

# データサイエンス 課題4

学籍番号2323050, 氏名 井上祐斗

## 問1

```
# 1.距離のみ
wells1 <- glm(switch ~ dist,family = binomial(link = "logit"),data = wells)
summary(wells1)
exp(coef(wells1))
```

```
(Intercept)    dist
    1.8330099  0.9938005
```

$$P = O/(1 + O) = 1.83/(1 + 1.83) = 64$$

Intercept, distはどちらも統計的に有意である。しかし、距離のオッズに関しては、ほぼ1に近く僅かに減少していくことがわかる。

上の計算式によって、オッズから確率を計算できる。dist=0mのとき、オッズは1.83であり、確率は64%であるので、distが1m増えると、僅かに64%から減少していくと考えられる。

つまり、このモデルで解釈するなら、安全な井戸が遠くなるほど、汲みに行く手間が面倒くさくなるため、それより近くの汚染された井戸を使い続けると言える。

## 問2

```
# 2.距離とヒ素
wells2 <- glm(switch ~ dist + arsenic,family = binomial(link = "logit"),data = wells)
summary(wells2)
exp(coef(wells2))
```

Coefficients:

```
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.002749   0.079448   0.035   0.972
dist        -0.008966   0.001043  -8.593 <2e-16 ***
arsenic      0.460775   0.041385  11.134 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4118.1 on 3019 degrees of freedom

```
Residual deviance: 3930.7 on 3017 degrees of freedom
AIC: 3936.7
```

```
Number of Fisher Scoring iterations: 4
```

```
> exp(coef(wells2))
(Intercept)    dist    arsenic
  1.0027525  0.9910736  1.5853020
```

dist, arsenicは統計有意であるが、Interceptは統計有意ではない。

ロジットの観点から、dist, arsenicはそれぞれ微妙な減少、増加をしていることがわかり、さらに、オッズの観点から、同じようにそれぞれ微妙な減少、増加を示していることがわかる。

特に、arsenicについては1(1Lあたり数百μg)増加するごとにオッズが1.5倍になる。つまり、井戸の交換に大きな影響を与えていることがわかる。

### 問3

```
# 3.全てで解析, 教育年数の影響について検証
wells3 <- glm(switch ~ dist + arsenic + assoc + educ,family = binomial(link = "logit"),data = wells)
summary(wells3)
exp(coef(wells3))

median_dist <- median(wells$dist)
median_arsenic <- median(wells$arsenic)

educ_values <- c(0, 4, 8, 12, 16)
newdata <- data.frame(
  educ = educ_values,
  dist = rep(median_dist, length(educ_values)),
  assoc = 1,
  arsenic = rep(median_arsenic, length(educ_values))
)
prob <- predict(wells3, newdata = newdata, type = "response") # responseは確率として返す
results <- data.frame(
  educ = educ_values,
  prob= prob
)
results
```

Coefficients:

```
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.156712  0.099601 -1.573  0.116
dist        -0.008961  0.001046 -8.569 < 2e-16 ***
```

```
arsenic 0.467022 0.041602 11.226 < 2e-16 ***
assoc -0.124300 0.076966 -1.615 0.106
educ 0.042447 0.009588 4.427 9.55e-06 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4118.1 on 3019 degrees of freedom  
Residual deviance: 3907.8 on 3015 degrees of freedom  
AIC: 3917.8

Number of Fisher Scoring iterations: 4

```
> exp(coef(wells3))
(Intercept) dist arsenic assoc educ
0.8549505 0.9910789 1.5952358 0.8831149 1.0433604
```

```
educ prob
1 0 0.4991732
2 4 0.5415240
3 8 0.5832832
4 12 0.6238806
5 16 0.6628093
```

教育年数	0	4	8	12	16
切り替えの確率	49%	54%	58%	62%	66%

教育年数が高まるにつれ、切り替えの確率が増えていることが示された。

## 問4

「教育年数が高いほど、ヒ素の危険性について理解しているため、高い濃度に対して敏感であり井戸の切り替えにつながりやすい。」ことが考えられるため、educとarsenicを交互作用として追加した。

```
# 交互作用項を追加
wells4 <- glm(switch ~ dist + arsenic + assoc + educ + arsenic * educ, family = binomial(link = "logit"))
summary(wells4)
exp(coef(wells4))
```

Coefficients:

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.028283 0.119711 0.236 0.81323
dist -0.008951 0.001048 -8.542 < 2e-16 ***
```

```
arsenic    0.348127  0.058902  5.910 3.42e-09 ***
assoc      -0.123399  0.077020 -1.602 0.10912
educ       -0.000392  0.018351 -0.021 0.98296
arsenic:educ 0.028551  0.010515  2.715 0.00662 **
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4118.1 on 3019 degrees of freedom

Residual deviance: 3900.3 on 3014 degrees of freedom

AIC: 3912.3

Number of Fisher Scoring iterations: 4

```
(Intercept)    dist    arsenic    assoc    educ arsenic:educ
1.0286871    0.9910894    1.4164128    0.8839109    0.9996080    1.0289625
```

教育レベルが1年上がるごとに、ヒ素レベルが井戸の切り替えに与えるオッズが1.02倍になり、交互作用項に交互作用が働くことがモデルで示された。z値を見ると統計的に有意である。

効果としては、arsenicのオッズは教育年数に応じて増加し、教育年数8年の場合の arsenic のオッズ比  $= 1.41 \times (1.02)^8$  として計算できる。