Sharif University Of Technology

Electrical Engineering

NeuroLab

Homework 03- Regression and Anova

Zahra Kavian - 98102121

1 Question 1

The data is experimental because we control the value of display size and training duration and we want to see their effect on subject search time.

TD and DS are discrete variable and we can treat them as a continuous or categorical variable. I consider as a continuous because 'fitlm' includes $L1$ indicator variables for categorical predictor that has L level (the first category use as a reference level and the model does not include the indicator variable for the reference level).

2 Question 2, Multiple Linear Regression

Now, I fit a multiple linear regression to the search time based on the two continues regressors. I use 'fitlm' matlab function. The main fitting algorithm is QR decomposition. I talk about it a little later.

❑ Regression Model Report

Coefficient Report:

	coefficients	standard error	t-test	p-value
Response Variable	142.9	8.94	15.97	3.3 e-56
DS	25.25	1.12	22.5	1.45e-1.7
TD	6.708	1.68	3.98	6.76e-5

Model Information:

Number of observations: 5709, Error degrees of freedom: 5706

Root Mean Squared Error: 189

R-squared: 0.0836, Adjusted R-Squared: 0.0833

F-statistic vs. constant model: 260, p-value = 6.27e-109

Statistics Analysis with 'ANOVA':

– For whole model:

	sum of squares	degrees of freedom	mean square	F-statistic	p-value
Total	2.22e+08	5708	38894		
Modal	1.85e+07	2	9.28e+06	260.35	6.26e-109
Residual	2.0344e+08	5706	35654		
Lack of fit	1.72e+07	13	1.32e+06	40.615	1.03e-99
Pure error	1.86e+08	5693	32702		

– For each variable:

	sum of squares	degrees of freedom	mean square	F-statistic	p-value
DS	1.8064e+07	1	1.8064e+07	506.65	1.4516e-107
TD	5.6687e+05	1	5.6687e+05	15.899	6.7638e-05
Error	2.0344e+08	5706	35654		

As you see, the variable's coefficients have low p-value and acceptable t-test value. So display size (DS) and training duration (TD) simultaneously have linear relationship with search-time.

❑ Plot Response and regressors variables

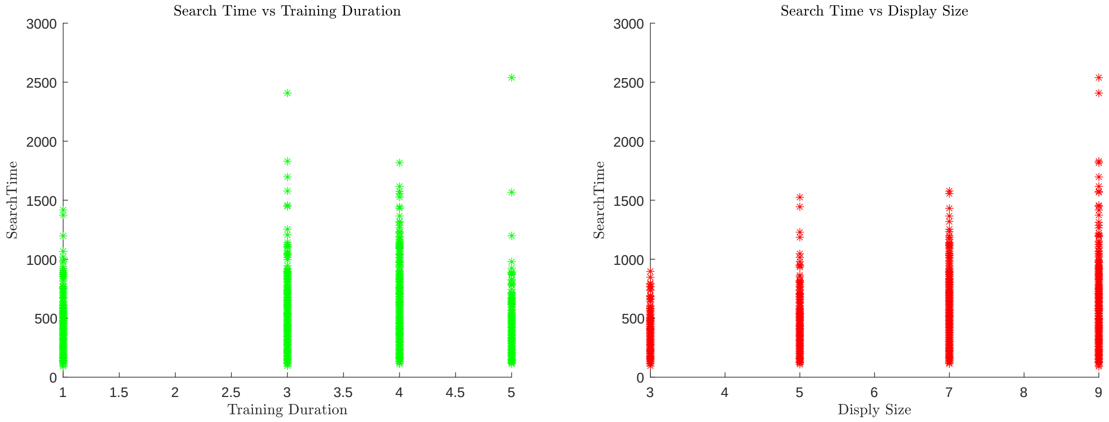


Figure 1: Search time vs each regressors individually

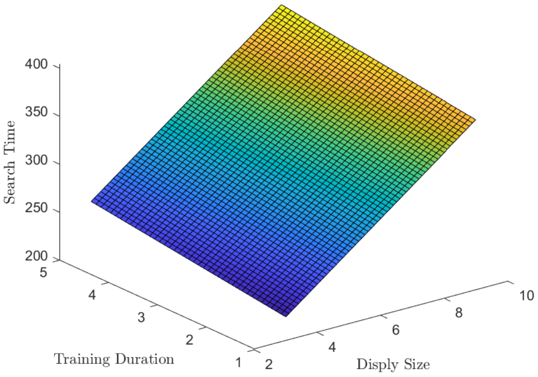


Figure 2: Plane fit on whole data

❑ 95% Confidence Bound

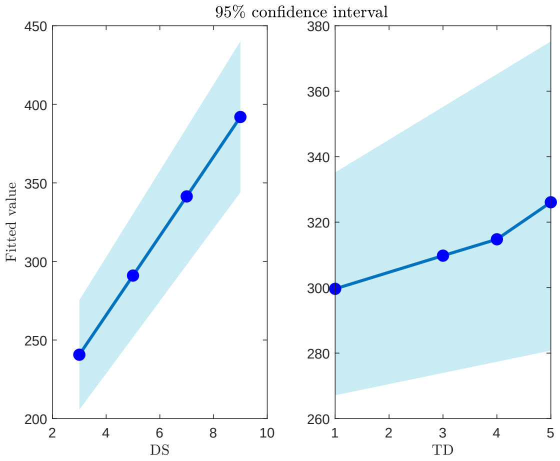


Figure 3: 95% Confidence Interval

❑ QR Decomposition *(Extra)*

In multiple regression, a single quantitative response variable is modeled as a linear combination of quantitative explanatory variables and error. There are n observations and p explanatory variables (including an intercept).

$y = b_0 + b_1 x_1 + ... + b_{p-1} x_{p-1} + error$
$y = Xb + error$

Letting X^T represent the transpose of the matrix X, the normal equations are formed in this way.

$X^T y = X^T Xb$

The solution to the normal equations is

$(X^T X)^{-1} X^T y = b$

where b is the estimate of the parameters that minimizes the residual sum of squares.

On the surface, it appears that this requires the explicit inversion of a matrix, which requires substantial computation. A better algorithm for regression is found by using the QR decomposition.

Here is the mathematical fact. If X is an n by p matrix of full rank (say $n > p$ and the rank = p), then $X = QR$ where Q is an n by p orthonormal matrix and R is a p by p upper triangular matrix. Since Q is orthonormal, $Q^T Q = I$, the identity matrix. Beginning with the normal equations, see how the QR decomposition simplifies them.

$X^T Xb = X^T y$
$(QR)^T (QR)b = (QR)^T y$
$R^T (Q^T Q)Rb = R^T Q^T y$
$R^T Rb = R^T Q^T y$
$(R^T)^{-1} R^T Rb = (R^T)^{-1} R^T Q^T y$
$Rb = Q^T y$
$Ifweletz = Q^T y, \quad Rb = z$

This is simply an upper triangular system of equations which may be quickly solved by back substitution.

