Homework 6+

Charles Hwang

4/29/2022

Charles Hwang

Dr. Whalen

STAT 410-001

29 April 2022

Problem 6.2

```
rm(list=ls())
all<-read.table("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/Alligators.d
all$y<-as.factor(all$y)
library(VGAM)
alll<-vglm(y~x,family=multinomial,data=all)
summary(alll)
```

```
##
## Call:
## vglm(formula = y ~ x, family = multinomial, data = all)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   1.6177     1.3073   1.237  0.21591
## (Intercept):2   5.6974     1.7937   3.176  0.00149 **
## x:1            -0.1101     0.5171  -0.213  0.83137
## x:2            -2.4654     0.8996      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 98.3412 on 114 degrees of freedom
##
## Log-likelihood: -49.1706 on 114 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## 'x:2'
##
##
## Reference group is level  3  of the response
```

```
exp(alll@coefficients["x:1"])+exp(alll@coefficients["x:2"])
```

```
##       x:1
## 0.9807074
```

Problem 6.4

```
al<-read.table("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/Afterlife.da
al$race<-as.factor(al$race)
al$gender<-as.factor(al$gender)
algr<-vglm(cbind(yes,undecided,no)~gender+race,family=multinomial,data=al)
summary(algr)
```

```
##
## Call:
## vglm(formula = cbind(yes, undecided, no) ~ gender + race, family = multinomial,
##     data = al)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   1.3016     0.2265   5.747  9.1e-09 ***
## (Intercept):2  -0.6529     0.3405  -1.918   0.0551 .
## gendermale:1   -0.4186     0.1713  -2.444   0.0145 *
## gendermale:2   -0.1051     0.2465  -0.426   0.6700
## racewhite:1     0.3418     0.2370   1.442   0.1493
## racewhite:2     0.2710     0.3541   0.765   0.4442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 0.8539 on 2 degrees of freedom
##
## Log-likelihood: -19.7324 on 2 degrees of freedom
##
## Number of Fisher scoring iterations: 3
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Reference group is level  3  of the response
```

```
algru<-vglm(cbind(yes,undecided,no)~gender+race,family=multinomial(refLevel="undecided"),data=al)
summary(algru)
```

```
##
## Call:
## vglm(formula = cbind(yes, undecided, no) ~ gender + race, family = multinomial(refLevel = "undecided
##     data = al)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   1.9546     0.2974   6.571 4.99e-11 ***
## (Intercept):2   0.6529     0.3405   1.918   0.0551 .
```

```
## gendermale:1   -0.3135     0.2083  -1.505   0.1324
## gendermale:2    0.1051     0.2465   0.426   0.6700
## racewhite:1     0.0708     0.3092   0.229   0.8189
## racewhite:2    -0.2710     0.3541  -0.765   0.4442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,2]), log(mu[,3]/mu[,2])
##
## Residual deviance: 0.8539 on 2 degrees of freedom
##
## Log-likelihood: -19.7324 on 2 degrees of freedom
##
## Number of Fisher scoring iterations: 3
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Reference group is level  2  of the response
```

```
exp(algr@coefficients["gendermale:2"])
```

```
## gendermale:2
##    0.9002671
```

```
exp(algru@coefficients["gendermale:1"])
```

```
## gendermale:1
##    0.7308942
```

(**a**) The estimated conditional odds ratio between gender and "undecided" vs. "no" belief in an afterlife (β_2^G), given race, is 0.9002671. The estimated the odds of an "undecided" response vs. a "no" response on belief in an afterlife for males are approximately 0.9002671 times the same estimated odds for females, adjusting for race.

(**b**) The estimated conditional odds ratio between gender and affirmative ("yes") vs. "undecided" belief in an afterlife (β_1^G), given race, is 0.7308942. The estimated the odds of an affirmative ("yes") response vs. an "undecided" response on belief in an afterlife for males are approximately 0.7308942 times the same estimated odds for females, adjusting for race.

Problem 6.5

Problem 6.5a

(i) more satisfied; (ii) less satisfied; (iii) less satisfied. This makes sense intuitively when looking at the equation, variables, and levels.

Problem 6.5b

We can clearly see that Y is maximized when $x_1 = 4$, $x_2 = 1$, and $x_3 = 1$. This also makes sense intuitively when looking at the equation, variables, and levels.

