COVID 案例分析报告

1 简介

COVID-19 疫情的爆发给全球带来了巨大的挑战，政府和卫生机构需要准确的数据和科学的分析来制定有效的防控措施。UQ4Covid 项目的目的就是通过整合和分析各种 COVID-19 相关数据，为决策者提供实时的、准确的信息和预测模型，以支持他们做出明智的决策。

通过建模和预测病例数据，可以提供有关 COVID-19 疫情的重要见解。期望我的模型可以预测疫情的传播趋势、病例数量的增长率以及可能的高风险区域。这些预测结果可以帮助政府和卫生机构制定相应的防控策略，优化资源分配，并及时采取措施来减缓疫情的蔓延。

本文的主要目的是针对 Covid 疾病病例数据进行分析，在不同需求条件下，对患病死亡人群进行预测，以期望可以得到某些维度对病例死亡的影响原因。

2 文献综述

高斯过程建模

高斯过程（Gaussian Process）是一种概率模型，常用于回归和分类问题。它基于贝叶斯推断的思想，通过对数据进行建模，可以为每个输入点提供一个输出值，并对预测结果提供置信度。

以下是高斯过程模型的基本原理：

假设有一组输入数据点 X 和对应的输出数据点 Y。输入数据点可以是任意维度的向量。高斯过程模型假设 Y 是由未知函数 f 在每个输入点上的取值加上一个噪声项 ε 生成的：$Y = f(X) + \varepsilon$。

高斯过程模型对函数 f 做了一个非参数化的假设，认为 f 是一个随机过程，并描述了 f 在每个输入点上的统计特性。具体而言，高斯过程模型假设任意有限个输出值的联合分布服从多元高斯分布。

对于给定的输入数据点 X，高斯过程模型构建了一个关于 X 的先验分布。这个先验分布可以看作是对函数的一种无限维度的概率分布。

在观测到部分输出数据点 Y 后，高斯过程模型可以利用贝叶斯定理来更新先验分布得到后验分布。后验分布是对函数 f 的条件分布，给出了已观测数据的情况下，函数在每个输入点上的取值的概率分布。

使用已观测数据点的后验分布，高斯过程模型可以进行预测。对于一个新的输入点 x，可以计算其在给定已观测数据的情况下，对应输出值的条件分布。常用的预测指标包括均值和方差，分别表示预测输出的期望值和不确定性度量。

总结来说，高斯过程模型通过对函数的先验分布和观测数据的后验分布建模，提供了一种灵活且有弹性的方式来处理回归和分类问题。它不仅可以进行预测，还能提供对预测结果的置信度估计，使得模型输出更具可解释性和可靠性。

高斯过程回归

高斯过程回归（Gaussian Process Regression，简称 GPR）是一种基于高斯过程的非参数回归方法。它可以用于建模输入和输出之间的非线性关系，并提供对预测结果的不确定性估计。

高斯过程回归具有很好的灵活性和非参数性质，因为它不需要事先对模型结构进行设定。它可以适应不同的数据分布和非线性关系，并且提供了对预测结果的置信度估计。这使得高斯过程回归成为一种强大的回归建模方法。

残差图

残差图是评估统计模型拟合效果的重要工具之一。在回归分析中，我们通过拟合模型来预测因变量的取值，而残差则代表了观测值和模型预测值之间的差异。

残差图能够展示出模型的拟合效果是否良好。一个良好的模型拟合应该具有以下特征：1）残差的平均值接近于零，这意味着模型对数据整体水平的拟合较好；2）残差的方差应该是恒定的，即在所有自变量的取值范围内保持稳定，没有明显的规律性变化；3）残差应该随机分布在零附近，没有明显的聚集或趋势。

当残差图显示以下情况时，可能表示模型存在问题：1）残差的平均值显著偏离零，说明模型整体上低估或高估了观测值；2）残差的方差不稳定，可能出现了异方差性，这可能意味着模型不能很好地适应不同区域的数据；3）残差出现了非随机的聚集或趋势，暗示模型未能捕捉到数据的某些重要特征，导致了系统性的偏差。

