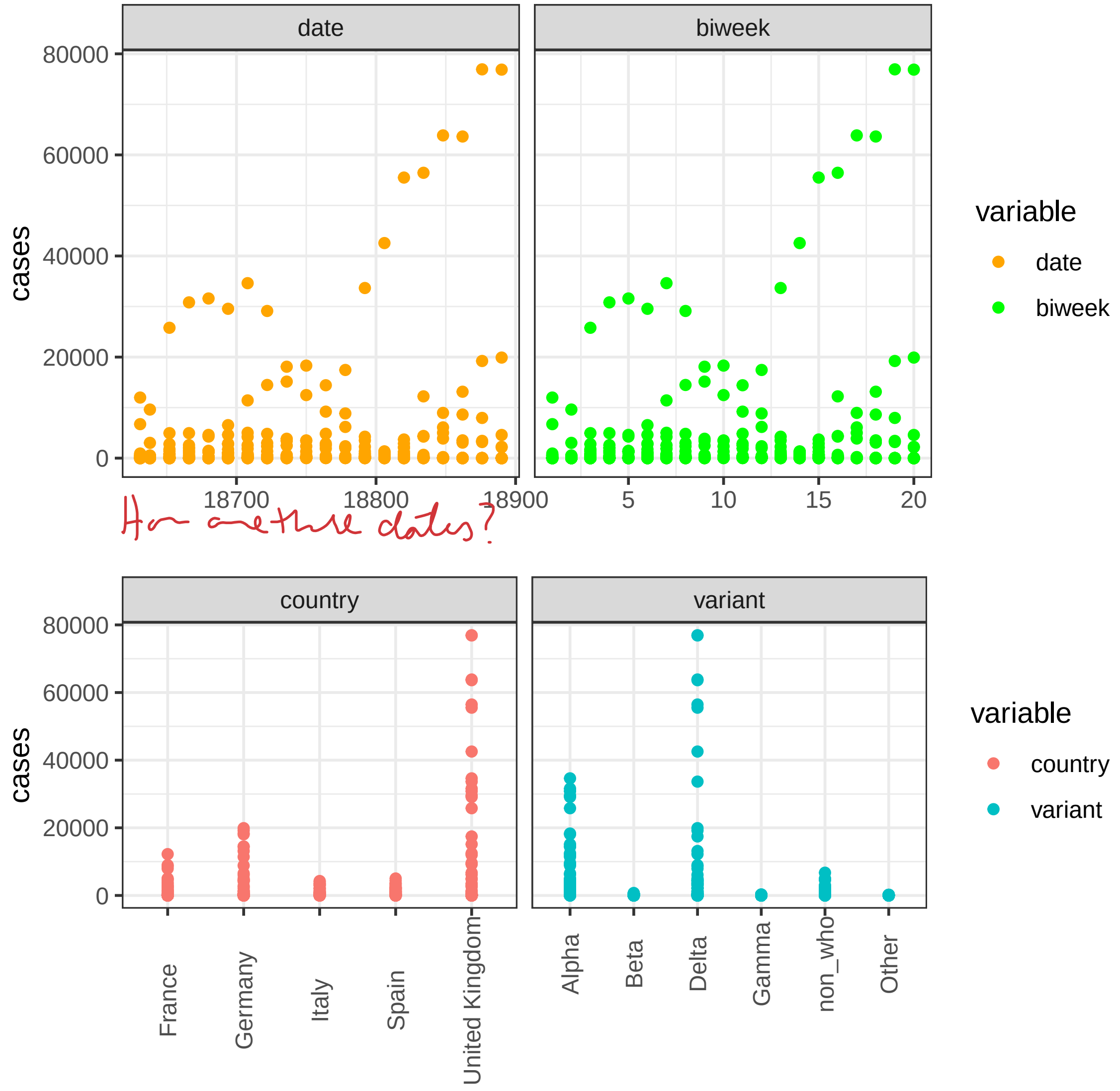
Bayesian Statistics, Philosophy and Practice: Coursework Assignment

Question 1

To begin the analysis a visualization of the data:



How are these dates?

Figure 1: First Data Visualization

The plot shows that there are considerable differences in case numbers between both variants and countries. Notably, the UK dominates case numbers with Germany following although some way behind as the maximum number of cases is 77000 in the UK compared to 20000 in Germany. Furthermore, Delta and Alpha variants see the highest case numbers with maximum values of 77000 and 35000 respectively. Other variants see very little cases in comparison. Date and biweek are the same variable as they both represent time. There are appears to be some relationship between biweek and case numbers. It seems plausible that a baseline model could be `cases ~ biweek + variant + country`. The cases are unbounded count data so a sensible family to use for the model would be the negatigve binomial. Therefore:

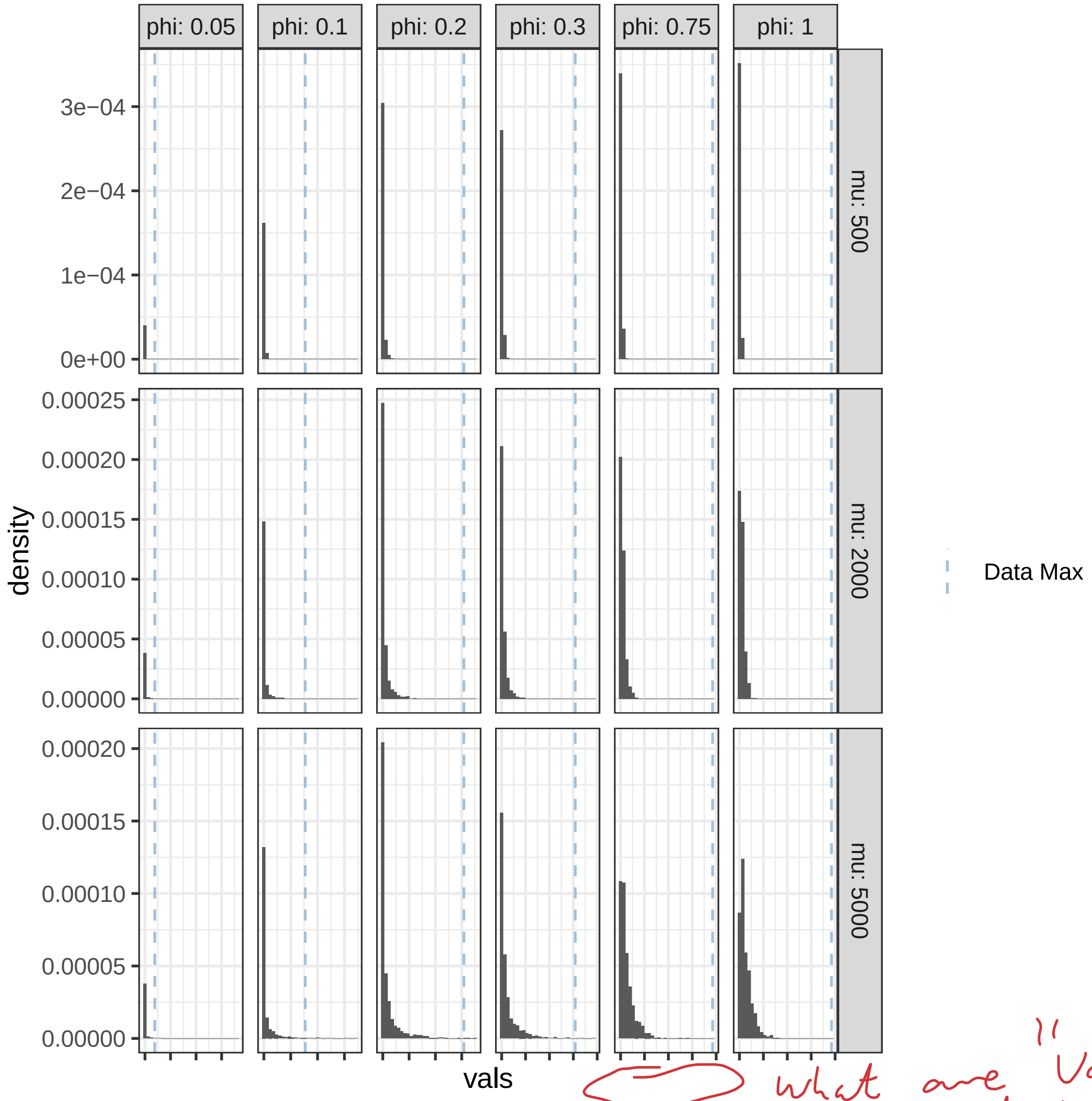
$$Y_i|\beta_0,\ldots,\beta_{10},\phi \sim \text{Negative Binomial}(\mu_i,\phi)$$

