PREDICTING COVID-19 CASE COUNTS

Jeremy Meyer
BYU Department of Statistics

April 20, 2020

1 Introduction

The recent coronavirus disease (COVID-19) is a respiratory illness that has caused a global pandemic. COVID-19 emerged as a novel coronavirus near Wuhan, China in December 2019. The virus spreads through droplets released in the air when an infected person coughs or sneezes. The droplets usually only spread a few feet, so physical distancing is an effective way to stop the spread [1]. Since there is currently no vaccine for the virus, physical and social distancing efforts have been made to "flatten the curve" of new cases so that hospitals are not overwhelmed. However, such efforts have caused major disruptions to our lives and the economy and the question remains as to how long and how much the disease will continue to spread.

In this project, we explore the effect of various demographics on the number of new COVID-19 cases reported each day in 29 different countries. We also generate predictions for the remainder of April 2020 and answer research questions such as when the number of new cases will peak and how many cases will be present at the end of the month. We collect and merge data from March 11 - April 17, 2020 using several internet sources.

Answering these questions can help hospital officials and governments prepare for the future. For example, knowing the amount of potential new cases can help hospitals plan for the number of ICUs or ventilators needed for the sudden influx of COVID-19 patients. Governments can also use this information to know how long the disease will spread and if necessary, instate social distancing mandates on their countries. These future predictions have the potential to save lives and keep the hospitals from being overwhelmed during the pandemic.

To answer the research questions, we build a negative binomial generalized linear mixed model (GLMM) and treat both time and the country as random effects. We consider alternative models (such as a Poisson GLMM) and perform backward AIC variable selection on the country demographics to create a final model. We explore different ways to get a GLMM R^2-like measure for evaluating model fit. We use the final model to determine both how the country demographics affect case counts and later use the model to generate predictions for future case counts.

2 The Data

The data used for the analysis were collected and merged from the following websites: *Kaggle* [2], *the World Bank* [3], and *Our World in Data* [4]. In addition to modeling the number of cases, we also wanted to determine if certain country demographics (such as country temperature) affected the number of case counts. These demographics will later be used to identify potential trends between countries in predicting the number of cases of COVID-19. The demographic covariates used in this analysis are shown in Table 1.

Since the COVID-19 outbreak is so recent, one challenge is finding data that is complete and well-maintained. For simplicity, we only kept countries that had complete data for all demographics. We also wanted to include information on the number of COVID-19 tests since a case can only be confirmed if a test is performed. However, many countries did not start reliably reporting test data until early April. We used the cumulative number of tests per country as of April 10, 2020 because it was complete for the most number of countries. We then divided each country's cumulative test count by the total population and multiplied the result by 1000. Thus, we created a number of tests per 1000 people (testCapita) variable that measures how rigorous a country does testing.

After merging the data together and removing all missing observations, we ended up with data from 29 different countries. The response variable in the analysis is the number of new cases of COVID-19 per day, which came from *Our World in Data*. We were able to have complete data for 37 days (March 11 - April 17, 2020) for the 29 countries. A pairs plot of the demographics against the log number of new cases per person on April 10, 2020 is shown in Figure 1.

Table 1: Table of demographic covariates used for analysis (fixed across time)

Source	Variable	Units	Demographic
All	Country	Name	Country Name (29 unique after removing missing data)
Kaggle	pop	persons	Country population
Kaggle	smokers	%	Percentage of smokers in the population
Kaggle	avgtemp	°C	Average yearly temperature
Kaggle	avghumid	%	Average yearly humidity in the country
Kaggle	urbanpop	%	Percentage of population that resides in urban areas
Kaggle	lung	deaths/100K people	Rate from lung diseases per 100,000 people
World Bank	perc65	%	Percentage of population over 65
Our World in Data	testCapita*	tests/1K people	Total COVID-19 tests per 1,000 people (April 10th 2020)

*This variable was created by dividing the cumulative number of tests on April 10th by the country population and multiplying by 1000.

