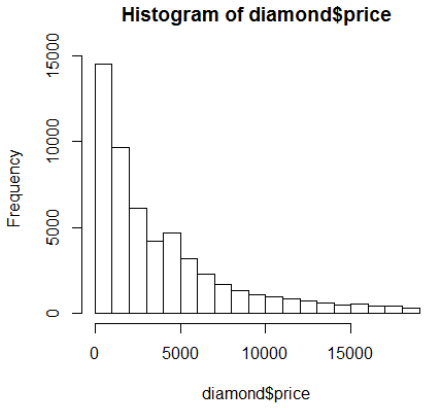
PATRICIA KAYE T. DUMANDAN **13 FEB 2019**
MS RAPTOR BIOLOGY **EEB 622 HW 3**

Q1: DIAMOND DATA

A. DATA EXPLORATION

```
> hist(diamond$price)  #appropriate for poisson
```




```
> str(diamond)
'data.frame':   53940 obs. of  3 variables:
 $ price: int  326 326 327 334 335 336 336 337 337 338 ...
 $ cut  : Factor w/ 5 levels "Fair","Good",..: 3 4 2 4 2 5 5 5 1 5 ...
 $ carat: num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
```

B. MODEL

```
> dmond.glm=glm(diamond$price~diamond$cut, family = "poisson")
> summary(dmond.glm)

Call:
glm(formula = diamond$price ~ diamond$cut, family = "poisson")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-82.42  -54.53  -25.95   20.69  181.67

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           8.3799424  0.0003775 22199.1   <2e-16 ***
diamond$cutGood      -0.1038367  0.0004409  -235.5   <2e-16 ***
diamond$cutIdeal     -0.2316292  0.0003949  -586.6   <2e-16 ***
diamond$cutPremium    0.0504411  0.0003979   126.8   <2e-16 ***
diamond$cutVery Good -0.0904632  0.0004041  -223.9   <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: 183127012  on 53939  degrees of freedom
Residual deviance: 180344852  on 53935  degrees of freedom
```

```
AIC: 180864021
Number of Fisher Scoring iterations: 5
> coef(dmond.glm)
          (Intercept)      diamond$cutGood     diamond$cutIdeal
            8.3799424           -0.1038367           -0.2316292
  diamond$cutPremium diamond$cutVery Good
            0.0504411           -0.0904632
> Anova(dmond.glm, type="II")
Analysis of Deviance Table (Type II tests)
Response: diamond$price
             LR Chisq Df Pr(>Chisq)
diamond$cut   2782159  4  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mae(dmond.glm)
[1] 2990.464
> rmse(dmond.glm)
[1] 3963.664
> confint(dmond.glm)
                             2.5 %      97.5 %
(Intercept)           8.37920242  8.38068216
diamond$cutGood      -0.10470072 -0.10297248
diamond$cutIdeal     -0.23240302 -0.23085517
diamond$cutPremium    0.04966133  0.05122103
diamond$cutVery Good -0.09125511 -0.08967112
```

C. GRAPH AND INTERPRETATION

```
> #intercept: fair
> exp(8.3799424)
[1] 4358.758
> #good
> exp(8.3799424+-0.1038367)-exp(8.3799424)
[1] -429.8935
> #ideal
> exp(8.3799424+-0.2316292)-exp(8.3799424)
[1] -901.2159
> #premium
> exp(8.3799424+0.0504411)-exp(8.3799424)
[1] 225.5
> #very good
> exp(8.3799424+ -0.0904632)-exp(8.3799424)
[1] -376.9979
> ggplot(diamond, aes(x = cut, y = price)) +
+   geom_boxplot(notch = FALSE) +
+   theme_classic() +
+   theme()
```

