EDDA_assign3_Group33

Wenbo Sun, Meifang Li, Yuhao Qian

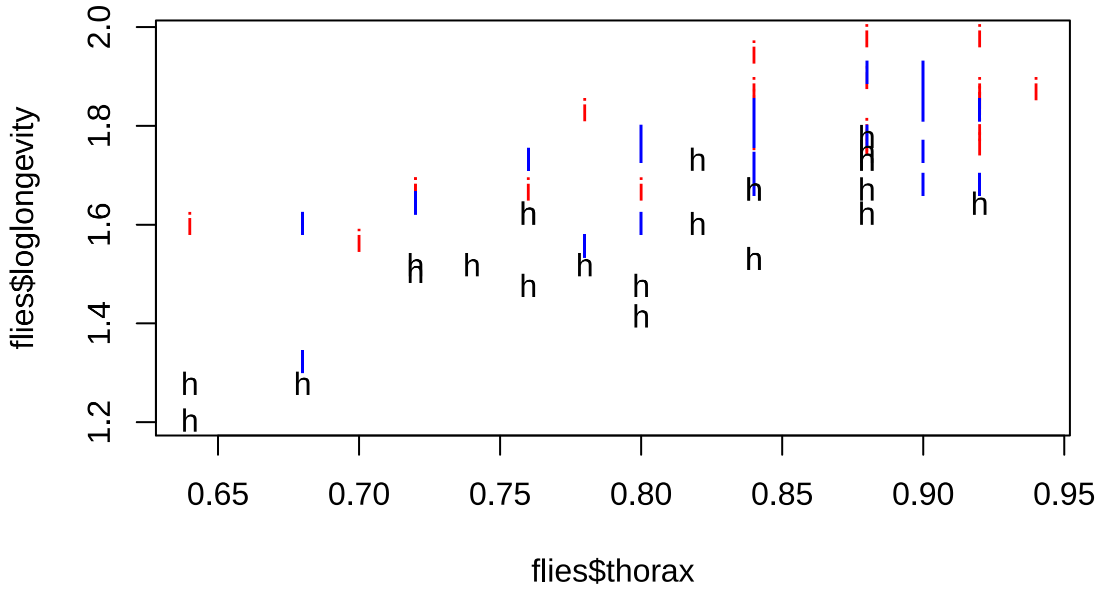
3/18/2021

Exercise 1

a) The informative plot of the data is as follows: "i" represents the isolated group, "l" and "h" represent the low and high group respectively. According to the summary of one-way anova, all p-values are smaller than 0.05, the sexual activity indeed influences longevity, longevity of low sexual activity is larger than that of high, longevity of isolated sexual activity is larger than that of low because isolated(0.22463)>low(0.17272)>high(0). And the longevities for the three conditions are in the following table. We can tell from the table that the longevities will decrease as the sexual activities of flies increase.

Category	isolated	low	high
longevity	61.52	54.59	36.68

```
flies=read.table("fruitflies.txt",header = TRUE)
loglongevity=log10(flies$longevity)
flies$loglongevity=loglongevity
flies$Colour="black"
flies$Colour[flies$activity=="isolated"]="red"
flies$Colour[flies$activity=="low"]="blue"
plot(flies$loglongevity~flies$thorax,col=flies$Colour,pch=as.character(flies$activity))
```



```
flies$activity=as.factor(flies$activity)
one_anov=lm(loglongevity~activity,data=flies)
summary(one_anov)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity, data = flies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41489 -0.05792  0.01108  0.09073  0.21377
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.56438    0.02669  58.621  < 2e-16 ***
## activityisolated  0.22463    0.03774   5.952 8.82e-08 ***
## activitylow       0.17272    0.03774   4.577 1.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1334 on 72 degrees of freedom
## Multiple R-squared:  0.3504, Adjusted R-squared:  0.3324
## F-statistic: 19.42 on 2 and 72 DF,  p-value: 1.798e-07
```

```
y1=10^(1.56438);y1
```

```
## [1] 36.67583
```

```r
y2=10^(1.56438+0.22463);y2
```

```
## [1] 61.5191
```

```r
y3=10^(1.56438+0.17272);y3
```

```
## [1] 54.58835
```

b) According to the estimated in the summary of two-way anova, the sexual activity decreases the longevity of the flies because isolated(0.17805)>low(0.12408)>high(0). The estimated longevities for the three groups, for a fly with average thorax length are in the following table.

