CS 4125 Seminar in Research Methodology for Data Science Coursework A

Dimitropoulos Georgios: 4727657
Manousogiannis Emmanouil :4727517
Mastoropoulou Emmeleia: 4743539
Group 13

March 19, 2018

Part 1-Design and set up true experiment

Experiment topic:The effect of alcohol consumption in human reaction time.

Motivation for the planned research

A lot of discussion has been made nowadays on how dangerous alcohol consumption is when people have to perform certain activities. For instance, alcohol consumption is one of the main reasons for serious car accidents, despite the fact that there are strict laws for this issue. However, most people tend to underestimate the effects of even a small dosage of alcohol in their body. The motivation beyond our experiment is to produce convincing results, pointing that even a very small dosage of alcohol can significantly reduce our reaction time, which can lead to very unpleasant events under certain circumstances, like a car accident or another occasion of emergency. Especially for the participants of the experiment this can be a very efficient way to convince them.

The theory underlying the research

In order to be able to investigate the relationship between the alcohol consumption and the body reaction time, we searched for relevant research reported in literature in order to confirm that our assumptions are valid. In [1] Kerr and Hundmarch studied the effects of alcohol on a variety of human mental or physical tasks like information processing, task performance etc, focusing on reaction time and activities related to driving. They claim that whether the dosage is small, the effect varies per person and occasion, however as it increases or as the tasks become more complex the effect of alcohol becomes significant.A similar research was published a few years earlier by the same researchers [2], comparing the effect of alcohol to other drugs in human psychomotor performance.
In addition, in [3] Maylor and Rabbitt presented a research showing that there is an almost linear effect of alcohol in human information processing in general, something that of course is related to human reaction time.

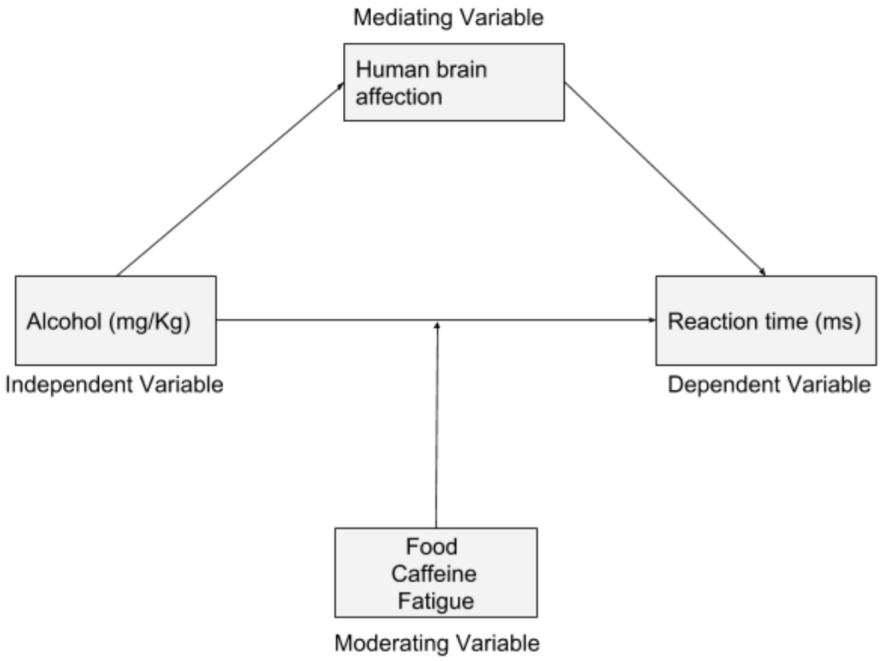
Hypothesis question

The hypothesis question that we are going to test is the following:
Does the amount of alcohol in mg/Kg bodyweight, affect the human reaction time on a certain reaction time test?

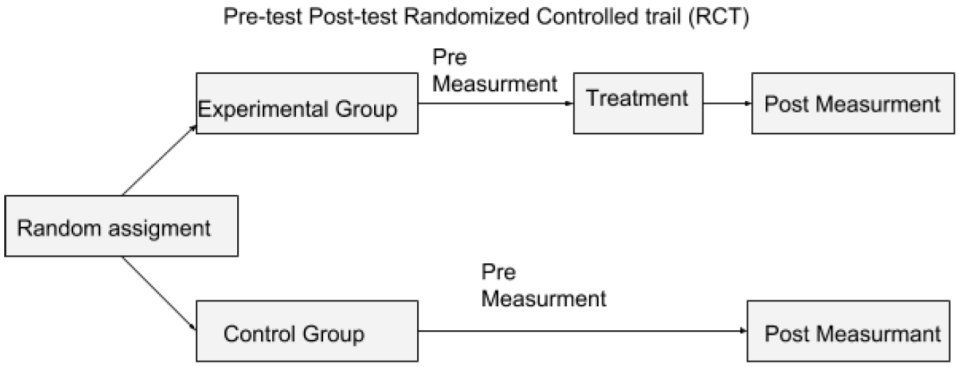
Conceptual model

Independent variable: The mg of alcohol per kg of body weight.
Dependent variable: Reaction time on a certain Reaction time test.
Mediating variable:Parts of brain related to physical performance are affected.
Moderating variable: Possible caffeine or food consumption before or after the drinking alcohol

and fatigue of the participant may affect the adult's performance on such a test, increasing or decreasing the correlation between the independent and the dependent variable.



Experimental design



After we have declared our **research question** ,defined our **hypothesis** and our **dependent and independent** variables, we have to select our experimental design.

Since we are performing a true experiment, we will implement a two-group experimental design, with pre-test post-test randomized control trail.
Initially, we will select **randomly** a number of participants for our experiment.Since, we will follow the two-group design, our sample will randomly be divided in two groups.The first group is the experimental group and the other is the **control group**. The experimental group is the one that receives treatment during our experiment,while the control group does not.
The next step is that both groups take a pre-measurement test in order to have an indicator of their performance in general, regardless of the independent variable.After that, the experimental group only will receive a treatment of our independent variable(alcohol) and then both groups will take a test again. We can now compare the performance of the two groups compared to the pre-measurement phase, and draw useful results about the effects of our independent variable on their performance.

Experimental procedure

The basic steps of our experimental procedure will be the following:

Step 1: We randomly select 100 participants for our experiment. We ask for volunteers,from universities companies etc, aiming to create a diverse random sample with people of different age,sex and physical condition.

Step 2:We assign a number to each participant and split them in two groups. The experimental and the control group. The categorization is done based on the numbers that are produced by a random number generator.

Step 3: Both the experimental and the control group take a pre-test without having received any treatment.In that way , we can keep track of the performance of each participant and each group,without the effect of the independent variable.The test that will measure the performance will be a human reaction time test, based on which, the participants will use have to react to certain 'sudden' events that occur in front of a computer screen.

Step 4:Both the experimental and the control group receive a 'treatment'.The control group will have a drink of no alcohol in order to ensure the 'placebo' effect. The experimental group will drink an alcoholic beverage, where the alcohol dosage will be determined as mg/kg of human weight.In this way we will ensure that the weight of each participant will not affect our results.

Step 5:Half an hour after the alcohol consumption, all the participants will take another reaction time test of similar difficulty to the pre-test. We now measure the performance of the two groups and compare it to their previous results in order to retrieve useful results.

Measures

In order to be able to assess the reaction time with respect to the amount of alcohol consumption, a human benchmark test which will measure the reaction time of the participants to certain events will take place. In that way, it would be possible to define a quantitative measure to assess the reaction time of our participants on the criteria. Examples of such tests are already available online.(https://www.humanbenchmark.com/)

Participants

The aforementioned group, namely these 100 people who are selected for our experiment come from universities companies etc, aiming to create a diverse random sample with people of different age,sex and physical condition are going to be the participants of our experiment.

Suggested Statistical Analysis

In order to be able to investigate the hypothesis question which we pose, statistical analysis should be performed. The first thing to do with the available data which we have is to summarize it, which means to present it in a way that best tells the story. For instance, a histogram could be plotted where the vertical axis corresponds to the absolute value of the difference in reaction time of the participants (people who belong to a different age-range) to a certain reaction test before and after the treatment which they received and the horizontal axis corresponds to the different categories of groups which we have, namely the experimental and the control group who consumed a different amount of alcohol in mg/Kg bodyweight.
In addition, more advanced statistical analysis could be examined which aims to identify patterns in data, for example, whether there is a link between two variables, or whether certain groups are more likely to show certain attributes through examining research hypotheses.Hence, after this we have to choose the right test that applies to your data. For instance, in our case, different models can be compared with an anova test. Finally, we have to compare the test statistic to that required to meet the desired level of significance. (usually 5% or 1%). This significance level is called the p

value. The smaller p-value is, the more unlikely that null hypothesis will hold. So if this p-value is less than the above threshold, we reject the null hypothesis and conclude that the two populations are different.

Part 2-Generalized Linear Models

Question 1: Twitter sentiment analysis (Between groups – single factor)

1)Conceptual Model

2)Homogeneity of variance of sentiments

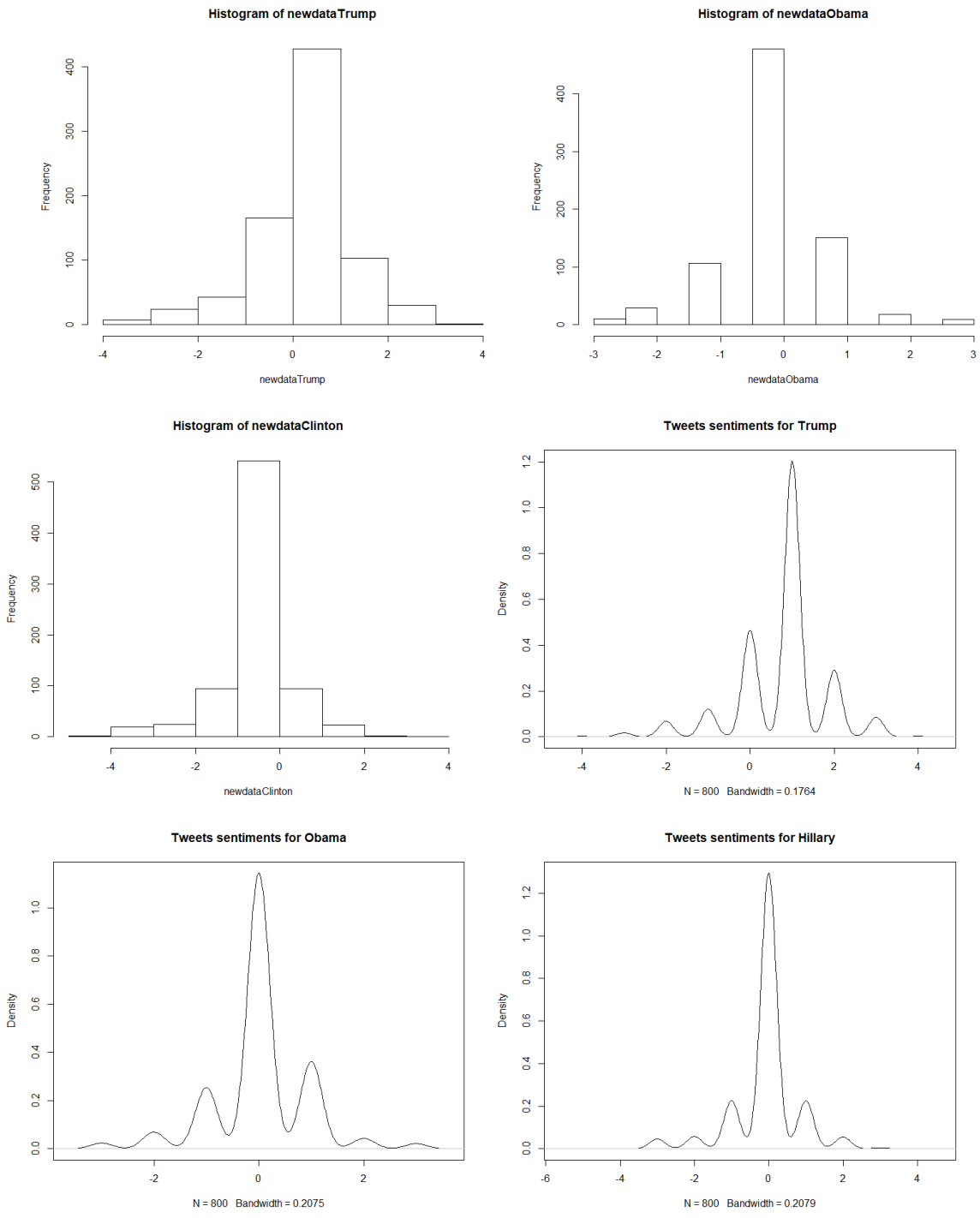
In order to test the homogeneity of variance We have implemented two different methods: Levene's test and Bartlett's test (as we think that our data comes from a normal or nearly normal distribution). If the resulting p-value of Levene's test is less than some significance level (typically 0.05), the obtained differences in sample variances are unlikely to have occurred based on random sampling from a population with equal variances. So in order to confirm our hypothesis testing techniques, we compare the p-value with 0.05. After testing the samples, for Levene's and Bartlett's tests, we noticed the following results (Table 1). Indeed p-values were smaller than 0.05, Hence, the null hypothesis is rejected. There is difference between our samples and there is no homogeneity of variance.

