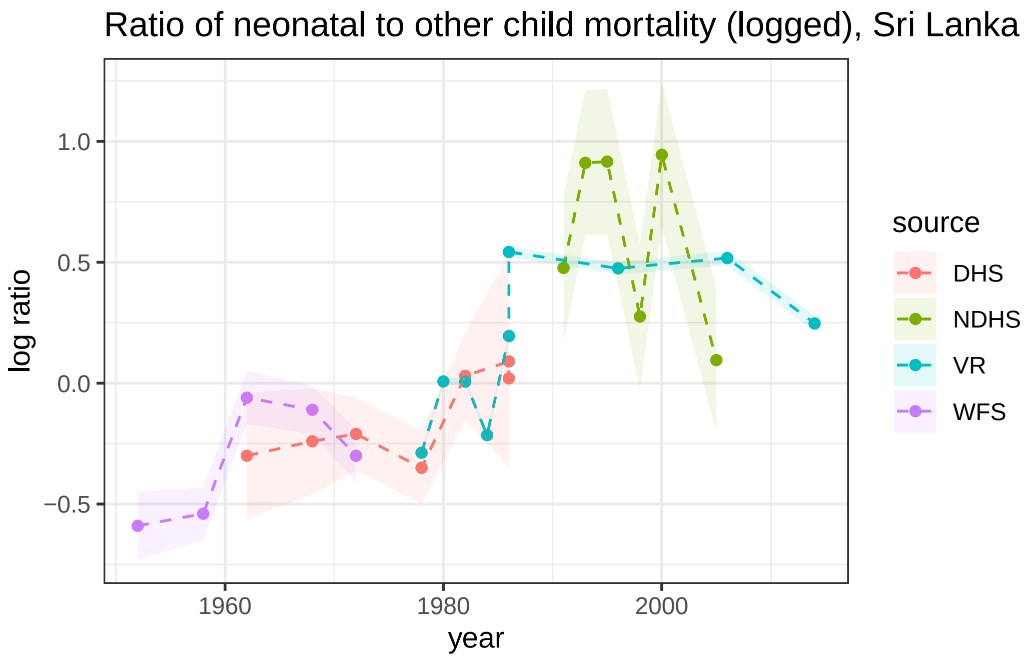
Lab 10: Temporal data

TJ Ayoub

24/03/23

Pre-lab Setup

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

lka <- read_csv(here("lka.csv"))
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
```



Ratio of neonatal to other child mortality (logged), Sri Lanka

Fitting a linear model

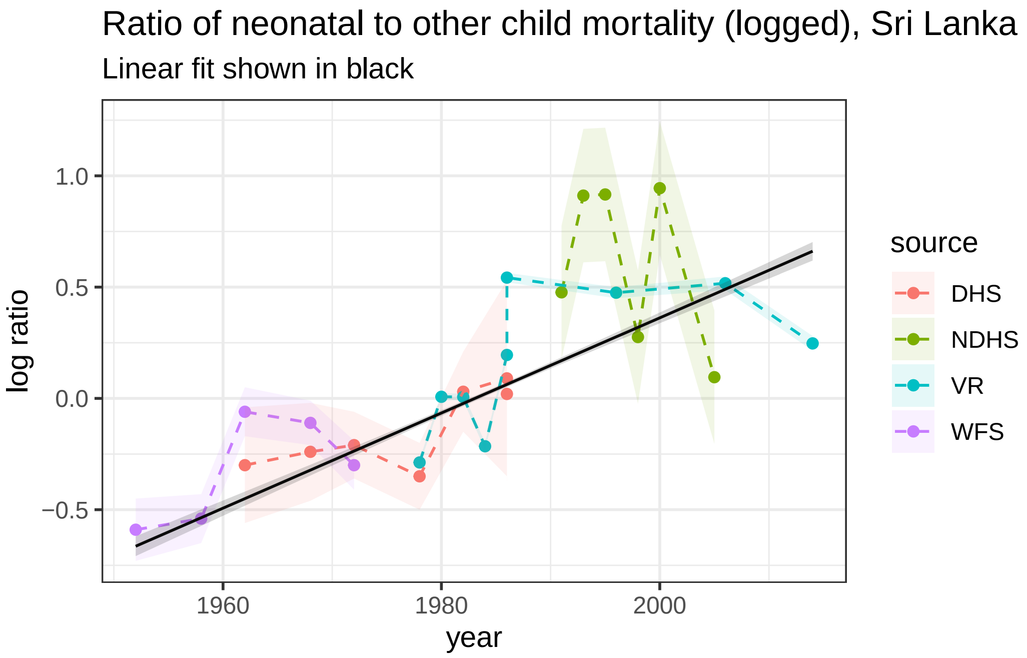
```
observed_years <- lka$year
years <- min(observed_years):max(observed_years)
nyears <- length(years)