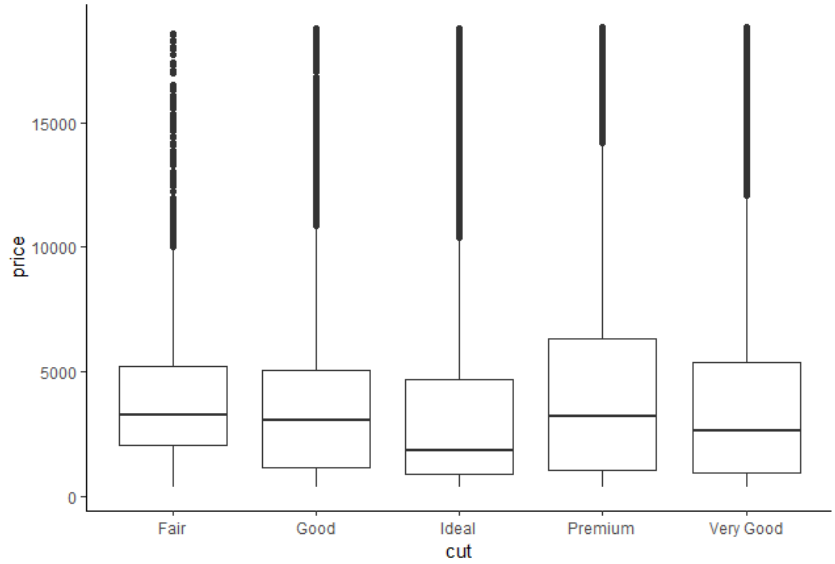
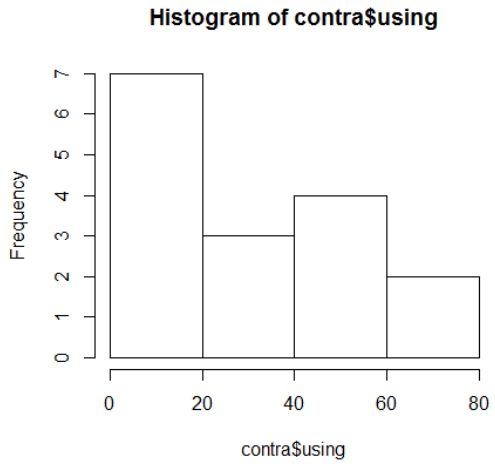


Fig.1. Effect of cut quality on price of diamond. A diamond with a "fair" cut costs $4358.76 which in this case, is the base cut. Among the different diamond cuts, the least expensive is an "ideal" cut which costs around $3457.54 and the most expensive is a "premium" cut which costs about $4584.26. A "good" diamond cut costs $3928.87 and "very good" cut diamonds cost $3981.78. Model seems appropriate with none of the values for the slope and intercept overlapping 0 (MAE=2990, *P<2.2e-16*).

Q2. CONTRACEPTION USE DATA

A. <u>DATA EXPLORATION</u>

```
> hist(contra$using)
```




```
> str(contra)
'data.frame':  16 obs. of  5 variables:
 $ age      : Factor w/ 4 levels "<25","25-29",..: 1 1 1 1 2 2 2 2 3 3 ...
 $ education: Factor w/ 2 levels "high","low": 2 2 1 1 2 2 1 1 2 2 ...
 $ notUsing : int  53 10 212 50 60 19 155 65 112 77 ...
 $ using    : int  6 4 52 10 14 10 54 27 33 80 ...
 $ Total    : int  59 14 264 60 74 29 209 92 145 157 ...
```

B.MODEL

```
> #make proportional data
> prop_using=cbind(contra$using, contra$Total-contra$using)
> contra.glm=glm(prop_using~contra$education, family="binomial")
> summary(contra.glm)
Call:
glm(formula = prop_using ~ contra$education, family = "binomial")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0868  -2.6566  -0.5529   2.1121   5.6674
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.81020    0.06871  -11.79   <2e-16 ***
contra$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
> Anova(contra.glm, type="II")
Analysis of Deviance Table (Type II tests)
Response: prop_using
                 LR Chisq Df Pr(>Chisq)
contra$education   0.7039  1     0.4015
> coef(contra.glm)
        (Intercept) contra$educationlow
        -0.81020374          0.09248529
> confint(contra.glm)
                         2.5 %     97.5 %
(Intercept)         -0.9460962 -0.6766394
contra$educationlow -0.1239481  0.3078275
> mae(contra.glm)
[1] 0.1403238
> rmse(contra.glm)
[1] 0.1725428
```

C.GRAPH AND INTERPRETATION

```
> #INTERCEPT
> plogis(-0.81)
[1] 0.3078905
> #SLOPE
> plogis(-0.81+0.09)-plogis(-0.81)
[1] 0.01950249
> ggplot(contra, aes(x = education, y = using)) +
+   geom_boxplot(notch = FALSE) +
+   theme_classic() +
+   theme()
```

