# **Covid case analysis report**

# **1 Introduction**

The outbreak of COVID-19 has caused huge shocks and challenges to the world. Governments and health institutions are under pressure to develop effective prevention and control measures, and accurate data and scientific analysis are needed to support decision-making [1].

In this context, the UQ4Covid project came into being. The project is dedicated to integrating and analyzing various COVID-19-related data to provide real-time and accurate information and predictive models for decision makers to help them make informed decisions.

By modeling and predicting case data, we can provide important insights about the COVID-19 pandemic. First, we expect to be able to predict the spread trend of the epidemic, understand the rate of increase in the number of cases and possible high-risk areas. These predictions will serve as the basis for governments and health agencies to develop corresponding prevention and control strategies, optimize resource allocation, and take timely measures to slow the spread of the epidemic [2].

The main purpose of this paper is to conduct an in-depth analysis of COVID-19 case data, especially to predict the patients and deaths under different demand conditions. We hope that through these analyses, we can reveal the reasons for the impact of certain dimensions on case deaths.

To do this, we will collect and integrate data on case data, demographics, health care resources, and other relevant factors by region. Then, we will use statistical and data modeling methods to build predictive models by analyzing the correlations and trends between these data.

Using these predictive models, we can explore the impact of different factors on the spread and fatality rate of the outbreak. For example, we can study the impact of climate factors on the number of cases and mortality, analyze the impact of population density and mobility on the speed of epidemic transmission, and the impact of medical resources and medical care level on mortality [3-9].

In conducting these analyses, we will apply a scientifically rigorous approach and pay great attention to the accuracy and credibility of the data. We will consider a variety of potential biases and errors and use a variety of correction and validation methods to ensure the stability and reliability of our models.

Ultimately, we hope that this study can provide useful information and insights for policy makers to help them formulate more scientific and effective prevention and control strategies. We believe that with accurate data and scientific analysis, we can jointly fight against the epidemic and safeguard the health and safety of the people.

# **2 Literature Review**

# **2.1 Gaussian process Modeling**

A Gaussian Process is a probabilistic model commonly used in regression and classification problems. It is based on the idea of Bayesian inference, and by modeling the data, it can provide an output value for each input point and provide confidence in the prediction result.

Here's how Gaussian process models work in a basic way:

Suppose you have a set of input data points X and corresponding output data points Y. The input data points can be vectors of any dimension. Gaussian process models assume that Y is generated by adding a noise term ε to the values of an unknown function f at each input point: Y = f(X) + ε.

Gaussian process models make a non-parametric assumption on the function f, consider f as a random process, and describe the statistical properties of f at each input point. Specifically, Gaussian process models assume that the joint distribution of any finite number of output values follows a multivariate Gaussian distribution.

For a given input data point X, the Gaussian process model constructs a prior distribution over X. This prior distribution can be thought of as a kind of infinite dimensional probability distribution over the function.

After observing some of the output data points Y, the Gaussian process model can use Bayes' theorem to update the prior distribution to obtain the posterior distribution. The posterior distribution is a conditional distribution over a function f that gives the probability distribution over the value of the function at each input point given the observed data.

Using the posterior distribution of the observed data points, a Gaussian process model can make a prediction. For a new input point x, we can calculate the conditional distribution of its corresponding output value, given the observed data. Common predictors include the mean and variance, which represent the expected value and the uncertainty measure of the predicted output, respectively.

To summarize, Gaussian process models provide a flexible and resilient way to deal with regression and classification problems by modeling the prior distribution of a function and the posterior distribution of the observed data. It can not only make predictions, but also provide confidence estimates of the prediction results, making the model output more interpretable and reliable.

# **2.2 Gaussian process regression**

# Gaussian Process regression (GPR) is a nonparametric regression method based on Gaussian process, which is widely used in modeling and prediction tasks in many fields. Compared with traditional parametric regression methods, GPR does not need to set the model structure when modeling the relationship between input and output, which has greater flexibility and adaptability.

# In Gaussian process regression, we assume that the observed data follow an unknown Gaussian process. A Gaussian process can be viewed as a random function, which can be used to describe the underlying relationship between the input and output. Specifically, for each input value, a Gaussian process can be used to obtain a probability distribution over the output, rather than a deterministic value. This probability distribution can be used to represent the uncertainty about the observed data.

# The core idea of Gaussian process regression is to infer the underlying Gaussian process from the observed data points and build a regression model based on it. In the training phase, we first collect a set of input-output training data. Then, using these training data, the prior distribution of the Gaussian process is established by calculating some statistics such as covariance matrix and mean vector.

# In the prediction stage, given a new input value, we can use the trained model to predict the corresponding output and obtain a probability distribution of the output. This probability distribution can provide an estimate of the confidence of the predicted result, that is, we can judge the degree of uncertainty of the predicted result by the width of the probability distribution. Generally, the wider the width of the probability distribution, the higher the uncertainty of the prediction result.

# Gaussian process regression has a variety of kernel functions to choose from, such as linear kernel, polynomial kernel, Radial basis function (RBF) kernel, and so on. The choice of kernel function will have an impact on the fitting ability of the model: linear kernel is suitable for linear relationships, polynomial kernel is suitable for polynomial relationships, and RBF kernel is suitable for nonlinear relationships. In addition, complex kernel functions can be constructed to better fit the data by using combinations of multiple kernel functions.

# In addition to flexibility and adaptability, Gaussian process regression has some other advantages. First, Gaussian process regression can handle noisy data because it models the predictions as a probability distribution. Second, Gaussian process regression can adapt to different data distributions and nonlinear relationships because it does not rely on specific assumptions and model structures. Finally, Gaussian process regression can also perform effective sequential prediction, as it can update and predict progressively based on the order of the observed data.

