Exam 2

Denis Ostroushko

2022-12-10

Problem 1

We begin this problem by summarizing the data available to us, and continue this summary into the **1 - A** Section. The data set contains 112 observations for 112 participants in the study.

One of the participants has a missing value of baseline NNAL measurement. This variable is important for our analysis, therefore we will omit this observation. The final data set includes 111 observations.

1 - A

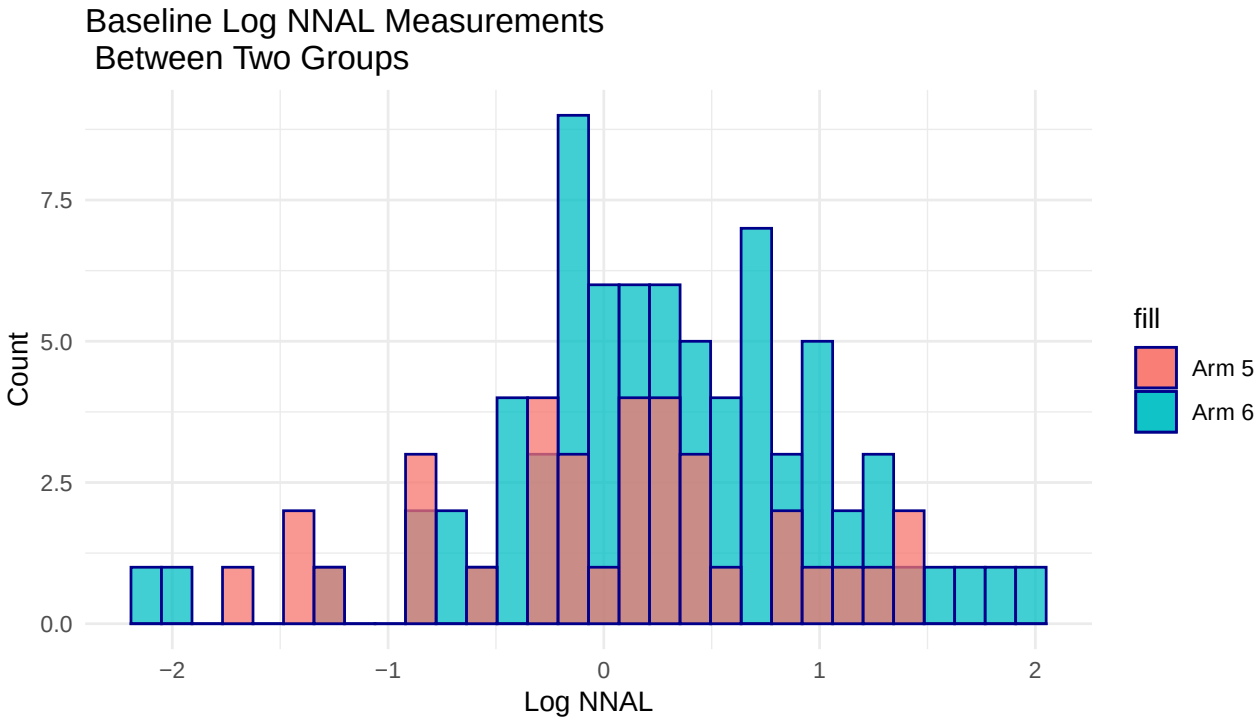
Problem 1-A asks us to fit a logistic regression model using two log-transformed baseline measurements. We are interested in evaluating how the two variables are balanced between the two experiment arms, arm 5 and arm 6.

