WMAN 633 Exam 2

Joe Kingsbury

Question 1

Import this dataset into R and inspect the first several rows of your data

```
# setwd(choose.dir())
exmdat <- read.csv("Exam 2 Data.csv")
head(exmdat)
```

```
##    y          x1          x2 x3
## 1  2  1.37034210 -0.66615843  b
## 2  2 -0.70417232 -0.03705622  c
## 3  4 -0.04223752 -1.53148692  c
## 4  9 -0.56711072 -0.06529335  b
## 5  1  0.31189773  0.81650472  a
## 6 13 -0.35994479 -0.80042308  b
```

Question 2

Fit a Poisson model that assumes your response is a function of x1, x2, and x3. Include an interaction between x1 and x2 only (i.e., do not include an interaction between your categorical variables and any other variables).

```
fit <- glm(y ~ x1 * x2 + x3, family = poisson, data = exmdat)
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x1 * x2 + x3, family = poisson, data = exmdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3620  -0.6973  -0.1007   0.5236   2.6779
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.13258    0.08540  13.262  < 2e-16 ***
## x1          -1.03491    0.05019 -20.620  < 2e-16 ***
## x2          -0.90839    0.06977 -13.021  < 2e-16 ***
## x3b          0.37532    0.09246   4.059 4.92e-05 ***
## x3c         -0.88354    0.12072  -7.319 2.50e-13 ***
## x1:x2       -0.28868    0.05142  -5.614 1.98e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 720.551  on 99  degrees of freedom
## Residual deviance:  89.088  on 94  degrees of freedom
## AIC: 392.86
##
## Number of Fisher Scoring iterations: 5
```

Question 3

Interpret the effect of variable x1 when x2 = -1

```
b <- coef(fit)

b[2] + b[6] * -1
```

```
##         x1
## -0.7462245
```

The log expected count decreases by -0.746 for each 1-unit increase in x1 when x2 = -1

Question 4

Plot expected counts ±90% confidence intervals over the observed range of variable x1. Assume variable when x2 = -1 and category "a"

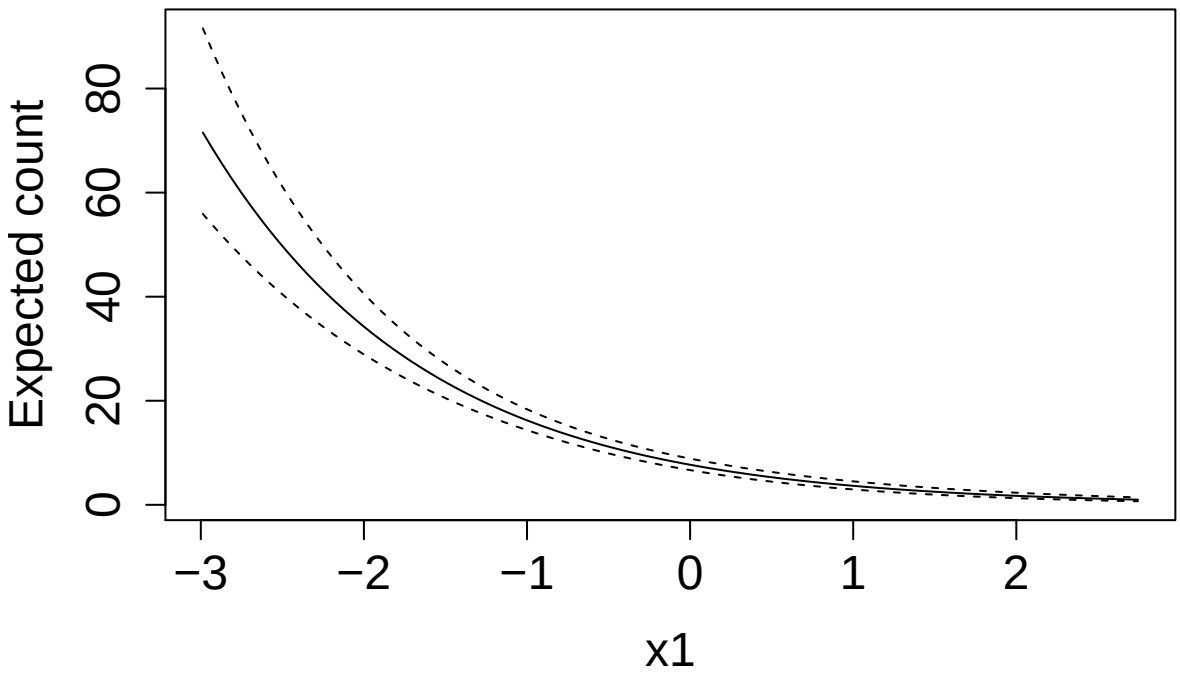
```
data2 <- data.frame(
x1 = seq(from = min(exmdat$x1), to = max(exmdat$x1),length.out = 100),
x2 = -1,
x3 = factor('a', levels = c('a','b','c')))

head(data2)
```

```
##          x1 x2 x3
## 1 -2.987603 -1  a
## 2 -2.929689 -1  a
## 3 -2.871776 -1  a
## 4 -2.813862 -1  a
## 5 -2.755948 -1  a
## 6 -2.698035 -1  a
```

```
prd1 <- predict.glm(object = fit, newdata = data2, se.fit = T)
low <- exp(prd1$fit - qnorm(0.95) * prd1$se.fit)
high <- exp(prd1$fit + qnorm(0.95) * prd1$se.fit)

plot(y = exp(prd1$fit), x = data2$x1, xlab = 'x1',
  ylab = 'Expected count', cex.axis = 1.5, cex.lab = 1.5,
  ylim = c(min(low), max(high)), type = 'l')
  lines(x = data2$x1, y = low, lty = 2)
  lines(x = data2$x1, y = high, lty = 2)
```



