Homework 3

Clara Buchholtz

2/19/2019

Question 1: What is the effect of cut quality on diamond price?

```
#Fit glm model
diamonds<-read.csv("diamond.csv")
di.mod <- glm(diamonds$price~diamonds$cut, family="poisson")
#Extract coefficients
coef(di.mod)
```

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

```
## [1] 4358.758
```

```
exp(-0.1038367) #Relationship of good to fair
```

```
## [1] 0.9013725
```

```
0.9013725*4358.758 #Avg. price of good cut
```

```
## [1] 3928.865
```

```
exp(-0.2316292) #Relationship of ideal to fair
```

```
## [1] 0.7932402
```

```
0.7932402*4358.758 #Avg. price of ideal cut
```

```
## [1] 3457.542
```

```
exp(0.0504411) #relationship of premium to fair
```

```
## [1] 1.051735
```

```
1.051735*4358.758 #Avg. price of preumium
```

```
## [1] 4584.258
```

```
exp(-0.0904632) #relationship of very good to fair
```

```
## [1] 0.913508
```

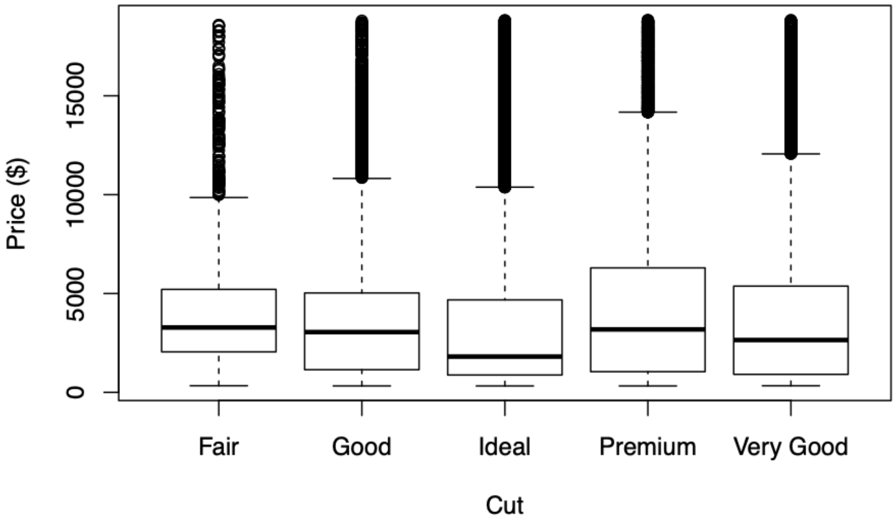
```
0.9135080*4358.758 #Avg. price of very good cut
```

```
## [1] 3981.76
```

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

```
##                             2.5 %      97.5 %
## (Intercept)            8.37920242  8.38068216
## diamonds$cutGood      -0.10470072 -0.10297248
## diamonds$cutIdeal     -0.23240302 -0.23085517
## diamonds$cutPremium    0.04966133  0.05122103
## diamonds$cutVery Good -0.09125511 -0.08967112
```

```
#Plot
boxplot(diamonds$price~diamonds$cut, xlab="Cut", ylab="Price ($)")
```



It looks like cut has a relationship with price, judging by the fact that none of the 95% confidence intervals cross zero. However, the small differences between mean prices, the number of outliers, and the strange relationship between higher/lower price and cut grade (e.g. ideal cuts are on average less expensive than fair and good cuts) makes me think that other factors may be more important than cut when it comes to price. A quick glance at the relationship between carat and price suggests that carat may be a thing to look at as another important variable.

Question 2: Does education have an impact on contraception use?

```
cuse<-read.csv("contraception.csv")
#create proportional response variable
response<- cbind(cuse$using, cuse$notUsing)
cuse.mod <-glm(response~cuse$education, family="binomial")
```

```
coef(cuse.mod) #parameter estimates
```

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

```
#relationship of low education to high education
plogis(-0.81020374)-plogis(-0.81020374+0.09248529)
```

