NUS National University of Singapore | Department of Chemical & Biomolecular Engineering
Faculty of Engineering

Statistical Analysis of Female Breast Cancer and Its Influential Factors

Module: SH5109 (Biostatistics and Epidemiology)

Group Members:

Xu Yun A0186603A

Wang Xinyi A0186672R

Song Beibei A0186806U

Cheng Yuejun A0184792N

CONTENTS

1. Background Information

1.1Introduction of Breast Cancer

Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer.

Breast cancer can begin in different parts of the breast. A breast is made up of three main parts: lobules, ducts, and connective tissue (figure 1.1). The lobules are the glands that produce milk. The ducts are tubes that carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together. Most breast cancers begin in the ducts or lobules.

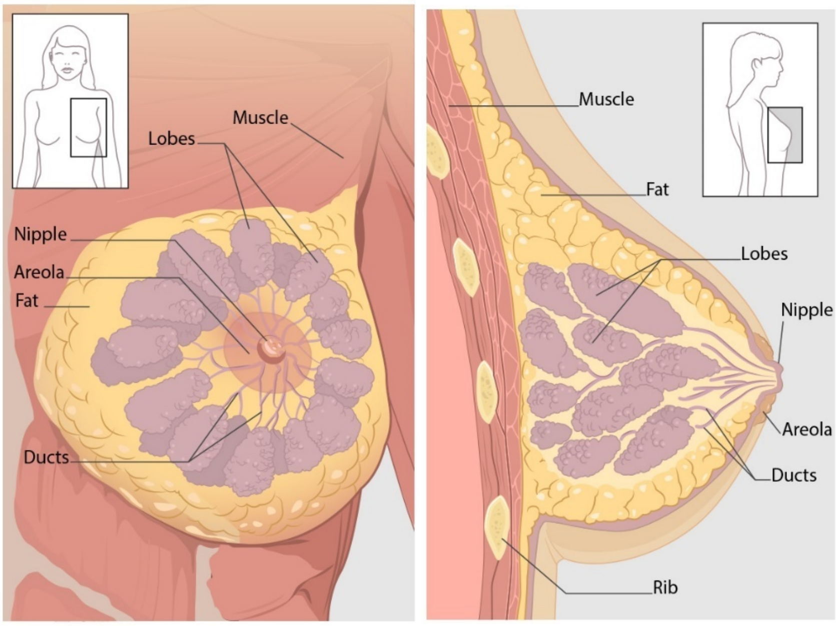


Figure 1.1 Basic Structure of Breast

Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized.

The most common kinds of breast cancer are—

- **Invasive ductal carcinoma**. The cancer cells grow outside the ducts into other parts of the breast tissue. Invasive cancer cells can also spread, or metastasize, to other parts of the body.
- **Invasive lobular carcinoma**. Cancer cells spread from the lobules to the breast tissues that are close by. These invasive cancer cells can also spread to other parts of the body.

Signs and symptoms of breast cancer may include:

- A breast lumps or thickening that feels different from the surrounding tissue
- Change in the size, shape or appearance of a breast
- Changes to the skin over the breast, such as dimpling
- A newly inverted nipple

- Peeling, scaling, crusting or flaking of the pigmented area of skin surrounding the nipple (areola) or breast skin
- Redness or pitting of the skin over your breast, like the skin of an orange

1.2The epidemiology of Breast Cancer

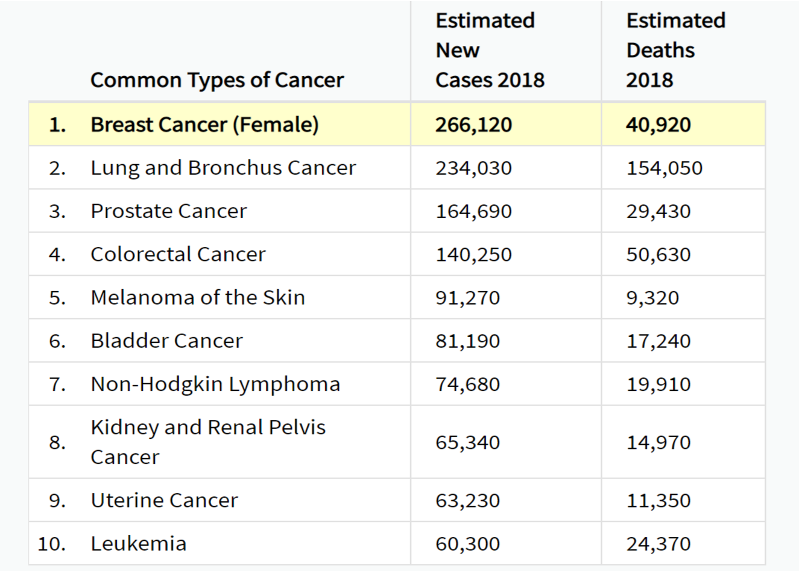
Breast cancer, the most frequently occurring cancer in women, is a major public health problem. It is the most common cancer in American women, except for skin cancers.

Currently, the average risk of a woman in the United States developing breast cancer sometime in her life is about 12%. This means there is a 1 in 8 chance she will develop breast cancer. This also means there is a 7 in 8 chance she will never have the disease. In the meantime, it is estimated that the percentage of survive 5 years or more after being diagnosed with female breast cancer is about 89.7%.

1.3Number of new cases and deaths

1.3.1 How common is breast cancer?

Compared to other cancers, female breast cancer is fairly common (figure 1.2), which represents 15.3% of all new cancer cases in the U.S. In 2018, it is estimated that there will be 266,120 new cases of female breast cancer and an estimated 40,920 people will die of this disease.



	Common Types of Cancer	Estimated New Cases 2018	Estimated Deaths 2018
1.	**Breast Cancer (Female)**	**266,120**	**40,920**
2.	Lung and Bronchus Cancer	234,030	154,050
3.	Prostate Cancer	164,690	29,430
4.	Colorectal Cancer	140,250	50,630
5.	Melanoma of the Skin	91,270	9,320
6.	Bladder Cancer	81,190	17,240
7.	Non-Hodgkin Lymphoma	74,680	19,910
8.	Kidney and Renal Pelvis Cancer	65,340	14,970
9.	Uterine Cancer	63,230	11,350
10.	Leukemia	60,300	24,370

Figure 1.2 Number of new cases and deaths of common type of cancer

1.3.2 Who get breast cancer?

Female breast cancer is most common in middle-aged and older women. Although rare, men can develop breast cancer as well. In our following research, we only considering female cases. The number of new cases of female breast cancer was 126.0 per 100,000 women per year based on 2011-2015 cases.

As for different races, while black women continue to have higher breast cancer death rates than whites nationally, death rates in several states are now statistically equivalent, perhaps reflecting an elimination of disparities in those states.

For different age group, the distribution of new cases occurring different (figure 1.3). Female breast cancer is most frequently diagnosed among women aged 55-64. The median age at diagnosis is 62.

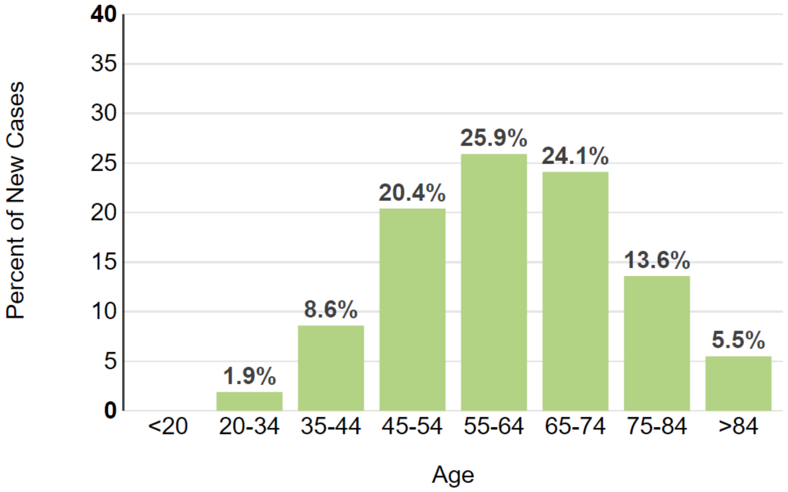


Figure 1.3 Percentage of new cases by age group of female breast cancer

1.3.3 Who dies from breast cancer?

Overall, female breast cancer survival is good. However, women who are diagnosed at an advanced age may be more likely than younger women to die of the disease. Female breast cancer is the fourth leading cause of cancer death in the United States. The number of deaths was 20.9 per 100,000 women per year based on 2011-2015.

For different age group, the percentages have shown below (figure 1.4), we can see that the percent of female breast cancer deaths is highest among women aged 65-74. And the median age at death is 68.

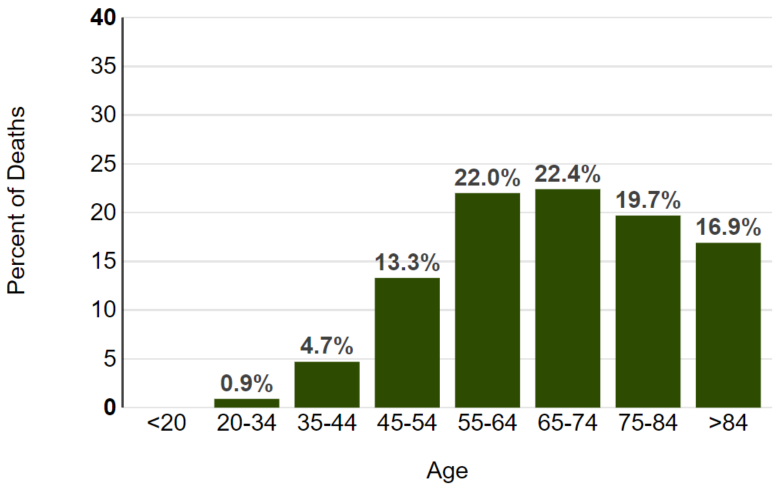


Figure 1.4 Percentage of deaths by age group of female breast cancer

1.4 Trends in rates

Keeping track of the number of new cases, deaths, and survival over time (trends) can help scientists understand whether progress is being made and where additional research is needed to address challenges, such as improving screening or finding better treatments.

