PROJECT REPORT

STATISTICAL ANALYSIS OF MOUTH CANCER

B.Sc. STATISTICS SEMESTER-VI 2024-2025

DEPARTMENT OF STATISTICS

FACULTY OF SCIENCE

THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

GUIDE_TEACHER:
Dr. DEEPA KANDPAL
Mr. KARAN PRMAR

SUBMITTED BY :

KEERTHI CHANDHANA

AMEYA TRIVEDI

TWINKLE MAKWANA

TANVI SOJITRA

VIREN PRAJAPATI

DECLARATION

We hereby declare that the Project Report entitled "**STATISTICAL ANALYSIS OF MOUTH CANCER**" submitted to the Department of Statistics, The Maharaja Sayajirao University of Baroda in partial fulfilment of the requirements for the reward of the degree of BACHELOR OF SCIENCE is a record of original project work done by us, under the guidance of Dr. Deepa Kandpal and Mr. Karan Parmar.

The results embodied in this project work has not been submitted to any other University or Institute.

CERTIFICATE

This is to certify that Keerthi Chandhana, Ameya Trivedi, Tanvi Sojitra, Twinkle Makwana, Viren Prajapati have satisfactorily completed the project entitled:

“STATISTICAL ANALYSIS OF MOUTH CANCER”

As a team in the academic year 2024-25 and submitted the work to the department as a fulfillment for degree of Bachelor of Science in Statistics.

Throughout the semester they carried out work with sincerity and have presented on time and with enthusiasm.

I wish them grand success in future.

Dr. V. A. Kalamkar

(Head, Department of Statistics)

Dr.Deepa Kandpal
Mr. Karan Parmar

(Guide teacher)

ACKNOWLEGEMENT

We would like to thank our guide teacher Dr. Deepa Kandpal and Mr. Karan Parmar who helped us very much in doing our project, particularly in part of analysis and spending their valuable time for us.

We are thankful to Professor V. A. Kalamkar, Head, Department of Statistics and other teachers for their encouragement.

Finally, we thank everybody who has directly or indirectly contributed to the success of this project.

CONTENT :

1 INTRODUCTION :

1.1 : Background:

Q) Why do we need to study mouth cancer?

Mouth cancer, or oral cancer, is a significant public health concern, particularly in regions with high tobacco, alcohol, or betel nut consumption. Studying mouth cancer is essential to raise awareness about its risk factors, improve early detection, and reduce mortality rates. It helps in identifying vulnerable groups, guiding effective treatment strategies, and supporting public health campaigns. Through research, we can promote timely diagnosis, enhance patient outcomes, and ultimately contribute to saving lives.

1.2 : Motivation:

Q) What is the main motive and why this project?

The primary objective of this project is to predict the stage of mouth cancer using a combination of demographic, clinical, and treatment-related variables. By analysing factors such as gender, age, tumor size, weight loss during treatment, tumor location (LBM/RBM), and the type of treatment received (chemotherapy, radiation, surgery), the project aims to uncover meaningful patterns that contribute to the advancement of the disease.

Understanding how these variables are associated with different stages of mouth cancer can lead to more informed clinical decisions, better risk stratification, and personalized treatment planning. The goal is to support early diagnosis, improve patient outcomes, and raise awareness about the importance of timely intervention. Additionally, the findings can be used to guide public health strategies, resource allocation, and educational campaigns focused on mouth cancer prevention and management.

1.3 : Parameters considered in study:

1. Age

- Definition: Age of the patient at the time of diagnosis.
- Role in Study: Age is a significant factor in cancer prognosis, with the risk of mouth cancer generally increasing with age. Older patients may experience different responses to treatment, and their overall survival rates can be affected by age-related health conditions. This variable helps identify age-based trends in disease progression and treatment outcomes.

2. Gender

- Definition: The biological sex of the patient (Male/Female).
- Role in Study: Gender plays a role in the risk of developing mouth cancer due to varying exposure to risk factors (such as tobacco and alcohol use) and potential biological differences in how the disease progresses. This variable helps assess whether the stage of cancer or treatment response differs significantly between males and females.

3. Stage

- Definition: The stage of mouth cancer at diagnosis, classified from Stage I to Stage IV, where Stage I represents early disease, and Stage IV represents advanced, metastatic disease.

- Role in Study: The stage of cancer is the primary outcome variable in this study. It indicates the severity of cancer, its spread, and is crucial for determining treatment options. Predicting the stage based on patient characteristics helps in early diagnosis and treatment planning, ultimately improving patient outcomes.

4. Tumor Size

- Definition: The size of the tumor at the time of diagnosis, measured in centimeters.

- Role in Study: Tumor size is directly related to cancer staging, with larger tumors generally indicating a more advanced stage. Tracking tumor size is essential for understanding disease progression and the effectiveness of various treatment methods in reducing tumour size.

5. Treatment Type

- Definition: The type of treatment the patient receives, categorized into chemotherapy, radiation, or surgery.

- Role in Study: Treatment type plays a crucial role in influencing tumour regression, patient recovery, and side effects. The analysis of treatment types allows for the evaluation of which modalities are most effective in treating mouth cancer at various stages and how different treatments impact side effects, tumour reduction, and recovery.

6. Tumor Location (LBM/RBM)

- Definition: The anatomical location of the tumor, specifically on the Left Buccal Mucosa (LBM) or Right Buccal Mucosa (RBM).

- Role in Study: The location of the tumour within the mouth can affect its accessibility for treatment and the ease with which it can be surgically removed. It may also influence disease progression, with some locations being more prone to complications or metastasis. Analysing tumour location can provide insights into treatment challenges and help tailor patient care strategies.

7. Patient’s Weight (Before, After Treatment)

- Definition: The weight of the patient before and after treatment.

- Role in Study: Weight loss during treatment is a common side effect, often indicating poor nutrition or severe disease progression. Significant weight loss may be associated with advanced cancer stages or adverse reactions to treatment. Monitoring weight changes throughout the treatment process can help assess the patient's health status, track treatment efficacy, and provide insights into the overall well-being of the patient.

8. SGPT Levels (Serum Glutamic Pyruvic Transaminase)

- Definition: SGPT, also known as ALT (Alanine Aminotransferase), is an enzyme primarily found in the liver. Its levels are often measured to assess liver function.
- Role in Study: SGPT levels can be elevated in response to cancer, metastasis, or treatment-related liver toxicity, particularly from chemotherapy or radiation. By monitoring SGPT levels, this study can assess the impact of treatment on liver health, detect potential treatment-related complications early, and adjust therapies as needed to prevent further liver damage.

This section of the report provides a comprehensive explanation of the variables used in the study. Each variable plays a pivotal role in understanding the progression of mouth cancer, predicting the disease stage, and analysing treatment outcomes. By examining these factors together, the project aims to contribute to improved patient care, earlier diagnosis, and better-targeted treatments for mouth cancer.

2 OBJECTIVES :

- To study association between treatment and stage by evaluating if stage influences treatment choice and effectiveness

- To analyze the effect of different treatment types (surgery, chemotherapy, radiation) on weight loss and SGPT levels.

- To study whether gender is associated with likelihood of receiving a particular treatment.

- To predict cancer treatment using clinical variables.

