Can the reported COVID-19 data tell us the truth?

Scrutinizing the data from the measurement error models perspective

Grace Y. Yi

Canada Research Chair in Data Science (Tier 1)
Department of Statistical and Actuarial Sciences
Department of Computer Science
University of Western Ontario

Joint work with Wenqing He, Qihuang Zhang, Yayuan Zhu, and GW-DSRG
Outline

- The website of COVID-19 Canada developed by the GW-DSRG (Grace-Wenqing Data Science Research Group)

- Different estimates from different studies
  - A meta-analysis

- Time series with measurement error: sensitivity analysis
  - An additive measurement error model
  - A multiplicative measurement error model
Part 1: The Website of COVID-19 Canada
developed by the GW-DSRG (Grace-Wenqing Data Science Research Group)

https://covid-19-canada.uwo.ca
Welcome!

Welcome to the web-page created by the GW/DSRG (Grace-Wening Data Science Research Group) at the Faculty of Science of the University of Western Ontario.

With the focus on Canadian COVID-19 data, this web-based interactive platform aims to report real-time development of COVID-19, visualize patterns, and provide state-of-the-art statistical insight from up-to-date accessible COVID-19 data. This website will be updated on a daily basis.
Heat map of cumulative confirmed cases by PHU in Canada
COVID-19 Canada Website: Died, Recovered and Infected

Death rate and recovery rate (%) in Canada

Infected cases (per 1,000 people) by province

Province

- Quebec
- Ontario
- Alberta
- Nova Scotia
- Saskatchewan
- Newfoundland and Labrador
- British Columbia
- Yukon
- Manitoba
- Prince Edward Island
- New Brunswick
- Northwest Territories

Rate (%)

Days

Feb 2020 Mar 2020 Apr 2020 May 2020
Proportion of confirmed cases by city in Ontario

Hospitalization, ICU, death rates (%) over time in Ontario

The hospitalized rate, ICU rate and death rate are computed by the number of people who are hospitalized, placed in ICU, and dead respectively divided by the number of confirmed cases in Ontario. Data comes from Government of Ontario.
Proportion of recovery by age and gender in Ontario

Recovery rate by age and gender in Ontario

Recovery rate for each age and gender group is computed by the number of recovered cases for particular age and gender group divided by number of confirmed cases in this age and gender group.
Objective

We implement the SIR (susceptible-infected-resolved) model to predict cumulative numbers of infected cases with COVID-19 in Ontario, British Columbia, Quebec and Alberta. We fit the model using the data from March 18 to May 11, 2020 (https://coronavirus.1point3acres.com/) to predict the cases from May 12 to May 18, 2020.

Assumption and Model

The population in each province is divided as three subpopulations defined as susceptible (S), infected (I), and resolved (R), respectively. The status for an individual in the population may change with time: a healthy individual may become infected, and an infected patient may recover or die of the disease. Figure 1 shows the process of status change, where $\beta$ represents the average number of contacts per person per time and $\gamma$ stands for the transition rate from I to R. We estimate both $\beta$ and $\gamma$ using the data in each province for the period of March 18 to May 11, 2020, during which all four provinces were in the "state of emergency". We assume that there are no inbound or outbound infected travellers during this period.
Findings and Discussion

The following figures present the predicted cumulative number of cases (in red) together with the reported cumulative number of confirmed cases (in blue) for the four provinces. A red solid curve reports the fitted number for the period of March 18 to May 11, 2020, and its differences from the blue curve show the performance of using the SIR model to do prediction. A red dashed curve is the predicted cumulative number of cases for the period of May 12 to May 18, 2020.

For ease of visualization, here we use connecting curves instead of isolated points to display the reported or predicted cumulative numbers of cases for the four provinces. For the next few days, British Columbia shows a somewhat possibly downward trend, whereas the other three provinces exhibit an upward trend.

The prediction here focuses on showing the trend in the near future using the SIR model. The validity of the results relies on the associated assumptions of the SIR model which may be violated. For instance, unreported cases as well as the inbound and outbound travelers would make the model fitting flawed. Due to the limited testing capacity, the daily reported cumulative number of confirmed cases can differ from the predicted number with a notable discrepancy.

