Homework 09
Joseph Blubaugh
jblubau1@tamu.edu
STAT 659-700

5.21

```
(alcohol = matrix(c(0, .5, 1.5, 5, 7, 48, 38, 5, 1, 1), nrow = 5))
```

```
     [,1] [,2]
[1,]  0.0   48
[2,]  0.5   38
[3,]  1.5    5
[4,]  5.0    1
[5,]  7.0    1
```

```
fisher.test(alcohol)
```

```
     Fisher's Exact Test for Count Data

data:  alcohol
p-value = 7.423e-14
alternative hypothesis: two.sided
```

5.22

Table 1: malformation

y	x
0	10
0	20
0	30
0	40
1	60
1	70
1	80
1	90

a. β is nearly 0 because there is a clear distinction between y = 0 and y = 1.

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred



Call:
glm(formula = y ~ x, family = binomial(), data = dta)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.045e-05  -2.110e-08   0.000e+00   2.110e-08   1.045e-05

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -118.158 296046.187       0        1
x                2.363   5805.939       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.1090e+01  on 7  degrees of freedom
Residual deviance: 2.1827e-10  on 6  degrees of freedom
AIC: 4

Number of Fisher Scoring iterations: 25
```

b. reported above

c. No because there is still a clear distinction in x when y = 0 or y = 1.

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred



Call:
glm(formula = y ~ x, family = binomial(), data = dta)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.177   0.000   0.000   0.000   1.177

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -98.158  39288.592  -0.002    0.998
x                1.963    785.772   0.002    0.998

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.8629  on 9  degrees of freedom
```

```
Residual deviance:  2.7726  on 8  degrees of freedom
AIC: 6.7726

Number of Fisher Scoring iterations: 21
```

    d. Yes, this is correct because now there is overlap between the point where y = 0 or y = 1 conditional on x. In the other two examples there is not "gray" area so there is nothing to model essentially.

```
Call:
glm(formula = y ~ x, family = binomial(), data = dta)

Deviance Residuals:
     Min        1Q     Median        3Q        Max
-1.19990  -0.00554   0.00000   0.00554   1.19990

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -26.3874    49.0989  -0.537    0.591
x             0.5277     0.9816   0.538    0.591

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.863  on 9  degrees of freedom
Residual deviance:  2.900  on 8  degrees of freedom
AIC: 6.9

Number of Fisher Scoring iterations: 10
```

5.23

penicillin	delay	cured	died
1/8	None	0	6
1/8	1.5hrs	0	5
1/4	None	3	3
1/4	1.5hrs	0	6
1/2	None	6	0
1/2	1.5hrs	2	4
1	None	5	1
1	1.5hrs	6	0
4	None	2	0
4	1.5hrs	5	0

    a. The software reports highly insignificant coefficients for both the first and last penicillin doses. The coefficients are large, but the standard errors are very large and include 0

```r
mdl = glm(cbind(dta$cured, dta$died) ~ penicillin + delay - 1, family = binomial(),
          data = dta)
summary(mdl)
```

```
Call:
glm(formula = cbind(dta$cured, dta$died) ~ penicillin + delay -
    1, family = binomial(), data = dta)

Deviance Residuals:
        1         2         3         4         5         6         7
-0.00011  -0.00003   0.28618  -0.84840   1.09179  -0.47412  -1.81284
        8         9        10
  1.41050   0.00002   0.00010

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
penicillin1/8  -20.7932  7857.3251  -0.003  0.99789
penicillin1/4   -0.2339     0.7737  -0.302  0.76238
penicillin1/2    2.2592     1.1236   2.011  0.04436 *
penicillin1      4.2626     1.5146   2.814  0.00489 **
penicillin4     23.1827  8074.0727   0.003  0.99771
delay1.5hrs     -2.5496     1.1752  -2.170  0.03004 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.4972  on 10  degrees of freedom
Residual deviance:  7.4943  on  4  degrees of freedom
AIC: 25.866

Number of Fisher Scoring iterations: 19
```

b. Based on a pvalue < .05 we would conclude that the odds ratios condtional on penicillin level are not equal.
$3.94 > 3.84 = qchisq(.95, 1)$

```r
library(lawstat)
library(reshape2)

dta = melt(data = dta, id.vars = c("penicillin", "delay"),
           variable.name = "status",
           value.name = "freq")
dta = xtabs(freq ~ delay + status + penicillin, data = dta)
mantelhaen.test(dta)
```

```
     Mantel-Haenszel chi-squared test with continuity correction

data:  dta
Mantel-Haenszel X-squared = 3.9286, df = 1, p-value = 0.04747
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
  1.026713 47.725133
sample estimates:
common odds ratio
               7
```

   c. The conditional odds ratios are not estimatable because every level of penicillin has at least one 0.

