Regression and Anova

Mahdi Tabatabaei

Neuroscience Lab

Javad Khodadoost

April 8, 2024

1. Loading the search data: The dependent variable is search time (ST) there are two independent variables: display size (DS) and training duration (TD).

a. Is this data observational or experimental?

Observational and Experimental Data

Observational and experimental data are two primary types of data used in scientific research.

Observational Data: Observational data is collected by observing subjects or phenomena without intervening or manipulating them. Researchers do not manipulate any variables; they simply observe and record what happens naturally. This type of data collection is often used in fields where it's unethical or impractical to conduct experiments or where the phenomenon under study cannot be manipulated. Observational studies can be cross-sectional (data collected at a single point in time) or longitudinal (data collected over a period of time).

Experimental Data: Experimental data is collected through controlled experiments where researchers deliberately manipulate one or more variables and observe the effect on another variable. Researchers apply treatments or interventions to a group (experimental group) and compare the results with another group that did not receive the treatment (control group). The aim of experimental data collection is to establish cause-and-effect relationships between variables. Experiments are designed to control for confounding variables that could influence the results, typically through randomization and control groups.

Since we can change experiment's situation in the way that we can adjust randomness of the samples, it is **experimental**.

b. Are the independent variables categorical or continuous?

Categorical and Continuous Variables

Categorical Variables: Categorical variables represent characteristics or qualities that do not have a numerical value. Instead, they are typically represented by labels or categories. Examples of categorical variables include gender, ethnicity, marital status, and type of vehicle. Categorical variables can be further divided into nominal and ordinal variables.

- Nominal Variables: Nominal variables are categorical variables that have no inherent order or ranking among the categories. For example, colors (red, blue, green) and types of fruit (apple, banana, orange) are nominal variables.
- Ordinal Variables: Ordinal variables are categorical variables that have a natural ordering or ranking among the categories. However, the intervals between the categories may not be uniform or meaningful. Examples of ordinal variables include education level (high school, bachelor's degree, master's degree) and satisfaction level (low, medium, high).

Continuous Variables: Continuous variables are variables that can take any numerical value within a certain range. They are measured, not counted, and can represent quantities that can be infinitely divided into smaller increments.Examples of continuous variables include height, weight, temperature, and income.

Independent variables (Ds, TD) are **categorical** because these are representing a group and are not continuous.

2. Fitting a multiple linear regression to the search time based on the two regressors:

a. Report regression coefficients, t-values and significance for each coefficient and the F statistic for the full model. What is your conclusion?

Regression Model

We use following model for our data:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad , \quad \sigma = \frac{SSE}{n-3}$$

We have 3 parameters and we lose 3 free degrees and for variance not to be biased we have the formula.
Now, we use `fitlm` function in MATLAB to find coefficients and p-values:

```
% Load data from .mat file
DS = Data.Data.DS;
TD = Data.Data.TD;
ST = Data.Data.SearchTime;

% Define independent variables (predictors) and dependent variable
X = [DS, TD];  % DS and TD are assumed to be column vectors
Y = ST;        % ST is assumed to be a column vector

% Fit multiple linear regression model
mdl = fitlm(X, Y, 'linear');

disp(mdl);
```

Results

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

	Coefficient	Standard Error	t-stat	p-value
Intercept	$\beta_0 = 142.9014$	8.9474	15.971	3.3032e-56
Display Size (DS)	$\beta_1 = 25.2588$	1.1222	22.509	1.4516e-107
Training Duration (TD)	$\beta_2 = 6.7084$	1.6824	3.9874	6.7638e-05

Conclusion

The conclusion drawn from the regression analysis typically involves examining the significance of the coefficients and the overall fit of the model. Here's what you would look for:

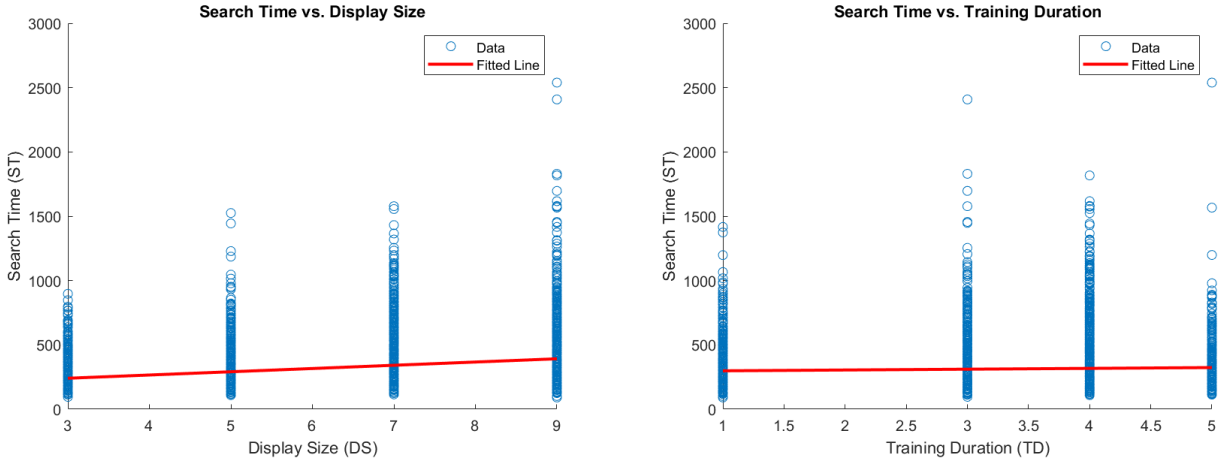
Coefficient Significance: Look at the p-values associated with each coefficient. If the p-value is less than the chosen significance level (usually 0.05), you can reject the null hypothesis that the coefficient is equal to zero, indicating that the predictor variable is significantly related to the response variable.

