HW\_3

Kristina Parker

2/13/2019

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

diamonds <- read.csv("Data/diamond.csv")  
head(diamonds[1:3, ])

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

dia\_mod <- glm(diamonds$price ~ diamonds$cut, family = "poisson")  
coef(dia\_mod)

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

exp(8.3799424)

## [1] 4358.758

exp(8.3799424 + -0.1038367) - exp(8.3799424) #Good effect from fair

## [1] -429.8935

exp(8.3799424 + -0.2316292) - exp(8.3799424) #Ideal effect from fair

## [1] -901.2159

exp(8.3799424 + 0.0504411) - exp(8.3799424) #Premium effect from fair

## [1] 225.5

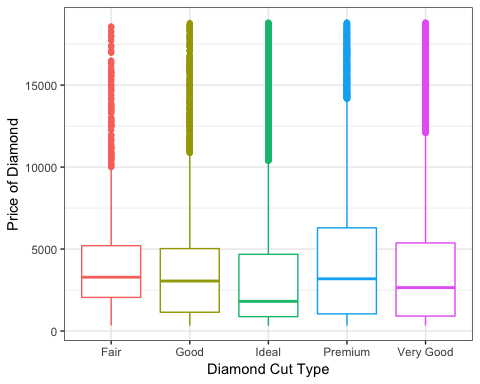
exp(8.3799424 + -0.0904632) - exp(8.3799424) #Very\_Good effect from fair

## [1] -376.9979

confint(dia\_mod) # None overlap zero = signficicant; 95% interval is very narrow = highly confident

## 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

ggplot(diamonds, aes(x = cut, y = price, colour = cut)) + geom\_boxplot(notch = FALSE) +   
 labs(x = "Diamond Cut Type", y = "Price of Diamond") + theme\_bw() +   
 theme(legend.position = "none")



The cut quality of diamonds effects the price of diamonds. The basline price of diamonds cut at fair quality is $4358.76. The price of diamonds cut at fair quality is higher than diamonds cut at good quality by $429.89, ideal quality by $901.22, and very good quality by $377.00. The price of diamonds cut at fair quality diamond is lower than a premium cut quality diamond by $225.50.

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

cont <- read.csv("Data/contraception.csv")  
head(cont[1:3, ])

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

# Ho: increased education does not effect contraception use  
# Ha: increased education promotes contraception use  
cont$prop\_use <- cont$using/cont$Total  
cont\_success <- cbind(cont$using, cont$notUsing)  
head(cont\_success)

## [,1] [,2]  
## [1,] 6 53  
## [2,] 4 10  
## [3,] 52 212  
## [4,] 10 50  
## [5,] 14 60  
## [6,] 10 19

cont\_mod <- glm(cont\_success ~ cont$education, family = "binomial")  
summary(cont\_mod) #no signifcance between groups

##   
## Call:  
## glm(formula = cont\_success ~ cont$education, family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.0868 -2.6566 -0.5529 2.1121 5.6674   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.81020 0.06871 -11.79 <2e-16 \*\*\*  
## cont$educationlow 0.09249 0.11011 0.84 0.401   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 165.77 on 15 degrees of freedom  
## Residual deviance: 165.07 on 14 degrees of freedom  
## AIC: 240.58  
##   
## Number of Fisher Scoring iterations: 4

coef(cont\_mod)

## (Intercept) cont$educationlow   
## -0.81020374 0.09248529

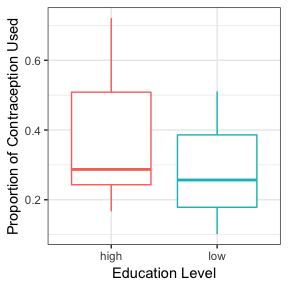
plogis(-0.81020374 + 0.09248529) - plogis(-0.81020374) # 2% effect size between high and low education

## [1] 0.02004851

confint(cont\_mod) # no significance because effect of education overlaps zero

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

ggplot(cont, aes(x = education, y = prop\_use, colour = education)) +   
 geom\_boxplot(notch = FALSE) + labs(x = "Education Level",   
 y = "Proportion of Contraception Used") + theme\_bw() + theme(legend.position = "none")



Women with a higher education is 2% more likely to use contraception. However, here is no significant effect that an increased education promotes contraception use.

# Question 3: Hurricanes vs Himmicanes

canes <- read.csv("Data/Hurricane Dataset.csv")  
head(canes)

## Year Name MasFem MinPressure\_before Minpressure\_Updated.2014  
## 1 1950 Easy 6.77778 958 960  
## 2 1950 King 1.38889 955 955  
## 3 1952 Able 3.83333 985 985  
## 4 1953 Barbara 9.83333 987 987  
## 5 1953 Florence 8.33333 985 985  
## 6 1954 Carol 8.11111 960 960  
## Gender\_MF Category alldeaths NDAM Elapsed.Yrs Source ZMasFem  
## 1 F 3 2 1590 63 MWR -0.00094  
## 2 M 3 4 5350 63 MWR -1.67076  
## 3 M 1 3 150 61 MWR -0.91331  
## 4 F 1 1 58 60 MWR 0.94587  
## 5 F 1 0 15 60 MWR 0.48108  
## 6 F 3 60 19321 59 MWR 0.41222  
## ZMinPressure\_A ZNDAM  
## 1 -0.35636 -0.43913  
## 2 -0.51125 -0.14843  
## 3 1.03765 -0.55047  
## 4 1.14091 -0.55758  
## 5 1.03765 -0.56090  
## 6 -0.25310 0.93174

canes <- na.omit(canes)  
canes\_mod <- glm(canes$alldeaths ~ canes$Gender\_MF, family = "poisson")  
summary(canes\_mod)

