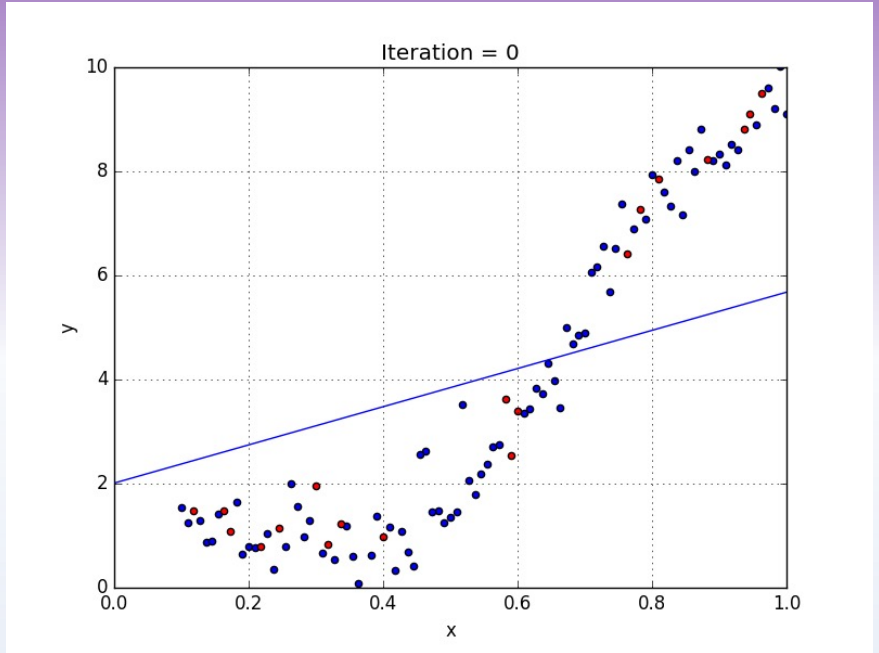
Directed Reading Program of
Statistics and Probability Association (SPA)

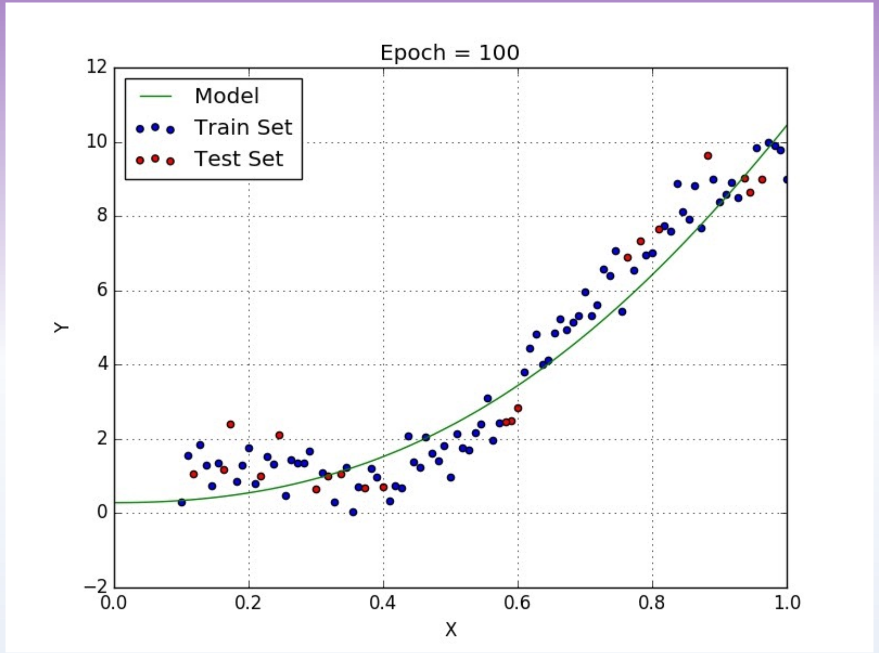
Nonlinear Regression on Covid-19 data

Muhammad Anas, mentored by Michael Pearce

Overview Nonlinear Regression

Linear vs Nonlinear Regression





Simple linear regression allows us to summarize the relationship between two continuous quantitative variables that are dependent on each other.

(Normally distributed data)

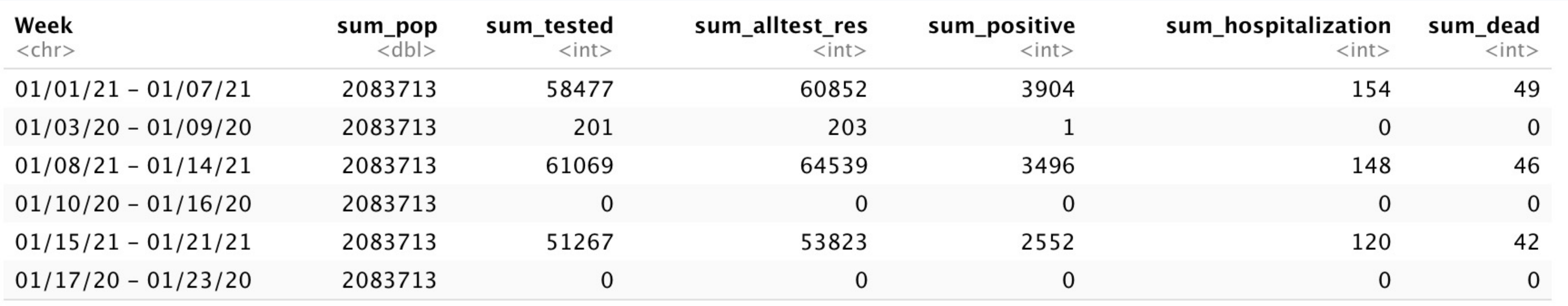
Note: Iteration use to find best fit lines base on smallest sum of squared residuals

Nonlinear regression is a form of regression in which is modeled by a function that is a nonlinear combination of the model parameter and depends on one or more independent variables.