Using statistical models for analysis (figure 1.5), rates for new female breast cancer cases have been rising on average 0.3% each year over the last 10 years. Death rates have been falling on average 1.8% each year over 2006-2015. 5-year survival trends are shown below. (figure 1.6).

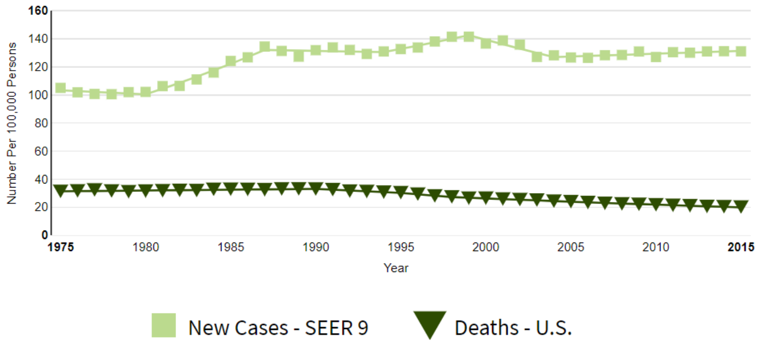


Figure 1.5 New cases and deaths trend

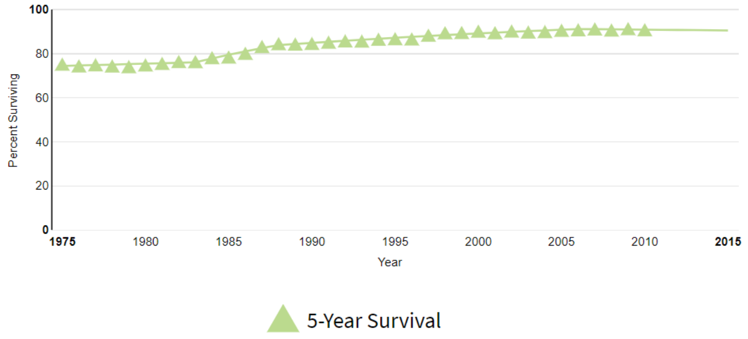


Figure 1.6 5-year survival trend

1.5The influential factors

Doctors know that breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more rapidly than healthy cells do and continue to accumulate, forming a lump or mass. Cells may spread (metastasize) through your breast to your lymph nodes or to other parts of your body. Researchers have identified hormonal, lifestyle and environmental factors that may increase your risk of breast cancer. But it is not clear why some people who have no risk factors develop cancer, yet other people with risk factors never do. It's likely that breast cancer is caused by a complex interaction of your genetic makeup and your environment. In all, a breast cancer risk factor is anything that makes it more likely you will get breast cancer. That comprises of many factors you can change and you cannot change.

For the risk factors you cannot change, like:

- Getting older. The risk of breast cancer increases with age.
- Genetic mutations. Inherited changes (mutations) to certain genes, such as BRCA1 and BRCA2. Women who have inherited these genetic changes are at higher risk of breast and ovarian cancer.
- Reproductive history. Early menstrual periods before age 12 and starting menopause after age 55 expose women to hormones longer, raising their risk of getting breast cancer.
- Personal history of breast cancer or certain non-cancerous breast diseases. Women who have had breast cancer are more likely to get breast cancer a second time. Some non-cancerous breast diseases such as atypical hyperplasia or lobular carcinoma in situ are associated with a higher risk of getting breast cancer.
- Family history of breast cancer. A woman’s risk for breast cancer is higher if she has a mother, sister, or daughter (first-degree relative) or multiple family members on either her mother’s or father’s side of the family who have had breast cancer. Having a first-degree male relative with breast cancer also raises a woman’s risk.
- Previous treatment using radiation therapy. Women who had radiation therapy to the chest or breasts (like for treatment of Hodgkin’s lymphoma) before age 30 have a higher risk of getting breast cancer later in life.

For the risk factors you cannot change, like:

- Not being physically active. Women who are not physically active have a higher risk of getting breast cancer.
- Being overweight or obese after menopause. Older women who are overweight or obese have a higher risk of getting breast cancer than those at a normal weight.
- Taking hormones. Some forms of hormone replacement therapy (those that include both estrogen and progesterone) taken during menopause can raise risk for breast

cancer when taken for more than five years. Certain oral contraceptives (birth control pills) also have been found to raise breast cancer risk.

- Reproductive history. Having the first pregnancy after age 30, not breastfeeding, and never having a full-term pregnancy can raise breast cancer risk.
- Drinking alcohol. Studies show that a woman's risk for breast cancer increases with the more alcohol she drinks.

But having one or even several breast cancer risk factors does not necessarily mean you will develop breast cancer. Many women who develop breast cancer have no known risk factors other than simply being women.

Early detection of this disease with screening and good standardized treatment has improved the prognoses of afflicted patients. However, approximately one-third of patients will develop distant metastases to the liver, bones, lungs, and brain. They eventually succumb to the disease.

2. Objects and Methods

2.1Objects

In this research, we are going to explore the relationships between final status of patients or time to death and several potential influential factors of breast cancer using data analysis. We choose **230** clinical breast cancer samples from **eBioPortal** that is a famous tumor database. The factors we choose are briefly explained below:

- **Invasive carcinoma diagnosis age.** The age of patient to be diagnosed with breast invasive ductal carcinoma.
- **Disease free (month).** In cancer, the length of time after primary treatment for a cancer ends that the patient survives without any signs or symptoms of that cancer. In a clinical trial, measuring the disease-free survival is one way to see how well a new treatment works.
- **HER2 primary status.** The status of the gene encoding human EGF-like receptor 2 (HER2) is an important prognostic and predictive marker in breast cancer. Only breast cancers with HER2 amplification respond to the targeted therapy with trastuzumab.
- **ER status of primary.** Estrogen receptor (ER) is an important biological marker for making decisions about breast cancer treatments. Estrogen-dependent growth of breast cancer can be blocked by anti-estrogens. Estrogen receptor (ER) presence in breast cancer implies responsiveness to endocrine therapy.
- **Menopausal status at diagnosis.** Whether the patients are in the menopausal status when they are diagnosed of breast cancer. Self-rated menopausal status appears to relate more closely to a women's endocrine status
- **Overall survival (month).** Overall survival time since the first diagnosis.
- **Overall survival status.** Overall patients' survival status, using 1 represent living status and 0 for deceased status.

- **Overall patient receptor status.** Overall patient tumor receptor subtype based on the review of the primary and metastatic tumors and the clinical history.
- **Overall patient HR status.** Overall patient tumor hormone receptor status based on the review of the primary and metastatic tumors and the clinical history
- **PR status of primary.** Progesterone receptor (PR) status of primary tumor. Progesterone is important in breast tumorigenesis, and its effects on the breast are mediated by t the progesterone receptor.
- **Sample type.** The type of sample (i.e., normal, primary, metastasis, recurrence)
- **Time to death (month).** Overall survival months of the patient.

2.2Methods

In our project, we use STATA to do t-test, chi quare test, two way ANOVA, logistic regression, Multinomial Logistic Regression.

3. Statistical Analysis

3.1Description of the data

This data set has a binary response outcome (dependent variable) called overall living status (living/deceased).

Variables:

1)Disease Free (Months)——continuous

2)ER Status of the Primary

3)HER2 Primary Status

4)Invasive Carcinoma Diagnosis Age——continuous

5)Menopausal Status at Diagnosis

6)Metastatic Recurrence Time——continuous

7)Mutation Count——continuous

8)Overall Survival (Months) ——continuous

9)Overall Survival Status

10)Overall Patient HR Status

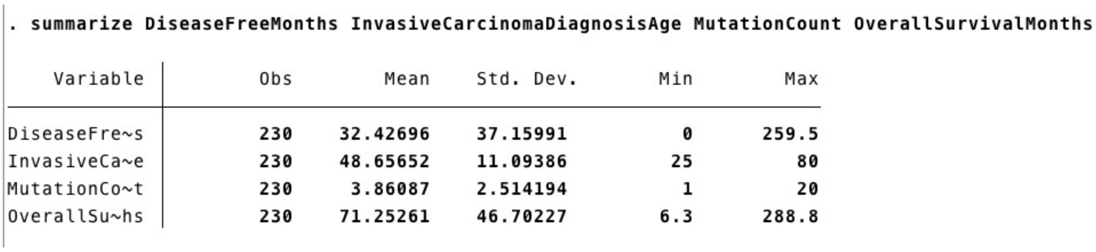
11)Overall Patient Receptor Status

12)PR Status of the Primary

13)Sample Type

14)Time To Death (Months) ——continuous

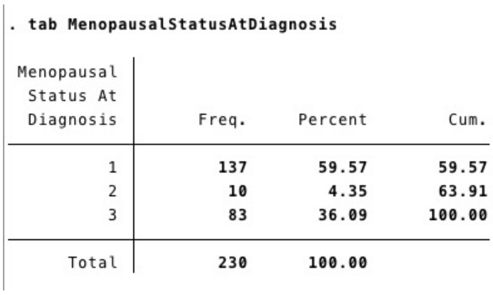
The variable ER status of the primary, HER2 Primary Status, overall patient HR status, PR Status of the Primary take on the values “1”/ “0”. Institutions with a status of “0” is negative, while those with a status of “1” is positive. The variable overall survival status, a status of “1” is living, while a status of “0” is deceased. As for sample type, “1” presents primary cancer, while “0” presents metastasis cancer.



```
. summarize DiseaseFreeMonths InvasiveCarcinomaDiagnosisAge MutationCount OverallSurvivalMonths

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
DiseaseFre~s |        230    32.42696    37.15991          0      259.5
InvasiveCa~e |        230    48.65652    11.09386         25         80
MutationCo~t |        230     3.86087    2.514194          1         20
OverallSu~hs |        230    71.25261    46.70227        6.3      288.8
```

