COVID-19 Project

190030150

14/05/2020

Contents

1 Introduction

In this analysis, spread of Covid-19 will be analysed, we are particularly interested in the evolution of number of confirmed cases and fatalities by country. Two datasets will be analysed. Using the first dataset fatality rate will be analysed and factors affecting fatality rate will be identified. The factors used are number of confirmed cases, population density, median age, urban population, number of hospital beds, health expenditure, GDP and death rate contributed to lung diseases. Using the second dataset we analyse number of confirmed cases per country.

2 Part 1

The first model that was fitted is a generalised linear model with a log link function. It assumes the response has a quasipoisson distribution which is often used for count data. For this model we assume independence since fatality rate of one country should not have a big effect on the fatality rate of other countries. The independence assumption will later be checked. Since we are modelling fatality rate we include confirmed cases as an offset. Quasipoisson is a right fit since the dispersion parameter is 123 which is significantly different from zero. Quasipoisson models have a different mean-variance relationship than poisson models. For quasipoisson models mean equals the dispersion parameter multiplied by the variance.

Table 1: Results from quasipoisson model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.5132098	0.7479525	-8.7080523	0.0000000
PopDensity	-0.0000432	0.0001134	-0.3806902	0.7040451
MedianAge	0.0847292	0.0106089	7.9866395	0.0000000
UrbanPop	0.0175560	0.0058240	3.0144008	0.0030884
Bed	-0.1404317	0.0224886	-6.2445641	0.0000000
Lung	-0.0059345	0.0044648	-1.3291827	0.1860805
HealthExp	-0.0000684	0.0000432	-1.5844039	0.1154947
GDP	-0.0000035	0.0000063	-0.5599929	0.5764334

[a] Null Deviance = 36778,139 DF
[b] Residual Deviance= 17033, 131 DF
[c] Dispersion parameter= 123

From table 1 it can be seen that the significant predictors are median age, urban population and number of hospital beds. For each one unit increase in percentage of urban population case fatality rate increases by a multiplicative effect of e^0.018 or 1.02. For each one unit increase in median age case fatality rate increases by a multiplicative effect of 1.09. For each one unit increase in number of beds per 1000 people the fatality rate increases by a factor of 0.87 which is a decrease by 13 percent. The deviance explained which measures how closely our models prediction are to the observed outcome is 46.2 % of the saturated model. That indicates our model fits the data fairly well. The overdispersion parameter is 123. It defines our standard errors and mean-variance relationship.

Next up a model is fitted only with countries with 10 or more deaths. These countries have gotten further in the virus process than other countries which makes the data more reliable. The fatality rate can fluctuate a lot in the first days of the virus in each country. The more data we have the more accurate the fatality rate gets. Another thing is that it takes time for people to die from the disease. First the cases appears and then some time later people can die from the disease. We might include countries with a lot of cases but no deaths, but the deaths will happen eventually. By including countries with more than 10 deaths we know that the virus has been affecting the country for long enough for people to die.

Table 2 shows the results after using a subset of the dataset. The significant parameters at the 95% level are number of hospital beds, median age and population density. The effect of bed increases to a multiplicative effect of 0.84. Population density has a multiplicative effect of 1.002 and median age 1.007. Dispersion parameter increases to 198 and deviance explained reduces to 38.2 %.

Table 2: Results from quasipoisson model on subsetted data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.7971404	0.9428704	-6.1483957	0.0000000
PopDensity	0.0017673	0.0005409	3.2670998	0.0017153
MedianAge	0.0730119	0.0136067	5.3658984	0.0000011
UrbanPop	0.0143504	0.0072498	1.9794233	0.0518793
Bed	-0.1734226	0.0300612	-5.7689843	0.0000002
Lung	-0.0120174	0.0061185	-1.9640991	0.0536692
HealthExp	0.0000035	0.0000663	0.0535121	0.9574831
GDP	-0.0000111	0.0000094	-1.1872024	0.2393400

[a] Disperion Parameter = 191.50
[b] Null Deviance = 35272,74 DF
[c] Residual Deviance = 13463, 66 DF

Now the assumptions of the model are checked. For a reliable model the model has to meet its assumptions.

The assumptions to check are collinearity, linearity of covariates, independence of response and the mean-variance relationship. First we check if the variables are too similar to each other by assessing collinearity. Looking at table 3 we see that two variables are highly collinear with higher VIF than 10. They are health expenditure and GDP. Including those results in increased variance because the parameters defining the plane are uncertain. This can be dealt with removing these variables, combining them in someway or fitting penalized regression models.

Table 3: Variance inflation factor results

	Variance inflation factor
PopDensity	1.765452
MedianAge	1.963916
UrbanPop	1.647581
Bed	1.349230
Lung	1.747160
HealthExp	13.647156
GDP	12.870191

Next up we check the linearity assumption by fitting pearson residual plots to each variable. Looking at the plots in figure 1 there seems to be some nonlinearity in all the plots however it can be hard to see. By using Tukey's test for nonaddivity seen in table 4 has the null hypothesis that the coefficient of the quadratic term is zero. The quadratic term is the blue line in the plot. All covariates have very low p values indicating a failure to reject the null hypothesis. This means that the quadratic term could be something other than zero indicating a nonlinear fit for all covariates. The assumption of linearity can not be fulfilled.

Table 4: Tukey's test results

	Test stat	Pr(>|Test stat|)
PopDensity	770.33566	0
MedianAge	704.28674	0
UrbanPop	68.61282	0
Bed	2094.73480	0
Lung	710.39597	0
HealthExp	1434.29351	0
GDP	650.06516	0

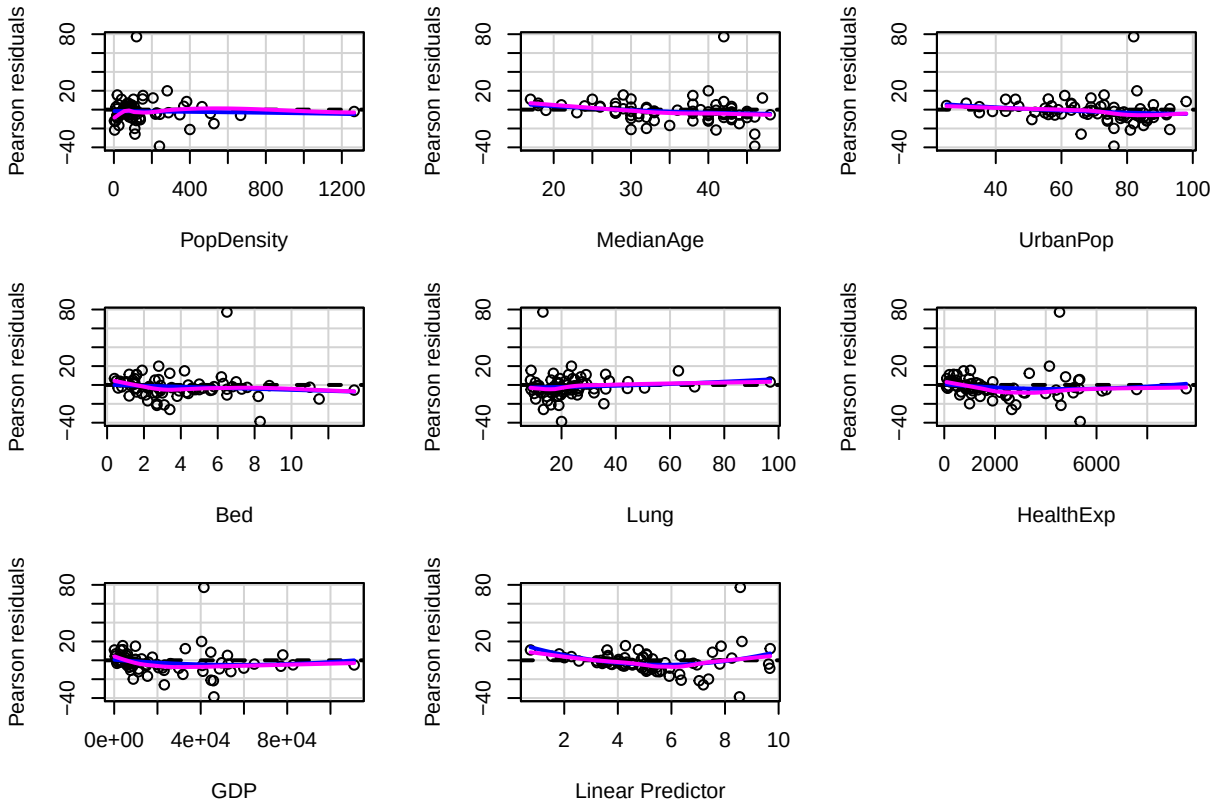


Figure 1: Residual plots for glm model

Next assumption is the assumption of independence. To check that we do a runs test and plot the autocorrelation function. The autocorrelation plot seen in figure 2 shows dont show any highly correlated residuals. The runs test compares the number of observed runs with what is expected under independence. The resulting test statistic is a standard normal distribution so values more extreme than the absolute value of 2 indicate autocorrelation. In our case the test statistic is -1.01 as seen in table 5 so there does not seem to be correlated residuals. The assumption of independence is fulfilled.

Table 5: Runs Test results

	Standardized Runs Statistic
Standardized Runs Statistic	-1.0449625
p-value	0.2960403

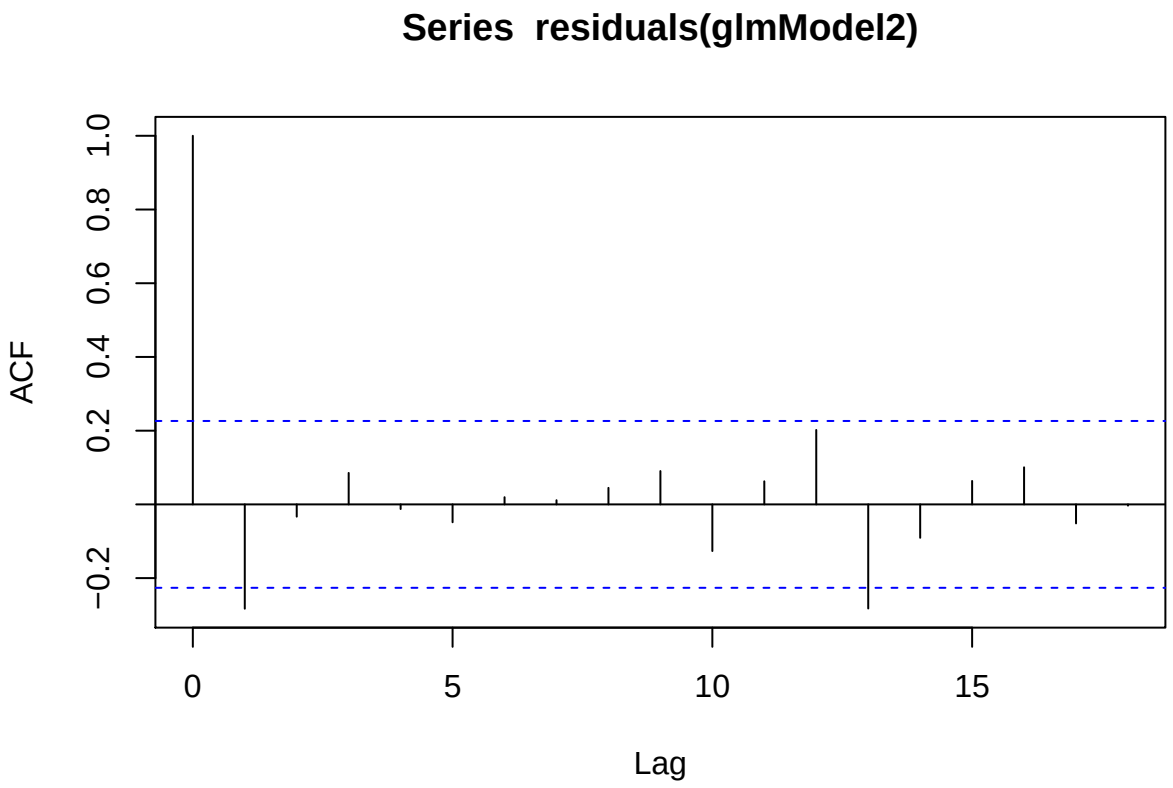


Figure 2: ACF plot for glm Model

Lastly we assess the mean-variance relationships by looking at the fitted values versus the residuals. A poisson model assumes that mean equals variance. The mean of a quasipoisson model equals variance multiplied by the dispersion parameter. By looking at the Pearson residuals we can analyse the mean-variance relationship, the Pearson residuals are standardised so they approximate a constant relationship. Looking at the pearson residual plot it is hard to tell if the relationship holds. Residual plots can be seen below. Not enough evidence to fulfill assumption according to Pearson residual plots but hard to tell.

