Bis620 Final Project

Team member: Yuchen Chang(yc925), Jia Wei (jw2844)

```
library(bis620.2022)
```

1. Background and Motivation

We searched for some statistics related to colorectal cancer, and found that colorectal cancer is the third most diagnosed cancer worldwide. An estimated 1,880,725 people were diagnosed with colorectal cancer in 2020[1], which means investigating in this field is quite necessary. We also found that when colorectal cancer is found early, it can often be cured.The 5-year survival rate of people with localized stage colorectal cancer is 91%[1], meaning that the use of treatment is efficient for colorectal cancer. In addition, Panitumumab in combination with Folfox is one of the preferred targeted strategies to deal with left-sided RAS wild-type metastatic colorectal cancer. The main members of the RAS gene family includes KRAS, HRAS, and NRAS[2]. The presence of these different mutations may have an impact on treatment decisions. Since KRAS and NRAS gene mutations are more commonly found in certain cancer types, we included KRAS and NRAS mutations in exons 3, 4, and 5 in our studies to compare the survival time of patients with different mutant types in different treatment uses.

2. Research Question

We want to investigate: Does the treatment of Chemotherapy alone or Panitumumab in combination with Chemotherapy have a different impact on patients' survival time, controlling for the effects of mutant type? We hypothesize that these two treatments will have different impacts on patients' survival time controlling for the effects of mutant type.

In our research project, we first do a Cox Proportional Hazards Model on the entire dataset to see the general impacts on how the differences in treatments affect patient survival time. Then, the dataset is divided into two categories based on mutant types "Mutant" and "Wild-Type". We conducted additional research on these two different datasets and developed two Cox Proportional Hazards Models.

3. Data cleaning and Exploration

3.1 Read the data and build the dataframe

```
library(haven)
library(purrr)
#> Warning: 程辑包'purrr'是用R版本4.2.2 来建造的
library(dplyr)
#>
#> 载入程辑包：'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
library(ggplot2)
#> Warning: 程辑包'ggplot2'是用R版本4.2.2 来建造的
library(tidyr)

dlfinal <- readRDS("../data-raw/dlfinal.rds")
dl <- dlfinal
```

Then we take the columns of `SUBJID`, `ATRT`, `PRSURG`, `DTHDY`, `DTH`, `LIVERMET`, `AGE`, `SEX`, `B_WEIGHT`, `B_HEIGHT`, `RACE`, `DIAGTYPE` in the `adsl` table of the original "PDS_DSA_20050203" dataset as our new dataframe for further analysis.

```
dat <- data.frame(dl$adsl$SUBJID, dl$adsl$ATRT,
                  dl$adsl$PRSURG, dl$adsl$DTHDY, dl$adsl$DTH,
                  dl$adsl$LIVERMET, dl$adsl$AGE, dl$adsl$SEX,
                  dl$adsl$B_WEIGHT, dl$adsl$B_HEIGHT,
                  dl$adsl$RACE, dl$adsl$DIAGTYPE)
names(dat) <- c("SUBJID", "ATRT", "PRSURG", "DTHDY", "DTH", "LIVERMET",
                "AGE", "SEX", "B_WEIGHT", "B_HEIGHT", "RACE", "DIAGTYPE")
```

3.2 Deal with biomarker

As our study will compare whether the treatment of Chemotherapy alone or Panitumumab in combination with Chemotherapy have a different impact on patients' with different mutant type for RAS gene, we also include the information from `biomarker` table.

Since KRAS and NRAS mutations are the main members of the RAS family, we will consider both KRAS and NRAS mutations in exons 3, 4, and 5 in our research study.

```
b <- dl$biomark |>
  select(SUBJID, BMMTNM1:BMMTR6, BMMTNM15, BMMTR15, BMMTNM16, BMMTR16) |>
  pivot_longer(-SUBJID) |>
  group_by(SUBJID) |>
  summarize(
    Mutant = sum(value == "Mutant"),
    Unknown = sum(value == "" | value == "Failure"),
    `Wild-type` = sum(value == "Wild-type")
  )
```

