**Module 6 Critical Thinking: Option 1**

Nolan Sharpe

Colorado State University – Global Campus

MIS 581: Capstone-Business Intelligence and Data Analytics

Dr. Ford

1/2/2022

**Module 6 Critical Thinking: Option 1**

Breast cancer affects numerous women in the United States and across the world every year. According to The American Cancer Society, an American woman has an approximately one in eight chance of developing breast cancer, which is the second most common cancer among women, with only skin cancers seen more (ACS, 2021). As of 2021, worldwide, breast cancer is the most common form of cancer, accounting for 12% of all new cancer cases annually (Breastcancer.org, 2021). The expectation is that this year in the United States, 281,550 women will be diagnosed with breast cancer, and 43,600 women will be killed by the disease (ACS, 2021). While breast cancer incident rates have increased slightly in the past few years, breast cancer death rates have remained steady for women under 50, and have decreased for women over 50, falling by 1% a year from 2013 to 2018 (ACS, 2021). Most of the improvements have been attributed to earlier breast cancer identification, and improved treatments.

Immune inhibitors, which block the interaction between immune checkpoint genes and T-cells, are being utilized more and more in a variety of cancer treatments (NCI, 2019). The body contains immune checkpoints, which prevent T-cells from attacking the body’s healthy tissues, but they can also prevent t-cells from attacking cancer, as tumors contain these genes as well (NCI, 2019). Using immune inhibitors limits the effectiveness of these genes, which allow the T-cells to more easily attack the cancer cells. A PD-1 inhibitor has been approved for treating early-stage triple negative breast cancer, in conjunction with chemotherapy, if the cancer has come back and is inoperable, or if it has spread to other parts of the body (ACS, 2020). But other than that, the use of immune inhibitors in breast cancer treatment is fairly limited. The final Capstone Project will look at the gene expressions of around 1,000 patients, to see if there is any correlation with immune checkpoint gene expression and treatment outcomes.

**Overview of National Cancer Institute**

The dataset that will be used for the final Capstone Project comes from the National Cancer Institute (NCI), which is the United States government's primary agency for cancer research and training. The NCI is made up of approximately 3,500 employees, and is part of the National Institutes of Health, which is one of 11 agencies that make up the Department of Health and Human Services (NCI, 2018). The goal of the NCI is to conduct and support cancer research across the United States, and to advance scientific knowledge that will help people live longer and healthier lives (NCI, 2018). The NCI, which is the largest funder of cancer research worldwide, manages a wide range of research, training, and information distribution activities (NCI, 2018). The agency strives to meet the needs of all demographics, whether they be rich, poor, urban or rural (NCI, 2018).

The NCI’s investments have contributed to declining rates of new cancer cases and total cancer deaths over the last few decades. These improvements have allowed the number of cancer survivors in the United States to double, from 7 million in 1992, to more than 15 million in 2016 (NCI, 2018). This trend is only expected to continue, with more than 26 million cancer survivors expected in the United States by 2040 (NCI, 2018). These improvements are the result of the advancements achieved in cancer detection, diagnosis, along with improvements in patient care, which have contributed to people living longer, and healthier lives. The NCI also maintains and curates publicly available datasets, such as The Cancer Genome Atlas (TCGA), which is where the data that was used in this project came from.

**Overview of Capstone Project Data**

Two datasets, which contain information on breast cancer patients, will be combined for analysis for the final Capstone Project. Each dataset contains well over 1,000 entries, and a total of 1,097 breast cancer patients have information contained in both datasets. The first dataset contains information on a patient’s tumor gene expression for four specific genes, CTLA4, PD1, PDL1 and TCRa. The CTLA4 gene is an immune inhibitor, that encodes a protein which transmits an inhibitory signal to the body’s T-cells (Gene Cards, 2021). T-cells are a type of white blood cell that are part of the immune system, which help to protect the body from infection, and may help fight cancer (NCI, n.d.). The CTLA4 gene helps to ensure that the T-cells do not attack the body’s healthy tissue, and only target the cells that are harmful to the body. Drugs that inhibit the effectiveness of the CTLA4 gene have been used to fight other cancers, including melanoma, and small cell lung cancer (Buchbinder, & Desai, 2016). This final Capstone Project will look to see if patients with a lower level of the CTLA4 gene, had a statistically significant difference in their treatment responses, then those with a higher expression of the gene.

Similar to CTLA4 gene, the PD1 gene is also an immune inhibitor. PD-1 is actually a protein located on the bodies T-cells, which helps keep the body’s immune system from going out of control (NCI, n.d). The PD1 protein binds to the PDL1 cell protein, another of the variables contained in the dataset, and they relay information to one another, which keeps the t-cell from attacking the cell the PDL1 protein it is on, but this can also prevent a t-cell from attacking cancer cells, as tumors can contain PDL1 genes as well (NCI, n.d.). There are drugs that have been approved to inhibit the effectiveness of both of these genes. For PDL-1 specifically, inhibitors have been approved for treating non-small cell lung cancer and Merkel cell carcinoma, after it was determined the drugs improved antitumor immunity, led to more durable clinical responses, and prolonged survival (Akinleye, & Rasool, 2019). This project will look at whether the level of the gene expression for these genes, had any impact on the effectiveness of the treatment.

TCRa is another key immune gene that is part of the T-cell receptor (TCR) which indicates to the body’s T-cells what the T-cells should attack. TCRs are a multimeric complex of integral membrane proteins that help to activate t-cells to respond to antigens in the body (Thermo Fisher, n.d.). Because only T cells express TCRa, a high level of TCRa indicates a high level of T-cells in the tumor, which just may not be working, due to the cancer cells containing a high level of PDL-1, which may indicate that a PDL-1 inhibitor might be beneficial. The final Capstone Project will explore whether or not a patients TCRa level had any effect on how well they responded to treatment.

The second dataset contains personal information on the patients, including a unique identifier, which correspond to the unique ID in the other data set. It also includes information on the age the patient was diagnosed, their ethnicity, gender and race, the year they were born, and the year they were diagnosed. This dataset also contains information relating to the patient’s diagnosis and treatment. This includes data points on whether or not the patient is alive, how long it took for them to die if they passed, how many days it has been since they were first diagnosed, the stage their cancer was when they were diagnosed, and whether or not they were treated with radiation or via pharmaceutical methods. Three variables in this dataset will be used as the dependent variables in the Capstone Project analysis. The main areas of focus will be on the patient’s vital status, and the length of time it has been since they have had a follow up.

**Objectives**

The goal of the final Research Projects is to try and determine if the gene expression from a patient’s tumor has a positive or negative effect on patient health outcomes. The first step will be to determine if there is any correlation between the four genes in the dataset, to three dependent variables, VitalStatus, whether the patient is alive or dead, DaysToDeath, how long it took for them to die, if they indeed die, and DaysSinceFirstDiagnosed, or the number of days that have passed since the patient was first diagnosed. The second step will be to split the patients into quartiles, based on the expressions for the four different genes. The top and bottom quartiles for each gene will then be compared to each other, to see if the means of the three variables described above were significantly different. The third step will be to use data visualizations to explore the relationships between the variables of the data set.

**Research Questions and Hypotheses**

There will be 7 research questions (RQ’s), and 7 null and alternative hypothesis that will be explored in this study.

RQ 1: Do the patient’s gene expression have any correlation with whether or not a patient lives or dies?  
 H0 1: The patients gene expressions do not have any correlation with patient survival.  
 Ha 1: A patient’s gene expressions are correlated with whether or not they survive.

RQ 2: Do the patients gene expressions have any correlation with how long they survived, before they passed away?  
 H0 2: The patients gene expressions do not have any correlation with how long a patient lived before they passed away.   
 Ha 2: A patient’s gene expression are corelated with the amount of time that a patient survived before passing away.

RQ 3: Do the patients gene expression have any correlation with how long a patient has lived, since they were first diagnosed with the disease?  
 H0 3: The patient’s gene expressions are not correlated with the amount of time it has been since the patient was first diagnosed.  
 Ha 3: A patient’s gene expressions are correlated with the amount of time it has been since the patient was first diagnosed.

RQ 4: Do patients with high and low CTLA4 gene expressions, have significant differences in their averages for the three variables, VitalStatus, DaysToDeath, and days since first diagnosed?   
 H0 4: There is no difference in the averages of the three variables, when the high and low CTLA4 groups are compared.   
 Ha 4: There is a statistically significant difference in the averages of three variables, when the high and low CTLA4 groups are compared.

RQ 5: Do patients with high and low PDL1 gene expressions, have significant differences in their averages for the three variables, VitalStatus, DaysToDeath, and DaysSinceFirstDiagnosed?   
 H0 5: There is no difference in the averages of the three variables, when the high and low PDL1 groups are compared.   
 Ha 5: There is a statistically significant difference in the averages of three variables, when the high and low PDL1 groups are compared.

RQ 6: Do patients with high and low PD1 gene expressions, have significant differences in their averages for the three variables, VitalStatus, DaysToDeath, and DaysSinceFirstDiagnosed?  
 H0 6: There is no difference in the averages of the three variables, when the high and low PD1 groups are compared.   
 Ha 6: There is a statistically significant difference in the averages of three variables, when the high and low PD1 groups are compared.

RQ 7: Do patients with high and low TCRa gene expressions, have significant differences in their averages for the three variables, VitalStatus, DaysToDeath, and DaysSinceFirstDiagnosed?  
 H0 7: There is no difference in the averages of the three variables, when the high and low TCRa groups are compared.   
 Ha 7: There is a statistically significant difference in the averages of three variables when the high and low TCRa groups are compared.

**Literature Review**

Here we will review four articles which looked at how immune checkpoint inhibitors are currently being used in breast cancer treatments. The first was a study from 2016, which looked at other studies that were being conducted at that time, all of which used the suppression of the PDL1, PD1, and/or the CTLA-4, in combination with chemotherapy, as part of the breast cancer treatment plan (Voutsadakis, 2016). The study concludes that more work needs to be done on how these types of inhibitors can be combined with other treatments to improve health outcomes, but that it initially appears like their success, can be attributed to an individual’s gene level in their tumor’s microenvironment (Voutsadakis, 2016). This means that more work needs to be done to determine the specific gene levels that will have a positive response to the treatment. Figuring this out may result in the development of more novel therapy’s that can be used to target breast cancers that may be resistant to current treatments (Voutsadakis, 2016). This study is pertinent to this paper because it looks at studies that used immune inhibiting treatment on three of the genes that are contained in the data used in this project. If a correlation is found between those gene expressions and higher treatment outcomes, it would support the idea that these types of treatments might be effective.

