データサイエンス 課題1 分散分析

再提出: 5月22日 10:25am以降
2323050 井上祐斗

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？

bmiを[(i) 0 - 25, (ii)>25 - 40, (iii) >40]の範囲で、それぞれラベルをlow, normal, highとして３分類した。それらのカテゴリー群における糖尿病血糖要因は以下であ流。

```
factor  diabetes
1    low 0.4067154
2 normal 0.4621494
3   high 0.6109896
```

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？

```
Response: diabetes
           Df Sum Sq Mean Sq F value    Pr(>F)
factor      2  2.432 1.21602  11.377 1.353e-05 ***
Residuals 765 81.768 0.10689
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

帰無仮説である「bmiのグループ間において糖尿病血統要因に差はない」は、F値が0.99998647(1 - 1.353e-05)であることから、棄却され、統計的に有意に差があると言える。つまり、bmiから糖尿病血統要因を予測可能である。

しかし、これはあくまで統計的にグループ間に差が優位であることをしめしているだけであるので、相関関係はあるといえるが因果関係があるとは言えない。研究では、原因と因果を説明するメカニズムをさらに明らかにする必要がある。具体的なメカニズムについては、親族間で食生活が似ているからbmiと糖尿病に相関関係が生まれたのかと考えられる。

ソースコード anova.r

```
load("yourpath/pima.RData")

factor ← cut(
  pima$bmi,
  breaks = c(-Inf, 25, 40, Inf),
  labels = c("low", "normal", "high"),
  right = TRUE
)

aggregate(diabetes ~ factor,data=pima, FUN=mean)

model ← lm(diabetes ~ factor, data = pima)

anova(model)
```