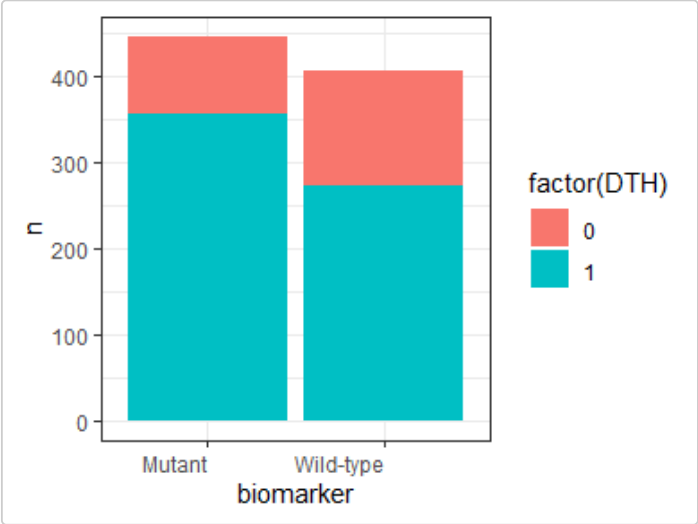
In our new dataframe, a patient is considered "Mutant" if at least one "Mutant" biomarker is found in KRAS or NRAS exons 3, 4, or 5. Patients will be considered "Wild-type" if they are not "Mutant" and they have more "Wild-type" markers than "Unknown" or "Failure" in KRAS or NRAS exons 3, 4, or 5.

```
get_biomarker <- function(x) {
  if (x[["Mutant"]] > 0) {
    return("Mutant")
  }
  if (x[["Wild-type"]] >  x[["Unknown"]]) {
    return("Wild-type")
  }
  return("Unknown")
}

b$biomarker <- map_chr(seq_len(nrow(b)), ~ get_biomarker(b[.x, ]))

db <- left_join(
    b |>
      filter(biomarker != "Unknown") |>
      select(SUBJID, biomarker),
    dat,
    by = "SUBJID"
  )
```

3.3 Data Visualization

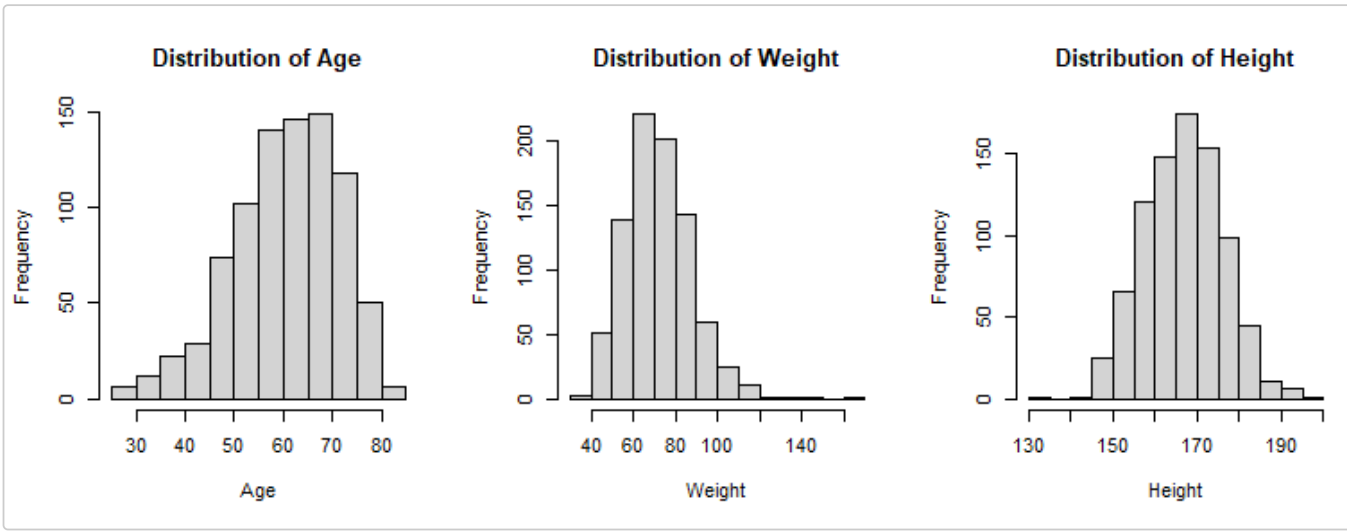
Then, we will draw some histograms for patients' `biomarker`, `AGE`, `B_Height`, and `B_Weight` to present an general view of the distribution of our data. It seems like the proportion of people died in the Mutant type tumors is more than it in the Wild-type tumors.

```
dbs <- db
dbs |>
  group_by(biomarker, DTH) |>
  summarize(n = n(), .groups = "drop") |>
  ggplot(aes(biomarker, n, fill = factor(DTH))) +
    geom_col() +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 360, hjust = 1))
```