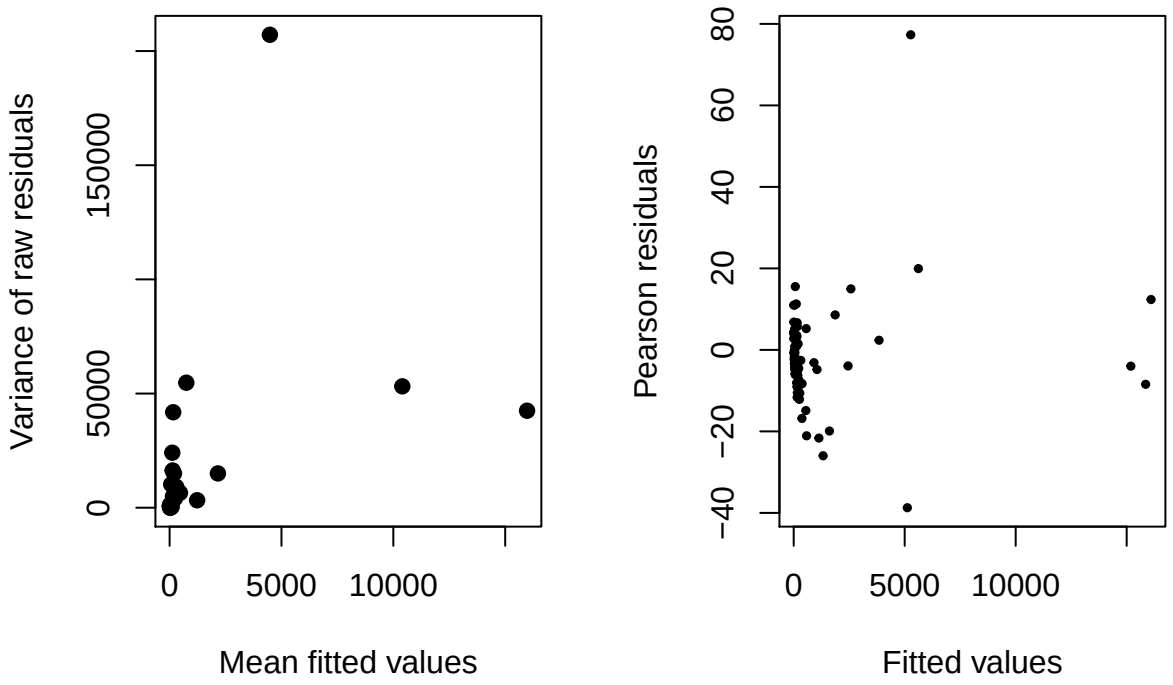


Figure 3: Pearson and raw residuals of glm

Using all-best subset selection the best models are identified using the QAIC to evaluate them which is a similar score as AIC which penalises for more parameters. The difference is that QAIC uses quasilikelihood rather than maximum likelihood. All Best-subset selction fits models with all possible combinations of covariates and calculates the QAIC score for each model. The lower the QAIC score the better the model fit. The top five models can be seen in the tables 6 and 7 below. The model with the lowest QAIC has all covariates except health expenditure. The next in line has all covariates except GDP. Third best has all covariates. Fourth best has all covariates except health expenditure and fifth best exlcudes health expenditure and lung diseases. Since the weights for the model with the lowest QAIC score is highest and the QAIC is lowest that model is the one that returns the best fit.

Table 6: First part of result of all-possible best subset selction

	(Intercept)	Bed	GDP	HealthExp	Lung	MedianAge	PopDensity
124	-5.797114	-0.1733684	-1.06e-05	NA	-0.0119197	0.0729067	0.0017522
126	-5.644515	-0.1739591	NA	-7.07e-05	-0.0106653	0.0693095	0.0014844
128	-5.797140	-0.1734226	-1.11e-05	3.50e-06	-0.0120174	0.0730119	0.0017673
60	-4.290366	-0.1720466	-7.00e-06	NA	-0.0163638	0.0619852	0.0018797
116	-7.174630	-0.1728255	-1.28e-05	NA	NA	0.0899675	0.0015639

Table 7: Second part of result of all-possible best subset selction

	UrbanPop	df	logLik	QAIC	delta	weight
124	0.0143785	7	-6980.612	86.48754	0.000000	0.4016366
126	0.0133571	7	-7132.521	88.02146	1.533916	0.1865295
128	0.0143504	8	-6980.328	88.48467	1.997129	0.1479661
60	NA	6	-7385.317	88.57410	2.086552	0.1414960
116	0.0208221	6	-7414.078	88.86451	2.376968	0.1223718

Since the model has correlated covariates a LASSO model might be a better fit. LASSO models reduce the weights of covariates which are not important in prediction. LASSO model can reduce the coefficents of the covariates to zero which effectively removes them from the model performing model selection. Penalised regression like LASSO adds a penalty term to the log-likelihood function we try to maximise. The reason a poisson model is fitted instead of quasipoisson is that the coefficients of the covariates stay the same for poisson and quasipoisson models. The only difference is the size of the standard errors. Before fitting the model we standardise the predictors by centering them and scale them by their standard deviation. This is done since the shrinkage for the predictors will have different contributions to the penalty term if the covariates have different scales.

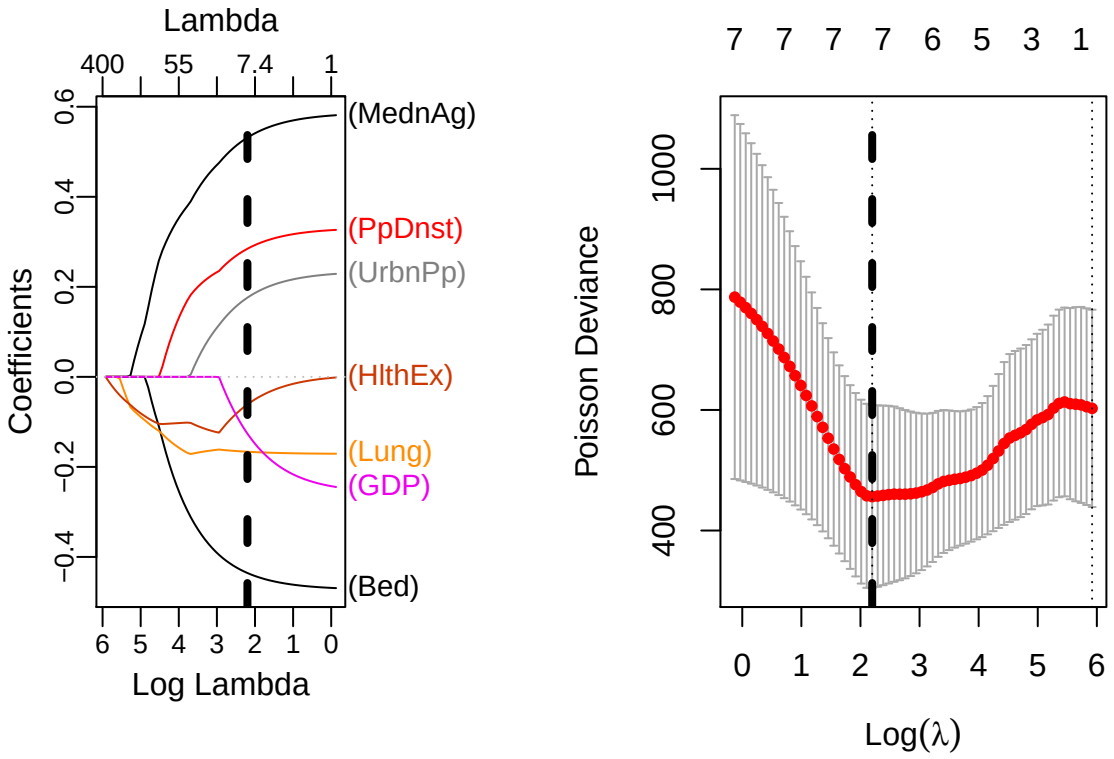


Figure 4: Results from Lasso Model

Table 8: LASSO coefficients at lambda min

Variables	Coefficients
Intercept	-3.05
PopDensity	0.29
MedianAge	0.54
UrbanPop	0.18
Bed	-0.44
Lung	-0.17
HealthExp	-0.06
GDP	-0.14

[a] Lambda min = 2.11

Figure 3 shows the log of the regularisation parameter which is called lambda that minimises the cross validation error as a dashed line. In the plot on the right hand side we see the evolution of the poisson deviance which is a goodness of fit measure. It starts decreasing until it hits the log lambda that minimises the cross validation error which is 2.11. The plot on the left shows how the coefficients of the covariates change by a change in the value of log lambda. The dashed line is again the log lambda that minimises the cross validation error. The best results using 10-fold cross validation returns a model where all covariates are kept in the model but most of them reduced by some amount. GDP seems to be the one affected the most with the greatest reduction, all of the others are quite close to the original estimate. In table 8 the coefficient estimates can be seen as well as the best log lambda.

Now we fit penalised regression splines with a smooth term for each covariate. We fit one model with five dimensions and another with 10 dimensions. For the first model all smooth functions are justified except population density and urban populations since they are not significant at the 95% level.The second model has all smooth functions justified except urban population and lung diseases. The first model has a much lower generalised cross validation score and higher deviance explained which indicates a better fit. This can be seen in table 9 and 10.

Table 9: Results from PRS model with 10 dimensions

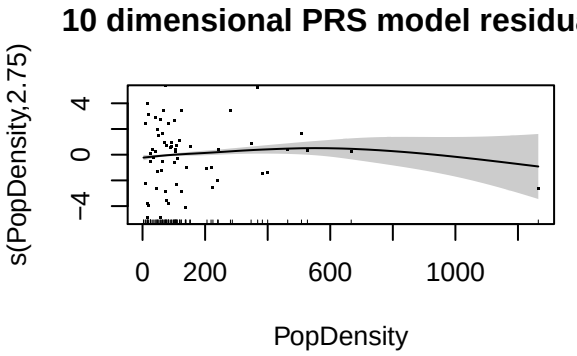
Covariates	edf	p.value
s(popDensity)	2.75	0.06
s(MedianAge)	1.00	0.05
s(UrbanPop)	1.00	0.24
s(Bed)	5.77	0.00
s(Lung)	7.50	0.00
s(HealthExp)	2.73	0.00
s(GDP)	8.85	0.00

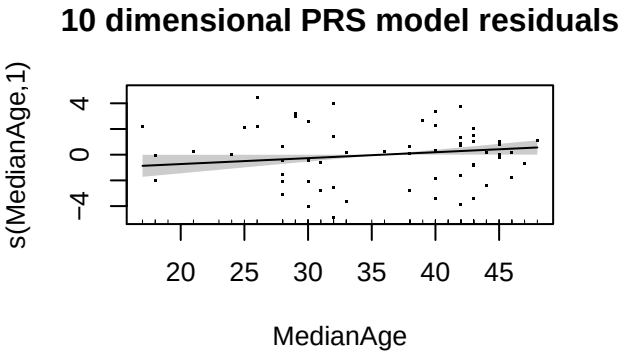
[a] Deviance Explained = 96.2%
[b] GCV = 51.255