cmh.test(dta)

```
     Cochran-Mantel-Haenszel Chi-square Test

data:  dta
CMH statistic = 5.657100, df = 1.000000, p-value = 0.017384, MH
Estimate = 7.000000, Pooled Odd Ratio = 1.846200, Odd Ratio of
level 1 = NaN, Odd Ratio of level 2 = Inf, Odd Ratio of level 3 =
Inf, Odd Ratio of level 4 = 0.000000, Odd Ratio of level 5 = NaN
```

5.24

Since we are dealing with several cells that have 0, we can pool the results over the levels and estimate that the odds ratio is 1.84.

6.1

   a. $log(\frac{\pi_R}{\pi_D}) = log(\frac{\pi_R}{\pi_I}) - log(\frac{\pi_D}{\pi_I}) = (1 + .3x) - (3.3 - .2x) = -2.3 + .5x$ Holding income constant the odds of someone prefering Republican vs Demovcrat is exp(-2.3) = .091

   b. $log(\frac{\pi_R}{\pi_D}) = 0 = -2.3 + .5x \rightarrow 2.3 = .5x \rightarrow x = 2.3/5 = .46$ The point at which the probability of preferring Republican is higher is when income reaches $46K.

   c. $\pi_I = 1/1 + e^{1+.3x} + e^{3.3-.2x}$

6.2

    a. Between Invertebrates or Other the odds of a Aligator preferring Invertibrates over Other increases exp(-2.46) = .085 for unit increase in length

```r
library(nnet)

summary(gator)

     length                    pref
 Min.   :1.240   Other      : 8
 1st Qu.:1.575   Fish       :31
 Median :1.850   Invertebrate:20
 Mean   :2.130
 3rd Qu.:2.450
 Max.   :3.890

mdl = multinom(pref ~ length, data = gator); summary(mdl)

# weights:  9 (4 variable)
initial  value 64.818125
iter  10 value 49.170785
final  value 49.170622
converged

Call:
multinom(formula = pref ~ length, data = gator)

Coefficients:
             (Intercept)      length
Fish            1.617952 -0.1101836
Invertebrate    5.697543 -2.4654695

Std. Errors:
             (Intercept)     length
Fish            1.307291 0.5170838
Invertebrate    1.793820 0.8996485

Residual Deviance: 98.34124
AIC: 106.3412
```

    b. The probability that the gator will prefer fish is .762

```r
predict(mdl, data.frame(length = 3.9), type = "probs")

      Other           Fish Invertebrate
 0.232494677   0.762882158   0.004623165
```

    c. $log(\frac{\pi_I}{\pi_O}) = 0 = 5.697 - 2.465x \rightarrow -5.697 = -2.465x \rightarrow x = 2.31$

6.3

```
(mdl = multinom(pref ~ lake + length, weights = count, data = gator))

# weights:  30 (20 variable)
initial  value 352.466903
iter  10 value 271.952836
iter  20 value 270.040244
final  value 270.040140
converged

Call:
multinom(formula = pref ~ lake + length, data = gator, weights = count)

Coefficients:
   (Intercept)       lakeG       lakeO       lakeT     lengthL
B   -1.2822801   0.1311702 -0.5278093 -0.2975959   0.9623943
F    0.7465176   0.8262827  0.8207557 -0.6900927   0.3314828
I   -1.0027527   2.4847427  3.4165041  2.0903366 -1.1268381
R   -1.6765868 -0.4168093  2.0369590  1.0024618  0.6826978

Residual Deviance: 540.0803
AIC: 580.0803
```

a. $log(\frac{\pi_B}{\pi_O}) = -1.28 + .131(lake_G) - .527(lake_O) - .297(lake_T) + .962(length_L)$

$log(\frac{\pi_F}{\pi_O}) = .7465 + .826(lake_G) + .820(lake_O) - .690(lake_T) + .332(length_L)$