Category	isolated	low	high
longevity	59.45	52.51	39.46

```r
flies$activity=as.factor(flies$activity)
flies$thorax=as.numeric(flies$thorax)
two_anov=lm(loglongevity~thorax+activity,data=flies)
summary(two_anov)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity, data = flies)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.210991 -0.069993  0.004518  0.065568  0.155204
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.52938    0.10799   4.902 5.79e-06 ***
## thorax            1.29376    0.13318   9.715 1.14e-14 ***
## activityisolated  0.17805    0.02536   7.021 1.07e-09 ***
## activitylow       0.12408    0.02540   4.885 6.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08804 on 71 degrees of freedom
## Multiple R-squared:  0.7211, Adjusted R-squared:  0.7093
## F-statistic:  61.2 on 3 and 71 DF,  p-value: < 2.2e-16
```

```r
mu_tho=mean(flies$thorax);mu_tho
```

```
## [1] 0.8245333
```

```r
y1=10^(0.52938+1.29376*mu_tho);y1
```

```
## [1] 39.45738
```

```
y2=10^(0.52938+1.29376*mu_tho+0.17805);y2
```

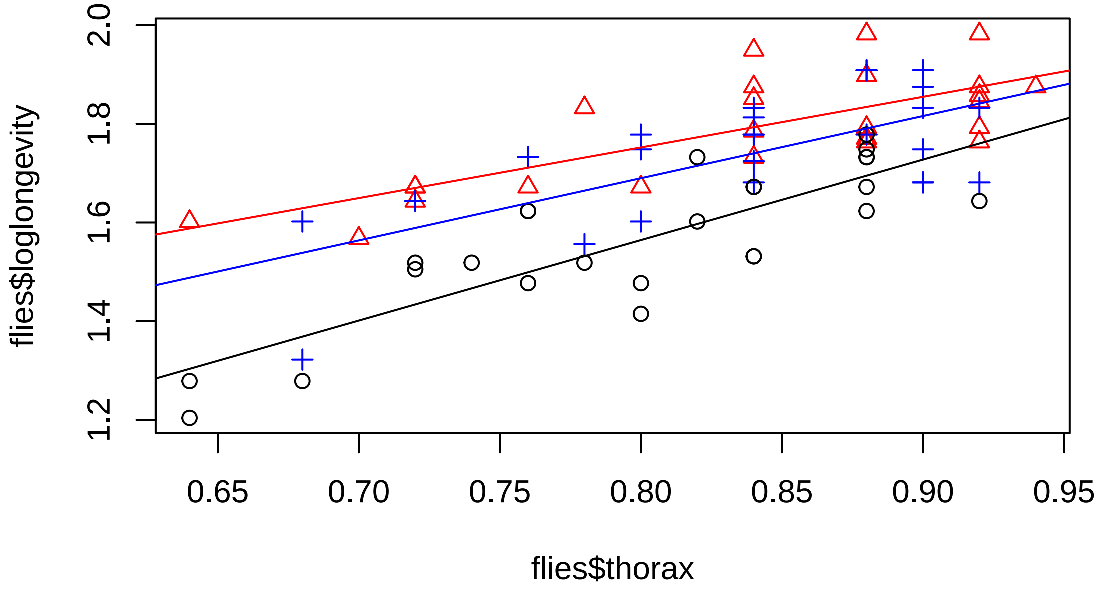
```
## [1] 59.45361
```

```
y3=10^(0.52938+1.29376*mu_tho+0.12408);y3
```

```
## [1] 52.50592
```

c) The red spots and line represent the isolated group, the blue spots and line represent the low group and the black spots and line represent the high group. As shown in the figure, all three groups show very similar slope in the line and the longevity is larger when the throax length is larger. Hence the longevity increases as the thorax length increases. And we perform a test testing for the interaction between factor activity and predictor thorax. The p-value testing for $H_0 : \beta_1 = \beta_2 = \beta_3$ is 0.1536 >0.05, so H_0 cannot be rejected, there is no interaction between activity and thorax and the dependence is similar under all three conditions of sexual activity.

```
plot(flies$loglongevity~flies$thorax,col=flies$Colour,pch=unclass(flies$activity))
flies_iso=subset(flies,activity=="isolated")
flies_low=subset(flies,activity=="low")
flies_high=subset(flies,activity=="high")
abline(lm(flies_iso$loglongevity~flies_iso$thorax,data=flies_iso),col="red")
abline(lm(flies_low$loglongevity~flies_low$thorax,data=flies_low),col="blue")
abline(lm(flies_high$loglongevity~flies_high$thorax,data=flies_high),col="black")
```