Figure 3.1: summarize of data

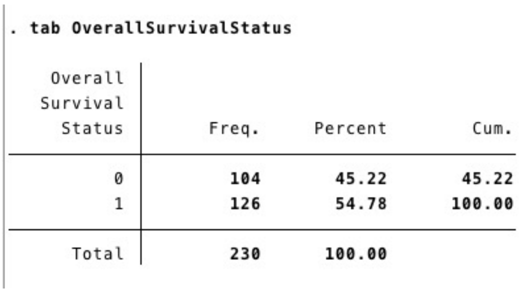
The total observation is 230 samples, we can get the mean, standards deviation, minimal& maximal values for each variable. The longest disease free time is 259.5 months, who was recovered from breast cancer almost 22 years. The youngest patient who was diagnosed breast cancer is only 25 years old, while the oldest patient is 80 years old. The average age is around 48 years old, when is around menopausal period. We also analyze “Menopausal Status At Diagnosis”， to see whether it’s pre-menopausal, peri-menopausal , or post-menopausal. The mutation counts range from 1 to 20. The worst case of our sample is only live 6.3 months, the luckiest patient lives over 24 years after diagnosed. The average surviving time is almost 6 years.



```
. tab MenopausalStatusAtDiagnosis

 Menopausal |
  Status At |
  Diagnosis |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        137       59.57       59.57
          2 |         10        4.35       63.91
          3 |         83       36.09      100.00
------------+-----------------------------------
      Total |        230      100.00
```

Figure 3.2: tab menopausal status

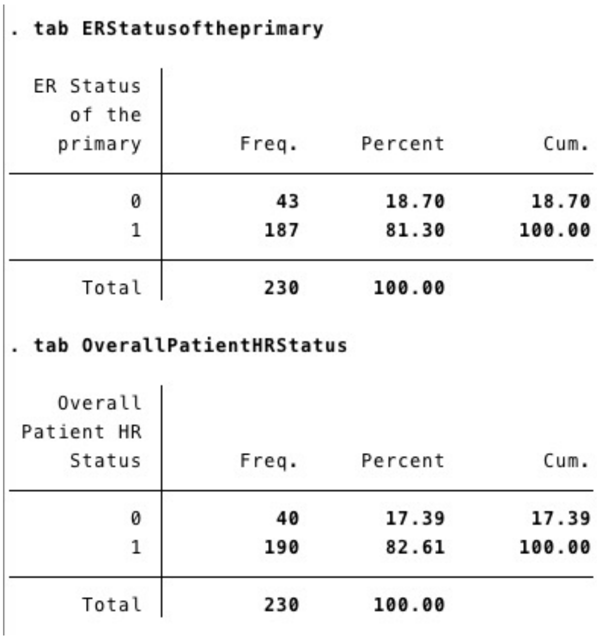
We set up three menopausal status at diagnosis as 1, 2, 3, respectively represents pre-menopausal, peri-menopausal, or post-menopausal. The data shows that 10 samples diagnose during menopausal, about 60% patients diagnose pre-menopausal, the rest (83 samples) is post-menopausal.



```
. tab OverallSurvivalStatus

    Overall |
   Survival |
     Status |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        104       45.22       45.22
          1 |        126       54.78      100.00
------------+-----------------------------------
      Total |        230      100.00
```

Figure 3.3: tab overall survival status

In our study, we are interested in the overall survival status. In the 230 samples, there are 126 patients still living accounting for 54.78%, and 126 cases deceased (45.22%).



```
. tab ERStatusoftheprimary

  ER Status |
     of the |
    primary |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         43       18.70       18.70
          1 |        187       81.30      100.00
------------+-----------------------------------
      Total |        230      100.00

. tab OverallPatientHRStatus

    Overall |
 Patient HR |
     Status |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         40       17.39       17.39
          1 |        190       82.61      100.00
------------+-----------------------------------
      Total |        230      100.00
```

Figure 3.4: tab ER/HER status

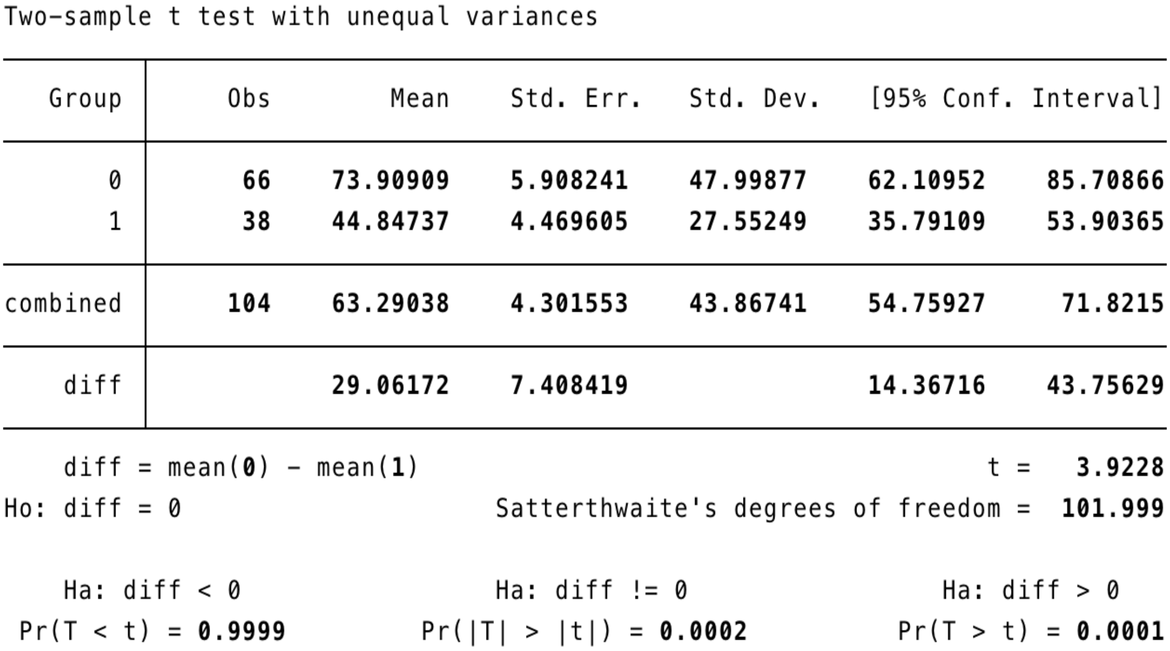
ER status of the primary and overall patient HR status are indicator variables. We can get the frequency and percent of total samples.

3.2T-test (two sample using group)

Mammary gland is not an important organ to maintain human life. But free cancer cells can spread throughout the body with blood or lymph, forming metastases that can be life-threatening

Here we used t-test (two sample using group) to compare the survival time of patients （already dead） with primary carcinoma and metastatic carcinoma.

Group: 0--metastatic; 1 – primary.



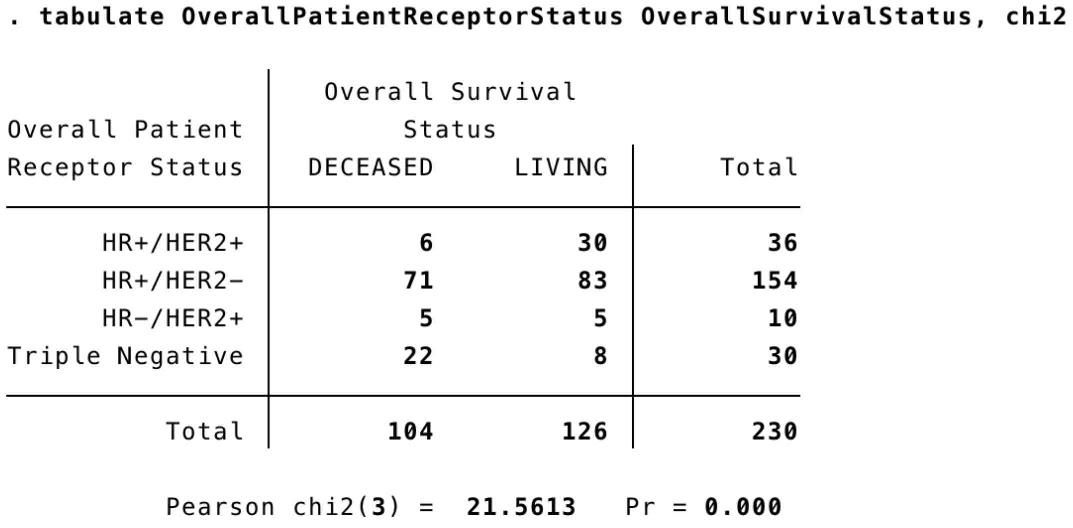
```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      66    73.90909    5.908241    47.99877    62.10952    85.70866
       1 |      38    44.84737    4.469605    27.55249    35.79109    53.90365
---------+--------------------------------------------------------------------
combined |     104    63.29038    4.301553    43.86741    54.75927     71.8215
---------+--------------------------------------------------------------------
    diff |            29.06172    7.408419                14.36716    43.75629
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   3.9228
Ho: diff = 0                     Satterthwaite's degrees of freedom =  101.999

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9999         Pr(|T| > |t|) = 0.0002          Pr(T > t) = 0.0001
```

Figure 3.5: t-test outcome

From the results, the P-value is 0.0002, smaller than 0.05. The number showed that the survival time between primary carcinoma and metastatic carcinoma has a significant difference.