3 DATA CONSIDERED FOR ANALYSIS :

The data considered for our project is for two years **2022-2024** ,there are **261** observations based on which we have done our analysis.

4 METHODOLOGY :

Patient records were manually reviewed.
Data was extracted from scanned hospital files.
Each record was manually entered into a structured dataset.

5 TOOLS AND SOFTWARE

STATISTICAL TOOLS

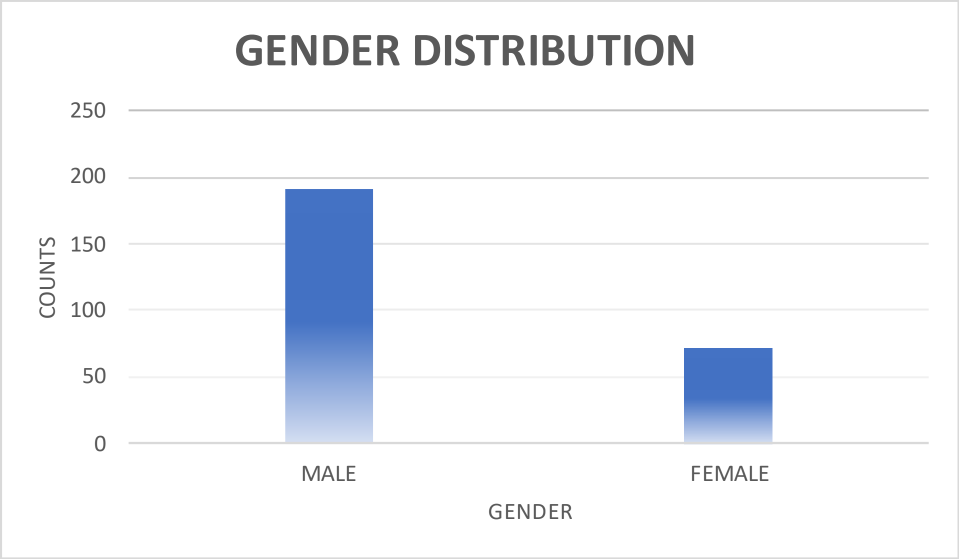
- Data Visualization
- Chi square test of association
- Odds ratio
- Non-Parametric Test
- Logistic and Multinomial Logistic Regression

SOFTWARE USED

- Microsoft Excel
- R Programming
- Python

6 DATA VISUALISATION

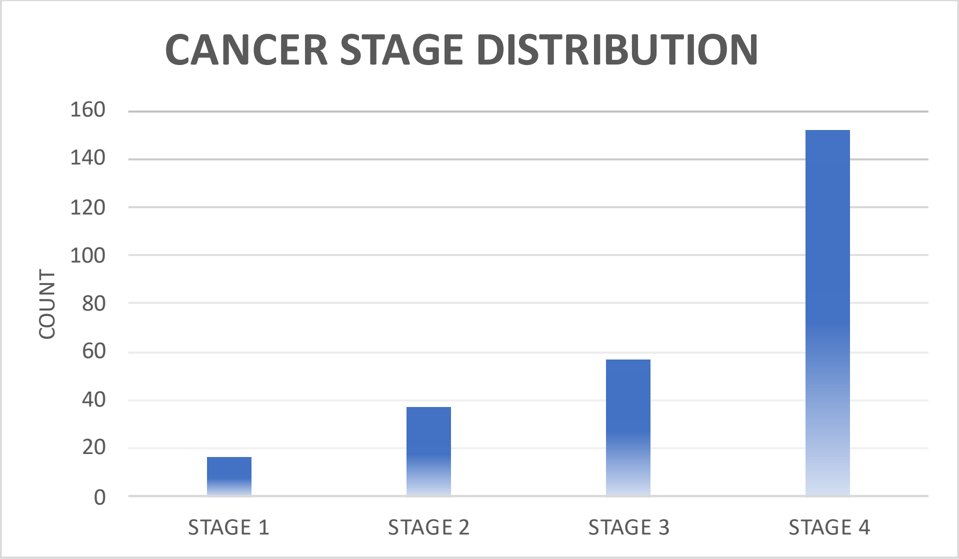
Gender Distribution of Mouth Cancer Patients



The bar chart illustrates the gender-wise distribution of mouth cancer cases. It is evident that males are disproportionately affected, accounting for nearly three times as many cases as females, chart confirms that mouth cancer is more prevalent in males than females in this dataset.

This significant gender gap may be attributed to higher exposure among males to risk factors such as tobacco use, alcohol consumption, and occupational hazards.

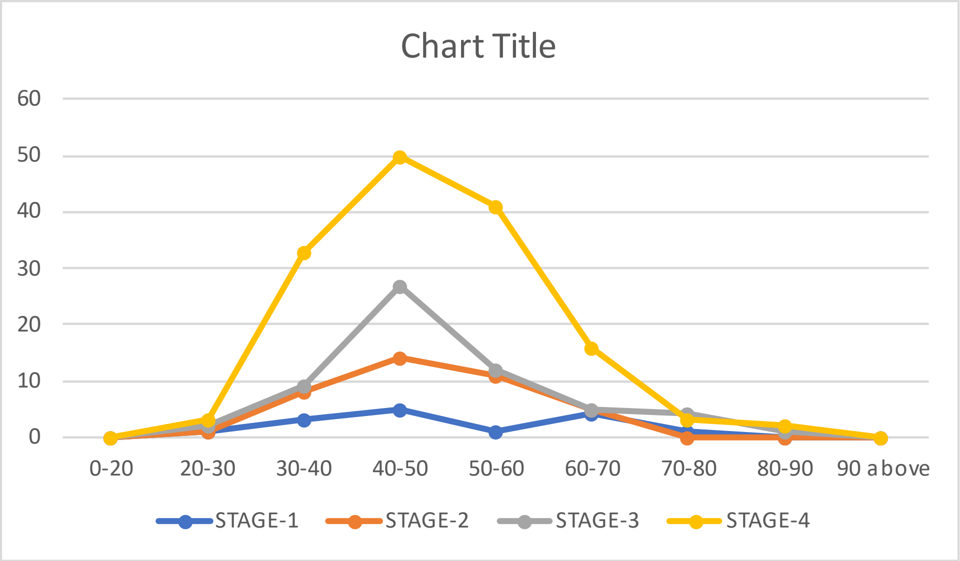
Distribution of Mouth Cancer by Stage



The chart shows a clear trend of late-stage diagnosis, with most patients diagnosed at Stage 4.