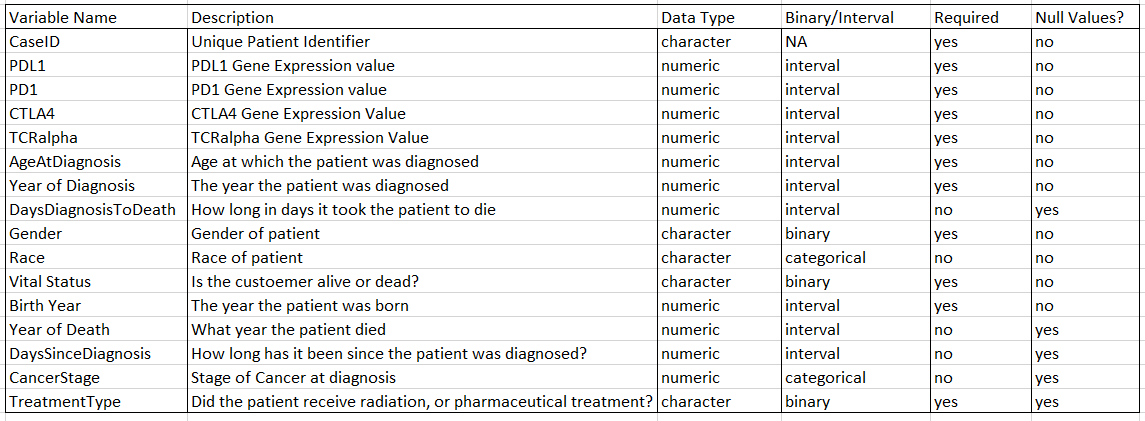
The second paper, published in 2019, discusses the fact that breast cancer tumors were historically shown modest responses to anti-PD1 and PDL1 therapy, but better tumor and immunological profiling has resulted in a better understanding of ways that breast cancer evades immune activity, as well as the tumor microenvironment, and how those things might be therapeutically targeted (Mays et al., 2019). Since there are immune cells in the tumor microenvironment, using immune receptor suppressors can shift the balance in the tumor microenvironment, which makes certain breast cancers more responsive to immunotherapy (Mays et al., 2019). This study also looks at how the PD1 and PDL1 genes, both of which are represented in the data used for this project, can me targeted by immune suppressants to improve the efficacy of breast cancer treatments for certain patients. Again, if a correlation is found between these gene expressions and better treatment outcomes, then it would help support the findings that the types of treatments discussed in these studies may be useful in treating breast cancer.

The next paper, which was published in 2020, looked at current studies that were using immune checkpoint inhibitors as a way of suppressing tumor immune cell suppression (Gaynor et al., 2020). The study provides an overview of trials that are using immune checkpoint inhibitors to target the PD1, PDL1, and/or CTLA4 genes for various disease types, discusses the various treatments that are being used, and provides a synopsis of how those treatments are being used to fight breast cancer specifically (Gayner et al., 2020). While the paper states that some of the studies that had inhibited the CTLA4 gene increased the immune response in patients with breast cancer, more work was still needed to examine the effectiveness of CTLA-4 inhibition in cancer treatment (Gayner et al., 2020). This paper also discusses inhibition of three of the genes that were represented in the data for this capstone project. So, again, establishing a connection between them and treatment outcomes would help to clarify if they may be an effective treatment.

The final paper was a meta-analysis, which was also published in 2020, that combined 27 studies with 1,746 patients. The study found that the tumor T-cell level was an ideal predictor of immune cell therapy response, and that PDL1 positive patients had a better 1-year progression-free survival, and a better 2-year overall survival, after receiving an immune checkpoint inhibitor (Zou et al., 2020). However, 25% of the patients who received an immune checkpoint inhibitor treatment, experienced a grade 3 or 4 treatment-related adverse event, and 15% experienced a grade 3 to 4 immune-related adverse event (Zou et al., 2020). While this paper only looked at inhibitors targeting the PDL1 gene, it is still useful as background for this project, as it also discusses the potential downsides of immune suppressant treatments. If this project fails to establish a correlation between the gene expressions and treatment outcomes, it will still be useful, as identifying patients who won’t benefit, will prevent them from having to pay the high costs, and experiencing the negative physical side effects that can come with immune inhibitory treatment.

A review of the literature examining immune checkpoint therapy effectiveness, especially in regards to blocking the PD1, PDL1, and CTLA4 genes, reveals that there is still a lot of information that needs to be gathered on these treatments’ effectiveness. While some initial signs are promising, the use of immune inhibitors is associated with significant adverse events, and significantly higher treatment cost, due to the fact that they often require a patient to have an in-person stay, or to have to make a trip to the emergency room, so their use needs to be closely monitored (George et al., 2021). This project will look to add a little more clarity to the immune checkpoint inhibitor conversation, by statistically analyzing if patient PDL1, PD1, CTLA4, and TCRa gene levels have an impact on the overall treatment outcome for individual patients. Evaluating the questions and hypotheses that are outlined above, may help to provide further insight into whether or not immune checkpoint inhibitors should be more broadly utilized, in regards to breast cancer treatment.

**Data Dictionary for Capstone Dataset**

**Figure 1***Data dictionary for combined dataset*****

**Research Design**

This paper will utilize a quantitative research approach to evaluate the questions and the hypotheses outlined above. This quantitative research will be conducted on two different statistical analysis software programs, SAS Studio, and RStudio. SAS Studio will be utilized to conduct the statistical analysis required to evaluate the questions and hypotheses laid out in this paper. RStudio will then be used to create data visualizations, to further explore some of the relationships in the data.

**Methodology and Methods Used**

To test to see if any of the variables can be used to predict how long a patient survived, or the length of time it had been since they were diagnosed, multiple regression was used. Multiple regression is a statistical technique that uses several continuous or categorical variables to predict the outcome of one continuous dependent variable (Hayes, & Estevez, 2021). The goal was to model the linear relationship between the dependent and independent variables, if indeed one existed. The analysis revealed p-values for each of the variables in the dataset, which were compared to a confidence interval of 5%, to determine if they were statistically significant in predicting treatment outcomes.

To test to see if any of the variables in the dataset can be used to predict the vital status, or whether a patient was alive or dead, logistic regression was used. Logistic regression models are classification models that can be used to predict a binary outcome, or a variable with only two possible values (Z\_ai, 2020). Since vital status has only two possible values, alive and dead, logistic regression seemed to be a sufficient way of determining if any of the variables had an impact on that outcome. The multiple logistic regression analysis again revealed p-values for all of the variables, which were again compared to the pre-selected confidence interval of 5%, to determine if any of them had a statistically significant impact, on whether a patient was alive or dead.

Finally, t-tests were used to determine if the top and bottom gene quartiles, had significantly different averages when it came to vital status, the length of time a patient survived, or how long it had been since the patient was diagnosed. A t-test is an inferential statistic that is used to determine if there is statistically significant difference between the averages of two groups, who are similar to one another when it comes to other features (Hayes et al, 2021). The t-test analysis also generates a p-value, which was again compared to a confidence interval of 5%, to determine if the differences of the means for the various high and low gene groups, was statistically significant.

RStudio was then used to create data visualizations to help further explore the relationships between the high or low gene groups. Scatterplots were created to look to see if the high and low groups could be visually identified based on the level of each gene expression. Then the same groups were compared using Kaplan-Meier curves, which are used to estimate the survival rate of two groups, and are especially useful when it is unknown if all of the subjects completed the entirety of the study (Rich et al., 2010). To create the Kaplan-Meier curve, the variables DaysToDeath, and DaysSinceFirstDiagnosed were combined, to create one continuous variable. This variable was then used, along with the VitalStatus, to create the Kaplan-Meier curves, which show the death expectancy percentage for the various high and low groups, over a certain period of time.

**Limitations**

There are a couple of limitations with the final capstone project. The first is that the DaysSinceFirstDiagnosed variable is quite vague, because it probably includes a number of individuals who did not fully complete the study. The reasons they may have left the study are also unknown, and their final health outcomes, some might have died after leaving the study for example, are also unclear. The second issue is that there are not enough Hispanic or Latino individuals contained in the data to draw any meaningful insights into how these groups gene expressions relate to individual patient health outcomes, so any of the results discussed, may not be applicable to that group. Also, while the studies population does reflect the population distribution nationwide, as the races are all adequately represented, there are still far more white and Caucasian people represented in the dataset, so the results may be most applicable to that group.

**Ethical Considerations**

While big data projects have been used to advance the fields of medicine and public health, there are still ethical considerations that need to be made, specifically in regards to respecting patient autonomy, ensuring equity, and respecting patient privacy (Howe III, & Ellenberg, 2020). Informed consent forms, which are intended to provide patients an understanding of the purpose, risks, and project methodology that will be used, are the main method medical researchers use to respect patient autonomy (Howe III, & Ellenberg, 2020). However, due to the differences between big data projects and traditional research, it’s possible the standard informed consent process is not an adequate way of protecting patient autonomy. Many participants may not fully realize who is actually able to access their data, which is essentially anybody once it is made public, or the inferences that can be made through the analysis of said data (Howe III, & Ellenberg, 2020). Since the point of a lot of big data research is to identify unknown correlations, associations and trends, it is difficult to fully identify all of the ways the data might be used, which means people’s information might be used in research that they morally oppose (Howe III, & Ellenberg 2020).

It can be difficult to achieve equity with big data medical research, as results that come from data that is concentrated too heavily with one demographic, can easily and inadvertently perpetuate social disparities (Howe III, & Ellenberg, 2020). When big data projects use data where a single group is over represented, whether it be race, ethnicity, country of origin, or socioeconomic class, the conclusions of these studies may only be relevant to a certain group, and may not translate, or may even have a negative impact, when the results are applied to other groups (Howe III, & Ellenberg, 2020). Genetic data is generally produced by either people with quality health insurance, or who have the means to afford genetic tests, so wealthy people are over represented, which means that findings from studies on genetic data may only be applicable to that group (Howe III, & Ellenberg, 2020).

There are numerous examples of how big data research on medical data can imperil patient privacy, but this section will focus on the ethical concerns around de-identified data. While research patients are often assured that it is not possible to identify them with the data they are providing, this if often not the case (Howe III, & Ellenberg, 2020). Researchers have demonstrated the ability to use genetic data to identify the person who provided the sample (Howe III, & Ellenberg, 2020). As advancements in the field continue, it’s likely that most of the public “de-identified” data will eventually become re-identifiable.

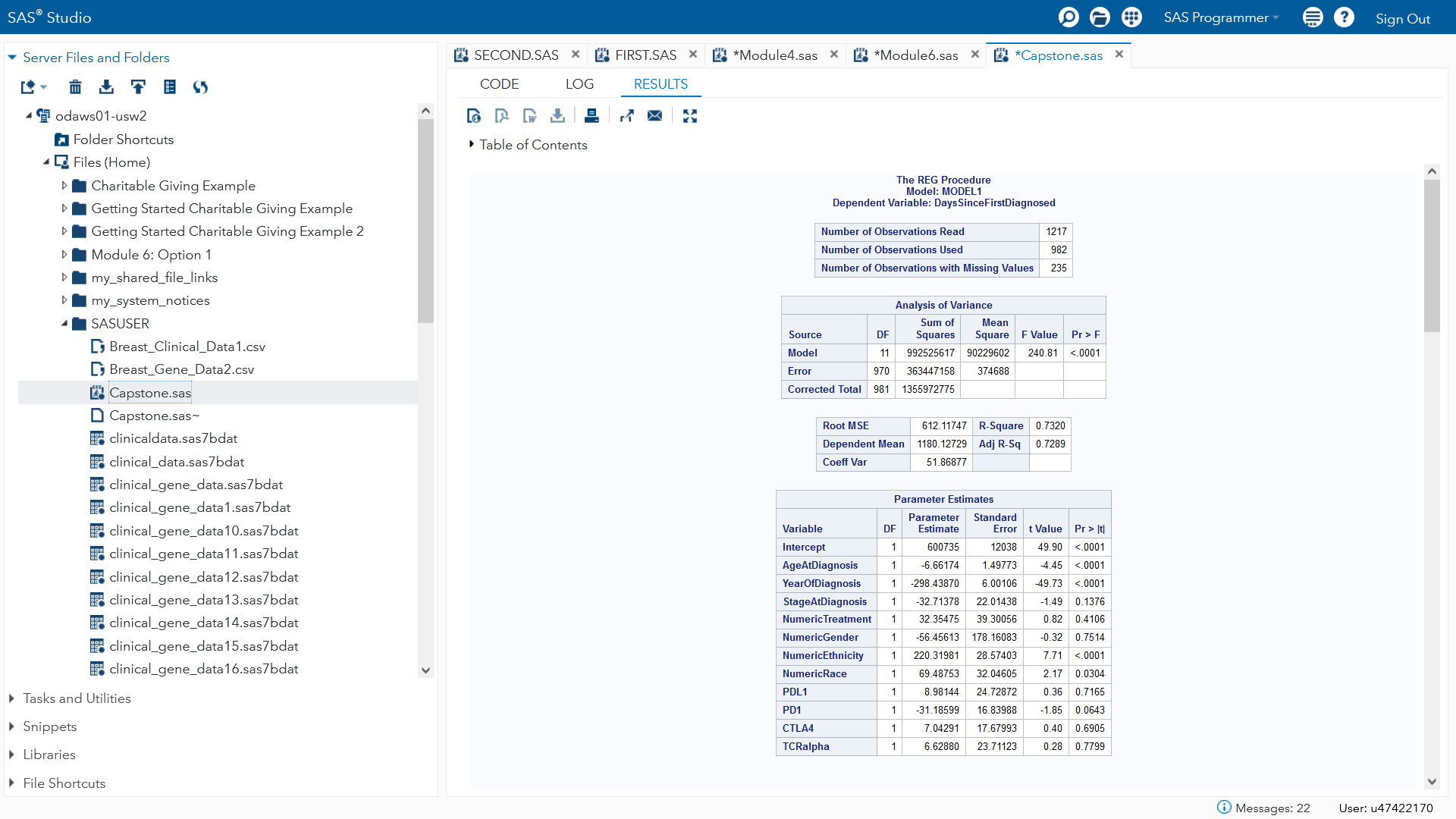
The ethical issues concerning medical data discussed above are all applicable to the data that was used for this project. It is unknow what the informed consent form process consisted of when the represented patients submitted their genetic information, or if they would be aware that their information would be used in a project of this nature. The initial analysis did reveal that Hispanic or Latino individuals were not adequately represented in the data, so the conclusions drawn may not be applicable to this demographic. Finally, it is possible, that since the data contains genetic information, it could be used to identify the individuals in the dataset.