$log(\frac{\pi_I}{\pi_O}) = -1.00 + 2.48(lake_G) + 3.41(lake_O) + 2.09(lake_T) - 1.12(length_L)$

$log(\frac{\pi_R}{\pi_O}) = -1.67 - .416(lake_G) + 2.04(lake_O) + 1.00(lake_T) + .682(length_L)$

b. If the length is short the probability that the gator will prefer Fish is .258, if the gator is long then the probability that it will prefer fish is .458

```
predict(mdl, data.frame(lake = c("O", "O"), length = c("S", "L")), type = "probs")


            O           B         F         I          R
1 0.05385937 0.008813521 0.2581781 0.6019220 0.07722696
2 0.06865543 0.029412068 0.4584523 0.2486419 0.19483832
```

6.4

```
      response     race      gender        count
 No        :4   White:6   Female:6   Min.    :  5.00
 Undecided:4   Black:6   Male  :6   1st Qu.: 14.50
 Yes       :4                       Median : 47.00
                                    Mean   : 82.58
                                    3rd Qu.: 71.75
                                    Max.   :371.00
```

a. $\pi_Y = exp(1.64 - .341)/(1 + exp(1.64 - .341) + exp(-.382 - .27)) = .707$

```
(mdl = multinom(response ~ race + gender, weights = count, data = afterlife))

# weights:  12 (6 variable)
initial  value 1088.724778
iter  10 value 773.731422
final  value 773.726510
converged

Call:
multinom(formula = response ~ race + gender, data = afterlife,
    weights = count)

Coefficients:
          (Intercept)   raceBlack genderMale
Undecided   -0.382000 -0.2709286 -0.1050516
Yes          1.643358 -0.3417948 -0.4185436

Residual Deviance: 1547.453
AIC: 1559.453

predict(mdl, data.frame(gender = "Female", race = "Black"), type = "probs")

       No Undecided       Yes
0.1924721 0.1001854 0.7073425
```

b. $log(\frac{\pi_U}{\pi_N}) = -.382 - .27(Race_B) - .105(Gender_M)$ The conditional odds of Undecided to No is multiplied by exp(-.105) = .9 when the gender is male vs female $log(\frac{\pi_Y}{\pi_U}) = (-.382 - 1.643) - (.2709 - .341)Race_B - (.105 - .418)Gender_M$ The conditional odds ratio of Yes to Undecided is multiplied by exp(-.105 - .418) = .592 when gender is male.

6.5

    a.

    b. x_1: Those who make more money relative to their peers are less satisfied with their job compared to those who make less money `exp(-.54 * 1:4) / (1 + exp(-.54 * 1:4))` = `[1] 0.3681876 0.2535060 0.1652049 0.1034005`

    ii. x_2: Employees have higher job satisfaction with less freedom `exp(.6 * 1:4)  / (1 + exp(.6 * 1:4))` = `[1] 0.6456563 0.7685248 0.8581489 0.9168273`

    iii. x_3: Work environments that are less productive have higher job satisfaction `exp(1.19 * 1:4) / (1 + exp(1.19 * 1:4))` = `[1] 0.7667411 0.9152894 0.9726152 0.9915071`

    b. x1 = 1, x2 = 4, x3 = 4

6.6

    a. $log(\frac{\pi_{H1}}{\pi_{H3}}) = -2.55 - .2275x$
       $log(\frac{\pi_{H2}}{\pi_{H3}}) = -.353 - .0962x$
       $log(\frac{\pi_{H1}}{\pi_{H2}}) = (-2.3 - .353) - (.2275 - .0962)x$

    b. The increase in odds for marital happiness increases exp(-.2275) = .79 for each increased level of income

    c. Wald: .9432, Pvalue: .624

    d. Based on the value/df ratio of 1.5 it appears that the mode does not fit adequately. The ratio should be close to 1 with a well fit model.

    e. $\pi_3 = 1/(1 + exp(-2.55 - .2275) + exp(-.351 - .0962)) = .587$

6.7

    a. 2 intercepts are fit because the response has 3 levels. There is only 1 income effect because it is assumed fixed across the 3 levels. If it were not fixed then there is no garuntee that the sum of the 3 probabilities of the response would add to 1.

    b. For each unit increase in income, odds of of increasing marital hapiness is multiplied times exp(.65) = 1.91

    c. Chi-square: .897, pvalue: .343, because the pvalue is > .05 we conclude that income does not have a significant affect on marital happiness

    d. Based on the deviance/df = 1.08, the model fits the data reasonably well.