and

$$\log(\mu_i) = \beta_0 + \beta_1(\text{variant}_{\text{Beta}}) + \beta_2(\text{variant}_{\text{Delta}}) + \beta_3(\text{variant}_{\text{Gamma}}) + \beta_4(\text{variant}_{\text{non\_who}}) + \beta_5(\text{variant}_{\text{Other}})$$
$$+\beta_6(\text{country}_{\text{Germany}}) + \beta_7(\text{country}_{\text{Italy}}) + \beta_8(\text{country}_{\text{Spain}}) + \beta_9(\text{country}_{\text{United Kingdom}}) + \beta_{10}(\text{biweek})$$
$$\therefore$$
$$\mu_i = e^{(\beta_0+\beta_1(\text{variant}_{\text{Beta}})+\cdots+\beta_{10}(\text{biweek}))}$$
$$\mu_i = e^{\beta_0}\ldots e^{\beta_{10}(\text{biweek}))}$$

Not tidy mathematics but gist is understandable.

Formulating μ_i in this way helps with setting weakly informative priors. For β_0 the mean rate $\mu_i \in (0, 77000)$ so if all other coefficients were 0 then $\beta_0 \in (log(1), log(77000)) \implies \beta_0 \in (0, 11.25)$. Therefore set $\underline{\beta_0 \sim (7,2)}$. Tight! For $\beta_1 \ldots \beta_9$ the values of variant and country are either 0 or 1. Therefore calculating e^{β} needs to be a sensible amount to multiply by the mean value by: e^2=7.3890561, e^3=20.0855369 and e^4=54.59815. Hence, let $\beta_{1\ldots9} \sim N(0,2)$ as 54 times the mean value would be an upper bound. Following the same logic for β_{10}, biweek can take values between 1 and 20. Given $e^{0.2*20}$=54.59815 let $\beta_{10} \sim N(0, 0.1)$. ✓ The final parameter to set a prior for is the ϕ, or the dispersion parameter. The cases are very overdispersed as $Var[Y] > E[Y]$ which means ϕ can be expected to be a relatively low number. Plotting a sequence of sensible means from the data and some ϕ values between 0 and 1 gives the following:



what are "vals"

I think I understand this and it's clever (could be better explained)

Figure 2: Shape Parameter Estimation plot

From the plot that it certainly would be sensible for ϕ to be between 0 and 1, furthermore the values closer to zero seem to give a distribution closer to that of the data hence however do have a larger range. The values of phi closer to one seem to cut the distrribution a little short. Therefore, I choose to set the mean of the prior to 0.2. The shape parameter is always positive so the exponential distribution is a good choice furthermore it is right skewed so the values closer to 1 will be less likely. The mean of the the $Exp(\lambda)$ distribution is $\frac{1}{\lambda}$. Therefore for a mean of 0.2, I set $\lambda = 5$. I am unsure if this prior is somewhat informative so will checkc it after the model fit. Therefore the priors for the model are:

- $\beta_0 \sim (7, 2)$
- $\beta_{1...9} \sim N(0, 2)$
- $\beta_{10} \sim N(0, 0.1)$
- $\phi \sim Exp(5)$

Before fitting the model testing and training data sets are created using the following code, which are equally weighted on variant and country:

```
testing <- variants%>%group_by(country,variant)%>%sample_frac(0.2)
training <- setdiff(variants,testing)
```

Using 20% for the testing set as this isn't a particularly large data set gives 480 data points on the training set and 120 on the testing set. Now fitting the specified model:

Going to make life hard but kudos!

```
#Baseline model
intercept_prior <- set_prior("normal(7,2)",class = "Intercept")
b_prior <- set_prior("normal(0,2)",class = "b")
b_prior_biweek <- set_prior("normal(0,0.1)", class = "b", coef = "biweek")
shape_prior <- set_prior("exponential(5)",class = "shape")
baseline <- brm(cases  ~ biweek + variant + country , family=negbinomial(),
                data=training, prior = c(intercept_prior,shape_prior,b_prior,b_prior_biweek))
```

I'm assuming 'baseline' = Not your actual model."