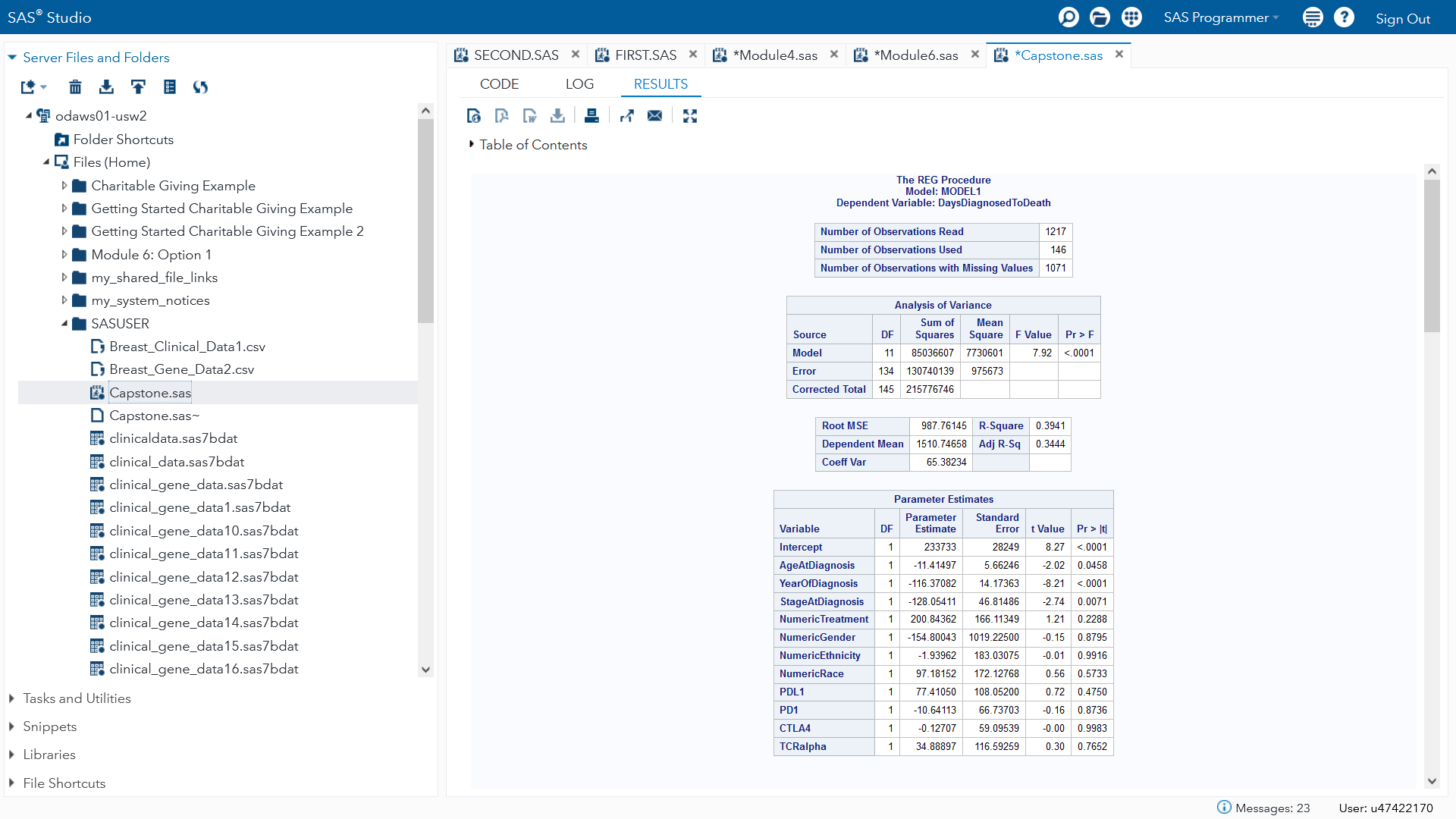
However, steps have been taken to help prevent a negative ethical situation from occurring. One would assume that the patients who submitted their genetic data would be fine with their data being used for this project, even if this type of research was not explicitly outlined in the informed consent process. The fact that there is not a sufficient representation of Hispanic and Latino individuals, which may mean the results are not applicable to this group, was explicitly mentioned above. And finally, no attempts were made to identify the individuals in the dataset using their genetic data, even though this may be a possibility.

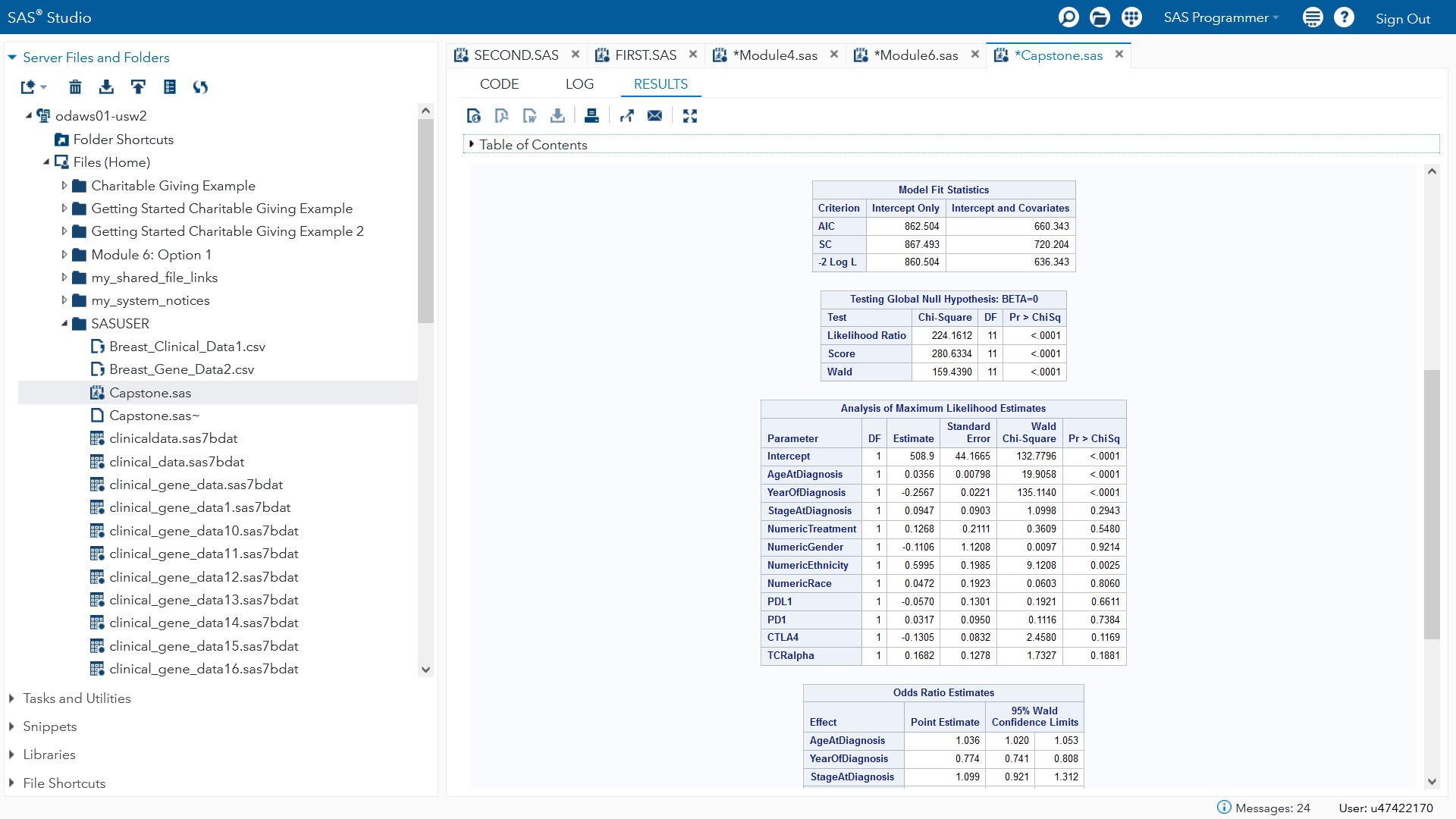
**Findings**

The statistical analysis preformed for this study did not reveal a statistically significant relationship between a patient’s gene expressions, and better or worse treatment outcomes. The regression analysis of the genes showed that none of them were statistically significant in predicting if a patient was alive or dead, how long they had survived before they died, or the amount of time it had been since they were diagnosed. This was determined by comparing the p-values that were generated by the regression analysis for each variable, to the predetermined confidence level of 95%.

The regression analysis did reveal that the year and the age of a patient were statistically relevant in predicting the previously discussed variables which makes sense, as the patient is less likely to survive as long if they are diagnosed at an older age, and cancer treatments have improved over time. Ethnicity was also revealed to be a statistically significant variable for predicting how long it had been since a patient was diagnosed, as well as a patients vital status, but this finding should be discounted, as there are not enough Hispanic or Latino patients represented in the data, to draw any significant conclusions about this group. The results of the multiple, and logistic regression models, can be seen in Figures 2 through 4.

