COVID-19 Forecasting Final Report

Thomas Wang, Qiuyue Wang

**1. Description of project and questions addressed (these will likely evolve over the course of the project).**

Our project aims to predict the total confirmed COVID-19 cases in the US on a daily basis. We want to predict total confirmed cases on daily basis and number of cases increased every day . We want to figure out a general evolution of the pandemic and give some insights on what stage we are currently in.

**2. Discussion of what methods were used, how they were used, and what motivated your choices. You will likely discuss things like data preprocessing, feature selection, feature extraction and transformation, regularization, model optimization, model selection and cross validation, etc.**

The first baseline we used is the simple linear regression. After plotting the total cases of each day, we see a slope and it looks like a simple linear regression. Thus, we hypothesize that daily increased cases are almost the same. The feature is just the number of days since March.15. The target variable is the total cases. We split the data into training dataset and testing dataset and squared loss is used to evaluate the model.

The first approach we use is to fit the data into a polynomial regression. We used a degree of 3. However, while the performance for the training dataset has been improved, the performance of the testing data even worked worse. We then determined to add ridge regularization to fix the overfitting problem.

The second approach we used is to use a delayed version of the features. Our hypothesis is that the total increased cases on one day would be related to the number of total increased cases in previous days. We looped from a lag of 1 day to 5 days to find the model that has the best performance.

The problem of the first two approaches is that we have a fixed point to split the training and testing data. Especially for the second approach, the lags are based on new predicted values. Thus, the performance would get worse and worse since the error will be accumulated. In order to fix this problem, we decided to update the feature of previous days to the actual number of confirmed cases. This will help change and improve accuracy. However, from this method, we can only predict one day ahead.

The third approach we tried is simple linear regression with previous day total cases as the feature. We want to predict next day’s total cases using the total cases of the current day. There is significant improvement of this method compared to the first baseline. We are going to set this approach as the second baseline. We will try data transformation based on that.

The fourth approach we tried is multiple linear regression with delay and regularization. Our hypothesis is that daily cases are related to other data such as death, number of tests conducted, number of hospitalization, number of patients on ventilator and etc. So we used those data from the previous days to predict how many new cases there will be on the next day. We used regularization to avoid overfitting and tried different delays.

For the previous approaches, we used total positive cases in the US as our target variable. However, it seems more intuitive to directly predict how many new cases will be on the next day. Also the number of total positive cases is too large to visualize the change on the graph. So we changed our target variable to daily increase of cases and tried the approaches above again. We will call it approach 5 and it has 3 parts.

For approach 5a, it was a simple linear regression with previous daily increase as the feature and we set this as the 3rd baseline.

In approach 5b, we did data transformation on the feature in 5a with LASSO regularization. We tried different degrees of transformation to find out the lowest loss we could get.

In approach 5c, we tried multiple linear using different features including: Total deaths, Daily death increase, Total cases, Daily new cases, Total test cases, Daily test, Currently hospitalized, New hospitalized cases in the US. For each set of features we chose, we tried different delays and alpha for regularization.

There are more different kinds of data we can access such as the number of patients on ventilators in the US. However, some features have fewer records than others. If we want to include those features that means our training and testing set will be smaller. Therefore, starting from approach 6, we will use a different set of data which included more features but fewer records. We need a new baseline because the testing size will be smaller than before and we shouldn’t compare the loss from this dataset to the previous baseline. Target variable will still be daily increased cases.

Approach 6a will be the new baseline for our new dataset. We will call it 4th baseline. It will be a simple linear regression with previous day increased cases as the feature.