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

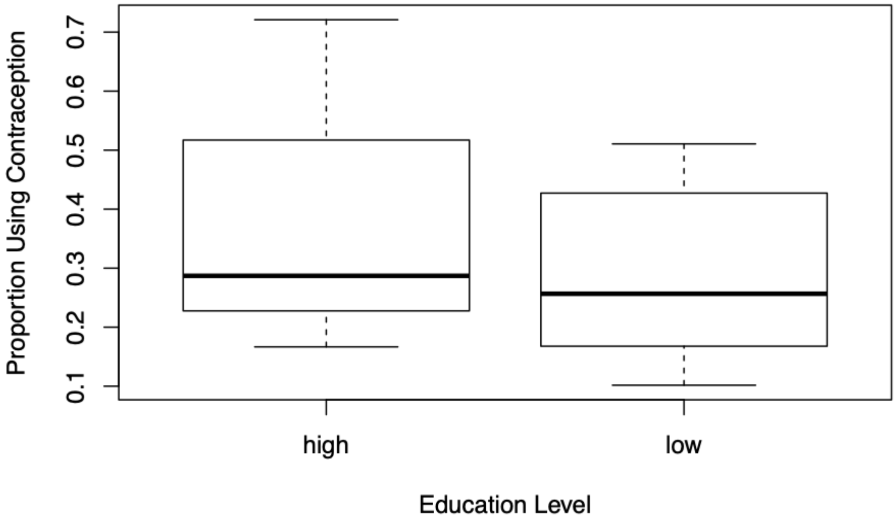
Those in the low education category are ~2% less likely to use contraception

```
#confidence interval
confint(cuse.mod)
```

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

The effect, however, does not appear to be significant, because the confidence interval crosses zero.

```
#plot
prop.using<-cuse$using/cuse$Total #proportion using contraception
boxplot(prop.using~cuse$education, xlab="Education Level", ylab="Proportion Using Contraception")
```



Because our confidence interval crosses zero, we fail to reject the null hypothesis. With this data, we are unable to state that education level plays a significant role in contraception use.

Question 3: Himmicanes

```
hurricanes<-read.csv("Hurricane Dataset.csv")
#Build model
hurs<-glm(hurricanes$alldeaths~hurricanes$Gender_MF, family="poisson")
#Parameter estimates
coef(hurs)
```

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

```
exp(3.1679220) #avg number of deaths for feminine names
```

```
## [1] 23.75806
```

```
exp(-0.5123354) #relationship of masculine deaths to feminine deaths
```

```
## [1] 0.5990948
```

```
23.75806*0.5990948 #Avg deaths for hurricanes with masculine names
```

```
## [1] 14.23333
```

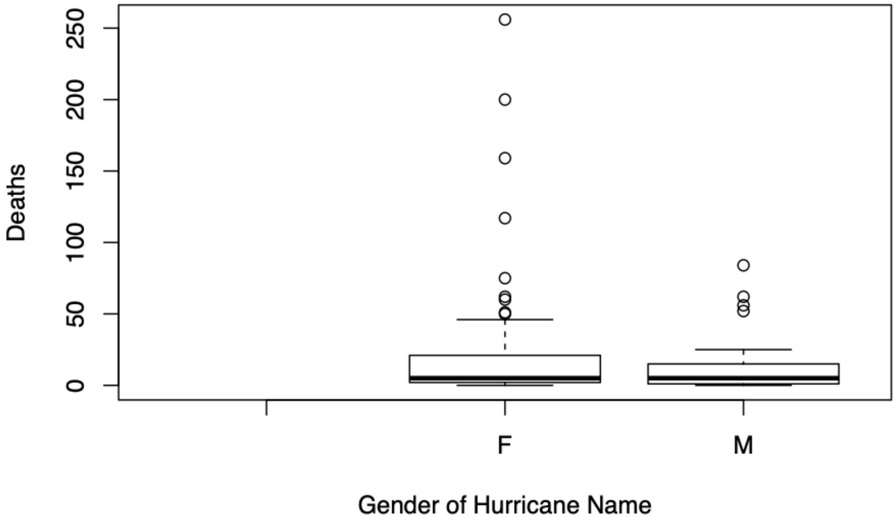
```
23.75806-14.23333 #effect of masculine name--number fewer of deaths
```

```
## [1] 9.52473
```

```
#confidence interval
confint(hurs)
```

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

```
#Plot
boxplot(hurricanes$alldeaths~hurricanes$Gender_MF,
        xlab="Gender of Hurricane Name", ylab="Deaths")
```



My results don't seem to differ too much from Jung et al's conclusions- the confidence interval doesn't cross zero. However, in their paper they seem to argue that a poisson distribution is not appropriate for this data, and they also seem to incorporate a lot of different variables into their analyses. The effect size of 9 more deaths for male named hurricanes doesn't seem especially large, and considering there are probably a lot of other factors much more important than the name of a hurricane that factor into why someone might die in the storm, I'd really be careful before I drew strong conclusions based on this outcome alone.