Test	p-value
Levene's test	9.589e-07
Bartlett's test	1.818e-07

Table 1: Tests For Homogeneity Of Variance

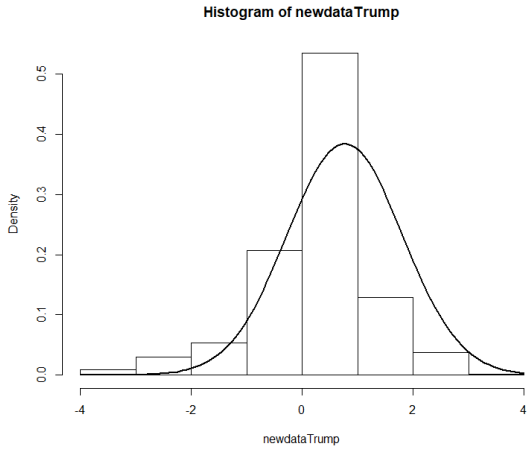
3)Variation in tweets' sentiments

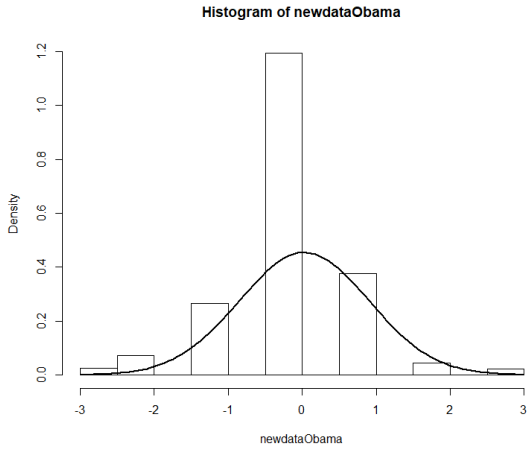
In order to examine the variation in tweets' sentiments we will make use of histograms and density plots. As it can be depicted from the graphs, the majority of the sentiment in the tweets of the three politicians is around zero (neutral). However, when the sentiment of the tweets that are related with the Trump are being considered, it is obvious that tweets with negative sentiment are more than the positive ones. On the contrary, in Obama's case the exactly opposite holds. Finally, there is a balance between positive and negative tweets for the Hilary Clinton.

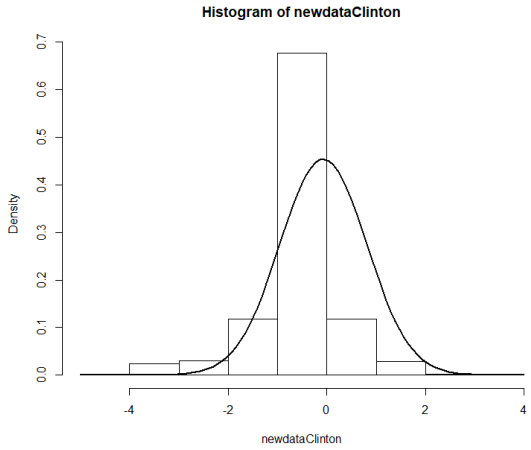


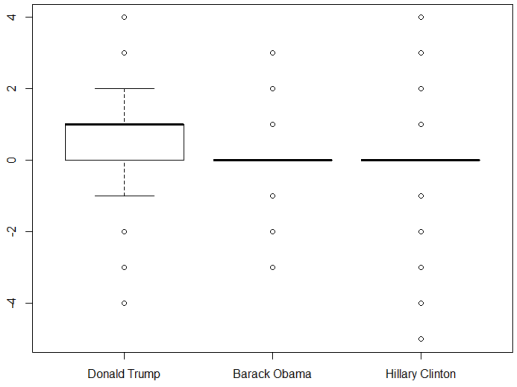
4)Mean sentiments

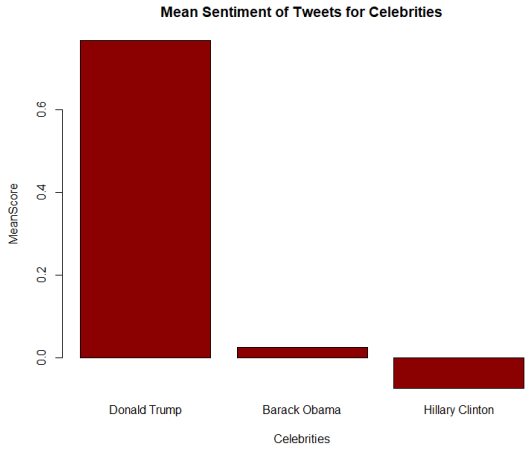
In order to examine the mean sentiments of tweets for each celebrity we consider to represent the histograms with the distribution (the mean and variation) of tweets for each celebrity and with a box plot. We can easily observe that the highest mean sentiment of tweets belongs to Donald Trump. Barack Obama comes secon and Hillary Clinton is in the third place.











5)Linear Model

In order to analyze whether the knowledge to which celebrity a tweet relates has a significant impact on explaining the sentiments of the tweets we make use of a linear model. First of all, We implement a model without predictor which is the null model and one model with candidate as predictor. After that, we used the anova function. The two factors that we compare are the Candidate and the score of the tweet. since, the P-value is smaller than 0.05 (Table 2) we can easily observe that the knowledge to which celebrity a tweet relates has a significant impact on explaining the sentiments of the tweets.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F) *
Model 1	2399	2433.8				
Model 2	2397	2096.1	2	337.71	193.1	< 2.2e-16

Table 2: Results from anova test

6)Post-Hoc Analysis

In this section, we used Bonferroni correction to examine which celebrity tweets differ from other celebrity tweets. As can be seen from the table below, Barack Obama and Hirary Clinton have a p-value larger than 0.05 and hence we do not have an indication if there is a significant difference in sentiment about those 2 celebrities. On the other hand, as it can be depicted in table 3, for all other combinations our p-value is below the threshold of 0.05 and we can conclude that there is a significant difference in the sentiment of their tweets.

	Donald Trump	Barack Obama
Barack Obama	<2e-16	-
Hillary Clinton	<2e-16	0.11

Table 3: Results from Post-hoc analysis

7)Section for a Scientific Publication

In order to confirm our hypothesis testing techniques, we compared the p-value with 0.05. After testing the samples, for Levene's and Bartlett's tests, we noticed that p-values were smaller than 0.05, Hence, the null hypothesis is rejected. There is difference between our samples and there is no homogeneity of variance.

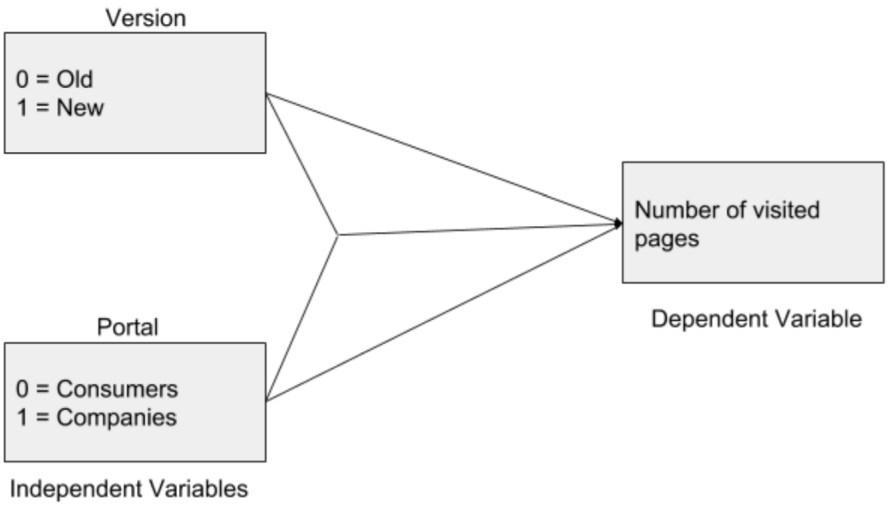
Through the graphical analysis which we performed, we conclude that the majority of the sentiment in the tweets of the three politicians is neutral. However, when the sentiment of the tweets that are related with the Trump are being considered, it is obvious that tweets with negative sentiment are more than the positive ones. On the contrary, in Obama's case the exactly opposite holds. Finally, there is a balance between positive and negative tweets for the Hilary Clinton.

Also, in order to analyze whether the knowledge to which celebrity a tweet relates has a significant impact on explaining the sentiments of the tweets we make use of a linear model and we used the anova function. The two factors that we compare are the Candidate and the score of the tweet. Since, the P-value is smaller than 0.05 we can observe that the knowledge to which celebrity a tweet relates has a significant impact on explaining the sentiments of the tweets.

Finally, through Bonferroni correction, we examined which celebrity tweets differ from other celebrity tweets. We found that, Barack Obama and Hirary Clinton have a p-value larger than 0.05 and hence we do not have an indication if there is a significant difference in sentiment about those 2 celebrities. On the other hand, for all other combinations our p-value is below the threshold of 0.05. Thus we conclude that there is a significant difference in the sentiment of their tweets.

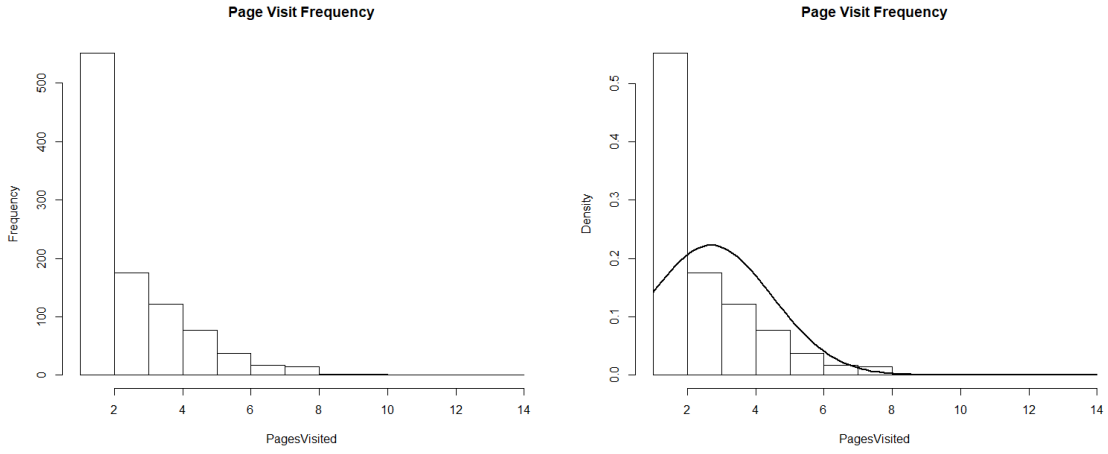
Question 2: Website visits (between groups – Two factors)

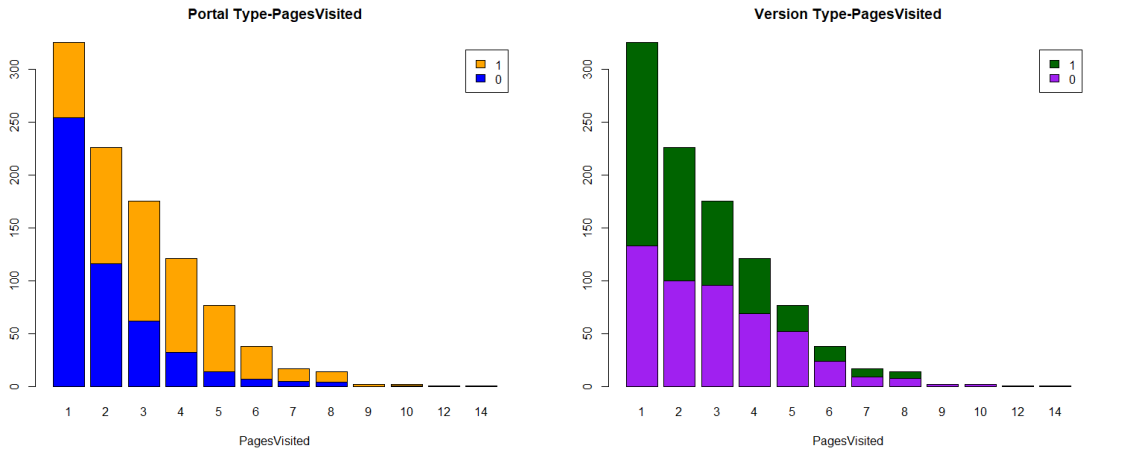
1)Conceptual Model



2)Variation in page visits

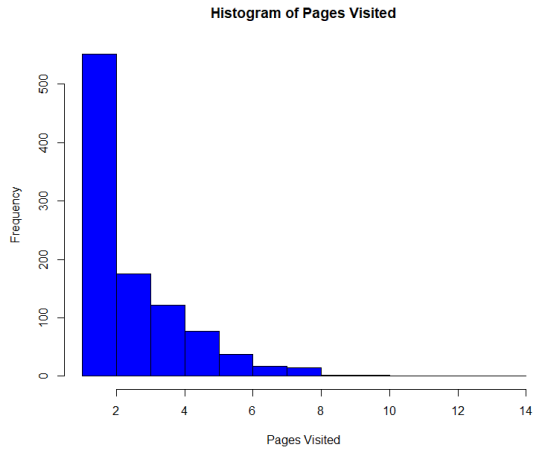
We can observe from the plots that new version is used more from the old version in total number of pages visited. Moreover, we can see that the portal from consumers is used more from the companies portal in in total number of pages visited.





3)Page Visits and normal distribution

The histogram confirms the non-normality of pages visited. Histograms of normal distributions show the highest frequency in the center of the distribution. From the histogram we can conclude that the the variable page visits is not a normal but a Poisson distribution because the events occur randomly at a constant rate.