Note: Epoch is term used to indicate the number of iteration used by machine learning algorithm on the training dataset

Dataset review

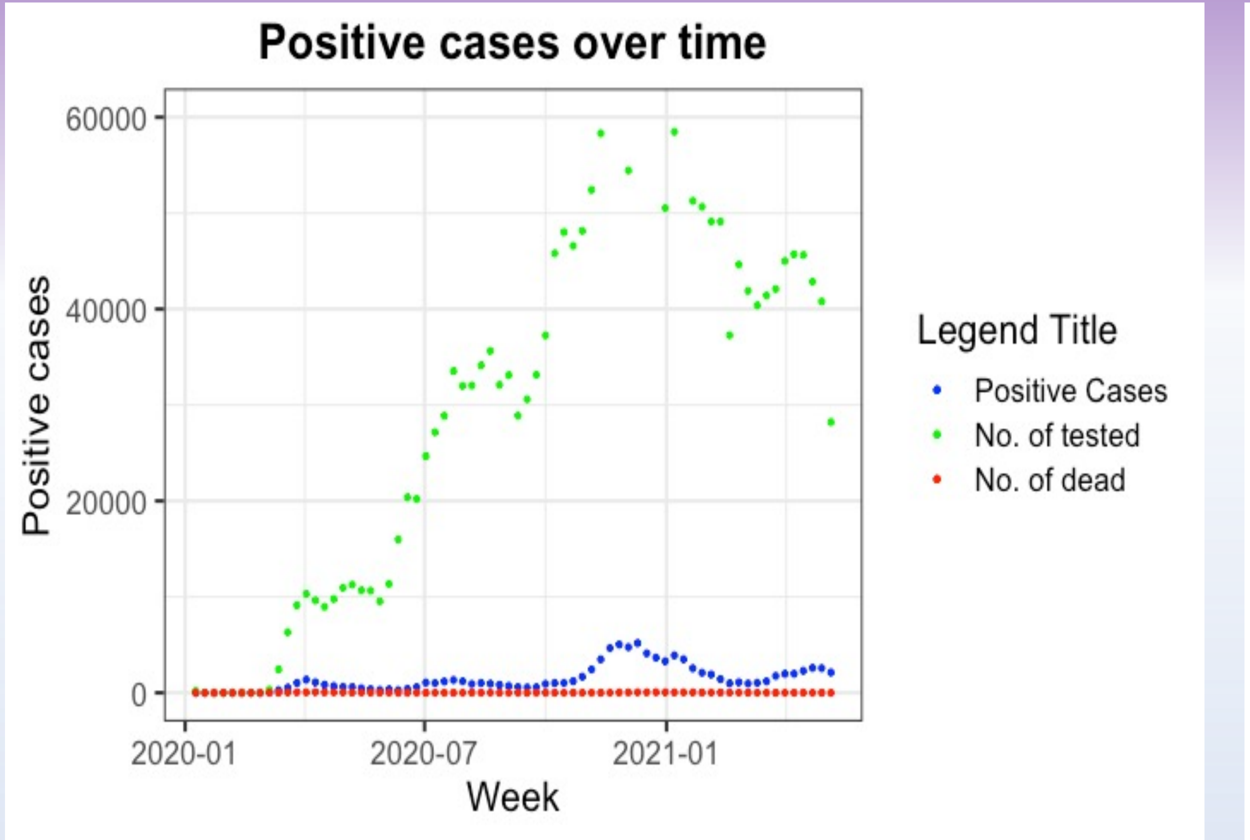
- Retrieved from official King County Government website.
- Dataset contained the weekly number of cases, number of tested, number of hospitalization, and number of death in King County, WA.
- Contained 700 observation based on age group by weekly count.