F-Statistic Significance: The F-statistic tests the overall significance of the model. If the p-value associated with the F-statistic is less than 0.05, you can reject the null hypothesis that all coefficients in the model are zero, suggesting that the model as a whole is significant.

Coefficient Signs and Magnitudes: Examine the signs and magnitudes of the coefficients to understand the direction and strength of the relationship between each predictor and the response variable.

Based on these factors, we can conclude whether the model is statistically significant and whether the predictors have a significant impact on the search time.

b. Plot 2D (search vs each regressor individually) and 3D (plane) fits.



(a) 2D plot (DS-TD)

(b) 2-D plot (ST-TD)

Figure 1: 2D plot (ST-each independent variable)

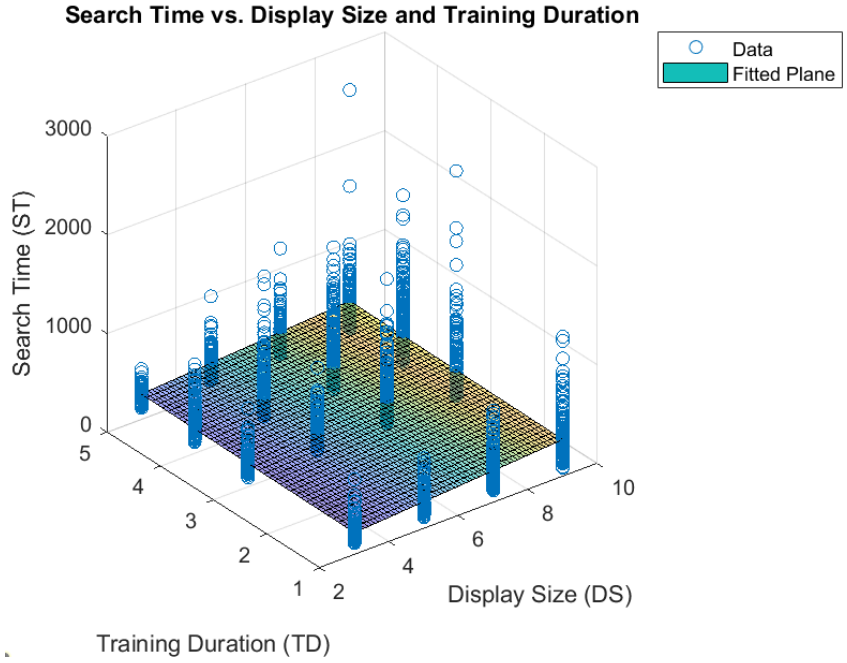


Figure 2: 3D plot

c. Plot y and its 95% confidence bound in the 2D plots.

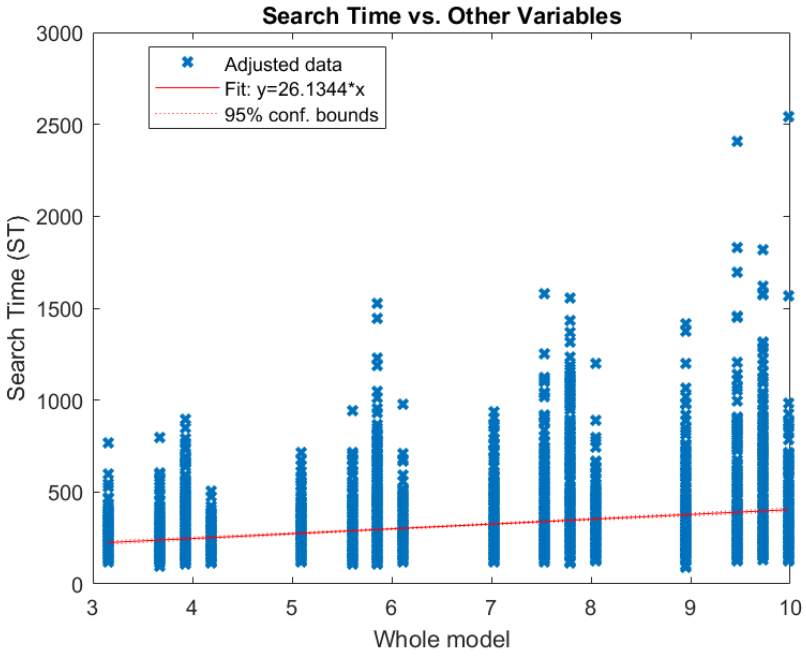


Figure 3: 2D plot (Y)

3. Quality control for the regression:

a. Check the assumption of residual normality with Q-Q plot.