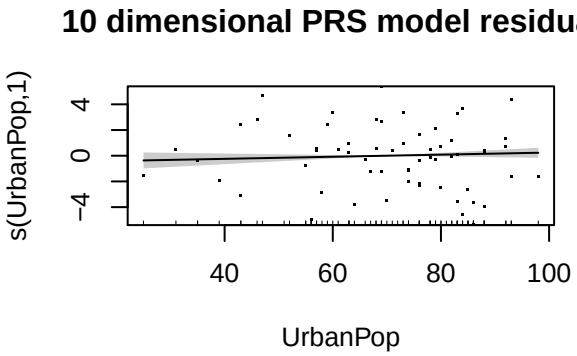
Table 10: Results from PRS model with 5 dimensions

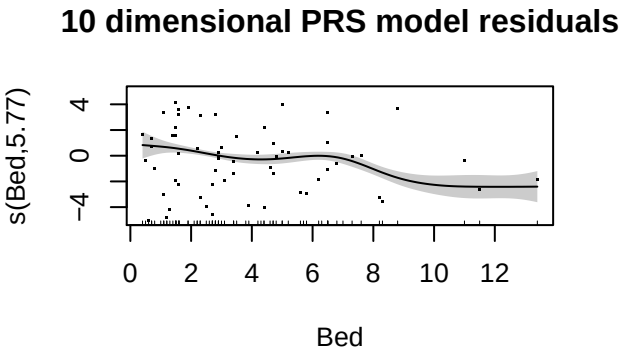
Covariates	edf	p.value
s(popDensity)	3.59	0.00
s(MedianAge)	1.00	0.00
s(UrbanPop)	1.55	0.02
s(Bed)	3.77	0.00
s(Lung)	1.41	0.17
s(HealthExp)	4.00	0.00
s(GDP)	2.12	0.00

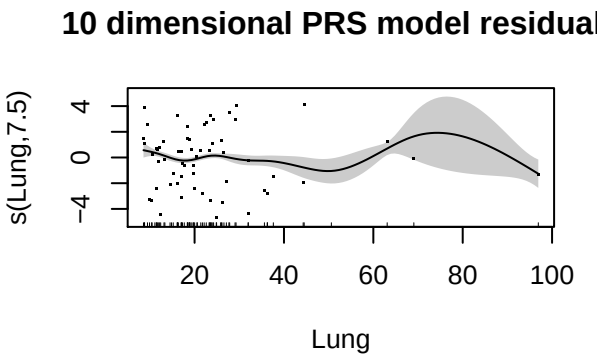
[a] Deviance Explained = 88.6%
[b] GCV = 94.605

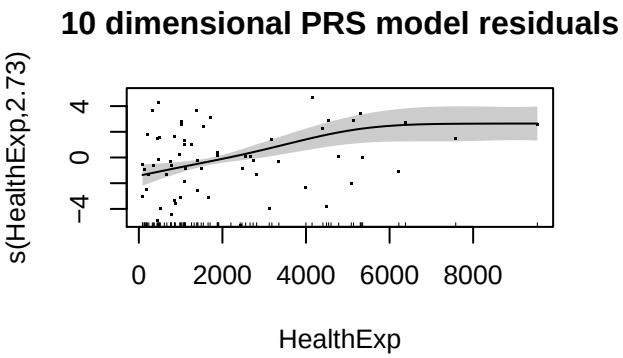


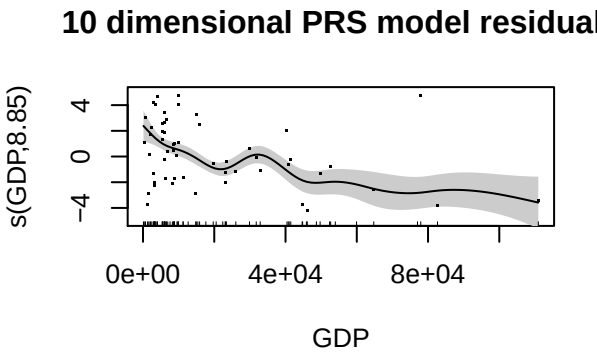


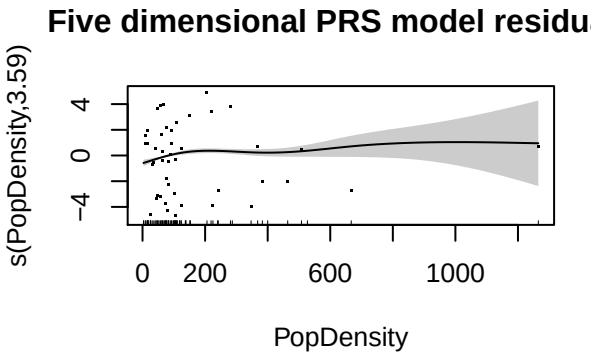


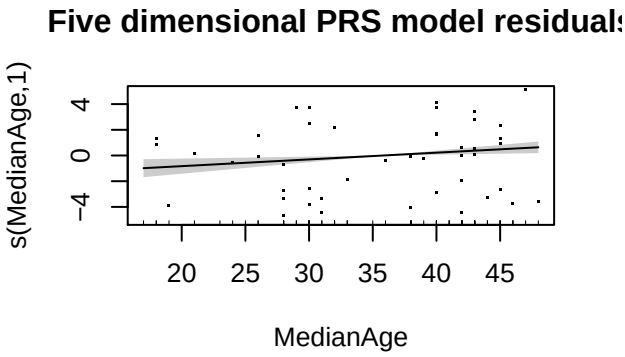


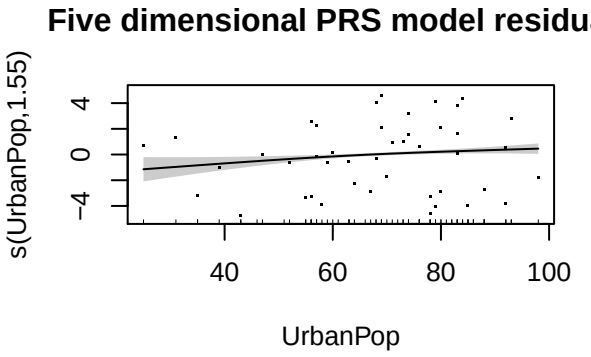

10 dimensional PRS model residuals
s(Lung,7.5)
4
0
−4
20
40
60
80
100
Lung

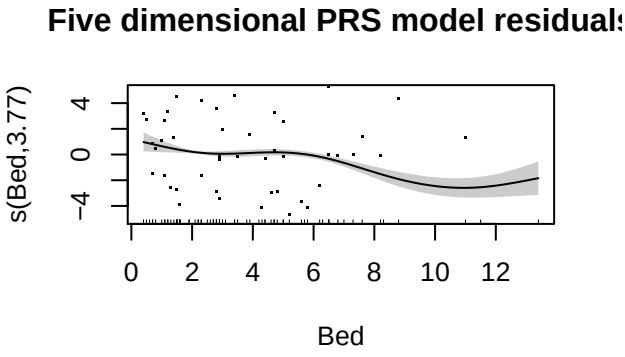

10 dimensional PRS model residuals
s(HealthExp,2.73)
4
0
−4
0
2000
4000
6000
8000
HealthExp

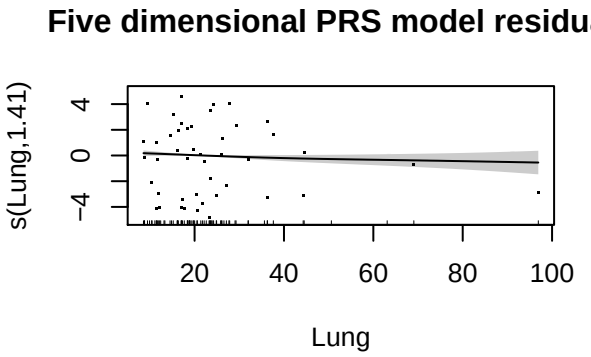

10 dimensional PRS model residuals
s(GDP,8.85)
4
0
−4
0e+00
4e+04
8e+04
GDP

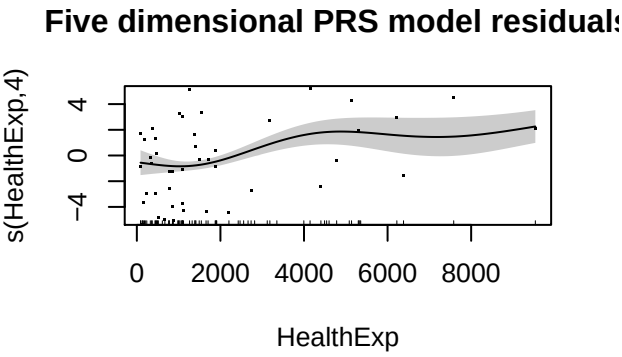

Five dimensional PRS model residuals
s(PopDensity,3.59)
−4
0
4
0
200
600
1000
PopDensity

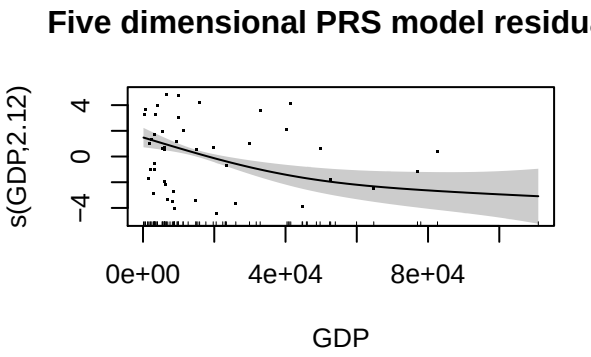

Five dimensional PRS model residuals
s(MedianAge,1)
−4
0
4
20
25
30
35
40
45
MedianAge


Five dimensional PRS model residuals
s(UrbanPop,1.55)
−4
0
4
40
60
80
100
UrbanPop