**ONTARIO**

The comparison of the predicted cumulative number of infected cases using the SIR model (in red) versus the reported cumulative infections (in blue) in Ontario. The red dashed curve represents the prediction for the next 7 days.
Part 2: Different Estimates from Different Studies

- Meta-Analysis

W He, G Y. Yi, and Y Zhu (2020). Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for COVID-19: Meta-analysis and sensitivity analysis.

To appear in *Journal of Medical Virology*. 
# Estimate of Average Incubation Time (in Day)

<table>
<thead>
<tr>
<th>Date of Study</th>
<th>Date Range</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tian et al.</td>
<td>12/31/2019 - 02/19/2020</td>
<td>Chinawide</td>
</tr>
<tr>
<td>Li et al.</td>
<td>up to 01/22/2020</td>
<td>425 cases in Wuhan</td>
</tr>
<tr>
<td>Backer et al.</td>
<td>01/20/2020 - 01/28/2020</td>
<td>88 Wuhan travellers</td>
</tr>
<tr>
<td>Jiang et al.</td>
<td>up to 02/08/2020</td>
<td>50 confirmed cases in Wuhan</td>
</tr>
<tr>
<td>Lauer et al.</td>
<td>up to 02/24/2020</td>
<td>mainly about China</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Effect Size</th>
<th>95% CI</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tian et al.</td>
<td>4.90</td>
<td>(4.32, 5.47)</td>
<td>0.30</td>
</tr>
<tr>
<td>Li et al.</td>
<td>5.20</td>
<td>(4.10, 7.10)</td>
<td>0.51</td>
</tr>
<tr>
<td>Backer et al.</td>
<td>6.40</td>
<td>(5.60, 7.70)</td>
<td>0.60</td>
</tr>
<tr>
<td>Jiang et al.</td>
<td>4.90</td>
<td>(4.40, 5.50)</td>
<td>0.28</td>
</tr>
<tr>
<td>Lauer et al.</td>
<td>5.10</td>
<td>(4.50, 5.80)</td>
<td>0.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary (Fixed)</th>
<th>Effect Size</th>
<th>95% CI</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.08</td>
<td>(4.77, 5.39)</td>
<td>0.18</td>
</tr>
</tbody>
</table>
## Estimate of the Fatality Rate (in Percent)

<table>
<thead>
<tr>
<th>Study</th>
<th>Date</th>
<th>Cohort</th>
<th>Effect Size</th>
<th>95% CI</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baud et al.</td>
<td>up to 03/01/2020</td>
<td>worldwide</td>
<td>5.60</td>
<td>(5.40,5.80)</td>
<td>0.10</td>
</tr>
<tr>
<td>Ruan</td>
<td>up to 03/21/2020</td>
<td>Chinawide</td>
<td>1.38</td>
<td>(1.23,1.53)</td>
<td>0.08</td>
</tr>
<tr>
<td>Verity et al.</td>
<td>01/04/2020 - 02/24/2020</td>
<td>outside Hubei province, China</td>
<td>3.67</td>
<td>(3.56,3.80)</td>
<td>0.06</td>
</tr>
<tr>
<td>Wu et al.</td>
<td>up to 02/29/2020</td>
<td>Wuhan, China</td>
<td>1.40</td>
<td>(0.90,2.10)</td>
<td>0.28</td>
</tr>
<tr>
<td>Sun et al.</td>
<td>a meta analysis with ten studies</td>
<td>50466 cases in China</td>
<td>4.30</td>
<td>(2.70,6.10)</td>
<td>0.87</td>
</tr>
<tr>
<td>Li et al.</td>
<td>12/2019 - 02/2020</td>
<td>China</td>
<td>5.00</td>
<td>(1.00,11.00)</td>
<td>2.55</td>
</tr>
<tr>
<td>Wang et al.</td>
<td>up to 02/27/2020</td>
<td>worldwide</td>
<td>3.46</td>
<td>(3.33,3.58)</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Summary (Random)</strong></td>
<td></td>
<td></td>
<td><strong>3.34</strong></td>
<td><strong>(2.18,4.49)</strong></td>
<td><strong>0.59</strong></td>
</tr>
</tbody>
</table>
# Estimate of the Asymptomatic Infection Rate (in Percent)

