678 Midterm Project- Breast Cancer Survival Duration

Runci Hu

2022-12-04

Abstract

This report is regarding the survival duration of breast cancer and the factors affecting it. Breast cancer is now the most commonly diagnosed cancer in women. When diagnosed with breast cancer, the most important part of the clinical decision for patients is the accurate estimation of prognosis and survival duration. So, my aim for this research is to determine what factors primarily affect one's survival time. To achieve my goal, I build a multilevel model of total survival month in two groups, integrative clusters and the Nottingham prognostic index. The result shows that components like age, tumor size, and cohort all significantly influence the result. The report will expand detail into four parts, Information, Method, Result, and Discussion.

Introduction

Breast cancer is now the most common cancer in women and the second leading cause of cancer death in women. In 2020, an estimated of over 680000 women across the world died from breast cancer. This year American, it is predicted that 43,780 deaths from breast cancer will occur. To be in my shoes, I have a matrilineal relative diagnosed with breast cancer, which increases the genetic risk score of myself diagnosed with cancer. These are the reason why I choose this topic. Moreover, cancers are associated with genetic abnormalities. Comparing the genes expressed in normal and diseased tissue can be a good way. During my research, I find BRCA1 and BRCA2 (BReast CAncer susceptibility gene) are two genes with high relation to breast cancer. Their mutations highly increase one's prevalence rate of having breast cancer. Beyond this, I can assume other relative factors. For example, one's age relates to survival time in high possibility because higher age means more body degeneration. Additionally, the distinctive subtypes of breast cancer show different clinical features. Based on the above discussion, I decided to exploit the multilevel model to discover the impacts of random and fixed effects.

Model

Data Preprocessing

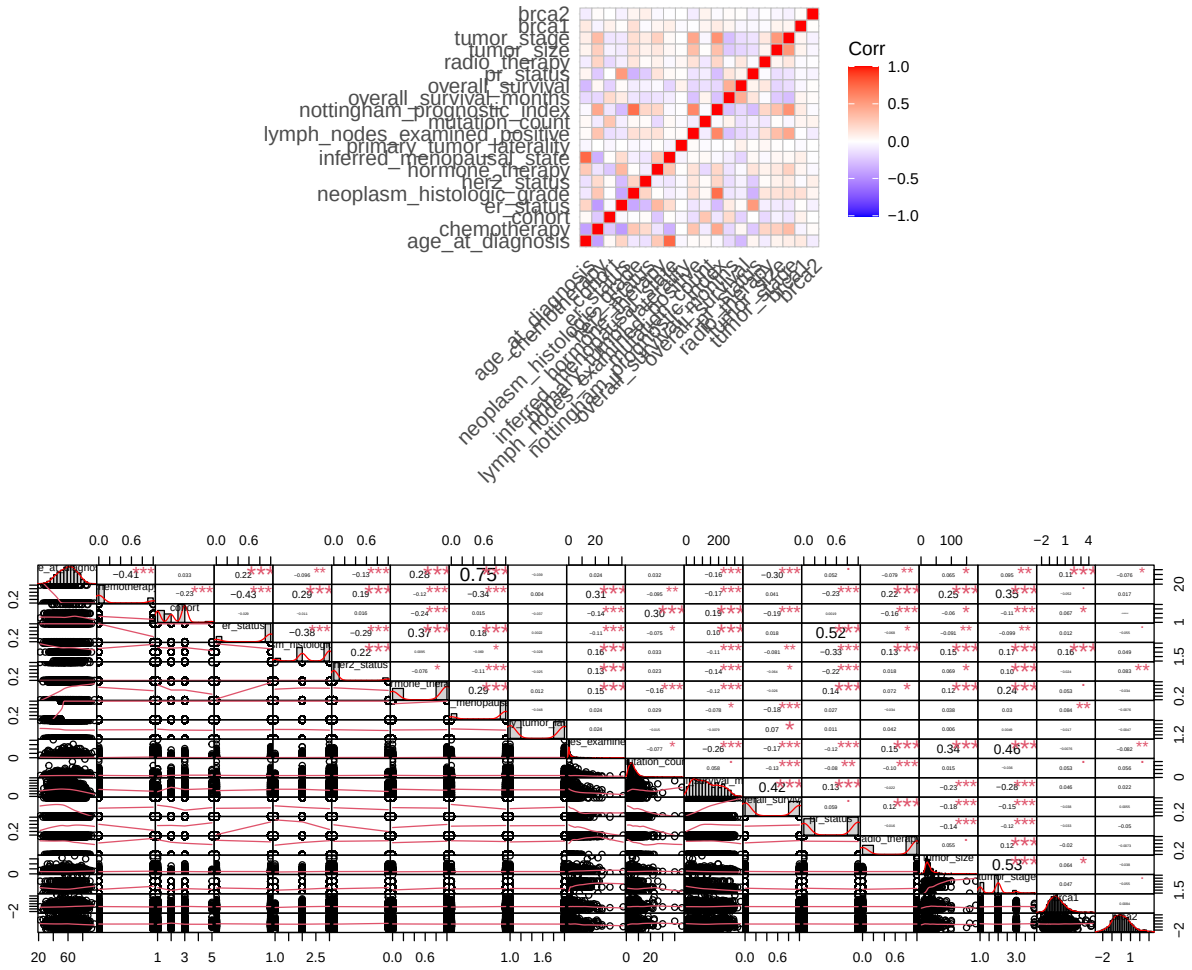
I found the dataset "Breast Cancer Gene Expression Profiles (METABRIC)" in Kaggle (https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric). This data from METABRIC (The Molecular Taxonomy of Breast Cancer International Consortium) contains targeted sequencing data of 1,980 primary breast cancer samples. The original dataset has 693 columns, about the different mRNA levels z-score, and mutation for a large number of genes after column 31. I only include two that I mentioned as symbolic, BRCA1 and BRCA2. The new dataset is named "bc". In addition, I deleted 'NA' data in the bc dataset. Because these missing variables may influent my model. From the rest of the numeric variables, I round them to integer. It is efficient in calculation. I also remove some meaningless columns or unusable columns. Plus, I changed alternative variables, such as positive or negative and left or right, to binary factors

