Math 150 - Methods in Biostatistics - Final Project

Barah Makhdum

Due: Friday, May 3rd, 2019

Table of Content

Introduction

We will talk here a brief information about data and how they collected.The data collected in 1997. Patients are randomized to receive open-label AZT and 3TC with or without indinavir sulfate for at least 48 weeks. There are 1750 participants but the researchers chose patients were eligible for the trial if they had no more than 200 CD4 cells per cubic millimeter and at least three months of prior zidovudine therapy. Randomization was stratified by CD4 cell count at the time of screening. The primary outcome measure was time to AIDS defining event or death. The primary purpose for this study is the treatment and how much it effect.
The patients age selected in this study started from 18 years and older. Also, it included both genders and all the Race/Ethnicity.

Statistical description of data

In this study, we have a sample is equal 851 and 16 variables serving the study axis. Where the scale used in the time variable is by number of day. That mean is time to death. From AIDS data we can see the range of this variable is between one day to 362 days and the average data is 243 days.
The variable censor_d is a Event indicator for death and it is a nominal variable (categorical variable). Thant mean the researchers made two choices, patients should choose one. (1) for death and (0) for Otherwise.

Notice that, there is a few people in the study dying.
the tx variable is the important variable because as I mentioned in the previous part is the primary purpose of this study. As we see in the data, this variable is also categorical variable define as a two groups. First group took 422 people who take the treatment includes IDV and the second group is 429 for who follow treatment regimen without IDV.
The strat2 is a CD4 stratum at screening and it was divided into two groups. (0) was for whom had CD4 less than or equal to 50 and (1) gaven to whom had grater than 50 at CD4 screening

From the pie chart (graph (3-1)) we can take a quick look at the percent of sex. Male took the highest percent which is 84% but the rest for Female. Also, (graph (4-1)) shows the highest and lowest percentage of the Race we have in the AIDS data.White (Non-Hispanic) is highest percentage. Hispanic and Black percentage are close to each other. However, Asian, took the lowest percentage.

what is ivdrug variable mean? Ivdrug was the question about the IV drug use history and the patients should choose one of these (never, currently and previously). graph (5-1) summarise the information in the graph by showing the number of said No they do not have Hemophiliac greater than to whom said yes they have it.

Power Analysis

(1) What is Power Analysis?

The power of any test of statistical significance is defined as the probability that it will reject a false null hypothesis.

	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision	Incorrect Decision Type I error (α)
H_0 is false	Incorrect Decision Type II error (β)	Correct Decision

α is the Probability when we reject H_0 when is true

β is the probability when we do not reject H_0 when is false

power is inversely related to beta or the probability of making a Type II error: $1-\beta$

75% power means you have an 75% chance of getting a significant result when the effect is real.

(2)How to Calculate the power?

To calculare the power there is a common formula we can used it.

$$1-\beta = 2\Phi(z - z_{1-\alpha}) - 1$$

$$z = (\delta - |ln(\theta)|\sqrt{(}nP_AP_BP_E)$$

,

$$n = \frac{1}{P_AP_BP_E}(\frac{z_{1-\alpha} + z_{1-\frac{\beta}{2}}}{\delta - |ln(\theta)|})^2$$

$1-\beta$ is our measure of power. $0 < \beta < 1$

Φ is the standard Normal distribution function.

δ is the testing margin.

θ is the hazard ration

$ln(\theta)$ is the natural logarithm of the hazard ratio, or the log-hazard ratio

n is sample size.

P_E is the overall probability of the event occurring within the study period

P_A and P_B are the proportions of the sample size allotted to the two groups, named 'A' and 'B'. Notice that $P_B = 1 - P_A$.

The assumption of power analysis is that sample random.

A simple random sample is a subset of a statistical population in which each member of the subset has an equal probability of being chosen. A simple random sample is meant to be an unbiased representation of a group.A simple random sample is a subset of a population in which each member of the subset has an equal probability of being chosen. A simple random sample is meant to be an unbiased representation of a group.

(3)What are ingredients of Statistical Power? There are three ingredients of that.

First,strength of the treatment. There is positive relationship between the power and strength of the treatment. That mean, when the strength of your treatment increases, the power of your experiment increases.

Second, background noise. There is opposite relationship between the power and background noise. That mean, when the background noise of your outcome variables increases, the power of your experiment decreases.

Third, experimental Design. Traditional power analysis focuses on one element of experimental design: the number of subjects in each experimental group.

The three ingredients of power connect to the survival analysis model by. Strength of the treatment, how much does the treatment effect how people behave. Backround noise, explaining by example.If we test the drugs there are many diseases kill people then if hard to see when drugs effect. For experimental design, how well the experimental setup or determined and how well good result.

(4)Power analysis for survival analysis

Survival analysis: The survival probability is the probability that an individual survives from the time origin to a specified future time t and is denoted by $S(t)$. Also, called the survivor function.

Hazard Function: Hazard Function is another idea in survival analysis. Also, called instantaneous death rate. It is usually denoted by $h(t)$ ot $\lambda(t)$ and is the probability that an individual who is under observation at a time t has an event at that time. In another word, it represents the instantaneous event rate for an individual who has already survived to time t.

In survival analysis, the power is directly related to the number of events observed in the study. The required sample size is therefore determined by the observed number of events. Survival data are commonly analyzed using the log-rank test or the Cox proportional hazards model.

Time to death is the event of interest

Null Hypothesis to be Tested

$$H_0 : HR = 1$$

where $HR = \frac{h_0(t)}{h_1(t)}$ for all t assuming proportional hazards

Alternative hypothesis

$$H_0 : HR \neq 1$$

Test Statistic

HR estimated from Cox model

Effect Size

HR = 1 implies no difference between treatments

HR > 1 implies "survival" is longer on treatment 2

HR < 1 implies "survival" is longer on treatment 1

Significance Level

$\alpha = 0.05$, $z_{\frac{\alpha}{2}} = 1.96$

Power

Typically desire power of at least 80%,90% or 95%. Recall that for means and proportions, power is a function of sample size.However, for survival data, power is entirely driven by number of events

Power	β	z_β
80%	0.20	0.842
90%	0.10	1.282
95%	0.05	1.645

Required Number of Events

$$events = \frac{(z_{\frac{\alpha}{2}} + z_\beta)^2}{\pi_1 \pi_2 (logHR)^2}$$

where $z_{\frac{\alpha}{2}}$ and z_β are dtandard normal percentiles, $\pi_1 and \pi_2$ are the proportion to be allocated to grpups 1 and 2

(for equal allocation π_1 and π_2=1/2)

And the $log(HR \neq 0)$ becouse in the power $HR \neq 1$. Thant mean, the hazard for the first group is not equal the hazard of the second group.
Probability of an Event

$$p(event) = 1 - (\pi_1 s_1(T) + \pi_2 s_2(T)$$

where $S_1(t)$ and $S_2(t)$ are Survival function of groups 1 and 2

We do the plot to estimate hazard rates for the dying data.