# However, there are some limitations and challenges associated with Gaussian process regression. Firstly, the computational complexity of Gaussian process regression is high, especially for large-scale datasets. In order to solve this problem, some approximation methods can be used, such as variational inference, approximation of kernel methods and sparsity based methods. Secondly, Gaussian process regression is sensitive to the dimensionality of the input space. When the input space has high dimensionality, it may face the problem of curse of dimensionality. To solve this problem, feature selection and dimensionality reduction methods can be used to deal with it. Finally, Gaussian process regression is sensitive to the selection of hyperparameters, which needs to be tuned by methods such as cross-validation.

# In the prediction of disease spread, Gaussian process regression can be applied to analyze and predict the trend and law of disease spread. By building a Gaussian process regression model, we can use the existing case data to infer the future number of cases and make predictions. At the same time, as Gaussian process regression provides the probability distribution of the prediction results, we can also evaluate the confidence and uncertainty of the prediction results [11].

# Besides Gaussian process regression, there are other forecasting models and methods to choose from. For example, support vector Regression (SVR) is a commonly used nonparametric regression method that can be used to model nonlinear relationships. Decision Tree Regression is a tree-based regression method that can be used to model data with multiple branches and nonlinear relationships. Deep learning models, such as neural networks, can also be applied to prediction problems, with strong fitting and expressive power.

# In summary, Gaussian process regression is a powerful nonparametric regression method with the advantages of flexibility, adaptability, and confidence estimation. In the prediction of disease spread, it can be used to model and predict the spread trend of the disease, providing uncertainty estimates of the prediction results. However, for complex data sets and problems, we can also consider other prediction models and methods to improve the accuracy and reliability of the prediction. With continuous research and efforts, we believe we can overcome the challenges in disease spread prediction and make positive contributions to social stability and people's well-being [12].

# **2.3 Residual plots**

Residual plots are one of the most important tools for evaluating the fit of a statistical model. In regression analysis, we fit a model to predict the value of the dependent variable, and the residuals represent the difference between the observed values and the values predicted by the model.

A residual plot shows how well the model fits the variable. A good model fit should have the following characteristics: 1) the mean of the residuals is close to zero, which means that the model fits the data well at the overall level; 2) the variance of the residuals should be constant, that is, stable across the range of all independent variables, with no obvious regular changes; 3) the residuals should be randomly distributed around zero with no obvious clustering or trend.

There may be a problem with the model when a residual plot shows: 1) the mean of the residuals deviates significantly from zero, indicating that the model as a whole is underestimating or overestimating the observations; 2) the variances of the residuals are unstable and may be heteroscedasticity, which may mean that the model does not fit the data well across regions; 3) the residuals exhibit non-random clusters or trends, suggesting that the model fails to capture some important features of the data, leading to systematic bias.

In addition to visually demonstrating the fit of a model, residual plots can also be used to test whether a model's assumptions hold. For example, if the residual plot shows significant nonlinear relationships or heteroscedasticity, it may be suggestive of unmodeled nonlinear relationships or other unaccounted for factors in the model.

# **3 Implementation**

The outbreak of COVID-19 has caused great shocks and challenges around the world. Governments and health agencies are under pressure to develop effective prevention and control measures, requiring accurate data and scientific analysis to support decision-making [13].

This paper is based on the dataset provided by the UQ4Covid project and uses linear models and Gaussian regression modeling to make regression predictions on the data. Our goal is to reveal the impact of different classifications on outcomes by making predictions from different classifications of the dataset (region) and case type (LAD).

To make regression predictions, we need a dataset as input to our model. The UQ4Covid project provided multiple categorical variables including region and case type (LAD). These variables can help us better understand the differences between different regions and different case types, and their impact on the spread of the outbreak.

Next, we'll use linear models and Gaussian regression models to model and predict this data. First of all, we will carry out data pre-processing and cleaning work to ensure the accuracy and consistency of the data. We will then select the appropriate regression model and use statistical methods to assess the fit and predictive power of the model. With these steps, we can get relatively accurate predictions and explore the impact of different classifications on the results.

When making regression predictions, we will pay special attention to the categorical variables of space (region) and case type (LAD). By modeling and forecasting data from different regions, we can understand the differences between different regions and identify key factors affecting the spread of the pandemic. At the same time, by modeling and predicting data from different case types, we can study the extent to which different case types affect the spread of the epidemic, as well as possible prevention and control strategies.

When making regression predictions, we will pay great attention to the quality and confidence of the data. We will carefully examine the source of the data and utilize appropriate analysis methods to deal with potential errors and biases. We will also use techniques for cross-validation and model evaluation to ensure model stability and reliability.

To summarize, this paper is based on the dataset provided by the UQ4Covid project and uses Gaussian regression modeling for regression prediction. By predicting datasets from different classifications of spatial (region) and case type (LAD), we hope to reveal the impact of different classifications on outcomes. Through this analysis and forecasting, we can provide policymakers with more accurate information and insights to support them in developing scientific and effective prevention and control strategies to jointly fight the COVID-19 epidemic [14-15].

# **4 Data overview and processing**

# **4.1.Data overview**

Input dataset Introduction:

|  |  |  |
| --- | --- | --- |
| R0 | Reproduction number | ,4.5 (2) |
| TE | Mean latent period | (0.1, 2) |
| TP | Mean pre-symptomatic infectious period | (1.2, 3) |
| TI1 | Mean infectious and symptomatic | (2.8, 4.5) |
| TI2 | Mean infectious and symptomatic | (0.0001, 0.5) |
| nuA | Numbers of infectious but asymptomatic; | (0, 1) |
| ns | Number of seeds | \ |
| p\_home\_weekend | \ | (0, 1) |
| alphaTH | parameter relating to probability of death given hospitalisation | \ |
| etaTH | hospitalised | \ |
| alphaEP | parameter relating to probability of death given hospitalisation | \ |
| alphaI1D | \ | \ |
| eta | \ | \ |

Table 4.1.1 Introduction to the input dataset (incomplete)

Data processing plays a crucial role in the process of building the model. Before using the dataset provided by the UQ4Covid project for Gaussian regression modeling, we need to preprocess and clean the data to ensure its accuracy and consistency.