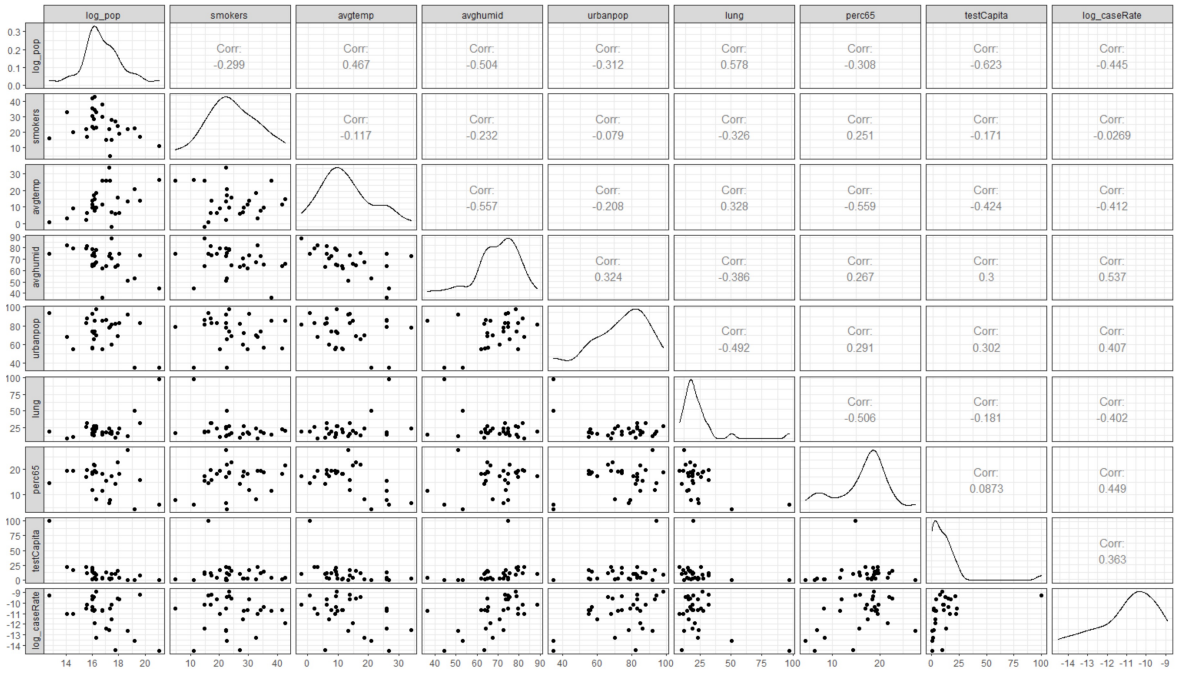

Figure 1: Plot of the covariates against the log of the number of new cases per person in the 29 countries. The number of new cases is the only variable that is not fixed over time. The number of new cases in this graph is fixed on April 10th, 2020.

Since the population size of the country has a large influence on the number of cases, we later control for population by modeling the number of new cases as a rate per person in the population.

Unlike the demographics, the case counts varied over time. Plots of a few of the 29 countries are shown in Figure 2. Although the countries are at different stages of case numbers, most countries first experience a sharp increase in the number of new cases per day, followed by a leveling off and decline of new cases. On the original scale, the variance increases with the number of case counts, thus we may be able to capture the variability with a negative binomial or Poisson model.

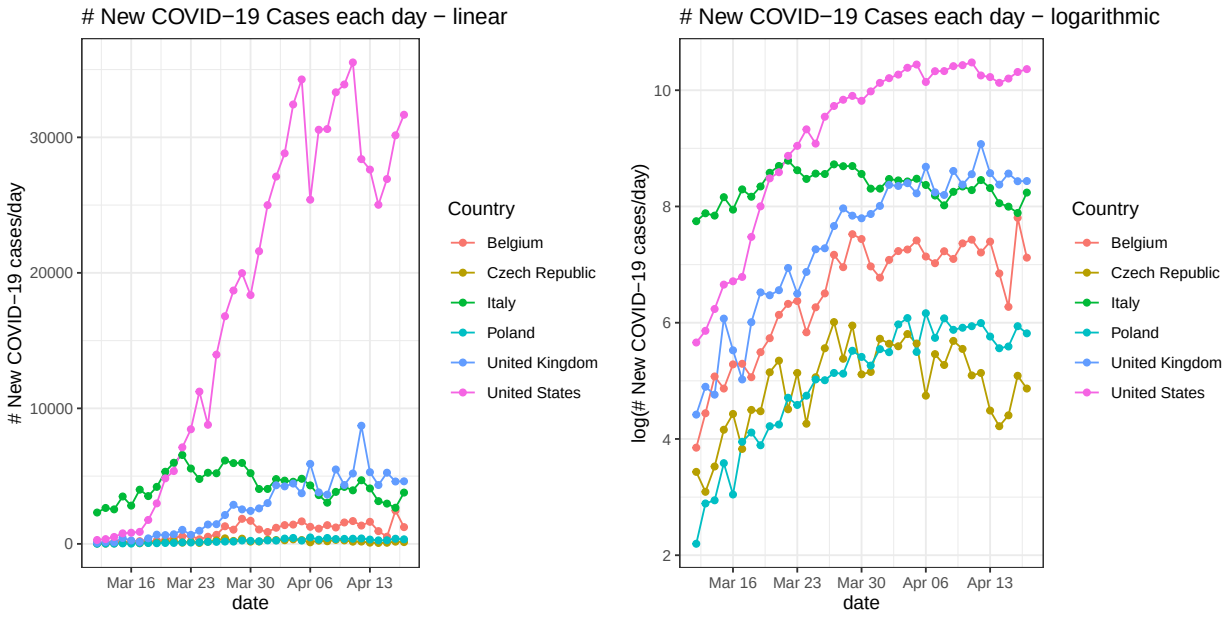

Figure 2: Example plots of the number of new cases in 6 countries. This was obtained from the *Our World in Data* website. Note the quadratic behavior on the logarithmic scale.