Five dimensional PRS model residuals
s(Bed,3.77)
−4
0
4
0
2
4
6
8
10
12
Bed





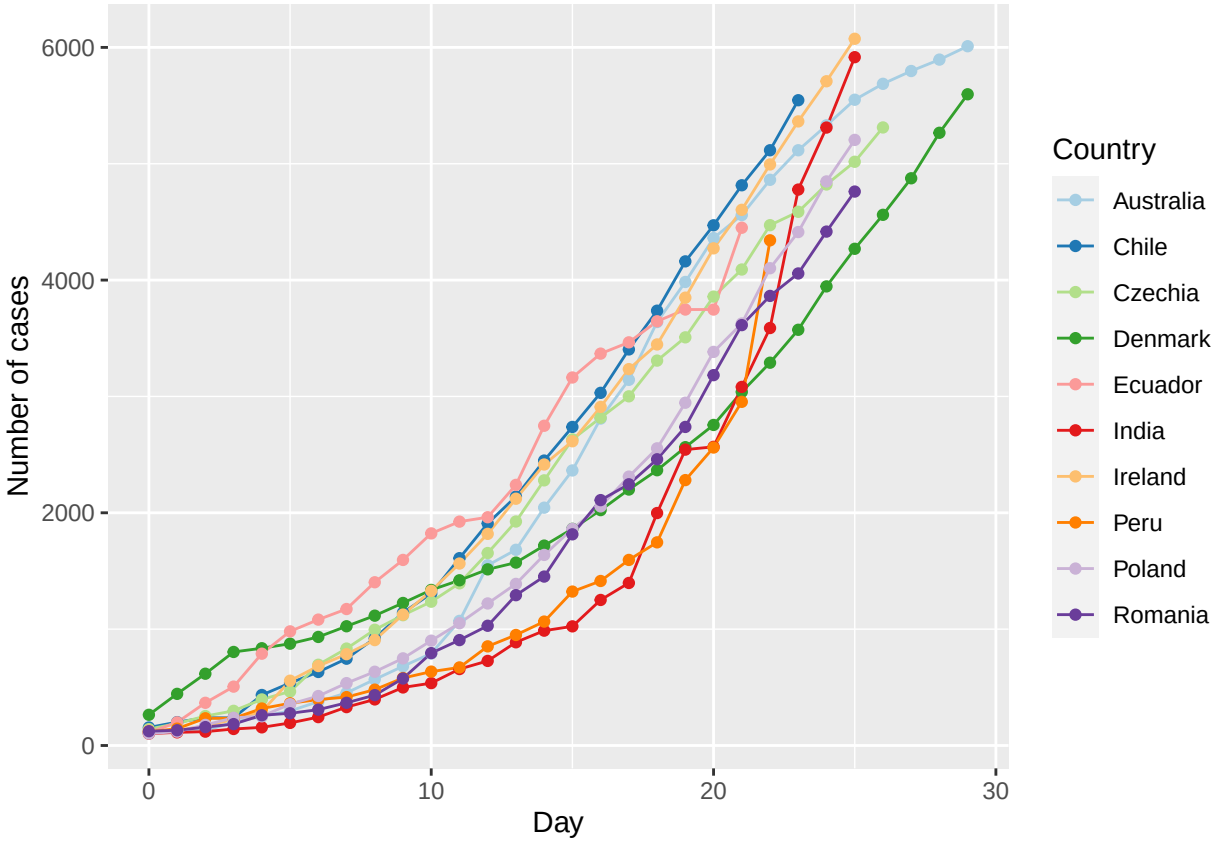


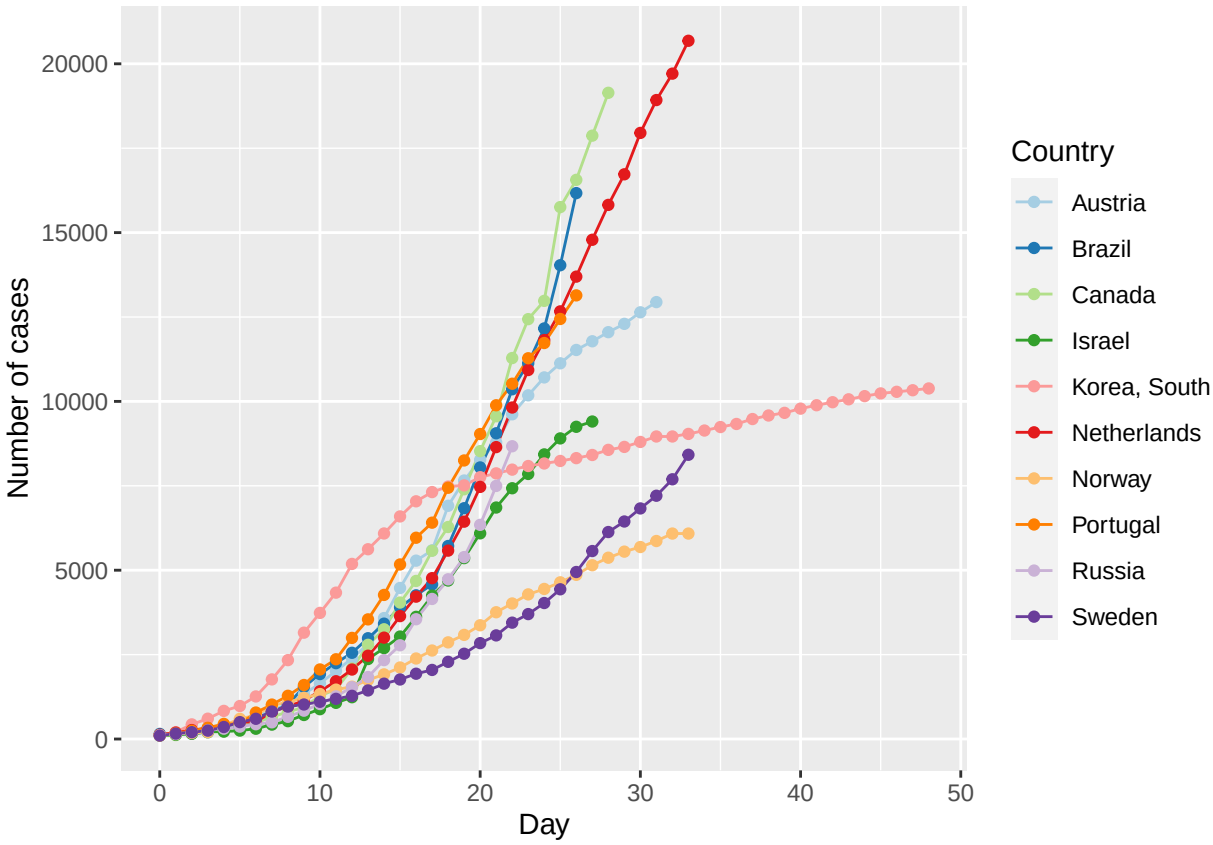
By looking at the residual plots of these models seen in the figures above it can be seen that the one with 10 dimensions fluctuates more and are more nonlinear. All covariates have similar directions in both models the main difference is that the first model fluctuates much more. The biggest difference can be seen in the residuals of lung diseases.

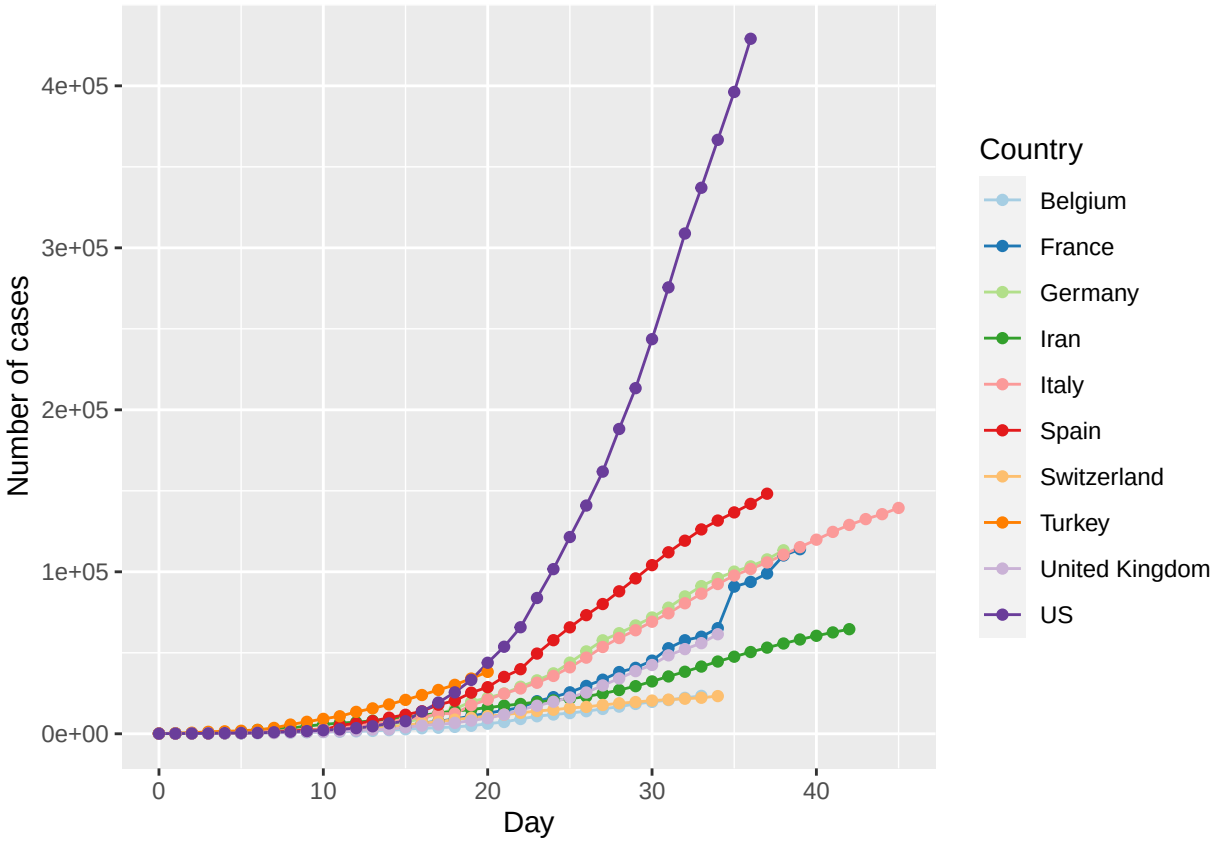
3 Part 2

Now we move on to another dataset that includes the number of confirmed cases of each country. We start by reading in the data and cleaning it.

Before fitting models the dataset will be explored with plots. The first plot shown in the figures below show the evolution of confirmed cases of each country by day. It can be seen that most countries follow a similar trajectory. The US has the steepest curve. The bar charts in figure 4 show total cases by country, USA are far ahead but Spain, Italy ,France and Germany have many cases also.


Number of cases
6000
4000
2000
0
0
10
20
30
Day
Country
Australia
Chile
Czechia
Denmark
Ecuador
India
Ireland
Peru
Poland
Romania


Number of cases
20000
15000
10000
5000
0
0
10
20
30
40
50
Day
Country
Austria
Brazil
Canada
Israel
Korea, South
Netherlands
Norway
Portugal
Russia
Sweden


Number of cases
4e+05
3e+05
2e+05
1e+05
0e+00
0
10
20
30
40
Day
Country
Belgium
France
Germany
Iran
Italy
Spain
Switzerland
Turkey
United Kingdom
US

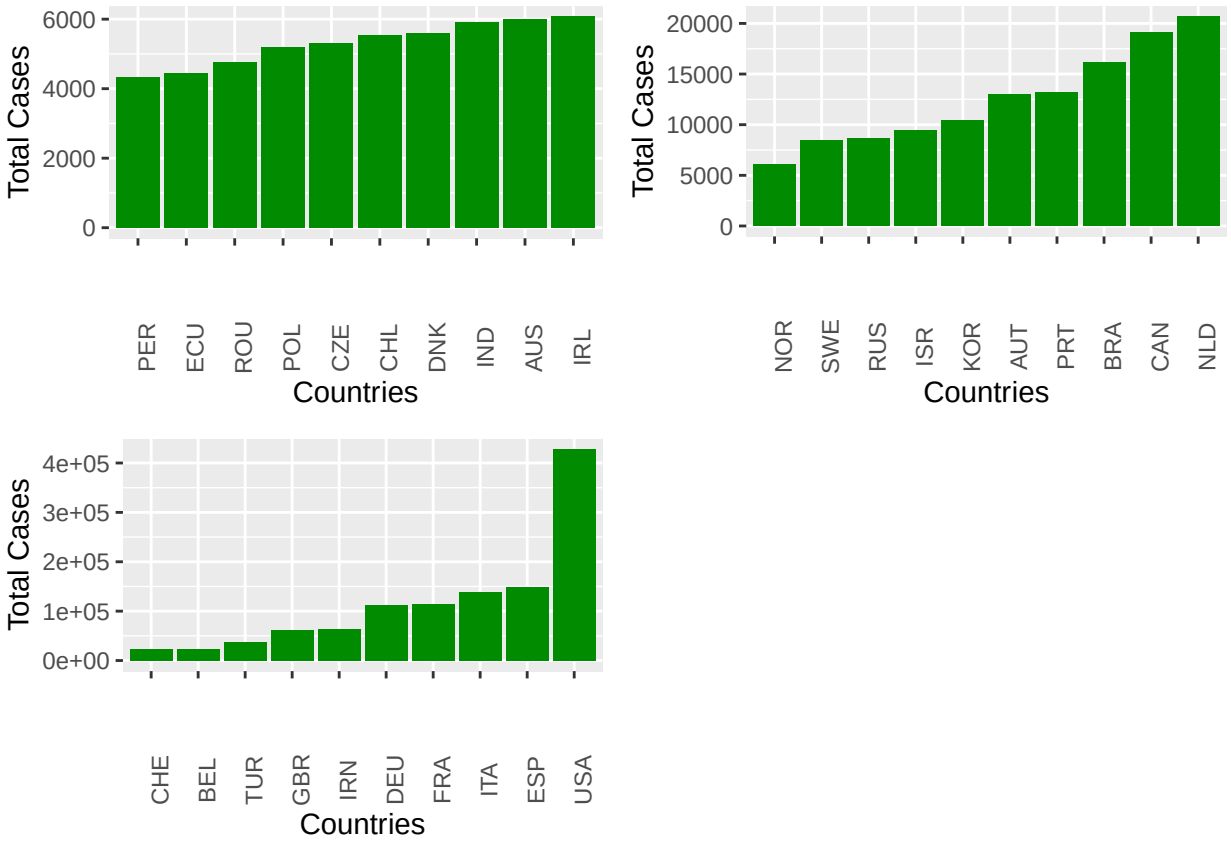


Figure 5: Bar Charts for number of cases by country

By looking at the plot in figure 12 we can see that Germany has a faster growth rate than the UK and the average country but seems to be slowing down. United Kingdom has a very similar growth rate as the average. By looking at the case fatality rates in table 11 we see that Germany has a much lower fatality rate. This seems to indicate that growth rate of cases and fatality rate are not related at least for these countries. A possible reason is that Germany might test more people even though they dont have serious symptoms. By having less restrictions on testing the more cases you have but many of the cases end up being not serious. This could mean a lower fatality rate.