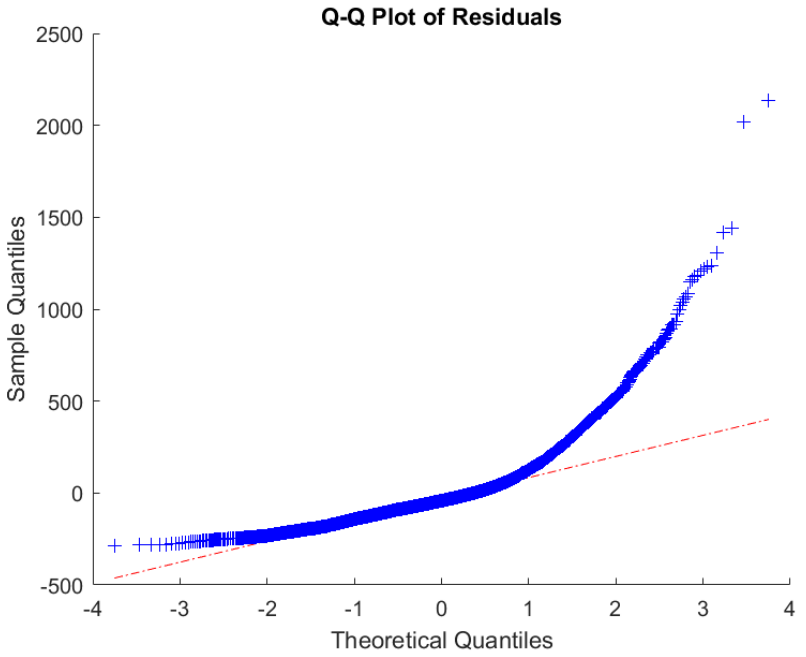


Figure 4: Q-Q plot

Residuals Normality

In a Q-Q plot (Quantile-Quantile plot), the residuals are plotted against the quantiles of a theoretical normal distribution. The diagonal line on the plot represents the line where the residuals would lie if they followed a perfect normal distribution. Therefore, if the residuals fall perfectly along this diagonal line, it indicates that they closely follow a normal distribution.

Figure 4, shows that the residuals deviate from the diagonal line, it means departures from normality. So, epsilon doesn't have a normal distribution.

As the points are consistently above the line, it suggests that the residuals have heavier tails than a normal distribution. This indicates positive skewness in the residuals.

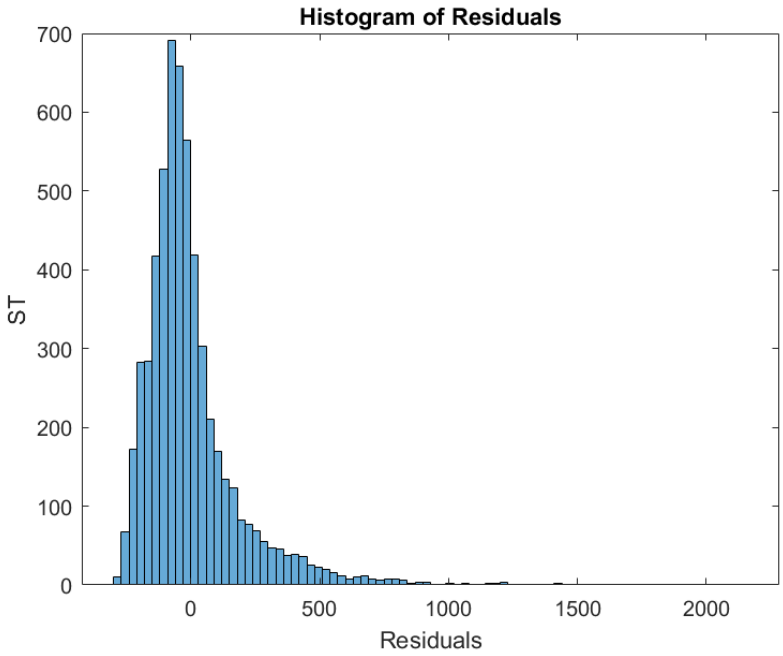


Figure 5: Histogram of Residuals

Histogram also shows our assumption for normality of residuals is not true.

b. Check the assumption of constant variance.

Constant Variance

Plotting the residuals versus the predicted values (fitted values) is a common method for checking the assumption of constant variance (homoscedasticity) in linear regression models. This plot is useful because it allows us to examine how the spread of residuals varies across different levels of the predicted values.
`Sum(residuals)` $= -2.7394e-09$ and it is almost 0. Now we should check the dispersion around the mean. As we observe the dispersion around 0 is not equal and above the red line we have more values. So, assumption of constant variance is not true.

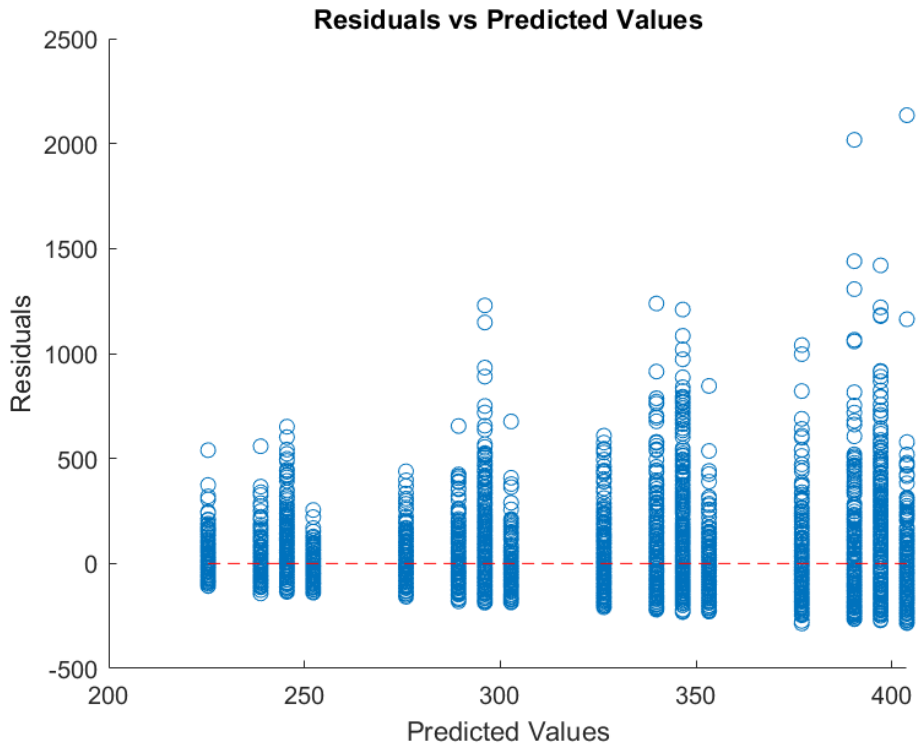


Figure 6: Residuals-Predicted Valuse

c. Check the assumption of residual independence.