<table>
<thead>
<tr>
<th>Study</th>
<th>Date</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nishiura et al.</td>
<td>up to 02/06/2020</td>
<td>565 Japanese nationals evacuated from Wuhan</td>
</tr>
<tr>
<td>Kimball et al.</td>
<td>03/13-20/2020</td>
<td>13 long-term care residents in King County, Washington</td>
</tr>
<tr>
<td>Song et al.</td>
<td>up to 03/06/2020</td>
<td>a single-centre study in Daofu county, Sichuan</td>
</tr>
<tr>
<td>Mizumoto et al.</td>
<td>up to 02/21/2020</td>
<td>3,711 people on the Diamond Princess cruise ship</td>
</tr>
<tr>
<td>Serra</td>
<td>04/02/2020</td>
<td>Northern Italy, 60 volunteer blood donors</td>
</tr>
<tr>
<td>Day</td>
<td>04/01/2020</td>
<td>166 new infections</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Effect Size</th>
<th>95% CI</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nishiura et al.</td>
<td>31.00</td>
<td>(7.70, 54.0)</td>
<td>23.30</td>
</tr>
<tr>
<td>Kimball et al.</td>
<td>56.50</td>
<td>(36.30, 76.80)</td>
<td>10.30</td>
</tr>
<tr>
<td>Song et al.</td>
<td>21.70</td>
<td>(12.80, 30.60)</td>
<td>4.50</td>
</tr>
<tr>
<td>Mizunoto et al.</td>
<td>17.90</td>
<td>(15.50, 20.20)</td>
<td>2.40</td>
</tr>
<tr>
<td>Serra</td>
<td>66.70</td>
<td>(54.70, 78.60)</td>
<td>6.10</td>
</tr>
<tr>
<td>Day</td>
<td>78.30</td>
<td>(72.00, 84.60)</td>
<td>3.20</td>
</tr>
<tr>
<td>Summary (Random)</td>
<td>46.00</td>
<td>(18.40, 73.60)</td>
<td>14.10</td>
</tr>
</tbody>
</table>
Part 3: Error-Prone COVID-19 Data
Some Error Sources of COVID-19 Data
Error-Contaminated COVID-19 Data

Confirmed cases may be under-reported:

- Asymptomatic patients
  - may not be detected
  - may be delayed to be found
  
  e.g.,
  Miszunoto et al. (2020): the asymptomatic infection rate \( \approx 17.90\% \)
  Day (2020): the asymptomatic infection rate \( \approx 78.30\% \)

- Limited test capacity does not allow infected patients with light symptoms to be tested.
  
  e.g., As of May 25, 2020
Error-Contaminated COVID-19 Data

Confirmed cases or discharged recoveries may be error-prone due to testing error:

- The accuracy of the current COVID-19 tests is not precisely known.
  
  Example:
  Based on the test performance in China and the performance of the influenza tests, Hutchison (2020) suggested that the sensitivity and specificity of COVID-19 tests were estimated to be 60% and 90%, respectively.

Example:
In China’s Guangdong province, 14% of people who recovered in the province were later retested to be positive. Similar cases have been reported in other countries (Guzman 2020).
Error-Contaminated COVID-19 Data

The incubation time varies from patient to patient.

Source: Wu and McGoogan (2020)
Heterogeneity:
Since the outbreak of the disease, a large body of research on COVID-19 has been done and many articles have been published in scientific journals or shared on platforms such as bioRxiir and medRxir.

Different studies have been carried out
- on different patients
- under different conditions,
- and different authors may make different model assumptions.

Simulations of the epidemic have been published under various assumptions to delineate hidden transmissions of the virus.

Interpretation of the available findings must be coupled with the associated features of the studies.
General Framework of Data with Measurement Error
Some Sources of Measurement Error (Yi 2017)

Measurement error may refer to random noise, sampling error, or uncertainty/variation in the measuring process.

Measurement Error = reading error + biological variability + sampling error + reporting error + detecting error + others

Flawed or mismanaged data collection procedures result in imprecise measurements.

Variables are not accurately measured due to reporting errors for sensitive questions.

Variables are not accurately measured due to recall bias.

Variables are impossible to measure precisely.

Variables represent averages of certain quantities over time.

Variables may be manipulated artificially.

