a3

Hanxiao Du, Jeffery Wei Xuan Su

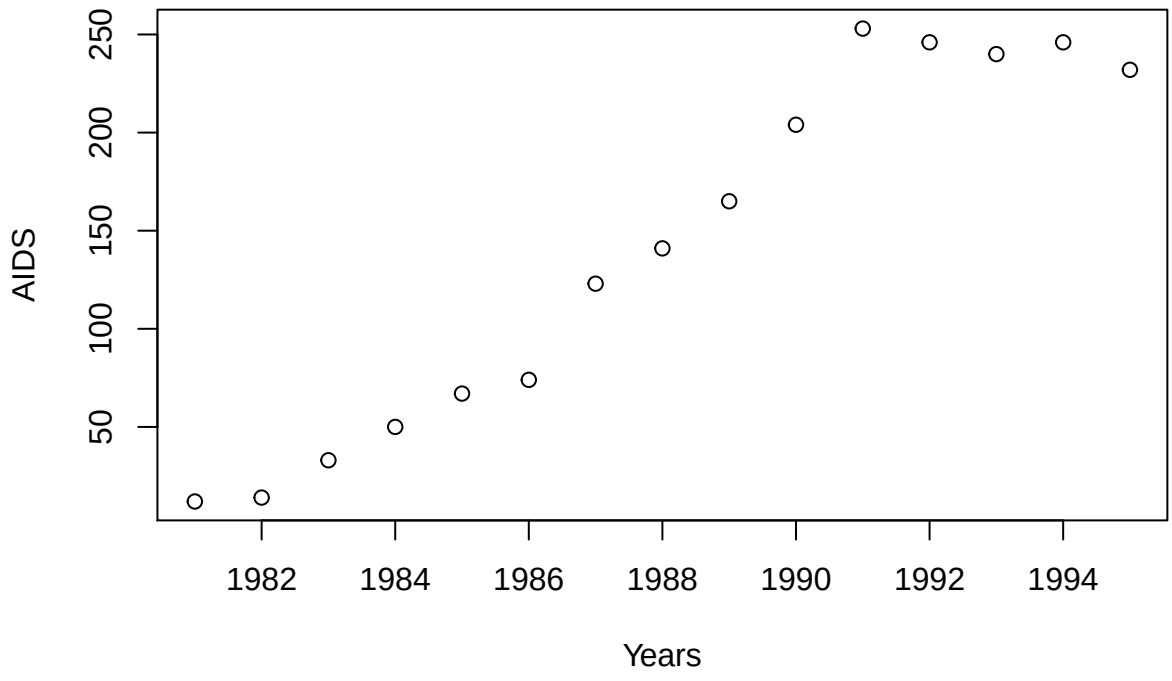
11/9/2020

Q1.

```
Year = c(1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995)
AIDS = c(12, 14, 33, 50, 67, 74, 123, 141, 165, 204, 253, 246, 240, 246, 232)
aids = data.frame(Year, AIDS)
```

(a)

```
plot(aids$Year, aids$AIDS,
     ylab = "AIDS",
     xlab = "Years")
```



Comment: From year 1981 to 1991, it is an exponential increase in the number of new AIDS each year. From year 1991 to 1995, it started to decrease gradually in the number of AIDS each year.

(b)