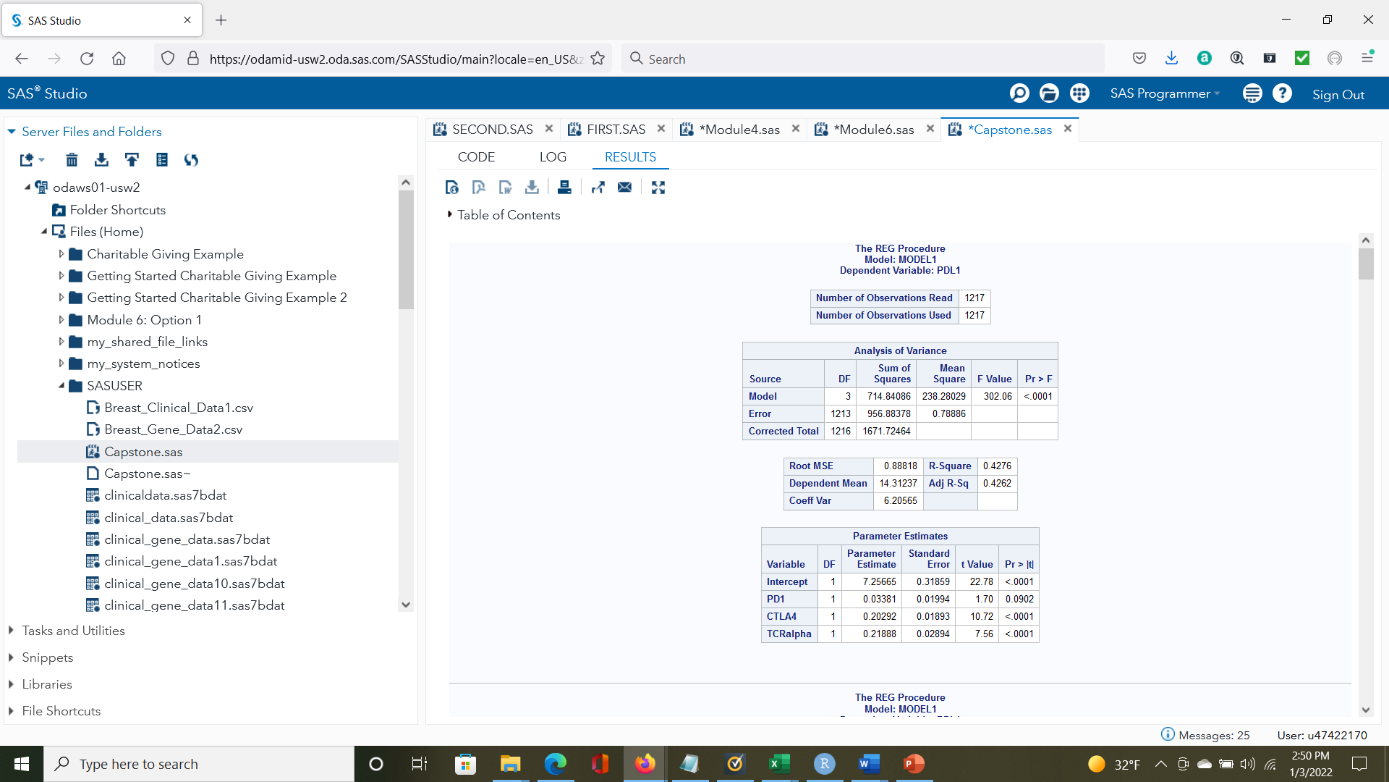
**Figure 2***Multiple regression analysis on DaysSinceFirstDiagnosed variable  
*

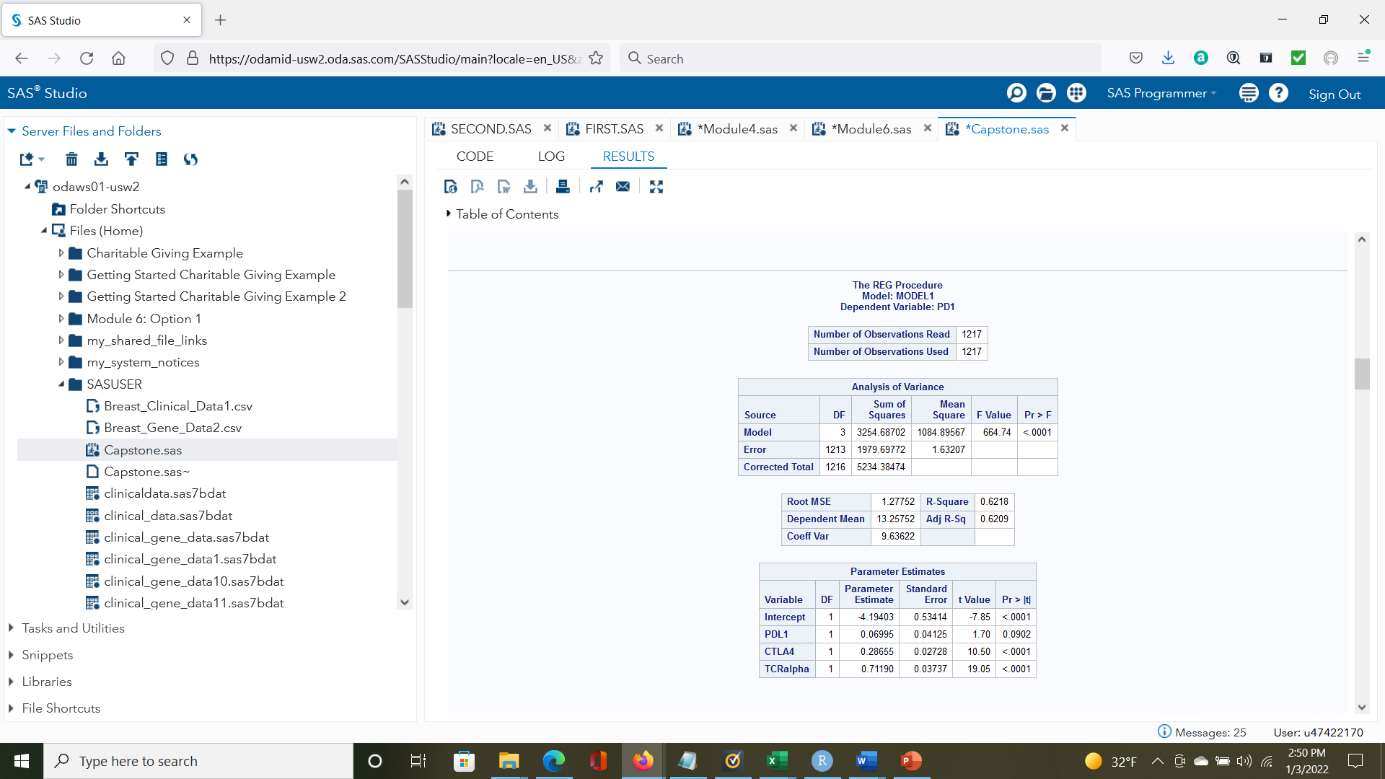
**Figure 3***Multiple regression analysis on DaysDiagnosedToDeath variable  
*

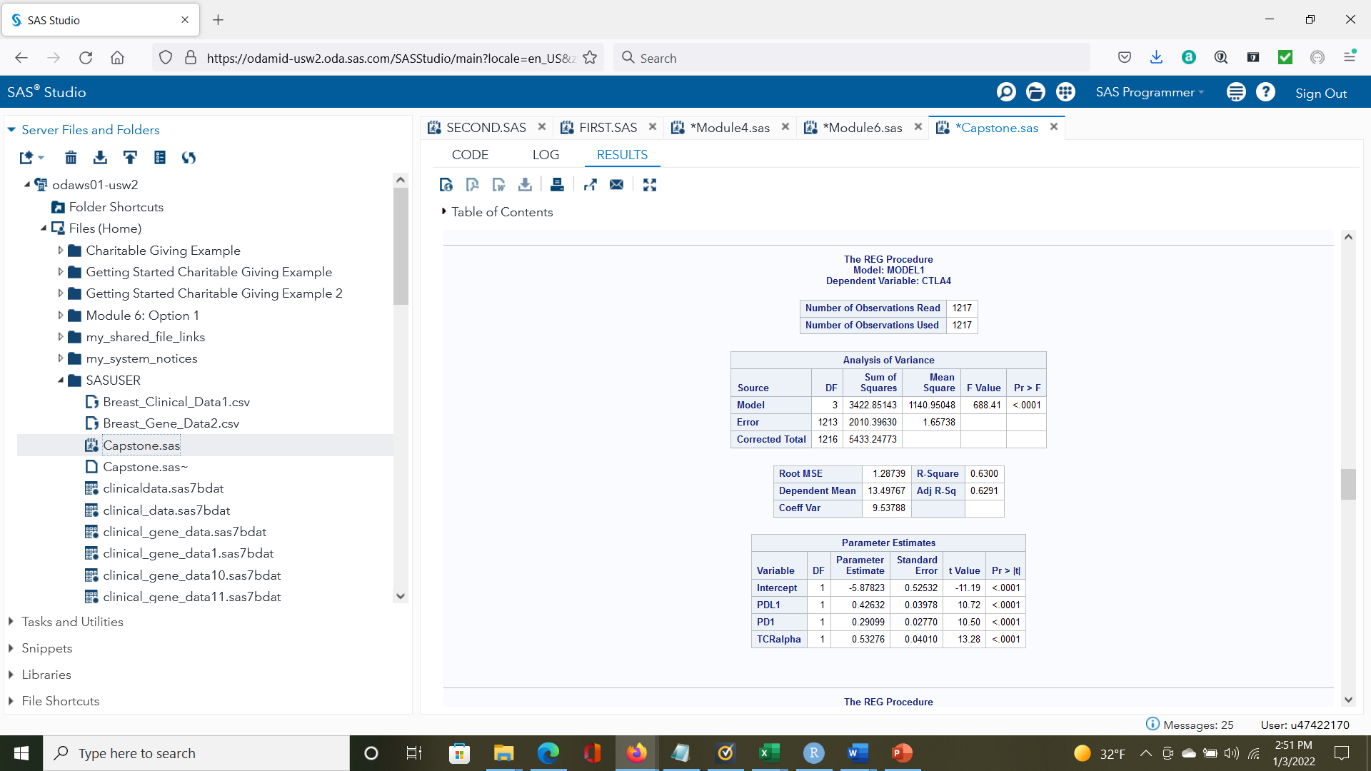
**Figure 4***Logistic regression on patient vital status  
*