A summary shows no issues with Rhats as they are all equal to 1 indicating convergence and more than adequate effective sample sizes for both the bulk and tail.

```
##  Family: negbinomial
##   Links: mu = log; shape = identity
## Formula: cases ~ biweek + variant + country
##    Data: training (Number of observations: 480)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept               8.17      0.36     7.47     8.94 1.00     2665
## biweek                 -0.02      0.03    -0.07     0.03 1.00     2602
## variantBeta            -3.24      0.31    -3.84    -2.61 1.00     2487
## variantDelta            0.53      0.36    -0.18     1.23 1.00     2403
## variantGamma           -3.53      0.30    -4.11    -2.93 1.00     2810
## variantnon_who         -1.53      0.30    -2.13    -0.95 1.00     2955
## variantOther           -4.06      0.29    -4.64    -3.47 1.00     2762
## countryGermany          0.04      0.27    -0.47     0.56 1.00     2634
## countryItaly           -0.72      0.28    -1.28    -0.18 1.00     2823
## countrySpain           -0.38      0.27    -0.89     0.16 1.00     2951
## countryUnitedKingdom    0.62      0.29     0.06     1.20 1.00     3046
##                     Tail_ESS
## Intercept               2657
## biweek                  2664
## variantBeta             2691
## variantDelta            3034
## variantGamma            3228
## variantnon_who          3042
## variantOther            2853
## countryGermany          3006
## countryItaly            3318
## countrySpain            3152
## countryUnitedKingdom    2783
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     0.30      0.02     0.27     0.33 1.00     3774     2776
```

```
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Further checking of the trace plots show the chains are well mixed with no obvious issues.

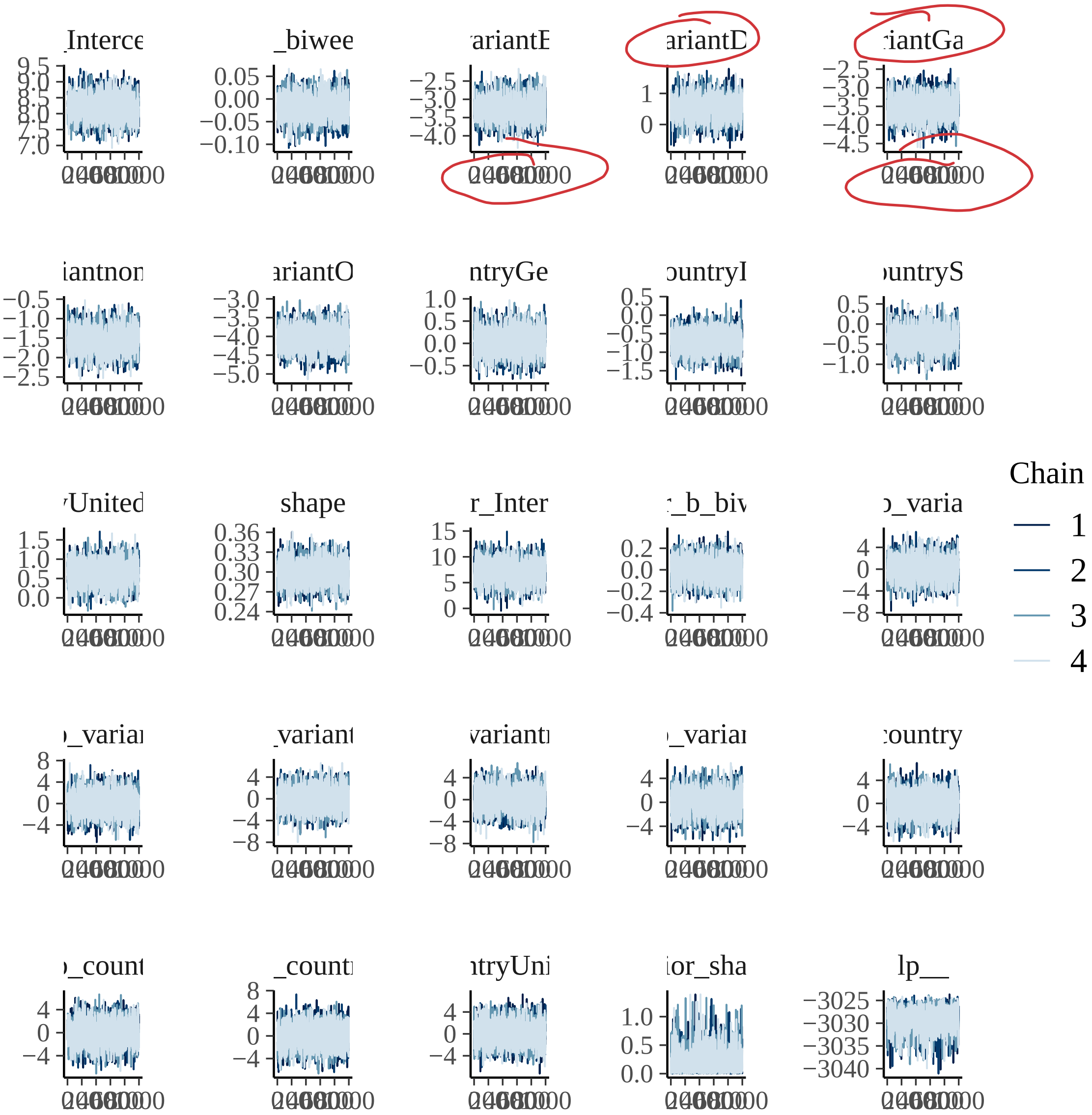


Figure 3: Baseline Model Trace Plots

At least can see traces clearly.

Plotting the prior and posterior of the shape parameter shows that the prior is somewhat informative therefore I will no change it further.

Very Nice.

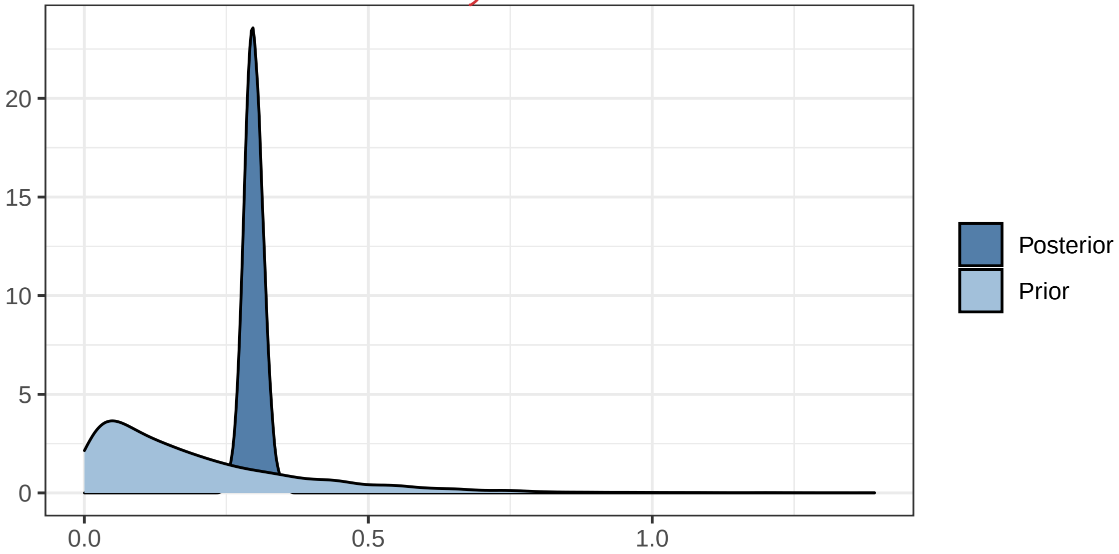


Figure 4: Prior and Posterior densities for Shape

From the summary there are potentially two intercepts that may warrant further investigation: Germany and biweek.

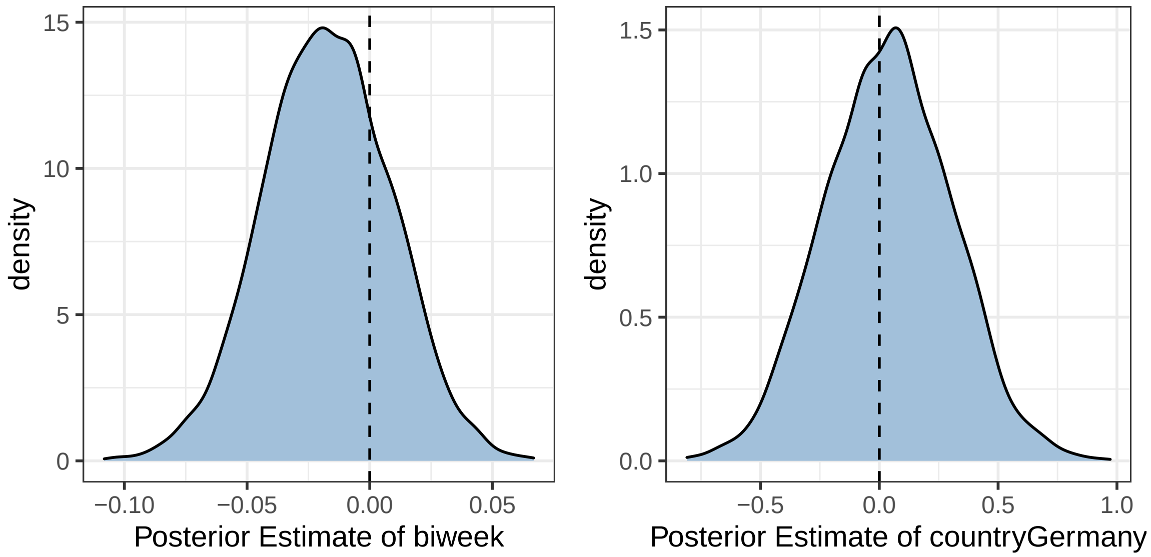


Figure 5: Baseline Model posterior checking

It was to be expected that Germany may be zero given how similar it's case numbers are to France in the original plot. We can conduct a hypothesis test in `brms` to confirm that Germany is no different from the mean level (France).

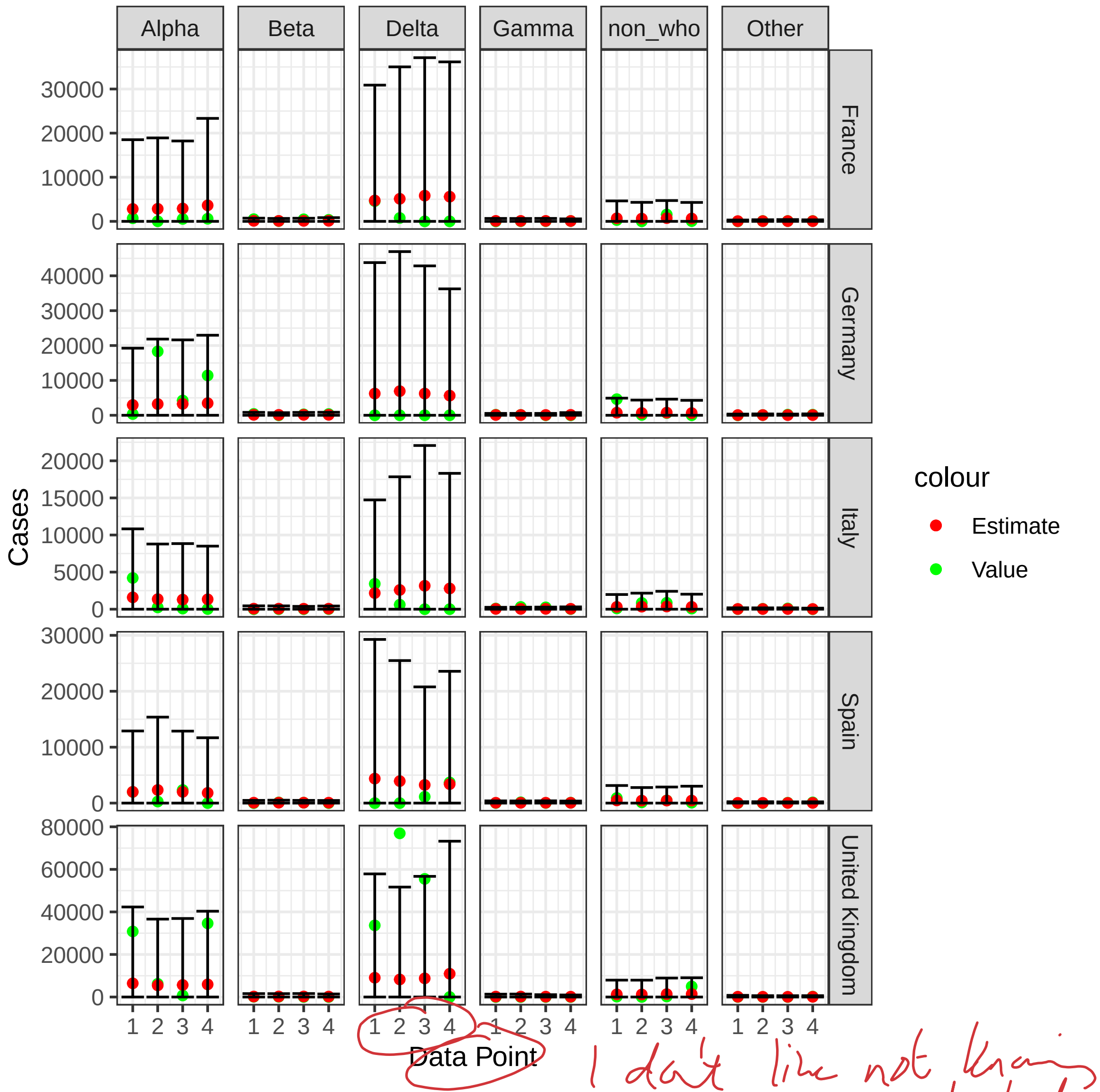
```
hypothesis(baseline, 'countryGermany < 0')
```