3 The Model

We now describe a GLMM that we can use to predict the number of new case counts per day and provide justification for our choice. The final GLMM we chose is shown in Equation (1). The response, Y_{ij} (new case counts for country i at time j), is an integer so we use a negative binomial(r, μ) model, but we later consider a Poisson. We use the negative binomial parameterization such that $E(Y_{ij}) = \mu_{ij}$ and $V(Y_{ij}) = \mu_{ij} + \mu_{ij}^2/r$. The negative binomial model is similar to the Poisson model in that it can model counts, but it offers greater flexibility in that the mean does not have to equal the variance. Using this parameterization, r acts somewhat like an overdispersion parameter under Poisson regression. The closer r gets to infinity, the more the model behaves like a Poisson model.

We include several demographics as fixed effects in the systematic component. The model shown in Equation (1) results from performing backwards AIC stepwise regression (more details later in Section 3.1). One assumption we make is that the demographics remain constant throughout the measured 37 days. Demographics such as average temperature and urban population percentage will likely remain very stable; however, the number of COVID-19 tests per person in the country will change. We would have included the number of tests for each day had the data been more complete; instead, we include the number of tests of a single day in the model simply as a measure of how much a country does COVID-19 testing. The asterisk (*) indicates that the covariates have been centered to improve convergence in estimating the model parameters. Thus, perc65$_i^*$ is the difference between country i relative to the mean of all perc65 values in the sample. The β parameters represent the various regression coefficients for the fixed effects.

To take into account the different population sizes, we adjust the expected case count (μ_{ij}) by dividing by the country population (pop$_i$) inside the link function. This is frequently referred to as an offset, and so we instead model the new case rate per person as a linear combination of covariates. This is in place so we can control for the different population sizes to make inferences about the fixed effects.

Since the countries are all at different stages of dealing with the virus and we want to generalize to countries outside the sample of complete data, we treat each country's behavior over time as a random effect. Recall that the case counts across time on the logarithmic scale (Figure 2) had quadratic-like behavior over time. Thus, we model μ_{ij} using a log link and add a quadratic effect for number of days past March 10. The $poly(time, 2)$ term in model (1) indicates an orthogonal 2nd order polynomial for the variable time. The random effects are denoted with the square brackets ([]), and a veritcal bar (|) indicates a random effect for each level in a factor. Thus, $[poly(\text{time}_{ij}, 2)|\ \text{country}_i]$ means that each country has a random intercept, slope, and quadratic term for time. Thus there are 3*29 = 87 different random effects in the model. We also include time as a fixed effect because it produces a slightly better fit (more on this in Section 3.1).

Thus, we model the number of new cases Y_{ij} for country i on j days after March 10 as follows:

$$\begin{aligned} Y_{ij} &\sim \text{Negative Binomial}(r, \mu_{ij}) \\ \log\left(\frac{\mu_{ij}}{pop_i}\right) &= \beta_0 + poly(\text{time}_{ij}, 2) + \beta_3(\text{lung}_i^*) + \beta_4(\text{perc65}_i^*) + \beta_5(\text{avghumid}_i^*) + \\ &\quad \beta_6(\text{testCapita}_i^*) + [poly(\text{time}_{ij}, 2) | \text{ country}_i]. \end{aligned} \tag{1}$$

3.1 Model Selection

We now proceed to justify our model choice. We justify our model through information criteria, residual plots, and statistical tests. We use AIC as the information criteria because we are interested in future predictions of case counts. To determine the type of model and how to account for time, we first use all demographics listed in Table 1 and later perform variable selection. We first consider 6 types of models that determine the model type and how to account for time. The models are listed below (where the color in parenthesis corresponds to the line color in Figure 3) with their corresponding AIC values in Table 2.

1. Poisson GLMM - full quadratic time fixed effect; only random intercepts and slopes over time for each country (gold)
2. Poisson GLMM - no time in fixed effects; random intercepts, slopes and quadratic terms over time for each country (orange)
3. Poisson GLMM - full quadratic time fixed effect; random intercepts, slopes and quadratic terms for each country (red)
4. Negative binomial GLMM - full quadratic time fixed effect; only random intercepts and slopes over time for each country (seagreen)
5. Negative binomial GLMM - no time in fixed effects; random intercepts, slopes and quadratic terms over time for each country (cyan)
6. Negative binomial GLMM - full quadratic time fixed effect; random intercepts, slopes and quadratic terms for each country (blue)

Table 2: AICs for the 6 models. The negative binomial (NB) models outperform the Poisson models.

Model	1-Pois	2-Pois	3-Pois	4-NB	5-NB	6-NB
df	13	14	16	14	15	17
AIC	81241.39	68121.99	68067.06	13114.61	13082.64	13035.52

Fitted lines of a few countries for these models is shown in Figure 3. From the results in Table 2, it appears that the negative binomial models fit much better than the Poisson models in terms of AIC. The random quadratic effect terms over time seem to make a differences, as can be seen from the higher AIC values for models 1 and 4. The AIC is slightly better when time is used as both a fixed and random effect (as opposed to just being a random effect). Since models 5 and 6 are nested, we can perform an analysis of deviance test to compare model fit. After performing the deviance test, we get a χ^2 value of 51.22 on 2 degrees of freedom and a p-value of 7.92×10^{-12}. Thus, we conclude that there is a difference in fit between models 5 and 6, and thus based on model deviances, we prefer model 6.