Problem 6.6

We can see from Table 6.6 (page 188) that the estimated odds of being "very" happy vs. "not" happy for those with "above average" or "average" incomes are approximately $e^{-0.22751} \approx 0.7965145$ times the same estimated odds for those in the next-lowest income category.

We can also see that the estimated odds of being "pretty" happy vs. "not" happy for those with "above average" or "average" incomes are approximately $e^{-0.09615} \approx 0.9083278$ times the same estimated odds for those in the next-lowest income category.

We can see from the fitted values that the range of the outcome variable increase with income. Among those who are "not" happy, those with "below average" income ($[1, y1]$) tended to be more happy than those with "above average" income ($[1, y3]$), with "average" income ($[1, y2]$) between the two. Among those who are "very" happy, those with "above average" income ($[3, y3]$) tended to be more happy than those with "below average" income ($[3, y1]$), with "average" income ($[3, y2]$) between the two. This may indicate that a higher income can be polarizing for happiness, whether positive or negative.

We can see the income variable itself is not very significant in the model ($p = 0.504907$, $p = 0.430694$), indicating there are other variables that may be important, as expected.

In conclusion, marital happiness may be associated with family income, but there are clearly other variables not in the model that are more closely associated or interact with family income. Additional analysis is needed to form a definitive conclusion.

Problem 6.7

Problem 6.7a

The cumulative logit model is able to account for ordinal variables, like we have in the happiness and income categories. Section 6.2 (page 167) of the textbook writes: "When response categories are ordered, logits can utilize the ordering. This results in models that have fewer parameters and potentially greater power and simpler interpretation than baseline-category logit models."

Problem 6.7b

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

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

The model fits adequately. With a residual deviance of 3.2472 on 3 degrees of freedom, we can see H_0 for the χ^2 goodness-of-fit test is not rejected at the $\alpha = 0.05$ level ($p = \chi^2_3(3.2472) = 0.3550593$). There are $c = 3$ response categories, which means there are $c - 1 = 3 - 1 = 2$ intercepts. There is only one income effect because it is the same for each cumulative probability level.

Problem 6.7c

```
1-pchisq(4.13476-3.2472,4-3)
```

```
## [1] 0.3461394
```

We can see from Tables 6.6 and 6.7 that the χ^2 test statistic is $4.13476 - 3.2472 = 0.88756$ with $4 - 3 = 1$ degree of freedom and the p-value is 0.3461394. We fail to reject H_0 at the $\alpha = 0.05$ level and there is insufficient evidence that there is an association between income and happiness. The estimate for the income effect is $\beta_I = -0.1117$, which indicates the odds of being less happy decreases as income increases.

Problem 8.1

```
mcnemar.test(matrix(c(159,8,22,14),nrow=2),correct=FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  matrix(c(159, 8, 22, 14), nrow = 2)
```

```
## McNemar's chi-squared = 6.5333, df = 1, p-value = 0.01059
```

We reject H_0 at the $\alpha = 0.05$ level. There is sufficient evidence ($\chi^2 = 6.5333333$, $p = 0.0105871$) that there is a relationship between smoking and birth weight.

Problem 8.3

```
log(359/785/(334/810))
```

```
## [1] 0.1035319
```

```
log(132/107)
```

```
## [1] 0.2099731
```

The $\hat{\beta}$ variable in (8.2) is for the entire sample, while the $\hat{\beta}$ variable in (8.3) is conditional on subject i. We can see that $\hat{\beta}$ from (8.2) is $\ln(\frac{\frac{n_{1\Sigma}}{n_{2\Sigma}}}{\frac{n_{\Sigma 1}}{n_{\Sigma 2}}}) = \ln(\frac{\frac{359}{785}}{\frac{334}{810}}) = \ln(\frac{29079}{26219}) \approx 0.1035319$, while $\hat{\beta}$ from (8.3) is $\ln(\frac{n_{12}}{n_{21}}) = \ln(\frac{132}{107}) \approx 0.2099731$.

Problem 8.4

We estimate that $e^{\hat{\beta}} = \frac{16}{37} \approx 0.4324324$. This estimate is valid because it is conditional on the subject (subject-specific), taken from (8.3).

Problem 8.6

McNemar's test is used when there are two binary factor variables and a two-way table is created, while the paired-difference t-test is used when the variables are numeric and normally distributed.