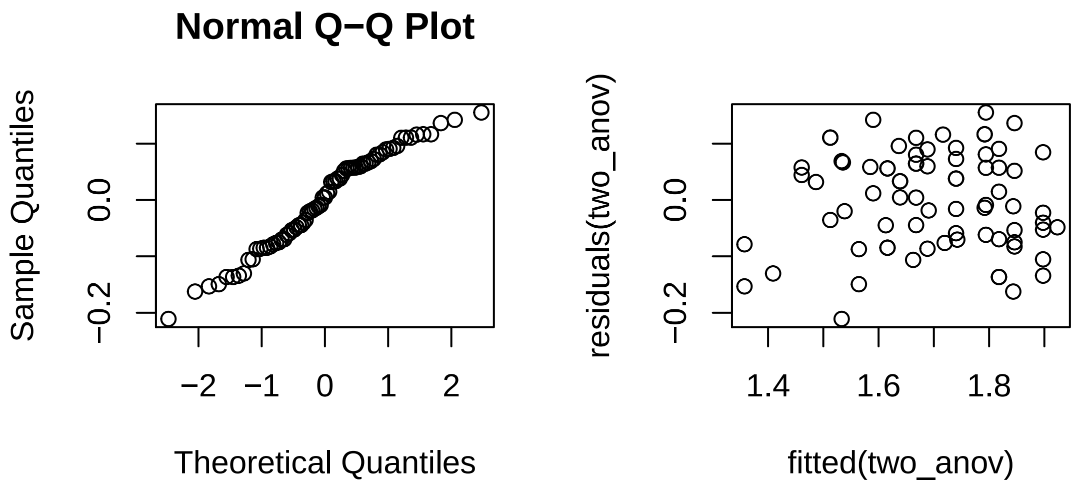
```
flies3=lm(loglongevity~activity*thorax,data=flies);anova(flies3)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## activity           2 0.69154 0.34577 45.7687 2.228e-13 ***
## thorax             1 0.73155 0.73155 96.8327 9.020e-15 ***
## activity:thorax    2 0.02909 0.01454  1.9251    0.1536
## Residuals         69 0.52128 0.00755
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) We prefer the analyses with thorax length because the p-value of thorax is $1.14 * 10^{-14}$<0.05 in the analyses, which means thorax has significant effect on the longevity. So it is wrong to analyze the longevity without considering the factor thorax.

e) The QQ-plot look like normal and the residuals don't change systematically with the fitted values.

```
par(mfrow=c(1,2))
qqnorm(residuals(two_anov));plot(fitted(two_anov),residuals(two_anov))
```

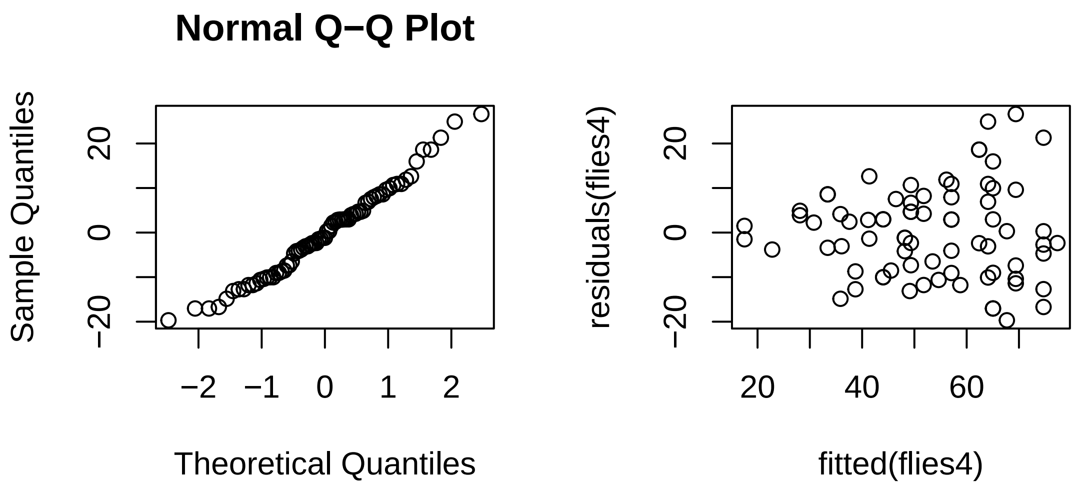


f) The p-values for both factors are smaller than 0.05 so they both have significant effects on the longevity. And the QQ-plot look like normal but the residuals in the fitted plot show some extreme values in the middle field compared to the residuals in the fitted plot of using logarithm. Thus it is wise to use the logarithm because it can reduce the gap between the extreme values, make the data more smooth.

