HW3 Edward Trout

Question 1: Delectable Diamonds

```
#read and inspect data
diamond <- read.csv("diamond.csv")
head(diamond,3)
```

```
##   price     cut carat
## 1   326   Ideal  0.23
## 2   326 Premium  0.21
## 3   327    Good  0.23
```

What is the effect of each cut on the price of a typical diamond?

```
#build model
di.mod <- glm(diamond$price~diamond$cut, family = "poisson")

#Coefficients
coef(di.mod)
```

```
##          (Intercept)      diamond$cutGood     diamond$cutIdeal
##            8.3799424           -0.1038367           -0.2316292
##   diamond$cutPremium diamond$cutVery Good
##            0.0504411           -0.0904632
```

These coefficients show the effect size of each cut against the comparative control of the model- which in this case is *Fair* (merely because it is the first cut alphabetically). To make this model more interpretable, let's reorder the cuts such that they're not alphabetical but in descending order of mean price (e.g. the first is *Premium*).

```
#reordering as a new categorical variable "cut2"
diamond$cut2 <- factor(diamond$cut, levels = c("Premium", "Fair", "Very Good", "Good", "Ideal"))

#new model
di.mod2 <- glm(diamond$price~diamond$cut2, family = "poisson")

#new coefficients
coef(di.mod2)
```

```
##           (Intercept)      diamond$cut2Fair diamond$cut2Very Good
##             8.4303835           -0.0504411           -0.1409043
##      diamond$cut2Good     diamond$cut2Ideal
##            -0.1542778           -0.2820703
```

Now we can see that every other cut besides the intercept (*Premium*) has a negative effect, which means all diamonds are less expensive than *Premium* diamonds. But by how much?

```
#interpret parameters
exp(coef(di.mod2)[1])                                          #premium
```

```
## (Intercept)
##    4584.258
```

```
exp(coef(di.mod2)[1]) - exp(coef(di.mod2)[1] + coef(di.mod2)[2])  #fair
```

```
## (Intercept)
```

```
##    225.4999
```

```
exp(coef(di.mod2)[1]) - exp(coef(di.mod2)[1] + coef(di.mod2)[3])  #very good
```

```
## (Intercept)
##    602.4978
```

```
exp(coef(di.mod2)[1]) - exp(coef(di.mod2)[1] + coef(di.mod2)[4])  #good
```

```
## (Intercept)
##    655.3933
```

```
exp(coef(di.mod2)[1]) - exp(coef(di.mod2)[1] + coef(di.mod2)[5])  #ideal
```