Problem 9.1

Problem 9.1(a)

$logit[P(Y_t = 1)] = \alpha + \beta_A z_A + \beta_C z_C + \beta_M z_M + \gamma x$, where β_A, β_C, and β_M are the coefficients for alcohol, cigarette, and marijuana and z_A, z_C, and z_M are binary variables for the usage of each respectively (yes = 1, no = 0). The hypotheses for testing for marginal homogeneity would be:

$H_0 : \beta_A = \beta_C = \beta_M$

H_A : At least one β_i is different

Problem 9.1(b)

$logit[P(Y_{it} = 1)] = \alpha_i + \beta_A z_A + \beta_C z_C + \beta_M z_M + \gamma_i x$

The interpretations for the β_i's and z_i's would generally be the same, but we can see the intercept and error terms α_i and γ_i now incorporate individual subjects i.

Problem 9.3

Problem 9.3(a)

$logit(\hat{\pi}) = -0.57 + 1.93(0) + 0.86(0) + 0.38r - 0.20g + 0.37g(0) + 0.22g(0)$; where $s_1 = 0$, $s_2 = 0$

$logit(\hat{\pi}) = -0.57 + 0.38r - 0.20g$

We can clearly see this equation is maximized when $r = 1$ (white) and $g = 0$ (male).

$logit(\hat{\pi}) = -0.57 + 1.93(1) + 0.86(0) + 0.38r - 0.20g + 0.37g(1) + 0.22g(0)$; where $s_1 = 1$, $s_2 = 0$

$logit(\hat{\pi}) = -0.57 + 1.93 + 0.38r - 0.20g + 0.37g$

$logit(\hat{\pi}) = 1.36 + 0.38r + 0.17g$

White females have the highest estimated probability of use of **alcohol**. We can clearly see this equation is maximized when $r = 1$ (white) and $g = 1$ (female).

$logit(\hat{\pi}) = -0.57 + 1.93(0) + 0.86(1) + 0.38r - 0.20g + 0.37g(0) + 0.22g(1)$; where $s_1 = 0$, $s_2 = 1$

$logit(\hat{\pi}) = -0.57 + 0.86 + 0.38r - 0.20g + 0.22g$

$logit(\hat{\pi}) = 0.29 + 0.38r + 0.02g$

White females have the highest estimated probability of use of **cigarettes**. We can clearly see this equation is maximized when $r = 1$ (white) and $g = 1$ (female).

Problem 9.3(b)

$logit(\hat{\pi}) = -0.57 + 1.93s_1 + 0.86s_2 + 0.38r - 0.20g + 0.37gs_1 + 0.22gs_2$ (original equation)

We can see that $e^{\beta_r} = e^{0.38} \approx 1.4622846$.

Problem 9.3(c)

$logit(\hat{\pi}) = 1.36 + 0.38r + 0.17g$; where $s_1 = 1$, $s_2 = 0$ (from Problem 9.3(a))

We can see that $e^{\beta_g} = e^{0.17} \approx 1.1853049$.

$logit(\hat{\pi}) = 0.29 + 0.38r + 0.02g$; where $s_1 = 0$, $s_2 = 1$ (from Problem 9.3(a))

We can see that $e^{\beta_g} = e^{0.02} \approx 1.0202013$.

$logit(\hat{\pi}) = -0.57 + 0.38r - 0.20g$; where $s_1 = 0$, $s_2 = 0$ (from Problem 9.3(a))

We can see that $e^{\beta_g} = e^{-0.20} \approx 0.8187308$.

Problem 9.3(d)

$logit(\hat{\pi}) = -0.57 + 1.93s_1 + 0.86s_2 + 0.38r - 0.20(1) + 0.37(1)s_1 + 0.22(1)s_2$; where $g = 1$

$logit(\hat{\pi}) = -0.77 + 2.30s_1 + 1.08s_2 + 0.38r$

We can see that $e^{\beta_{s_1}} = e^{2.30} \approx 9.9741825$ and that $e^{\beta_{s_2}} = e^{1.08} \approx 2.9446796$.

Problem 9.3(e)

$logit(\hat{\pi}) = -0.57 + 1.93s_1 + 0.86s_2 + 0.38r - 0.20(0) + 0.37(0)s_1 + 0.22(0)s_2$; where $g = 0$

$logit(\hat{\pi}) = -0.57 + 1.93s_1 + 0.86s_2 + 0.38r$

We can see that $e^{\beta_{s_1}} = e^{1.93} \approx 6.8895102$ and that $e^{\beta_{s_2}} = e^{0.86} \approx 2.3631607$.