(0 or 1) so that these variables could be used for analysis. There is a table for the remaining variables and their definitions.

column names	explanation
patient_id	Patient ID, represent different patient
age_at_diagnosis	Age of the patient at diagnosis time
type_of_breast_surgery	2 types of breast cancer surgery
cancer_type	1. Breast Cancer or 2. Breast Sarcoma
cancer_type_detailed	5 types of detailed Breast cancer types
cellularity	The amount of tumor cells in the specimen
chemotherapy	Whether or not the patient had chemotherapy
cohort	Groups share a defining characteristic
er_status	Positive or negative for estrogen receptors
neoplasm_histologic_grade	Aggressive level of nature of the cells
her2_status	Positive or negative for HER2
hormone_therapy	Whether or not the patient had hormonal therapy
inferred_menopausal_state	Patient is pre menopausal or post menopausal
integrative_cluster	Molecular subtype ('4ER+', '3', '9', '7', '4ER-', '5', '8', '10', '1', '2', '6')
primary_tumor_laterality	Cancer on right breast or the left breast
lymph_nodes_examined_positive	Whether lymph node involved by the cancer
mutation_count	Number of gene that has relevant mutations
nottingham_prognostic_index	A calculation determines prognosis following surgery for breast cancer
overall_survival_months	Survival Duration from intervention to death
overall_survival	Whether the patient is alive of dead
pr_status	Positive or negative for progesterone receptor
radio_therapy	Whether or not the patient had radio therapy
tumor_size	Tumor size
tumor_stage	Stage of the cancer
death_from_cancer	Wether the patient's death was due to cancer
brca1	BReast CAncer gene 1
brca2	BReast CAncer gene 2

The two groups of random effects I set up are nottingham_prognostic_index and integrative_cluster. In my research, Nottingham Prognostic Index is used to determine the prognosis following surgery for breast cancer. It is calculated by tumor size, the number of involved lymph nodes, and the grade of the tumor. More importantly, it is used to determine one's 5-year survival probability. NPI between 2 to 2.4 has a 93% chance of surviving five years. NPI in 2.4 to 3.4, 3.4 to 5.4, and over 5.4 correspond respectively to 85%, 70%, and 50% probability. When I group the bc data, there is no NPI variable between the 2.4-3 group and the 5-5.4 group. So, I rounded them to an integer with no bias in grouping them. Another random effect is integrative_cluster. It is about 11 different types of Molecular subtypes of breast cancer. The 11 groups are '4ER+', '3', '9', '7', '4ER-', '5', '8', '10', '1', '2', and '6'.

Check Correlations



I first deliberate correlation between my y, overall survival month, with different x variables. Based on the upper graph, overall survival month has nearly 0 relations with primary tumor laterality, radio therapy, and brca2. As a result, I am not considering these variables in my model. Moreover, in the bottom plot, I compare the correlation between different x variables, with a setup of 0.5 as the criteria. Age/inferred menopausal state is highly correlated because menopausal transition most often begins between ages 45 and 55. The menopausal state is closely related to women's age. Hence, I decide to keep age but delete inferred menopausal state. Pr status/er-status also has a higher correlation than 0.5. Based on the above plots of their relations to the groups, I choose to drop er-status. Another high correlation group is tumor size/tumor stage. I am indecisive about these two, so I try both in my function. The final decision is to keep tumor size. Above all, I drop primary tumor laterality, radio therapy, brca2, er-status, inferred menopausal state, tumor stage.

Compare one variable to two subsets

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

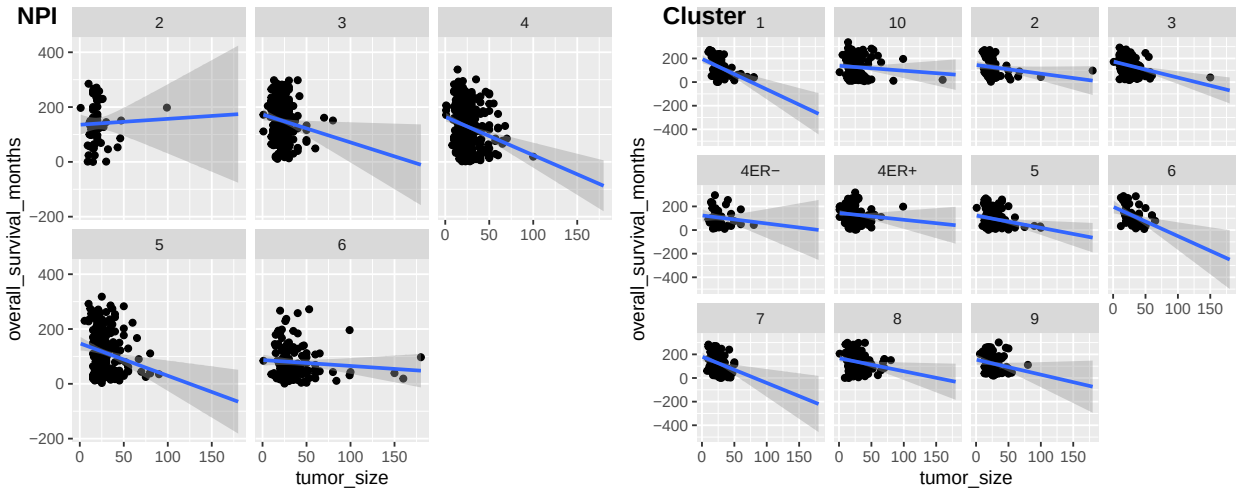


Figure 1 plots tumor size associations for two subsets, NPI and integrative cluster, of each group in it. The overall relation between tumor size and overall survival month is negative, in which one's survival month is shorter than others with smaller tumor sizes. On the other hand, in each group, the slope of tumor size is quite different from others. I can describe group differences in two subsets and then determine if those differences are related to differences in overall survival time.

