HW6

Nikhil Gopal

4/13/2021

Question 1

```
setwd("C:/Users/d/Google Drive/Notability/Categorical Data/psets/HW6")

rm(list = ls())
```

1a:

In cumulative logit models with the proportional odds property, the interpretation of the coefficients must be flipped since the regression output gives the log odds of being in a higher category. With this in mind

For x1, the coefficient was -0.54, indicating that the response level increases with an increase in x1. Thus, job satisfaction increases at higher x1 (earnings compared to similar positions = much less) increase.

For x2, the coefficient was 0.60, indicating that the response level will increase as x2 decreases Thus, job satisfaction increases at lower x2 (freedom to make decisions).

For x3, the coefficient was 1.19, again indicating that the response level will increase as x3 decreases. Thus job satisfaction increases at lower x3 (work environment allows productivity).

1b:

Since the coefficient is negative for x1, job satisfaction is highest at the highest category (x1 = 4), when earnings are much more than others in similar positions.

Since the coefficient is positive for x2, job satisfaction is highest at the lowest category (x2 = 1), when an individual is very free to make decisions on how to do a job.

Since the coefficient is positive for x3, job satisfaction is highest at the lowest category (x3 = 1), when they strongly agree that the work environment allows productivity.

Thus the settings for x1,x2 and x3 where an individual will have the highest job satisfaction are x1 =4, x2 =1 and x3=1.

1c:

This equation is the reciprocal of the prediction equation, so it will be the same equation with the signs of the slopes being flipped.

Question 2

2a:

```
library(VGAM)
```

```
## Warning: package 'VGAM' was built under R version 4.0.4
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
Not <- c(6,6,6);
Pretty <- c(43,113,57);
Very <- c(75,178,117);
Income <- c("Below", "Average", "Above")
data.frame(Income, Not, Pretty, Very)
```

```
##    Income Not Pretty Very
## 1   Below   6     43   75
## 2 Average   6    113  178
## 3   Above   6     57  117
```

```
scores<-c(1,2,3)

logit_mod <- vglm(cbind(Not, Pretty, Very) ~ scores, family = cumulative(parallel=TRUE))

summary(logit_mod)
```

```
##
## Call:
## vglm(formula = cbind(Not, Pretty, Very) ~ scores, family = cumulative(parallel = TRUE))
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -3.2466     0.3404  -9.537   <2e-16 ***
## (Intercept):2  -0.2378     0.2592  -0.917    0.359
## scores         -0.1117     0.1179  -0.948    0.343
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 3.2472 on 3 degrees of freedom
##
## Log-likelihood: -15.4146 on 3 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##    scores
## 0.8942746
```

Prediction equations:

j = 1:

$$logit[P(\hat{Y} \leq 1)] = -3.2466 - 0.1117 * score$$

j = 2:

$$logit[P(\hat{Y} \leq 2)] = -0.2378 - 0.1117 * score$$

2b:

For any fixed J, an increase in income from a lower to a higher score group will increase the odds of being happy by exp(-0.1117423) = 1.118225.

2c:

$$Ho = \beta = 0$$

$$Ha = \beta! = 0$$

A slope of zero would indicate that income has no effect on marital happiness, and a non-zero slope would indicate that income has has an effect on marital happiness, and that the two variables are not independent.

```
lrtest(logit_mod)
```

```
## Likelihood ratio test
##
## Model 1: cbind(Not, Pretty, Very) ~ scores
## Model 2: cbind(Not, Pretty, Very) ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -15.415
## 2   4 -15.858  1 0.8876     0.3461
```

The likelihood ratio test returned a Chi-Square statistic of 0.8876 on 1 DF and a p-value of 0.3461. Since our p value is higher than the typical 0.05% significance level, we do not have enough evidence to reject our null hypothesis, and cannot conclude that income has an effect on marital happiness.

2d:

The residual deviance of our model is 3.2472 on 3 DF. We can run a deviance goodness of fit test:

Ho : Model fits adequately

Ha: Model does not fit adequately

```
1-pchisq(3.2472, 3)
```

```
## [1] 0.3550593
```

Our deviance goodness of fit test returned a p value of 0.3550593. Since the p value is greater than 0.05, we can confidently say that the model fits adequately. 2e:

```
predict(logit_mod, data.frame(scores = 2), type="response")
```

```
##          Not    Pretty      Very
## 1 0.03017419 0.3565176 0.6133082
```

The probability that someone with average family income reports a very happy marriage is 0.6133082.

Question 3

```r
data <- matrix(c(833,125,2,160),ncol=2,byrow=TRUE)
colnames(data) <- c("Hell (Yes)","Hell (No)")
rownames(data) <- c("Heaven (Yes)","Heaven (No)")
Heaven_Hell <- as.table(data)
data
```

```
##              Hell (Yes) Hell (No)
## Heaven (Yes)        833       125
## Heaven (No)           2       160
```

3a

Ho: Population proportions answering yes are the same for heaven/hell.

