C240424_9

YOUSEF IBRAHIM GOMAA MAHMOUD MABROUK

2025-01-11

Exercise 7-3

a.
```
MBTI <- read.table("MBTI_Ex7_3.dat", header=TRUE)
```

b.
```
M1.fit <- glm(y ~ factor(EI) + factor(SN) + factor(TF) + factor(JP),
                    family = binomial(link="logit"),
                    data = MBTI)
summary(M1.fit)
```

```
##
## Call:
## glm(formula = y ~ factor(EI) + factor(SN) + factor(TF) + factor(JP),
##     family = binomial(link = "logit"), data = MBTI)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1140     0.2715  -7.788 6.82e-15 ***
## factor(EI)i  -0.5550     0.2170  -2.558  0.01053 *
## factor(SN)s  -0.4292     0.2340  -1.834  0.06663 .
## factor(TF)t   0.6873     0.2206   3.116  0.00184 **
## factor(JP)p   0.2022     0.2266   0.893  0.37209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 646.83  on 1049  degrees of freedom
## Residual deviance: 627.49  on 1045  degrees of freedom
## AIC: 637.49
##
## Number of Fisher Scoring iterations: 5
```

- Used factor(x) for every categorical variable x in order to fit the model. As such, the indicators carry values of 0 or 1, where 1 denotes the latter value of character in alphabetical order by default. Since each variable consists of 2 levels, there exists 1 indicator for each variable.
(e.g. The indicator EI is "i" at EI=1 and EI is "e" at EI=0, while SN is "s" at SN=1 and SN is "n" at SN=0... and so on)
- The estimated prediction equation is as follows:
$\hat{\pi} = \hat{P}(Y = 1) = -2.114 - 0.5550EI - 0.4292SN + 0.6873TF + 0.2022JP$

- Which means that the log odds of success are subtracted by 0.555 per 1 unit increase of EI, subtracted by 0.4292 per 1 unit increase of SN, added by 0.6873 unit increase of TF and 0.2022 by per unit increase of JP.
- In this case, since all variables are categorical and binary. The indicator values can only be either 0 or 1.

c.
```
predicted_probability <- predict(M1.fit,
                                 newdata = data.frame(EI = "e",
                                                      SN = "s",
                                                      TF = "t",
                                                      JP = "j"),
                                 type = "response")
predicted_probability
```

```
##        1
## 0.135186
```

- $\hat{\pi} = \hat{P}(Y = 1|EI = e, SN = n, TF = T, JP = J)$=0.135186

d. The personality type, Extroversion, Intuitive, Thinking and Perceiving (ENTP), has the highest predicted probability as Extroversion and Intuitive remove the effect of the negative indicators EI and SN when substituted in the estimated prediction equation as they are both reference values. On the other hand, Thinking and Perceiving coefficients increase the log-odds, which in turn maximizes the predicted probability of Y=1 (Drinking alcohol frequently).

e. $\because$ The log odds of success (Y=1) increase by 0.6873 per 1 unit increase of TF.
$\therefore$ The odds of success (Y=1) are multiplied by 1.9883398 per 1 unit increase of TF.

f.
```
M2.fit <- glm(y ~ factor(EI) + factor(SN),
                    family = binomial(link="logit"),
                    data = MBTI)
summary(M2.fit)
```

```
##
## Call:
## glm(formula = y ~ factor(EI) + factor(SN), family = binomial(link = "logit"),
##     data = MBTI)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.7605     0.1969  -8.940   <2e-16 ***
## factor(EI)i  -0.5169     0.2155  -2.399   0.0165 *
## factor(SN)s  -0.3902     0.2222  -1.756   0.0790 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 646.83  on 1049  degrees of freedom
## Residual deviance: 637.37  on 1047  degrees of freedom
## AIC: 643.37
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(M2.fit, M1.fit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ factor(EI) + factor(SN)
## Model 2: y ~ factor(EI) + factor(SN) + factor(TF) + factor(JP)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1047     637.37
## 2      1045     627.49  2   9.8877 0.007127 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $\chi^2(1) = 9.89, p = .007$:
- Since the p-value is less than .05, we conclude that including the factors TF and JP significantly improve the model fit. Based on significance alone, M1 is preferred.

g.
```r
M3.fit <- glm(y ~ factor(EI) + factor(SN) + factor(EI):factor(SN),
                    family = binomial(link="logit"),
                    data = MBTI)
summary(M3.fit)
```

