Assignment1_BTC1877H

Leen Madani

2023-09-19

Question 1

In this question, our objective is to construct a mapping linear regression model that bridges the gap between two instruments measuring the quality of life and health-related outcomes. Specifically, we aim to establish a connection between the UCLA-PCI (University of California, Los Angeles Prostate Cancer Index), a descriptive instrument assessing the Health-Related Quality of Life in prostate cancer patients, and the PORPUS (Patient-Oriented Prostate Utility Scale) utility instrument.

Utility scores, which range from 1 (indicating perfect health) to 0 (representing deceased), are critical for conducting cost-effectiveness analyses. These scores can even assume values less than 0, indicating health states worse than being "dead." Obtaining utility scores typically involves using specific instruments like EQ5D or HUI. However, in some scenarios, data from such instruments may not be available, whereas data from other quality of life instruments that don't directly yield utility scores might be accessible.

Our goal is to create a "map" or relationship that allows us to estimate utility scores from the UCLA-PCI descriptive instrument when PORPUS data is unavailable. The dataset we will utilize, named "study3.csv," comprises various variables, including components of the UCLA-PCI and the final utility scores obtained from PORPUS-U.

Through this analysis, we will establish a model that provides valuable insights into the relationship between these instruments, enabling us to derive utility scores when needed.

Task 1 & 2:

Dataset Description and Data Preparation

Dataset Description and Data Preparation

The dataset used for this analysis, known as "study3," originally consisted of 676 rows and 42 columns, encompassing various aspects of patients' health and well-being. However, to align with the specific objectives outlined in the instructions, a subset of variables was selected for analysis: age, bowel and urinary function, sexual health, comorbidity index, and the utility score PORPUS

To prepare the dataset for analysis, the following key steps were undertaken:

1. Handling Missing Data: It was observed that missing values were represented by the character "." in the dataset. To address this, all instances of "." were replaced with "NA" to signify missing data.

2. Filtering for Complete Cases: To ensure the reliability of the dataset, rows with missing values in any of the selected variables were removed. It's noteworthy that 86 observations (equivalent to 12.7% of the dataset) had missing values. Importantly, each of the selected variables had less than 30% missingness, with the maximum being 8%. This justified the inclusion of all selected variables in the analysis.

3. Data Type Conversion: All variables, except for the comorbidity index, were converted from their original character data type to numeric data types. This transformation enhanced the dataset's analytical capabilities, enabling numerical calculations and statistical tests on these variables.

4. Comorbidity Index Categorization: A new variable was introduced to categorize the comorbidity index into four groups: 0, 1, 2, and 3+. This categorization facilitated the interpretation of findings related to comorbidity and allowed for exploration of its relationships with other selected variables.

5. Multiple Imputation and Statistical Testing: In addition to the complete case analysis, multiple imputation was employed to address missing data. Five imputed datasets were generated, and subsequent t-tests were conducted to assess whether statistically significant differences existed between the means of imputed and complete case datasets for each selected variable. Importantly, no statistically significant differences were observed, supporting the decision to proceed with the complete case analysis.

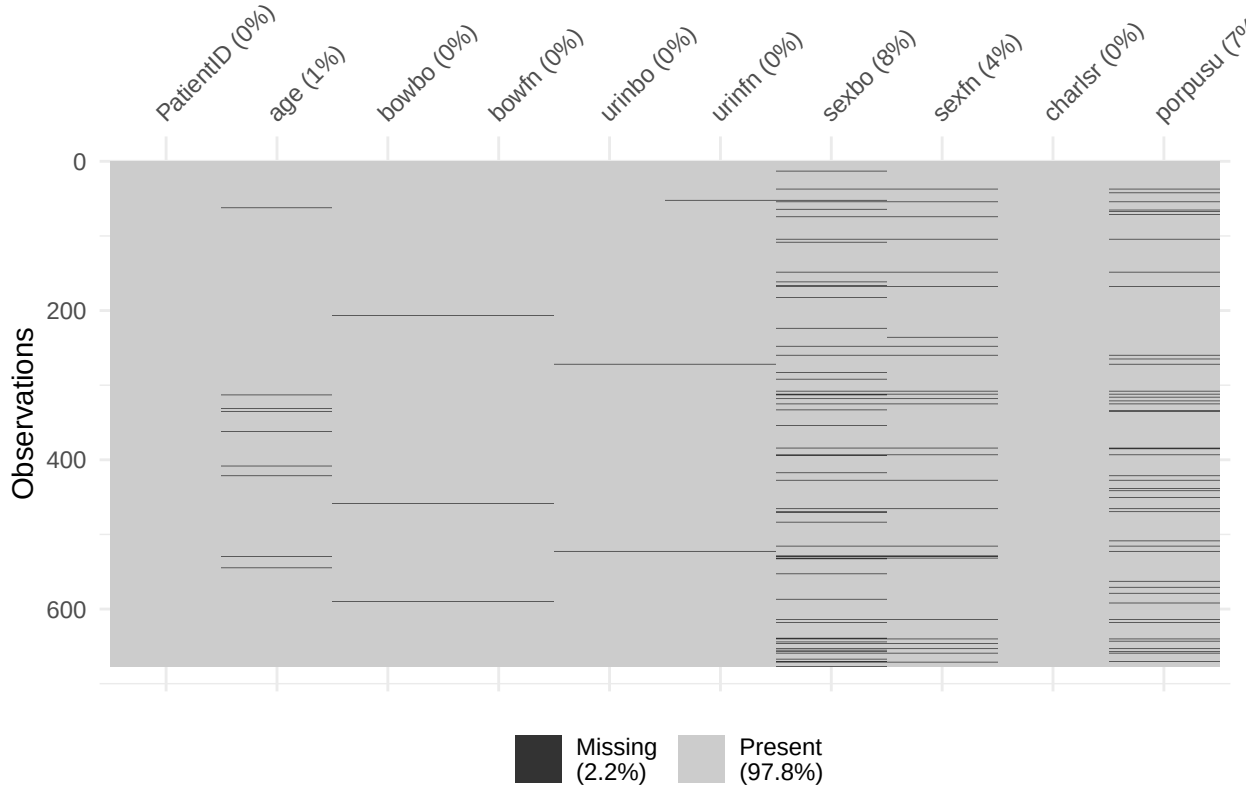


Figure 1: Visualization of the Amount of Missing Data.

6. Visualizing Missing Data: As an additional step, a visualization plot (figure 1) was generated to provide a visual representation of missing data distribution across the selected variables. This plot aided in assessing the extent of missing values in the chosen variables, and aided in the process of handling missing data.

Finally, these data preparation steps resulted in a refined dataset consisting of 590 rows and 11 columns, encompassing the specific variables of interest: age, bowel and urinary function, sexual health, comorbidity index, and the utility score PORPUS. This streamlined dataset is now well-suited for in-depth analysis and exploration of the relationships between these selected variables.