3.3Chi square test (contingency table)

In the field of medicine, the survival status of patient is closely related to receptor status(HR/HER2). In this test, our purpose is to determine the correlation between overall patient receptor status(HR/HER2) and survival status(deceased/living).



```
. tabulate OverallPatientReceptorStatus OverallSurvivalStatus, chi2

                   |   Overall Survival
    Overall Patient|        Status
   Receptor Status |  DECEASED     LIVING |     Total
-------------------+----------------------+----------
         HR+/HER2+ |         6         30 |        36
         HR+/HER2- |        71         83 |       154
         HR-/HER2+ |         5          5 |        10
   Triple Negative |        22          8 |        30
-------------------+----------------------+----------
             Total |       104        126 |       230

          Pearson chi2(3) =  21.5613   Pr = 0.000
```

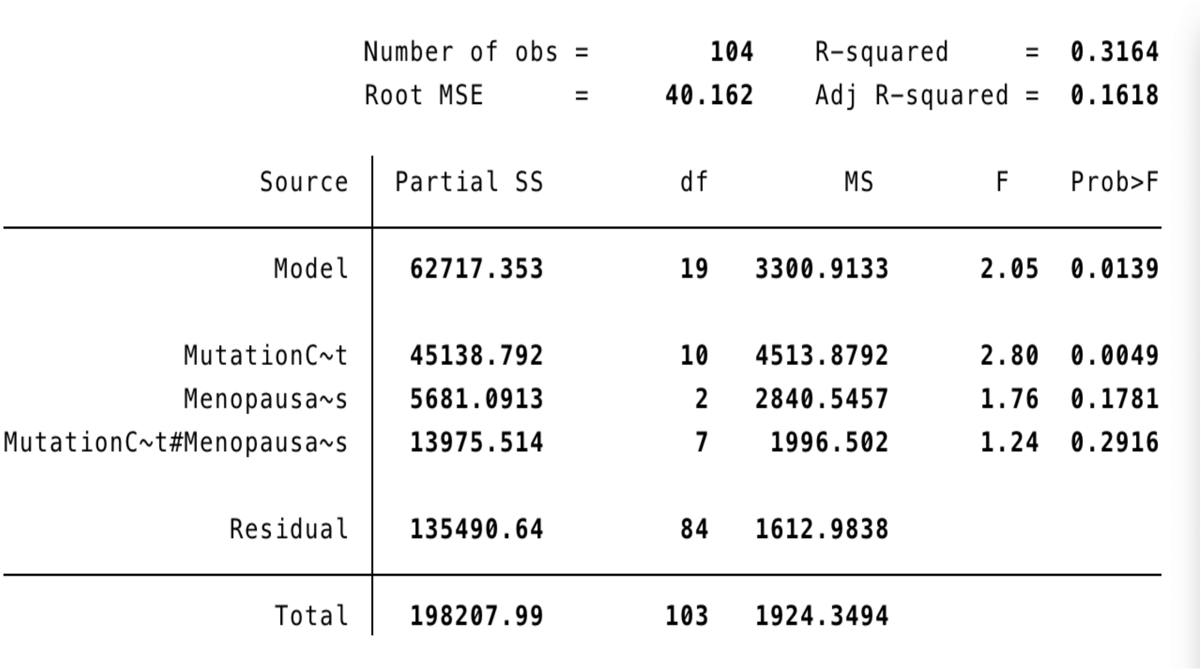
Figure 3.6: chi square test outcome

From the table above, we can see the P-value is smaller than 0.05, which means there is a strong correlation between overall patient receptor status(HR/HER2) and survival status(deceased/living). In other words, patient receptor status is an important impact factor for survival status of patient.

3.4Two-way ANOVA

We use two-way ANOVA to study the relationship between survival time（already dead）and mutation count, menopause status and their interaction.

The results are shown in the following table, the significance value is 0.0002, smaller than 0.05, which means there is a correlation between survival time of patient and mutation count, or the mutation count is an impact factor of survival time.



```
                         Number of obs =        104    R-squared     =  0.3164
                         Root MSE      =     40.162    Adj R-squared =  0.1618

                  Source | Partial SS         df         MS        F    Prob>F
  -----------------------+----------------------------------------------------
                   Model |  62717.353         19   3300.9133      2.05  0.0139
                         |
              MutationC~t|  45138.792         10   4513.8792      2.80  0.0049
              Menopausa~s|  5681.0913          2   2840.5457      1.76  0.1781
 MutationC~t#Menopausa~s |  13975.514          7    1996.502      1.24  0.2916
                         |
                Residual |  135490.64         84   1612.9838
  -----------------------+----------------------------------------------------
                   Total |  198207.99        103   1924.3494
```

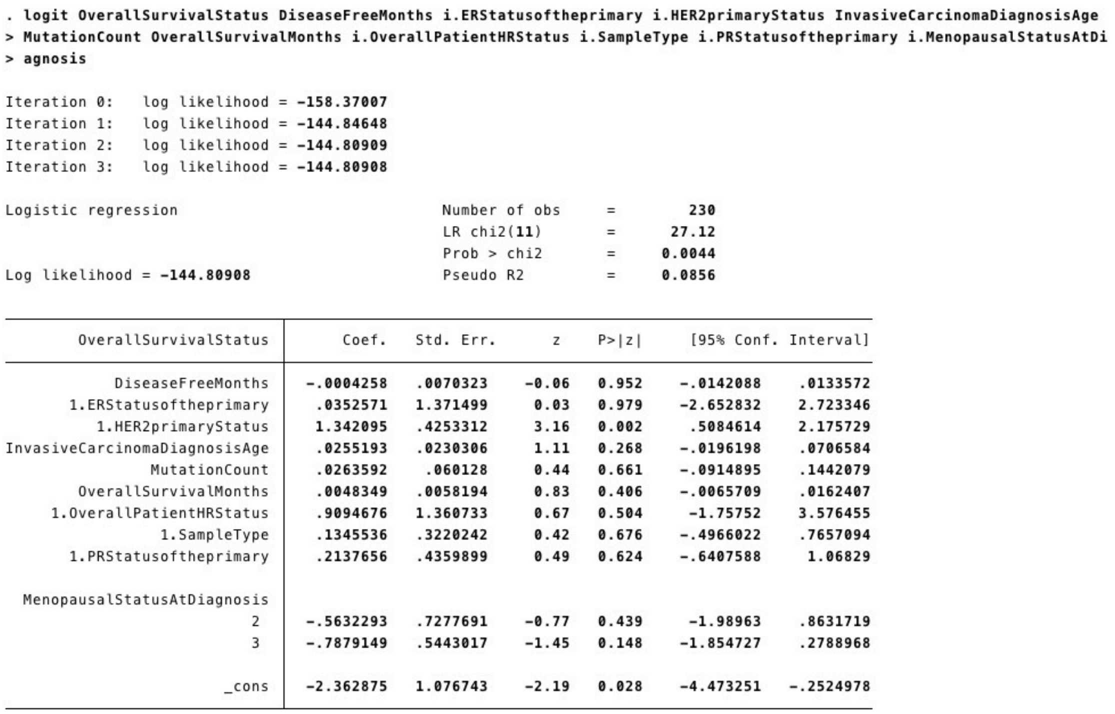
Figure 3.7: two-way ANOVA outcome

3.5Logistic Regression:

Logistic regression, also called a logit model, is used to model dichotomous outcome variables. In the logit model the log odds of the outcome is modeled as a linear combination of the predictor variables. At the center of the logistic regression analysis is the task estimating the log odds of an event. Mathematically, logistic regression estimates a multiple linear regression function defined as:

$$\log\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_n * x_n$$

We are interested in how variables, such as disease free (months), ER status of the primary, time to diagnosis invasive carcinoma, metastatic recurrence time, mutation count, overall patient HR status, effect the overall living status. The response variable, living or deceased, is a binary variable.

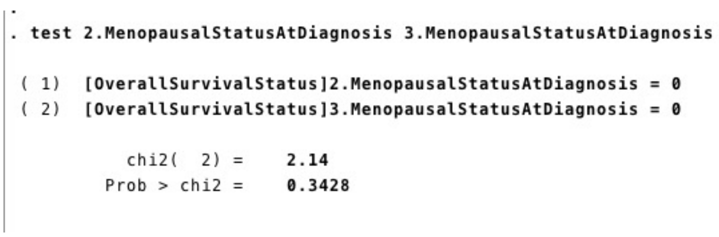


```
. logit OverallSurvivalStatus DiseaseFreeMonths i.ERStatusoftheprimary i.HER2primaryStatus InvasiveCarcinomaDiagnosisAge
> MutationCount OverallSurvivalMonths i.OverallPatientHRStatus i.SampleType i.PRStatusoftheprimary i.MenopausalStatusAtDi
> agnosis

Iteration 0:   log likelihood = -158.37007
Iteration 1:   log likelihood = -144.84648
Iteration 2:   log likelihood = -144.80909
Iteration 3:   log likelihood = -144.80908

Logistic regression                             Number of obs     =        230
                                                LR chi2(11)       =      27.12
                                                Prob > chi2       =     0.0044
Log likelihood = -144.80908                     Pseudo R2         =     0.0856

------------------------------------------------------------------------------------------------
       OverallSurvivalStatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------------+------------------------------------------------------------------
           DiseaseFreeMonths |  -.0004258   .0070323    -0.06   0.952    -.0142088    .0133572
      1.ERStatusoftheprimary |   .0352571   1.371499     0.03   0.979    -2.652832    2.723346
         1.HER2primaryStatus |   1.342095   .4253312     3.16   0.002     .5084614    2.175729
InvasiveCarcinomaDiagnosisAge |   .0255193   .0230306     1.11   0.268    -.0196198    .0706584
               MutationCount |   .0263592    .060128     0.44   0.661    -.0914895    .1442079
       OverallSurvivalMonths |   .0048349   .0058194     0.83   0.406    -.0065709    .0162407
    1.OverallPatientHRStatus |   .9094676   1.360733     0.67   0.504     -1.75752    3.576455
                1.SampleType |   .1345536   .3220242     0.42   0.676    -.4966022    .7657094
      1.PRStatusoftheprimary |   .2137656   .4359899     0.49   0.624    -.6407588     1.06829
                             |
  MenopausalStatusAtDiagnosis |
                          2  |  -.5632293   .7277691    -0.77   0.439     -1.98963    .8631719
                          3  |  -.7879149   .5443017    -1.45   0.148    -1.854727    .2788968
                             |
                       _cons |  -2.362875   1.076743    -2.19   0.028    -4.473251   -.2524978
------------------------------------------------------------------------------------------------
```

Figure 3.8: logistic regression

- In the output above, we see the iteration log, indicating how quickly the model converged. The log likelihood (-158.37007) can be use in comparisons of nested models.
- Also, at the top of the output we see that all 230 observations in our data set were used in the analysis.
- The likelihood ratio chi-square of 27.12 with a p-value of 0.0044(<0.05) tells us that our models as a whole fit significantly.
- In the table we see the coefficients, their standard errors, the z-statistic, associated p-values, and the 95% confidence interval of the coefficients. Only variable HER2 primary status statistically significant, as are two indicator variables for menopausal status at diagnosis. The logistic regression coefficient give the change in the log odds of the outcome for a one unit increase in the predictor variable.
  - For every one unit change in disease free time. The log odds of living (versus deceased) decrease by 0.004.
  - For a one unit increase in mutation counts, the log odds of living increase by 0.0263592.
  - The indicator variables for menopausal status at diagnosis have a slightly different interpretation. Still living with peri-menopausal, versus pre-menopausal, decreases the log odds of living by 0.563.