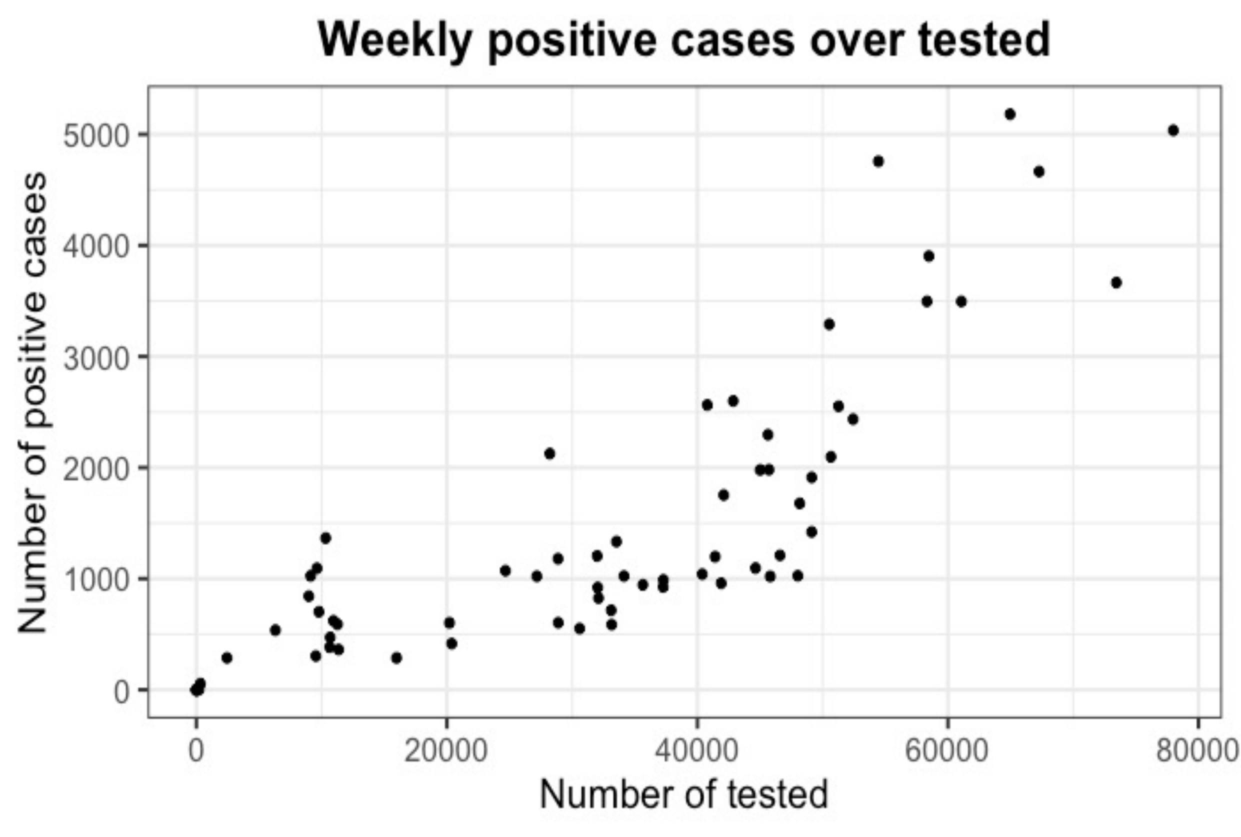


Week <chr>	sum_pop <dbl>	sum_tested <int>	sum_alltest_res <int>	sum_positive <int>	sum_hospitalization <int>	sum_dead <int>
01/01/21 – 01/07/21	2083713	58477	60852	3904	154	49
01/03/20 – 01/09/20	2083713	201	203	1	0	0
01/08/21 – 01/14/21	2083713	61069	64539	3496	148	46
01/10/20 – 01/16/20	2083713	0	0	0	0	0
01/15/21 – 01/21/21	2083713	51267	53823	2552	120	42
01/17/20 – 01/23/20	2083713	0	0	0	0	0

Project overview

Predict weekly rate of cases in forthcoming week in King County





Project Goals

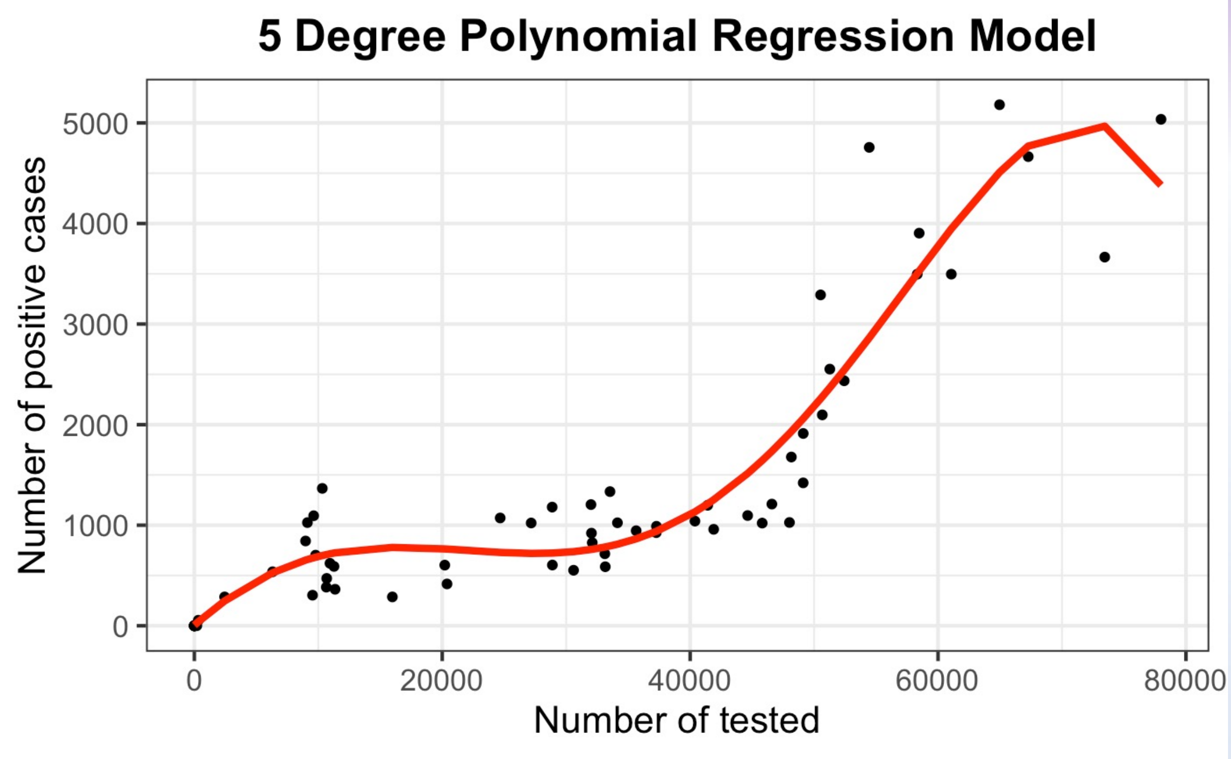
Choosing the best model to predict the future rate of positive cases by comparing the average least residual error and the Residual Sum Square.