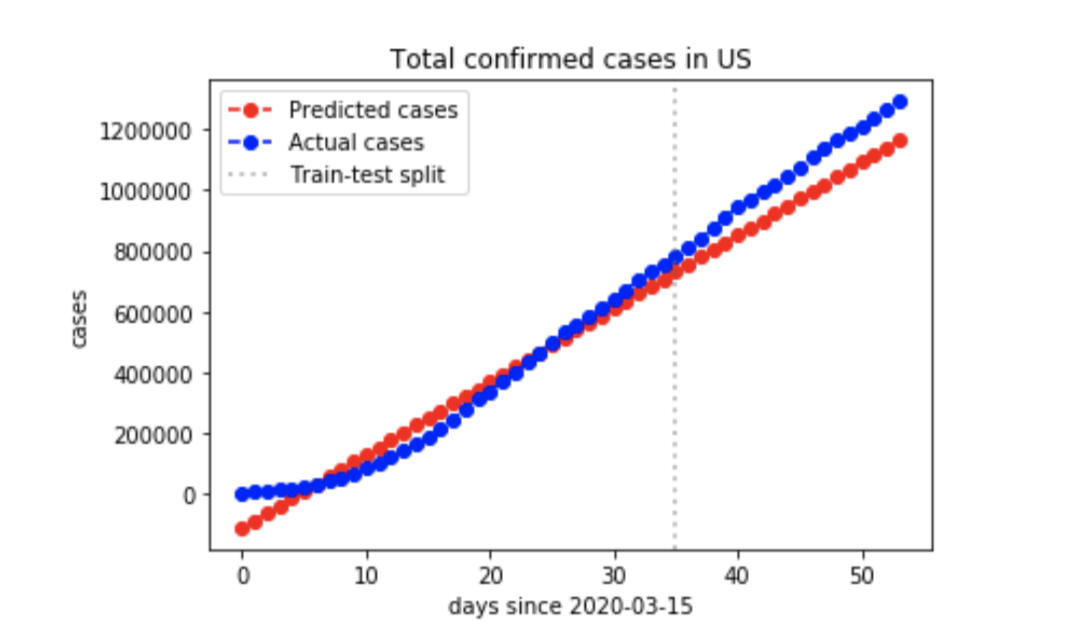
Approach 6b will be using data transformation on the feature in 6a. We tried different degrees and regularization alpha.

Approach 6c will be using multiple linear regression. In addition to the features we used in 5c, there was Number of cases on ventilators currently, Number of cases on ventilators cumulatively and Total recovered cases. We tuned regularization and tried different sets of features.

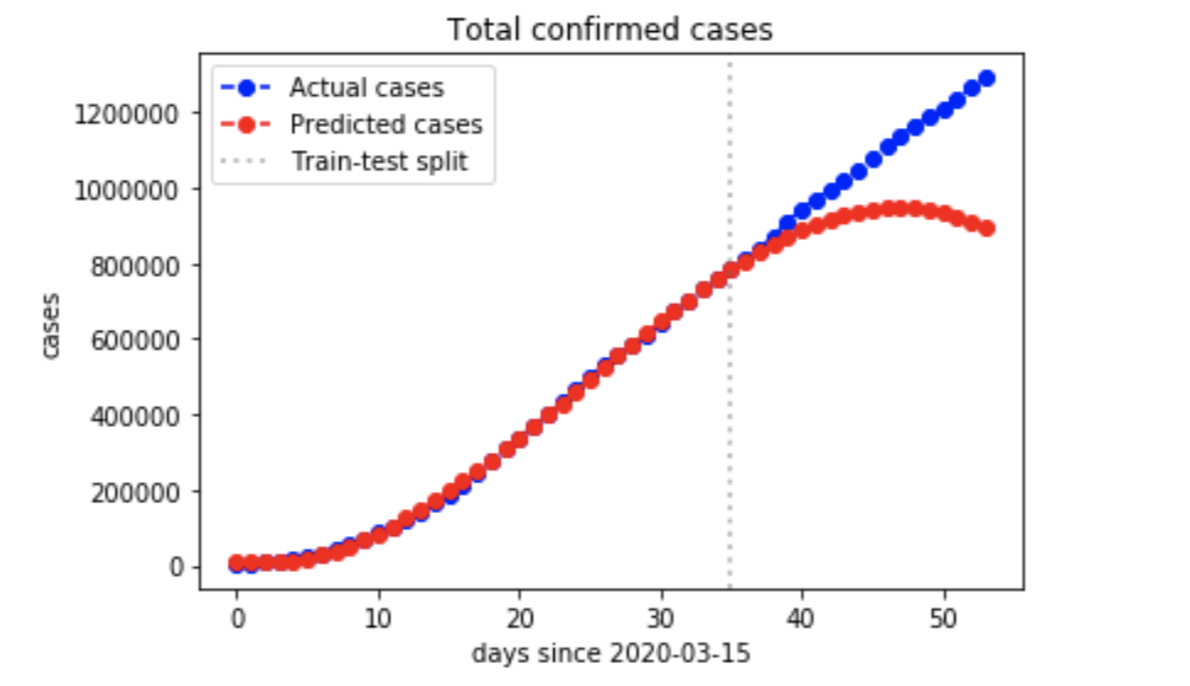
**3. Discussion of final model performance, including a full comparison to simple baseline methods. You will likely want to include plots, tables, or other figures, but do so judiciously. We don't need to see every experiment run.**