Variables are too expensive or time consuming to measure precisely.
Impact of Measurement Error

Remarks  (Carroll et al. 2006; Yi 2017)

- Measurement error in variables may
  - change the structure of the response model
  - cause bias in parameter estimation
  - lead to a loss of power for detecting interesting relationship among variables
  - mask the features of the data

- The effects of measurement error are very complex, depending on the form of
  - the inference method
  - the measurement error model
  - the response model
  - the association of the covariates
Part 4: Time Series Data with Measurement Error

- Sensitivity Analysis
Objectives and Challenges

Objectives

- Model the daily quantities of interest
  - daily confirmed cases
  - daily deaths/fatality rates
  - daily recoveries
- Forecast the trends using historical data

Challenges

- Reported data are error-contaminated due to multiple reasons including
  - limited test capacity / shortage of healthcare personnel
  - undetected asymptomatic infections / unreported infections with light symptoms
  - delayed detection due to the incubation period
  - reporting errors and recall bias
- It is difficult to characterize the measurement error process due to
  - unknown error form
  - unknown error degree
Time Series Data in Error-Free Settings
**Autoregressive Model**

**Notation**

discrete time series: \( \{X_t : t = 1, \ldots, T\} \)

**AR\((p)\) Model**

\[
X_t = \phi_0 + \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t
\]

where

\( p \) is an integer smaller than \( T \)
the \( \epsilon_t \) are independent of each other and of the \( X_t \)
each \( \epsilon_t \) has zero mean and variance \( \sigma^2_\epsilon \)

**Interest**

estimation of \( \phi = (\phi_1, \ldots, \phi_p)^T \) and \( \phi_0 \)
prediction for a short period after \( T \)

**Remark**

The parameters are constrained for stationary time series.
Least Squares Method

Estimation of $\phi = (\phi_1, \ldots, \phi_p)^T$

- Let $S(\phi) = \sum_{t=p+1}^{T} \{ X_t - (\phi_0 + \sum_{j=1}^{p} \phi_j X_{t-j}) \}^2$

- Estimator:

  $$\hat{\phi}^{(LS)} = \text{argmin}_\phi S(\phi)$$

  $$= \left( \sum_{t=p+1}^{T} \{ \tilde{X}_{t-1} - E(\tilde{X}_{t-1}) \} \{ \tilde{X}_{t-1} - E(\tilde{X}_{t-1}) \}^T \right)^{-1}$$

  $$\times \sum_{t=p+1}^{T} \{ \tilde{X}_{t-1} - E(\tilde{X}_{t-1}) \} \{ X_t - E(X_t) \},$$

where $\tilde{X}_{t-1} = (X_{t-1}, \ldots, X_{t-p})^T$ and $E(X_t)$ is estimated by $\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} X_t$

- $\hat{\phi}_0^{(LS)} = \hat{\mu}(1 - \sum_{j=1}^{p} \hat{\phi}_j)$

- $\hat{\sigma}_\epsilon^2^{(LS)} = \frac{1}{T-p} S(\hat{\phi})$
Estimating Equation

Notation
\[ \gamma_k = \text{Cov}(X_t, X_{t-k}) \text{ for } t > p \text{ and } k = 0, \ldots, p \]
\[ \gamma = (\gamma_1, \ldots, \gamma_p)^T \]
\[ \Gamma: \text{the } p \times p \text{ autocovariance matrix} \]

Alternative Method
Estimators \( \hat{\phi}, \hat{\phi}_0 \) and \( \hat{\sigma}^2_\varepsilon \) solve
\[
\begin{align*}
\phi &= \hat{\Gamma}^{-1}\hat{\gamma}; \\
\phi_0 &= \left(1 - \sum_{i=1}^{p} \phi_i\right) \hat{\mu}; \\
\sigma^2_\varepsilon &= \hat{\gamma}_0 - 2\phi^T\hat{\gamma} + \phi^T\hat{\Gamma}\phi
\end{align*}
\]

Remark
- Asymptotic equivalence:
  \( \hat{\phi} - \hat{\phi}^{\text{LS}} \xrightarrow{p} 0, \hat{\phi}_0 - \hat{\phi}_0^{\text{LS}} \xrightarrow{p} 0 \) and \( \hat{\sigma}^2_\varepsilon - \hat{\sigma}^2_\varepsilon^{\text{LS}} \xrightarrow{p} 0 \) as \( T \rightarrow \infty \)
- This offers a unified estimation framework in its connections with the least squares estimation, the maximum likelihood method with the assumption of Gaussian error, and the Yule-Walker method.
Measurement Error Models
Measurement Error Models