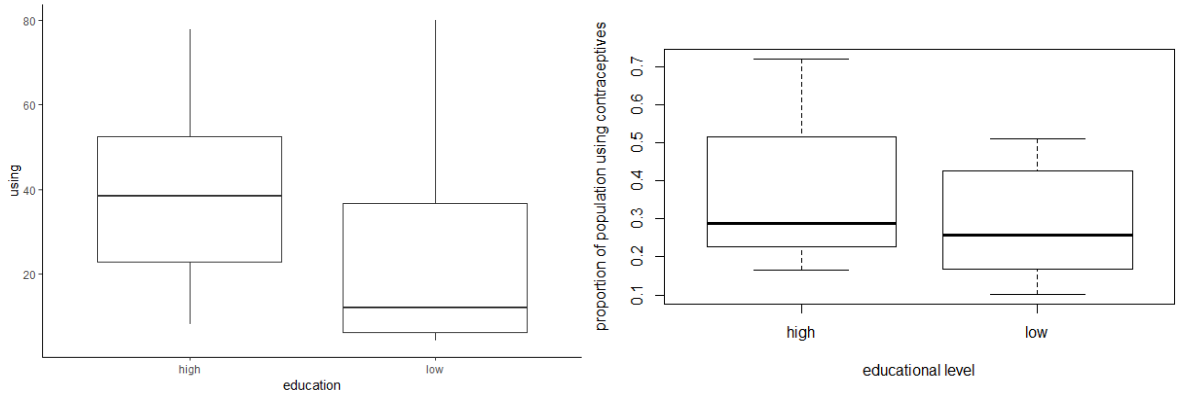
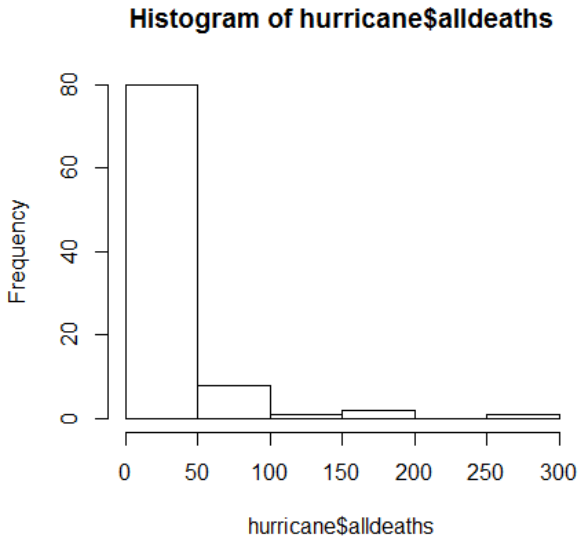


Fig.2. Association between educational level and likelihood of using contraception in Fiji women. Fiji women are 30% likely to use contraception and there is a 1.9% increase in likelihood of these women to use contraception with increased educational level. However, the model seems to be not a good fit for the data as values for the confidence intervals of intercept and slope overlap 0 (P=0.4015, MAE=0.14).

Q3: HURRICANES AND HIMMICANES DATA

A.DATA EXPLORATION

```
> hist(hurricane$alldeaths)   #most values between 0 and 50
```




```
> str(hurricane)
'data.frame':  98 obs. of  14 variables:
 $ Year                    : Factor w/ 55 levels "","," $ Name
: Factor w/ 84 levels "","Able","Agnes",..: 39 78
 $ MasFem                  : num  6.78 1.39 3.83 9.83 8.33 ...
 $ MinPressure_before      : int  958 955 985 987 985 960 954 938 962 987 ...
 $ Minpressure_Updated.2014: int  960 955 985 987 985 960 954 938 962 987 ...
 $ Gender_MF               : Factor w/ 3 levels "","F","M": 2 3 3 2 2 2 2 2 2
 $ Category                : int  3 3 1 1 1 3 3 4 3 1 ...
```

```
 $ alldeaths               : int  2 4 3 1 0 60 20 20 0 200 ...
 $ NDAM                    : int  1590 5350 150 58 15 19321 3230 24260 2030
$ Elapsed.Yrs             : int  63 63 61 60 60 59 59 59 58 58 ...
 $ Source                  : Factor w/ 4 levels "","http://www.nhc.noaa.gov/
 $ ZMasFem                 : num  -0.00094 -1.67076 -0.91331 0.94587 0.48108
 $ ZMinPressure_A          : num  -0.356 -0.511 1.038 1.141 1.038 ...
 $ ZNDAM                   : num  -0.439 -0.148 -0.55 -0.558 -0.561 ...
```

B.MODEL

```
> hur.glm=glm(hurricane$alldeaths~hurricane$Gender_MF, family="poisson")
> summary(hur.glm)

Call:
glm(formula = hurricane$alldeaths ~ hurricane$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 ***
hurricane$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(hur.glm)
       (Intercept) hurricane$Gender_MFM
         3.1679220         -0.5123354
> Anova(hur.glm)
Analysis of Deviance Table (Type II tests)
Response: hurricane$alldeaths
                  LR Chisq Df Pr(>Chisq)
noblank$Gender_MF   94.851  1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mae(hur.glm)
[1] 23.65316
> rmse(hur.glm)
[1] 40.4355
> confint(hur.glm)
                        2.5 %     97.5 %
(Intercept)         3.1164152  3.2185581
hurricane$Gender_MFM -0.6211542 -0.4056501
```