		Baseline Log NNAL		Baseline Log TNE	
Arm	N	Mean	SD	Mean	SD
5	35	0.0265297	0.8009900	3.963363	0.7431000
6	76	0.2704350	0.7677647	4.097586	0.5577623

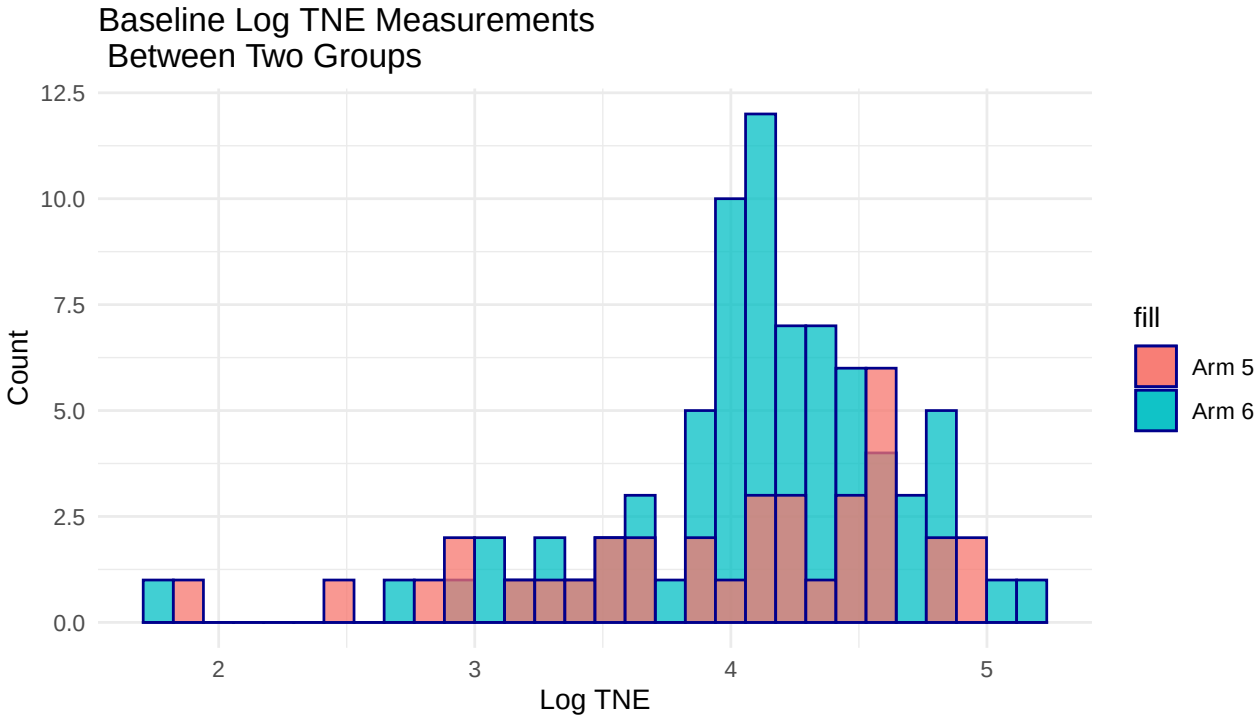
Table above provides mean and standard deviation for the two variables we want to use in the propensity score model. We can see that the means of the baseline NNAL measurements appear to be visually different between the two groups, although, with a high standard deviation difference may be due to the noise and variation of the data.

Average TNE measurements appear to be quite similar for both treatment arms.

In order to investigate the difference in distribution of two variables further we refer to the histograms.



It appears that the shape of two distributions is quite similar for the two treatment arms. There are quite more observations with the log-NNAL measurements above 1.5 for treatment arm 6, which may be the reason as to why the mean of observations is higher for this group. We also have about twice the amount of observations in the treatment group 6, so, perhaps, if we are able to observe 30 more people who can qualify to be in treatment arm 5, we can observe more extreme values, and more values that would tend toward the center of the distribution, making the two distributions very similar.



The distributions of log-TNE values between the two groups appear quite different, but also have similar

features. We can see that the distribution have heavy tails on the left, with the 'center' of each distribution being on the right side. Perhaps, this is caused by the logarithmic transformation. Overall, it is not easy to gauge the similarity of two distributions here due to varying sample size and natural variations of these biomarkers.

We are now ready to fit the logistic regression model to obtain propensity scores for each subject. Our response variable is Z, a binary, where $Z = 1$ if a study participant is in the treatment arm 6, and 0 if the participant is in arm 5.