Notation

- $X_t$: the true variable of interest on day $t$, which is unobserved
- $X^*_t$: the reported value of $X_t$

Goal

use the observed $\{X^*_t : t = 1, \ldots, T\}$ to understand / infer the true process of $\{X_t : t = 1, \ldots, T\}$
Additive Measurement Error Model

$$X_t^* = \alpha_0 + \alpha_1 X_t + e_t$$

for $t = 1, \ldots, T$, where

- $e_t$: independent of each other as well as of $X_t$; has mean 0 and variance $\sigma_e^2$
- $\alpha_0$ represents the systematic error
- $\alpha_1$ represents the constant inflation (or shrinkage) due to the measurement error

Remark

- If $\alpha_0 = 0$ and $\alpha_1 < 1$, the measurement error process implies that $X_t^*$ tends to be smaller than $X_t$ if the noise term is ignored.
- If $\alpha_0 = 0$ and $\alpha_1 > 1$, the measurement error process implies that $X_t^*$ tends to be greater than $X_t$ if the noise term is ignored.
- This model generalizes the classical additive model considered by Staudenmayer & Buonaccorsi (2005) who considered the case with $\alpha_0 = 0$ and $\alpha_1 = 1$.
Multiplication Measurement Error Model

Multiplicative Measurement Error Model

\[ X^*_t = \beta_0 u_t X_t \]

for \( t = 1, \ldots, T \), where

\( \beta_0 \): a scaling parameter

\( u_t \): independent of each other and of the \( X_t \); has mean 1 and variance \( \sigma_u^2 \)

Remark

These models are examples of describing the discrepancies between \( X_t \) and \( X^*_t \).

Other flexible models may be considered. For example,
- two-piece additive model:
  \[ X^*_t = \alpha_{10} + \alpha_{11} X_t + e_{1t} \quad \text{for } t = 1, \ldots, T_1 \]
  \[ X^*_t = \alpha_{20} + \alpha_{21} X_t + e_{2t} \quad \text{for } t = T_1 + 1, \ldots, T \]
  where \( T_1 \) represents the day on which measures of curbing the virus start

- mixture of additive and multiplicative models:
  \[ X^*_t = \alpha_0 + \beta_0 u_t X_t + e_t \quad \text{for } t = 1, \ldots, T \]
Bias Analysis for Naive Estimators
Bias Analysis for the AR(p) model

Naive Analysis

If naively replacing $X_t$ in the AR(p) model with $X_t^*$, then the working model is

$$X_t^* = \phi_0^* + \sum_{j=1}^{p} \phi_j^* X_{t-j}^* + \epsilon_t^*$$

where

$(\phi_0^*, \phi_1^*)^\top$ and $\epsilon_t^*$ show possible difference from the corresponding quantity in the AR(p) model.

Naive Estimator

$\hat{\phi}_1^*$ and $\hat{\phi}_0^*$ are the least squares estimators of $\phi_1^*$ and $\phi_0^*$
Bias Analysis with AR(1) Model

Theorem 0  Assume that the times series \( \{ X_t : t = 1, \ldots, T \} \) is AR(1) stationary. Then

\[
\hat{\phi}_1^* \xrightarrow{p} \phi_1^* \quad \text{and} \quad \hat{\phi}_0^* \xrightarrow{p} \phi_0^* \quad \text{as} \quad T \to \infty
\]

where

(a) Under the additive measurement error model:

\[
\hat{\phi}_1^* = \phi_1 \omega_1 \quad \text{and} \quad \hat{\phi}_0^* = \left( \alpha_0 + \frac{\alpha_1 \phi_0}{1 - \phi_1} \right) (1 - \phi_1 \omega_1)
\]

with \( \omega_1 = \left\{ 1 + \frac{\sigma_e^2 (1 - \phi_1^2)}{\alpha_1 \sigma_e^2} \right\}^{-1} \)

(b) Under the multiplicative measurement error model:

\[
\hat{\phi}_1^* = \phi_1 \omega_2 \quad \text{and} \quad \hat{\phi}_0^* = \frac{\beta_0 \phi_0}{1 - \phi_1} (1 - \omega_2 \phi_1)
\]

with \( \omega_2 = \left\{ 1 + \sigma_u^2 + \frac{\sigma_u^2 \phi_0^2}{\sigma_e^2} \frac{1 + \phi_1}{1 - \phi_1} \right\}^{-1} \)
Bias Analysis with AR($p$) Model with $p \geq 2$

