Stats Exam 2016: Question 4

Sergii Gladchuk

December 27, 2016

Background

It is well known that lemmings are suicidal. According to locals, they jump off cliffs into deep water when they are too stressed. Together with his colleague prof W. Disney, prof. Plupp designed an experiment to test this "common knowledge". They randomly collected 20 lemmings from the field. In the lab, they took blood samples from each lemming to measure its stress level (variable stress). They also determined the sex of all individuals (sex = 0 means female, sex = 1 means male). Next, each lemming was put on a high platform (12m) above a barrel of water. It was noted whether it jumped (jump=1) or not (jump=0). You find the data in the file "lemmingjump" (.sav and .csv).

Since dependent variable *jump* is binary, Logistic regression analyses should be applied to answer questions about tendency to jump based on all independent variables.

Reading and checking of the data.

```
jump.data <- read.csv('~/Documents/courses/stats/Exam/lemmingjump.csv', sep=';')
str(jump.data)
```

```
## 'data.frame':    20 obs. of  3 variables:
##  $ stress: num  3.32 2.15 2.68 2.32 0.3 1.25 4.34 2.93 4 1.55 ...
##  $ sex   : int  0 0 1 0 0 0 1 1 1 0 ...
##  $ jump  : int  1 0 1 0 0 0 1 1 0 1 ...
```

There is no need to convert integer 1/0 variable *sex* to factor yet, because first sub-question is analysis of frequencies.

Sub-question a)

Ignoring the stress level, is there a difference between the sexes in the tendency to jump?

To test hypothesis that *jump* is dependent on *sex*, next steps should be performed:

1. Frequency data to be converted into contingency table
2. Calculation of expected counts to define test, which should be used.
3. All or some of χ^2-test, G-test and Fishers' exact tests should be performed.

```
counts <- table(jump.data$sex,jump.data$jump)
rownames(counts) <- c('female','male')
colnames(counts) <- c('no jump', 'jump')
counts
```

```
##
##          no jump jump
##   female       8    2
##   male         2    8
```

```r
#expected counts
hab.exp <- matrix(NA,2,2)
for (i in 1:length(counts[1,])) {
  for (j in 1:length(counts[,1])){
    hab.exp[i,j] <- sum(counts[i,])/sum(counts) *
      (sum(counts[,j])/sum(counts)) * sum(counts);
  }
}
hab.exp
```

```
##      [,1] [,2]
## [1,]    5    5
## [2,]    5    5
```

```r
chisq.test(counts)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  counts
## X-squared = 5, df = 1, p-value = 0.02535
```

```r
library(DescTools)
GTest(counts)
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  counts
## G = 7.7098, X-squared df = 1, p-value = 0.005492
```

```r
fisher.test(counts)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  counts
## p-value = 0.02301
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    1.309537 239.395560
## sample estimates:
## odds ratio
##   13.25038
```