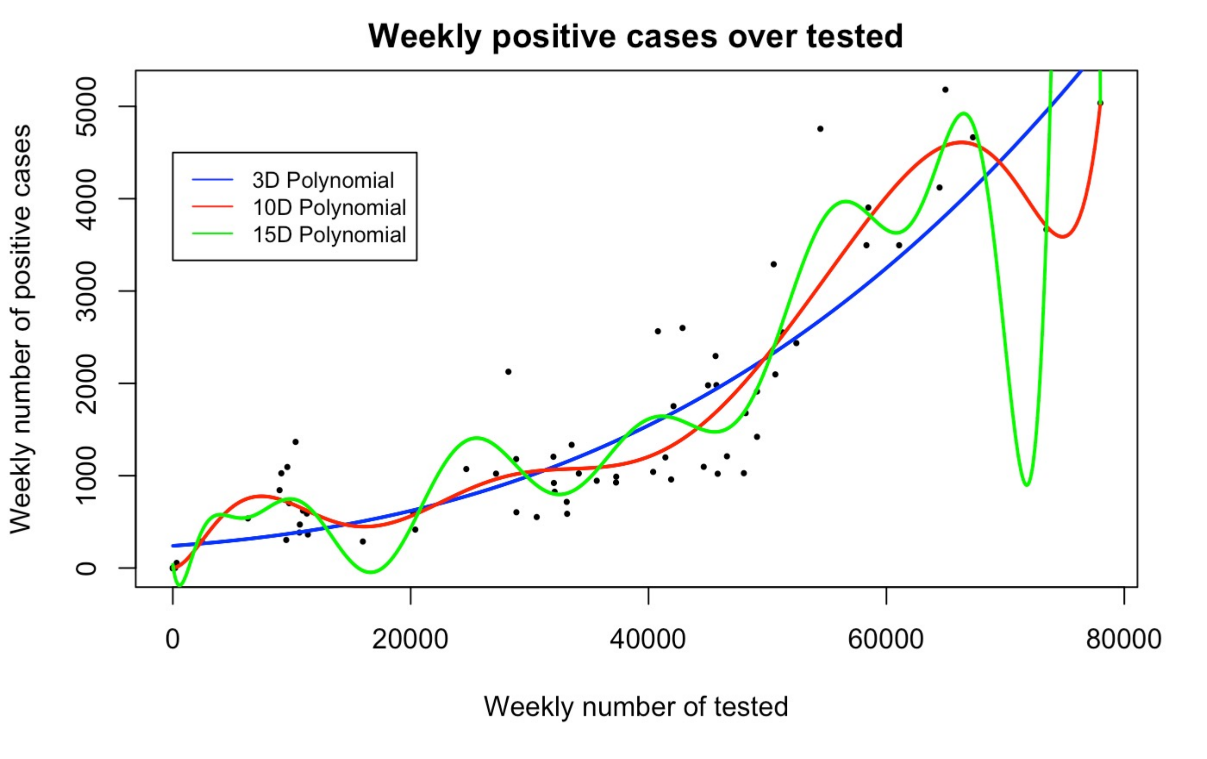
Method:

- Separate dataset into 10 time points
- Fit the model in first time point, predict the next time point
- Keep track of the residual error
- Compare the model and choose model with least residual error