Table 4.1.2 Overview of input data (first five rows)



Table 4.1.3 Overview of output

First, note that there are repeats in the input data set. To avoid duplicate samples affecting the model, we need to remove these duplicates from the dataset. This can be done by checking the unique identifiers or other relevant fields in the dataset. After removing these duplicate samples, we can ensure that each sample is unique and does not create conflicts for the model's training and prediction.

For the output data set, we split the data set using the mean. Specifically, we split the output dataset into a training set and a prediction set, where 70% of the data is used to train the model, and the remaining 30% is used to validate the predictive ability of the model. This splitting ensures that the model has enough data to learn and validate during both training and prediction.

Here's the final use of the later training dataset with the same region, LAD, etc. :

Table 4.1.4 Overview of train\_used

In the data preview, we analyzed the relationship between R0 and Death. R0 represents the base number of virus infections, and Deaths represents the log number of deaths. We plotted R0 versus Death to see how they were distributed.

Looking at the scatter plot, we see that there is no significant heteroscedasticity in the distribution of the data points. Heteroscedasticity refers to the situation where the variance of the data varies significantly under different value ranges. If there is heteroscedasticity, it may affect the fitting effect and prediction accuracy of the model. However, in our dataset, we did not observe this, which provides a positive sign for our modeling efforts.

In the following work, we will conduct Gaussian regression modeling based on these preprocessed data and use statistical methods to evaluate the fit and predictive power of the model. We will continue to explore the relationships between other variables and use appropriate feature selection and model optimization techniques to improve the performance of the model.

To summarize, data processing is critical to model building. Before using the dataset provided by the UQ4Covid project for Gaussian regression modeling, we need to perform data preprocessing and cleaning, including removing duplicate samples and splitting the dataset. Through the analysis of the relationship between R0 and Death, we observed that the distribution of the data points was not significantly heteroscedastic. In the following work, we will conduct modeling and prediction based on the preprocessed data, and further optimize the model to improve its performance and accuracy.

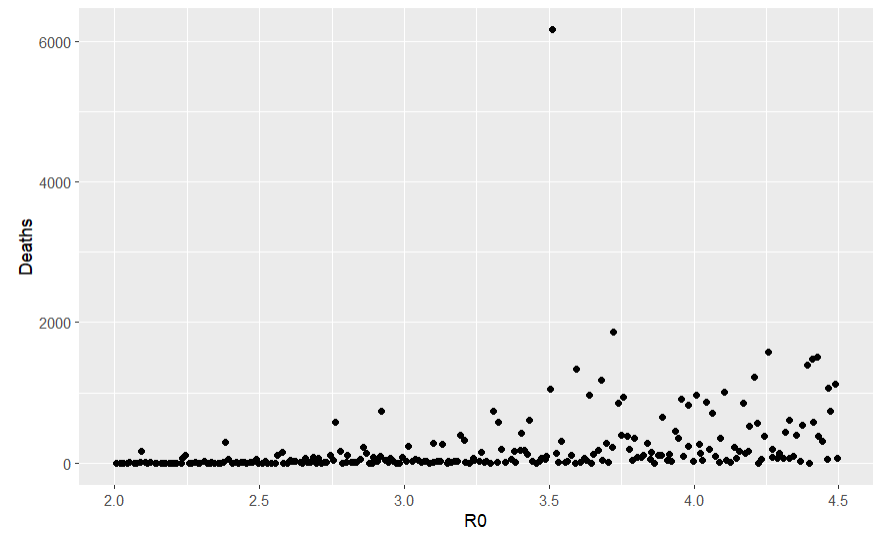


Figure 4.1.1 Scatter plot of deaths-R0

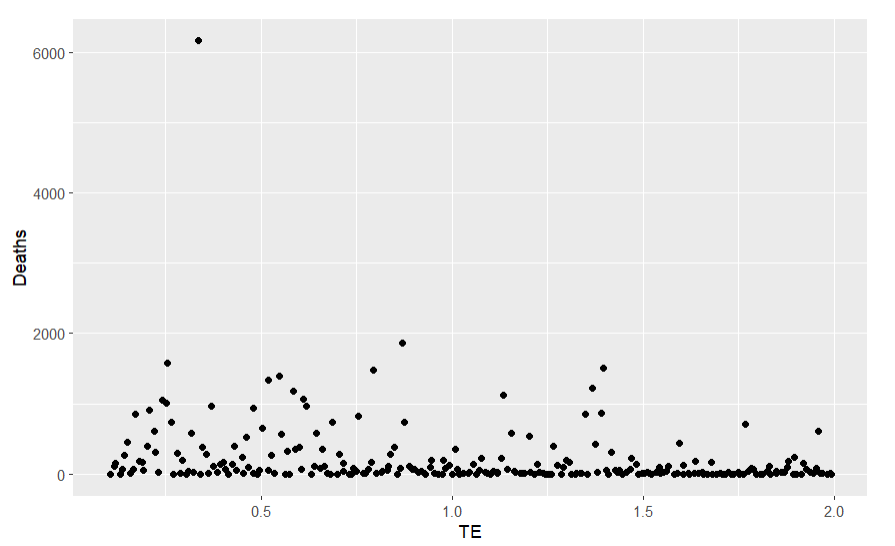


Figure 4.1.2 deaths-TE scatter plot

Looking at the distribution of data points is crucial to understanding the characteristics and nature of the data. In this problem, we observe and analyze the normality of the Deaths data point. The normal distribution plays an important role in statistics because many statistical inferences are based on the assumption that the data follow a normal distribution.