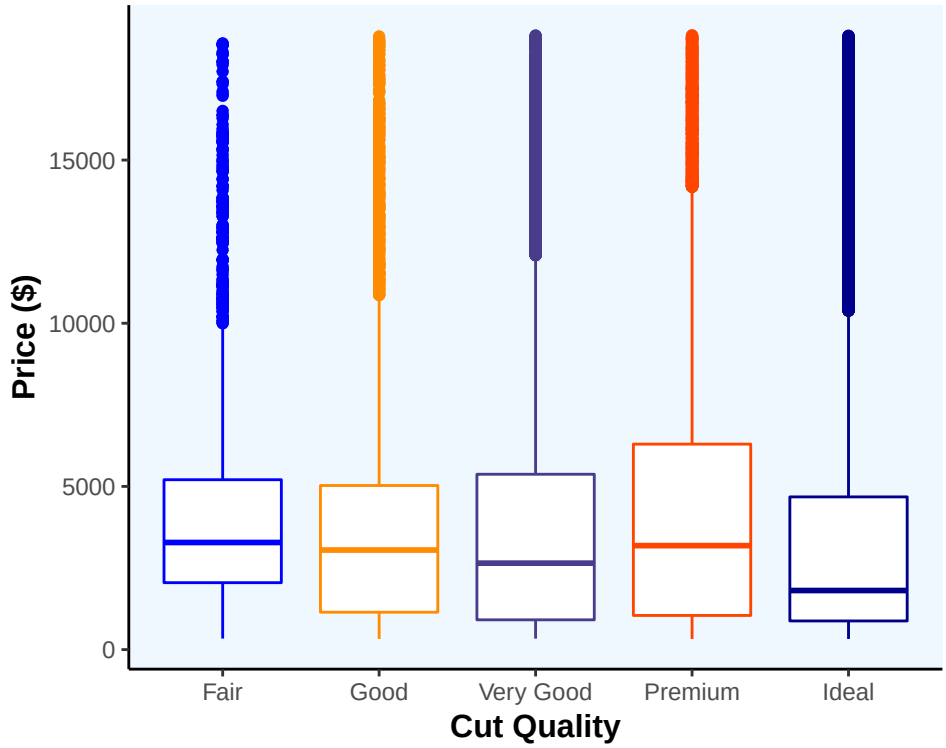
```
## (Intercept)
##    1126.716
```

Premium diamonds cost an average of $4584.26. *Fair* diamonds cost $225.50 less so $4358.76. *Very Good* diamonds cost $602.50 less, $3981.76. *Good* diamonds cost $3928.87, and *Ideal* diamonds cost $3457.54.

```
#confidence intervals
confint(di.mod2)
```

```
##                              2.5 %      97.5 %
## (Intercept)             8.43013697  8.43062996
## diamond$cut2Fair       -0.05122103 -0.04966133
## diamond$cut2Very Good -0.14127928 -0.14052931
## diamond$cut2Good       -0.15478774 -0.15376782
## diamond$cut2Ideal      -0.28240541 -0.28173514
```

All the slopes have very small confidence intervals, and none overlap zero, indicating a good effect of cut quality on diamond price.



Overall there seem to be a large amount of high-priced outliers for every cut type. This suggests that while

most diamonds are sold at lower prices, there are a considerable number highly over-sold, perhaps to unwitting buyers. It is interesting that the highest quality cut- *Ideal*, has the lowest effect on price, perhaps due to lower demand for the highest quality. The second rating of *Fair* suggests that this is the most sought after diamond after *Premium*.

Question 2: Cautious Contraception

```
#read in data
contra <- read.csv("contraception.csv")
head(contra, 3)
```

```
##   age education notUsing using Total
## 1 <25       low       53     6    59
## 2 <25       low       10     4    14
## 3 <25      high      212    52   264
```

Does higher education increase contraception use?

```
# create proportion variable and response variable
contra$prop <- (contra$using/contra$Total)
response <- cbind(contra$using, contra$notUsing)

# build model
contra.mod <- glm(response ~ contra$education, family = "binomial")

# coefficients
coef(contra.mod)
```

```
##         (Intercept) contra$educationlow
##         -0.81020374          0.09248529
```

```
#interpretation
plogis(-.81020374) - plogis(-.81020374 + .09248529)
```