Therefore, the model statement is:

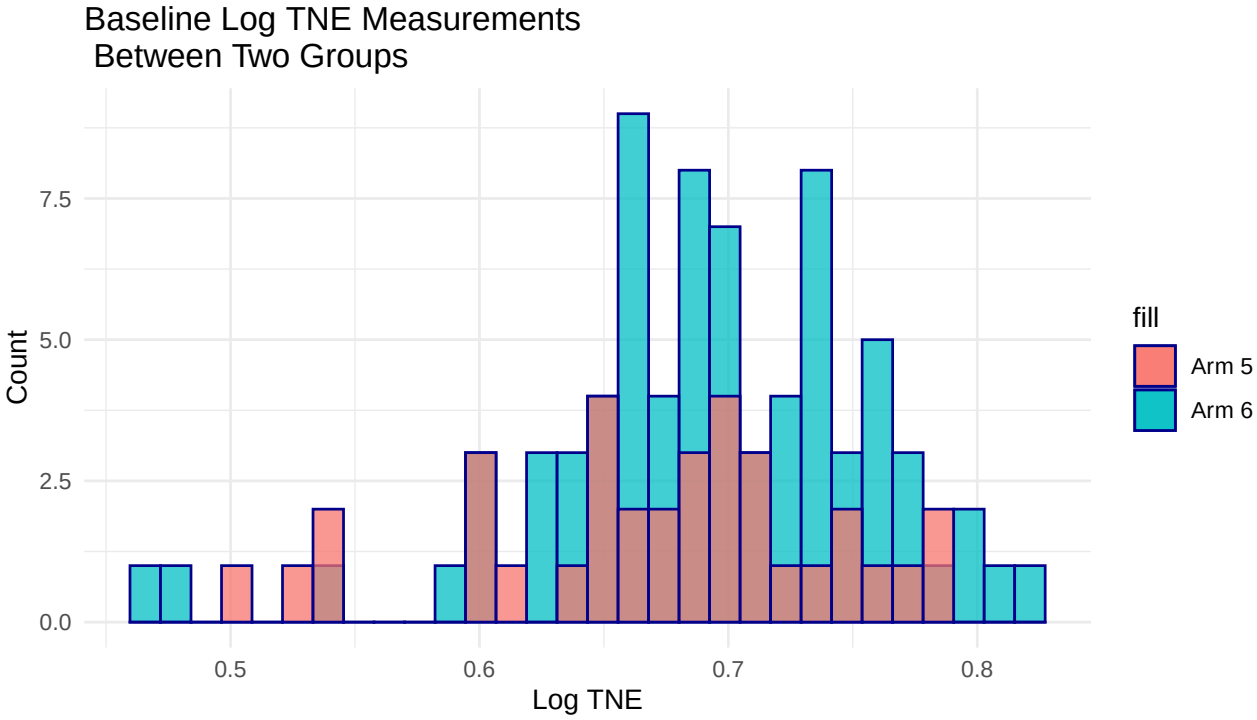
$$ln\frac{P(Z=1)}{1-P(Z=1)} = \hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2$$

Where X_1 is a baseline measurement of NNAL on the natural logarithmic scale, and X_2 is a baseline measurement of TNE on the natural logarithmic scale

After fitting the model, we summarize obtained propensity scores for each experiment arm.

		Propensity Score	
Arm	N	Mean	SD
5	35	0.6702007	0.0710263
6	76	0.6913550	0.0660584

It appears that the two samples have mean and standard deviations that are quite similar. We can also investigate the shape of distributions for each treatment group.



We have fairly balanced distributions of propensity scores for the two samples.