Residual Independence

Plot Residuals Over Time or Observation Index: If your data is time series data or has a natural ordering (e.g., in longitudinal studies), you can plot the residuals against time or observation index. If the residuals exhibit any systematic patterns, such as trends, cycles, or seasonality, it suggests the presence of autocorrelation, which violates the assumption of residual independence. We can't see a pattern So there is no autocorrelation. So, residuals are independant.

Examine Residuals Scatter Plot: Create a scatter plot of residuals against the index of observations (e.g., the order in which they were collected). If there are any discernible patterns, trends, or clusters in the scatter plot, it may indicate lack of independence in the residuals. We can't see a pattern So there is no autocorrelation. So, residuals are independant.

Statistical Tests: Conduct formal statistical tests for autocorrelation, such as the Durbin-Watson test or the Ljung-Box test. These tests can provide quantitative measures of autocorrelation in the residuals.

Durbin-Watson test statistic	1.7524
Ljung-Box test statistic	728.3524

Durbin-Watson test is close to 2. So it shows no autocorrelation which means residuals are independent.

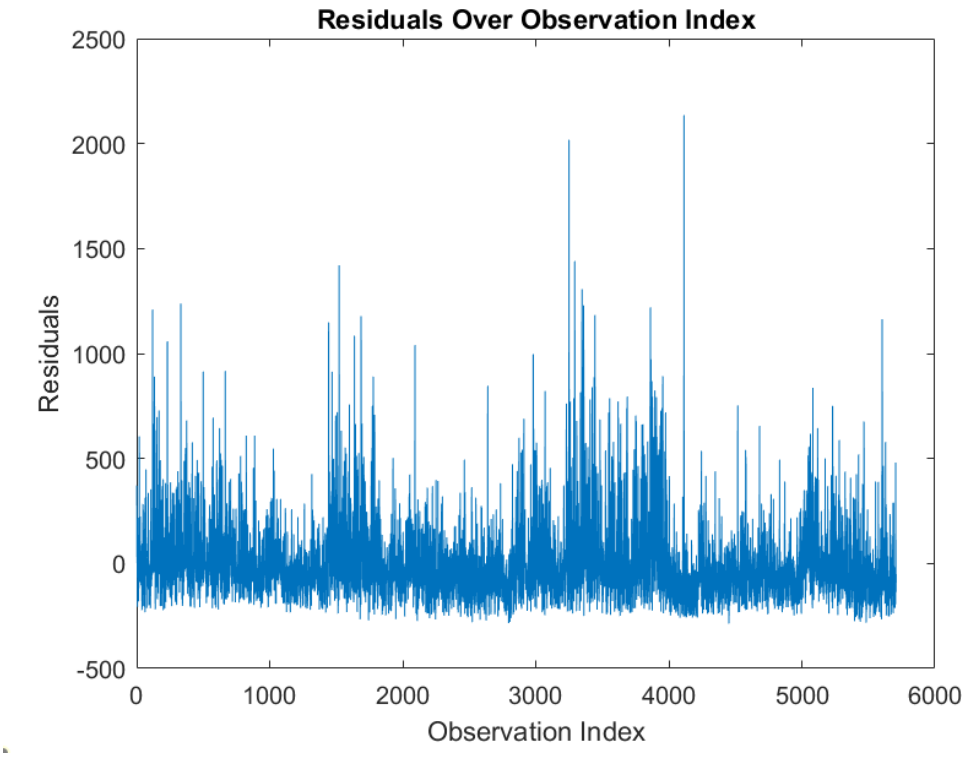


Figure 7: Residuals Against Observation Index Plot

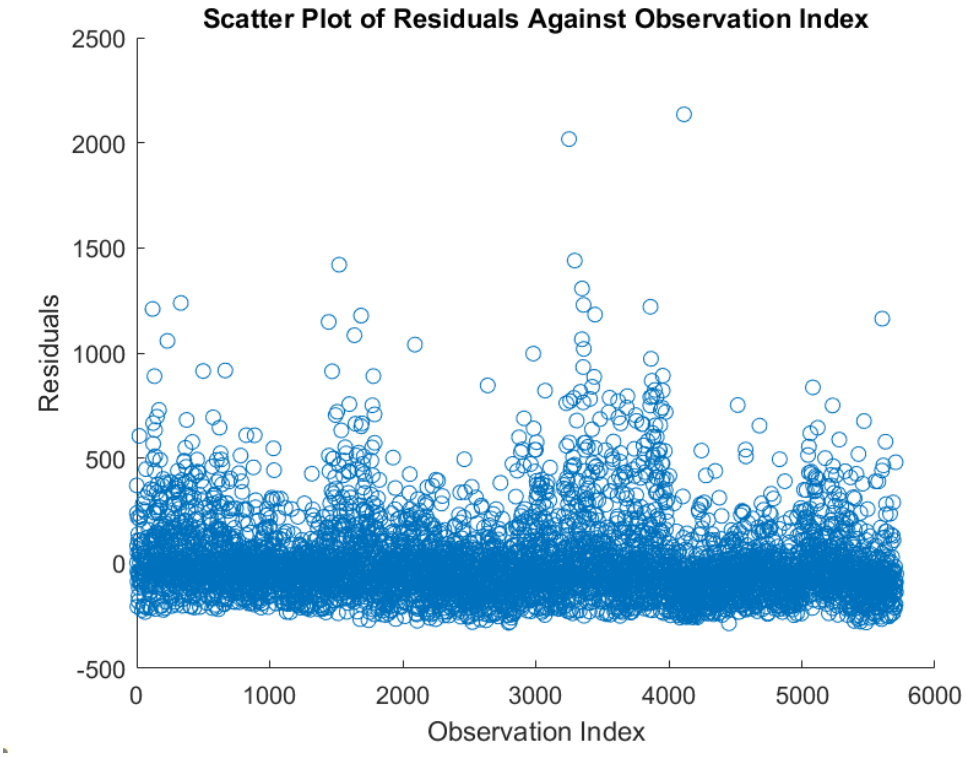


Figure 8: Residuals Scatter Plot

4. Do step-wise regression first fit ST with DS then use the residual to fit TD. Also do the opposite. Are the beta coefficients comparable with when you fit the full model?

DS - TD VS. Full Model

First, fitting ST with DS:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

Number of observations: 5709, Error degrees of freedom: 5707
Root Mean Squared Error: 189
R-squared: 0.0811, Adjusted R-Squared: 0.0809
F-statistic vs. constant model: 503, p-value = 6.23e-107

	Coefficient	Standard Error	t-stat	p-value
Intercept	$\beta_0 = 164.98$	7.0381	23.441	4.3287e-116
X_1	$\beta_1 = 25.211$	1.1236	22.438	6.2343e-107

Then, using residuals to fit TD:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

Number of observations: 5709, Error degrees of freedom: 5707
Root Mean Squared Error: 1.48
R-squared: 0.00278, Adjusted R-Squared: 0.0026
F-statistic vs. constant model: 15.9, p-value = 6.76e-05

