Final Exam

Mohammad Mottahedi

May 5, 2016

Problem 1

(a) **F** Overdispersion and latent variables can be used to deal with unobserved but predictive variables in a model.

(b) **F** The quasi-symmetry model requires that marginal distributions be equal.

(c) **F** The adjacent-category logit model and proportional-odds cumulative-logit model are reparameterizations of each other and give equivalent conclusions.

(d) **T** A loglinear model with a structural zero fit by adding an indicator function, and the same model with the same indicator function but a 7 in place of the structural zero, will give the same estimate for the rest of the parameters (those not corresponding to the indicator function).

(e) **F** In a proportional-odds cumulative-logit model for a 2 × 5 table (where the second variable is ordinal), the odds ratio comparing levels 1 and 2 of the second variable is the same as that comparing levels 2 and 4.

(f) **T** A negative binomial GLM is often used as an alternative to Poisson regression with an overdispersion parameter

Problem 2

part a

```
##             year10
## year1        birch oak-others
##   birch        201        154
##   oak-others   114        266
```

```
## 
##  McNemar's Chi-squared test
## 
## data:  tree_a
## McNemar's chi-squared = 5.9701, df = 1, p-value = 0.01455
```

```
             | year10 
       year1 |      birch | oak-others |  Row Total | 
-------------|------------|------------|------------|
       birch |        201 |        154 |        355 | 
             |     15.689 |     11.767 |            | 
             |      0.566 |      0.434 |      0.483 | 
             |      0.638 |      0.367 |            | 
             |      0.273 |      0.210 |            | 
-------------|------------|------------|------------|
  oak-others |        114 |        266 |        380 | 
             |     14.657 |     10.993 |            |
             |      0.300 |      0.700 |      0.517 |
             |      0.362 |      0.633 |            |
             |      0.155 |      0.362 |            |
-------------|------------|------------|------------|
Column Total |        315 |        420 |        735 |
             |      0.429 |      0.571 |            |
-------------|------------|------------|------------|
```

$H_0 = \pi_{1,2} = \pi_{1,2}$

$\chi^2 = 5.6754$, $df = 1$, $p - value = 0.0172$

based on the p-value $P(\chi^2 \geq 5.6) \approx 0$ there is a strong evidence that number of birch and non birch trees has change over time.

part b

The quasi-independence model for 3x3 table:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.58255    0.23940  19.142  < 2e-16 ***
year1oak     0.04225    0.23332   0.181  0.85629
year1other  -2.09618    0.22385  -9.364  < 2e-16 ***
year10oak    0.28512    0.24294   1.174  0.24054
year10other -1.40522    0.19915  -7.056 1.71e-12 ***
ibirch       0.76456    0.24915   3.069  0.00215 **
ioak         0.25486    0.24995   1.020  0.30789
iother       2.83088    0.28423   9.960  < 2e-16 ***
---
Null deviance: 563.476  on 8  degrees of freedom
Residual deviance:  17.156  on 1  degrees of freedom
AIC: 84.56
```

The model doesn't seem to fit well $P(\chi^2 \geq 17.156)$, $df = 1$, $P(\chi^2 \geq 17.156) = 3.4432007 \times 10^{-5}$.

part c

fitting quasi-symmetry:

```
Coefficients: (5 not defined because of singularities)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        4.1589     0.1250  33.271  < 2e-16 ***
year1_birchTRUE   -1.6015     0.2891  -5.540 3.03e-08 ***
year1_oakTRUE     -1.7521     0.3047  -5.751 8.88e-09 ***
year1_otherTRUE        NA         NA      NA       NA
year10_birchTRUE  -1.8465     0.3147  -5.867 4.44e-09 ***
year10_oakTRUE    -1.9726     0.3285  -6.005 1.91e-09 ***
year10_otherTRUE       NA         NA      NA       NA
symm1              2.4672     0.4828   5.110 3.22e-07 ***
symm4              0.3767     0.1039   3.624  0.00029 ***
```

```
symm6                   NA         NA      NA       NA
symm2               1.0654     0.2712   3.928 8.55e-05 ***
symm5                   NA         NA      NA       NA
symm3                   NA         NA      NA       NA

Null deviance: 1.0221e+02  on 8  degrees of freedom
Residual deviance: 6.0515e-03  on 1  degrees of freedom
AIC: 56.198
```

The model seem to fit well $P(\chi^2 \geq 0.00605)$, $df = 1$, $P(\chi^2 \geq 0.00605) = 0.9380017$.

part d

fitting the symmetry model:

goodness of fit for symmetry model:

```
Null deviance: 102.21480  on 8  degrees of freedom
Residual deviance:   0.59276  on 3  degrees of freedom
AIC: 52.784
```

goodness-of-fit for marginal homogeneity:

$G^2(marginalhomogeneity) = G^2(symmetry) - G^2(quasisymmetry) = 0.59276 - 6.0515e - 03 = 0.5867085$, $df = I - 1 = 2$

$p - value = 0.7457579$

the marginal homogeneity fits moderately well.

part e

model	df	G^2	p-value
quasi-independence	1	17.156	3.4432007x 10−5
symmetry	3	0.59276	0.8980878
Marginal Homogeneity	2	0.5867085	0.7457579
quasi-symmetry	1	0.00605	0.9380017

There seems we have a strong symmetric association. The best model fit belongs to quasi-symmetry model. The marginal homogeneity fits moderately well and difference between the two likelihood ratios is not significant. This is in contrast with conclusion from part a, that there has been a change in number of birch trees after 10 years.

Problem 3

part a

fitting passion regression model:

```
Coefficients:
```

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.9071756  0.2512779  15.549  < 2e-16 ***
Price_movementup  1.6155856  0.2453932   6.584 4.59e-11 ***
SecurityB        -1.3110020  0.2417294  -5.423 5.85e-08 ***
SecurityC         0.6848235  0.1490582   4.594 4.34e-06 ***
Milisecond       -0.0019280  0.0005142  -3.750 0.000177 ***

Null deviance: 159.489  on 5  degrees of freedom
Residual deviance:  11.466  on 1  degrees of freedom
AIC: 53.025
```

fitted model:

$log(\hat{\mu_i}) = 3.91 + 1.62 \times PriceMmovemenet - 1.311 \times Security_B + 0.68 \times Security_C - 0.002 \times Miliseconds$

model fit: $P - value = 7.0881079 \times 10^{-4}$

the $exp(intercept) = 49.7494794$ indicates the mean of orders when other variables are zero.

Price movement: indicates when price movement is up number of order is going to increase by the order of compared to down movement $exp(1.6155) = 5.0304025$

security B: indicates when security is type B compared to type A decreases the number of orders by factor of $exp(-1.311) = 0.2642129$

security C: indicates when security is type C compared to type A increase the number of orders by factor of $exp(0.68) = 1.9738777$

Milisecond: as length of time between transactions increases the number of orders decreases slightly by factor $exp(-0.0019280) = 0.9980739$

part b

model fit:

$G^2 = 11.466$, $df = 1$

$P - value = 7.0881079 \times 10^{-4}$

all of the variables are significant and the model doesn't fit well.

part c

setting the orders with row with (B, Down) to zero and adding numerical dummy variables for the cell with zero.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.528e+00  2.816e-01  12.527  < 2e-16 ***
Price_movementup  1.692e+00  2.423e-01   6.985 2.85e-12 ***
SecurityB        -1.711e+00  3.021e-01  -5.662 1.49e-08 ***
SecurityC         7.061e-01  1.473e-01   4.793 1.64e-06 ***
Milisecond       -1.487e-03  5.245e-04  -2.835  0.00458 **
delta            -2.410e+01  4.225e+04  -0.001  0.99954


Null deviance: 2.3158e+02  on 5  degrees of freedom
```

```
Residual deviance: 4.1224e-10  on 0  degrees of freedom
AIC: 38.623
```

The model fits is well, the coefficients are changed slightly but the our conclusion didn't change about the effect of main effects on the response.

part d

after treating the cell (B, Down) as anomalous the, the model fit got better that confirm that the cell was the reason for poor fit. but the model parameters didn't change significantly.

Problem 4

part a

complete independence model

the $G^2 = 240.94$, $df = 8$ and the complete independence model does not fit.

part b

Since we only have one sampling zero in the table the MLE estimates exists and it's always positive.

part c

Model_name	G2	df	Pvalue	AIC
Complete Independence	181.2	7	0	250.8
Joint Independence XY,Z	142.8	5	0	216.4
Joint Independence XZ,Y	165.2	6	0	236.7
Joint Independence ZY,X	74.47	5	1.199e-14	148
Conditional Independence XY,XZ	126.8	4	0	202.3
Conditional Independence XY,ZY	36.06	3	7.262e-08	113.6
Conditional Independence ZY,ZX	58.45	4	6.147e-12	134
Homogeneous assocition	32.2	2	1.016e-07	111.7
Saturated	0	0	0	83.55

comparing homogeneous model with conditional independence model (XY,ZY):

$\Delta G^2 = 3.86$, $df = 1$, $p - value = 0.0495$

based on AIC and and p-value in the table above the homogeneous Association model best fit the data.

The model indicates that XY does not depend on level of A, and association between YZ does not depend on X.

part d

model equation:

$log(\mu_{ij}) = \lambda + \lambda_i^Y + \lambda_j^Z + \lambda_k^X + \lambda_{ij}^{YZ} + \lambda_{jk}^{ZX} + \delta_{212} I(2, 1, 2)$

part e

model equation:

$X_1 = 1$ if X=2

$X_1 = 0$ otherwise

$X_2 = 1$ if Y=2

$X_2 = 0$ otherwise

$X_3 = 1$ if Y=3

$X_3 = 0$ otherwise

$log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \delta_{212} I(2,1,2)$

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.605e-01  1.190e+00  -0.219    0.827
X2           5.210e-01  1.262e+00   0.413    0.680
Y2           7.906e-01  1.614e+00   0.490    0.624
Y3          -3.167e-16  1.426e+00   0.000    1.000
delta       -1.862e+01  3.956e+03  -0.005    0.996
Null deviance: 16.636  on 11  degrees of freedom
Residual deviance: 14.737  on  7  degrees of freedom
AIC: 24.737
```