Additionally, we can examine the residual plots of the Poisson and negative binomial GLMMs for dispersion and see how well the models are capturing the variation in the data. We use the Pearson residuals, which for the ith observed value y_i, are calculated as

$$r_k = \frac{y_k - \hat{\mu}_k}{\sqrt{\text{var}(\hat{\mu}_k)}}, \tag{2}$$

where $\hat{\mu}_i$ is the fitted value for the kth observation. These are very similar to the standardized residuals in normal linear regression. Plots of Pearson residuals for models 3 and 6 are shown in Figure 4. In the Poisson GLMM, we have a clear case of overdispersion because the residuals funnel out as the fitted values increase. Although they are centered at 0, some of the residuals are more than 40 standard deviations above what is expected. The negative binomial model's residuals are much more stable and live within a much more reasonable range (94.5% are within [-2,2]). There are a few outliers, which is expected with this kind of data as sometimes countries may suddenly do mass testing or batches

of case reports may be reported all at once. There might actually be a slight case of underdispersion with the negative binomial model because the residuals have less variance for larger fitted values. However, it is not severe; it will just mean the results may be somewhat conservative.

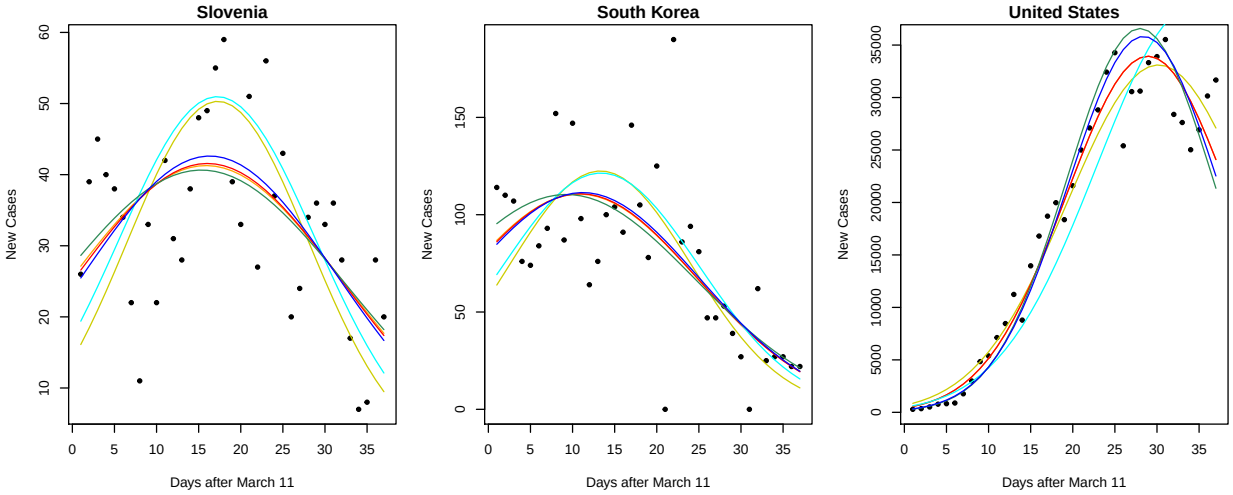

Figure 3: Fitted lines for various model choices in 3 different countries. The colors correspond to the 6 different GLMMs as listed in Section 3.1. Models 2 (orange) and 3 (red) have an almost identical fit.