4)Model Analysis

We create the following models: First of all, Model 0 the predictor version (null model). Model 1, just adding to Model 0 the predictor version. After that, Model 2 has as a predictor the portal. Furthermore, Model 3 has as predictors both version and portal. Finally,Model 4 has as predictors both version and portal as well as their combination
Furthermore we compared the previous models using anova function with test Chisq. In model 4 we observed the greatest Deviance and the lowest p-value(Table 4,5). Hence, we use anova function just for our best predictor model 4 to test if the analysis shows a significant two-way interaction effect. Easily can be observed from the results above that null hypothesis is rejected because the Pr(>Chi) value is less than 0.05 (Table 6).

	RResid. Df	Resid. Dev	Df	Deviance	Pr(>Chi) *
Model 0	998	1067.00			
Model 1	997	1032.8	1	34.249	4.85e-09

Table 4: Results from anova test between model 0 and model 1

	RResid. Df	Resid. Dev	Df	Deviance	Pr(>Chi) *
Model 0	998	1067.00			
Model 2	997	898.85	1	168.16	< 2.2e-16

Table 5: Results from anova test between model 0 and model 2

	RResid. Df	Resid. Dev	Df	Deviance	Pr(>Chi) *
Model 0	998	1067.00			
Model 3	996	861.98	2	205.02	< 2.2e-16

Table 6: Results from anova test between model 0 and model 3

5)Simple Effect Analysis

We can see from the following tables (7-10) that the t-value of version on companies portal has a larger distance from 1 than t-value of version on consumers portal. Moreover, p-value of version

	RResid. Df	Resid. Dev	Df	Deviance	Pr(>Chi) *
Model 0	998	1067.00			
Model 4	995	833.97	3	233.03	< 2.2e-16

Table 7: Results from anova test between model 0 and model 4

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi) *
NULL			998	1067.00	
version	1	34.249	997	1032.75	4.850e-09
portal	1	170.773	996	861.98	< 2.2e-16
version:portal	1	28.013	995	833.97	1.205e-07

Table 8: Results from anova model 4

on companies portal is less than p-value of version on consumers portal. Hence, the version on the consumers portal has a less significant effect than the version on the companies portal.

We can observe that p-value of portal on the old version is lower than p-value of portal on the new version and also that the t-value of portal of the old version has greater distance from 1 than t-value of portal of the new version. So, the portal on the new version has a less significant effect from the portal on the old version.

	Estimate	Std. Error	t value	Pr(>|t|) *
(Intercept)	2.69936	0.05050	53.455	<2e-16
factorsver_consumers	-0.03051	0.07166	-0.426	0.67
factorsver_comp	0.65634	0.07117	9.222	<2e-16
factors	1.35364	0.10100	13.403	<2e-16

Table 9: Results from consumers companies portal

	Estimate	Std. Error	t value	Pr(>|t|) *
(Intercept)	2.69936	0.05050	53.455	<2e-16
factorsnew_ver	0.33339	0.07124	4.680	3.27e-06
factorsold_ver	1.02024	0.07159	14.252	< 2e-16
factors	-0.62582	0.10100	-6.196	8.44e-10

Table 10: Results from old and new version

6)Section for a scientific publication

In this section, website visits analysis was conducted in order to examine whether the version of the website, the portal, or combination of the two had an impact on the number of pages visited. We used the data file webvisit[1] where the dependent variable is the number of pages visited which is affected by the version and the portal as well as by their combination. We observed that the variable page visits is not a normal distribution but a Poisson distribution. For the model analysis, we created 5 models. One was the null model and the other was Model 1, just adding to Model 0 the predictor version. Next, Model 2 has as a predictor the portal. After that, Model 3 has as predictors both version and portal, Finally Model 4 has as predictors both version and portal as well as their combination. The result of the analysis, indicated that in model 4 we observed the greatest Deviance and the lowest p-value. The analysis shows a significant two-way interaction effect since Pr(>Chi) value is less than 0.05. Hence, we conduct a simple effect analysis and we can conclude that the portal on the new version has a less significant effect from the portal on the old version.

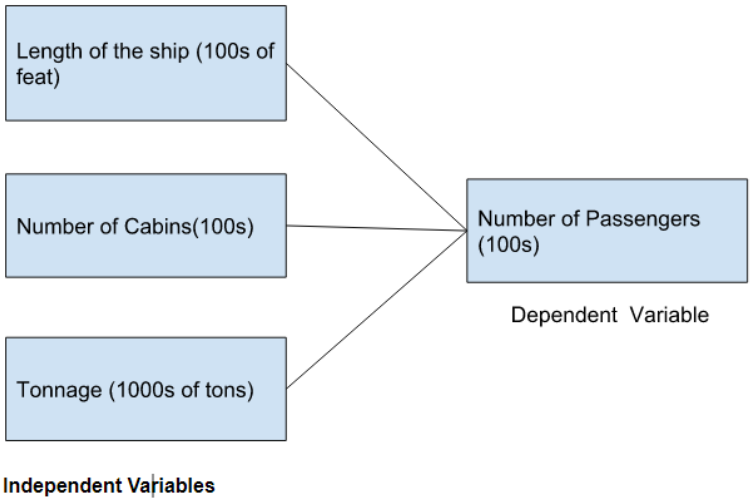
Question 3: Linear regression analysis

Setting: In order to conduct this question the following dataset obtained: Measurements for 158 cruise ships. This dataset includes measurements for 158 cruise ships. The variables that this

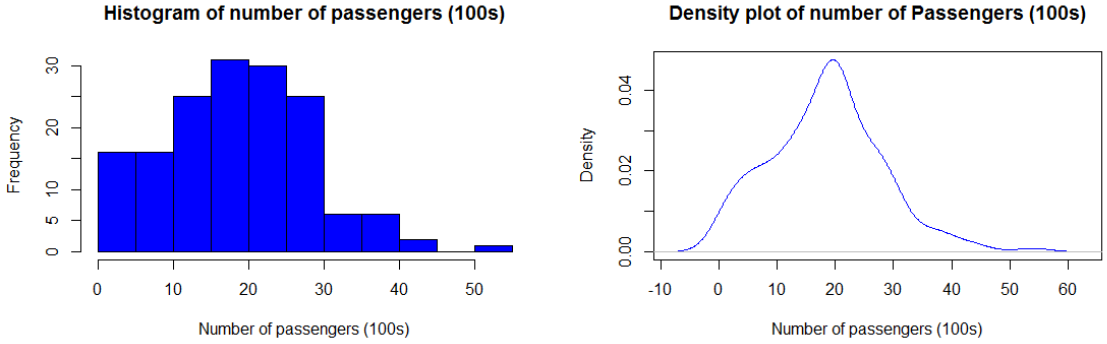
dataset has are the following: Ship Name, Cruise Line, Age (as of 2013), Tonnage (1000s of tons), Passengers (100s) ,Length (100s of feet), Cabins(100s), Passenger Density, and Crew (100s). In the model in which we conduct our analysis, the dependent variable is the number of Passengers (100s) that a ship can carry. The independent variables are the Length of the ship (100s of feat), the Tonnage (1000s of tons) of the ship, and the number of Cabins (100s) that the ship has. Other variables like the Ship Name, Cruise Line, and the Age of the ship (as of 2013) do not affect our model. Hence they are not going to be investigated in our analysis. This data set meets all the requirements as it consists of 158 measurements ($n > 100$), 3 independent variables namely (Length of the ship (100s of feat), the Tonnage (1000s of tons) of the ship, and the number of Cabins (100s) that the ship has) which are of interval (or ratio) level, a dependent variable namely (number of Passengers (100s) that a ship can carry) which is of interval (or ratio) level and which is reasonable normally distributed and the observations are independent.

1)Conceptual model

Our conceptual model of the underlying research question we will work on is presented in the below figure.

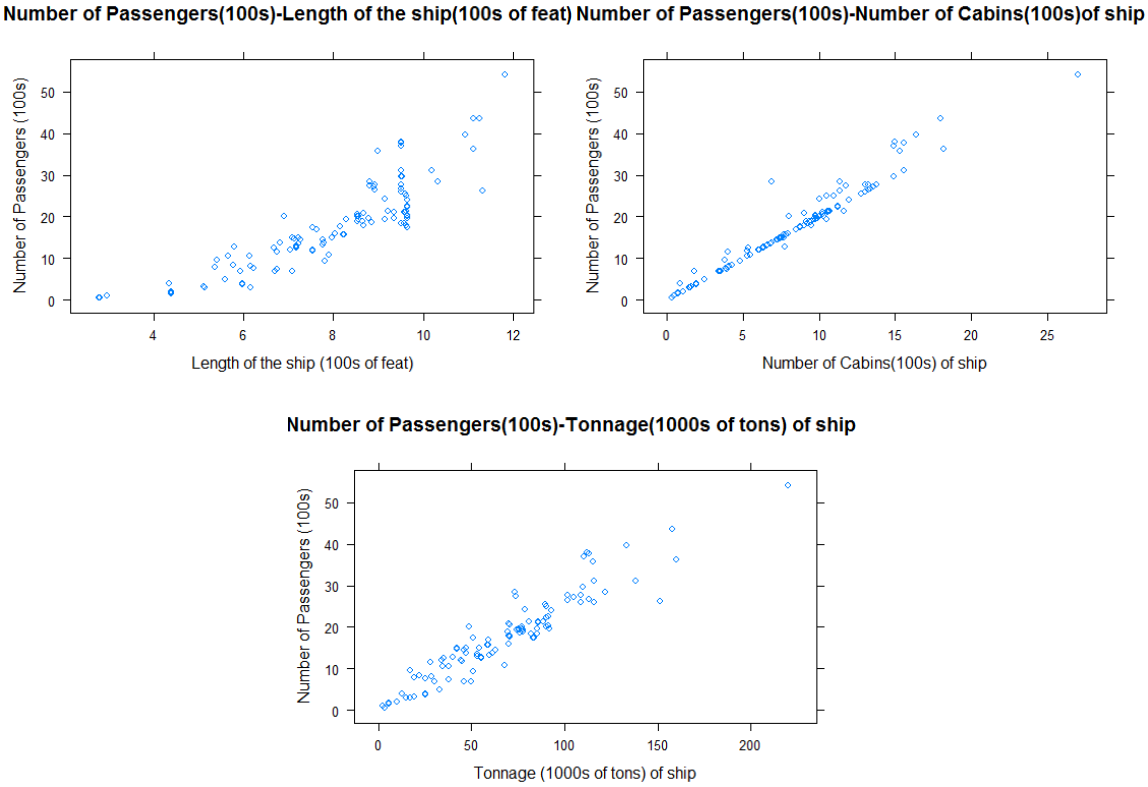


2)Graphical analysis of the distribution of the dependent variable



In the above figures the histogram of number of passengers (100s) that a ship can carry and the corresponding density plot can be depicted. It is evident from these two figures that our dependent variable, namely the number of Passengers (100s) that a ship can carry) tends to follow a normal distribution with a mean of 18.45 and a standard deviation 0f 9.67.

3)Scatter plots between dependent variable and the predictor variables



In the above figures, scatter plots between the dependent variable (number of passengers (100s) that a ship can carry) and the predictor variables (Length of the ship (100s of feat), Number of Cabins(100s) of ship and Tonnage (1000s of tons) of the ship) can be depicted. In all these plots it is evident that there is a linear relationship between our dependent variable and the three independent variables. This assumption, that the relation between independent variable(s) and dependent variable is linear, is crucial in order to be able to perform Linear regression analysis in the sequel.

4)Multiple linear regression

In order to conduct a multiple linear regression, seven linear models have been created. Model0 is the null model. Model1 is identical to model0 with the difference that the Length of the ship (100s of feat) has been added as a predictor. Model2 is identical to model0 with the difference that the Tonnage (1000s of tons) of the ship has been added as a predictor whereas in model3 the Number of Cabins(100s) has been added as a predictor to model0. In the sequel, model4 has both Length of the ship (100s of feat) and Cabins (100s) as predictors. Model5 has cabins (100s) and Tonnage (1000s of tons) as predictors, whereas model6 has Length of the ship (100s of feat) and Tonnage (1000s of tons) as predictors. Finally, model7 has all the above (Length of the ship (100s of feat), Number of Cabins(100s) of ship and Tonnage (1000s of tons) of the ship) as predictors. Now, that we created our models we want to compare them. For this reason the ANOVA function is used and the result are shown in the table 11.
It can be observed from this table that the last model, namely model7(fit) which has all three predictors, has a p-value <0.05. Hence, this model (model7) has a significant impact on the prediction. More specifically, it is a very good model for prediction. We can also conclude to this point, based on the the fact that F value of this fit model (223.2466) has a great distance from 1 and the value of Sum of Sq (921.6) has a great deviance from the mean.

After that we take a summary of the fit model and the results are shown in the table 12. Hence, the number of passengers are given by the following equation:

$$\textbf{Passengers} = -0.296 + 0.0042 * length + 1.7272 * Cabins + 0.04863 * Tonnage + 2.032$$

We found that the adjusted R squared equals 0.9559 which is really close to 1. So, we conclude that

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
0	157	14702.4				
1	156	3225.2	1	11477.2	2780.3302	< 2e-16
2	156	1571.1	0	1654.1		
3	156	687.5	0	883.6		
4	155	672.0	1	15.4	3.7399	0.05496
5	155	635.7	0	36.3		
6	155	1557.3	0	-921.6		
7	154	635.7	1	921.6	223.2466	< 2e-16

Table 11: Results from ANOVA function

our model is very good. Also, we can conclude that the Number of Cabins(100s) is the predictor with has the most influence on the outcome for three reasons. Firstly, it has a p-value which is lower than 0.05. Secondly, it has a t-value which equals 14.941 and the has the greatest distance from 1 in comparison with the other predictors. Finally, it has the biggest estimate (1.727292), which also verifies the fact that this predictor (Number of Cabins(100s)) has the most influence on the outcome.