**Theorem 0** Assume that the times series \{X_t : t = 1, \ldots, T\} is AR($p$) stationary with $p \geq 2$. Then

$$\hat{\phi}_1^{*} \xrightarrow{p} \phi_1^{*} \quad \text{and} \quad \hat{\phi}_0^{*} \xrightarrow{p} \phi_0^{*} \quad \text{as} \quad T \to \infty$$

where

(a) Under the additive measurement error model:

$$\phi^{*} = \alpha_1^2 (\alpha_1^2 \Gamma + \sigma_e^2 I_p)^{-1} \gamma$$

$$\phi_0^{*} = (1 - \phi_1^{*} 1_p) (\alpha_0 + \alpha_1 \mu)$$

(b) Under the multiplicative measurement error model:

$$\phi^{*} = \{\Gamma + \sigma_u^2 (\gamma_0 + \mu^2) I_p\}^{-1} \gamma$$

$$\phi_0^{*} = \beta_0 (1 - \phi_1^{*T} 1_p) \mu$$
Adjustment of Measurement Error Effects
Adjustment of Measurement Error Effects

Basic Idea
find $\tilde{\mu}$ and the $\tilde{\gamma}_k$ which are expressed in terms of the observed data $X_i^*$ such that

$\tilde{\mu}$ and $\hat{\mu}$ have the same limit in probability
$\tilde{\gamma}_k$ and $\hat{\gamma}_k$ have the same limit in probability for $k = 0, \ldots, p$

Estimating Equation
Estimators $\tilde{\phi}, \tilde{\phi}_0$ and $\tilde{\sigma}_\epsilon^2$ are obtained by solving

$\phi = \tilde{\Gamma}^{-1} \tilde{\gamma}$

$\phi_0 = (1 - \sum_{i=1}^P \phi_i) \tilde{\mu}$

$\sigma_\epsilon^2 = \tilde{\gamma}_0 - 2\phi^T \tilde{\gamma} + \phi^T \tilde{\Gamma} \phi$

Theorem 0  Assume regularity conditions. Then as $T \to \infty$

(1) $\tilde{\phi} \xrightarrow{p} \phi, \tilde{\phi}_0 \xrightarrow{p} \phi_0, \text{ and } \tilde{\sigma}_\epsilon^2 \xrightarrow{p} \sigma_\epsilon^2$;

(2) $\sqrt{n}(\tilde{\phi} - \phi) \xrightarrow{d} N(0, GQG^T)$,

where $G$ is the matrix of derivatives of $\tilde{\phi}$ with respect to the components of $(\tilde{\gamma}_0^*, \tilde{\gamma}_0^{*T})^T$, and the form of $Q$ depends on the measurement error model.
Adjustment of Measurement Error Effects

Forecasting

\[ \hat{X}_{T+h} = \hat{\phi}_0 + \hat{\phi}_1 \hat{X}_{T+h-1} + \ldots + \hat{\phi}_p \hat{X}_{T+h-p} \]

with \( \hat{X}_j \) estimated from the measurement error model

Prediction Error

the \( h \)-step prediction error with the AR(1) model is

\[
E\{(\hat{X}_{T+h} - X_{T+h})^2\} = \begin{cases} 
\frac{\phi_1^{2h} \sigma_e^2}{1 - \phi_1^2} + \frac{1 - \phi_1^{2h}}{1 - \phi_1^2} \sigma^2_\epsilon & \text{for the additive error model} \\
\phi_1^{2h} \left( \frac{\sigma_e^2}{1 - \phi_1^2} + \frac{\phi_0^2}{(1 - \phi_1)^2} \right) \sigma^2_u + \frac{1 - \phi_1^{2h}}{1 - \phi_1^2} \sigma^2_\epsilon & \text{for the multiplicative error model}
\end{cases}
\]

Remark

- These procedures are developed under the assumption that the measurement error model and the associated parameters are known.
- This assumption is usually untrue in applications. But the developed procedures are useful for conducting sensitivity analyses to understand the impact of measurement error on inference results.
Sensitivity Analysis for the Fatality Rate in Canadian Provinces: Ontario and British Columbia
Uncertainty in Defining the Fatality Rate

Definition 1: \[ Y_t = \frac{\text{the \# of deaths on day } t}{\text{the \# of cases up to day } t} \]
Definition 2: \[ Y_t = \frac{\text{the \# of deaths on day } t}{\text{the \# of cases up to day } t - 10} \]
Definition 3: \[ Y_t = \frac{\text{the \# of deaths on day } t}{\text{the \# of cases up to day } t - 14} \]

Remark

- Definition 1 is conventionally used.
- Definition 2 is proposed based on the fact that the median time from onset of symptoms to ICU admission is around 10 days (Huang et al. 2020).
- Definition 3 is taken by Baud et al. (2020) with the consideration that the maximum incubation period is assumed to be up to 14 days.