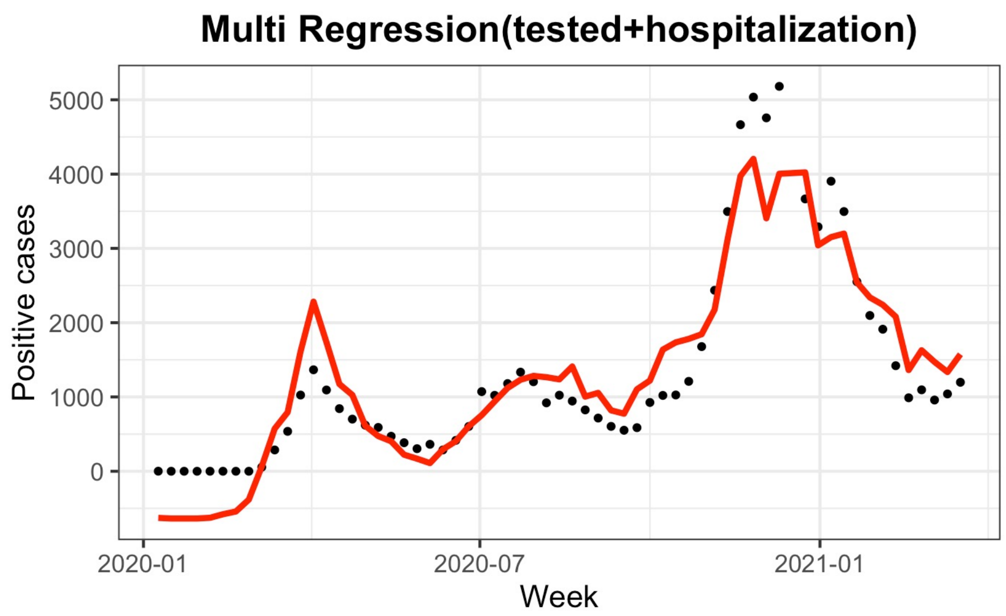
Polynomial Regression

Extension of Linear Regression

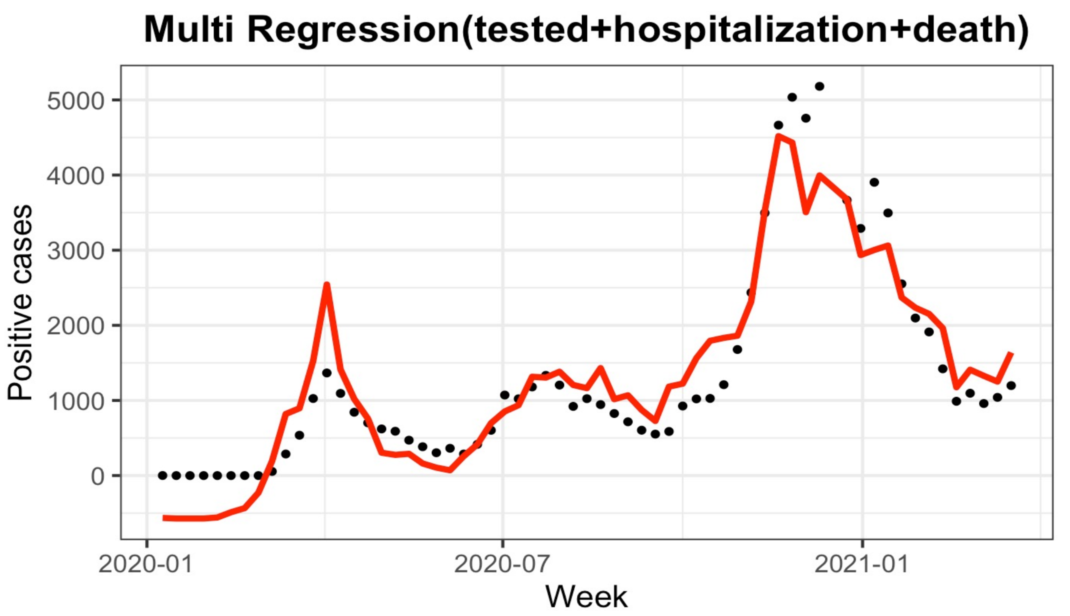




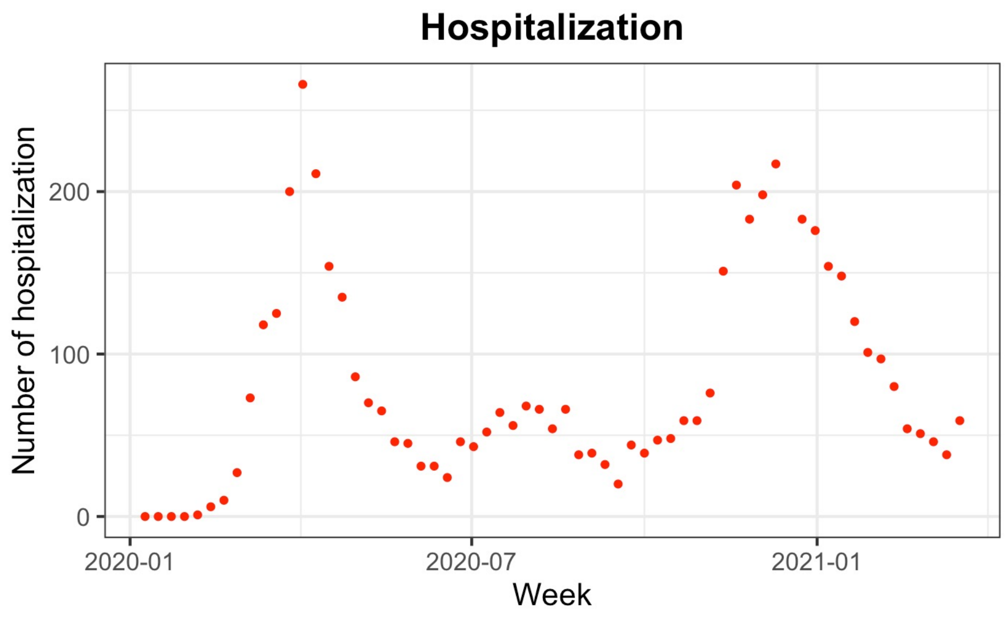
Multiple Regression Model



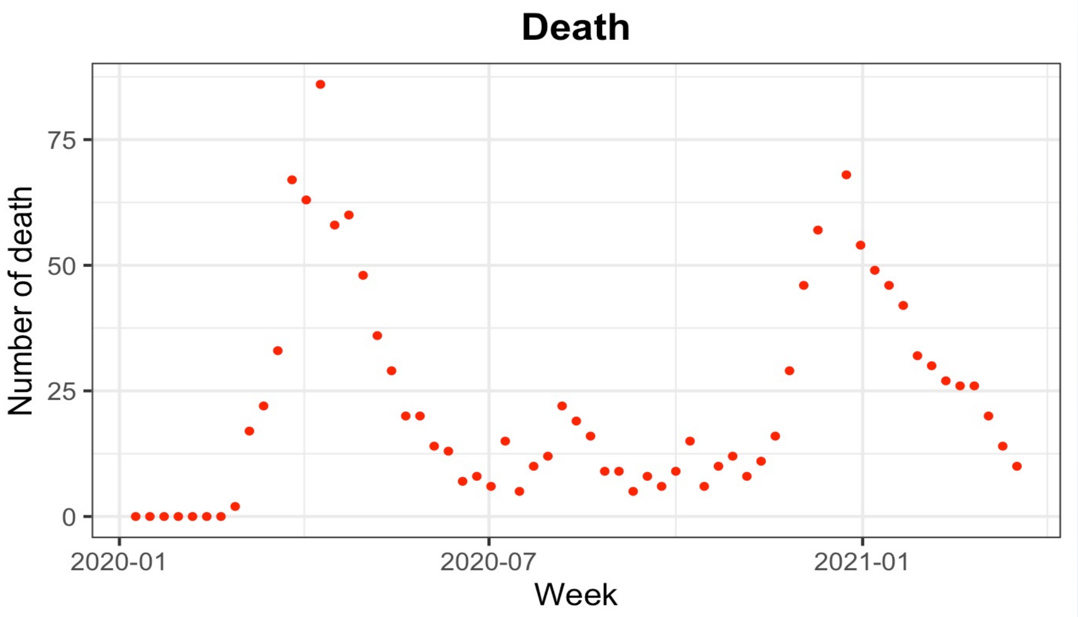
Multi Regression(tested+hospitalization)



Multi Regression(tested+hospitalization+death)