3 Question 3, Quality Control for The Regression

Stochastic just means unpredictable. In statistics, the error is the difference between the expected value and the observed value. Let's put these terms together—the gap between the expected and observed values must not be predictable. Or, no explanatory power should be in the error. If you can use the error to make predictions about the response, your model has a problem. This issue is where residual plots play a role.

How do you determine whether the residuals are random in regression analysis? Just check that they are randomly scattered around zero for the entire range of fitted values. When the residuals center on zero, they indicate that the model's predictions are correct on average rather than systematically too high or low. Regression also assumes that the residuals follow a normal distribution and that the degree of scattering is the same for all fitted values.

❑ Assumption of Residual Normality With Q-Q Plot

A histogram of residuals is not a good way to check for normality, since histograms of the same data but using different bin sizes and/or different cut-points between the bins may look quite different. Instead, use a probability plot (quantile plot or Q-Q plot). If the residuals follow the straight line on this type of graph, they are normally distributed.

As you see in figure 4, data approximately fit on straight line. I also show four type residual distribution in box-plot. The distribution is almost same.

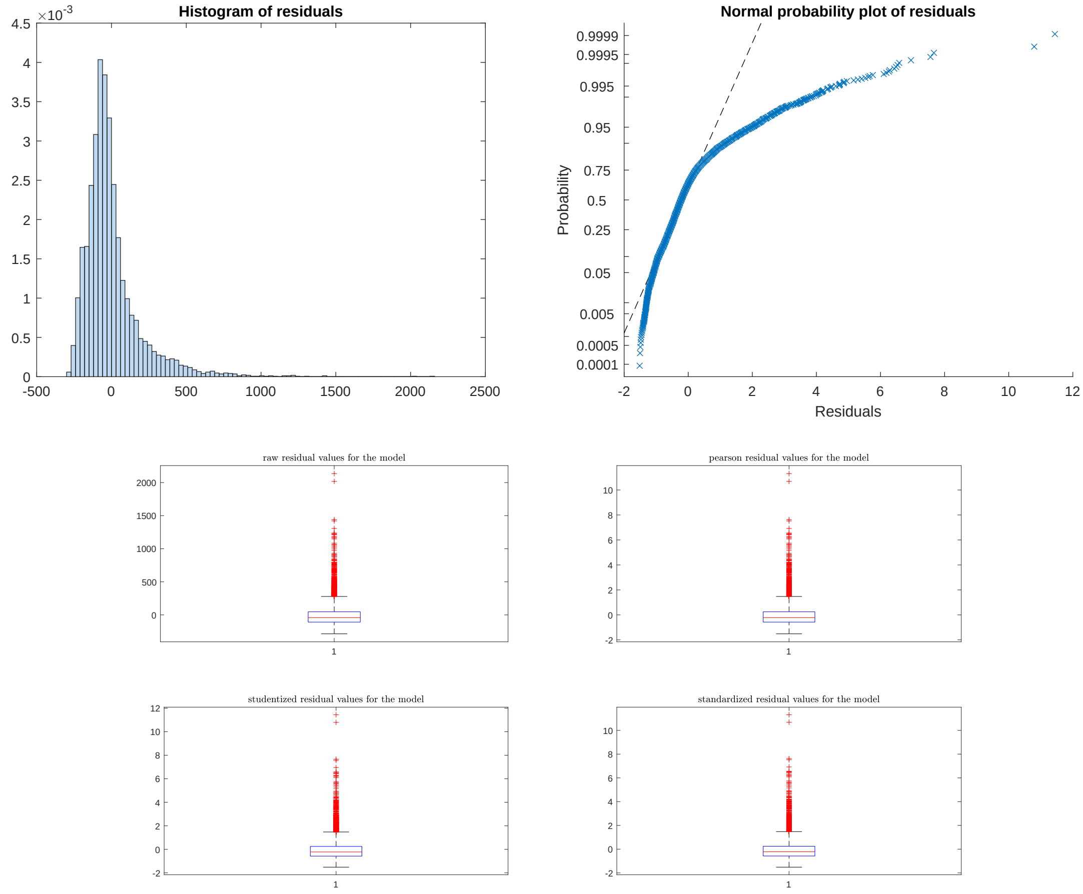


Figure 4: Check Residual Normality

❑ Assumption of Constant Variance

The variance of the errors should be consistent for all observations. In other words, the variance does not change for each observation or for a range of observations. This preferred condition is known as homoscedasticity (same scatter). If the variance changes, we refer to that as heteroscedasticity (different scatter).

The easiest way to check this assumption is to create a residuals versus fitted value plot. On this type of graph, heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction.

As you see in figure 5, the spread of residuals don't increase as the fitted value increase. I plot residuals versus dependent variable. The main distribution is around zero. Also, You could see the residuals versus each independent variable.

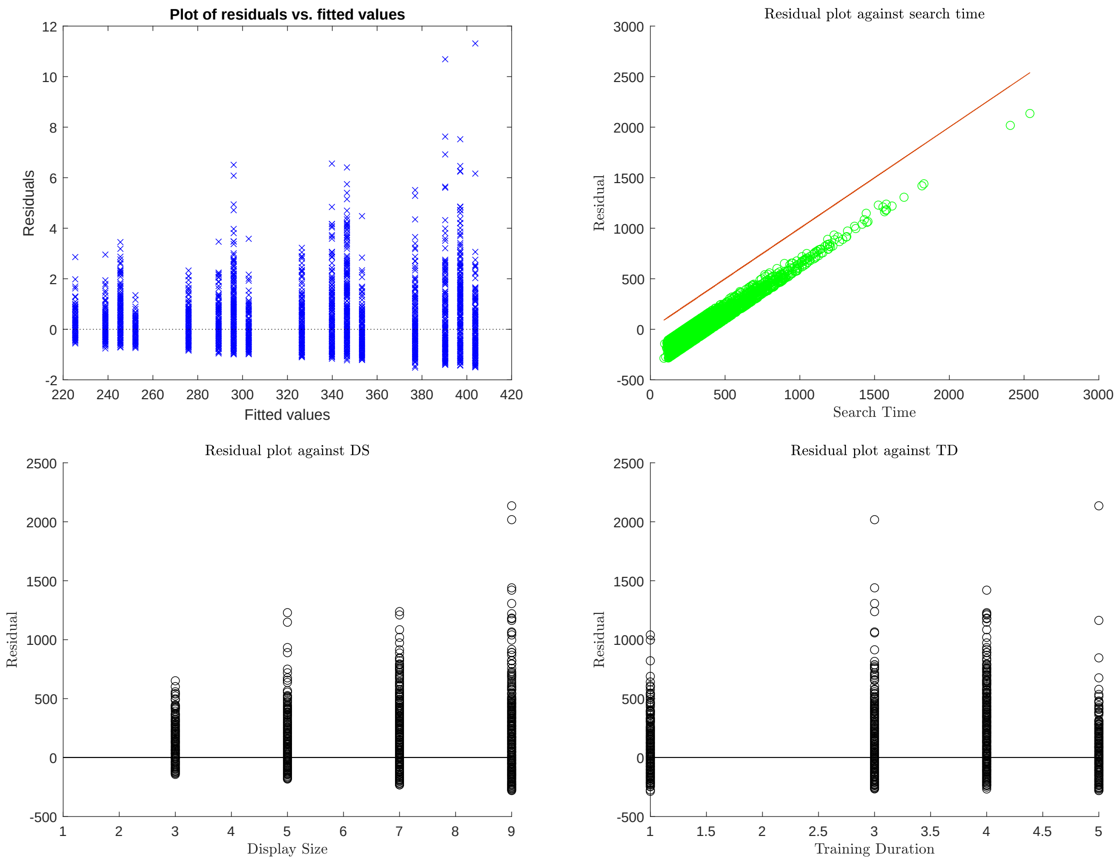


Figure 5: Check Consistency of Residual Variance

❑ Assumption of Residual Independence

One observation of the error term should not predict the next observation. For instance, if the error for one observation is positive and that systematically increases the probability that the following error is positive, that is a positive correlation. If the subsequent error is more likely to have the opposite sign, that is a negative correlation. This problem is known both as serial correlation and autocorrelation.