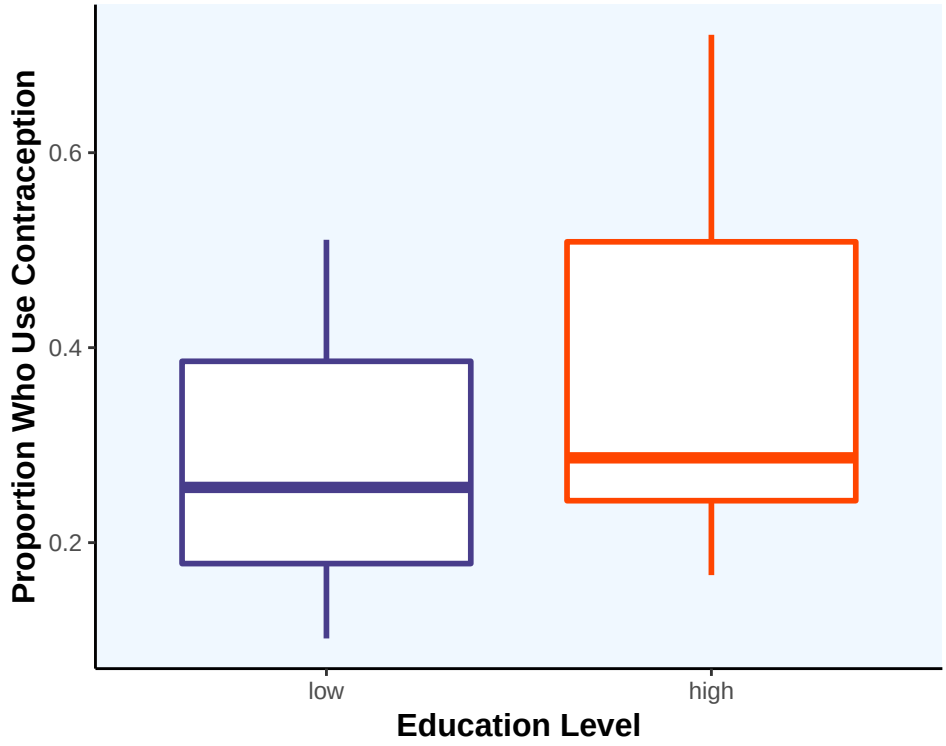
```
## [1] -0.02004851
```

This suggests a 2% difference in proportion of individuals using contraception based on education level. This is a very small difference, amounting to about 32 individuals in this dataset.

```
#confidence intervals
confint(contra.mod)
```

```
##                            2.5 %     97.5 %
## (Intercept)         -0.9460962 -0.6766394
## contra$educationlow -0.1239481  0.3078275
```

This reveals that the confidence interval for this effect crosses zero, which indicates that this is not a significant effect. We can reject the null that education has an effect on contraception use.



Question 3: Himmicanes and Hurricanes

```
# read and inspect data
hur_him <- read.csv("Hurricane Dataset.csv")

# there is a gender type of blank "" due to some header rows, make subset
hur_him <- subset(hur_him, hur_him$Gender_MF!="")
```

Does the gender of a hurricane's name affect its death toll?

```
#build model
cane.mod <- glm(hur_him$alldeaths ~ hur_him$Gender_MF, family = "poisson")

#coefficients
coef(cane.mod)
```

```
##        (Intercept) hur_him$Gender_MFM
##          3.1679220         -0.5123354
```

```
#interpretation
exp(3.1679220) - exp(3.1679220 - 0.5123354)
```

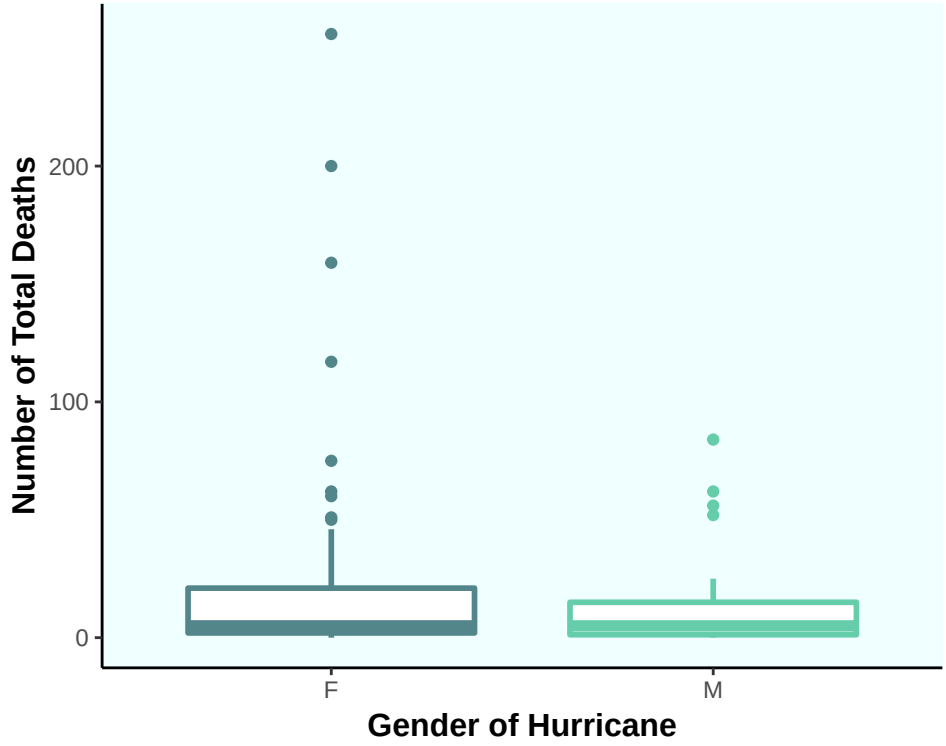
```
## [1] 9.524731
```

These parameters do suggest that there are less deaths from a male-named "himmicane", an effect of 10 less deaths.