We discovered that the distribution for patients' `AGE` does not follow a normal distribution. It is a left-skewed distribution, implying that more patients in the study are older.The median of the `AGE` data is 62.

```
par(mfrow = c(1, 3))
hist(db$AGE, main = "Distribution of Age", xlab = "Age")
hist(db$B_WEIGHT, main = "Distribution of Weight", xlab = "Weight")
hist(db$B_HEIGHT, main = "Distribution of Height", xlab = "Height")
```



4. Analysis and Interpretations

4.1 Full Model

In order to find out which variables have influence on the survival days of patients, we first build a Cox Proportional-Hazards Model(Model1) to get an overall view of relationships between different variables.

```
library(survival)
library(survminer)
#> 载入需要的程辑包：ggpubr
#>
#> 载入程辑包：'survminer'
#> The following object is masked from 'package:survival':
#>
#>     myeloma
model1 <- coxph(Surv(db$DTHDY, db$DTH) ~  biomarker + ATRT +
                  PRSURG + LIVERMET + AGE + SEX + B_WEIGHT +
                  B_HEIGHT + factor(RACE) + DIAGTYPE, data = db)
model1
#> Call:
#> coxph(formula = Surv(db$DTHDY, db$DTH) ~ biomarker + ATRT + PRSURG +
#>     LIVERMET + AGE + SEX + B_WEIGHT + B_HEIGHT + factor(RACE) +
#>     DIAGTYPE, data = db)
#>
#>                                         coef exp(coef)  se(coef)      z
#> biomarkerWild-type                   -0.387382  0.678832  0.082766 -4.680
#> ATRTPanitumumab + FOLFOX              0.007817  1.007848  0.080652  0.097
#> PRSURGY                              -0.393804  0.674486  0.142702 -2.760
#> LIVERMETY                            -0.037312  0.963376  0.125440 -0.297
#> AGE                                   0.008019  1.008052  0.004191  1.913
#> SEXMale                              -0.104181  0.901062  0.113540 -0.918
#> B_WEIGHT                             -0.005905  0.994113  0.003359 -1.758
#> B_HEIGHT                              0.013320  1.013409  0.006608  2.016
#> factor(RACE)Black or African American 0.710826  2.035671  0.531453  1.338
#> factor(RACE)Hispanic or Latino        0.511704  1.668132  0.449091  1.139
#> factor(RACE)Other                     0.559463  1.749733  0.578989  0.966
#> factor(RACE)White or Caucasian        0.249580  1.283487  0.413132  0.604
#> DIAGTYPERectal                       -0.213845  0.807473  0.088770 -2.409
#>                                             p
#> biomarkerWild-type                   2.86e-06
#> ATRTPanitumumab + FOLFOX             0.92279
#> PRSURGY                              0.00579
#> LIVERMETY                            0.76613
#> AGE                                  0.05569
#> SEXMale                              0.35885
#> B_WEIGHT                             0.07877
#> B_HEIGHT                             0.04383
#> factor(RACE)Black or African American 0.18106
#> factor(RACE)Hispanic or Latino       0.25453
#> factor(RACE)Other                    0.33391
#> factor(RACE)White or Caucasian       0.54577
#> DIAGTYPERectal                       0.01600
#>
#> Likelihood ratio test=45.31  on 13 df, p=1.859e-05
#> n= 853, number of events= 628
#>    (因为不存在，1个观察量被删除了)
```