Group fixed effects

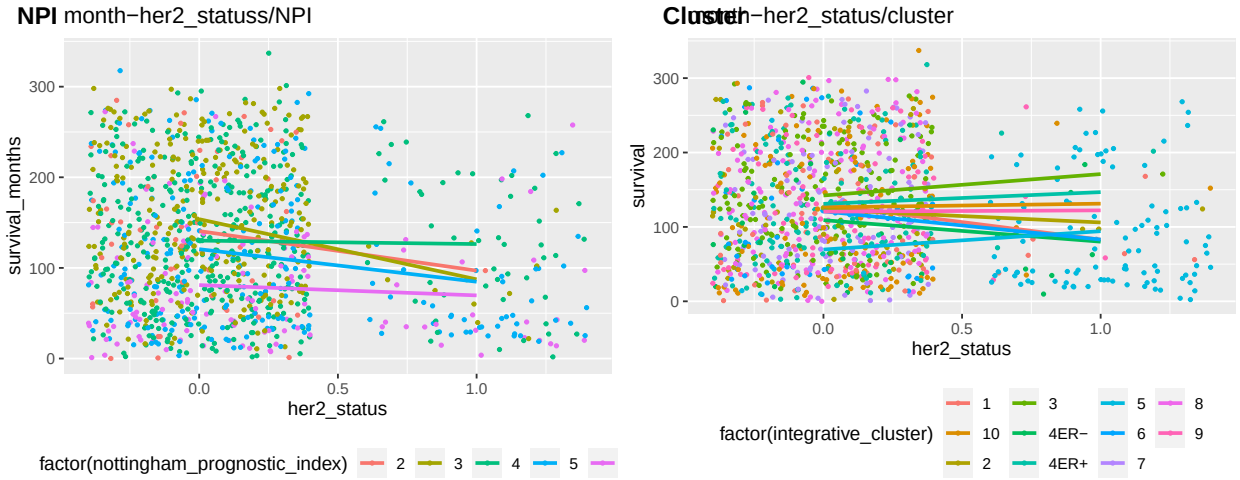


Figure 1: Relation between overall_survival_months with her2-status in 2 groups

Figure 1 and Figure 2 are two sets of plots representing relations between overall survival months with her2-status and hormone therapy. The left plot in each Figure is the relation in the NPI groups, and the right one is in the integrative cluster groups. By viewing the slope differences in each group factor, we are able to decide for what group we should put this fixed variable. If the slopes in a figure have the same tendency, downward for example, the variable perfectly corresponds to this group. I compare the two groups, choose the one with more significant difference in its slope tendency, and then this variable should fix in this chosen group. Taking the above two graphs as examples, her2_status should be put in the integrative cluster group, whereas hormone therapy will be fixed in the NPI group. Similarly, I plot the relations of each variable with overall survival months. For detailed information, please see in the appendix.

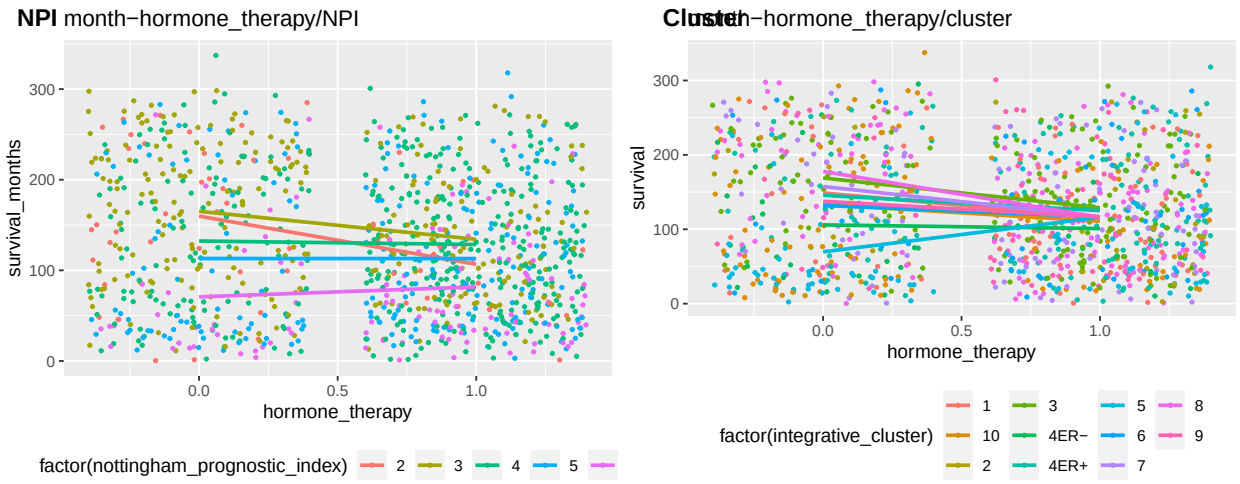


Figure 2: Relation between overall_survival_months with hormone therapy in 2 groups