However, we also notice some deviations from the normal distribution at the tail end of the data. Deviations can include a thick tail, a long tail, or a small tail. This deviation may be due to special circumstances or outliers that cause the data to deviate from normality for certain ranges of values. This may have a certain impact on the establishment of the model and the inference results.

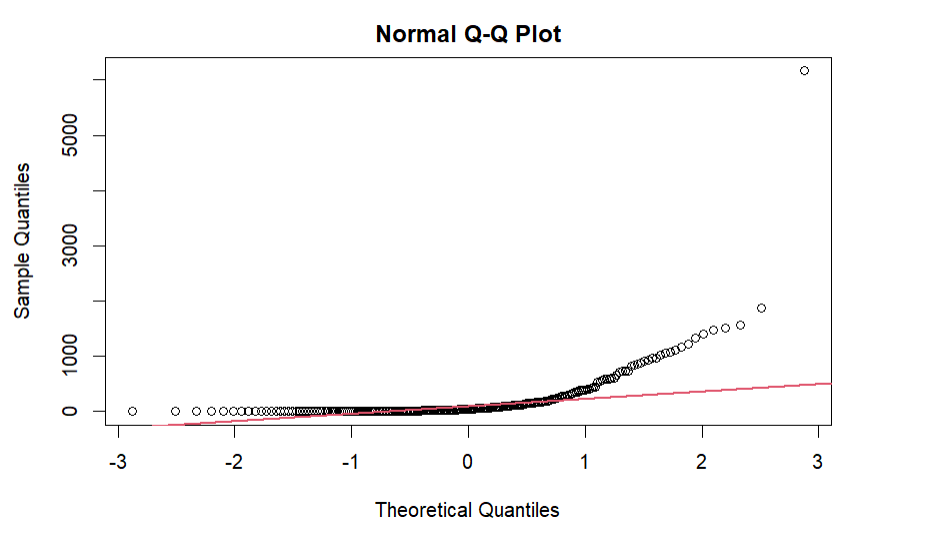


Figure 4.1.3 Normal QQ plot

# **5 Model construction and prediction**

**5.1Linear regression Modeling and Gaussian Process Modeling**:

5.1.1 Overview of linear models

A linear regression model is a common statistical model used to establish a linear relationship between an independent variable (or predictor variable) and a dependent variable. It assumes that the relationship between the independent variable and the dependent variable can be described by a linear equation.

The general form of a linear regression model can be expressed as follows.

Y = β0 + β1X1 + β2X2 +... + βn\*Xn + ε

Where Y is the dependent variable, X1, X2,... , Xn are independent variables, β0, β1, β2,... , βn are the coefficients of the model and ε is the error term.

The goal of a linear regression model is to find the best coefficient estimates that give the model the best fit to the observed data. The coefficients are usually estimated using the least squares method, which determines the best coefficient estimates by minimizing the squared difference between the observed values and the values predicted by the model.

Linear regression models have a wide range of applications and are often used to predict and explain changes in the dependent variable. It can be used to predict a continuous dependent variable, to infer causality, and to analyze associations between variables. Linear regression models can also be extended by introducing nonlinear terms, interaction terms, etc., to accommodate more complex data patterns.

It should be noted that the applicability and accuracy of linear regression models rely on some assumptions, such as linear relationship, independence, homoscedasticity, etc. Before applying a linear regression model, it is necessary to properly test and preprocess the data to ensure the validity and reliability of the model.

We fit a linear regression model for Deaths on the remaining variables to this data:

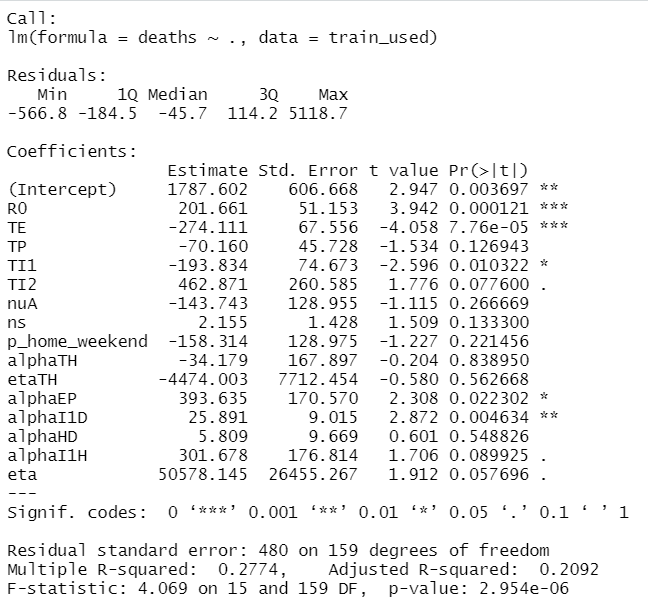


Figure 5.1.1.1 Linear regression fit results

You can see that in the pure linear regression model, the fit of the data is poor, with an adjusted R-square of 0.2092. Because there is no heteroscedasticity in the data, the effect of weighted linear regression is small. In the case of generalized linear regression, you can see that the likelihood ratio test also fails.

