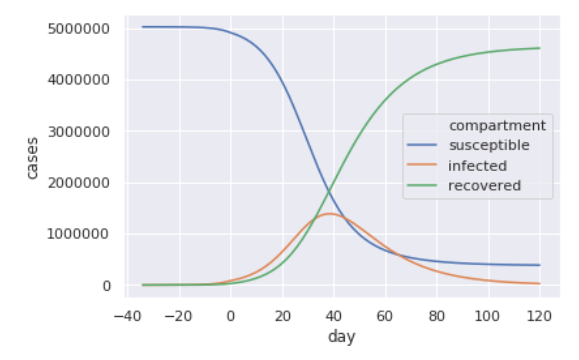
This uses concepts and techniques we learned in the Golf Clubs Pricing Example we went through in class. In fact, feel free to use the file from that example as a starting template. **However, you’ll be making significant changes to it as this problem differs from that one in several ways (such as using RMSE as the error metric instead of MAPE – details below).**

## Analysis – Modeling the early stages of the Covid-19 pandemic

Modeling and analyzing the dynamics of pandemics such as Covid-19 can be quite difficult due to numerous issues such as:

* lack of reliable data on key pandemic parameters such as transmission rates, infectious time, recovery time, and whether or not immunity exists after recovery,
* changes in testing rates as tests are developed and ramped up for widescale use,
* changes in pandemic dynamics due to changes in behavior and policy.

Nevertheless, epidemiologists and other scientists and engineers have been hard at work trying to build useful and sufficiently accurate models to help guide policy and decision making. Many of these models are based on what are known as [compartment models](https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology). One very useful, yet simple, such model is known as the SIR model in which people move between three states: Susceptible, Infected, Removed. The rates at which people move between states are important parameters whose values affect how the pandemic will play out. The simplest, deterministic versions of SIR models lead to a set of differential equations from which quantitative estimates of disease dynamics can be obtained. Here’s what a typical plot might look like from the analysis of an SIR model.



It can be shown that for the SIR model, early stage epidemic growth is exponential – i.e. growth is at a constant rate of the current number of infections. As we saw in class, exponential growth of any population can quickly lead to massive population size. The number of susceptible people is reduced as more and more get infected, and if one assumes at least some temporary immunity, the infection will start to slow and then peak and decrease as there is less “fuel for the fire”. Of course, this is just one, quite simple, epidemic model. There are numerous variants of this model (see the link above to information on compartment models, if you are interested).

In this assignment, we’ll be looking at **very simple** models for use during the early stages of the spread of a pandemic when infections are spreading quickly. Specifically, we’ll just be statistically fitting some simple functions such as linear, exponential, power, and polynomial functions to actual Covid-19 data from the state of Michigan to see if they seem to exhibit exponential growth. As we will see, one must be **VERY CAREFUL** when doing such “curve fitting” in terms of how the models are used.

**1 – Time Series Plot**) Use the **covid\_mich\_period1.csv** data file provided with this assignment. On the first sheet, named Period1, it contains the reported number of cases from mid-March to early April. Create a scatter plot which shows sales over time for the 2020-03-12 to 2020-04-01 time period. **Suggestion (strong)**: Do NOT use the Date column directly as your X-axis for the scatter. Instead, create a Day column which starts at 1 (for 2020-03-12) and use Day in your scatter plot. Move your scatter plot to a chart sheet and make sure it is properly formatted and includes a title, axis labels, and whatever else you think it needs. Use principles of graphing excellence that we covered in the Data Visualization module. You can use this chart sheet as the basis for the trend fit charts in the next part (i.e. you can copy the sheet as needed).

**1a – Trend Line Fitting**) Using Excel’s Fit Trend Line feature of graphs, fit six different trend lines: linear, exponential, power, logarithmic, quadratic, and cubic. Which fits best? To determine which fits best, use the same approach as used in the Golf Clubs Pricing example we covered in class. However, **do NOT use MAPE** as your error metric. The error metric you are to use for choosing between the six candidate models is *root mean square error* (RMSE). RMSE is commonly used in predictive modeling to evaluate the relative predictive accuracy of competing models. For a dataset of size n, the RMSE is defined as:

where and . So, basically, we are just squaring the forecast errors, adding them all up, dividing by the sample size to get an average or *mean squared error (MSE)*, and then taking the square root. You can easily do this by including “helper columns” (much like we did in the Golf Clubs problem in which we included columns with the absolute value of the errors – of course, we do NOT need the absolute value of the errors in this problem because we are using RMSE instead of MAPE as our error metric). However, for a little **extra credit**, you could also compute RMSE for each model using array formulas based solely on the range of values containing the raw errors for each model.

Describe how each of the models does in terms of fitting the data with respect to RMSE. Which models seem to fit the best? Which fit the worst?

**For a little more extra credit, use array formulas such as LINEST and LOGEST along with the INDEX function to automatically pull the model coefficients into cells without having to copy and paste them from the trend line label or without manually typing them. See TrendFittingFunctions.xlsx in the Downloads for ideas.**

**1b – Future predictions**) Use all six models that you fit Part 1a to PREDICT the number of cases for 2020-04-02 through 2020-04-08. You’ll end up with six sets (of 7 days each) of predictions (one set for each model). You are using the models you fit with the period1 data to make predictions (forecasts) for 2020-04-02 through 2020-04-08.

**1c – Evaluating the predictions**) Now, let’s add some more recent data to see how our models did. In the **covid\_mich\_period2.csv file,** you’ll find data from 2020-04-02 through 2020-04-08 for the actual number of cases.

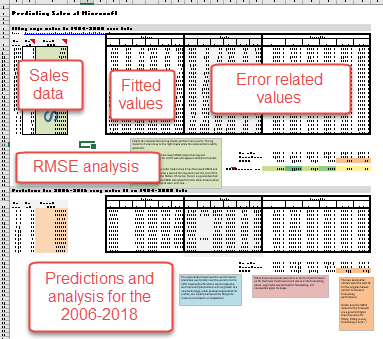
Compute the RMSE for your predictions.

Compare the performance of the various models in terms of how well they did in predicting the number of future cases compared to how well they did when being fit to historical data. Do the same models that did well in the fitting phase also do well in the prediction phase or has the relative performance of the models changed? Discuss.

Do the models perform better or worse during the prediction phase than they did during the model fitting phase? Why do you think this is the case?

Why do you think that cases in early April 2020 seem to be growing slower than the exponential growth predicted the SIR model?

Here’s what my main analysis sheet looks like. **IMPORTANT**: **My screenshot below is from a similar problem but with different data and a different number of years in Period1 and Period2**. I’m just showing you the overall layout of what my solution would look like. Obviously, the graphs and fitted parameter values are on other sheets.



**Again, the screenshot above is from a different problem but the layout of my spreadsheet is similar for this problem.**

**DELIVERABLE: You’ll be submitting your Excel file containing all of the work above.**