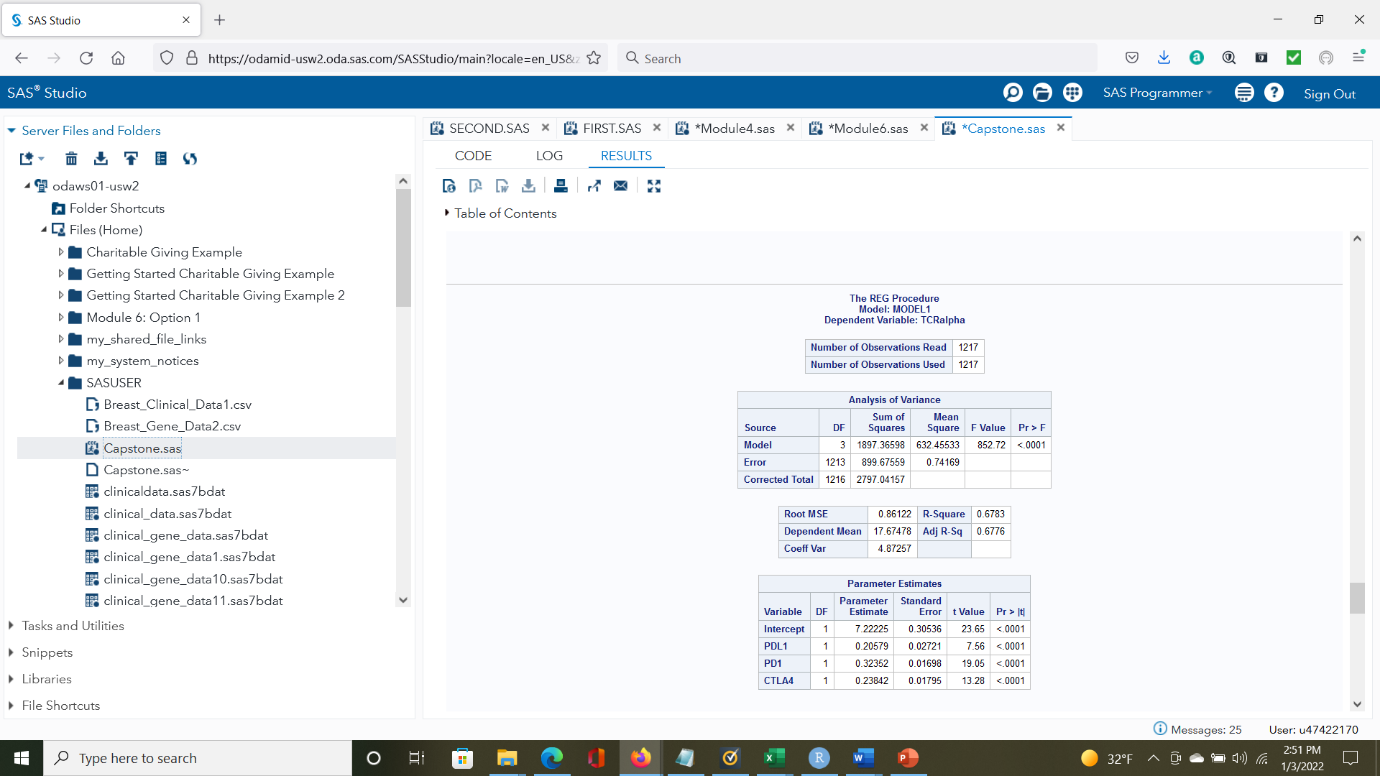
**Regression on Gene Expressions**

However, when regression was performed on the gene expressions, with the other gene expression used as the independent variables, the analysis dis show that they were correlated with one another. PD1 and PDL1 expressions were not found to be significant for predicting the expressions of the other, but they were significant in predicting both CTLA4, and TCRa. Ctla4 and TCRa were found to be significant in all of the gene expression regression models. All of the coefficients generated for the regression equations were positive, which means that patients with high gene expressions for one gene, can be expected to have higher gene expressions for the other genes in the dataset. The patients with the highest gene expressions, can be expected to respond best to immune suppressant therapy, if it is ever determined to be an effective treatment. The results of this regression analysis can be seen in Figures 5 through 8.

**Figure 5***Regression analysis on PDL1 gene expression  
*

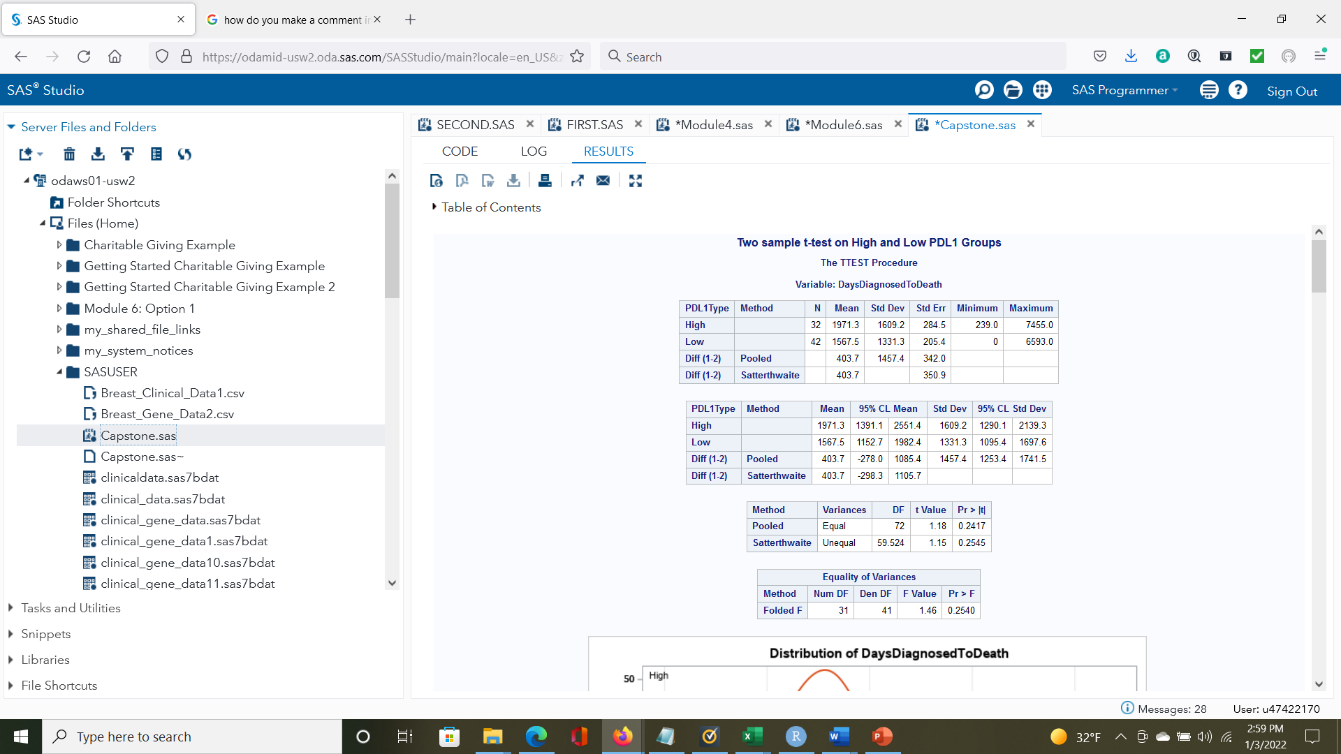
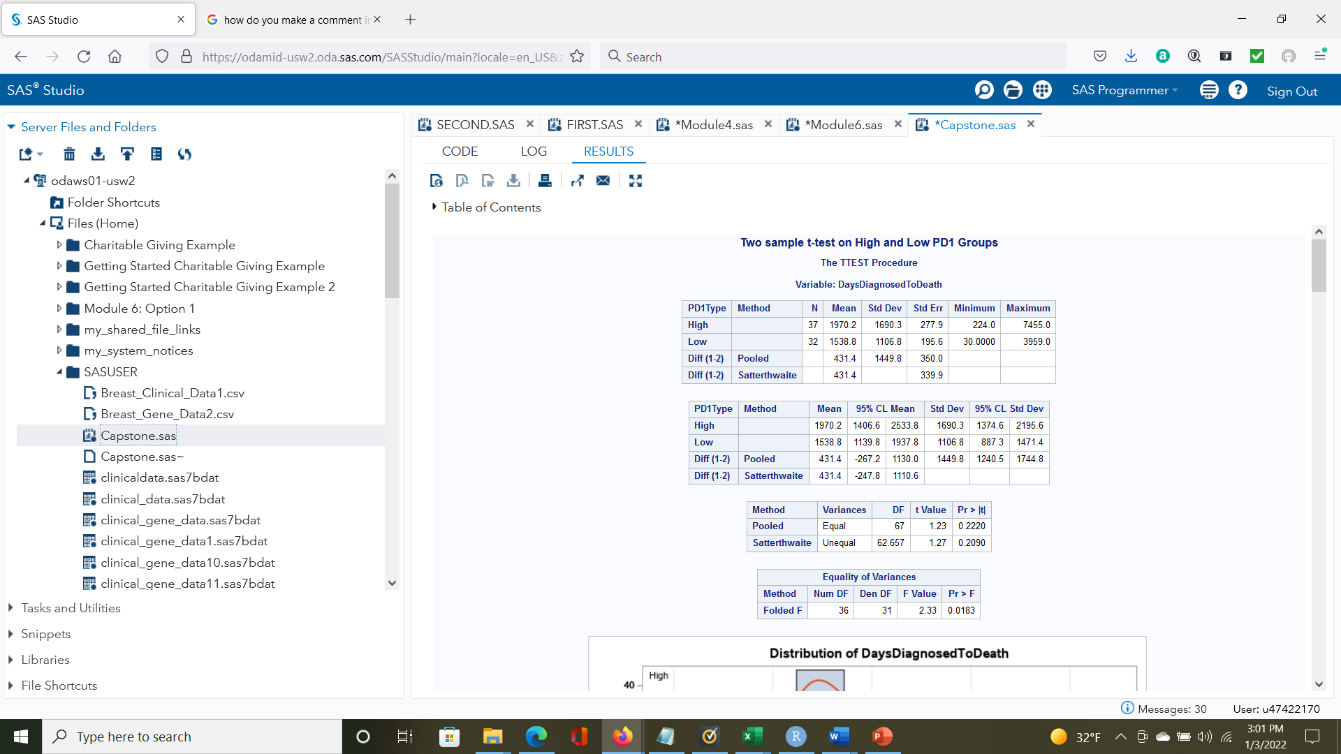
**Figure 6***Regression analysis on PD1 gene expression*