Problem 9.4

Problem 9.4(a)

$logit(\hat{\pi}) = -0.02810 - 1.31391s - 0.05927(1) + 0.048246t + 1.01719(1)t$; where $d = 1$

$logit(\hat{\pi}) = -0.02810 - 1.31391s - 0.05927 + 0.048246t + 1.01719t$

$logit(\hat{\pi}) = -0.08737 - 1.31391s + 1.065436t$

We can see that $e^{\beta_t} = e^{1.065436} \approx 2.902104$. The estimated odds of a normal classification of a subject's level of mental depression after t weeks of using the new drug are approximately $e^{1.065436} \approx 2.902104$ times

the same estimated odds after $\frac{t}{2}$ weeks of using the new drug, where $t = 1, 2$, or 4, holding initial severity constant.

$logit(\hat{\pi}) = -0.02810 - 1.31391s - 0.05927(0) + 0.048246t + 1.01719(0)t$; where $d = 0$

$logit(\hat{\pi}) = -0.02810 - 1.31391s + 0.048246t$

We can see that $e^{\beta_t} = e^{0.048246} \approx 1.0494288$. The estimated odds of a normal classification of a subject's level of mental depression after t weeks of using the standard drug are approximately $e^{0.048246} \approx 1.0494288$ times the same estimated odds after $\frac{t}{2}$ weeks of using the standard drug, where $t = 1, 2$, or 4, holding initial severity constant.

Problem 9.4(b)

$logit(\hat{\pi}) = -0.02810 - 1.31391s - 0.05927d + 0.048246t + 1.01719dt$ (original equation)

We can see when using the coefficients with drug (d) that $e^{\beta_d + \beta_{dt} t} = e^{-0.05927 + 1.01719t}$. The estimated odds of a normal classification of a subject's level of mental depression after t weeks of using the new drug are approximately $e^{-0.05927 + 1.01719 \log_2 t}$ times the same estimated odds after t weeks of using the standard drug, where $t = 1, 2$, or 4, holding initial severity constant.

Problem 9.11

Problem 9.11(a)

```
i<-read.table("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/Insomnia2.dat
library(VGAM)
names(i)<-c("Treatment","Initial","F10","F25","F45","F75")
tm<-vglm(cbind(F10,F25,F45,F75)~Treatment*Initial,family=cumulative(parallel=TRUE),data=i)
summary(tm)
```

```
##
## Call:
## vglm(formula = cbind(F10, F25, F45, F75) ~ Treatment * Initial,
##     family = cumulative(parallel = TRUE), data = i)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept):1      1.107400   0.299122   3.702 0.000214 ***
## (Intercept):2      2.810496   0.319678   8.792  < 2e-16 ***
## (Intercept):3      4.326758   0.374342  11.558  < 2e-16 ***
## Treatment         -0.217123   0.421879  -0.515 0.606792
## Initial           -0.052835   0.005467  -9.665  < 2e-16 ***
## Treatment:Initial  0.021737   0.007425   2.927 0.003417 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3])
##
## Residual deviance: 30.7719 on 18 degrees of freedom
##
## Log-likelihood: -48.5192 on 18 degrees of freedom
##
## Number of Fisher scoring iterations: 15
##
## No Hauck-Donner effect found in any of the estimates
```

```
##
##
## Exponentiated coefficients:
##         Treatment           Initial Treatment:Initial
##         0.8048307         0.9485368         1.0219751
```

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 0.052835y_1 + 0.021737xy_1$ (original equation)

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 0.052835(10) + 0.021737x(10)$; where $y_1 = 10$

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 0.52835 + 0.21737x$

$logit[P(Y_2 \leq j)] = \alpha_j - 0.52835 + 0.000247x$

We can see the estimated treatment effect at $y_1 = 10$ is approximately $\hat{\beta}_2 = 0.000247$.

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 0.052835(25) + 0.021737x(25)$; where $y_1 = 25$

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 1.320875 + 0.543425x$

$logit[P(Y_2 \leq j)] = \alpha_j - 1.320875 + 0.326302x$

We can see the estimated treatment effect at $y_1 = 25$ is approximately $\hat{\beta}_2 = 0.326302$.