	Coefficient	Standard Error	t-stat	p-value
Intercept	$\beta_0 = 3.2491$	0.01936	165.46	0
X_1	$\beta_1 = 0.00041421$	0.0001038	3.9875	1

Now, for full model we have:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

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

	Coefficient	Standard Error	t-stat	p-value
Intercept	$\beta_0 = 142.9014$	8.9474	15.971	3.3032e-56
Display Size (DS)	$\beta_1 = 25.2588$	1.1222	22.509	1.4516e-107
Training Duration (TD)	$\beta_2 = 6.7084$	1.6824	3.9874	6.7638e-05

We can observe that coefficient of DS in the model we linearize without TD, is almost equal to coefficient of it in full model.
P-Values indicates that there is nothing related betwwen residuals and TD.

TD - DS VS. Full Model

First, fitting ST with DTDS:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

Number of observations: 5709, Error degrees of freedom: 5707
Root Mean Squared Error: 197
R-squared: 0.00226, Adjusted R-Squared: 0.00208
F-statistic vs. constant model: 12.9, p-value = 0.000331

	Coefficient	Standard Error	t-stat	p-value
Intercept	$\beta_0 = 292.1$	6.2707	46.581	0
X_1	$\beta_1 = 6.3044$	1.7553	3.5917	0.00033128

Then, using residuals to fit DS:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

Number of observations: 5709, Error degrees of freedom: 5707
Root Mean Squared Error: 189
R-squared: 0.0815, Adjusted R-Squared: 0.0814
F-statistic vs. constant model: 507, p-value = 1.43e-107

	Coefficient	Standard Error	t-stat	p-value
Intercept	$\beta_0 = -147.87$	7.0283	-21.039	1.0539e-94
X_1	$\beta_1 = 25.256$	1.122	22.51	1.4318e-107

Now, for full model we have:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

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

	Coefficient	Standard Error	t-stat	p-value
Intercept	$\beta_0 = 142.9014$	8.9474	15.971	3.3032e-56
Display Size (DS)	$\beta_1 = 25.2588$	1.1222	22.509	1.4516e-107
Training Duration (TD)	$\beta_2 = 6.7084$	1.6824	3.9874	6.7638e-05

We can observe that coefficient of TD in the model we linearize without DS, is almost equal to coefficient of it in full model.
P-Values indicates that there is nothing related betwwen residuals and DS.

5. Is the search time normally distributed? If not apply a transformation to make it normal (confirm success with Q-Q plot) and redo the regression. Compare the results with previous version of regression.