Coefficients	Estimate	Std. Error	t.value	Pr(>|t|)
(Intercept)	-0.296000	1.214610	-0.244	0.80779
Length	0.004222	0.235752	0.018	0.98574
Cabins	1.727292	0.115604	14.941	<2e-16
Tonnage	0.048637	0.016402	2.965	0.00351

Table 12: Summary of the fit model

Next the confidence intervals (95%) are calculated where we can depict the lower and the upper limit of the confidence intervals for each of the three predictors. The results are shown in table 13. It can be drawn as a conclusion from the results that the Number of Cabins(100s) has the biggest lower and biggest upper bound of coefficients' values.

	2.5%	97.5%
(Intercept)	-2.6954489	2.10344794
Length	-0.4615031	0.46994631
Cabins	1.4989178	1.95566703
Tonnage	0.0162354	0.08103908

Table 13: Confidence Intervals

Length	0.0007823944
Cabins	0.7981160818
Tonnage	0.1871162826

Table 14: Beta values (standardized regression coefficients)

Finally, the beta values (standardized regression coefficients) are calculated. The results are shown in the table 14. It can be observed from this figure that the Number of Cabins(100s) have the greater beta value in comparison to the two other predictors (Length of the ship (100s of feat) and Tonnage (1000s of tons)). Hence, again we can conclude that the Number of Cabins(100s) have a strong influence to the number of Passengers (100s) in comparison with the two other predictors (Length of the ship (100s of feat) and Tonnage (1000s of tons)) which do not have a strong influence on our dependent variable (the number of Passengers (100s).

5)Examination of assumptions underlying linear regression

Collinearity of Variables

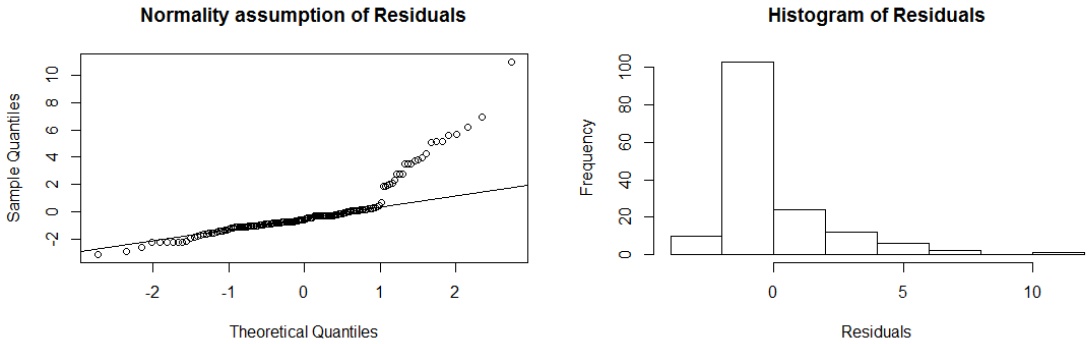
First, a collinearity analysis is being performed. The results of this analysis are shown in the

table 15. Our goal is to assess the presence of multicollinearity. For this reason we compute the Variance Inflation Factor (VIF). If the value of the Variance Inflation Factor is greater than 10 for an independent variable, then this indicates presence of multicollinearity. Hence, in our case there is a presence of multicollinearity, since the Number of Cabins(100s) and Tonnage (1000s of tons)) have a VIF value greater than 10.

Length	6.799215
Cabins	10.162402
Tonnage	14.181589

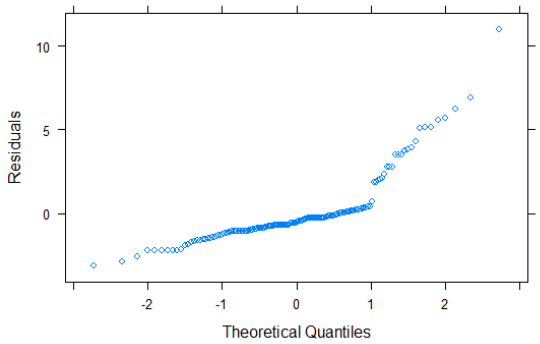
Table 15: Collinearity Analysis

Normality Analysis of Residuals



It can be depicted from the above figures that the residuals do not follow a normal distribution.

Linearity Assumption of residuals



In the above figure the linearity assumption is being considered where we depict that there is not a perfect linear relationship.

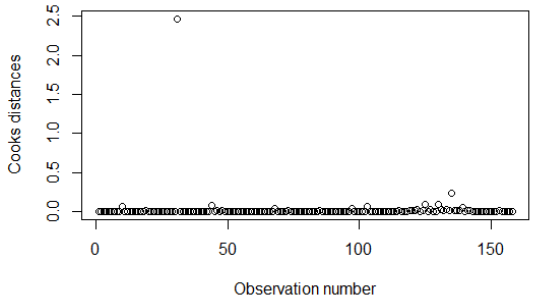
Homogeneity of variance assumption

The results of the Levene's test for the homogeneity of variance can be depicted in table 16. Our dependent variable is the number of Passenger (100s) and it is the first argument. It can be observed that the p-value is smaller than the 0.05. Hence, we can conclude that the assumption of homogeneity of variance does not hold.

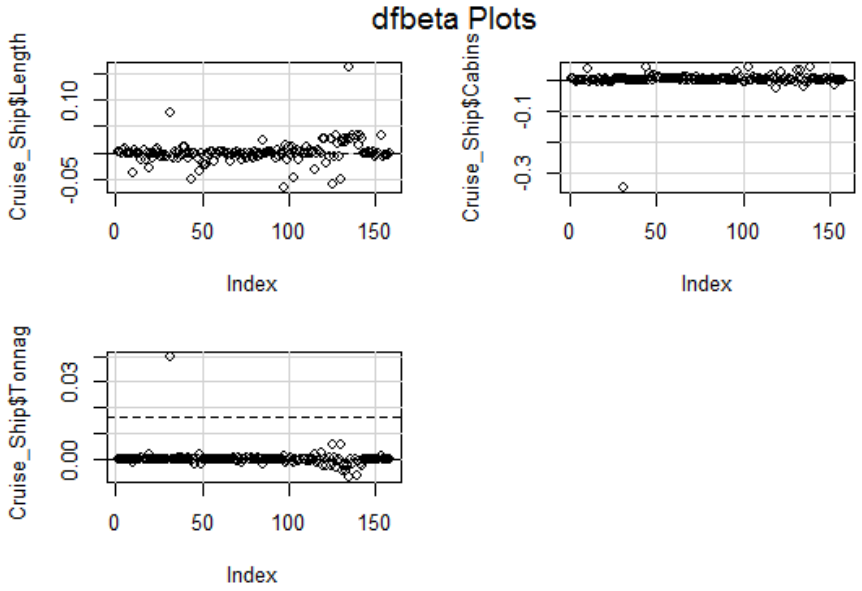
	Df	F value	Pr(>F)
group	79	1.6893	0.01071

Table 16: Levene's Test

6)Examination of the effect of single cases on the predicted values



In the above figure, the results of the formula of Cooks' distance can be depicted.We can conclude that since the observation 31 is greater than 2, it is of a high influence in comparison to other observations.



Finally, in the above figure the dfbeta plots can be depicted for each of our three independent predictors.We did it in order to find the influential points (points whose deletion has a large effect on the parameter estimates). Hence, we can conclude again that the observation 31 is of a high influence in comparison to other observations.

7)Section for a scientific publication

In the model in which we conduct our analysis, the dependent variable is the number of Passengers (100s) that a ship can carry. The independent variables are the Length of the ship (100s of feat), the Tonnage (1000s of tons) of the ship, and the number of Cabins (100s) that the ship has. Our dataset contains 158 measurements of cruise ships.

Initially, a histogram of number of passengers (100s) that a ship can carry and the corresponding density plot can be depicted. It is evident that our dependent variable, namely the number of Passengers (100s) that a ship can carry) tends to follow a normal distribution with a mean of 18.45 and a standard deviation 0f 9.67.

After that, scatter plots between the dependent variable (number of passengers (100s) that a ship can carry) and the predictor variables (Length of the ship (100s of feat), Number of Cabins(100s) of ship and Tonnage (1000s of tons) of the ship) can be depicted. In all these plots it is evident that there is a linear relationship between our dependent variable and the three independent variables.

In the sequel, a multiple linear regression is conducted. For this reason seven linear models have been created and we compare them with the ANOVA function. The last model, model7, has the (Length of the ship (100s of feat), Number of Cabins(100s) of ship and Tonnage (1000s of tons) of the ship) as predictors. It can be observed that this model, has a p-value <0.05. Hence, this model (model7) has a significant impact on the prediction. More specifically, it is a very good model for prediction. We can also conclude to this point, based on the the fact that F value of this fit model (223.2466) has a great distance from 1 and the value of Sum of Sq (921.6) has a great deviance from the mean.

In addition, we found that the adjusted R squared equals 0.9559 which is really close to 1. So, we conclude that our model is very good. Also, we can conclude that the Number of Cabins(100s) is the predictor with has the most influence on the outcome for three reasons. Firstly, it has a p-value which is lower than 0.05. Secondly, it has a t-value which equals 14.941 and the has the greatest distance from 1 in comparison with the other predictors. Finally, it has the biggest estimate (1.727292), which also verifies the fact that this predictor (Number of Cabins(100s)) has the most influence on the outcome.

Furthermore, the beta values (standardized regression coefficients) are calculated. It can be observed that the Number of Cabins(100s) have the greater beta value in comparison to the two other predictors (Length of the ship (100s of feat) and Tonnage (1000s of tons)). Hence, again we can conclude that the Number of Cabins(100s) have a strong influence to the number of Passengers (100s) in comparison with the two other predictors.

As far as the collinearity analysis is being considered, there is a presence of multicollinearity, since the Number of Cabins(100s) and Tonnage (1000s of tons)) have a VIF value greater than 10.

Also, the residuals do not follow a normal distribution here is not a perfect linear relationship.As far as the homogeneity of variance is being considered, results of the Levene's test reveal that the p-value is smaller than the 0.05. Hence, we can conclude that the assumption of homogeneity of variance does not hold.

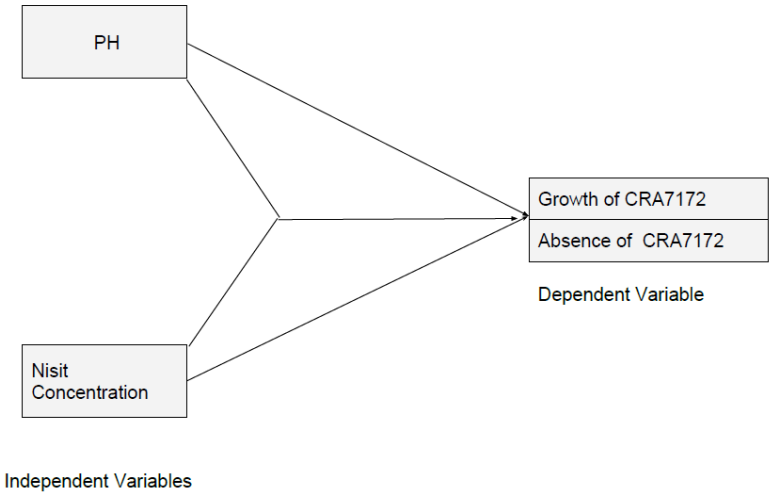
Finally, the results of the formula of Cooks' distance can be depicted which show that observation 31 is of a high influence in comparison to other observations. This can also be verified through the dfbeta plots.

Question 4: Logistic regression analysis

1)Conceptual Model

For this question we will use a apple juice dataset ($http://www.stat.ufl.edu/\ winner/data/apple_juice.dat$), where the the Presence or the Absence of Growth of CRA7152 in Apple juice is presented as a function of PH and and Nisin concentration. The presence of Growth of the CRA7152 is denoted with 1, while the absence with 0.

Below we present the conceptual model of the underlying research question we will work on.



2)Logistic Regression, Pseudo-R square, Odd ratio,Confidence interval

logistic regression model

Following logistic regression model, we will be able to determine whether and how much can each of the predictor variables contribute, in order to determine the binary value of CRA7152 Growth. Initially we implemented the NULL model (model 0) and after that , we tried to examine how the PH predictor impacts our results (model1).Apart from that, we also examined the case where both predictor variables were part of our model to identify how the combination of those two can add to the prediction of the dependent variable(model2). The results of our analyses are attached below. As we can see from the table, the addition of PH predictor variable contributes to the prediction

	RResid. Df	Resid. Dev	Df	Deviance	Pr(>Chi) *
Model 0	73	95.945			
Model 1	72	87.248	1	8.6977	0.003186
Model 2	71	64.049	1	23.1983	0.000001461

Table 17: Results between model 0, model 1 and model 2

of the dependent variable. As we can see the p_{value} is much lower than 0.05 so the null hypothesis is rejected.If we add the Nisin concentration to our model, then the p value becomes even smaller, and the deviance is increased. Thus, we can conclude that the combination of those two variables is the most useful predictor and the most suitable model is model2.

Final Model

Now that we have concluded to our model, we will demonstrate this final model which will let us identify which of the predictor variables contributes more in the performance of the whole model in predicting the DV.

Min	1Q	Median	3Q	Max
-2.3566	-0.6756	-0.2462	0.5301	1.8431

Table 18: Final model Deviance Residuals

As can be seen, the Nisin concentration, has the lowest p-value, and hence it plays an important role in the prediction of our model.