```
flies4=lm(longevity~thorax+activity,data=flies)
drop1(flies4,test="F")
```

```
## Single term deletions
##
## Model:
## longevity ~ thorax + activity
##           Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>                  7673 355.10
## thorax     1    7686.8 15360 405.15  71.127 2.624e-12 ***
## activity  2    4966.7 12640 388.53  22.979 2.016e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
qqnorm(residuals(flies4));plot(fitted(flies4),residuals(flies4))
```

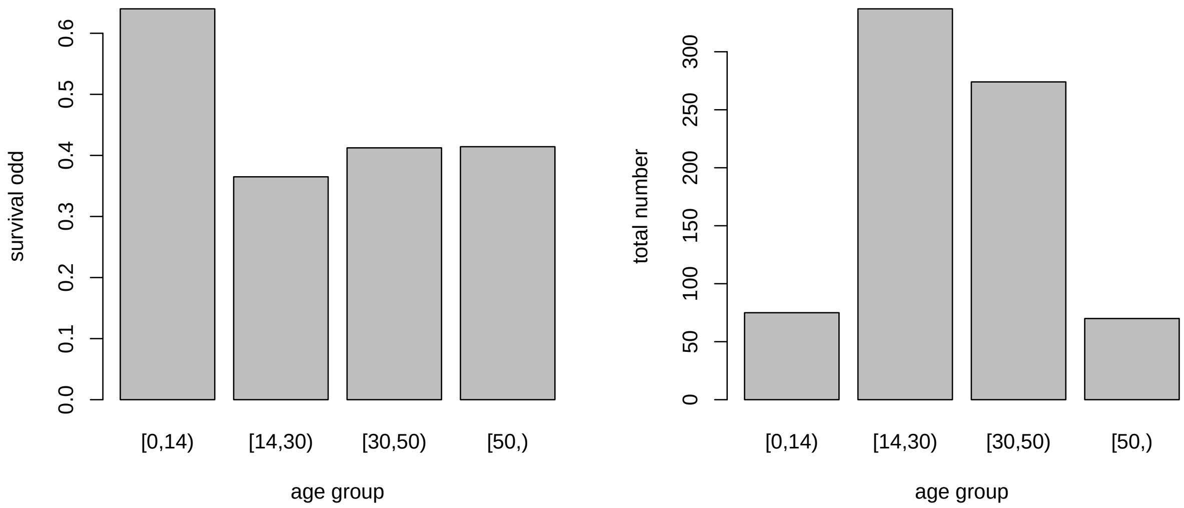


Exercise 2

a) We first examine the raw data and find more than half of age data are missing, denoted by NA. Hence, we omit NA data for generating a quantitative summary of age distribution. We set 4 buckets for ages as (0,14), [14,30), [30,50) and [50,], each of which represents younger, teenager, middle-aged and elder. Then we count the survivor number and survival rates of each group as follows:

```
titan$PClass=as.factor(titan$PClass)
titan$Sex=as.factor(titan$Sex)

tot_age=xtabs(~age_group,data=na.omit(titan))
par(mfrow=c(1,2))
barplot(xtabs(Survived~age_group,data=na.omit(titan))/tot_age,xlab="age group", ylab="survival odd")
barplot(tot_age,xlab="age group",ylab="total number")
```



We find that youngers have a higher survival rate than other age groups. Besides, we also generate a 2-dimensional survivor counting table regarding PClass and Sex and compute each class-sex combination's survival rate. Here we use the entire dataset because missing ages does not affect the statistics.

