Experimental Design and Data Analysis

Assignment 3

Tommaso Castiglione Ferrari 2673807

Daniyal Selani 2692551

Simone Korteling 2671463

Group 71

Exercise 1.

A.

Make an informative plot of the data. Investigate whether sexual activity influences longevity by performing a statistical test, without taking the thorax length into account. What are the estimated longevities for the three conditions? Comment.

```
df$loglongevity = log(df$longevity)
plot(df$longevity~df$thorax,pch=as.character(df$activity))
```

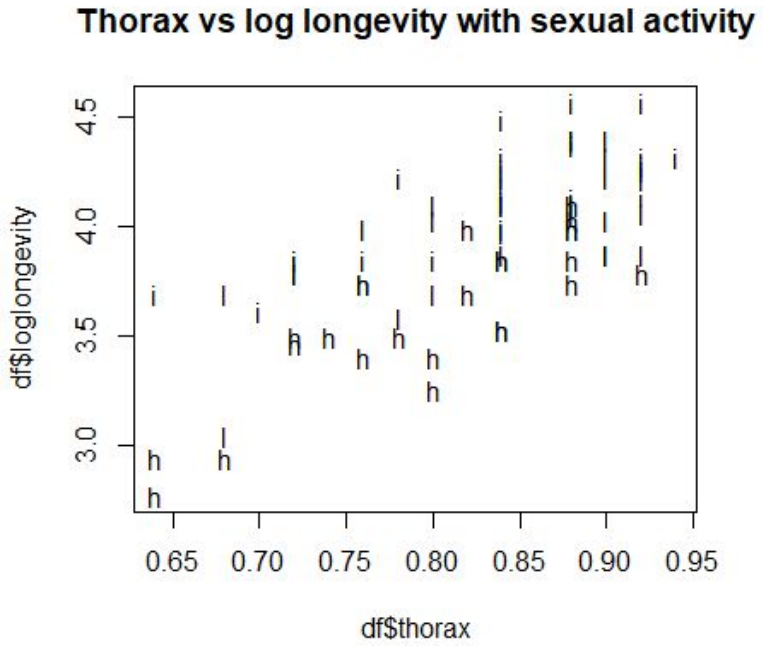


Fig 1.1

```
m1 = lm(loglongevity~activity, data=df)
print(anova(m1))

Analysis of Variance Table

Response: loglongevity
          Df Sum Sq Mean Sq F value    Pr(>F)
activity   2 3.6665  1.8333  19.421 1.798e-07 ***
Residuals 72 6.7966  0.0944
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P value < 0.05, and hence we can reject H0. Sexual activity is significant for loglongevity. However 1-way anova test with a factor is not correct. We must consider the numeric variable, or atleast investigate the interaction between numeric and factor variable.

```
summary(m1)
Call:
lm(formula = loglongevity ~ activity, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95531 -0.13338  0.02552  0.20891  0.49222

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.60212    0.06145  58.621  < 2e-16 ***
activityisolated  0.51722    0.08690   5.952 8.82e-08 ***
activitylow       0.39771    0.08690   4.577 1.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3072 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
```

mean loglongevity for high activity = 3.6
isolated activity = 3.6+0.52 = 4.12
low activity = 3.6 + 0.40 = 4.0

B. Investigate whether sexual activity influences longevity by performing a statistical test, now including thorax length as an explanatory variable into the analysis. Does sexual activity increase or decrease longevity? What are the estimated longevities for the three groups, for a fly with average thorax length?

```
m2 = lm(loglongevity~thorax+activity,data=df)
drop1(m2,test='F')
```

```
Model:
loglongevity ~ thorax + activity
         Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                 2.9180 -235.50
thorax    1    3.8786 6.7966 -174.08  94.374 1.139e-14 ***
activity  2    2.1129 5.0309 -198.64  25.705 4.000e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P value < 0.05. We can reject H0, sexual activity does affect log longevity

```
summary((m2))
```

```
Call:
lm(formula = loglongevity ~ thorax + activity, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4858 -0.1612  0.0104  0.1510  0.3574

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.21893    0.24865   4.902 5.79e-06 ***
thorax            2.97899    0.30665   9.715 1.14e-14 ***
activityisolated  0.40998    0.05839   7.021 1.07e-09 ***
activitylow       0.28570    0.05849   4.885 6.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2027 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
```

The estimates from the summary show that high sexual activity results in shorter longevity.

