STAT504_Q2

Houjie Wang & Shuang Wu

2/19/2021

Problem Recap

Are more physical activities related to a healthier life?

Related Variables Recap

Variables

HEALTH: If -9, not ascertained; if -8, I don't know; if -7, prefer not to answer; if 1, excellent; if 2, very good; if 3, good; if 4, fair; if 5, poor

Predictor

LPACT Count of times of light or moderate physical activity in past week

VPACT Count of times of vigorous physical activity in past week

PHYACT If -9, not ascertained; if -8, I don't know; if -7, prefer not to answer; if 1, rarely or never conduct any physical activities; if 2, some light or moderate physical activities; 3. Some vigorous physical activities

Data Cleaning

To simplify, observations only with positive values in HEALTH and PHYCT are taken.

```
data <- read.csv("DataforUse.csv")
# data_full <- read.csv("perpub.csv")
valid_health_ind <- data$HEALTH %in% 1: 5 + data$PHYACT %in% 1: 3 +
  as.vector(data$LPACT >= -1) + as.vector(data$VPACT >= -1) == 4
data_health <- cbind.data.frame(HEALTH = data$HEALTH[valid_health_ind],
                                LPACT = as.integer(data$LPACT[valid_health_ind]),
                                VPACT = as.integer(data$VPACT[valid_health_ind]),
                                PHYACT = as.factor(data$PHYACT[valid_health_ind]))
nrow(data_health)
```

```
## [1] 262437
```

```
# Verify if LPACT=VPACT=-1, when PHYACT=1 (TRUE)
sum(rowsums(as.matrix(data_health[data_health$PHYACT == 1, 2: 3])) != -2) == 0
```

```
## [1] TRUE
```

```
# Verify if VPACT=-1, when PHYACT=2 (TRUE)
sum(data_health[data_health$PHYACT == 2, 3] != -1) == 0
```

```
## [1] TRUE
```

```
# Verify if LPACT=-1, when PHYACT=3 (TRUE)
sum(data_health[data_health$PHYACT == 3, 2] != -1) == 0
```

```
## [1] TRUE
```

```r
# Set all "-1" in LPACT and VPACT to "0" since they both represent no LPACT or VPACT last week
data_health$LPACT[data_health$LPACT == -1] = 0
data_health$VPACT[data_health$VPACT == -1] = 0
```

Model Proposal

$$HEALTH = \beta_0 + \beta_1 LPACT + \beta_2 VPACT$$

Here is our multinomial logistic regression model. Y represents HEALTH and X represents the design matrix.

$$\begin{aligned}
P(Y_i = 1) &= 1 - \sum_{k=2}^{5} P(Y_i = k) = \frac{1}{1 + \sum_{k=2}^{5} \exp\{\boldsymbol{\beta}_k' \boldsymbol{X}_i\}} \\
P(Y_i = 2) &= P(Y_i = 1) \exp\{\boldsymbol{\beta}_2' \boldsymbol{X}_i\} = \frac{\exp\{\boldsymbol{\beta}_2' \boldsymbol{X}_i\}}{1 + \sum_{k=2}^{5} \exp\{\boldsymbol{\beta}_k' \boldsymbol{X}_i\}} \\
P(Y_i = 3) &= P(Y_i = 1) \exp\{\boldsymbol{\beta}_3' \boldsymbol{X}_i\} = \frac{\exp\{\boldsymbol{\beta}_3' \boldsymbol{X}_i\}}{1 + \sum_{k=2}^{5} \exp\{\boldsymbol{\beta}_k' \boldsymbol{X}_i\}} \\
&\vdots \\
P(Y_i = 5) &= P(Y_i = 1) \exp\{\boldsymbol{\beta}_5' \boldsymbol{X}_i\} = \frac{\exp\{\boldsymbol{\beta}_5' \boldsymbol{X}_i\}}{1 + \sum_{k=2}^{5} \exp\{\boldsymbol{\beta}_k' \boldsymbol{X}_i\}}.
\end{aligned}$$

So we obtiant he optimization problem:

$$\min_{\boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_5}} \sum_{i=1}^{n} \sum_{k=1}^{5} \log P(Y_i = k) \mathbf{1}_{\{Y_i = k\}}.$$

Now we fit such a model:

```r
fit <- multinom(HEALTH ~ LPACT + VPACT, data = data_health)
```

```
## # weights:  20 (12 variable)
## initial  value 422376.057424
## iter  10 value 350198.489695
## iter  20 value 340538.315501
## final  value 340536.151421
## converged
```

```r
(results <- summary(fit))
```

```
## Call:
## multinom(formula = HEALTH ~ LPACT + VPACT, data = data_health)
##
## Coefficients:
##   (Intercept)       LPACT      VPACT
## 2   0.4196178  0.01817174 -0.1369965
## 3   0.2977002 -0.01621469 -0.3287100
## 4  -0.4084865 -0.11911179 -0.6138016
## 5  -1.1498592 -0.37261323 -1.0038904
##
## Std. Errors:
##   (Intercept)       LPACT       VPACT
## 2 0.008278215 0.001797869 0.002041267
```

```
## 3 0.009091518 0.002019297 0.003121447
## 4 0.012382896 0.003300016 0.007399451
## 5 0.018540585 0.007489716 0.022520209
##
## Residual Deviance: 681072.3
## AIC: 681096.3
```

After the model fit, we would like to evaluate the significance of coefficients. Here we use Wald test.

```
(1 - pnorm(abs(results$coefficients/results$standard.errors), 0, 1)) * 2
```

```
##   (Intercept)        LPACT VPACT
## 2           0 0.000000e+00     0
## 3           0 8.881784e-16     0
## 4           0 0.000000e+00     0
## 5           0 0.000000e+00     0
```

All p-values are extremely small and we conclude that the model is significant. Instead of evaluating coefficient significance, we are also interested in testing the robustness of the model as a whole. Here we use likelihood ratio test by comparing the proposed model to an empty model:

```
fit2 <- multinom(HEALTH ~ 1, data = data_health)
```

```
## # weights:  10 (4 variable)
## initial  value 422376.057424
## iter  10 value 357412.922574
## final  value 357412.751928
## converged
```

```
chi_stat <- -fit$deviance - (-fit2$deviance)
1 - pchisq(chi_stat, df = 2)
```

```
## [1] 0
```

As we can see that the p-value is also small. The proposed model is strong.