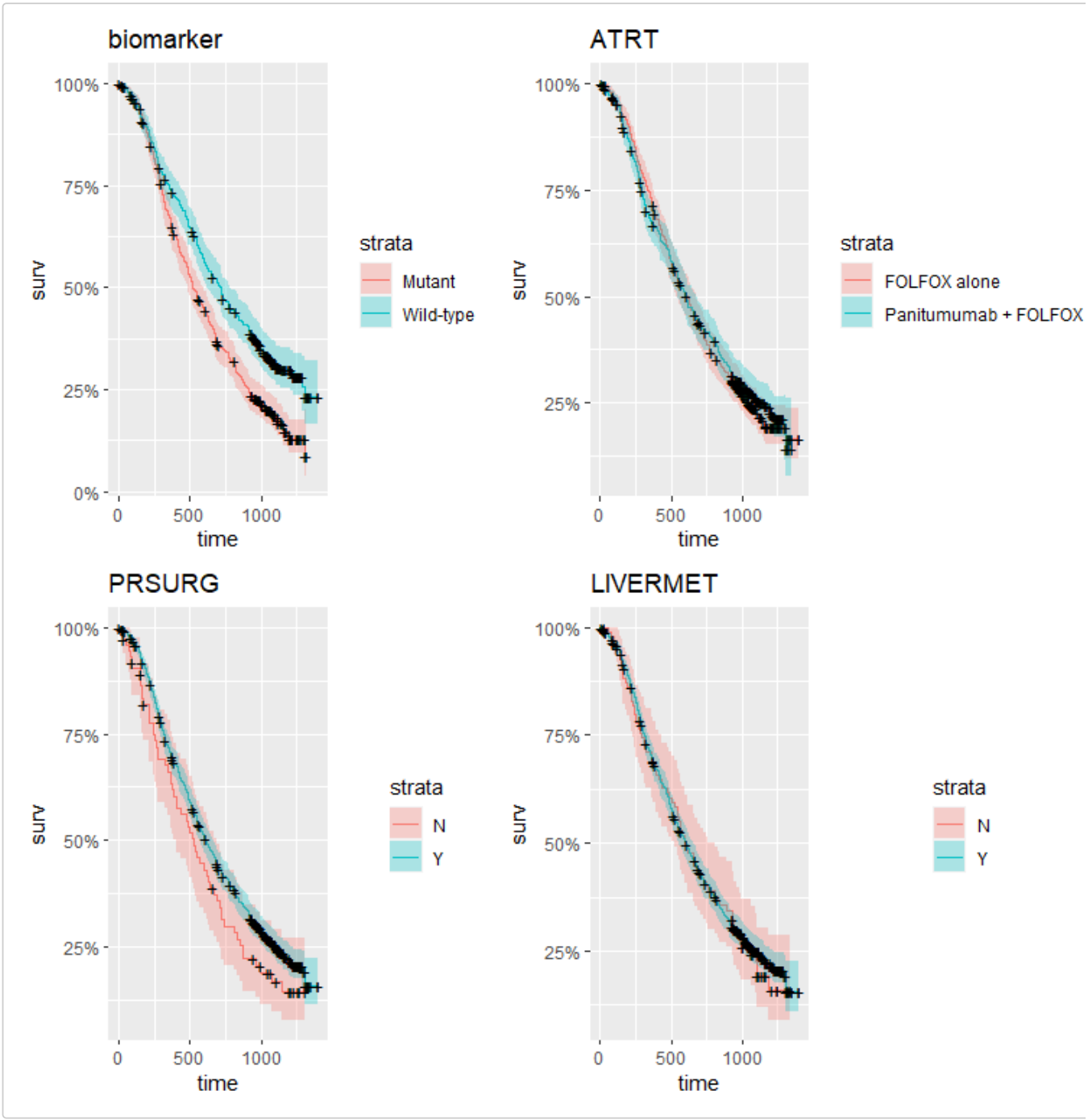
At the significance level of $\alpha = 0.05$, the p-value of the coefficient of variable `biomarker`,`PRSURGY`,`DIAGTYPE` is largely smaller than α, we can say that these variables have significant influence on the survival days of patients.

We can also observe that the p-value of coefficient of variable `AGE`, `B_WEIGHT`, `B_HEIGHT` are close to α, we think they might also have influence on the survival days, but the influence is less significant that the variables we mentioned before. The p-value of `ATRT` is not significant.