All these graphs for the hazard function and all of them gave same idea. Which is the hazard of dying is increasing by time. For the last graph we notice that, there are three lines. The two (-) lines are the confidence interval for the hazard function.

Find the best model for AIDS data

There are many way to find the best model for the data. We can find the best model by using cox and adding all explanatory variables. Then we see which variables are significant and which aren't.

Full Model

```
library(readr)
AIDSdata <- read_csv("~/Math-150/AIDSdata.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   time = col_double(),
##   censor = col_double(),
##   time_d = col_double(),
##   censor_d = col_double(),
##   tx = col_double(),
##   txgrp = col_double(),
##   strat2 = col_double(),
##   sex = col_double(),
##   raceth = col_double(),
##   ivdrug = col_double(),
##   hemophil = col_double(),
##   karnof = col_double(),
##   cd4 = col_double(),
##   priorzdv = col_double(),
##   age = col_double()
## )
```

```
library(coxed)
```

```
## Loading required package: rms
## Loading required package: Hmisc
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
## Loading required package: SparseM
##
## Attaching package: 'SparseM'
## The following object is masked from 'package:base':
##
##     backsolve
## Loading required package: mgcv
## Loading required package: nlme
## This is mgcv 1.8-28. For overview type 'help("mgcv-package")'.
```

```
cop1<-coxph(Surv(time_d,censor_d==1)~cd4+tx+strat2+sex+raceth+ivdrug+hemophil+karnof+priorzdv+age,data=A
cop1
```

```
## Call:
## coxph(formula = Surv(time_d, censor_d == 1) ~ cd4 + tx + strat2 +
##     sex + raceth + ivdrug + hemophil + karnof + priorzdv + age,
##     data = AIDSdata)
##
##               coef exp(coef)  se(coef)      z        p
## cd4      -0.011255  0.988809  0.009008 -1.249 0.211498
## tx       -0.835826  0.433516  0.495609 -1.686 0.091707
## strat2   -0.566918  0.567271  0.803798 -0.705 0.480624
## sex       0.659749  1.934307  0.567115  1.163 0.244690
## raceth    0.197628  1.218509  0.238724  0.828 0.407755
## ivdrug   -0.039250  0.961510  0.297013 -0.132 0.894865
## hemophil  0.987971  2.685780  1.089686  0.907 0.364588
## karnof   -0.077120  0.925779  0.027610 -2.793 0.005220
## priorzdv -0.005682  0.994335  0.009919 -0.573 0.566774
## age       0.079219  1.082442  0.023888  3.316 0.000912
##
## Likelihood ratio test=39.14  on 10 df, p=2.4e-05
## n= 851, number of events= 20
```

For the first output, it seems the variabile (ivdrug) is not significant. We will do a likelihood ratio test to confirm after we take out this variable. Before doing that step, let's do the model without variable (ivdrug) and see the reult with compare by likelihood.

Since the Chi-square test is not significant with one degree of freedom, we do not reject the null hypothesis. Therefore, we feel comfortable removing (ivdrug) variable from the model. Now, We continue testing nested models:Because the variable (priorzdv) has the biggest p-value and is not significant. We will take out and see what will be the model?

What we did for removing the variable (priorzdv) confirm becouse chi-square test is not significant. For next step, we keep removing the highest p-value until we get the best model. And here, we can see from the above output the variable (Strat2) will remove it. Every time we do chi-square to confirm that what variable take out the is right decision.

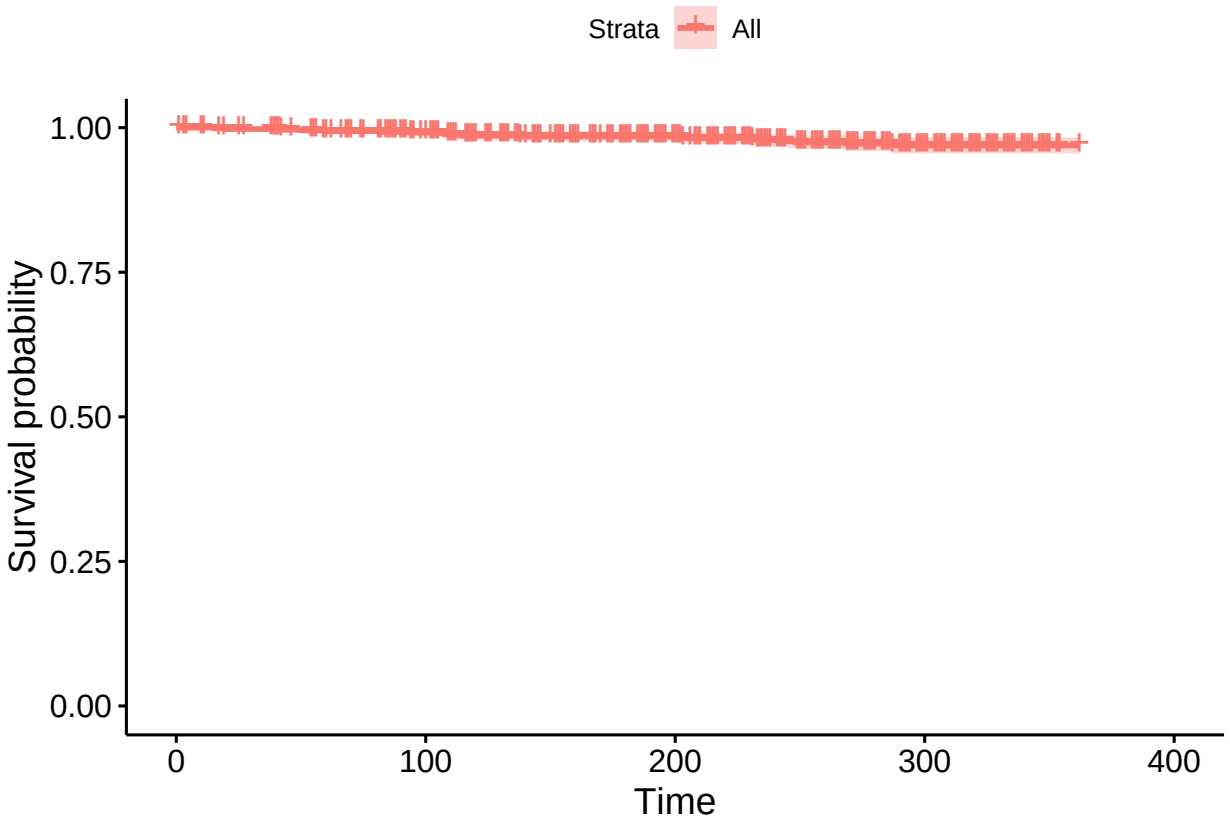
We Notice, in one of that steps we did all the variables in the model have p-value < 0.05. That mean, all of them are significant except the variable tx (treatment) is not significant. However, we can not able to remove it because this is the variable we interesting for the study.

```
library(survival)
library(survminer)
```

```
## Loading required package: ggpubr
```

```
## Loading required package: magrittr
```

```
ggsurvplot(survfit(Surv(time_d, censor_d)~1, data=AIDSdata)) #graph (9-1)
```



