MA 576 - Take Home Exam

Benjamin Draves

```
#load necessary packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
```

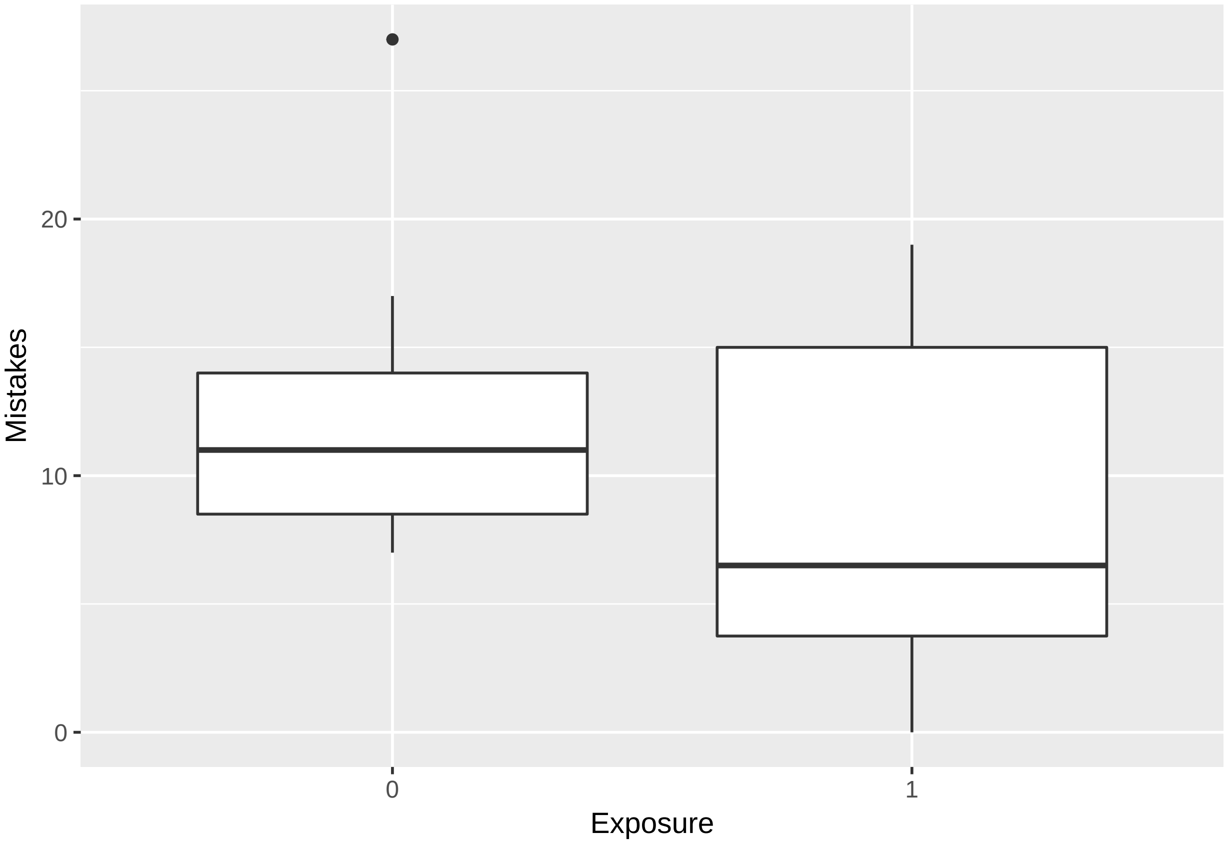
Problem 2

(a)

```
#read in data
mem = read.csv("~/Desktop/Courses/MA 576/data/memory.csv", header = F)
colnames(mem) = c("Exposure","Age","Mistakes")
mem$Exposure = as.factor(mem$Exposure)
head(mem)
```

```
##   Exposure Age Mistakes
## 1        0  23        8
## 2        0  47       17
## 3        0  47       27
## 4        0  23        8
## 5        0  32       14
## 6        0  29       10
```

```
#Make boxplot
ggplot(mem, aes(y = Mistakes, x = Exposure)) + geom_boxplot()
```



Here we see that the new study method reduced the number of median mistakes but simulatenoulsly increased the range, or spread, of the number of mistakes made. We see that the IQR is about three times the size of that of the number of mistakes under no exposure to the new study method. #### (b)

```
#Fit poisson glm
model = glm(Mistakes~Exposure, data = mem, family  = poisson)
summary(model)
```

```
##
## Call:
## glm(formula = Mistakes ~ Exposure, family = poisson, data = mem)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1952  -1.4161  -0.5589   0.5558   3.4826
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.54160    0.08874  28.642  < 2e-16 ***
## Exposure1   -0.36685    0.13870  -2.645  0.00817 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 81.476  on 19  degrees of freedom
## Residual deviance: 74.362  on 18  degrees of freedom
## AIC: 156.87
##
## Number of Fisher Scoring iterations: 5
```