To check independence, plot residuals against any time variables present (e.g., order of observation), any spatial variables present, and any variables used in the technique (e.g., factors, regressors). A pattern that is not random suggests lack of independence.

Also, Durbin-Watson test returns the p-value for the Durbin-Watson test of the null hypothesis that the residuals from a linear regression are uncorrelated. The alternative hypothesis is that there is autocorrelation among the residuals.

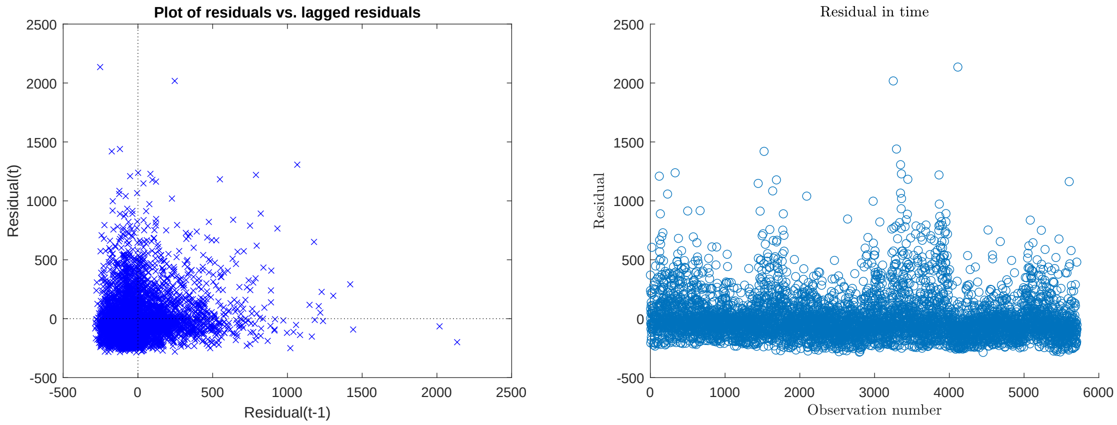


Figure 6: Check independence of residual

Durbin-Watson test value: $7.5594e - 21$

As you see in figure 6 and the Durbin-Watson test value, there is autocorrelation among the residuals.

4 Question 4, Step-Wise Regression

First Method: At first, fit DS to search time and then fit TD to model output. Other time, fit TD to search time and then fit DS to model output. There is significant different in second model coefficient.

	coefficients
DS	25.25
TD	6.708

First TD, then DS:

	coefficients
DS	0.0028
TD	6.30

First DS, then TD:

	coefficients
DS	25.21
TD	-0.40

Second Method:

At first, I fit Ds (or TD) to search time and then use model residual. Then fit TD (or DS) to this. Use DS or TD first, here is no significant difference between coefficient, if first use DS or TD.

Full fit model:

First TD, then DS:

	coefficients
DS	25.21
TD	6.70

First DS, then TD:

	coefficients
DS	6.30
TD	25.25

There is not comparable different between coefficient in step-wise regression and full fit model.

5 Question 5

At first, I fit the best normal distribution on search time. Then use mean and variance model to transform data to normal distribution. Use 'qqplot' to check.

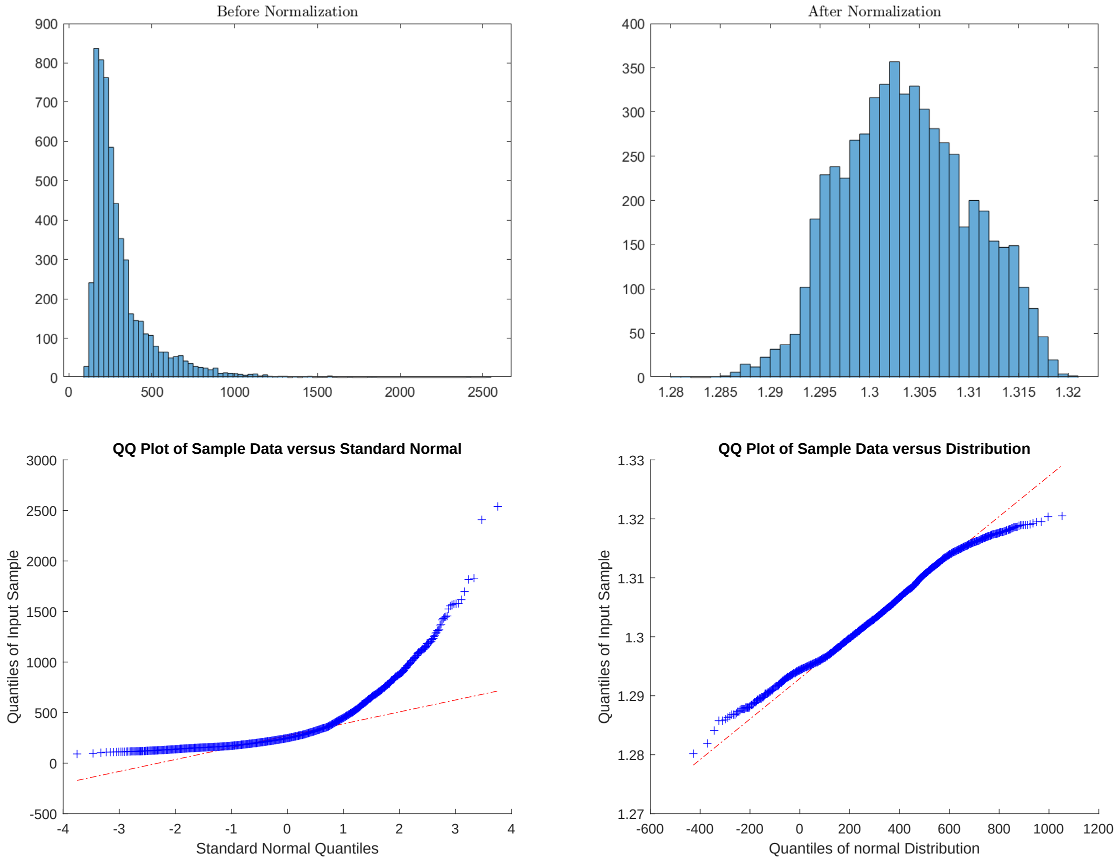


Figure 7: Normalization response variable

Run the model again:

	coefficients	standard error	t-test	p-value
Response Variable	1.299	0.00029	4409.3	0
DS	0.000725	3.69e-05	19.62	5.192e-83
TD	0.000165	5.53e-05	2.97	0.00290