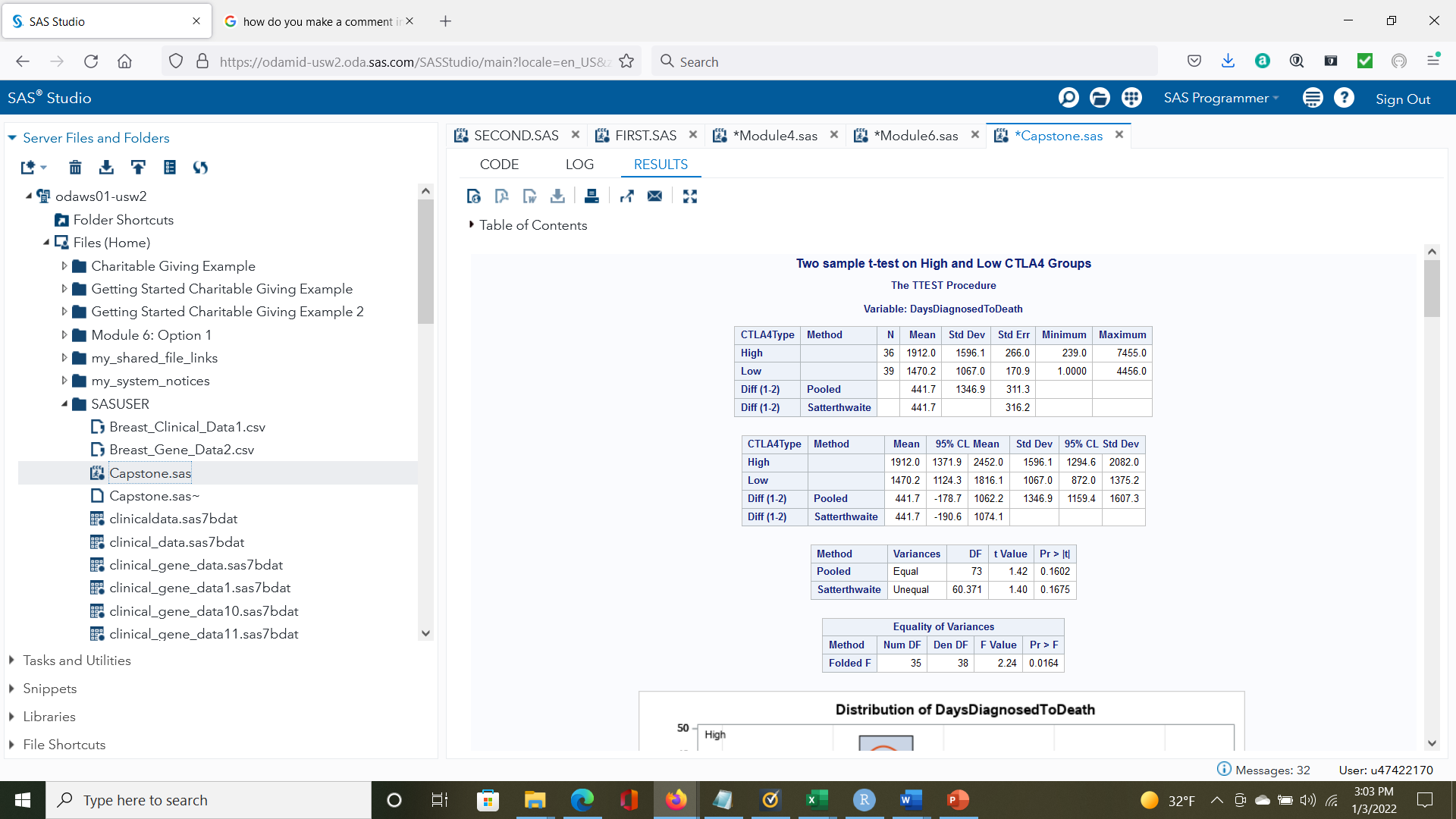
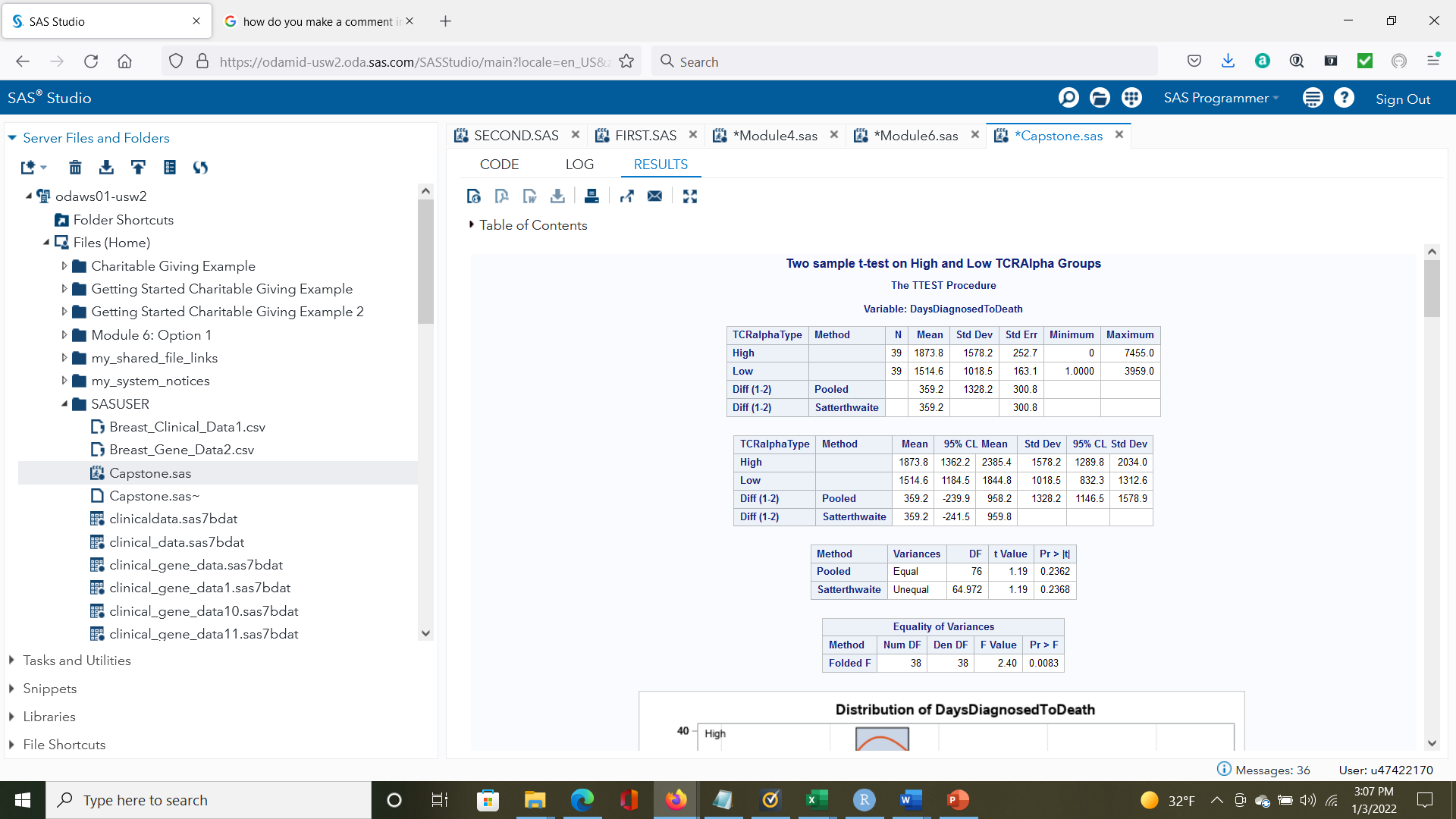
**Figure 7***Regression analysis on CTLA4 gene expression*****

**Figure 8***Regression analysis on* TCRa *gene expression  
*

**T-test results**

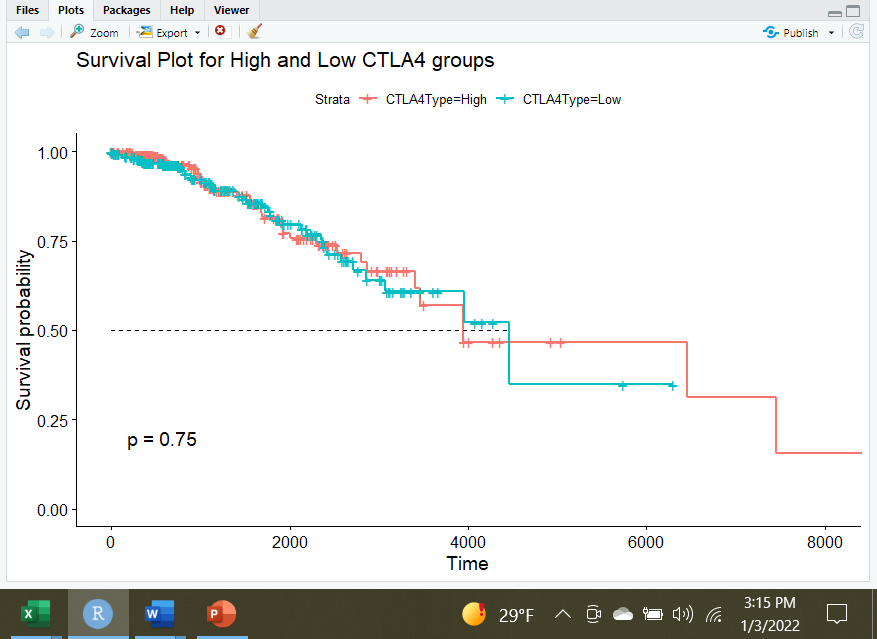
The results of the t-test were the same as the other statistical analysis, where no statistical significant difference was discovered in the average time it took for a patient to die, the amount of time it had been since they were diagnosed, or a patient’s vital status, when the low and high gene expression groups were compared. It is interesting to note that the high gene expression groups for all the genes did live longer on average before death, when compared to the low groups, but the differences in the means was determined to not be statistically significant, because the p-values where all higher than the pre-established confidence interval of .05. To save space, only the results of the t-tests for the DaysDiagnosedToDeath variable were included in this paper. The results of those tests can be seen in Figures 9 through 12.

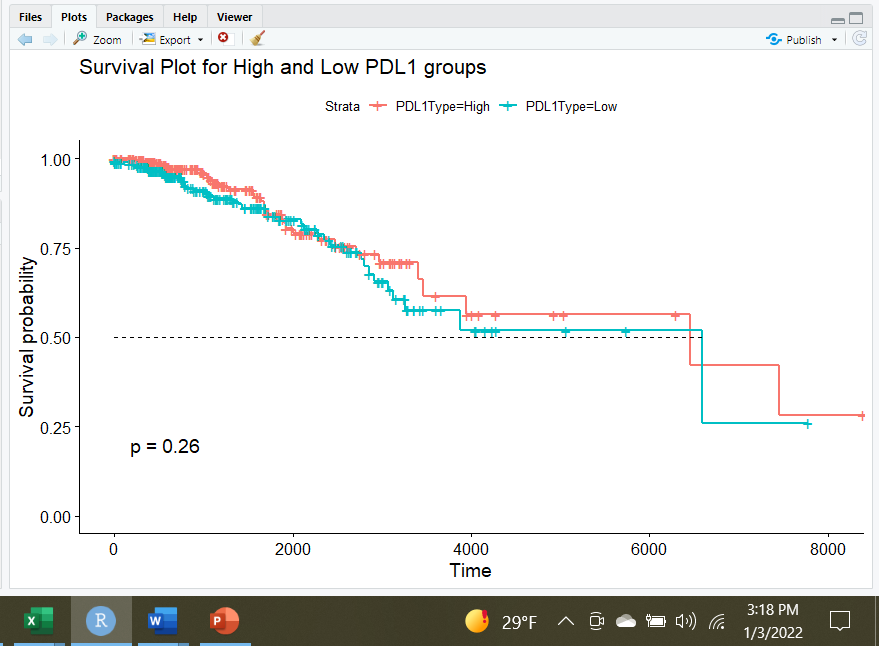
**Figure 9***T-test on DaysDiagnosedToDeath for High and Low PDL1 groups  
***Figure 10***T-test on DaysDiagnosedToDeath for High and Low PD1 groups  
*

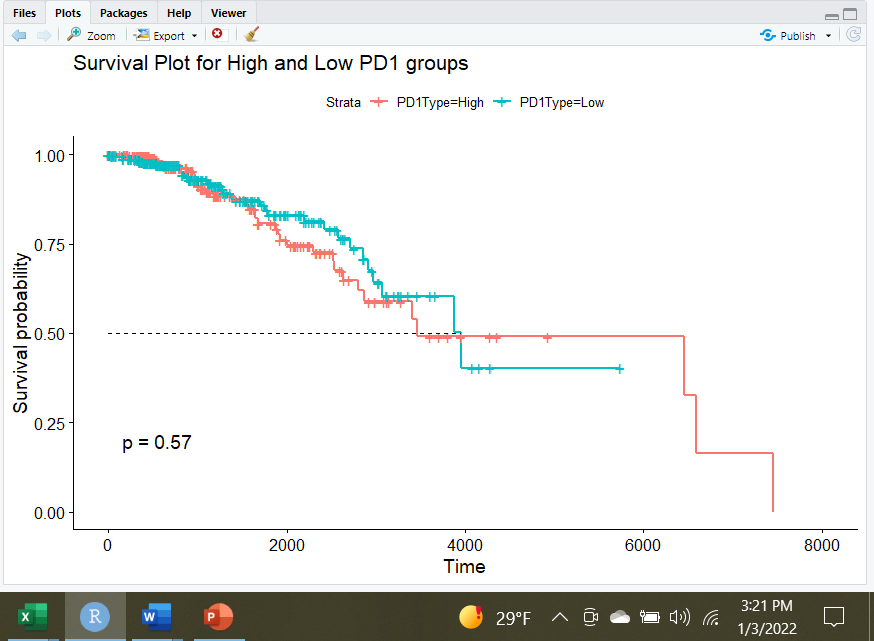
**Figure 11***T-test on DaysDiagnosedToDeath for High and Low CTLA4 groups  
* **Figure 12***T-test on DaysDiagnosedToDeath for High and Low TCRa groups*******