This is a model with two parameters (β_0, β_1) correpsonding to the model given by

$$Mistakes = \exp(\beta_0 + \beta_1 1_{Expsoure})$$

where $1_{Expsoure}$ is the indicator for weather the subject followed the new study method. Hence, setting $1_{Expsoure} = 0$, we model $Mistakes = \exp(\beta_0)$. Therefore, under no exposure to the new study method, we expect a subject to make

$$\widehat{Mistakes} = \exp(\hat{\beta}_0) = \exp(2.54160) = 12.69997$$

miskates. Now when $1_{Expsoure} = 1$, we expect the subects to make

$$\widehat{Mistakes} = \exp(\hat{\beta}_0)\exp(\hat{\beta}_1) = \exp(2.54160)\exp(-0.36685) = 8.799985$$

many mistakes. In other words, we expect a subject subject to the new study method to make $\exp(\hat{\beta}_1) = 0.6929136$ times less mistakes than those subjects who were not exposed.

```r
#Build Exposure 95% T-Confidence interval
t = qt(1 - .05/2,model$df.residual)
se = 0.13870 #mined from summary above
left=-0.36685 - (se * t)
right = -0.36685 + (se * t)
c(left, right)
```

```
## [1] -0.65824789 -0.07545211
```

(c)

There is evidence for overdisperion. We see that the residual deviance is 74.362 on 18 degrees of freedom. With the fact that $D \overset{asy}{\sim} \chi^2(n-p)$ for $p = 2$ in this case, we expect the residual deviance to be near 18. While 74.362 does not seem near 18, we will complete a χ^2 test for signficance.

```r
#get chisq test statistic
sigma2 = sum(residuals(model, type = "pearson")^2)/(model$df.residual)
ts = sigma2*model$df.residual
#define rejection region
cuttoff = qchisq(0.95, model$df.residual)
#preform test
cuttoff < ts
```

```
## [1] TRUE
```

Therefore, having rejected the hypothesis that $\hat{\sigma}^2 < 1$ we expect that this model is overdispersed.

```r
#Fit QuasiPoisson
model2 = glm(Mistakes ~ Exposure, data = mem, family = quasipoisson)
summary(model2)
```

```
##
## Call:
## glm(formula = Mistakes ~ Exposure, family = quasipoisson, data = mem)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1952  -1.4161  -0.5589   0.5558   3.4826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5416     0.1782  14.265 2.98e-11 ***
## Exposure1    -0.3669     0.2785  -1.317    0.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 4.031755)
##
##     Null deviance: 81.476  on 19  degrees of freedom
## Residual deviance: 74.362  on 18  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
#Build Exposure 95% T-Confidence interval
t = qt(1 - .05/2,model$df.residual)
se = 0.2785 #mined from summary above
left=-0.36685 - (se * t)
right = -0.36685 + (se * t)
c(left, right)
```

```
## [1] -0.9519568  0.2182568
```

Here we note that only the estimates of the standard error change. This change however, now includes the value 0. In words, having accounted for dispersion in this model, we now cannot say with certainity that the new study method has an effect on the number of mistakes a subject makes.

(d)

```r
#fit with both exposure and age
model3 = glm(Mistakes ~ Exposure + Age, data = mem, family = poisson)
summary(model3)
```

```
##
## Call:
## glm(formula = Mistakes ~ Exposure + Age, family = poisson, data = mem)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7393  -0.9205  -0.1355   0.3474   2.2305
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.341156   0.204232   6.567 5.14e-11 ***
## Exposure1   -0.749200   0.157730  -4.750 2.04e-06 ***
## Age          0.031753   0.004673   6.795 1.09e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 81.476  on 19  degrees of freedom
## Residual deviance: 30.598  on 17  degrees of freedom
## AIC: 115.11
##
## Number of Fisher Scoring iterations: 4
```

Here we see that this model is parameterized by three values $(\beta_0, \beta_{Exposure}, \beta_{Age})$. We can interpret $\hat{\beta}_0 = 1.341156$ as follows. For an age zero subject with no exposure to the treatment, we expect them to make $\exp(1.341156) = 3.823461$ mistakes. As an "age zero" subject doesn't necessariliy exist, the age zero subject with no exposure to the new study method will serve as the baseline group. Next, we can interpret $\hat{\beta}_{Exposure} = -0.749200$ as follows. For an age zero patient with exposure to the new study method, we expect them to make $\exp(1.341156 - 0.749200) = 1.80752$ mistakes or $\exp(-0.749200) = 0.4727446$ times fewer mistakes than the age zero subject with no exposure to the new study method. Lastly, we can interpret the $\hat{\beta}_{Age} = 0.031753$ estimate as follows. For a subject with no exposure to the new study method, we expect them to make $\exp(0.031753) = 1.032263$ times more mistake each year.

The AIC in this model is 115.11 compared to the 156.87 AIC in the exposure only model. This statistic suggests that this model is more consistent with the data than the expsore model.