Build the Model

```
model1 <- lmer( overall_survival_months ~  overall_survival + age_at_diagnosis
              + tumor_size + chemotherapy + cohort + her2_status + brca1
              + lymph_nodes_examined_positive
              + (1 + cohort | nottingham_prognostic_index)
              + (age_at_diagnosis  + brca1 + chemotherapy + her2_status |
                   integrative_cluster),
              data = bc)

summary(model1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: overall_survival_months ~ overall_survival + age_at_diagnosis +
##     tumor_size + chemotherapy + cohort + her2_status + brca1 +
##     lymph_nodes_examined_positive + (1 + cohort | nottingham_prognostic_index) +
##     (age_at_diagnosis + brca1 + chemotherapy + her2_status |
##         integrative_cluster)
##    Data: bc
##
## REML criterion at convergence: 12157.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.84617 -0.72004 -0.09319  0.66772  3.04698
##
## Random effects:
##  Groups                      Name             Variance  Std.Dev. Corr
##  integrative_cluster         (Intercept)       890.5113 29.841
##                              age_at_diagnosis    0.1452  0.381   -1.00
##                              brca1               4.1419  2.035    0.49 -0.49
##                              chemotherapy       22.6351  4.758   -0.89  0.89
```

```
##                               her2_status        35.0257  5.918   -1.00  1.00
##  nottingham_prognostic_index (Intercept)        81.7426  9.041
##                               cohort             30.1056  5.487   -1.00
##  Residual                                      4075.6321 63.841
##
##
##
##
##  -0.04
##  -0.43  0.91
##
##
##
## Number of obs: 1092, groups:
## integrative_cluster, 11; nottingham_prognostic_index, 5
##
## Fixed effects:
##                                 Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)                     113.6997    17.4688   16.6430   6.509 5.94e-06
## overall_survival                 63.4456     4.3145 1064.9274  14.705  < 2e-16
## age_at_diagnosis                 -0.6288     0.2152   12.0114  -2.922  0.01279
## tumor_size                       -0.3779     0.1421 1075.5041  -2.660  0.00793
## chemotherapy                    -15.8312     6.2474   53.7354  -2.534  0.01422
## cohort                           17.9617     3.3719    4.1559   5.327  0.00536
## her2_status                     -21.1685     6.4867   31.5733  -3.263  0.00265
## brca1                             5.0787     2.2042   14.0086   2.304  0.03705
## lymph_nodes_examined_positive    -1.4989     0.5681  346.0057  -2.638  0.00871
##
## (Intercept)                     ***
## overall_survival                ***
## age_at_diagnosis                *
## tumor_size                      **
## chemotherapy                    *
## cohort                          **
## her2_status                     **
## brca1                           *
## lymph_nodes_examined_positive **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) ovrll_ ag_t_d tmr_sz chmthr cohort hr2_st brca1
## ovrll_srvvl -0.382
## age_t_dgnss -0.877  0.236
## tumor_size  -0.132  0.108 -0.085
## chemotherpy -0.452  0.053  0.464 -0.195
## cohort      -0.439  0.145  0.067  0.006  0.123
## her2_status -0.236  0.056  0.232  0.000 -0.030 -0.031
## brca1        0.150 -0.020 -0.141 -0.060  0.030 -0.036 -0.022
## lymph_nds__  0.029  0.120 -0.073 -0.240 -0.211  0.044 -0.068  0.017
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

Exam the ramdon effects

This tables are the summary of random effects. The first one is random effect of integrative_cluster groups and the second is NPI groups.

```
## $integrative_cluster
##      (Intercept) age_at_diagnosis       brca1 chemotherapy her2_status
## 1    -15.2372319      0.194603227 -0.61873068    2.0208966  2.99101694
## 10   -18.3953287      0.235563202 -1.68672994    1.2946716  3.42405627
## 2     21.5835293     -0.275705664  0.94417630   -2.7794789 -4.22340738
## 3     21.7685334     -0.277909442  0.92581546   -2.8515127 -4.27032054
## 4ER- -14.0772153      0.180099590 -0.98775396    1.3559292  2.67724398
## 4ER+  -0.2812738      0.003502815  0.05848774    0.1273455  0.07158237
## 5    -25.2370460      0.321808347 -0.30411417    4.2287123  5.10087465
## 6     -4.0781915      0.051823046  0.21812303    1.0092094  0.87726354
## 7     -5.4078131      0.068610761  0.33740423    1.4054814  1.17611139
## 8     37.3320042     -0.476511615  1.10070523   -5.4573972 -7.41244447
## 9      2.0300333     -0.025884267  0.01261675   -0.3538572 -0.41197677
##
## $nottingham_prognostic_index
##   (Intercept)     cohort
## 2 -0.09804375  0.0594804
## 3 -9.02139469  5.4750378
## 4 -4.03642204  2.4495870
## 5  4.55888931 -2.7666191
## 6  8.59697118 -5.2174860
##
## with conditional variances for "integrative_cluster" "nottingham_prognostic_index"
```

Result

Interpretation

By the model summary, we can write the formula:

$$overall\,survival\,months = 113.70 + 63.45 \times overall\,survival - 0.63 \times age\,at\,diagonsis - 0.38 \times tumor\,size$$

$$-15.83 \times chemotherapy + 17.96 \times (1 + cohort) - 21.17 \times her2\,status + 5.07 \times brca1 - 1.50 \times lymph\,nodes\,examined\,positive$$

For example, patient with id 121,

$$113.70 + 63.45 \times 1 - 0.63 \times 79 - 0.38 \times 30 - 15.83 \times 0 + 17.96 \times (1 + 1) - 21.16 \times 0 + 5.08 \times (-0.2284) - 1.50 \times 6$$

, the answer is not exactly the same but nearly. From the formula, we can understand the influence of each variable on patients' survival duration. The overall survival has a great impact on survival duration. Because it is a binary variable, patients can be alive or dead. If one is dead, the survival duration also stops increasing. Besides, age negatively affects survival time, just like I predicted before. People's body degeneration increases when they get older. Chemotherapy is the most surprising result. The coefficient is negative and not a small number, which means that chemotherapy may shorten survival duration. As a result, I would recommend patients take other therapy methods. At least, the BRCA1 gene indeed shortens one's survival duration. The interception is positive, but the variables of BRCA1 are all negative. Hence, the influence on the total survival months is negative.

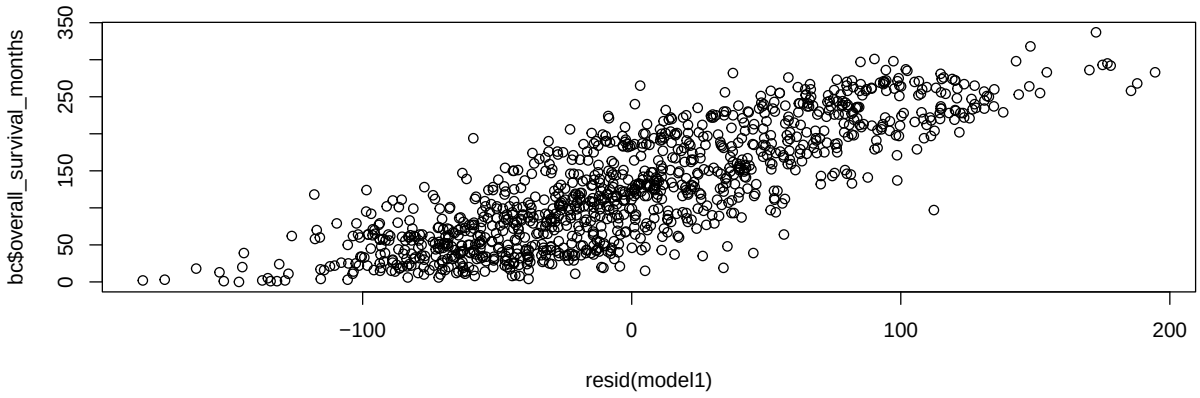
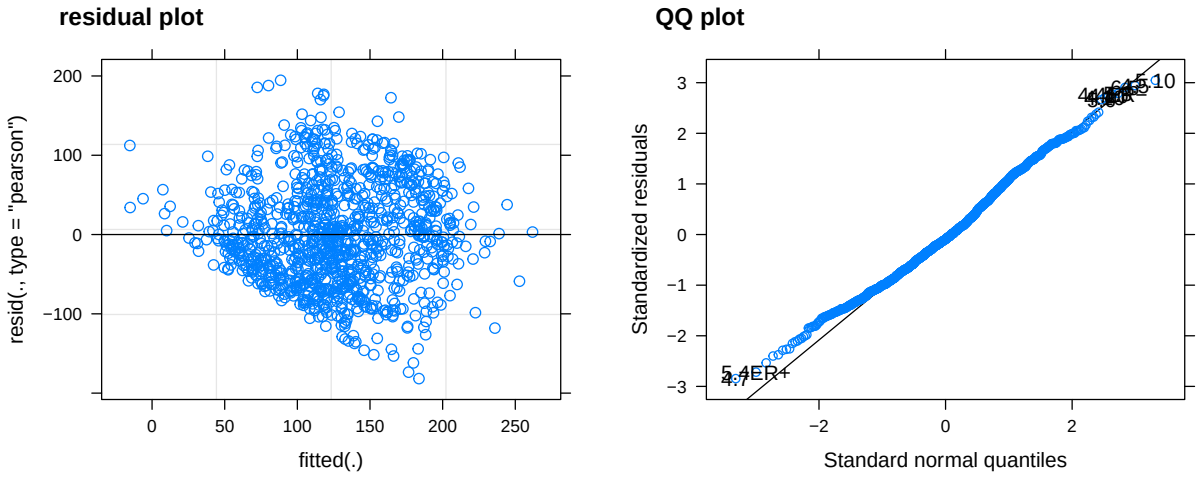


Figure 3: Model Checking1

Model Checking