Task 2

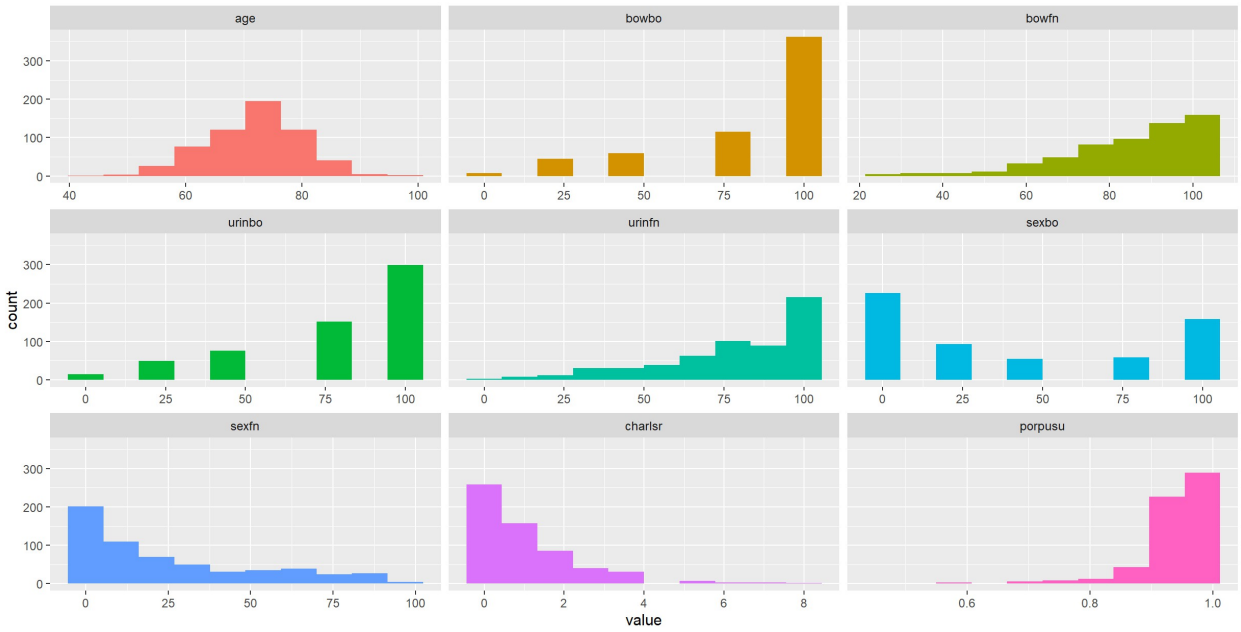


Figure 2: Visualization of the Nine Continuous Variables. "Age" represents patients' age in years. "Bowel Bother" (bowbo) and "Bowel Function" (bowfn) assess bowel discomfort and functionality on scales from 0 to 100, respectively, where higher values indicate improved function and less bother. "Urine Bother" (urinbo) and "Urine Function" (urinfn) measure urinary discomfort and functionality, with lower scores indicating less bother and better function. "Sexual Bother" (sexbo) and "Sexual Function" (sexfn) gauge sexual discomfort and functionality, with lower scores indicating reduced bother and higher functionality. The "Comorbidity Index" (charlsr) quantifies overall health comorbidity, while "PORPUS Utility" (porpusu) evaluates overall health and quality of life on a scale from 0 to 1, where higher values signify better well-being. Each histogram represents the values of the specific variables along the x-axis, with a default bin size of 10. The y-axis denotes the counts or frequency of occurrences. The dataset consists of 590 patients after dealing with missing entries.

Table 1: Descriptive Statistics for Key Variables

Variable	Mean	Standard Deviation	Variation Coef.	1st Quartile	Median	3rd Quartile	Skewness
Age	71.98	7.95	0.11	67.00	73.00	77.00	-0.32
Bowel Bother	83.05	25.50	0.31	75.00	100.00	100.00	-1.42
Bowel Function	84.71	16.26	0.19	75.00	91.71	100.00	-1.33
Urine Bother	78.52	26.97	0.34	75.00	100.00	100.00	-1.13
Urine Function	78.85	23.26	0.29	65.42	85.50	100.00	-1.07
Sex Bother	42.80	41.70	0.97	0.00	25.00	100.00	0.31
Sex Function	24.61	27.06	1.10	0.00	13.67	41.62	0.99
PORPUS	0.94	0.06	0.07	0.92	0.95	0.99	-2.88

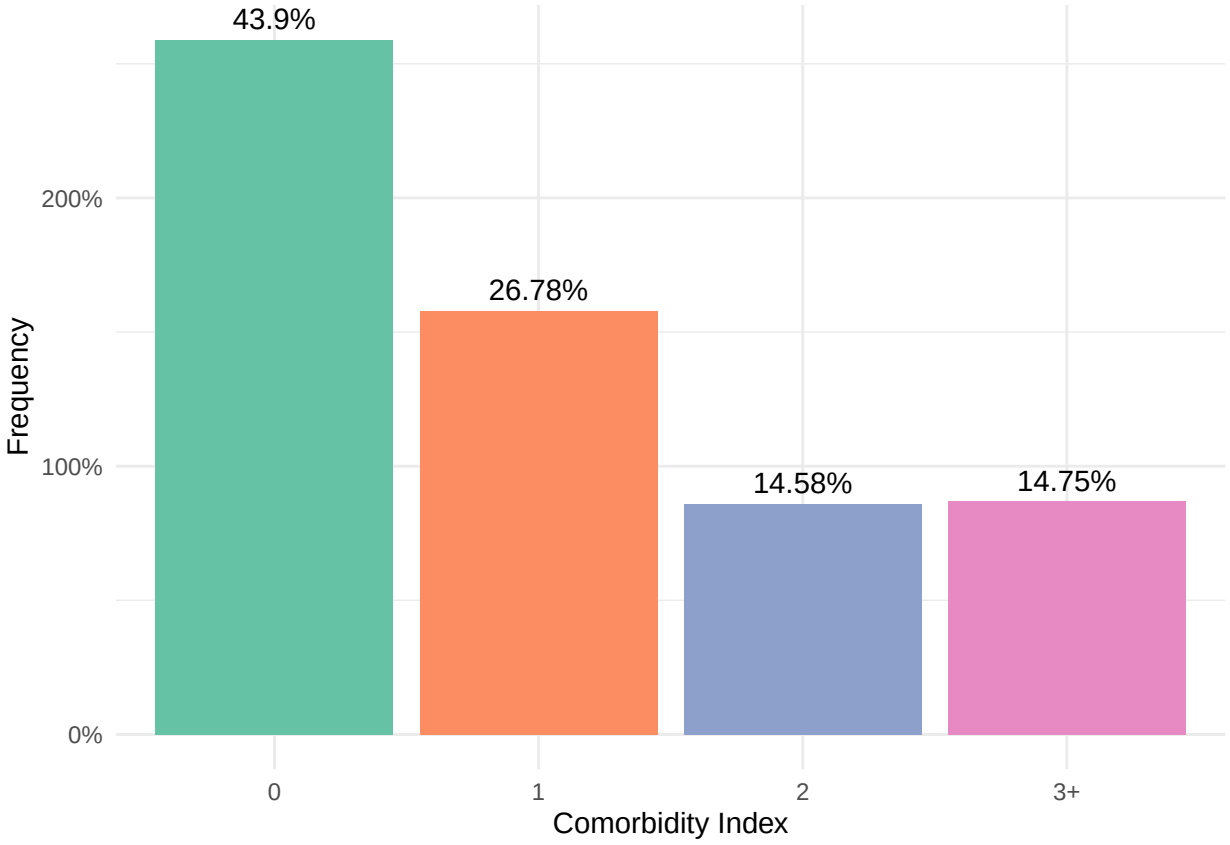


Figure 3: Frequency Distribution of Comorbidity Index. The comorbidity index has been categorized into four groups: 0, 1, 2, and 3+. Each group represents varying degrees of comorbidity, with '0' indicating no comorbid conditions and '3+' suggesting the presence of three or more comorbid conditions.

Descriptive Statistics Insights

Age: The participants in the study have an average age of approximately 72 years, with a minumum age of 43 years. This suggests that the sample consists mainly of older individuals. The standard deviation of around 8 indicates that the age values are relatively concentrated around the mean, with relatively low variability. The variation coefficient of 0.11 further supports this observation, indicating low variability relative to the mean. When looking at quartiles, we find that 25% of the participants are aged 67 or below, while 50% are aged 73 or below, and 75% are aged 77 or below. The slight negative skewness (-0.32) of the age distribution suggests a slightly longer tail on the left side, meaning that there are relatively fewer participants with ages significantly below the mean.It's important to note that there are 9 missing entries in the age variable.

Bowel Bother: On average, participants reported a Bowel Bother score of approximately 83, with higher values indicating more bother related to bowel function. The standard deviation of about 25 suggests considerable variability in Bowel Bother scores, indicating that responses vary widely. The variation coefficient of 0.31 indicates moderate variability relative to the mean. When examining quartiles, we find that scores range from a minimum of 75 (first quartile) to a maximum of 100 (third quartile), with some participants reporting the maximum bother score of 100. The distribution is negatively skewed (-1.42), indicating that more participants reported higher Bowel Bother scores, with a tail on the left side of the distribution. There are 3 missing entries in the Bowel Bother variable.

Bowel Function: Participants have an average Bowel Function score of approximately 85, implying better function. However, these scores exhibit moderate variability, with a standard deviation of about 16 and a variation coefficient of 0.19, suggesting moderate variability relative to the mean. Quartile analysis positions the first quartile at 75, the median at 91.71, and the third quartile at 100. The distribution is negatively skewed (-1.33), indicating that more participants have higher scores, reflecting better bowel function. Like Bowel Bother, there are 3 missing entries in the Bowel Function variable.