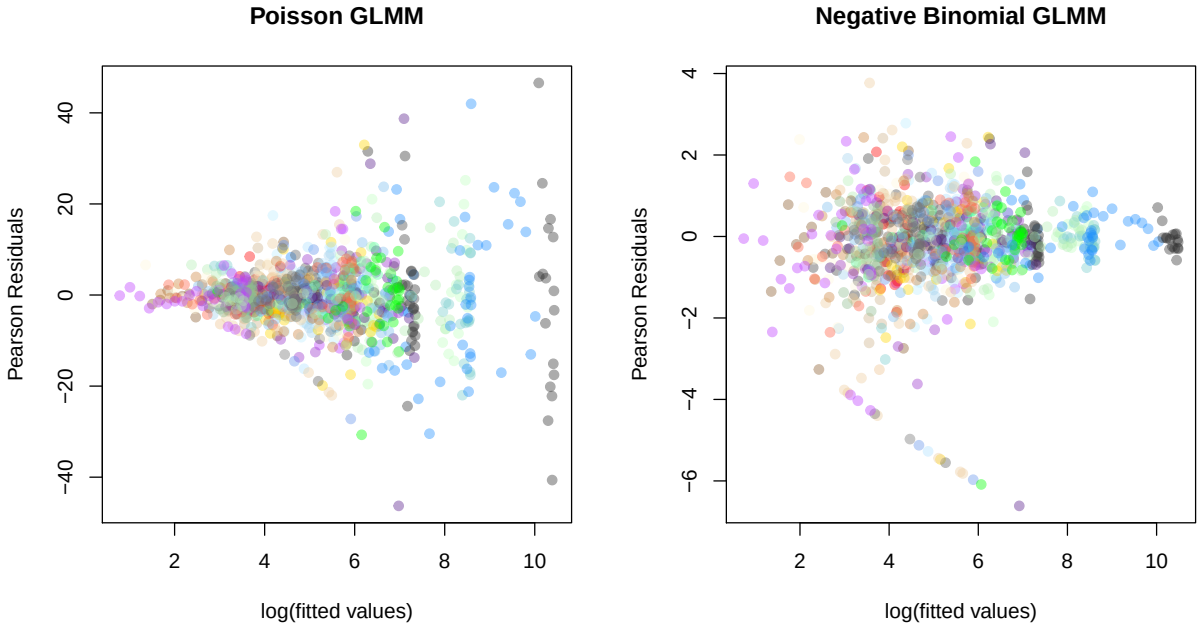

Figure 4: Residual plots for both the Poisson (model 3) and negative binomial (model 6) GLMM with quadratic fixed and random effects for time. The different colors represent the 29 different countries.

Overall, the overdispersion from the Poisson model, deviance test results, and smaller AICs from the model with time as both a fixed and random effet suggest that we go with model 6. Since many covariates were insignificant in the full model, we performed backwards AIC variable selection to determine if including all the fixed effects is necessary. Using model 6, we eliminated the variables from the model one at a time until doing nothing resulted in the lowest AIC. The variables urbanpop, smoker, and avgtemp were eliminated (in that order). The resulting final model had 4 demographic covariates plus the fixed effects for time, an AIC of 13032.4, and a deviance χ^2 p-value of .53 when compared to the full model before variable selection. This is not significant, and so we do not reject the null hypothesis

that there is no difference in fit between the two models. Thus, we use the resulting model after backwards selection, which is shown in Equation (1).

3.2 Model Fit: Psuedo-R^2 Measures for GLMMs

A frequently used goodness-of-fit metric in multiple regression is the R^2 metric. It can also be used to get an idea of the percentage of total variation explained by a model. However, in mixed models, it becomes unclear how to compute this metric because there is variation from both the random effects and likelihood. In order to get a statistic comparable to an R^2 measurement, it helps to think about R^2 as a full and reduced model comparison. Xu [5] generalizes R^2 to linear mixed models by first defining a meaningful "null" model and then comparing the residual variance to a full model.

In a repeated measurements mixed model, there are repeated measurements on several individual subjects. Differences between subjects are a given and sometimes it is more meaningful to get a sense of global fit from the covariates within each subject. However, if the subject variable itself is of interest for the analysis, then it might be meaningful to get a sense of overall predictive power including the subject variable. Thus, Xu suggests computing two different measurements analogous to R^2 where the full model is compared to 2 different "null" models (or baseline). We use 2 different baseline models: one model with intercepts for both fixed and random effects and another model with just a global intercept. Xu denotes the pseudo-R^2 measurement for the fixed/random intercept model as the baseline to be Ω^2 and the psuedo-R^2 for the model with only a global intercept to be Ω_0^2.

The equations to calculate Ω^2 and Ω_0^2 are shown in (3) and (4). Denote $\boldsymbol{\beta}$ as the vector of fixed effects and $\boldsymbol{z}$ as a vector of random effects. Let β_0 be the intercept in the fixed effects and $\boldsymbol{z_0}$ be a vector of only random intercepts. We can think of the model residuals as the "leftover" variance of the response after taking into account the covariates. Thus, $\text{Var}(Y|\boldsymbol{\beta}, \boldsymbol{z})$ can be estimated by taking the variance of the residuals after fitting the the model with covariates $\boldsymbol{\beta}$ and $\boldsymbol{z}$. We extend this idea to GLMMs by using the Pearson residuals. Since the variance is not constant in a negative binomial GLMM, we use the Pearson residuals, which have been standardized. Thus, each residual will receive equal weight so it can be compared. Let $\hat{V}(\epsilon)$ be the variance of the full model Pearson residuals, $\hat{V}(\epsilon_0)$ be the variance of the fixed/random intercept only model Pearson residuals, and $\hat{V}(\epsilon_{00}) = \hat{\sigma}_Y^2$ be the overall variance of the response.

$$\Omega^2 = 1 - \frac{\text{Var}(Y|\boldsymbol{\beta}, \boldsymbol{z})}{\text{Var}(Y|\beta_0, \boldsymbol{z_0})} \approx 1 - \frac{\hat{V}(\epsilon)}{\hat{V}(\epsilon_0)} \tag{3}$$

$$\Omega_0^2 = 1 - \frac{\text{Var}(Y|\boldsymbol{\beta}, \boldsymbol{z})}{\text{Var}(Y)} \approx 1 - \frac{\hat{V}(\epsilon)}{\hat{V}(\epsilon_{00})} \tag{4}$$

Table 3: Psuedo R^2 measurements using Equation (1) as the full model.

metric	Psuedo-R^2
Ω^2	0.6572
Ω_0^2	0.9796

The psuedo-R^2 measurements for our selected model are shown in Table 3. Offsets were excluded from the null models. The interpretation is that our model explains about 65.72% of the total variation in case counts, given we know the country, and 97.96% of the total variation in entire dataset. Overall, our model fits quite well, as can be seen from the lines in Figure 3.