C.GRAPH AND INTERPRETATION

```
> #intercept
> exp(3.1679220)
[1] 23.75806
> #slope
> exp(3.1679220)-exp(3.1679220+-0.5123354)
[1] 9.524731
```

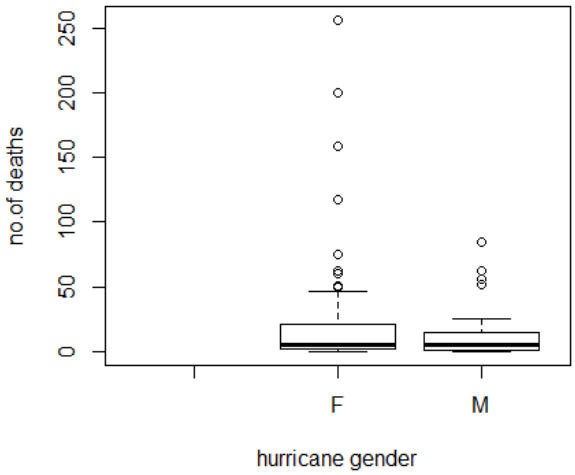
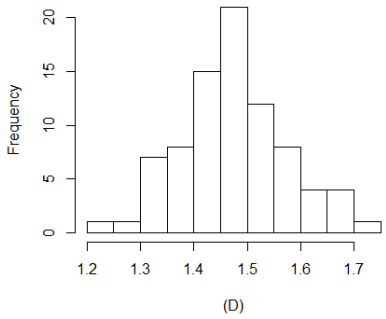


Fig.3. Number of deaths per type of hurricane. Hurricanes named after females are expected to result to 10 more deaths than hurricanes named after males ("himmicanes") (MAE=23.65, df=91, *P<2.2e-16*). Given the MAE, this model seems appropriate as it only suggests that approximately 30% of the individual values is different from the true mean. These results are similar to the results published by Jung et al. However, a more appropriate distribution model (i.e. negative binomial) would have been more appropriate as overdispersion is observed on the histogram of the response variable(alldeaths).

Q4. SPECIES RICHNESS DATA

A. DATA EXPLORATION

```
> hist(t2$Margalef.s.index..S...1....LN..n...excel.formula., xlab="(D)" )
```



```
> #use annual countsraw file
> #remove non-Aug to Dec counts
> t2=read.csv(file.choose(), h=T)
> str(t2)
'data.frame':  82 obs. of  19 variables:
 $ YR                                                : int  1934 1935 1936  $
HRS                                               : num  325 523 593 650 561
 $ obs.effort.categ                                 : int  1 2 2 2 2 2 2 3 2
 $ n..number.of.species..taxa.                      : int  16 15 16 15 14 15
$ Margalef.s.index..S...1....LN..n...excel.formula.: num  1.67 1.46 1.55 1.45
 $ Margalef.s.index..R.formula.log.                 : num  1.67 1.46 1.55 1.
$ H..sum.Pi.logPi..                                : num  1.01 1.76 1.46 1.48
$ J..H..ln.S..                                     : num  0.363 0.651 0.528 0
 $ J..R.comp.                                      : num  0.363 0.651 0.528
$ Persecution                                      : Factor w/ 2 levels "no",
$ DDT                                              : Factor w/ 2 levels "no",
"yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Deforestation                                   : Factor w/ 2 levels "no"
,"yes": 1 1 1 1 1 1 1 1 1 1 ...
```

B.MODEL

1.per time period

```
> rich.glm=glm(t2$Margalef.s.index..S...1....LN..n...excel.formula.~t2$DDT, f
amily="gaussian")
> summary(rich.glm)
Call:
glm(formula = t2$Margalef.s.index..S...1....LN..n...excel.formula. ~
    t2$DDT, family = "gaussian")
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.275824  -0.041573   0.007647   0.044012   0.233556
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.43820    0.01580  91.025  < 2e-16 ***
t2$DDTyes    0.06273    0.02087   3.006  0.00354 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.008737434)
    Null deviance: 0.77794  on 81  degrees of freedom
Residual deviance: 0.69899  on 80  degrees of freedom
AIC: -152.01
Number of Fisher Scoring iterations: 2
> coef(rich.glm)
(Intercept)   t2$DDTyes
 1.43820172  0.06273072
> Anova(rich.glm)
Analysis of Deviance Table (Type II tests)

Response: t2$Margalef.s.index..S...1....LN..n...excel.formula.
       LR Chisq Df Pr(>Chisq)
t2$DDT    9.035  1   0.002649 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> confint(rich.glm)
                 2.5 %    97.5 %
(Intercept) 1.40723423 1.4691692
t2$DDTyes   0.02182689 0.1036346
```