Question 4: Does the # of galls/leaf influence min, avg, or max gall volume?

```
gallsize <- read.csv("MinMaxGallSize.csv")
#Max size
m1<-glm(MaxSize~GallsPerLeaf,data=gallsize,family=Gamma(link="log")) #model
coef(m1) #parameter estimates
```

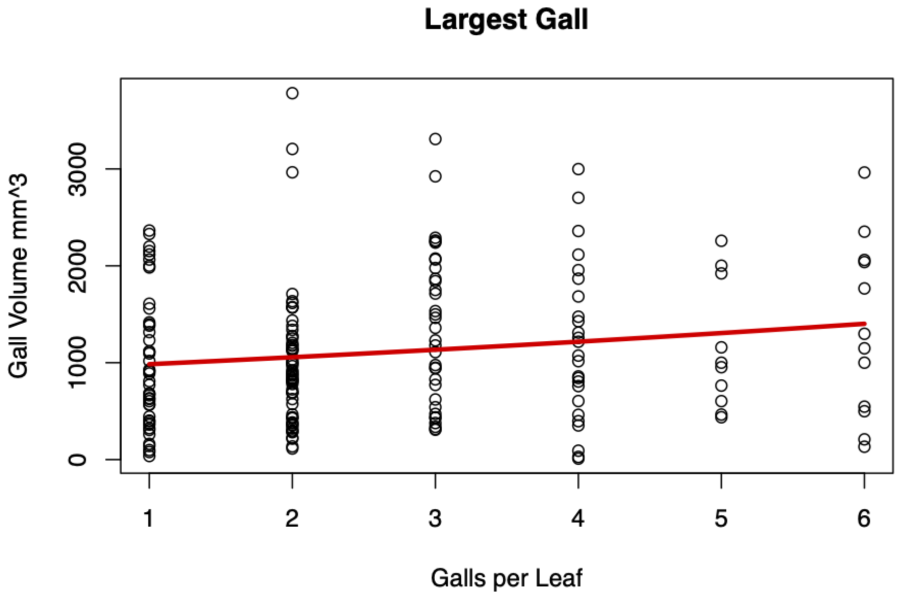
```
##  (Intercept) GallsPerLeaf
##   6.82121934   0.07067095
```

```
exp(0.07067095) #multiplicative effect of an increase in one gall/leaf
```

```
## [1] 1.073228
```

An increase of 1 gall per leaf corresponds with a ~7% increase in the average size of the largest gall on a leaf.

```
#plot
plot(gallsize$MaxSize~gallsize$GallsPerLeaf, xlab="Galls per Leaf",
     ylab="Gall Volume mm^3", main="Largest Gall")
curve(exp(6.82121934 +0.07067095*x), col="red3",lwd=3, add=T)
```



```
#confidence interval
confint(m1)
```

```
##                   2.5 %    97.5 %
## (Intercept)  6.62299307 7.0236954
## GallsPerLeaf 0.00414067 0.1389038
```

This appears to be a significant effect, because the confidence interval does not cross zero. With an increase in the number of galls per leaf, there is a modest increase in the average volume of the largest gall on a leaf. This is not what I initially expected to see when I first started looking at this data during my master's, because the general line of thinking about gall sizes on a leaf in the gall literature is that the more galls you have on a leaf, the more competition for plant space and resources there is, usually resulting in smaller galls.

```
#avg size
m2<-glm(AvgSize~GallsPerLeaf,data=gallsize,family=Gamma(link="log")) #model
coef(m2) #parameter estimates
```

```
##  (Intercept) GallsPerLeaf
##   6.94518441  -0.09629729
```

```
exp(-0.09629729) #effect of an increase in 1 gall/leaf on avg gall volume
```

```
## [1] 0.908194
```