Ha: Population proportions answering yes are not the same for heaven/hell.

Since the data is paired and nominal, a McNemar's test is appropriate.

```r
mcnemar.test(data, correct = F)
```

```
##
##  McNemar's Chi-squared test
##
## data:  data
## McNemar's chi-squared = 119.13, df = 1, p-value < 2.2e-16
```

Our hypothesis test returned a chi-square statistic of 119.13 and a p value of < 2.2 e^-16 or near extremely close to zero. We therefore reject our null hypothesis, and have enough evidence to conclude that the population proportions are not the same for answering yes for Heaven and Hell.

3b:

```r
SE <- 1/1120 * sqrt( 125 + 2 - (125-2)^2 / 1120)

lower_bound  <- ((125 - 2) / 1120) - 1.64 * SE

upper_bound  <- ((125 - 2) / 1120) + 1.64 * SE

lower_bound
```

```
## [1] 0.09422201
```

```r
upper_bound
```

```
## [1] 0.1254208
```

The 90 % confidence interval for the true difference in proportions for those who believe in Heaven and Hell is [0.09422201, 0.1254208]. We are 90% confident that the true difference in proportions lies within the interval.

Question 4

4a:

```r
dat <- read.table("http://users.stat.ufl.edu/~aa/cat/data/DeathPenalty.dat", header = TRUE)

log_lin <- glm(count ~ (D+V+P)^2, family = "poisson", data = dat)

summary(log_lin)
```

```
##
## Call:
## glm(formula = count ~ (D + V + P)^2, family = "poisson", data = dat)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  0.02505  -0.00895  -0.05463   0.03000  -0.60362   0.04572   0.09251  -0.01545
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.93578    0.08471  58.265  < 2e-16 ***
## Dwhite        -2.17465    0.26377  -8.245  < 2e-16 ***
## Vwhite        -1.32980    0.18479  -7.196 6.19e-13 ***
## Pyes          -3.59610    0.50691  -7.094 1.30e-12 ***
## Dwhite:Vwhite  4.59497    0.31353  14.656  < 2e-16 ***
## Dwhite:Pyes   -0.86780    0.36707  -2.364   0.0181 *
## Vwhite:Pyes    2.40444    0.60061   4.003 6.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1225.07955  on 7  degrees of freedom
## Residual deviance:    0.37984  on 1  degrees of freedom
## AIC: 52.42
##
## Number of Fisher Scoring iterations: 3
```

Goodness of fit test:

```r
1-pchisq(0.37984, 1)
```

```
## [1] 0.5376889
```

The deviance Goodness of Fit test returned a p value of 0.5376889, using a residual deviance of 0.37984 on 1 DF. Since the p value is above 0.05, we are confident that the model fits the data adequately.

4b:

The predicted conditional odds ratio between D and P at each category of V is exp(-0.8678) = 0.4198743. If the race variable is kept constant, the estimated odds of a white defendant getting the death penalty is 0.4198743 that of a black defendant.

4c:

odds ratio calculation:

```r
((53+0) / (414+16)) / ((11+4) / (139+37))
```

```
## [1] 1.446202
```

The marginal odds ratio between D and P is 1.446202, meaning that the estimated odds of a white person receiving the death penalty is 1.446202 times higher compared to a black person.

Simpson's paradox is when trends are observed in groups of data but disappear or reverse when groups are combined. This is exactly what happened in this case, as marginally whites have a higher chance of receiving the death penalty, but when controlled for race, it is clear that blacks have a higher chance of receiving the death penalty.

4d:

```r
dataaa <- data.frame("D" = c("white", "black", "white", "black"),
                     "V" =c("white", "white", "black", "black"),
                     "Y" = c(53, 11, 0, 4),
                     "N" = c(414, 37, 16, 139))


logistic <- glm(cbind(Y, N) ~ D+V, data = dataaa, family = "binomial")

summary(logistic)
```

```
##
## Call:
## glm(formula = cbind(Y, N) ~ D + V, family = "binomial", data = dataaa)
##
## Deviance Residuals:
##        1         2         3         4
##  0.02660  -0.06232  -0.60535   0.09379
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.5961     0.5069  -7.094 1.30e-12 ***
## Dwhite       -0.8678     0.3671  -2.364   0.0181 *
## Vwhite        2.4044     0.6006   4.003 6.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22.26591  on 3  degrees of freedom
## Residual deviance:  0.37984  on 1  degrees of freedom
## AIC: 19.3
##
## Number of Fisher Scoring iterations: 4
```

Prediction equation:

$$logit(\hat{\pi}) = -3.5961 - 0.8678 * Dwhite + 2.4044 * Vwhite$$

The coefficient for defendant race = white was -0.8678 and the coefficient for victim race = white was 2.4044. These are the same as the coefficients in the loglinear model.