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("lka_linear_me.stan"))
```

Extract the results:

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

Plot the results:

```
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 other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear fit shown in black")
```



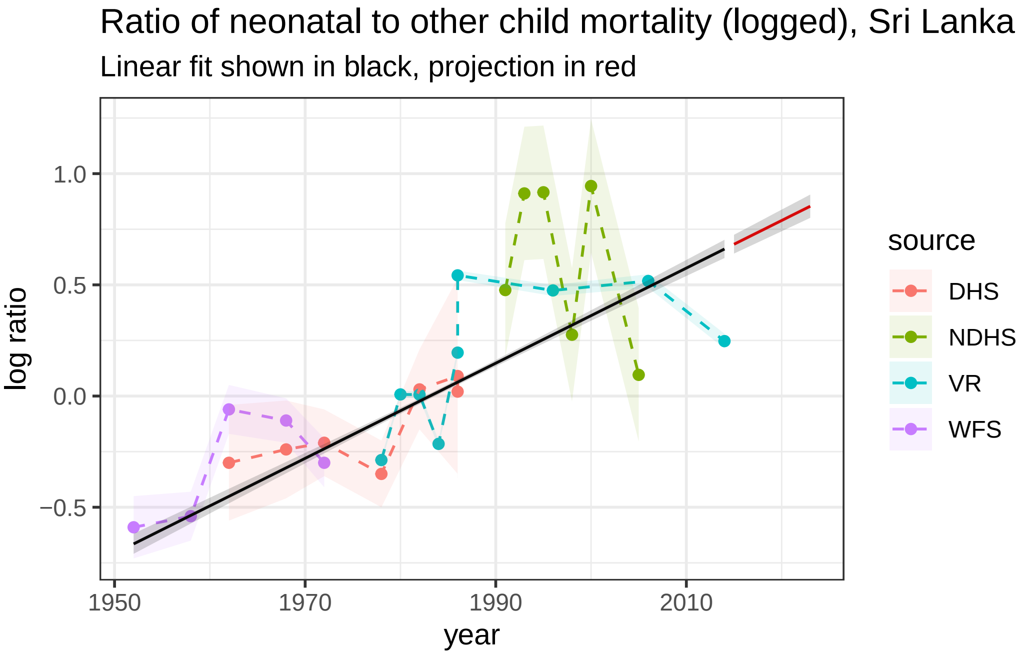
Question 1

- Project the linear model above out to 2023 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.

```
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,
                  P = 9)

mod2 <- stan(data = stan_data,
             file = here("10q1.stan"))
```

```
res_q1 <- mod2 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])

res_p_q1 <- mod2 %>%
  gather_draws(mu_p[p]) %>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

```
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_q1, aes(year, .value)) +
  geom_ribbon(data = res_q1, aes(y = .value,
                                 ymin = .lower,
                                 ymax = .upper),
              alpha = 0.2) +
  geom_line(data = res_p_q1, aes(year, .value), col='red') +
  geom_ribbon(data = res_p_q1, aes(y = .value,
                                         ymin = .lower,
                                         ymax = .upper),
                  alpha = 0.2) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear fit shown in black, projection in red")
```