Hospitalization



Death

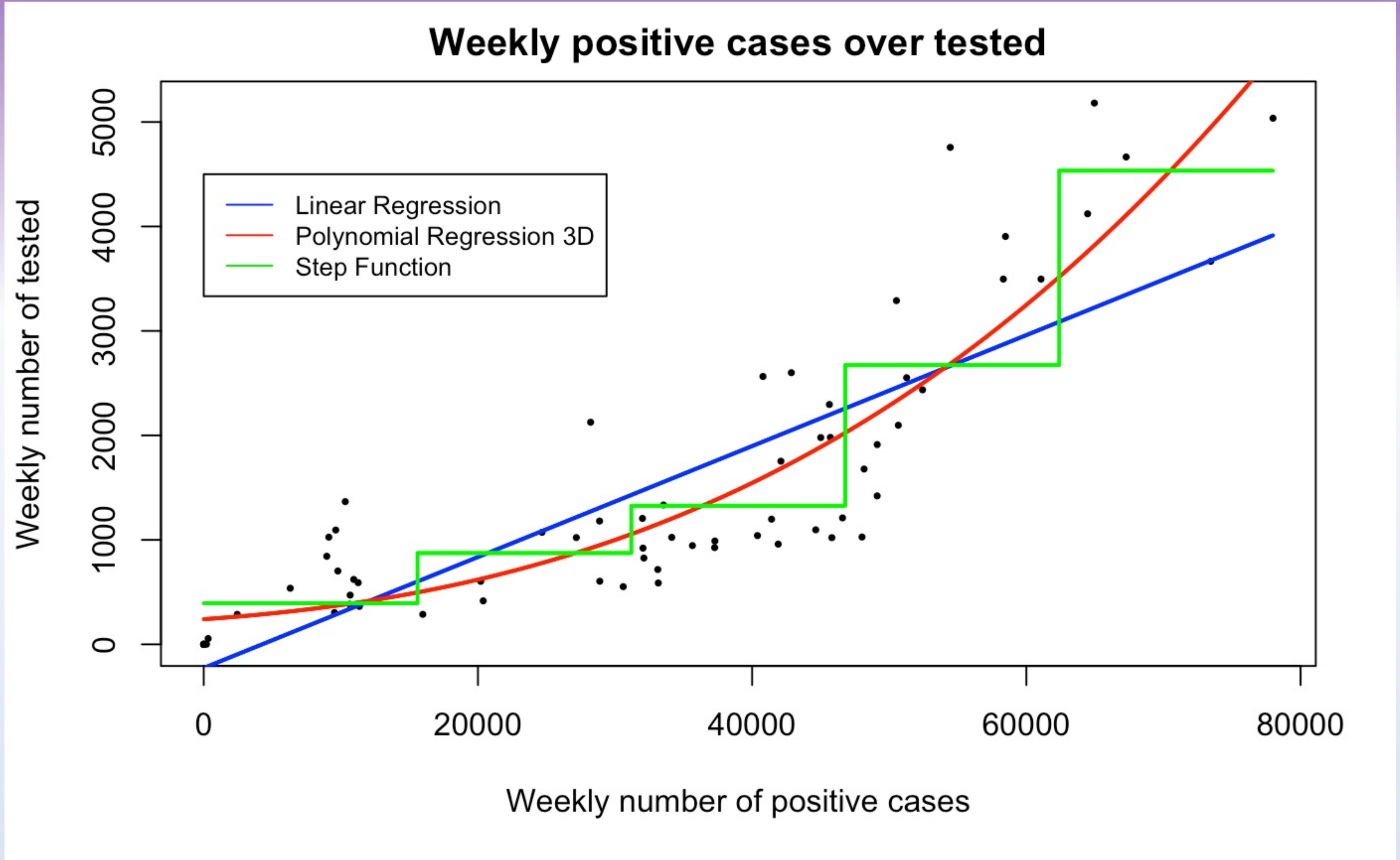
Cross comparison

Step Function

- Avoid global structure
- Break range of X into bins
- Fit different constant in each bin
- Continuous variable to ordered categorical

Piecewise Function

- Fit separate low-degree function over different regions of X
- Knots: Point of coefficient change