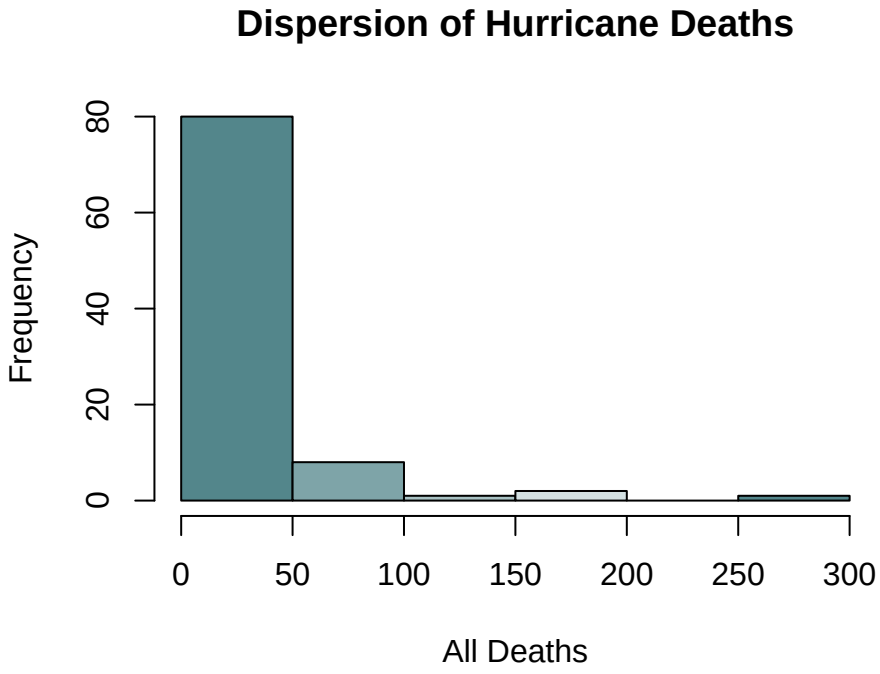
```
confint(cane.mod)
```

```
##                          2.5 %     97.5 %
## (Intercept)          3.1164152  3.2185581
## hur_him$Gender_MFM -0.6211542 -0.4056501
```

The confidence interval does not cross zero, therefore we can take the result to be significant, there are less deaths from male-named "himmicanes". 10 is still a very small effect size.



What could have Jung et al. done differently to inspire more confidence in their analyses?



As we can see from the boxplot above and this histogram, the death toll of hurricanes is heavily overdispersed. Perhaps a negative binomial distribution would be better to estimate these data rather than the poisson.

```
#build new model
cane.mod2 <- glm.nb(hur_him$alldeaths~hur_him$Gender_MF)

#coefficients
coef(cane.mod2)
```

```
##        (Intercept) hur_him$Gender_MFM
##          3.1679220         -0.5123354
```

```
summary(cane.mod2)[12]
```

```
## $coefficients
##                       Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)          3.1679220  0.1900111 16.672297 2.084301e-62
## hur_him$Gender_MFM -0.5123354  0.3341535 -1.533234 1.252182e-01
```

```
summary(cane.mod2)[18]
```

```
## $theta
## [1] 0.4552964
```

```
confint(cane.mod2)
```

```
##                         2.5 %    97.5 %
## (Intercept)          2.816448 3.5640722
## hur_him$Gender_MFM -1.149166 0.1720959
```

Once the analysis is run with a more appropriate probability distribution, it is clear that the effect found (9.5 or 10 more deaths) is *not* significant. The coefficient slope has a p-value of .125 and the confidence interval now overlaps zero. The dispersion parameter for the negative binomial (k for us in class, θ in R) is .4553. In assuming before that the deaths conformed to a poisson distribution, it gave the high outlier death toll of female-named hurricanes too much weight.