1 - B

Using propensity scores we calculate the odds of being in the treatment arm 6 for each experiment subject on the logarithmic scale. We give the summary of these odds for each sample in the table below:

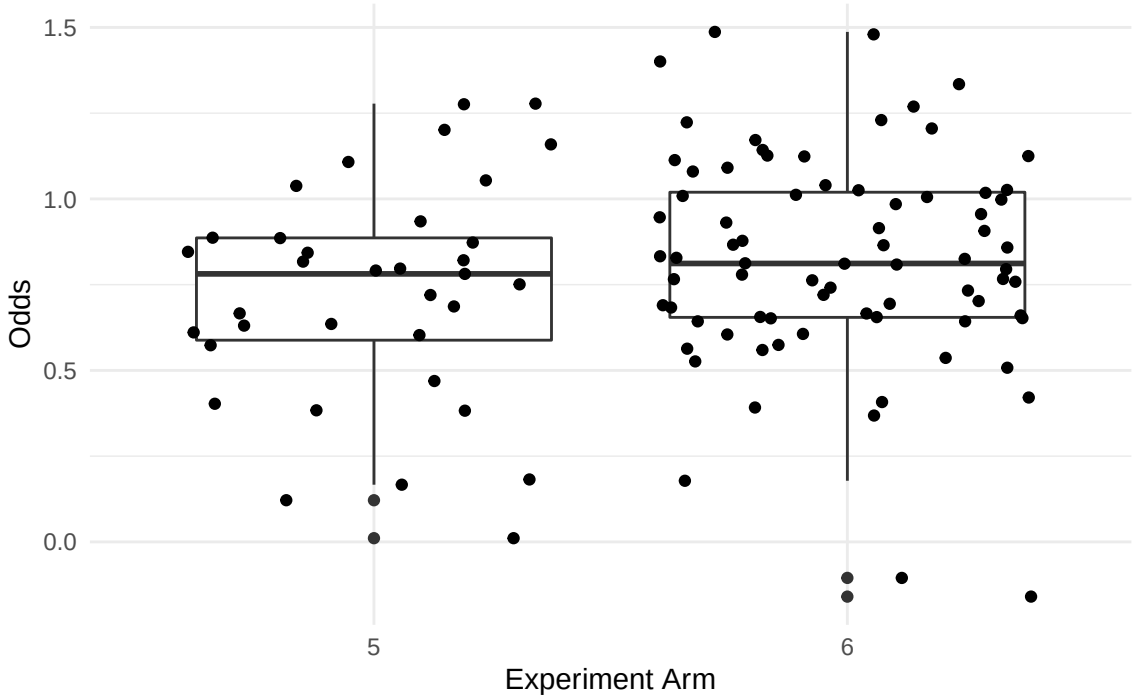
		Log-Odds	
Arm	N	Mean	SD
5	35	0.7252930	0.3213682
6	76	0.8231971	0.3082685

Because log-odds are a function of propensity scores, same conclusions about the mean, standard deviations, and the overall shape of the two distributions should apply here again.

We can compare the average Y values using a t-test. Before doing so, we can do two checks:

1. Use box plots to make sure there are no influential outliers. In case there are many influential outliers, we can pivot to a non-parametric Wilcoxon test. However, I suspect that we will not observe many influential outliers on the histograms above
2. Compare the variances of the two samples using an $F-$test. We will statistically test if the ratio of variances is greatly different from 1. In case the ratios are statistically different we can pivot to the Wilcoxon test as well.

Boxplot below shows no visual evidence of greatly influential outliers.



We can now carry out an F test to check the difference between variances of treatment odds between the two groups. We perform the test on the $\alpha = 0.05$ significance level.

Formal statement for $F-$test on the $\alpha = 0.05$ level is below:

- Variance of treatment arm 5 is $s_1^2 = 0.1033$ and variance of treatment arm 6 is $s_2^2 = 0.095$. The ratio s_2^2/s_1^2 is 0.9201
- Null Hypothesis: $H_0 : s_1^2 = s_2^2$

- Alternative Hypothesis: $H_a : s_1^2 \neq s_2^2$
- Test statistic $F = 0.9201$
- Critical cutoff $F-$value on 75 and 34 degrees of freedom is 1.8456
- $P(F^* > F) = 0.6261$
- Conclusion: $F-$ statistic is smaller than the critical value, p-value is much bigger than the accepted cutoff 0.05, so there is not enough evidence to reject the null hypothesis and conclude that the the variances of two samples is not equal.

So, we confirmed that the variances do not differ greatly using a statistical test and visually confirmed that there are no influential outliers in the two samples. Therefore, we can conduct a $t-$ test to compare the average treatment odds.