Question 5

5a:

```
#loglinear model of mutual independence
mbti <- read.table("http://users.stat.ufl.edu/~aa/intro-cda/data/MBTI.dat",
header = T)

log_lin_mbti <- glm(n ~ EI + SN + TF + JP, family = poisson, data = mbti)

summary(log_lin_mbti)
```

```
##
## Call:
## glm(formula = n ~ EI + SN + TF + JP, family = poisson, data = mbti)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3550  -2.1182  -1.0628   0.8506   5.7457
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.79255    0.07674  49.422  < 2e-16 ***
## EIi          0.26439    0.06226   4.246 2.17e-05 ***
## SNs          0.87008    0.06765  12.861  < 2e-16 ***
## TFt         -0.48551    0.06355  -7.640 2.17e-14 ***
## JPp         -0.12971    0.06185  -2.097    0.036 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 399.94  on 15  degrees of freedom
## Residual deviance: 135.87  on 11  degrees of freedom
## AIC: 238.7
##
## Number of Fisher Scoring iterations: 4
```

residual deviance GOF test:

```
1 - pchisq(135.87, 11)
```

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

pearson chisquare:

```
1 - pchisq(sum(residuals(log_lin_mbti, type =
"pearson")^2), 11)
```

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

Ho: The model fits adequately

Ha: The model does not fit adequately

The residual deviance of the model is 135.87 on 11 DF. The pearson chi square statistic is 145.1028. Both goodness of fit tests returned p values of 0. Since this is below the 0.05 significance level, we fail to reject the null hypothesis. The model does not fit adequately.

5b:

```
#model of homogenous association
mbti2 <- glm(formula = n ~ (EI + SN + TF + JP)^2, family = "poisson", data
= mbti)

summary(mbti2)
```

```
##
## Call:
## glm(formula = n ~ (EI + SN + TF + JP)^2, family = "poisson",
##     data = mbti)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
## -0.72826   1.00215   0.05168  -0.01429   1.49947  -1.29325  -0.07596   0.00231
##        9        10        11        12        13        14        15        16
##  0.56850  -0.82975  -0.04948   0.01728  -1.57051   1.09960   0.08587  -0.00804
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.44760    0.13793  24.994  < 2e-16 ***
## EIi         -0.02907    0.15266  -0.190 0.848952
## SNs          1.21082    0.14552   8.320  < 2e-16 ***
## TFt         -0.64194    0.16768  -3.828 0.000129 ***
## JPp          0.93417    0.14594   6.401 1.54e-10 ***
## EIi:SNs      0.30212    0.14233   2.123 0.033780 *
## EIi:TFt      0.19449    0.13121   1.482 0.138258
## EIi:JPp      0.01766    0.13160   0.134 0.893261
## SNs:TFt      0.40920    0.15243   2.684 0.007265 **
## SNs:JPp     -1.22153    0.14547  -8.397  < 2e-16 ***
## TFt:JPp     -0.55936    0.13512  -4.140 3.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 399.944  on 15  degrees of freedom
## Residual deviance:  10.162  on  5  degrees of freedom
## AIC: 125
##
## Number of Fisher Scoring iterations: 4
```

residual deviance GOF:

```r
1 - pchisq(10.162 , 5)
```

```
## [1] 0.07077304
```

Pearson Chi Square GOF:

```r
1 - pchisq(sum(residuals(mbti2, type = "pearson")^2), df.residual(mbti2))
```

```
## [1] 0.07235899
```

Ho: The model fits adequately

Ha: The model does not fit adequately

The residual deviance is 10.16171 and the Pearson Chi-squared statistic is 10.10336.The residual deviance test returned a p value of 0.07077304 and the pearson test returned a p value of 0.07235899. Since our p value is greater than our significance level of 0.05, we can conclude that this model fits adequately.

5bi:

Since we are now certain that our model fits adequately, we can be certain that the conditional association is highest between S/N and J/P scales, since this coefficient -1.22153 is higher than any other coefficient, and it is also statistically significant.

5bii:

The z statistic for the ratio between the EI/TF scale was 1.482 (p =0.138258) and for the ratio between the EI/JP scale was 0.134 (p = 0.893261). These p values are greater than the standard significance level of 0.05, indicating that we cannot be certain that the coefficients for these variables are not zero. This might mean that the conditional association between these scales are insignificant.