- More knots, more flexible



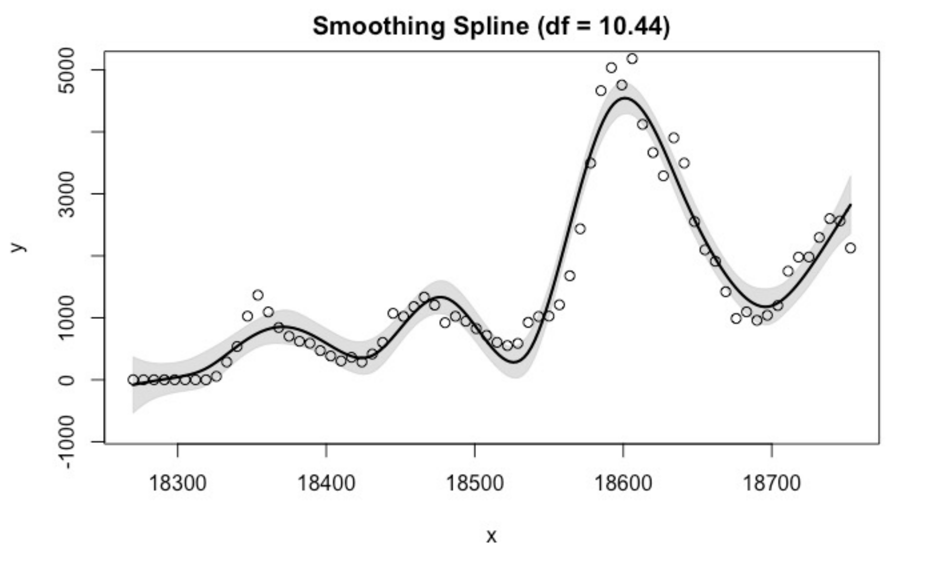
Splines

Overview of Spline

- Divide the predictor variable into sections
- Fit separate model in each section
- Constraint: key difference between type of spline model

Smoothing Splines

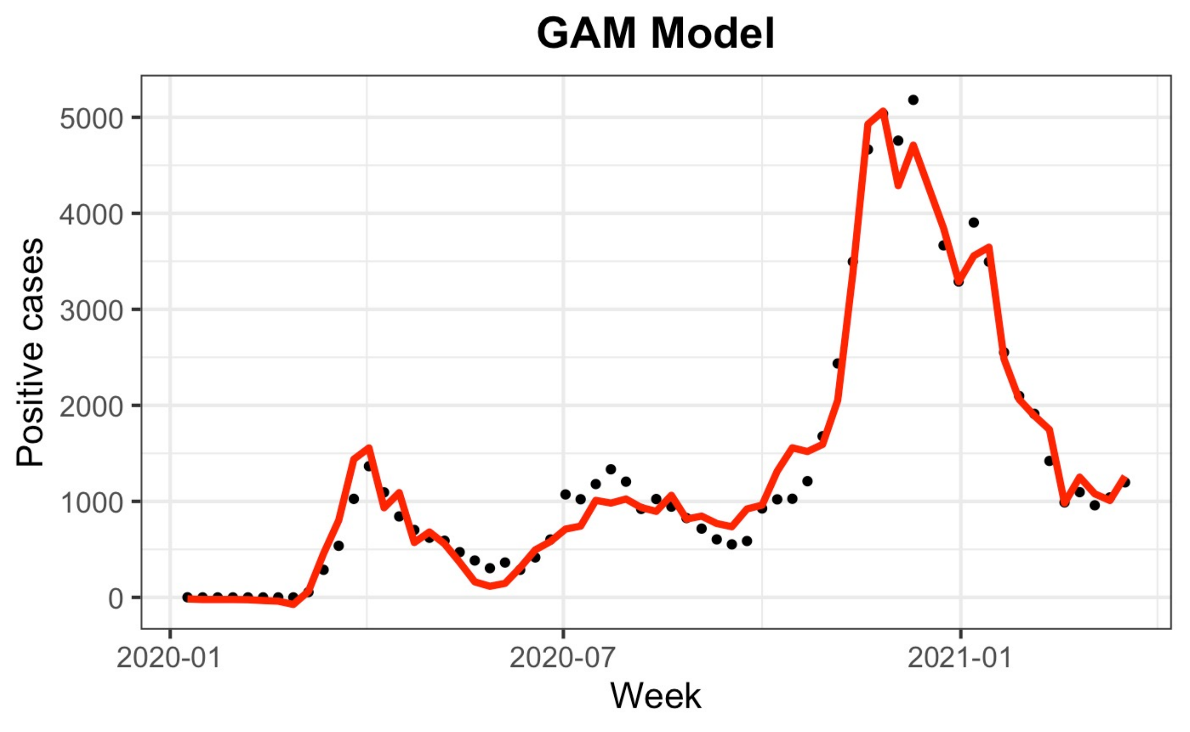
- Pick many knots
- Penalized the roughness(2nd derivative)

