A. C. Davison **Modern Regression Methods Project 2020** MATH EPFL

Problem

Data on the progress of the COVID-19 epidemic are widely available. Cantonal data for Switzerland (and some nice graphics) can be found here. Global data can be downloaded from

- the World Health Organisation,
- the European Centre for Disease Control (ECDC),
- or Johns Hopkins University.

A comparison of these sources may be found here. The ECDC data are updated each afternoon and the site helpfully provides code to get the latest version into `R`. The data for 20 April 2020 are provided on the Moodle page for the course, as is code to download them and wrangle them into a matrix whose rows correspond to countries and whose columns to days since (and including) 31 December 2019, plus estimated populations of countries in 2020. Read the file `basic_analysis.R` carefully, and use it to get the data (preferably the most recent data) into `R`. You may need to edit the `R` file, since the ECDC occasionally re-order the columns of their files.

Problems with the data are well-known: there are reporting delays (very visible in the Swiss data, which have successive days with numbers below and then above the general pattern); testing regimes vary from country to country (and maybe over time), so the numbers of confirmed cases for different countries are not comparable; many cases (maybe 80%, according to some analyses) may show no symptoms; some countries are said to be massaging the figures; there may be mis-attribution of coronavirus deaths to other causes, or vice versa; and in some countries coronavirus deaths in nursing homes are not included in overall totals. The function `round.moving.average` is used by default to produce a lightly smoothed version of the data which can then be plotted using statements such as

```
par(mfrow=c(2,3),pty="s")
plot.Country("Switzerland",plot=T, plot.cumul=T, xmin=50)
plot.Country("Italy",plot=T, plot.cumul=T, xmin=50)
```

The purpose of the project is to use the tools and methods learned during the course to study some aspect of the coronavirus epidemic. There are various different possibilities:

- studying the initial (so-called exponential) growth phase for deaths and comparing it for different countries;
- studying how cases are linked to deaths, for a single country or a subset of countries; or
- an issue of your choice.

Initial phase

In the early phase of an epidemic the growth in cases can usually be well-modelled by an exponential curve, so one might expect that the mean on day t after the start is of the form $\mathrm{E}(Y_t) = m\exp(\alpha+\beta t)$, where m is the population of the country. One could assume that Y_t follows a Poisson distribution, or allow for overdispersion using quasi-likelihood or a negative binomial model, and attempt to estimate the parameters, including the doubling time s, which satisfies $\mathrm{E}(Y_{t+s})/\mathrm{E}(Y_t) = 2$. This approach could also be applied to deaths, but won't work after interventions have started to take effect. It could be interesting to compare α, β (or equivalently s) for different countries, providing suitable measures of uncertainty, and/or to use a random effects model in which one supposes that if α_c and β_c are the parameters for country c, then $\alpha_c \overset{\text{iid}}{\sim} \mathcal{N}(\alpha, \sigma^2_\alpha)$ and $\beta_c \overset{\text{iid}}{\sim} \mathcal{N}(\beta, \sigma^2_\beta)$, and attempt to estimate the overall means α and β and the variances. Then identifying outliers, i.e., countries with unusual values of the parameters relative to the others, could be of interest.

One could complicate such a model by adding smooth terms, e.g., $\log \mathrm{E}(Y_t) = \log m + \alpha + \beta t + g(t)$, where $g(t)$ is a spline, or a low-order polynomial, to allow better estimation of β by allowing the use of data further into the epidemic; presumably g would have to vary from country to country.

Cases and deaths

Some fraction of cases leads to subsequent deaths, so one might relate the mean number of deaths on day t in terms of the numbers of cases declared on previous days as

$$\mu_t^{\text{deaths}} = \sum_{l=0}^{l_{\max}} \lambda_l \text{Cases}_{t-l}, \quad t = n, n-1, \ldots$$

where $\lambda_l \geq 0$, and we would expect that $l_{\max} < 30$. This leaves a large number of parameters to be estimated, so a simpler formulation would be to smooth the numbers of cases to obtain its mean μ_{t-l}^{cases} and then to use the simplistic formulation

$$\log \mu_t^{\text{deaths}} = \eta_0 + \eta_1 \log \mu_{t-l}^{\text{cases}},$$

where l is chosen to get the best-fitting model for the number of deaths in terms of the past cases. We would hope that $\eta_1 \approx 1$, in which case $\eta_0 = \log \lambda_l$, where l would represent the typical lag between the cases and the deaths; λ_l would vary between countries but could be expected to be around 0.01. Such a model could then be used to predict μ_t^{deaths} for l steps into the future. Uncertainty analysis for this would require one to include the variation of the 'covariate', i.e., the smoothed number of cases, as well as the prediction uncertainty for the new deaths.