The figuresays that the majority of people is not dead. And that make sense because if we look at the data to see how many people die is 20 out of 851.

simulation

Simulation means, created new data under condition we put. For example, what is the Number of observations we want?, how much the latest time point during which an observation may fail?, what is your beta?, what is your mean and standard deviations, and other things.

Here, we simulated data to compare it with AISD data under the particular hypothese of $H_0 : \beta = 0$. $H_0 : h_1(t) = h_2(t)$ By using same sample size in AISD data (N=851) and use all the value of β we got from the best model. Since the censor in the AISD data is not equal 0.1. we need first to know how much by using the code below please see (5-1)

Note that 831/851=0.9764982. Now we have censor=0.976, it shows that the power will be small. Also, we need to find the mean and standard deviations for all 4 variables we selected. please see (6-1) and (7-1)

Now, we have all these information we can able to do the simulation.

```
library(simsurv)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:nlme':
##
##     collapse

## The following objects are masked from 'package:Hmisc':
##
```

```
##     src, summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(broom)

set.seed(1234)
n.reps<-100
a<-c()
for(i in 1:n.reps){
  simdata<-sim.survdata(N=851,num.data.frames=1,xvars=4,beta=c(-0.016659,-0.867409,-0.071620,0.073674),
model<-coxph(Surv(y,failed)~X1+X2+X3+X4,data=simdata$data)
a<-rbind(a,cbind(rep=rep(i,4),model %>% tidy()))
}
```

Now we have three samples size

$H_0 : \beta_1 = 0$

$H_0 : \beta_2 = 0$

$H_0 : \beta_3 = 0$

All the null hypothese should be fales and all the alternative hypothesis is true
$H_a : \beta_1 \neq 0$

$H_a : \beta_2 \neq 0$

$H_a : \beta_3 \neq 0$

The proportion of the first sample size(cd4 variable) is 0.03, the power will be very small here. The proportion of (tx variable) is 0.49 and the variables (karnof + age) have small proportion.

The varaiable (tx) treatment has bigger power. That mean, this variable has bigger chance to correcte the reject H_0. In other word, the variable treatment has bigger effect.

References:

http://egap.org/methods-guides/10-things-you-need-know-about-statistical-power https://www.investopedia.com/terms/s/simple-random-sample.asp Yulia Marchenko, 2007. "Power analysis and sample-size determination in survival models with the new stpower command

Appendix: R code programs

```
library(readr)
AIDSdata <- read_csv("~/Math-150/AIDSdata.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   time = col_double(),
##   censor = col_double(),
##   time_d = col_double(),
##   censor_d = col_double(),
##   tx = col_double(),
##   txgrp = col_double(),
##   strat2 = col_double(),
##   sex = col_double(),
##   raceth = col_double(),
##   ivdrug = col_double(),
##   hemophil = col_double(),
##   karnof = col_double(),
##   cd4 = col_double(),
##   priorzdv = col_double(),
##   age = col_double()
## )
```

```
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: purrr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## collapse():  dplyr, nlme
## filter():    dplyr, stats
## lag():       dplyr, stats
## src():       dplyr, Hmisc
## summarize(): dplyr, Hmisc
```

