mod3 |>
    gather_draws(mu_new[p])|>
    median_qi() |>
    mutate(year = years[nyears]+p)
```

```
ggplot() +
  geom_point(data = lka, aes(x = year, y = logit_ratio, color = source)) +
  geom_line(data = lka, aes(x = year, y = logit_ratio), linetype = 2) +
  geom_ribbon(data = lka, aes(x = year, ymin = logit_ratio - se, ymax = logit_ratio + se,
  geom_line(data = res2, aes(x = year, y = .value), color = "black") +
  geom_line(data = proj3, aes(x = year, y = .value), color = "red") + # Projected logit ra
  geom_ribbon(data = res2, aes(x = year, ymin = .lower, ymax = .upper),fill="lightblue", a
  geom_ribbon(data = proj3, aes(x = year, ymin = .lower, ymax = .upper), fill="lightblue",
  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
       y = "logit ratio", subtitle = "RW1 fit shown in black") +
  theme_bw()
```



Question 4

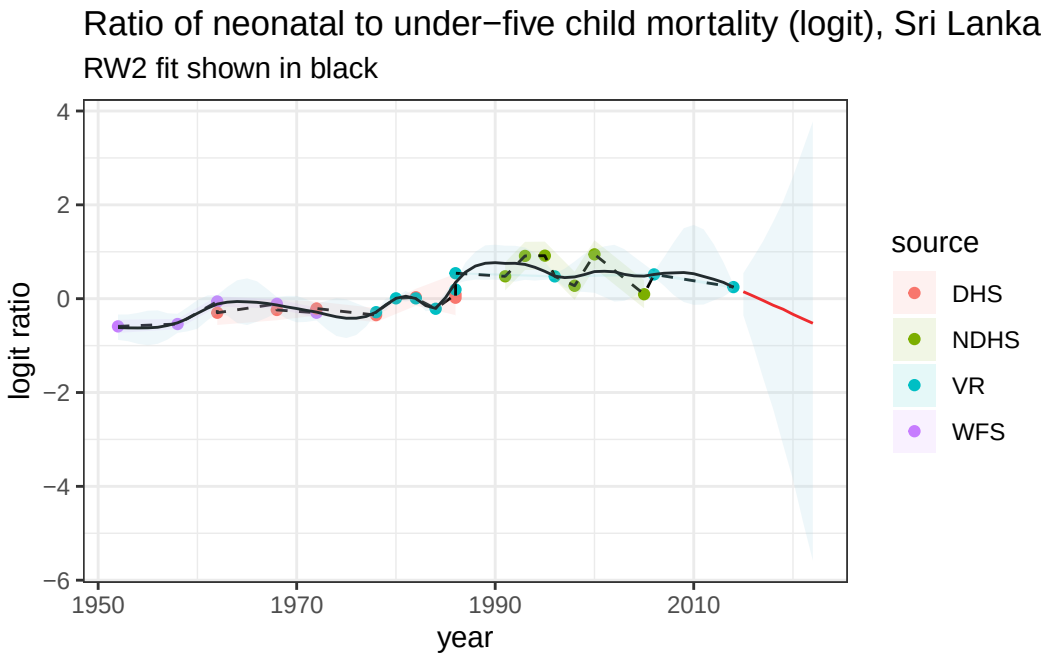
Now alter your model above to estimate and project a second-order random walk model (RW2).

The figure below shows the estimate of first order random walk. The redline is the projected line, whereas the black line is the fitted RW1, and the blue shadows represents the measurement errors.

```
#mod4<-stan(data=stan_data2,
#           seed=100,
#           file ="~/Documents/2024/STA2201-Methods-of-Applied-Statistics/labs/lab10_mod3.
#save(mod4, file = "/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics/l
load("/Users/hainanxu/Documents/2024/STA2201-Methods-of-Applied-Statistics/labs/model_obje
res3 <- mod4 |>
  gather_draws(mu[t]) |>
  median_qi() |>
  mutate(year = years[t])

proj4 <- mod4 |>
    gather_draws(mu_new[p])|>
    median_qi() |>
    mutate(year = years[nyears]+p)