```
## Analysis of Variance Table
##
## Response: sqr_residuals
##              Df     Sum Sq   Mean Sq F value  Pr(>F)
## patient_id    1 1.6357e+08 163565366  5.9219 0.01511 *
## Residuals  1090 3.0106e+10  27620407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



To test the assumption, I firstly plot the model residuals versus the predictor y. By seeing figure 1, I do not have enough evidence to tell the randomness between them. So I do one more step of testing the homogeneity of variance. I pick and calculate the residuals from the model and put them in a new column named residuals. Next, I create two new columns, one with the absolute value of the residuals and another with squares of absolute values, to provide a more accurate estimate. Finally, I ANOVA the squared residuals for each patient residuals. In the result, the P value is 0.02, which is smaller than 0.05. Therefore, the variance of the residuals is equal, and the assumption is met. Now, let us see two visualizations of the model. The points in the residual plot show no pattern; they are randomly dispersed. Thus, the model

is appropriate. QQ plot result can also prove the model is appropriate because most points fall on the 45-degree reference line.

Discussion

For what to improve, I need more careful consideration of picking variables. The current model shows a lot of factors that will shorten the survival duration. However, the more important information for patients should be the reasons that can increase their living time. My point of view is seeing different impacts of other genes. Genes express in normal and disease cells are different. After we find the gene type with the worst effect on survival duration, we can treat that gene as targeted therapy. My best wish for this experiment is to find a solution that could rapidly increase the patient's survival duration. Therefore, this motivates my next step in improving the model.

Reference

Breast Cancer Gene Expression Profiles (METABRIC). (2020, May 26). Kaggle. https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric Palmeri, M. (n.d.). Chapter 18: Testing the Assumptions of Multilevel Models. https://ademos.people.uic.edu/Chapter18.html R Bootcamp: Introduction to Multilevel Model and Interactions | QuantDev Methodology. (n.d.). https://quantdev.ssri.psu.edu/tutorials/r-bootcamp-introduction-multilevel-model-and-interactions

Appendix

Groups

```
count(bc, nottingham_prognostic_index)
```

```
##   nottingham_prognostic_index   n
## 1                           2  56
## 2                           3 276
## 3                           4 407
## 4                           5 234
## 5                           6 119
```

```
count(bc,integrative_cluster)
```

```
##    integrative_cluster   n
## 1                    1  77
## 2                   10 123
## 3                    2  46
## 4                    3 171
## 5                 4ER-  40
## 6                 4ER+ 142
## 7                    5 107
## 8                    6  53
## 9                    7 110
## 10                   8 147
## 11                   9  76
```

```
## count variables in each group
```

