Exercise for Lecture 9-10

Rei Monden

Exercise 7-3

a.

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

b.

```
M1 <- glm(y ~ EI + SN + TF + JP,
          family  = binomial,
          data    = MBTI)
M1
```

```
##
## Call:  glm(formula = y ~ EI + SN + TF + JP, family = binomial, data = MBTI)
##
## Coefficients:
## (Intercept)          EIi          SNs          TFt          JPp
##     -2.1140      -0.5550      -0.4292       0.6873       0.2022
##
## Degrees of Freedom: 1049 Total (i.e. Null);  1045 Residual
## Null Deviance:       646.8
## Residual Deviance: 627.5     AIC: 637.5
```

R internally created indicator variables with 1s for Introversion (variable EI), Sensing (variable SN), Thinking (variable TF), and Perceiving (variable JP).

The estimated prediction equation for $\hat{\pi}$ is

$$\hat{\pi} = -2.11 - 0.56\text{Introversion} - 0.43\text{Sensing} + 0.69\text{Thinking} + 0.20\text{Perceiving}.$$

c. $\hat{\pi} = \frac{\exp(-2.11-0.56(0)-0.43(1)+0.69(1)+0.20(0))}{1+\exp(-2.11-0.56(0)-0.43(1)+0.69(1)+0.20(0))} = 0.14.$

Also, using the `predict()` function:

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

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

d. Consider which coefficients are positive and which are negative. Each negative coefficient should be multiplied by 0 from the corresponding indicator variable. This means we need to consider groups Extroversion (variable EI) and Intuitive (variable SN). Similarly, each positive coefficient should be multiplied by 1 from the corresponding indicator variable. This means we need to consider groups Thinking (variable TF) and Perceiving (variable JP). Therefore, the highest estimated probability corresponds to group (Extroversion, Intuitive, Thinking, Perceiving) and is equal to

$$\hat{\pi} = \frac{\exp(-2.11 - 0.56(0) - 0.43(0) + 0.69(1) + 0.20(1))}{1 + \exp(-2.11 - 0.56(0) - 0.43(0) + 0.69(1) + 0.20(1))} = 0.23.$$

Or via `predict()`:

```
predict(M1,
        newdata = data.frame(EI = "e", SN = "n", TF = "t", JP = "p"),
        type    = "response")
```

```
##         1
## 0.2271486
```

e. $\exp(0.6873) = 1.99$, thus conditional on the other three predictors (i.e., keeping them fixed), the estimated odds that a person in the Thinking group reports drinking alcohol frequently is 1.99 times the estimated odds that a person in the Feeling group reports drinking alcohol frequently.

f.

```
M2 <- glm(y ~ EI + SN,
          family  = binomial,
          data    = MBTI)
anova(M2, M1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ EI + SN
## Model 2: y ~ EI + SN + TF + 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
```

At 5% significance level, the test result is statistically significant ($\chi^2(2) = 9.89$, $p = .007$). We conclude that, controlling for EI and SN, predictors TF and JP have a significant effect on reporting to drink alcohol frequently.

g.

```
M3 <- glm(y ~ EI * SN,
          family  = binomial,
          data    = MBTI)
anova(M2, M3, test = "LRT")
```

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

At 5% significance level, the test result is not statistically significant ($\chi^2(1) = 0.58$, $p = .45$). We conclude that, controlling for EI and SN, their interaction effect does not have a significant effect on reporting to drink alcohol frequently.

h.

```
library(pROC)
M1.ROC <- roc(y ~ fitted(M1),
              data = MBTI)
```

```
auc(M1.ROC)
```

```
## Area under the curve: 0.6405
```

Exercise 10

1.

```
MBTI <- read.table("Datasets/MBTI.dat", header = TRUE)

M1 <- glm(y ~ EI + SN + TF + JP,
          family  = binomial,
          data    = MBTI)
M2 <- glm(y ~ EI + SN,
          family  = binomial,
          data    = MBTI)
AIC(M1,M2)
```

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

2. We select the smallest AIC, therefore M_1 is preferable over M_2

3.

```
library(MASS)
stepAIC(M1)
```

```
## Start:  AIC=637.49
## y ~ EI + SN + TF + JP
##
```

```
##        Df Deviance    AIC
## - JP    1   628.28 636.28
## <none>      627.49 637.49
## - SN    1   630.77 638.77
## - EI    1   634.08 642.08
## - TF    1   637.14 645.14
##
## Step:  AIC=636.28
## y ~ EI + SN + TF
##
##        Df Deviance    AIC
## <none>      628.28 636.28
## - SN    1   632.74 638.74
## - EI    1   634.81 640.81
## - TF    1   637.37 643.37
```

```
##
## Call:  glm(formula = y ~ EI + SN + TF, family = binomial, data = MBTI)
##
## Coefficients:
## (Intercept)          EIi          SNs          TFt
##     -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
```

Based on backward elimination, the selected predictors are EI, SN and TF.