Coefficients:	Estimate	Std.Error	z value	Pr(>Chi) *
Intercept	-6.31911	2.07827	-3.041	0.002361
PH	1.64210	0.49171	3.340	0.000839
Nisit Concentration	-0.05819	0.01498	-3.884	0.000103

Table 19: Final model coefficients

Pseudo-R squared:
For model 2, with PH and Nisin concentration predictors, we now calculate the Pseudo-R squared,as requested.The table is presented below.

Pseudo R^2 for logistic regression	
Hosmer and Lemshow R^2	0.332
Cox and Snell R^2	0.35
Nagelkerke R^2	0.482

Table 20: Pseudo R^2 for logistic regression

Odd ratio and confidence interval of the predictors:
As we can notice, the odd ratio of PH is really high,almost 5.17, which means that for 1 unit increase above the average, the odds of presence of CRA7152 growth will increase significantly.

(Intercept)	PH	Nisin Con.
0.00180154	5.16602001	0.94346707

Table 21: Odd ratio of the predictor(s)

	2.5%	97.5%
(Intercept)	0.00002005301	0.07865834
PH	2.14732861143	15.24082097
Nisin Con.	0.91259180182	0.96873926

Table 22: confidence interval of the predictor(s)

3)Crosstable

In the tables below we are presenting the cross-table of our prediction model, yielding how many observations of either Presence or Absence of CRA7152 Growth are correctly predicted and how many of them are misclassified.

	CRAinAppleJuice$Growth		
CRAinAppleJuice$Growpred	0	1	Row Total
Absence	42	8	50
	0.840	0.160	0.676
Presence	6	18	24
	0.250	0.750	0.324
Column Total	48	26	74

As we can see, in a total of 74 observations, our model achieves a prediction accuracy of 84% in the Absence case, while on the Presence the accuracy of prediction is 75%.

4)Scientific publication section

In this section, logistic regression analysis was conducted in order to compare the effect of PH and Nisit concentration on the absence or presence of Alicyclobacillus Acidoterrestris(CRA7152) in Apple Juice. Our dataset consisted of 54 independent observations with the related measurements for our variables and an indication for the absence or presence of CRA7152.

For the logistic regression analysis, we created 3 models. One was the null model and the other two considered the effect of PH as a dependent variable,as well as the combination of PH and Nisit Concentration as a predictor for the value of the CRA7152 in apple juice. The result of the analysis, indicated that there was a significant effect of PH and Nisit concentration of sugar on presence of CRA 7152 , as the p-value level was lower than 0.05.

If we summarize the model described above, with both dependent variables as predictors, we will conclude that Nisit Concentration , having the lowest p-value level, plays a significant role as a main predictor of our model.In addition, after the calculation of the odd ratio of our predictors we concluded that 1 unit change of PH above the average, can result in significant increase of the odds of CRA7152 growth.

In order to test our prediction model we finally presented the crosstable indicating the true positives/negatives and the false positive/negatives of our model based on our initial dataset. The results were quite satisfying as the accuracy was between 75 and 84 per cent.

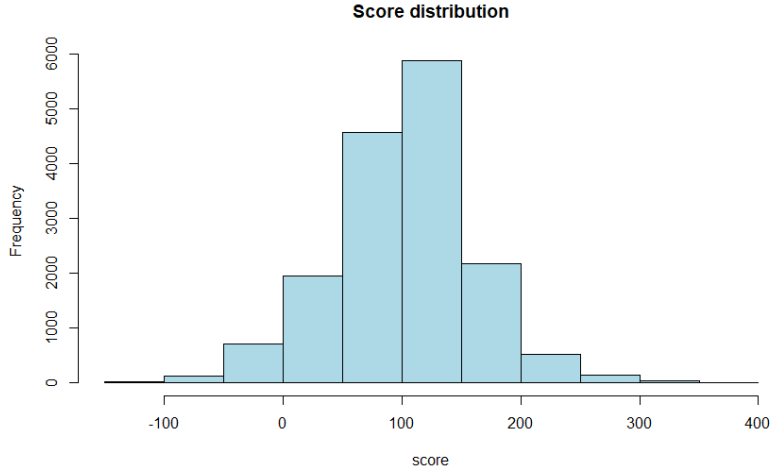
Part 3–Multilevel model

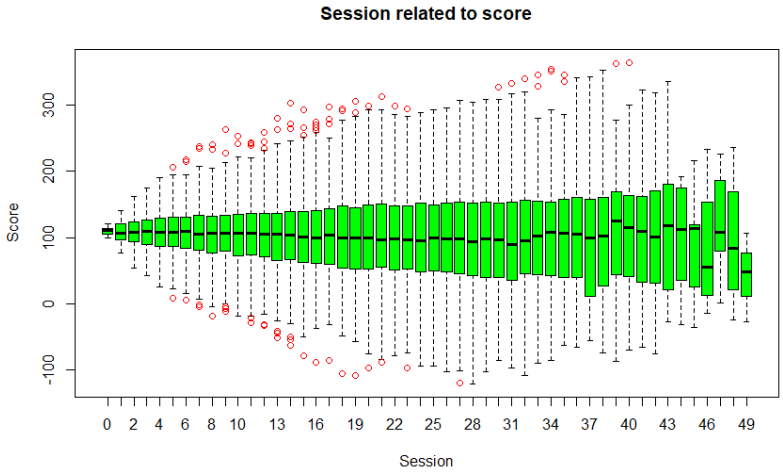
1)Graphical Analysis

Setting: The dataset includes the longitudinal data collected from a large group of participants that in multiple sessions completed a learning exercise for which exercise score was collected (set1.cvs). In the model in which we conduct our analysis, the dependent variable is the score. The independent variables are the subject and sessions.

In order to determine the distribution of the score an histogram has been implemented. The histogram confirms that the distribution in the dataset is likely normal. Histograms of normal distributions show the highest frequency in the center of the distribution.

Furthermore, in order to observe if there is significant variance between the participants in their score a boxplot has been implemented. Then we may conclude, from the boxplot figure, that for many sessions the mean remains steady at score approximately 100. Moreover, as the sessions increases the score's variance becomes higher until score 43 and after that and a small amount of fluctuations, the mean decreases.





2)

Multilevel Analysis

a)

In this section, we are going to conduct multilevel analysis. Initially a baseline model(intercept) is created for comparison reasons. After that we created a model that includes session as a fixed factor, and uses a random intercept for the participants(subject). In the table 23 the output of our final model is presented. From this table we can observe that the session has a strong influence on the score. We draw this conclusion as the p value of the fixed variable session is smaller than 0.05.

	Value	Std.Error	DF	t-value	p-value
Intercept	108.66574	2.1398997	15626	50.78076	0
session	-0.42533	0.0282743	15626	-15.04304	0

Table 23: Analysis of the model

b)

Finally, we are going to examine if there is a significant variance between the participants in their score. It can be depicted for the table 24 that the standard deviation of the subjects is 46.44301. Hence we can conclude that indeed there is a significant variance between the participants in their score.

Random effects	(Intercept)	Residual
StdDev	46.443	34.96826

Table 24: Standard Deviation Of Random Effects

95% Confidence Intervals

The 95% confidence interval of the sessionRI model that includes both the fixed and the random factors have been analyzed in the next section. We can easily observe (table 25-26) that the estimate of session in the fixed effect has significant smaller value than the value of estimate of subject in the random effect(46.4301). Hence, we can conclude that the subject has impact on the dependent variable.

Fixed effects	lower	est.	upper
Intercept	104.4715439	108.6657350	112.8599262
session	-0.4807492	-0.4253317	-0.3699142

Table 25: 95% Confidence Interval Of Fixed Effect

Random effects	lower	est	upper
sd((Intercept))	43.60687	46.443	49.4636

Table 26: 95% Confidence Interval Of Random Effect

3)Scientific publication section

In order to determine the distribution of the score an histogram has been implemented. The histogram confirms that the distribution in the dataset is likely normal as the highest frequency in the center of the distribution. Furthermore, in order to observe if there is significant variance between the participants in their score a box plot has been implemented. Then we may conclude, that for many sessions the mean remains steady at score approximately 100. After that multilevel analysis has been conducted. We can conclude that the session has a strong influence on the score. We draw this conclusion as the p value of the fixed variable session is smaller than 0.05. Also, we found that the standard deviation of the subjects is 46.44301. Hence we can conclude that indeed there is a significant variance between the participants in their score. Finally, the 95% confidence interval of the sessionRI model that includes both the fixed and the random factors has been analyzed. We can easily observe that the estimate of session in the fixed effect has significant smaller value than the value of estimate of subject in the random effect(46.4301). So, the subject has impact on the dependent variable.

References

[1] *Kerr, J. S. and Hindmarch, I. (1998), The effects of alcohol alone or in combination with other drugs on information processing, task performance and subjective responses. Hum. Psychopharmacol. Clin. Exp., 13: 1–9. doi:10.1002/(SICI)1099-1077(199801)13:1<1::AID-HUP939>3.0.CO;2-0*

[2] *I. HINDMARCH, J. S. KERR, N. SHERWOOD; The effects of alcohol and other drugs on psychomotor performance and cognitive function, Alcohol and Alcoholism, Volume 26, Issue 1, 1 January 1991, Pages 71–79, https://doi.org/10.1093/oxfordjournals.alcalc.a045085.*