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 0.052835(45) + 0.021737x(45)$; where $y_1 = 45$

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 2.377575 + 0.978165x$

$logit[P(Y_2 \leq j)] = \alpha_j - 2.377575 + 0.761042x$

We can see the estimated treatment effect at $y_1 = 45$ is approximately $\hat{\beta}_2 = 0.761042$.

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 0.052835(75) + 0.021737x(75)$; where $y_1 = 75$

$logit[P(Y_2 \leq j)] = \alpha_j - 0.217123x - 3.962625 + 1.630275x$

$logit[P(Y_2 \leq j)] = \alpha_j - 3.962625 + 1.413152x$

We can see the estimated treatment effect at $y_1 = 75$ is approximately $\hat{\beta}_2 = 1.413152$.

Problem 9.11(b)

```
i$Initial<-as.factor(i$Initial)
q<-vglm(cbind(F10,F25,F45,F75)~Treatment+Initial,family=cumulative(parallel=TRUE),data=i)
summary(q) # (i)
```

```
##
## Call:
## vglm(formula = cbind(F10, F25, F45, F75) ~ Treatment + Initial,
##     family = cumulative(parallel = TRUE), data = i)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -0.3057     0.2950  -1.036 0.299978
## (Intercept):2   1.4254     0.3011   4.733 2.21e-06 ***
## (Intercept):3   2.9048     0.3320   8.749  < 2e-16 ***
## Treatment       0.9108     0.1666   5.466 4.60e-08 ***
## Initial25       0.3662     0.3618   1.012 0.311526
## Initial45      -1.1543     0.3103  -3.720 0.000199 ***
## Initial75      -2.3068     0.3100  -7.442 9.90e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3])
##
## Residual deviance: 30.2314 on 17 degrees of freedom
##
## Log-likelihood: -48.249 on 17 degrees of freedom
##
## Number of Fisher scoring iterations: 15
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##  Treatment  Initial25  Initial45  Initial75
## 2.48630206 1.44222146 0.31527540 0.09957979
```

```
x<-vglm(cbind(F10,F25,F45,F75)~Treatment*Initial,family=cumulative(parallel=TRUE),data=i)
summary(x) # (ii)
```

```
##
## Call:
## vglm(formula = cbind(F10, F25, F45, F75) ~ Treatment * Initial,
##     family = cumulative(parallel = TRUE), data = i)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept):1        -0.1459     0.3712  -0.393 0.694236
## (Intercept):2         1.6359     0.3791   4.315 1.60e-05 ***
## (Intercept):3         3.1651     0.4104   7.712 1.24e-14 ***
## Treatment             0.5284     0.5590   0.945 0.344539
## Initial25             1.0206     0.5195   1.965 0.049453 *
## Initial45            -1.5576     0.4178  -3.728 0.000193 ***
## Initial75            -2.7247     0.4136  -6.589 4.44e-11 ***
## Treatment:Initial25  -1.3732     0.7409  -1.853 0.063837 .
## Treatment:Initial45   0.7733     0.6298   1.228 0.219508
## Treatment:Initial75   0.7775     0.6125   1.269 0.204316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3])
##
## Residual deviance: 20.8646 on 14 degrees of freedom
##
## Log-likelihood: -43.5655 on 14 degrees of freedom
##
## Number of Fisher scoring iterations: 15
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##           Treatment           Initial25           Initial45           Initial75
##          1.69624620          2.77472153          0.21065035          0.06556479
```

```
## Treatment:Initial25 Treatment:Initial45 Treatment:Initial75
##          0.25329694          2.16701251          2.17597479
```

(**i**) We can see the estimated treatment log odds ratio is approximately $\ln(e^{\beta_x}) = \beta_x = 0.9108$, assuming no interaction. The estimated "log" odds of a patient receiving the active drug falling asleep after n minutes are approximately 0.9108 times the estimated "log" odds of a patient receiving the placebo falling asleep after n minutes, where $n = 10, 25, 45$, or 75 and no interaction is assumed, holding all other variables constant.

(**ii**) We can see the estimated treatment log odds ratio is approximately $\ln(e^{\beta_x}) = \beta_x = 0.5284$ when allowing for interaction. The estimated "log" odds of a patient receiving the active drug falling asleep after n minutes are approximately 0.5284 times the estimated "log" odds of a patient receiving the placebo falling asleep after n minutes, where $n = 10, 25, 45$, or 75 and interaction is assumed, holding all other variables constant.