```
## Hypothesis Tests for class b:
##              Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio
## 1 (countryGermany) < 0     0.04      0.27     -0.4     0.47       0.79
##   Post.Prob Star
## 1      0.44
## ---
## 'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
## '*': For one-sided hypotheses, the posterior probability exceeds 95%;
## for two-sided hypotheses, the value tested against lies outside the 95%-CI.
## Posterior probabilities of point hypotheses assume equal prior probabilities.
```

This gives the posterior probability that the coefficient for Germany is less than 0 to be 44%, suggesting that it probably would be worth removing from the model. However, it seems pointless to group Germany and France together given all the other countries give evidence of effect. I am not going to group France and Germany together. Applying the hypothesis test to biweek:

```
## Hypothesis Tests for class b:
##      Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob Star
## 1 (biweek) < 0    -0.02      0.03    -0.06     0.02       2.97      0.75
## ---
## 'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
## '*': For one-sided hypotheses, the posterior probability exceeds 95%;
## for two-sided hypotheses, the value tested against lies outside the 95%-CI.
## Posterior probabilities of point hypotheses assume equal prior probabilities.
```

This shows there is less evidence to remove biweek from the model as the posterior probability that the coefficient is less than zero is 75%. Although, does it make sense that the coefficient is estimated to be negative the original plot showed a weak positive correlation. Naturally, it would make sense to group biweek on variant, as variants are present for different time periods. Before looking at another model lets look at the predictions:



I don't like not knowing the biweek, but otherwise the plot is cool.

Figure 6: Baseline Predictions

The estimates for low number of cases appear very accurate but not so with the higher number of cases. However, the credible intervals do capture the countries and variants where there are a larger number of cases.

Grouping biweek over variant yields the following model:

$$
\begin{aligned}
&Y_i|\alpha,\beta,\phi \sim \text{NegativeBinomial}(\mu_i,\phi), \text{ where } \beta \text{ represents the regression cofficients,}\\
&\quad \log(\mu_i) = \alpha_{j[i]} + \beta_{1j[i]}(\text{biweek}) + \beta_2(\text{country}_{\text{Germany}}) + \beta_3(\text{country}_{\text{Italy}}) + \beta_4(\text{country}_{\text{Spain}}) + \beta_5(\text{country}_{\text{United Kingdom}})\\
&\qquad\qquad + \beta_6(\text{variant}_{\text{Beta}}) + \beta_7(\text{variant}_{\text{Delta}}) + \beta_8(\text{variant}_{\text{Gamma}}) + \beta_9(\text{variant}_{\text{non\_who}}) + \beta_{10}(\text{variant}_{\text{Other}})\\
&\quad \begin{pmatrix}\alpha_j\\ \beta_{1j}\end{pmatrix} \sim N\left(\begin{pmatrix}\epsilon_{\alpha_j}\\ \epsilon_{\beta_{1j}}\end{pmatrix}, \begin{pmatrix}\sigma^2_{\alpha_j} & \rho_{\alpha_j\beta_{1j}}\\ \rho_{\beta_{1j}\alpha_j} & \sigma^2_{\beta_{1j}}\end{pmatrix}\right), \text{ for variant j} = 1,\ldots,6
\end{aligned}
$$

j's not in distribution think $\epsilon_j \sim N(0, \sigma^2)$

ρ no j.

This means the priors relating to the σ's in the covariance matrix have to be specified. To begin with I set the SD of the intercept to be Normal(0,2) as a best guess of the variation between variants and set the SD of the regression coefficients to be Normal(0,0.1) to match that of the prior on the regression coefficients. I believe these will be too large. Fitting the model:

So why fit this? What is your prior coming?