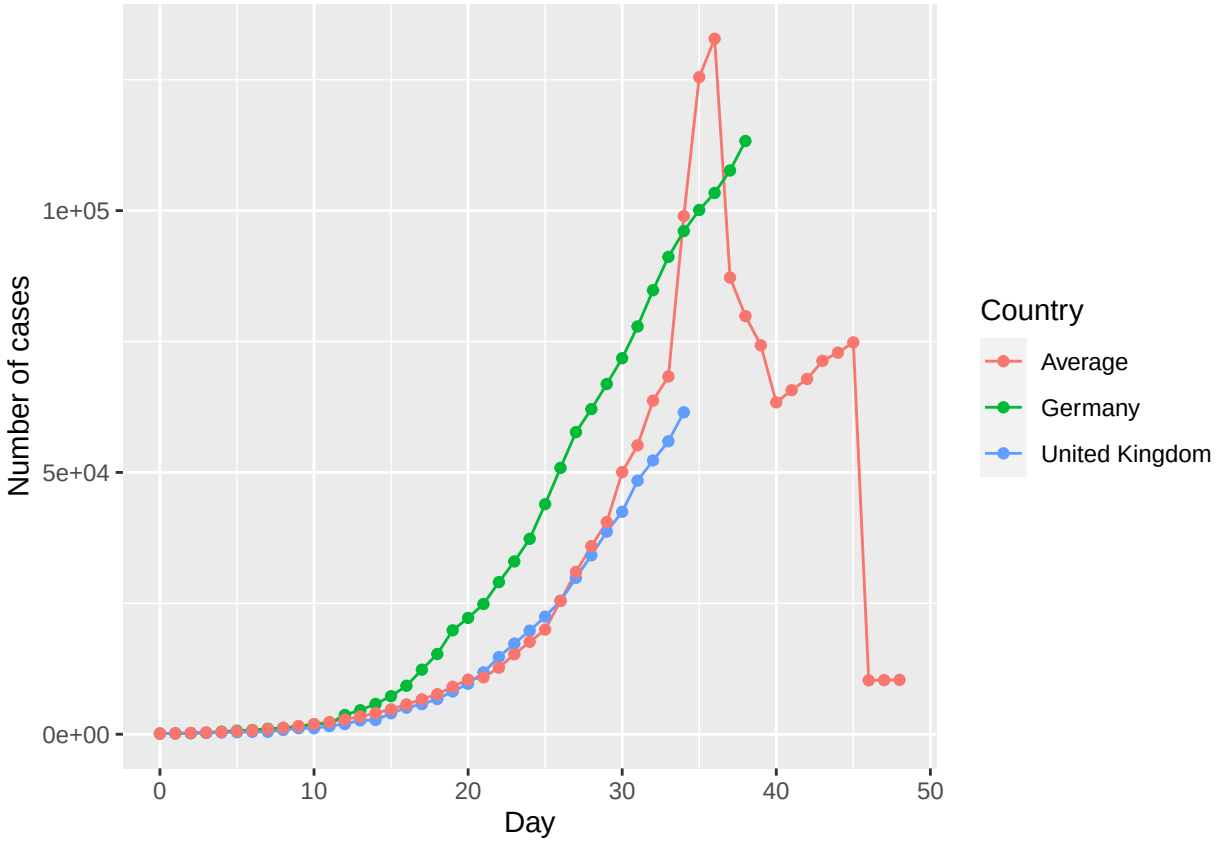


Figure 6: Comparison of observed trajectories

Table 11: Fatality rate of UK and Germany

Country	fatalityRate
Germany	0.0207333
United Kingdom	0.1156749

This model has day as a single predictor but allows each country and intercept to have its own intercept and slope. We use a AR(1) within group error structures since present values for cases include past values. We use a poisson model since we are modelling counts of confirmed cases. We have population average parameters and country specific parameters. The population average parameters show the expected amount of confirmed cases by day for the average country. From table 12 the results of the model can be seen. The link function is log since it is a poisson model so the average country has a baseline number of cases as 566 and each one day increase mulitplies the cases by 1.13. The country specific paramaters seen in table 13 show the difference in the coefficent by day and intercept.So for example for Australia the difference in intercept is -0.41 and -0.015 for day. That means the coefficients or australia is 5.93 for the intercept and 0.11 for the slope.

Table 12: Results from slope and intercept model

Covariate	Estimate	p.value
Intecept	6.34	0
Day	0.12	0

[a] Variance for intercept = 0.82

[b] Variance for slope = 0.001, Correlation of fixed effects = -0.508

Table 13: Country-Specific coefficients for slope and intercept model

	(Intercept)	Day
Australia	-0.4138851	-0.0150413
Austria	0.2898791	-0.0198177
Belgium	-0.3510642	0.0082267
Brazil	-0.4641785	0.0278509
Canada	-0.4569007	0.0269779
Chile	-0.5275739	0.0070920
Czechia	-0.3128216	-0.0162453
Denmark	-0.0230255	-0.0414509
Ecuador	-0.1126347	-0.0125940
France	0.3988252	0.0077800
Germany	1.1377545	-0.0050754
India	-1.7804805	0.0442171
Iran	1.5502307	-0.0419456
Ireland	-0.4540766	-0.0021814
Israel	-0.6380897	0.0166668
Italy	1.8193406	-0.0333026
Korea, South	1.6150102	-0.0898353
Netherlands	0.0751099	-0.0081611
Norway	0.0046718	-0.0428291
Peru	-1.3683717	0.0241520
Poland	-0.9140245	0.0083827
Portugal	-0.0275303	0.0092118
Romania	-1.0707109	0.0125578
Russia	-1.2280287	0.0607679
Spain	1.4299970	-0.0022382
Sweden	-0.4036710	-0.0253159
Switzerland	0.6015743	-0.0218563
Turkey	0.7658582	0.0581584
United Kingdom	-0.2217763	0.0274365
US	1.0810438	0.0383922

To get a better view of the country specific coefficients we plot the slopes in a bar chart below. The three biggest deviations are for South Korea, Russia and Turkey. Those countries are analysed further by plotting the slopes for those countries and compare to the average country fit. It can be seen in figure 14 that turkey has the biggest growth rate while south korea has the smallest growth rate of these countries. Comparing to the actual observations it seems similar to them but hard to see since we only have limited observations for turkey and russia. We make thea axis smaller to see the comparison better. Comparing figure 15 of fitted values and figure 16 of actual values the fits seem reasonably close to the actual values.