[3] *Maylor, E. A. and Rabbitt, P. M. A. (1993), Alcohol, reaction time and memory: A meta-analysis. British Journal of Psychology, 84: 301–317. doi:10.1111/j.2044-8295.1993.tb02485.x*

```
            # CS4125 Seminar Research Methodology for Data Science
# Coursework assignment A - Part 2, Question 1 - Twitter sentiment analysis
# 2017
#
# This code requires the following file:
# sentiment3.R, negative-words.txt, and positive-words.txt.
#
#
# this is based on youtube https://youtu.be/adIvt_luO1o
# also see
    https://silviaplanella.wordpress.com/2014/12/31/sentiment-analysis-twitter-and-r/
##############################################################################################


setwd("C:\\Users\\pc1\\Desktop\\Seminar")
# apple , note use / instead of \, which used by windows



#install.packages("twitteR", dependencies = TRUE)
library(twitteR)
#install.packages("RCurl", dependencies = T)
library(RCurl)

#install.packages("bitops", dependencies = T)
library(bitops)
#install.packages("plyr", dependencies = T)
library(plyr)
#install.packages('stringr', dependencies = T)
library(stringr)
#install.packages("NLP", dependencies = T)
library(NLP)
library(openssl)
library(httpuv)
#install.packages("tm", dependencies = T)
library(tm)

#install.packages("wordcloud", dependencies=T)
#install.packages("RColorBrewer", dependencies=TRUE)
library(RColorBrewer)
library(wordcloud)
#install.packages("reshape", dependencies=T)
library(reshape)

################### functions


clearTweets <- function(tweets, excl) {

  tweets.text <- sapply(tweets, function(t)t$getText()) #get text out of tweets


  tweets.text = gsub('[[:cntrl:]]', '', tweets.text)
  tweets.text = gsub('\\d+', '', tweets.text)
  tweets.text <- str_replace_all(tweets.text,"[^[:graph:]]", " ") #remove graphic


  corpus <- Corpus(VectorSource(tweets.text))

  corpus_clean <- tm_map(corpus, removePunctuation)
  corpus_clean <- tm_map(corpus_clean, content_transformer(tolower))
  corpus_clean <- tm_map(corpus_clean, removeWords, stopwords("english"))
```

```
  corpus_clean <- tm_map(corpus_clean, removeNumbers)
  corpus_clean <- tm_map(corpus_clean, stripWhitespace)
  corpus_clean <- tm_map(corpus_clean, removeWords, c(excl,"http","https","httpst"))


  return(corpus_clean)
}


## capture all the output to a file.

sink("output.txt")

################## Collect from Twitter

# for creating a twitter app (apps.twitter.com) see youtube https://youtu.be/lT4Kosc_ers
#consumer_key <-'your key'
#consumer_scret <- 'your secret'
#access_token <- 'your access token'
#access_scret <- 'your access scret'

source("your_twitter.R") #this file will set my personal variables for my twitter app,
    adjust the name of this file. use the provide template your_twitter.R

setup_twitter_oauth(consumer_key,consumer_scret, access_token,access_scret) #connect to
    twitter app


##### This example uses the following 3 celebrities: Donald Trump, Hillary Clinton, and
    Bernie Sanders
##  You should replace this with your own celebrities, at least 3, but more preferred
##  Note that it will take the computer some to collect the tweets

tweets_T <- searchTwitter("#trump", n=800, lang="en", resultType="recent") #1000 recent
    tweets about Donald Trump, in English (I think that 1500 tweets is max)
tweets_C <- searchTwitter("#BarackObama", n=800, lang="en", resultType="recent") #1000
    recent tweets about Hillary Clinton
tweets_B <- searchTwitter("#hillary", n=800, lang="en", resultType="recent") #1000 recent
    tweets about Bernie Sanders



######################### WordCloud
### This not requires in the assignment, but still fun to do

# based on https://youtu.be/JoArGkOpeU0
corpus_T<-clearTweets(tweets_T,
    c("trump","amp","realdonaldtrump","trumptrain","donald","trumps","alwaystrump"))
    #remove also some campain slogans
wordcloud(corpus_T, max.words=50)

corpus_C<-clearTweets(tweets_C, c("Obama","BarackObama","obamalegacy","44thPresident"))
wordcloud(corpus_C, max.words=50)

corpus_B<-clearTweets(tweets_B, c("hillary","amp","clinton","hillarys"))
wordcloud(corpus_B, max.words=50)
##############################
```

```
######################### Sentiment analysis

tweets_T.text <- laply(tweets_T, function(t)t$getText()) #get text out of tweets
tweets_C.text <- laply(tweets_C, function(t)t$getText()) #get text out of tweets
tweets_B.text <- laply(tweets_B, function(t)t$getText()) #get text out of tweets



#taken from https://github.com/mjhea0/twitter-sentiment-analysis
pos <- scan('positive-words.txt', what = 'character', comment.char=';') #read the
    positive words
neg <- scan('negative-words.txt', what = 'character', comment.char=';') #read the
    negative words

source("sentiment3.R") #load algoritm
# see sentiment3.R form more information about sentiment analysis. It assigns a intereger
    score
# by substracitng the number of occurrence of negative words from that of positive words

analysis_T <- score.sentiment(tweets_T.text, pos, neg)
analysis_C <- score.sentiment(tweets_C.text, pos, neg)
analysis_B <- score.sentiment(tweets_B.text, pos, neg)


sem<-data.frame(analysis_T$score, analysis_C$score, analysis_B$score)


semFrame <-melt(sem, measured=c(analysis_T.score,analysis_C.score, analysis_B.score ))
names(semFrame) <- c("Candidate", "score")
semFrame$Candidate <-factor(semFrame$Candidate, labels=c("Donald Trump", "Barack Obama",
    "Hillary Clinton")) # change the labels for your celibrities

################### Below insert your own code to answer question 1. The data you need can
    be found in semFrame
#Question1.2 Homogeneity of Variance
install.packages('lawstat')
library(car)
leveneTest(semFrame$score, semFrame$Candidate, center = median)
bartlett.test(semFrame$score, semFrame$Candidate, center=median)

#leveneTest(semFrame$score, semFrame$Candidate=="Donald Trump", center=median)
#leveneTest(semFrame$score, semFrame$Candidate=="Barack Obama", center=median)
#leveneTest(semFrame$score, semFrame$Candidate=="Hillary Clinton", center=median)



#bartlett.test(semFrame$score, semFrame$Candidate=="Donald Trump", center=median)
#bartlett.test(semFrame$score, semFrame$Candidate=="Barack Obama", center=median)
#bartlett.test(semFrame$score, semFrame$Candidate=="Hillary Clinton", center=median)


#Question1.3

newdataTrump <- subset(semFrame$score, semFrame$Candidate=="Donald Trump")
newdataObama <- subset(semFrame$score, semFrame$Candidate=="Barack Obama")
newdataClinton <- subset(semFrame$score, semFrame$Candidate=="Hillary Clinton")

# plot the results histogram
hist(newdataTrump)
hist(newdataObama)
hist(newdataClinton)
```

```
# plot the results density plot
d <- density(newdataTrump) # returns the density data
plot(d , main="Tweets sentiments for Trump")
d <- density(newdataObama) # returns the density data
plot(d,main="Tweets sentiments for Obama")
d <- density(newdataClinton) # returns the density data
plot(d,main="Tweets sentiments for Hillary")

#Question1.4

meanScoreCancidate= tapply(semFrame$score, semFrame$Candidate, mean)

barplot((meanScoreCancidate),
       main="Mean Sentiment of Tweets for Celebrities",
       xlab="Celebrities",
       ylab="MeanScore",
       border="black",
       col="darkred",
       density=10)


hist(newdataTrump,prob=T )

m<-mean(newdataTrump);std<-sqrt(var(newdataTrump))
hist(newdataTrump,prob=T )
curve(dnorm(x, mean=m, sd=std),lwd=2, add=TRUE)


hist(newdataObama,prob=T )

m<-mean(newdataObama);std<-sqrt(var(newdataObama))
hist(newdataObama,prob=T )
curve(dnorm(x, mean=m, sd=std),lwd=2, add=TRUE)


hist(newdataClinton,prob=T )

m<-mean(newdataClinton);std<-sqrt(var(newdataClinton))
hist(newdataClinton,prob=T )
curve(dnorm(x, mean=m, sd=std),lwd=2, add=TRUE)

boxplot(semFrame$score~semFrame$Candidate)


#boxplot(semFrame$score~semFrame$Candidate=="Donald Trump")
#boxplot(semFrame$score~semFrame$Candidate=="Barack Obama")
#boxplot(semFrame$score~semFrame$Candidate=="Hillary Clinton")

#Question1.5
#model without predictor
model0<- lm(score ~ 1, data = semFrame)
#model with Candidate predictor
model1 <- lm(score ~ Candidate, data = semFrame)
anova(model0,model1, test = "F")

summary(model1)

#Question1.6
pairwise.t.test(semFrame$score, semFrame$Candidate, paired = FALSE, p.adjust.method =
    "bonferroni")

######### stop redireting output.
```

```
####################Question 2 Website Visits(between Groups- Two
    Factors)#########################################################################################################
#Question2.2
library(Rcmdr)
library(foreign)
library(ggplot2)
#read from the webvisit0.csv
WebVisit<-read.csv("webvisit1.csv", header = TRUE)

#install.packages("sm")
library(sm)

hist(WebVisit$pages, main="Page Visit Frequency", xlab ='PagesVisited')



hist(WebVisit$pages,prob=T )

m<-mean(WebVisit$pages);std<-sqrt(var(WebVisit$pages))
hist(WebVisit$pages,prob=T )
curve(dnorm(x, mean=m, sd=std),lwd=2, add=TRUE)



counter <- table(WebVisit$portal, WebVisit$pages)
barplot(counter, main="Portal Type-PagesVisited",
       xlab="PagesVisited", col=c("blue","orange"),
       legend = rownames(counter))


counter <- table(WebVisit$version, WebVisit$pages)
barplot(counter, main="Version Type-PagesVisited",
       xlab="PagesVisited", col=c("purple","darkgreen"),
       legend = rownames(counter))

#Question2.3
hist(WebVisit$pages, main="Histogram of Pages Visited",xlab = "Pages Visited", col="blue")

#Question2.4
#adding two factors version and portal
WebVisit$version <- factor(WebVisit$version)
WebVisit$portal <- factor(WebVisit$portal)


model0 = glm(WebVisit$pages~1, data = WebVisit, family = "poisson")
model1 = glm(WebVisit$pages~version, data = WebVisit, family = "poisson")
model2 = glm(WebVisit$pages~portal, data = WebVisit, family = "poisson")
model3 = glm(WebVisit$pages~version+portal, data = WebVisit, family = "poisson")
model4 = glm(WebVisit$pages~version+portal+version:portal, data = WebVisit, family =
    "poisson")


anova(model0,model1, test="Chisq")
anova(model0,model2, test="Chisq")
anova(model0,model3, test="Chisq")
anova(model0,model4, test="Chisq")
summary(model4)
anova(model4,test="Chisq")



#Question2.5
```

```
########version on consumers and companies portal
factors <- interaction(WebVisit$version, WebVisit$portal)
levels(factors)
ver_consumers <-c(1,-1,0 , 0) #test version with consumers portal
ver_comp <-c(0, 0, 1, -1) #test version with companies portal
SimpleEff <- cbind(ver_consumers,ver_comp)
SimpleEff
contrasts(factors) <- SimpleEff
contrasts(factors)
simpleEffectModel <-lm(WebVisit$pages ~ factors , data = WebVisit, na.action = na.exclude)
summary.lm(simpleEffectModel)


########effect of portal in the old and mew version
factors <- interaction(WebVisit$version, WebVisit$portal)
levels(factors)
new_ver <-c(0,-1,0 , 1) #portal new version
old_ver <-c(-1, 0, 1, 0) #portal old version
SimpleEff <- cbind(new_ver,old_ver)
SimpleEff
contrasts(factors) <- SimpleEff
contrasts(factors)
simpleEffectModel <-lm(WebVisit$pages ~ factors , data = WebVisit, na.action = na.exclude)
summary.lm(simpleEffectModel)



#Part 2: Question 3: Linear regression
    analysis#########################################################################################
#######################################


#load libraries
library(Rcmdr)
library(foreign)
library(ggplot2)
library(lattice)
library(QuantPsyc)
library(car)


#Measurements for 158 cruise ships
colnames = c("Ship_Name", "Cruise_Line", "Age",
             "Tonnage", "Passengers", "Length",
             "Cabins","Passengers_Density","Crew")

#Import the dataset
Cruise_Ship <- read.fwf(
  file=url("http://www.stat.ufl.edu/~winner/data/cruise_ship.dat"),
  widths=c(20, 20, 9, 8, 8, 8, 8, 9, 9 ),
  col.names = colnames)


#2)Graphical analysis of the distribution of the dependent variable


#Histogram of number of Passengers (100s)
hist(Cruise_Ship$Passengers, main="Histogram of number of passengers (100s)", col =
    "blue",xlab = "Number of passengers (100s)")

#Density plot of number of Passengers (100s)
```

```
density_passengers <-density(Cruise_Ship$Passengers)
plot(density_passengers , main = "Density plot of number of Passengers (100s)", col =
    "blue", xlab = "Number of passengers (100s)")

#computation of the mean and the standar deviation of the number of Passengers (100s)
mean(Cruise_Ship$Passengers)
sd(Cruise_Ship$Passengers)


#3)Scatter plots between dependent variable and the predictor variables

#Scatter plot between Number of Passengers(100s)-Length of the ship(100s of feat)
xyplot(Cruise_Ship$Passengers ~ Cruise_Ship$Length, data = Cruise_Ship,
      xlab = "Length of the ship (100s of feat)",
      ylab = "Number of Passengers (100s)",
      main = "Number of Passengers(100s)-Length of the ship(100s of feat)"
)

#Scatter plot between Number of Passengers(100s)-Number of Cabins(100s) of ship
xyplot(Cruise_Ship$Passengers ~ Cruise_Ship$Cabins, data = Cruise_Ship,
      xlab = "Number of Cabins(100s) of ship",
      ylab = "Number of Passengers (100s)",
      main = "Number of Passengers(100s)-Number of Cabins(100s)of ship"
)


#Scatter plot between Number of Passengers(100s)-Tonnage (1000s of tons) of ship
xyplot(Cruise_Ship$Passengers ~ Cruise_Ship$Tonnage, data = Cruise_Ship,
      xlab = "Tonnage (1000s of tons) of ship",
      ylab = "Number of Passengers (100s)",
      main = "Number of Passengers(100s)-Tonnage(1000s of tons) of ship"
)


#4)Multiple linear regression



#creation of 7 models
model0<-lm(Cruise_Ship$Passengers~1, data=Cruise_Ship)
model1<-lm(Cruise_Ship$Passengers~Cruise_Ship$Length, data=Cruise_Ship)
model2<-lm(Cruise_Ship$Passengers~Cruise_Ship$Tonnage, data=Cruise_Ship)
model3<-lm(Cruise_Ship$Passengers~Cruise_Ship$Cabins, data=Cruise_Ship)
model4<-lm(Cruise_Ship$Passengers~Cruise_Ship$Length+Cruise_Ship$Cabins, data=Cruise_Ship)
model5<-lm(Cruise_Ship$Passengers~Cruise_Ship$Cabins+Cruise_Ship$Tonnage,
    data=Cruise_Ship)
model6<-lm(Cruise_Ship$Passengers~Cruise_Ship$Length+Cruise_Ship$Tonnage,
    data=Cruise_Ship)