Question 4: Our Own Data

My data consists of camera trapping results- perfect count data for a poisson distribution

```
# a record of every capture event from last season
cam1min <- read.csv("record_table_1min_deltaT_2018-12-13.csv")

# a list of the camera stations and their gps locations
cameras <- read.csv("camtraps.csv")
```

Using package camtrapR, we can make a data frame of each camera's capture events by species.

```
captures <- detectionMaps(CTtable     = cameras,
                          recordTable  = cam1min,
                          Xcol         = "long",
                          Ycol         = "lat",
                          stationCol   = "Station",
                          speciesCol   = "Species",
                          printLabels  = FALSE,
                          richnessPlot = FALSE,
                          speciesPlots = FALSE,
                          addLegend    = FALSE)
```

For this analysis, lets see if latitude has any effect on the number of deer captured.

```
lat.mod <- glm(captures$Deer~captures$lat, family = "poisson")
summary(lat.mod)[12]
```

```
## $coefficients
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  160.312281 16.8713977  9.502015 2.058667e-21
## captures$lat  -3.608754  0.3865022 -9.336954 9.914501e-21
```

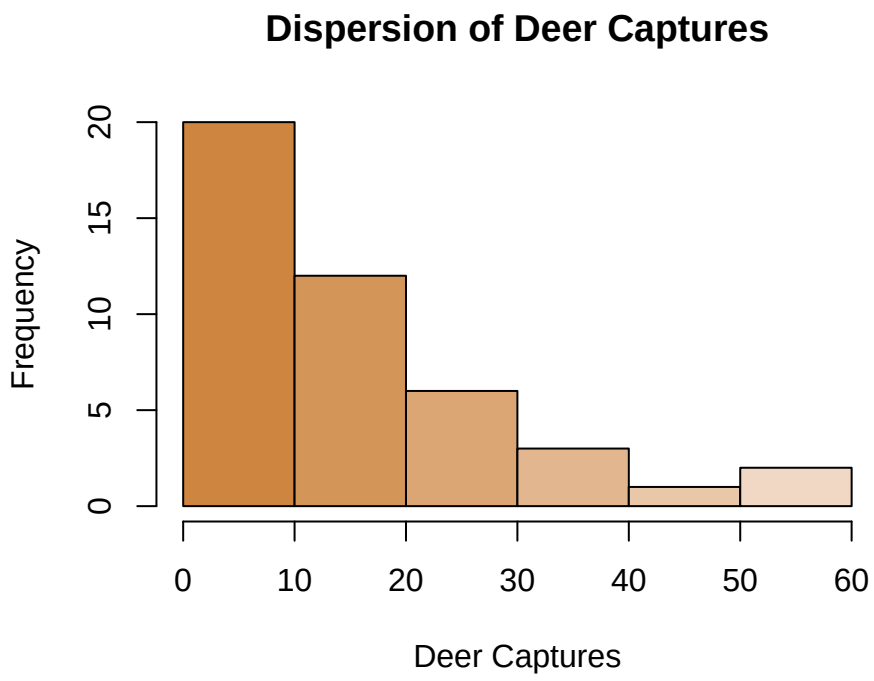
```
confint(lat.mod)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept)  127.470977 193.619843
## captures$lat  -4.371854  -2.856467
```

Alright, this shows that there is a good negative correlation with deer captures and latitude. Let's check dispersion though, and run a Negative Binomial glm.

```
tans <- colorRampPalette(c("tan3", "snow"))
hist(captures$Deer, col = tans(8), xlab = "Deer Captures", main = "Dispersion of Deer Captures")
```




```
lat.mod2 <- glm.nb(captures$Deer~captures$lat)
summary(lat.mod2)[11]
```

```
## $coefficients
##                Estimate Std. Error   z value    Pr(>|z|)
## (Intercept)  157.137308  56.138494  2.799101 0.005124513
## captures$lat  -3.536071   1.285007 -2.751792 0.005927014
```

```
summary(lat.mod2)[18]
```

```
## $SE.theta
## [1] 0.3103552
```