```
##
## Call:
## glm(formula = y ~ factor(EI) + factor(SN) + factor(EI):factor(SN),
##     family = binomial(link = "logit"), data = MBTI)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.6740     0.2224  -7.526 5.25e-14 ***
## factor(EI)i               -0.7378     0.3651  -2.021   0.0433 *
## factor(SN)s               -0.5380     0.2940  -1.829   0.0673 .
## factor(EI)i:factor(SN)s    0.3446     0.4540   0.759   0.4478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 646.83  on 1049  degrees of freedom
## Residual deviance: 636.79  on 1046  degrees of freedom
## AIC: 644.79
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(M2.fit, M3.fit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ factor(EI) + factor(SN)
## Model 2: y ~ factor(EI) + factor(SN) + factor(EI):factor(SN)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1047     637.37
## 2      1046     636.79  1  0.58208   0.4455
```

- $\chi^2(1) = 0.582, p = .45$:
- Since the p-value is higher than .05, we are better off not including the interaction effect. Based on significance alone, M2 is preferred.

Exercise 10

1. AIC for Models M1 and M2:

```
AIC(M1.fit)
```

```
## [1] 637.4865
```

```
AIC(M2.fit)
```

```
## [1] 643.3742
```

```
# another method:
loglikm1 <- logLik(M1.fit)
-2*loglikm1+2*attr(loglikm1, "df")
```

```
## 'log Lik.' 637.4865 (df=5)
```

```
loglikm2 <- logLik(M2.fit)
-2*loglikm2+2*attr(loglikm2, "df")
```

```
## 'log Lik.' 643.3742 (df=3)
```

2. Comparison

```
AIC(M1.fit, M2.fit)
```

```
##        df      AIC
## M1.fit  5 637.4865
## M2.fit  3 643.3742
```

- M1 has lesser AIC value, therefore it is the better/more preferable model to use.

3.
```
library(MASS)
stepAIC(M1.fit, direction="backward")
```

```
## Start:  AIC=637.49
## y ~ factor(EI) + factor(SN) + factor(TF) + factor(JP)
##
##              Df Deviance    AIC
## - factor(JP)  1   628.28 636.28
## <none>            627.49 637.49
## - factor(SN)  1   630.77 638.77
## - factor(EI)  1   634.08 642.08
## - factor(TF)  1   637.14 645.14
```

```
##
## Step:  AIC=636.28
## y ~ factor(EI) + factor(SN) + factor(TF)
##
##              Df Deviance    AIC
## <none>            628.28 636.28
## - factor(SN)  1   632.74 638.74
## - factor(EI)  1   634.81 640.81
## - factor(TF)  1   637.37 643.37
```

```
##
## Call:  glm(formula = y ~ factor(EI) + factor(SN) + factor(TF), family = binomial(link = "logit")
##     data = MBTI)
##
## Coefficients:
## (Intercept)  factor(EI)i  factor(SN)s  factor(TF)t
##     -1.9678      -0.5518      -0.4843       0.6601
##
## Degrees of Freedom: 1049 Total (i.e. Null);  1046 Residual
## Null Deviance:       646.8
## Residual Deviance: 628.3     AIC: 636.3
```

- Using backward elimination, the factors that should be selected in the final model are: SN, EI and TF. As the AIC value has reached a minimum of 636.3 using said factors. That is assuming no interactions are made.
- The resultant estimated prediction equation is:
  $\hat{\pi} = -1.9678 - 0.5518EI - 0.4843SN + 0.6601TF$