Formal statement for $t-$test at the $\alpha = 0.05$ significance level is below:

- Average log-odds for treatment arm 5 is $\bar{X}_1 = 0.7253$, average log-odds for treatment arm 6 is $\bar{X}_2 = 0.8232$, observed difference $\bar{X}_2 - \bar{X}_1$ is 0.0979
- Null Hypothesis $H_0 : \bar{X}_1 = \bar{X}_2$
- Alternative Hypothesis $H_a : \bar{X}_1 \neq \bar{X}_2$
- Test $T-$ statistics is 1.5105
- $P(T^* > T) = 0.1359$
- Conclusion: $T-$ statistics is smaller than the critical cutoff, so we can not reject the Null Hypothesis and conclude that the treatment log-odds are different between the two groups. However, p-value is only about 2.5 times greater than the accepted significance level, so these results can be suggestive that $\bar{X}_2$ is greater. We should not disregard these results easily.

  This is also supported by the confidence interval for the difference estimate, which contains mostly positive values. C.I. is given by (-0.0316, 0.2274).

1 - C

Based on some literature review, many psychology and social science data analysis methods refer to this method as Cohen's d. All sources I reviewed state that in practice effect sizes between 0.2 and 0.5 are considered as Medium size.

A good explanation is given here: https://datatab.net/tutorial/effect-size-independent-t-test

We calculate the effect size as the difference divided by the pooled variance, where pooled variance is given by:

$$s_p^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_2 + n_1 - 2}$$

And effect size is given by:

$$E.S. = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{S_p^2}}$$

I will rely on a test instead because we have more accessible and straightforward way to get a confidence interval, that is widely known and accepted among the research community.

However, effect size is also a good method that gives us an intuitive tool for inference. We took the ratio of a difference to the common, or pooled, standard deviation. With our estimate of 0.3134, we have a moderate

effect size. Also, difference between samples is approximately 31% of the common standard deviation. So, the effect is , perhaps, "hidden" due to great variance in the sample, or it is "diluted" due to the variance that occurs in the sample.

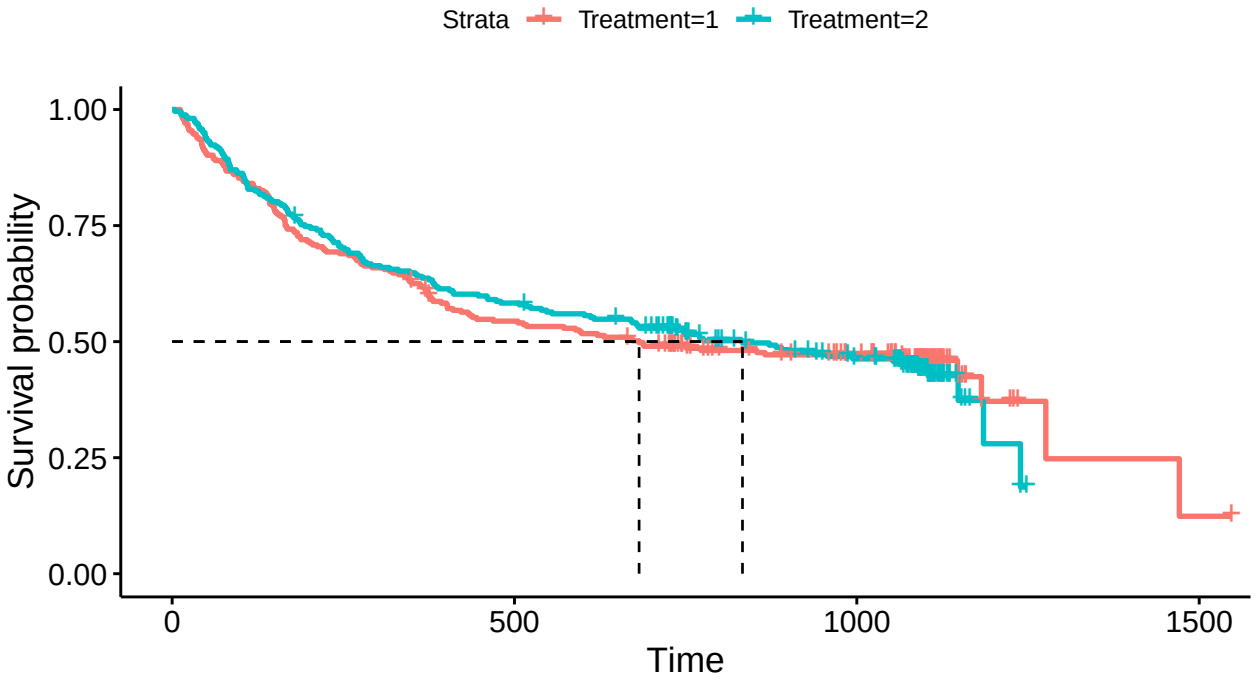
Another interpretation is that the difference between the two groups is small when compared to the average, or pooled, variance in the sample. There is too much variance to make conclusive statements.