5c:

```r
#model assumes conditional independence
mbti3 <- glm(n ~ EI + SN + TF + JP + EI:SN + SN:TF + SN:JP + TF:JP,
family = "poisson", data = mbti)

summary(mbti3)
```

```
##
## Call:
## glm(formula = n ~ EI + SN + TF + JP + EI:SN + SN:TF + SN:JP +
##     TF:JP, family = "poisson", data = mbti)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.65487  -0.46916   0.00529   0.54208   1.47431
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.41362    0.12930  26.402  < 2e-16 ***
## EIi          0.03871    0.11361   0.341 0.733287
## SNs          1.19414    0.14548   8.208 2.24e-16 ***
## TFt         -0.54137    0.15282  -3.543 0.000396 ***
## JPp          0.94292    0.13064   7.218 5.28e-13 ***
```

```
## EIi:SNs      0.32190    0.13598   2.367 0.017922 *
## SNs:TFt      0.42366    0.15200   2.787 0.005318 **
## SNs:JPp     -1.22021    0.14513  -8.408  < 2e-16 ***
## TFt:JPp     -0.55853    0.13497  -4.138 3.50e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 399.944  on 15  degrees of freedom
## Residual deviance:  12.369  on  7  degrees of freedom
## AIC: 123.2
##
## Number of Fisher Scoring iterations: 4
```

```
anova(mbti3, mbti2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: n ~ EI + SN + TF + JP + EI:SN + SN:TF + SN:JP + TF:JP
## Model 2: n ~ (EI + SN + TF + JP)^2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         7     12.369
## 2         5     10.162  2    2.207   0.3317
```

The difference in deviance is 12.369 - 10.162 = 2.207, and the df is 2. This gives a p-value of 0.3317. Since the p value is greater than 0.05, we can be confident that the simpler model has a better fit than the model of homogenous association.

```
exp(confint(mbti3, method="profile"))
```

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

```
##                  2.5 %     97.5 %
## (Intercept) 23.4067710 38.8667659
## EIi          0.8319067  1.2992469
## SNs          2.4923799  4.4102180
## TFt          0.4299779  0.7832723
## JPp          1.9947113  3.3306215
## EIi:SNs      1.0568719  1.8015497
## SNs:TFt      1.1363105  2.0630723
## SNs:JPp      0.2214587  0.3913279
## TFt:JPp      0.4385878  0.7446507
```

5cii:

The 95% likelihood-ratio confidence interval for the conditional odds ratio between the S/N and J/P scales is [0.2214587, 0.3913279]. We are 95% confident that the that the odds of being N given the odds of being J are within the interval.

5d:

```r
mbti4 <- glm(formula = n ~ (EI + SN + TF + JP)^3, family = "poisson", data
= mbti)


summary(mbti4)
```

```
##
## Call:
## glm(formula = n ~ (EI + SN + TF + JP)^3, family = "poisson",
##     data = mbti)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
## -0.4805   0.6865   0.4228  -0.4746   0.9580  -0.9412  -0.7382   0.4889
##       9       10       11       12       13       14       15       16
##  0.3666  -0.5798  -0.3618   0.4228  -1.0803   0.7577   0.8099  -0.4746
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.56370    0.16197  22.002  < 2e-16 ***
## EIi         -0.27880    0.23337  -1.195   0.2322
## SNs          1.05839    0.18535   5.710 1.13e-08 ***
## TFt         -0.63483    0.25356  -2.504   0.0123 *
## JPp          0.76316    0.19243   3.966 7.31e-05 ***
## EIi:SNs      0.61460    0.25451   2.415   0.0157 *
## EIi:TFt      0.20026    0.30833   0.650   0.5160
## EIi:JPp      0.37430    0.26332   1.421   0.1552
## SNs:TFt      0.41081    0.27510   1.493   0.1353
## SNs:JPp     -0.96288    0.22994  -4.187 2.82e-05 ***
## TFt:JPp     -0.58773    0.29782  -1.973   0.0484 *
## EIi:SNs:TFt -0.02364    0.30704  -0.077   0.9386
## EIi:SNs:JPp -0.51039    0.29275  -1.743   0.0813 .
## EIi:TFt:JPp  0.02440    0.27403   0.089   0.9290
## SNs:TFt:JPp  0.01922    0.30880   0.062   0.9504
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 399.9439  on 15  degrees of freedom
## Residual deviance:   7.0963  on  1  degrees of freedom
## AIC: 129.93
##
## Number of Fisher Scoring iterations: 4
```

#numbers of parameters

```r
#loglinear model of mutual independence
length(coef(log_lin_mbti))
```

```
## [1] 5
```

```r
#model of homogenous association
length(coef(mbti2))
```

```
## [1] 11
```

```r
#three factor interactionsL
length(coef(mbti4))
```

```
## [1] 15
```

The loglinear model of mutual independence has 5 parameters, the model of homogenous association has 11 parameters, and the three factor interaction terms model has 15 parameters.