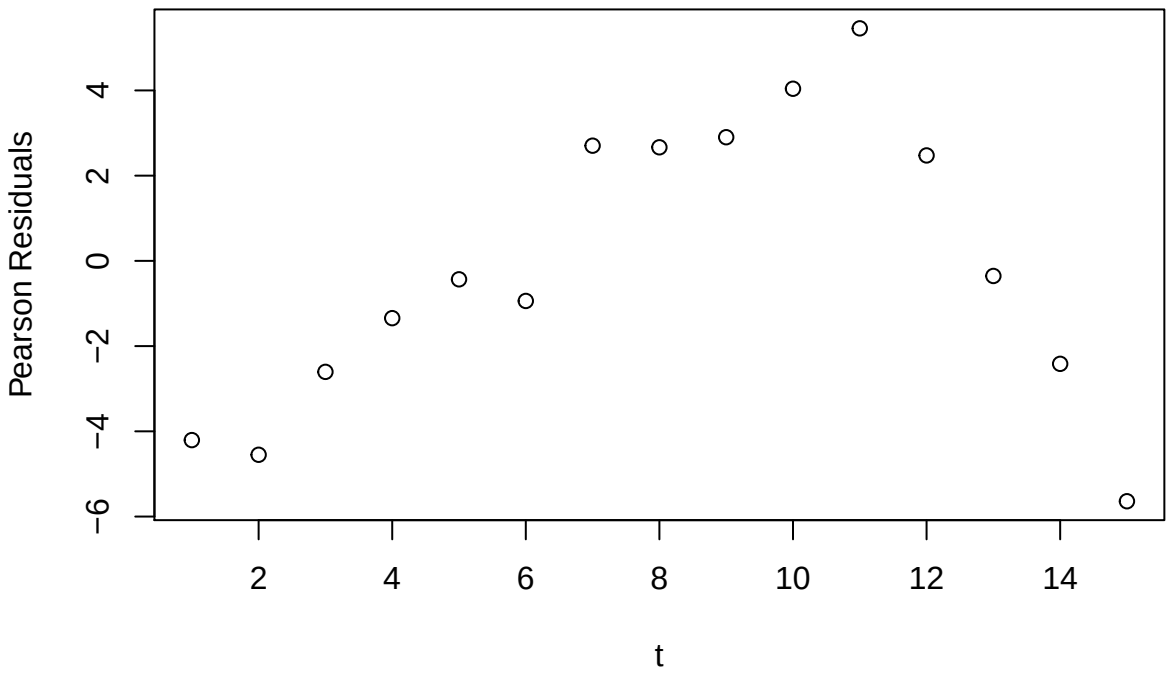
```
aids.trimmed = sweep(aids, 2, c(1980, 0))
colnames(aids.trimmed) = c("t", "AIDS")

aids.trimmed.log = glm(formula = aids.trimmed$AIDS ~ aids.trimmed$t,
                       family = poisson(link="log"))
summary(aids.trimmed.log)
```

```
##
## Call:
## glm(formula = aids.trimmed$AIDS ~ aids.trimmed$t, family = poisson(link = "log"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9751  -2.6345  -0.4367   2.5776   5.1378
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.478884   0.064975   53.54   <2e-16 ***
## aids.trimmed$t 0.155739   0.005735   27.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1010.27  on 14  degrees of freedom
## Residual deviance:  173.33  on 13  degrees of freedom
## AIC: 273.65
##
## Number of Fisher Scoring iterations: 4
```

(c)

```
plot(aids.trimmed$t, residuals(aids.trimmed.log, type="pearson"), xlab="t",
     ylab="Pearson Residuals")
```



Comment: By the plot above, there is overdispersion, since the value of Pearson's residual is not around 1.
(d)

```
t.tbar = aids.trimmed$t - mean(aids.trimmed$t)
aids.quad = glm(formula = aids.trimmed$AIDS ~ poly(t.tbar, 2, raw = TRUE),
                family=poisson(link = "log"))
summary(aids.quad)
```

```
##
## Call:
## glm(formula = aids.trimmed$AIDS ~ poly(t.tbar, 2, raw = TRUE),
##     family = poisson(link = "log"))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.45122  -0.54143   0.03733   0.56349   1.54168
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   4.982168   0.032545  153.08   <2e-16 ***
## poly(t.tbar, 2, raw = TRUE)1  0.214565   0.008816   24.34   <2e-16 ***
## poly(t.tbar, 2, raw = TRUE)2 -0.021221   0.001775  -11.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 1010.2722  on 14  degrees of freedom
## Residual deviance:    9.2446  on 12  degrees of freedom
## AIC: 111.56
##
## Number of Fisher Scoring iterations: 4
```

(e)

```
aids.affine = glm(formula = aids.trimmed$AIDS ~ poly(t.tbar, 1, raw = TRUE),
                  family=poisson(link = "log"))
anova(aids.affine, aids.quad, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: aids.trimmed$AIDS ~ poly(t.tbar, 1, raw = TRUE)
## Model 2: aids.trimmed$AIDS ~ poly(t.tbar, 2, raw = TRUE)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        13    173.335
## 2        12      9.245  1   164.09 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let $\alpha = 0.05$, since the p-value $= 2.2 \times 10^{-16} < 0.05 = \alpha$, we reject $H_0 : \beta_2 = 0$, thus we believe that the simpler model (i.e. $log(\mu(t)) = \alpha + \beta_1(t - \bar{t})$) does not fit the data well compare to the complex model (i.e. $log(\mu(t)) = \alpha + \beta_1(t - \bar{t}) + \beta_2(t - \bar{t})^2$).

Q2.