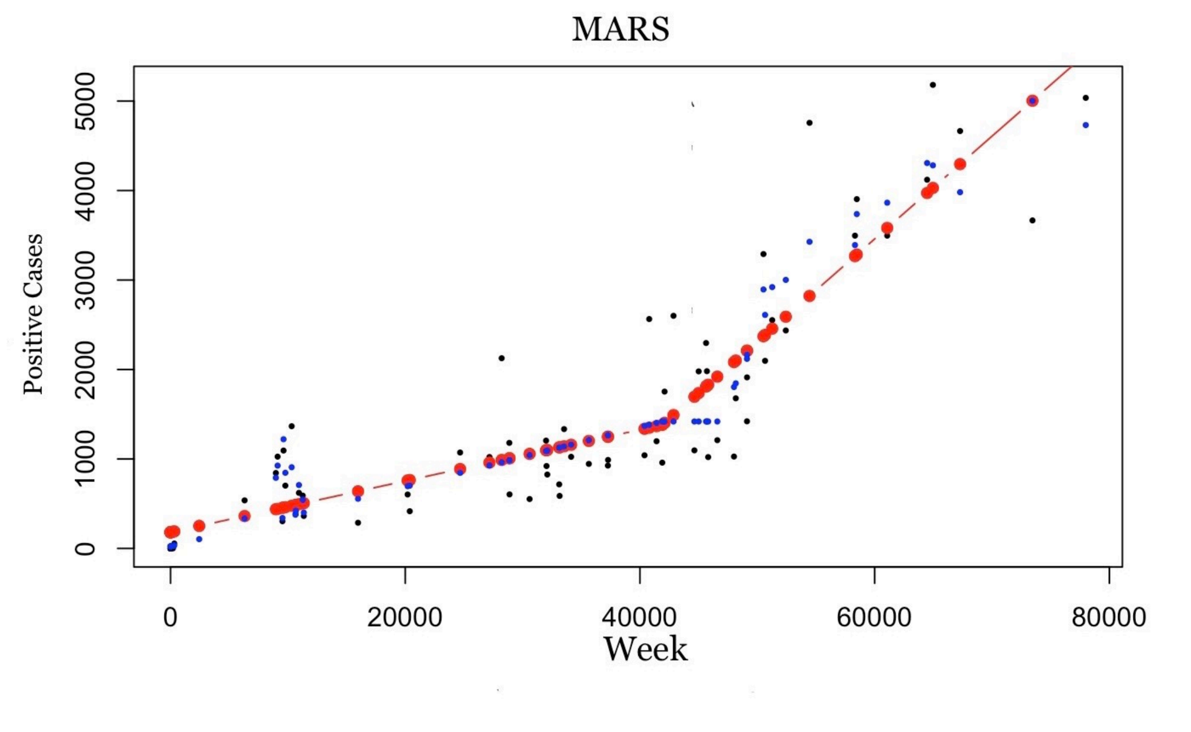


Generalized Addictive Model (GAM) and MARS

- Framework derived from linear model
- Maintain additivity
- Allow to examine each predictor effect
- Could miss important interaction

- Use forward and backward pass
- Eliminate unnecessary functions by GCV
- Repeat until reach predefined term
- Produce optimal fit





Choosing a model

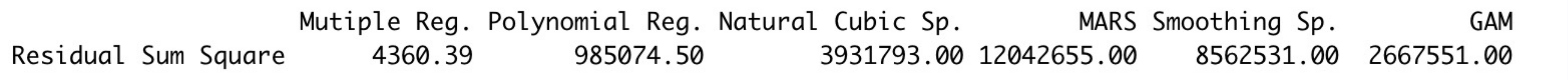
- **Assess the accuracy of the model**
- RSS =

$$RSS = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$$

- RSE =

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Results:



```
                    Mutiple Reg. Polynomial Reg. Natural Cubic Sp.        MARS Smoothing Sp.         GAM
Residual Sum Square      4360.39       985074.50        3931793.00 12042655.00      8562531.00  2667551.00
```

Question?

THANK YOU!!!!