This is corroborated by the loglongevity for the average thorax length for each activity level.

```
avgthorax = mean(df$thorax)
for (i in c('low', 'isolated', 'high')){


ndata = data.frame(thorax=avgthorax, activity=i)
ndata$activity = as.factor(ndata$activity)
print(i)
print(predict(m2, ndata, type='response'))
}
```

```
[1] "low"
      1
3.96091
[1] "isolated"
      1
4.08519
[1] "high"
       1
3.675209
```

C.

How does thorax length influence longevity? Investigate graphically and by using an appropriate test whether this dependence is similar under all three conditions of sexual activity.

```
plot(loglongevity~thorax,pch=unclass(activity), data=df)
for (i in c('high', 'low', 'isolated')) abline(lm(loglongevity~thorax,data=df[df$activity==i,]))
```

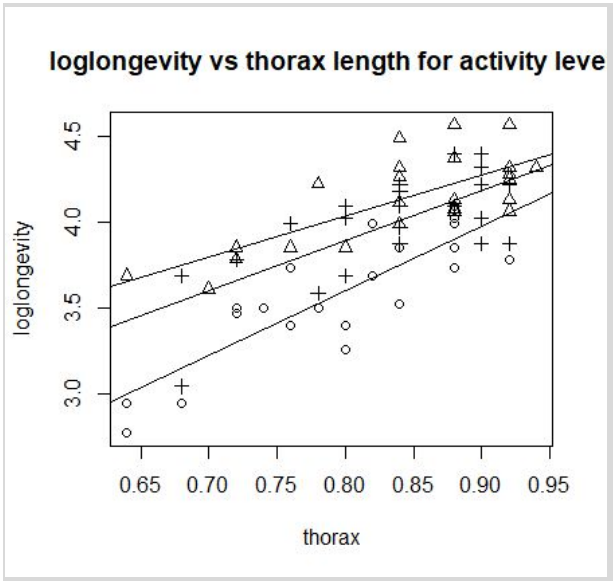


Fig 1.2

```
m3 = lm(loglongevity~activity*thorax,data=df)
anova(m3)
```

Analysis of Variance Table

Response: loglongevity

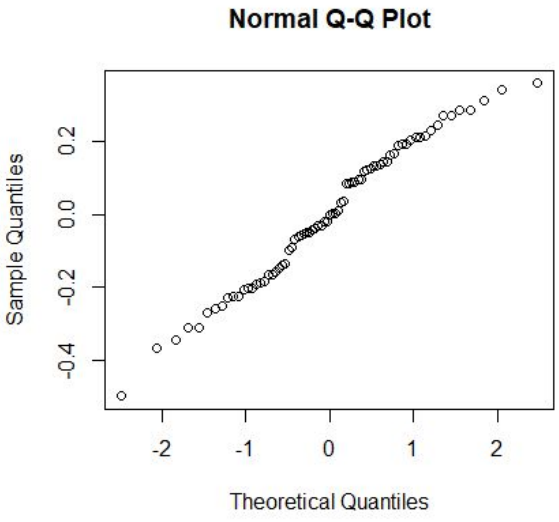
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
activity	2	3.6665	1.8332	45.7687	2.228e-13	***
thorax	1	3.8786	3.8786	96.8327	9.020e-15	***
activity:thorax	2	0.1542	0.0771	1.9251	0.1536	
Residuals	69	2.7638	0.0401			

P value for thorax is < 0.05. Hence we can reject H0, and deduce that thorax is indeed significant for loglongevity.
P value for interaction between activity and thorax is not significant, p > 0.05. H0 cannot be rejected, and there is no significant interaction between the two variables.

D.
Which of the two analyses, without or with thorax length, do you prefer? Is one of the analyses wrong?

As there is no interaction between the predictor variables, analysis with and without thorax is valid. However, because r squared error for model without thorax is considerably lower, we prefer that analysis.

E.
Verify normality and heteroscedasticity by making a normal QQ-plot of the residuals, and a residuals versus fitted plot, for the analysis that includes thorax length.