(1)First Baseline Performance

loss: 191063786160.99817

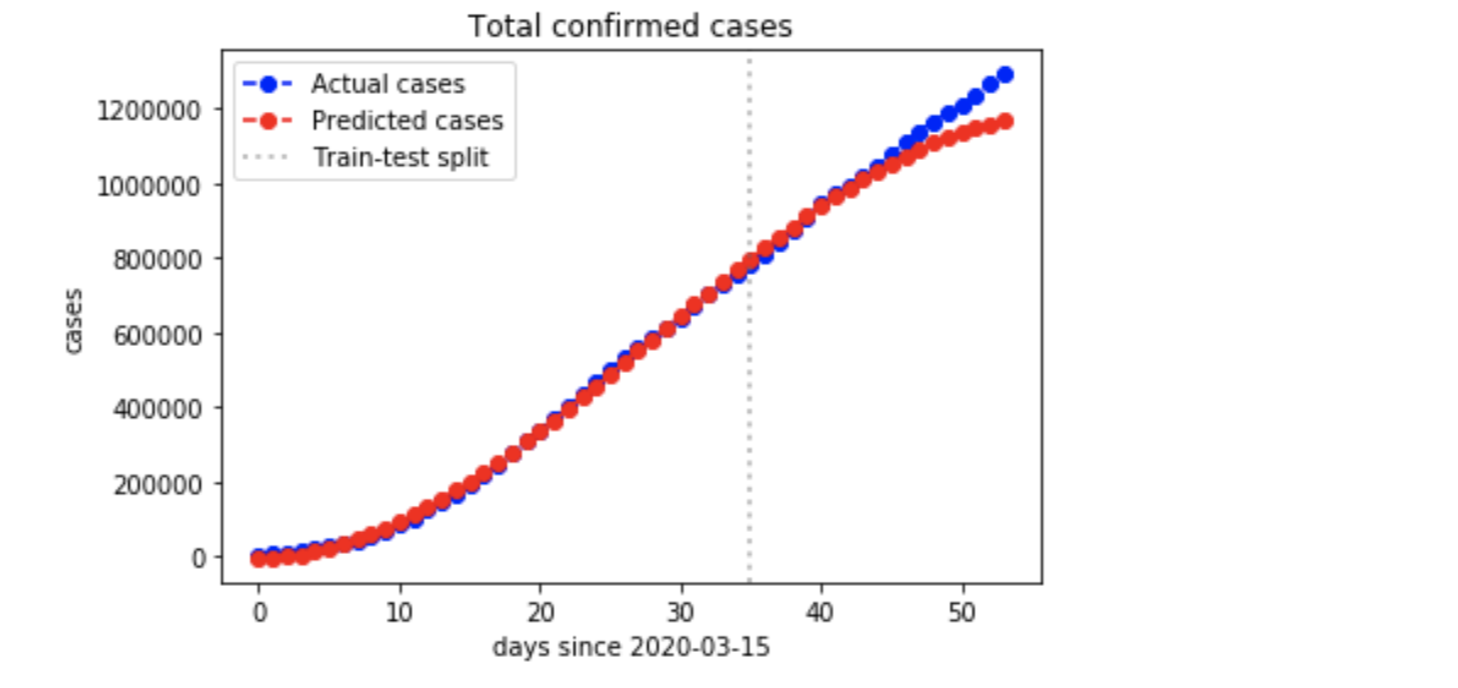


(2) Polynomial regression

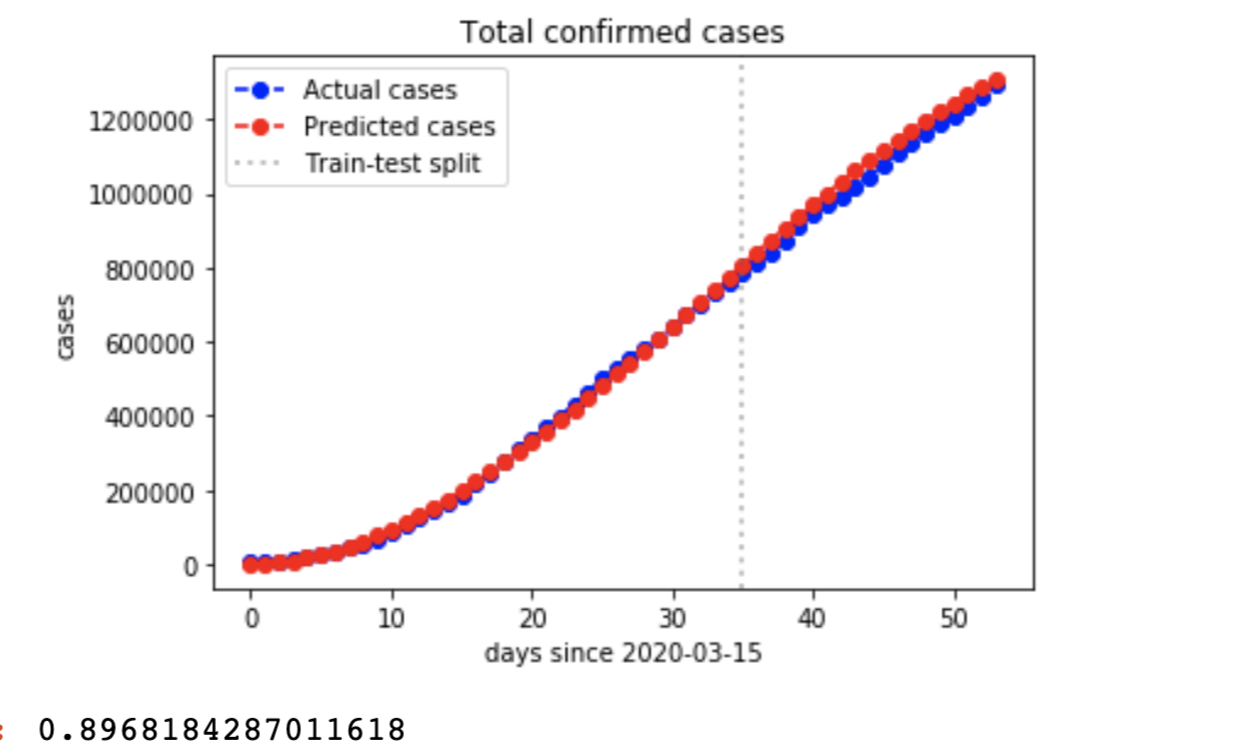


Result: 256% worse than the first baseline

(3) Polynomial Regression with ridge regularization



alpha = 1000 improved 82%



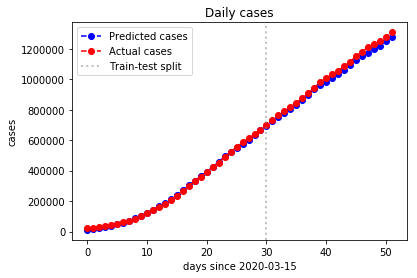
alpha = 10000 improved 90%

(4) Predicting with delays

|  |  |  |
| --- | --- | --- |
|  | lag = 1 | Worsen 320% |
|  | lag = 2 | improved 56% |
|  | lag = 3 | improved 51% |
|  | lag =4 | improved 45% |
|  | lag= 5 | improved 40% |

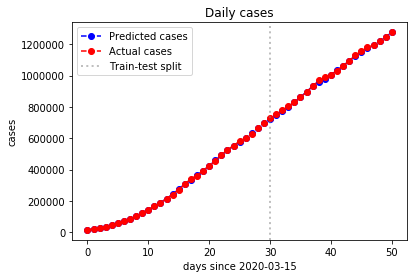
(5) Second Baseline Performance(3rd approach):

99.4% better than first baseline



(6) Increase delay to 3(3rd approach):

98% better than second baseline

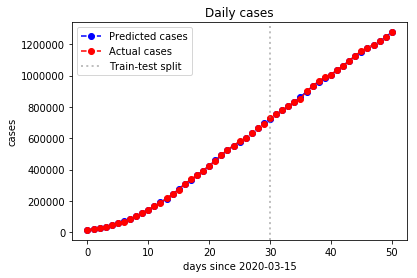


(7) Multiple linear regression(4th approach):

Features: total positive cases, daily positive increase, daily death increase, daily test increase.

96.8% better than 2nd baseline. Not better than the best performance of 3rd approach.

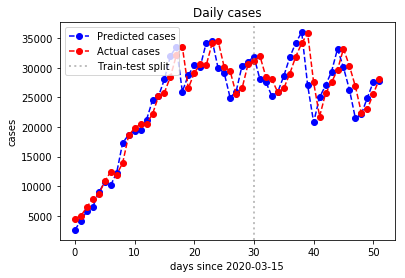
*Conclusion:* multiple linear regression doesn’t seem to do a better job than simple linear regression.