```
intercept_prior <- set_prior("normal(7,2)",class = "Intercept")
b_prior <- set_prior("normal(0,2)",class = "b")
b_prior_biweek <- set_prior("normal(0,0.1)", class = "b", coef = "biweek")
shape_prior <- set_prior("normal(30000,10000)",class = "shape")
sd_prior <- set_prior("normal(0,0.1)",class="sd")
sd_prior_intercept <- set_prior("normal(0,2)",class = "sd", coef="Intercept",group = "variant")
model_1 <- brm(cases ~ biweek + variant + country + (biweek|variant), family=negbinomial(), data=trainin
```

The models trace plots are well mixed, Rhats are all equal to one and effective sample sizes are all large. (Omitted as not necassary) However, a posterior predictive check(below) shows the model is predicting some extremely large values over 1.5 million cases. This is simply not feasible based on the data and is indicative of improper prior choice.

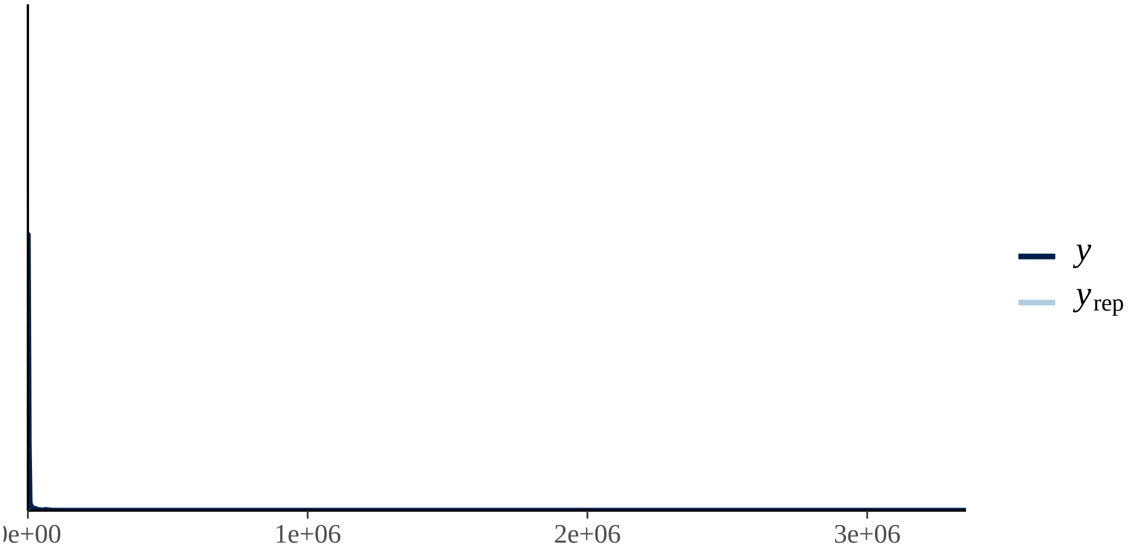


Figure 7: PPC of model 1

Adjusting the standard deviation priors to Normal(0,0.5) for the intercept and Normal(0,0.006) for the coefficients in an attempt to reduce the range of the posterior yields the following results:

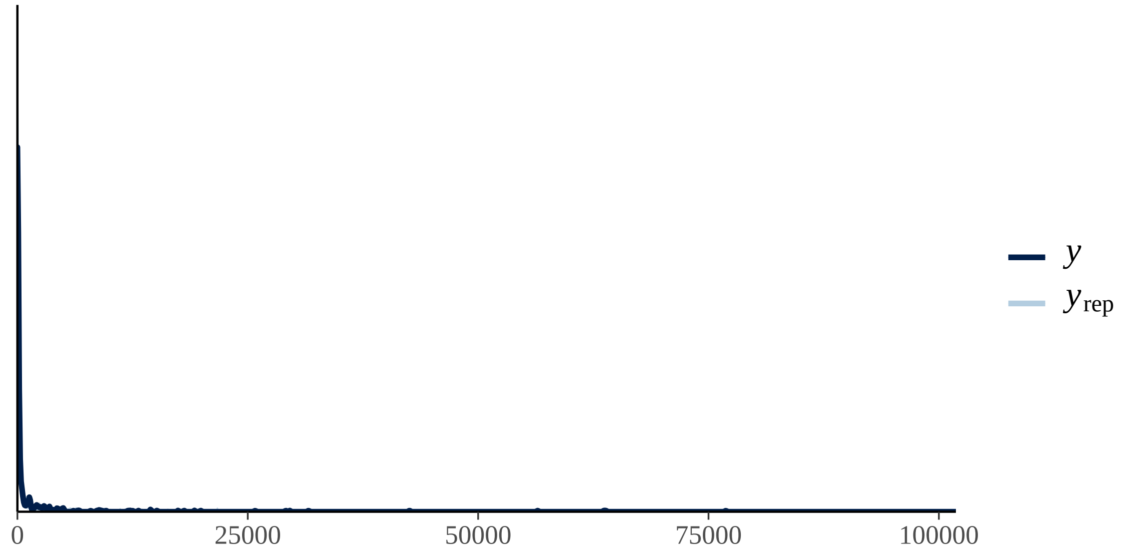


Figure 8: PPC of model 2

This is significantly better than before, with posterior mimiming the data very well.

```
##  Family: negbinomial
##   Links: mu = log; shape = identity
## Formula: cases ~ biweek + variant + country + (biweek | variant)
##    Data: training (Number of observations: 480)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~variant (Number of levels: 6)
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)             0.87      0.29     0.29     1.47 1.00     1125
## sd(biweek)                0.05      0.01     0.04     0.06 1.00     2286
## cor(Intercept,biweek)    -0.57      0.26    -0.97    -0.02 1.00      753
##                       Tail_ESS
## sd(Intercept)              937
## sd(biweek)                1550
## cor(Intercept,biweek)     1317
##
## Population-Level Effects:
##                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                6.90      0.85     5.20     8.45 1.00     1038
## biweek                   0.01      0.03    -0.04     0.07 1.00     2457
## variantBeta             -2.01      1.09    -3.87     0.33 1.00     1489
## variantDelta            -0.37      1.80    -3.81     3.07 1.00     1446
## variantGamma            -2.27      1.06    -4.01    -0.03 1.00     1219
## variantnon_who          -1.12      1.15    -3.11     1.35 1.00     1192
## variantOther            -2.72      1.09    -4.57    -0.39 1.00     1183
## countryGermany          -0.03      0.24    -0.49     0.46 1.00     2548
## countryItaly            -0.65      0.26    -1.15    -0.15 1.00     2471
## countrySpain            -0.42      0.25    -0.90     0.06 1.00     2564
## countryUnitedKingdom     0.58      0.27     0.06     1.11 1.00     2670
##                      Tail_ESS
## Intercept                1670
## biweek                   2628
## variantBeta              2171
## variantDelta             1280
## variantGamma             1981
## variantnon_who           2318
## variantOther             1898
## countryGermany           2319
## countryItaly             2311
## countrySpain             2861
## countryUnitedKingdom     2695
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     0.38      0.03     0.33     0.44 1.00     2903     2127
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The summary shows Rhats are all equal to one indicating convergence however, the bulk ess is a little low for `cor(Intercept,biweek)` but the tail ess is adequate.