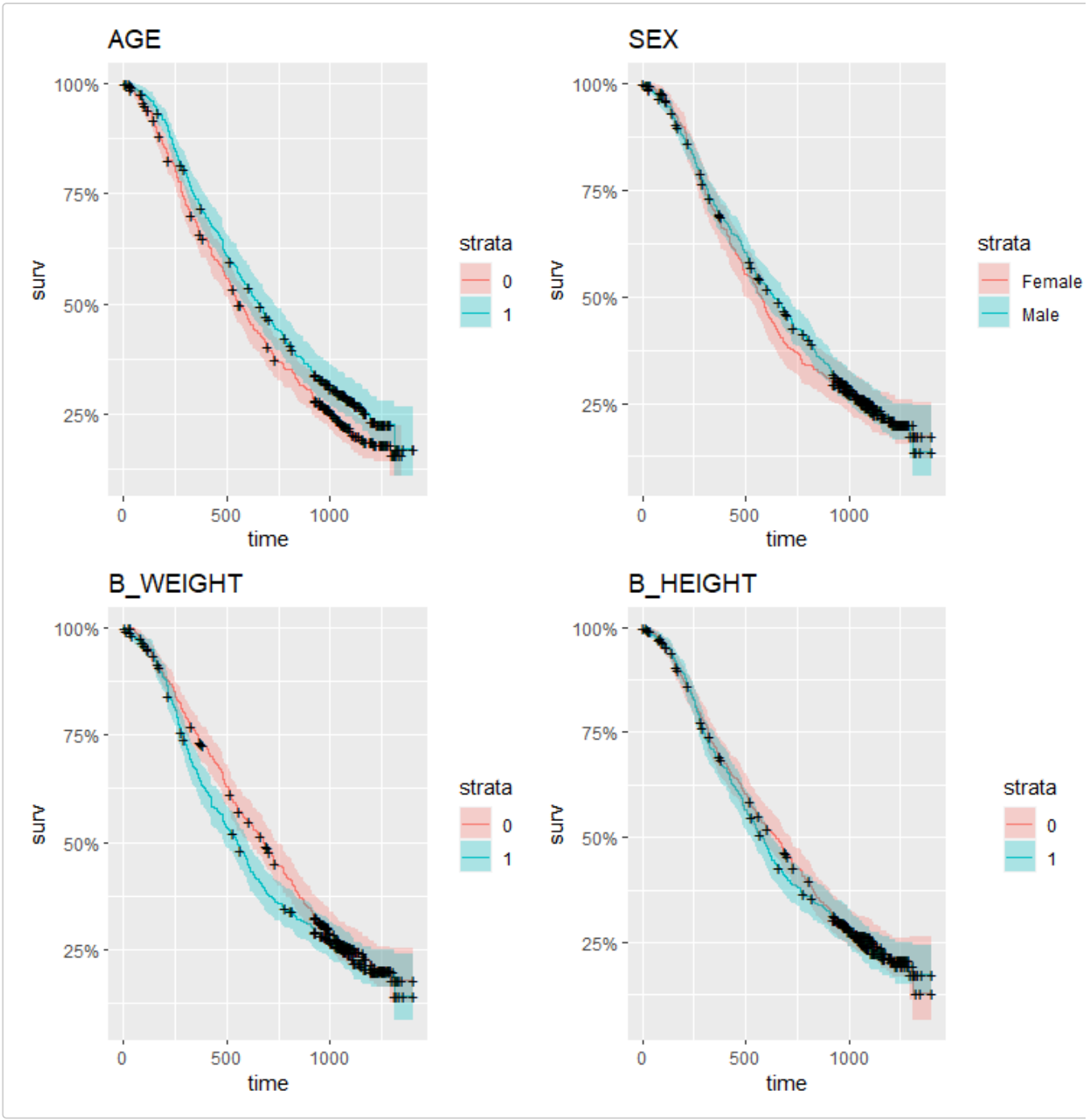
In addition to analyzing the statistical significance of each coefficient in the model, we also draw Kaplan-Meier curve over each variable as well as conduct Log-rank tests to compare curves with different labels. This process can be realized by the function we built, `plotkm_lrt()`.

As the value of variables`AGE`,`B_HEIGHT`,`B_WEIGHT` are consecutive numbers, we simply split the sample into two groups according to the median of each variable.