  We can test for an overall effect of menopausal status at diagnosis using the test command. Below we see that the overall effect of menopausal status at diagnosis is not statistically significant.



```
.
. test 2.MenopausalStatusAtDiagnosis 3.MenopausalStatusAtDiagnosis

 ( 1)  [OverallSurvivalStatus]2.MenopausalStatusAtDiagnosis = 0
 ( 2)  [OverallSurvivalStatus]3.MenopausalStatusAtDiagnosis = 0

           chi2(  2) =    2.14
         Prob > chi2 =    0.3428
```

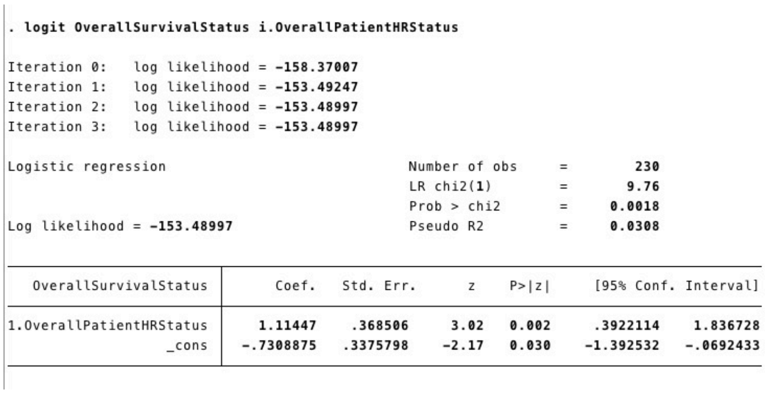
Figure 3.9:test menopausal status

We also can see the column “Overall Patient Receptor Status”, including HR status and HER2 status.

Firstly, we can focus on overall patient HR status.

$$\log\left(\frac{p(y=1)}{1-p(y=1)}\right) = 1.114 * x_1 - 0.731$$

The p-value of overall patient HR status is 0.002(<5%), which is significant to

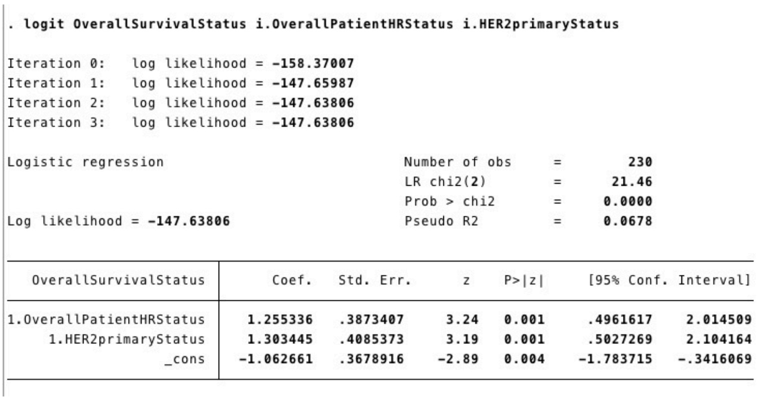


```
. logit OverallSurvivalStatus i.OverallPatientHRStatus

Iteration 0:   log likelihood = -158.37007
Iteration 1:   log likelihood = -153.49247
Iteration 2:   log likelihood = -153.48997
Iteration 3:   log likelihood = -153.48997

Logistic regression                             Number of obs     =        230
                                                LR chi2(1)        =       9.76
                                                Prob > chi2       =     0.0018
Log likelihood = -153.48997                     Pseudo R2         =     0.0308

------------------------------------------------------------------------------------------
   OverallSurvivalStatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------+----------------------------------------------------------------
1.OverallPatientHRStatus |    1.11447     .368506     3.02   0.002     .3922114    1.836728
                   _cons |  -.7308875    .3375798    -2.17   0.030    -1.392532   -.0692433
------------------------------------------------------------------------------------------
```

Figure 3.10: logit overall survival status

Considering of both overall patient HR status and HER2 primary status.

$$\log\left(\frac{p(y=1)}{1-p(y=1)}\right) = 1.255 * x_1 + 1.303 * x_2 - 1.062$$



```
. logit OverallSurvivalStatus i.OverallPatientHRStatus i.HER2primaryStatus

Iteration 0:   log likelihood = -158.37007
Iteration 1:   log likelihood = -147.65987
Iteration 2:   log likelihood = -147.63806
Iteration 3:   log likelihood = -147.63806

Logistic regression                             Number of obs     =        230
                                                LR chi2(2)        =      21.46
                                                Prob > chi2       =     0.0000
Log likelihood = -147.63806                     Pseudo R2         =     0.0678

------------------------------------------------------------------------------------------
   OverallSurvivalStatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------+----------------------------------------------------------------
1.OverallPatientHRStatus |   1.255336    .3873407     3.24   0.001     .4961617    2.014509
     1.HER2primaryStatus |   1.303445    .4085373     3.19   0.001     .5027269    2.104164
                   _cons |  -1.062661    .3678916    -2.89   0.004    -1.783715   -.3416069
------------------------------------------------------------------------------------------
```

Figure 3.11: add on HER2 status

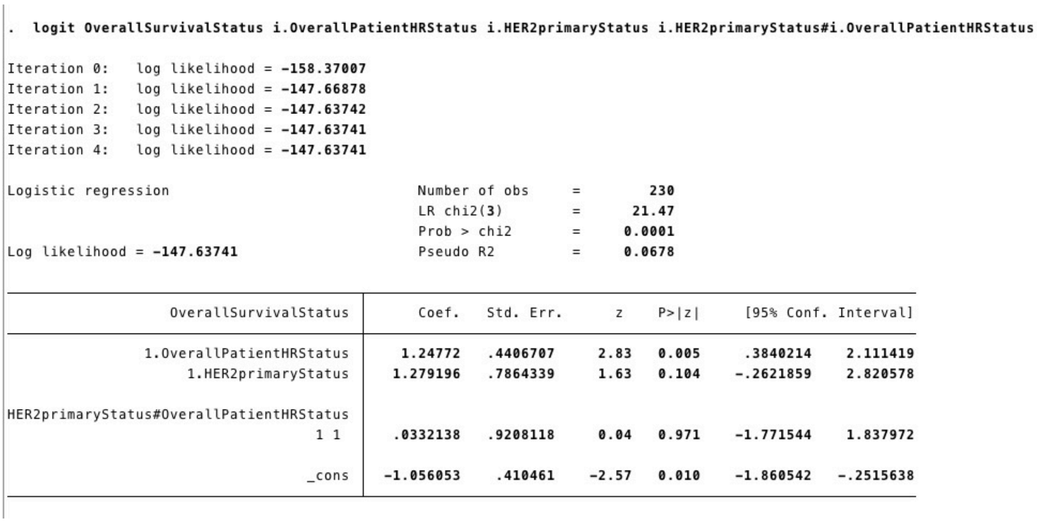
In the presence of interaction term of overall patient HR status by HER2 primary status, we can no longer talk about the effect of overall patient HR status, holding all other variables at certain value, since it does not make sense to fix overall patient HR status and HR status* HER2 primary status at certain value and still allow HR status change from 0 to 1

$$\log\left(\frac{p(y=1)}{1-p(y=1)}\right) = 1.247 * x_1 + 1.279 * x_2 + 0.033 * x_1x_2 - 1.056$$

x_1 = overall patient HR status

x_2 = HER2 primary status

x_1x_2 = the interaction of overall patient HR status and HER2 primary status



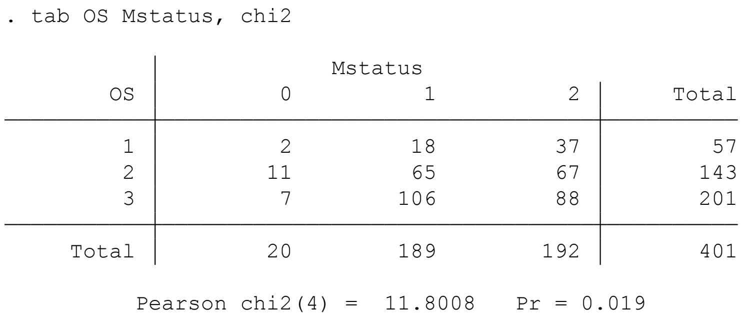
```
.  logit OverallSurvivalStatus i.OverallPatientHRStatus i.HER2primaryStatus i.HER2primaryStatus#i.OverallPatientHRStatus

Iteration 0:   log likelihood = -158.37007
Iteration 1:   log likelihood = -147.66878
Iteration 2:   log likelihood = -147.63742
Iteration 3:   log likelihood = -147.63741
Iteration 4:   log likelihood = -147.63741

Logistic regression                             Number of obs     =        230
                                                LR chi2(3)        =      21.47
                                                Prob > chi2       =     0.0001
Log likelihood = -147.63741                     Pseudo R2         =     0.0678

------------------------------------------------------------------------------------------------------
                  OverallSurvivalStatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------------------------------+-------------------------------------------------------------
               1.OverallPatientHRStatus |    1.24772   .4406707     2.83   0.005     .3840214    2.111419
                    1.HER2primaryStatus |   1.279196   .7864339     1.63   0.104    -.2621859    2.820578
                                        |
HER2primaryStatus#OverallPatientHRStatus |
                                    1 1 |   .0332138   .9208118     0.04   0.971    -1.771544    1.837972
                                        |
                                  _cons |  -1.056053    .410461    -2.57   0.010    -1.860542   -.2515638
------------------------------------------------------------------------------------------------------
```

Figure 3.12: add on interaction

3.6Multinomial Logistic Regression

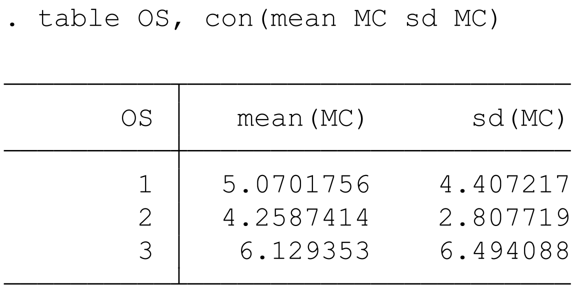
In cancer research, the survival time is an important indicator to evaluate the treatment effectiveness. For example, five-year absolute survival rates describe the percentage of patients alive five years after the disease is diagnosed, normally calculated from the point of diagnosis. If a patient is generally cured and survive for up to 5 years after the radical operation, typically, this means that he has a better chance to survive in disease free condition for a longer time. For this reason, we do not use the continuous survival months as your dependent values, instead, the survival months are divided into 3 groups: 1. The short survival group, the survival time of which are less than 24 months. 2. The medium survival group, the survival time of which are from 24 months to 60 months (5 years). 3, The long survival group, the survival months of which are more than 60.