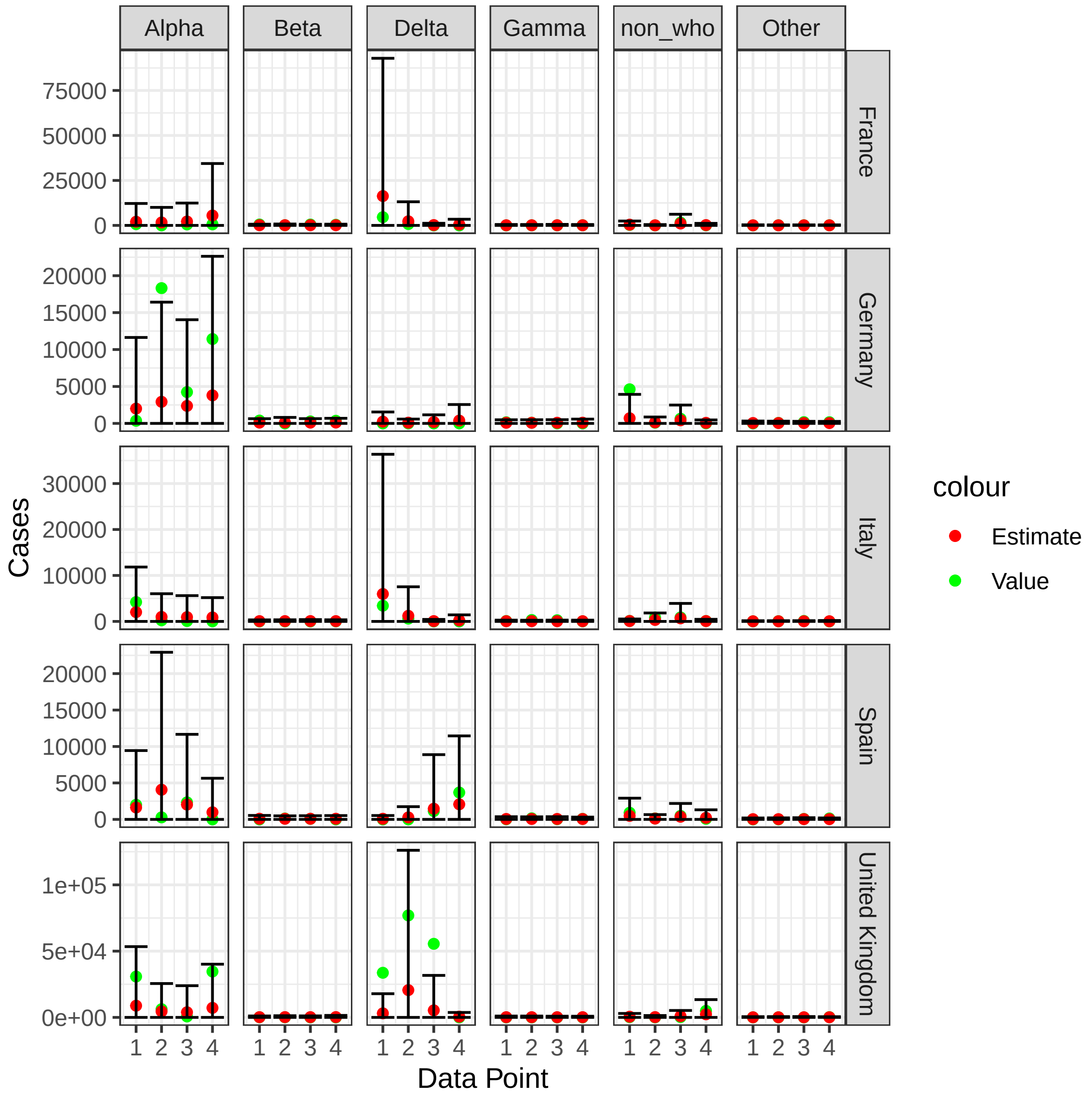


Figure 9: Model 2 Predictions

The predictions look marginally better for the larger values as thet seem to follow trend a lot better, this is visible in the UK predictions for Alpha and Delta. It would appear that the predictions are just a little low. The issue may be that the model is not considering country beyond the mean for each country. Perhaps adding an additional intercept grouped over country and variant - `cases ~ biweek + variant + country + (biweek|variant)+(1|country:variant)`. Personally, I don't think this would be beneficial as it is just adding more complexity to the model and more priors to specify. Furthermore, biweek could be grouped over country and variant - `cases ~ biweek + variant + country + (biweek|variant:country)`. This model may solve the problem but the variant and country terms become pointless as they are just means and they are both now considered in the grouping. Hence, I propose the simpler model - `cases ~ biweek +(biweek|country:variant)`. Writing this mathematically:

This seems sensible!

$$
\begin{aligned}
&Y_i|\alpha,\beta,\phi \sim \text{NegativeBinomial}(\mu_i,\phi), \text{ where } \beta \text{ represents the regression cofficients,}\\
&\quad \log(\mu_i) = \alpha_{j[i]} + \beta_{1j[i]}(\text{biweek})\\
&\quad \begin{pmatrix}\alpha_j\\ \beta_{1j}\end{pmatrix} \sim N\left(\begin{pmatrix}\epsilon_{\alpha_j}\\ \epsilon_{\beta_{1j}}\end{pmatrix}, \begin{pmatrix}\sigma^2_{\alpha_{\overset{0}{j}}} & \rho_{\alpha_{\overset{0}{j}}\beta_{1\overset{0}{j}}}\\ \rho_{\beta_{1\overset{0}{j}}\alpha_{\overset{0}{j}}} & \sigma^2_{\beta_{1\overset{0}{j}}}\end{pmatrix}\right), \text{ for variant:country j} = 1,\ldots,30
\end{aligned}
$$

Once again I set the prior on the standard deviation of the intercept as Normal(0,2) but I choose the standard deviation of the coefficients to be Normal(0,0.01), this is because on the previous model Normal(0,0.1) and produced a posterior with a huge range. Fitting the model:

```
intercept_prior <- set_prior("normal(7,2)",class = "Intercept")
b_prior <- set_prior("normal(0,2)",class = "b")
b_prior_biweek <- set_prior("normal(0,0.1)", class = "b", coef = "biweek")
shape_prior <- set_prior("exponential(5)",class = "shape")
sd_prior_intercept <- set_prior("normal(0,2)",class = "sd", coef="Intercept",group = "variant:country")
sd_prior_biweek <- set_prior("normal(0,0.01)",class="sd",coef="biweek",group="variant:country")
model_3 <- brm(cases ~ biweek  + (biweek|variant:country), family=negbinomial(), data=training, prior =
```

