HW1

Joe Zhang

```
library(dplyr)
library(ggplot2)
library(mclust)
library(survival)
```

Problem 5

```
data <- read.csv("~/DataMining/CLV_CRM/cell.csv")
```

a)

It looks like the customers are split evenly across the months and that the customers each month are pretty close to splitting evenly between churn and no churn. According to the crosstable result, we should use stratified sampling method.

```
#data <- cell
table(data[,c('churn','billmonth')])
```

```
##      billmonth
## churn Apr 08 Aug 07 Dec 07 FEB 08 JAN 08 Jul 07 Jun 07 Mar 08 May 08
##     0    209    207    246    244    191    205    207    203    207
##     1    205    213    224    191    229    194    217    205    183
##      billmonth
## churn Nov 07 Oct 07 Sep 07
##     0    206    190    209
##     1    203    189    223
```

b)

2476 people canceled. The retention rate equals 0.02589308.

```
sum(data$churn)
```

```
## [1] 2476
```

```
r_hat = sum(data$churn,na.rm = TRUE)/sum(data$t2,na.rm = TRUE)
r_hat
```

```
## [1] 0.02589308
```

c)

We have to do stratetified sampling. As the retention rate is different across account types or line counts. And the distribution of the value in those two features are not uniform. If we just randomly drwa sample from the dataset, it may give a lot of bias.

d)

The average monthly revenue is 55.7633. E(CLV) = m(1+d)/(1+d-r) = 57.23051

```
mean.rm.na <- function(x){
return(mean(x,na.rm = TRUE))}
function()mean(x,na.rm = TRUE)
```

```
## function()mean(x,na.rm = TRUE)
```

```
tmp = apply(data[,7:21], 2, mean.rm.na)
m = mean(tmp)
d = 0.01
mean(tmp)
```

```
## [1] 55.7633
```

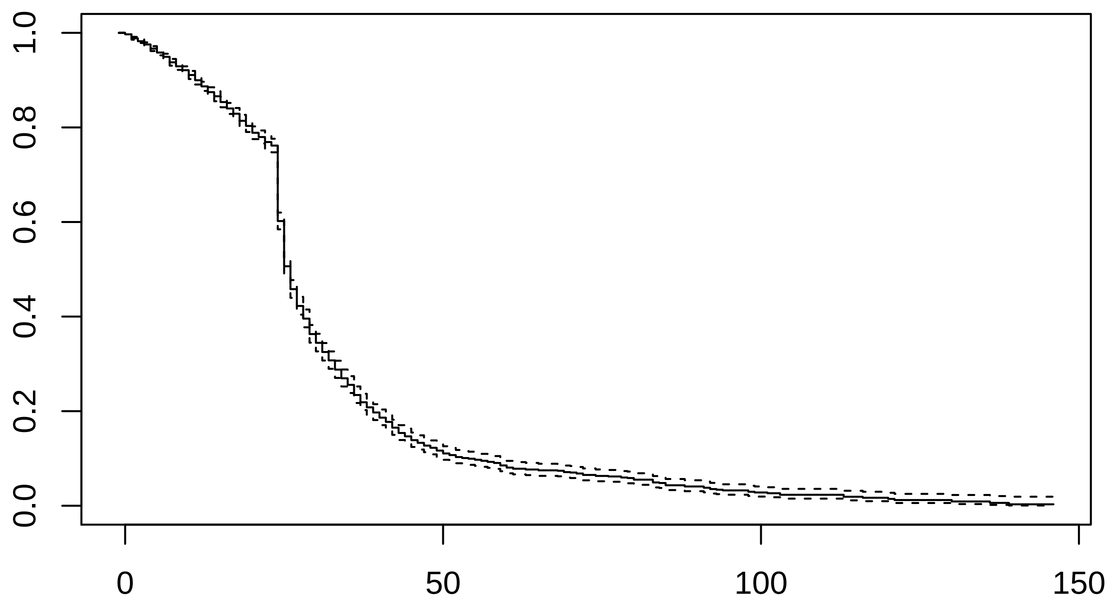
```
m*(1+d)/(1+d-r_hat)
```

```
## [1] 57.23051
```

e)

There is a significant drop at the 23-25 month mark which may indicate that there is a 2 year contract in place.