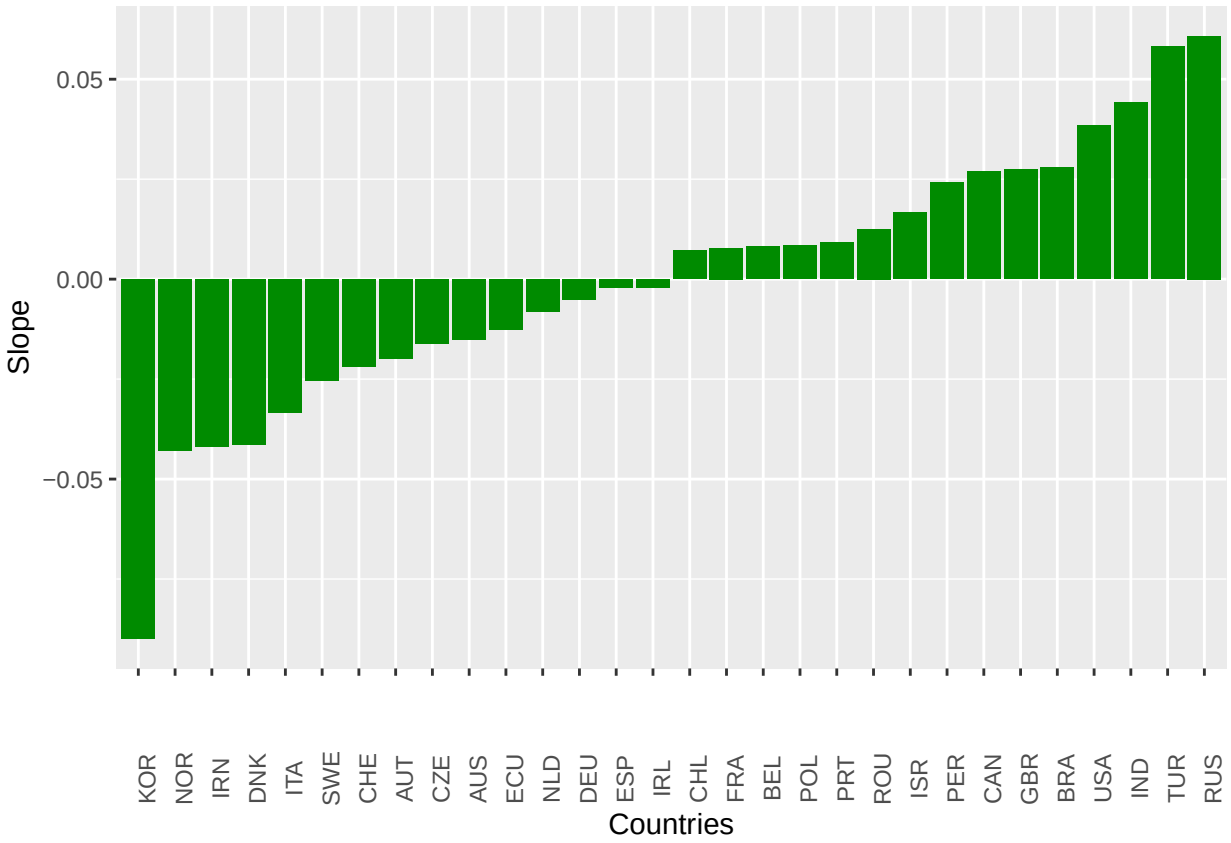


Figure 7: Bar chart for slope and intercept model

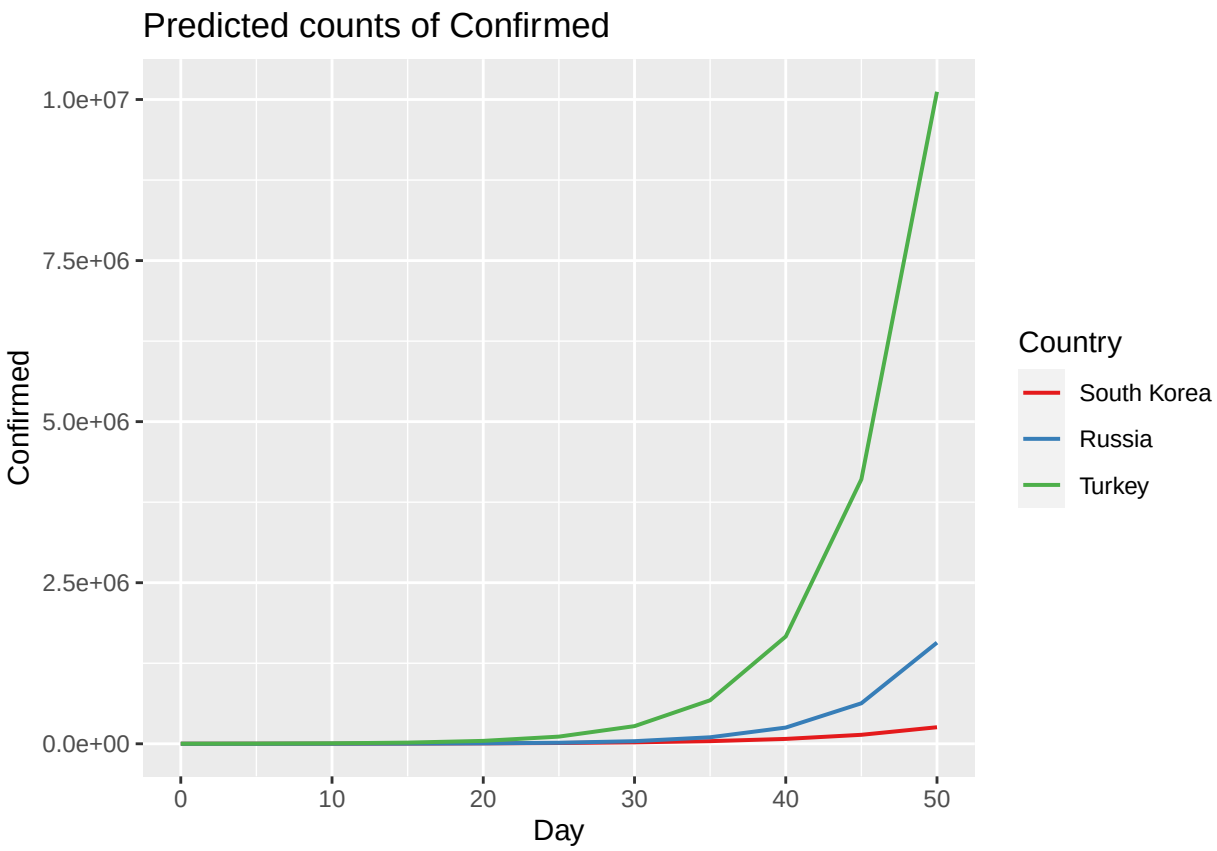


Figure 8: Fitted trajectories for slope and intercept model

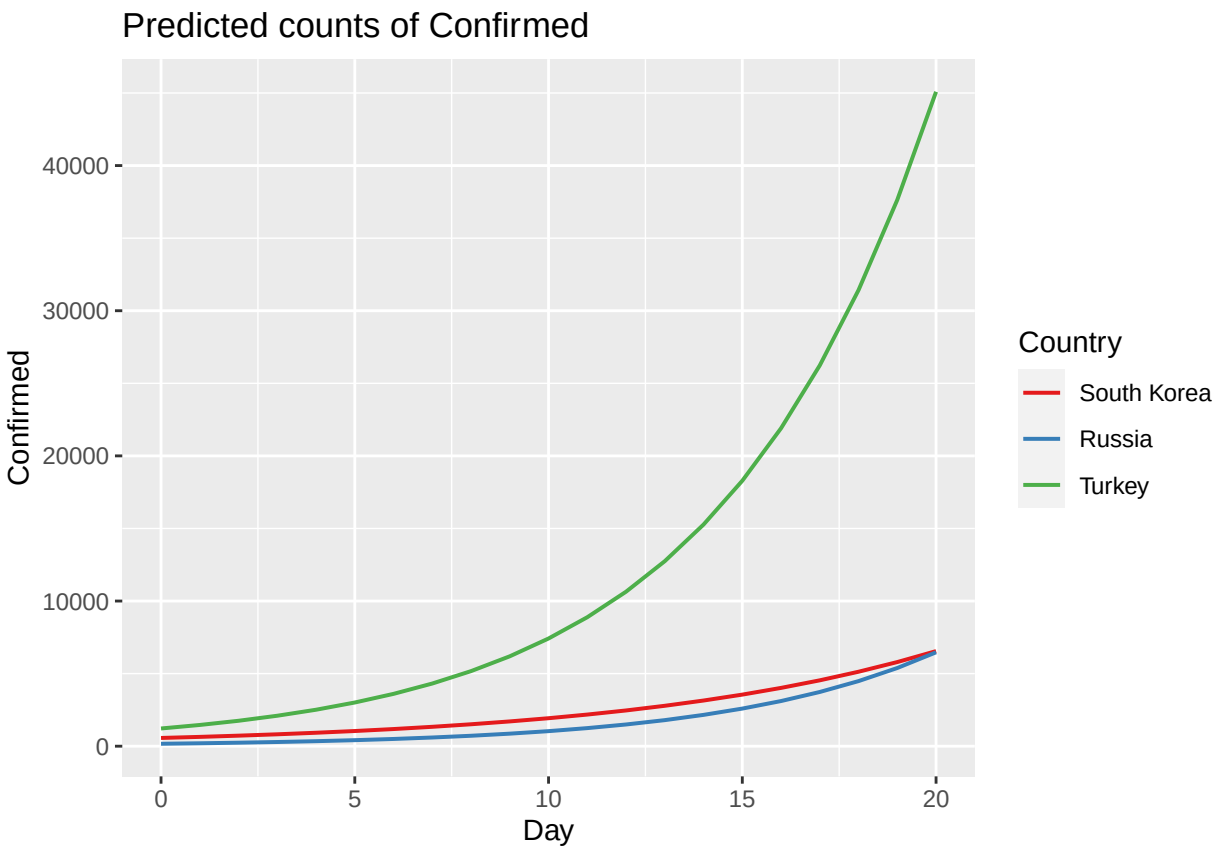


Figure 9: Fitted trajectories for slope and intercept model, tighter axis

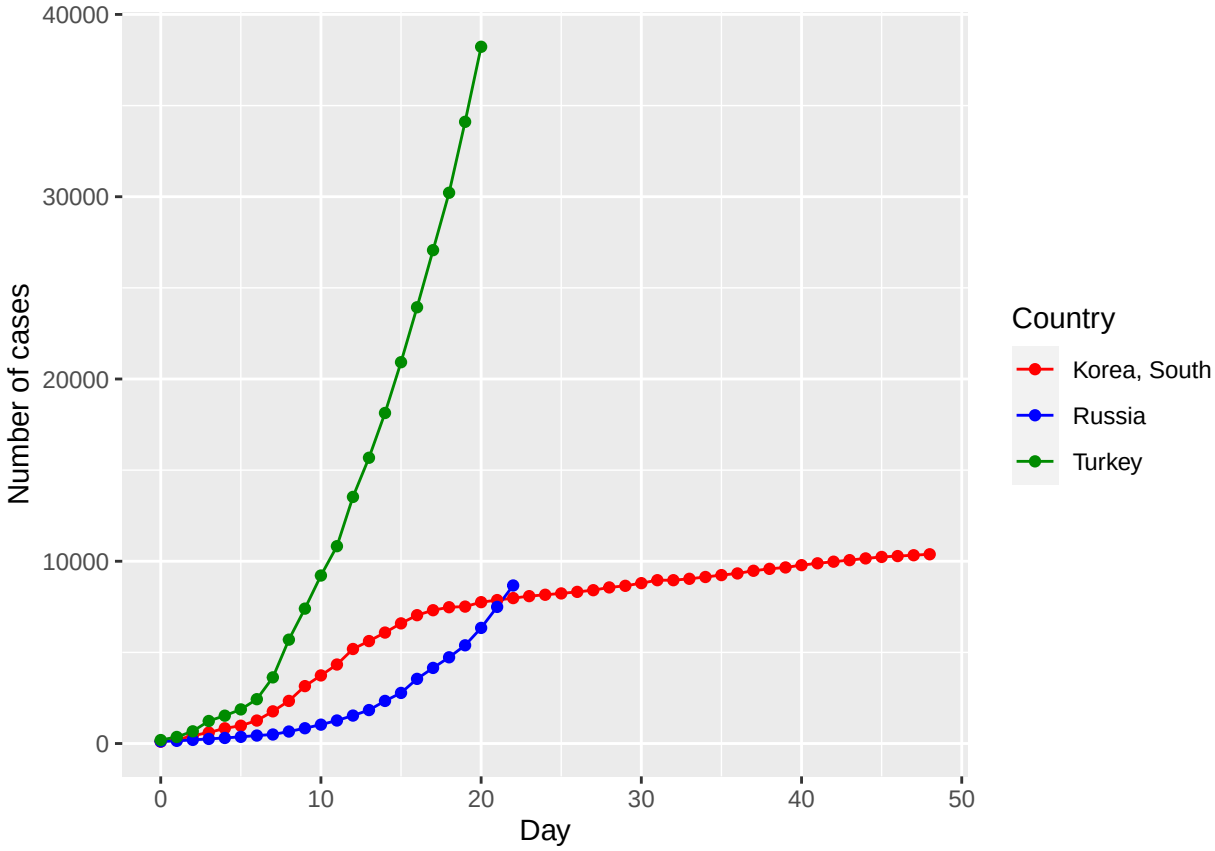


Figure 10: Observed trajectories

```
## $Day
```

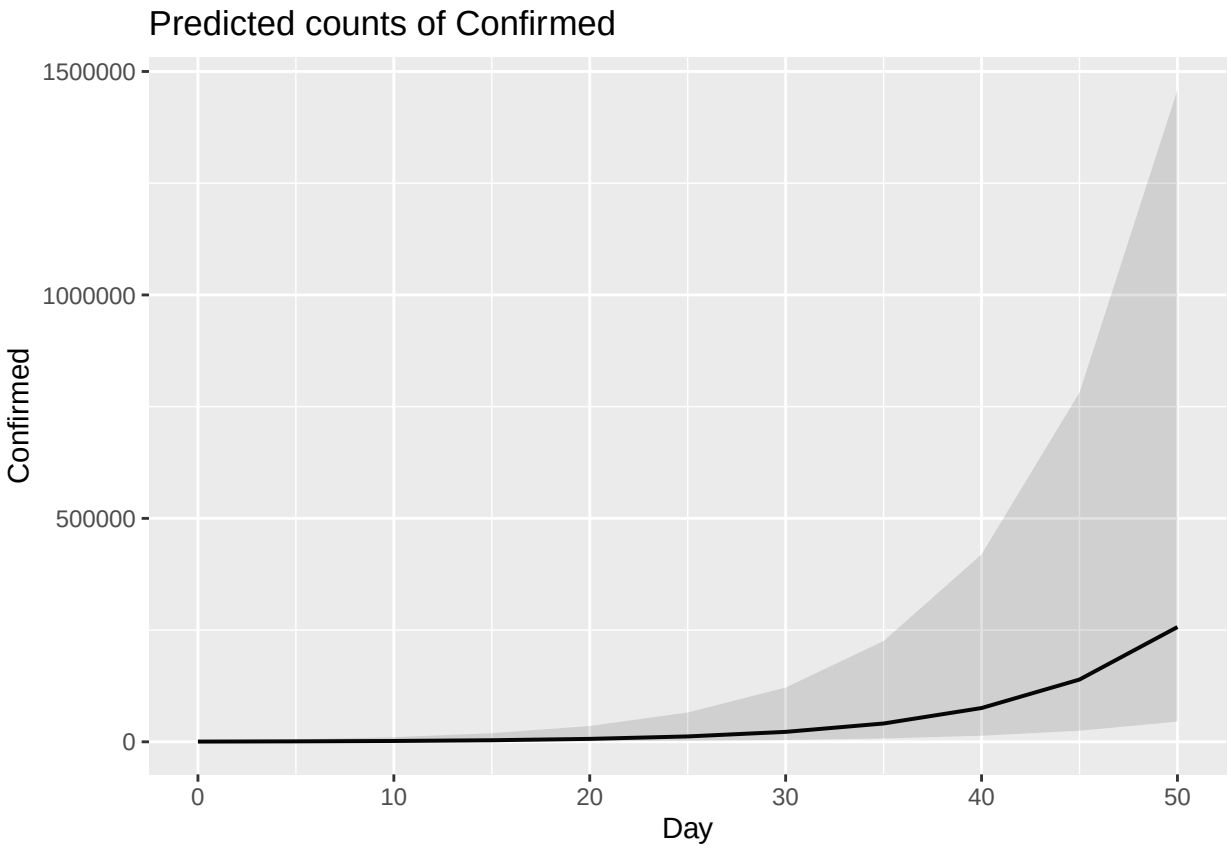


Figure 11: Model for average country

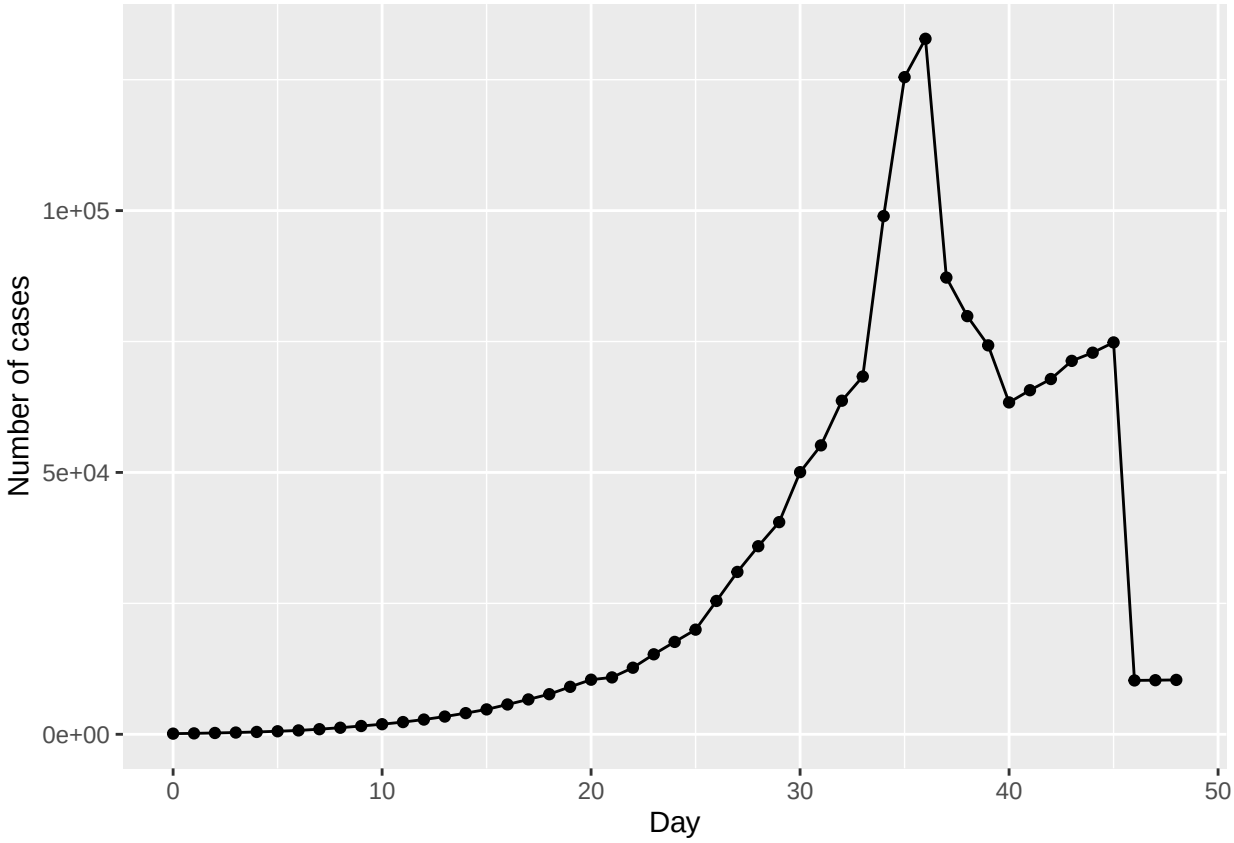


Figure 12: Observed values for average country

Figure 17 and 18 show average country fit and average country actual values. By comparing figure 14 and figure 17 it can be seen that Turkey and Russia have a higher growth rate than the average. Hard to see from the plot for South Korea but looking at the country specific parameters in table 13 it can be seen that South Korea has a lower growth rate than the average. Comparing figure 17 and 18 it can be seen that fittend and actual values start at a similar trajectory but the actual values go down while the fitted values are always rising.

The last model that is fitted has a common intercept and different slopes for each country. Results in table 14 indicate that the intercept is slightly larger than in the previous model while the slopes are slightly smaller. The country specific slopes can be seen in table 15. We plot the same plots as for the previous model