```
sx_tot=xtabs(~PClass+Sex,data=titan)
sx_svv=xtabs(Survived~PClass+Sex,data=titan)
sx_svv;sx_svv/sx_tot
```

```
##       Sex
## PClass female male
##    1st    134   59
##    2nd     94   25
##    3rd     78   58
```

```
##       Sex
## PClass    female      male
##    1st 0.9370629 0.3296089
##    2nd 0.8785047 0.1445087
##    3rd 0.3768116 0.1176471
```

Generally speaking, females have higher survival rates than males in all classes; both male and female passengers from 1st class have a higher probability of surviving; females' survival rate from 3rd class is much lower than the other two higher classes.

b) We fit a logistic model for association between survival status and variable PClass, Age and Sex. Taking PClass and Age as factors and omitting rows missing age data, we have a model as follows:

```
lr1=glm(Survived~PClass+Age+Sex,family="binomial",data=na.omit(titan))
summary(lr1)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Age + Sex, family = "binomial",
##     data = na.omit(titan))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7226  -0.7065  -0.3917   0.6495   2.5289
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.759662   0.397567   9.457  < 2e-16 ***
## PClass2nd   -1.291962   0.260076  -4.968 6.78e-07 ***
## PClass3rd   -2.521419   0.276657  -9.114  < 2e-16 ***
## Age         -0.039177   0.007616  -5.144 2.69e-07 ***
## Sexmale     -2.631357   0.201505 -13.058  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  695.14  on 751  degrees of freedom
## AIC: 705.14
##
## Number of Fisher Scoring iterations: 5
```

```
drop1(lr1,test="Chisq")
```

```
## Single term deletions
##
## Model:
## Survived ~ PClass + Age + Sex
##        Df Deviance    AIC     LRT  Pr(>Chi)
## <none>      695.14 705.14
## PClass  2   795.59 801.59 100.445 < 2.2e-16 ***
## Age     1   723.59 731.59  28.454 9.595e-08 ***
## Sex     1   909.92 917.92 214.776 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Without considering the interaction between predictors, we apply drop1 to examine our model and find all three variables are statistically significant because p-values are all less than 0.05. Regarding the coefficients of variables, we find that PClass 2nd, PClass 3rd and Sex Male have negative coefficients, which means all these three factors lower the odd. Also, as the age increases, the odd decreases. Specifically, a deviance Δ in a single variable changes the odds multiplied by e^{Δ}. For example, PClass changes from 1st to 2nd, the odds $\frac{P(Y=1)}{P(Y=0)}$ multiplies $e^{-1.292}$; Age, as a predictor add 1 years, the odds multiplies $e^{-0.04}$. A multiplier less than 1 indicates the $P(Y=1)$ is getting smaller. These statistics conform to our observation in question a).

c) We further investigate the interactions of Age-Sex and Age-PClass by changing the formula in the generic linear model.

```
lr2=glm(Survived~Age*Sex+PClass,family="binomial")
lr3=glm(Survived~Age*PClass+Sex,family="binomial")
drop1(lr3,test="Chisq")
```

```
## Single term deletions
##
## Model:
## Survived ~ Age * PClass + Sex
##            Df Deviance    AIC    LRT Pr(>Chi)
## <none>          689.64 703.64
## Sex         1   908.75 920.75 219.11  < 2e-16 ***
## Age:PClass  2   695.14 705.14   5.50  0.06393 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(lr2,test="Chisq")
```

```
## Single term deletions
##
## Model:
## Survived ~ Age * Sex + PClass
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       667.08 679.08
## PClass   2   770.56 778.56 103.479 < 2.2e-16 ***
## Age:Sex  1   695.14 705.14  28.064 1.174e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to results from drop1, the p-value of Age-Sex interaction is less than 0.05, indicating the statistical significance, whereas Age-PClass doesn't show such significance.

Based on this conclusion, we predict survival probabilities of new data using Age*PClass+Sex because of the significance. The result is demonstrated as follows.