(8) Set daily increased cases as target variable(5th approach)

1. Simple linear regression with previous daily increase as feature(Third baseline)

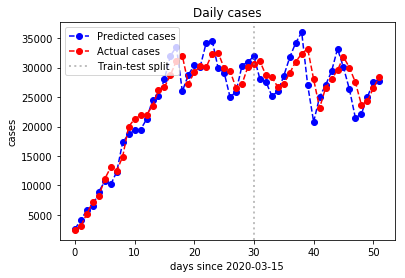
Loss: 259001827.52229908



1. With data transformation

Loss: 222175226.44708955

14.2% better than 3rd baseline

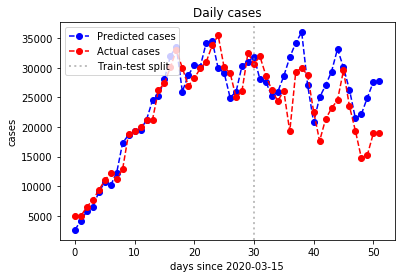


1. Multiple linear regression

Features: total positive cases, daily positive increase, daily death increase, daily test increase.

Loss: 393708307.7036844

52% *worse* than 3rd baseline

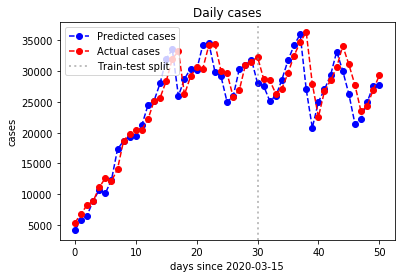


1. Multiple linear regression with less features

Features: total positive cases, daily positive increase

Loss: 277080663.04101866

7% worse than 3rd baseline

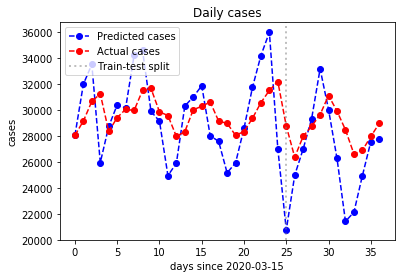


*Conclusion*: Still multiple linear regression loses to simple linear regression. Among different sets of features we use, it seems less features predicted better as long as the feature is more relevant to the target variable. However, all methods seemed to predict the correct trend.

(9) Use more features and less data(6th approach)

1. Simple linear regression with previous daily increase as feature(4th baseline)

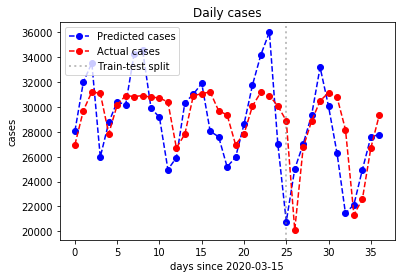
Loss: 168767390.8360201



1. With data transformation

Loss: 173186994.7628799

2.6% worse than 4th baseline

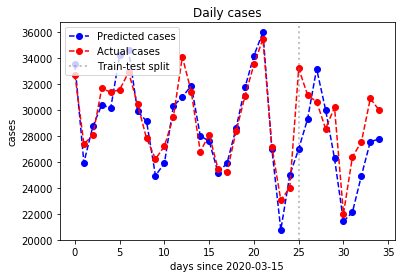


1. Multiple linear regression

Features: Number of cases on ventilators currently, Number of cases on ventilators cumulatively and Total recovered cases, total positive cases, daily positive increase, daily death increase, daily test increase.

Loss: 106505801.31091526

36.9% better than 4th baseline.



Conclusion: Multiple linear regression out performances simple linear regression this time. However, we need to use very large alpha for regularization this time because there are less data records, the model tends to overfit.

4. A conclusion describing what you would have done with more time: are there modifications of your approach worth trying? Other questions to address? Different data sets or features?

If we had time, we would have trained a model from other countries and test the data of the US. This needs more dataset research and we might also need to do some data processing and feature extraction.

Another thing we might do is normalizing the data. Since different countries might have different populations, in order to better compare the error, we might do some normalization.

Also, there must be more informative data for us to explore and experiment with if we had more time such as mobility, policies etc.

Reference:

GitHub repo: <https://github.com/MengranWang123/Machine-Learning-Project>