除了直观地展示模型拟合效果，残差图还可以用来检验模型的假设是否成立。例如，如果残差图显示了明显的非线性关系或者异方差性，可能暗示着模型中存在未建模的非线性关系或其他未考虑的因素。

3 实现方法

本文基于 UQ4covid 项目提供的数据集，使用高斯回归建模，对模型进行回归预测，并将数据集从空间（region）和病例类型（LAD）不同分类进行预测，期望得到不同分类对结果的影响。

4 数据

输入数据集介绍:

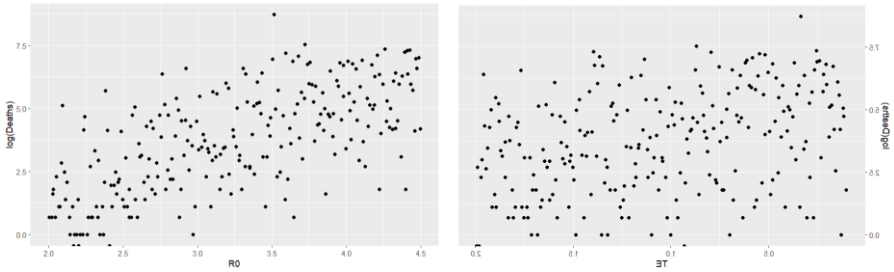
R0	Reproduction number	(2,4.5)
TE	Mean latent period	(0.1,2)
TP	Mean pre-symptomatic infectious period	(1.2,3)
TI1	Mean infectious and symptomatic	(2.8,4.5)
TI2	Mean infectious and symptomatic	(0.0001,0.5)
nuA	Numbers of infectious but asymptomatic;	(0,1)
ns	Number of seeds	\
p_home_weekend	\	(0,1)
alphaTH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

数据处理：由于在 input 输入集中存在 repeat 重复试验，所以在使用时，需要删去 repeat 集，并且在对 output 集处理时，使用 mean 均值来进行数据集拆分。

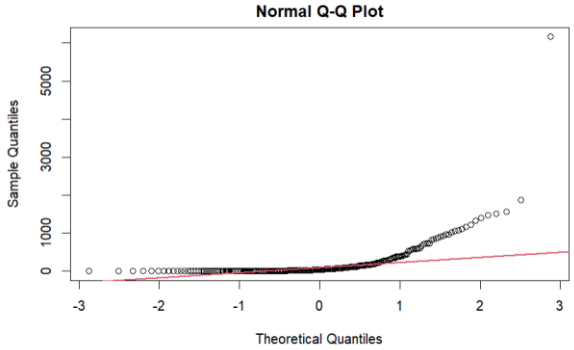
将数据集拆分为 70%的训练集和 30%的预测集。

数据预览:

R0 与 log（Death）的分布如下，数据点的分布不存在一个明显的异方差。



观察对 Deaths 数据点的正态性，可以看到数据点的分布满足正态性，但是数据尾端有偏离的趋势。



5 模型构建与预测

5.1 高斯过程模型:

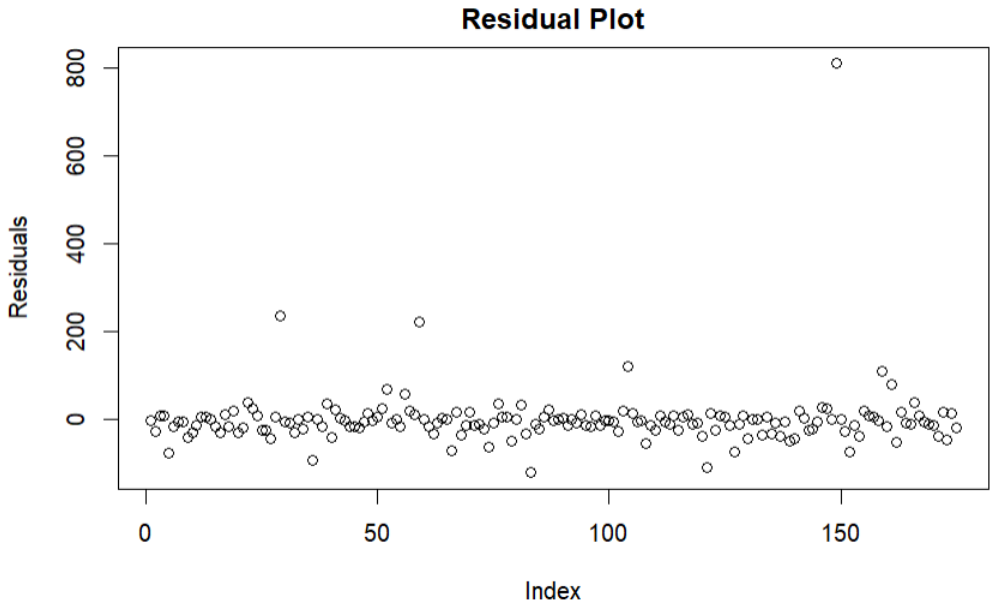
5.1.1 总体数据进行建模，并且根据模型拟合，使用 R 语言中的 gausspr 函数可以方便地构建高斯过程模型。该函数接受自变量和因变量作为输入，并通过拟合数据来推断数据之间的关系。

在构建高斯过程模型之后，我们可以利用该模型进行预测。通过输入新的自变量，模型可以生成对应的概率分布，即预测的因变量取值范围。这使得我们能够对未知数据进行预测，并估计其不确定性。

除了预测功能，高斯过程模型还可以提供其他有用的信息。例如，可以计算边际似然函数来评估模型的拟合程度，或者通过计算残差来检验模型的拟合优度。此外，还可以通过调整模型的超参数来优化模型的性能，例如核函数的选择和参数的设置。

5.1.2

对于整体模型，我们可以从残差图看出模型的拟合效果:



上图中可以看出，除了一些数据点明显的异常值，其他残差数据点都均匀分布在 0 附近，模型的拟合效果较好。

针对测试集的预测，可以得到测试集的预测效果 RMSE=326，R^2=0.36，预测效果并不理想。

5.2 对于 region 分辨率的建模预测

将训练集根据 region 变量划分集合，循环建模并进行预测，得各个模型的评价。

见下图：

	rmse <dbl>	mse <dbl>	r2 <dbl>
E12000001	232.4125	54015.56	0.3815607
E12000002	305.1065	93089.95	0.3373736
E12000003	377.8127	142742.41	0.3452525
E12000004	285.3310	81413.77	0.3701621
E12000005	401.9725	161581.93	0.3637832
E12000006	310.9329	96679.26	0.3654153
E12000007	612.8088	375534.57	0.3616315
E12000008	291.6243	85044.73	0.3503383
E12000009	284.7738	81096.10	0.3552679