Problem 2

The data set for this problem contains 551 observations. In the context of the survival analysis problem, we look at the time to event, which is defined by the `duration` variable. There are 25 observations in the data set with the missing time to event measure, so they will be removed from the analysis. Therefore, all survival models will be fit using a sample of 526 study participants.

2 - A

Before obtaining the model estimate it is helpful to develop intuition using graphical methods. We can see that the two groups have very similar fitted survival curves. Therefore, we should not expect to see a big estimate of hazard ratio. However, we can see that the median survival time is quite larger for the Peripheral Blood Stem Cells (treatment 2) treatment group.



Now we can fit a Cox Proportional Hazard Regression Model to compare survival likelihood between the two groups. Model estimate is given below:

Predictor	Estiamte	Exponentiated Estiamte	Standard Error	Z Value	P value
Treatment	0.002623	1.002626	0.118131	0.022204	0.982285

- We defined group who received Bone Marrow (BM) treatment as a reference level, and a group who received Peripheral Blood Stem Cells (PBSC) treatment as a comparison group
- As we can see, exponentiated coefficient, a hazard ratio, is 1.0026, which means that the PBSC treatment group has approximately 0.26% higher chance of having a leukemia related death at any point in time.
- High p-value indicates that this difference is not statistically significant, so we cannot reject null hypothesis. Therefore, PBSC treatment does not improve survival chances when compared to Bone Marrow treatment for the sample of leukemia patients.
- In addition, standard error is quite large in comparison to the coefficient, the coefficient itself is close to zero, therefore there is a high degree of uncertainty in the estimate.

2 - B

Before fitting the model we can summarize the data.

Variable	Bone Marrow	Peripheral Blood Stem Cells
N	264	262
Median Survival Time	647.5	720.5
% Female	40.91%	46.56%
% White	90.53%	90.08%
% HLA 6 Score	87.88%	91.6%
Avg. Age	41.9	41.6

It appears that the two samples are pretty balanced in terms of all covariates.

We can how fit a Cox Proportional Hazard Regression Model with all predictors. A summary table with model estimates is given below:

Predictor	Estiamte	Exponentiated Estiamte	Standard Error	Z Value	P value
Treatment	0.032936	1.033485	0.118417	0.278139	0.780906
GENDER2	-0.087586	0.916140	0.120380	-0.727583	0.466869
age	0.017906	1.018067	0.003979	4.499976	0.000007
Race	-0.202888	0.816370	0.196934	-1.030232	0.302901
'HLA-Match'	-0.508781	0.601228	0.185197	-2.747244	0.006010

- Age and HLA-Match are two statistically significant predictors of the survival chance for leukemia patients
- Exponentiated coefficient for Age is 1.0181 which means that one additional year of age multiplicatively increases the chance of dying at any point in time by 1.0181, or by approximately 1.81%

  However, it makes more sense to assess age effects on a large time frame. For example, additional 10 years of age increase the chance of having a leukemia related death at point in time by 18.1%, after adjusting for the effects of other predictors

  This estimate is bounded by the (10.2%, 26%) 95% confidence interval. Due to exponentiation of the coefficient, confidence interval appears to be skewed to the right.
- HLA-Match has two possible values. Patients with score 5 are chosen as a reference level for this modeling exercise, while patients with score 6 were used as the comparison group.