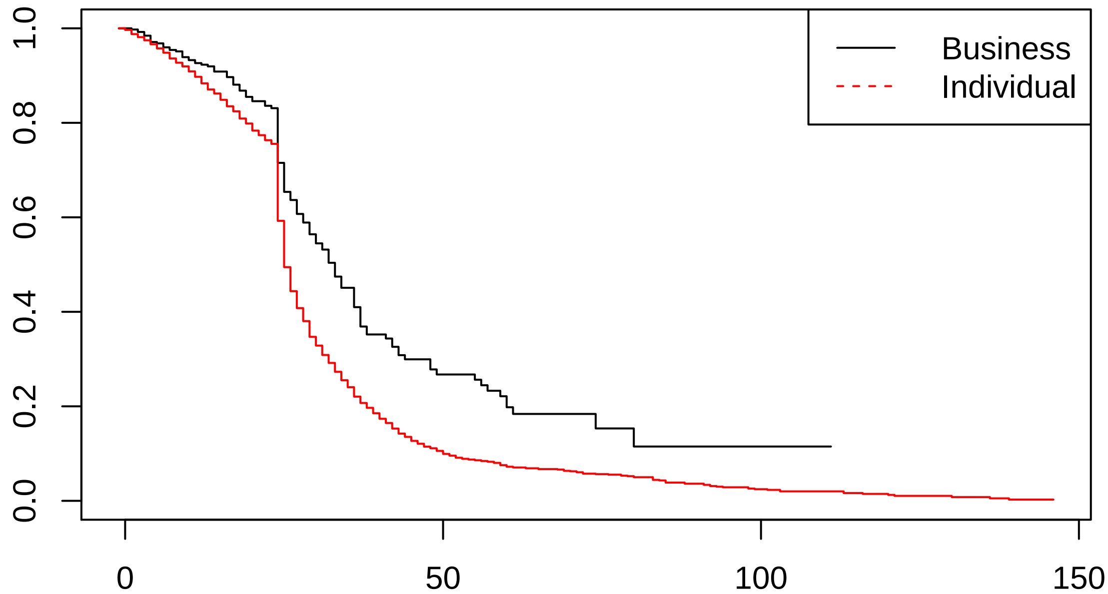
```
fit = survfit(Surv(t2, churn) ~ 1, data=data)
#summary(fit)
plot(fit)
```



f)

The individual is more likely to churn.

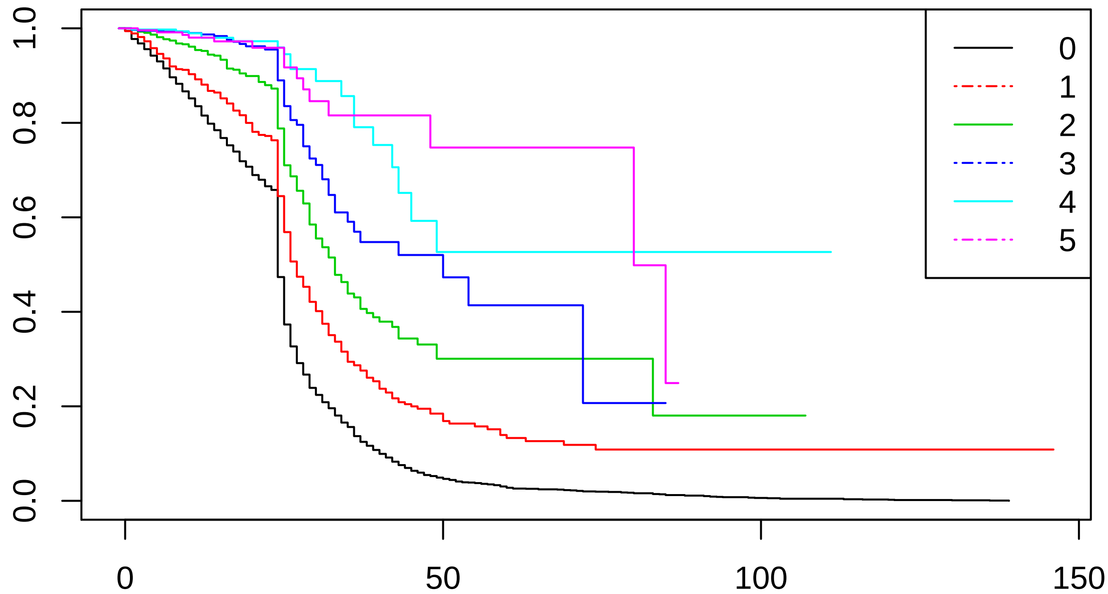
```
fit = survfit(Surv(t2, churn) ~ ACCOUNT_TYPE, data=data)
#summary(fit)
plot(fit, col=1:2)
legend("topright", paste(" ",c("Business","Individual")), col=1:2, lty=c(1,2))
```



g)

The fewer lines a customer/business may have, the more likely they will churn. Thos with more lines tend to churn less especially at the 24 month time.

```
data$LINE_COUNT2 = data$LINE_COUNT
data$LINE_COUNT2[data$LINE_COUNT2>5]=5
fit = survfit(Surv(t2, churn) ~ LINE_COUNT2, data=data)
#summary(fit)
#plot(fit)
plot(fit, col=1:6)
legend("topright", paste(" ",c(0,1,2,3,4,5)), col=1:6, lty=c(1,6))
```



0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150