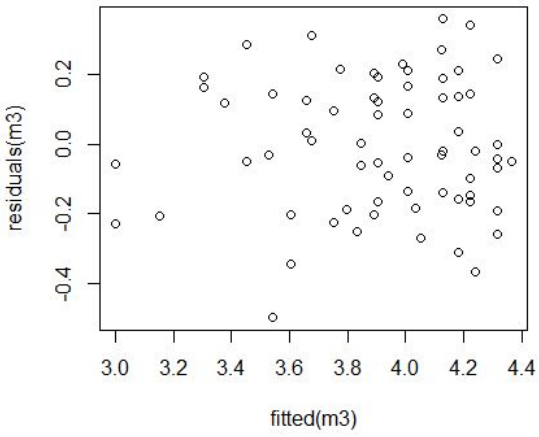


Fig 1.3

Fig 1.4
The qqplot shows residuals are normally distributed and the residuals vs fitted plot is highly heteroscedastic.

F.
Perform the ancova analysis with the number of days as the response, rather than its logarithm. Verify normality and homoscedasticity of the residuals of this analysis. Was it wise to use the logarithm as response?

```
m4 = lm(longevity~thorax+activity,data=df)
drop1(m4, test='F')
```

```
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 ***
---
```

ANCOVA reveals that both thorax and activity are significant ($p < 0.05$). Hence, we can reject H0 for both variables.

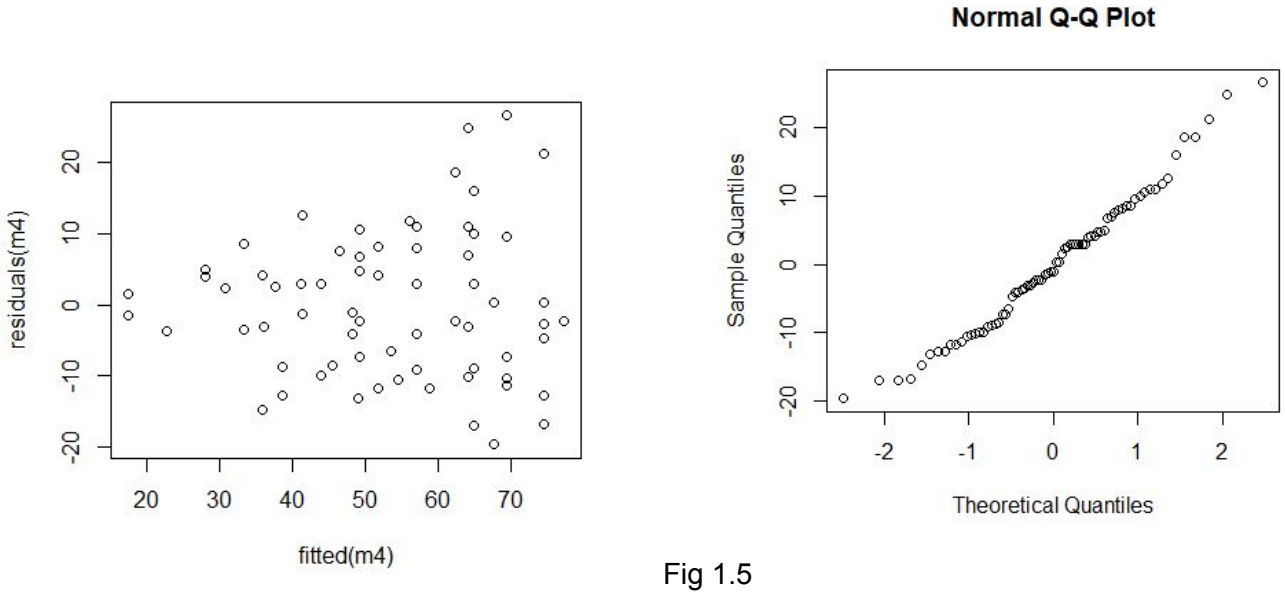


Fig 1.5

Fig 1.6

The qqplot of the residuals reveals a slightly less ideal normal distributions, and the fitted vs residuals plot reveals slightly more homoscedasticity.

Exercise 2. Military coups in Africa

To study the influence of different political and geographical variables on the number of military coups, these data are collected for several Sub Saharan African countries in the file africa.txt.

a) Study the data and give a few (>1) summaries (graphics or tables).