ggplot() +
  geom_point(data = lka, aes(x = year, y = logit_ratio, color = source)) +
  geom_line(data = lka, aes(x = year, y = logit_ratio), linetype = 2) +
  geom_ribbon(data = lka, aes(x = year, ymin = logit_ratio - se, ymax = logit_ratio + se,
  geom_line(data = res3, aes(x = year, y = .value), color = "black") +
  geom_line(data = proj4, aes(x = year, y = .value), color = "red") + # Projected logit ra
  geom_ribbon(data = res3, aes(x = year, ymin = .lower, ymax = .upper),fill="lightblue", a
  geom_ribbon(data = proj4, aes(x = year, ymin = .lower, ymax = .upper), fill="lightblue",
  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
       y = "logit ratio", subtitle = "RW2 fit shown in black") +
  theme_bw()
```



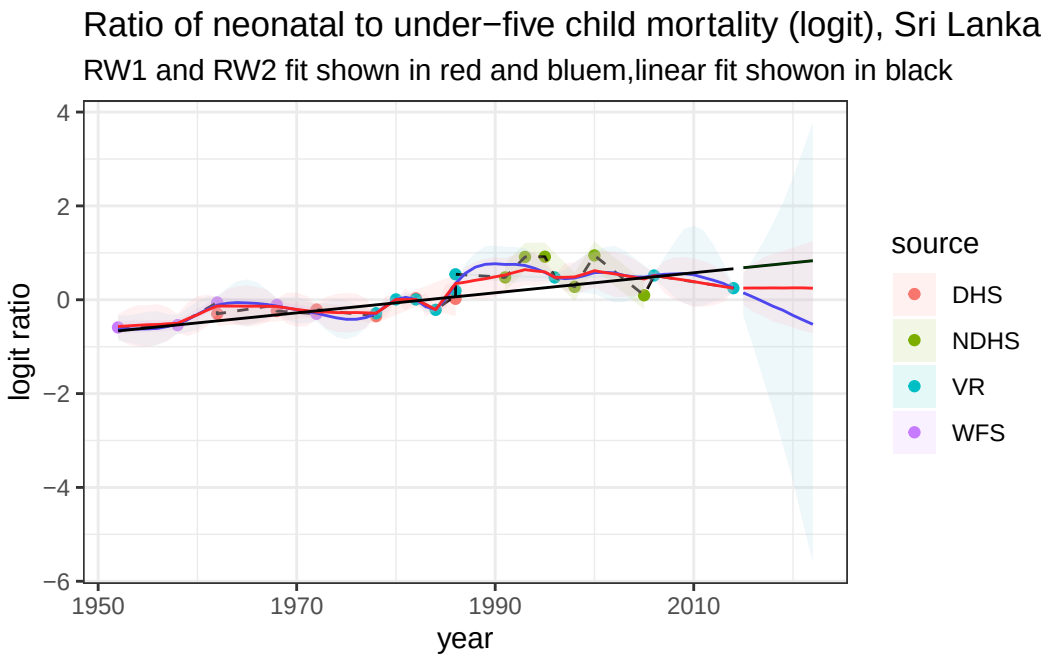
Question 5

Run the first order and second order random walk models, including projections out to 2022. Compare these estimates with the linear fit by plotting everything on the same graph.

```
ggplot() +
  geom_point(data = lka, aes(x = year, y = logit_ratio, color = source)) +
  geom_line(data = lka, aes(x = year, y = logit_ratio), linetype = 2) +
  geom_ribbon(data = lka, aes(x = year, ymin = logit_ratio - se, ymax = logit_ratio + se,
  geom_line(data = res3, aes(x = year, y = .value), color = "blue") +
  geom_line(data = proj4, aes(x = year, y = .value), color = "blue") + # Projected logit r
  geom_ribbon(data = res3, aes(x = year, ymin = .lower, ymax = .upper),fill="lightblue", a
  geom_ribbon(data = proj4, aes(x = year, ymin = .lower, ymax = .upper), fill="lightblue",
  geom_line(data = res2, aes(x = year, y = .value), color = "red") +
  geom_line(data = proj3, aes(x = year, y = .value), color = "red") + # Projected logit ra
  geom_ribbon(data = res2, aes(x = year, ymin = .lower, ymax = .upper),fill="lightpink", a
  geom_ribbon(data = proj3, aes(x = year, ymin = .lower, ymax = .upper), fill="lightpink",
    geom_line(data = proj, aes(x = year, y = logit_ratio), color = "black") + # Projected
  geom_line(data = res, aes(year, .value))+
  geom_ribbon(data = proj, aes(x = year, ymin = logit_ratio - se, ymax = logit_ratio + se)

  labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
```

```
      y = "logit ratio", subtitle = "RW1 and RW2 fit shown in red and bluem,linear fit sh
  theme_bw()
```



Question 6

Briefly comment on which model you think is most appropriate, or an alternative model that would be more appropriate in this context.

The figure above shows the the first order and second order random walk models, including projections out to 2022. We can see that all tree model captures the general trend. The black line represents the linear fit, the red line represents RW1 whereas the blue line represents RW2. We observe that the RW2 model have the most broad confidence interval whereas the linear fit has smallest CI. We think RW1 is the most appropriate model, since it not only captures the general trend, but also have a comparatively smaller CI. If we compare the three model in a smaller scale, we can see that the projection for RW1 is more realistic. Based on the projection line of RW2, it is indicating that the neonatal mortality rate in 2022 is very likely to fall back to 1950, which is very unlikely. Therefore, RW1 is the most appropriate model asmong the three, it not only captures the fluctuations in the data, but also gives a realistic projection line.

```
ggplot() +
  geom_point(data = lka, aes(x = year, y = logit_ratio, color = source)) +
```

```
 geom_line(data = lka, aes(x = year, y = logit_ratio), linetype = 2) +
 geom_ribbon(data = lka, aes(x = year, ymin = logit_ratio - se, ymax = logit_ratio + se,
 geom_line(data = res3, aes(x = year, y = .value), color = "blue") +
 geom_line(data = proj4, aes(x = year, y = .value), color = "blue") + # Projected logit r
 geom_ribbon(data = res3, aes(x = year, ymin = .lower, ymax = .upper),fill="lightblue", a
 #geom_ribbon(data = proj4, aes(x = year, ymin = .lower, ymax = .upper), fill="lightblue"
 geom_line(data = res2, aes(x = year, y = .value), color = "red") +
 geom_line(data = proj3, aes(x = year, y = .value), color = "red") + # Projected logit ra
# geom_ribbon(data = res2, aes(x = year, ymin = .lower, ymax = .upper),fill="lightpink",
# geom_ribbon(data = proj3, aes(x = year, ymin = .lower, ymax = .upper), fill="lightpink"
   geom_line(data = proj, aes(x = year, y = logit_ratio), color = "black") + # Projected
 geom_line(data = res, aes(year, .value))+
 #geom_ribbon(data = proj, aes(x = year, ymin = logit_ratio - se, ymax = logit_ratio + se

 labs(title = "Ratio of neonatal to under-five child mortality (logit), Sri Lanka",
      y = "logit ratio", subtitle = "RW1 and RW2 fit shown in red and bluem,linear fit sh
 theme_bw()
```