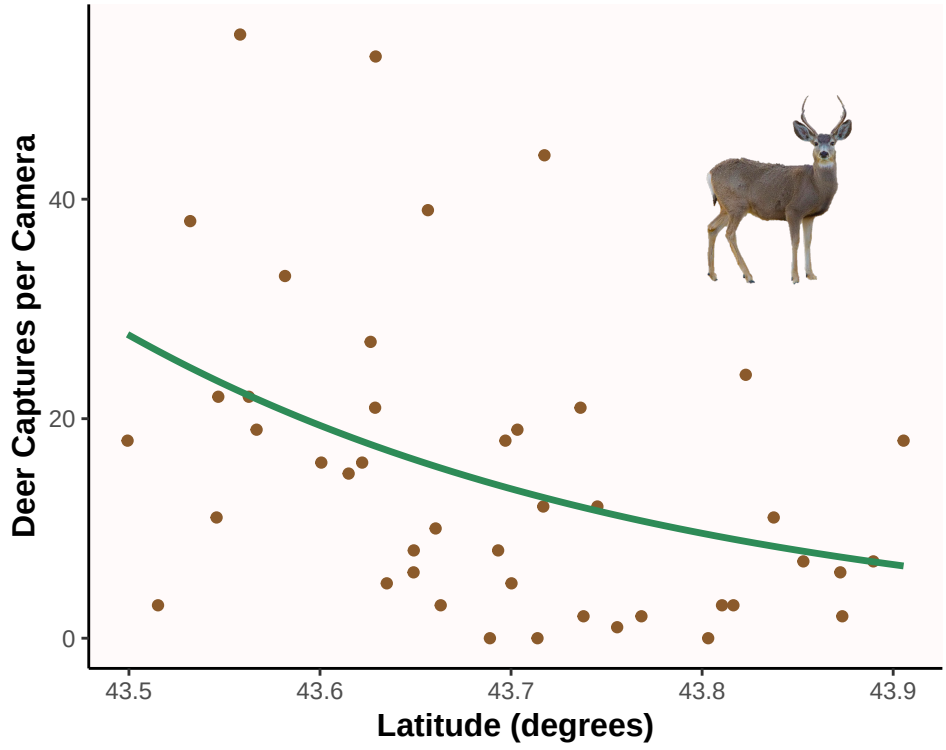
```
confint(lat.mod2)
```

```
##                   2.5 %     97.5 %
## (Intercept)  48.404388 263.815687
## captures$lat -5.977307  -1.046498
```

```
exp(157.137308 + (-3.536071 * max(captures$lat))) - exp(157.137308 + (-3.536071 * min(captures$lat)))
```

```
## [1] -21.09834
```

Well still a good confidence interval and slope significance, a microscopically less slope. The best way to interpret these parameters is to compute the effect of the northern and southern bounds of the study area, as the intercept (the equator) has no relevance and the effect per unit difference would be by 1 degree, which is more than the entire expanse. Thre result is about 21 less capture events for deer across our study area. This tracks with our study design as well- as we camera-trapped more heavily in the south than in the north.



Does this trend track to richness (i.e. did we perceive less species in the north where we trapped less heavily)?

```
rich.mod <- glm.nb(captures$n_species~captures$lat)
summary(rich.mod)[11]
```

```
## $coefficients
##                Estimate Std. Error   z value    Pr(>|z|)
## (Intercept)  85.284660 29.6751109  2.873946 0.004053786
## captures$lat -1.912083  0.6794641 -2.814104 0.004891337
```

```
confint(rich.mod)
```

```
##                   2.5 %      97.5 %
## (Intercept)  27.424631 144.1705503
## captures$lat -3.260502  -0.5873979
```

```
exp(85.284660 + (-1.912083 * max(captures$lat))) - exp(85.284660 + (-1.912083 * min(captures$lat)))
```

```
## [1] -4.456336
```

From this we can see that the trend does track, there are 4 less species found from south to north in the study area.