Problem 10.3

Problem 10.3(a)

Sample proportion estimates: $\frac{2}{5}, \frac{4}{5}, \frac{1}{5}, \frac{3}{5}, \frac{3}{5}, \frac{5}{5}, \frac{4}{5}, \frac{2}{5}, \frac{3}{5}, \frac{1}{5}$

Mode: $logit(\pi_i) = u_i + \alpha$

Problem 10.3(b)

```
library(lme4)
h<-c(2,4,1,3,3,5,4,2,3,1)
f<-rep(5,10)
c<-1:10
rem<-glmer(h/f~(1|c),family=binomial,weights=f,nAGQ=100)
summary(rem)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 100) [glmerMod]
##  Family: binomial  ( logit )
## Formula: h/f ~ (1 | c)
## Weights: f
##
##      AIC      BIC   logLik deviance df.resid
##     18.5     19.1     -7.3     14.5        8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1792 -0.5314  0.1168  0.6250  1.5334
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  c      (Intercept) 0.3098   0.5566
## Number of obs: 10, groups:  c, 10
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2589     0.3452    0.75    0.453
```

```
exp(rem@beta)/(1+exp(rem@beta))
```

```
## [1] 0.5643724
```

```
rem@beta
```

```
## [1] 0.2589267
```

```
rem@theta
```

```
## [1] 0.5566224
```

```
round(fitted(rem),5)
```

```
##       1       2       3       4       5       6       7       8       9      10
## 0.51874 0.62830 0.46296 0.57418 0.57418 0.68014 0.62830 0.51874 0.57418 0.46296
```

We can see the maximum-likelihood estimates are $\hat{\pi}_i = 0.5643724$, $\hat{\alpha} = 0.2589267$, and $\hat{\sigma} = 0.5566224$. We can also see the predicted values of the probability of a head for each of the ten coins.

Problem 10.4

Problem 10.4(a)

```
sub<-read.table("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/Substance.da
library(dplyr)
library(tibble)
library(tidyr)
s<-sub %>%
uncount(count) %>%
rowid_to_column(var="ID") %>%
pivot_longer(cols=-ID,names_to="Drug",values_to="u") %>%
mutate(Drug=factor(Drug,levels=c("cigarettes","alcohol","marijuana"),labels=c(1,2,3)),u=ifelse(u=="yes"
m<-glmer(u~(1|ID)+Drug,family=binomial,nAGQ=100,data=s)
summary(m)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 100) [glmerMod]
##  Family: binomial  ( logit )
## Formula: u ~ (1 | ID) + Drug
##    Data: s
##
##      AIC      BIC   logLik deviance df.resid
##   6629.1   6656.4  -3310.6   6621.1     6824
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.1038 -0.1989  0.1359  0.4133  5.0284
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 12.6     3.549
## Number of obs: 6828, groups:  ID, 2276
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6208     0.1207   13.43   <2e-16 ***
## Drug2         2.6017     0.1421   18.31   <2e-16 ***
## Drug3        -2.3958     0.1272  -18.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation of Fixed Effects:
##       (Intr) Drug2
## Drug2 -0.043
## Drug3 -0.635 -0.092
```

```
library(geepack)
anova(geeglm(u~Drug,id=ID,family="binomial",corstr="exchangeable",data=s))
```

```
## Analysis of 'Wald statistic' Table
## Model: binomial, link: logit
## Response: u
## Terms added sequentially (first to last)
##
##      Df     X2 P(>|Chi|)
## Drug  2 1276.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that $\hat{\beta}_A = 2.6016869$ and $\hat{\beta}_M =$ -2.3957814.

We reject H_0 at the $\alpha = 0.05$ level. There is strong evidence ($p = \chi_2($ 1276.6701834 $) < 0.000000000000001$) that at least one β_i is different.

Problem 10.4(b)

We can see that $\hat{\sigma} = 3.5494429$.

(**i**) The large value implies that the probability of an affirmative response has a high variance as ID changes.

(**ii**) A large positive value for u_i for a particular student implies they "start out" with a high probability of an affirmative response prior to taking variables into account.

Problem 10.4(c)