```
newdata=data.frame(Sex=c("male","female"),PClass=c("1st","2nd","3rd"),Age=rep(53,times=6))
pd=predict(lr2,newdata,type="response")
newdata$Predict=pd
newdata
```

```
##      Sex PClass Age    Predict
## 1   male    1st  53 0.16404203
## 2 female    2nd  53 0.79293376
## 3   male    3rd  53 0.01362089
## 4 female    1st  53 0.94715339
## 5   male    2nd  53 0.04023998
## 6 female    3rd  53 0.55776269
```

d) Data preprocessing is the first step to building a predictive model.Then, we need to decide feasible variables. For example, we can use drop1 function to choose the variable according to significance value. In this question, we need to omit the na value because we will use Age as a variable to train the model. Subsequently, organize the data into dataframe to feed the glm function. After obtaining the model, we need to check the intra-sample accuracy using 'fitted' function. We can take the response value larger than 0.5 as 1 (Survived) to predict the binary survival status. As for quality justification, we can use F-score. The F-score is a measurement of binary classification's accuracy.

The first step of computing the F-score is collecting and splitting sample data into a test set and a training set. We use the training set to fit a model then predict the survival status by the test set.

Since we have the actual status of the test set, we then can compare the results of prediction and raw data, constructing a confusion matrix. A confusion matrix consists of four elements, True-Positive (TP), True-Negative(TN), False-Positive(FP) and False-Negative(FN), which are counted from the comparison between prediction and actual value. The F-score is computed by

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

The highest possible value of F-score is 1.0, the higher the better.

e) Since the contingency table does not use the age data, we can use the entire data set. To examine the effect of PClass and Sex on survival status, we count the survivor numbers among survived passengers. Then apply the contingency table test on this matrix. Since we focus on the effect of two factors, we can combine each factor with survival status and apply the test two times to examine whether each category comes from the same distribution given survival status.

We first test PClass and Survived. Computing the row sum and column sum to examine whether the chisq.test is sufficient for this contingency table.

```
svv_cnt=xtabs(~Survived+PClass,data=titan)
svv_cnt
```

```
##         PClass
## Survived 1st 2nd 3rd
##        0 129 161 564
##        1 193 119 136
```

```
checkE=svv_cnt
rowsum=apply(svv_cnt,1,sum);a=as.vector(rowsum/sum(rowsum))
colsum=apply(svv_cnt,2,sum);b=as.vector(colsum/sum(colsum))
checkE[,1]<-checkE[,2]*a;checkE[,2]<-checkE[,2]*a;
checkE[,2]<-checkE[,2]*a
checkE[1,]<-checkE[1,]*b;checkE[2,]<-checkE[2,]*b;
checkE
```

```
##         PClass
## Survived        1st        2nd        3rd
##        0  26.116661  14.895898 303.225806
##        1  10.126489   3.029898  73.118280
```

More than 80% of expected counts are larger than 5, the Chi-square test is reliable.

```
z=chisq.test(svv_cnt)
z;residuals(z)
```

```
##
##  Pearson's Chi-squared test
##
## data:  svv_cnt
## X-squared = 170.71, df = 2, p-value < 2.2e-16
```

```
##         PClass
## Survived       1st       2nd       3rd
##        0 -5.656440 -1.671780  4.893711
##        1  7.809677  2.308177 -6.756600
```

The results show that the survival distribution of each passenger class has a significant difference because the p=value is less than 0.05. We can see the 1st and 2nd class have higher survival number than 3rd class.

Then we test the gender factor. The contingency table of gender and survived is 2x2. We can use Fisher's test to compute exact p-value.

```
svp_cnt=xtabs(~Survived+Sex,data=titan)
svp_cnt
```

```
##         Sex
## Survived female male
##        0    151  703
##        1    306  142
```

```
fisher.test(svp_cnt)
```

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  svp_cnt
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.07577233 0.13112456
## sample estimates:
## odds ratio 
## 0.09989048
```

The fisher test is also get p-value less than 0.05, which conforms the conclusion from a) and b), gender also presents significant effect on survival status regarding the p-value: female has higher survival rate.

f) If we only care about the qualitative effect of gender and class, the contingency table is also applicable. The relative advantage of the contingency table over logistic regression is that the method only relies on counting numbers, which means the implementation can be fast, and the result is intuitive. But the contingency table can not give a quantitative conclusion. We do not know to what extent a factor affect the survival rate.