❑ Assumption of Constant Variance

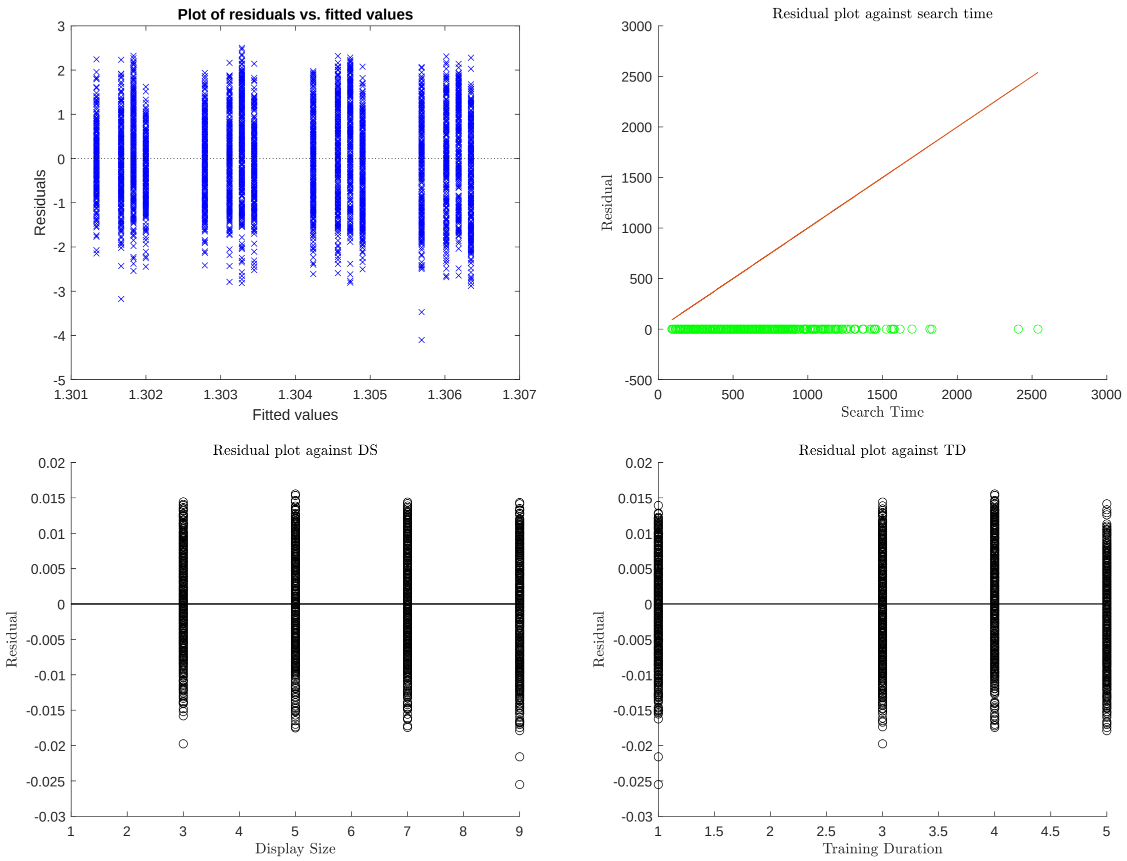


Figure 8: Check Consistency of Residual Variance

❑ Assumption of Residual Independence

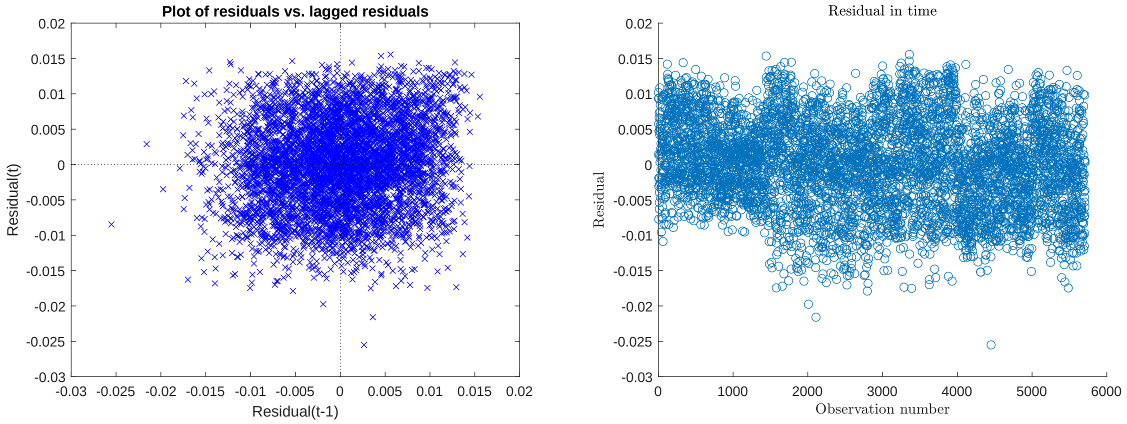


Figure 9: Check independence of residual

❑ Assumption of Residual Normality With Q-Q Plot

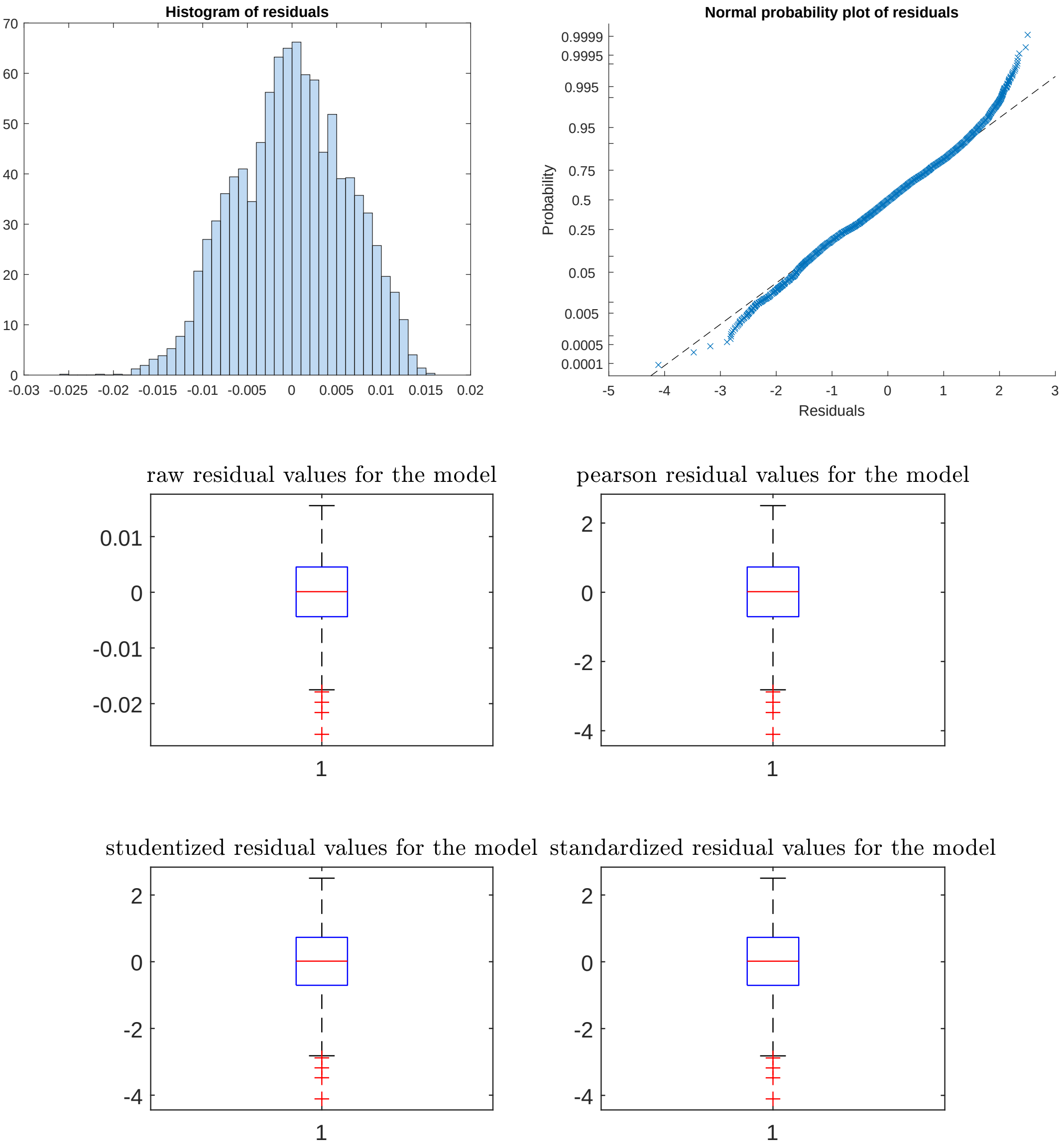


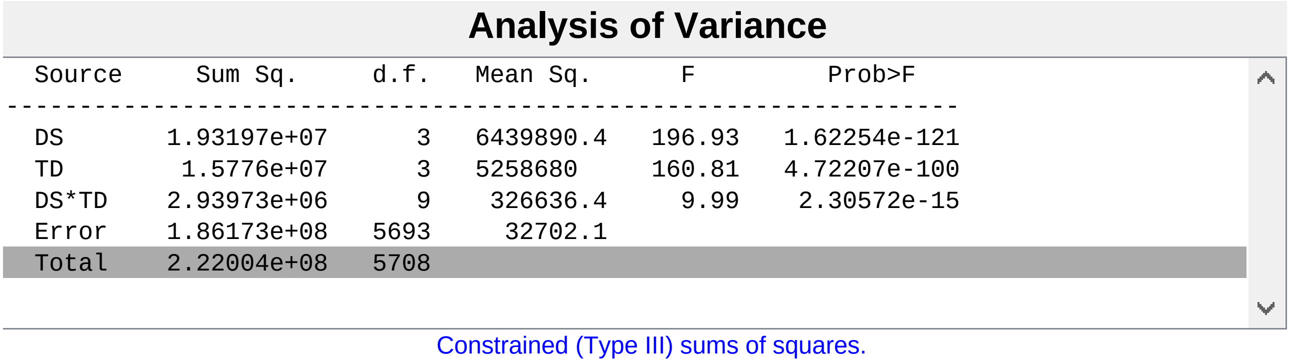
Figure 10: Check Residual Normality

6 Question 6, ANOVA Analysis

a. It is a fixed-effect model. ”The fixed-effects model (class I) of analysis of variance applies to situations in which the experimenter applies one or more treatments to the subjects of the experiment to see whether the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole.”