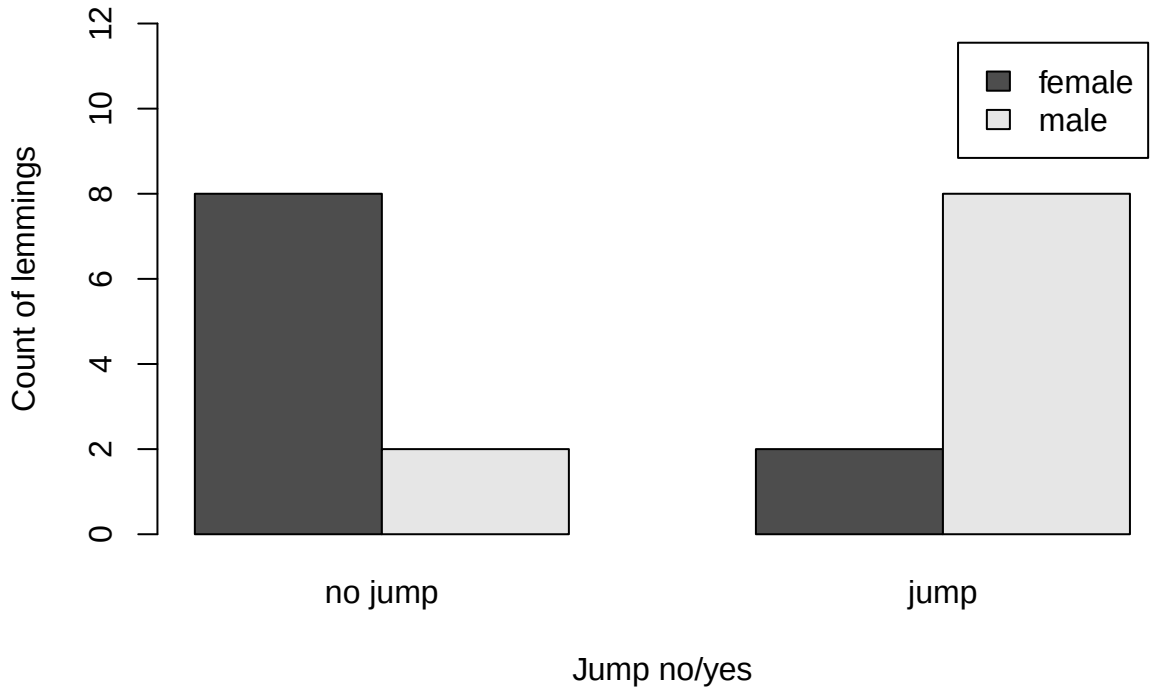
Expected values are all 5 so all three test are valid to take into account and all three tests have p-value less than 0.05 so null-hypothesis that there is no tendency to jump between sexes should be rejected.

Sub-question b)

Illustrate your result in a)

Bar-plot to visualize proportion.

```
barplot(counts, xlab='Jump no/yes', ylab='Count of lemmings',
        beside=TRUE, legend=rownames(counts), ylim=c(0,12))
```



The bar-plot supports the fact that proportion is not the same (fully reversed one actually) and that male tendency to jump is much higher.

Sub-question c)

Taking all data into account, can you find an effect of the stress level on the jumping propensity?

Now is the time for logistic regression analysis. The family for generalized linear model should be binomial because dependent variable is binomial. Also no interaction between stress and sex should be included into the model.

```
jump.data$sex <- factor(jump.data$sex)
str(jump.data)
```

```
## 'data.frame':    20 obs. of  3 variables:
##  $ stress: num  3.32 2.15 2.68 2.32 0.3 1.25 4.34 2.93 4 1.55 ...
##  $ sex   : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 2 2 1 ...
##  $ jump  : int  1 0 1 0 0 0 1 1 0 1 ...
```

```r
full.model <- glm(jump~stress+sex, data=jump.data, family = binomial)
summary(full.model)
```

```
##
## Call:
## glm(formula = jump ~ stress + sex, family = binomial, data = jump.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.94815  -0.62309  -0.00267   0.71590   1.80010
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5637     1.3676  -1.875   0.0609 .
## stress        0.7510     0.6167   1.218   0.2233
## sex1          1.2950     1.5280   0.848   0.3967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 18.397  on 17  degrees of freedom
## AIC: 24.397
##
## Number of Fisher Scoring iterations: 4
```

```r
library(glmulti)
```

```
## Loading required package: rJava
```

```r
model.sel <- glmulti(full.model, level=2, crit='aic')
```

```
## Initialization...
## TASK: Exhaustive screening of candidate set.
## Fitting...
## Completed.
```

```r
summary(model.sel)
```

```
## $name
## [1] "glmulti.analysis"
##
## $method
## [1] "h"
##
## $fitting
## [1] "glm"
##
## $crit
## [1] "aic"
```

```
##
## $level
## [1] 2
##
## $marginality
## [1] FALSE
##
## $confsetsize
## [1] 100
##
## $bestic
## [1] 25.73783
##
## $icvalues
## [1] 25.73783 26.10591 26.40550 26.93155 27.57398 33.03165
##
## $bestmodel
## [1] "jump ~ 1 + stress"
##
## $modelweights
## [1] 0.283770998 0.236069680 0.203228442 0.156226656 0.113305882 0.007398342
##
## $includeobjects
## [1] TRUE
```

```
weightable(model.sel)
```

```
##                                   model      aic     weights
## 1                    jump ~ 1 + stress 25.73783 0.283770998
## 2                       jump ~ 1 + sex 26.10591 0.236069680
## 3               jump ~ 1 + sex + stress 26.40550 0.203228442
## 4 jump ~ 1 + sex + stress + sex:stress 26.93155 0.156226656
## 5       jump ~ 1 + stress + sex:stress 27.57398 0.113305882
## 6                              jump ~ 1 33.03165 0.007398342
```

```
best.model <- glm(jump~stress, data=jump.data, family = binomial)
summary(best.model)
```

```
##
## Call:
## glm(formula = jump ~ stress, family = binomial, data = jump.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8872  -0.6935  -0.0042   0.8013   1.6647
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8022     1.3520  -2.073   0.0382 *
## stress        1.0996     0.4792   2.295   0.0217 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 19.134  on 18  degrees of freedom
## AIC: 23.134
##
## Number of Fisher Scoring iterations: 4
```

```
AIC(full.model)
```

```
## [1] 24.3973
```

```
AIC(best.model)
```

```
## [1] 23.13377
```