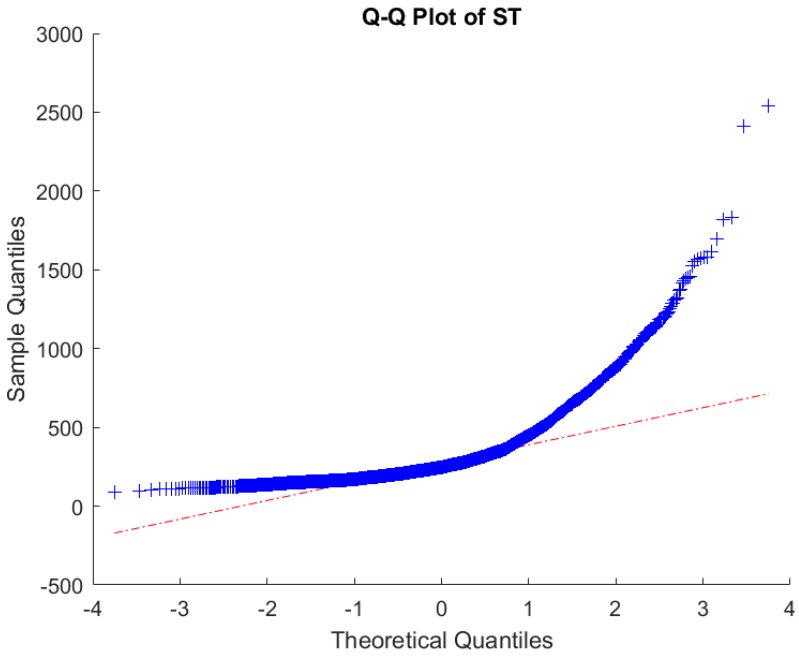


Figure 9: Q-Q Plot of ST

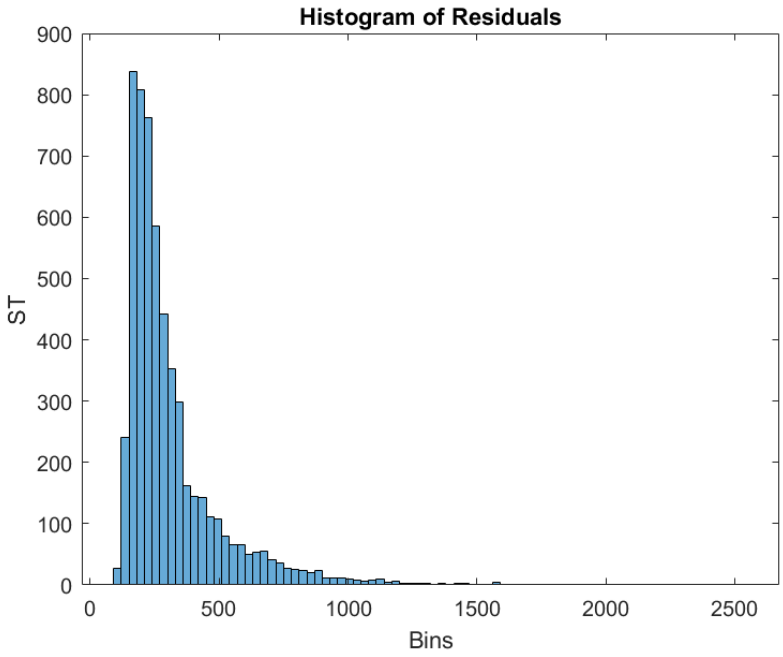


Figure 10: Histogram of ST

Normality of ST

We learned that we can use Q-Q plot and histogram to see if ST is normal or not. By plotting the Q-Q plot we see that values are not on the line. So ST is not normal. The histogram will show that ST is not normally distributed.

Transformed ST

Now, we apply `log, sqrt, inv` into ST and will compare the results on Q-Q plots and histograms.

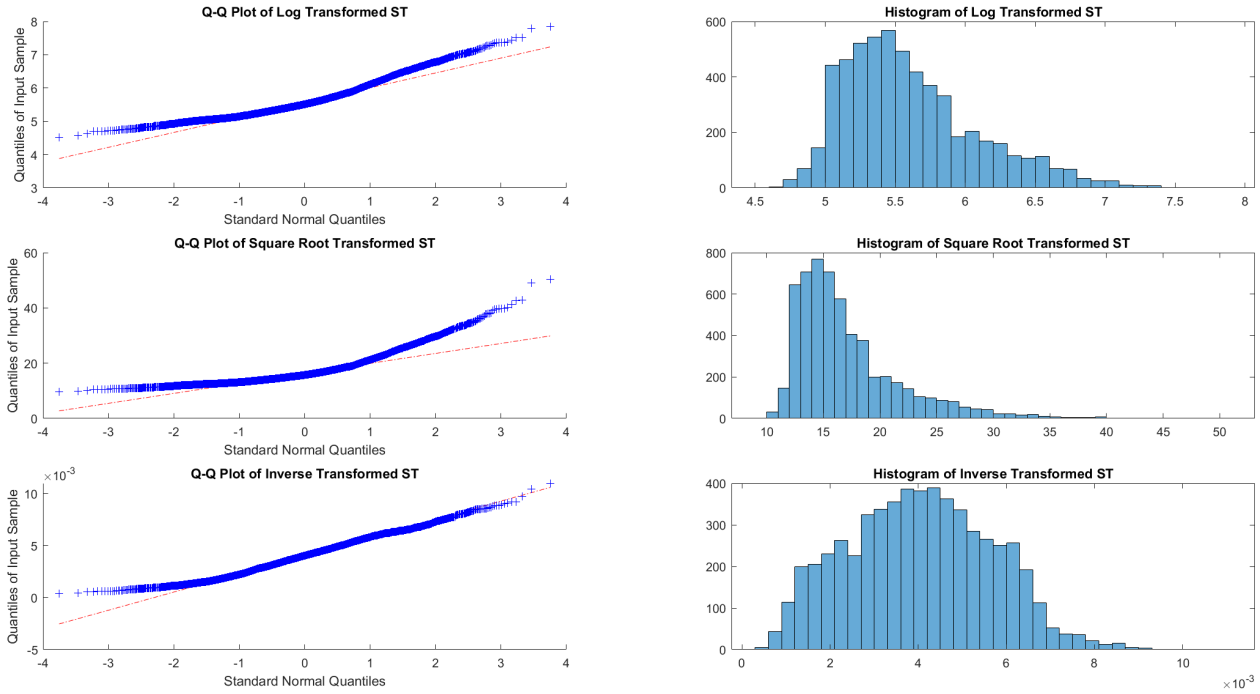


Figure 11: Q-Q plots and Histograms of Transformed ST

Normality of ST

As we can observe, inverse transformation will have a great effect on normality of ST. Now we can see the results again:

$$Y_i = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2} + \epsilon_i$$

Number of observations: 5709, Error degrees of freedom: 5706
Root Mean Squared Error: 0.00159
R-squared: 0.0584, Adjusted R-Squared: 0.0581
F-statistic vs. constant model: 177, p-value = 2.72e-75