Urine Bother: The average Urine Bother score is approximately 78, with higher values indicating more bother. Scores display moderate variability, with a standard deviation of around 27 and a variation coefficient of 0.34, signifying moderate variability relative to the mean. Quartile analysis demonstrates scores ranging from 75 (first quartile) to 100 (third quartile), with some participants reporting the maximum bother score. The distribution is negatively skewed (-1.13), suggesting that more participants reported higher scores, indicating more bother.There are 2 missing entries in the Urine Bother variable.

Urine Function: Urine Function: Participants have an average Urine Function score of approximately 79, indicating better function. Scores vary moderately, with a standard deviation of about 23 and a variation coefficient of 0.29, indicating moderate variability relative to the mean. Quartile analysis positions the first quartile at 65.42, the median at 85.50, and the third quartile at 100. The distribution is slightly negatively skewed (-1.07), indicating that more participants have higher scores, reflecting better urine function. There are 3 missing entries in the Urine Function variable.

Sexual Bother: On average, participants reported a Sexual Bother score of approximately 43. However, these scores exhibit a high degree of variability, with a standard deviation of about 42, resulting in a variation coefficient of 0.97, indicating high variability relative to the mean. Quartile analysis places the first quartile at 0, the median at 25, and the third quartile at 100. The distribution is positively skewed (0.31), suggesting that more participants reported lower scores, implying significant sexual bother for some individuals. Notably, this variable has the highest level of missingness, with 8% of the total entries missing (55 entries in total). This observation could be explained by the fact that individuals are more likely to skip questions about sexual function due to feelings of embarrassment or discomfort, which are not captured in your dataset, and so the missingness can be considered MNAR (missing not at random).

Sexual Function: Participants have an average Sexual Function score of approximately 25. These scores vary widely, with a standard deviation of about 27 and a variation coefficient of 1.10, indicating high variability relative to the mean. Quartile analysis positions the first quartile at 0, the median at 13.67, and the third quartile at 41.62. The distribution is positively skewed (0.99), indicating that more participants reported lower scores, suggesting challenges in sexual function for some individuals. There are 27 missing entries in the Sexual Function variable.

Comorbidity Index: Based on figure 3, the most common comorbidity index among the participants is 0, accounting for 43.90% of the sample. This indicates that a significant proportion of participants in the study do not have any comorbid conditions, reflecting relatively good overall health. The second-largest group is 1, representing 26.78% of participants, suggesting the presence of one comorbid condition for this subgroup. Lastly, 14.25% of the participants have 2 comborbid conditions and 14.75% of the participants have 3+ comorbid conditions.

PORPUS Utility: The average PORPUS Utility score is approximately 0.94, signifying a relatively high health-related quality of life on average. Scores are closely clustered around the mean, with a standard deviation of 0.06 and a low variation coefficient of 0.07, indicating low variability relative to the mean. Quartile analysis positions scores ranging from 0.92 (first quartile) to 0.95 (median) to 0.99 (third quartile), showing that most participants report high PORPUS Utility scores. However, it's important to note that the distribution is negatively skewed (-2.88), indicating that some participants reported lower scores, potentially reflecting lower health-related quality of life for a subset of individuals. Notably, there are 44 missing entries in the PORPUS Utility scores.

Task 4 & 5:

Linear Regression Model Findings and Interpretation:

Null Hypothesis (H0): There is no relationship between the predictors (bowbo, bowfn, urinbo, urinfn, sexbo, sexfn, age, charlsr_catog) and the log-transformed PORPUS Utility scores.

Alternative Hypothesis (H1): There is a relationship between the predictors (bowbo, bowfn, urinbo, urinfn, sexbo, sexfn, age, charlsr_catog) and the log-transformed PORPUS Utility scores.

- Intercept (β_0): The estimated intercept is approximately -0.2388. This represents the estimated log-transformed PORPUS Utility score when all other predictor variables are zero.

Coefficients:

- bowbo (β_1): The coefficient for "bowbo" is approximately 0.0002012. This implies that for every one-unit increase in the "bowbo" score, the estimated expected log-transformed PORPUS Utility score increases by approximately 0.0002012 units, holding other predictors constant.
- bowfn (β_2): The coefficient for "bowfn" is approximately 0.00112. This means that for every one-unit increase in the "bowfn" score, the estimated expected log-transformed PORPUS Utility score increases by approximately 0.00112 units, keeping other predictors constant.
- urinbo (β_3): The coefficient for "urinbo" is very small: approximately 0.00001887. This suggests that changes in "urinbo" have a minimal impact on the log-transformed PORPUS Utility score.
- urinfn (β_4): The coefficient for "urinfn" is approximately 0.0005508. For every one-unit increase in "urinfn," the estimated expected log-transformed PORPUS Utility score increases by approximately 0.0005508 units, holding other predictors constant.

- sexbo (β_5): The coefficient for "sexbo" is approximately 0.00003428, signifying a positive relationship. A one-unit increase in "sexbo" leads to an estimated increase of approximately 0.00003428 units in the log-transformed PORPUS Utility score, with other predictors held constant.
- sexfn (β_6): The coefficient for "sexfn" is approximately 0.0008862, signifying a positive relationship. For every one-unit increase in "sexfn," the estimated expected log-transformed PORPUS Utility score increases by approximately 0.0008862 units, keeping other predictors constant
- age (β_7): The coefficient for "age" is approximately -0.00001977. This suggests that for every one-unit increase in age, the estimated expected log-transformed PORPUS Utility score decreases by approximately 0.00001977 units, holding other predictors constant.
- charlsr_categ1(β_8): The coefficient is 0.002967. For individuals with a comorbidity index of 1, the estimated expected log-transformed PORPUS Utility score increases by approximately 0.002967 units, compared to those with a comorbidity index of 0 (the reference category), holding other predictors constant.
- charlsr_categ2 (β_9): The coefficient is approximately -0.005722. For individuals with a comorbidity index of 2, the estimated expected log-transformed PORPUS Utility score decreases by approximately 0.005722 units, compared to those with a comorbidity index of 0 (the reference category), holding other predictors constant. This implies that having a comorbidity index of 2 is associated with a slight decrease in the expected health utility score.
- charlsr_categ3+ (β_9): The coefficient is approximately -0.02876. For individuals with a comorbidity index of 3 or higher, the estimated expected log-transformed PORPUS Utility score decreases by approximately 0.02876 units, compared to those with a comorbidity index of 0 (the reference category), holding other predictors constant. This indicates that having a comorbidity index of 3 or higher is associated with a notable decrease in the expected health utility score.

P-values:

> The p-values associated with each coefficient test the null hypothesis that the respective coefficient is equal to zero (H0: $\beta_i = 0$). If the p-value is less than the chosen significance level (commonly 0.05), we reject the null hypothesis.

Based on the p-values: Bowel Bother (bowbo), Bowel Function (bowfn), Urinary Function (urinfn), Sexual Function (sexfn), Comorbidity Index 1 (charlsr_categ1), and Comorbidity Index 3+ (charlsr_categ3+) have p-values less than 0.05, indicating that they are statistically significant predictors of the log-transformed PORPUS Utility score.

Urinary Bother (urinbo), Sexual Bother (sexbo), Age (age), and Comorbidity Index 2 (charlsr_categ2) have p-values greater than 0.05, suggesting that they are not statistically significant predictors.