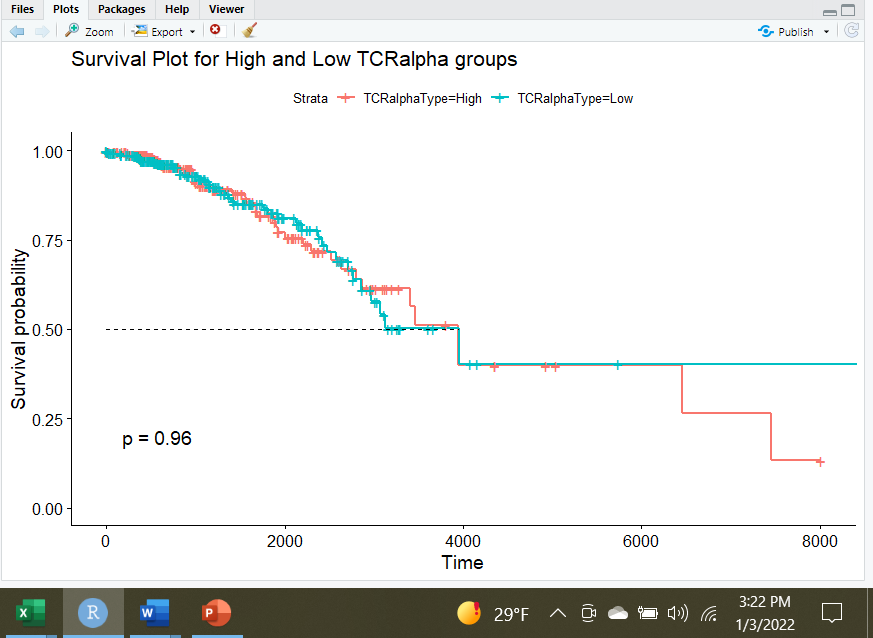
**Kaplan-Meier Curves**

The Kaplan-Meier curves that were produced using R-Studio visually show that there is little difference between the low and high gene groups when it comes to the statistical probability of experiencing an adverse event, in this case, whether a person dies or not. There is some separation between the curves that can be observed between the low and high groups, especially for the PDL1 groups, but the generated p-value shows that the differences that can be observed are not statistically significant. The Kaplan-Meier curves for each of the high and low gene expression groups can be seen below, in Figures 13 through 16.

**Figure 13***Kaplan-Meier curve for high and low CTLA4 groups*****

**Figure 14***Kaplan-Meier curve for high and low PDL1 groups  
*

**Figure 15***Kaplan-Meier curve for high and low PD1 groups  
*

**Figure 16***Kaplan-Meier curve for high and low TCRa groups  
*

**Conclusions**

The analysis conducted for this project did not reveal a statistically significant relationship between a patient’s gene expressions, and better or worse treatment outcomes. This means that for the seven questions research questions and sets of hypotheses outlined above, we fail to reject the null hypothesis, for all of the seven scenarios. While no statistical relationships were found more work is still needed to determine if any breast cancer patient could benefit from immune receptor suppressants. It is possible that using quartiles as a grouping mechanism for the t-tests, did not significantly differentiate the groups. It may be the case that if the top and bottom 5% were compared to one another, statistical differences might be found in the health-outcome averages. Just because this research did not reveal any statistical relationships between the gene expression in the data and better or worse health outcomes does not mean it isn’t still useful, as identifying patients who will not benefit from immune suppression therapies is still valuable, as they will not be subject to the high costs, and numerous adverse effects that are associated with that type of treatment (Zou et al, 2020).

**Recommendations**

More work is still needed to determine if immune response repressors, especially ones that target the genes discussed in this project, would be effective for treating breast cancer patients. The results of this project show that a patient’s gene expressions for the four genes in the data don’t necessarily correlate with better health outcomes, but it’s possible that the groups that were created were not sufficiently different from one another. It’s possible that certain breast cancer patients could benefit from immune receptor suppression treatment, but the number of people in this group is likely very small. Further work needs to go into identifying the individuals who may benefit, however small that group may be, as this may be their only chance at improving their chance at survival. However, identifying the individuals who won’t benefit will still be beneficial, because as discussed before, it prevents unnecessary costs and adverse side effects.

**Project Code**

The code that was used to conduct the analysis in SAS Studio, and create the data visualizations in RStudio, is available below.

**SAS Code**

\* import data sets;

libname MyFiles '/home/u47422170/SASUSER';

data MyFiles.Gene\_Data;

infile '/home/u47422170/SASUSER/Breast\_Gene\_Data2.csv' dlm=',' firstobs=2;

input CaseID $12. PDL1 $ PD1 $ CTLA4 $ TCRalpha ;

libname MyFiles '/home/u47422170/SASUSER';

data MyFiles.Clinical\_Data;

infile '/home/u47422170/SASUSER/Breast\_Clinical\_Data1.csv' dlm=',' firstobs=2;

input CaseID $12. AgeAtDiagnosis $ YearOfBirth $ Gender $ Ethnicity $ Race $ VitalStatus $ DaysDiagnosedToDeath $ YearOfDeath $ DaysSinceFirstDiagnosed $ YearOfDiagnosis $ TreatmentType $ StageAtDiagnosis;

\*merge datasets;

data MyFiles.Clinical\_Gene\_Data;

merge MyFiles.Gene\_Data MyFiles.Clinical\_Data;

run;

proc print data=MyFiles.Clinical\_Gene\_Data;

where ethnicity ="not repo";

title 'Data Set COMBINED';

run;

proc means data = myfiles.clinical\_gene\_data;

run;

\*change variables in the dataset from character to numeric, create new variables;

data MyFiles.Clinical\_Gene\_Data1;

set MyFiles.Clinical\_Gene\_Data(rename=(CaseID=old));

CaseID = old;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data2;

set MyFiles.Clinical\_Gene\_Data1(rename=(AgeAtDiagnosis=old));

AgeAtDiagnosis = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data3;

set MyFiles.Clinical\_Gene\_Data2(rename=(YearOfBirth=old));

YearOfBirth = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data4;

set MyFiles.Clinical\_Gene\_Data3(rename=(Gender=old));

Gender = old;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data5;

set MyFiles.Clinical\_Gene\_Data4(rename=(Gender=Gender));

If Gender ='female' then NumericGender = 1; else NumericGender = 2;

run;

data MyFiles.Clinical\_Gene\_Data6;

set MyFiles.Clinical\_Gene\_Data5(rename=(Ethnicity=old));

Ethnicity = old;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data7;

set MyFiles.Clinical\_Gene\_Data6(rename=(Ethnicity=Ethnicity));

if Ethnicity ="not hisp" then NumericEthnicity =2; else if Ethnicity ='hispanic' then NumericEthnicity = 1; else NumericEthnicity = 0;

run;

data MyFiles.Clinical\_Gene\_Data8;

set MyFiles.Clinical\_Gene\_Data7(rename=(Race=old));

Race = old;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data9;

set MyFiles.Clinical\_Gene\_Data8(rename=(Race=Race));

if race = "white" then NumericRace = 1; else if race = "black or" then NumericRace = 2; else if race = "asian" then NumericRace = 3; else NumericRace = 0 ;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data10;

set MyFiles.Clinical\_Gene\_Data9(rename=(VitalStatus=old));

VitalStatus = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data11;

set MyFiles.Clinical\_Gene\_Data10(rename=(DaysDiagnosedToDeath=old));

DaysDiagnosedToDeath = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data12;

set MyFiles.Clinical\_Gene\_Data11(rename=(YearOfDeath=old));

YearOfDeath = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data13;

set MyFiles.Clinical\_Gene\_Data12(rename=(DaysSinceFirstDiagnosed=old));

DaysSinceFirstDiagnosed = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data14;

set MyFiles.Clinical\_Gene\_Data13(rename=(VitalStatus=VitalStatus));

if VitalStatus = "1" then LiveDieTime = DaysDiagnosedToDeath; else LiveDieTime = DaysSinceFirstDiagnosed;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data15;

set MyFiles.Clinical\_Gene\_Data14(rename=(YearOfDiagnosis=old));

YearOfDiagnosis = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data16;

set MyFiles.Clinical\_Gene\_Data15(rename=(TreatmentType=old));

TreatmentType = old;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data17;

set MyFiles.Clinical\_Gene\_Data16(rename=(TreatmentType=TreatmentType));

If TreatmentType ='PT' then NumericTreatment = 1; else NumericTreatment = 2;

run;

data MyFiles.Clinical\_Gene\_Data18;

set MyFiles.Clinical\_Gene\_Data17(rename=(StageAtDiagnosis=old));

StageAtDiagnosis = old;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data19;

set MyFiles.Clinical\_Gene\_Data18(rename=(PDL1=old));

PDL1 = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data20;

set MyFiles.Clinical\_Gene\_Data19(rename=(PD1=old));

PD1 = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data21;

set MyFiles.Clinical\_Gene\_Data20(rename=(CTLA4=old));

CTLA4 = input(old, 8.);

drop old;

run;

data MyFiles.Clinical\_Gene\_Data\_Final;

set MyFiles.Clinical\_Gene\_Data21(rename=(TCRalpha=old));

TCRalpha = old;

drop old;

run;

data MyFiles.Clinical\_Gene\_Data\_Final;

set myfiles.clinical\_gene\_data\_final;

if Gender = "" then delete;

run;

proc means data =myfiles.Clinical\_Gene\_Data\_Final N Nmiss mean std min P1 P5 P10 P25 P50 P75 P90 P95 P99 max;

run;

\*Run regression equations on data;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model DaysSinceFirstDiagnosed = PDL1 PD1 CTLA4 TCRalpha;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model DaysSinceFirstDiagnosed = AgeAtDiagnosis YearOfDiagnosis StageAtDiagnosis NumericTreatment NumericGender NumericEthnicity NumericRace;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model DaysDiagnosedToDeath = PDL1 PD1 CTLA4 TCRalpha;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model DaysDiagnosedToDeath = AgeAtDiagnosis YearOfDiagnosis StageAtDiagnosis NumericTreatment NumericGender NumericEthnicity NumericRace;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model DaysSinceFirstDiagnosed = AgeAtDiagnosis YearOfDiagnosis StageAtDiagnosis NumericTreatment NumericGender NumericEthnicity NumericRace PDL1 PD1 CTLA4 TCRalpha;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model DaysDiagnosedToDeath = AgeAtDiagnosis YearOfDiagnosis StageAtDiagnosis NumericTreatment NumericGender NumericEthnicity NumericRace PDL1 PD1 CTLA4 TCRalpha;

run;

proc freq data=MyFiles.Clinical\_Gene\_Data\_Final;

table Ethnicity;

run;

proc logistic data=MyFiles.Clinical\_Gene\_Data\_Final descending;

model VitalStatus = AgeAtDiagnosis YearOfDiagnosis StageAtDiagnosis NumericTreatment NumericGender NumericEthnicity NumericRace PDL1 PD1 CTLA4 TCRalpha ;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model PDL1 = PD1 CTLA4 TCRalpha;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model PD1 = PDL1 CTLA4 TCRalpha;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model CTLA4 = PDL1 PD1 TCRalpha;

run;

proc reg data=myfiles.clinical\_gene\_Data\_final PLOTS(MAXPOINTS=NONE);

model TCRalpha = PDL1 PD1 CTLA4;

run;

proc export data=work.high\_low\_tcralpha\_labled

outfile='/home/u47422170/SASUSER\highlowtcralpha.csv'

dbms=csv

replace;

run;

\*Determine quartiles using proc univariate, and seperate patients into high and low gene groups, run t-tests on those groups;

proc univariate data =myfiles.clinical\_gene\_data\_final;

var TCRAlpha;

run;

DATA High\_PDL1;

SET myfiles.clinical\_gene\_data\_final;

IF (PDL1 GE 15.09 ) THEN OUTPUT;

RUN;

DATA Low\_PDL1;

SET myfiles.clinical\_gene\_data\_final;

IF (PDL1 LE 13.8) THEN OUTPUT;