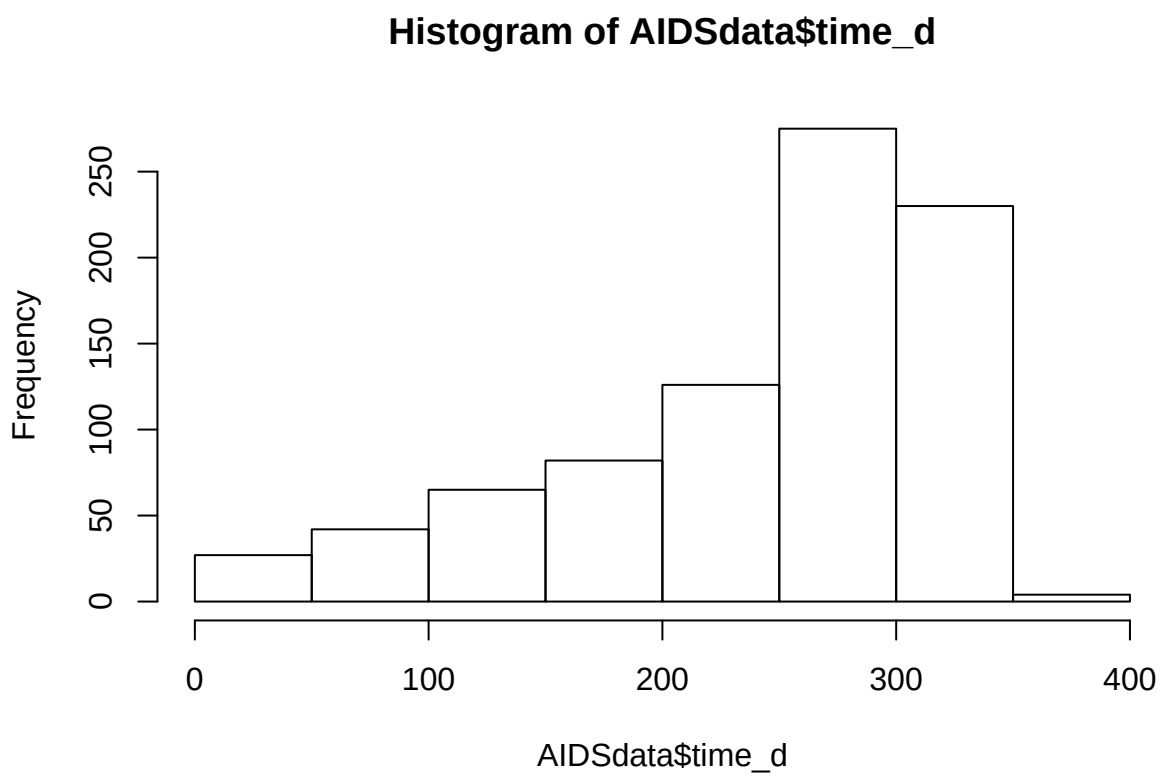
```
library(broom)
library(dplyr)
library(survival)
library(simsurv)
library(survminer)
library(FDRsampsize)
library(powerSurvEpi)
library(coxed)

dim(AIDSdata)  #(1-1)
```

```
## [1] 851  16
```

```r
summary(AIDSdata$time_d)  #(2-1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   199.5   266.0   243.4   306.0   362.0
```

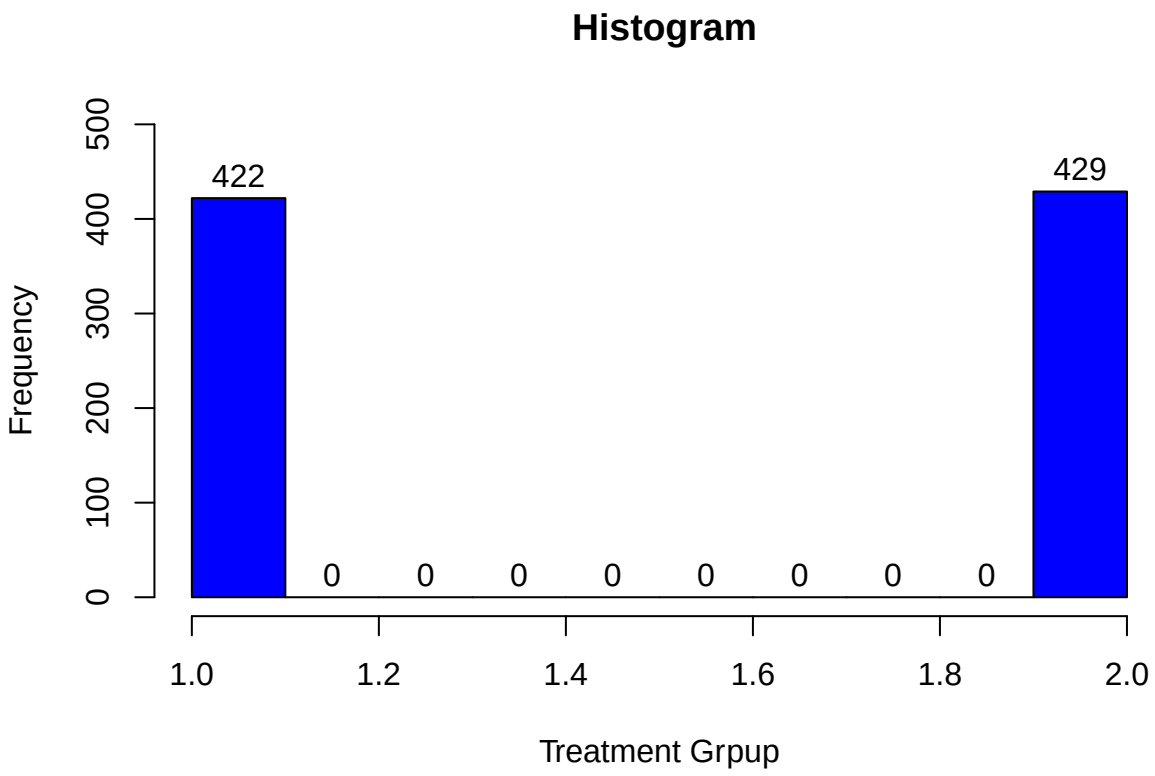
```r
hist(AIDSdata$time_d)   #graph (1-1)
```




```r
table(AIDSdata$censor_d)   #(3-1)
```

```
##
##   0   1
## 831  20
```

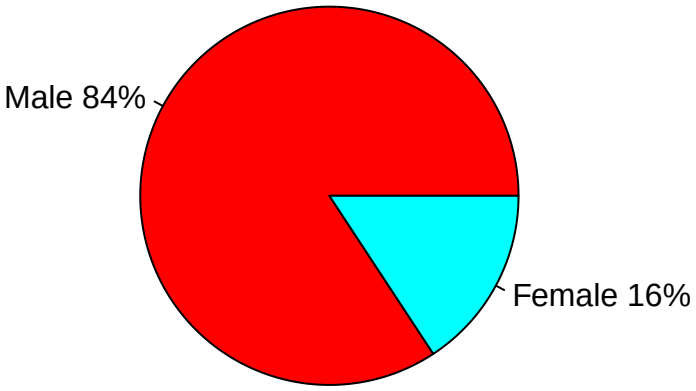
```r
hist(AIDSdata$txgrp, col="blue", xlab="Treatment Grpup", main=" Histogram",ylim =c(0,500), labels = TRU
```