Even in small full model (with two parameters) non of the parameters was significant - model selection procedure advised to remove least significant parameter (*sex*) in order to get better fit. This is confirmed by slight decrease of AIC value.

Sub-question d)

Taking the stress level into account, is there now a difference between the sexes in their tendency to jump?

Since *sex* factor was excluded from best fitted model the *stress* is only significant effect which plays role in jump decision. That leaves *sex* as other dependent variable on *stress*. This can be confirmed by creating another model.

```
sex.model <- glm(sex~stress, data=jump.data, family=binomial)
summary(sex.model)
```

```
##
## Call:
## glm(formula = sex ~ stress, family = binomial, data = jump.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.94636  -0.35097  -0.00165   0.32697   1.68306
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -5.771      2.701  -2.137   0.0326 *
## stress         2.260      1.017   2.221   0.0263 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 11.327  on 18  degrees of freedom
## AIC: 15.327
##
## Number of Fisher Scoring iterations: 6
```

The p-value for *stress* parameter in this model for sex prediction has almost the same significance as our previous model for jump prediction. In simple words if lemming is very stressful there is high probability that this lemming is male (based on sex prediction model) and that he will jump (based on jump prediction model). That is why *sex* variable does not add any value to the jump prediction model.

Sub-question e)

Make a suitable illustration of your results in c)

Graph of 'actual values diferentiated by sex' and 'values predicted by model' so as 'actual model curve'.

```
jump.data$pred <- predict(best.model, jump.data, type='response')

#add some jitter to the points
jump.data$jump.jit <- jitter (jump.data$jump, amount = 0.02)
females <- subset(jump.data, jump.data$sex == '0')
males <- subset(jump.data, jump.data$sex == '1')

coefs <- coef(best.model)
coefs
```