```
library(gee)
gee(u~Drug,id=ID,family=binomial(link="logit"),data=s)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)       Drug2       Drug3
##   0.6493063   1.1358052  -0.9647252
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Independent
##
## Call:
## gee(formula = u ~ Drug, id = ID, data = s, family = binomial(link = "logit"))
##
## Number of observations :  6828
```

```
##
## Maximum cluster size   :  3
##
##
## Coefficients:
## (Intercept)       Drug2       Drug3
##   0.6493063   1.1358052  -0.9647252
##
## Estimated Scale Parameter:  1.00044
## Number of Iterations:  1
##
## Working Correlation[1:4,1:4]
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
##
##
## Returned Error Value:
## [1] 0
```

We can see the $\{\beta_t\}$ for the GLMM are $\hat{\beta}_A = 2.6016869$ and $\hat{\beta}_M$ = -2.3957814 and the $\{\beta_t\}$ for the GEE are $\hat{\beta}_A = 1.1358052$ and $\hat{\beta}_M$ = -0.9647252. There are notable differences in method and approach between the two models that would lead them to produce different parameter estimates (Section 10.2.5, pages 283-284).

Problem 10.4(d)

```
summary(glm(count~alcohol*cigarettes+alcohol*marijuana+cigarettes*marijuana,family=poisson,data=sub))
```

```
##
## Call:
## glm(formula = count ~ alcohol * cigarettes + alcohol * marijuana +
##     cigarettes * marijuana, family = poisson, data = sub)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  0.02044  -0.02658  -0.09256   0.02890  -0.33428   0.09452   0.49134  -0.03690
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  5.63342    0.05970  94.361  < 2e-16 ***
## alcoholyes                   0.48772    0.07577   6.437 1.22e-10 ***
## cigarettesyes               -1.88667    0.16270 -11.596  < 2e-16 ***
## marijuanayes                -5.30904    0.47520 -11.172  < 2e-16 ***
## alcoholyes:cigarettesyes     2.05453    0.17406  11.803  < 2e-16 ***
## alcoholyes:marijuanayes      2.98601    0.46468   6.426 1.31e-10 ***
## cigarettesyes:marijuanayes   2.84789    0.16384  17.382  < 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: 2851.46098  on 7  degrees of freedom
## Residual deviance:    0.37399  on 1  degrees of freedom
```

```
## AIC: 63.417
##
## Number of Fisher Scoring iterations: 4
```

We can see the random effects and marginal models take a binary response variable and use a logit while the loglinear model uses count data and the Poisson distribution. The focus for the random effects and marginal models appears to be on the independent variable (drug) since it has multiple levels, while the focus in the loglinear model appears to be on the dependent count variable.

Problem 10.4(e)

```
s2<-read.table("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/Substance2.da
s2<-s2 %>%
uncount(count) %>%
rowid_to_column(var="ID") %>%
pivot_longer(cols=-c(ID,R,G),names_to="Drug",values_to="Use") %>% # Excluding "R" and "G"
mutate(Drug=factor(Drug,levels=c("C","A","M"),labels=c(1,2,3)),Use=ifelse(Use==1,1,0))
mrg<-glmer(Use~(1|ID)+factor(Drug,c(3,1,2))*G+R,family=binomial,nAGQ=100,data=s2)
summary(mrg)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 100) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Use ~ (1 | ID) + factor(Drug, c(3, 1, 2)) * G + R
##    Data: s2
##
##      AIC      BIC   logLik deviance df.resid
##   6615.4   6670.0  -3299.7   6599.4     6820
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -8.3451 -0.2074  0.1198  0.4203  5.8731
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 12.7     3.564
## Number of obs: 6828, groups:  ID, 2276
##
## Fixed effects:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -0.5031     0.4841  -1.039 0.298715
## factor(Drug, c(3, 1, 2))1     3.2485     0.3523   9.221  < 2e-16 ***
## factor(Drug, c(3, 1, 2))2     6.3684     0.4779  13.325  < 2e-16 ***
## G                             0.5114     0.2111   2.423 0.015391 *
## R                            -0.9748     0.3312  -2.944 0.003245 **
## factor(Drug, c(3, 1, 2))1:G  -0.5549     0.2096  -2.647 0.008117 **
## factor(Drug, c(3, 1, 2))2:G  -0.8868     0.2684  -3.304 0.000954 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##                 (Intr) fc(D,c(3,1,2))1 fc(D,c(3,1,2))2 G      R
## fc(D,c(3,1,2))1 -0.321
## fc(D,c(3,1,2))2 -0.290  0.566
```