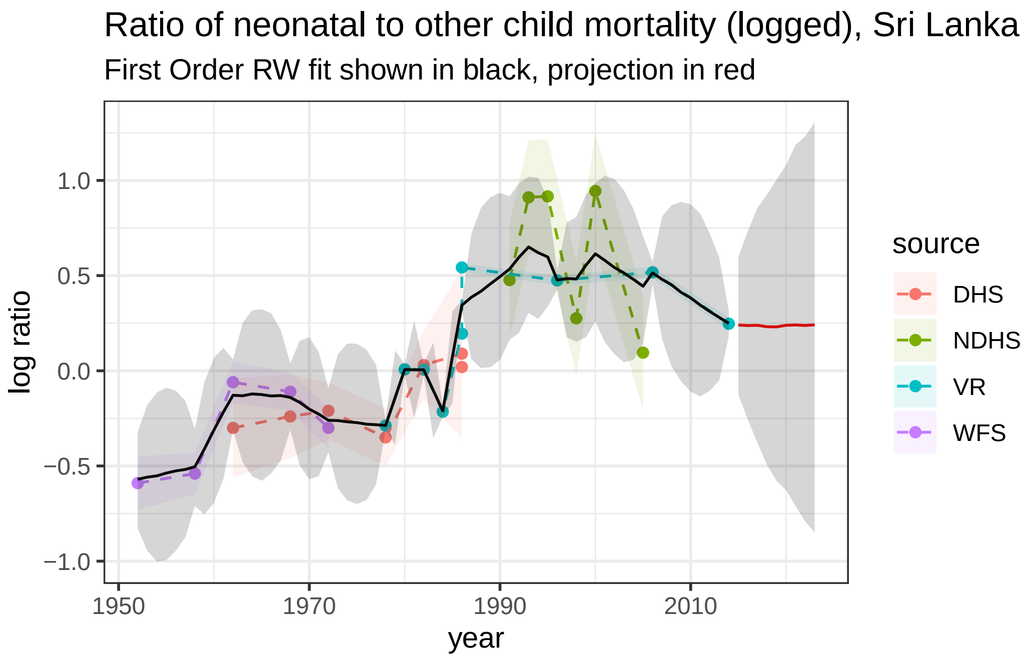
Question 2

- Code up and estimate a first order random walk model to fit to the Sri Lankan data, taking into account measurement error, and project out to 2023.

```
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,
                  P = 9)

mod3 <- stan(data = stan_data,
             file = here("10q2.stan"))
```

```r
res_q2 <- mod3 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])

res_p_q2 <- mod3 %>%
  gather_draws(mu_p[p]) %>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

```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_q2, aes(year, .value)) +
  geom_ribbon(data = res_q2, aes(y = .value, ymin = .lower,
                                 ymax = .upper), alpha = 0.2) +
  geom_line(data = res_p_q2, aes(year, .value), col='red') +
  geom_ribbon(data = res_p_q2, aes(y = .value, ymin = .lower,
                                   ymax = .upper), alpha = 0.2) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "First Order RW fit shown in black, projection in red")
```



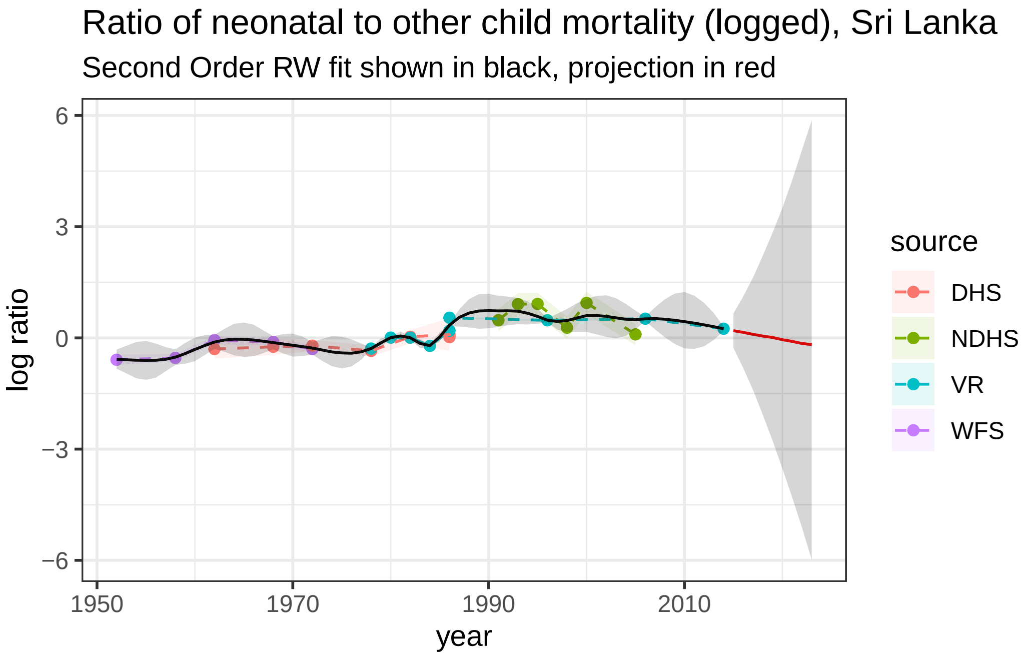
Question 3

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