Code:

```
titanic$Age = as.numeric(titanic$Age)
titanic$PClass = as.numeric(titanic$PClass)
hist(titanic[,3],main="Age")
hist(titanic[,2], main="PClass")
```

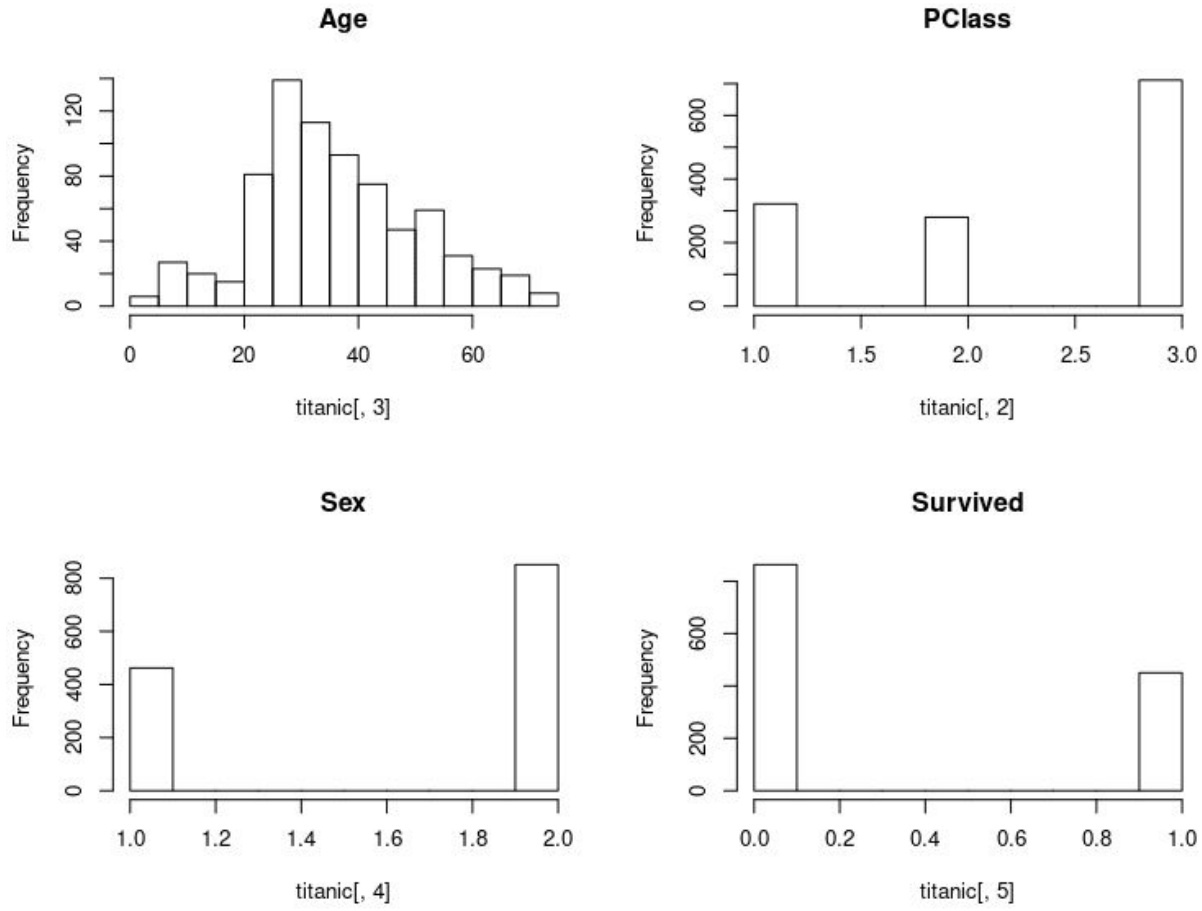


Fig. 2.1

In the images Fig. 2.1 we can observe the number of data available related to the different variables. As we can see, in our data, most of the people that were onboard of the Titanic were male, in their late 20s and from and residing in the 3 passenger class. Unfortunately, the available data shows, as we know, that most of these people did not survive when the Titanic sank.