Task 6

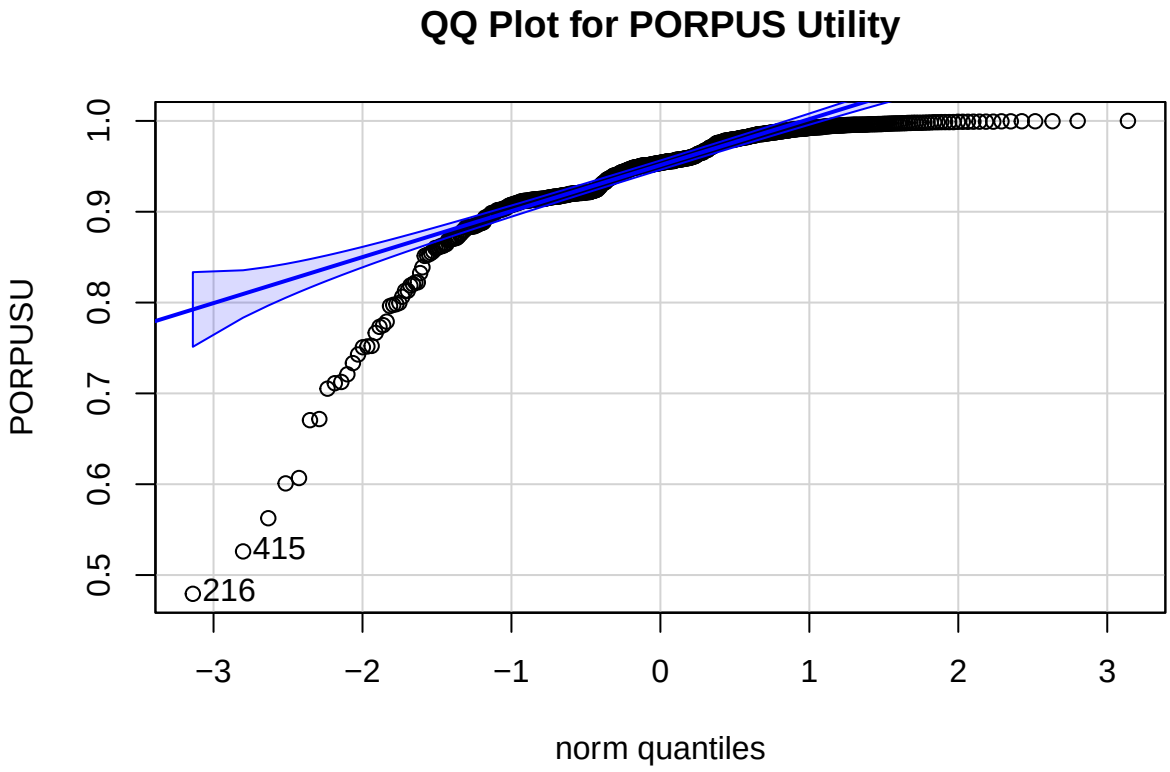
Prediction for a new patient

The predicted mean expected health utility for the patient with the specified characteristics is approximately 0.9074. The corresponding 95% confidence interval for this prediction, represented by the lower and upper bounds of 0.922023 and 0.8929, respectively, provides a range within which we can reasonably expect the true mean expected health utility to fall. In other words, if we were to conduct this analysis multiple times with different samples of patients who share the same characteristics, we would anticipate that about 95% of those intervals calculated would contain the true population parameter, which is the mean expected health utility for patients with these specific attributes.

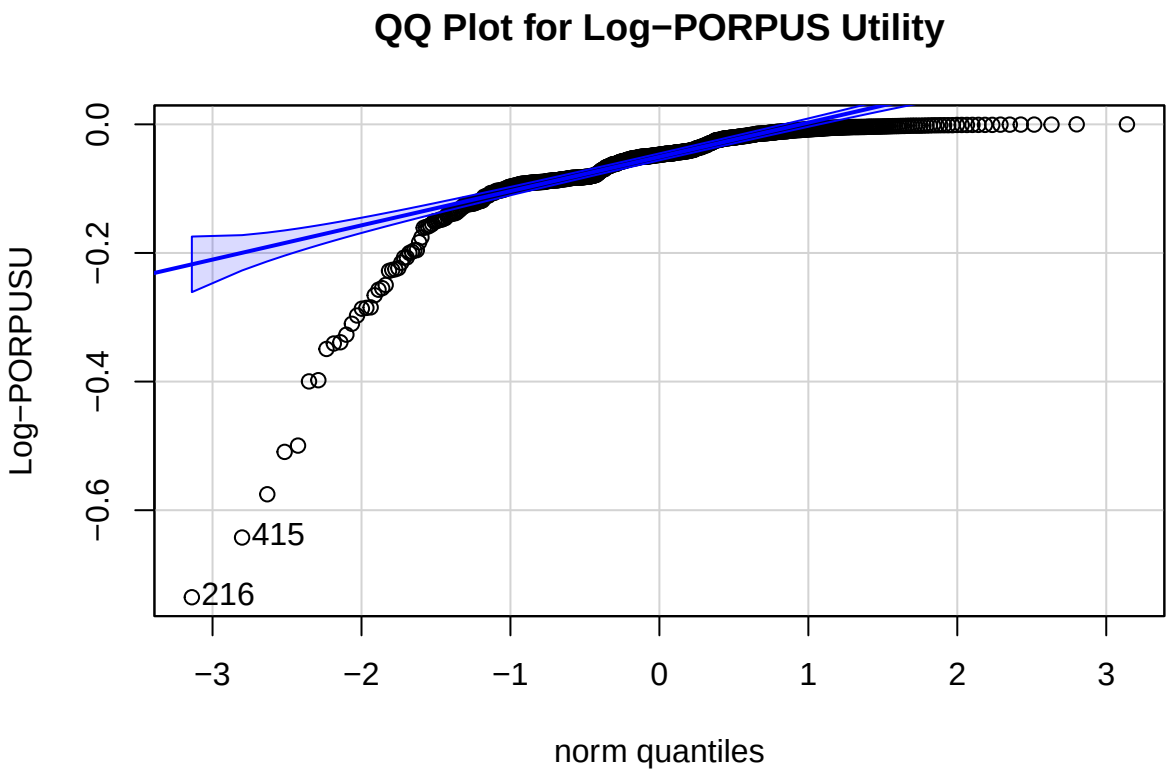
Task 7

Discuss different issues around the appropriateness of linear regression for this application.

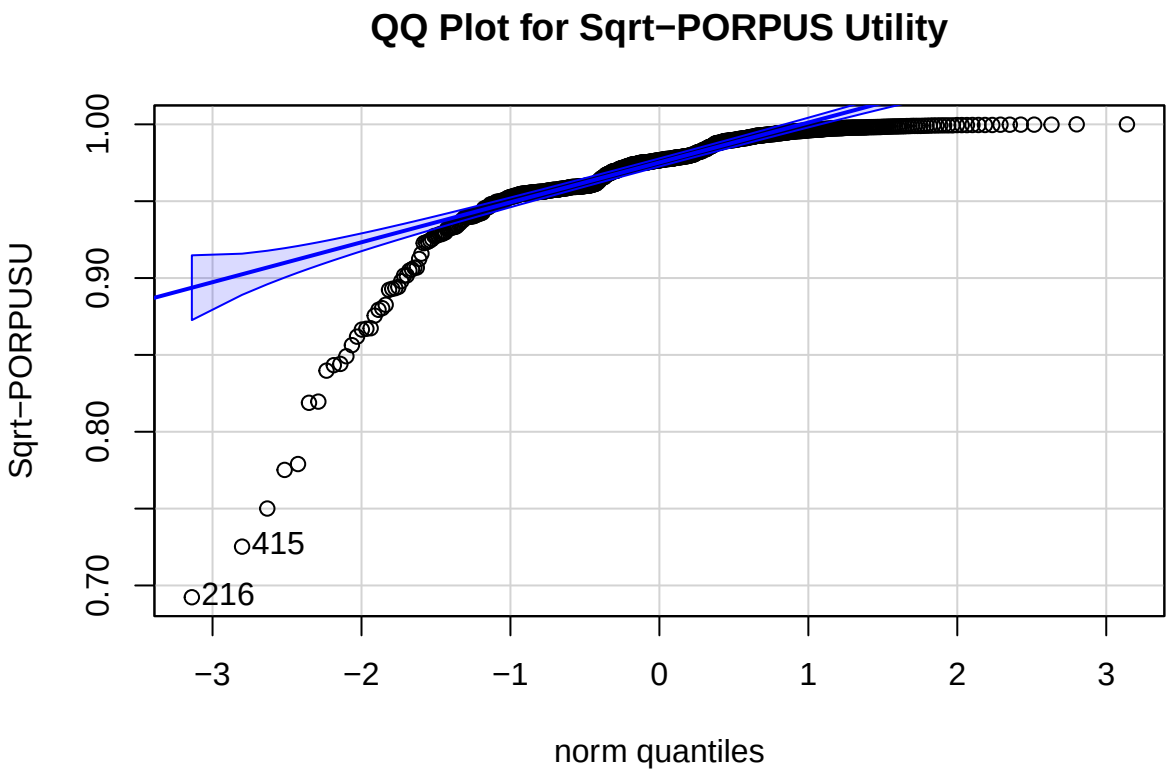
- In this study, we employed linear regression to model the relationship between PORPUS utility and various predictor variables in prostate cancer patients. While linear regression is a widely used statistical technique, its appropriateness for this application is subject to certain considerations.
- Firstly, linear regression assumes a linear relationship between the dependent and independent variables. In the context of Porpusu, this assumption may not always hold true, as the relationship can be more complex. To address this, we should explore the linearity assumption through diagnostic plots and consider alternative models if needed. . I have conducted a Shapiro-Wilk test for PORPUSU, which yielded a p-value of less than 0.05 (2.2e-16), indicating that the distribution significantly departs from normality. To further investigate, I performed QQ plots for PORPUS Utility in its original form, the selected log-transformed form, and square root-transformed form. These plots show that not all data points fall within the Q-line, suggesting that the data may not follow a normal distribution.




```
## [1] 216 415
```