#creation of model7 that has all the three above predictors length, cabins and tonnage

fit<-lm( Cruise_Ship$Passengers ~ Cruise_Ship$Length + Cruise_Ship$Cabins +
    Cruise_Ship$Tonnage, data = Cruise_Ship)

#comparison of the models using ANOVA function
anova(model0,model1,model2,model3,model4,model5,model6,fit)

#summary of the fit model
summary(fit)

#Determination of the confidence intervals (95%)
confint(fit)
```

```
#Computation of the the beta values (standardised regression coefficients):
lm.beta(fit)

#5)Examination of assumptions underlying linear regression



#checking the collinearity assumption

vif(fit)

#checking the normality assumption of residuals
shipres = residuals(fit)#
qqnorm(shipres, main="Normality assumption of Residuals")
qqline(shipres)
hist(shipres,main="Histogram of Residuals", xlab = "Residuals")



#checking the linearty assumption
qqmath( ~ resid(fit),
       xlab = "Theoretical Quantiles",
       ylab = "Residuals"
)

#checking the homogeneity of variance with levene test


leveneTest(Cruise_Ship$Passengers~ factor(Cruise_Ship$Length)*factor(Cruise_Ship$Cabins),
    center = mean)
leveneTest(Cruise_Ship$Passengers, factor(Cruise_Ship$Cabins), center = mean)
leveneTest(Cruise_Ship$Passengers, factor(Cruise_Ship$Length), center = mean)
leveneTest(Cruise_Ship$Passengers, factor(Cruise_Ship$Tonnage), center = mean)

#6)Examination of the effect of single cases on the predicted values

#application of Cooks' distance
plot(cooks.distance(fit),ylab="Cooks distances",xlab="Observation number")

#application of DFBeta
dfbetaPlots(fit)


#####################################################Question 4 Logistic
    Regression########################################
library(foreign)
library(car)  #Package includes Levene's test
library(tidyr) # for wide to long format transformation of the data
library(ggplot2)
library(QuantPsyc) #include lm.beta()
library(class)
library(pscl)
library(gmodels)

columns = c("PH","NisinConcentration","Temoerature","BrixConcetration",
           "Growth"
)

#reading the dataset
CRAinAppleJuice <- read.fwf(
  file=url("http://www.stat.ufl.edu/~winner/data/apple_juice.dat"),
  widths=c(9, 9, 9, 8, 8),
  col.names = columns)
```

```
model0 <- glm(CRAinAppleJuice$Growth ~ 1, data=CRAinAppleJuice, family=binomial())
model1 <- glm(CRAinAppleJuice$Growth ~ PH, data=CRAinAppleJuice, family=binomial())
#model2 <- glm(CRAinAppleJuice$Growth ~NisinConcentration , data=CRAinAppleJuice,
    family=binomial())
model2<-glm(CRAinAppleJuice$Growth ~ PH + NisinConcentration, data=CRAinAppleJuice,
    family=binomial())
anova(model0,model1,model2,test = "Chisq")

summary(model2)

##Pseudo R2
logisticPseudoR2s <- function(LogModel)
  #taken from Andy Fields et al. book on R, p.334
{
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <- 1 - dev / nullDev
  R.cs <- 1 - exp(-(nullDev - dev) / modelN)
  R.n <- R.cs / (1 - (exp(-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemshow R^2: ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2:    ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2:       ", round(R.n, 3), "\n")
}
logisticPseudoR2s(model2)
##Odd and conf
exp(model2$coefficients)
exp(confint(model2))
##crosstable
CRAinAppleJuice$Growpred[fitted(model2) <=0.5] <- 0
CRAinAppleJuice$Growpred[fitted(model2) > 0.5] <- 1
CRAinAppleJuice$Growpred<-factor(CRAinAppleJuice$Growpred, levels = c(0:1), labels =
    c("Absence","Presence"))
table(CRAinAppleJuice$Growth, CRAinAppleJuice$Growpred)
CrossTable(CRAinAppleJuice$Growpred, CRAinAppleJuice$Growth, prop.c=FALSE, prop.t=FALSE,
    prop.chisq=FALSE, fisher=FALSE, chisq=FALSE, expected = FALSE)


############################################################Part
    3####################################
#############################################################################################################
library(MASS)
library(foreign)
library(car)
library(ggplot2)
library(nlme)
library(reshape)
library(graphics)
require(lattice)
library(lattice)
require(Matrix)
library(lme4)

#question 1#
set1 <- read.csv2("set1.csv", header = TRUE, sep = ",",
                  dec = ",", fill = TRUE, comment.char = "")
hist(set1$score,col="lightblue",main="Score distribution",xlab="score")
boxplot(set1$score ~ set1$session , xlab="Session", ylab="Score", main="Session related
    to score",col="green",outcol="red")

#question 2
##We present thr intercdept only model which violates the independent assumption. It will
```

```
    only allow us to comparre its results with multilevel analysis.
interceptOnly <- gls(score ~ 1, data = set1, method = "ML")
summary(interceptOnly)

randomInterceptOnly <- lme(score ~ 1, data = set1, random = ~1|Subject, method = "ML")
summary(randomInterceptOnly)

#adding session to our model
sessionMod<-lme(score ~(1+ session), data = set1, random = ~1|Subject, method = "ML")
summary(sessionMod)

##compute confidence intervals
intervals(sessionMod, 0.95)

sink()
```

```
          ####################Question 2 Website Visits(between Groups- Two
              Factors)#########################################################################################
> #Question2.2
> library(Rcmdr)
Loading required package: splines
Loading required package: RcmdrMisc
Loading required package: car
Loading required package: sandwich
Loading required package: effects
Loading required package: carData




    Guyer, UN, Vocab

lattice theme set by effectsTheme()
See ?effectsTheme for details.
RcmdrMsg: [1] NOTE: R Commander Version 2.4-1: Mon Mar 19 16:12:59 2018

Rcmdr Version 2.4-1

> library(foreign)
> library(ggplot2)

Attaching package: 'ggplot2'

The following object is masked from 'package:NLP':

    annotate

> #read from the webvisit0.csv
> WebVisit<-read.csv("webvisit1.csv", header = TRUE)
> #install.packages("sm")
> library(sm)
Package 'sm', version 2.2-5.4: type help(sm) for summary information
> hist(WebVisit$pages, main="Page Visit Frequency", xlab ='PagesVisited')
> hist(WebVisit$pages,prob=T )
> m<-mean(WebVisit$pages);std<-sqrt(var(WebVisit$pages))
> hist(WebVisit$pages,prob=T )
> curve(dnorm(x, mean=m, sd=std),lwd=2, add=TRUE)
> counter <- table(WebVisit$portal, WebVisit$pages)
> barplot(counter, main="Portal Type-PagesVisited",
+         xlab="PagesVisited", col=c("blue","orange"),
+         legend = rownames(counter))
> counter <- table(WebVisit$version, WebVisit$pages)
> barplot(counter, main="Version Type-PagesVisited",
+         xlab="PagesVisited", col=c("purple","darkgreen"),
+         legend = rownames(counter))
> #Question2.3
> hist(WebVisit$pages, main="Histogram of Pages Visited",xlab = "Pages Visited",
    col="blue")
> #Question2.4
> #adding two factors version and portal
> WebVisit$version <- factor(WebVisit$version)
> WebVisit$portal <- factor(WebVisit$portal)
> model0 = glm(WebVisit$pages~1, data = WebVisit, family = "poisson")
> model1 = glm(WebVisit$pages~version, data = WebVisit, family = "poisson")
> model2 = glm(WebVisit$pages~portal, data = WebVisit, family = "poisson")
> model3 = glm(WebVisit$pages~version+portal, data = WebVisit, family = "poisson")
> model4 = glm(WebVisit$pages~version+portal+version:portal, data = WebVisit, family =
```

```
    "poisson")
> anova(model0,model1, test="Chisq")
Analysis of Deviance Table

Model 1: WebVisit$pages ~ 1
Model 2: WebVisit$pages ~ version
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       998     1067.0
2       997     1032.8 1   34.249 0.00000000485 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(model0,model2, test="Chisq")
Analysis of Deviance Table

Model 1: WebVisit$pages ~ 1
Model 2: WebVisit$pages ~ portal
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       998    1067.00
2       997     898.85 1   168.16 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(model0,model3, test="Chisq")
Analysis of Deviance Table

Model 1: WebVisit$pages ~ 1
Model 2: WebVisit$pages ~ version + portal
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       998    1067.00
2       996     861.98 2   205.02 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(model0,model4, test="Chisq")
Analysis of Deviance Table

Model 1: WebVisit$pages ~ 1
Model 2: WebVisit$pages ~ version + portal + version:portal
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       998    1067.00
2       995     833.97 3   233.03 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model4)

Call:
glm(formula = WebVisit$pages ~ version + portal + version:portal,
    family = "poisson", data = WebVisit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8100  -0.7783  -0.4582   0.4642   5.5300

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.68916    0.04472  15.410    < 2e-16 ***
version[T.1]           0.03018    0.06315   0.478      0.633
portal[T.1]            0.70524    0.05485  12.859    < 2e-16 ***
version[T.1]:portal[T.1] -0.42399  0.08017  -5.289 0.000000123 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1067.00 on 998 degrees of freedom
```

```
Residual deviance: 833.97 on 995 degrees of freedom
AIC: 3553.5

Number of Fisher Scoring iterations: 5

> anova(model4,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: WebVisit$pages

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          998    1067.00
version       1   34.249      997    1032.75 0.00000000485 ***
portal        1  170.773      996     861.98    < 2.2e-16 ***
version:portal 1  28.013      995     833.97 0.00000012050 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> ########version on consumers and companies portal
> factors <- interaction(WebVisit$version, WebVisit$portal)
> levels(factors)
[1] "0.0" "1.0" "0.1" "1.1"
> ver_consumers <-c(1,-1,0 , 0) #test version with consumers portal
> ver_comp <-c(0, 0, 1, -1) #test version with companies portal
> SimpleEff <- cbind(ver_consumers,ver_comp)
> SimpleEff
     ver_consumers ver_comp
[1,]             1        0
[2,]            -1        0
[3,]             0        1
[4,]             0       -1
> contrasts(factors) <- SimpleEff
> contrasts(factors)
    ver_consumers ver_comp
0.0             1        0 -0.5
1.0            -1        0 -0.5
0.1             0        1  0.5
1.1             0       -1  0.5
> simpleEffectModel <-lm(WebVisit$pages ~ factors , data = WebVisit, na.action =
    na.exclude)
> summary.lm(simpleEffectModel)

Call:
lm(formula = WebVisit$pages ~ factors, data = WebVisit, na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0325 -1.0325 -0.7198 0.9469 12.0080

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.69936    0.05050 53.455  <2e-16 ***
factorsver_consumers -0.03051 0.07166 -0.426    0.67
factorsver_comp     0.65634    0.07117  9.222  <2e-16 ***
factors             1.35364    0.10100 13.403  <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.596 on 995 degrees of freedom
```

```
Multiple R-squared: 0.2079, Adjusted R-squared: 0.2056
F-statistic: 87.08 on 3 and 995 DF, p-value: < 2.2e-16

> ########effect of portal in the old and mew version
> factors <- interaction(WebVisit$version, WebVisit$portal)
> levels(factors)
[1] "0.0" "1.0" "0.1" "1.1"
> new_ver <-c(0,-1,0 , 1) #portal new version
> old_ver <-c(-1, 0, 1, 0) #portal old version
> SimpleEff <- cbind(new_ver,old_ver)
> SimpleEff
     new_ver old_ver
[1,]       0      -1
[2,]      -1       0
[3,]       0       1
[4,]       1       0
> contrasts(factors) <- SimpleEff
> contrasts(factors)
    new_ver old_ver
0.0       0      -1 -0.5
1.0      -1       0  0.5
0.1       0       1 -0.5
1.1       1       0  0.5
> simpleEffectModel <-lm(WebVisit$pages ~ factors , data = WebVisit, na.action =
     na.exclude)
> summary.lm(simpleEffectModel)

Call:
lm(formula = WebVisit$pages ~ factors, data = WebVisit, na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0325 -1.0325 -0.7198 0.9469 12.0080

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.69936    0.05050 53.455 < 2e-16 ***
factorsnew_ver 0.33339    0.07124  4.680 3.27e-06 ***
factorsold_ver 1.02024    0.07159 14.252 < 2e-16 ***
factors       -0.62582    0.10100 -6.196 8.44e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.596 on 995 degrees of freedom
Multiple R-squared: 0.2079, Adjusted R-squared: 0.2056
F-statistic: 87.08 on 3 and 995 DF, p-value: < 2.2e-16

> #load libraries
> library(Rcmdr)
> library(foreign)
> library(ggplot2)
> library(lattice)
> library(QuantPsyc)
Loading required package: boot

Attaching package: 'boot'

The following object is masked from 'package:lattice':

    melanoma

The following object is masked from 'package:sm':
```

```
    dogs

The following object is masked from 'package:car':

    logit

Loading required package: MASS

Attaching package: 'MASS'

The following object is masked from 'package:sm':

    muscle


Attaching package: 'QuantPsyc'

The following object is masked from 'package:base':

    norm

> library(car)
> #Measurements for 158 cruise ships
> colnames = c("Ship_Name", "Cruise_Line", "Age",
+              "Tonnage", "Passengers", "Length",
+              "Cabins","Passengers_Density","Crew")
> #Import the dataset
> Cruise_Ship <- read.fwf(
+   file=url("http://www.stat.ufl.edu/~winner/data/cruise_ship.dat"),
+   widths=c(20, 20, 9, 8, 8, 8, 8, 9, 9 ),
+   col.names = colnames)
> #Histogram of number of Passengers (100s)
> hist(Cruise_Ship$Passengers, main="Histogram of number of passengers (100s)", col =
    "blue",xlab = "Number of passengers (100s)")
> #Density plot of number of Passengers (100s)
> density_passengers <-density(Cruise_Ship$Passengers)
> plot(density_passengers , main = "Density plot of number of Passengers (100s)", col =
    "blue", xlab = "Number of passengers (100s)")
> #computation of the mean and the standar deviation of the number of Passengers (100s)
> mean(Cruise_Ship$Passengers)
[1] 18.45741
> sd(Cruise_Ship$Passengers)
[1] 9.677095
> #Scatter plot between Number of Passengers(100s)-Length of the ship(100s of feat)
> xyplot(Cruise_Ship$Passengers ~ Cruise_Ship$Length, data = Cruise_Ship,
