**データサイエンス　課題1　分散分析**

**締め切り: 5月15日10:25am**

各質問について、関連するRコードと出力を、文書に貼り付けてください。

**Part 1**pima.RDataを用いて以下の問いに答えよ。

* 国立糖尿病・消化器・腎臓病研究所がフェニックス近郊に住むピマ・インディアンの女性768人を対象に実施した研究
* データセットは以下の変数を含む
  + pregnant 妊娠回数
  + Glucose 経口ブドウ糖負荷試験における2時間後の血漿中グルコース濃度
  + Diastolic 拡張期血圧 (mm Hg)
  + Triceps 上腕三頭筋 皮膚襞厚(mm)
  + insulin 2時間後血清インスリン
  + Bmi 肥満度
  + Diabetes 糖尿病血統要因
  + age 年齢（歳）
  + Test 患者が糖尿病の兆候を示すかどうか（0（陰性）、1（陽性）)

問 1

bmiを3つのカテゴリーに分類する新しい変数を作成せよ[(i) 0 - 25, (ii)>25 - 40, (iii) >40]　カテゴリーのラベル名は好きなようにつけてください。 この新しい変数をfactorとして設定します。bmiの各カテゴリーにおける糖尿病血統要因の平均は何ですか？

問 2

分散分析を実行して、糖尿病血統要因に関する bmi カテゴリ変数の予測能力をテストしなさい。 その結果からどのような結論が得られますか？

尚、糖尿病血統要因(diabetes)は、両親、祖父母、兄弟姉妹、叔父叔母、そして最初のいとこからの情報である点に注意してください。これは、被験者の糖尿病リスクに対して予想される遺伝的影響の尺度を示すものです。 つまり、diabetesという変数は、被験者の家族の間での糖尿病の有病率についてであり、bmiは被験者のbmiです。 この2つの変数の間には、表面的には関係がないはずなのに、なぜ関連があるのでしょうか（言い換えれば、なぜ被験者のbmiと親族の糖尿病の状態が関連するのでしょうか）？ 「偽りの発見」なのでしょうか？