RUN;

proc sort data = High\_PDL1;

by CaseID;

run;

proc sort data = Low\_PDL1;

by CaseID;

run;

data High\_Low\_PDL1;

merge Low\_PDL1 High\_PDL1 ;

by CaseID;

run;

data High\_Low\_PDL1\_labled;

set High\_Low\_PDL1;

if (PDL1 GE 15.09 ) then PDL1Type = 'High';

if (PDL1 LE 13.8) then PDL1Type = 'Low';

run;

proc ttest data = High\_Low\_PDL1\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low PDL1 Groups";

class PDL1Type;

var DaysDiagnosedToDeath;

run;

proc ttest data = High\_Low\_PDL1\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low PDL1 Groups";

class PDL1Type;

var DaysSinceFirstDiagnosed;

run;

proc ttest data = High\_Low\_PDL1\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low PDL1 Groups";

class PDL1Type;

var VitalStatus;

run;

DATA High\_PD1;

SET myfiles.clinical\_gene\_data\_final;

IF (PD1 GE 14.74 ) THEN OUTPUT;

RUN;

DATA Low\_PD1;

SET myfiles.clinical\_gene\_data\_final;

IF (PD1 LE 12.28) THEN OUTPUT;

RUN;

proc sort data = High\_PD1;

by CaseID;

run;

proc sort data = Low\_PD1;

by CaseID;

run;

data High\_Low\_PD1;

merge Low\_PD1 High\_PD1 ;

by CaseID;

run;

data High\_Low\_PD1\_labled;

set High\_Low\_PD1;

if (PD1 GE 14.74 ) then PD1Type = 'High';

if (PD1 LE 12.28) then PD1Type = 'Low';

run;

proc ttest data = High\_Low\_PD1\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low PD1 Groups";

class PD1Type;

var DaysDiagnosedToDeath;

run;

proc ttest data = High\_Low\_PD1\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low PD1 Groups";

class PD1Type;

var DaysSinceFirstDiagnosed;

run;

proc ttest data = High\_Low\_PD1\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low PDL1 Groups";

class PD1Type;

var VitalStatus;

run;

DATA High\_CTLA4;

SET myfiles.clinical\_gene\_data\_final;

IF (CTLA4 GE 15.06 ) THEN OUTPUT;

RUN;

DATA Low\_CTLA4;

SET myfiles.clinical\_gene\_data\_final;

IF (CTLA4 LE 12.58) THEN OUTPUT;

RUN;

proc sort data = High\_CTLA4;

by CaseID;

run;

proc sort data = Low\_CTLA4;

by CaseID;

run;

data High\_Low\_CTLA4;

merge Low\_CTLA4 High\_CTLA4 ;

by CaseID;

run;

data High\_Low\_CTLA4\_labled;

set High\_Low\_CTLA4;

if (CTLA4 GE 15.06) then CTLA4Type = 'High';

if (CTLA4 LE 12.58) then CTLA4Type = 'Low';

run;

proc ttest data = High\_Low\_CTLA4\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low CTLA4 Groups";

class CTLA4Type;

var DaysDiagnosedToDeath;

run;

proc ttest data = High\_Low\_CTLA4\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low CTLA4 Groups";

class CTLA4Type;

var DaysSinceFirstDiagnosed;

run;

proc ttest data = High\_Low\_CTLA4\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low CTLA4 Groups";

class CTLA4Type;

var VitalStatus;

run;

DATA High\_TCRalpha;

SET myfiles.clinical\_gene\_data\_final;

IF (TCRalpha GE 18.74 ) THEN OUTPUT;

RUN;

DATA Low\_TCRalpha;

SET myfiles.clinical\_gene\_data\_final;

IF (TCRalpha LE 16.94) THEN OUTPUT;

RUN;

proc sort data = High\_TCRalpha;

by CaseID;

run;

proc sort data = Low\_TCRalpha;

by CaseID;

run;

data High\_Low\_TCRalpha;

merge Low\_TCRalpha High\_TCRalpha ;

by CaseID;

run;

data High\_Low\_TCRalpha\_labled;

set High\_Low\_TCRalpha;

if (TCRalpha GE 18.74) then TCRalphaType = 'High';

if (TCRalpha LE 16.94) then TCRalphaType = 'Low';

run;

proc ttest data = High\_Low\_TCRalpha\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low TCRAlpha Groups";

class TCRalphaType;

var DaysDiagnosedToDeath;

run;

proc ttest data = High\_Low\_TCRalpha\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low TCRalpha Groups";

class TCRalphaType;

var DaysSinceFirstDiagnosed;

run;

proc ttest data = High\_Low\_TCRalpha\_labled sides = 2 alpha = 0.05 h0 = 0;

title "Two sample t-test on High and Low TCRalpha Groups";

class TCRalphaType;

var VitalStatus;

run;

**R Code**

setwd("C:/Users/nolan/OneDrive/Documents/MIS581")

getwd()

Clinical<-read.csv("ClinicalGeneData.csv", header=TRUE)

PD1<-read.csv("highlowpd1.csv", header=TRUE)

PDL1<-read.csv("highlowpdl1.csv", header=TRUE)

CTLA4<-read.csv("highlowctla4.csv", header=TRUE)

TCRalpha<-read.csv("highlowtcralpha.csv", header=TRUE)

head(Clinical)

install.packages("survival")

install.packages("ggplot2")

install.packages("survminer")

library(ggplot2)

library(survival)

library(survminer)

##Survival curve code for the entire dataset

Clinical$LiveDieTime

surv\_object <- Surv(time = Clinical$LiveDieTime, event = Clinical$VitalStatus)

surv\_object

fit1 <- survfit(surv\_object ~ Race, data = Clinical)

summary(fit1)

ggsurvplot(fit1, data = Clinical, pval = TRUE, surv.median.line ="h", title= "Survival Plot by Race" )

##Survival Curve for high and low CTLA4 count

CTLA4$LiveDieTime

CTLA4\_object <- Surv(time = CTLA4$LiveDieTime, event = CTLA4$VitalStatus)

CTLA4\_object

fitCTLA4 <- survfit(CTLA4\_object ~ CTLA4Type, data = CTLA4)

summary(fitCTLA4)

ggsurvplot(fitCTLA4, data = CTLA4, pval = TRUE, surv.median.line ="h", title="Survival Plot for High and Low CTLA4 groups")

##Survive curve for high and low PDL1

PDL1$LiveDieTime

PDL1\_object <- Surv(time = PDL1$LiveDieTime, event = PDL1$VitalStatus)

PDL1\_object

fitPDL1 <- survfit(PDL1\_object ~ PDL1Type, data = PDL1)

summary(fitPDL1)

ggsurvplot(fitPDL1, data = PDL1, pval = TRUE, surv.median.line ="h", title="Survival Plot for High and Low PDL1 groups")

##Survive curve for high and low PD1

PD1$LiveDieTime

PD1$PD1Type

PD1\_object <- Surv(time = PD1$LiveDieTime, event = PD1$VitalStatus)

PD1\_object

fitPD1 <- survfit(PD1\_object ~ PD1Type, data = PD1)

summary(fitPD1)

ggsurvplot(fitPD1, data = PD1, pval = TRUE, surv.median.line ="h", title="Survival Plot for High and Low PD1 groups")

##Survive curve for high and low TCRalpha

TCRalpha$LiveDieTime

TCRalpha\_object <- Surv(time = TCRalpha$LiveDieTime, event = TCRalpha$VitalStatus)

TCRalpha\_object

fitTCRalpha <- survfit(TCRalpha\_object ~ TCRalphaType, data = TCRalpha)

summary(fitTCRalpha)

ggsurvplot(fitTCRalpha, data = TCRalpha, pval = TRUE, surv.median.line ="h", title ="Survival Plot for High and Low TCRalpha groups")

**References**

ACS. (2021). How common is breast cancer? *American Cancer Society*. <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html>

Akinlye, A., Rasool, Z. (2019). Immune checkpoint inhibitors of PD-L1 as cancer therapeutics. *Journal of Hematology and Oncology*, 12(92). <https://jhoonline.biomedcentral.com/articles/10.1186/s13045-019-0779-5>

Breastcancer.org. (2021). U.S. breast cancer statistics. *Breastcancer.org*. <https://www.breastcancer.org/symptoms/understand_bc/statistics>

Desai, A., & Buchcinder, E. (2016). CTLA-4 and PD-1 pathways: Similarities, differences, and implications of their inhibition. *National Library of Medicine*. <https://pubmed.ncbi.nlm.nih.gov/26558876/>

Gatti-Mays, M., Balko, J., Gameiro, S., Bear, H., Prabhakarni, S., Fukui, J., Disis, M., Nanda, R., Gulley, J., Kalinsky, K., Sater, H., Sparano, J., Cescon, D., Page, D., McArthur, H., Adams, S., Mittendorf, E. (2019). If we build it they will come: targeting the immune response to breast cancer. npj Breast Cancer, 5(37). <https://www.nature.com/articles/s41523-019-0133-7>

Gaynor, N., Crown, J., Collins, D. (2020). Immune checkpoint inhibitors: Key trials and a emerging role in breast cancer. *Seminars in Cancer Biology*. <https://www.sciencedirect.com/science/article/pii/S1044579X20301528>

Gene Cards. (2021). CTLA4 gene- Cytotoxic T-lymphocyte associated protein 4. *Gene Cards- The Human Gene Database*. <https://www.genecards.org/cgi-> bin/carddisp.pl?gene=CTLA4

George, S., Bell, E., Zheng, Y., Kim, R., White, J., Devgan, G., Smith, J., Lal, L., Engel-Nitz, N., Liu, F. (2021). The impact of adverse events on health care resource utilization, costs, and mortality among patients treated with immune checkpoint inhibitors. *Oncologist*, 26(7), 1205-1215. <https://pubmed.ncbi.nlm.nih.gov/33955118/>

Hayes, A., & Estevez, E. (2021). What is multiple linear regression (MLR)? *Investopedia*. <https://www.investopedia.com/terms/m/mlr.asp>

Hayes, A., Westfall, P., & Beer, K. (2021). T-test. *Investopedia*. <https://www.investopedia.com/terms/t/t-test.asp>

Howe III, E., & Elenberg, F. (2020). Ethical challenges posed by big data. *Innovations in Clinical Neuroscience*, 17(10-12), 24-30. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7819582/>

NCI. (2018). National Cancer Institute overview and mission. *NIH National Cancer Institute*. <https://www.cancer.gov/about-nci/overview>

NCI. (n.d.). T cell. *National Institute of Health- National Cancer Institute*. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/t-cell>

NCI. (n.d.). PD-1. *National Institute of Health- National Cancer Institute*. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pd-1>

Rich, J., Neeley, J., Paneillo, R., Voelker, C., Nussenbaum, B., & Wang, E. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology-Head and Neck Surgery*, 143(3), 331-336. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959/>

Thermo Fisher. (n.d.). TCR cell signaling pathway. *Thermo Fisher Scientific*. <https://www.thermofisher.com/us/en/home/life-science/cell-analysis/signaling-> pathways/t-cell-receptor-tcr/t-cell-receptor-tcr-overview.html

Voutsadakis, I. (2016). Immune blockade inhibition in breast cancer. *Aticancer Research*, 36, 5607-5622. <https://ar.iiarjournals.org/content/anticanres/36/11/5607.full.pdf>

Z\_ai. (2020). Logistic regression explained. *Towards Data Science*. <https://towardsdatascience.com/logistic-regression-explained-9ee73cede081>

Zou, Y., Zou, X., Zheng, S., Tang, H., Zhang, L., Liu, P., & Xie, X. (2020). Efficiency and predictive factors of immune checkpoint inhibitors in metastatic breast cancer: a systematic review and meta-analysis. *Therapeutic Advances in Medical Oncology*, 12(1), 1-17. <https://journals.sagepub.com/doi/pdf/10.1177/1758835920940928>