+       xlab = "Length of the ship (100s of feat)",
+       ylab = "Number of Passengers (100s)",
+       main = "Number of Passengers(100s)-Length of the ship(100s of feat)"
+ )
> #Scatter plot between Number of Passengers(100s)-Number of Cabins(100s) of ship
> xyplot(Cruise_Ship$Passengers ~ Cruise_Ship$Cabins, data = Cruise_Ship,
+       xlab = "Number of Cabins(100s) of ship",
+       ylab = "Number of Passengers (100s)",
+       main = "Number of Passengers(100s)-Number of Cabins(100s)of ship"
+ )
> #Scatter plot between Number of Passengers(100s)-Tonnage (1000s of tons) of ship
> xyplot(Cruise_Ship$Passengers ~ Cruise_Ship$Tonnage, data = Cruise_Ship,
+       xlab = "Tonnage (1000s of tons) of ship",
+       ylab = "Number of Passengers (100s)",
+       main = "Number of Passengers(100s)-Tonnage(1000s of tons) of ship"
+ )
> #creation of 7 models
> model0<-lm(Cruise_Ship$Passengers~1, data=Cruise_Ship)
```

```
> model1<-lm(Cruise_Ship$Passengers~Cruise_Ship$Length, data=Cruise_Ship)
> model2<-lm(Cruise_Ship$Passengers~Cruise_Ship$Tonnage, data=Cruise_Ship)
> model3<-lm(Cruise_Ship$Passengers~Cruise_Ship$Cabins, data=Cruise_Ship)
> model4<-lm(Cruise_Ship$Passengers~Cruise_Ship$Length+Cruise_Ship$Cabins,
    data=Cruise_Ship)
> model5<-lm(Cruise_Ship$Passengers~Cruise_Ship$Cabins+Cruise_Ship$Tonnage,
    data=Cruise_Ship)
> model6<-lm(Cruise_Ship$Passengers~Cruise_Ship$Length+Cruise_Ship$Tonnage,
    data=Cruise_Ship)
> fit<-lm( Cruise_Ship$Passengers ~ Cruise_Ship$Length + Cruise_Ship$Cabins +
    Cruise_Ship$Tonnage, data = Cruise_Ship)
> #comparison of the models using ANOVA function
> anova(model0,model1,model2,model3,model4,model5,model6,fit)
Analysis of Variance Table

Model 1: Cruise_Ship$Passengers ~ 1
Model 2: Cruise_Ship$Passengers ~ Cruise_Ship$Length
Model 3: Cruise_Ship$Passengers ~ Cruise_Ship$Tonnage
Model 4: Cruise_Ship$Passengers ~ Cruise_Ship$Cabins
Model 5: Cruise_Ship$Passengers ~ Cruise_Ship$Length + Cruise_Ship$Cabins
Model 6: Cruise_Ship$Passengers ~ Cruise_Ship$Cabins + Cruise_Ship$Tonnage
Model 7: Cruise_Ship$Passengers ~ Cruise_Ship$Length + Cruise_Ship$Tonnage
Model 8: Cruise_Ship$Passengers ~ Cruise_Ship$Length + Cruise_Ship$Cabins +
   Cruise_Ship$Tonnage
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    157 14702.4
2    156  3225.2 1   11477.2 2780.3302 < 2e-16 ***
3    156  1571.1 0    1654.1
4    156   687.5 0     883.6
5    155   672.0 1      15.4   3.7399 0.05496 .
6    155   635.7 0      36.3
7    155  1557.3 0    -921.6
8    154   635.7 1     921.6 223.2466 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #summary of the fit model
> summary(fit)

Call:
lm(formula = Cruise_Ship$Passengers ~ Cruise_Ship$Length + Cruise_Ship$Cabins +
    Cruise_Ship$Tonnage, data = Cruise_Ship)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0910 -1.0304 -0.5211 0.0814 10.9521

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.296000  1.214610 -0.244 0.80779
Cruise_Ship$Length 0.004222  0.235752  0.018 0.98574
Cruise_Ship$Cabins 1.727292  0.115604 14.941 < 2e-16 ***
Cruise_Ship$Tonnage 0.048637 0.016402  2.965 0.00351 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.032 on 154 degrees of freedom
Multiple R-squared: 0.9568, Adjusted R-squared: 0.9559
F-statistic: 1136 on 3 and 154 DF, p-value: < 2.2e-16

> #Determination of the confidence intervals (95%)
> confint(fit)
                      2.5 %    97.5 %
(Intercept)       -2.6954489 2.10344794
```

```
Cruise_Ship$Length -0.4615031 0.46994631
Cruise_Ship$Cabins 1.4989178 1.95566703
Cruise_Ship$Tonnage 0.0162354 0.08103908
> #Computation of the the beta values (standardised regression coefficients):
> lm.beta(fit)
 Cruise_Ship$Length Cruise_Ship$Cabins Cruise_Ship$Tonnage
      0.0007823944       0.7981160818       0.1871162826
> vif(fit)
 Cruise_Ship$Length Cruise_Ship$Cabins Cruise_Ship$Tonnage
           6.799215          10.162402          14.181589
> #checking the normality assumption of residuals
> shipres = residuals(fit)#
> qqnorm(shipres, main="Normality assumption of Residuals")
> qqline(shipres)
> hist(shipres,main="Histogram of Residuals", xlab = "Residuals")
> #checking the linearty assumption
> qqmath( ~ resid(fit),
+         xlab = "Theoretical Quantiles",
+         ylab = "Residuals"
+ )
> leveneTest(Cruise_Ship$Passengers~
    factor(Cruise_Ship$Length)*factor(Cruise_Ship$Cabins), center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value      Pr(>F)
group 113 5.0802 0.00000001382 ***
       44
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> leveneTest(Cruise_Ship$Passengers, factor(Cruise_Ship$Cabins), center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value    Pr(>F)
group 97 2.8778 0.00001039 ***
      60
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> leveneTest(Cruise_Ship$Passengers, factor(Cruise_Ship$Length), center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group 79 1.6893 0.01071 *
      78
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> leveneTest(Cruise_Ship$Passengers, factor(Cruise_Ship$Tonnage), center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group 93 5.9238 1.73e-12 ***
      64
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #application of Cooks' distance
> plot(cooks.distance(fit),ylab="Cooks distances",xlab="Observation number")
> #application of DFBeta
> dfbetaPlots(fit)
> ##################################################Question 4 Logistic
    Regression#######################################
> library(foreign)
> library(car)  #Package includes Levene's test
> library(tidyr) # for wide to long format transformation of the data

Attaching package: 'tidyr'

The following objects are masked from 'package:reshape':
```

```
    expand, smiths

The following object is masked from 'package:RCurl':

    complete

> library(ggplot2)
> library(QuantPsyc) #include lm.beta()
> library(class)

Attaching package: 'class'

The following object is masked from 'package:reshape':

    condense

> library(pscl)
Classes and Methods for R developed in the
Political Science Computational Laboratory
Department of Political Science
Stanford University
Simon Jackman
hurdle and zeroinfl functions by Achim Zeileis
> library(gmodels)
> columns = c("PH","NisinConcentration","Temoerature","BrixConcetration",
+            "Growth"
+ )
> #reading the dataset
> CRAinAppleJuice <- read.fwf(
+   file=url("http://www.stat.ufl.edu/~winner/data/apple_juice.dat"),
+   widths=c(9, 9, 9, 8, 8),
+   col.names = columns)
> model0 <- glm(CRAinAppleJuice$Growth ~ 1, data=CRAinAppleJuice, family=binomial())
> model1 <- glm(CRAinAppleJuice$Growth ~ PH, data=CRAinAppleJuice, family=binomial())
> #model2 <- glm(CRAinAppleJuice$Growth ~NisinConcentration , data=CRAinAppleJuice,
    family=binomial())
> model2<-glm(CRAinAppleJuice$Growth ~ PH + NisinConcentration, data=CRAinAppleJuice,
    family=binomial())
> anova(model0,model1,model2,test = "Chisq")
Analysis of Deviance Table

Model 1: CRAinAppleJuice$Growth ~ 1
Model 2: CRAinAppleJuice$Growth ~ PH
Model 3: CRAinAppleJuice$Growth ~ PH + NisinConcentration
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        73     95.945
2        72     87.248 1   8.6977   0.003186 **
3        71     64.049 1  23.1983 0.000001461 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model2)

Call:
glm(formula = CRAinAppleJuice$Growth ~ PH + NisinConcentration,
    family = binomial(), data = CRAinAppleJuice)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.3566 -0.6756 -0.2462  0.5301  1.8431

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -6.31911    2.07827  -3.041 0.002361 **
```

```
PH                1.64210   0.49171  3.340 0.000839 ***
NisinConcentration -0.05819 0.01498 -3.884 0.000103 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.945 on 73 degrees of freedom
Residual deviance: 64.049 on 71 degrees of freedom
AIC: 70.049

Number of Fisher Scoring iterations: 5

> ##Pseudo R2
> logisticPseudoR2s <- function(LogModel)
+   #taken from Andy Fields et al. book on R, p.334
+ {
+   dev <- LogModel$deviance
+   nullDev <- LogModel$null.deviance
+   modelN <- length(LogModel$fitted.values)
+   R.l <- 1 - dev / nullDev
+   R.cs <- 1 - exp(-(nullDev - dev) / modelN)
+   R.n <- R.cs / (1 - (exp(-(nullDev / modelN))))
+   cat("Pseudo R^2 for logistic regression\n")
+   cat("Hosmer and Lemshow R^2: ", round(R.l, 3), "\n")
+   cat("Cox and Snell R^2:    ", round(R.cs, 3), "\n")
+   cat("Nagelkerke R^2:       ", round(R.n, 3), "\n")
+ }
> logisticPseudoR2s(model2)
Pseudo R^2 for logistic regression
Hosmer and Lemshow R^2: 0.332
Cox and Snell R^2:      0.35
Nagelkerke R^2:         0.482
> ##Odd and conf
> exp(model2$coefficients)
      (Intercept)              PH NisinConcentration
       0.00180154       5.16602001       0.94346707
> exp(confint(model2))
Waiting for profiling to be done...
                         2.5 %     97.5 %
(Intercept)      0.00002005301 0.07865834
PH               2.14732861143 15.24082097
NisinConcentration 0.91259180182 0.96873926
> ##crosstable
> CRAinAppleJuice$Growpred[fitted(model2) <=0.5] <- 0
> CRAinAppleJuice$Growpred[fitted(model2) > 0.5] <- 1
> CRAinAppleJuice$Growpred<-factor(CRAinAppleJuice$Growpred, levels = c(0:1), labels =
    c("Absence","Presence"))
> table(CRAinAppleJuice$Growth, CRAinAppleJuice$Growpred)

   Absence Presence
  0      42        6
  1       8       18
> CrossTable(CRAinAppleJuice$Growpred, CRAinAppleJuice$Growth, prop.c=FALSE,
    prop.t=FALSE, prop.chisq=FALSE, fisher=FALSE, chisq=FALSE, expected = FALSE)


   Cell Contents
|-------------------------|
|                       N |
|          N / Row Total |
|-------------------------|
```

```
Total Observations in Table: 74


                     | CRAinAppleJuice$Growth
CRAinAppleJuice$Growpred |      0 |        1 | Row Total |
-------------------------|----------|----------|-----------|
              Absence |       42 |        8 |       50 |
                      |    0.840 |    0.160 |    0.676 |
-------------------------|----------|----------|-----------|
             Presence |        6 |       18 |       24 |
                      |    0.250 |    0.750 |    0.324 |
-------------------------|----------|----------|-----------|
         Column Total |       48 |       26 |       74 |
-------------------------|----------|----------|-----------|


> ############################################################Part
    3####################################
>
    ##########################################################################################################
> library(MASS)
> library(foreign)
> library(car)
> library(ggplot2)
> library(nlme)
> library(reshape)
> library(graphics)
> require(lattice)
> library(lattice)
> require(Matrix)
Loading required package: Matrix

Attaching package: 'Matrix'

The following object is masked from 'package:tidyr':

    expand

The following object is masked from 'package:reshape':

    expand

> library(lme4)

Attaching package: 'lme4'

The following object is masked from 'package:nlme':

    lmList

> #question 1#
> set1 <- read.csv2("set1.csv", header = TRUE, sep = ",",
+                   dec = ",", fill = TRUE, comment.char = "")
> hist(set1$score,col="lightblue",main="Score distribution",xlab="score")
> boxplot(set1$score ~ set1$session , xlab="Session", ylab="Score", main="Session related
    to score",col="green",outcol="red")
> #question 2
> ##We present thr intercdept only model which violates the independent assumption. It
    will only allow us to comparre its results with multilevel analysis.
> interceptOnly <- gls(score ~ 1, data = set1, method = "ML")
> summary(interceptOnly)
Generalized least squares fit by maximum likelihood
```

```
  Model: score ~ 1
  Data: set1
       AIC     BIC   logLik
  177112.8 177128.1 -88554.38

Coefficients:
             Value Std.Error t-value p-value
(Intercept) 102.2758 0.4619305 221.4095   0

Standardized residuals:
       Min         Q1        Med         Q3        Max
-3.80616985 -0.56725058 0.04643939 0.57489464 4.46159777

Residual standard error: 58.66154
Degrees of freedom: 16128 total; 16127 residual
> randomInterceptOnly <- lme(score ~ 1, data = set1, random = ~1|Subject, method = "ML")
> summary(randomInterceptOnly)
Linear mixed-effects model fit by maximum likelihood
 Data: set1
       AIC     BIC   logLik
  162676.1 162699.1 -81335.04

Random effects:
 Formula: ~1 | Subject
       (Intercept) Residual
StdDev:   46.38355 35.22182

Fixed effects: score ~ 1
             Value Std.Error  DF  t-value p-value
(Intercept) 102.0279 2.091487 15627 48.78248   0

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.135330461 -0.641203765 0.008247302 0.642200771 4.009769404

Number of Observations: 16128
Number of Groups: 501
> #adding session to our model
> sessionMod<-lme(score ~(1+ session), data = set1, random = ~1|Subject, method = "ML")
> summary(sessionMod)
Linear mixed-effects model fit by maximum likelihood
 Data: set1
       AIC     BIC  logLik
  162453.4 162484.1 -81222.7

Random effects:
 Formula: ~1 | Subject
       (Intercept) Residual
StdDev:     46.443 34.96826

Fixed effects: score ~ (1 + session)
               Value Std.Error  DF   t-value p-value
(Intercept) 108.66574 2.1398997 15626 50.78076   0
session     -0.42533 0.0282743 15626 -15.04304   0
 Correlation:
       (Intr)
session -0.206

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.114815156 -0.647994395 0.007899928 0.638213193 4.160353413

Number of Observations: 16128
```

```
Number of Groups: 501
> ##compute confidence intervals
> intervals(sessionMod, 0.95)
Approximate 95% confidence intervals

 Fixed effects:
                lower       est.      upper
(Intercept) 104.4715439 108.6657350 112.8599262
session      -0.4807492 -0.4253317 -0.3699142
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                 lower  est.  upper
sd((Intercept)) 43.60687 46.443 49.4636

 Within-group standard error:
   lower    est.   upper
34.58273 34.96826 35.35809
```