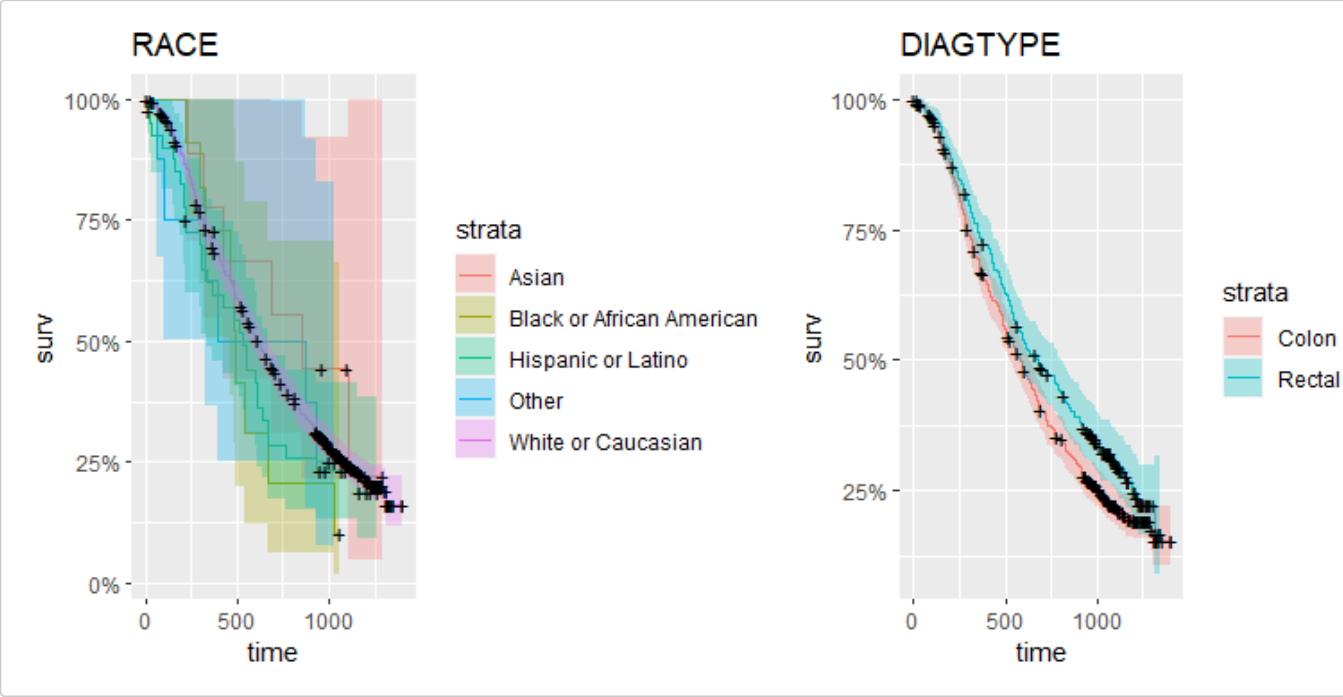
```
db2 <- db

db2$AGE <- as.numeric(db2$AGE < median(db2$AGE))
db2$B_HEIGHT <- as.numeric(db2$B_HEIGHT < median(db2$B_HEIGHT, na.rm = TRUE))
db2$B_WEIGHT <- as.numeric(db2$B_WEIGHT < median(db2$B_WEIGHT, na.rm = TRUE))


library(ggfortify)
#> Warning: 程辑包'ggfortify'是用R版本4.2.2 来建造的
library(broom)
plotkm_lrt <- function(db, datcol) {
  plot <- autoplot(survfit(Surv(db$DTHDY, db$DTH) ~ datcol),
                   title = names(datcol))
  lrt_p <- broom::glance(survdiff(Surv(db$DTHDY, db$DTH) ~ datcol))$p.value
  return(list(plot, lrt_p))
}


library(patchwork)

res <- sapply(db2[, c(2:4, 7:13)], plotkm_lrt, db = db2)
res[[1]] + ggtitle("biomarker") + res[[3]] + ggtitle("ATRT") +
  res[[5]] + ggtitle("PRSURG") + res[[7]] + ggtitle("LIVERMET") +
  plot_layout(ncol = 2)
```




```
res[[9]] + ggtitle("AGE") +
  res[[11]] + ggtitle("SEX") +
  res[[13]] + ggtitle("B_WEIGHT") +
  res[[15]] + ggtitle("B_HEIGHT") + plot_layout(ncol = 2)
```




```
res[[17]] + ggtitle("RACE") + res[[19]] + ggtitle("DIAGTYPE")
```



By analyzing the plot, the differences between curves with different labels in plots of `biomarker`, `PRSURG`, `DIAGTYPE` is visible, which provides the same conclusion as the p-value, while that of `AGE` and `B_WEIGHT` is not that obvious.