Source: Baud et al. (2020)
Comparison of the Fatality Rate from Different Definitions

Fatality Rate of Ontario and British Columbia (BC)
Sensitivity Analysis of the Fatality Rate

Data
We use the COVID-19 data in British Columbia and Ontario containing the daily confirmed cases and deaths from April 10 to May 4, 2020.

Goal
forecast the fatality rate for May 5-9 of 2020 in British Columbia and Ontario

Sensitivity Analysis

Starting Point:
If under-reported cases are only caused from undetected asymptomatic cases, then

\[ Y_{t-i} = (1 - \tau)Y_t^* \quad \text{for } i = 0, 10, 14 \]

where \( \tau \) represents the average rate of asymptomatic infections; set \( \tau = 46\% \) as the estimate from the meta-analysis of He, Yi and Zhu (2020).

Multiplicative Model: \( \beta_0 = \frac{1}{1-\tau} \); \( \sigma_u^2 = 0.2^2 \) or \( 0.5^2 \)

Additive Model: \( \alpha_0 = 0 \) and \( \alpha_1 = \frac{1}{1-\tau} \); \( \sigma_e^2 = 0.1^2 \) or \( 0.3^2 \)

Note: \( Y_t \) may be transformed by “differencing" to make the series (nearly) stationary.
Sensitivity Analysis: Definition 1 - BC vs ON

Definition 1 for British Columbia: AR(1), order–1 differencing

Definition 1 for Ontario: AR(4), order–1 differencing
Sensitivity Analysis: Definition 2 - BC vs ON

Definition 2 for British Columbia: AR(2), order-1 differencing

Definition 2 for Ontario: AR(2), no differencing
Sensitivity Analysis: Definition 3 - BC vs ON
Concluding Remarks
Take Home Messages

Confirmed cases are commonly error-contaminated:

- under-reported:
  - limited test capacity
  - asymptomatic patients are not detected.

- testing error:
  - false positive
  - false negative

Incubation times varies from individual to individual:
- recall bias
- reporting bias

Fatality rate estimates based on Definition 1 may not be representative of the actual death rate.
- reported confirmed cases on a given day usually contracted the virus at an earlier time
Take Home Messages

COVID-19 data are available from multiple sources, and many methods have been developed:

- statistical validity/efficiency
- modeling complexity/validity
- computation complexity/feasibility
**Take Home Messages**

COVID-19 data are available from multiple sources, and many methods have been developed:
- statistical validity/efficiency
- modeling complexity/validity
- computation complexity/feasibility

An issue is often overlooked:
- EXAMINING DATA PROVENANCE AND QUALITY IS CRUCIAL!
COVID-19 data are available from multiple sources, and many methods have been developed:
- statistical validity/efficiency
- modeling complexity/validity
- computation complexity/feasibility

An issue is often overlooked:
- EXAMINING DATA PROVENANCE AND QUALITY IS CRUCIAL!

Reported COVID-19 data are not automatically useful:
- Ignoring error in COVID-19 data can yield misleading results!
- Carefully scrutinizing COVID-19 is needed!
Take Home Messages

COVID-19 data are available from multiple sources, and many methods have been developed:
- statistical validity/efficiency
- modeling complexity/validity
- computation complexity/feasibility

An issue is often overlooked:
- EXAMINING DATA PROVENANCE AND QUALITY IS CRUCIAL!

Reported COVID-19 data are not automatically useful:
- Ignoring error in COVID-19 data can yield misleading results!
- Carefully scrutinizing COVID-19 is needed!

Characterizing the form and degree of error in COVID-19 data can be difficult:
- Conducting sensitivity analyses may be a viable strategy!
Acknowledgements

Funding Agency:

NSERC
Canada Research Chairs Program