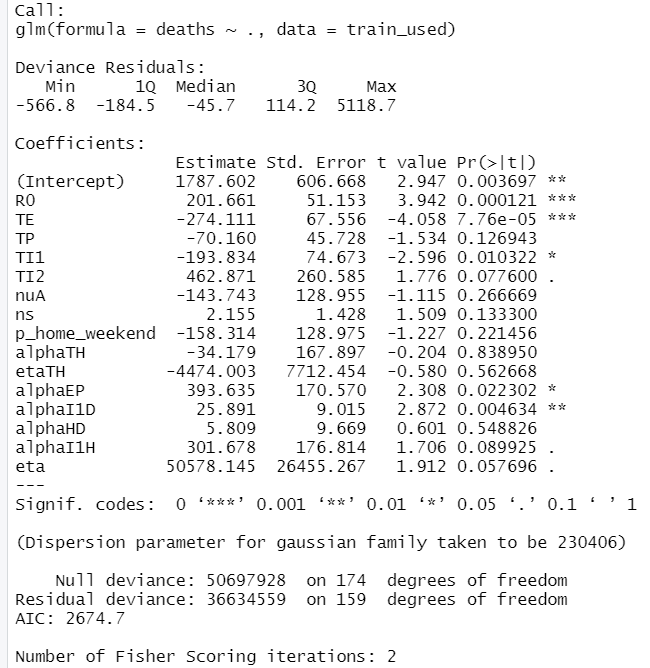


Figure 5.1.1.2 Generalized linear regression results

5.1.2 Gaussian process models

Gaussian Process Model is a non-parametric statistical model based on probability theory, which is suitable for modeling the dependence and uncertainty of data. When constructing the Gaussian process model, we assume that the observed data obey the multivariate normal distribution, and use the kernel function to describe the correlation between the data.

First, we need to choose an appropriate kernel function to represent the correlation between the data. Common kernel functions include linear kernel functions, polynomial kernel functions, Gaussian kernel functions, etc. The choice of kernel function should be made according to the characteristics of the data and the needs of the problem. For example, if the data varies smoothly, a Gaussian kernel may be a good choice, while a polynomial kernel may be more appropriate for data with nonlinear relationships.

By using the gausspr function in R language, we can conveniently construct and fit the Gaussian process model. This function accepts independent and dependent variables as input and fits the data by maximum likelihood estimation based on the given kernel function parameters. During the fitting process, the function will infer the relationship between the data based on the data points in the training set and generate the parameters of the model.

After building a Gaussian process model, we can use it to make predictions. Given new independent variables, the model generates a probability distribution, which is the predicted range of values for the dependent variable. This allows us to make predictions on unknown data and estimate their uncertainty. The mean of the probability distribution can be used as the predicted value, while the variance represents the uncertainty of the prediction.

In addition to its predictive power, Gaussian process models can provide other useful information. For example, we can calculate the marginal likelihood function to assess how well the model fits. The marginal likelihood function represents the probability density function of the observed data given the observed data and the model parameters. By maximizing the marginal likelihood function, we can choose the model that best fits the data.

In addition, we can test the goodness of fit of the model by calculating the residuals. By analyzing the distribution and characteristics of the residuals, which are the differences between the observed values and the values predicted by the model, we can assess whether the model is doing a good job of explaining the variability in the data.

Tuning the hyperparameters of the model is also an important way to optimize the performance of Gaussian process models. Hyperparameters include the parameters of kernel functions and other model-related parameters. By adjusting the values of the hyperparameters, we can change the complexity of the model and the degree of fit, thereby improving the prediction accuracy of the model.

In summary, Gaussian process models are a powerful modeling tool that can capture both the relationship and the uncertainty between data. By utilizing Gaussian process models, we can make predictions about the data and evaluate the fit and performance of the model. This technique is widely used in many fields, such as financial prediction, weather prediction, medical diagnosis, etc. However, when using Gaussian process model, we also need to pay attention to the selection of appropriate kernel function and hyperparameters, and reasonable interpretation and verification to ensure the reliability and effectiveness of the model.

5.1.2.Model fitting

For the overall Gaussian random process model, we can see how well the model fits from the residual plot:

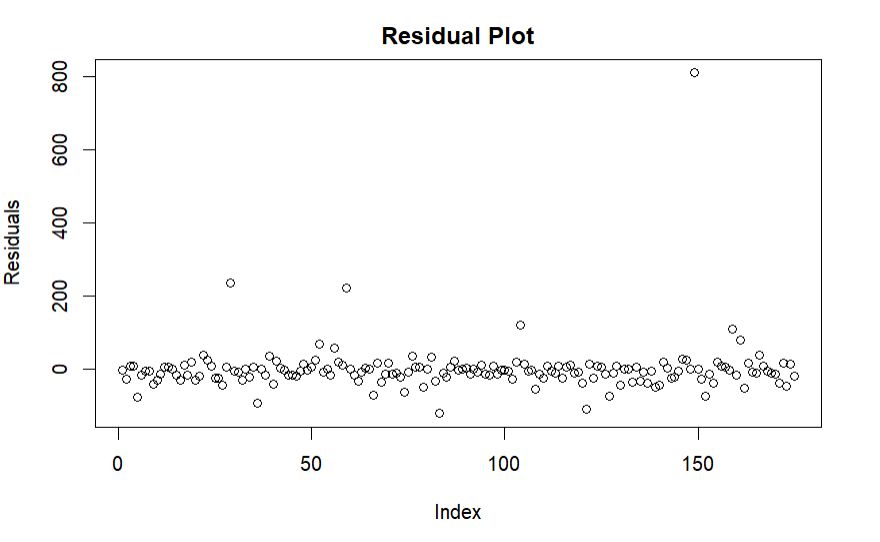


Figure 5.1.2 Residual plot of Gaussian process modeling for deaths prediction

As you can see from the above figure, except for some significant outliers, the residual data points are uniformly distributed around 0, and the model fits well. This means that the model is fairly good at explaining the variability in the observed data and that there is no significant bias or model error.

However, we also need to evaluate how well the model predicts on the test set. We use the model to predict the value of the dependent variable on a given test dataset and calculate the difference between the predicted value and the true value. The Root Mean Squared Error (RMSE) and Coefficient of Determination (R-²) are commonly used to assess the accuracy and explanatory power of predictions.