Problem 5

part a

$log(\mu_{ijkl}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} + \lambda_{kl}^{ZW} + \lambda_{jk}^{YZ}$

$log(\mu_{ijkl}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} + \lambda_{jkl}^{YZW}$

df = # of cells - # unique parameters in the model

model XY, YZ, ZW:

# cell = 48, # unique parameters = 13

$df = 35$

model XY,YZW:

# cell = 48, # unique parameters = 13

$df = 35$

$\Delta G^2 = 35 - 35 = 0$

part part b

conditional odds ratio for fixed level of Z and W

$\theta_{XY(ZW)} = \frac{\mu_{11jl}\mu_{22jl}}{\mu_{12jl}\mu_{21jl}}$

for Z=2, W=1

$\theta_{XY(ZW)} = \frac{\mu_{1121}\mu_{2221}}{\mu_{1221}\mu_{2121}}$

we are testing conditional conditional independence between X and Y given Z and W are equal to 2 and 1, respectively.

(XZW, YZW)

part c

$log(\mu_{ijkl}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$

$log(\mu_{3221}) = \lambda + \lambda_3^X + \lambda_1^W$

CODE

Problem 2

```
tree <- matrix(c(201, 110, 4, 122, 175, 24, 32, 17, 50), ncol = 3,
               dimnames = list(year1 = c("birch", "oak", "others"),
                               year10 = c("birch", "oak", "others")))


tree_a <- matrix(c(tree[1,1], tree[2,1]+ tree[3,1] ,
                   tree[1,2]+ tree[1,3], tree[2,2] +
                       tree[2,3] + tree[3,2]+
                       tree[3,3] ), ncol=2,
                  dimnames = list(year1 = c("birch", "oak-others"),
                               year10 = c("birch", "oak-others")))
print(tree_a)
mcnemar.test(tree_a, correct = F)


year1 <- rep(c("birch","oak","other"),c(3,3,3))
year10 = rep(c("birch","oak","other"),3)
count=c(210,122,32,110,175,17,4,24,50)
year1=factor(year1)
year10=factor(year10)

#-------create indicator variables--------
ibirch=(year1=="birch")*(year10=="birch")
year1_birch=(year1=="birch")
year1_oak=(year1=="oak")
year1_other=(year1=="other")
year10_birch=(year10=="birch")
year10_oak=(year10=="oak")
year10_other=(year10=="other")
ioak=(year1=="oak")*(year10=="oak")
iother=(year1=="other")*(year10=="other")
symm3=3*(year1=="other")*(year10=="other")
```

```
symm1=1*(year1=="birch")*(year10=="birch")
symm4=4*(year1=="oak")*(year10=="birch")+4*(year1=="birch")*(year10=="oak")
symm6=6*(year1=="birch")*(year10=="other")+6*(year1=="other")*(year10=="birch")
symm2=2*(year1=="oak")*(year10=="oak")
symm5=5*(year1=="oak")*(year10=="other")+5*(year1=="other")*(year10=="oak")
symm=symm3+symm1+symm4+symm6+symm2+symm5

#----generate dataset movies-----------
data = data.frame(year1,year10,count,ibirch,ioak,iother,symm)

#-------contigency table------------------
### Contigency Table
table=xtabs(count~year1 + year10)

options(contrast=c("contr.treatment","contr.poly"))

model_2_b <- glm(count ~ year1 + year10 + ibirch + ioak + iother, family=poisson(link=log))

model_2_c <- glm(count ~ year1_birch + year1_oak + year1_other +
                     year10_birch + year10_oak + year10_other +
                     symm1+symm4+symm6+symm2+symm5+symm3,
                     family=poisson(link=log))

model_2_d <- glm(count ~ symm1+symm4+symm6+symm2+symm5+symm3,
                     family=poisson(link=log))
```