4 Results

We now address the research questions mentioned in the introduction by interpreting the model coefficients, predicting when the number of cases will peak, and predicting how many new cases will occur through the end of April 2020. The model coefficients are shown in Table 4.

Using the `glm.nb` function in the `lmer` library in `R`, we estimated r to be 3.7875. The standard deviation of the random effects of each country over time were 0.926 for the intercepts, 18.608 for the slopes, and 6.419 for the quadratic terms. The size of these effects suggest that there were sizable differences between the countries in terms of how the number of new cases developed over time.

Table 4: Fixed effect parameter estimates, 95% CIs, and exponentiated values for interpretation.

	Est	2.5 %	97.5 %	exp(Est)	exp(2.5 %)	exp(97.5 %)
(Intercept) (+)	-11.5558	-11.8908	-11.2208	9.58×10^{-6}	6.85×10^{-6}	1.33×10^{-5}
poly(time, 2)1 (+)	19.0562	14.0029	24.1094	–	–	–
poly(time, 2)2 (-)	-12.5413	-14.9604	-10.1221	–	–	–
lung*	-0.0196	-0.0441	0.0049	0.9806	0.9568	1.0049
perc65* (+)	0.0903	0.0202	0.1604	1.0945	1.0204	1.1740
avghumid* (+)	0.0507	0.0196	0.0817	1.0520	1.0198	1.0852
testCapita* (+)	0.0288	0.0089	0.0488	1.0293	1.0089	1.0500

*Indicates variable was centered

The plus and minus signs next to the covariate indicate the direction of significance on case counts. All demographics eliminated by variable selection are assumed to have negligible effect in the model. Since we centered many covariates and included an offset, β_0 represents the number of new cases per person when the country demographics are at their global mean on March 11, 2020. Thus, the expected number of new cases of COVID-19 per person on March 11 for a country with demographics at the global mean is $\exp(-11.5556)$ or $\exp(-11.556) \times 10^6 = 9.58$ cases per million people. The time effects are difficult to interpret because they come from orthogonal polynomials, but they represent the overall, "center" effect across time between all countries. However, both terms are significant, so they are important. The quadratic term for time, poly(time,2)2, is negative so the number of new cases over time tends to follow a concave-like pattern between countries.

The interpretation of the other β coefficients is the same regardless of centering or offset (since we used a log link). For every additional death per 100k people, we expect the average number of cases per person (or just simply the number of cases) to be multiplied by exp(-0.0196)=0.9806 with 95% CI [.958, 1.0049], holding all other variables constant. A similar interpretation holds for testCapita. For every percentage increase in humidity, we expect the average number of new cases in a country to be multiplied by 1.0520 with 95% CI of [1.019, 1.085]. A similar interpretation holds for perc65. The percentage of persons over 65, the average humidity, and the number of tests done per 1K people in a country all had a statistically significant positive association on the number of new case counts.

In general, humidity has a negative impact on COVID-19 spread [6], although the pairs plot in Figure 1 indicate a marginal positive relationship between the number of new cases per person and humidity. Thus, there may be some confounding (e.g., most countries in the sample with high humidity are European, who happen to have a high number of cases) or the yearly estimate of humidity does not reflect March conditions.

4.1 Predicting Future Case Counts

For the remainder of the analysis, we exclude Japan and Peru as outliers because case count predictions outside the bounds of the data resulted in a forever increasing curve. These countries have too much noise in the recent days and/or they are not far enough in the virus outbreak for the model to get a good sense of when the number of new cases will curve back down. We proceed to answer the other 2 research questions. Table 5 contains the latest metrics (as of April 17th 2020) of the case counts, along with predicted peaks, and predicted case count by the end of April 2020. The first 3 columns contain data on the number of cumulative cases, cases per million people in the population, and number of new cases on April 17. The next two columns contain the estimated highest number of cases in a day and the corresponding peak day from the model. Finally, the last 4 columns indicate the number of additional cases that are predicted to happen from April 17 through April 30, how many new cases are predicted to happen on April 30, the total number of cases by the end of April, and the total number of cases per million people.