For the test set prediction, the RMSE of the test set prediction effect is 326 and the R-squared is 0.36. RMSE measures the average difference between the predicted value and the true value, with lower values indicating more accurate predictions. The r-squared is a measure of how well the model explains variation in the data and ranges from 0 to 1, with values closer to 1 indicating better performance.

For the test set, an RMSE of 326 indicates that the average difference between the predicted value and the true value is large, which means that the prediction is not ideal. At the same time, an r-squared of 0.36 indicates that the model explains only 36% of the variability in the dependent variable, leaving much of the variability unexplained.

Such predictions may suggest some limitations or inadequacies in the model. Possible reasons for this are:

Insufficient data features: The model needs sufficient data features to capture the inherent regularities of the data. If the features in the test set are significantly different from the features in the training set, the model's prediction performance will decrease.

Inappropriate model Complexity: The complexity of Gaussian process models is determined by the parameters of the kernel function and other hyperparameters. If the model is too simple, it may not capture complex relationships in the data. On the other hand, if the model is too complex, it may easily overfit the training data, resulting in poor prediction performance on the test set.

Impact of outliers: Outliers are observations that are significantly different from other data points and usually have a large impact on the fitting effect of the model in data analysis. If there are many outliers in the test set, the predictive accuracy of the model will decrease.

To address these issues, we can take the following steps to improve our model's predictions:

Data preprocessing: Data preprocessing, including feature selection, dimensionality reduction, standardization, etc. is performed to improve the input quality and stability of the model. At the same time, the detection and correction of outliers can reduce the interference of outliers on the model.

Hyperparameter tuning: By adjusting the parameters of the kernel function and other hyperparameters, the most appropriate model complexity is found to balance the fitting ability and generalization ability. Techniques such as cross-validation can be employed to evaluate the performance of different parameter combinations and select the optimal parameter configuration.

Model ensemble: We can try to ensemble multiple Gaussian process models, and use ensemble learning methods to improve the robustness and prediction accuracy of the model. Common ensemble methods include Bagging, Boosting, and so on.

More domain knowledge is introduced: According to the characteristics of the specific domain, more relevant domain knowledge is introduced to assist modeling. For example, the results of the model can be modified or adjusted by combining the experience and prior knowledge of domain experts, so as to improve the prediction effect.

In summary, to solve the problem of unsatisfactory prediction results on the test set, we can improve the performance of the model by data preprocessing, hyperparameter tuning, model integration and introducing domain knowledge. At the same time, an in-depth analysis of the limitations and shortcomings of the model is also needed to better understand the data and provide more accurate predictions.

**5.2 Modeling predictions for region resolution**

The training set is divided into a set according to the region variable, and the cyclic modeling and prediction are performed to obtain the evaluation of each model.

See the following figure:

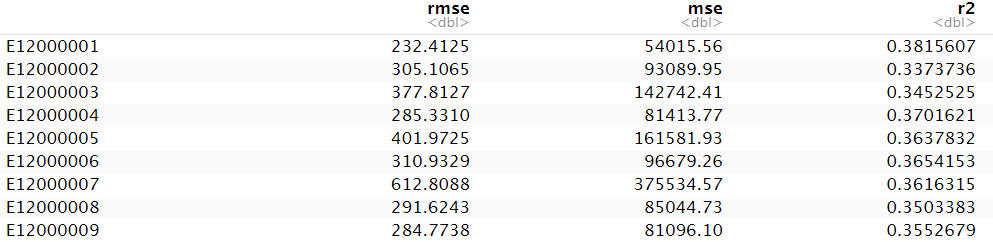


Figure 5.2.1 Showing the prediction results for different regions

Given the data, we can further analyze the prediction performance when the first column shows the number of regions selected. We observe that E120000001 is the region with the best prediction performance, with the smallest RMSE of 232.4, but even then, the prediction is still not ideal.

First, let's recall what the R-squared (coefficient of determination) is: it measures how well a model explains the variation in the data and ranges from 0 to 1, with values closer to 1 indicating better explanatory power. Therefore, an r-squared of 0.38 means that the model explains only 38% of the variation in the dependent variable, leaving most of the variability unexplained.

To better understand the unsatisfactory predictions, let's consider some possible causes and ways to improve them.

First of all, one possible reason is the inadequacy of data features. Models need sufficient data features to capture the inherent regularities of the data. If the features in the test set are significantly different from the features in the training set, or if the features in the test set do not provide enough information to predict the dependent variable, the model will be less predictive. Therefore, we need to carefully review the data set to ensure the quality and adequacy of the features, and try to use other methods to fill in the missing features in the test set or find alternative features.

Second, the complexity of the model may not be appropriate. The complexity of such Gaussian process models is determined by the parameters of the kernel function and other hyperparameters. If the model is too simple, it may fail to capture the complex relationships in the data, resulting in poor predictions. Conversely, if the model is too complex, it may overfit on the training set and thus perform poorly on the test set. Therefore, we need to carefully tune the hyperparameters of the model and find the best model complexity to balance the fitting ability and generalization ability.

Another possible reason is the effect of outliers. Outliers are observations that are significantly different from other data points. In data analysis, outliers usually have a large impact on the fit of the model. If there are many outliers in the test set, the prediction accuracy of the model will decrease. To solve this problem, we can either detect and correct for outliers or choose to use a more robust model to reduce the impact of outliers.

There are a number of steps we can take to improve our model's predictions. First of all, data preprocessing is very important. We can perform feature selection, dimensionality reduction and normalization on the data to improve the input quality and stability of the model. At the same time, for the case where there are outliers, we can use the outlier detection method to identify them and take corresponding treatment measures.

Secondly, hyperparameter tuning is a very critical step. By adjusting the parameters of the kernel function and other hyperparameters, we can find the model complexity that best fits the data, thus improving the fitting ability and generalization ability of the model. Common approaches include grid search, random search, and Bayesian optimization. Through these methods, we can evaluate the performance of different parameter combinations and select the optimal parameter configuration [16].