QQ Plot for Log−PORPUS Utility
Log−PORPUSU
0.0
−0.2
−0.4
−0.6
415
216
−3
−2
−1
0
1
2
3
norm quantiles

```
## [1] 216 415
```




```
## [1] 216 415
```

- Secondly, linear regression assumes that the residuals are normally distributed and have constant variance (homoscedasticity). Violations of these assumptions can lead to biased or inefficient parameter estimates. It is crucial to check for homoscedasticity in the residuals and apply transformations or consider different models if necessary. In this assignment, I tried different transformations and concluded that log may be the best option for the negative skewness of the PORPUSU variable.
- Finally, issues such as multicollinearity, outliers, and missing data can affect the reliability of linear regression results. Using techniques like VIF or consider reducing the number of predictors if necessary for multicollinearity and proper handling of missing values are essential steps in mitigating these issues.

Question 2

In question 2, we turn our attention to predicting the risk of developing acute graft-versus-host disease (GvHD) in leukemia patients who have received nondepleted allogeneic bone marrow transplants. GvHD is a common complication for allogeneic hematopoietic stem cell transplants, as it arises when the immune cells from the donor graft perceive the recipient's tissues as foreign and mount an attack (Socié, 2014).

One of the variables of interest in our investigation is the "index," a numeric variable representing mixed epidermal cell-lymphocyte reactions (MLR). MLR is rooted in the observation that when normal peripheral blood leukocytes from different donors are combined, they stimulate each other to proliferate. This index holds potential as a predictor of GvHD risk (Zhou, 2014).

Additionally, we consider other variables such as recipient age (rcpage), donor age (donage), leukemia type (type), and whether the donor has been pregnant (preg). GvHD, a binary outcome variable (0 for absence, 1 for presence), serves as our primary focus. Through logistic regression analysis, we aim to gain insights into the factors influencing the occurrence of GvHD.

Task 1, 2, & 3

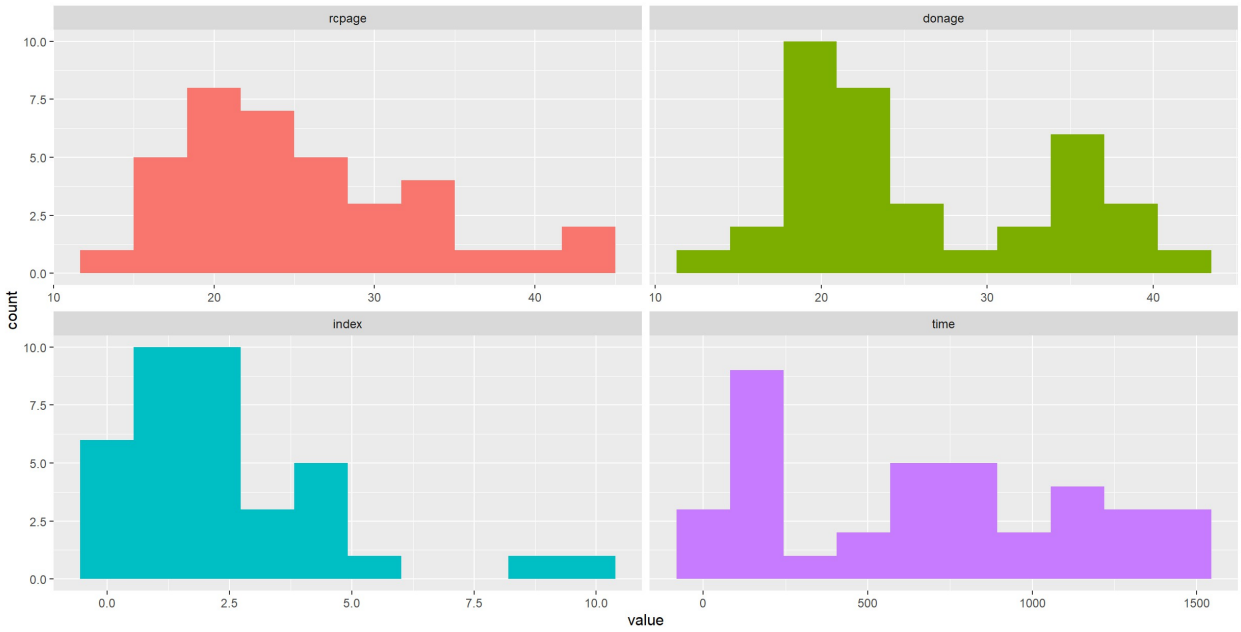


Figure 4: Visualization of the Four Continuous Variables. "Rcpage" corresponds to the age of the recipient (in years); "donage" corresponds to the age of the donor (in years); "index" refers to the index of mixed epidermal cell-leukocyte reaction; "time" is follow-up time in days. Each histogram represents the values of the specific variables along the x-axis, with a default bin size of 10. The y-axis denotes the counts or frequency of occurrences. The dataset consists of 37 patients.