The logistic regression, as a generic linear model, shows qualitative effects by signs of coefficients. Moreover, we can use the model to predict new data out of our samples. But the modeling method is relatively more complicated compared with the contingency table. Fitting a model needs data processing and model justification, which can not be handled by a simple p-value test as we discussed in question d).

Exercise 3

a) We first perform the Poisson regression with all the explanatory variables. As we can see in the summary, only the numerical variables 'oligarchy', 'parties' and the categorical variable 'pollib' are significant, while the remaining variables are not(p-value larger than 0.05). 'oligarchy' and 'parties' have positive effect(0.073 and 0.031) on the number of military coups, whilst political liberalization has negative effect.

```
po=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,
       data = mili,family = poisson)
summary(po)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson, data = mili)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5075  -0.9533  -0.3100   0.4859   1.6459
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.2334274  0.9976112  -0.234  0.81500
## oligarchy    0.0725658  0.0353457   2.053  0.04007 *
## pollib1     -1.1032439  0.6558114  -1.682  0.09252 .
## pollib2     -1.6903057  0.6766503  -2.498  0.01249 *
## parties      0.0312212  0.0111663   2.796  0.00517 **
## pctvote      0.0154413  0.0101027   1.528  0.12641
## popn         0.0109586  0.0071490   1.533  0.12531
## size        -0.0002651  0.0002690  -0.985  0.32444
## numelec     -0.0296185  0.0696248  -0.425  0.67054
## numregim     0.2109432  0.2339330   0.902  0.36720
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 28.249  on 26  degrees of freedom
## AIC: 113.06
##
## Number of Fisher Scoring iterations: 5
```

b) Then we apply the step-down method to reduce the unnecessary variables, removing the variable having the highest p-value in each round. The variables are removed in the order of ‘numlec(p=0.67054)’, ‘numregim(p=0.42075)’, ‘size(p=0.33262)’, ‘popn(p=0.30204)’, ‘popvote(p=0.18383)’. The final model is $log\lambda = 0.207 + 0.091 * oligarchy + 0.022 * parties - 0.495 * pollib1 - 1.112 * pollib2$, where ‘pollib1’ and ‘pollib2’ are binomial and mutually exclusive. Besides ‘pollib1’, all the remaining variables are significant. But we can not drop ‘pollib1’ solely since the result of ANCOVA indicated that ‘pollib’ is a significant variable(p-value=0.02727<0.05).

When comparing the models in a) and b), we notice that both models have the null deviance of 65.945, but the residual deviances are different: the residual deviance of the second model(32.822) is closer to its degrees of freedom(31), as the first model is 28.249 and its degrees of freedom is 26. Hence, we prefer the model with fewer explanatory variables.

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,
            data=mili,family=poisson))$coefficients#remove numlec
```

```
##                 Estimate  Std. Error    z value    Pr(>|z|)
## (Intercept) -0.233427404 0.997611246 -0.2339863 0.814995584
## oligarchy    0.072565776 0.035345686  2.0530306 0.040069605
## pollib1     -1.103243890 0.655811418 -1.6822578 0.092518855
## pollib2     -1.690305731 0.676650306 -2.4980492 0.012487887
## parties      0.031221156 0.011166298  2.7960166 0.005173674
## pctvote      0.015441251 0.010102737  1.5284226 0.126407643
## popn         0.010958554 0.007148991  1.5328812 0.125305131
## size        -0.000265074 0.000269008 -0.9853760 0.324439426
## numelec     -0.029618512 0.069624834 -0.4254015 0.670543980
## numregim     0.210943219 0.233932967  0.9017251 0.367202938
```

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numregim,
            data=mili,family=poisson))$coefficients #numerigm
```

```
##                  Estimate   Std. Error    z value    Pr(>|z|)
## (Intercept) -0.4577458289 0.8602344719 -0.5321175 0.594644608
## oligarchy    0.0812015376 0.0288154207  2.8179890 0.004832547
## pollib1     -0.9642975542 0.5620939337 -1.7155452 0.086245314
## pollib2     -1.5149509438 0.5269441006 -2.8749747 0.004040599
## parties      0.0293409468 0.0103100564  2.8458571 0.004429207
## pctvote      0.0139115305 0.0094653971  1.4697250 0.141636255
## popn         0.0099592030 0.0067248724  1.4809505 0.138619769
## size        -0.0002687704 0.0002686512 -1.0004436 0.317095873
## numregim     0.1804415213 0.2241166271  0.8051233 0.420748524
```