Code:
```
xtabs(~titanic$PClass+titanic$Sex, data=titanic)
```

	Sex	
PClass	1	2
1st	143	179
2nd	107	173
3rd	212	499

In the table above we can see the combination of people divided by gender and passenger class. As already seen in the histograms above, the vast majority of the population resides in the third class, with an overwhelming difference between males (2) and females (1)

Code:
```
xtabs(~titanic$Survived+titanic$Sex, data=titanic)
```

	Sex	
Survived	1	2
0	154	709
1	308	142

In the table above we can see the combination between who survived and the gender of the passengers, and we can determine that there is a correlation between the gender of each passenger and its probability of surviving the catastrophe.

Code:
```
xtabs(~titanic$Survived+titanic$PClass, data=titanic)
```

	PClass		
Survived	1st	2nd	3rd
0	129	161	573
1	193	119	138

In the table above we can see the combination between who survived and the residing class of the passengers, and we can determine that there is a correlation between the class of each passenger and its probability of surviving the catastrophe.

Code:

```
tot = xtabs(titanic$Survived~titanic$PClass+titanic$Sex, data=titanic)
print(round(tot/xtabs(~titanic$PClass+titanic$Sex), 2))
```

	Sex	
PClass	1	2
1st	0.94	0.33
2nd	0.88	0.14
3rd	0.38	0.12

Finally, in the table above we can see an early probability result of the chance of survival of a passenger, considering its gender and class on the cruise. As already anticipated, we can see a clear increase in the probability of surviving the event when moving to a more prestigious class, and moving from male to female. We can see, in fact, that the highest number of survivors in percentage were women from the 1st class with a 94% survival rate, while men from the 3rd class had an overall survival rate of 12%.

b) Fit a logistic regression model (no interactions yet) to investigate the association between the survival status and the predictors PClass, Age and Sex. Interpret the results in terms of odds, comment.

Considering the variables PClass, Age and Sex as predictors, we obtain the following logistic regression model:

Code:

```
titanic$Sex = as.numeric(titanic$Sex)
titanic$PClass = as.numeric(titanic$PClass)
titanic$Age = as.numeric(titanic$Age)
call <- glm(Survived~PClass+Age+Sex,family=binomial, data=titanic)
summary(call)
```

The estimated odds from this model are the following:

$$\hat{o}_k = Pr(Y_k =1) \setminus Pr(Y_k =0) \approx$$
$$\exp\{7.88 -1.26\ PClass_k - 0.04\ age_k - 2.63\ Sex_k \}$$

Analyzing the odds obtained from this model, we can preliminary determine some interesting features. First of all, we can see that by increasing the passenger class (as a remainder, 1st class is the most prestigious, while the 3rd is the lowest), the probability of surviving sensibly decreases. Moreover, being a male (or Sex = 2) has the most drastic change in the linear predictor, by reducing the chance of surviving of a factor of -2.63. Finally, it seems that the age variable has the minimum effect on the possibility of surviving.

c) Investigate for interaction of predictor Age with factors PClass and Sex. From this and b), choose (and justify) a resulting model. For this model, report the estimate for the probability of survival for each combination of levels of the factors PClass and Sex for a person of age 53.

For investigating the interaction between the variable Age and, separately, the two variables PClass and Sex (i.e. Age:PClass and Age:Sex) we can implement the following logistic regression model:

Code:
```
titanic$Sex = factor(titanic$Sex)
titanic$PClass = factor(titanic$PClass)
call1 <- glm(Survived~Age*(PClass+Sex),family=binomial, data=titanic)
anova(call1_alternativa, test="Chisq")
```

```
          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      755    1025.57
Age            1    2.972       754      1022.60    0.08471 .
PClass         2  113.028       752      909.57   < 2.2e-16 ***
Sex            1  214.986       751      694.59   < 2.2e-16 ***
Age:PClass  2    5.477         749       689.11     0.06467 .
Age:Sex        1   26.979       748      662.13      2.057e-07 ***
```

As we can see from the result obtained by the Anova executed with the Chi-squared test, there is no meaningful interaction between Age and PClass, while it seems that a strong interaction between Age and Sex is present. Considering this, and the importance represented by the variable PClass from the odds analysis in the previous sub-question, we can implement a resulting model on the basis of this information:

Code:
```
titanic$PClass = factor(titanic$PClass)
titanic$Sex = factor(titanic$Sex)
titanic$Age = as.numeric(titanic$Age)
model = glm(Survived~PClass+Age*Sex,data=titanic,family=binomial)
```