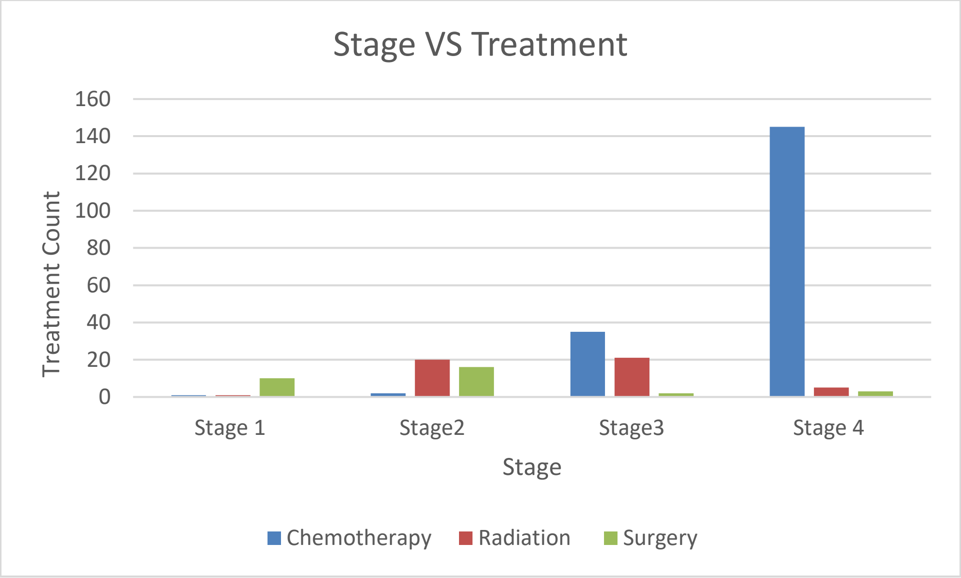
Early stages (Stage 1 and 2) have significantly fewer cases, indicating possible delays in detection and diagnosis.

Trend of Cancer Stages Across Age Groups



The line graph illustrates the distribution of cancer stages across different age groups.
The highest number of Stage 4 cases is observed in the 40–60 age range, highlighting a peak in advanced-stage diagnoses during middle age.
There is a sharp decline in cases after age 60, possibly due to lower survival or reduced reporting.

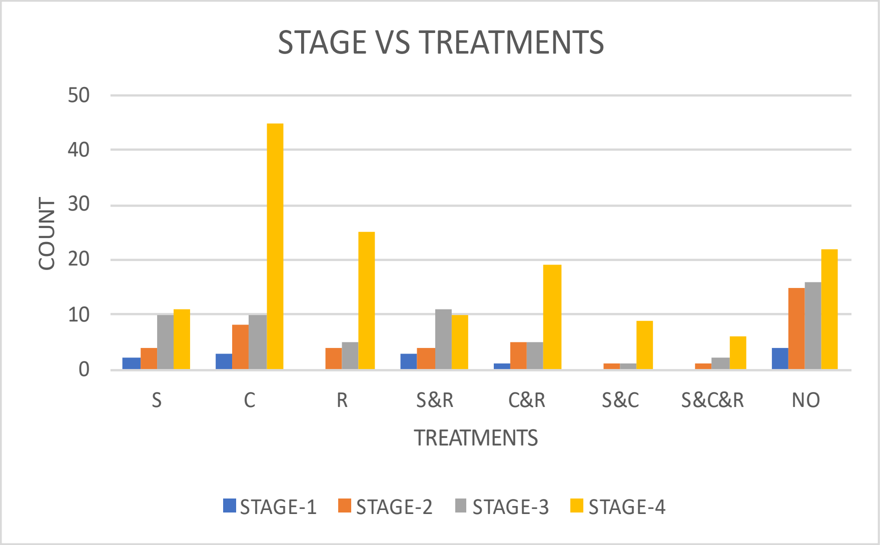
Distribution of Cancer Stages Across Treatment Types



The Stage vs Treatment chart illustrates the distribution of treatments—Chemotherapy, Radiation, and Surgery across four stages of a disease.

- Stage 1 and Stage 2- Minimal use of Chemotherapy, with moderate use of Radiation and Surgery.
- Stage 3-Chemotherapy usage increases noticeably, alongside a steady count for Surgery, while Radiation declines.
- Stage 4- Chemotherapy dominates as the primary treatment, reflecting its importance in advanced stages.

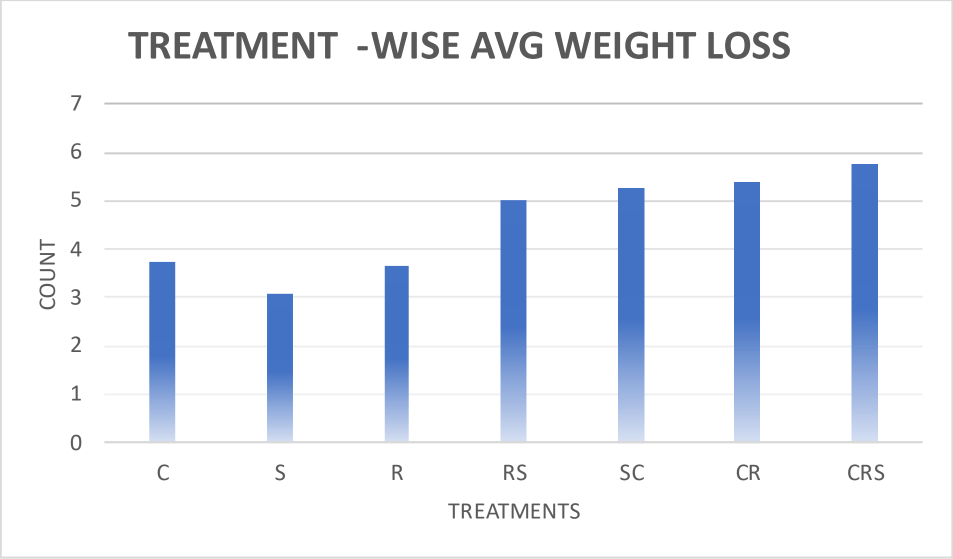
Distribution of Cancer Stages Across Treatment Types



The graph compares treatment effectiveness across four stages (Stage-1 to Stage-4).

- Combined treatments (especially S&C&R) perform best, with the highest bars in all stages.
- Effectiveness declines as stages progress suggesting early intervention is critical.
- Effectiveness declines as stages progress suggesting early intervention is critical.
- Single treatments (S, C, R) are less effective than combinations.

Average Weight Loss by Treatment Type

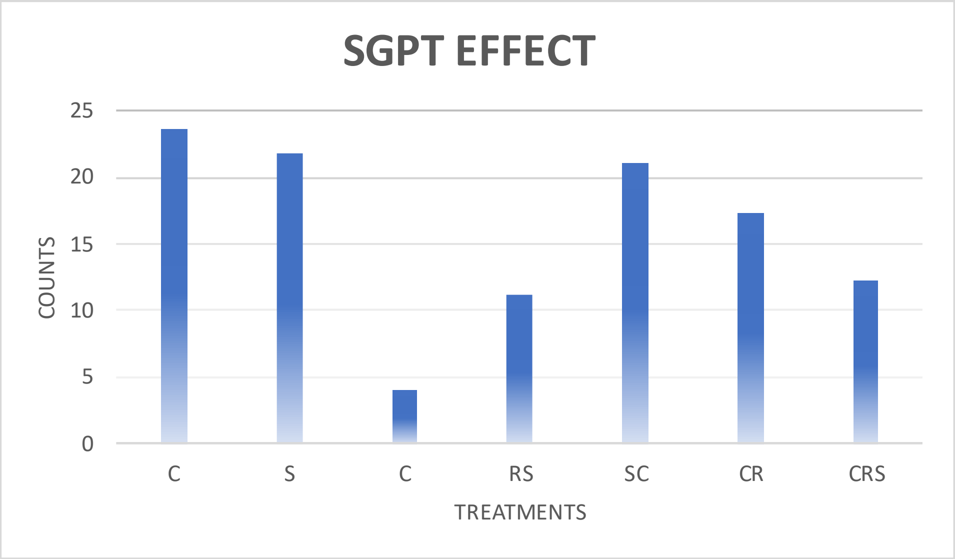


Average weight loss is higher in combination treatments (RS, SC, CR, CRS) compared to individual treatments (C, S, R).