Checking the summary shows no issues with Rhats all at one and ESS are all adequate.

```
##  Family: negbinomial
##   Links: mu = log; shape = identity
## Formula: cases ~ biweek + (biweek | variant:country)
##    Data: training (Number of observations: 480)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~variant:country (Number of levels: 30)
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)             2.02      0.23     1.63     2.53 1.00     1461
## sd(biweek)                0.10      0.01     0.08     0.11 1.00     1547
## cor(Intercept,biweek)    -0.49      0.10    -0.66    -0.28 1.00     1936
##                       Tail_ESS
## sd(Intercept)             2360
## sd(biweek)                1843
## cor(Intercept,biweek)     2396
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     4.97      0.43     4.12     5.82 1.00     1304     2229
## biweek        0.02      0.03    -0.03     0.07 1.00     2500     3416
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     0.49      0.04     0.41     0.57 1.00     1628     2204
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The trace plots are well mixed indicating convergence. ✓

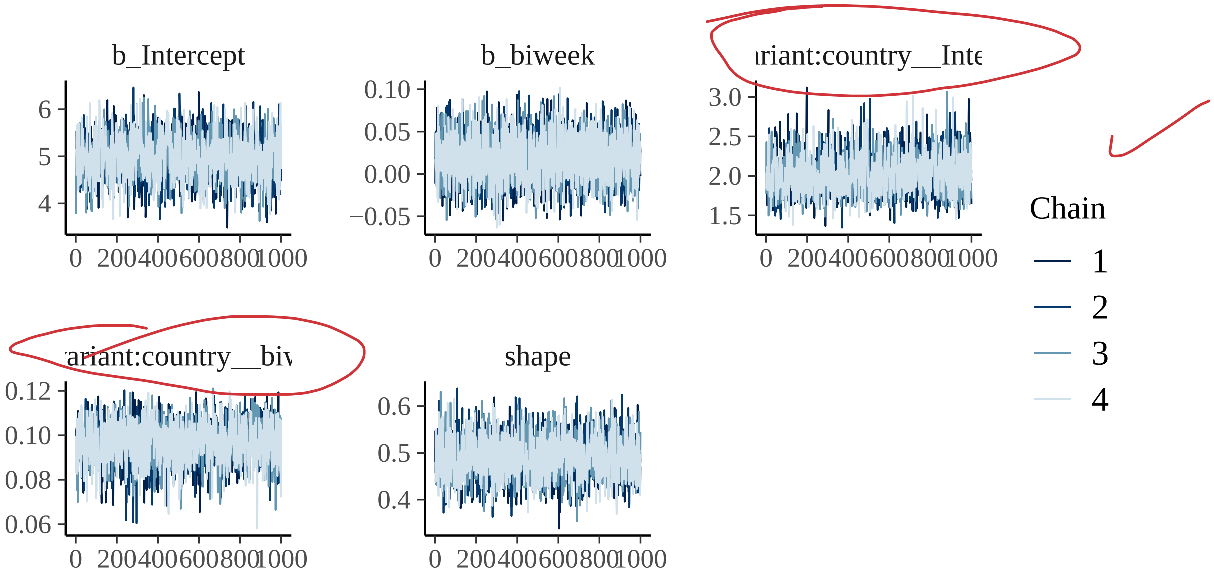


Figure 10: Model 3 Trace Plots

The predictions look better, there are no points outside of the credible interval and point estimates appear to be more accurate. The model seems to still be struggling with variation of the Alpha variant and the credible intervals of some of the delta predictions are larger than hoped for which could become a problem when predicting on new data. I am happy enough to use this model for inference.

Modelling: Overall very good. The ton of models you didn't believe was a little confusing. I wanted to know what you thought. Model made sense. Slight mathematical error with indices 13/14

Priors: This was really exceptional. The analysis leading to the exponential prior was better than anything I've seen on this course. I have never done this before but since I have given many deserved (14/14)'s I am going to overmark. 17/14 (not a typo)

Convergence: 14/14.

Validation: Mostly excellent. Everything done was clear and look at the high values of the predictions interesting. Out of sample validation tough. Would have liked to see in sample if only to establish temporal structure of fits vs data 12/14

Critique: Fine. Could have said more about what was going on, but the prior/posterior for ϕ was worth the entry fee 6/7.

Presentation: 1 mark docked for untidy typesetting in linear predictor equations and for title/axis labels occasionally missing/unreadable 5/7.

Q1: 67/70.

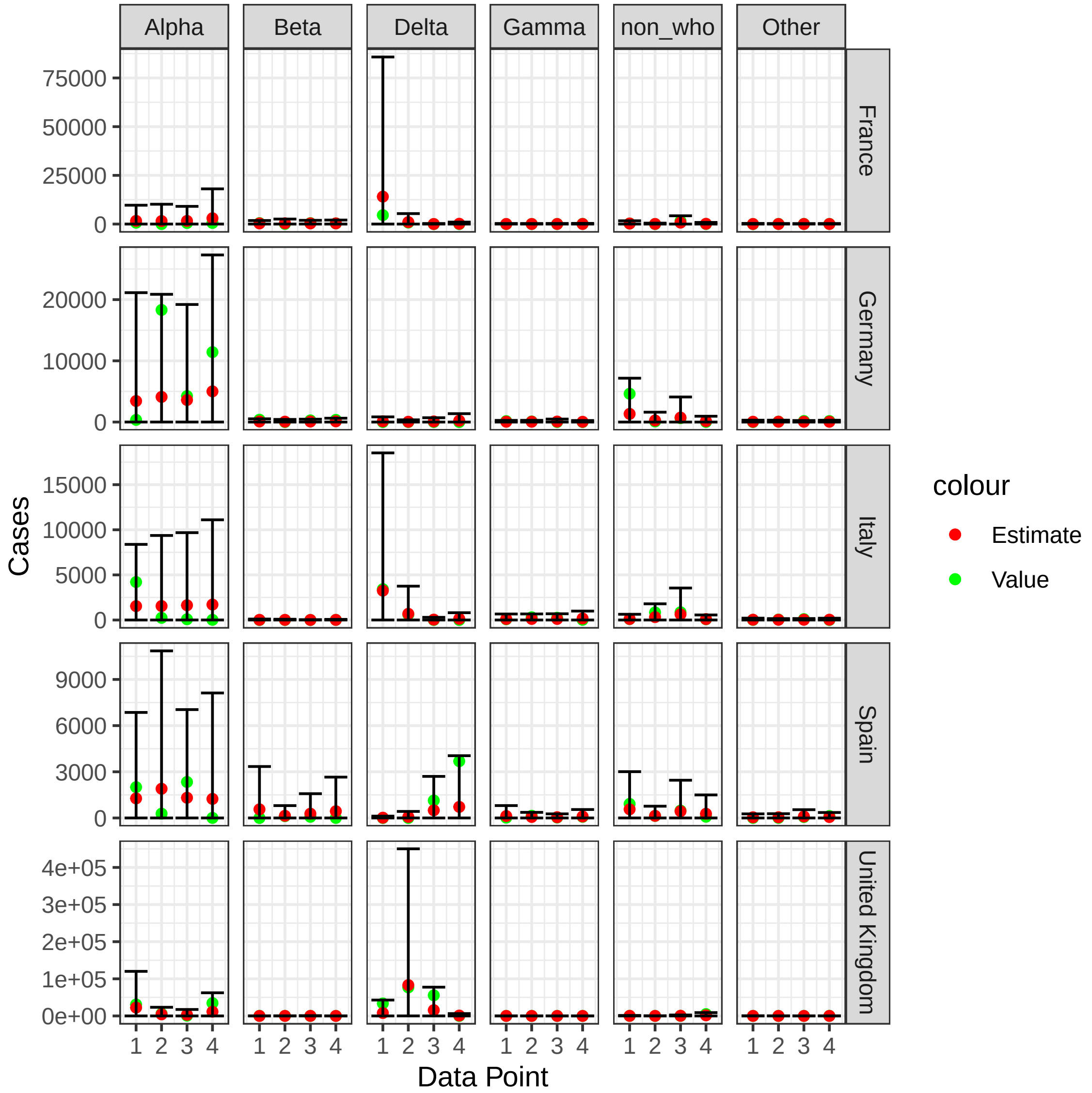


Figure 11: Model 3 Predictions

Question 2

Firstly, the last recorded date in the is 2021-09-20, which means there are 102 days between then and the 2021-31-12 therefore there are 7 fortnights between left in 2021. Creating a new data frame with the following code:

```
uk_delta_pred_data <- data.frame(date = seq(as_date("2021-10-04"),by = "2 week",length.out = 7),
                                 biweek=seq(21,27,1),country=rep("United Kingdom",7),
                                 variant=rep("Delta",7))
```

Now predicting with this data set with the probs set to 5% and 95% to represent a 90% confidence interval. Also adding in the dates:

```
uk_delta_predictions <- data.frame(predict(model_3,uk_delta_pred_data, probs = c(0.05,0.95),
summary = TRUE))
%>%mutate(biweek=seq(21,27,1),date = seq(as_date("2021-09-27"),by = "2 week",length.out = 7))
```

What about weeks 1:20.

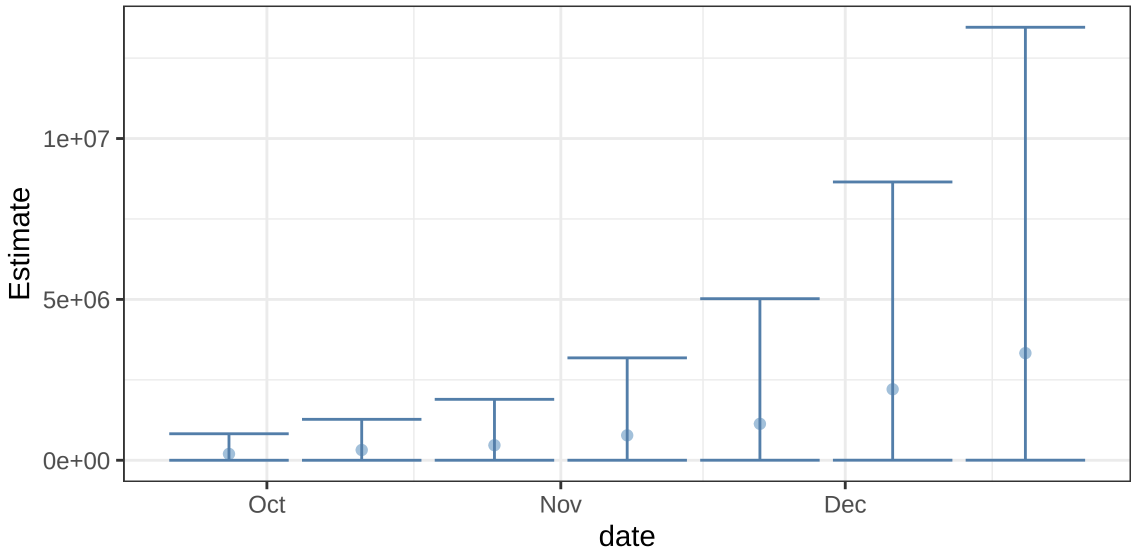
This gives the following results:

Table 1: UK Delta Variant Predictions

Estimate	Est.Error	Q5	Q95	biweek	date
198176.0	435162.8	352.65	823949.6	21	2021-09-27
320252.4	733547.7	627.95	1272300.6	22	2021-10-11
467990.4	1036915.4	874.40	1894438.6	23	2021-10-25
774977.3	1930369.9	1257.75	3183217.9	24	2021-11-08
1133039.2	2918225.0	1670.90	5021988.8	25	2021-11-22
2207423.1	7741979.2	2693.95	8649648.4	26	2021-12-06
3330295.8	13848763.5	3274.75	13458067.8	27	2021-12-20

Could peak delta have already happened!

5



✓5

Figure 12: UK Delta Variant Predcitions Plot

✓3

This gives an estimate as 3425932 cases occurring on `2021-12-20` with a 90% credible interval of (3760,13359540). The efficacy of the model is very poor as it is just predicting exponential growth of case numbers. We would expect the case numbers to probably go up more and then begin to flatline and decrease (parabola) as this is the trend of all the previous variants. The model doesn't predict this because the relationship between time and cases is non-linear so a linear regression model is simply not going to work in predicting this.

Not true, but only having linear "in time" is an issue

13/15

Question 3

For the calculations of the monte-carlo errors I will use the minimum bulk effective sample size across the all the parameters. This was 1300 (see previous summary). After completing a few calculations the monte-carlo error was not below 0.01, so I re-ran the model with double the number of iterations.