Given this model, we can estimate the probability of survival of a 53 years old person for each combination of levels of the factors PClass and Sex:

Code:
```
for (pc in levels(titanic$PClass)){
  for (sex in levels(titanic$Sex)) {
    newdata = data.frame(Sex=sex, PClass=pc, Name="Mr. Mc Buttface", Age=53)
    print(newdata)
    print(predict(model, newdata, type="response"))
    }}
 Sex PClass Name Age
   1      1        test    53   = 0.95 %

  Sex PClass Name Age
   2      1        test    53   = 0.23 %

  Sex PClass Name Age
```

```
 1     2        test    53   = 0.79 %

Sex PClass Name Age
 2     2        test    53   = 0.06 %

 Sex PClass Name Age
 1     3        test    53   = 0.55 %

 Sex PClass Name Age
 2     3        test    53    = 0.02 %
```

Remembering that Sex = 1 is female and Sex = 2 is male, and PClass = 1 is first class (most luxurious) and PClass = 3 is third class (most humble).

d) Propose a method to predict the survival status and a quality measure for your prediction and describe how you would implement that method (you do not need to implement it).

Fitting the observed data $(X_1, Y_1), \ldots, (X_N, Y_N)$ in logistic regression, we have

$$\Pr(Y_k = 1) = 1 / (1 + e^{-x^{(T)}_{(k)} \theta}) \quad \text{for } k = 1, \ldots, N,$$

And we obtain (by the maximum likelihood) an estimate $\hat{\theta}$ of the parameter θ.
For a new predictor vector X_{new}, we can predict its success probability

$$\hat{P}_{new} = 1 / (1 + e^{-x^{(T)}_{(new)} \hat{\theta}})$$

Now we can use $\hat{P}_{new}$ to predict the new label $\hat{Y}_{new}$ as

$$\hat{Y}_{new} = \begin{cases} 1 & \text{if } \hat{P}_{new} \geq p_0 \\ 0 & \text{if } \hat{P}_{new} < p_0 \end{cases}$$

for some threshold $p_0 \in [0, 1]$.

Considering this, because we used the maximum likelihood, we can consider the value of $\hat{P}_{new}$ as a measure of the quality of the prediction. In fact, if $\hat{P}_{new} \gg p_0$ or $\hat{P}_{new} \ll p_0$ we can have a more “safe” prediction of $\hat{Y}_{new}$ - i.e. we can assert with more certainty that $\hat{Y}_{new}$ is either equal to 1 or to 0, namely if $\hat{P}_{new} \gg p_0$ or $\hat{P}_{new} \ll p_0$ -, than if $\hat{P}_{new}$ is close to the value of p_0.

e) Another approach would be to apply a contingency table test to investigate whether factor passenger class (and gender) has an effect on the survival status. Implement the relevant test(s).

As an alternative approach, we can use the contingency table tests Chi-squared and Fisher tests, for studying the dependency of namely the class and the survival rate, and the gender and the survival status.

Code:
```
chisq.test(titanic$PClass, titanic$Survived)
fisher.test(titanic$Sex, titanic$Survived)
```

The data obtained is:

```
data:  titanic$PClass and titanic$Survived
X-squared = 172.3, df = 2, p-value < 2.2e-16
```

And

```
data:  titanic$Sex and titanic$Survived
p-value < 2.2e-16
```

So it looks like there is a strong dependency relationship for both the passenger class and the survival rate, and for the gender and the survival rate - as we already assessed -.

f) Is the second approach in e) wrong? Name both an advantage and a disadvantage of the two approaches,relative to each other.

The second approach is not wrong, but it retrieves different information, and has some dataset sizes limitations.
In fact, the output obtained using the logistic regression model uncovers only the probability of survival.
Conversely, the contingency tests retrieve a different type of information, namely the dependency between different variables.
Moreover, the Fisher test can only be used on tables that are 2x2 or, in other words, one variable against another variable. Therefore, the Fisher test is more precise when applied to small datasets, which is not true for logistic regression. But, the disadvantage of the fisher test with regards to the logistic regression model resides exactly in its size limitations, that the logistic regression does not have.
On the other side, also the Chi-squared test is size bounded, being more accurate in its approximation with a wider dataset. This problem arises because chi-squared follows the chi-squared distribution only approximately. The more observations we have, the better the approximation is going to be. This, once again, is not true for the logistic regression model.
Finally, the latter allows for predictions to be made, because it has an intercept and coefficients.
For this analysis, the logistic regression approach seems to be more accurate and an overall better choice, considering the dataset size and the prediction task of the exercise.