identifying the three countries that deviate the most. This time they are Peru, Turkey and USA as seen in figure 19. They seem to follow a similar trajectory as the observed values with Turkey with the highest growth rate and Peru the smallest. This can be seen by comparing figure 21 and figure 22. Comparing figure 23 and 20 we see that Turkey and USA have a much higher growth rate than the average country but hard to tell for Peru because of the range of the y-axis. Comparing the plots for the average country we get similar results as in the previous model. The fitted values are always rising while the observed values go down after some time.

Table 14: Results form common intercept model

Covariate	X..Estimate	X.p.value
Intecept	7.15	0
Day	0.81	0

[a] Random effect variance = 0.0015
[b] Correaltion of fixed effects = -0.007

Table 15: Country-Specific coefficients for common intercept model

	Day
Australia	-0.0272454
Austria	0.0006991
Belgium	0.0070422
Brazil	0.0092266
Canada	0.0134491
Chile	-0.0257308
Czechia	-0.0303575
Denmark	-0.0375285
Ecuador	-0.0283353
France	0.0367465
Germany	0.0463244
India	-0.0447359
Iran	0.0205295
Ireland	-0.0252011
Israel	-0.0084047
Italy	0.0343634
Korea, South	-0.0268789
Netherlands	0.0059909
Norway	-0.0328716
Peru	-0.0630864
Poland	-0.0382940
Portugal	0.0108187
Romania	-0.0421411
Russia	-0.0126406
Spain	0.0586186
Sweden	-0.0305495
Switzerland	0.0119652
Turkey	0.0969469
United Kingdom	0.0330376
US	0.0882541