Type I ANOVA only ask about the differences among these treatments. In our experiment, It only say how these specific display size and training duration would effect on search time. But if we had various factor levels (sampled from a larger population), we could use type II. Then ANOVA ask about the effect of parameters in general.

b.



Analysis of Variance

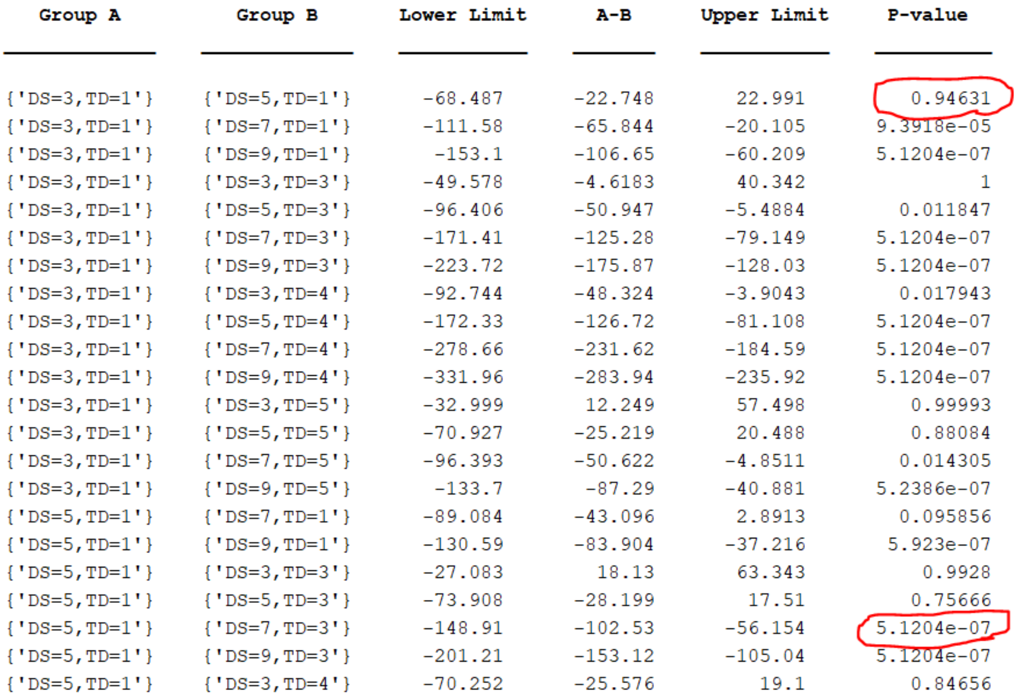
```
  Source      Sum Sq.      d.f.    Mean Sq.       F          Prob>F
-------------------------------------------------------------------
  DS        1.93197e+07      3    6439890.4    196.93   1.62254e-121
  TD         1.5776e+07      3    5258680      160.81   4.72207e-100
  DS*TD     2.93973e+06      9     326636.4      9.99     2.30572e-15
  Error     1.86173e+08   5693      32702.1
  Total     2.22004e+08   5708
```

Constrained (Type III) sums of squares.