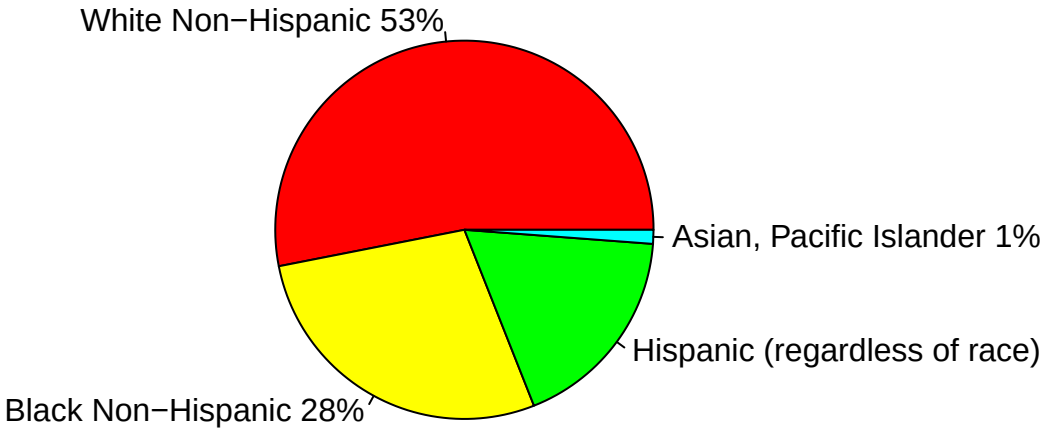

```
matrix(c("CD4<=50",sum(AIDSdata$strat2==0),"CD4>50",sum(AIDSdata$strat2==1)),ncol=2,byrow=TRUE)  #(4-1)
```

```
##      [,1]      [,2]
## [1,] "CD4<=50" "327"
## [2,] "CD4>50"  "524"
```

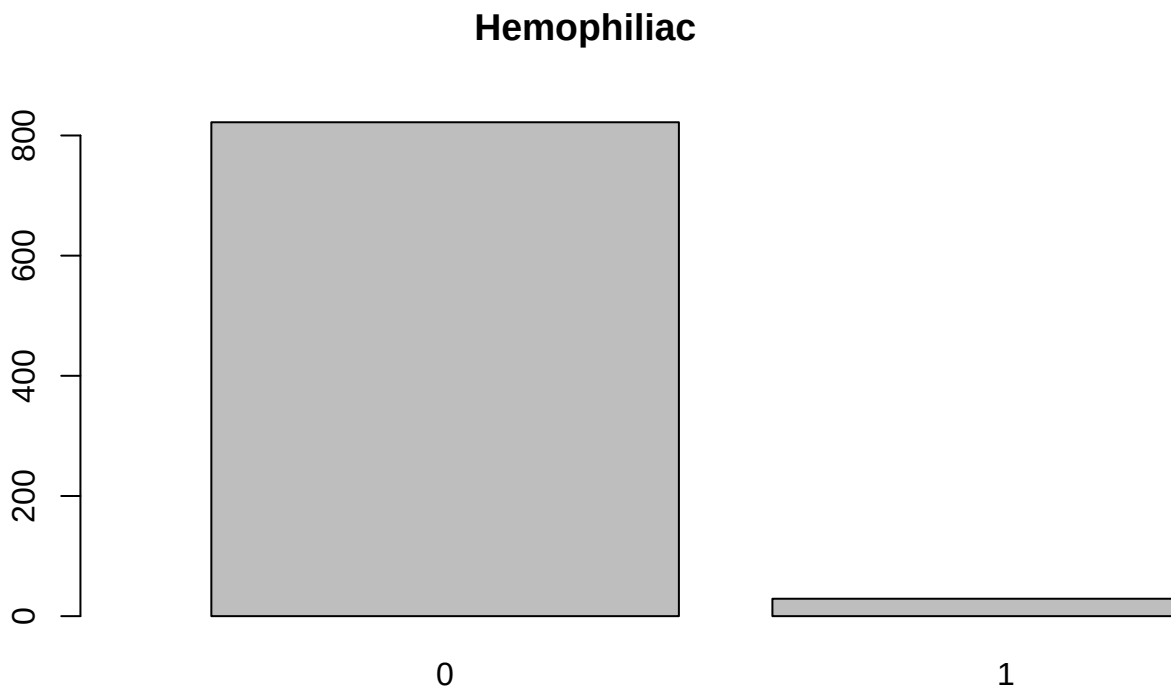
```
a<-sum(AIDSdata$sex==1)
b<-sum(AIDSdata$sex==2)
slices<-c(a,b)
labs<-c("Male","Female")
pct<-round(slices/sum(slices)*100)
lbls<-paste(labs,pct)
lbls<-paste(lbls,"%",sep="")
pie(slices,lbls,col=rainbow(length(lbls))) #graph (3-1)
```