  Patients with HAL score 6 were have a hazard ratio of 0.6012, which means that patients in this group had a much lower chance of dying. In fact, at any point in time, patients in group with score 6 had approximately -39.88% lower chance of dying, after adjusting for other predictors.
- Looking at raw coefficients for these predictors, we can also have more confidence in these predictors because standard errors are small in comparison with the coefficient magnitude and size. Of course, confidence intervals are constructed using these standard errors, but it is also nice to see this information in the table next to the model output.

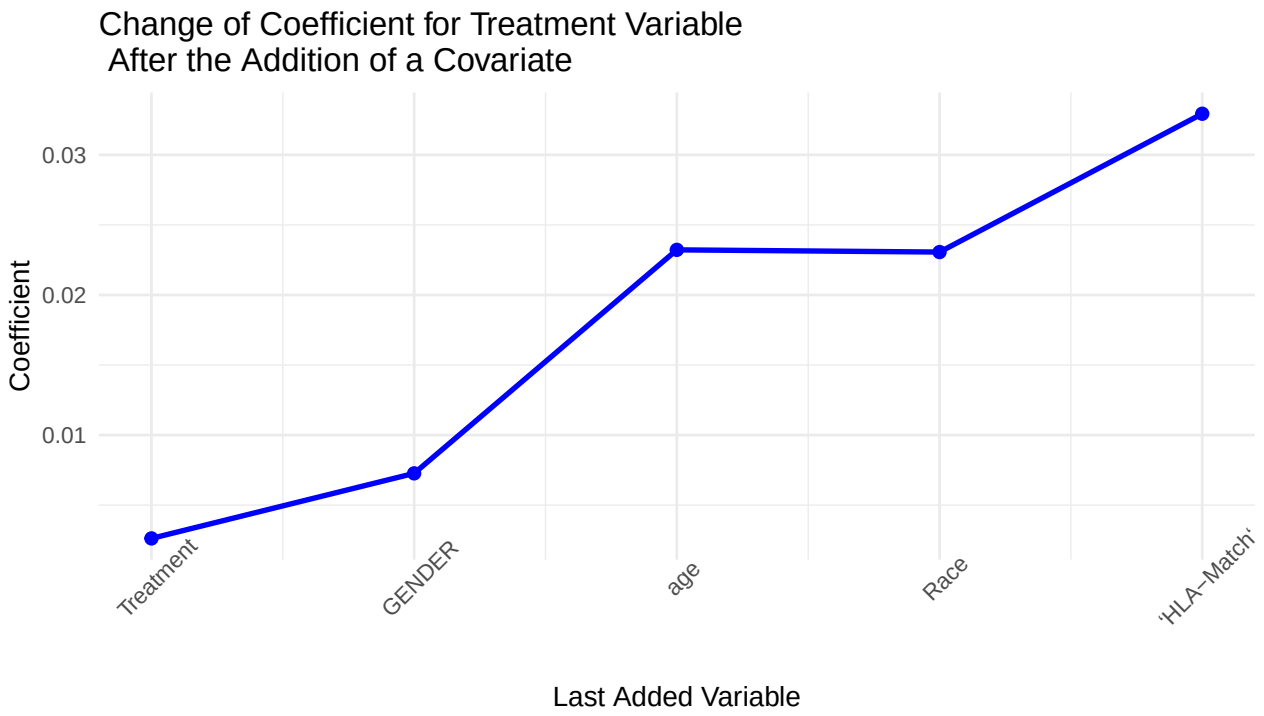
2 - C

Table below provides a side by side comparison of the Treatment coefficient from the two regression models. Recall that the coefficient is given for the comparison of PBSC group to the BM group.

Model	Predictor	Estiamte	Exponentiated Estiamte	Standard Error	Z Value	P value
Full	Treatment	0.032936	1.033485	0.118417	0.278139	0.780906
Treatment Only	Treatment	0.002623	1.002626	0.118131	0.022204	0.982285

- The two estimates are not the same, and they do not have the same p-value. However, they are also not drastically different and do not tell a different story.
- It is interesting to observe that the coefficient actually increased in magnitude after the addition of extra predictors into the model. Additionally, it is quite interesting to see that the standard error estimate remains very similar in the two models.