Relation Plot

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

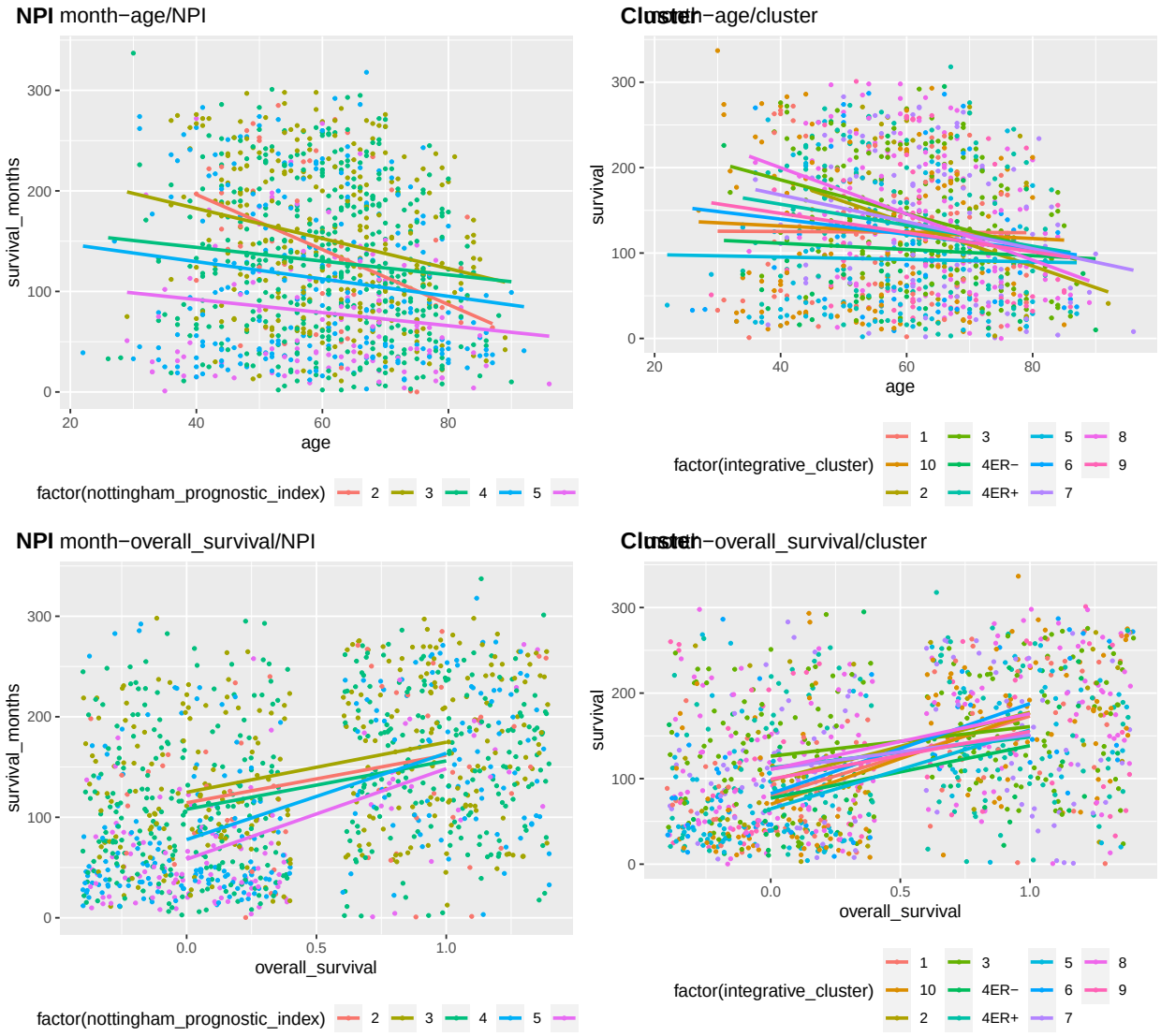


Figure 4: Relation Plot 1