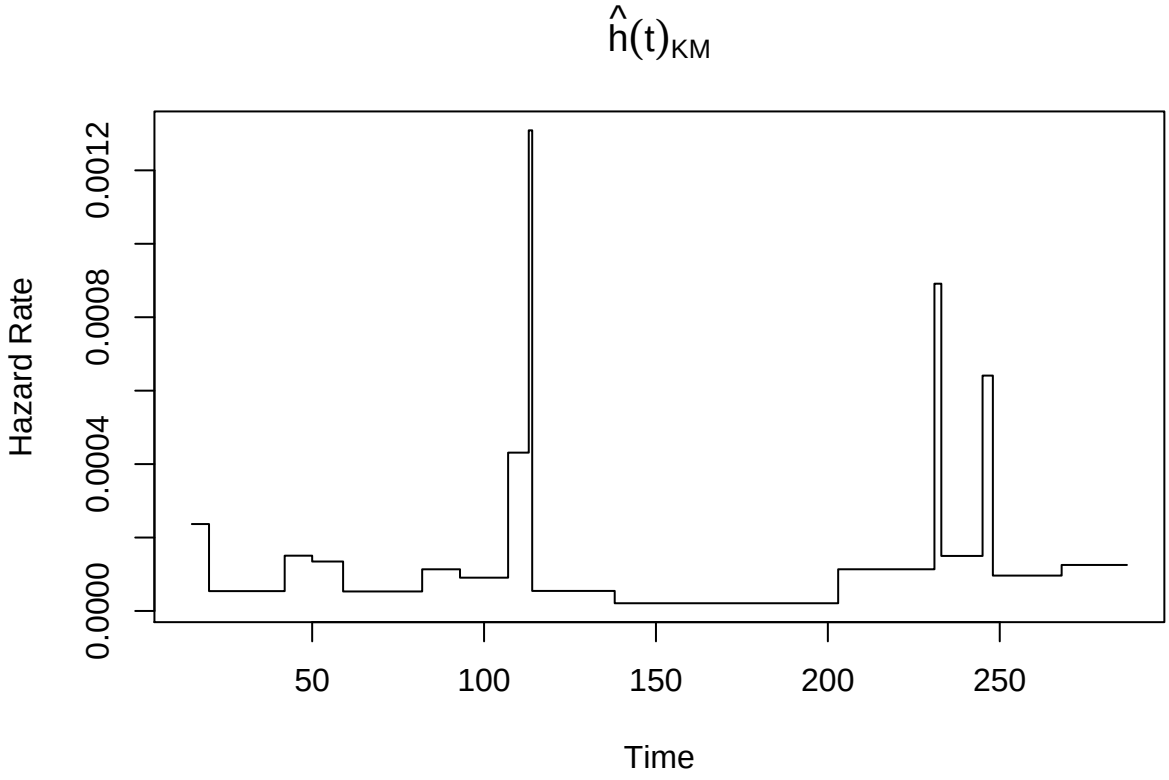

```r
a<-sum(AIDSdata$raceth==1)
b<-sum(AIDSdata$raceth==2)
c<-sum(AIDSdata$raceth==3)
d<-sum(AIDSdata$raceth==4)
slices<-c(a,b,c,d)
labs<-c("White Non-Hispanic","Black Non-Hispanic","Hispanic (regardless of race)","Asian, Pacific Islanc
pct<-round(slices/sum(slices)*100)
lbls<-paste(labs,pct)
lbls<-paste(lbls,"%",sep="")
pie(slices,lbls,col=rainbow(length(lbls)))  #graph(4-1)
```

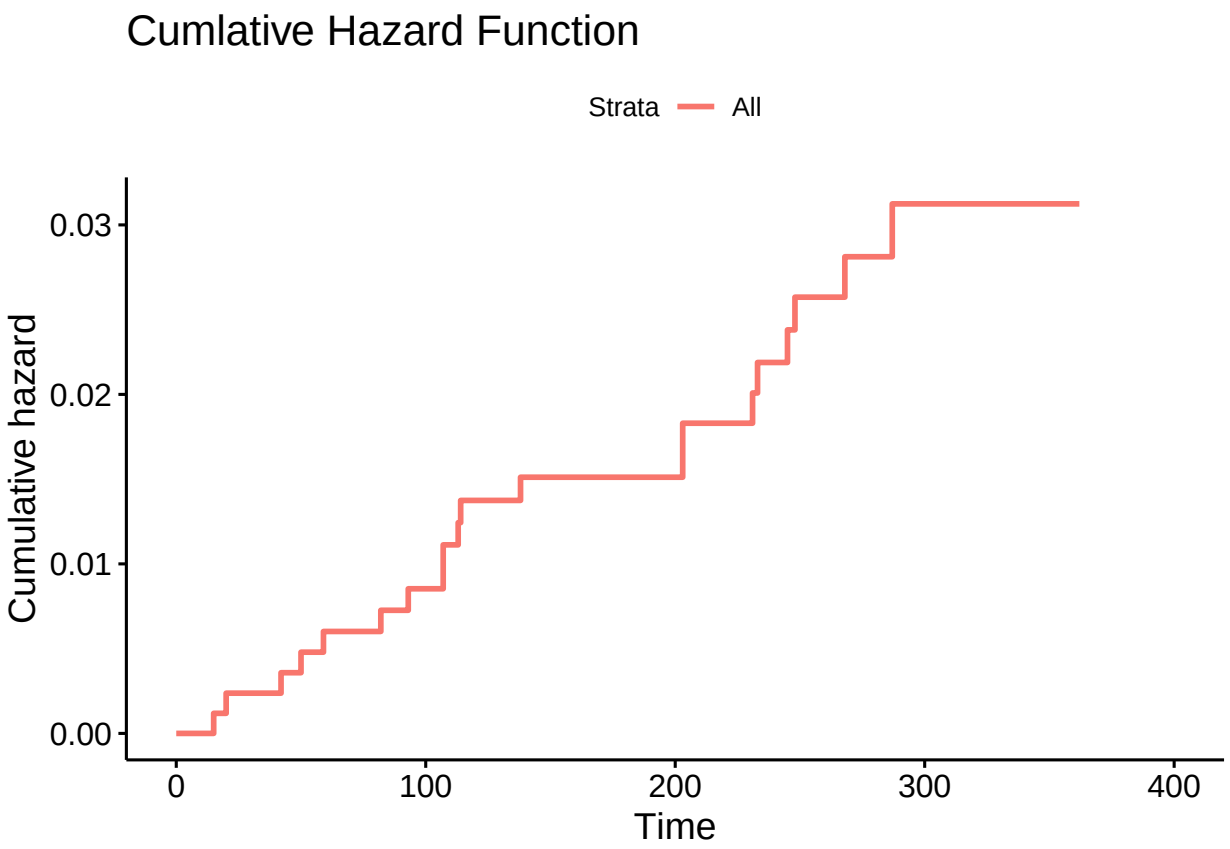



```
counts <- table(AIDSdata$hemophil)
barplot(counts, xlim=c(0,2), ylim = c(0,850),main="Hemophiliac") #graph (5-1)
```