```
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,
                  P = 9)

mod4 <- stan(data = stan_data,
             file = here("10q3.stan"))
```

```
res_q3 <- mod4 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])

res_p_q3 <- mod4 %>%
  gather_draws(mu_p[p]) %>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

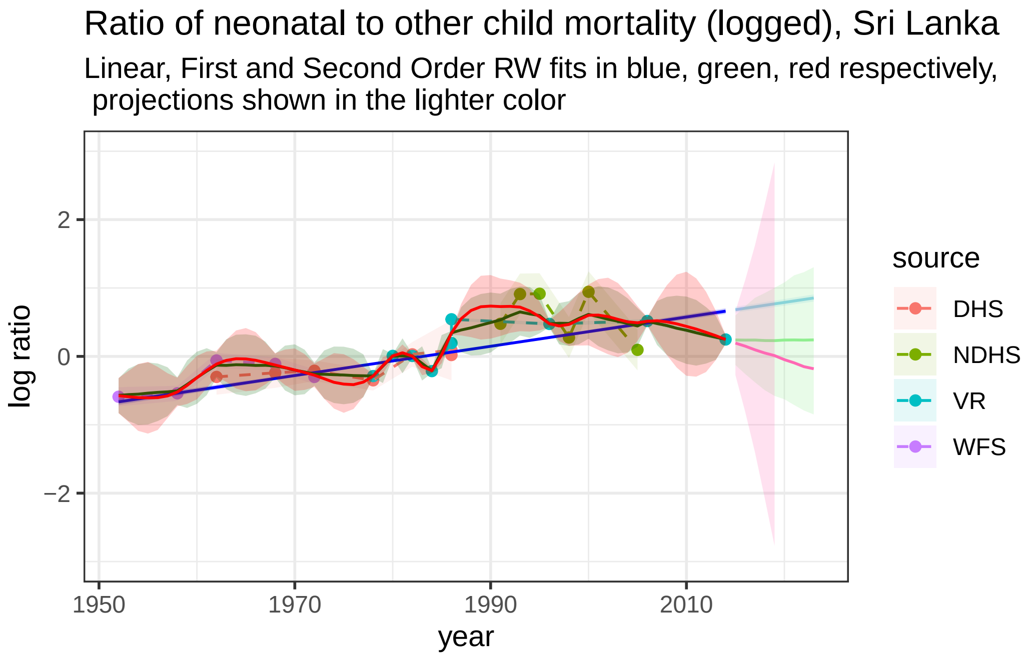
```
ggplot(lka, aes(year, logit_ratio)) +
  geom_point(aes( color = source)) +
  geom_line(aes( color = source), lty = 2) +
  geom_ribbon(aes(ymin = logit_ratio - se,
                  ymax = logit_ratio + se,
                  fill =  source), alpha = 0.1) +
  theme_bw()+
  geom_line(data = res_q3, aes(year, .value)) +
  geom_ribbon(data = res_q3, aes(y = .value, ymin = .lower,
                                 ymax = .upper), alpha = 0.2) +
  geom_line(data = res_p_q3, aes(year, .value), col='red') +
  geom_ribbon(data = res_p_q3, aes(y = .value, ymin = .lower,
                                   ymax = .upper), alpha = 0.2) +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Second Order RW fit shown in black, projection in red"
```