Exercise 3. Military coups in Africa

To study the influence of different political and geographical variables on the number of military coups, these data are collected for several Sub Saharan African countries in the file africa.txt.

a) Perform Poisson regression on the full data set africa, taking miltcoup as response variable, Comment on your findings.

In this research the response variable *Y* is a count, namely = miltcoup (numerical). The explanatory variables are both numerical and factorial. Before we performed the Poisson regression we changed the variables onto the right data type, resulting in the following variables: oligarchy(numerical), pollib (factor, with three levels 0-2), parties (numerical), pctvote (numerical), popn (numerical), size (numerical), numelec (numerical), numregim (numerical).

We performed the Poisson regression with the following code:

Code:

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

Output:

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

With an alpha of <0.05 the variables ‘oligarchy’, ‘pollib2’ and ‘parties’ have a significant effect on the number of military coups. The coefficient for oligarchy is 0.073, which indicates that the expected log count for the number of military coups for a one-unit increase in oligarchy is 0.073. The coefficient for pollib2 is -1.69, which indicates that the expected log count for the number of military coups for a one-unit increase in pollib2 is -1.69. The coefficient for parties is 0.031, which indicates that the expected log count for the number of military coups for a one-unit increase in parties is 0.031. The interpreting of the output results in the following model for estimating the number of military coups: -0.23intercept + 0.073oligrchy - 1.69pollib2 + 0.031parties.

b) Use the step down approach (using output of the function summary) to reduce the number of explanatory variables. Compare the resulting model with your findings in a).

By using the step down method we iteratively excluded the variable with the highest p-value, resulting in a final model with only significant variables explaining the outcome. We respectively excluded the following variables to improve the model: numregim, numelec,size, popn, pctvote. The step down procedure results

in a model with the following variables: oligarchy, pollib and parties, as significant explanatory variables for the response variable.

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

Output:
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.825480   0.527632   1.565  0.11770
oligarchy    0.092622   0.021779   4.253 2.11e-05 ***
pollib      -0.574103   0.204383  -2.809  0.00497 **
parties      0.022059   0.008955   2.463  0.01377 *
```

Comparing this with the model of question 3a, we see a difference in the number of variables. Where we used 8 variables to explain the outcome variable in the first model, we now have 3 variables (oligarchy + pollib + parties) that are essential in explaining *Y*. The first model has a degrees of freedom of 27, which is lower than the top-down model with a degrees of freedom of 32. The p-values of the variables in the first model are (except for parties) higher than our final top-down model. So for most of the variables the statistics increase (lower p-values), when we exclude non-significant variables from the model.

c) Predict the number of coups for a hypothetical country for all the three levels of political liberalization and the averages (over all the counties in the data) of all the other (numerical) characteristics. Comment on your findings.

First we changed the data such that we use the average of the numerical explanatory variables. The response variable *Y* is the number of military coups. We then predicted the number of military coups for the three levels of political liberalization.

Code:
```
pred_lib_data = data.frame(oligarchy=avg_oligarchy, pollib=0, parties=avg_parties)  #level 0
output = predict(topdown_model, pred_lib_data, type="response"); output

pred_lib_data = data.frame(oligarchy=avg_oligarchy, pollib=1, parties=avg_parties)  #level 1
output = predict(topdown_model, pred_lib_data, type="response"); output

pred_lib_data = data.frame(oligarchy=avg_oligarchy, pollib=2, parties=avg_parties)  #level 2
output = predict(topdown_model, pred_lib_data, type="response"); output
```

The output shows that for level 0 on ‘political liberalization’ the number of predicted military coups of a hypothetical country is: 3.04. For level 1 on ‘political liberalization’ the number of predicted military coups of a hypothetical country is: 1.71. Finally for level 2 on ‘political liberalization’ the number of predicted military coups of a hypothetical country is: 0.96. We observe that the more civil rights a country has (level 2) the lower the corresponding number of military coups is.