```
PayYes = c(24,10,5,16,7,47,45, 57,54,59)
PayNo = c(9,3,4,7,4,12,8,9,10,12)
District = rep(c("NC", "NE", "NW", "SE", "SW"),2)
Race = c(rep("Blacks",5), rep("Whites",5))
merit = data.frame(Race, District, PayYes, PayNo)
print(merit)
```

```
##      Race District PayYes PayNo
## 1  Blacks       NC     24     9
## 2  Blacks       NE     10     3
## 3  Blacks       NW      5     4
## 4  Blacks       SE     16     7
## 5  Blacks       SW      7     4
## 6  Whites       NC     47    12
## 7  Whites       NE     45     8
## 8  Whites       NW     57     9
## 9  Whites       SE     54    10
## 10 Whites       SW     59    12
```

(a)

```
total = PayYes+PayNo
Y = PayYes/total
merit.fit = glm(Y ~ Race+District, weight=total, family=binomial(link="logit"))
summary(merit.fit)
```

```
##
## Call:
## glm(formula = Y ~ Race + District, family = binomial(link = "logit"),
##     weights = total)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  0.60191   0.30311  -0.97042  -0.09608  -0.30707  -0.53319  -0.18583   0.47422
##        9        10
##  0.07216   0.15054
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.74947    0.29581   2.534  0.01129 *
## RaceWhites   0.79129    0.28532   2.773  0.00555 **
## DistrictNE   0.25837    0.42067   0.614  0.53909
## DistrictNW   0.13836    0.40517   0.341  0.73273
## DistrictSE   0.12087    0.37287   0.324  0.74581
## DistrictSW   0.00445    0.38486   0.012  0.99077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10.665  on 9  degrees of freedom
## Residual deviance:  2.071  on 4  degrees of freedom
## AIC: 49.437
##
## Number of Fisher Scoring iterations: 4
```

```
drop1(merit.fit, test="Chisq")
```

```
## Single term deletions
##
## Model:
## Y ~ Race + District
##          Df Deviance    AIC    LRT Pr(>Chi)
## <none>        2.0710 49.437
## Race      1   9.4624 54.828 7.3915 0.006553 **
## District  4   2.5876 41.953 0.5167 0.971859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With H_0 :independent. Since the p-value from both test for Race is much smaller then $\alpha = 0.05$, we reject the H_0, so the merit pay increase is independent of race. Since the p-value from both test for District, is much bigger then $\alpha = 0.05$, we don't have enough information to reject H_0, so the merit pay is confitional on the district.

(b) The estimate of the common odds ratio between Merit Pay and Race is $\frac{e^{\alpha+\beta_W+\beta_{NE}+\beta_{NW}+\beta_{SE}+\beta_{SW}}}{e^{\alpha+\beta_{NE}+\beta_{NW}+\beta_{SE}+\beta_{SW}}} = e^{\beta_W} = e^{0.79129} = 2.206241$

```
# 95% Wald C.I. for common odds ratio
exp(confint.default(merit.fit))
```

```
##                 2.5 %   97.5 %
## (Intercept) 1.1849325 3.778236
## RaceWhites  1.2612097 3.859347
## DistrictNE  0.5677217 2.953110
## DistrictNW  0.5190500 2.540811
## DistrictSE  0.5433854 2.343579
## DistrictSW  0.4724281 2.135648
```

```
# 95% LR C.I. for common odds ratio
exp(confint(merit.fit))
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %   97.5 %
## (Intercept) 1.2015339 3.853896
## RaceWhites  1.2519737 3.844984
## DistrictNE  0.5755038 3.033447
## DistrictNW  0.5224551 2.585144
## DistrictSE  0.5437179 2.363742
## DistrictSW  0.4725999 2.154622
```

The 95% Wald C.I. for the common odds ratio between Merit Pay and Race is (1.2612097, 3.859347) which does not contains 1, so there is a statistically significant relationship between races and the probability of getting a merit pay increase. The 95% LR C.I. for the common odds ratio between Merit Pay and Race is (1.2519737 3.844984) which does not contains 1, so there is a statistically significant relationship between races and the probability of getting a merit pay increase.

(c)

```
merit.fit1 = glm(cbind(PayYes, PayNo) ~ Race+District,
                 family=binomial(link="logit"))
merit.fit2 = glm(cbind(PayYes, PayNo) ~ Race+District + District:Race,
                 family=binomial(link="logit"))
anova(merit.fit1, merit.fit2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(PayYes, PayNo) ~ Race + District
## Model 2: cbind(PayYes, PayNo) ~ Race + District + District:Race
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         4      2.071
## 2         0      0.000  4    2.071   0.7227
```

Since the p-value = 0.7227 is greater then $\alpha = 0.05$, so we fail to reject the H_0. Therefore homogeneous association is valid.

Q3.