To analysis the association between survival time and other factors (in this part we choose the “Menopausal Status At Diagnosis” and “Mutation Count”, for easier description purpose, in the following article, we will use “**Mstatus**” to represent “Menopausal Status At Diagnosis”, use “**MC**” to represent “Mutation Count”, and use “**OS**” to represent “Overall survival time”), the size of our data source is expanded to 401 samples. After expanding, the descriptive statistic for overall survival time, Mstatus and MC are shown as below using a simple data description command. From the below table you can also read the mean and standard deviation of MC for different OS groups. From this preliminary analysis, no obvious differences are showed in the means of MC for OS group 1 to 3, which means more advanced analysis tool is needed.



```
. tab OS Mstatus, chi2

           |              Mstatus
        OS |         0          1          2 |     Total
-----------+---------------------------------+----------
         1 |         2         18         37 |        57 
         2 |        11         65         67 |       143 
         3 |         7        106         88 |       201 
-----------+---------------------------------+----------
     Total |        20        189        192 |       401 

          Pearson chi2(4) =  11.8008   Pr = 0.019
```

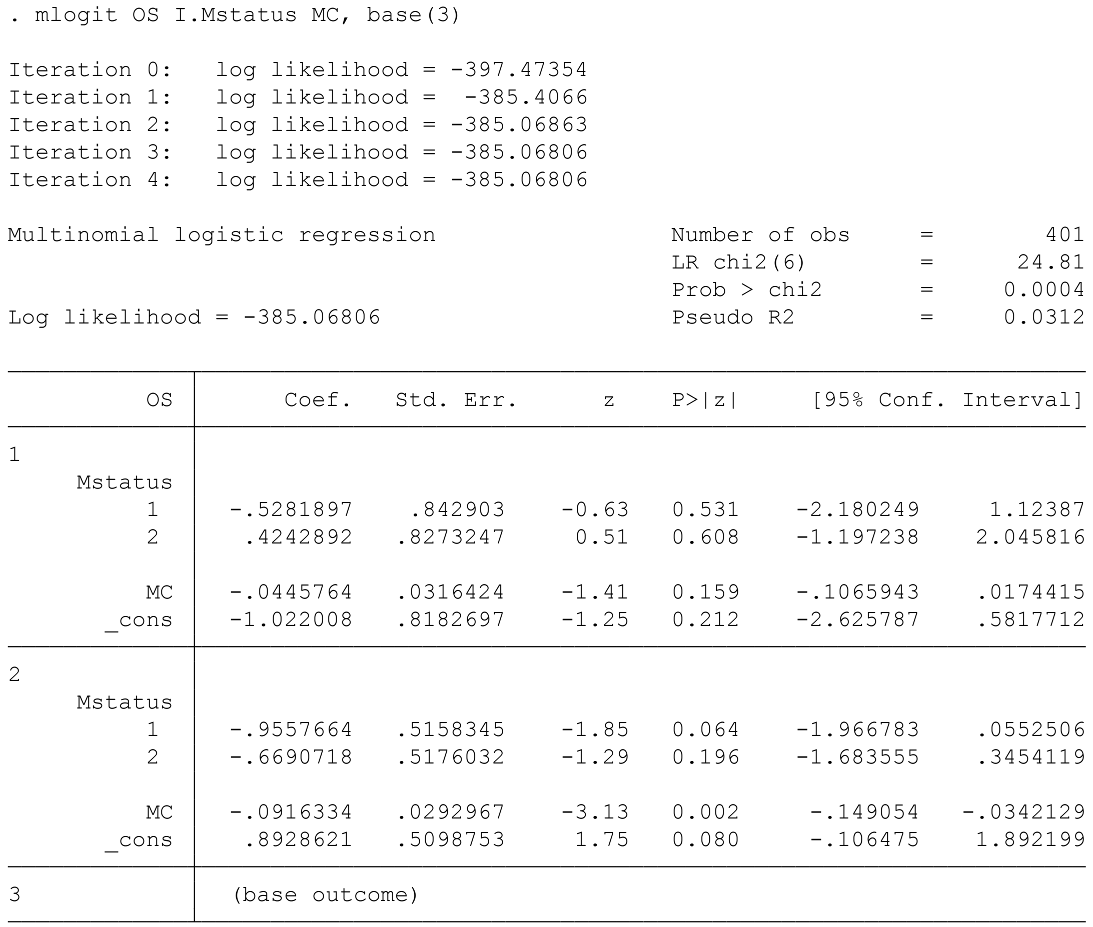
Figure 3.13: Statistic description 1



```
. table OS, con(mean MC sd MC)

----------------------------------
       OS |    mean(MC)      sd(MC)
----------+-----------------------
        1 |   5.0701756    4.407217
        2 |   4.2587414    2.807719
        3 |    6.129353    6.494088
----------------------------------
```

Figure 3.14: Data description 2

Below we use the **mlogit command** to estimate a multinomial logistic regression model. The i. before Mstatus indicates that Mstatus is an indicator variable (i.e., categorical variable), and that it should be included in the model. We have also used the option “base” to indicate the category we would want to use for the baseline comparison group. In the model below, we have chosen to use the “long survival group” as the baseline category.



```
. mlogit OS I.Mstatus MC, base(3)

Iteration 0:   log likelihood = -397.47354
Iteration 1:   log likelihood =  -385.4066
Iteration 2:   log likelihood = -385.06863
Iteration 3:   log likelihood = -385.06806
Iteration 4:   log likelihood = -385.06806

Multinomial logistic regression                 Number of obs     =        401
                                                LR chi2(6)        =      24.81
                                                Prob > chi2       =     0.0004
Log likelihood = -385.06806                     Pseudo R2         =     0.0312

------------------------------------------------------------------------------
          OS |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
     Mstatus |
          1  |  -.5281897    .842903    -0.63   0.531    -2.180249     1.12387
          2  |   .4242892   .8273247     0.51   0.608    -1.197238    2.045816
             |
          MC |  -.0445764   .0316424    -1.41   0.159    -.1065943    .0174415
       _cons |  -1.022008   .8182697    -1.25   0.212    -2.625787    .5817712
-------------+----------------------------------------------------------------
2            |
     Mstatus |
          1  |  -.9557664   .5158345    -1.85   0.064    -1.966783    .0552506
          2  |  -.6690718   .5176032    -1.29   0.196    -1.683555    .3454119
             |
          MC |  -.0916334   .0292967    -3.13   0.002     -.149054   -.0342129
       _cons |   .8928621   .5098753     1.75   0.080     -.106475    1.892199
-------------+----------------------------------------------------------------
3            |  (base outcome)
------------------------------------------------------------------------------
```

Figure 3.15: Mlogit command outcome

The p value tells us that our model as a whole fit significantly better than an empty model which means a model with no predictors, because the P value shown here is very small: 0.0004.

They correspond to the two equations below:

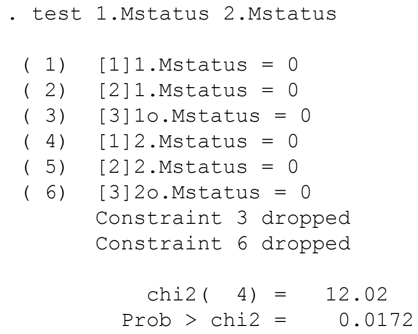
$$ln(\frac{p(OS=1)}{p(OS=3)}) = b10 + b11(Mstatus = 1) + b12(Mstatus = 2) + b13 * MC$$

$$ln(\frac{p(OS=2)}{p(OS=3)}) = b20 + b21(Mstatus = 1) + b22(Mstatus = 2) + b23 * MC$$

Where b is the regression coefficients.

This means that a one unit increase in the variable MC (Mutation Count) is associated with a 0.44 decrease in the relative log odds of being in the short survival group versus being in the long survival group. A one unit increase in the variable MC (Mutation Count) is associated with a 0.91 decrease in the relative log odds of being in the medium survival group versus being in the long survival group.