第一列是 region 的不同选择量，可以看到，预测效果最好的是 E120000001 下的预测，R^2为 0.38，但是预测效果也并不理想。

5.3 对于度量指标的分类建模预测

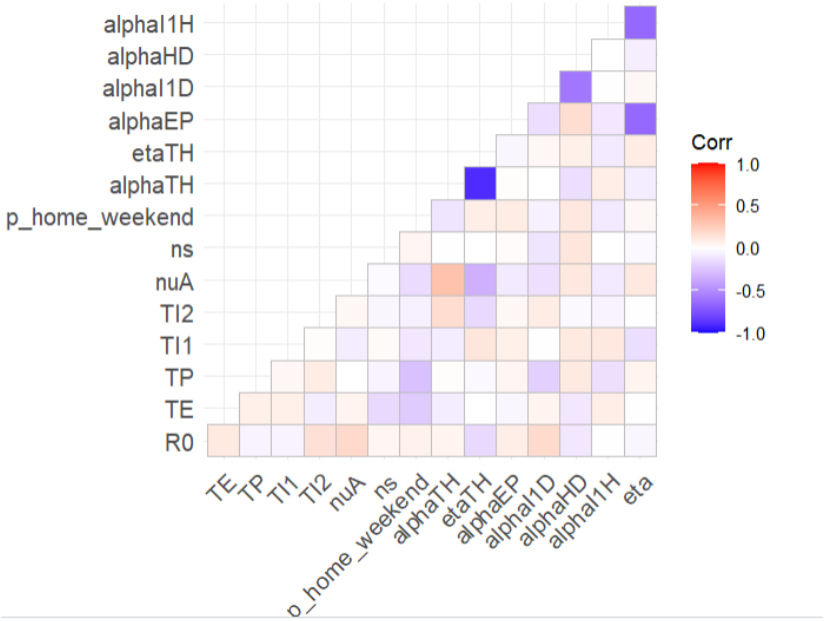
将训练集根据 LAD19CD 变量划分集合，循环建模并进行预测，得各个模型的评价。

输出了 R^2的最大的输入集：

	rmse <dbl>	mse <dbl>	r2 <dbl>
E06000012	216.7424	46977.25	0.417521

有 R^2将近 0.42，相较于 5.2 的预测又有一部分的提升，说明变量的选择是有效的。

基于上述结论，我们可以来查看一下建模的变量重要性与相关性：



可以看到变量间相关性最强的是 etaTH 与 alphaTH 之间，其他都比较弱，但是将其他变量筛去不符合逻辑，考虑模型问题。

6 结论

在已知病例数据集的情况下，高斯过程建模是一种常见的方法，用于对疾病传播进行预测和分析。这种方法可以通过拟合已有数据的分布来建立模型，并利用该模型进行连续变量的预测，以评估预测结果的准确性。

在疾病传播预测中，常用的评价指标包括均方根误差（RMSE）、均方误差（MSE）和确定系数（R^2）。这些指标可以衡量模型预测结果与实际数据之间的差异程度。如果模型的预测结果与实际数据越接近，则评价指标的值越小，预测效果越好。

在高斯过程建模中，我们假设疾病传播的数据具有正态分布的特征，因此假设正态性是合理的。这意味着我们可以利用正态分布的性质来推断未来的病例数目，并预测疾病传播的趋势。

另外，考虑到病例数据具有时效性，我们可以采用时间序列分析的方法来更好地理解疾病的传播趋势和演变模式。通过对时间序列数据进行建模和分析，可以获得关于疾病传播的动态信息，如季节性变化、长期趋势等，并据此进行未来的预测。

除了时间序列分析，我们还可以划定特定的空间范围，例如国家或城市，以更好地模拟疫情的传播。利用已知的空间和疾病传播参数（如传播速度、感染概率等），我们可以通过建立模型进行进一步的模拟，以预测疫情在特定地区的传播情况。

然而，需要注意的是，对于包含更多特征的 Covid 数据，简单的高斯过程建模可能无法满足需求。因为更多的特征可能导致数据的复杂性增加，模型的预测效果可能会下降。在这种情况下，我们可以考虑使用其他预测模型或采用更高级的机器学习方法，以处理数据中的更多特征和复杂性。

此外，现实情况中的疾病传播通常比理论模型更加复杂和具有挑战性。除了传播速度和感染概率等基本因素外，还有许多其他因素会影响疾病的传播，如个人行为、社交距离、隔离措施和疫苗覆盖率等。因此，在疫情建模和预测中，我们需要综合考虑各种因素，并不断更新和优化模型，以提高对现实情况的准确性和可靠性。

最后，我们需要认识到疾病传播确实是一项复杂而严峻的挑战。它涉及到许多未知的因素和不确定性，因此，我们需要密切关注疫情的发展，定期收集和更新数据，并与公共卫生部门和专家紧密合作，以制定科学、合理的防控策略，保护人民的健康和安全。