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疾患になりやすくなる」というものである。

データセットとして、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」を含むものを使用する。「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。

```
library(ggplot2)
data ← read.csv("heart.csv")
summary(data)
str(data)

# ロジスティック回帰モデルの構築
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死亡率（心臓病発生率）のオッズ比
exp(coef(model))
exp(confint(model))

# 視覚化 (オプション: LDLコレステロールと死亡率の関係をプロット)
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 = FALSE) +
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 **
---
```

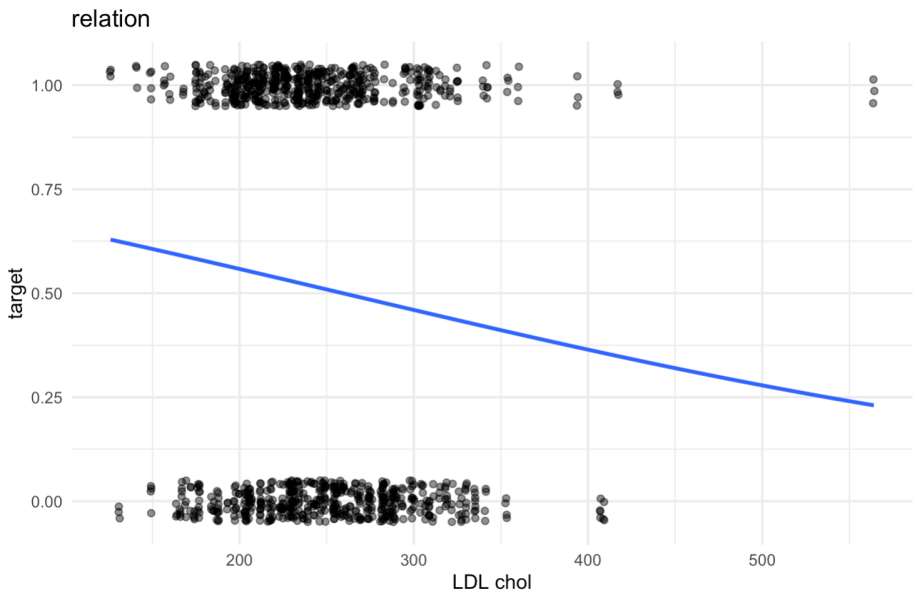
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(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