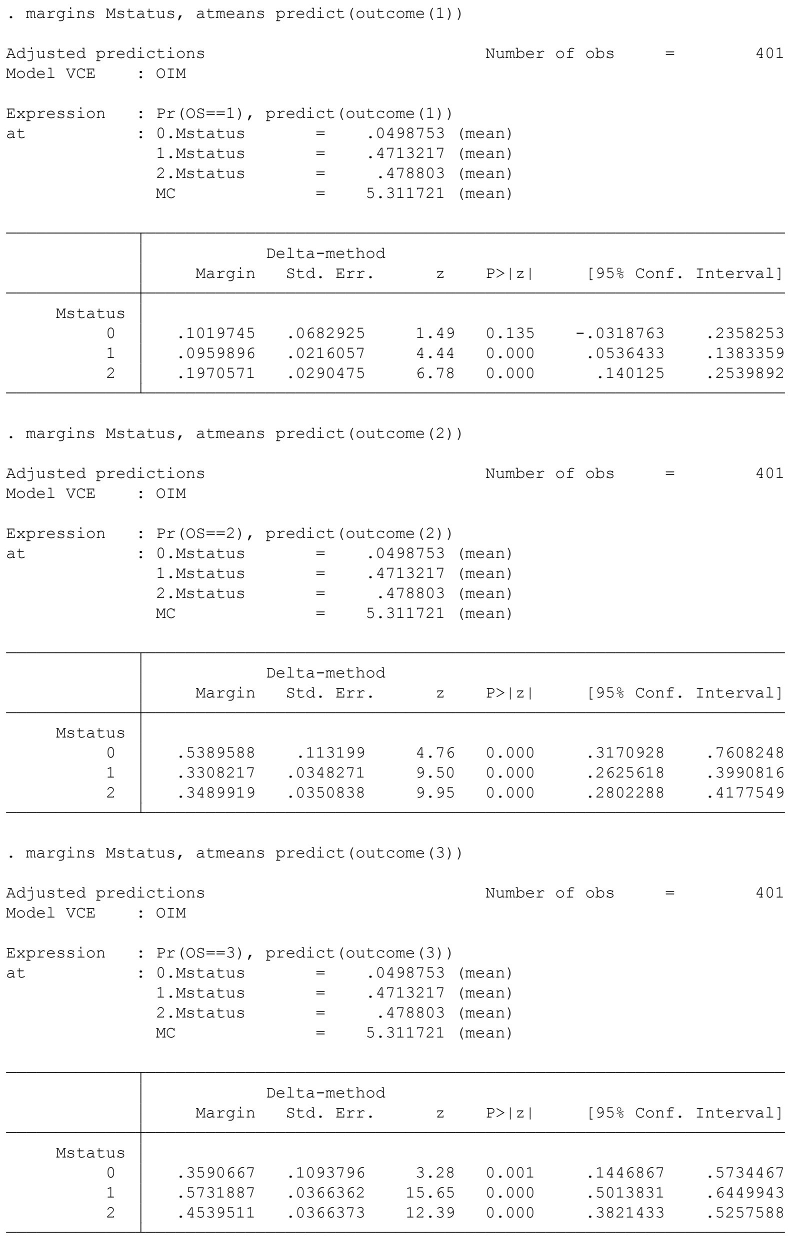
We can also test for an overall effect of Mstatus using test command. We can see from below results that the Mstatus is statistically significant to the OS as the P value here is 0.0172.



```
. test 1.Mstatus 2.Mstatus

 ( 1)  [1]1.Mstatus = 0
 ( 2)  [2]1.Mstatus = 0
 ( 3)  [3]1o.Mstatus = 0
 ( 4)  [1]2.Mstatus = 0
 ( 5)  [2]2.Mstatus = 0
 ( 6)  [3]2o.Mstatus = 0
       Constraint 3 dropped
       Constraint 6 dropped

           chi2(  4) =   12.02
         Prob > chi2 =    0.0172
```

Figure 3.16: overall effect of Mstatus

We can also use predicted probabilities to help you understand the model. You can calculate predicted probabilities using the margins command. Below we use the margins command to calculate the predicted probabilities of being in each OS group at each level of Mstatus, holding all other variables in the model at their means. Since there are 3 possible outcomes, we will need to use the margins command 3 times, one for each outcome value.



```
. margins Mstatus, atmeans predict(outcome(1))

Adjusted predictions                            Number of obs     =        401
Model VCE    : OIM

Expression   : Pr(OS==1), predict(outcome(1))
at           : 0.Mstatus       =    .0498753 (mean)
               1.Mstatus       =    .4713217 (mean)
               2.Mstatus       =     .478803 (mean)
               MC              =    5.311721 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Mstatus |
          0  |   .1019745   .0682925     1.49   0.135    -.0318763    .2358253
          1  |   .0959896   .0216057     4.44   0.000     .0536433    .1383359
          2  |   .1970571   .0290475     6.78   0.000      .140125    .2539892
------------------------------------------------------------------------------

. margins Mstatus, atmeans predict(outcome(2))

Adjusted predictions                            Number of obs     =        401
Model VCE    : OIM

Expression   : Pr(OS==2), predict(outcome(2))
at           : 0.Mstatus       =    .0498753 (mean)
               1.Mstatus       =    .4713217 (mean)
               2.Mstatus       =     .478803 (mean)
               MC              =    5.311721 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Mstatus |
          0  |   .5389588    .113199     4.76   0.000     .3170928    .7608248
          1  |   .3308217   .0348271     9.50   0.000     .2625618    .3990816
          2  |   .3489919   .0350838     9.95   0.000     .2802288    .4177549
------------------------------------------------------------------------------

. margins Mstatus, atmeans predict(outcome(3))

Adjusted predictions                            Number of obs     =        401
Model VCE    : OIM

Expression   : Pr(OS==3), predict(outcome(3))
at           : 0.Mstatus       =    .0498753 (mean)
               1.Mstatus       =    .4713217 (mean)
               2.Mstatus       =     .478803 (mean)
               MC              =    5.311721 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Mstatus |
          0  |   .3590667   .1093796     3.28   0.001     .1446867    .5734467
          1  |   .5731887   .0366362    15.65   0.000     .5013831    .6449943
          2  |   .4539511   .0366373    12.39   0.000     .3821433    .5257588
------------------------------------------------------------------------------
```

Figure 3.17: Margin command outcome of the predicted probabilities

We can use **marginsplot command** to plot predicted probabilities by Menopausal status for each category of survival length. We also use the combine command to combine the 3 marginsplots into one graph to facilitate comparison.

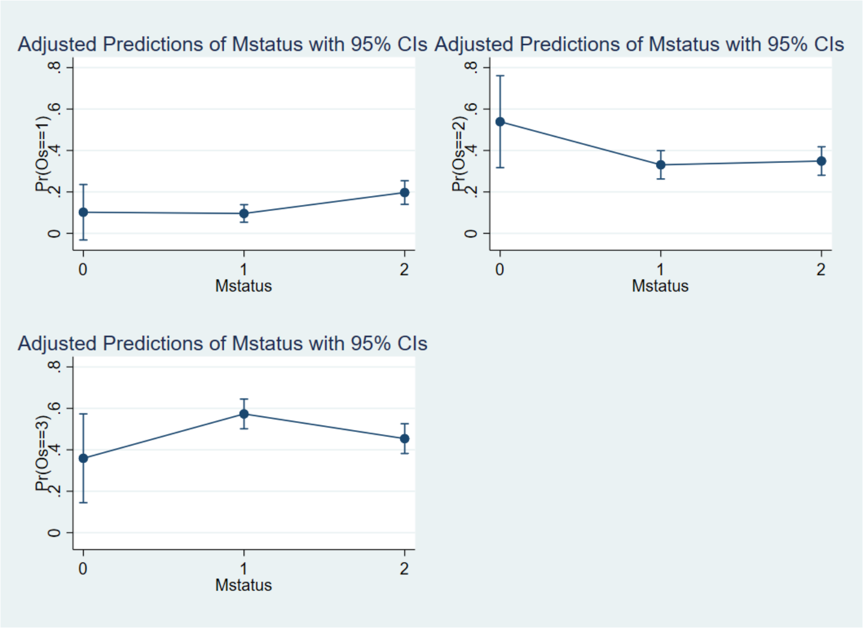


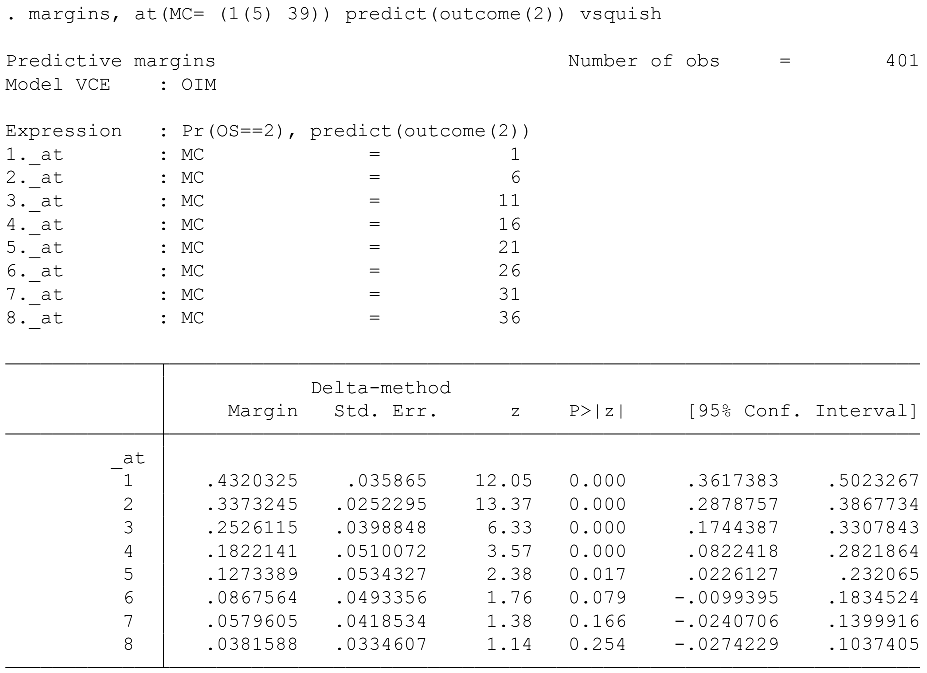
Figure 3.18: predicted probabilities plot

It's showed that for example, the probability of being survival for less than 2 years for the Peri M status women is 10.2%, while the probability for post Mstatus women and the pre-M status women to survival for less than 2 years are also not high, being 9.6% and 19. 7% respectively. For other survival groups, for example, the peri-Mstatus women have 53.9% probabilities to stay in the mid time survival group (high than the pre and post Mstatus women), while the pre-Mstatus women has the highest probabilities of 57.3% to survival for a long time (over 5 years).

Another way to understand the model using the predicted probabilities is to look at the averaged predicted probabilities for different values of the continuous predictor variable MC, averaging across the level of OS.