	Coefficient	Standard Error	t-stat	p-value
Intercept	$\beta_0 = 0.005202$	7.5396e-05	68.995	0
Display Size (DS)	$\beta_1 = -0.00017625$	9.4559e-06	-18.639	2.5742e-75
Training Duration (TD)	$\beta_2 = -3.9002e-05$	1.4177e-05	-2.7511	0.0059585

Now if we plot Q-Q plot and histogram for residuals, we can observe there are normally distributed.

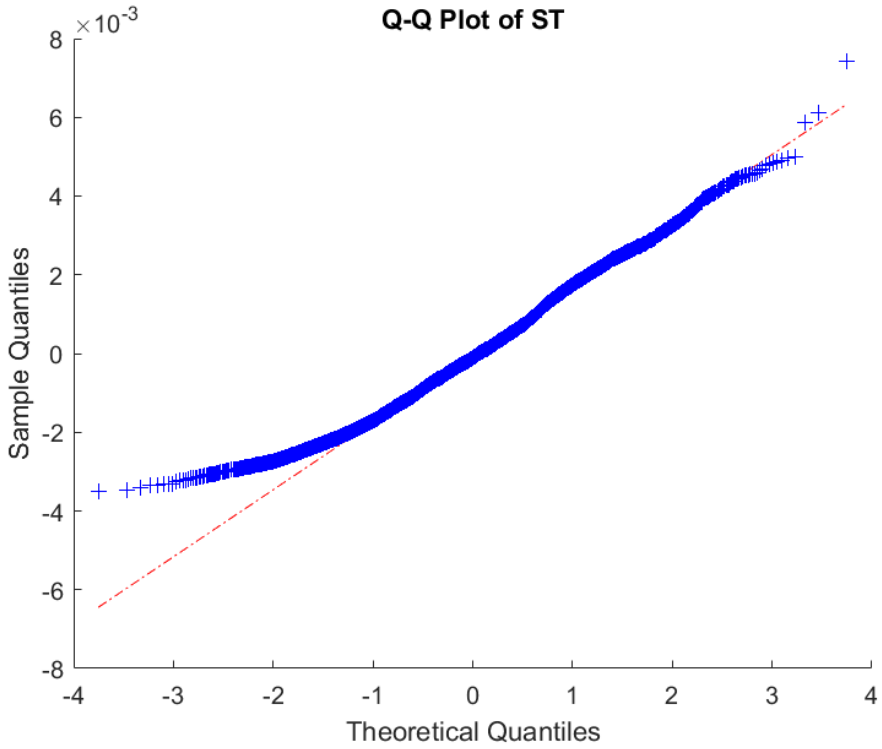


Figure 12: Q-Q plot of Residuals for Transformed ST

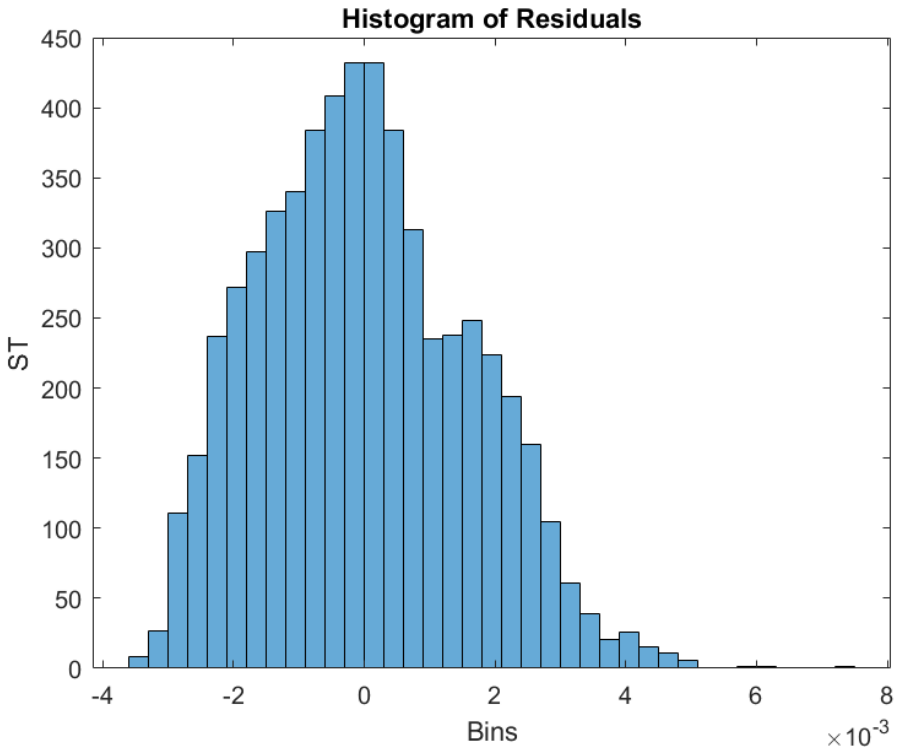


Figure 13: Histogram of Residuals for Transformed ST

6. ANOVA analysis for DS and TD:

a. Is this a fixed or random effect model? How may way ANOVA should be done?