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size,
            data=mili,family=poisson))$coefficients #size
```

```
##                  Estimate   Std. Error     z value    Pr(>|z|)
## (Intercept)  0.0419756813 0.5774100214  0.07269649 0.942047643
## oligarchy    0.0894950616 0.0270440021  3.30923882 0.000935500
## pollib1     -0.9673252804 0.5605601269 -1.72564054 0.084412101
## pollib2     -1.5321125526 0.5232779163 -2.92791365 0.003412448
## parties      0.0288170423 0.0102172799  2.82042211 0.004796051
## pctvote      0.0149215757 0.0093762023  1.59143065 0.111512690
## popn         0.0071646561 0.0056842444  1.26044124 0.207510232
## size        -0.0002579079 0.0002662008 -0.96884720 0.332621435
```

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn,
            data=mili,family=poisson))$coefficients #popn
```

```
##                 Estimate  Std. Error    z value    Pr(>|z|)
## (Intercept) -0.231434917 0.528887463 -0.4375882 0.661684821
## oligarchy    0.083467586 0.025829007  3.2315445 0.001231231
## pollib1     -0.683589192 0.495822322 -1.3786979 0.167987918
## pollib2     -1.320568052 0.490268490 -2.6935609 0.007069322
## parties      0.029769711 0.010309890  2.8874907 0.003883281
## pctvote      0.013924684 0.009370609  1.4859956 0.137280288
## popn         0.005659313 0.005483446  1.0320723 0.302038232
```

```
summary(glm(miltcoup~oligarchy+pollib+parties+pctvote,
            data=mili,family=poisson))$coefficients #popvote
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept) -0.11649888 0.513750518 -0.2267616 8.206091e-01
## oligarchy    0.09471197 0.023183825  4.0852606 4.402738e-05
## pollib1     -0.62075614 0.487525525 -1.2732793 2.029190e-01
## pollib2     -1.31037384 0.489017399 -2.6796058 7.370891e-03
## parties      0.02574472 0.009552367  2.6951146 7.036443e-03
## pctvote      0.01205704 0.009071988  1.3290409 1.838345e-01
```

```
summary(glm(miltcoup~oligarchy+pollib+parties,data=mili,family=poisson))
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
##     data = mili)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3609  -1.0407  -0.3153   0.6145   1.7536
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.207981   0.445679   0.467   0.6407
## oligarchy    0.091466   0.022563   4.054 5.04e-05 ***
## pollib1     -0.495414   0.475645  -1.042   0.2976
## pollib2     -1.112086   0.459492  -2.420   0.0155 *
## parties      0.022358   0.009098   2.458   0.0140 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 32.822  on 31  degrees of freedom
## AIC: 107.63
##
## Number of Fisher Scoring iterations: 5
```

```
drop1(glm(miltcoup~oligarchy+pollib+parties,data=mili,family=poisson),test='Chisq')
```

```
## Single term deletions
##
## Model:
## miltcoup ~ oligarchy + pollib + parties
##           Df Deviance    AIC     LRT  Pr(>Chi)
## <none>          32.822 107.63
## oligarchy  1   49.458 122.27 16.6363 4.528e-05 ***
## pollib     2   40.025 110.83  7.2038   0.02727 *
## parties    1   38.162 110.97  5.3401   0.02084 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

c)

```
bestmili=glm(miltcoup~oligarchy+pollib+parties,data=mili,family=poisson)
tt=apply(mili[,c(2,4,5,6,7,8,9)],2,mean)
```

```r
newmili2=data.frame(oligarchy=rep(tt[1],times=3),pollib=c(0,1,2),parties=rep(tt[2],times=3))
newmili2$pollib=as.factor(newmili2$pollib)
pred2=predict(bestmili,newmili2,type = 'response');pred2
```

```
##         1         2         3
## 2.9083520 1.7721126 0.9564757
```

The predicted ‘miltcoup’ for the three ‘pollib’ is 2.908, 1.772 and 0.956 respectively. This result is coherent with the negative coefficients in the summary in b), as ‘pollib1’ and ‘pollib2’ are supposed to have fewer coups than ‘pollib0’. In other words, more civil rights leads to smaller number of successful military coups.