CRS (Chemotherapy + Radiation + Surgery) shows the maximum weight loss, while Surgery (S) alone shows the least weight loss.

Thus, combined therapies lead to greater weight loss than single treatments.

Impact of Treatments on SGPT levels



Surgery and chemotherapy individually cause significant elevation in SGPT levels, indicating substantial liver stress. Radiation, when combined with surgery or chemotherapy, appears to moderate this effect, leading to lower SGPT levels.

The combination of chemotherapy, radiation, and surgery (CRS) shows the most notable reduction, suggesting a potential protective interaction of radiation against liver damage

7 STATISTICAL ANALYSIS

Chi Square Test of Association:

The Chi-Square Test of Association (also called the Chi-Square Test of Independence) is a statistical test used to determine whether there is a significant association between two categorical variables in a contingency table.

To examine whether there is a significant association between the type of treatment (chemotherapy, radiation, surgery) and the stage of mouth cancer (Stage I–IV), the Chi-Square Test of Association was used.

Since both variables are categorical, this test helps determine if the distribution of treatment types varies significantly across different cancer stages.

A significance level of 0.05 (alpha) will be applied uniformly across all tests. Our rejection criterion remains consistent across all tests: we will reject the null hypothesis if the p value does not exceed the alpha value.

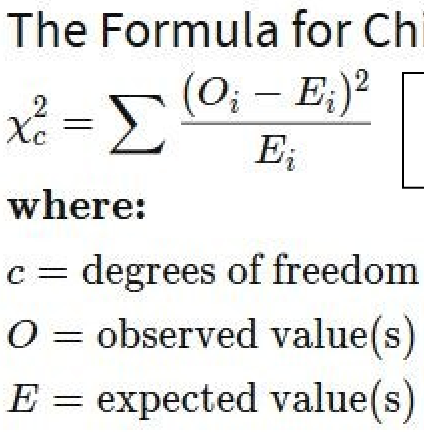
HO: There is no association between treatment & stage.
H1: There is association between treatment & stage.

Observed Frequency Table:

	Stage 1	Stage 2	Stage 3	Stage 4	Total
Chemotherapy	1	2	35	145	183
Radiation	1	20	21	5	47
Surgery	10	16	2	3	31
Total	12	38	58	153	261

Expected Frequency Table:

	Stage 1	Stage 2	Stage 3	Stage 4
Chemotherapy	8.41	26.64	40.67	107.27
Radiation	2.16	6.84	10.44	27.55
Surgery	1.42	4.51	6.88	18.17



The Formula for Chi Square Is

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = degrees of freedom

O = observed value(s)

E = expected value(s)

$\boxed{=195.677}$

χ^2 tab= χ^2 (r-1)(c-1),alpha

= χ^2 (3-1)(4-1),0.05

= χ^2 6,0.05

= 12.5916

Rej HO: When χ^2 calculated > χ^2 tabulated

Since 195.67>12.5916 hence we reject HO.
Hence we can conclude that there is association between treatment and stage at 5% level of significance.

Tschuprow's T Coefficient of association: Tschuprow's T is a measure of association between two categorical variables in a contingency table.

$$\hat{T} = \sqrt{\frac{\chi^2/n}{\sqrt{(r-1)(c-1)}}}. \qquad \boxed{=0.5532}$$

- χ2= Chi-square statistic
- n = Total number of observations
- r = Number of rows
- c = Number of columns

After confirming a statistically significant association between the stage of cancer and the type of treatment using the Chi-square test, we applied Tschuprow’s T to measure the strength of this association.

The resulting Tschuprow’s T value was **0.5532**, which indicates a strong association between the stage of cancer (Stage 1 to 4) and the treatment type (chemotherapy, radiation, surgery).

This means that as the cancer stage changes, there is a meaningful and strong pattern in how treatment types vary, validating that treatment decisions are likely influenced by the stage of cancer.

Wilcoxon Signed Rank Test :

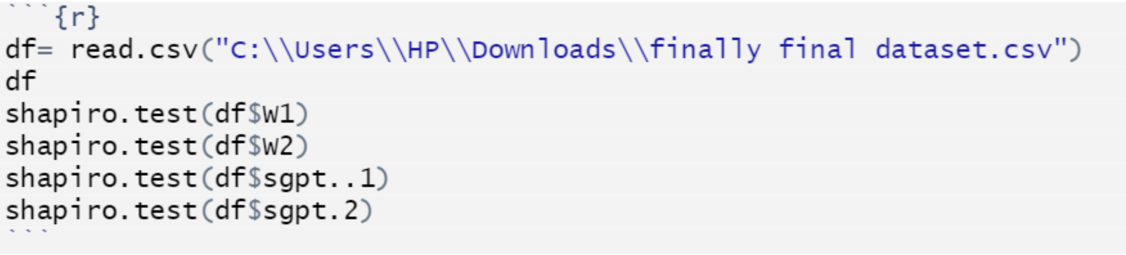
The Wilcoxon Signed-Rank Test is a non-parametric statistical test used to compare two related (paired) samples, such as before-and-after measurements, or matched subjects. It's the non-parametric alternative to the paired t-test and is used when the differences between paired data are not normally distributed.

Since the data for weight and SGPT levels before and after treatment did not follow a normal distribution, the Wilcoxon Signed-Rank Test was used. This test was used to assess whether there is a statistically significant change in patients weight and SGPT levels following treatment. This test was used to assess whether there is a statistically significant change in patients' weight and SGPT levels following treatment.

To check for normality, we conducted the Shapiro Test, result are as follows:

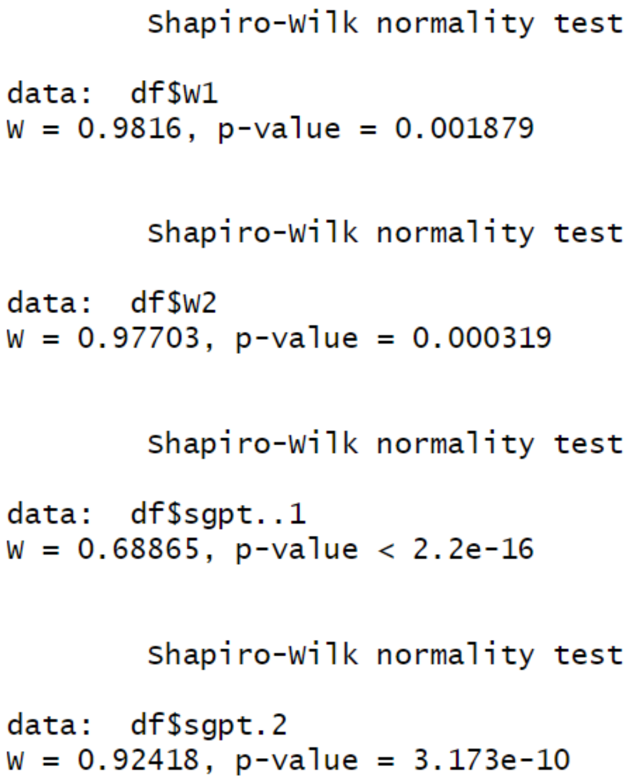
HO: Data follows Normal Distribution
H1: Data does not follows Normal Distribution