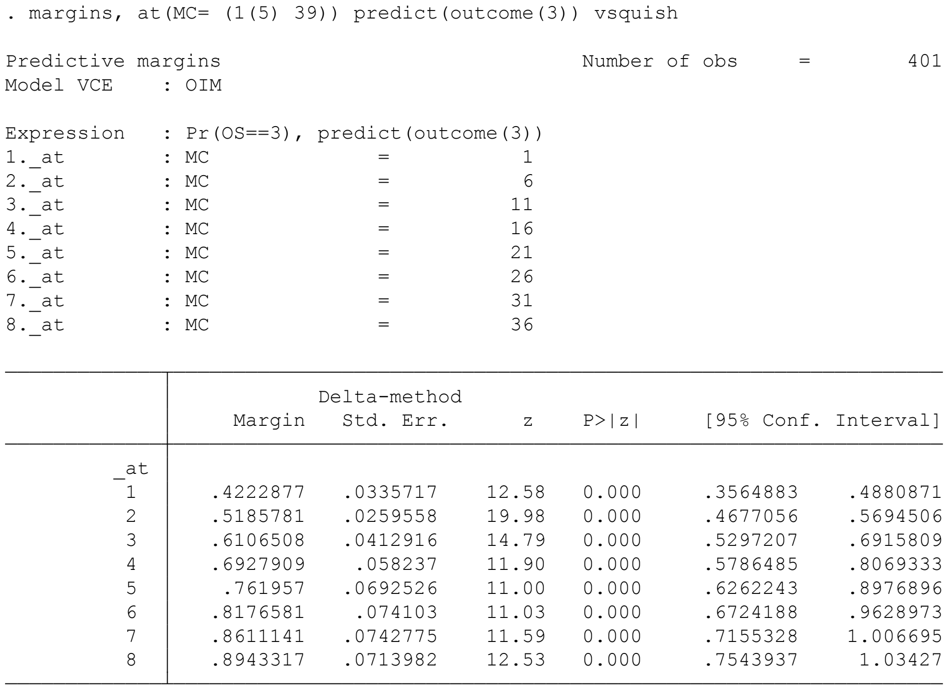
```
. margins, at(MC= (1(5) 39)) predict(outcome(1)) vsquish

Predictive margins                              Number of obs     =        401
Model VCE      : OIM

Expression     : Pr(OS==1), predict(outcome(1))
1._at          : MC              =           1
2._at          : MC              =           6
3._at          : MC              =          11
4._at          : MC              =          16
5._at          : MC              =          21
6._at          : MC              =          26
7._at          : MC              =          31
8._at          : MC              =          36

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .1456797   .0236494     6.16   0.000     .0993278    .1920317
          2  |   .1440974   .0178779     8.06   0.000     .1090574    .1791374
          3  |   .1367377   .0270576     5.05   0.000     .0837058    .1897696
          4  |   .1249951   .0393249     3.18   0.001     .0479196    .2020705
          5  |   .1107042   .0495681     2.23   0.026     .0135525    .2078559
          6  |   .0955855   .0565519     1.69   0.091    -.0152543    .2064252
          7  |   .0809253   .0601557     1.35   0.179    -.0369777    .1988284
          8  |   .0675095   .0607939     1.11   0.267    -.0516444    .1866634
------------------------------------------------------------------------------
```

Figure 3.19: predicted probabilities for different values of MC for short survival group (OS=1)



```
. margins, at(MC= (1(5) 39)) predict(outcome(2)) vsquish

Predictive margins                              Number of obs     =        401
Model VCE    : OIM

Expression   : Pr(OS==2), predict(outcome(2))
1._at        : MC              =           1
2._at        : MC              =           6
3._at        : MC              =          11
4._at        : MC              =          16
5._at        : MC              =          21
6._at        : MC              =          26
7._at        : MC              =          31
8._at        : MC              =          36

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .4320325    .035865    12.05   0.000     .3617383    .5023267
          2  |   .3373245   .0252295    13.37   0.000     .2878757    .3867734
          3  |   .2526115   .0398848     6.33   0.000     .1744387    .3307843
          4  |   .1822141   .0510072     3.57   0.000     .0822418    .2821864
          5  |   .1273389   .0534327     2.38   0.017     .0226127     .232065
          6  |   .0867564   .0493356     1.76   0.079    -.0099395    .1834524
          7  |   .0579605   .0418534     1.38   0.166    -.0240706    .1399916
          8  |   .0381588   .0334607     1.14   0.254    -.0274229    .1037405
------------------------------------------------------------------------------
```

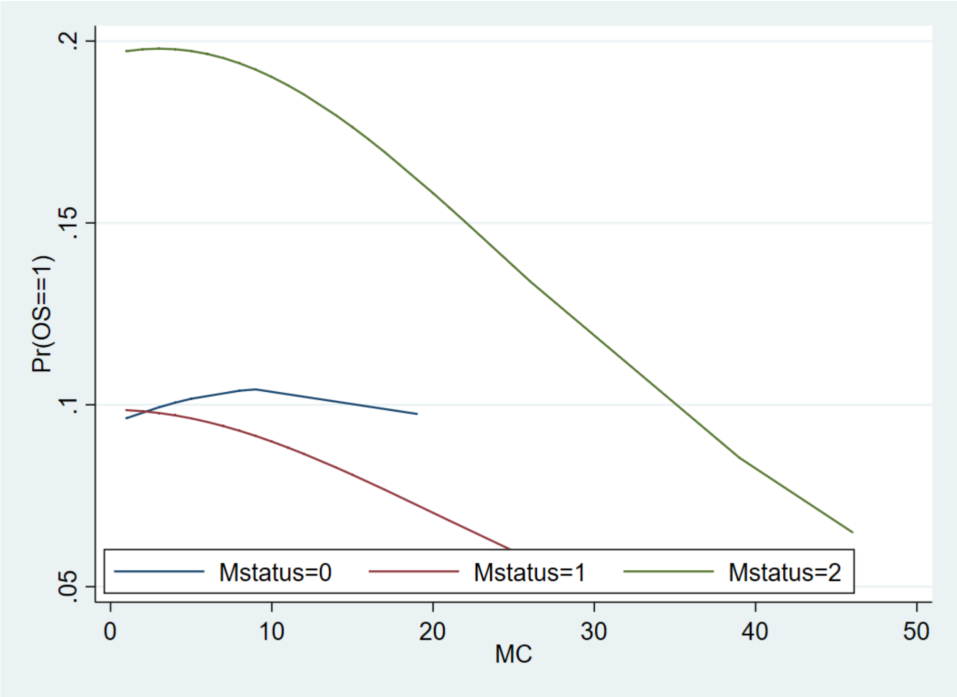
Figure 3.20: predicted probabilities for different values of MC for medium survival group (OS=2)

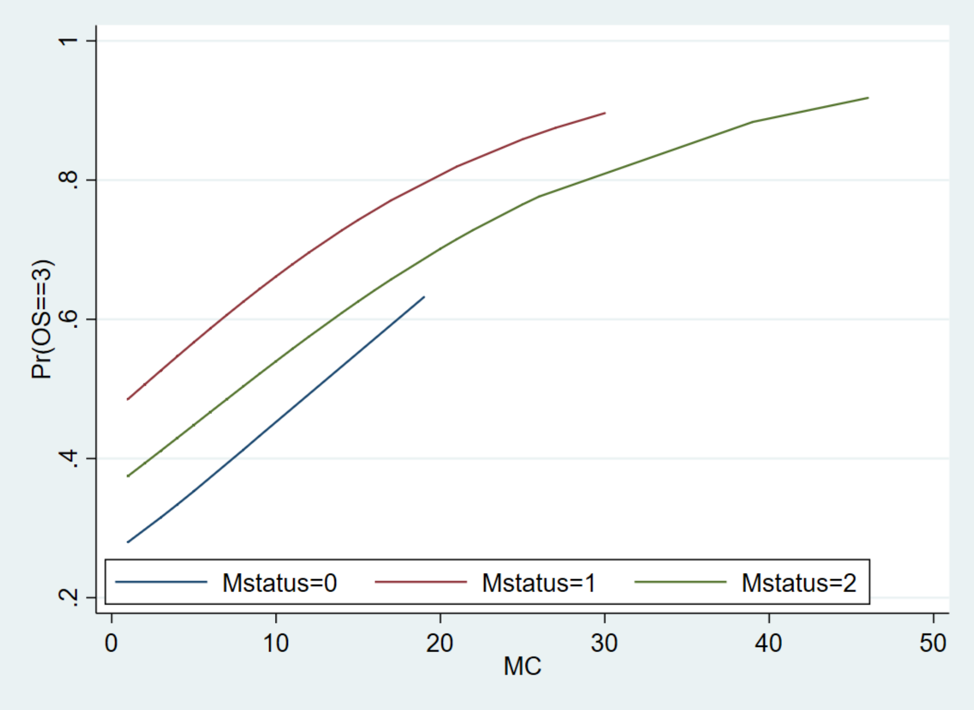


```
. margins, at(MC= (1(5) 39)) predict(outcome(3)) vsquish

Predictive margins                              Number of obs     =        401
Model VCE    : OIM

Expression   : Pr(OS==3), predict(outcome(3))
1._at        : MC              =           1
2._at        : MC              =           6
3._at        : MC              =          11
4._at        : MC              =          16
5._at        : MC              =          21
6._at        : MC              =          26
7._at        : MC              =          31
8._at        : MC              =          36

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .4222877   .0335717    12.58   0.000     .3564883    .4880871
          2  |   .5185781   .0259558    19.98   0.000     .4677056    .5694506
          3  |   .6106508   .0412916    14.79   0.000     .5297207    .6915809
          4  |   .6927909     .058237    11.90   0.000     .5786485    .8069333
          5  |    .761957   .0692526    11.00   0.000     .6262243    .8976896
          6  |   .8176581     .074103    11.03   0.000     .6724188    .9628973
          7  |   .8611141   .0742775    11.59   0.000     .7155328    1.006695
          8  |   .8943317   .0713982    12.53   0.000     .7543937     1.03427
------------------------------------------------------------------------------
```

Figure 3.21: predicted probabilities for different values of MC for long survival group (OS=3)

Below, we plot the predicted probabilities against the writing score by the level of Mstatus for different levels of the outcome variables: OS.





We can see from the plots that or all the 3 status of Menopausal, with MC getting larger, the probability of being survival only for less than 2 years goes down; the probability of being survival for more than 5 years goes up. This may sound like a somewhat counterintuitive fact but the possible reason may be that patient with higher mutation will have a high rate of objective response to some gene therapy such as PD-1 Inhibition.

4. Limitation

In this study, we try to apply all the analysis tool we have learnt to find the relationship between different cancer indicators and the survival condition. However, the outcome is very limited mainly because of the lack of knowing some necessary cofounders. If we want to conduct further analysis, more information is needed, like how the patients are treated after diagnosed, how the treatment method changes the patients' status, etc. In addition, the database is still not large enough. We can make an assumption that after expanding the database, we can see more obvious variation tendency in our models. We cannot then come to the conclusion that this study result is meaningless, since many unnatural facts found in the pure statistical estimation can actually provide a valuable research direction for professional clinic researchers.