The following is the p-value of the Log-Rank Tests carried out on each variables. The null hypothesis is: There's no difference between the survival curve between the two/multiple groups.

	p-value
biomarker	0.0000013
ATRT	0.7374731
PRSURG	0.0651667
LIVERMET	0.7421902
AGE	0.0205762
SEX	0.4767210
B_WEIGHT	0.0736488
B_HEIGHT	0.5142676
RACE	0.5797985
DIAGTYPE	0.0143764

From this form, we know that the p-value of `biomarker`, `AGE`, `DIAGTYPE` is less than $\alpha = 0.05$, while the p-value of `PRSURG` is slightly greater than α, which partially confirmed our previous conclusion. However, the p-value of `ATRT` is still not significant.

In conclusion, we have made a preliminary analysis on the dataset, and found out that variables like `biomarker`, `PRSURG`, `DIAGTYPE`, `AGE` might have influence on the survival time of patients, but what we are interested, the treatment type `ATRT` seems not significant in this preliminary analysis.

As the significance of `biomarker` is the greatest, we will then split the dataset according to patients' `biomarker`, and make further analysis on each type, and find out whether `ATRT` will be influential when controlling for biomarker type.

4.2 Models controlling for the different Mutant types

Based on the above results, we discovered that there is a significant differences for the survival time of patient in different `AGE` groups in the Kaplan-Meier curve, and the p-value for variable `AGE` classified into groups "lower than Median age 62" and "higher than the Median age 62" in the Log-Rank Tests is less than $\alpha = 0.05$, we will classify the patients into two age group based on the Median value.

The variables B_HEIGHT and B_WEIGHT are also being classified into two groups, but they do not show a great difference in the survival plots and are not significant in the Log-Rank Tests, so we determine not to classify the patients' B_HEIGHT and B_WEIGHT into two groups in our next model.

According to the result of Model1, we deleted variableLIVERMET,SEX,B_WEIGHT,RACE from our further analysis, as they are both insignificant in both Cox model and Log rank test.

4.2.1 Models for Mutant Type

In our model2, we included all the rest variables and build a model for the patients with biomarker=="Mutant".

```
db3 <- db
db3$AGE <- as.numeric(db3$AGE < median(db3$AGE))
db3 <- db3 |> filter(biomarker == "Mutant")
model2 <- coxph(Surv(db3$DTHDY, db3$DTH) ~  ATRT +
                  PRSURG + AGE + B_HEIGHT + DIAGTYPE, data = db3)
model2
#> Call:
#> coxph(formula = Surv(db3$DTHDY, db3$DTH) ~ ATRT + PRSURG + AGE +
#>     B_HEIGHT + DIAGTYPE, data = db3)
#>
#>                               coef exp(coef)  se(coef)      z        p
#> ATRTPanitumumab + FOLFOX  0.198629  1.219729  0.106672  1.862 0.062595
#> PRSURGY                  -0.754152  0.470409  0.195874 -3.850 0.000118
#> AGE                      -0.153133  0.858016  0.107108 -1.430 0.152801
#> B_HEIGHT                  0.007313  1.007340  0.005703  1.282 0.199728
#> DIAGTYPERectal           -0.224298  0.799077  0.121204 -1.851 0.064229
#>
#> Likelihood ratio test=21.16  on 5 df, p=0.0007567
#> n= 447, number of events= 356
```

We can found that only the p-value for the variable PRSURGY is smaller than the significance level $\alpha = 0.05$. The p-value for left variables ATRT AGE B_HEIGHT and DIAGTYPE are all quiet large and not significant, which means they have no impacts to patients' survival time.

To exclude the insignificant variables in the model2, we build a reduced model (model3) with the only significant variable PRSURG.

```
model3 <- coxph(Surv(db3$DTHDY, db3$DTH) ~ PRSURG, data = db3)
model3
#> Call:
#> coxph(formula = Surv(db3$DTHDY, db3$DTH) ~ PRSURG, data = db3)
#>
#>            coef exp(coef) se(coef)      z        p
#> PRSURGY -0.6711    0.5112   0.1916 -3.502 0.000461
#>
#> Likelihood ratio test=10.25  on 1 df, p=0.001366
#> n= 447, number of events= 356
```

The p-value for PRSURGY equals to 0.000461 is really small, which means the variable PRSURGY will have significant impacts on the survival time for the patients with 'Mutant' type biomarker. The variable ATRT we expected to be significant does not actually have impacts on the survival time for patients with "Mutant" type gene.