Figure 11: ANOVA result

The p-value 1.6e-121 indicate the mean responses for levels of the factor DS are significantly different. The p-value 4.7e-100 indicate the mean responses for levels of the factor TD are significantly different. The last entry is the p-values for two-way interactions. The p-value of 2.3e-15 indicates that the interaction between DS and TD is significant.

I perform multiple comparison tests to find out which groups of the factors TD and DS are significantly different. I save the result in ’anova.xlsx’. You can see some of these comparsion in figure 12.

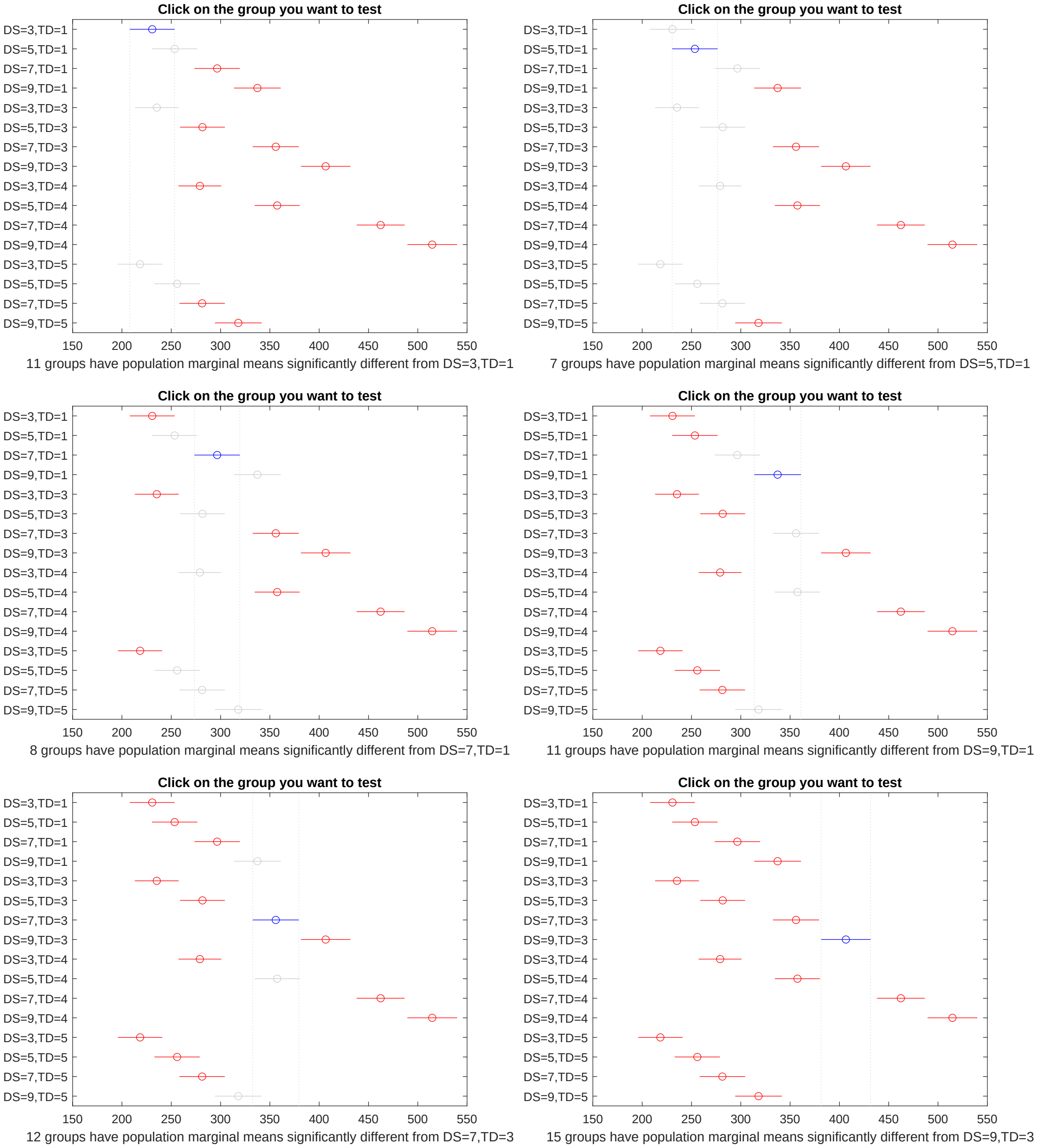


Group A	Group B	Lower Limit	A-B	Upper Limit	P-value
{'DS=3,TD=1'}	{'DS=5,TD=1'}	-68.487	-22.748	22.991	0.94631
{'DS=3,TD=1'}	{'DS=7,TD=1'}	-111.58	-65.844	-20.105	9.3918e-05
{'DS=3,TD=1'}	{'DS=9,TD=1'}	-153.1	-106.65	-60.209	5.1204e-07
{'DS=3,TD=1'}	{'DS=3,TD=3'}	-49.578	-4.6183	40.342	1
{'DS=3,TD=1'}	{'DS=5,TD=3'}	-96.406	-50.947	-5.4884	0.011847
{'DS=3,TD=1'}	{'DS=7,TD=3'}	-171.41	-125.28	-79.149	5.1204e-07
{'DS=3,TD=1'}	{'DS=9,TD=3'}	-223.72	-175.87	-128.03	5.1204e-07
{'DS=3,TD=1'}	{'DS=3,TD=4'}	-92.744	-48.324	-3.9043	0.017943
{'DS=3,TD=1'}	{'DS=5,TD=4'}	-172.33	-126.72	-81.108	5.1204e-07
{'DS=3,TD=1'}	{'DS=7,TD=4'}	-278.66	-231.62	-184.59	5.1204e-07
{'DS=3,TD=1'}	{'DS=9,TD=4'}	-331.96	-283.94	-235.92	5.1204e-07
{'DS=3,TD=1'}	{'DS=3,TD=5'}	-32.999	12.249	57.498	0.99993
{'DS=3,TD=1'}	{'DS=5,TD=5'}	-70.927	-25.219	20.488	0.88084
{'DS=3,TD=1'}	{'DS=7,TD=5'}	-96.393	-50.622	-4.8511	0.014305
{'DS=3,TD=1'}	{'DS=9,TD=5'}	-133.7	-87.29	-40.881	5.2386e-07
{'DS=5,TD=1'}	{'DS=7,TD=1'}	-89.084	-43.096	2.8913	0.095856
{'DS=5,TD=1'}	{'DS=9,TD=1'}	-130.59	-83.904	-37.216	5.923e-07
{'DS=5,TD=1'}	{'DS=3,TD=3'}	-27.083	18.13	63.343	0.9928
{'DS=5,TD=1'}	{'DS=5,TD=3'}	-73.908	-28.199	17.51	0.75666
{'DS=5,TD=1'}	{'DS=7,TD=3'}	-148.91	-102.53	-56.154	5.1204e-07
{'DS=5,TD=1'}	{'DS=9,TD=3'}	-201.21	-153.12	-105.04	5.1204e-07
{'DS=5,TD=1'}	{'DS=3,TD=4'}	-70.252	-25.576	19.1	0.84656

Figure 12: comparison of combinations of groups (levels) of the two variables, TD and DS.