The predictions for each country can also be seen visually in Figure 5. The predicted counts correspond to the red line on each country graph, with the black dots being the observed case count values. Displayed below each country name is the Mean Absolute Percent Error (MAPE) of the predicted line on the observed values, the estimated peak date, and the total number of cumulative cases by the end April 30. MAPE is calculated by taking the mean of $100 * |$(Observed-Expected)/Expected$|$ for all observed values. It's a good measure of error for time series data that is multiplicative (variance increases with the fitted value). As can be seen from the plots, some countries fit better than others. Countries like the US seem to fit really well, while countries like Denmark have unusual trends and do not fit the bell-shaped curve well. For the most part, even with some outliers, most countries fit the predicted curve well. As of April 17, almost all of the predicted curves are going down, which means the virus's spread is starting to slow down.

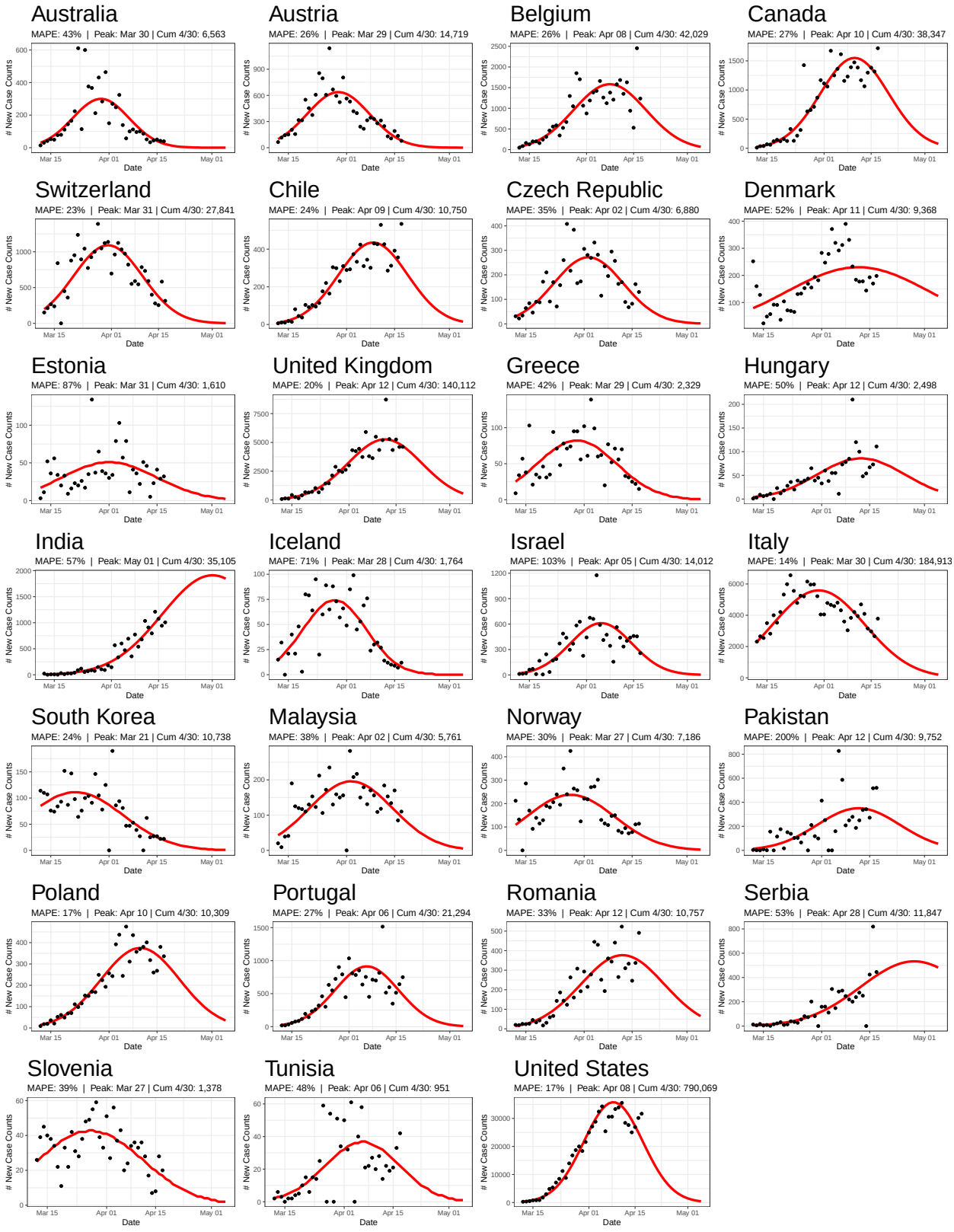

Figure 5: Model fits to all countries with mean absolute percent error, the peak outbreak, and the predicted total number of COVID-19 cases by the end of April 30 2020.

Table 5: Current case counts, peak values, and predicted case counts by the end of April 2020.