In addition, model ensemble is also an effective strategy. Ensemble learning can improve the overall prediction accuracy and stability by combining the prediction results of multiple models. Common ensemble methods include Bagging, Boosting, and so on. By integrating multiple models, we can make full use of their respective advantages and compensate for the shortcomings of a single model.

Finally, introducing more domain knowledge is also a way to improve the prediction effect. According to the characteristics of the specific domain, we can use the experience and prior knowledge of domain experts to assist the modeling. This domain knowledge can help us better understand the background of the data and the problem, so as to propose more accurate hypotheses or adjust the results of the model.

In conclusion, when analyzing poor prediction results of a model, we should comprehensively consider factors such as data characteristics, model complexity, outliers, and domain knowledge, and take corresponding measures to improve the prediction results. Through methods such as data preprocessing, hyperparameter tuning, model integration, and the introduction of domain knowledge, we can gradually improve the performance of the model and make it better adapt to actual application scenarios. Through continuous adjustment and optimization, we can find the most suitable model configuration and obtain more accurate prediction results.

**5.3.Model and predict the classification of the metrics**

The training set was divided into sets according to the LAD19CD variable, and the cyclic modeling and prediction were performed to obtain the evaluation of each model.

The best output set is output:



Figure 5.3.1 Prediction results of different metrics for LAD19CD

Based on the given data, we can further analyze the modeled variable importance and correlation. It is observed that the RMSE is around 216, which is a certain improvement compared with the previous prediction results, indicating that the selected variables are effective for the prediction target.

**5.4.Model Optimization**

Now let's dive into the importance and correlation of variables. To better understand these concepts, let's first review their definitions:

Variable importance: Variable importance is a measure of how much a variable contributes to the predictive power of a model. In the modeling process, we often encounter a large number of feature variables, but not all of them have a significant impact on the predictive power of the model. Variable importance can help us to determine which variables are more critical to the prediction target, so that we can select features or optimize the model.

Correlation: Correlation is a measure of the strength of the linear relationship between two variables. Correlation can help us determine whether two variables have some kind of connection or dependence and how much they are related to each other. In the modeling process, we often want to select variables that are highly correlated with the target variable, as they are likely to have a large impact on the predictive power of the model.

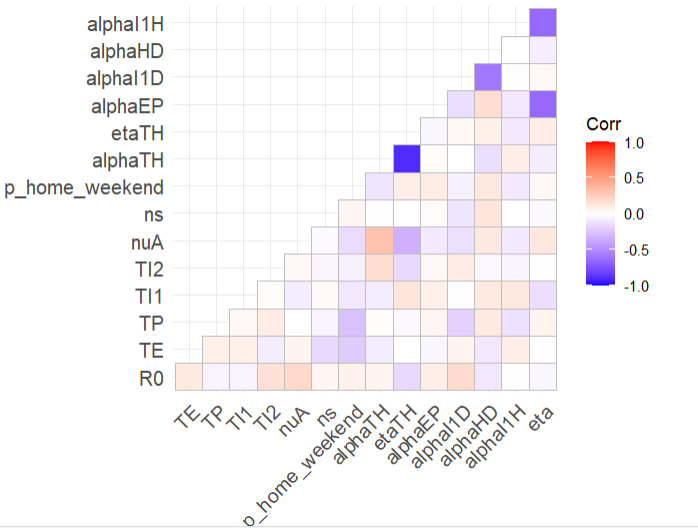


Figure 5.4.1 Correlation heatmap among dimensions of the data

Next, let's look at the correlations between the variables. Correlation can be assessed by calculating the correlation coefficient between variables.Commonly used correlation coefficients include the Pearson correlation coefficient and the Spearman correlation coefficient.

The Pearson correlation coefficient is applied to variables that have a linear relationship, and it ranges from -1 to 1, with values closer to 1 or -1 indicating a stronger correlation and closer to 0 indicating a weaker or no linear relationship.

Spearman's correlation is for variables with non-linear relationships and is based on the rank of the variables rather than the raw values. Spearman's correlation coefficient also ranges from -1 to 1, similar to Pearson's correlation coefficient.

By calculating these correlation coefficients, we can obtain the correlation matrix between the variables and perform visual analysis. The correlation matrix can help us find linear or nonlinear relationships between variables and serve as a basis for feature selection.

Based on the importance and correlation of variables, we can further optimize the modeling process. By ranking the variables by importance, we can select the variables with higher importance and use them as the input of the model. At the same time, we can also analyze the correlation between the variables to avoid duplicate or highly correlated variables being included in the model at the same time.

In addition, we can also consider performing feature engineering to extract more meaningful features. Feature engineering involves transforming, aggregating, and combining raw data to extract features that are more informative and predictive. Feature engineering can be performed through methods such as domain knowledge, statistical analysis and machine learning.

Finally, we should note that the choice of variables in the modeling process is closely related to the performance of the model. Proper selection and use of important and relevant variables can improve the predictive accuracy and explanatory power of a model. Therefore, we should comprehensively consider the importance, correlation and domain knowledge of the variables, and flexibly adjust and optimize them according to the characteristics of the data and the problem.

To sum up, based on the optimal results in 5.3, we can conclude that the selected variables have some validity in predicting the target. By evaluating the importance and correlation of the variables, we can further refine the modeling process, select the important and relevant variables, and perform feature engineering to extract more meaningful features. These steps will help to improve the predictive and explanatory power of the model and make it more suitable for practical application requirements.

# **6 Conclusions**

Gaussian process modeling is a common method to predict and analyze the spread of diseases when the case data set is known. In this method, a model can be built by fitting the distribution of existing data, and the model can be used to predict continuous variables to evaluate the accuracy of the prediction results [17].