Question 5

Interpret the effect of variable x3

```
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x1 * x2 + x3, family = poisson, data = exmdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3620  -0.6973  -0.1007   0.5236   2.6779
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.13258    0.08540  13.262  < 2e-16 ***
## x1          -1.03491    0.05019 -20.620  < 2e-16 ***
## x2          -0.90839    0.06977 -13.021  < 2e-16 ***
## x3b          0.37532    0.09246   4.059 4.92e-05 ***
## x3c         -0.88354    0.12072  -7.319 2.50e-13 ***
## x1:x2       -0.28868    0.05142  -5.614 1.98e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 720.551  on 99  degrees of freedom
## Residual deviance:  89.088  on 94  degrees of freedom
## AIC: 392.86
##
## Number of Fisher Scoring iterations: 5
```

The difference in the log expected count between 'a' and 'b' is 0.38 and the difference between 'a' and 'c' for the expected count is -0.88.

Question 6

Use contrasts to evaluate the null hypothesis that the difference in log expected count between levels "b" and "c" = 0. Fix x1 and x2 at their means

```
library(multcomp) #For the glht function and hypothesis test
```

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

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Warning: package 'TH.data' was built under R version 4.0.4
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
b[5] - b[4]
```

```
##       x3c
## -1.258863
```

```
#Constrast Matrix

x <- matrix(c(0, mean(exmdat$x1), mean(exmdat$x2), -1, 1, ((mean(exmdat$x1) * mean(exmdat$x2)) )), nrow
x
```

```
##      [,1]       [,2]       [,3] [,4] [,5]         [,6]
## [1,]    0 -0.1320536 0.03782269   -1    1 -0.004994622
```

```
cntr <- glht(model = fit, linfct = x)
summary(cntr, test = adjusted('none'))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = y ~ x1 * x2 + x3, family = poisson, data = exmdat)
##
## Linear Hypotheses:
##        Estimate Std. Error z value Pr(>|z|)
## 1 == 0  -1.1551     0.1184  -9.759   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

When x1 and x2 are set at their means the 1-unit change in log expected count between 'b' and 'c' is significantly different from 0. The conclused is bason on the very small p-value and we reject the null hypothesis based on this value.

Question 7

Derive the test statistic and p-value associated with the interaction between x1 and x2. What is the null hypothesis? Do we reject or fail to reject this null hypothesis? Defend your answer.

```
#Test Statistic
ts <- b[6] / summary(fit)[['coefficients']]['x1:x2','Std. Error']
ts
```

```
##     x1:x2
## -5.614074
```

```
#P-value
pnorm(-1 * abs(ts)) * 2
```

```
##        x1:x2
## 1.976182e-08
```

The null hypothesis is b5 = 0. We reject the null hypothesis based on a very small p-vlaue of 1.98e-08. What this means is that there is sufficient evidence that the effect of x1 is dependent on the level of x2.

Question 8

assume you have the following realizations of random variable Y : y = (1, 0) Further assume realizations of the random variable Y are Bernoulli distributed: y ~ Bernoulli(p). What is the probability of observing each of these random variables assuming the log odds of success = -2?

```
p <- plogis(-2)
x_1 <- c(1,0)
dbinom(x_1, size = 1, p = p)
```

```
## [1] 0.1192029 0.8807971
```

Question 9

What is the "support" of a Bernoulli random variable? What are the acceptable values of it's sole parameter? To which quantity do we apply a link function, and why do we do this? What is the principle link function we use in binomial (i.e., logistic) regression, and what it it's inverse function?

Answer The "support" of a Bernoulli random variables are the real numbers 0 and 1 representing success or failure. A link function is applied to the parameters which allows us to map things like temperautre, cloud cover, precipitation, etc to an unconstrained number line. The principle link function we use in binomial is logit link (logistic). The inverse function of logit link is "plogis()" which takes a an unconstrained vector and squishes it between 0 and 1.

Question 10

What is a fundamental assumption we make to derive inference when comparing two levels of a categorical random variable?

Answer The fundamental assumption we make when deriving inferences while comparing two levels of a categorical variable is that linear combinations of Gaussian random variables are themselves random variables. This assumption enables us to perform a hypothesis test.