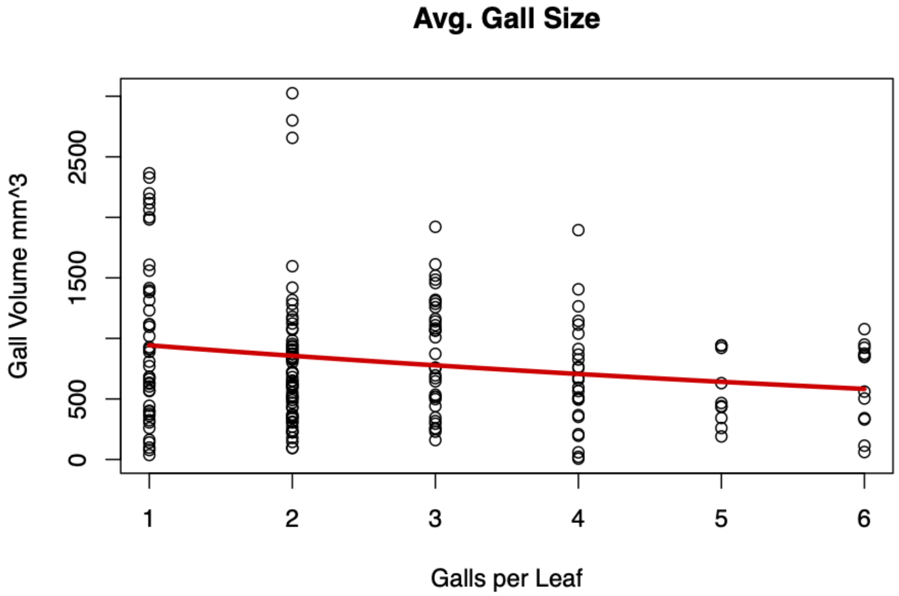
With each additional gall per leaf, we see a ~9% decrease in average gall volume

```
#confidence interval
confint(m2)
```

```
##                   2.5 %      97.5 %
## (Intercept)   6.7535172  7.14098486
## GallsPerLeaf -0.1602624 -0.03068792
```

This appears to be a significant effect, because the confidence interval does not cross zero.

```
#plot
plot(gallsize$AvgSize~gallsize$GallsPerLeaf, xlab="Galls per Leaf",
     ylab="Gall Volume mm^3", main="Avg. Gall Size")
curve(exp(6.94518441 +-0.09629729 *x),col="red3",lwd=3,add=T)
```



```
#Min size
m3<-glm(MinSize~GallsPerLeaf,data=gallsize,family=Gamma(link="log")) #model
coef(m3) #parameter estimates
```

```
##  (Intercept) GallsPerLeaf
##    7.2098593   -0.3769824
```

```
exp(-0.3769824) #effect of increase of 1 gall/leaf on avg. volume of smallest gall
```

```
## [1] 0.6859281
```

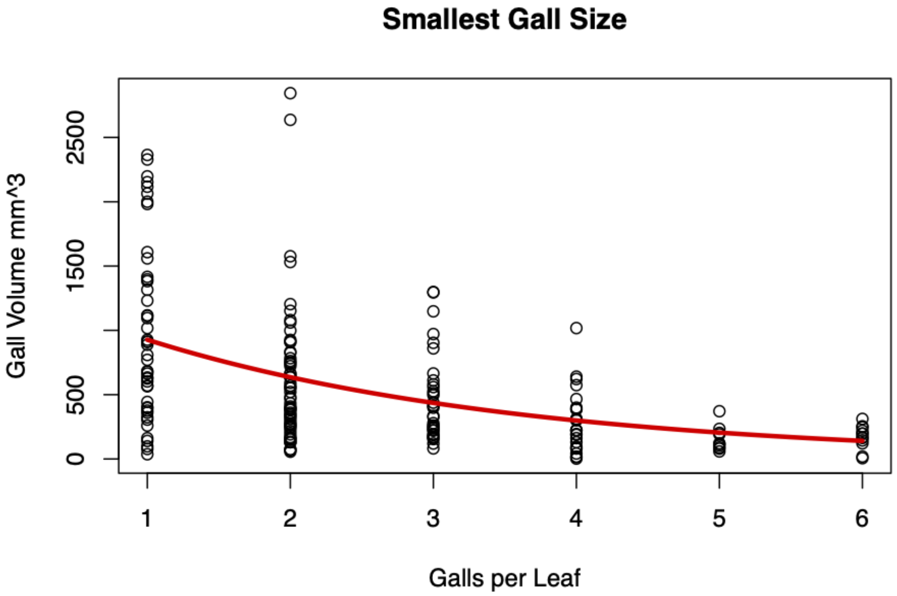
With each additional gall/leaf, there is a ~31% decrease in the average volume of the smallest gall on a leaf

```
#confidence interval
confint(m3)
```

```
##                   2.5 %     97.5 %
## (Intercept)   6.9897187  7.4355040
## GallsPerLeaf -0.4500356 -0.3016323
```

This appears to be a significant effect, becasue the confidence interval does not cross zero.

```
#plot
plot(gallsize$MinSize~gallsize$GallsPerLeaf, xlab="Galls per Leaf",
     ylab="Gall Volume mm^3", main="Smallest Gall Size")
curve(exp(7.209+-0.37698*x),col="red3",lwd=3,add=T)
```


Smallest Gall Size
Gall Volume mm^3
0
500
1500
2500
1
2
3
4
5
6
Galls per Leaf