Question 4

- Run the first order and second order random walk models, including projections out to 2023. Compare these estimates with the linear fit by plotting everything on the same graph.

```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_q1, aes(year, .value), col='blue') +
  geom_ribbon(data = res_q1, aes(y = .value, ymin = .lower,
                                  ymax = .upper), alpha = 0.2, fill='blue') +
  geom_line(data = res_p_q1, aes(year, .value), col='skyblue') +
  geom_ribbon(data = res_p_q1, aes(y = .value, ymin = .lower,
                                    ymax = .upper), alpha = 0.2, fill='skyblue')+
  geom_line(data = res_q2, aes(year, .value), col = "darkgreen") +
  geom_ribbon(data = res_q2, aes(y = .value, ymin = .lower,
                                  ymax = .upper), alpha = 0.2, fill = "darkgreen") +
  geom_line(data = res_p_q2, aes(year, .value), col='lightgreen') +
  geom_ribbon(data = res_p_q2, aes(y = .value, ymin = .lower,
                                    ymax = .upper), alpha = 0.2, fill='lightgreen') +
  geom_line(data = res_q3, aes(year, .value), col='red') +
  geom_ribbon(data = res_q3, aes(y = .value, ymin = .lower,
                                  ymax = .upper), alpha = 0.2, fill = "red") +
  geom_line(data = res_p_q3, aes(year, .value), col='hotpink') +
  geom_ribbon(data = res_p_q3, aes(y = .value, ymin = .lower,
                                    ymax = .upper), alpha = 0.2, fill = 'hotpink') +
  theme_bw()+
  labs(title = "Ratio of neonatal to other child mortality (logged), Sri Lanka",
       y = "log ratio", subtitle = "Linear, First and Second Order RW fits in blue, green,
  ylim(-3,3)
```