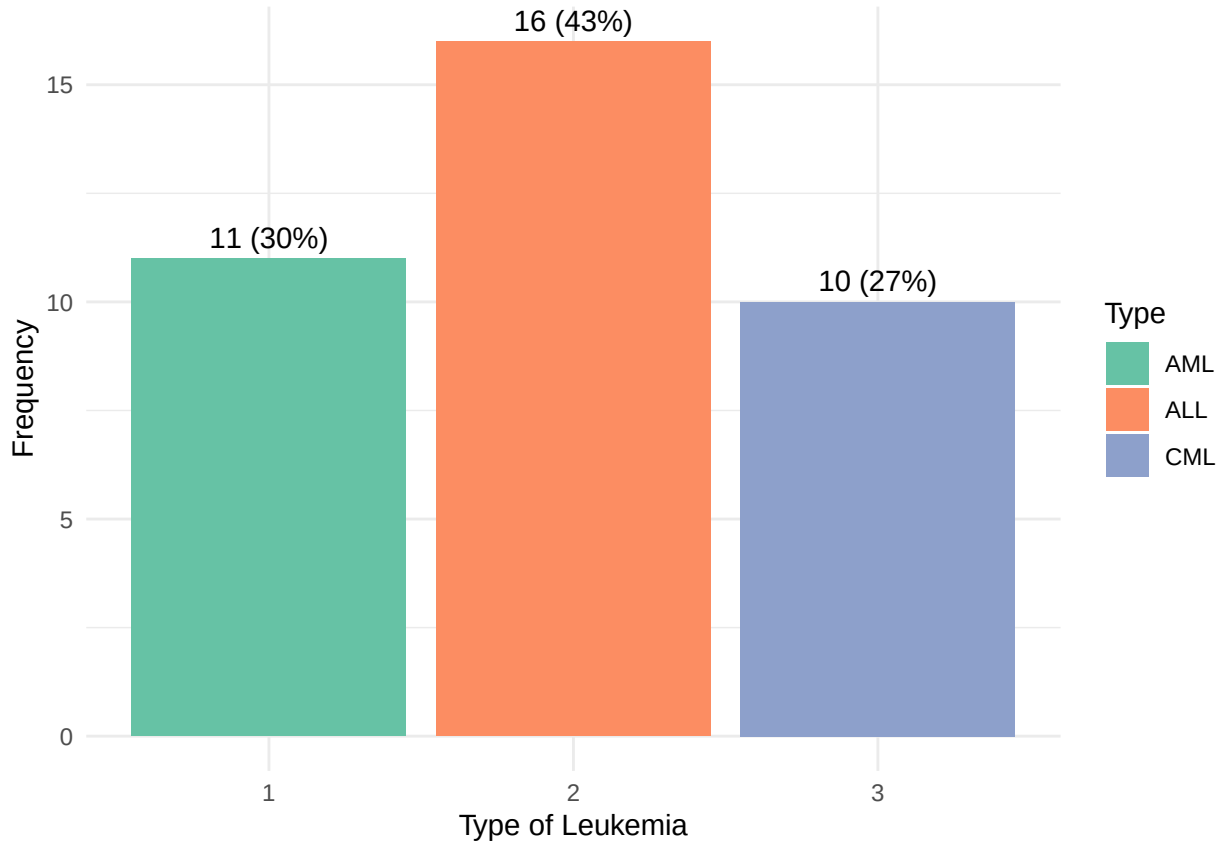


Figure 5: Frequency Distribution of the Types of Leukemia. The type of leukemia, with "1" for AML (Acute Myeloid Leukemia), "2" for ALL (Acute Lymphoblastic Leukemia), and "3" for CML (Chronic Myeloid Leukemia).

The distribution of leukemia types reveals that ALL is the most prevalent subtype, accounting for 43% of cases, followed by AML at 30%, and CML at 27%.

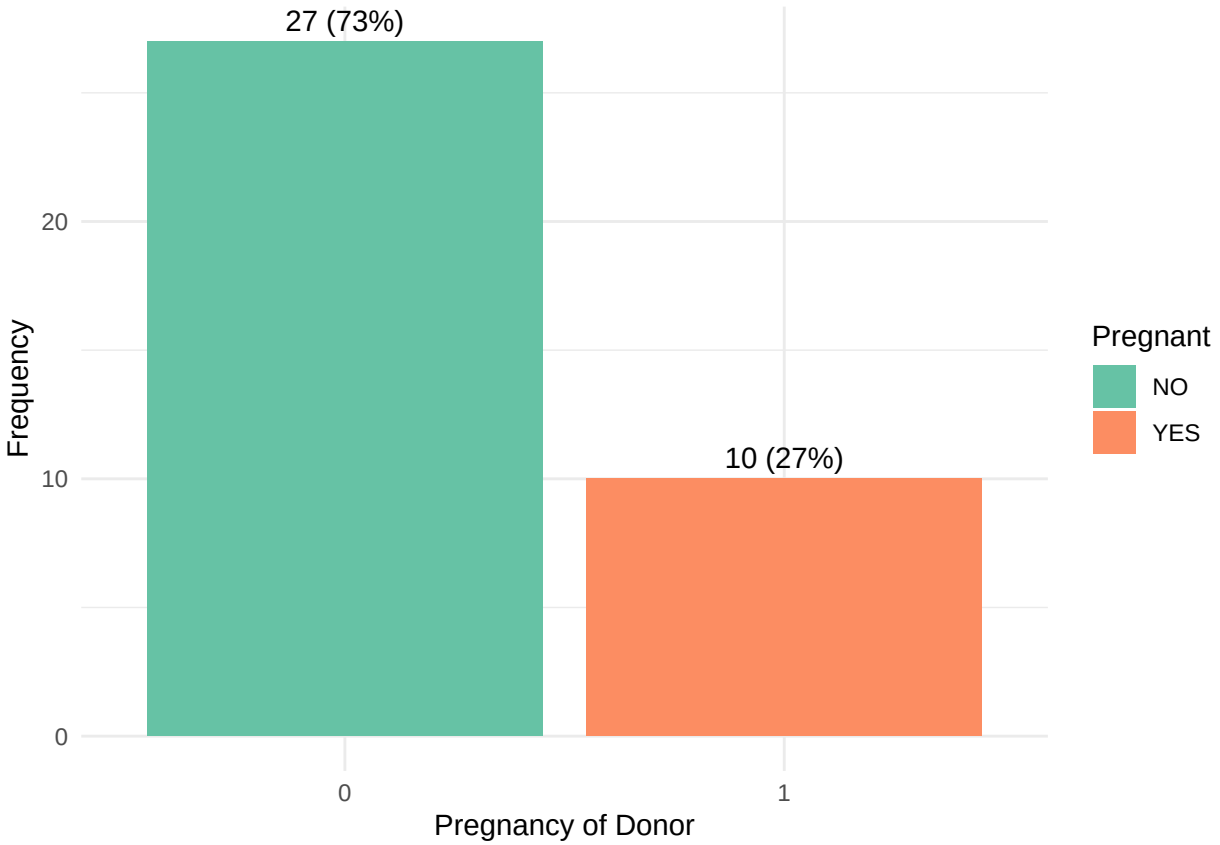


Figure 6: Frequency Distribution of the Donor's Pregnancy History The pregnancy variable is categorically encoded as 0 for "not pregnant" and 1 for "have been pregnant before".

Based on figure 6, 27% of participants experienced pregnancy, while the majority (73%) did not.

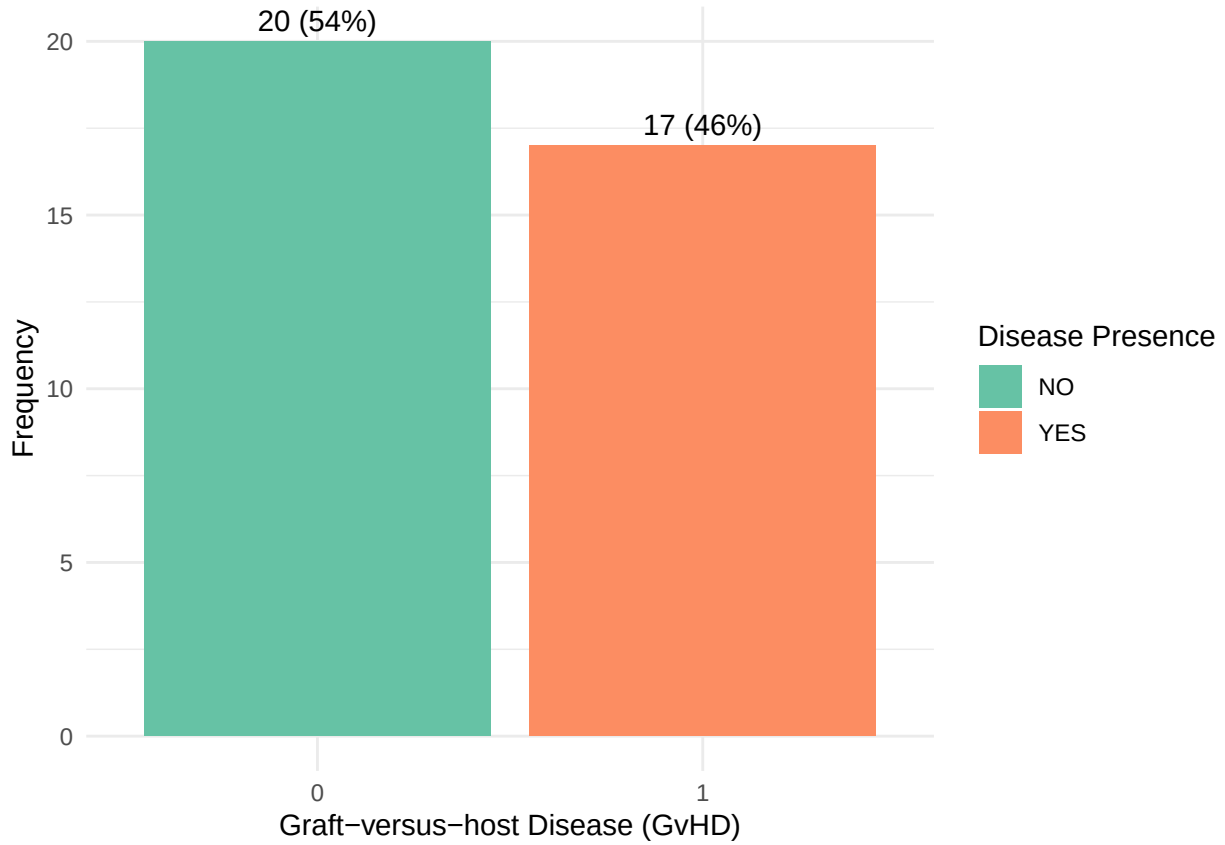


Figure 7: Frequency Distribution of Graft versus Host Disease The GvHD variable is categorically coded as "0" for its absence and "1" for its presence.