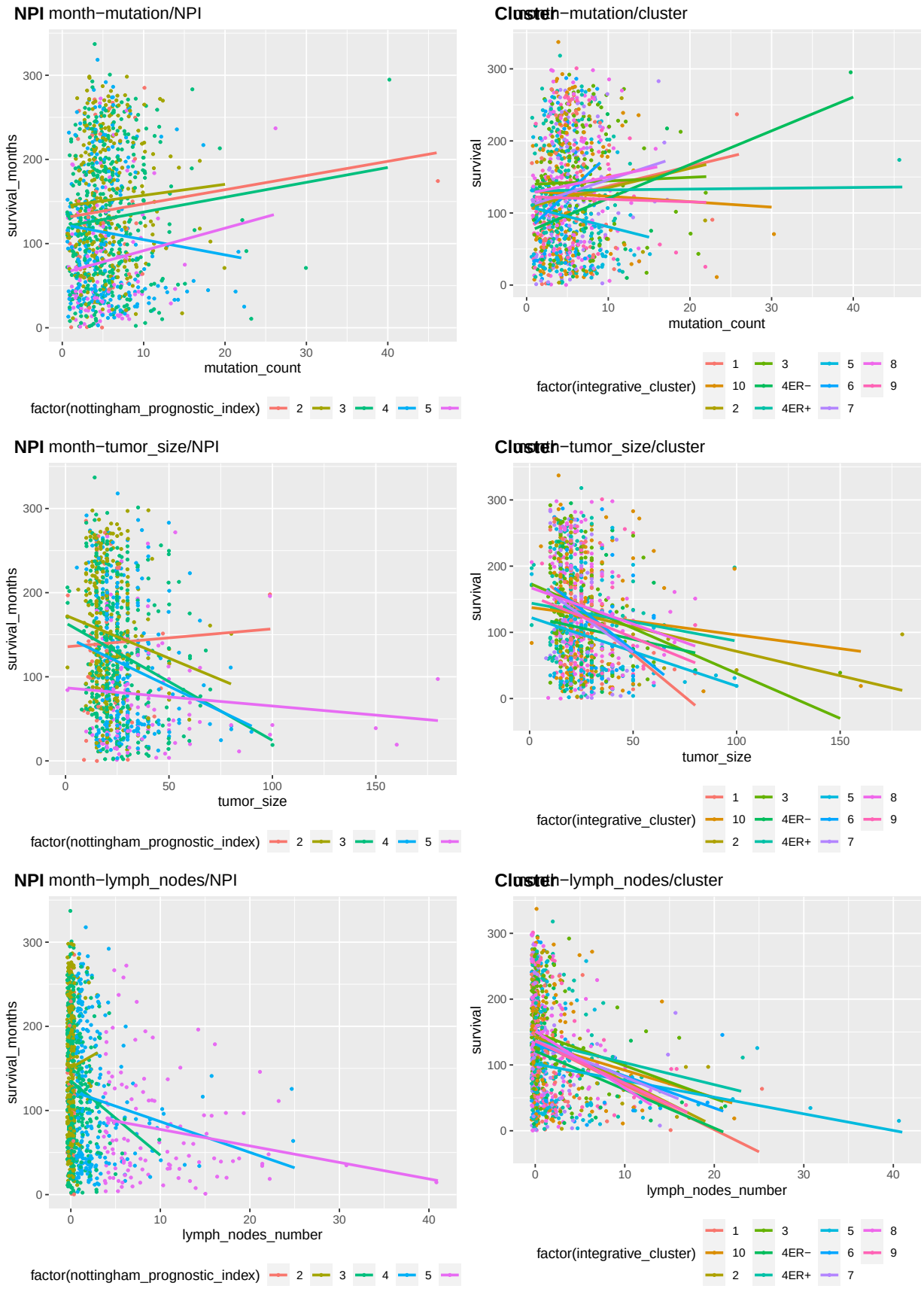


Figure 5: Relation Plot 2

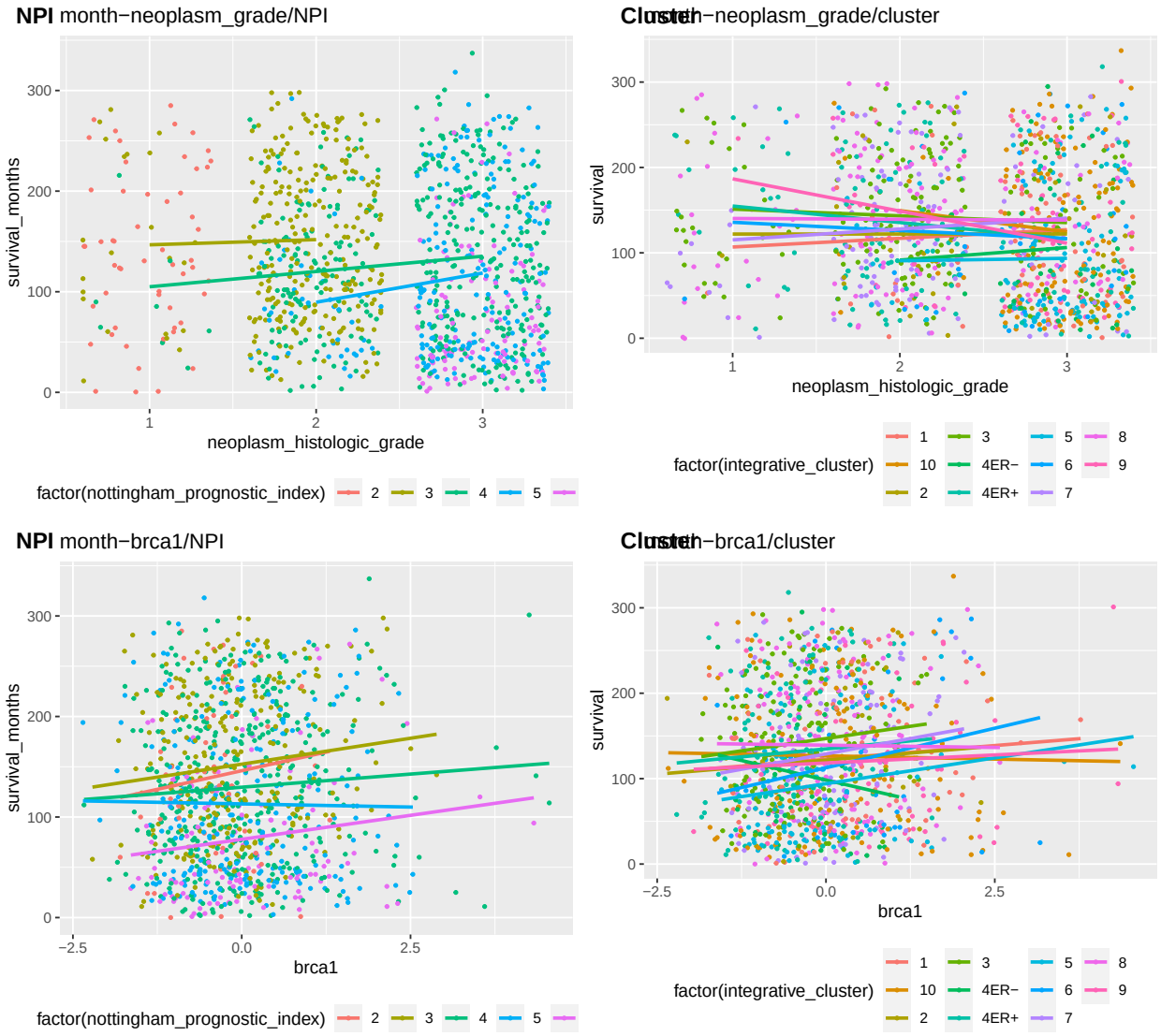


Figure 6: Relation Plot 3