Rej HO: p-vale<alpha (0.05)



~~~
```{r}
df= read.csv("C:\\Users\\HP\\Downloads\\finally final dataset.csv")
df
shapiro.test(df$W1)
shapiro.test(df$W2)
shapiro.test(df$sgpt..1)
shapiro.test(df$sgpt.2)
```
~~~

Results:

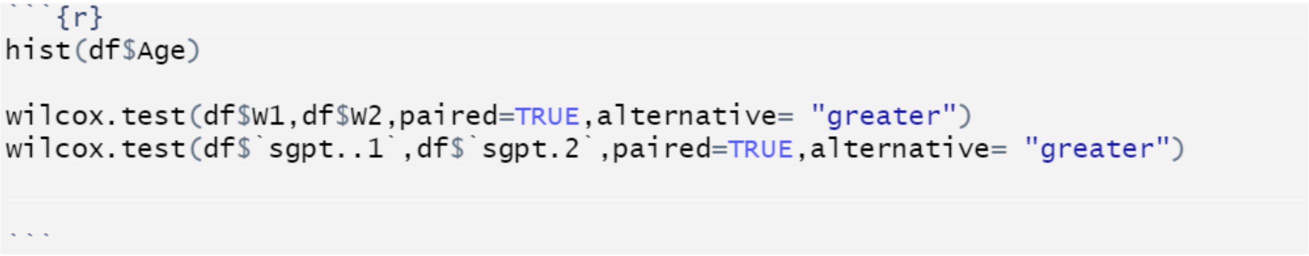


```
        Shapiro-Wilk normality test

data:  df$W1
W = 0.9816, p-value = 0.001879


        Shapiro-Wilk normality test

data:  df$W2
W = 0.97703, p-value = 0.000319


        Shapiro-Wilk normality test

data:  df$sgpt..1
W = 0.68865, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  df$sgpt.2
W = 0.92418, p-value = 3.173e-10
```

Interpretation:

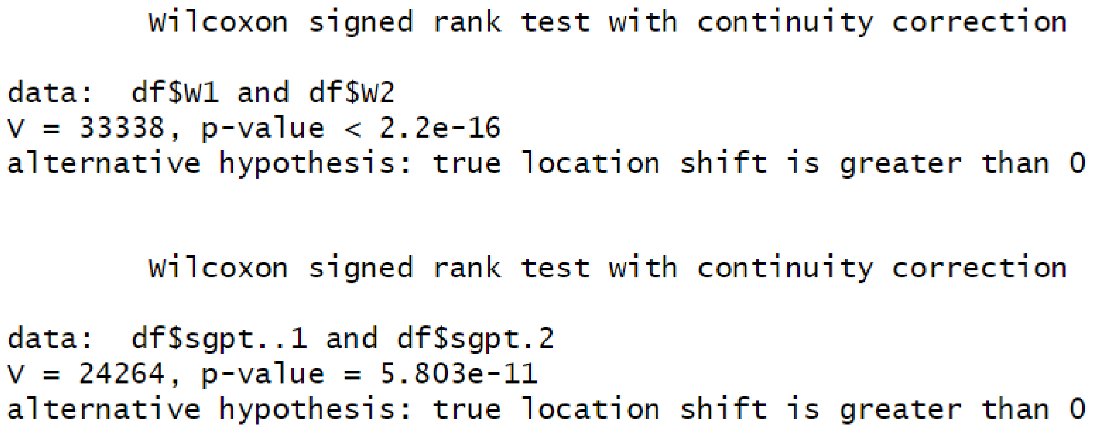
Since all the p- values are less than 0.05 therefore we reject H0 at 5% Level of Significance hence the data does not follow normal distribution.

Wilcoxon Signed rank test:



~~~
```{r}
hist(df$Age)

wilcox.test(df$W1,df$W2,paired=TRUE,alternative= "greater")
wilcox.test(df$`sgpt..1`,df$`sgpt.2`,paired=TRUE,alternative= "greater")


```
~~~

Results:



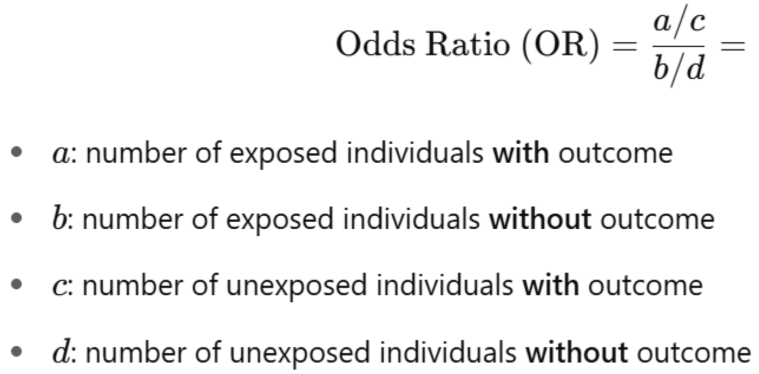
```
        Wilcoxon signed rank test with continuity correction

data:  df$W1 and df$W2
V = 33338, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0


        Wilcoxon signed rank test with continuity correction

data:  df$sgpt..1 and df$sgpt.2
V = 24264, p-value = 5.803e-11
alternative hypothesis: true location shift is greater than 0
```

We summarize our null hypothesis, alternative hypothesis, and test conclusions in the following table

NULL HYPOTHESIS	ALTERNATE HYPOTHESIS	p-value	CONCLUSION
There is no effect of treatment on SGPT levels S1<=S2	There is effect of treatment on SGPT levels S1>S2	5.803e-11	Hence we reject HO therefore there is effect of treatment on SGPT.
There is no effect of treatment on Weight loss W1<=W2.	There is effect of treatment on weight loss W1>W2	2.2e-16	Hence we reject HO therefore there is effect of treatment on weight loss.

Odds Ratio: Gender vs Treatment type

The Odds Ratio (OR) is a measure of association between an exposure and an outcome. It tells you how much more likely (or less likely) the outcome is to occur in the exposed group compared to the unexposed group.



$$\text{Odds Ratio (OR)} = \frac{a/c}{b/d} = \frac{a \cdot d}{b \cdot c}$$

- a: number of exposed individuals **with** outcome
- b: number of exposed individuals **without** outcome
- c: number of unexposed individuals **with** outcome
- d: number of unexposed individuals **without** outcome

To examine the relationship between gender and type of treatment received, the odds ratio (OR) was calculated. Both variables are categorical, and the odds ratio helps measure the strength and direction of association between them. Specifically, it indicates how the odds of receiving a particular treatment (e.g., chemotherapy) differ between male and female patients. An OR > 1 suggests higher odds for one gender, while an OR < 1 indicates lower odds, helping to uncover potential gender-based differences in treatment patterns.