(a)

```
MBTI = read.table("MBTI.txt", header = T)
MBTI$drink_false = MBTI$n - MBTI$drink
MBTI.logit = glm(cbind(drink, drink_false) ~ EI+SN+TF+JP,
                 family = binomial(link= "logit"), data = MBTI)
summary(MBTI.logit)
```

```
##
## Call:
## glm(formula = cbind(drink, drink_false) ~ EI + SN + TF + JP,
##     family = binomial(link = "logit"), data = MBTI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2712  -0.8062  -0.1063   0.1124   1.5807
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1140     0.2715  -7.788 6.82e-15 ***
## EIi          -0.5550     0.2170  -2.558  0.01053 *
## SNs          -0.4292     0.2340  -1.834  0.06664 .
## TFt           0.6873     0.2206   3.116  0.00184 **
## JPp           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: 30.488  on 15  degrees of freedom
## Residual deviance: 11.149  on 11  degrees of freedom
## AIC: 73.99
##
## Number of Fisher Scoring iterations: 4
```

The prediction equation is $\hat{\pi}(x) = \frac{e^{-2.1140-0.5550I-0.4292S+0.6873T+0.2022P}}{1+e^{-2.1140-0.5550I-0.4292S+0.6873T+0.2022P}}$

The indicator variables are $I = \mathbb{I}(EI = i)$, $S = \mathbb{I}(SN = s)$, $= T = \mathbb{I}(TF = t)$ and $P = \mathbb{I}(JP = p)$

(b) ENTP personality type has the highest estimated probability, to maximize $\hat{\pi}(x)$, let the indicator variables with negative coefficients be 0 and others be 1.

Q4.

(a)

```
p.value = pchisq(MBTI.logit$deviance, df=11, lower.tail = F)
print(p.value)
```

```
## [1] 0.4308605
```

Since p-value $= 0.4308605 > 0.05 = \alpha$, we have no strong evidence to reject H_0, thus the model fits the data well. I would remove JP since the p-value of JP is the greatest.

(b)

```
MBTI.fit1 = glm(cbind(drink, drink_false) ~
                  EI+SN+TF+JP+EI:SN+EI:TF+EI:JP+SN:TF+SN:JP+TF:JP,
                family = binomial(link= "logit"), data = MBTI)
summary(MBTI.fit1)
```

```
##
## Call:
## glm(formula = cbind(drink, drink_false) ~ EI + SN + TF + JP +
##     EI:SN + EI:TF + EI:JP + SN:TF + SN:JP + TF:JP, family = binomial(link = "logit"),
##     data = MBTI)
##
## Deviance Residuals:
##       1         2         3         4         5         6         7         8
## -0.09316   0.56452  -0.43696  -0.03168   0.13900  -0.84856   0.62519   0.00661
##       9        10        11        12        13        14        15        16
##  0.11129  -0.79249   0.37773   0.11962  -0.34909   0.77692  -0.75044  -0.07286
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.25933    0.47184  -4.788 1.68e-06 ***
## EIi         -0.45894    0.55219  -0.831    0.406
## SNs         -0.55200    0.50880  -1.085    0.278
## TFt          0.27522    0.56135   0.490    0.624
## JPp          0.79110    0.49089   1.612    0.107
## EIi:SNs      0.01767    0.50769   0.035    0.972
## EIi:TFt      0.30405    0.46929   0.648    0.517
## EIi:JPp     -0.54072    0.48426  -1.117    0.264
## SNs:TFt      0.66547    0.50842   1.309    0.191
## SNs:JPp     -0.29800    0.51578  -0.578    0.563
## TFt:JPp     -0.29654    0.48354  -0.613    0.540
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.4880  on 15  degrees of freedom
## Residual deviance:  3.7409  on  5  degrees of freedom
## AIC: 78.582
##
## Number of Fisher Scoring iterations: 4
```

```
MBTI.fit2 = MBTI.logit
summary(MBTI.fit2)
```

```
##
## Call:
## glm(formula = cbind(drink, drink_false) ~ EI + SN + TF + JP,
##     family = binomial(link = "logit"), data = MBTI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.2712  -0.8062  -0.1063   0.1124   1.5807
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1140     0.2715  -7.788 6.82e-15 ***
## EIi          -0.5550     0.2170  -2.558  0.01053 *
## SNs          -0.4292     0.2340  -1.834  0.06664 .
## TFt           0.6873     0.2206   3.116  0.00184 **
## JPp           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: 30.488  on 15  degrees of freedom
## Residual deviance: 11.149  on 11  degrees of freedom
## AIC: 73.99
##
## Number of Fisher Scoring iterations: 4
```

```
MBTI.fit3 =  glm(cbind(drink, drink_false) ~ 1,
                 family = binomial(link= "logit"), data = MBTI)
summary(MBTI.fit3)
```

```
##
## Call:
## glm(formula = cbind(drink, drink_false) ~ 1, family = binomial(link = "logit"),
##     data = MBTI)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.19328  -1.02200   0.07511   0.99097   2.61603
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2849     0.1066  -21.44   <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: 30.488  on 15  degrees of freedom
## Residual deviance: 30.488  on 15  degrees of freedom
## AIC: 85.329
##
## Number of Fisher Scoring iterations: 4
```

The AIC value for model with four main effects and six interaction terms is 78.582. The AIC value for model with only four main effect is 73.99. The AIC value for model with no predictors is 85.329.