##   
## Call:  
## glm(formula = canes$alldeaths ~ canes$Gender\_MF, family = "poisson")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.8932 -5.3945 -3.7551 -0.3653 27.4348   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.16792 0.02606 121.584 <2e-16 \*\*\*  
## canes$Gender\_MFM -0.51234 0.05496 -9.322 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 4031.9 on 91 degrees of freedom  
## Residual deviance: 3937.1 on 90 degrees of freedom  
## AIC: 4266  
##   
## Number of Fisher Scoring iterations: 6

coef(canes\_mod)

## (Intercept) canes$Gender\_MFM   
## 3.1679220 -0.5123354

exp(3.167922 + -0.5123354) - exp(3.167922)

## [1] -9.524731

confint(canes\_mod)

## 2.5 % 97.5 %  
## (Intercept) 3.1164152 3.2185581  
## canes$Gender\_MFM -0.6211542 -0.4056501

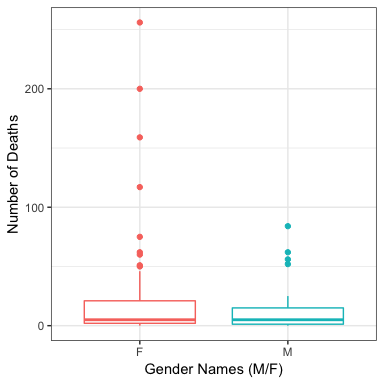
# negative binomial  
canes\_negbin <- glm.nb(canes$alldeaths ~ canes$Gender\_MF)  
coef(canes\_negbin)

## (Intercept) canes$Gender\_MFM   
## 3.1679220 -0.5123354

confint(canes\_negbin)

## 2.5 % 97.5 %  
## (Intercept) 2.816448 3.5640722  
## canes$Gender\_MFM -1.149166 0.1720959

ggplot(canes, aes(x = Gender\_MF, y = alldeaths, colour = Gender\_MF)) +   
 geom\_boxplot(notch = FALSE) + labs(x = "Gender Names (M/F)",   
 y = "Number of Deaths") + theme\_bw() + theme(legend.position = "none")



The hurricanes with a female names has 9.5 more deahts than male names. The poisson generalized linear model shows signficances that hurricanes have more deaths than himmicanes. The results of the poisson glm states that the gender names of hurricanes has significances on the total number of deaths, this is similar to Jung et al.’s conclusion. However, the poisson glm is not the best glm to use for this data set due to the overdispersed toward the female names. The negative binomial results indicate there is no significant differences (95% CI overlap zero) between female and male hurricane names and total deaths.

# Question 4:

## 'data.frame': 48 obs. of 3 variables:  
## $ Study\_Area: chr "NCA" "NCA" "NCA" "NCA" ...  
## $ Array\_ID : Factor w/ 48 levels "FD1","FD2","FD3",..: 12 13 14 15 16 17 18 19 20 21 ...  
## $ Richness : num 1 2 1 3 2 1 1 2 3 2 ...

hist(rich$Richness)



rich\_mod <- glm(rich$Richness ~ rich$Study\_Area, family = "poisson")  
summary(rich\_mod)

##   
## Call:  
## glm(formula = rich$Richness ~ rich$Study\_Area, family = "poisson")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3214 -0.4435 -0.1260 0.7414 1.7308   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.4055 0.2357 1.720 0.0854 .  
## rich$Study\_AreaOCTC 0.5857 0.2566 2.282 0.0225 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 50.145 on 47 degrees of freedom  
## Residual deviance: 44.215 on 46 degrees of freedom  
## AIC: 167.92  
##   
## Number of Fisher Scoring iterations: 5

coef(rich\_mod)

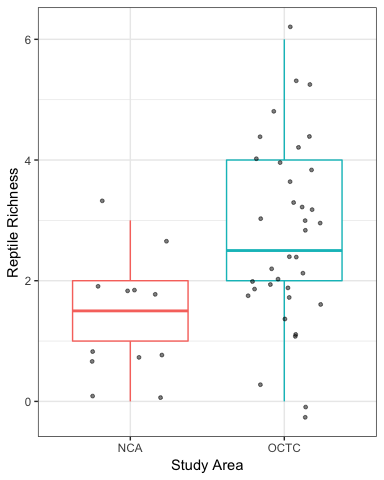
## (Intercept) rich$Study\_AreaOCTC   
## 0.4054651 0.5857269

exp(0.4054651 + 0.5857269) - exp(0.4054651)

## [1] 1.194444

confint(rich\_mod)

## 2.5 % 97.5 %  
## (Intercept) -0.09503459 0.8343866  
## rich$Study\_AreaOCTC 0.10879789 1.1209811



The reptile richness is significantly different between two study areas. There is a 119% difference in reptile richness between study areas. The 95% CI does not overlap zero; however, the interval is wide for the effect size suggesting low confindence in the significance. The reptile richness data set is does not fit normal or binomial regression; however, the poisson regression is not a good fit either due to the overdispersion of the data. A negative binomial to account for the overdispersion would be a better fit, but the theta is too large causing errors.