	Chemotherapy	No Chemotherapy
Male	130	60
Female	51	20

Odds Ratio	0.8497

Interpretation: The odds ratio of 0.8497 indicates that males have lower odds of receiving chemotherapy as compared to females.

	Surgery	No Surgery
Male	21	169
Female	11	60

Odds Ratio	0.6778

Interpretation: The odds ratio of 0.6778 indicates that males have lower odds of receiving chemotherapy as compared to females.

	Radiation	No Radiation
Male	39	151
Female	9	62

Odds Ratio	1.7792

Interpretation: The odds ratio of 1.7792 indicates that males have higher odds of receiving chemotherapy as compared to females.

Predictive Modeling:

What is Multinomial Logistic Regression?

We use multinomial logistic regression when the response variable is categorical with more than two levels (i.e., more than two classes), and there’s no natural order among the categories.

Why are We using Multinomial Logistic Regression in our study?

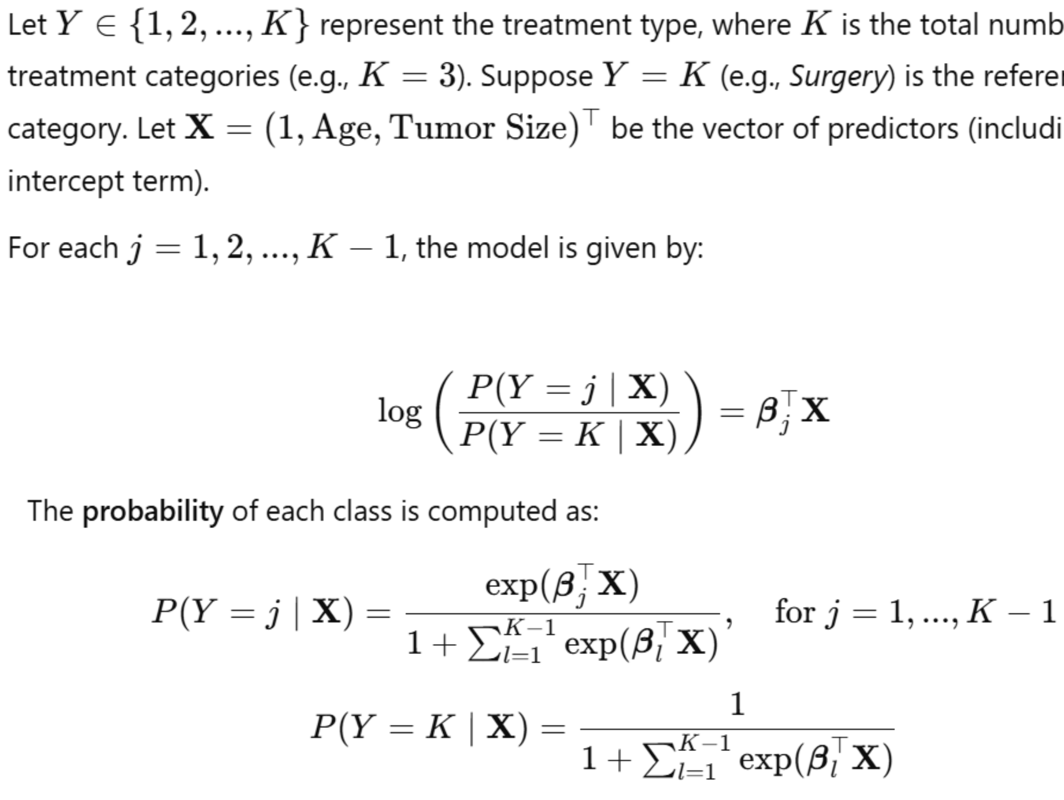
In this study, we utilize multinomial logistic regression to model the relationship between treatment types and patient-specific predictors in the context of cancer diagnosis. The response variable in our model is the type of treatment administered (i.e., Chemotherapy, Radiation, or Surgery), which is categorical and has no inherent order. Since the outcome variable involves more than two unordered categories, binary logistic regression is not applicable, and a linear model would violate the assumptions of normality and constant variance. Therefore, multinomial logistic regression serves as the most appropriate and interpretable approach.

We include two key independent variables:

- **Age**: The age of the patient (in years)
- **Tumor Size**: Volume of the tumor (in cm³)

These predictors are chosen based on their clinical relevance, as age and tumor size often influence treatment planning decisions.

Mathematical Formulation:



Let $Y \in \{1, 2, ..., K\}$ represent the treatment type, where K is the total number of treatment categories (e.g., $K = 3$). Suppose $Y = K$ (e.g., *Surgery*) is the reference category. Let $\mathbf{X} = (1, \mathrm{Age}, \mathrm{Tumor\ Size})^\top$ be the vector of predictors (including the intercept term).

For each $j = 1, 2, ..., K - 1$, the model is given by:

$$\log\left(\frac{P(Y = j \mid \mathbf{X})}{P(Y = K \mid \mathbf{X})}\right) = \boldsymbol{\beta}_j^\top \mathbf{X}$$

The **probability** of each class is computed as:

$$P(Y = j \mid \mathbf{X}) = \frac{\exp(\boldsymbol{\beta}_j^\top \mathbf{X})}{1 + \sum_{l=1}^{K-1} \exp(\boldsymbol{\beta}_l^\top \mathbf{X})}, \quad \text{for } j = 1, ..., K - 1$$

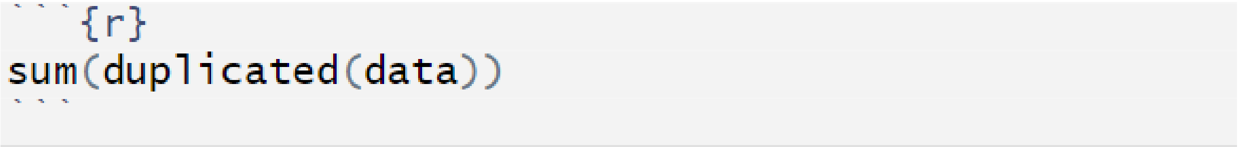
$$P(Y = K \mid \mathbf{X}) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\boldsymbol{\beta}_l^\top \mathbf{X})}$$

Before interpreting the model results, it is essential to verify whether the key assumptions of multinomial logistic regression are met. Below, we outline each assumption, how it was tested, and what we found.