Issue of your choice

You could also choose another approach with the aims of the analysis being to gain insight into the time-space dynamics of the disease and to make short-term predictions useful for the implementation of public health measures. For example

- Study the distribution of the final proportion infected in susceptible-infected-recovered (SIR) epidemics in asymptotically large closed populations.
- Study the space-time dynamics using a Bayesian spatial hierarchical model (using Markov chain Monte Carlo method).
- Study different temporal trends in rate of the Poisson distribution using parametric or semi-parametric models.

In your analysis you could focus on one country of your choice or use data for selected countries and use the statistical tools from the course to discuss why death rates can look so starkly different from place to place.

If you want to do this, it would be wise to discuss your idea with us beforehand.

Report

Submission and deadlines

A report of at most **12 pages** (excluding cover page, table of content and bibliography) is to be **submitted on Moodle** as a PDF file (nothing else will be accepted) by midnight on Friday 19 June 2020. Later submission will not be accepted. Reports must be written in **pairs**; you can discuss with your classmates, but the code and writing should be your own. You should upload your code in a separate file and ensure the output is reproducible (not proper to your laptop or computer).

Discussion points

Exploratory data analysis : nature of covariates and distribution of response, presence of outliers or missing values, range of variables.

Modelling : key variables, with fitting using GLMs and/or GAMs, Test statistics and goodness-of-fit diagnostics. Interpretation of the final model on a meaningful scale.

Discussion : discuss the results, showing (e.g.,) histograms of fitted values, comparisons of fitted and predicted probabilities, analysis of deviance and/or AIC for model comparisons, give estimates and their standard errors, explain any disagreement between the models. Ensure that you carefully interpret the fitted models in terms of the original problem.

Structure of the report

The report should be typed in English. Some notes on report-writing can be found here and there is an example report posted on Moodle.

Introduction : Briefly state the purpose of the analysis, discuss the main features of the data (e.g., via exploratory data analysis), and outline what will follow.

Analysis : describe the model(s) fitted, using your own words. Give the key elements only: you can refer to the lecture notes and to books, but should give careful references (to pages and equations etc.). It is not enough simply to give a list of sources at the end of the work: references should be mentioned in the text, and only those mentioned in the text should be listed at the end. Use BibTeX or similar to ensure that the references appear properly; check a book or journal article to see what details should appear in the bibliography.

Discussion of the results in more detail. Include crucial graphs and tables only, make sure that their contents are understandable without reference to the text, and that their axis labels and captions are clear and informative; each graph and/or table should tell the reader a coherent story. Give appropriate numbers of digits for tables. The text should give detailed interpretations of the plots and tables, with more details, if they are needed, and should show where the graph/table fits into the overall picture.

Conclusions : the take-away message from your analysis. Convince the reader that you know what you did and are aware of its strengths and limitations. Sketch what more you might do, if you had more time.

Suggestions and caveats

1. We recommend that you use LaTeX.
2. Your report should be sufficiently detailed that a reader can reproduce your results after reading it.

3. Figures and tables should be numbered and have captions briefly explaining their contents. Reference should be made to each figure/table from within the text.
4. Read your report carefully before handing it in and use a spell checker to find any spelling mistakes.
5. Mention any references you have used and provide a detailed bibliography. References should be made to scientific articles or books. Detailed (chapter, section, page, equation) references to books are usually needed, so that the reader does not have to figure out which page(s) of a large book you are referring to.
6. Pasting plain computer output is not acceptable.
7. Due to space limitations, you should provide only relevant output. Your code can however contain exploratory data analysis and other model fits and diagnostics that are not reported in the text.
8. Your code file should be commented (but self-explanatory commands need not be commented on).
9. When writing a report, you should not answer questions directly. Instead, make sure your report covers the material discussed in each point, but structure your report as a scientific paper.
10. Common problems: many students don't give enough (or sometimes any!) interpretation of fitted models; often tables have too many digits; often figures are too small to be read properly or don't use the page layout well; often captions to tables or figures are uninformative; often the discussion section is insufficiently detailed; often the bibliography has missing details; often references to publications are inadequate; often equations are not (or are incorrectly) punctuated; often the English has persistent spelling errors.

Marking scheme

Correctness	Accurate, appropriate use of statistical tools	Incorrect, many errors	
	10 9 8 7 6 5	4 3 2 1 0	Score ______
Discussion	Thoughtful, detailed, apposite	Banal, obvious, thin	
	10 9 8 7 6 5	4 3 2 1 0	Score ______
Graphics and tables	Clearly labelled, well-chosen, good captions and discussion, appropriate numbers of digits	Poorly labelled, no discussion, unmotivated, unedited output	
	10 9 8 7 6 5	4 3 2 1 0	Score ______
Originality and scope	Wide range of tools/ideas	Limited range of tools/ideas	
	5 4 3	2 1 0	Score ______
Quality of writing	Good grammar and punctuation, including mathematics	Poor grammar and punctuation	
	5 4 3	2 1 0	Score ______
Referencing	Full, accurate, and detailed references given	Inadequate citation of sources	
	5 4 3	2 1 0	Score ______
		Grand total (max 45)	______