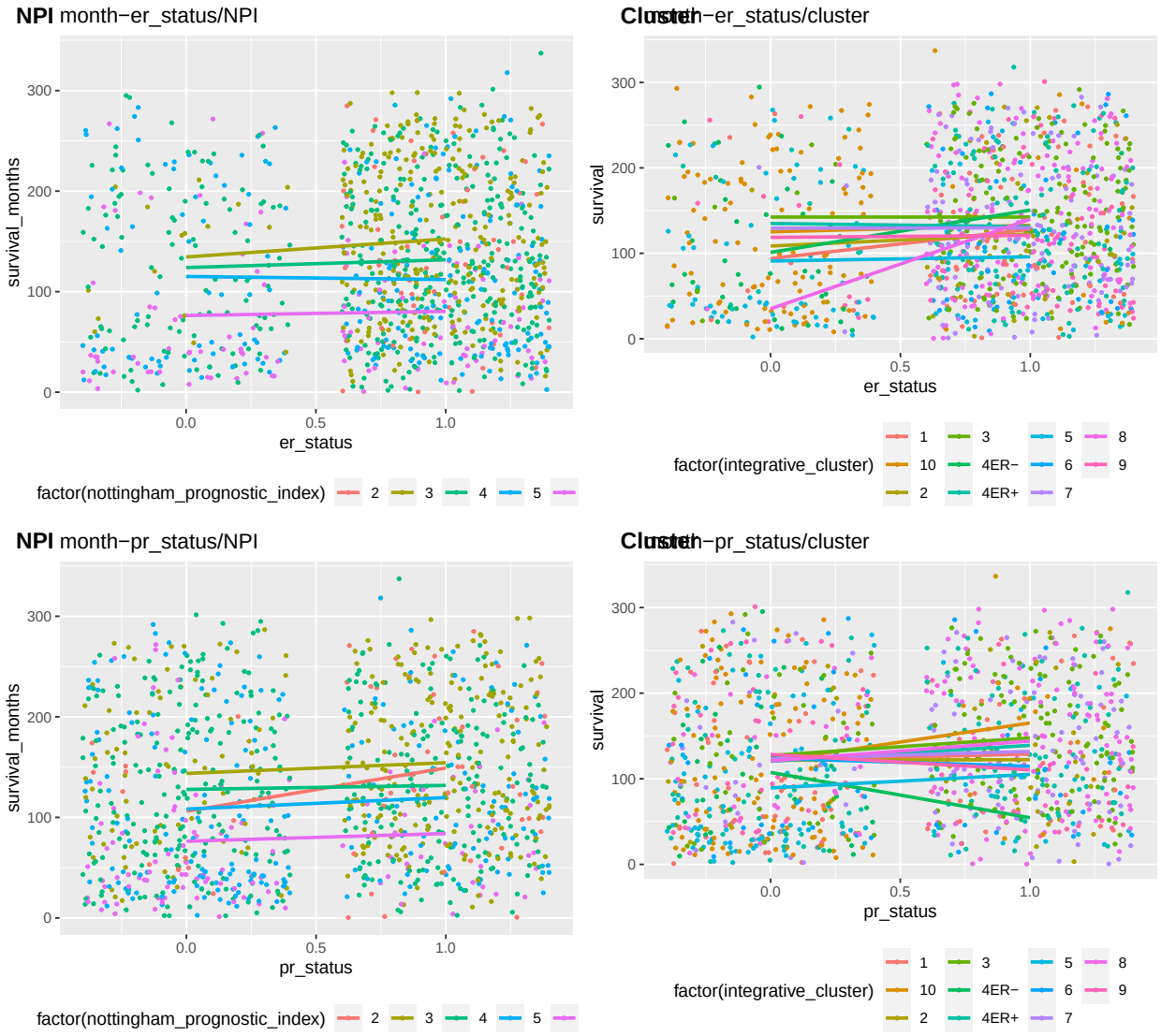


Figure 7: Relation Plot 4

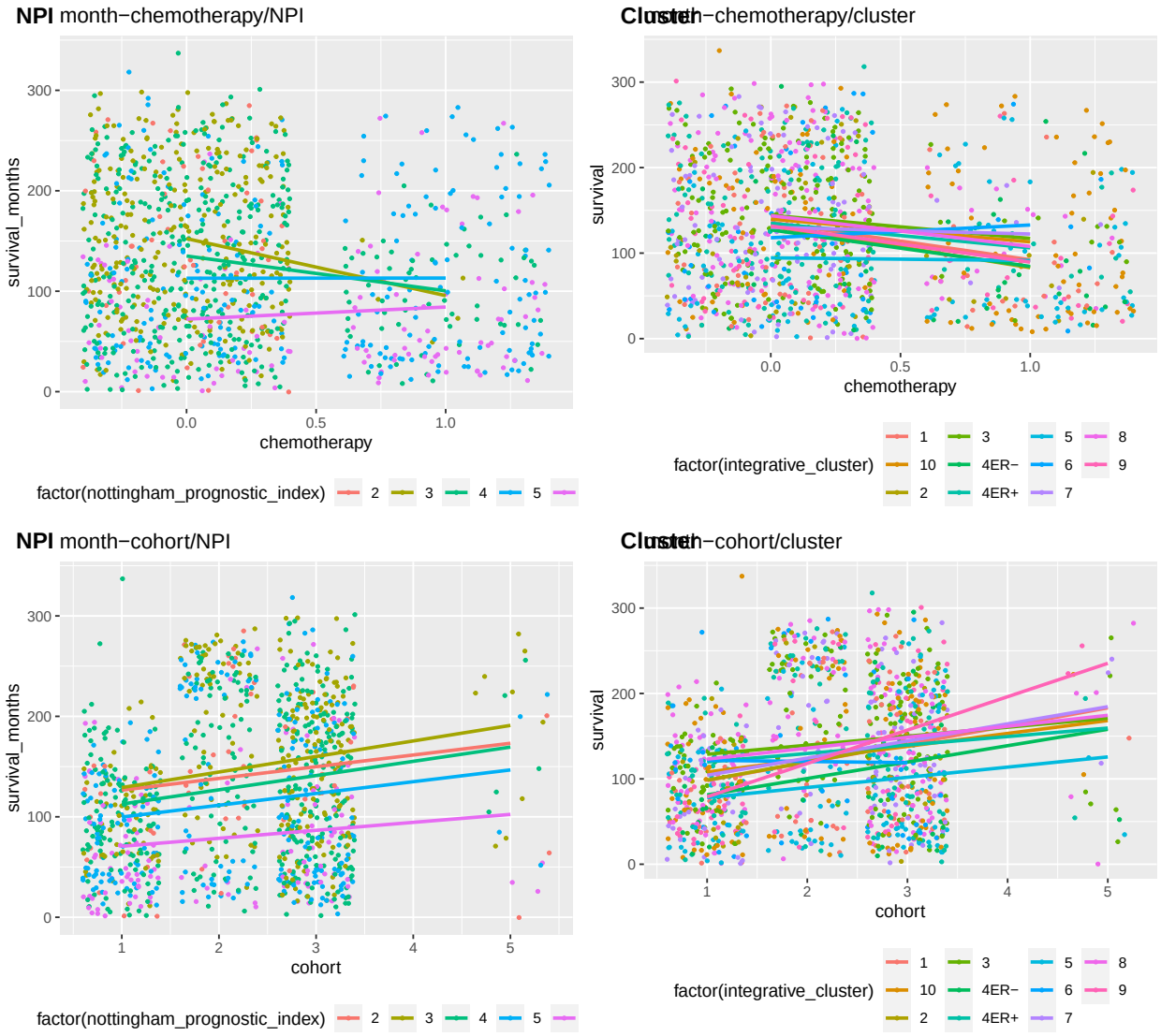


Figure 8: Relation Plot 5