Based on figure 7, 46% of participants exhibited graft versus host disease, indicating a significant occurrence, while 54% did not manifest this condition.

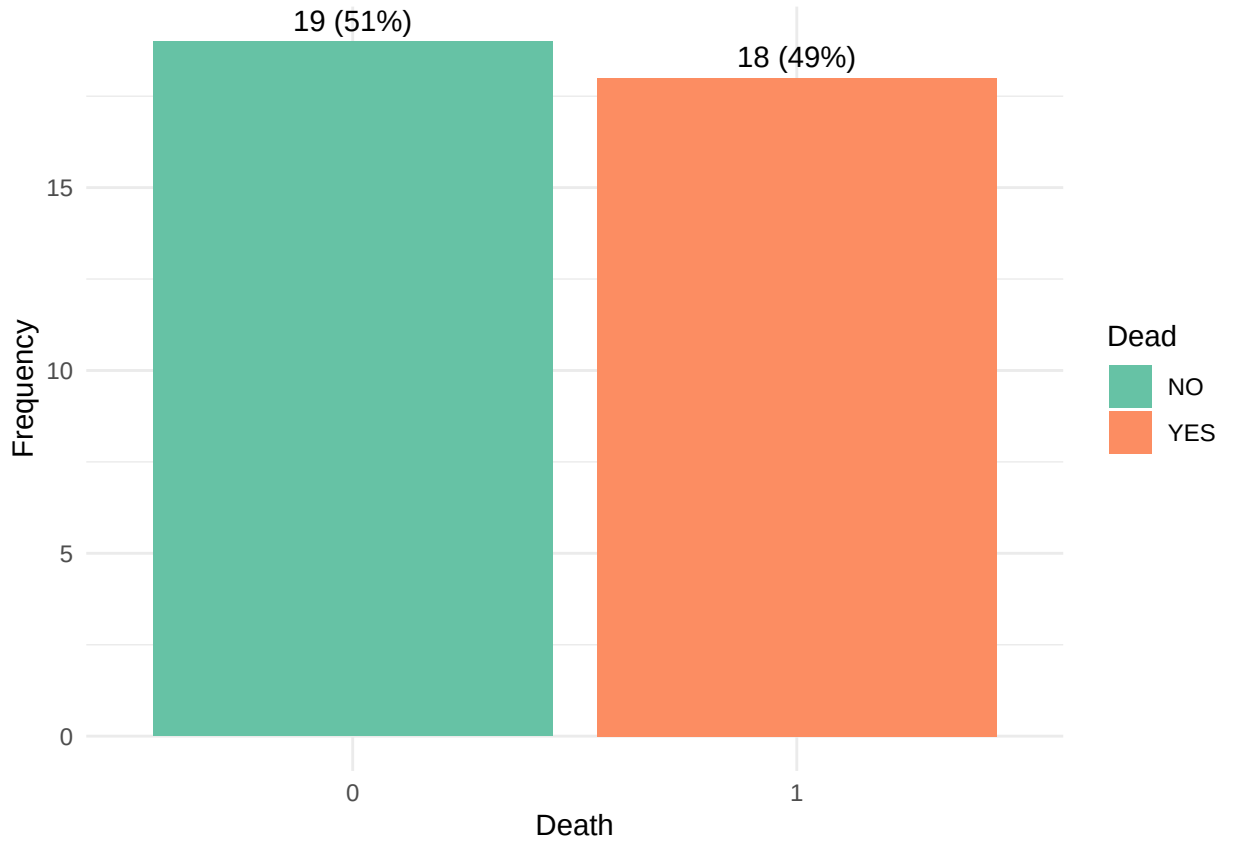


Figure 8: Frequency Distribution of Mortality. The death variable captures the vital outcome of patients, with "0" signifying survival and "1" representing mortality.

Based on figure 8, 49% of study participants died, while the remaining 51% survived.

Table 2: Descriptive Statistics for Key Variables

Variable	Mean	Standard Dev.	Variation Coef.	1st Quartile	Median	3rd Quartile	Skewness
Recipient Age (Years)	25.43	7.50	0.30	20.00	23.00	29.00	0.68
Donor Age (Years)	25.81	7.84	0.30	20.00	23.00	34.00	0.54
MLR Index	2.56	2.23	0.87	0.92	2.01	3.73	1.69
Follow-up Time (Days)	669.78	483.72	0.72	177.00	667.00	1105.00	0.20

Descriptive Statistics Insights

Recipient Age: The mean recipient age is 25.43 years, with a standard deviation of 7.50 years. The coefficient of variation is relatively low at 0.30, indicating moderate variability. The distribution is slightly right-skewed (skewness = 0.68), and most recipients fall within the 20 to 29-year age range.

Donor Age: Donor age exhibits similar characteristics to recipient age. The mean donor age is 25.81 years, with a standard deviation of 7.84 years. The coefficient of variation is 0.30, indicating moderate

variability. The distribution is slightly right-skewed (skewness = 0.54), and most donors are within the 20 to 34-year age range.

MLR Index: The mean MLR Index is 2.56, but it stands out with a relatively high standard deviation of 2.23 and a high coefficient of variation (0.87), suggesting substantial variability in the data. The distribution is strongly right-skewed (skewness = 1.69), indicating potential outliers on the higher end. Figure 4 further confirms this with an individual having an MLR Index in the 8.5-10 range, highlighting the need for attention to potential outliers in this variable.

Follow-up Time: The mean follow-up time is 669.78 days, with a standard deviation of 483.72 days. The coefficient of variation is moderate at 0.72, indicating moderate variability. The distribution is slightly right-skewed (skewness = 0.20), with most follow-up times falling within the lower range. However, some longer follow-up times are present in the data.

Task 4 & 5

Logistic Regression Model Findings and Interpretation

Null Hypothesis (H0): There is no relationship between the predictors (rcpage, index, type) and the occurrence of GvHD.

Alternative Hypothesis (H1): There is a relationship between the predictors (rcpage, index, type) and the occurrence of GvHD.

- Intercept (β_0): The estimated log-odds of developing GvHD when all other predictor variables are zero is approximately -5.11502.

Coefficients:

- rcpage (Recipient Age, β_1): For each one-unit increase in recipient age (rcpage), the estimated log-odds of developing GvHD increase by approximately 0.13061, holding other predictors constant. This variable is statistically significant (p = 0.04467).
- index (Mixed Epidermal Cell-Lymphocyte Reactions, β_2): For each one-unit increase in the index, the estimated log-odds of developing GvHD increase by approximately 0.61024, keeping other predictors constant. This variable is marginally significant (p = 0.06118).
- type2 (Leukemia Type 2 - ALL, β_3): This coefficient represents the change in log-odds of GvHD risk for patients with leukemia type 2 (ALL) compared to leukemia type 1 (AML). However, it is not statistically significant (p = 0.63744), suggesting that the difference in GvHD risk between these two leukemia types is not significant.
- type3 (Leukemia Type 3 - CML, β_4): This coefficient represents the change in log-odds of GvHD risk for patients with leukemia type 3 (CML) compared to leukemia type 1 (AML). Like type2, it is not statistically significant (p = 0.35741), indicating that the difference in GvHD risk between CML and AML is not significant.

P-values: Recipient age is a statistically significant predictor of GvHD risk, with older recipients having higher odds of developing GvHD. The index of MLR variable shows marginal significance, suggesting that it may also influence GvHD risk, but further investigation may be needed to confirm its significance. Leukemia types 2 and 3 do not significantly impact GvHD risk in this analysis.