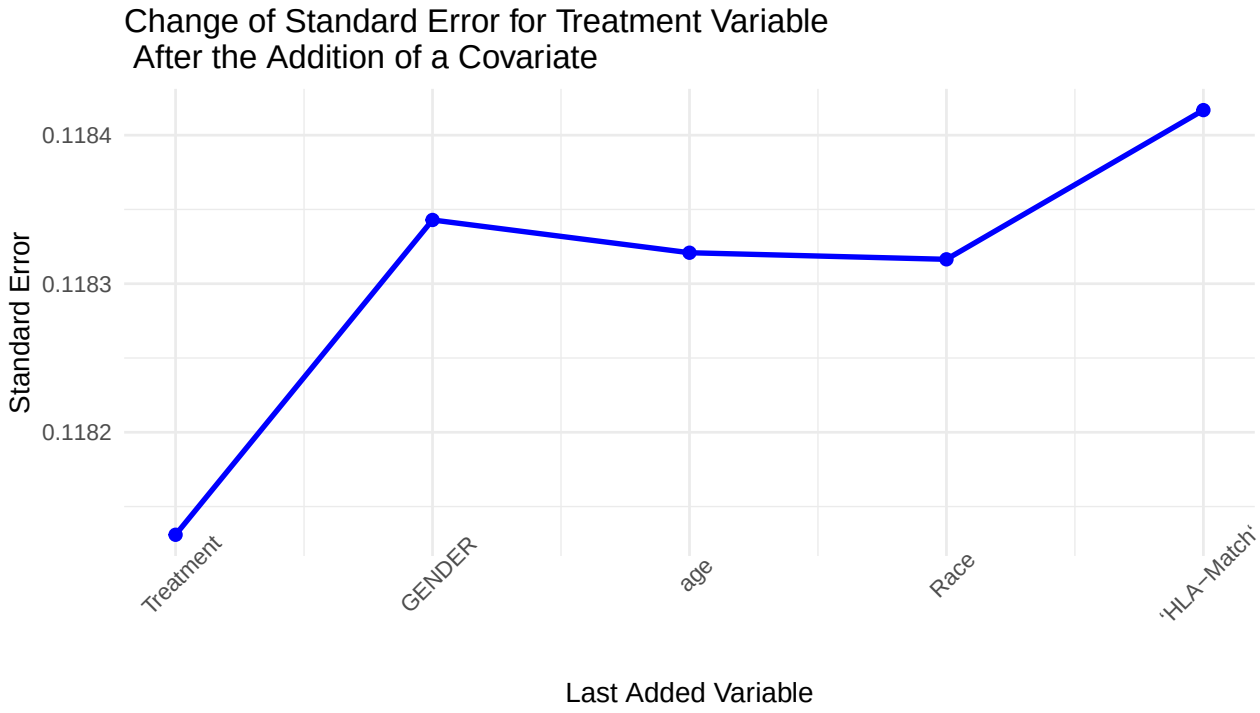
We use iterative process to add more variables into the model, one by one, to see what variables impact the change in coefficient of the treatment variable



Conclusion

It appears that the largest hikes in coefficient for the treatment variable happen after the addition of Age and HLA-Match variables. These two are the only two statistically significant variables that help meaningfully explain variation in survival times and likelihood. Therefore, it is likely that by adjusting for variables that are statistically related to the target variable of interest, we account for effects that are caused by Age and HLA-Match, and therefore the model highlights, or isolates, the effect of PBSC vs BM treatment better.

In a similar fashion we can check how adding more variables into the model affects a standard error for the treatment variable. We saw that the model with just one treatment variable and the full model produced almost identical standard error.



It is quite remarkable that the standard error does not change drastically. A change from 0.118417 to 0.118131 is minimal for all practical purposes and applications. This result verifies that treatment is not heavily correlated with the other predictors.