課題1

4.モデルの選択理由は、残差の分布が過剰に偏っていないことや多重共線性がないことを確認した上で、第一に、自由度調整済み決定係数が高くAICやBICが低いことが示すように、フィッテイングと複雑さのバランスが良いためである。第二に、p値が示す有意差検定における有意な変数の数が多いためである。第三に、標準偏回帰係数が医学的知識と整合的なためである。それぞれの変数を選択した理由について、まず、従来株と変異株の与える影響は異なるためこれらを分離した。人の動きに関しては、緊急事態宣言は情報効果を人々に与え、実質GDPは経済の動きを示し、長期休暇期間は気のゆるみを表すとして説明変数の中に取り入れた。時期としての要因について、ウイルスの生存に適した環境かどうかとして平均気温や冬であるかを組み込み、土日は検査数が少なくなることから週の前半後半とのダミー変数を作成した。感染から陽性の報告までの発症、認知、報告ラグなどを一週間と仮定して一部の変数にラグを置いた。結果として、新規感染者数に対して、平均気温、緊急事態宣言、実質GDPは負の影響を与え、週の後半、従来株、正月、検査数、冬、変異株は正の影響を与えることが確認された。

課題3

1.

最小二乗法の利点は、第一に、誤差を複数の解が現れうる絶対値誤差や複雑な四乗誤差としたときよりも計算が容易でありかつ解が一意に定まることである。一階条件より、それぞれのパラメータで損失関数を偏微分したものが0となり、それらの連立方程式を解くことで解を求めることができる。第二に、方法がシンプルであるために、企業で専門知識のない人に説明する際に、説明責任を果たすことができることが利点である。

2.

分散分析は、データから得られた関係や結果が偶然ではないかを確かめる有意差検定の一部であり、特にグループごとの平均値の差の検定を分析するものである。2グループの差を検定するt検定を拡張したもので、三つ以上の水準や多元配置の分析にも対応している。また、三つ以上の水準で差が有意であった場合、水準の数をnとすると、通りの具体的な組み合わせでの差の検証を行う多重比較とともに用いられる。

3.

因子分析は、それぞれの変数から共通する因子を取り出すことを目的としたものであり、使用する際は抽出する因子をあらかじめ決定する必要があり、解釈がしやすい一方で実行時の手法の選択が難しく、現象分析を目的としたものに使用される。主成分分析は、測定された変数から主成分を表現することを目的とし、主成分数は後から決定できるが、解釈が難しい一方で、実行は容易であり、機械学習の全処理などで効力を発揮する。