```
##  Family: negbinomial
##   Links: mu = log; shape = identity
## Formula: cases ~ biweek + (biweek | variant:country)
##    Data: training (Number of observations: 480)
##   Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
##          total post-warmup draws = 8000
##
## Group-Level Effects:
## ~variant:country (Number of levels: 30)
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)             2.03      0.23     1.62     2.55 1.00     2620
## sd(biweek)                0.10      0.01     0.08     0.11 1.00     3194
## cor(Intercept,biweek)    -0.48      0.10    -0.66    -0.26 1.00     3383
##                       Tail_ESS
## sd(Intercept)             3885
## sd(biweek)                3017
## cor(Intercept,biweek)     4544
##
## Population-Level Effects:
```

```
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     4.99      0.44     4.14     5.85 1.00     2636     4121
## biweek        0.02      0.03    -0.03     0.07 1.00     4855     6253
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     0.49      0.04     0.41     0.57 1.00     3733     4219
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The effective sample size is now 2600. Below is the code for the calculations for part a:

```
#a
#Getting vector of dates
dates <- unique(variants$date)
#Selecting Case 1
d1 <- variants%>%filter(variant=="Beta",country=="France")%>%select(country,variant,biweek)
#Predicting Case 1
p1 <- predict(model_4,d1, summary = FALSE)
#Selecting Case 2
d2 <- variants%>%filter(variant=="non_who",country=="France")%>%select(country,variant,biweek)
#Predicting Case 2
p2 <- predict(model_4,d2, summary = FALSE)
#Creating estimate stores
estimate <- numeric(20)
mc_error <- numeric(20)
dominance <- numeric(20)
#Looping over each biweek
for (i in 1:length(estimate)) {
estimate[i] <-  mean(p1[,i]>p2[,i]) #monte carlo of dominance
mc_error[i] <- sd(p1[,i]>p2[,i])/sqrt(2600) #monte carlo error
dominance[i] <- ifelse({0.5<estimate[i]},{1},{0}) #dominance true or false
}
tibble(estimate,mc_error,dominance,dates)
```

If I was being picky $\sqrt{\hat{p}(1-\hat{p})/N_{eff}}$ is more accurate.

This produces the following table:

Table 2: Beta dominance over non_who in France

estimate	mc_error	dominance	dates
0.323875	0.0091779	0	2021-01-04
0.333750	0.0092485	0	2021-01-11
0.356375	0.0093931	0	2021-01-25
0.398125	0.0096007	0	2021-02-08
0.410750	0.0096489	0	2021-02-22
0.438375	0.0097317	0	2021-03-08
0.470125	0.0097889	0	2021-03-22
0.500125	0.0098064	1	2021-04-05
0.517375	0.0098005	1	2021-04-19
0.539250	0.0097762	1	2021-05-03
0.569125	0.0097123	1	2021-05-17
0.593125	0.0096348	1	2021-05-31
0.619750	0.0095210	1	2021-06-14
0.638875	0.0094206	1	2021-06-28
0.667250	0.0092415	1	2021-07-12
0.682875	0.0091270	1	2021-07-26
0.711750	0.0088836	1	2021-08-09
0.720000	0.0088061	1	2021-08-23
0.740000	0.0086029	1	2021-09-06
0.747000	0.0085263	1	2021-09-20

5

Therefore the first date that beta dominates non_who is `2021-04-19`. A similar code produces the dominance of Gamma and Delta in Italy:

Table 3: Gamma dominance over Delta in Italy

estimate	mc_error	dominance	dates
0.759750	0.0083793	1	2021-01-04
0.737125	0.0086335	1	2021-01-11
0.706750	0.0089288	1	2021-01-25
0.665125	0.0092562	1	2021-02-08
0.642625	0.0093990	1	2021-02-22
0.597250	0.0096191	1	2021-03-08
0.555750	0.0097453	1	2021-03-22
0.510750	0.0098042	1	2021-04-05
0.460000	0.0097750	0	2021-04-19
0.412125	0.0096538	0	2021-05-03
0.370000	0.0094692	0	2021-05-17
0.332375	0.0092389	0	2021-05-31
0.294000	0.0089355	0	2021-06-14
0.258500	0.0085867	0	2021-06-28
0.223375	0.0081689	0	2021-07-12
0.190125	0.0076961	0	2021-07-26
0.180125	0.0075370	0	2021-08-09
0.145375	0.0069131	0	2021-08-23
0.122875	0.0064388	0	2021-09-06
0.114625	0.0062480	0	2021-09-20

Therefore the last date that Gamma dominates Delta is `2021-04-05`. For the Delta variance over all over 5 strains, the following code loops the estimates over all countries and variants and produces the dates of delta dominance:

```
country <- c("France","Germany","Italy","Spain","United Kingdom")
biweek <- numeric(length(country))
c_variants <- c("Alpha","Beta","Gamma","non_who","Other")
for (k in 1:length(country)) {
country1 <- country[k]
for (j in 1:length(c_variants)) {
variant1 <- c_variants[j]
d1 <- variants%>%filter(variant=="Delta",country==country1)%>%select(country,variant,biweek)
p1 <- predict(model_4,d1, summary = FALSE)
d2 <- variants%>%filter(variant==variant1,country==country1)%>%select(country,variant,biweek)
p2 <- predict(model_4,d2, summary = FALSE)
estimate <- numeric(20)
dominance <- numeric(20)
for (i in 1:length(estimate)) {
estimate[i] <-  mean(p1[,i]>p2[,i])
dominance[i] <- ifelse({0.5<estimate[i]},{1},{0})
}
df <- tibble(dominance)%>%select_all(list(~paste0(.,sep="_",variant1)))
ifelse({j==1},{country_df <- df},{country_df <- cbind(country_df,df)})
}
biweek[k] <- min(which(rowSums(country_df)==5))
}

knitr::kable(data.frame(country,dates[biweek]),caption = "Delta dominance")
```

Table 4: Delta dominance

country	Date.of.Delta.Dominance
France	2021-07-12
Germany	2021-08-09
Italy	2021-08-09
Spain	2021-08-09
United Kingdom	2021-05-31

5

15/15

Tot 95/100

Course record!!

(well I'm half way through marking, so... record so far!)