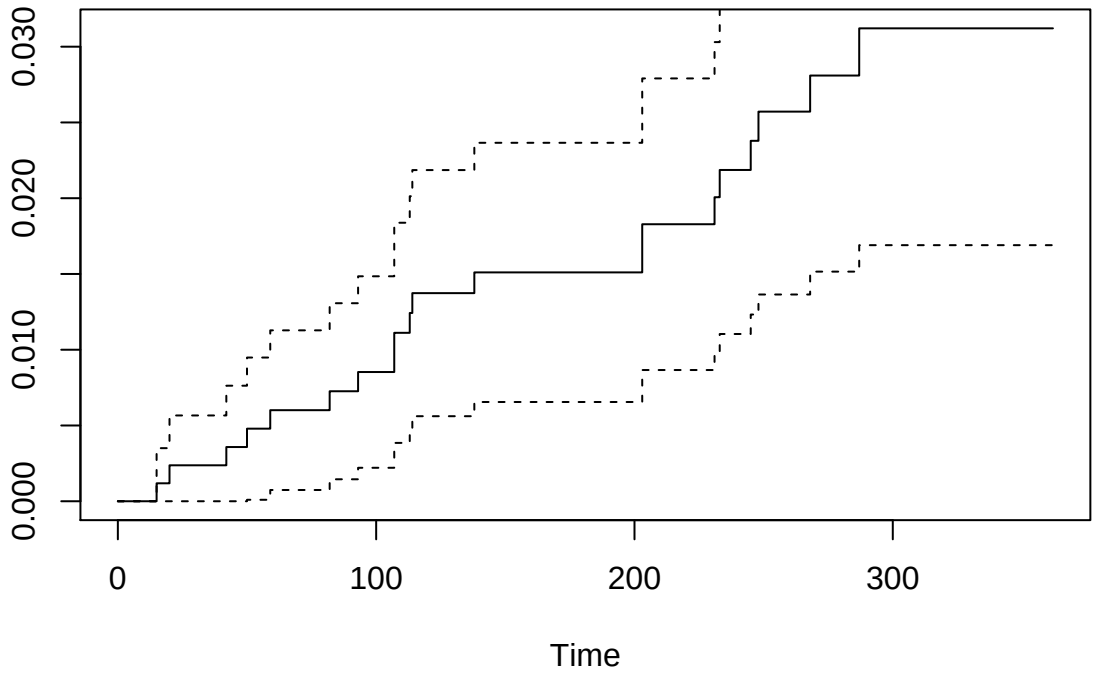



```
plot.haz <- function(KM.obj,plot="TRUE") {
ti <- summary(KM.obj)$time
di <- summary(KM.obj)$n.event
ni <- summary(KM.obj)$n.risk
#Est Hazard Function
est.haz <- 1:(length(ti))
for (i in 1:(length(ti)-1))
est.haz[i] <- di[i]/(ni[i]*(ti[i+1]-ti[i]))
est.haz[length(ti)] <- est.haz[length(ti)-1]
if (plot=="TRUE") {
plot(ti,est.haz,type="s",xlab="Time", ylab="Hazard Rate",
main=expression(paste(hat(h),(t)[KM])))
}
#return(list(est.haz=est.haz,time=ti))
}

KM.obj <- survfit(Surv(time_d,censor_d)~1,data=AIDSdata,conf.type="plain")
plot.haz(KM.obj)  #graph (6-1)

ggsurvplot(survfit(Surv(time_d,censor_d) ~ 1, data=AIDSdata),
censor=F, conf.int=F, fun="cumhaz") +
  ggtitle("Cumlative Hazard Function") #graph (7-1)
```


