データサイエンス最終課題

2323050 井上祐斗

検証する仮説とデータセット

このレポートで検証する仮説は、「LDLコレステロールが増加すると心臓病になりやすくなる」というものである。

データセットとして、kaggleで見つけたものを使用した。これは、1988年のもので、個人の総コレステロールcholを含む様々な要因と、心臓病の有無を示す「target」を含むものである。

以下の表にデータがどのようなものであるかをまとめる。

項目名	説明	値の範囲・種類
Age	患者の年齢	29〜77歳
Sex	患者の性別	0 = 女性, 1 = 男性
Serum Cholestoral	血清コレステロール値	単位: mg/dl 範囲: 126〜564 mg/dl
Target	予測対象：心臓病の有無	0 = 心臓病なし, 1 = 心臓病あり
...(17列ある)		

「target」は0（心臓病なし）または1（心臓病あり）の整数値である。 なお、LDLコレステロールを調べるデータセットが見つからなかったため、非常に強い相関を持つ総コレステロールを今回は使用する。

また、1025行のデータセットで、心臓病の有無の割合がおおよそ半々になっている。

ロジスティック回帰モデルについて

今回の仮説検証には、ロジスティック回帰を用いる。ロジスティック回帰は、ある事象が発生するかが2値で表されていて、複数の要因がどう影響を与えているかを検証できる。そのため、ある要因によって何かを分類するような問題に適している。そのため、今回の場合に最適である。

以下のロジット（logit）関数がモデルの式になる。

$$log(\frac{P}{1-P}) = \beta_0 + \beta_1 X$$

- Pは、心臓病である確率を示す
- 1−Pは、心臓病でない確率を示す
- P/1-Pはオッズと呼ばれ、心臓病がある確率とない確率の比率を表す
- β0 は、切片
- β1 は、総コレステロール値 cholの係数

分析 Rコード

```
library(ggplot2)
# データの読み込み
data ← read.csv("heart.csv")
summary(data)
str(data)

# 分析
model ← glm(target ~ chol , data = data, family = binomial(link = "logit"))
summary(model)

# オッズ比の計算
# LDLコレステロールが1単位増加したときの心臓病発生率のオッズ比
exp(coef(model))
exp(confint(model))

# グラフの描画
ggplot(data, aes(x = chol, y = target)) +
geom_point(position = position_jitter(height = 0.05), alpha = 0.5) +
stat_smooth(method = "glm", method.args = list(family = "binomial"), se = FAL
SE) +
```

```
labs(title = "relation", x = "LDL chol", y = "target") +
theme_minimal()
```

結果

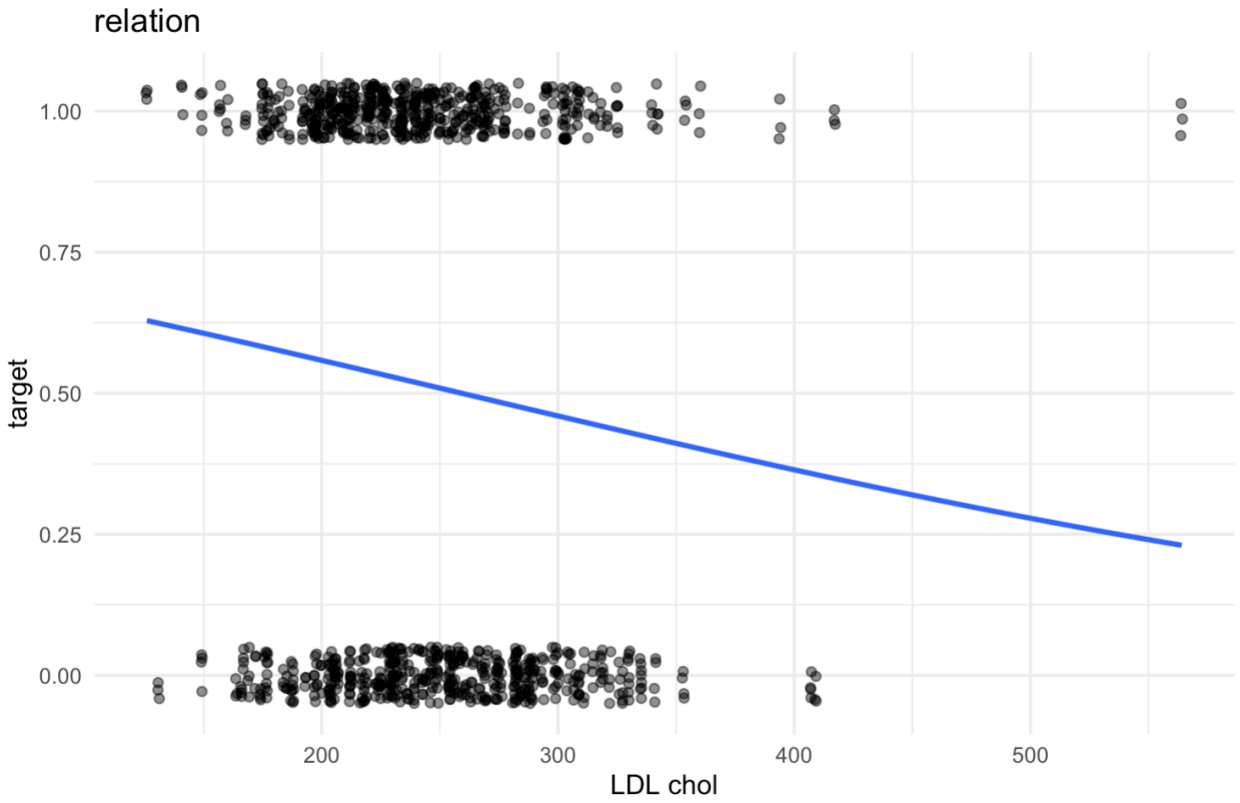
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.025595   0.313193   3.275  0.00106 **
chol        -0.003956   0.001248  -3.168  0.00153 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1420.2  on 1024  degrees of freedom
Residual deviance: 1409.9  on 1023  degrees of freedom
AIC: 1413.9

Number of Fisher Scoring iterations: 4
```

```
> exp(coef(model))
(Intercept)        chol
  2.7887533   0.9960523
> exp(confint(model))
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) 1.5179073 5.1881940
chol        0.9935874 0.9984679
```



解釈

- z値の絶対値である3.275, 3.168が、1.96より大きいため95%信頼区間で有意である。
- 総コレステロール値が1単位（1 mg/dL）増加するごとに、心臓病になるオッズが約 0.996 倍になる。つまり、減少する。

今回は、従来の医学で信じられてきたコレステロールによる心臓病の予測についての検証を行った。しかし、結果は仮説と反対するものであった。

現在、LDLコレステロールはマーカーとしての側面が強く、より広い文脈で総合的に判断されるべきだと考えるのが最新の動向となっている。今回は、そのマーカー機能としてすら使えないというの反対の結果が出てしまったが、これはデータセットが古く、コレステロールの内訳がデータセットに含まれていない、複数のデータベースから集めているので測定にばらつきがある、などが考えられる。複雑な要因が組み合わさって起きる問題については、分析が難しいことがわかった。

参考文献

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data