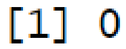
Assumptions of Multinomial Logistic Regression and Their Validation

1. Independence of Observations

Assumption: Each observation (i.e., patient) must be independent of the others. There should be no repeated measures or clustering.



~~~
```{r}
sum(duplicated(data))
```
~~~

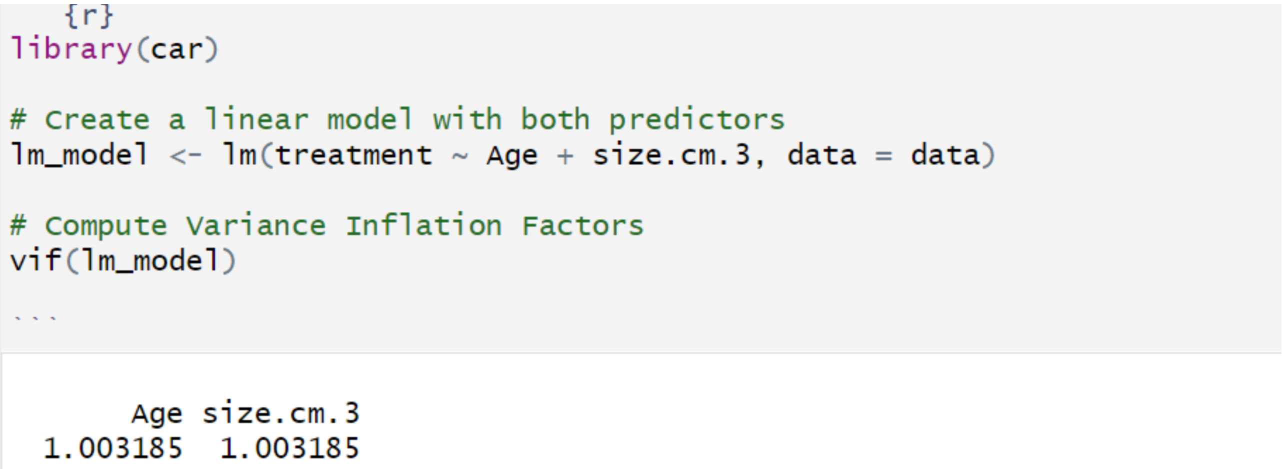


```
[1] 0
```

Result: We checked the dataset for duplicate rows using the duplicated() function in R and found zero duplicates. Each patient record is unique, and there is no evidence of clustering. Thus, the independence assumption is satisfied.

2. No Multicollinearity Between Predictors

Assumption: Predictors should not be highly correlated with each other. High multicollinearity can distort the estimated coefficients.



```
   {r}
library(car)

# Create a linear model with both predictors
lm_model <- lm(treatment ~ Age + size.cm.3, data = data)

# Compute Variance Inflation Factors
vif(lm_model)

```
```

```
      Age size.cm.3
 1.003185  1.003185
```

Result: The VIF values for both predictors were well below the threshold of 5, indicating that multicollinearity is not a concern. Therefore, this assumption is met
```

3. Linearity of the Log-Odds

Assumption: The relationship between each continuous predictor and the log-odds of the outcome should be linear.

Test: We visualized the relationship between predictors (Age and Tumor Size) and treatment categories by plotting smooth loess curves.

4. Sufficient Sample Size per Category

Assumption: Each treatment category should have a sufficient number of cases relative to the number of predictors.

Result: All treatment categories had a sufficient number of observations, ensuring model stability. Thus, this assumption is satisfied.

5. Outcome Variable is Nominal

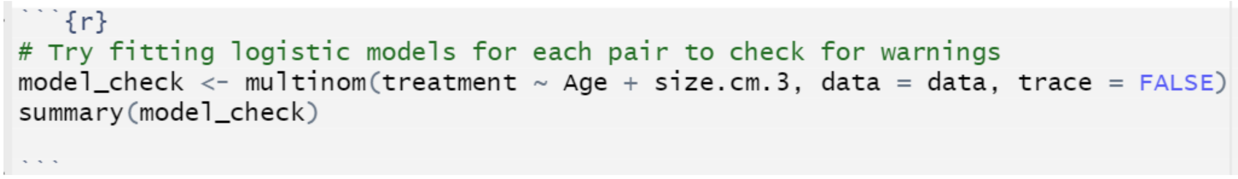
Assumption: The dependent variable should be categorical with no inherent order (nominal).

Result: Our response variable Treatment includes categories such as surgery, radiation, and chemotherapy which are distinct and unordered. This assumption is met.

6. No Perfect Separation

Assumption: No predictor should perfectly separate the treatment categories, which would make estimation unreliable.

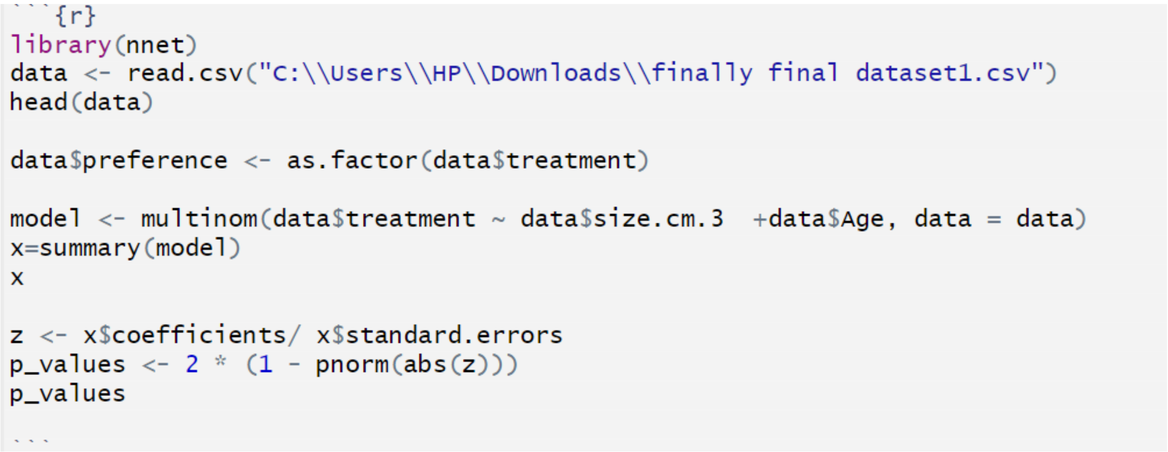
Test: The model was fit using the multinom() function from the nnet package. No warnings about perfect separation or convergence issues were observed.



~~~
```{r}
# Try fitting logistic models for each pair to check for warnings
model_check <- multinom(treatment ~ Age + size.cm.3, data = data, trace = FALSE)
summary(model_check)

```
~~~