$\hat{h}(t)_{KM}$
Hazard Rate
0.0000
0.0004
0.0008
0.0012
50
100
150
200
250
Time




```
survfit <- survfit(Surv(time_d,censor_d) ~ 1, data=AIDSdata)
plot(survfit, fun="cumhaz", xlab="Time") #graph (8-1)
```




```
#Full Model
cop1<-coxph(Surv(time_d,censor_d==1)~cd4+tx+strat2+sex+raceth+ivdrug+hemophil+karnof+priorzdv+age,data=A
cop1
```

```
## Call:
## coxph(formula = Surv(time_d, censor_d == 1) ~ cd4 + tx + strat2 +
##     sex + raceth + ivdrug + hemophil + karnof + priorzdv + age,
##     data = AIDSdata)
##
##               coef exp(coef)  se(coef)      z        p
## cd4      -0.011255  0.988809  0.009008 -1.249 0.211498
## tx       -0.835826  0.433516  0.495609 -1.686 0.091707
## strat2   -0.566918  0.567271  0.803798 -0.705 0.480624
## sex       0.659749  1.934307  0.567115  1.163 0.244690
## raceth    0.197628  1.218509  0.238724  0.828 0.407755
## ivdrug   -0.039250  0.961510  0.297013 -0.132 0.894865
## hemophil  0.987971  2.685780  1.089686  0.907 0.364588
## karnof   -0.077120  0.925779  0.027610 -2.793 0.005220
## priorzdv -0.005682  0.994335  0.009919 -0.573 0.566774
## age       0.079219  1.082442  0.023888  3.316 0.000912
##
## Likelihood ratio test=39.14  on 10 df, p=2.4e-05
## n= 851, number of events= 20
```

```
cop2<-coxph(Surv(time_d,censor_d==1)~cd4+tx+strat2+sex+raceth+hemophil+karnof+priorzdv+age,data=AIDSdata
cop=2*(cop1$loglik[2]-cop2$loglik[2])
1-pchisq(cop,1)
```

```
## [1] 0.8941714
```

```
cop3<-coxph(Surv(time_d,censor_d==1)~cd4+tx+strat2+sex+raceth+hemophil+karnof+age,data=AIDSdata)
cop=2*(cop2$loglik[2]-cop3$loglik[2])
1-pchisq(cop,1)
```

```
## [1] 0.537843
```

```
cop4<-coxph(Surv(time_d,censor_d==1)~cd4+tx+sex+raceth+hemophil+karnof+age,data=AIDSdata)
copp=2*(cop3$loglik[2]-cop4$loglik[2])
1-pchisq(copp,1)
```

```
## [1] 0.4930001
```

```
cop5<-coxph(Surv(time_d,censor_d==1)~cd4+tx+sex+raceth+karnof+age,data=AIDSdata)
coppp=2*(cop4$loglik[2]-cop5$loglik[2])
coppp
```

```
## [1] 0.5026622
```

```
1-pchisq(coppp,1)
```

```
## [1] 0.4783327
```

```
cop6<-coxph(Surv(time_d,censor_d==1)~cd4+tx+sex+karnof+age,data=AIDSdata)
cop6
```

```
## Call:
## coxph(formula = Surv(time_d, censor_d == 1) ~ cd4 + tx + sex +
##     karnof + age, data = AIDSdata)
##
##             coef exp(coef)  se(coef)      z       p
## cd4    -0.016431  0.983703  0.006357 -2.585 0.00975
## tx     -0.869705  0.419075  0.490311 -1.774 0.07610
## sex     0.636651  1.890139  0.561470  1.134 0.25684
## karnof -0.072203  0.930342  0.026420 -2.733 0.00628
## age     0.075065  1.077954  0.023576  3.184 0.00145
##
## Likelihood ratio test=36.85  on 5 df, p=6.409e-07
## n= 851, number of events= 20
```

```
coppp1=2*(cop5$loglik[2]-cop6$loglik[2])
coppp1
```

```
## [1] 0.9167282
```

```
1-pchisq(coppp1,1)
```

```
## [1] 0.3383355
```

```
cop7<-coxph(Surv(time_d,censor_d==1)~cd4+tx+karnof+age,data=AIDSdata)
cop7
```

```
## Call:
## coxph(formula = Surv(time_d, censor_d == 1) ~ cd4 + tx + karnof +
##     age, data = AIDSdata)
##
##             coef exp(coef)  se(coef)      z       p
## cd4    -0.016659  0.983479  0.006408 -2.600 0.00933
## tx     -0.867409  0.420038  0.490243 -1.769 0.07684
```

```
## karnof -0.071620  0.930884  0.025831 -2.773 0.00556
## age     0.073674  1.076456  0.023613  3.120 0.00181
##
## Likelihood ratio test=35.72  on 4 df, p=3.297e-07
## n= 851, number of events= 20
```

```
# Example for Simulation
simdata <- sim.survdata(N=1000, T=100, num.data.frames=1,beta=c(0.01,0.05,0.08))
head(simdata$data,10)
```

```
##             X1          X2         X3  y failed
## 1  -0.03615590  1.29218194 -0.1665820 52   TRUE
## 2   0.17822389  0.53606372  0.5437421 73   TRUE
## 3   0.42264791  0.69437823 -0.5044286 30   TRUE
## 4   0.02722508  0.26220397 -0.4404290 51   TRUE
## 5  -0.64084552  1.00725079  0.6039851 78   TRUE
## 6   0.65930635 -0.71426069 -1.4900195 99   TRUE
## 7   0.19029150  0.71647479  1.9074511 78   TRUE
## 8   0.68731498  1.38020861 -0.6543776  8   TRUE
## 9   0.54860825 -0.88261818 -0.1686744 19   TRUE
## 10 -1.87836836  0.03027097 -1.2639493 99   TRUE
```

```
simdata$betas
```

```
##      [,1]
## [1,] 0.01
## [2,] 0.05
## [3,] 0.08
```

```
head(simdata$baseline,10)
```

```
##    time  failure.PDF  failure.CDF  survivor       hazard
## 1     1 5.670163e-06 5.670163e-06 0.9999943 5.670163e-06
## 2     2 3.969114e-05 4.536131e-05 0.9999546 3.969137e-05
## 3     3 1.077331e-04 1.530944e-04 0.9998469 1.077380e-04
## 4     4 2.097960e-04 3.628904e-04 0.9996371 2.098282e-04
## 5     5 3.458800e-04 7.087704e-04 0.9992912 3.460055e-04
## 6     6 5.159848e-04 1.224755e-03 0.9987752 5.163508e-04
## 7     7 7.201107e-04 1.944866e-03 0.9980551 7.209938e-04
## 8     8 9.582576e-04 2.903124e-03 0.9970969 9.601249e-04
## 9     9 1.230425e-03 4.133549e-03 0.9958665 1.234008e-03
## 10   10 1.536614e-03 5.670163e-03 0.9943298 1.542992e-03
```

```
library(dplyr)
library(broom)

table(AIDSdata$censor_d) # (5-1)
```

```
##
##   0   1
## 831  20
```

```
m<-c(mean(AIDSdata$cd4),mean(AIDSdata$tx),mean(AIDSdata$karnof),mean(AIDSdata$age)) # (6-1)
s<-c(sd(AIDSdata$cd4),sd(AIDSdata$tx),sd(AIDSdata$karnof),
     sd(AIDSdata$age)) #(7-1)

# My Simulation
set.seed(1234)
```

```
n.reps<-100
a<-c()
for(i in 1:n.reps){
  simdata<-sim.survdata(N=851,num.data.frames=1,xvars=4,beta=c(-0.016659,-0.867409,-0.071620,0.073674),
model<-coxph(Surv(y,failed)~X1+X2+X3+X4,data=simdata$data)
a<-rbind(a,cbind(rep=rep(i,4),model %>% tidy()))
}
str(simdata$data)
```

```
## 'data.frame':    851 obs. of  6 variables:
##  $ X1    : num  -0.5408 -0.0373 0.7524 -0.0815 -0.7385 ...
##  $ X2    : num  0.13507 -0.71183 -0.2093 -0.65512 -0.00651 ...
##  $ X3    : num  -2.7651 0.0358 0.4747 -1.5919 0.3266 ...
##  $ X4    : num  -0.815 0.43 0.167 -0.576 -0.211 ...
##  $ y     : int  67 71 77 85 67 72 85 82 84 85 ...
##  $ failed: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
x1pvalue<-a%>% dplyr::filter(term=="X1") %>%
  dplyr::summarize(sum(p.value<0.05))
x2pvalue<-a%>% dplyr::filter(term=="X2") %>%
  dplyr::summarize(sum(p.value<0.05))
x3pvalue<-a%>% dplyr::filter(term=="X3") %>%
  dplyr::summarize(sum(p.value<0.05))
x4pvalue<-a%>% dplyr::filter(term=="X4") %>%
  dplyr::summarize(sum(p.value<0.05))
c(x1pvalue,x2pvalue,x3pvalue,x4pvalue) # (8-1)
```

```
## $`sum(p.value < 0.05)`
## [1] 3
##
## $`sum(p.value < 0.05)`
## [1] 49
##
## $`sum(p.value < 0.05)`
## [1] 4
##
## $`sum(p.value < 0.05)`
## [1] 6
```