```
## G               -0.645  0.456           0.420
## R               -0.716 -0.034          -0.046          -0.025
## f(D,c(3,1,2))1:  0.316 -0.932          -0.472          -0.478  0.010
## f(D,c(3,1,2))2:  0.301 -0.493          -0.909          -0.453  0.010
##                 f(D,c(3,1,2))1:
## fc(D,c(3,1,2))1
## fc(D,c(3,1,2))2
## G
## R
## f(D,c(3,1,2))1:
## f(D,c(3,1,2))2:  0.507
```

```
mrg@beta
```

```
## [1] -0.5030971  3.2485132  6.3684362  0.5113938 -0.9748350 -0.5548959 -0.8867907
```

```
mrg@theta
```

```
## [1] 3.564027
```

(**i**) We can see from the GLMM that $\hat{\alpha}$ = -0.5030971, $\hat{\beta}_C = 3.2485132$, $\hat{\beta}_A = 6.3684362$, $\hat{\beta}_r = 0.5113938$, $\hat{\beta}_g$ = -0.974835, $\hat{\beta}_{gC}$ = -0.5548959, $\hat{\beta}_{gA}$ = -0.8867907, and $\hat{\sigma} = 3.564027$.

(**ii**) We can see that the two models are different. The signs for the $\hat{\beta}_r$'s and $\hat{\beta}_g$'s are different between the models which illustrates the difference between the two methods.

Problem 10.6

```
d<-read.table("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/Depression.da
mD<-glmer(outcome~(1|case)+severity+drug*time,family=binomial,nAGQ=100,data=d)
summary(mD)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 100) [glmerMod]
##  Family: binomial  ( logit )
## Formula: outcome ~ (1 | case) + severity + drug * time
##    Data: d
##
##      AIC      BIC   logLik deviance df.resid
##   1173.9   1203.5   -581.0   1161.9     1014
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2835 -0.8264  0.2324  0.7963  2.0191
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  case   (Intercept) 0.004331 0.06581
## Number of obs: 1020, groups:  case, 340
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.02795    0.16412  -0.170    0.865
## severity    -1.31521    0.15464  -8.505  < 2e-16 ***
## drug        -0.05970    0.22245  -0.268    0.788
## time         0.48284    0.11596   4.164 3.13e-05 ***
```

```
## drug:time    1.01842    0.19240   5.293 1.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) sevrty drug   time
## severity  -0.385
## drug      -0.614 -0.004
## time      -0.671 -0.133  0.522
## drug:time  0.460 -0.134 -0.740 -0.552
```

```
mD@beta
```

```
## [1] -0.02795299 -1.31521160 -0.05969884  0.48284384  1.01841530
```

The estimated odds of a subject with initial severe depression having their level of mental depression being classified as normal after t weeks are approximately 0.2684175 times the same estimated odds for a subject with initial mild depression, holding all other variables constant.

The estimated odds of a normal classification of a subject's level of mental depression after t weeks of using the new drug are approximately $e^{-0.05969884+1.01841530 \log_2 t}$ times the same estimated odds after t weeks of using the standard drug, where $t = 1, 2$, or 4, holding all other variables constant. (From Problem 9.4(b))

We can see when using the coefficients with time that $e^{\beta_t + \beta_{dt} d} = e^{-0.05969884+1.01841530d}$. The estimated odds of a normal classification of a subject's level of mental depression after t weeks of treatment are approximately $e^{-0.05969884+1.01841530d}$ times the same estimated odds after $\frac{t}{2}$ weeks of treatment, where $t = 1, 2$, or 4, holding all other variables constant.

We can see the interaction term is present if $d = 1$ and $t = 1$ or 2. The estimated odds of a normal classification of a subject's level of mental depression after 4 weeks of using the new drug are approximately 2.7688036 times the same estimated odds after 2 weeks of using the new drug, holding all other variables constant. Additionally, the estimated odds of a normal classification of a subject's level of mental depression after 2 weeks of using the new drug are approximately 2.7688036 times the same estimated odds after *either* 1 week of using the new drug *or any* duration using the standard drug, holding all other variables constant.

We can see this model is almost identical to the marginal model.