Country	4-17 Cases	Cases/Mil	Case/day	Peak	Peak Day	+ AprilCases	4-30 Case/day	Total Cases	Cases/Mil
Australia	6497	254.8	39	300	Mar 30	66	0	6563	257.4
Austria	14448	1604.2	78	637	Mar 29	271	2	14719	1634.3
Belgium	34809	3003.5	1236	1583	Apr 08	7220	194	42029	3626.4
Canada	30081	797.0	1717	1551	Apr 10	8266	225	38347	1016.0
Chile	8807	460.7	534	435	Apr 09	1943	15	10750	562.4
Czech Republic	6433	600.7	130	272	Apr 02	447	47	6880	642.5
Denmark	6879	1187.6	198	230	Apr 11	2489	7	9368	1617.3
Estonia	1434	1081.0	32	51	Mar 31	176	157	1610	1213.7
Greece	2207	211.7	15	82	Mar 29	122	6	2329	223.4
Hungary	1763	182.5	111	86	Apr 12	735	1315	2498	258.6
Iceland	1739	5096.1	12	74	Mar 28	25	2	1764	5169.3
India	13387	9.7	1007	1911	May 01	21718	33	35105	25.4
Israel	12758	1474.0	257	611	Apr 05	1254	1906	14012	1618.8
Italy	168941	2794.2	3786	5588	Mar 30	15972	0	184913	3058.3
Malaysia	5182	160.1	110	196	Apr 02	579	478	5761	178.0
Norway	6791	1252.7	114	237	Mar 27	395	3581	7186	1325.5
Pakistan	7025	31.8	520	351	Apr 12	2727	2	9752	44.1
Poland	7918	209.2	336	375	Apr 10	2391	9	10309	272.4
Portugal	18841	1847.8	750	913	Apr 06	2453	112	21294	2088.3
Romania	7707	400.6	491	377	Apr 12	3050	6565	10757	559.2
Serbia	5318	608.6	445	533	Apr 28	6529	83	11847	1355.9
Slovenia	1268	609.9	20	43	Mar 27	110	40	1378	662.8
South Korea	10635	207.4	22	111	Mar 21	103	129	10738	209.4
Switzerland	26651	3079.4	315	1092	Mar 31	1190	529	27841	3216.9
Tunisia	822	69.6	42	37	Apr 06	129	4	951	80.5
United Kingdom	103093	1518.6	4617	5263	Apr 12	37019	3	140112	2063.9
United States	671331	2028.2	31667	35740	Apr 08	118738	1948	790069	2386.9

5 Conclusion

We built a negative binomial GLMM that allowed us to both make inferences on country demographics and make future predictions for the rest of the month of April. We considered 6 different models and found that the negative binomial GLMM with time as both a fixed and random effect fit the best in terms of AIC, deviance tests, and residual plots. We found that the model does a great job at fitting the data; given the country, it had a pseudo-R^2 value of 65.72%. We found that of the 7 original demographic variables, the percentage of people over 65, the average yearly humidity in the country, and the number of tests done per 100K people all had significant, positive associations with case counts. When making future predictions, we threw out Japan and Peru because the predictions did not extrapolate well. We created a table (Table 5) and a collection of graphs (Figure 5) for each country that summarizes the future predictions.

Due to computational and technical challenges of bootstrapping a sample of repeated measures data, we were not able to generate prediction bands for the number of new case counts per day for the GLMM selected. This also meant we were not able to generate confidence intervals for the cumulative number of cases or the predicted peak dates. A point estimate sometimes does not tell the full story because there could be a lot of variability in the predictions. Uncertainty intervals could give health and government officials a range of reasonable values to work with. Uncertainty intervals could easily be calculated under a Bayesian framework, but in this example, we had 87 random effects alone. A Bayesian model would get computational very quickly.

Another shortcoming is that our analysis is only as good as the data it came from. For simplicity, we omitted several countries like Spain and Germany because of missing data related to COVID-19 tests or demographics. We might be missing information that could impact the model coefficients that we used to make inferences on what demographics contribute to higher case counts. This data also only contains *reported* cases and tests; so we are very likely underrepresenting the true amount of COVID-19 cases and tests. Further experimental and statistical work could be done to try and estimate the amount of true cases in a population. Moving forward, analysis done later in time with data that is more complete will help us more fully understand the virus.

The good news is that according to our analysis, many countries are already past their peak of new COVID-19 cases per day. This shows that the physical distancing and quarantine measures are working. As long as such measures are kept as long as they are needed and we avoid a second outbreak, the world is currently on track to exiting the COVID-19 pandemic and resuming normal life.

References

[1] Sauer, Lauren (2020). What is Coronavirus? from *John Hopkins Medicine* www.hopkinsmedicine.org

[2] My Koryto (2020) *Country Info* dataset on *Kaggle*. `https://www.kaggle.com/koryto/countryinfo`

[3] World Bank (2018) Population ages 65 and above (% of total population) on `data.worldbank.org`

[4] Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell (2020) Coronavirus Disease (COVID-19) – Statistics and Research published on *Our World in Data* `https://ourworldindata.org/coronavirus`

[5] Xu, R. (2003). Measuring explained variation in linear mixed effects models. pp.3-4. Published in *Statistics in Medicine*

[6] G. Kamf (2020). Persistence of Coronaviruses on Inanimate Surfaces and their Inactivation with Biocidal Agents. Published in *Journal of Hospital Infection*