The p-value corresponding to this test is 5.1e-07, which indicates that the mean responses of DS=5, TD=1 and DS=7, TD=3 are significantly different. The p-value equal 0.94 means the mean response of these two group is almost the same.

In figure 13, I show which groups are significantly different from each other. The bars for the groups that are significantly different are red. The bars for the groups that are not significantly different are gray.



Click on the group you want to test

11 groups have population marginal means significantly different from DS=3,TD=1

Click on the group you want to test

7 groups have population marginal means significantly different from DS=5,TD=1

Click on the group you want to test

8 groups have population marginal means significantly different from DS=7,TD=1

Click on the group you want to test

11 groups have population marginal means significantly different from DS=9,TD=1

Click on the group you want to test

12 groups have population marginal means significantly different from DS=7,TD=3

Click on the group you want to test

15 groups have population marginal means significantly different from DS=9,TD=3

Figure 13: comparison of combinations of groups (levels) of the two variables, TD and DS.

I do post-hoc comparison using ’Tukey’, ’Sheffe’ and ’Bonferroni’ comparison and save their result in separate excel file (by their name). In figure 14 and figure 15, I show comparison of two group among these three methods.

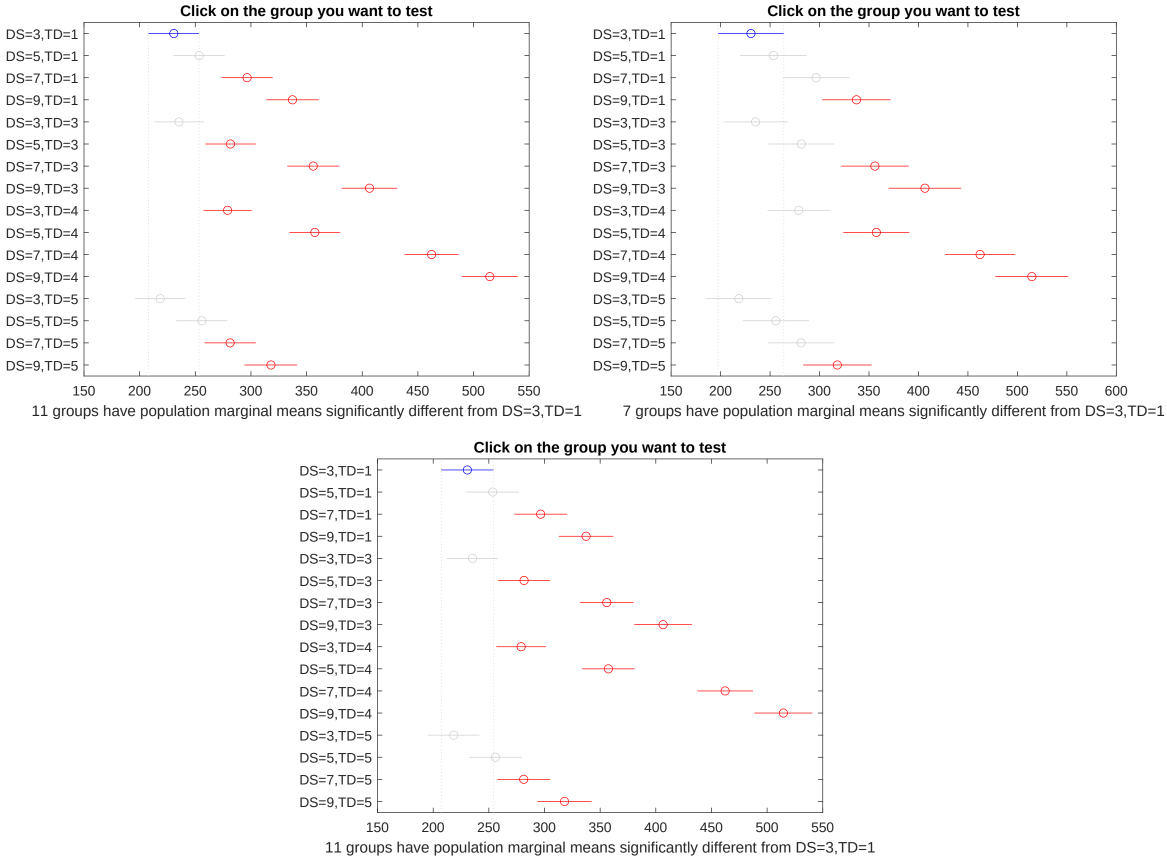


Figure 14: comparison of combinations of groups (levels) of the two variables;tukey-kramer,scheffe,bonferroni.

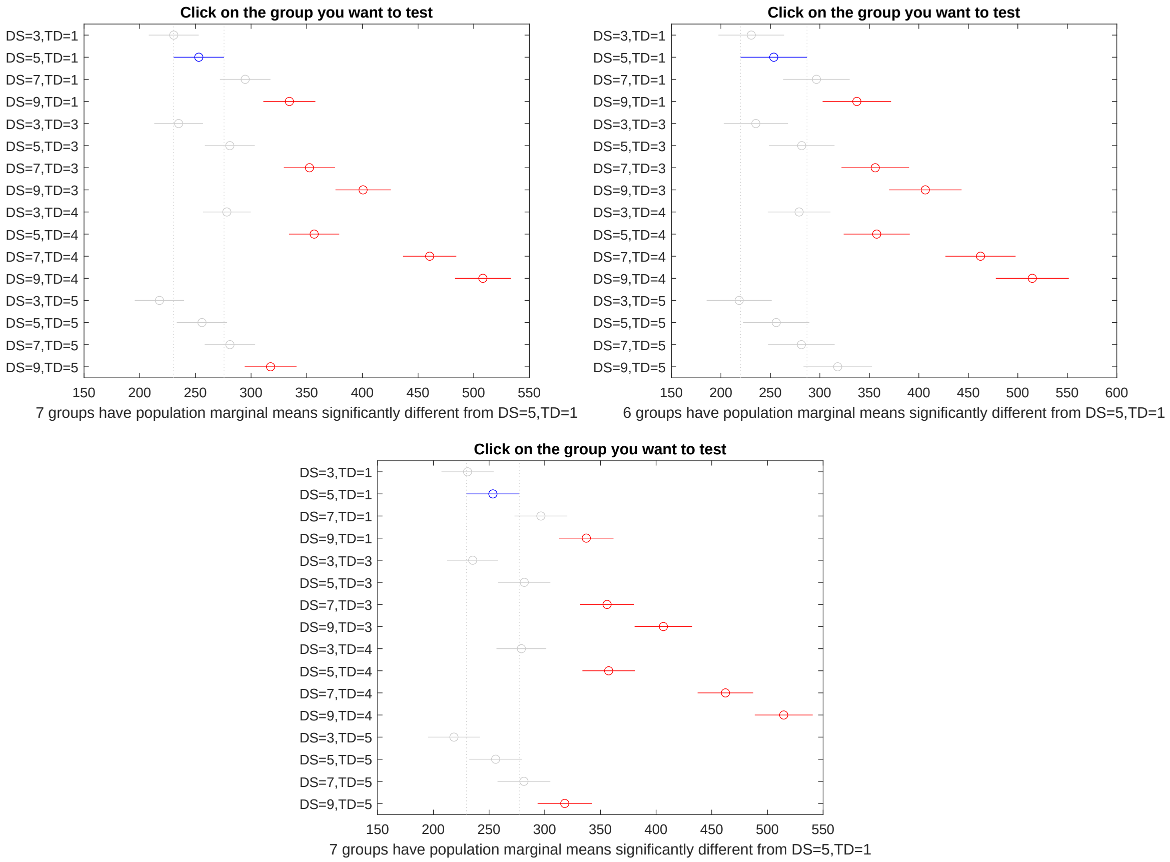
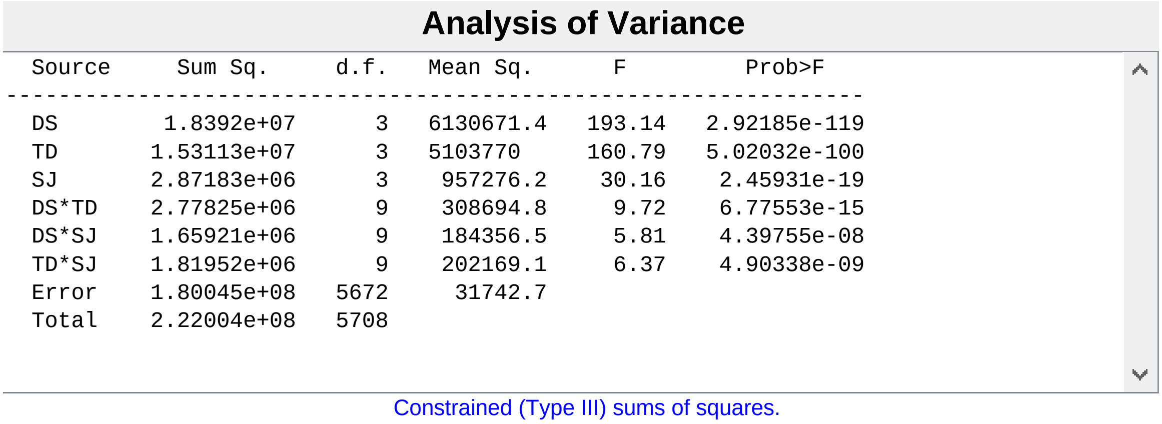