Gaussian process modeling is a method based on probability and statistics, which is widely used in various fields, including finance, meteorology, environment, etc. In the prediction of disease spread, Gaussian process modeling is used to analyze and predict the trend and law of disease spread.

In Gaussian process modeling, we assume that the data of disease spread has the characteristics of normal distribution, so it is reasonable to assume normality. This means that we can use the properties of the normal distribution to infer the number of cases in the future and predict the trend of the disease spread. By building a Gaussian process model, we can infer the number of cases in the future and make predictions based on the existing case data.

In the prediction of disease spread, the commonly used evaluation indicators include root mean square error (RMSE) and mean square error (MSE). These indicators can measure the degree of difference between the prediction results of the model and the actual data. If the prediction result of the model is closer to the actual data, the smaller the value of the evaluation index is, the better the prediction effect is [18].

In addition to using Gaussian process modeling to predict the spread of diseases, we can also use time series analysis methods to better understand the spread trend and evolution pattern of diseases. By modeling and analyzing time series data, dynamic information about disease spread, such as seasonal changes and long-term trends, can be obtained, and future predictions can be made based on this information.

Time series analysis (TDA) is a modeling and forecasting method for time series data. In epidemic prediction, we can view the time series data of disease transmission as a process changing over time, and infer the future change trend based on historical data. Through time series analysis, we can grasp the law and trend of disease spread, and then make more accurate predictions.

In addition, considering the timeliness of case data, we can use the method of time series analysis to better understand the transmission trend and evolution pattern of the disease. By modeling and analyzing time series data, we can obtain dynamic information about disease transmission, such as seasonal changes and long-term trends, and make future predictions based on this information.

In addition to time series analysis, we can also delineate specific spatial extences, such as countries or cities, to better model the spread of epidemics. Using known spatial and disease spread parameters (such as spread speed, infection probability, etc.), we can conduct further simulations by building models to predict the spread of the epidemic in a specific area.

However, it is important to note that for Covid data containing more features, simple Gaussian process modeling may not meet the needs. Because more features may lead to an increase in the complexity of the data, the predictive effect of the model may decrease. In such cases, we can consider using other prediction models or adopting more advanced machine learning methods to handle more features and complexity in the data [19].

In addition, disease transmission in real-world situations is usually more complex and challenging than theoretical models. In addition to basic factors such as speed of transmission and probability of infection, there are many other factors that influence the spread of a disease, such as individual behavior, social distancing, quarantine measures, and vaccine coverage. Therefore, in epidemic modeling and prediction, we need to comprehensively consider various factors, and continuously update and optimize the model to improve the accuracy and reliability for real-world situations.

Finally, we need to recognize that disease transmission is indeed a complex and serious challenge. It involves many unknown factors and uncertainties. Therefore, we need to pay close attention to the development of the epidemic, collect and update data regularly, and work closely with public health departments and experts to formulate scientific and rational prevention and control strategies to protect people's health and safety. Through continuous research and efforts, we believe that we can effectively deal with the challenges brought by the spread of the disease and ensure social stability and people's well-being [20].

**References**

1. Alimadadi A et al (2020) Artificial intelligence and machine learning to fight COVID-19. American Physiological Society, Bethesda
2. Amar LA, Taha AA, Mohamed MY (2020) Prediction of the final size for COVID-19 epidemic using machine learning: a case study of Egypt. Infect Dis Model 5:622–634
3. Anastasopoulos C et al (2020) Development and clinical implementation of tailored image analysis tools for COVID-19 in the midst of the pandemic: the synergetic effect of an open, clinically embedded software development platform and machine learning. Eur J Radiol 131:109233
4. Arvind V et al (2020) Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. J Crit Care 62:25–30
5. Assaf D et al (2020) Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med 15(8):1435–1443
6. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
7. Brinati D et al (2020) Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. J Med Syst 44:135
8. Brunese L et al (2020) Machine learning for coronavirus COVID-19 detection from chest x-rays. Proced Comput Sci 176:2212-2221
9. Burdick H et al (2020a) Prediction of respiratory decompensation in Covid-19 patients using machine learning: the READY trial. Comput Biol Med 124:103949
10. Burdick H et al (2020b) Is machine learning a better way to identify COVID-19 patients who might benefit from hydroxychloroquine treatment? The Identify Trial. J Clin Med 9(12):3834
11. Chekole YA et al (2020) Perceived Stress and Its Associated Factors during COVID-19 among Healthcare Providers in Ethiopia: a cross - sectional study. Adv Public Health. https://doi.org/10.1155/2020/5036861
12. Choi S et al (2020) Predicting psychological distress amid the COVID-19 pandemic by machine learning: discrimination and coping mechanisms of Korean Immigrants in the US. Int J Environ Res Public Health 17(17):6057
13. Day M (2020) Covid-19: identifying and isolating asymptomatic people helped eliminate virus in Italian village. BMJ 368:135
14. Deng X et al (2020) A classification–detection approach of COVID-19 based on chest X-ray and CT by using keras pre-trained deep learning models. Comput Model Eng Sci 125(2):579–596
15. Doanvo A et al (2020) Machine learning maps research needs in covid-19 literature. Patterns 1(9):100–123
16. Gao K et al (2020) Julia language in machine learning: algorithms, applications, and open issues. Comput Sci Rev 37:100254
17. Garcia-Ordas MT et al (2020) Evaluation of country dietary habits using machine learning techniques in relation to deaths from COVID-19. Healthcare 8:371
18. Hastie TR, Tibshirani JF (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science and Business Media, Berlin
19. Hermans JJ et al (2020) Chest CT for triage during COVID-19 on the emergency department: myth or truth? Emerg Radiol 27(6):641–651
20. Hossain B et al (2019) Surgical outcome prediction in total knee arthroplasty using machine learning. Intell Autom Soft Comput 25(1):105–115