Probelem 3

```
dat_3 <- data.frame(Milisecond = c(673, 539, 843, 974, 11, 350),
                    Orders = c(68, 15,107, 8, 22, 41),
                    Price_movement = c("up","up","up","down","down","down"),
                    Security = c("A","B","C", "A","B","C"))


model_3_a <- glm(Orders ~ Price_movement + Security + Milisecond,
                 data = dat_3, family = poisson(link=log) )

dat_3_c <- dat_3

dat_3_c$Orders[5] <- 0

delta <- rep(0,6)

delta[5] <- 1

model_3_c <- glm(Orders ~ Price_movement + Security + Milisecond + delta,
                 data = dat_3_c, family = poisson(link = log) )
```

Probelem 4

```r
X <- c(1,2)
Y <- c(1,2,3)
Z <- c(1,2)

table <- expand.grid(Y=Y,Z=Z,X=X)
count <- c(28,67,93,84,20,74, 7,0,46,77,9,23)
table <- cbind(table,count =count)
table <- xtabs(count ~ Y + Z + X, data = table)
df <- data.frame(table)

model_4_a <- glm(Freq~ X + Y + Z, data = df, family  = poisson(link=log))



# complete indendenet model
model_4_a_1 <- glm(Freq~ X + Y + Z, data = df, family  = poisson(link=log))
# Join independent model
## (XY,Z)
model_4_a_2 <- glm(Freq~ X + Y + Z + X*Y,
                    data = df, family  = poisson(link=log))
##(XZ,Y)
model_4_a_3 <- glm(Freq~ X + Y + Z + X*Z,
                    data = df, family  = poisson(link=log))
##(ZY,X)
model_4_a_4 <- glm(Freq~ X + Y + Z + Z*Y,
                    data = df, family  = poisson(link=log))
# conditional independence model
##(XY,XZ)
model_4_a_5 <- glm(Freq~ X + Y + Z + X*Y + X*Z,
                    data = df, family  = poisson(link=log))
##(XY,ZY)
model_4_a_6 <- glm(Freq~ X + Y + Z + X*Y + Z*Y,
                    data = df, family  = poisson(link=log))

##(ZY,ZX)
model_4_a_7 <- glm(Freq~ X + Y + Z + Z*Y + Z*X,
                    data = df, family  = poisson(link=log))
# Homogeneous association model
model_4_a_8 <- glm(Freq~ X + Y + Z + X*Y + Y*Z + X*Z,
                    data = df, family  = poisson(link=log))
# Saturated model
model_4_a_9 <- glm(Freq~ X + Y + Z + X*Y + Y*Z + X*Z + X*Y*Z,
                   data = df, family  = poisson(link=log))

model_name <- c("Complete Independence","Joint Independence XY,Z ",
                "Joint Independence XZ,Y", "Joint Independence ZY,X",
                "Conditional Independence XY,XZ", "Conditional Independence XY,ZY",
                "Conditional Independence ZY,ZX", "Homogeneous assocition", "Saturated")

G2 <- c(model_4_a_1$deviance,model_4_a_2$deviance,model_4_a_3$deviance,
        model_4_a_4$deviance,model_4_a_5$deviance,model_4_a_6$deviance,
        model_4_a_7$deviance,model_4_a_8$deviance,model_4_a_9$deviance)

model_df <- c(model_4_a_1$df.residual,model_4_a_2$df.residual,model_4_a_3$df.residual,
        model_4_a_4$df.residual,model_4_a_5$df.residual,model_4_a_6$df.residual,
        model_4_a_7$df.residual,model_4_a_8$df.residual,model_4_a_9$df.residual)

AIC <- c(model_4_a_1$aic,model_4_a_2$aic,model_4_a_3$aic,
        model_4_a_4$aic,model_4_a_5$aic,model_4_a_6$aic,
        model_4_a_7$aic,model_4_a_8$aic,model_4_a_9$aic)

pvalue <- rep(0,9)


for (i in 1:9){
    pvalue[i] <- 1 - pchisq(G2[i], model_df[i])
}

model_summary <- data.frame(Model_name =model_name,
                            G2=round(G2,2), df=model_df, Pvalue=pvalue, AIC=AIC)

pander(model_summary)



delta <- rep(0,12)
delta[8] <- 1

model_4_e <- glm(Z ~ X + Y + delta,data = df, family  = binomial(link="logit"))
```