Figure 15: comparison of combinations of groups (levels) of the two variables;tukey-kramer,scheffe,bonferroni.

7 Question 7 (Bonus assignment)

a. Subject is a fixed effect, we just have specific level (subject number). It is not repeatable measurement.

b.



Analysis of Variance

```
Source       Sum Sq.      d.f.   Mean Sq.       F          Prob>F
-------------------------------------------------------------------
DS         1.8392e+07       3   6130671.4    193.14    2.92185e-119
TD        1.53113e+07       3   5103770      160.79    5.02032e-100
SJ        2.87183e+06       3     957276.2    30.16      2.45931e-19
DS*TD     2.77825e+06       9     308694.8     9.72      6.77553e-15
DS*SJ     1.65921e+06       9     184356.5     5.81      4.39755e-08
TD*SJ     1.81952e+06       9     202169.1     6.37      4.90338e-09
Error     1.80045e+08    5672      31742.7
Total     2.22004e+08    5708
```

Constrained (Type III) sums of squares.

Figure 16: comparison of combinations of groups (levels) of the two variables, TD and DS.

The comparison between each independent variable save in separate excel file. In figure 17, I show comparison for 'TD' and 'subject'.

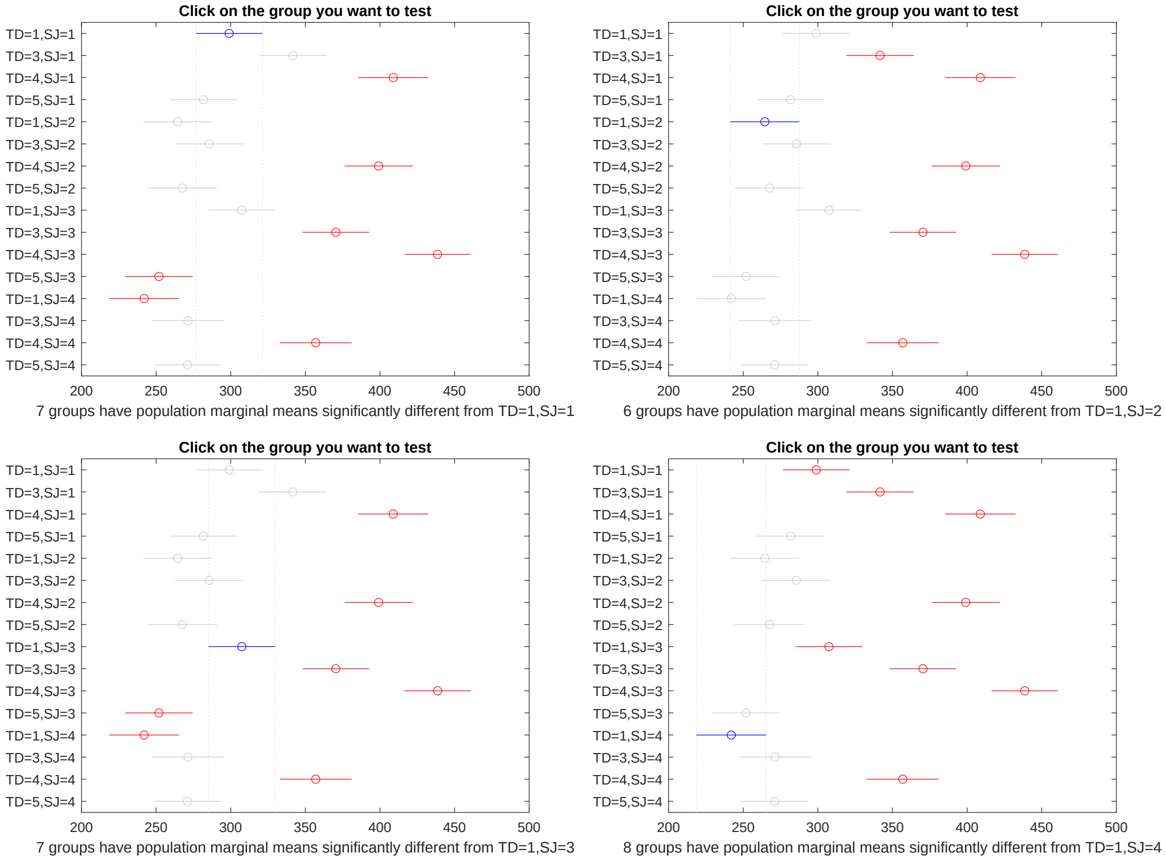


Figure 17: comparison of combinations of groups (levels) of the two variables, TD and Subject.

8 jh

	Pros	
ICA	ability to separate mixed signals does not require assumptions about the underlying probability distribution of the data unsupervised learning Feature extraction	a Co