(e)

As in part *c* we will complete a χ^2 test to see if this model is overdispersed.

```r
#get chisq test statistic
sigma2 = sum(residuals(model2, type = "pearson")^2)/(model$df.residual)
ts = sigma2*model$df.residual
#define rejection region
cuttoff = qchisq(0.95, model$df.residual)
#preform test
cuttoff < ts
```

```
## [1] TRUE
```

Here, we fail to reject the null hypothesis that $\sigma^2 \leq 1$ and hence do not have evidence of overdispersion.

With this, under regularity of the residuals, that we do not need to rescale our standard errors. Thus, we see that both age and exposure are significant in explaining varition in the the number of mistakes recorded. Moreover, we see that the exposure to the new study *decreases* the expected number of miskates made while increase in age *increases* the expected number of mistakes made by the subject.

Problem 3

(a)

We'll define some thing we'll need in the IRLS procedure here. With the log-link we see that $g^{-1}(\mu) = e^{\mu}$ and $g'(\mu)^2 = \frac{1}{\mu^2}$. From problem 1 we see that $V(\mu) = \mu^2 + \mu$. Together we see that

$$W = \text{diag}([V(\mu)g'(\mu)^2]^{-1}) = \text{diag}([\frac{\mu(1+\mu)}{\mu^2}]^{-1}) = \text{diag}[\frac{\mu}{1+\mu}]$$

Lastly, we define $g'(\mu) = \frac{1}{\mu}$. From here we are ready to implement this IRLS procedure.

```
IRLS = function(df){
  #IRLS------------------------------

  #set up dataframes
  Y = as.matrix(as.numeric(as.character(df[,1])), ncol = 1)
  X = as.numeric(as.character(df[,2:ncol(df)]))
  X = cbind(rep(1, length(X)),X)

  #initialize eta
  beta = matrix(0, nrow = ncol(X))
  eta = X%*%beta

  #threshold values
  diff = Inf
  thres = 0.01

  #iterate
  while(diff>thres){

    #update mu
    mu = exp(eta)

    #update weights
    W = diag(c((mu)/(mu+1)))

    #updates Z
    Z = eta + diag(c(1/(mu))) %*% (Y-mu)

    #Fisher matrix
    I = t(X) %*% W %*% X
    #update beta
    beta.new = solve(I)%*%t(X)%*%W%*%Z

    #record difference - l^2 distance for stopping criterion
    diff = sqrt(sum((beta.new -beta)^2))
    beta = beta.new

    #update eta
    eta = X%*%beta

  }

  ret = list(beta, I)

  return(ret)
}
```

(b)

```
#Set up data frame
df = mem[,c(3,1)]

#Get estimates
tmp = IRLS(df)
MLEst = tmp[[1]]
Fisher = tmp[[2]]

#print out ML
MLEst
```

```
##           [,1]
##     2.5416020
## X -0.3668503
```

Here we see that the ML estimates are very close to that of in problem 2 of part *c*.

(c)

We can find the standard error estimates from the inverse of the Fisher Information matrix. That is $se(\hat{\beta}_i) = (\sqrt{I^{-1}})_{ii}$.

```
se = sqrt(diag(solve(Fisher)))
se
```

```
##                    X
## 0.3284412 0.4682278
```

Here we see that $se(\hat{\beta}_0) = 0.3284412$ and $se(\hat{\beta}_1) = 0.4682278$ which is higher than both of the estimates in the poisson and quasipoisson model. From these we we can build confidence interval for the exposure random variable as follows.

```
#Beta Exposure 95% CI
t = qt(1 - .05/2,20-2)
left = MLEst[2,1] - (se[2] * t)
right = MLEst[2,1] + (se[2] * t)
c(left, right)
```

```
##          X          X 
## -1.3505603  0.6168597
```

(d)

To calculate $AIC = 2p - 2\log(\hat{\mathcal{L}})$ for $\log(\hat{\mathcal{L}}) = \sum_{i=1}^{n} y_i\hat{\theta}_i - b(\hat{\theta}_i)$ for $\hat{\theta}_i = \log(\frac{\hat{\mu}_i}{1+\hat{\mu}_i})$ and $b(\hat{\theta}_i) = \log(\frac{1}{1-e^{\hat{\theta}_i}}) = \log(1+\hat{\mu}_i)$ (see derivation in problem 1 for these values).

```
p = 2
y = as.numeric(as.character(mem$Mistakes))
x = as.numeric(as.character(mem$Exposure))

#get theta hat and mu hat
mu_hat = exp(2.5416020 - 0.3668503*x)
theta_hat = log(mu_hat/(1+mu_hat))

l = sum(y*theta_hat -log(1 + mu_hat))

AIC = 2*p - 2*l
AIC
```

```
## [1] 140.1902
```

Here we see that this model reduces this AIC significantly; from 115.11 from part (e) and 156.87 from part (b) to 140.1902. As this is the model corresponding to the hierachial framework, we implicitly model a stochastic rate parameter in this model. By taking this into account, we see that this model fits the data better than the Poisson Regression that ignores this variability in the ML procedure.