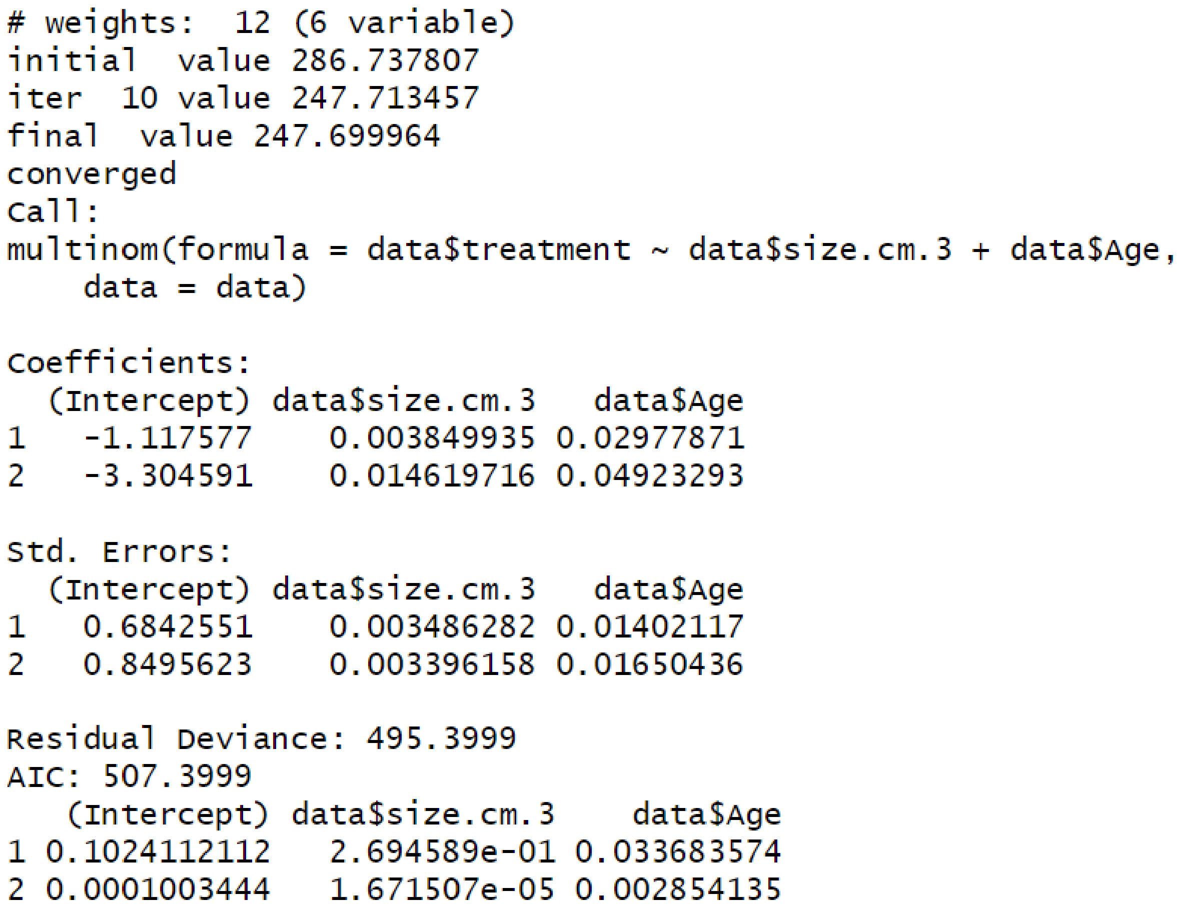
Result: There is no evidence of perfect separation. This assumption is met.

MODELLING:



~~~
```{r}
library(nnet)
data <- read.csv("C:\\Users\\HP\\Downloads\\finally final dataset1.csv")
head(data)

data$preference <- as.factor(data$treatment)

model <- multinom(data$treatment ~ data$size.cm.3  +data$Age, data = data)
x=summary(model)
x

z <- x$coefficients/ x$standard.errors
p_values <- 2 * (1 - pnorm(abs(z)))
p_values

```
~~~

RESULTS:



```
# weights:  12 (6 variable)
initial  value 286.737807
iter  10 value 247.713457
final  value 247.699964
converged
Call:
multinom(formula = data$treatment ~ data$size.cm.3 + data$Age,
    data = data)

Coefficients:
  (Intercept) data$size.cm.3   data$Age
1   -1.117577    0.003849935 0.02977871
2   -3.304591    0.014619716 0.04923293

Std. Errors:
  (Intercept) data$size.cm.3   data$Age
1   0.6842551    0.003486282 0.01402117
2   0.8495623    0.003396158 0.01650436

Residual Deviance: 495.3999
AIC: 507.3999
   (Intercept) data$size.cm.3    data$Age
1 0.1024112112   2.694589e-01 0.033683574
2 0.0001003444   1.671507e-05 0.002854135
```

Interpretation:

Category 1 vs. Category 0

- Intercept:

- -1.1176
  → Baseline log-odds of being in Treatment 1 compared to Treatment 0 when both predictors are 0.
- Tumor Size:
- 0.00385
  → For each 1 cm³ increase in tumor size, the log-odds of choosing Treatment 1 over Treatment 0 increases by 0.00385 (small but positive effect).
- Age:
- 0.02978
  → For each 1 year increase in age, the log-odds of choosing Treatment 1 over Treatment 0 increases by 0.0298.

Category 2 vs. Category 0

- Intercept:
- -3.3046
  → Lower baseline odds of choosing Treatment 2 compared to Treatment 0 when predictors are 0.
- Tumor Size:
- 0.01462
  → A stronger positive effect than Treatment 1: a 1 cm³ increase in tumor size increases the log-odds of choosing Treatment 2 over 0 by 0.0146.

- Age:
- 0.04923
  → A year increase in age increases log-odds of choosing Treatment 2 over 0 by 0.0492 — again, a stronger effect than in Treatment 1.

8 CONCLUSION AND FINDINGS:

This project conducted a comprehensive statistical analysis on 261 patient records diagnosed with mouth cancer between 2022 and 2024. The primary aim was to explore the association between demographic and clinical variables with the stage of cancer and to assess the effectiveness and side effects of various treatment types.

- Late-stage diagnosis dominates, with Stage 4 comprising the majority of cases. This suggests delayed detection and emphasizes the urgent need for awareness and screening programs.
- Chi-square analysis confirmed a strong association between cancer stage and treatment type, with chemotherapy being predominantly used in advanced stages. Tschuprow's T coefficient of 0.5532 further supported this finding.
- Wilcoxon Signed-Rank Tests showed statistically significant changes in both weight loss and SGPT levels post-treatment, highlighting the physiological toll of treatment and the need for supportive care.
- Odds ratio analysis uncovered gender-based disparities in treatment allocation, with males showing higher odds of receiving radiation but lower odds for chemotherapy and

surgery, indicating potential biases or clinical differences in treatment selection.

- Multinomial logistic regression confirmed that age and tumor size significantly predict the type of treatment administered, reinforcing the relevance of these clinical parameters in therapeutic decision-making.

In conclusion, this analysis emphasizes the importance of early diagnosis, tailored treatment planning, and continuous monitoring of patient health indicators such as weight and liver function. The results not only support clinical decision-making but also provide actionable insights for public health policies targeting mouth cancer management and prevention.

9 REFERENCES :

1. Tranby EP, Heaton LJ, Tomar SL, Kelly AL, Fager GL, Backley M, Frantsve-Hawley J. Oral Cancer Prevalence, Mortality, and Costs in Medicaid and Commercial Insurance Claims Data. Cancer Epidemiol Biomarkers Prev. 2022 Sep 2;31(9):1849-1857. doi: 10.1158/1055-9965.EPI-22-0114. PMID: 35732291; PMCID: PMC9437560.
2. Muthuvel, Marimuthu & .V, Keerthika & Professors, Assistant & Poobalan, C.P. Sri Chidambaram & Sreenath, Chidambaram & Kumar, R. (2018). Statistical Analysis of Cancer Data.
3. Jiang H, An L, Baladandayuthapani V, Auer PL. Classification, predictive modelling, and statistical analysis of cancer data (a). Cancer Inform. 2014 Sep 21;13(Suppl 2):1-3. doi: 10.4137/CIN.S19328. PMID: 25288874; PMCID: PMC4179642.
4. Blair, E. and R. Tibshirani (2003) Machine learning methods applied to DNA microarray data can improve the diagnosis of cancer, *SIGKDD Explorations*, **5** (2), 48–55.