Modeling COVID-19 Cases

On Jan. 22, 2020 the first 9 cases of the novel Coronavirus (COVID-19) were diagnosed outside of China. As an expert in mathematics, modeling, and data science, you have been called by the World Health Organization (WHO) as a consultant to model the growth of the COVID-19 cases outside China. Billion dollar decisions on deployment of resources, limitations on travel, and restrictions on trade will be based on your modeling and predictions.

Today is Feb. 7, 2020 and you are about to board a private jet to WHO Headquarters in Geneva to make your presentation. Your primary objective is to predict the cumulative number of cases outside China for the coming two weeks, until Feb. 21st (T=31) and beyond. The only data that you have are the number of cases (Y) reported daily starting from Jan. 22nd ($T = 1$) until today, Feb. 7th (T=17):

T	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Y	9	6	15	10	16	10	18	18	29	28	14	13	4	31	27	30	52

You slam this data into Matlab just before takeoff and after looking at the data on a semilog plot `figure;semilogy(T,Y,'.');` you quickly decide on an exponential growth model.

$$Y = \exp(a_1 + a_2 T)$$

You choose this model because the semi-log plot looks somewhat linear, and you have a clever trick to easily fit this model. Taking log of both sides you have,

$$\log Y = a_1 + a_2 T$$

which is a simple linear model.

Question 1: To find the best fit line you first try a **least squares** fit. Code this up and plot the data alongside your model prediction for `T=1:17;` on a normal plot and a semilog plot. (To check your answer: $a_1 \approx 2.12$ and $a_2 \approx 0.072$).

Question 2: Compute the standard deviation of your fit using the `std` command on the differences $\log Y - a_1 - a_2 T$. Now add 95% confidence intervals to your plots using:

```
upper = exp(a1+a2*T+2*stddev);
lower = exp(a1+a2*T-2*stddev);
fill([T fliplr(T)],[upper fliplr(lower)],[.8 .8 .9])
```

(you may need to reorder your plot commands to see all the curves, if fill doesn't work use `plot(T,upper)`)

Question 3: Use the `cumsum` command to plot the cumulative cases (with error bounds) out to March 11th ($T = 50$).

The fit looks pretty reasonable, but then you remember that least squares assumes Gaussian noise. This is worrisome because when T is small your model is predicting a very small number of cases, and a Gaussian distribution would actually say there is a probability of a negative number of cases, which is clearly nonsense. Also, Gaussian noise would allow fractions of a cases, but the number of cases is always an integer. Then you remember your stats professor who told you about the Poisson distribution, which models a random number of arrivals with a known average arrival rate λ. Using Poisson, the probability of k patients showing up at the hospital is:

$$P(k \,|\, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}$$

So you assume that at each time T the expected arrival rate is given by your model

$$\lambda = Y = \exp(a_1 + a_2 T)$$

so given the parameters a_1 and a_2, the probability of observing Y_i cases at time T_i is

$$P(Y_i \,|\, a_1, a_2) = e^{-\exp(a_1 + a_2 T_i)} \frac{(\exp(a_1 + a_2 T_i))^{Y_i}}{Y_i!}$$

You assume that the Poisson random variables with different T values are all independent, so that the probability of the 17 data points you have is the product of their individual probabilities:

$$P(Y_1, ..., Y_{17} \,|\, a_1, a_2) = \prod_{i=1}^{17} e^{-\exp(a_1 + a_2 i)} \frac{(\exp(a_1 + a_2 i))^{Y_i}}{Y_i!}.$$

Your goal is to maximize this by choosing the best a_1, a_2. In other words, we want to pick the model such that the observed data is the most likely among all possible models. This is the Maximum Likelihood Estimator (MLE) of the parameters a_1, a_2. In order to maximize this likelihood, you recall that it is equivalent to maximize the logarithm (because the log function is monotone increasing), so you compute the log-likelihood:

$$\begin{aligned} L(a_1, a_2) = \log(P(Y_1, ..., Y_{17} \,|\, a_1, a_2)) &= \log\left(\prod_{i=1}^{17} e^{-\exp(a_1 + a_2 i)} \frac{(\exp(a_1 + a_2 i))^{Y_i}}{Y_i!}\right) \\ &= \sum_{i=1}^{17} \log\left(e^{-\exp(a_1 + a_2 i)} \frac{(\exp(a_1 + a_2 i))^{Y_i}}{Y_i!}\right) \\ &= \sum_{i=1}^{17} \log(e^{-\exp(a_1 + a_2 i)}) + \log((\exp(a_1 + a_2 i))^{Y_i}) - \log(Y_i!) \\ &= \sum_{i=1}^{17} -\exp(a_1 + a_2 i) + Y_i * (a_1 + a_2 i) - \log(Y_i!) \end{aligned}$$

Question 4: To find the maximum likelihood estimates for a_1 and a_2 you need to maximize $L(a_1, a_2)$ which means computing the gradient and setting it equal to zero. First, compute the gradient:

$$\nabla L = \begin{bmatrix} \frac{\partial L}{\partial a_1} \\ \frac{\partial L}{\partial a_2} \end{bmatrix} =$$

Question 5: In order to solve $\nabla L(\vec{a}) = \vec{0}$ for $\vec{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ you want to use Newton's method, which requires the derivative of ∇L which is the 'second derivative' or Hessian matrix of L, compute the Hessian matrix:

$$H(L) = \begin{bmatrix} \frac{\partial^2 L}{\partial a_1 \partial a_1} & \frac{\partial^2 L}{\partial a_2 \partial a_1} \\ \frac{\partial^2 L}{\partial a_1 \partial a_2} & \frac{\partial^2 L}{\partial a_2 \partial a_2} \end{bmatrix} =$$

Question 6: Now apply the multivariable Newton's method,

$$\vec{a}_{k+1} = \vec{a}_k - H(L)(\vec{a}_k)^{-1}\nabla L(\vec{a}_k)$$

to find the solution to $\nabla L(\vec{a}) = \vec{0}$ which will give you the a_1 and a_2 that maximize L and so are the most likely values. Use the least squares solution $\vec{a}_0 \approx \begin{bmatrix} 2.12 \\ 0.07 \end{bmatrix}$ as your initial guess.

(You should get $a_1 \approx 2.11$ and $a_2 \approx 0.085$)

Question 7: Repeat the plots from Questions 1-3 using your new a_1 and a_2. The Poisson model has variance λ, so the confidence intervals become:

```
upper = exp(a(1)+a(2)*T)+2*sqrt(exp(a(1)+a(2)*T));
lower = exp(a(1)+a(2)*T)-2*sqrt(exp(a(1)+a(2)*T));
```

Fast forward to today, your model had the best performance and WHO has asked for an updated fit (for a modest additional consulting fee). Go to

`https://www.worldometers.info/coronavirus/coronavirus-cases/`

and grab the latest data (view source and search for "total cases outside of china" and you can find the cumulative cases data in the HTML source).

Question 8: Add the latest data to your cumulative sum plots (the data is cumulative cases), still comparing to your model that you fit using only days 1-17. How did the two models compare between Feb. 7th and today?

Question 9: Refit both models using all the data currently available and run your forecast out for 200 days. The data from the internet is cumulative sums so we need to convert that to daily new cases using the following command (assuming the internet data is stored as a column vectored named `Y2`)

```
Y2= [Y2(1); diff(Y2)];
```

Check to make sure that the first 17 days of `Y2` agree with the original `Y` data in the table on page 1. When do you predict there will be a million cumulative cases? Notice that the curve will just continue to infinity which is obviously impossible, so the model must break down at some point in the future. How long do you think this model will give accurate forecasts?