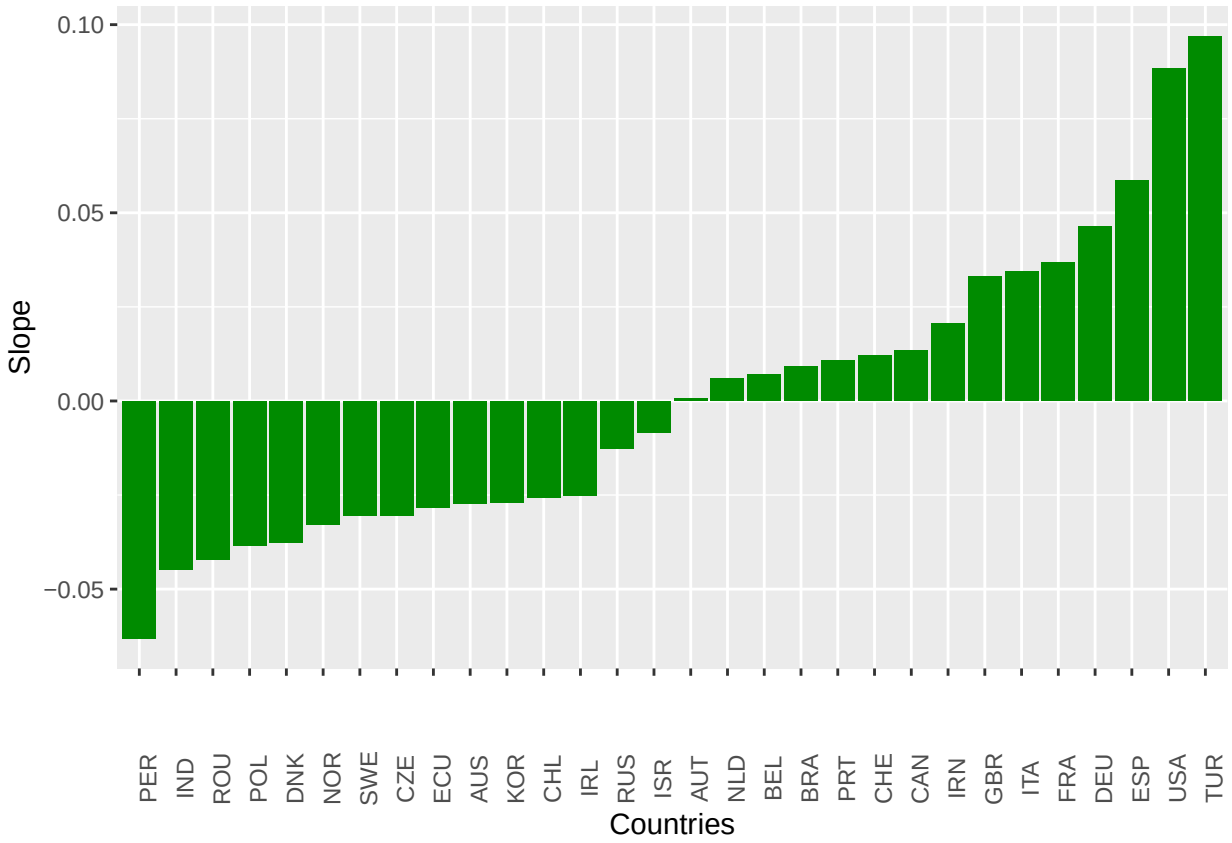


Figure 13: Bar charts for common intercept model

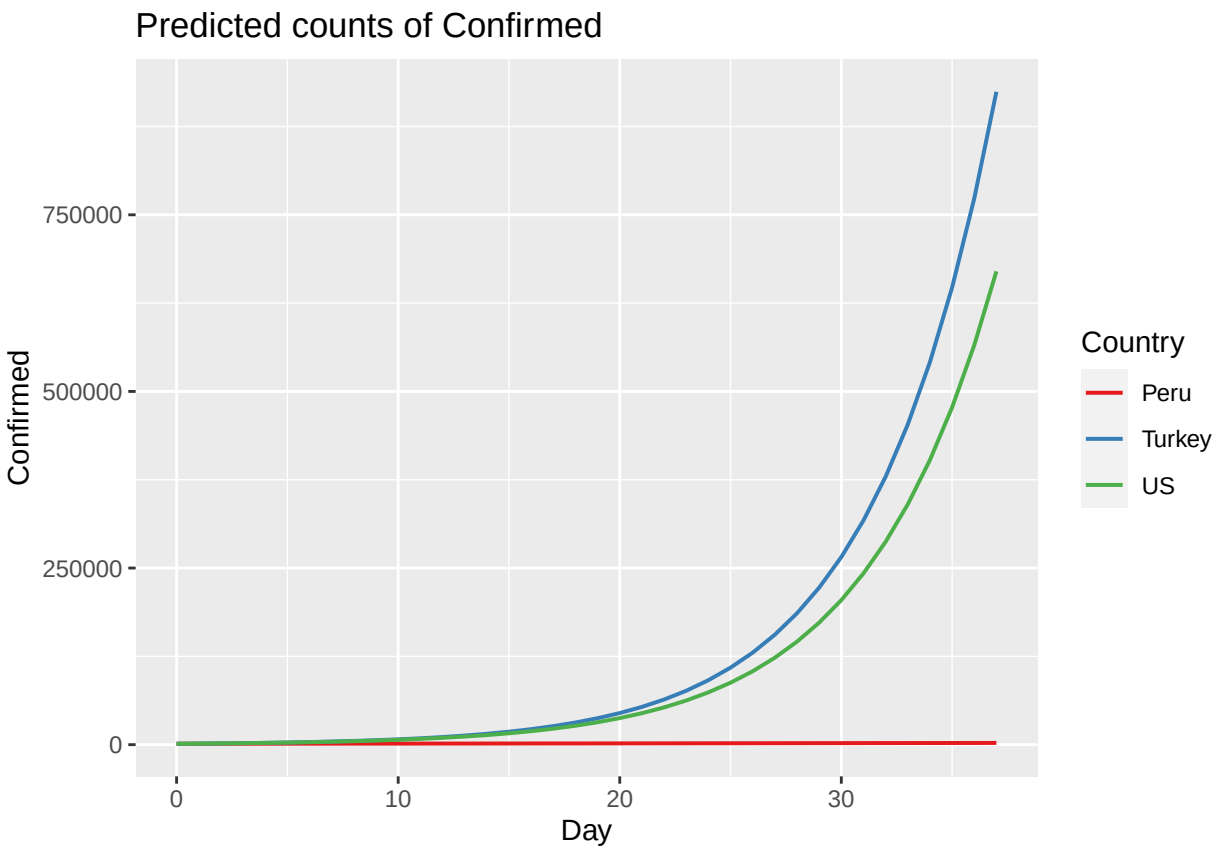


Figure 14: Trajectory for common intercept model, tighter axis

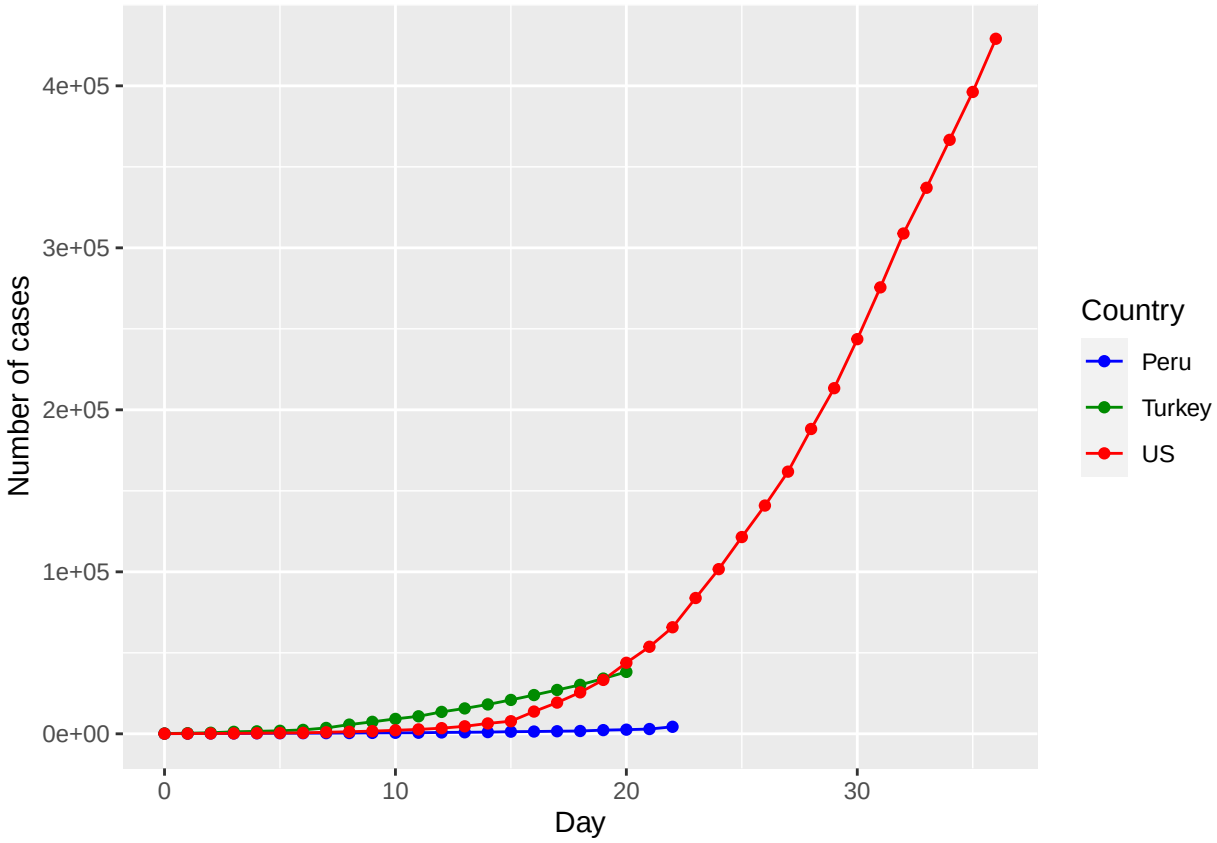


Figure 15: Observed trajectory for countries

```
## $Day
```

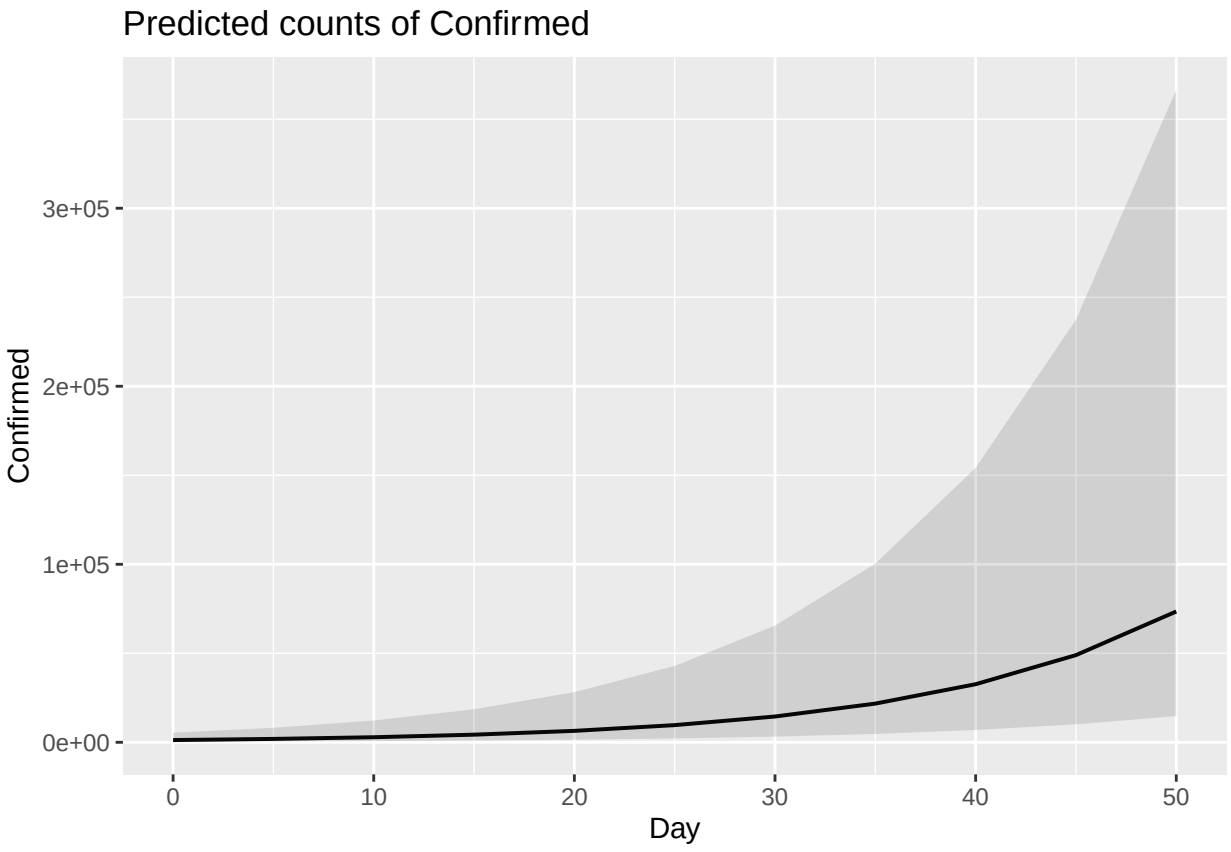


Figure 16: Average country plot for common intercept model

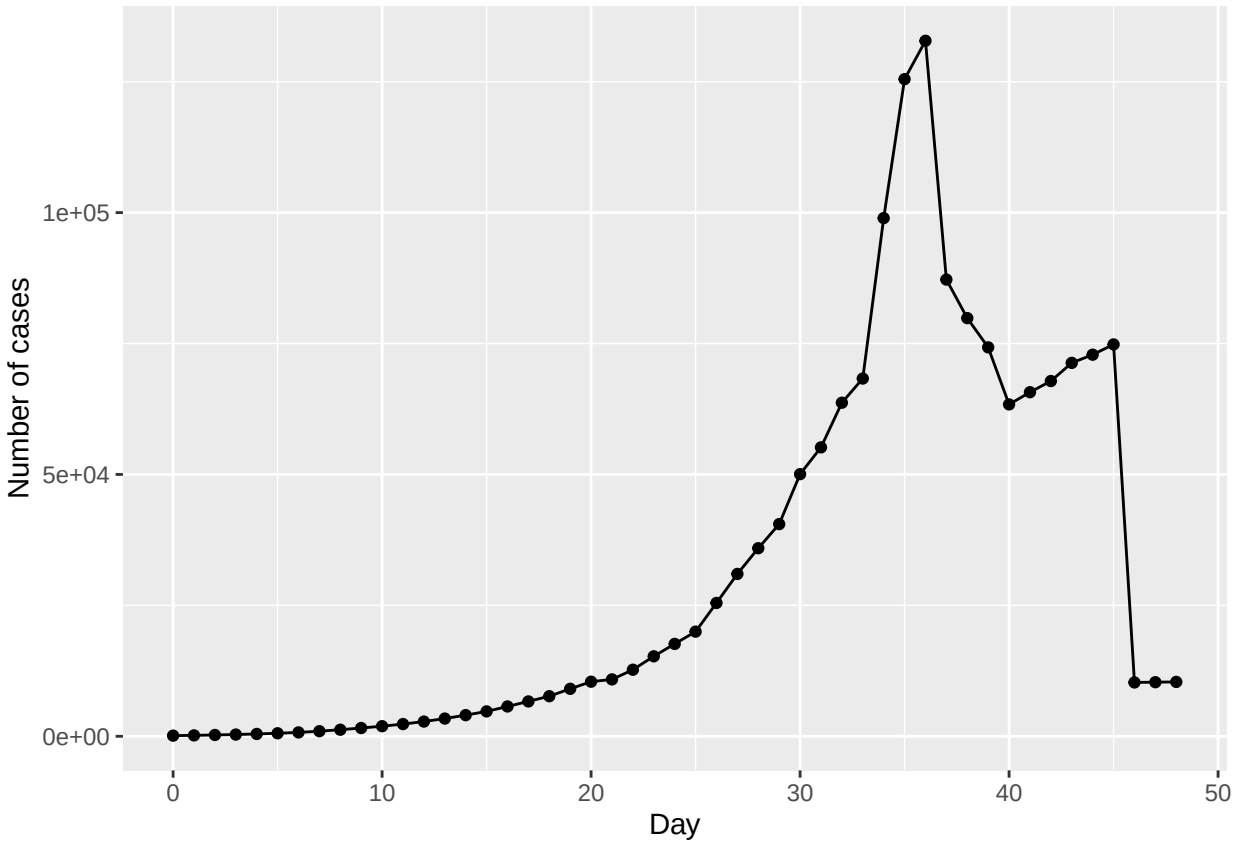


Figure 17: Observed average country trajectory

By looking at the comparison between average country fitted and the observed values. It can be seen that the models dont deal well with values after the peak of the outbreak. The reason is that the models have values that continue to rise while the observed values will eventually go down after the peak. This can be seen by comparing the model fits with figure 24 of the actual observations.