Fixed Effect or Random Effect

Fixed Effects Model: In a fixed effects model, the levels of the factors are treated as fixed and specific to the study. The focus is on making inferences about the specific levels included in the analysis. Fixed effects models are appropriate when the levels of the factors are of direct interest, and the goal is to generalize to those specific levels.

Random Effects Model: In a random effects model, the levels of the factors are treated as a random sample from a larger population of possible levels. The focus is on making inferences about the broader population of levels from which the sample was drawn. Random effects models are appropriate when the levels of the factors are considered to be a random sample from a larger population, and the goal is to generalize beyond the specific levels included in the analysis.

This model is **fixed effect** because DS and TD had been tested in all their modes.

ANOVA

One-Way ANOVA: This is used when you have one categorical independent variable (factor) and one continuous dependent variable. It assesses whether there are statistically significant differences in the means of the dependent variable across the levels of the independent variable.

Two-Way ANOVA: This is used when you have two categorical independent variables (factors) and one continuous dependent variable. It assesses whether there are statistically significant main effects of each independent variable, as well as whether there is an interaction effect between the two independent variables.

We have two categorical independent variables (factors) and one continuous dependent variable, So we use **Two-Way ANOVA**

b. Do the ANOVA and report significance of each factor and their interaction.

anovan - MATLAB

We use `anpvan()` in MATLAB to see the results.

Source	SS	df	MS	F	Prob>F
X_1 (DS)	$1.9194e+07$	3	$6.3979e+06$	192.9043	$3.7419e-119$
X_2 (TD)	$1.4837e+07$	3	$4.9457e+06$	149.1180	$4.9268e-93$
Error	$1.8911e+08$	5702	$3.3166e+04$		
Total	$2.2200e+08$	5708			

At least one factor or interaction is significant ($p < 0.05$).

c. Based on the group level effect do post-hoc comparison using Tukey, Sheffe and Bonferroni comparison.

Comparisons Multcompare

We use `m = multcompare()` to do post-hoc comparison.

Multiple comparison procedure results, returned as a table with the following variables:

- Group1 - Values of the factors in the first comparison group
- Group2 - Values of the factors in the second comparison group
- MeanDifference - Difference in the marginal mean response between the observations in Group1 and the observations in Group2
- Lower - 95% lower confidence bound on the marginal mean difference
- Upper - 95% lower confidence bound on the marginal mean difference
- pValue - p-value corresponding to the null hypothesis that the marginal mean of Group1 is not statistically different from the mean of Group2. Based on P-Values all groups are significantly separable.

Tukey Post-Hoc Comparison

1.0000	2.0000	-64.2675	-47.2158	-30.1642	0.0000
1.0000	3.0000	-125.6388	-108.3826	-91.1265	0
1.0000	4.0000	-170.4724	-152.8363	-135.2002	0
2.0000	3.0000	-78.6462	-61.1668	-43.6874	0.0000
2.0000	4.0000	-123.4749	-105.6205	-87.7661	0
3.0000	4.0000	-62.4943	-44.4537	-26.4131	0.0000

Sheffe Post-Hoc Comparison

1.0000	2.0000	-65.7761	-47.2158	-28.6556	0.0000
1.0000	3.0000	-127.1654	-108.3826	-89.5998	0.0000
1.0000	4.0000	-172.0327	-152.8363	-133.6399	0.0000
2.0000	3.0000	-80.1926	-61.1668	-42.1410	0.0000
2.0000	4.0000	-125.0545	-105.6205	-86.1864	0.0000
3.0000	4.0000	-64.0903	-44.4537	-24.8170	0.0000

Bonferroni Post-Hoc Comparison

1.0000	2.0000	-64.7331	-47.2158	-29.6986	0.0000
1.0000	3.0000	-126.1099	-108.3826	-90.6553	0.0000
1.0000	4.0000	-170.9540	-152.8363	-134.7187	0.0000
2.0000	3.0000	-79.1234	-61.1668	-43.2102	0.0000
2.0000	4.0000	-123.9624	-105.6205	-87.2786	0.0000
3.0000	4.0000	-62.9869	-44.4537	-25.9205	0.0000

7. Add subject as a factor and repeat the ANOVA analysis (bonus):

a. Is subject fixed or random effect? Is it repeated measures?

Subject

It is fixed effect and repeated measures. because all 4 modes had been used in all experiments.

b. Do ANOVA and report significance of each factor and all interactions.

Anova

Source	**Sum Sq.**	**d.f.**	**Mean Sq.**	**F**	**Prob>F**
DS	$2.8718e+6$	3	$9.5728e+5$	30.1574	2.4593×10^{-19}
TD	$1.8392e+7$	3	$6.1307e+6$	193.1364	2.9218×10^{-119}
subject	$1.5311e+7$	3	$5.1038e+6$	160.7856	5.0203×10^{-100}
DS*TD	$1.6592e+6$	9	$1.8436e+5$	5.8078	4.3975×10^{-8}
DS*subject	$1.8195e+6$	9	$2.0217e+5$	6.3690	4.9034×10^{-9}
TD*subject	$2.7783e+6$	9	$3.0869e+5$	9.7249	6.7755×10^{-15}
Error	$1.8004e+8$	5672	$3.1743e+4$		
Total	$2.2200e+8$	5708			