Task 6

Odd Ratios and their Interpretation

- Recipient Age (rcpage): The odds ratio for recipient age is approximately 1.14. Therefore, since the odds ratio is greater than 1, this means that for each one-year increase in recipient age, the odds of developing GvHD are approximately 14% higher for patients with higher ages compared to younger patients, holding other predictors constant. In addition, 95% of the time, we expect the odds of developing GvHD to change within a range of approximately 1.01 to 1.32 for each one-year increase in recipient age.
- Mixed Epidermal Cell-Lymphocyte Reactions (index): The odds ratio for the index is approximately 1.84. Thus, as the odds ratio is greater than 1, this indicates that for each one-unit increase in the index, the odds of developing GvHD are approximately 84% higher for patients with higher mixed epidermal cell-lymphocyte reactions compared to those with lower reactions, while keeping other predictors constant. Thus, 95% of the time, we anticipate the odds of developing GvHD to change within a range of approximately 1.10 to 3.97 for each one-unit increase in the index.
- Leukemia Type 2 (ALL) vs. Leukemia Type 1 (AML): The odds ratio for leukemia type 2 (ALL) compared to leukemia type 1 (AML) is approximately 0.59. However, since the odds ratio is less than 1, this means that the odds of developing GvHD are approximately 41% lower for patients with leukemia type 2 (ALL) compared to patients with leukemia type 1 (AML), but this difference is not statistically significant. In addition, 95% of the time, we expect the odds of developing GvHD to change within a range of approximately 0.06 to 5.49 for patients with leukemia type 2 (ALL) compared to patients with leukemia type 1 (AML),
- Leukemia Type 3 (CML) vs. Leukemia Type 1 (AML): The odds ratio for leukemia type 3 (CML) compared to leukemia type 1 (AML) is approximately 2.86. Since the odds ratio is greater than 1, this suggests that the odds of developing GvHD are approximately 186% higher for patients with leukemia type 3 (CML) compared to patients with leukemia type 1 (AML). However, it's important to note that this difference is not statistically significant.95% of the time, we expect the odds of developing GvHD to change within a range of approximately 0.32 to 32.90 for patients with leukemia type 3 (CML) compared to patients with leukemia type 1 (AML).

Task 7

Derive mathematically the risk of developing graft-vs-host disease for a patient with type 2 leukemia and index = 4.

- β_0 (Intercept) = -5.11502
- β_2 (Index) = 0.61024
- β_3 (Type 2) = -0.52425
- β_4 (Recipient Age) = 0.13061
- Mean recipient age is around 25, so will select 25 years.

$$\hat{p} = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X + \ldots)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X + \ldots)} = \frac{exp(-5.11502 + 0.61024 \cdot 4 + -0.52425 \cdot 1 + 0.13061(25)}{1 + exp(-5.11502 + 0.61024 \cdot 4 + -0.52425 \cdot 1 + 0.13061(25)}$$

The estimated probability of developing graft-vs-host disease (GvHD) for a patient with Type 2 leukemia, an index of 4, and a recipient age of 25 mathematically is approximately 0.52.

Task 8

Using the predict function in R, the estimated probability for developing GvHD for an an individual who is 25 years old, with type 2 Leukemia (ALL), and MLR index of 4 is also 0.52.

Task 9

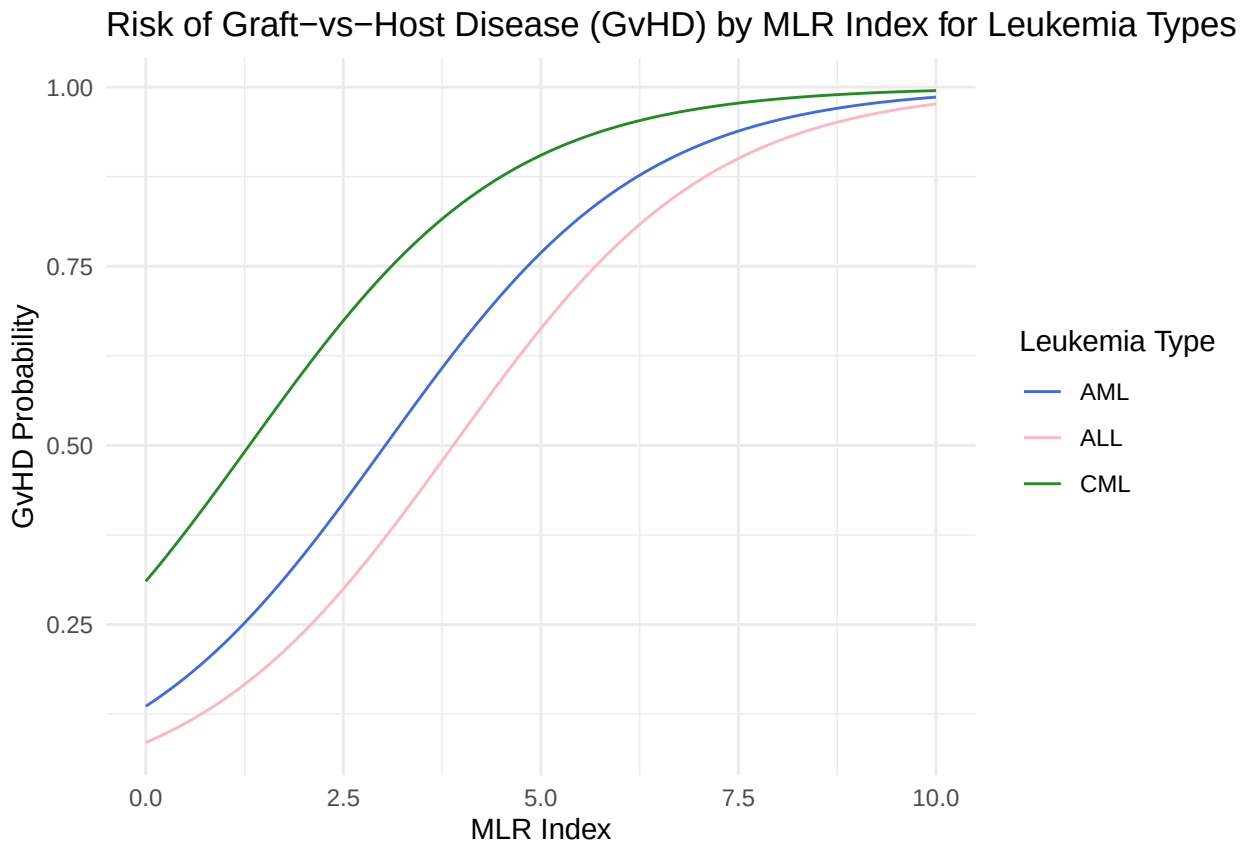


Figure 9: Risk of Graft-vs-Host Disease (GvHD) by MLR Index for Leukemia Types. The plot illustrates the relationship between the MLR Index and the probability of developing GvHD for different types of leukemia (AML, ALL, and CML). The x-axis represents the MLR Index, while the y-axis represents the GvHD probability.

Task 10

A researcher is also interested in predicting the risk of death at 2 years, using these data. A junior analyst suggested using logistic regression with death as the response variable for this analysis. Do you agree or disagree and why? Is your answer the same for the 1-year risk of death?

Predicting the risk of death at 2 years appears to be a well-suited approach, given the average follow-up time of approximately 2 years (665 days). This choice aligns with the research objective of modeling mortality within this time frame. However, it's crucial to select relevant predictor variables and ensure data quality, especially considering that the risk of death is nearly 50%, indicating a balanced dataset in terms of outcomes.

Evaluating the model's performance and clinically interpreting coefficients and odds ratios will be essential for understanding its predictive capabilities in the context of 2-year mortality.

Predicting death at 1 year is also a valid option but comes with considerations. While it may align with certain research questions focused on shorter-term outcomes, researchers should assess data availability, as the 2-year average follow-up time results in a dataset with a substantial number of deaths within the first year. Careful selection of predictors that align with shorter-term outcomes and rigorous model validation will be necessary. Both time frames offer valuable insights, but the choice should reflect the research goals and the clinical relevance of the selected prediction.

References

Socié, G., & Ritz, J. (2014). Current issues in chronic graft-versus-host disease. Blood, 124(3), 374–384.https://doi.org/10.1182/blood-2014-01-514752

Zhou, J., He, W., Luo, G., & Wu, J. (2014). Mixed lymphocyte reaction induced by multiple alloantigens and the role for IL-10 in proliferation inhibition. Burns & trauma, 2(1), 24–28. https://doi.org/10.4103/2321-3868.126088