2.per year

```
> rich.glm1=glm(t2$Margalef.s.index..S...1....LN..n...excel.formula.~t2$YR, f
amily="gaussian"(link=identity))
> summary(rich.glm1)

Call:
glm(formula = t2$Margalef.s.index..S...1....LN..n...excel.formula. ~
    t2$YR, family = gaussian(link = identity))
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.250472  -0.040653  -0.002203   0.039826   0.271735
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.9210895  0.8051265  -2.386   0.0194 *
t2$YR        0.0017172  0.0004072   4.217 6.46e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 0.007955507)
    Null deviance: 0.77794  on 81  degrees of freedom
Residual deviance: 0.63644  on 80  degrees of freedom
AIC: -159.7
Number of Fisher Scoring iterations: 2
> coef(rich.glm1)
 (Intercept)        t2$YR
-1.921089482  0.001717225
> Anova(rich.glm1)
Analysis of Deviance Table (Type II tests)
Response: t2$Margalef.s.index..S...1....LN..n...excel.formula.
      LR Chisq Df Pr(>Chisq)
t2$YR   17.786  1  2.472e-05 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> confint(rich.glm1)
                    2.5 %       97.5 %
(Intercept) -3.4991084858 -0.343070477
t2$YR        0.0009191648  0.002515285
```

C.GRAPH AND INTERPRETATION

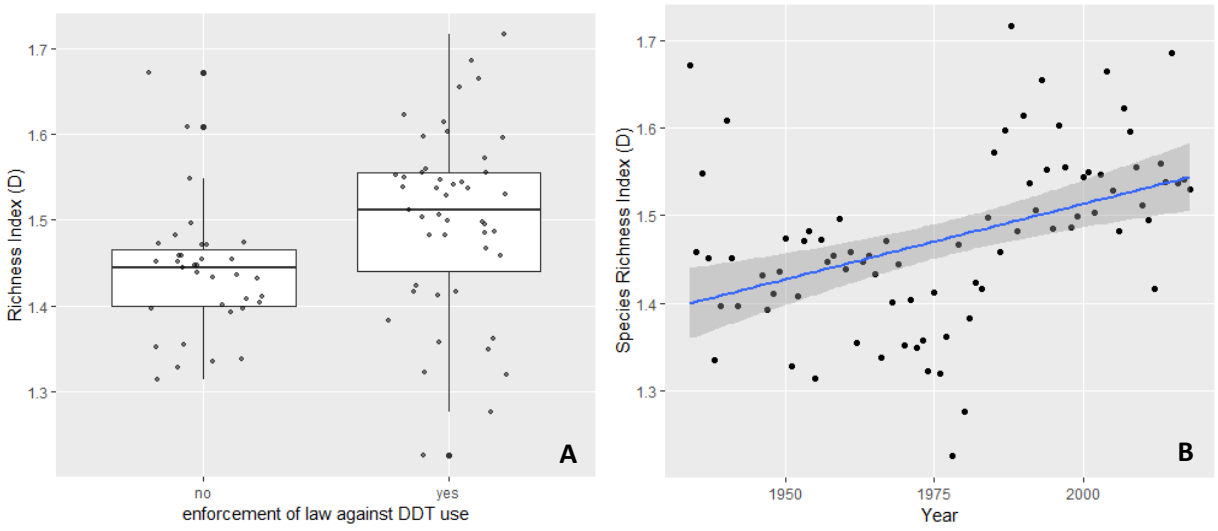


Fig.4. Effect of implementation of ban against DDT use (in 1972) on species richness of migrating raptors recorded at Hawk Mountain Sanctuary,PA. Regardless of presence or absence of law, the richness index is at 1.44. Species richness increased by 0.06 with the enforcement of law against DDT use (*P=0.0026; Fig4a)*. Effect is significant given that the confidence interval values do not overlap 0. Consistently, species richness index increases by 0.002 units annually (*P*=2.472e-05; Fig.4b). Similarly, confidence interval values do not overlap 0. Interaction between year of implementation(time period) and year should probably be looked at instead of interpreting each as separate predictors.