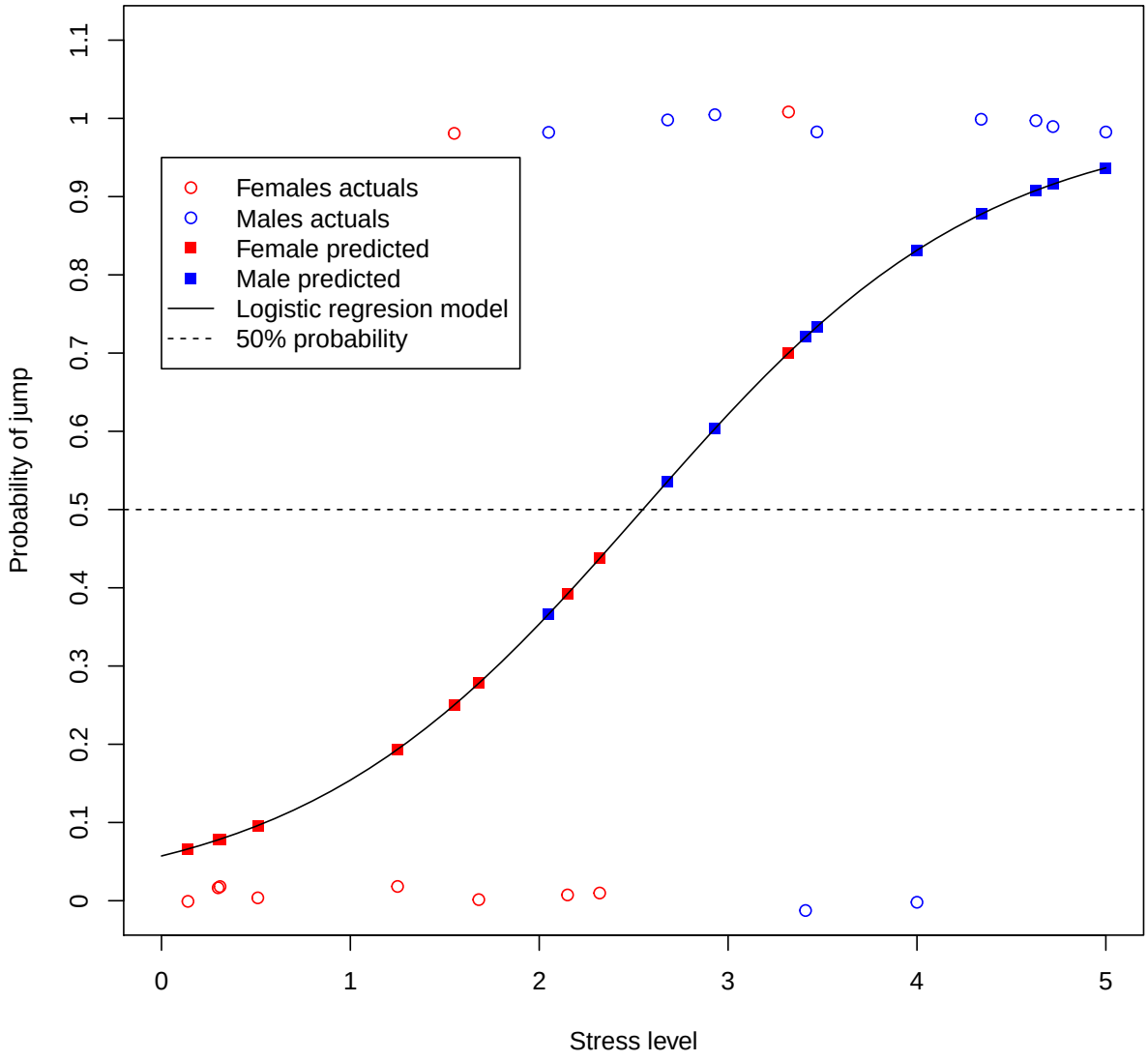
```
## (Intercept)      stress
##   -2.802188    1.099595
```

```
x <- seq(0,5, by=0.1)
y <- coefs[1] + coefs[2]*x
pred <- exp(y)/(1+exp(y))

plot(females$jump.jit~females$stress, xlab='Stress level',
     ylab='Probability of jump', xlim=c(0,5), ylim=c(0,1.1),
     col='red', yaxt="n")
axis(2,at=seq(0,1.1,by=0.1), labels = seq(0,1.1,by=0.1))
points(males$jump.jit~males$stress, col='blue')

points(females$pred~females$stress, col='red', pch=15)
points(males$pred~males$stress, col='blue', pch=15)

lines(x,pred)
abline(0.5,0, lty=2)
legend(0,0.95,legend = c('Females actuals','Males actuals',
                         'Female predicted','Male predicted',
                         'Logistic regresion model',
                         '50% probability'),
       col=c('red','blue','red','blue','black','black'),
       pch=c(21,21,15,15,NA,NA),
       lty=c(0,0,0,0,1,2))
```



Conclusion:

Based on the graph only 4 actual points (2 males which did not jumped and 2 lemmings of each sex which actually jumped) are not predicted correctly out of 20, which is not so bad. Also it is seen that no better logistic regression line can be build taking into account sexes, so decision to exclude *sex* parameter from final model was correct.