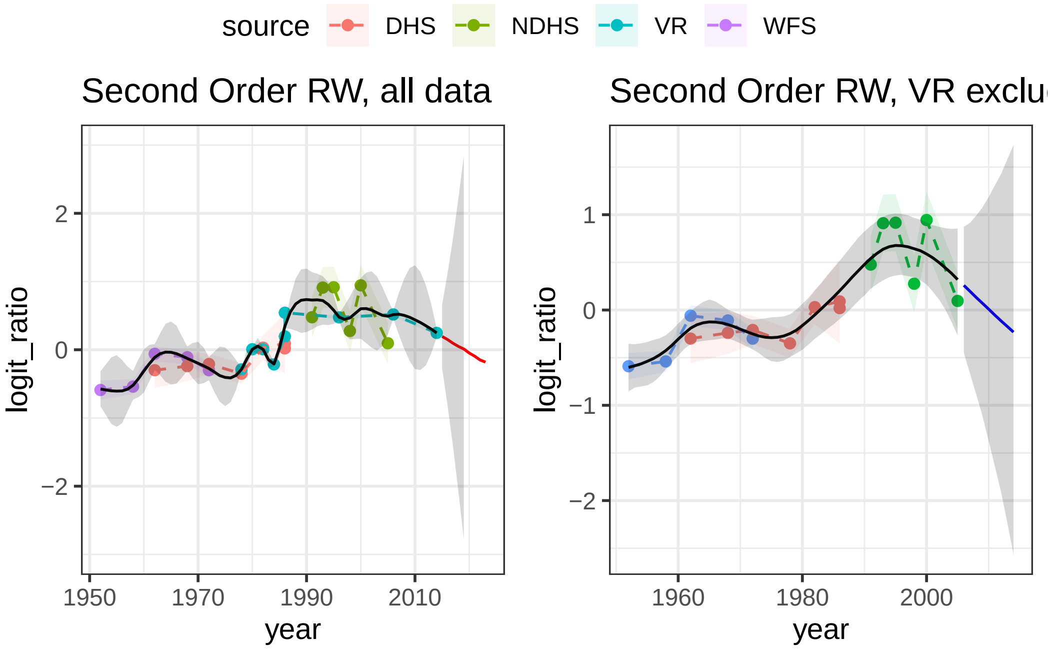
Question 5

- Rerun the RW2 model excluding the VR data. Briefly comment on the differences between the two data situations.

```
lka_exclude = lka |> filter(source != "VR")
observed_years <- lka_exclude$year
years <- min(observed_years):max(observed_years)
nyears <- length(years)

stan_data <- list(y = lka_exclude$logit_ratio,
                  year_i = observed_years - years[1]+1,
                  T = nyears, years = years,
                  N = length(observed_years),
                  mid_year = mean(years), se = lka_exclude$se,
                  P = 9)

mod5 <- stan(data = stan_data,
             file = here("10q3.stan"))
```

```
res_q5 <- mod5 %>%
  gather_draws(mu[t]) %>%
  median_qi() %>%
  mutate(year = years[t])

res_p_q5 <- mod5 %>%
  gather_draws(mu_p[p]) %>%
  median_qi() %>%
  mutate(year = years[nyears]+p)
```

```
p1 = 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()+ ylim(-3,3)+
  geom_line(data = res_q3, aes(year, .value)) +
  geom_ribbon(data = res_q3, aes(y = .value, ymin = .lower,
                                 ymax = .upper), alpha = 0.2) +
  geom_line(data = res_p_q3, aes(year, .value), col='red') +
  geom_ribbon(data = res_p_q3, aes(y = .value, ymin = .lower,
                                   ymax = .upper), alpha = 0.2) +
  labs(title = "Second Order RW, all data")

p2 = ggplot(lka_exclude, 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_q5, aes(year, .value)) +
  geom_ribbon(data = res_q5, aes(y = .value, ymin = .lower,
                                 ymax = .upper), alpha = 0.2) +
  geom_line(data = res_p_q5, aes(year, .value), col='blue') +
  geom_ribbon(data = res_p_q5, aes(y = .value, ymin = .lower,
                                   ymax = .upper), alpha = 0.2) +
  labs(title = "Second Order RW, VR excluded")

ggarrange(p1,p2, common.legend = TRUE)
```



When all of the data is being considered the second order RW model is projecting a downwards trend, contrary to the linear positive slope fit, and the first order sideways projection. However, the uncertainty of this projection is “equal” on both sides. That is the model is equally uncertain about whether the true ratio will be larger or smaller than its projection.

Now when we omit the VR data, the second order model still projects downwards, but the error ribbons are no longer symmetric and skew downwards. The model looks to be more sure that the true value is smaller when compared to the model fit to all of the data. This means the omitted VR data holds leverage in fitting to larger projections.

Question 6

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

Looking at the data with a very coarse eye one would likely say that there is a positive relationship. For capturing the macro trend, the linear model is doing the best job. The RW models, especially being second and first order models, are really only modeling and projecting on such a small time dependence. One or two year deviations can greatly influence how these models work and lack some robustness on that front. I think that ideally, we would want to be able to tap into some longer range time dependencies and wouldn’t want our model to be quickly swayed after a year or two off trend. This then raises the question of well, what order model is needed. What is that balance we wish to strike. How many years of a new trend is required in succession for it to no longer be considered an “off” couple of years. It is hard to give that answer up front and will depend on the dynamics of what is being modeled. So to suggest a specific model would be hard to do without a field expert consultation, but as a general direction, I think a higher order RW model would be more appropriate.