Base on the AIC model selection criterion, we should the model with the lowest AIC. So model with only four main effect is preferred. By using AIC, we can compare models that neither is a special case of the other.

(c)

```
step(glm(cbind(drink, drink_false) ~ 1,
         family = binomial(link= "logit"), data = MBTI),
     scope=~EI*SN*TF*JP, direction = "forward", test="Chisq")
```

```
## Start:  AIC=85.33
## cbind(drink, drink_false) ~ 1
##
##        Df Deviance    AIC    LRT Pr(>Chi)
## + TF    1   23.683 80.523 6.8054 0.009088 **
## + EI    1   24.036 80.877 6.4521 0.011082 *
## + SN    1   26.832 83.673 3.6555 0.055885 .
## <none>      30.488 85.329
## + JP    1   29.508 86.348 0.9804 0.322095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=80.52
## cbind(drink, drink_false) ~ TF
##
##        Df Deviance    AIC    LRT Pr(>Chi)
## + EI    1   16.398 75.239 7.2847 0.006955 **
## + SN    1   18.469 77.310 5.2135 0.022413 *
## + JP    1   21.631 80.472 2.0519 0.152021
## <none>      23.683 80.523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=75.24
## cbind(drink, drink_false) ~ TF + EI
##
##         Df Deviance    AIC    LRT Pr(>Chi)
## + SN     1   11.945 72.786 4.4525  0.03485 *
## <none>       16.398 75.239
## + JP     1   14.436 75.277 1.9618  0.16132
## + EI:TF  1   14.984 75.825 1.4136  0.23446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=72.79
## cbind(drink, drink_false) ~ TF + EI + SN
##
##         Df Deviance    AIC    LRT Pr(>Chi)
## + SN:TF  1   8.2328 71.074 3.7127   0.0540 .
## <none>      11.9455 72.786
## + EI:TF  1  10.5461 73.387 1.3994   0.2368
## + JP     1  11.1491 73.990 0.7964   0.3722
## + EI:SN  1  11.3814 74.222 0.5641   0.4526
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=71.07
## cbind(drink, drink_false) ~ TF + EI + SN + TF:SN
```

```
##
##         Df Deviance    AIC     LRT Pr(>Chi)
## <none>       8.2328 71.074
## + EI:TF  1   7.0895 71.930 1.14332   0.2850
## + JP     1   7.4797 72.321 0.75310   0.3855
## + EI:SN  1   7.8198 72.661 0.41304   0.5204
```

```
##
## Call:  glm(formula = cbind(drink, drink_false) ~ TF + EI + SN + TF:SN,
##     family = binomial(link = "logit"), data = MBTI)
##
## Coefficients:
## (Intercept)          TFt          EIi          SNs      TFt:SNs
##    -1.76795      0.07959     -0.55499     -0.86844      0.89962
##
## Degrees of Freedom: 15 Total (i.e. Null);  11 Residual
## Null Deviance:       30.49
## Residual Deviance: 8.233     AIC: 71.07
```

Thus the selected model is:

```
fit = glm(formula = cbind(drink, drink_false) ~ TF + EI + SN + TF:SN, family = binomial(link = "logit")
summary(fit)
```

```
##
## Call:
## glm(formula = cbind(drink, drink_false) ~ TF + EI + SN + TF:SN,
##     family = binomial(link = "logit"), data = MBTI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2673  -0.5975  -0.2545   0.5777   1.0198
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.76795    0.22630  -7.812 5.61e-15 ***
## TFt          0.07959    0.38550   0.206  0.83644
## EIi         -0.55499    0.21731  -2.554  0.01065 *
## SNs         -0.86844    0.30356  -2.861  0.00423 **
## TFt:SNs      0.89962    0.47632   1.889  0.05894 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.4880  on 15  degrees of freedom
## Residual deviance:  8.2328  on 11  degrees of freedom
## AIC: 71.074
##
## Number of Fisher Scoring iterations: 4
```