    e. $\pi_2 = \frac{exp(-.237+(2*-.117))-exp(-3.24+(2*-.111))}{(1+exp(-3.24+(2*-.111)))*(1+exp(-.237+(2*-.117)))} = .35$

6.8

a. Sequential Effect: $exp(.5807) = 1.78$ The increase in odds for Sequential over Alternating is a multiplier of 1.78 Gender Effect: $exp(.541) = 1.71$ The increase in odds for Male over Female is a multiplier of 1.78

```
        therapy      gender                    response       count
 Alternating:8   Female:8   Progressive       :4   Min.   : 1.00
 Sequential :8   Male  :8   No Change         :4   1st Qu.: 4.75
                            Partial Remission :4   Median :16.00
                            Complete Remission:4   Mean   :18.69
                                                   3rd Qu.:28.25
                                                   Max.   :45.00
```

```
library(MASS)
mdl = polr(response ~ therapy + gender, weights = count, data = cancer); summary(mdl)
```

```
Re-fitting to get Hessian


Call:
polr(formula = response ~ therapy + gender, data = cancer, weights = count)

Coefficients:
                  Value Std. Error t value
therapySequential 0.5807    0.2121   2.737
genderMale        0.5414    0.2872   1.885

Intercepts:
                                       Value    Std. Error t value
Progressive|No Change                  -0.1960  0.2893     -0.6774
No Change|Partial Remission             1.3713  0.3000      4.5706
Partial Remission|Complete Remission    2.4221  0.3224      7.5119

Residual Deviance: 789.0566
AIC: 799.0566
```

b. Interaction: $exp(-.59) = .55$ The treatment effect for a Male vs a female is a multiplier of .55 so the treatment is more effective for females than it is for males.

```
mdl = polr(response ~ therapy*gender, weights = count, data = cancer); summary(mdl)
```

```
Re-fitting to get Hessian


Call:
polr(formula = response ~ therapy * gender, data = cancer, weights = count)

Coefficients:
                              Value Std. Error t value
therapySequential            1.0785     0.5336   2.021
genderMale                   0.8646     0.4316   2.003
therapySequential:genderMale -0.5904     0.5792  -1.019

Intercepts:
                                    Value    Std. Error t value
Progressive|No Change               0.0770   0.3988     0.1931
No Change|Partial Remission         1.6484   0.4099     4.0211
Partial Remission|Complete Remission 2.6978  0.4257     6.3380

Residual Deviance: 788.0098
AIC: 800.0098
```

c. Based on the lower AIC the interaction model is better than the model with no interaction

6.9

a. Political ideology is an ordered factor so you expect the coefficients to be ordered as well. There are only 4 because the 5th logit is redundant and can dervied from the other 4.
b. The group by that is the most liberal is the relgious group: none because all of the other coefficients for religion are negative. The group that is the most conservative is the religion group: Protestant because the have the smallest coefficient.
c. Protestant: $exp(-1.03 - 1.27)/(1 + exp(-1.03 - 1.27)) = .091$
   None: $1/(1 + exp(-1.03 - .44)) = .81$
d. Protestant/None: $exp(-1.27) = .28$ The effect of being more conservative between Protestant/None is a multiplier of .28 Protestant/Catholic: $exp(-1.27 - 1.22) = .082$ The effecto of being more conservative between Protestant/Catholic is .082

Additional

$log(\frac{\pi_1}{\pi_3}) = \alpha_1 + \beta_1 x \rightarrow \frac{\pi_1}{\pi_3} = exp(\alpha_1 + \beta_1 x) \rightarrow \pi_1 = \frac{exp(\alpha_1 + \beta_1 x)}{(1 + exp(\alpha_1 + \beta_1 x)) + (1 + exp(\alpha_2 + \beta_2 x))}$
$log(\frac{\pi_2}{\pi_3}) = \alpha_2 + \beta_2 x \rightarrow \frac{\pi_2}{\pi_3} = exp(\alpha_2 + \beta_2 x) \rightarrow \pi_2 = \frac{exp(\alpha_2 + \beta_2 x)}{(1 + exp(\alpha_1 + \beta_1 x)) + (1 + exp(\alpha_2 + \beta_2 x))}$
$\pi_3 = 1 - \pi_1 - \pi_2$