4.2.2 Models for Wild Type

We will then build a model(Model4) for patients with biomarker="Wild-Type".

```
db4 <- filter(db, db$biomarker == "Wild-type")
db4$AGE <- as.numeric(db4$AGE < median(db4$AGE))
```

Same as what we have done before, we first build a full model(Model4) with all the rest variables, and then find out whether treatment makes difference as well as which variables are most influential on the survival days,then we get the reduced model(Model5).

```
model4 <- coxph(Surv(db4$DTHDY, db4$DTH) ~ ATRT + PRSURG +
                  AGE + B_HEIGHT + DIAGTYPE, data = db4)

model4
#> Call:
#> coxph(formula = Surv(db4$DTHDY, db4$DTH) ~ ATRT + PRSURG + AGE +
#>     B_HEIGHT + DIAGTYPE, data = db4)
#>
#>                               coef exp(coef)  se(coef)      z      p
#> ATRTPanitumumab + FOLFOX -0.258109  0.772511  0.122221 -2.112 0.0347
#> PRSURGY                  -0.122448  0.884752  0.201897 -0.606 0.5442
#> AGE                      -0.219127  0.803220  0.122786 -1.785 0.0743
#> B_HEIGHT                 -0.004363  0.995647  0.006922 -0.630 0.5285
#> DIAGTYPERectal           -0.210785  0.809948  0.128373 -1.642 0.1006
#>
#> Likelihood ratio test=10.84  on 5 df, p=0.05463
#> n= 406, number of events= 272
#>    (因为不存在，1个观察量被删除了)
```

From model4, we know that only the variable `ATRT`is significant in the model, which means that for people with Wild-type mutant on their RAS gene, the treatment of Panitumumab with Folfox has influence on their survival times.

And we get the reduced model(Model5) based on previous findings.

```
model5 <- coxph(Surv(db4$DTHDY, db4$DTH) ~ ATRT, data = db4)

model5
#> Call:
#> coxph(formula = Surv(db4$DTHDY, db4$DTH) ~ ATRT, data = db4)
#>
#>                             coef exp(coef) se(coef)     z      p
#> ATRTPanitumumab + FOLFOX -0.2527    0.7767   0.1215 -2.08 0.0375
#>
#> Likelihood ratio test=4.34  on 1 df, p=0.03733
#> n= 407, number of events= 273
```

We can see that the p-value of coefficient significance for `ATRT` is less than $\alpha = 0.05$, the p-value of a Likelihood ratio test of Model5 is less than $\alpha = 0.05$, while that of Model4, which is 0.05463, is not significant.

As a result, for patients with Wild-type mutant, we think the type of treatment have the most significant influence on their survival time.

5. Conclusions

In conclusion, treatment of Chemotherapy alone or Panitumumab in combination with Chemotherapy have a different impact on patients' survival time when controlling for their mutant type.

As for patients with Mutant type, it is whether the patients received prior surgery or not that influence their survival time.

As for patients with Wild type, it is the treatment type that has the most significant influence on their survival time.

Comparing with the result on gov, our primary analysis gets the same result. The result on the gov shows that different treatments will have impacts on patients progression-free survival with Wild-type genes, and have no impacts on patients with Mutant type genes. This study further confirms the efficacy of the treatment panitumumab-FOLFOX in treating the patients with Wild-type KRAS and NRAS.

References

[1] “Colorectal Cancer - Statistics,” Cancer.Net, Jun. 25, 2012. https://www.cancer.net/cancer-types/colorectal-cancer/statistics (accessed Dec. 07, 2022).

[2] “Panitumumab Combined With FOLFOX Emerges As the New First-Line PARADIGM for Left-Sided *RAS* Wild-Type Metastatic Colorectal Cancer,” ASCO Daily News. https://dailynews.ascopubs.org/do/10.1200/ADN.22.201008/full (accessed Dec. 07, 2022).